Cleavage Factor I$_m$ as a key regulator of 3' UTR length

Andreas R. Gruber, Georges Martin, Walter Keller, Mihaela Zavolan

Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Basel, Switzerland.

Keywords: cleavage factor I$_m$, CF I$_m$, CPSF5, CPSF6, CPSF7, alternative polyadenylation, 3' end processing, mRNA processing

Abbreviations: CF I$_m$, cleavage factor I$_m$; CP, cleavage and polyadenylation; UTR, untranslated region; CPSF, cleavage and polyadenylation specificity factor; CstF, cleavage stimulation factor; RRM, RNA recognition motif; CLIP, cross-linking and immunoprecipitation

Abstract

In eukaryotes, the 3' ends of RNA polymerase II-generated transcripts are generated in the majority of cases by site-specific endonucleolytic cleavage, followed by the addition of a poly(A) tail. Through alternative polyadenylation, a gene can give rise to multiple mRNA isoforms that differ in the length of their 3' UTRs and hence in their susceptibility to post-transcriptional regulatory factors such as microRNAs. A series of recently conducted, high-throughput studies of poly(A) site usage revealed an extensive tissue-specific control and drastic changes in the length of mRNA 3' UTRs upon induction of proliferation in resting cells. To understand the dynamics of poly(A) site usage, we recently identified binding sites of the major pre-mRNA 3' end processing factors - cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), and cleavage factor I$_m$ (CF I$_m$) - and mapped polyadenylation sites in HEK293 cells. Our present study extends previous findings on the role of CF I$_m$ in alternative polyadenylation and reveals that subunits of the CF I$_m$ complex generally control 3' UTR length. More specifically, we demonstrate that the loss-of-function of CF I$_m$68 and CF I$_m$25 but not of CF I$_m$59 leads to a transcriptome-wide increase of the use of proximal polyadenylation sites.

Introduction

Generation of mature eukaryotic mRNAs from pre-mRNAs includes addition of a 7-methylguanosine cap, splicing out of introns and cleavage and polyadenylation of the 3' end [1, 2, 3, 4]. Most of these processes are carried out co-transcriptionally by a number of protein complexes and are completed before the transcription complex reaches the end of the gene. The process of cleavage and polyadenylation which is the focus of our work, involves a complex that contains up to 85 proteins [5]. At the core however, are a few smaller subcomplexes: the cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), cleavage factors I_m and II_m (CF I_m and CF II_m), a poly(A) polymerase (PAP) [4], and the nuclear poly(A) binding protein 1 (PABPN1) [6].

CF I_m is a tetramer composed of two 25 kDa (CF I_m25) subunits and two proteins of either 59 or 68 kDa (CF I_m59 or CF I_m68) [7, 8]. It was previously hown through SELEX analysis to preferentially bind UGUA subsequences in the pre-mRNAs [9]. The molecular basis of this interaction emerged from recently solved crystal structures of CF I_m25 in complex with the RNA recognition motif (RRM) of CF I_m68 [10, 11]. Surprisingly, it is the Nudix hydrolase domain of CF I_m25 that specifically recognizes UGUA, whereas CF I_m68 appears to increase the binding affinity of the complex. These structure models further revealed that a CF I_m25 dimer binds two UGUA sequences in an antiparallel manner forcing the looping of the RNA sequence between the UGUA motifs. Yang and colleagues proposed that looping might facilitate alternative polyadenylation via CF I_m [10]. The composition of individual CF I_m complexes that bind pre-mRNA molecules is not known and it is unclear whether CF I_m59 and CF I_m68 are functionally interchangeable. CF I_m25, CF I_m59 and CF I_m68 share many interaction partners and structures of the CF I_m25/CF I_m59-RRM and CFI_m25/CFI_m68-RRM complexes suggest a nearly identical overall domain architecture [12]. However, subtle differences between the sequences of CF I_m59 and CF I_m68 or amino acid modifications not obvious in the structure could enable these proteins to establish distinct interactions with target RNAs and carry out somewhat different functions. Consistent with this hypothesis are observations that CF I_m59 and CF I_m68 also have distinct interaction partners. CF I_m68 has been shown to interact with the SR proteins hTra2b, Srp20 and 9G8 [13] and CF I_m59 with U2AF65 [14]. In both cases these interactions take place via serine/arginine rich (SR) domains. In addition, CF I_m59 interacts with the arginine methyltransferase PRMT2 [15, 16].

By cross-linking and immunoprecipitation (CLIP) followed by deep sequencing we recently mapped the transcriptome-wide binding sites of RNA-binding proteins of the core polyadenylation machinery including CF I_m25, CF I_m59, and CF I_m68 [17]. By further quantifying cleavage and polyadenylation (CP) site usage in HEK293 cells in which we mapped the binding sites, we showed that binding of CF I_m is predictive for the choice of a polyadenylation site, and that knock-down of CF I_m68 causes a transcriptome-wide increase in proximal CP site use. Here we report the results of follow-up experiments, in which we explored the effects of CF I_m25 and CF I_m59 knock-down, and discuss the general question of how CF I_m acts in the regulation of polyadenylation.

