Advanced Review

Means to an end: mechanisms of alternative polyadenylation of messenger RNA precursors

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

Expression of mature messenger RNAs (mRNAs) requires appropriate transcription initiation and termination, as well as pre-mRNA processing by capping, splicing, cleavage, and polyadenylation. A core 3′-end processing complex carries out the cleavage and polyadenylation reactions, but many proteins have been implicated in the selection of polyadenylation sites among the multiple alternatives that eukaryotic genes typically have. In recent years, high-throughput approaches to map both the 3′-end processing sites as well as the binding sites of proteins that are involved in the selection of cleavage sites and in the processing reactions have been developed. Here, we review these approaches as well as the insights into the mechanisms of polyadenylation that emerged from genome-wide studies of polyadenylation across a range of cell types and states.

How to cite this article:
WIREs RNA 2014, 5:183–196. doi: 10.1002/wrna.1206

INTRODUCTION

All eukaryotic messenger RNAs (mRNAs) as well as many noncoding RNAs are synthesized by the nuclear DNA-dependent RNA polymerase II (Pol II), whose catalytic activity resides in Rpb1, the largest of its 12 subunits. The C-terminal domain (CTD) of Rpb1 coordinates most RNA processing events.[1] It not only recruits histone-modifying factors and chromatin remodeling complexes to assist the start of transcription but also, following controlled phosphorylation and dephosphorylation of specific serines or threonines in its heptad repeats, recruits capping, splicing, and 3′-end processing factors at different stages of the transcription cycle.[1–3] Thus, the maturation of pre-mRNAs to mRNAs occurs mostly cotranscriptionally by addition of a 7-methyl guanosine cap to the 5′ end, removal of intronic sequences by splicing, endonucleolytic cleavage, and polyadenylation. The site of pre-mRNA 3′ end cleavage is determined by the interaction of specific sequence elements within the pre-mRNA with a multiprotein complex whose core component is the cleavage and polyadenylation specificity factor (CPSF) composed of 160 (CPSF1), 100 (CPSF2), 73 (CPSF3), and 30 (CPSF4) kDa subunits, Fip1 (FIP1L1), and WDR33. Other members of the assembly are cleavage factors I (CF I$_m$), composed of 25 (CPSF5), 59 (CPSF7), and 68 (CPSF6) kDa proteins, and II (CF II$_m$), composed of Pcf11 and Clp1, as well as the cleavage stimulation factor (CstF), which consists of 50 (CSTF1), 64 (CSTF2 and CTSF2T), and 77 (CSTF3) kDa proteins (Figure 1; see also recent reviews[4,5]). Nuclear poly(A) polymerases α (PAPOLA), β (PAPOLB), or γ (PAPOLG) further add a poly(A) tail of up to 250 nucleotides, the precise length being determined by the nuclear

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: mihaela.zavolan@unibas.ch

Computational and Systems Biology, Biozentrum, University of Basel, Basel, Switzerland

Conflict of interest: The authors have declared no conflicts of interest for this article.

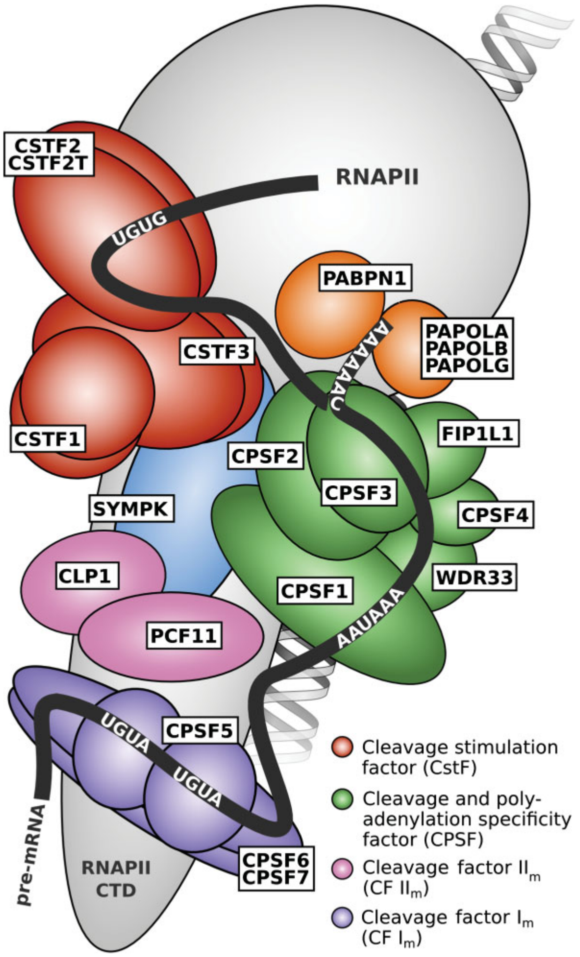
CSTF2
CSTF2T
RNAPII
UGUG
PABPN1
PAPOLA
PAPOLB
PAPOLG
AAAAAC
CSTF3
CSTF1
FIP1L1
CPSF2
CPSF3
CPSF4
SYMPK
CLP1
WDR33
CPSF1
AAUAAA
PCF11
UGUA
CPSF5
UGUA
pre-mRNA
RNAPII
CTD
CPSF6
CPSF7
● Cleavage stimulation factor (CstF)
● Cleavage and poly-adenylation specificity factor (CPSF)
● Cleavage factor II_m (CF II_m)
● Cleavage factor I_m (CF I_m)

FIGURE 1 | Composition of the human cleavage and polyadenylation complex. Different colors indicate individual protein subcomplexes. Components of the cleavage and polyadenylation specificity factor (CPSF) complex are depicted in close proximity to the cleavage and polyadenylation site, where CPSF1 recognizes the polyadenylation signal AAUAAA and CPSF3 is the endonuclease responsible for cleavage of the pre-messenger RNA (mRNA). CF I_m (cleavage factor) is depicted binding to UGUA motifs upstream of the cleavage site, while the cleavage stimulation factor (CstF) complex specifically interacts with a UG-rich region downstream of the cleavage site.

poly(A)-binding protein 1 (PABPN1).[6] Upon export of the mRNA from the nucleus PABPN1 is replaced by the cytoplasmic poly(A)-binding protein PABPC.[7] Its interaction with the translation initiation factor eIF4G at the cap complex leads to the formation of a pseudo-circular, translation-competent mRNA.

A few exceptions to the canonical 3′-end processing mechanism described above have been identified. For example, although transcribed by Pol II, replication-dependent histone mRNAs are not polyadenylated (with a few exceptions, reported in a recent study[8]). Instead, their pre-mRNAs contain a stem-loop element downstream of the stop codon, followed by a purine-rich histone downstream element (HDE).[9] Base pairing of the HDE with the U7 small nuclear RNA (snRNA), which is part of the Sm class U7 snRNP, and recognition of the stem-loop by the stem-loop-binding protein (SLBP) lead to the assembly of a complex containing a subset of proteins of the canonical pre-mRNA 3′-end processing apparatus, namely CPSF1, -2, -3, and -4 and Fip1, CstF-64 and -77, symplekin (SYMPK),[10] and CF $\text{I}_{\text{m.}}$[11] Cleavage occurs at a CA dinucleotide between the stem-loop and the HDE. Interaction between the SLBP and eIF4G then leads to the formation of the pseudo-circular, translation-competent form of the mRNA. Some long noncoding RNAs such as the metastasis-associated lung adenocarcinoma transcript 1 (MALAT1)[12] and the multiple endocrine neoplasia 1 (MEN1-ϵ/β)[13] appear to be processed by yet another alternative mechanism, ribonuclease P (RNase P).

