

1 Cleavage Factor I_m as a key regulator of 3' UTR length

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

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

4 Keywords: cleavage factor I_m , CF I_m , CPSF5, CPSF6, CPSF7, alternative polyadenylation, 3' end
5 processing, mRNA processing

6 Abbreviations: CF I_m , cleavage factor I_m ; CP, cleavage and polyadenylation; UTR, untranslated region;
7 CPSF, cleavage and polyadenylation specificity factor; CstF, cleavage stimulation factor; RRM, RNA
8 recognition motif; CLIP, cross-linking and immunoprecipitation

9 Abstract

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

1 **Introduction**

2 Generation of mature eukaryotic mRNAs from pre-mRNAs includes addition of a 7-methylguanosine cap,
3 splicing out of introns and cleavage and polyadenylation of the 3' end [1, 2, 3, 4]. Most of these processes
4 are carried out co-transcriptionally by a number of protein complexes and are completed before the
5 transcription complex reaches the end of the gene. The process of cleavage and polyadenylation which is
6 the focus of our work, involves a complex that contains up to 85 proteins [5]. At the core however, are a
7 few smaller subcomplexes: the cleavage and polyadenylation specificity factor (CPSF), cleavage
8 stimulation factor (CstF), cleavage factors I_m and II_m (CF I_m and CF II_m), a poly(A) polymerase (PAP) [4],
9 and the nuclear poly(A) binding protein 1 (PABPN1) [6].
10 CF I_m is a tetramer composed of two 25 kDa (CF I_m 25) subunits and two proteins of either 59 or 68 kDa
11 (CF I_m 59 or CF I_m 68) [7, 8]. It was previously shown through SELEX analysis to preferentially bind
12 UGUA subsequences in the pre-mRNAs [9]. The molecular basis of this interaction emerged from recently
13 solved crystal structures of CF I_m 25 in complex with the RNA recognition motif (RRM) of CF I_m 68 [10,
14 11]. Surprisingly, it is the Nudix hydrolase domain of CF I_m 25 that specifically recognizes UGUA,
15 whereas CF I_m 68 appears to increase the binding affinity of the complex. These structure models further
16 revealed that a CF I_m 25 dimer binds two UGUA sequences in an antiparallel manner forcing the looping
17 of the RNA sequence between the UGUA motifs. Yang and colleagues proposed that looping might
18 facilitate alternative polyadenylation via CF I_m [10]. The composition of individual CF I_m complexes that
19 bind pre-mRNA molecules is not known and it is unclear whether CF I_m 59 and CF I_m 68 are functionally
20 interchangeable. CF I_m 25, CF I_m 59 and CF I_m 68 share many interaction partners and structures of the
21 CF I_m 25/CF I_m 59-RRM and CF I_m 25/CF I_m 68-RRM complexes suggest a nearly identical overall domain
22 architecture [12]. However, subtle differences between the sequences of CF I_m 59 and CF I_m 68 or amino
23 acid modifications not obvious in the structure could enable these proteins to establish distinct interactions
24 with target RNAs and carry out somewhat different functions. Consistent with this hypothesis are
25 observations that CF I_m 59 and CF I_m 68 also have distinct interaction partners. CF I_m 68 has been shown to

1 interact with the SR proteins hTra2b, Srp20 and 9G8 [13] and CF I_m59 with U2AF65 [14]. In both cases
2 these interactions take place via serine/arginine rich (SR) domains. In addition, CF I_m59 interacts with the
3 arginine methyltransferase PRMT2 [15, 16].

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

12 **Transcriptome-wide analyses reveal extensive alternative polyadenylation**

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

1 of polyadenylation site usage in specific libraries or tissues, towards precise quantification of alternative
2 polyadenylation site use. This in turn allows exploration of the processing mechanism in various
3 conditions and for various classes of transcripts such as the still poorly understood noncoding RNAs.