Transcriptome-wide analyses reveal extensive alternative polyadenylation

Alternative polyadenylation is a fundamental mechanism underlying eukaryotic mRNA diversity. Both computational and biochemical approaches have been used to map pre-mRNA 3’ ends and to characterize the proteins involved in 3’ end formation (for reviews, see [18, 19]). The recent work of Sandberg and colleagues [20], demonstrating that proliferating cells express transcripts whose 3’ UTRs are systematically shorter compared to those of resting cells, incited an upsurge of interest in this field. Several protocols to capture polyadenylation sites via deep sequencing have been developed, including 3SEQ [21], direct RNA sequencing (DRS) [22], 3P-Seq [23], MAPS [24], PAS-Seq [25], SAPAS [26], A-seq [17], and PolyA-Seq [27]. A systematic effort to combine the data generated in all of these studies has not been undertaken. However, the recent study of Babak and colleagues [27] alone resulted in a list of 280,000 human CP sites compared to a mere 150,000 sites that were known from previous work. The advantage of these deep sequencing-based methods is that they enable us to move away from a binary (present/absent), EST-based description [28], or a semi-quantitative, microarray-based measurement [29] of polyadenylation site usage in specific libraries or tissues, towards precise quantification of alternative polyadenylation site use. This in turn allows exploration of the processing mechanism in various conditions and for various classes of transcripts such as the still poorly understood noncoding RNAs.

Relationship between tissue-specific alternative polyadenylation and proliferation rate

Babak and colleagues [27] were the first to quantitatively determine CP site usage over a broad set of tissues as well as in actively proliferating cells. To determine whether differences in CP site use between individual tissues follow a systematic pattern, we obtained pre-processed read mappings from the NCBI GEO archive (GSE30198), and inferred CP sites using our computational pipeline that was previously described [17]. In total, we identified 1,047 genes with two tandem CP sites that show expression of at least 5 tags per million in each sample investigated. Following the approach of Sandberg et al. [20], we further computed a cell type-specific "proliferation index". For a given sample, the proliferation index was defined as the median z-score of the expression level of a cell cycle-associated gene [20] in the respective sample relative to all others. The scatter plot of the proximal/distal site usage ratio against the proliferation index for the samples in Fig. 1A shows the expected trend. First, replicate samples prepared from the same type of cells have very similar proliferation index as well as proximal/distal CP usage. Further, tissues with a low proliferation index such as the brain have low proximal/distal CP usage ratios compared with samples prepared from cells with high proliferation index such as the mixture of ten human cancer cell lines (MAQC-UHR samples from the Stratagene Universal Human Reference RNA). The correlation is however far from perfect. Proximal/distal CP site usage ratio differs quite strongly for tissues that have a comparable proliferation index (median $\log_{10}$ proximal/distal ratio of -0.53 for the brain and -0.31 for liver). Strikingly, the tissue-to-tissue differences appear to be systematic. This is illustrated more clearly in Fig. 1B, which shows that the scatter of proximal/distal CP site usage ratios for individual genes in pairs of brain samples forms a narrow band around the diagonal, while the brain against liver scatter shows a clear off-diagonal shift. This systematic, transcriptome-wide shift in CP site usage would be most parsimoniously explained by a "master regulator" that alters the CP site usage of most genes, rather than by many individual regulators that operate on small subsets of genes. The simplest lead to follow is the core polyadenylation machinery or a factor that directly interacts with it. We recently demonstrated that knock-down of CF I_m68, a key component of the mammalian polyadenylation apparatus, induces a systematic, transcriptome-wide shift to increased proximal CP site usage [17]. In this report, we further explore the role of the individual components of CF I_m in alternative polyadenylation.

Cleavage factor I as a key regulator of 3' UTR length

New advances in high-throughput technologies also fueled the investigation of binding patterns of RNA-binding proteins. UV crosslinking and immunoprecipitation followed by sequencing of the bound RNA fragments allow the identification of RNA molecules targeted by the protein of interest. These methods enable the mapping of binding sites with nucleotide level resolution, either by exploiting crosslink-diagnostic mutations (in PAR-CLIP [30, 31] and HITS-CLIP [32]) or the propensity of reverse transcriptase to stop at crosslinked sites [33].

We recently mapped by PAR-CLIP the transcriptome-wide binding sites for CF I_m25, CF I_m59, and CF I_m68 proteins in HEK293 cells. We found that all components of CF I_m exhibit very specific positioning 40-50 nucleotides (nt) upstream of cleavage and polyadenylation sites. The underlying cause of this positional specificity seems to be two-fold. In half of the CP sites investigated the binding profile of CF I_m components can be explained by the density profile of UGUA sequence motifs, which also peaks 40-50 nt upstream of the CP site. However, even CP sites that do not have any UGUA within the 100 upstream nucleotides exhibit the same peak in the CF I_m read density at 40-50 nt. This suggests that positioning of CF I_m on the pre-mRNA is not only governed by sequence-specific binding, but also by interactions with other factors such as CPSF. Motif analysis revealed that CF I_m CLIP reads were enriched in the UGUA tetramer. Detailed investigation of the cross-linking pattern further showed a positional bias of individual components of CF I_m with respect to the crosslinked nucleotide. Despite the presence of two U residues that could act as crosslinking sites when replaced by 4-thio-U in the UGUA motif, none of the CF I_m components cross-linked efficiently directly to UGUA. The weak crosslinking efficiency of CF I_m59 and

