# A comprehensive analysis of 3′ end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation

Andreas J. Gruber,[1] Ralf Schmidt,[1] Andreas R. Gruber,[1] Georges Martin,[1] Souvik Ghosh,[1] Manuel Belmadani,[1,2] Walter Keller,[1] and Mihaela Zavolan[1]

[1] *Computational and Systems Biology, Biozentrum, University of Basel, 4056 Basel, Switzerland*

**Alternative polyadenylation (APA) is a general mechanism of transcript diversification in mammals, which has been recently linked to proliferative states and cancer. Different 3′ untranslated region (3′ UTR) isoforms interact with different RNA-binding proteins (RBPs), which modify the stability, translation, and subcellular localization of the corresponding transcripts. Although the heterogeneity of pre-mRNA 3′ end processing has been established with high-throughput approaches, the mechanisms that underlie systematic changes in 3′ UTR lengths remain to be characterized. Through a uniform analysis of a large number of 3′ end sequencing data sets, we have uncovered 18 signals, six of which are novel, whose positioning with respect to pre-mRNA cleavage sites indicates a role in pre-mRNA 3′ end processing in both mouse and human. With 3′ end sequencing we have demonstrated that the heterogeneous ribonucleoprotein C (HNRNPC), which binds the poly(U) motif whose frequency also peaks in the vicinity of polyadenylation (poly(A)) sites, has a genome-wide effect on poly(A) site usage. HNRNPC-regulated 3′ UTRs are enriched in ELAV-like RBP 1 (ELAVL1) binding sites and include those of the *CD47* gene, which participate in the recently discovered mechanism of 3′ UTR–dependent protein localization (UDPL). Our study thus establishes an up-to-date, high-confidence catalog of 3′ end processing sites and poly(A) signals, and it uncovers an important role of HNRNPC in regulating 3′ end processing. It further suggests that U-rich elements mediate interactions with multiple RBPs that regulate different stages in a transcript's life cycle.**

**[Supplemental material is available for this article.]**

The 3′ ends of most RNA polymerase II–generated transcripts are generated through endonucleolytic cleavage and the addition of a polyadenosine tail of 70–100 nucleotides (nt) median length (Subtelny et al. 2014). Recent studies have revealed systematic changes in 3′ UTR lengths upon changes in cellular states, either those that are physiological (Sandberg et al. 2008; Berg et al. 2012) or those during pathologies (Masamha et al. 2014). 3′ UTR lengths are sensitive to the abundance of specific spliceosomal proteins (Kaida et al. 2010), core pre-mRNA 3′ end processing factors (Gruber et al. 2012; Martin et al. 2012), and polyadenylation factors (Jenal et al. 2012). Because 3′ UTRs contain many recognition elements for RNA-binding proteins (RBPs) that regulate the subcellular localization, intracellular traffic, decay, and translation rate of the transcripts in different cellular contexts (see, e.g., Nam et al. 2014), the choice of polyadenylation (poly(A)) sites has important regulatory consequences that reach up to the subcellular localization of the resulting protein (Berkovits and Mayr 2015). Studies of presumed regulators of polyadenylation would greatly benefit from the general availability of comprehensive catalogs of poly(A) sites such as PolyA_DB (Zhang et al. 2005; Lee et al. 2007), which was introduced in 2005 and updated 2 yr later.

Full-length cDNA sequencing offered a first glimpse on the pervasiveness of transcription across the genome and on the complexity of gene structures (Kawai et al. 2001). Next-generation sequencing technologies, frequently coupled with the capture of transcript 5′ or 3′ ends with specific protocols, enabled the quantification of gene expression and transcript isoform abundance (Katz et al. 2010). By increasing the depth of coverage of transcription start sites and mRNA 3′ ends, these protocols aimed to improve the quantification accuracy (de Hoon and Hayashizaki 2008; Ozsolak et al. 2009; Beck et al. 2010; Shepard et al. 2011). Sequencing of mRNA 3′ ends takes advantage of the poly(A) tail, which can be captured with an oligo-dT primer. More than 4.5 billion reads were obtained with several protocols from human or mouse mRNA 3′ ends in a variety of cell lines (Shepard et al. 2011; Lin et al. 2012), tissues (Derti et al. 2012; You et al. 2014), developmental stages (Li et al. 2012; Ulitsky et al. 2012), and cell differentiation stages (Hoque et al. 2013), as well as following perturbations of specific RNA processing factors (Gruber et al. 2012; Jenal et al. 2012; Martin et al. 2012; Almada et al. 2013; Ji et al. 2013). Although some steps are shared by many of the proposed 3′ end sequencing protocols, the studies that employed these methods have reported widely varying numbers of 3′ end processing sites. For example, 54,686 (Lee et al. 2007), 439,390 (Derti et al. 2012), and 1,287,130 (Lin et al. 2012) sites have been reported in the human genome.

[2] **Present address: University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada**
**Corresponding author: mihaela.zavolan@unibas.ch**

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.202432.115. Freely available online through the *Genome Research* Open Access option.

The current knowledge about sequence motifs that are relevant to cleavage and polyadenylation (for review, see Proudfoot 2011) goes back to studies conducted before next-generation sequencing technologies became broadly used (Proudfoot and Brownlee 1976; Beaudoing et al. 2000; Tian et al. 2005). These studies revealed that the AAUAAA hexamer, which recently was found to bind the WDR33 and CPSF4 subunits of the cleavage and polyadenylation specificity factor (CPSF) (Chan et al. 2014; Schönemann et al. 2014) and some close variants, is highly enriched upstream of the pre-mRNA cleavage site. The A[AU]UAAA *cis*-regulatory element (also called poly(A) signal) plays an important role in pre-mRNA cleavage and polyadenylation (Tian and Graber 2012) and is found at a large proportion of pre-mRNA cleavage sites identified in different studies (Graber et al. 1999; MacDonald and Redondo 2002; Tian et al. 2005). However, some transcripts that do not have this poly(A) signal are nevertheless processed, indicating that the poly(A) signal is not absolutely necessary for cleavage and polyadenylation. The constraints that functional poly(A) signals have to fulfill are not entirely clear, and at least 10 other hexamers have been proposed to have this function (Beaudoing et al. 2000).

Viral RNAs as, for example, from the simian virus 40 have been instrumental in uncovering RBP regulators of polyadenylation and their corresponding sequence elements. Previous studies revealed modulation of poly(A) site usage by U-rich element binding proteins such as the heterogeneous nuclear ribonucleoprotein (hnRNP) C1/C2 (Wilusz et al. 1988; Zhao et al. 2005), the polypyrimidine tract binding protein 1 (Castelo-Branco et al. 2004; Zhao et al. 2005), FIP1L1, and CSTF2 (Zhao et al. 2005), and by proteins that bind G-rich elements—cleavage stimulation factor CSTF2 (Alkan et al. 2006) and HNRNPs F and H1 (Arhin et al. 2002)—or C-rich elements—poly(rC)-binding protein 2 (Ji et al. 2013). Some of these proteins are multifunctional splicing factors that appear to couple various steps in pre-mRNA processing, such as splicing, cleavage, and polyadenylation (Millevoi et al. 2009). The sequence elements to which these regulators bind are also frequently multifunctional, enabling positive or negative regulation by different RBPs (Alkan et al. 2006). A first step toward understanding the regulation of poly(A) site choice is to construct genome-wide maps of poly(A) sites, which can be used to investigate differential polyadenylation across tissues and the response of poly(A) sites to specific perturbations.