4 **Relationship between tissue-specific alternative polyadenylation and proliferation rate**

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

1 core polyadenylation machinery or a factor that directly interacts with it. We recently demonstrated that
2 knock-down of CF I_m68, a key component of the mammalian polyadenylation apparatus, induces a
3 systematic, transcriptome-wide shift to increased proximal CP site usage [17]. In this report, we further
4 explore the role of the individual components of CF I_m in alternative polyadenylation.

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

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

12 We recently mapped by PAR-CLIP the transcriptome-wide binding sites for CF I_m25, CF I_m59, and CF
13 I_m68 proteins in HEK293 cells. We found that all components of CF I_m exhibit very specific positioning
14 40-50 nucleotides (nt) upstream of cleavage and polyadenylation sites. The underlying cause of this
15 positional specificity seems to be two-fold. In half of the CP sites investigated the binding profile of CF I_m
16 components can be explained by the density profile of UGUA sequence motifs, which also peaks 40-50 nt
17 upstream of the CP site. However, even CP sites that do not have any UGUA within the 100 upstream
18 nucleotides exhibit the same peak in the CF I_m read density at 40-50 nt. This suggests that positioning of
19 CF I_m on the pre-mRNA is not only governed by sequence-specific binding, but also by interactions with
20 other factors such as CPSF. Motif analysis revealed that CF I_m CLIP reads were enriched in the UGUA
21 tetramer. Detailed investigation of the cross-linking pattern further showed a positional bias of individual
22 components of CF I_m with respect to the crosslinked nucleotide. Despite the presence of two U residues
23 that could act as crosslinking sites when replaced by 4-thio-U in the UGUA motif, none of the CF I_m
24 components cross-linked efficiently directly to UGUA. The weak crosslinking efficiency of CF I_m59 and

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

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

1 **Master regulators of 3' UTR length**

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

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

22 **Is 3'UTR length actively regulated?**

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

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

12 Possible mechanisms by which CF I_m alone may modulate alternative polyadenylation are depicted in Fig
13 3. One alternative is that the composition of the CF I_m complex is condition-dependent. Data collected so
14 far suggest that CF I_m is a heterotetramer consisting of a CF I_{m25} dimer in complex with either CF I_{m59}
15 or CF I_{m68}. Coimmunoprecipitation of FLAG-CF I_{m59} or FLAG-CF I_{m68} indicates that CF I_{m59} and CF
16 I_{m68} can be present in the same complex (Fig. 3D) with a CF I_{m25} dimer. In addition, a 72 kDa
17 alternatively spliced form of the CF I_{m68} protein (CF I_{m72}) [7] found in mammals could also take part in
18 and change the functionality of the CF I_m complex. Thus, one way to modulate the choice of poly(A) sites
19 could be by changing the composition of the CF I_m complex.

20 Another related possibility is that binding of CF I_m to its RNA targets or to protein-binding partners is
21 modulated by posttranslational modifications. In fact, phosphorylation of a purified cleavage factor
22 fraction (containing CF I_m and CF II_m) was found to be required for in vitro cleavage and polyadenylation
23 [38]. Ser166 in the RRM of CF I_{m68} is subject to phosphorylation, and mutation studies replacing Ser166
24 by aspartate, a phosphate mimic, revealed a twofold increase in RNA binding affinity of the CF I_{m25}/CF
25 I_{m68} complex [12]. Moreover, CF I_{m68} from HeLa cells, but not CF I_{m59}, was found to contain

1 symmetrically dimethylated arginines and that it could be methylated at a glycine-arginine rich (GAR)
2 motif *in vitro* by the methyltransferase PRMT5 [15]. CF I_m59 from HeLa cell nuclei is more strongly
3 modified by asymmetrical dimethylation than CF I_m68 and both proteins can be dimethylated by the
4 methyltransferase PRMT1 *in vitro* mainly at the C-terminus that is rich in arginines. However, no effects
5 of these modifications on protein-protein interactions or RNA binding capacity of the CF I_m factors were
6 so far identified [15].