While pre-mRNA polyadenylation takes place in the nucleus, cytoplasmic deadenylation and readenylation of mature mRNAs has also been observed (see also recent reviews[14,15]), initially in *Xenopus* oocytes, where dormant mRNAs containing short oligo(A) tails of 20–40 nucleotides are reactivated for translation by addition of long poly(A) tails during oocyte maturation. The cytoplasmic polyadenylation element (CPE, consensus UUUUUAU) located in the 3′ untranslated regions (UTRs) of these mRNAs binds the CPE-binding protein (CPEB), which in turn interacts with a poly(A) ribonuclease (PARN) and the cytoplasmic poly(A) polymerase GLD-2. The composition of this complex, which is modulated in response to signals, results in either short poly(A) tails and no translation or long poly(A) tails and protein production.[16] CPEB was also detected in postsynaptic structures. Its phosphorylation after calcium entry into the synapse leads to polyadenylation and subsequent translation of CPE-containing RNAs. The resulting proteins act as tags, marking experienced synapses and providing a cellular basis for learning and memory.[16]

A multitude of factors is involved in the selection of the 3′-end processing site among the many poly(A) sites that a gene typically has.[17,18] With the advent of high-throughput sequencing methods, transcriptome-wide polyadenylation sites have been determined in a variety of conditions to reveal a very dynamic landscape and systematic changes in poly(A) site use that point to yet unidentified global regulators. Four types of alternative polyadenylation (APA) patterns are generally distinguished (Figure 2). They either only modulate the length of the 3′ UTR or result in distinct protein isoforms. APA at coding region-proximal poly(A) sites has been observed in cellular states associated with increased proliferation

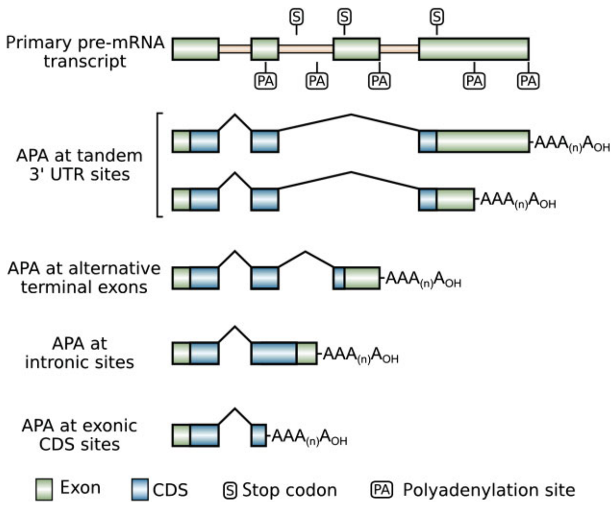

FIGURE 2 | Outline of the main alternative polyadenylation (APA) patterns. One of the most studied patterns, tandem poly(A) sites, corresponds to multiple poly(A) sites being located in the 3′ untranslated region (UTR) of the terminal exon. Cleavage and polyadenylation at any of these sites will only lead to transcript isoforms that differ in the length of the 3′ UTR, but will not affect the protein-coding region of the messenger RNA (mRNA). Although referred to as an APA event, cleavage and polyadenylation at a different terminal exon is rather governed by alternative splicing decisions than APA. APA at cryptic poly(A) sites located in introns or exons can lead to truncated transcript isoforms with an altered coding sequence (CDS).

(e.g. cancer cells), where the short 3′ UTRs, devoid of microRNA-binding sites, have been associated with an increased protein output.[19–21] Here, we summarize the insights into 3′-end processing that emerged from recent, high-throughput experimental and computational studies. The molecular mechanism of 3′-end processing and its regulation and relationship with other cellular processes have been covered in a few recent reviews.[4,22–24]

PRE-mRNA 3′-END PROCESSING THROUGH THE LENS OF HIGH-THROUGHPUT EXPERIMENTS

Approaches to the Genome-Wide Mapping of Poly(A) Sites

The accumulation of substantial numbers of cDNA and EST sequences in public sequence repositories such as Genbank[25] allowed the construction of genome-wide maps of poly(A) sites.[26,27] These data then enabled inferences on global trends, such as the preferential use of distal poly(A) sites in cells from the nervous system compared to those from blood.[28] The most recent release (version 2) of the polyA_DB database of 3′-end processing sites contains more than 54,000 poly(A) sites mapped to the human genome.[27] In an alternative approach, researchers took advantage of gene expression microarrays to estimate the signal intensities of probes mapping to alternatively processed 3′ UTRs of mRNAs and thereby quantify poly(A) site use under different experimental conditions.[19,20] These studies led to the striking observation that dividing cells systematically express mRNAs with shortened 3′ UTRs compared to resting cells, prompting a flurry of investigations into the underlying mechanisms. Several laboratories then developed 3′-end sequencing protocols, which simultaneously allow the mapping and quantification of poly(A) site use on a genome-wide scale (see Refs 29 and 30 for a detailed comparison of the methods). Currently, more than 4.5 billion reads, generated with 14 different protocols, can be retrieved from public data repositories such as NCBI's Gene Expression Omnibus (GEO).[31]

The bulk of the data was contributed by methods, such as PAS-Seq,[8] PolyA-seq,[17] A-seq,[32] or 3′-seq,[33] that rely on reverse transcription with an oligo(dT) primer. These methods differ in the length of the oligo(dT) primer, the way second-strand synthesis is accomplished, and from which strand of the cDNA molecule the sequence is read (Figure 3). PolyA-seq and PAS-seq libraries can be generated with relative ease but custom sequencing primers need to be used to avoid sequencing through long oligo(T) tracts. On the other hand, A-seq and 3′-seq sequence in the sense direction, but to capture the beginning of the poly(A) tail, which allows identification of the cleavage site, they require a precise size selection of the RNA fragments. The main complication with oligo(dT)-primed libraries is that annealing of the primer at poly(A)-rich sequences that are internal to the mRNAs can yield false-positive 3′-end processing sites. A typical solution is to discard putative poly(A) sites that are followed by genome-encoded poly(A) stretches during computational analysis. Alternatively, long oligo(dT) primers (e.g., 45 Ts in 3′READS[18]) are annealed to mRNA under stringent hybridization conditions, thus preventing priming at shorter internal A-rich stretches. Finally, the 3P-seq[34] method has been designed to capture specifically the poly(A) tails of mRNAs through an initial ligation to the intact 3′ ends of polyadenylated transcripts. This protocol identifies true poly(A) sites, but has the drawback that it is lengthier and more complex. A direct RNA sequencing method in which poly(A)-containing RNA molecules are hybridized to poly(dT)-coated flow cell surfaces where antisense strand synthesis is initiated has also been used to map poly(A) sites.[35,36] This has the advantage that no prior reverse transcription or cDNA amplification is needed, but on the other