## Results

### Preliminary processing of 3′ end sequencing data sets

Protocol-specific biases as well as vastly different computational data processing strategies may explain the discrepancy in the reported number of 3′ end processing sites, which ranges from less than 100,000 to over 1 million (Lee et al. 2007; Derti et al. 2012; Lin et al. 2012) for the human genome. By comparing the 3′ end processing sites from two recent genome-wide studies (Derti et al. 2012; You et al. 2014), we found that a substantial proportion was unique to one or the other of the two studies (Supplemental Table 1). This motivated us to develop a uniform and flexible processing pipeline that facilitates the incorporation of all published sequencing data sets, yielding a comprehensive set of high-confidence 3′ end processing sites. From public databases we obtained 78 human and 110 mouse data sets of 3′ end sequencing reads (Supplemental Tables 2, 3), generated with nine different protocols, for which sufficient information to permit the appropriate preprocessing steps (trimming of 5′ and 3′ adapter sequences, reverse-complementing the reads, etc., as appropriate) was available. We preprocessed each sample as appropriate given the underlying protocol and then subjected all data sets to a uniform analysis as follows. We mapped the preprocessed reads to the corresponding genome and transcriptome and identified unique putative 3′ end processing sites. Because many protocols employ oligo-dT priming to capture the pre-mRNA 3′ ends, internal priming is a common source of false-positive sites, which we tried to identify and filter out as described in the Methods section. From the nearly 200 3′ end sequencing libraries, we thus obtained an initial set of 6,983,499 putative 3′ end processing sites for human and 8,376,450 for mouse. The majority of these sites (76% for human and 71% for mouse) had support in only one sample, consistent with our initial observations of limited overlap between the sets of sites identified in individual studies and mirroring also the results of transcription start site mapping with the CAGE technology (FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014). Nevertheless, we developed an analysis protocol that aimed to identify bona fide, independently regulated poly(A) sites, including those that have been captured in a single sample. To do this, we used not only the sequencing data but also information about poly(A) signals, which we therefore set to comprehensively identify in the first step of our analysis.

### Highly specific positioning with respect to the pre-mRNA cleavage site reveals novel poly(A) signals

To search for signals that may guide polyadenylation, we designed a very stringent procedure to identify high-confidence 3′ end processing sites. Pre-mRNA cleavage is not completely deterministic but occurs with higher frequency at “strong” 3′ end processing sites and with low frequency at neighboring positions (Tian et al. 2005). Therefore, a common step in the analysis of 3′ end sequencing data is to cluster putative sites that are closely spaced and to report the dominant site from each cluster (Tian et al. 2005; Martin et al. 2012; Lianoglou et al. 2013). To determine an appropriate distance threshold, we ranked all the putative sites first by the number of samples in which they were captured and then by the normalized number of reads in these samples. By traversing the list of sites from those with the strongest to those with the weakest support, we associated lower-ranking sites located up to a specific distance from the higher-ranked site with the corresponding higher-ranking site. We scanned the range of distances from 0 to 25 nt upstream of and downstream from the high-ranking site, and we found that the proportion of putative 3′ end processing sites that are merged into clusters containing more than one site reached 40% at ~8 nt and changed little by further increasing the distance (for details, see Methods). For consistency with previous studies (Tian et al. 2005), we used a distance of 12 nt. To reduce the frequency of protocol-specific artifacts, we used only clusters that were supported by reads derived with at least two protocols, and to allow unambiguous association of signals to clusters, for the signal inference we only used clusters that did not have another cluster within 60 nt. This procedure resulted in 221,587 3′ end processing clusters for human and 209,345 for mouse.

By analyzing 55-nt-long regions located immediately upstream of the center of these 3′ end processing clusters (as described in the Methods section), we found that the canonical poly(A) signals AAUAAA and AUUAAA were highly enriched and had a strong positional preference, peaking at 21 nt upstream of cleavage sites (Fig. 1A), as reported previously (Beaudoing et al.

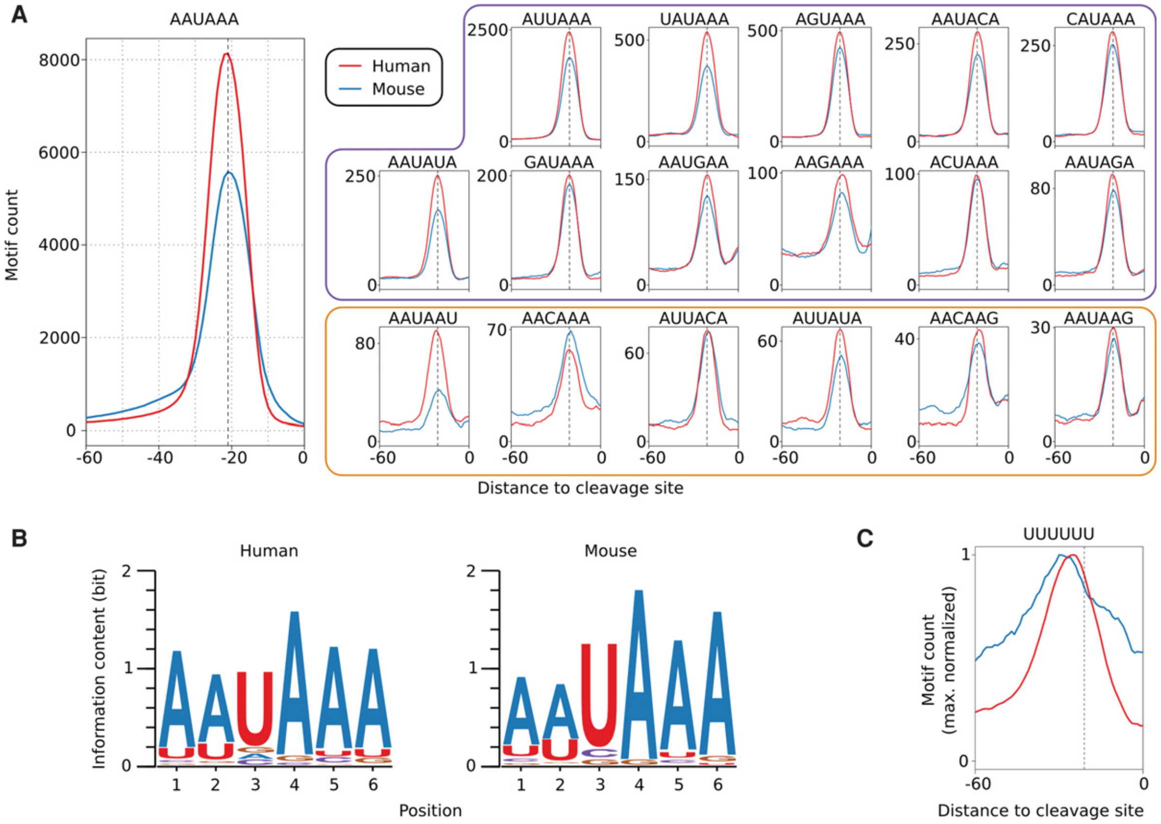


**Figure 1.** Hexamers with highly specific positioning upstream of human and mouse pre-mRNA 3′ end cleavage sites. (*A*) The frequency profiles of the 18 hexamers that showed the positional preference expected for poly(A) signals in both human and mouse. The known poly(A) signal, AAUAAA, had the highest frequency of occurrence (*left*). Apart from the 12 signals previously identified (AAUAAA and motifs with the purple frame) (Beaudoing et al. 2000), we have identified six additional motifs (orange frame) whose positional preference with respect to poly(A) sites suggests that they function as poly(A) signals and are conserved between human and mouse. (*B*) Sequence logos based on all occurrences of the entire set of poly(A) signals from the human (*left*) and mouse (*right*) atlas. (*C*) The $(U)_6$ motif, which is also enriched upstream of pre-mRNA cleavage sites, has a broader frequency profile and peaks upstream of the poly(A) signals, which are precisely positioned 20–22 nt upstream of the pre-mRNA cleavage sites (indicated by the dashed, vertical line).