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

13 Finally, Shimazu et al. [39] also found that acetylation of CF I_m25 decreases the interaction of CF I_m with
14 poly(A) polymerase. This suggests that it may be the polyadenylation rather than the cleavage step that is
15 modulated by CF I_m and other factors. Indeed, direct interactions of the U1 snRNP proteins U1A and U1-
16 70K [40, 41] as well as of the U2 snRNP-associated protein U2AF65 [14] with poly(A) polymerase were
17 shown to inhibit polyadenylation of the newly cleaved pre-mRNA. This suggests that the presence of
18 factors that are involved in pre-mRNA processing steps that precede cleavage and polyadenylation
19 suppresses polyadenylation of transcripts that were prematurely cleaved. This in turn would also suppress
20 the export and translation of these abortive transcripts because they would lack poly(A) tails. Northern
21 blots with total RNA upon RNAi-mediated knock-down of CF I_m68 appear to show shortening of the
22 transcripts to proximal cleavage sites [15], although it can still be that the long, non-polyadenylated
23 transcripts are unstable.

24 The availability of technologies for exploring the entire transcriptome of a cell at once brought a new
25 appreciation of the complexity of regulation of gene expression. At the same time, they allow us to
26 identify biologically relevant patterns, taking advantage of the possibly very small responses of a large

1 number of genes. It will be exciting to see new applications of this approach in the field of RNA 3' end
2 processing.

3 **Methods**

4 **A-seq**

5 A-seq was carried out as described [17] with the exception of the partial RNA fragmentation step, which
6 consisted of alkaline hydrolysis instead of RNase I digestion. To this end, poly(A) containing RNA was
7 released from (dT)₂₅-Dynabeads in 35 µl 5 mM Tris-Cl pH 8.0. 70 µl alkaline hydrolysis buffer was
8 added. Hydrolysis buffer is 50 mM Na₂CO₃, 1 mM EDTA, pH 9.2 and was prepared by mixing 1 ml 0.1 M
9 Na₂CO₃ with 9 ml 0.1 M NaHCO₃, adding EDTA to 1 mM, adjusting the pH to 9.2 and the volume to 20
10 ml with H₂O. The reactions are incubated for exactly 7 minutes at 95 °C. Reactions were chilled on ice
11 and 500 µl lysis buffer of the mRNA-DIRECT kit (Invitrogen) were added. The Dynabeads from the first
12 step were recycled to bind the fragmented RNA that still contains poly(A). After washing the beads with
13 buffers A and B, the protocol continues with 5' end phosphorylation as described [17]. The Gene
14 Expression Omnibus (GEO) accession number for the A-seq data is GSEXXXXXX.

15 **RNAi**

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

22 **Western blots**

23 Flp-In-293 cells either without transgene or stably transformed with either Flag-CF I_m59 or Flag-CF I_m68
24 fusion constructs in pcDNA5 plasmids (Invitrogen) were grown to 70% confluency, harvested and frozen

1 at -80 °C as pellets. Pellets were lysed in PND buffer (1xPBS, 0.5% NP-40, 1 mM DTT and “cOmplete”
2 protease inhibitor (Roche) and sonicated for 10-20 sec. 30 µg of protein from the lysates of was loaded
3 onto 10% SDS gels. In addition, lysate containing 50 µg of protein was co-immunoprecipitated with anti-
4 Flag antibody (M2 monoclonal from Sigma) coupled to magnetic protein-G Dynabeads (Invitrogen).
5 Beads were washed 3x with PND buffer containing 0.1% NP-40. Bound proteins were released by heating
6 to 90 °C in NUPAGE LDS sample buffer (Invitrogen) containing 0.1 M DTT. Lysates and supernates from
7 co-IP after magnetic retention were loaded on the SDS gel, blotted to ECL membrane (GE Healthcare),
8 filters were probed with anti-CFI_m antibody [7] and further processed with the ECL system (Invitrogen).

9 **Acknowledgements**

10 The work was supported by the University of Basel, the Louis-Jeantet Foundation for Medicine and the
11 Swiss National Science Foundation (grant 31003A_133145 to W.K.). We acknowledge the technical
12 support of Béatrice Dimitriadis.