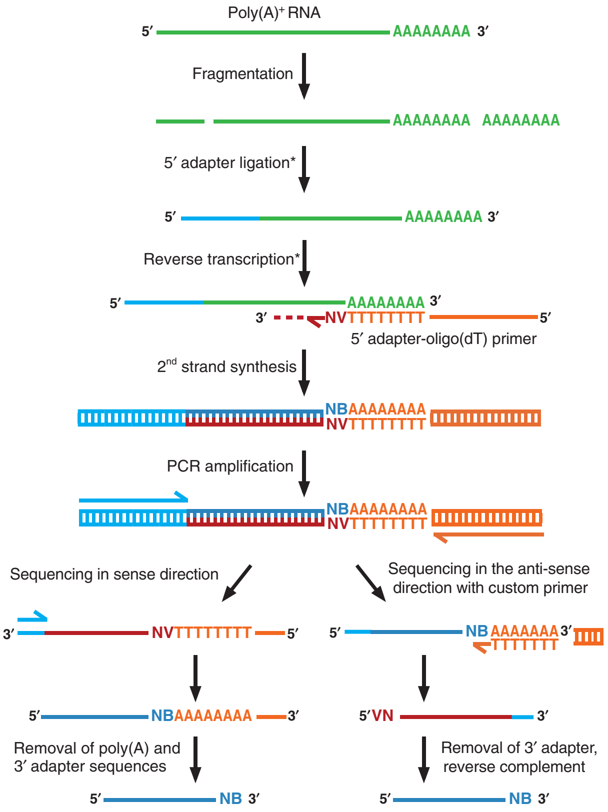

FIGURE 3 | General outline of oligo(dT)-based 3′-end sequencing protocols (e.g., A-seq,[32] PAS-seq,[8] 3′-seq,[33] and PolyA-seq[17]). Poly(A)$^+$ RNA is usually isolated with oligo(dT)-coated beads, fragmented by alkaline hydrolysis, ribonuclease (RNase) treatment, or sonication, and oligo(dT)-adapter primers are used to reverse transcribe the RNA. Second-strand synthesis is accomplished with primers complementary to 5′ adapters, random hexamer-adapter primers, or by the Eberwine method (SMARTer kit by Clontech) where the reverse transcriptase (RT) adds a CCC tag to the cDNA that can be primed by an adapter-GGG molecule leading to a template switch. 5′ Adapter ligation can be omitted when the template switch method is used or second-strand synthesis after RT is performed with hexamer-5′ adapter primers (*). N is any nucleotide, B is any but A, and V is any but T.

hand it requires specialized instruments that are not widely accessible. Current estimates of the number of poly(A) sites in the human genome, based on data of different sequencing depths and somewhat different computational analyses, range from 280,000[17] to 1,287,130,[36] with up to 58% of human genes having multiple poly(A) sites. Given the limited accuracy of current 3′ UTR annotations, this latter number may be an underestimate.[37]

In addition to cataloging poly(A) sites, high-throughput studies have also attempted to quantify their relative use across tissues,[17] cell lines,[8,36] developmental stages,[38,39] during cell differentiation,[18] or following the knockdown of a specific factor.[32,33,40–42] An aspect that became apparent from differential analysis of poly(A) site use in nuclear and cytoplasmic RNA fractions is that promoter-proximal poly(A) sites are preferentially used in the cytoplasmic fraction.[43] This implies that the relative stability of mRNA isoforms affects the estimation of APA site use and that the frequency of polyadenylation at distal sites may be underestimated, presumably to different extents depending on the cytoplasmic-to-nuclear RNA ratio of specific samples.

Approaches to the Identification of Binding Sites of RNA-Binding Proteins

Many of the factors that are involved in pre-mRNA 3′-end processing have been extensively studied and their binding specificities are known.[44–49] However, the discovery of systematic, condition-dependent changes in polyadenylation at the transcriptome level points toward yet uncharacterized regulatory interactions. A powerful method to map the sites of interaction of RNA-binding proteins (RBPs) in RNAs at close to nucleotide resolution consists of cross-linking and immunoprecipitation (CLIP) of proteins of interest followed by deep sequencing of the protein-bound RNA fragments. Since the method was introduced by Ule and Darnell,[50,51] a number of variants have emerged (see Figure 4 for a sketch). To cross-link RBPs to their RNA targets UV-C light (254 nm) is used in HITS-CLIP[52] and iCLIP[53] and UV-A (365 nm), after incorporation of photoreactive 4-thiouridine into the RNAs, in PAR-CLIP.[54] The nucleotide-level resolution of the methods stems from the propensity of the reverse transcriptase (RT) to stop at the cross-linked nucleotide, which presumably still carries a peptide stub that fails to be removed by proteolytic treatment. It has been estimated that 80% of the RT reactions generate such truncated products that are specifically captured by RNA circularization in iCLIP.[53] When the reverse transcription does continue through the cross-linked nucleotide, frequent skipping or misrecognition of this nucleotide leads to cross-link diagnostic mutations that are exploited in PAR-CLIP and HITS-CLIP. CLIP methods have been reviewed recently[55] and a comparative assessment of their accuracy has been provided in two recent studies.[56,57] The specificity of the antibodies is a limiting factor and identification of *bona fide* binding sites from a large pool of unspecifically captured and amplified RNAs remains challenging, especially when the protein has a relatively small number of binding sites.

To identify the RBP-binding sites, computational methods that take advantage of the cross-link diagnostic mutations in PAR-CLIP have been proposed.[54,58–61] These methods can also be applied to HITS-CLIP, taking advantage of the mutations,

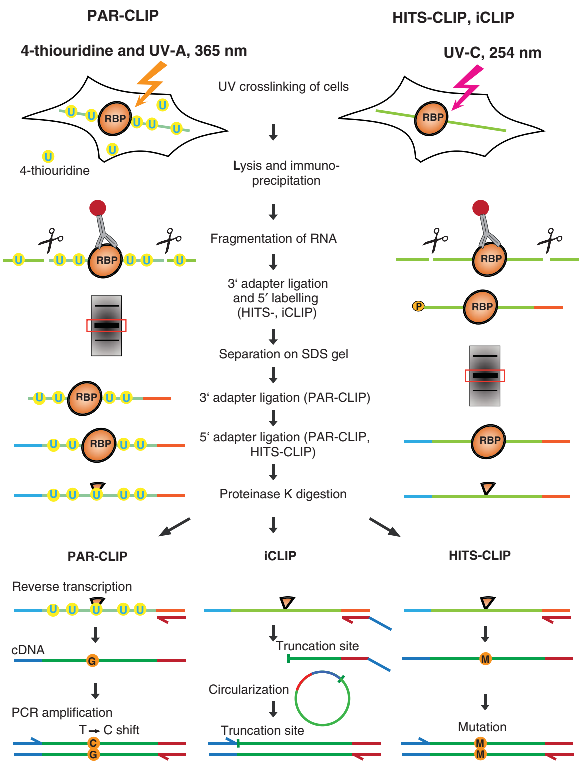

FIGURE 4 | Schematic outline of cross-linking and immunoprecipitation (CLIP) protocols for inferring protein interactions sites in RNAs. Many steps are interchangeable between protocols. Blotting after the sodium dodecyl sulfate (SDS) gel electrophoresis is frequently used to remove contaminating RNAs that are not cross-linked to proteins. Diagnostic mutations (substitutions, deletions, or insertions in all protocols, T→C mutations specifically in PAR-CLIP) are indicated.

deletions, or insertions that are introduced, albeit with much lower frequency.[56,62] It should be noted, however, that the frequency of cross-link diagnostic mutations in PAR-CLIP does not simply reflect the residence time of the RBP on individual sites. 4-Thiouridine is randomly incorporated in the RNA and its cross-linking to the RBP depends on its occurrence in a favorable configuration in the RBP-binding site. Thus, the frequency of cross-link diagnostic mutations does not need to be strongly correlated with the affinity of interaction between the RBP and the binding site. Indeed, some evidence suggests that a better indicator of the site's affinity for the protein is the enrichment of RNA fragments originating from putative binding sites relative to the overall transcript expression.[56]

Sequence Elements That Direct 3′-End Processing

Biochemical and computational analyses of a restricted number of genes have already yielded the core set of sequence motifs that are recognized by various 3′-end processing factors,[63,64] and thus the more recent transcriptome-wide analyses of poly(A) sites did not identify strikingly novel elements.[17,65] As summarized in Figure 5(a), the frequency of adenosine nucleotides is high in the region upstream of the cleavage site, with a peak at approximately −21 nucleotides (nt). The peak in A's is followed by a U-rich stretch close to the site of cleavage, which in human most often is between a C and an A nucleotide. A peak of G nucleotides follows immediately downstream of the cleavage site, followed by a peak of U's at approximately +25 nt. The sequence motif most reproducibly found at poly(A) sites is the A-rich hexamer polyadenylation signal, which has the canonical form AAUAAA (Figure 5(b)) and has been shown to be bound by the CPSF1 3′-end processing factor.[44] Slight variations are tolerated[64] and the frequency with which these polyadenylation signals appear upstream of 3′-end processing sites roughly corresponds to their *in vitro* determined polyadenylation efficiency[63] (Figure 5(b)). However, at the level of individual genes, point mutations in the polyadenylation signal can lead to altered relative expression of transcript isoforms[66] and ultimately to genetic diseases.[67] Although most conserved across genes, the hexamer polyadenylation signal may also be dispensable. This is indicated both by a study in which CF I_m was sufficient to direct sequence-specific, AAUAAA-independent poly(A) addition *in vitro*[68] as well as by a recent report that an A-rich upstream sequence combined with potent downstream signals is sufficient to direct cleavage and polyadenylation.[69]