2000; Tian et al. 2005). We therefore asked whether other hexamers have a similarly peaked frequency profile, which would be indicative of their functioning as poly(A) signals. The 12 signals that were identified in a previous study (Beaudoing et al. 2000) served as controls for the procedure. In both mouse and human data, the motif with the highest peak was, as expected, the canonical poly(A) signal AAUAAA, which occurred in 46.82% and 39.54% of the human and mouse sequences, respectively. Beyond this canonical signal, we found 21 additional hexamers, the second most frequent being the close variant of the canonical signal AUUAAA, which was present in 14.52% and 12.28% of the human and mouse 3′ sequences, respectively. All 12 known poly(A) signals (Beaudoing et al. 2000) were recovered by our analysis in both species, demonstrating the reliability of our approach. Further supporting this conclusion is the fact that six of the 10 newly identified signals in each of the two species are shared. All of the conserved signals are very close variants (1 nt difference except for AACAAG) of one of the two main poly(A) signals, AAUAAA and AUUAAA. Strikingly, all of these signals peak in frequency at 20–22 nt upstream of the cleavage site (Fig. 1A). Experimental evidence for single-nucleotide variants of the AAUAAA signal (including the AACAAA, AAUAAU, and AAUAAG motifs identified here) functioning in polyadenylation was already provided by Sheets et al. (1990). The four signals identified in only one of each species also had a clear peak at the expected position with respect to the poly(A) site, but they had a larger variance (Supplemental Fig. 1). Altogether, these results indicate a genuine role of the newly identified signals in the process of cleavage and polyadenylation.

Of the 221,587 high-confidence 3′ end processing clusters in human and 209,345 in mouse, 87% and 79%, respectively, had at least one of the 22 signals identified above in their upstream region. Even when considering only the 18 signals that are conserved between human and mouse, 86% of the human clusters and 75% of the mouse clusters had a poly(A) signal. Thus, our analysis almost doubles the set of poly(A) signals and suggests that the vast majority of poly(A) sites does indeed have a poly(A) signal that is positioned very precisely with respect to the pre-mRNA cleavage site. The dominance of the canonical poly(A) signal is reflected in the sequence logos constructed based on all annotated hexamers in the human and mouse poly(A) site atlases, generated as described in the following section and in the Methods section (Fig. 1B).

## A comprehensive catalog of high-confidence 3′ end processing sites

Based on all of the 3′ end sequencing data sets available (for more details about the protocols that were used to generate these data sets, see Supplemental Material) and the conserved poly(A) signals that we inferred as described above, we constructed a comprehensive catalog of strongly supported 3′ end processing sites in both the mouse and human genomes. We started from the 6,983,499 putative cleavage sites for human and 8,376,450 for mouse. Although in many data sets a large proportion of putative sites was supported by single reads and did not have any of the expected poly(A) signals in the upstream region, the incidence of upstream poly(A) signals increased with the number of reads supporting a putative site (Supplemental Fig. 2). Thus, we used the frequency of occurrence of poly(A) signals to define sample-specific cutoffs for the number of reads required to support a putative cleavage site. We then clustered all putative sites with sufficient read support, associating lower-ranked sites with higher-ranking sites that were located within at most 12 nt upstream or downstream, as described above. Because in this set of clusters we found cases where the pre-mRNA cleavage site appeared located in an A-rich region upstream of another putative cleavage site, we specifically reviewed clusters in which a putative cleavage site was very close to a poly(A) signal, as these likely reflect internal priming events (Shepard et al. 2011; Derti et al. 2012; Gruber et al. 2014). These clusters were either associated with a downstream cluster, retained as independent clusters, or discarded, according to the procedure outlined in the Methods section. By reasoning that distinct 3′ end processing sites should have independent signals to guide their processing, we merged clusters that shared all poly(A) signals within 60 nt upstream of their representative sites, clusters whose combined span was <25 nt, and clusters without annotated poly(A) signals that were closer than 12 nt to each other and had a combined span of at most 50 nt. Clusters >50 nt and without poly(A) signals were excluded from the atlas. This procedure (for details, see the Methods section) resulted in 392,912 human and 183,225 mouse 3′ end processing clusters. Of note, even though 3′ end processing sites that were within 25 nt of each other were merged into single clusters, the median cluster span was very small, 7 and 3 nt for mouse and human, respectively (Supplemental Fig. 3). Supplemental Figures 4A and 5A show the frequency of occurrence of the four nucleotides as a function of the distance to the cleavage sites for sites that were supported by a decreasing number of protocols. These profiles exhibited the expected pattern (Tian et al. 2005; Ozsolak et al. 2010; Martin et al. 2012), indicating that our approach identified bona fide 3′ end processing sites, even when they had limited experimental support.

The proportion of clusters located in the terminal exon increased with an increasing number of supporting protocols (Supplemental Fig. 4B, 5B), probably indicating that the canonical poly(A) sites of constitutively expressed transcripts are identified by the majority of protocols, whereas poly(A) sites that are only used in specific conditions were captured only in a subset of experiments. Although in constructing our catalog we used most of the reads generated in two recent studies (>95% of the reads that supported human 3′ end processing sites in these two data sets mapped within the poly(A) site clusters of our human catalog) (Derti et al. 2012; You et al. 2014), only 61.82% (You et al. 2014) and 41.38% (Derti et al. 2012) of the unique processing sites inferred in these studies were located within poly(A) clusters from our human catalog. This indicated that a large fraction of the sites that were cataloged in previous studies is supported by a very small number of reads and lacks canonically positioned poly(A) signals. We applied very stringent rules to construct an atlas of high-confidence poly(A) sites, and the entire set of putative cleavage sites that resulted from mapping all of the reads obtained in these 3′ end sequencing studies is available as Supplemental Data S1 (human) and S2 (mouse), as well as online at http://www.polyasite.unibas.ch, where users can filter sites of interest based on the number of supporting protocols, the identified poly(A) signals, and/or the genomic context of the clusters.