13 **References**

- 14 [1] Shatkin AJ, Manley JL. The ends of the affair: capping and polyadenylation. *Nat Struct Biol* 2000;
15 7:838–42.
- 16 [2] Carrillo Oesterreich F, Bieberstein N, Neugebauer KM. Pause locally, splice globally. *Trends Cell*
17 *Biol* 2011; 21:328–35.
- 18 [3] Mandel CR, Bai Y, Tong L. Protein factors in pre-mRNA 3'-end processing. *Cell Mol Life Sci* 2008;
19 65:1099–122.
- 20 [4] Millevoi S, Vagner S. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation.
21 *Nucleic Acids Res* 2010; 38:2757–74.
- 22 [5] Shi Y, et al. Molecular architecture of the human pre-mRNA 3' processing complex. *Mol Cell* 2009;
23 33:365–76.
- 24 [6] Wahle E. A novel poly(A)-binding protein acts as a specificity factor in the second phase of
25 messenger RNA polyadenylation. *Cell* 1991; 66:759–68.
- 26 [7] Rügsegger U, Beyer K, Keller W. Purification and characterization of human cleavage factor Im

- 1 involved in the 3' end processing of messenger RNA precursors. *J Biol Chem* 1996; 271:6107–13.
- 2 [8] Yang Q, Gilmartin GM, Doublie S. Structural basis of UGUA recognition by the Nudix protein
3 CFI(m)25 and implications for a regulatory role in mRNA 3' processing. *Proc Natl Acad Sci U S A*
4 2010; 107:10062–7.
- 5 [9] Brown KM, Gilmartin GM. A mechanism for the regulation of Pre-mRNA 3' processing by human
6 cleavage factor Im. *Mol Cell* 2003; 12:1467 – 1476.
- 7 [10] Yang Q, Coseno M, Gilmartin GM, Doublie S. Crystal structure of a human cleavage factor
8 CFI(m)25/CFI(m)68/RNA complex provides an insight into poly(A) site recognition and RNA
9 looping. *Structure* 2011; 19:368–77.
- 10 [11] Li H, et al. Structural basis of pre-mRNA recognition by the human cleavage factor Im complex. *Cell*
11 *Res* 2011; 21:1039–51.
- 12 [12] Yang Q, Gilmartin GM, Doublie S. The structure of human cleavage factor Im hints at functions
13 beyond UGUA-specific RNA binding: A role in alternative polyadenylation and a potential link to 5'
14 capping and splicing. *RNA Biol* 2011; 8:748–53.
- 15 [13] Dettwiler S, Aringhieri C, Cardinale S, Keller W, Barabino SML. Distinct sequence motifs within
16 the 68-kDa subunit of cleavage factor Im mediate RNA binding, Protein-Protein interactions, and
17 subcellular localization. *J Biol Chem* 2004; 279:35788–35797.
- 18 [14] Millevoi S, et al. An interaction between U2AF 65 and CF I(m) links the splicing and 3' end
19 processing machineries. *EMBO J* 2006; 25:4854–64.
- 20 [15] Martin G, et al. Arginine methylation in subunits of mammalian pre-mRNA cleavage factor I. *RNA*
21 2010; 16:1646–59.
- 22 [16] Rual JF, et al. Towards a proteome-scale map of the human protein-protein interaction network.
23 *Nature* 2005; 437:1173–8.
- 24 [17] Martin G, Gruber AR, Keller W, Zavolan M. Genome-wide analysis of pre-mRNA 3' end processing
25 reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Rep* 2012;
26 1:753–763.
- 27 [18] Chan S, Choi EA, Shi Y. Pre-mRNA 3'-end processing complex assembly and function. *Wiley*
28 *Interdiscip Rev RNA* 2011; 2:321–35.