Transcriptome-Wide Mapping of Binding Sites of Core 3′-End Processing Factors

Although application of CLIP approaches to 3′-end processing factors[32,70] largely confirmed the sequence specificities inferred with biochemical methods, it further revealed that the subunits of the CstF that binds in the U/G-rich region 10–30 nt downstream of the cleavage site exhibited the strongest positional preference. Binding of cleavage factor I (CF I_m) occurred within the −100 to −30 nucleotide regions upstream of the cleavage site, in a region typically containing UGUA motifs, which are recognized by the CF I_m 25 (CPSF5) subunit of the complex (Figure 6). Surprisingly, the CLIP data indicated that the positioning of CF I_m is similar on RNAs that lack UGUA motifs, pointing toward yet unknown recruitment mechanisms.[32,70]

Unexpectedly, CPSF and especially its largest subunit CPSF1, which in a previous study was

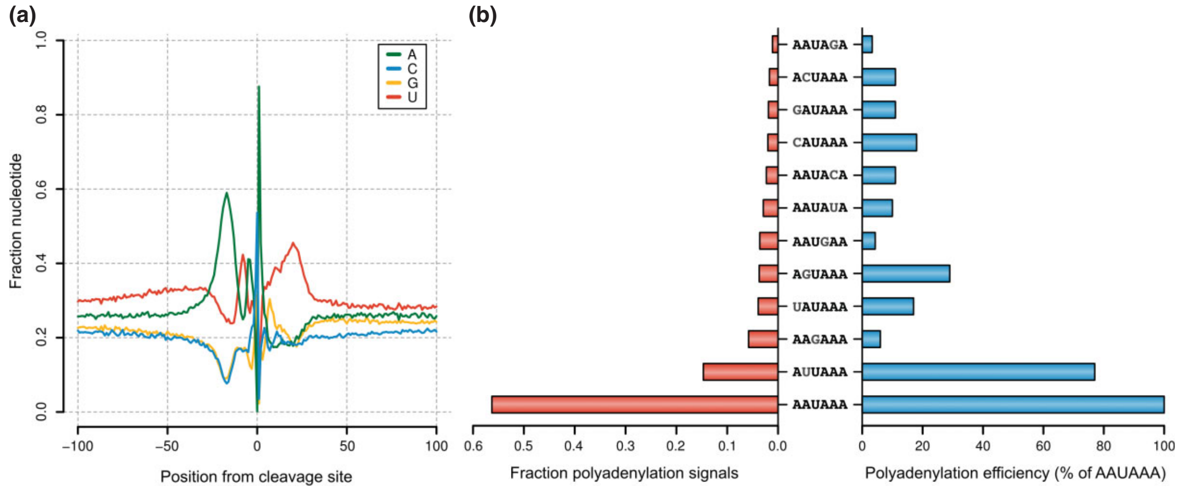

FIGURE 5 | Sequence composition around poly(A) sites. Poly(A) sites were determined based on publicly available 3′-end sequencing data (NCBI GEO entry GSE30198[17]), which we processed as described previously.[32] (a) Position-dependent mononucleotide frequencies around the 10,000 poly(A) sites most frequently used in human cells. (b) Comparison of the frequency of occurrence of hexameric motifs at the same human poly(A) sites and their *in vitro* measured efficiency in polyadenylation.[63]

shown to bind the conserved polyadenylation signal AAUAAA,[44] did not exhibit a strong positional preference with respect to the cleavage site (Figure 6, middle panel). Although some mechanistic hypotheses were proposed,[32] the reason for this discrepancy remains to be determined.

DYNAMIC MODULATION OF POLY(A) SITE USE

Systematic Changes in Poly(A) Site Use in Physiological Conditions

One of the most surprising recent findings has been the preferential use of proximal poly(A) sites in dividing cells.[19] Eighty-seven percent of the poly(A) sites that showed a significant change in use in dividing compared to resting cells were located proximal to coding regions. Other cellular states associated with increased proliferation such as malignancy show a similar pattern of APA.[21,36] Because proximal sites are typically ‘weaker’ than distal sites,[32] the simplest models that would explain these observations are that either (1) specific factors are recruited at the ‘weak’ sites to promote their use or (2) the core factors have a decreased specificity in their recognition of poly(A) sites in dividing cells. This can be caused, for example, by an increase in the abundance of these factors. Both of these models would require an increased expression of the relevant proteins in proliferating compared to resting cells. Indeed, across the samples of the human gene expression atlas,[71] the mRNA expression level of many factors that have been implicated in the regulation of 3′ UTR length is positively correlated with the proliferative potential of cells (Figure 7). Surprisingly, however, the same pattern of increased use of proximal poly(A) sites was also observed in studies in which the expression of individual 3′-end processing factors was *reduced* by siRNA-mediated knockdown.[32,33,40,72] Thus, the molecular mechanisms underlying systematic 3′ UTR shortening associated with proliferative states remain to be uncovered.

The factor whose impact was studied in most detail is the U1 snRNP, which, although necessary for splicing in equimolar concentration to the other snRNPs, is much more abundant within cells. The Dreyfuss group demonstrated that the marked knockdown of U1 snRNA leads to premature cleavage and polyadenylation at cryptic poly(A) sites located close (<5 kb) to the transcription start site,[73] whereas a moderate knockdown leads to various types of transcript shortening, from alternative splicing of 3′ terminal exons located proximally to the transcription start sites to shortened 3′ UTRs.[74] Interestingly, a study from the same group showed that 3′ UTR shortening could be a consequence of transiently limiting U1 snRNA levels.[74] Specifically, it was found that activation of transcription upon neural activation leads to a spike in the ‘RNA load’ that is not matched by a corresponding spike in U1 snRNA levels, resulting in an effective knockdown of the U1 snRNA and premature polyadenylation. 3′-End processing factors may similarly become limiting in conditions associated with increased cell proliferation. The U1 snRNP and polyadenylation events have also been implicated recently in the control of promoter directionality.[41,75]