CF I_m68 to UGUA may be explained in terms of the mode of interaction of CF I_m inferred from recent structural studies [10, 11], that rather suggests that CF I_m25 specifically recognizes UGUA. However, the reason for the rather weak cross-linking of CF I_m25 to UGUA remains unclear; a possible explanation may be that the substitution of U with 4-thio-U decreases the affinity of interaction between the UGUA sequence and CF I_m25. In a comparison of CF I_m59 and CF I_m68 in complex with CF I_m25 and RNA Yang and colleagues describe the overall architecture of both complexes as nearly identical, but also point out that the minor differences observed could lead to different ways RNA is bound by each of these complexes [12]. Indeed, we observed differences in the cross-linking patterns of CF I_m59 and CF I_m68 as well. CF I_m68 was most efficiently cross-linked immediately downstream of UGUA motifs, whereas CF I_m59 cross-linking at this position was only slightly above background. Intersection of binding profiles of 3' end processing factors with CP site usage showed CF I_m68 and CstF-64 as the most predictive factors for CP site choice. We used A-seq to quantify the effect of the knock-down of these two factors on CP site choice and found that CF I_m68 but not CstF-64 loss-of-function led to a transcriptome-wide increase in the use of proximal CP sites (Fig. 2A) [17]. To further clarify the role of CF I_m in the regulation of 3'UTR length, we generated four additional A-seq libraries from HEK293 cells that were either grown under standard conditions without treatment, treated with a control siRNA, or treated with siRNAs directed against the CF I_m25 and CF I_m59 components of CF I_m. We also obtained an additional A-seq sample from a more efficient CF I_m68 knock-down relative to our initial study [17] (Fig. 2D) as well as a paired A-seq sample from cells treated with control siRNA.

We found that reduced levels of CF I_m25 and CF I_m68, but not of CF I_m59 lead to a transcriptome-wide increase in proximal CP site usage. These findings generalize the results of [34] to the entire transcriptome (Fig. 2A,B) and demonstrate that the CF I_m25/CF I_m68 complex globally controls 3' UTR length by suppression of proximal CP sites. The precise molecular mechanism underlying these observations remains to be elucidated.

Master regulators of 3' UTR length

The search for master regulators of 3'UTR length has revealed additional candidates. In a recent report, Berg and colleagues [35] proposed that the U1 snRNP, that normally protects pre-mRNAs from premature cleavage and polyadenylation [36], becomes limiting when cells divide rapidly, leading to a general shortening of 3' UTRs. They illustrated this phenomenon in neurons, in which the rapid transcriptional boost induced by activation led to a relative decrease in U1 snRNP availability, which in turn caused increased usage of proximal CP sites. The mechanism behind this effect remains, like in the case of CF I_m, to be characterized.

Another recent study found that knock-down of the nuclear poly(A) binding protein PABPN1 leads to increased usage of proximal CP sites transcriptome-wide [37]. The authors proposed a model whereby under normal conditions, PABPN1 competes with the polyadenylation machinery for weak or non-canonical CP sites, which in the absence of PABPN1 are unmasked and processed. To investigate this hypothesis and more specifically to test whether the CF I_m component of the cleavage and polyadenylation machinery specifically increases the selection of weak CP sites, we grouped genes according to the relative strength of the most proximal relative to the most distal CP site (Fig. 2C; for the calculation of the hexamer score, see [17]). In our previous work [17] we showed that distal sites are on average, stronger and they are preferentially used in polyadenylation. We determined the change in proximal/distal ratio that different categories of genes undergo upon CF I_m68 and CF I_m25 knock-down and found that the knock-downs induce a similar increase in proximal/distal ratio irrespective of the relative strength of the proximal sites. This indicates that suppression of proximal CP sites by the CF I_m25/CF I_m68 complex is not biased by the "strength" of the CP site, as has been proposed for PABPN1.

Is 3'UTR length actively regulated?

The question now arises how downregulation of CF I_m25/68, U1 snRNP or PABPN1 promotes selection of the proximal instead of distal poly(A) sites for cleavage.

One explanation may be that cleavage is the default behavior of the 3’ end processing machinery, most of the factors in the complex serving to mask polyadenylation sites or to prevent the interaction of the cleavage factor with the putative polyadenylation site. This hypothesis is consistent with observations that systematic shifts in polyadenylation sites are induced by the knock-down of several, very different factors, but it is difficult to reconcile with observations that binding of many factors of the 3’ end processing complex occurs predominantly at the sites where 3’ end reads are also most abundant. To explain this paradox, we proposed in our previous study [17] that the cleavage sites that are used for cleavage under normal conditions promote formation of specific 3’ end processing complex conformations that allow cleavage in spite of the fact that cleavage-inhibitory factors are present. It will be very interesting to determine whether the different factors that have been shown to suppress the use of proximal CP sites act on different subsets of genes, whose expression is thereby coordinately regulated in specific conditions.

Possible mechanisms by which CF I_m alone may modulate alternative polyadenylation are depicted in Fig 3. One alternative is that the composition of the CF I_m complex is condition-dependent. Data collected so far suggest that CF I_m is a heterotetramer consisting of a CF I_m25 dimer in complex with either CF I_m59 or CF I_m68. Coimmunoprecipitation of FLAG-CF I_m59 or FLAG-CF I_m68 indicates that CF I_m59 and CF I_m68 can be present in the same complex (Fig. 3D) with a CF I_m25 dimer. In addition, a 72 kDa alternatively spliced form of the CF I_m68 protein (CF I_m72) [7] found in mammals could also take part in and change the functionality of the CF I_m complex. Thus, one way to modulate the choice of poly(A) sites could be by changing the composition of the CF I_m complex.