## 3′ end processing regions are enriched in poly(U)

Of the human and mouse 3′ end processing sites from our poly(A) atlases, 76% and 75%, respectively, possessed a conserved poly(A) signal in their 60 nt upstream region. That ~25% did not may support the hypothesis that pre-mRNA cleavage and polyadenylation do not absolutely require a poly(A) signal (Venkataraman et al. 2005). Nevertheless, we asked whether these sites possess other signals, with a different positional preference, which may contribute to their processing. To answer this question, we searched for hexamers that were significantly enriched in the 60 nt upstream of cleavage sites without an annotated poly(A) signal. The two most enriched hexamers were poly(A) (*P*-value of binomial test $<1.0 \times 10^{-100}$), which showed a broad peak in the region of −20 to −10 upstream of cleavage sites, and poly(U) (*P*-value $<1.0 \times 10^{-100}$), which also has a broad peak around −25 nt upstream of cleavage sites, particularly pronounced in the human data set (Fig. 1C). The poly(U) hexamer is very significantly enriched (*P*-value of binomial test $<1.0 \times 10^{-100}$) in the 60 nt upstream regions of all poly(A) sites, not only in those that do not have a common poly(A) signal (11th most enriched hexamer in the human atlas and 60th most enriched hexamer in the mouse atlas) (Supplemental Tables 4, 5). Although the A- and U-richness of pre-mRNA 3′ end processing regions have been observed before (Tian and Graber 2012), their relevance for polyadenylation and the regulators that bind these motifs have been characterized only partially. For example, the core 3′ end processing factor FIP1L1 can bind poly(U) (Kaufmann et al. 2004; Lackford et al. 2014), and its knock-down causes a systematic increase in 3′ UTR lengths (Lackford et al. 2014; Li et al. 2015).

## *HNRNPC* knock-down causes global changes in alternative cleavage and polyadenylation

Several proteins (ELAVL1, TIA1, TIAL1, U2AF2, CPEB2 and CPEB4, HNRNPC) that regulate pre-mRNA splicing and polyadenylation, as well as mRNA stability and metabolism, have also been reported to bind U-rich elements (Ray et al. 2013). Of these, HNRNPC has been recently studied with crosslinking and immunoprecipitation (CLIP) and found to bind the majority of protein-coding genes (König et al. 2010), with high specificity for poly(U) tracts (König et al. 2010; Ray et al. 2013; Zarnack et al. 2013; Cieniková et al. 2014; Liu et al. 2015). HNRNPC appears to nucleate the formation of ribonucleoprotein particles on nascent transcripts and to regulate pre-mRNA splicing (König et al. 2010; Zarnack et al. 2013) and polyadenylation at *Alu* repeats (Tajnik et al. 2015). We therefore hypothesized that HNRNPC binds to the U-rich regions in the vicinity of poly(A) sites and globally regulates not only splicing but also pre-mRNA cleavage and polyadenylation.

To test this hypothesis, we generated two sets of pre-mRNA 3′ end sequencing libraries from HEK 293 cells that were transfected either with a control siRNA or with an siRNA directed against *HNRNPC*. The siRNA was very efficient, strongly reducing the HNRNPC protein expression, as shown in Supplemental Figure 6. To evaluate the effect of *HNRNPC* knock-down on polyadenylation, we focused on exons with multiple poly(A) sites. We identified 12,136 such sites in 4405 exons with a total of 22,698,094 mapped reads (Supplemental Table 6). We calculated the relative usage of a poly(A) site in a given sample as the proportion of reads that mapped to that site among the reads mapping to any 3′ end processing site in the corresponding exon. We then computed the change in relative use of each poly(A) site in si-*HNRNPC*–treated cells compared with control siRNA-treated cells. We found that *HNRNPC* knock-down affects a large proportion of transcripts with multiple poly(A) sites, reminiscent of what we previously reported for the 25- and 68-kDa subunits of the cleavage factor I ($\mathrm{CFI_m}$) (Gruber et al. 2012; Martin et al. 2012). Out of the 5152 poly(A) sites that showed consistent behavior across replicates, we found 1402 poly(A) sites (27.2%) to increase in usage, 1378 poly(A) sites (26.7%) to decrease in usage, and 2372 poly(A) sites (46.0%) to undergo only a minor change in usage upon knock-down of *HNRNPC*. To find out whether HNRNPC systematically increases or decreases 3′ UTR lengths, we examined the relative position of poly(A) sites whose usage increases or decreases most strongly in response to *HNRNPC* knock-down, within 3′ UTRs. The results indicated that poly(A) sites whose usage increased and decreased upon *HNRNPC* knock-down tended to be located distally and proximally, respectively, within exons (Fig. 2A). We confirmed the overall increase in 3′ UTR lengths upon *HNRNPC* knock-down by comparing the proximal-to-distal poly(A) site usage ratios of exons that had exactly two polyadenylation sites (replicate 1 *P*-value: $1.1 \times 10^{-19}$; replicate 2 *P*-value: $3.1 \times 10^{-61}$; one-sided Wilcoxon signed-rank test) (Supplemental Figs. 7, 8). It was noted before that distal poly(A) sites are predominantly used in HEK 293 cells (Martin et al. 2012). Indeed, the proportion of dominant (>50% relative usage) distal sites was 61.75% and 62.58%, respectively, in the two control siRNA-treated samples. However, this proportion increased further in the si-*HNRNPC*–treated samples to 64.16% and 65.67%, respectively, consistent with HNRNPC decreasing, on average, the lengths of 3′ UTRs. Nevertheless, many 3′ UTRs became shorter upon this treatment as will be discussed in more detail in the analysis of terminal exons with exactly two poly(A) sites (tandem poly(A) sites) below.

As HNRNPC binds RNAs in a sequence-specific manner, one expects an enrichment of HNRNPC binding sites in the vicinity of poly(A) sites whose usage is affected by the *HNRNPC* knock-down. Indeed, this is what we observed. The density of $(U)_5$ tracts, previously reported to be the binding sites for HNRNPC (König et al. 2010; Ray et al. 2013; Liu et al. 2015), was markedly higher around poly(A) sites whose usage increased upon *HNRNPC* knock-down compared with sites whose relative usage did not change or decreased upon *HNRNPC* knock-down (Fig. 2B). No such enrichment emerged from a similar analysis of untransfected versus si-Control transfected cells (Supplemental Fig. 9). To exclude the possibility that this profile is due to a small number of regions that are very U-rich, we also determined the fraction of poly(A) sites that contained $(U)_5$ tracts among the poly(A) sites whose usage increased, decreased, or did not change upon *HNRNPC* knock-down (Supplemental Fig. 10). We found, consistent with the results shown in Figure 2B, a higher proportion of $(U)_5$ tract-containing poly(A) sites among those whose usage increased upon *HNRNPC* knock-down compared with those whose usage decreased or was not changed. To further validate HNRNPC binding at the derepressed poly(A) sites, we carried out HNRNPC CLIP and found, indeed, that derepressed sites have a higher density of HNRNPC CLIP reads compared with other poly(A) sites (Supplemental Fig. 11). Finally, we found that poly(A) sites with the highest density of $(U)_5$ tracts in the 100-nt region centered on the cleavage site were reproducibly used with increased frequency upon *HNRNPC* knock-down relative to poly(A) sites that did not contain any binding sites within 200 nt upstream or downstream (replicate 1 *P*-value: $2.4 \times 10^{-36}$; replicate 2 *P*-value: $1.9 \times 10^{-42}$; one-sided Mann-Whitney *U* test) (Fig. 2C). We therefore concluded that HNRNPC's binding in close proximity of 3′ end processing sites likely masks them from cleavage and polyadenylation.