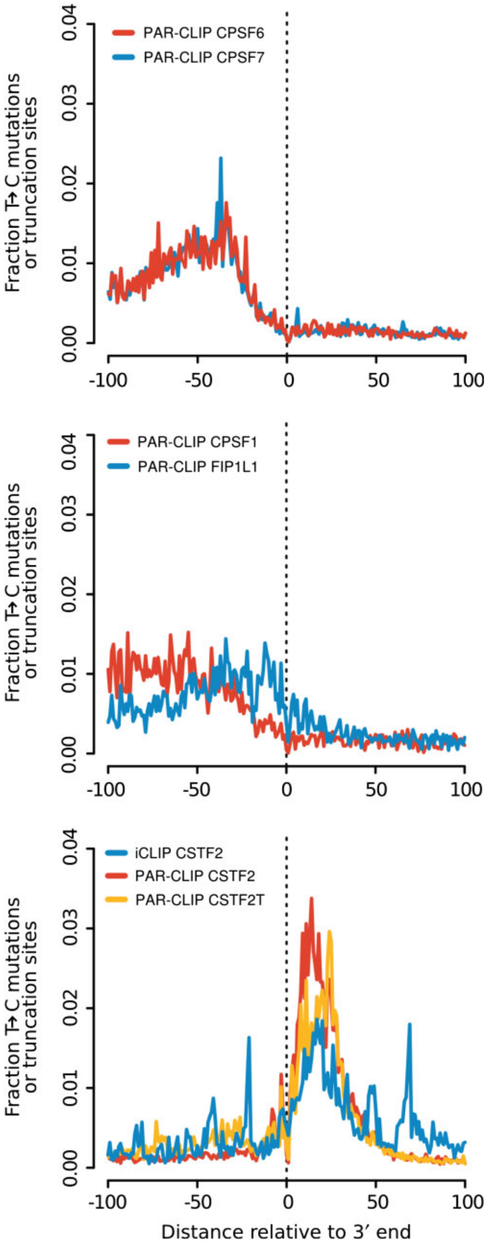

FIGURE 6 | Positional preferences of 3′-end processing subcomplexes. Profiles show the densities of T → C mutations (PAR-CLIP[32]) or reverse transcriptase (RT) truncation sites (iCLIP[70]) obtained in various cross-linking and immunoprecipitation (CLIP) experiments, relative to the 1000 most abundantly used 3′-end processing sites in the human genome.[17]

Although once recruited to a transcription start site (TSS) Pol II can initiate transcription in either direction, antisense transcripts generally terminate within 1 kb of the TSS, probably through a canonical termination mechanism that involves the typical polyadenylation signal. These short antisense transcripts are then degraded by the exosome.[75] The frequency of occurrence of the poly(A) signal is increased immediately downstream of the TSS in the antisense direction compared to the sense direction. Conversely, the U1 snRNP-binding motif is strongly enriched in the sense direction, antisense morpholino-mediated U1 snRNA knockdown causing a marked increase in premature termination of sense transcripts with only a small effect on antisense transcripts.[41]

The 25- and 68-kDa components of CF I_m (CPSF5 and CPSF6) have also been reported to lead to shortened 3′ UTRs upon knockdown,[32,40] and consistently, the depletion of Thoc5, which is presumably involved in the recruitment of CPSF6 to the polymerase at transcription start sites, resulted in the same phenotype.[76] In contrast to U1 snRNA, however, knockdown of CPSF5 and CPSF6 did not increase the use of intronic poly(A) sites. Interestingly, the expression of CPSF5 and CPSF6 most closely tracks the proliferative potential (Figure 7), indicating that cells are highly sensitive to the level of these factors and that it would be very informative to obtain detailed measurements of the concentrations of regulatory factors in relation to the RNA load in various cell states. Alternatively, it may be the post-translational modifications or the composition of CF I_m components that contribute to poly(A) site selection. For example, phosphorylation of Ser166 in the RRM of CPSF6 modulates its RNA-binding affinity.[77] The precise composition of the CF I_m tetramer composed of CPSF5 and CPSF6 and/or CPSF7[40] may be another factor that influences the choice of poly(A) sites. Similarly, the competition between hnRNPK and CPSF6 in binding to CPSF5 has recently been found to determine the choice of poly(A) site in the lncRNA NEAT1.[78]

3′ UTR shortening can also be brought about by the knockdown of the PABPN1 component of the 3′-end processing machinery,[33,72] which has so far been known to control polymerization of the poly(A) tail. The proposed model is that PABPN1 masks weak poly(A) sites that are more readily recognized by CPSF upon depletion of PABPN1.[33] Finally, cold-induced and circadian clock-regulated RBPs Cirbp and Rbm3 also induce APA, binding between tandem poly(A) sites to mask the proximal and promote the use of the distal poly(A) site.[79] In this case, however, the use of distal poly(A) sites is accompanied by an induction of cell proliferation, at least in immature germ cells in mice.[80]

Coupling Between Polyadenylation and Other Steps of Gene Expression

The effects of kinetic parameters such as the rate of Pol II-dependent transcription on the structure of mature

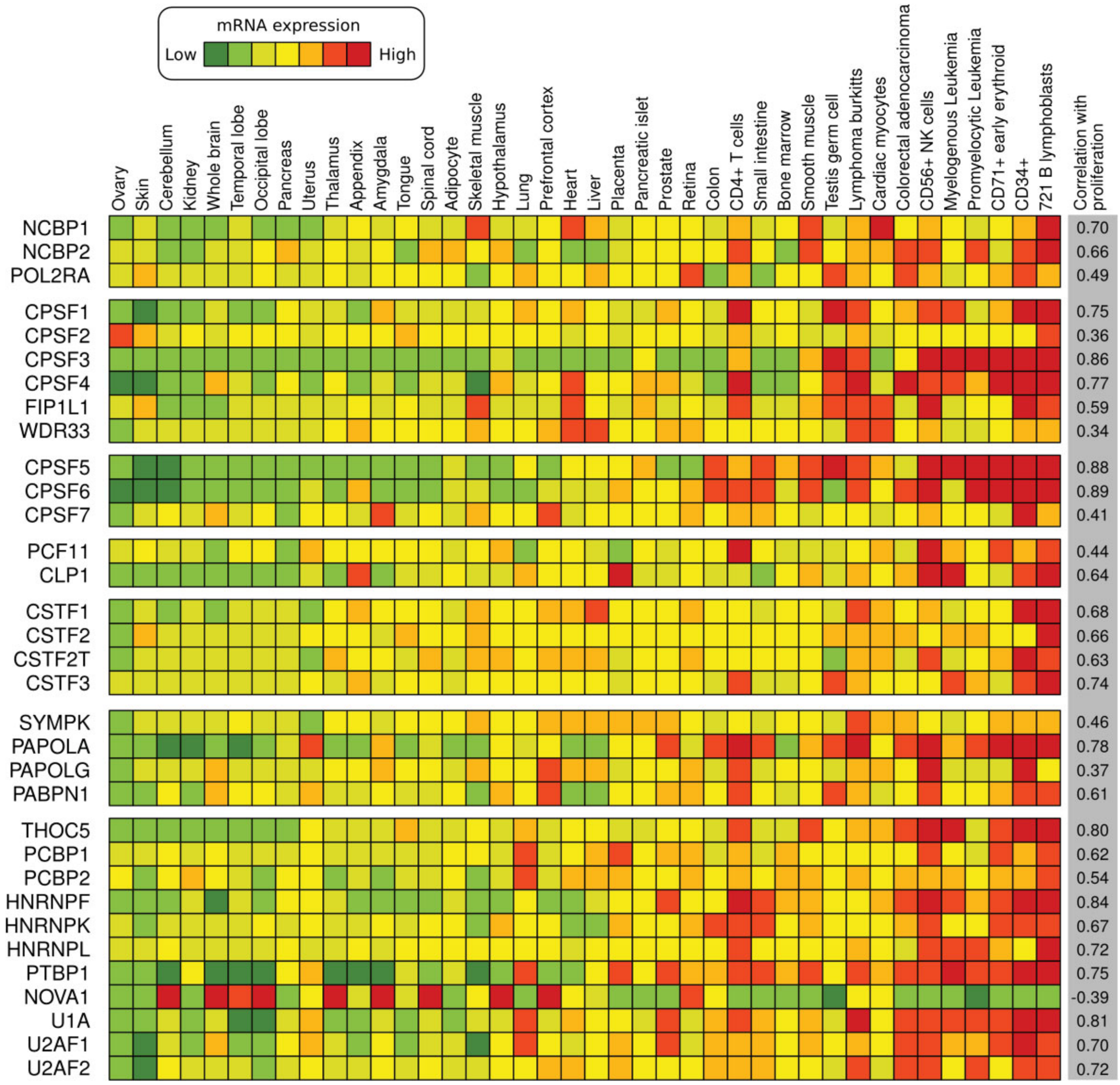

FIGURE 7 | Expression profiles of core and modulatory 3′-end processing factors in human tissues. The tissues are sorted from left to right in the order of increasing proliferation index (defined as in Ref 40). Expression data[71] were obtained from BioGPS (http://biogps.org) and processed as described in the online Supporting Information. The numbers on the right side of each line represent the Spearman correlation coefficient between the expression levels of the indicated gene and the proliferative potential estimated from individual samples.