Another related possibility is that binding of CF I_m to its RNA targets or to protein-binding partners is modulated by posttranslational modifications. In fact, phosphorylation of a purified cleavage factor fraction (containing CF I_m and CF II_m) was found to be required for in vitro cleavage and polyadenylation [38]. Ser166 in the RRM of CF I_m68 is subject to phosphorylation, and mutation studies replacing Ser166 by aspartate, a phosphate mimic, revealed a twofold increase in RNA binding affinity of the CF I_m25/CF I_m68 complex [12]. Moreover, CF I_m68 from Hela cells, but not CF I_m59, was found to contain symmetrically dimethylated arginines and that it could be methylated at a glycine-arginine rich (GAR) motif *in vitro* by the methyltransferase PRMT5 [15]. CF I_m59 from Hela cell nuclei is more strongly modified by asymetrical dimethylation than CF I_m68 and both proteins can be dimethylated by the methyltransferase PRMT1 *in vitro* mainly at the C-terminus that is rich in arginines. However, no effects of these modifications on protein-protein interactions or RNA binding capacity of the CF I_m factors were so far identified [15].

CF I_m68 is not the only component of CF I_m that has been found to be post-translationally modified. Lysine residue 23 of CF I_m25 is acetylated by CREB-binding protein and knock-down of CF I_m68 reduced CF I_m25 acetylation suggesting that CF I_m68 is needed for efficient acetylation [39]. Modulation of CF I_m binding affinity could be consistent with the RNA looping model proposed by Yang and colleagues [10]. Reduced binding of CF I_m would prevent looping of alternative CP sites and enable the CP site to be recognized and cleaved by CPSF.

Finally, Shimazu et al. [39] also found that acetylation of CF I_m25 decreases the interaction of CF I_m with poly(A) polymerase. This suggests that it may be the polyadenylation rather than the cleavage step that is modulated by CF I_m and other factors. Indeed, direct interactions of the U1 snRNP proteins U1A and U1-70K [40, 41] as well as of the U2 snRNP-associated protein U2AF65 [14] with poly(A) polymerase were shown to inhibit polyadenylation of the newly cleaved pre-mRNA. This suggests that the presence of factors that are involved in pre-mRNA processing steps that precede cleavage and polyadenylation suppresses polyadenylation of transcripts that were prematurely cleaved. This in turn would also suppress the export and translation of these abortive transcripts because they would lack poly(A) tails. Northern blots with total RNA upon RNAi-mediated knock-down of CF I_m68 appear to show shortening of the transcripts to proximal cleavage sites [15], although it can still be that the long, non-polyadenylated transcripts are unstable.

The availability of technologies for exploring the entire transcriptome of a cell at once brought a new appreciation of the complexity of regulation of gene expression. At the same time, they allow us to identify biologically relevant patterns, taking advantage of the possibly very small responses of a large number of genes. It will be exciting to see new applications of this approach in the field of RNA 3' end processing.

Methods

A-seq

A-seq was carried out as described [17] with the exception of the partial RNA fragmentation step, which consisted of alkaline hydrolysis instead of RNase I digestion. To this end, poly(A) containing RNA was released from $(dT)_{25}$-Dynabeads in 35 µl 5 mM Tris-Cl pH 8.0. 70 µl alkaline hydrolysis buffer was added. Hydrolysis buffer is 50 mM Na-CO_3, 1 mM EDTA, pH 9.2 and was prepared by mixing 1 ml 0.1 M Na_2CO_3 with 9 ml 0.1 M $NaHCO_3$, adding EDTA to 1 mM, adjusting the pH to 9.2 and the volume to 20 ml with H_2O. The reactions are incubated for exactly 7 minutes at 95 °C. Reactions were chilled on ice and 500 µl lysis buffer of the mRNA-DIRECT kit (Invitrogen) were added. The Dynabeads from the first step were recycled to bind the fragmented RNA that still contains poly(A). After washing the beads with buffers A and B, the protocol continues with 5' end phosphorylation as described [17]. The Gene Expression Omnibus (GEO) accession number for the A-seq data is GSEXXXXX.

RNAi

Silencer Select siRNAs (Ambion) were used for knock-downs of CF I_m25 (S224836) and CF I_m59 (S21772). For RNAi with CF I_m68 a double stranded RNA oligo with sequence 5'-NNG ACC GAG AUU ACA UGG AUA-3' was obtained from Dharmacon. As a negative control the oligo 5'-AGG UAG UGU AAU CGC CUU GTT-3' (1491991) from Microsynth was used. RNAiMax transfection agent (Invitrogen) was employed according to the forward transfection method of the supplier. Cells were harvested after 3 days.

Western blots

Flp-In-293 cells either without transgene or stably transformed with either Flag-CF I_m59 or Flag-CF I_m68 fusion constructs in pcDNA5 plasmids (Invitrogen) were grown to 70% confluency, harvested and frozen at -80 °C as pellets. Pellets were lysed in PND buffer (1xPBS, 0.5% NP-40, 1 mM DTT and "cOmplete" protease inhibitor (Roche) and sonicated for 10-20 sec. 30 µg of protein from the lysates of was loaded onto 10% SDS gels. In addition, lysate containing 50 µg of protein was co-immunoprecipitated with anti-Flag antibody (M2 monoclonal from Sigma) coupled to magnetic protein-G Dynabeads (Invitrogen). Beads were washed 3x with PND buffer containing 0.1% NP-40. Bound proteins were released by heating to 90 °C in NUPAGE LDS sample buffer (Invitrogen) containing 0.1 M DTT. Lysates and supernates from co-IP after magnetic retention were loaded on the SDS gel, blotted to ECL membrane (GE Healthcare), filters were probed with anti-CFI$_m$ antibody [7] and further processed with the ECL system (Invitrogen).