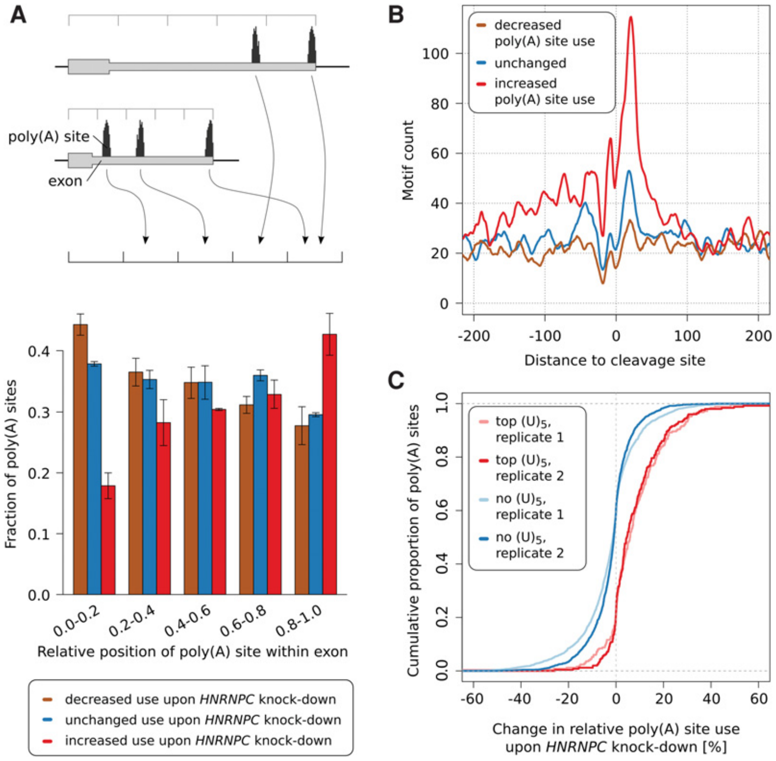


**Figure 2.** siRNA-mediated knock-down of *HNRNPC* leads to increased use of distal poly(A) sites. (*A*) Relative location of sites whose usage decreased (brown), did not change (blue) or increased (red) in response to *HNRNPC* knock-down within 3′ UTRs. We identified the 1000 poly(A) sites whose usage increased most, the 1000 whose usage decreased most, and the 1000 whose usage changed least upon *HNRNPC* knock-down; divided the associated terminal exons into five bins, each covering 20% of the exon's length; and computed the fraction of poly(A) sites that corresponded to each of the three categories within each position bin independently. Values represent means and SDs from the two replicate *HNRNPC* knock-down experiments. (*B*) Smoothened (±5 nt) density of nonoverlapping $(U)_5$ tracts in the vicinity of sites with a consistent behavior (increased, unchanged, decreased use) in the two *HNRNPC* knock-down experiments. (*C*) Cumulative density function of the percentage change in usage of the 250 poly(A) sites with the highest number of $(U)_5$ motifs within ±50 nt around their cleavage site (red) and of poly(A) sites that do not contain any $(U)_5$ tract within ±200 nt (blue), upon *HNRNPC* knock-down.

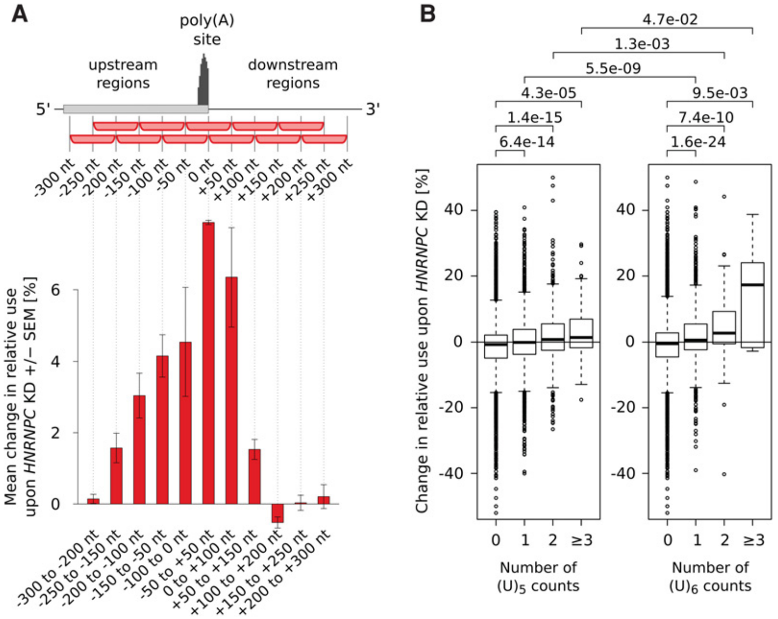


**Figure 3.** The length, number, and location of poly(U) tracts with respect to poly(A) sites influence the change in poly(A) site use upon *HNRNPC* knock-down. (*A*) Mean change in the use of sites containing the highest number of $(U)_5$ motifs within 100-nt-long regions located at specific distances from the cleavage site (indicated on the *x*-axis) upon *HNRNPC* knock-down (KD). Shown are mean ± SEM in the two knock-down experiments. Two hundred fifty poly(A) sites with the highest density of $(U)_5$ motifs at each particular distance were considered. (*B*) Mean changes in the relative use of poly(A) sites that have 0, 1, 2, or more (≥3) nonoverlapping poly(U) tracts within ±50 nt from their cleavage site. Distributions of relative changes in the usage of specific types of sites were compared, and the *P*-values of the corresponding one-sided Mann-Whitney *U* tests are shown at the *top* of the panel.

## Both the number and the length of the uridine tracts contribute to the HNRNPC-dependent poly(A) site usage

If the above conclusions were correct, the effect of *HNRNPC* knock-down should decrease with the distance between the poly(A) site and the HNRNPC binding sites. Thus we determined the mean change in usage of sites with high densities of poly(U) tracts at different distances with respect to the cleavage site, upon *HNRNPC* knock-down. As shown in Figure 3A, we found that the largest change in poly(A) site use is observed for poly(A) sites that have a high density of poly(U) tracts in the 100-nt window centered on the cleavage site. The apparent efficacy of HNRNPC binding sites in modulating polyadenylation decreased with their distance to poly(A) sites and persisted over larger distances upstream (approximately −200 nt) of the poly(A) site compared with regions downstream (approximately +100 nt) from the poly(A) site (Fig. 3A).

Although the minimal RNA recognition motif of HNRNPC consists of five consecutive uridines (Ray et al. 2013; Cieniková et al. 2014; Liu et al. 2015), longer uridine tracts are bound with higher affinity (König et al. 2010; Zarnack et al. 2013; Cieniková et al. 2014). Consistently, we found that, for a given length of the presumed HNRNPC binding site, the effect of the *HNRNPC* knock-down increased with the number of independent sites and that, given the number of nonoverlapping poly(U) tracts, the effect of *HNRNPC* knock-down increased with the length of the sites (Fig. 3B).