RNAs are only starting to emerge. Highly transcribed genes tend to be processed at proximal poly(A) sites and lowly transcribed genes at distal sites.[81] A possible mechanistic model is suggested by the study of Nagaike et al.,[82] who found that strong transcriptional activators recruit the PAF1c component of the transcription elongation complex to the promoter, PAF1c recruiting the 3′-end processing complex and promoting polyadenylation at proximal sites. Such sites are overlooked in genes whose promoters lack strong transcriptional activation elements.[82] Similar Pol II rate-dependent effects have also been described for splicing, where slow transcription elongation (window of opportunity model[83]) is believed to allow the assembly of spliceosomal complexes at exons with 'weak' splicing signals and their subsequent inclusion in the mature mRNAs.[84]

It is believed that nucleosome occupancy along the gene can modulate the Pol II elongation rate and thereby affect poly(A) site choice. The region immediately flanking poly(A) sites has been reported to be depleted of nucleosomes,[81,85–87] which may be explained in part by the AT-rich sequence that is resistant to curvature.[85] However, regions further downstream of the cleavage sites have higher nucleosome occupancy at frequently used poly(A) sites compared to those that are infrequently used.[85,86] Differences in histone modification around these two

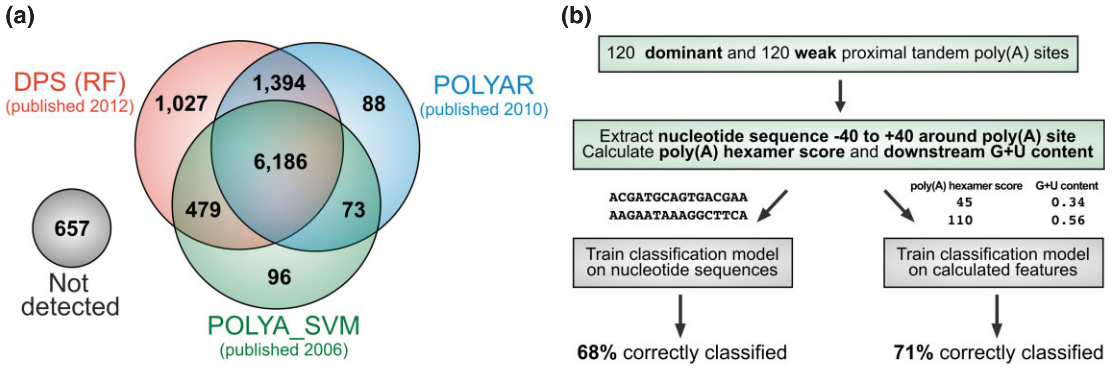
(a)
DPS (RF)
(published 2012)
POLYAR
(published 2010)
1,027
1,394
88
6,186
479
73
96
657
Not detected
POLYA_SVM
(published 2006)
(b)
120 dominant and 120 weak proximal tandem poly(A) sites
Extract nucleotide sequence -40 to +40 around poly(A) site
Calculate poly(A) hexamer score and downstream G+U content
ACGATGCAGTGACGAA
AAGAATAAAGGCTTCA
poly(A) hexamer score
G+U content
45
0.34
110
0.56
Train classification model on nucleotide sequences
Train classification model on calculated features
68% correctly classified
71% correctly classified

FIGURE 8 | Evaluation of computational poly(A) site prediction tools. (a) Prediction of poly(A) sites: the 10,000 most frequently processed 3′ ends of human genes[17] were used as the positive set and mononucleotide randomized variants of these sequences were used as the negative set to test the ability of POLYA_SVM,[88] POLYAR,[89] and Dragon PolyA spotter[90] (DPS). Sequences and program outputs are available online as Supporting Information. (b) Prediction of relative use of tandem poly(A) sites in the human brain.[17] We trained support vector classification models using either a string kernel on the nucleotide sequence at positions −40 to +40 around the poly(A) site or a RBF kernel using the poly(A) hexamer score and the G + U content in the 40 nt window downstream of the poly(A) site as input. Reported values are averaged accuracy values from a fourfold cross-validation.

types of poly(A) sites have also been reported,[86] but it remains to be determined whether these chromatin marks are established to guide 3′-end processing as opposed to being triggered by the 3′-end cleavage process itself.

PREDICTION OF POLY(A) SITES

Sequence-Based Computational Prediction of Poly(A) Sites and Relative Poly(A) Site Usage

Several methods that take advantage of local sequence composition biases to predict poly(A) sites have been proposed.[88–90] By evaluating their accuracy in predicting the 10,000 most frequently used human poly(A) sites relative to randomized sequences (Figure 8(a)), we found a relatively good performance, with 6,186 of the 10,000 poly(A) sites being predicted by all three methods. The most recently published method, Dragon PolyA spotter,[90] performs best, identifying about 90% of the genuine poly(A) sites at a false-positive rate of 19%. The availability of large data sets of binding sites of RBP modulators of 3′-end processing[32,70] will help to further improve prediction accuracy. Predicting the relative use of alternative poly(A) sites of individual genes, however, is more challenging. On the basis of a recently published data set of APA in the human brain,[17] we evaluated the ability of standard machine learning algorithms to predict dominant or weak use of the proximal site (defined as at least 75% and less than 25%, respectively, of all reads mapped to poly(A) sites associated with the gene being assigned to the proximal site) in genes with tandem poly(A) sites located in the same terminal exon. As input to the algorithm we either used the nucleotide sequence around the cleavage site (−40 to +40 nt) or the combination of the poly(A) hexamer score (defined as the sum over all hexamers detected in the 40 nt window upstream of the poly(A) site of the *in vitro* polyadenylation efficiency weighted by the distance of the hexamer to the cleavage site[32]), and the G + U content in the 40 nt window downstream of the cleavage site. We found that both approaches have limited accuracy, achieving at most 71% correctly classified instances (Figure 8(b)). This may indicate that poly(A) site selection not only depends on sequence motifs located in close proximity to the cleavage site, but that motifs that are located further away and bind auxiliary factors with tissue- or condition-specific expression,[42,91] RNA secondary structure,[92] and chromatin marks[86] also contribute significantly to the selection process.

Understanding the functional impact of protein–RNA interactions is challenging, because many proteins interact with pre-mRNAs and can modulate their processing.[93] Identifying these interactions with CLIP, which is applied to one protein and one particular condition, is very time consuming. An interesting alternative is now taking shape with the availability of large collections of binding motifs of RBPs that were determined *in vitro*.[94] These could be used in an approach that was already developed to identify key transcription regulatory interactions.[95,96] Namely, RBP-binding sites could be predicted with methods based on comparative genomics, and the number and quality of RBP-binding sites could be related to the use of 3′-end processing sites transcriptome-wide to identify the regulators that have high activity in the choice of poly(A) sites in specific states or conditions. Similarly, other types of modulation, for example, via the rate of transcription or the density of various chromatin marks can be included as the data become available.