Acknowledgements

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.). We acknowledge the technical support of Béatrice Dimitriades.

References

[1] Shatkin AJ, Manley JL. The ends of the affair: capping and polyadenylation. Nat Struct Biol 2000; 7:838–42.

[2] Carrillo Oesterreich F, Bieberstein N, Neugebauer KM. Pause locally, splice globally. Trends Cell Biol 2011; 21:328–35.

[3] Mandel CR, Bai Y, Tong L. Protein factors in pre-mRNA 3'-end processing. Cell Mol Life Sci 2008; 65:1099–122.

[4] Millevoi S, Vagner S. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. Nucleic Acids Res 2010; 38:2757–74.

[5] Shi Y, et al. Molecular architecture of the human pre-mRNA 3' processing complex. Mol Cell 2009; 33:365–76.

[6] Wahle E. A novel poly(A)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation. Cell 1991; 66:759–68.

[7] Rüegsegger U, Beyer K, Keller W. Purification and characterization of human cleavage factor Im

involved in the 3’ end processing of messenger RNA precursors. J Biol Chem 1996; 271:6107–13.

[8] Yang Q, Gilmartin GM, Doublié S. Structural basis of UGUA recognition by the Nudix protein CFI(m)25 and implications for a regulatory role in mRNA 3’ processing. Proc Natl Acad Sci U S A 2010; 107:10062–7.

[9] Brown KM, Gilmartin GM. A mechanism for the regulation of Pre-mRNA 3’ processing by human cleavage factor Im. Mol Cell 2003; 12:1467 – 1476.

[10] Yang Q, Coseno M, Gilmartin GM, Doublié S. 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 2011; 19:368–77.

[11] Li H, et al. Structural basis of pre-mRNA recognition by the human cleavage factor Im complex. Cell Res 2011; 21:1039–51.

[12] Yang Q, Gilmartin GM, Doublié S. The structure of human cleavage factor Im hints at functions beyond UGUA-specific RNA binding: A role in alternative polyadenylation and a potential link to 5’ capping and splicing. RNA Biol 2011; 8:748–53.

[13] Dettwiler S, Aringhieri C, Cardinale S, Keller W, Barabino SML. Distinct sequence motifs within the 68-kDa subunit of cleavage factor Im mediate RNA binding, Protein-Protein interactions, and subcellular localization. J Biol Chem 2004; 279:35788–35797.

[14] Millevoi S, et al. An interaction between U2AF 65 and CF I(m) links the splicing and 3’ end processing machineries. EMBO J 2006; 25:4854–64.

[15] Martin G, et al. Arginine methylation in subunits of mammalian pre-mRNA cleavage factor I. RNA 2010; 16:1646–59.

[16] Rual JF, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature 2005; 437:1173–8.

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

[18] Chan S, Choi EA, Shi Y. Pre-mRNA 3’-end processing complex assembly and function. Wiley Interdiscip Rev RNA 2011; 2:321–35.

[19] Tian B, Graber JH. Signals for pre-mRNA cleavage and polyadenylation. Wiley Interdiscip Rev RNA 2012; 3:385–96.

[20] Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science 2008; 320:1643–7.

[21] Beck AH, et al. 3'-end sequencing for expression quantification (3SEQ) from archival tumor samples. PLoS One 2010; 5:e8768.

[22] Ozsolak F, et al. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell 2010; 143:1018–29.

[23] Jan CH, Friedman RC, Ruby JG, Bartel DP. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. Nature 2011; 469:97–101.

[24] Fox-Walsh K, Davis-Turak J, Zhou Y, Li H, Fu XD. A multiplex RNA-seq strategy to profile poly(A(+)) RNA: application to analysis of transcription response and 3' end formation. Genomics 2011; 98:266–71.

[25] Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA 2011; 17:761–72.

[26] Fu Y, et al. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome Res 2011; 21:741–7.

[27] Derti A, et al. A quantitative atlas of polyadenylation in five mammals. Genome Res 2012; 22:1173–83.

[28] Zhang H, Lee JY, Tian B. Biased alternative polyadenylation in human tissues. Genome Biol 2005; 6:R100.

[29] Ji Z, Lee JY, Pan Z, Jiang B, Tian B. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc Natl Acad Sci U S A 2009; 106:7028–33.

[30] Hafner M, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 2010; 141:129–41.

[31] Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat Methods 2011; 8:559–64.

[32] Zhang C, Darnell RB. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nat Biotechnol 2011; 29:607–14.

[33] König J, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol 2010; 17:909–15.

[34] Kim S, et al. Evidence that cleavage factor Im is a heterotetrameric protein complex controlling alternative polyadenylation. Genes Cells 2010; 15:1003–13.

[35] Berg MG, et al. U1 snRNP determines mRNA length and regulates isoform expression. Cell 2012; 150:53–64.

[36] Kaida D, et al. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature 2010; 468:664–8.

[37] Jenal M, et al. The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. Cell 2012; 149:538–53.

[38] Ryan K. Pre-mRNA 3' cleavage is reversibly inhibited in vitro by cleavage factor dephosphorylation. RNA Biol 2007; 4:26–33.

[39] Shimazu T, Horinouchi S, Yoshida M. Multiple histone deacetylases and the CREB-binding protein regulate pre-mRNA 3'-end processing. J Biol Chem 2007; 282:4470–8.

[40] Gunderson SI, Beyer K, Martin G, Keller W, Boelens WC, Mattaj LW. The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase. Cell 1994; 76:531–41.