## Altered transcript regions contain ELAVL1 binding sites that mediate UDPL

As demonstrated above, binding of HNRNPC to U-rich elements that are located preferentially distally in terminal exons seems to promote the use of proximal 3′ end processing sites. Analysis of a conservative set of tandem poly(A) sites showed that among the poly(A) sites that were derepressed upon *HNRNPC* knock-down and that had at least one $(U)_5$ motif within −200 to +100 nt, two-thirds (390 sites, 67.2%) were located distally, leading to longer 3′ UTRs, whereas the remaining one-third (190 sites, 32.8%) were located proximally leading to shorter 3′ UTRs (for examples, see Supplemental Figs. 12, 13). The altered 3′ UTRs contain U-rich elements with which a multitude of RBPs such as ELAVL1, (also known as Hu Antigen R, or HuR) could interact to regulate, among others, the stability of mRNAs in the cytoplasm (Brennan and Steitz 2001). To determine whether the HNRNPC-dependent alternative 3′ UTRs indeed interact with ELAVL1, we determined the number of ELAVL1 binding sites (obtained from a previous ELAVL1 CLIP study) (Kishore et al. 2011) that are located in the 3′ UTR regions between tandem poly(A) sites. As expected, we found a significant enrichment of ELAVL1 binding sites in 3′ UTR regions whose inclusion in transcripts changed in response to *HNRNPC* knock-down compared with regions whose inclusion did not change (Fig. 4A). Moreover, the density of ELAVL1 binding sites and not only their absolute number was enriched across these 3′ UTR regions (Fig. 4B). Our results thus demonstrate that the HNRNPC-regulated 3′ UTRs are bound and probably susceptible to regulation by ELAVL1.

Recently, a new function has been attributed to the already multifunctional ELAVL1 protein. Work from the Mayr laboratory (Berkovits and Mayr 2015) showed that 3′ UTR regions that contain ELAVL1 binding sites can mediate 3′ UTR–dependent protein localization (UDPL). The ELAVL1 binding sites in the 3′ UTR of the CD47 molecule (*CD47*) transcript were found to be necessary and sufficient for the translocation of the CD47 transmembrane protein from the endoplasmic reticulum (ER) to the plasma membrane, through the recruitment of the SET protein to the site of translation. SET binds to the cytoplasmic domains of the CD47 protein, translocating it from the ER to the plasma membrane via active RAC1 (Fig. 4C; ten Klooster et al. 2007; Berkovits and Mayr 2015). By inspecting our data, we found that the region of the *CD47* 3′ UTR that mediates UDPL is among those that responded to *HNRNPC* knock-down (Fig. 4D). Sashimi plots generated based on mRNA-seq experiments of HEK 293 cells transfected with si-Control or si-*HNRNPC*, respectively, confirmed the increased abundance of the long 3′ UTR isoform of *CD47* upon knock-down of *HNRNPC*. This analysis also verified that the increased relative usage of distal poly(A) sites cannot be explained by alternative splicing events (Supplemental Fig. 14) but are the consequence of increased usage of the distal poly(A) site upon knock-down of *HNRNPC* (Fig. 4D). To find out whether HNRNPC can act as an upstream regulator of UDPL, we quantified

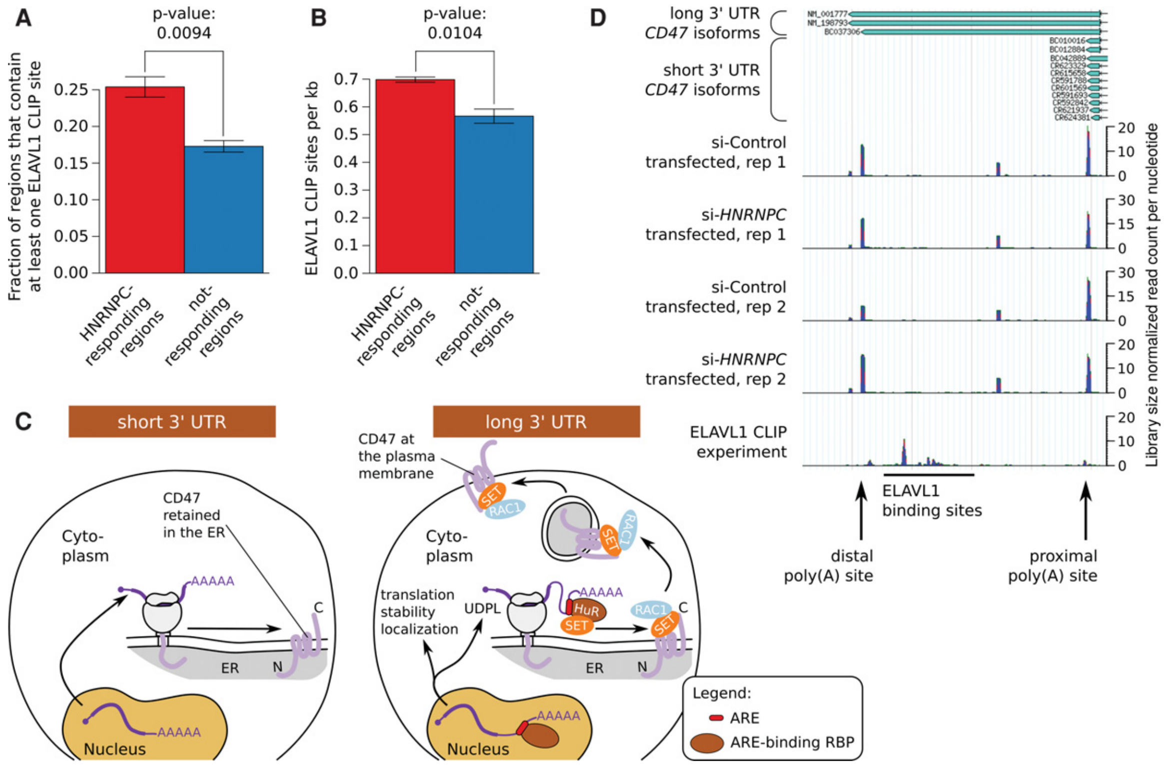
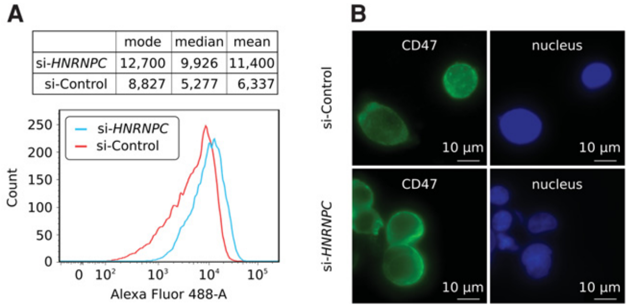


**Figure 4.** HNRNPC-responsive 3′ UTRs are enriched in ELAVL1 binding sites. (*A*) Fraction of HNRNPC-responding and not-responding 3′ UTR regions that contain one or more ELAVL1 CLIP sites. The *P*-value of the one-sided *t*-test is shown. (*B*) Density of ELAVL1 CLIP sites per kilobase (kb) in the 3′ UTR regions described above. The *P*-value of the one-sided *t*-test is shown. (*C*) Model of the impact of A/U-rich elements (ARE) in 3′ UTR regions on various aspects of mRNA fate (Berkovits and Mayr 2015). (*D*) Density of A-seq2 reads along the *CD47* 3′ UTR in cells, showing the increased use of the distal poly(A) site in si-*HNRNPC* compared with si-Control transfected cells. The density of ELAVL1 CLIP reads in this region is also shown.

the level of CD47 at the plasma membrane of cells that underwent siRNA-mediated knock-down of *HNRNPC* and cells that were treated with a control siRNA. Strikingly, we found that the CD47 level at the plasma membrane increased upon *HNRNPC* knock-down (Fig. 5A; Supplemental Fig. 15). Western blots for CD47 that were performed in *HNRNPC* and control siRNA-treated cells ruled out the possibility that the increase in membrane-associated CD47 upon *HNRNPC* knock-down was due to an increase in total CD47 levels (Supplemental Fig. 16). We also carried out an independent immunofluorescence analysis of CD47 in these two conditions and again observed that the *HNRNPC* knock-down led to an increase in the plasma membrane CD47 levels (Fig. 5B). Overall, our results suggest that HNRNPC can function as an upstream regulator of UDPL.