CONCLUSION

The proper generation of mRNA 3′ ends requires the recognition of sequence elements in the pre-mRNAs by the cognate protein factors. Recently developed high-throughput methods enabled the mapping of both RBP–RNA interaction sites as well as of the 3′ ends that are used in specific cell types in specific conditions. Although substantial 'static' information has been gathered, it remains nontrivial to predict sites of pre-mRNA processing and their quantitative usage under specific conditions for a number of reasons. For example, the recruitment of many splicing and 3′-end processing factors occurs already at the transcription start site, through the interaction with the Pol II CTD, which in turn depends on post-translational modifications of the CTD such as phosphorylation and methylation. Furthermore, much of the RNA processing occurs cotranscriptionally, putative poly(A) sites emerging sequentially from the RNA polymerase. Thus, the efficiency of processing of individual poly(A) sites is a reflection of not only the relative affinity of the 3′-end processing complexes for the RNA but also of the rates at which various steps of RNA processing proceed. Consequently, a model that satisfactorily explains the experimental data is missing.

The fact that knockdown of the U1 snRNA and of the core components of the 3′-end processing machinery almost always leads to the more frequent use of promoter proximal poly(A) sites suggests that several safeguard mechanisms operate to prevent premature cleavage and polyadenylation. Indeed, premature cleavage and polyadenylation sites are effectively and reproducibly used when the expression of individual factors is reduced, suggesting that safeguard mechanisms suppress the use of promoter-proximal sites, which emerge first from the RNA polymerase, rather than actively promote the use of distal sites. That dividing cells, which have a high expression of 3′-end processing factors, express short 3′ UTRs is in apparent contradiction with the similar phenotype caused by the siRNA-mediated reduction in the expression of these factors. The argument that has been made, namely that the 'load' of RNA to be processed changes as a function of the cell's state leading to an imbalance between the number of targets and the number of processing complexes, suggests a very promising avenue of future research.

Systematic changes in 3′-end processing in specific conditions such as during the cell cycle, in proliferating compared to resting cells, and during development have been uncovered in a variety of studies. Although in some circumstances shorter 3′ UTRs have been associated with increased protein output, it remains to be determined how general this relationship is, because 3′ UTRs harbor not only destabilizing but also stabilizing elements. Furthermore, additional work is necessary to quantify how large the contribution of APA to the protein output is, relative to other regulatory mechanisms such as condition-dependent transcription. In the coming years, we therefore expect a very active 3′-end processing field.

ACKNOWLEDGMENTS

The work was supported by the University of Basel, the Louis-Jeantet Foundation for Medicine, and the Swiss National Science Foundation (grant no. 31003A-143977 to W.K.).

REFERENCES

1. Hsin J-P, Manley JL. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev* 2012, 26:2119–2137.
2. Proudfoot N. New perspectives on connecting messenger RNA 3′ end formation to transcription. *Curr Opin Cell Biol* 2004, 16:272–278.
3. Maniatis T, Reed R. An extensive network of coupling among gene expression machines. *Nature* 2002, 416:499–506.
4. Danckwardt S, Hentze MW, Kulozik AE. 3′ end mRNA processing: molecular mechanisms and implications for health and disease. *EMBO J* 2008, 27:482–498.

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

6. Wahle E. A novel poly(A)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation. *Cell* 1991, 66:759–768.

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

8. Shepard PJ, Choi E-A, Lu J, Flanagan LA, Hertel KJ, Shi Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* 2011, 17:761–772.

9. Marzluff WF. Metazoan replication-dependent histone mRNAs: a distinct set of RNA polymerase II transcripts. *Curr Opin Cell Biol* 2005, 17:274–280.

10. Kolev NG, Steitz JA. Symplekin and multiple other polyadenylation factors participate in 3′-end maturation of histone mRNAs. *Genes Dev* 2005, 19:2583–2592.

11. Ruepp M-D, Vivarelli S, Pillai RS, Kleinschmidt N, Azzouz TN, Barabino SML, Schümperli D. The 68 kDa subunit of mammalian cleavage factor I interacts with the U7 small nuclear ribonucleoprotein and participates in 3′-end processing of animal histone mRNAs. *Nucleic Acids Res* 2010, 38:7637–7650.

12. Wilusz JE, Freier SM, Spector DL. 3′ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* 2008, 135:919–932.

13. Sunwoo H, Dinger ME, Wilusz JE, Amaral PP, Mattick JS, Spector DL. MEN ϵ/β nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res* 2009, 19:347–359.

14. Richter JD. CPEB: a life in translation. *Trends Biochem Sci* 2007, 32:279–285.

15. Charlesworth A, Meijer HA, de Moor CH. Specificity factors in cytoplasmic polyadenylation. *Wiley Interdiscip Rev RNA* 2013, 4:437–461.

16. Darnell JC, Richter JD. Cytoplasmic RNA-binding proteins and the control of complex brain function. *Cold Spring Harb Perspect Biol* 2012, 4:a012344.

17. Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. A quantitative atlas of polyadenylation in five mammals. *Genome Res* 2012, 22:1173–1183.

18. Hoque M, Ji Z, Zheng D, Luo W, Li W, You B, Park JY, Yehia G, Tian B. Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nat Methods* 2013, 10:133–139.

19. 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, 320:1643–1647.

20. 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, 106:7028–7033.

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

22. Mandel CR, Bai Y, Tong L. Protein factors in pre-mRNA 3′-end processing. *Cell Mol Life Sci* 2008, 65:1099–1122.

23. Shi Y, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, Yates JR III, Frank J, Manley JL. Molecular architecture of the human pre-mRNA 3′ processing complex. *Mol Cell* 2009, 33:365–376.

24. Millevoi S, Decorsière A, Loulergue C, Iacovoni J, Bernat S, Antoniou M, Vagner S. A physical and functional link between splicing factors promotes pre-mRNA 3′ end processing. *Nucleic Acids Res* 2009, 37:4672–4683.

25. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res* 2013, 41:D36–D42.

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

27. Lee JY, Yeh I, Park JY, Tian B. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res* 2007, 35:D165–D168.

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

29. Mueller AA, Cheung TH, Rando TA. All's well that ends well: alternative polyadenylation and its implications for stem cell biology. *Curr Opin Cell Biol* 2013, 25:222–232.

30. Sun Y, Fu Y, Li Y, Xu A. Genome-wide alternative polyadenylation in animals: insights from high-throughput technologies. *J Mol Cell Biol* 2012, 4: 352–361.

31. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2013, 41:D991–D995.

32. 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.

33. Jenal M, Elkon R, Loayza-Puch F, van Haaften G, Kühn U, Menzies FM, Oude Vrielink JAF, Bos AJ, Drost J, Rooijers K, et al. The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell* 2012, 149:538–553.

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

35. Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* 2010, 143:1018–1029.

36. Lin Y, Li Z, Ozsolak F, Kim SW, Arango-Argoty G, Liu TT, Tenenbaum SA, Bailey T, Monaghan AP, Milos PM, et al. An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res* 2012, 40:8460–8471.

37. Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC. Widespread and extensive lengthening of 3′ UTRs in the mammalian brain. *Genome Res* 2013, 23:812–825.

38. Ulitsky I, Shkumatava A, Jan CH, Subtelny AO, Koppstein D, Bell GW, Sive H, Bartel DP. Extensive alternative polyadenylation during zebrafish development. *Genome Res* 2012, 22:2054–2066.

39. Li Y, Sun Y, Fu Y, Li M, Huang G, Zhang C, Liang J, Huang S, Shen G, Yuan S, et al. Dynamic landscape of tandem 3′ UTRs during zebrafish development. *Genome Res* 2012, 22:1899–1906.

40. Gruber AR, Martin G, Keller W, Zavolan M. Cleavage factor I_m is a key regulator of 3′ UTR length. *RNA Biol* 2012, 9:1405–1412.

41. Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* 2013, 499:360–363.

42. Ji X, Wan J, Vishnu M, Xing Y, Liebhaber SA. αCP poly(C) binding proteins act as global regulators of alternative polyadenylation. *Mol Cell Biol* 2013, 33:2560–2573.

43. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. Landscape of transcription in human cells. *Nature* 2012, 489:101–108.