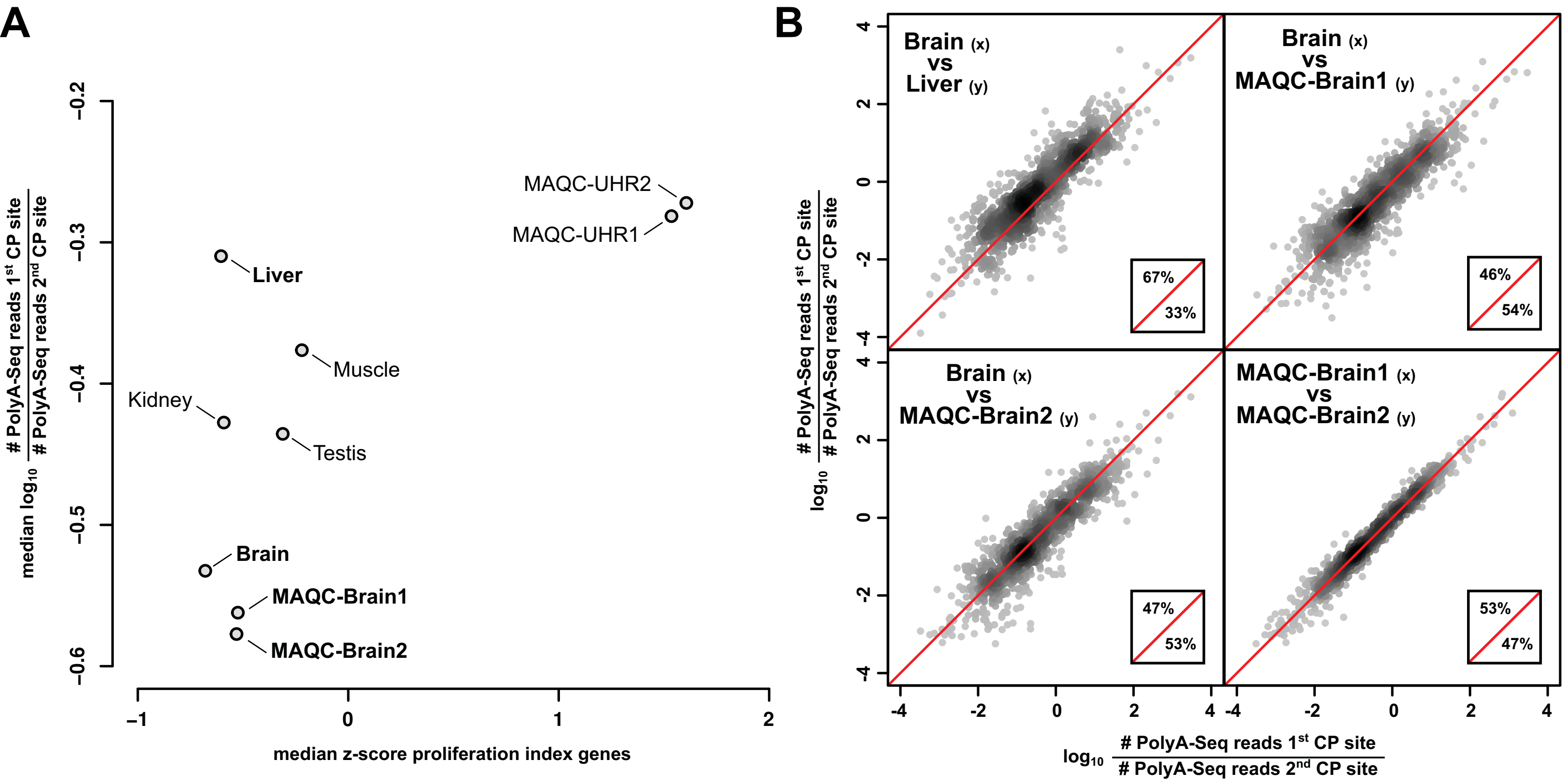
[41] Gunderson SI, Polycarpou-Schwarz M, Mattaj IW. U1 snRNP inhibits pre-mRNA polyadenylation through a direct interaction between U1 70K and poly(A) polymerase. Mol Cell 1998; 1:255–64.