## HNRNPC represses cleavage and polyadenylation at intronic, transcription start site-proximal poly(A) sites

Up to this point, we focused on alternative polyadenylation (APA) sites that are located within single exons. However, given that HNRNPC binds to nascent transcripts, we also asked whether HNRNPC affects other types of APA, specifically at sites located in regions that in the GENCODE v19 set of transcripts (Harrow et al. 2012) are annotated as intronic. Indeed, we found that the *HNRNPC* knock-down increased the use of intronic poly(A) sites that are most enriched in putative HNRNPC-binding $(U)_5$ motifs within ±50 nt compared with sites that do not have $(U)_5$ tracts within ±200 nt (*P*-values of the one-sided Mann-Whitney *U* test for the data from the two replicate knock-down experiments are $1.4 \times 10^{-30}$ and $5.1 \times 10^{-29}$) (Fig. 6A). These sites are predominantly associated with cryptic exons that are spliced in upon *HNRNPC* knock-down as opposed to exons whose splice site fails to be recognized by the spliceosome leading to exon extension in HNRNPC-depleted cells (Supplemental Fig. 17). Importantly, only the intronic sites that responded to *HNRNPC* knock-down were



| | mode | median | mean |
|---|---|---|---|
| si-*HNRNPC* | 12,700 | 9,926 | 11,400 |
| si-Control | 8,827 | 5,277 | 6,337 |

**Figure 5.** The knock-down of *HNRNPC* affects CD47 protein localization. (*A*) Indirect immunophenotyping of membrane-associated CD47 in HEK 293 cells that were treated either with an si-*HNRNPC* (blue) or with si-Control (red) siRNA. Mean, median, and mode of the Alexa Fluor 488 intensities computed for cells in each transfection set (*top*), with histograms shown in the *bottom* panel. (*B*) Immunofluorescence staining of permeabilized HEK 293 cells with CD47 antibody (*left*) or nuclear staining with Hoechst (*right*). *Top* and *bottom* panels correspond to cells that were treated with control siRNA and si-*HNRNPC*, respectively.

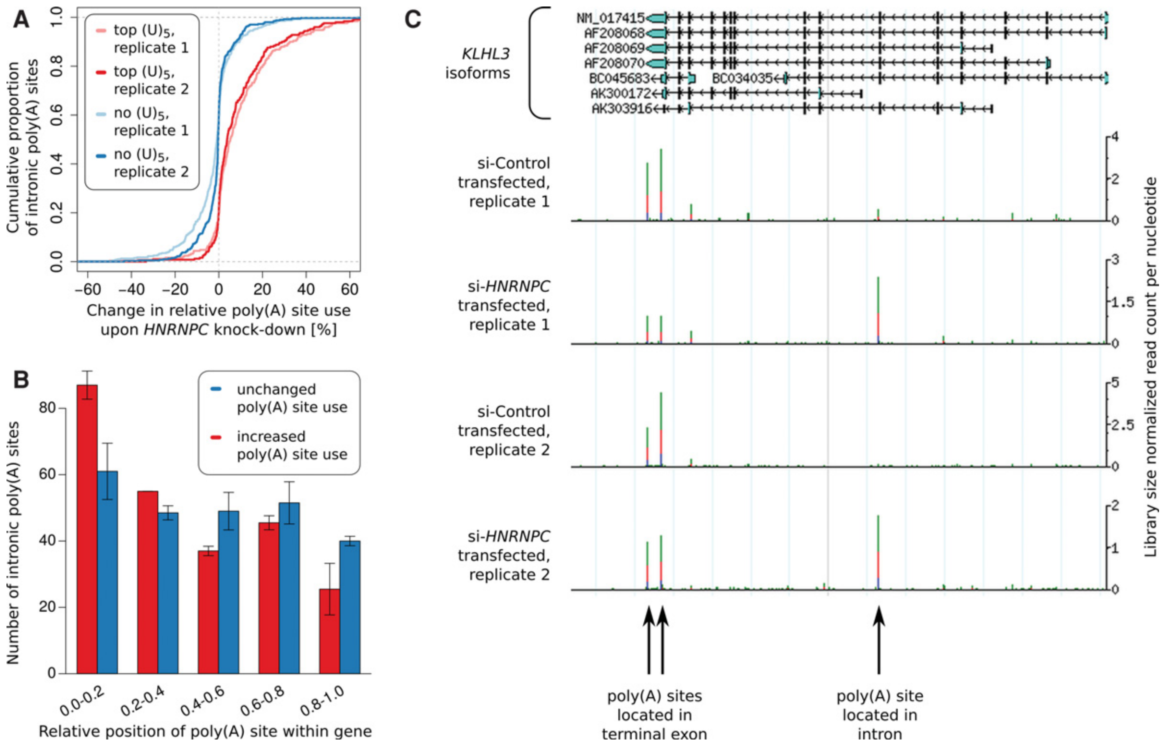


**Figure 6.** *HNRNPC* knock-down leads to increased usage of intronic poly(A) sites. (*A*) The change in the relative use of intronic poly(A) sites that did not contain any $(U)_5$ within ±200 nt and of the top 250 intronic poly(A) sites according to the number of $(U)_5$ motifs within ±50 nt around the cleavage site, upon *HNRNPC* knock-down. (*B*) Relative location within the gene of the top 250 most-derepressed intronic poly(A) sites that have HNRNPC binding motifs within −200 to +100 nt around their cleavage site and of the 250 intronic poly(A) sites that changed least upon *HNRNPC* knock-down. (*C*) Screenshot of the *KLHL3* gene, in which intronic cleavage and polyadenylation was strongly increased upon *HNRNPC* knock-down.

strongly enriched in $(U)_5$ tracts immediately downstream from the poly(A) site (Supplemental Fig. 18). This indicates that these poly(A) site–associated motifs contribute to the definition of these terminal exons. To further characterize the "masking" effect of HNRNPC on intronic poly(A) sites, we binned poly(A) sites into five groups based on their relative position within the host gene and asked how the position of sites within genes relates to their usage upon *HNRNPC* knock-down. As shown in Figure 6B, we found that intronic poly(A) sites that are most derepressed upon *HNRNPC* knock-down are preferentially located toward the 5′ ends of genes. We conclude that HNRNPC tends to repress the usage of intronic cleavage and polyadenylation sites whose usage leads to a strong reduction of transcript length. Figure 6C shows the example of the kelch like family member 3 (*KLHL3*) gene, which harbors one of the most derepressed intronic poly(A) sites.