- 1 [19] Tian B, Graber JH. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip Rev RNA*
2 2012; 3:385–96.
- 3 [20] Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with
4 shortened 3' untranslated regions and fewer microRNA target sites. *Science* 2008; 320:1643–7.
- 5 [21] Beck AH, et al. 3'-end sequencing for expression quantification (3SEQ) from archival tumor
6 samples. *PLoS One* 2010; 5:e8768.
- 7 [22] Ozsolak F, et al. Comprehensive polyadenylation site maps in yeast and human reveal pervasive
8 alternative polyadenylation. *Cell* 2010; 143:1018–29.
- 9 [23] Jan CH, Friedman RC, Ruby JG, Bartel DP. Formation, regulation and evolution of *Caenorhabditis*
10 *elegans* 3'UTRs. *Nature* 2011; 469:97–101.
- 11 [24] Fox-Walsh K, Davis-Turak J, Zhou Y, Li H, Fu XD. A multiplex RNA-seq strategy to profile
12 poly(A(+)) RNA: application to analysis of transcription response and 3' end formation. *Genomics*
13 2011; 98:266–71.
- 14 [25] Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. Complex and dynamic landscape of
15 RNA polyadenylation revealed by PAS-Seq. *RNA* 2011; 17:761–72.
- 16 [26] Fu Y, et al. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and
17 normal cells by high-throughput sequencing. *Genome Res* 2011; 21:741–7.
- 18 [27] Derti A, et al. A quantitative atlas of polyadenylation in five mammals. *Genome Res* 2012; 22:1173–
19 83.
- 20 [28] Zhang H, Lee JY, Tian B. Biased alternative polyadenylation in human tissues. *Genome Biol* 2005;
21 6:R100.
- 22 [29] Ji Z, Lee JY, Pan Z, Jiang B, Tian B. Progressive lengthening of 3' untranslated regions of mRNAs
23 by alternative polyadenylation during mouse embryonic development. *Proc Natl Acad Sci U S A*
24 2009; 106:7028–33.
- 25 [30] Hafner M, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target
26 sites by PAR-CLIP. *Cell* 2010; 141:129–41.
- 27 [31] Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M. A quantitative analysis of
28 CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods* 2011; 8:559–64.

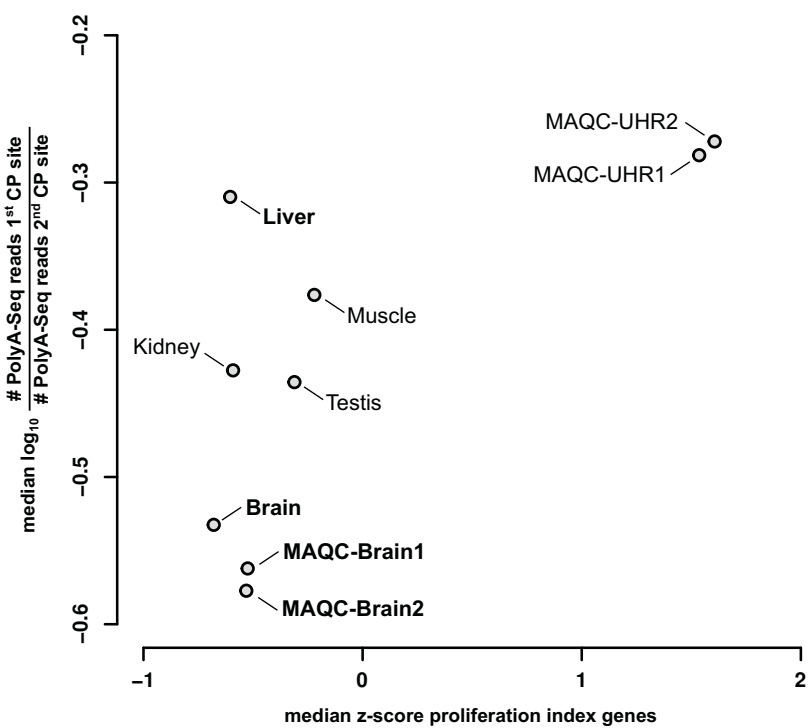
- 1 [32] Zhang C, Darnell RB. Mapping in vivo protein-RNA interactions at single-nucleotide resolution
2 from HITS-CLIP data. *Nat Biotechnol* 2011; 29:607–14.
- 3 [33] König J, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide
4 resolution. *Nat Struct Mol Biol* 2010; 17:909–15.
- 5 [34] Kim S, et al. Evidence that cleavage factor Im is a heterotetrameric protein complex controlling
6 alternative polyadenylation. *Genes Cells* 2010; 15:1003–13.
- 7 [35] Berg MG, et al. U1 snRNP determines mRNA length and regulates isoform expression. *Cell* 2012;
8 150:53–64.
- 9 [36] Kaida D, et al. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation.
10 *Nature* 2010; 468:664–8.
- 11 [37] Jenal M, et al. The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and
12 polyadenylation sites. *Cell* 2012; 149:538–53.
- 13 [38] Ryan K. Pre-mRNA 3' cleavage is reversibly inhibited in vitro by cleavage factor dephosphorylation.
14 *RNA Biol* 2007; 4:26–33.
- 15 [39] Shimazu T, Horinouchi S, Yoshida M. Multiple histone deacetylases and the CREB-binding protein
16 regulate pre-mRNA 3'-end processing. *J Biol Chem* 2007; 282:4470–8.
- 17 [40] Gunderson SI, Beyer K, Martin G, Keller W, Boelens WC, Mattaj LW. The human U1A snRNP
18 protein regulates polyadenylation via a direct interaction with poly(A) polymerase. *Cell* 1994;
19 76:531–41.
- 20 [41] Gunderson SI, Polycarpou-Schwarz M, Mattaj IW. U1 snRNP inhibits pre-mRNA polyadenylation
21 through a direct interaction between U1 70K and poly(A) polymerase. *Mol Cell* 1998; 1:255–64.
- 22
- 23