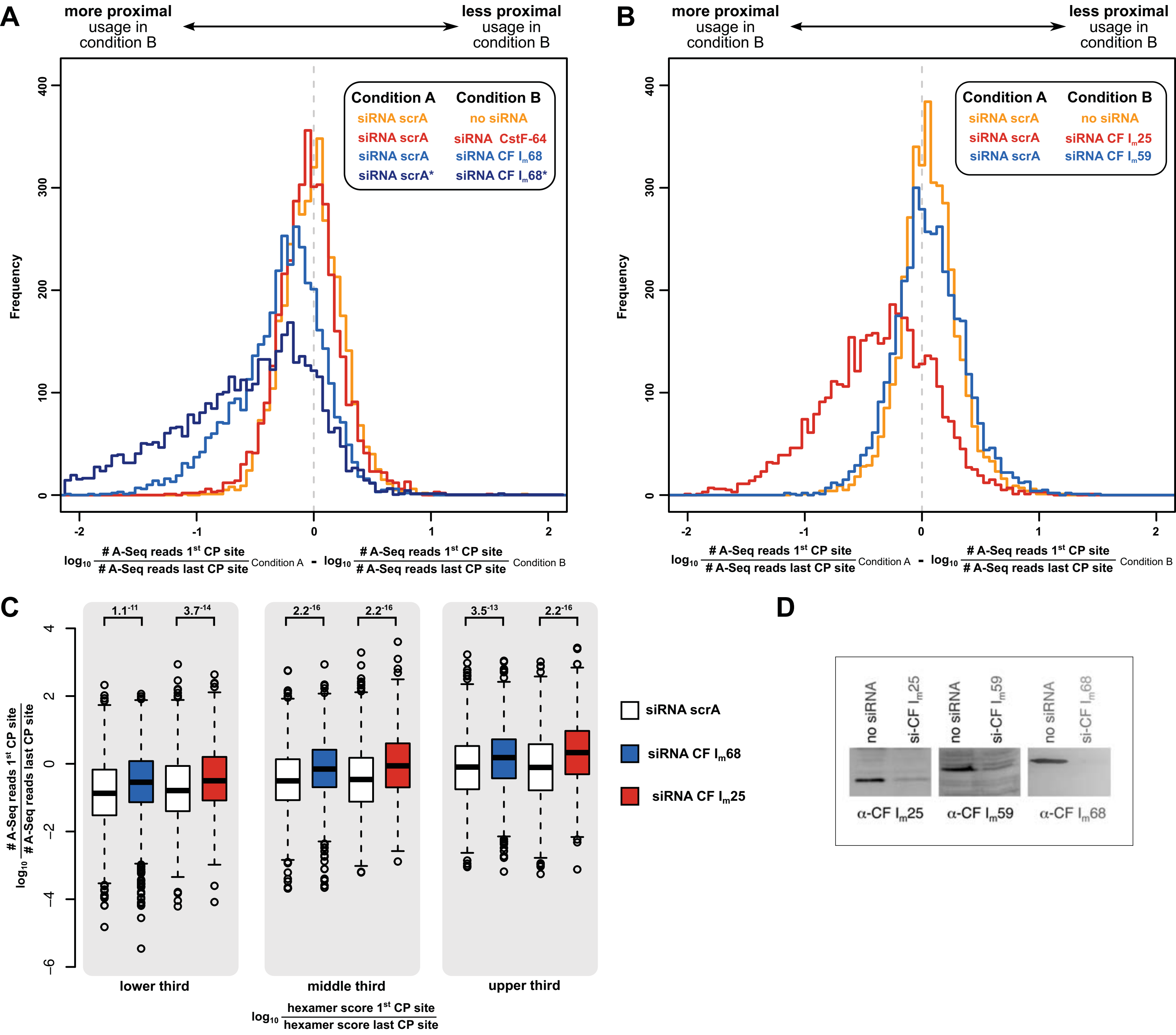
Figure 1. Comparison of proximal/distal CP usage ratios of 1,047 human genes with two tandem CP sites in tissues covered by the data set of Derti and colleagues [27]. (A) Scatter plot relating proliferative index (x-axis) to CP site usage (y-axis) (see text for the computation of these quantities). (B) Scatter plots of proximal/distal CP usage ratios in brain, liver and MAQC-brain samples. The grey scale indicates the density of data points representing individual genes. Numbers in the insets represent the proportion of points above and below the diagonal that indicates identical proximal/distal CP usage ratio for a gene in the two tissues.

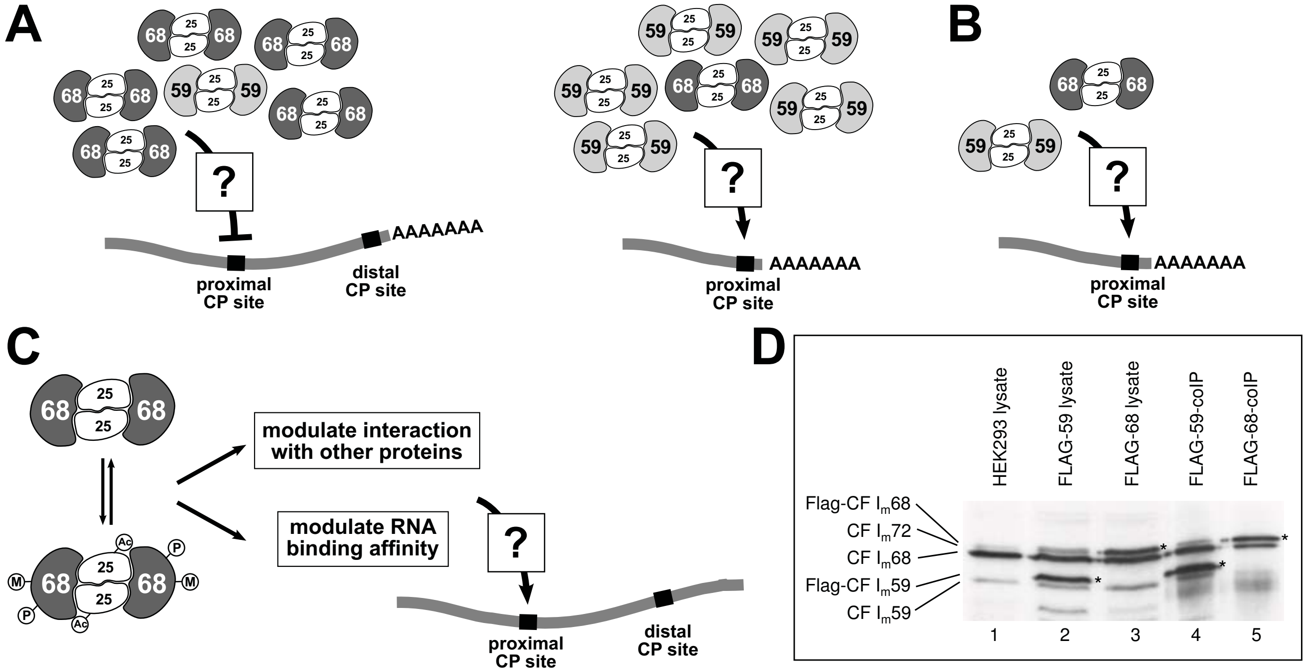
Figure 2. Changes in cleavage and polyadenylation site usage upon knock-down of CF I_m components and of CstF-64 in HEK 293 cells. A total of 3,821 transcripts with 2, 3 or 4 tandem CP sites (inferred based on the A-seq sequence data [17] and located in the same 3' UTR exon) whose expression was estimated to be at least five A-seq tags per million in both untreated samples were selected. (A) Data sets were described in [17] . An additional CF I_m68 knock-down data set (marked by the asterisk) was generated in this study. (B) Comparison of CP site usage in CF I_m25 and CF I_m59 knock-down sample relative to a control siRNA. (C) Proximal shift in CP site usage under CF I_m25 and CF I_m68 knock-down conditions as a function of the relative strength of the proximal CP site. Genes were divided into three subsets based on the ratio of hexamer scores [17] of the most proximal and most distal CP sites. Within each subset, we computed the proximal/distal CP usage ratio in a knock-down compared to the corresponding control siRNA-treated sample. Box-plots summarize the distribution of proximal/distal CP usage ratio for all genes within a particular subset and a particular sample. P-values of the t-test comparing the means of the two distributions are shown above the box-plots. (D) Western blots showing the efficiency of CF I_m25, CF I_m68 and CF I_m59 knock-downs.