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

45. Beyer K, Dandekar T, Keller W. 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* 1997, 272:26769–26779.

46. Takagaki Y, Manley JL. RNA recognition by the human polyadenylation factor CstF. *Mol Cell Biol* 1997, 17:3907–3914.

47. Brown KM, Gilmartin GM. A mechanism for the regulation of pre-mRNA 3′ processing by human cleavage factor I_m. *Mol Cell* 2003, 12:1467–1476.

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

49. Proudfoot N, O'Sullivan J. Polyadenylation: a tail of two complexes. *Curr Biol* 2002, 12:R855–R857.

50. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. CLIP identifies Nova-regulated RNA networks in the brain. *Science* 2003, 302:1212–1215.

51. Ule J, Jensen K, Mele A, Darnell RB. CLIP: a method for identifying protein–RNA interaction sites in living cells—post-transcriptional regulation of gene expression. *Methods* 2005, 37:376–386.

52. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 2008, 456:464–469.

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

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

55. König J, Zarnack K, Luscombe NM, Ule J. Protein–RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet* 2012, 13:77–83.

56. 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, 8:559–564.

57. Sugimoto Y, König J, Hussain S, Zupan B, Curk T, Frye M, Ule J. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol* 2012, 13:R67.

58. Jaskiewicz L, Bilen B, Hausser J, Zavolan M. Argonaute CLIP—a method to identify in vivo targets of miRNAs—microRNA methods. *Methods* 2012, 58:106–112.

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

60. Corcoran DL, Georgiev S, Mukherjee N, Gottwein E, Skalsky RL, Keene JD, Ohler U. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol* 2011, 12:R79.

61. Sievers C, Schlumpf T, Sawarkar R, Comoglio F, Paro R. Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic Acids Res* 2012, 40:e160.

62. Zhang C, Darnell RB. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol* 2011, 29:607–614.

63. Sheets MD, Ogg SC, Wickens MP. Point mutations in AAUAAA and the poly (A) addition site: effects on the

accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res* 1990, 18:5799–5805.

64. Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res* 2000, 10:1001–1010.

65. Wang L, Dowell RD, Yi R. Genome-wide maps of polyadenylation reveal dynamic mRNA 3′-end formation in mammalian cell lineages. *RNA* 2013, 19:413–425.

66. Yoon OK, Hsu TY, Im JH, Brem RB. Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. *PLoS Genet* 2012, 8:e1002882.

67. Higgs DR, Goodbourn SE, Lamb J, Clegg JB, Weatherall DJ, Proudfoot NJ. α-Thalassaemia caused by a polyadenylation signal mutation. *Nature* 1983, 306:398–400.

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

69. Nunes NM, Li W, Tian B, Furger A. A functional human poly(A) site requires only a potent DSE and an A-rich upstream sequence. *EMBO J* 2010, 29:1523–1536.

70. Yao C, Biesinger J, Wan J, Weng L, Xing Y, Xie X, Shi Y. Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. *Proc Natl Acad Sci U S A* 2012, 109:18773–18778.

71. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* 2004, 101:6062–6067.

72. de Klerk E, Venema A, Anvar SY, Goeman JJ, Hu O, Trollet C, Dickson G, den Dunnen JT, van der Maarel SM, Raz V, et al. Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation. *Nucleic Acids Res* 2012, 40:9089–9101.

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

74. Berg MG, Singh LN, Younis I, Liu Q, Pinto AM, Kaida D, Zhang Z, Cho S, Sherrill-Mix S, Wan L, et al. U1 snRNP determines mRNA length and regulates isoform expression. *Cell* 2012, 150:53–64.

75. Ntini E, Järvelin AI, Bornholdt J, Chen Y, Boyd M, Jørgensen M, Andersson R, Hoof I, Schein A, Andersen PR, et al. Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nat Struct Mol Biol* 2013, 20:923–928.

76. Katahira J, Okuzaki D, Inoue H, Yoneda Y, Maehara K, Ohkawa Y. Human TREX component Thoc5 affects alternative polyadenylation site choice by recruiting mammalian cleavage factor I. *Nucleic Acids Res* 2013, 41:7060–7072.

77. Yang Q, Gilmartin GM, Doublié S. The structure of human cleavage factor I(m) 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, 8:748–753.

78. Naganuma T, Nakagawa S, Tanigawa A, Sasaki YF, Goshima N, Hirose T. Alternative 3′-end processing of long noncoding RNA initiates construction of nuclear paraspeckles. *EMBO J* 2012, 31:4020–4034.

79. Liu Y, Hu W, Murakawa Y, Yin J, Wang G, Landthaler M, Yan J. Cold-induced RNA-binding proteins regulate circadian gene expression by controlling alternative polyadenylation. *Sci Rep* 2054, 2013:3.

80. Masuda T, Itoh K, Higashitsuji H, Higashitsuji H, Nakazawa N, Sakurai T, Liu Y, Tokuchi H, Fujita T, Zhao Y, et al. Cold-inducible RNA-binding protein (Cirp) interacts with Dyrk1b/Mirk and promotes proliferation of immature male germ cells in mice. *Proc Natl Acad Sci U S A* 2012, 109:10885–10890.

81. Ji Z, Luo W, Li W, Hoque M, Pan Z, Zhao Y, Tian B. Transcriptional activity regulates alternative cleavage and polyadenylation. *Mol Syst Biol* 2011, 7:534.

82. Nagaike T, Logan C, Hotta I, Rozenblatt-Rosen O, Meyerson M, Manley JL. Transcriptional activators enhance polyadenylation of mRNA precursors. *Mol Cell* 2011, 41:409–418.

83. Perales R, Bentley D. "Cotranscriptionality": the transcription elongation complex as a nexus for nuclear transactions. *Mol Cell* 2009, 36:178–191.

84. de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, Cramer P, Bentley D, Kornblihtt AR. A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* 2003, 12:525–532.

85. Spies N, Nielsen CB, Padgett RA, Burge CB. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell* 2009, 36:245–254.

86. Khaladkar M, Smyda M, Hannenhalli S. Epigenomic and RNA structural correlates of polyadenylation. *RNA Biol* 2011, 8:529–537.

87. Lee C-Y, Chen L. Alternative polyadenylation sites reveal distinct chromatin accessibility and histone modification in human cell lines. *Bioinformatics* 2013, 29:1713–1717.

88. Cheng Y, Miura RM, Tian B. Prediction of mRNA polyadenylation sites by support vector machine. *Bioinformatics* 2006, 22:2320–2325.

89. Akhtar MN, Bukhari SA, Fazal Z, Qamar R, Shahmuradov IA. POLYAR, a new computer program for prediction of poly(A) sites in human sequences. *BMC Genomics* 2010, 11:646.

90. Kalkatawi M, Rangkuti F, Schramm M, Jankovic BR, Kamau A, Chowdhary R, Archer JAC, Bajic VB. Dragon PolyA Spotter: predictor of poly(A) motifs within

human genomic DNA sequences. *Bioinformatics* 2012, 28:127–129.

91. Darmon SK, Lutz CS. Novel upstream and downstream sequence elements contribute to polyadenylation efficiency. *RNA Biol* 2012, 9:1255–1265.
92. Hans H, Alwine JC. Functionally significant secondary structure of the simian virus 40 late polyadenylation signal. *Mol Cell Biol* 2000, 20:2926–2932.
93. Baltz AG, Munschauer M, Schwanhäusser B, Vasile A, Murakawa Y, Schueler M, Youngs N, Penfold-Brown D, Drew K, Milek M, et al. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell* 2012, 46:674–690.
94. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 2013, 499:172–177.
95. Arnold P, Schöler A, Pachkov M, Balwierz PJ, Jørgensen H, Stadler MB, van Nimwegen E, Schübeler D. Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. *Genome Res* 2013, 23:60–73.
96. FANTOM Consortium. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet* 2009, 41:553–562.