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

8 **Figure 2.** Changes in cleavage and polyadenylation site usage upon knock-down of CF I_m components and
9 of CstF-64 in HEK 293 cells. A total of 3,821 transcripts with 2, 3 or 4 tandem CP sites (inferred based on
10 the A-seq sequence data [17] and located in the same 3' UTR exon) whose expression was estimated to be
11 at least five A-seq tags per million in both untreated samples were selected. (A) Data sets were described
12 in [17]. An additional CF I_m 68 knock-down data set (marked by the asterisk) was generated in this study.
13 (B) Comparison of CP site usage in CF I_m 25 and CF I_m 59 knock-down sample relative to a control siRNA.
14 (C) Proximal shift in CP site usage under CF I_m 25 and CF I_m 68 knock-down conditions as a function of
15 the relative strength of the proximal CP site. Genes were divided into three subsets based on the ratio of
16 hexamer scores [17] of the most proximal and most distal CP sites. Within each subset, we computed the
17 proximal/distal CP usage ratio in a knock-down compared to the corresponding control siRNA-treated
18 sample. Box-plots summarize the distribution of proximal/distal CP usage ratio for all genes within a
19 particular subset and a particular sample. P-values of the t-test comparing the means of the two
20 distributions are shown above the box-plots. (D) Western blots showing the efficiency of CF I_m 25, CF
21 I_m 68 and CF I_m 59 knock-downs.

22 **Figure 3.** Possible models of modulation of alternative polyadenylation by CF I_m . (A) High concentration
23 of CF I_m 68 relative to CF I_m 59 leads to suppression of proximal CP sites. (B) Overall low levels of CF I_m
24 and hence low abundance of CF I_m 25/CF I_m 68 promote cleavage and polyadenylation at proximal sites.

1 (C) Post-translational modifications modulate RNA and protein interactions of CF I_m. (D) Co-
2 immunoprecipitation experiments with FLAG-CF I_m59 and FLAG-CF I_m68 indicate that the FLAG-
3 tagged CFI_m proteins can randomly bind both CF I_m59 and CF I_m68 (and in addition CF I_m72) and
4 possibly also form dimers of the Flag-tagged versions. Asterisks mark FLAG-tagged proteins.

A**B**