Figure 3. Possible models of modulation of alternative polyadenylation by CF I_m. (A) High concentration of CF I_m68 relative to CF I_m59 leads to suppression of proximal CP sites. (B) Overall low levels of CF I_m and hence low abundance of CF I_m25/CF I_m68 promote cleavage and polyadenylation at proximal sites.

(C) Post-translational modifications modulate RNA and protein interactions of CF I$_m$. (D) Co-immunoprecipitation experiments with FLAG-CF I$_m$59 and FLAG-CF I$_m$68 indicate that the FLAG-tagged CFI$_m$ proteins can randomly bind both CF I$_m$59 and CF I$_m$68 (and in addition CF I$_m$72) and possibly also form dimers of the Flag-tagged versions. Asterisks mark FLAG-tagged proteins.


A
median log10 (# PolyA-Seq reads 1st CP site / # PolyA-Seq reads 2nd CP site)
−0.2
−0.3
−0.4
−0.5
−0.6
MAQC-UHR2
MAQC-UHR1
Liver
Muscle
Kidney
Testis
Brain
MAQC-Brain1
MAQC-Brain2
−1
0
1
2
median z-score proliferation index genes
B
log10 (# PolyA-Seq reads 1st CP site / # PolyA-Seq reads 2nd CP site)
Brain (x) vs Liver (y)
67%
33%
Brain (x) vs MAQC-Brain1 (y)
46%
54%
Brain (x) vs MAQC-Brain2 (y)
47%
53%
MAQC-Brain1 (x) vs MAQC-Brain2 (y)
53%
47%
4
2
0
-2
-4
log10 (# PolyA-Seq reads 1st CP site / # PolyA-Seq reads 2nd CP site)


A
more proximal usage in condition B
less proximal usage in condition B
Condition A
Condition B
siRNA scrA
no siRNA
siRNA scrA
siRNA CstF-64
siRNA scrA
siRNA CF Im68
siRNA scrA*
siRNA CF Im68*
Frequency
0
100
200
300
400
-2
-1
0
1
2
log10 # A-Seq reads 1st CP site / # A-Seq reads last CP site Condition A - log10 # A-Seq reads 1st CP site / # A-Seq reads last CP site Condition B
B
more proximal usage in condition B
less proximal usage in condition B
Condition A
Condition B
siRNA scrA
no siRNA
siRNA scrA
siRNA CF Im25
siRNA scrA
siRNA CF Im59
Frequency
0
100
200
300
400
-2
-1
0
1
2
log10 # A-Seq reads 1st CP site / # A-Seq reads last CP site Condition A - log10 # A-Seq reads 1st CP site / # A-Seq reads last CP site Condition B
C
1.1-11
3.7-14
2.2-16
2.2-16
3.5-13
2.2-16
4
2
0
−2
−4
−6
log10 # A-Seq reads 1st CP site / # A-Seq reads last CP site
lower third
middle third
upper third
log10 hexamer score 1st CP site / hexamer score last CP site
□ siRNA scrA
🟦 siRNA CF Im68
🟥 siRNA CF Im25
D
no siRNA
si-CF Im25
no siRNA
si-CF Im59
no siRNA
si-CF Im68
α-CF Im25
α-CF Im59
α-CF Im68


A
68
25
59
?
AAAAAAA
proximal CP site
distal CP site
B
C
modulate interaction with other proteins
modulate RNA binding affinity
Ac
P
M
D
HEK293 lysate
FLAG-59 lysate
FLAG-68 lysate
FLAG-59-coIP
FLAG-68-coIP
Flag-CF Im68
CF Im72
CF Im68
Flag-CF Im59
CF Im59
*
1
2
3
4
5