## Discussion

Studies in recent years have shown that pre-mRNA cleavage and polyadenylation is a dynamically regulated process that yields transcript isoforms with distinct interaction partners, subcellular localization, stability, and translation rate (for review, see, e.g., Davis and Shi 2014). Specific polyadenylation programs seem to have evolved in relation with particular cell types or states. For example, APA and 3′ UTR lengths are developmentally regulated (Ji and Tian 2009; Ji et al. 2009; Miura et al. 2013), and short 3′ UTRs are generated in proliferating and malignant cells (Lee et al. 2007; Sandberg et al. 2008; Xia et al. 2014). The key regulators of these polyadenylation programs are unknown. Reduced expression of the U1 snRNP (Berg et al. 2012) or of the mammalian cleavage factor I ($CFI_m$) components NUDT21 and CPSF6 (Gruber et al. 2012; Martin et al. 2012) can cause a systematic reduction in 3′ UTR lengths, but only limited evidence about the relevance of these factors in physiological conditions has been provided (Berg et al. 2012; Masamha et al. 2014). Other factors that are part of the 3′ end processing machinery and have systematic effects on polyadenylation are the poly(A) binding protein nuclear 1 (Jenal et al. 2012), which suppresses cleavage and polyadenylation; the 64-kDa cleavage stimulation factor subunit 2 (CSTF2) component of the 3′ end cleavage and polyadenylation complex, whose expression correlates with the preferential use of short 3′ UTRs in cancer cells (Xia et al. 2014); and the retinoblastoma binding protein 6, whose reduced expression results in reduced transcript levels and increased use of distal poly(A) sites (Di Giammartino et al. 2014).

Many experimental protocols to capture transcript 3′ ends and enable studies of the dynamics of polyadenylation have been developed (for review, see de Klerk et al. 2014), and consequently, a few databases of 3′ end processing sites are available (Lee et al. 2007; Derti et al. 2012; You et al. 2014). However, none of these databases has used the entire set of 3′ end sequencing data available to date, and thus, their coverage is limited. In this study, we have developed a procedure to automatically process heterogeneous data sets generated with one of nine different protocols, aiming to identify bona fide poly(A) sites that are independently regulated. Although most of the reads that were used to construct the currently available databases (Derti et al. 2012; You et al. 2014) map within the poly(A) site clusters that we constructed, the differences at the level of reported processing sites are quite large. This is largely due to the presence of many sites with very limited read support and no upstream poly(A) signals in previous data sets. For example, focusing on the terminal exons of protein-coding genes and lincRNAs from the UCSC GENCODE v19 Basic Set annotation, the human atlas that we constructed has a higher fraction of exons with assigned poly(A) sites compared with previous databases; 71.12% of all terminal exons of protein coding genes in our atlas have at least one annotated poly(A) site in contrast to 66.26% and 62.69% for the studies of Derti et al. (2012) and You et al. (2014), respectively. The coverage of the terminal exons of lincRNAs is smaller overall but is clearly higher in our atlas (37.59%) compared with those of Derti et al. (2012) and You et al. (2014) (29.57% and 24.51%, respectively) (Supplemental Fig. 19). The lower coverage of lincRNAs is probably due to their lower expression in comparison with protein-coding genes (Wu et al. 2014) and to the fact that some of them are bimorphic, appearing in both the poly(A)$^+$ and poly(A)$^-$ fraction (Hangauer et al. 2013), and cannot be captured efficiently with protocols that require the presence of a poly(A) tail.

Although for the mouse we did not have lincRNA annotations, the general trend of higher coverage in our atlas compared with existing ones holds also for mouse genes (Supplemental Fig. 20; for detailed numbers, see Supplemental Tables 7, 8).

The 3′ end processing sites reported by other studies (Derti et al. 2012; You et al. 2014) but missing from our atlas have, on average, a substantially lower read support. Some were only documented by multimapping reads, had features indicative of internal priming, or originated in regions from which broadly scattered reads were generated.

By building upon a large set of 3′ end sequencing samples, we have analyzed the sequence composition around high-confidence poly(A) sites to identify elements that may recruit RBPs to modulate polyadenylation. We have identified sequence motifs that exhibit a positional preference with respect to 3′ end cleavage sites almost identical to the canonical poly(A) signal AAUAAA. Six of the 10 novel motifs that we found in each human and mouse data set are shared. Not all the poly(A) sites in the atlas that we constructed have one of the 18 conserved signals, which suggests that the set of poly(A) signals is still incomplete. However, with a more comprehensive set of poly(A) signals, we have been able to more efficiently use data from many heterogeneous experiments, thereby achieving a higher coverage of terminal exons and annotated genes by poly(A) sites. Even though the poly(A) and poly(U) motifs are also strongly enriched around poly(A) sites, they were not annotated as poly(A) signals due to positional profiles divergent from what is expected for poly(A) signals. The general A- and U-richness in the vicinity of cleavage and polyadenylation sites has been observed before (Tian and Graber 2012), but the RBP interactors and their role in polyadenylation remain to be characterized.

Here we hypothesized that HNRNPC, a protein that binds poly(U) tracts (Ray et al. 2013; Cieniková et al. 2014; Liu et al. 2015) and has a variety of functions including pre-mRNA splicing (König et al. 2010) and mRNA transport (McCloskey et al. 2012), also modulates the processing of pre-mRNA 3′ ends. HNRNPC has originally been identified as a component of the HNRNP core particle (Beyer et al. 1977; Choi and Dreyfuss 1984) and found to form stable tetramers that bind to nascent RNAs (Whitson et al. 2005). Systematic evolution of ligands by exponential enrichment (SELEX) experiments have shown that HNRNPC particles bind to uninterrupted tracts of five or more uridines (Görlach et al. 1994), and studies employing CLIP indicated that longer tracts are bound with higher affinity (König et al. 2010). By sequencing mRNA 3′ ends following the siRNA-mediated knock-down of *HNRNPC*, we found that transcripts that contain poly(U) tracts around their poly(A) sites respond in a manner indicative of HNRNPC masking poly(A) sites. This is reminiscent of the U1 snRNP protecting nascent RNAs from premature cleavage and polyadenylation, in a mechanism that has been called “telescripting” (Kaida et al. 2010; Berg et al. 2012). Indeed, HNRNPC seems to have at least in part a similar function, because the knock-down of *HNRNPC* increased the incidence of cleavage and polyadenylation at intronic sites, with a preference for intronic sites close to the transcription start. It should be noted that these intronic sites are not spurious but have experimental support as well as polyadenylation signals. Thus, the short transcripts that terminate at these sites could be functionally relevant, either through the production of truncated proteins or through an effective down-regulation of the functional, full-length transcript forms. In terminal exons, U-rich poly(A) sites whose usage increased upon *HNRNPC* knock-down tended to be located distally. In these transcripts, HNRNPC may function to “mask” the distal, “stronger” signals, allowing the “weaker” proximal poly(A) sites to be used (Shi 2012). Interestingly, the competition between HNRNPC and U2AF2 appears to regulate exonization of *Alu* elements (Zarnack et al. 2013) and, furthermore, impacts polyadenylation at *Alu* exons (Tajnik et al. 2015). These studies have emphasized the complex cross-talk between regulators that come into play during RNA splicing and polyadenylation (Proudfoot 2011). They also illustrate the striking multifunctionality of U-rich and A/U-rich elements that are bound by various proteins at different stages to modulate processes ranging from transcription termination (Almada et al. 2013) up to protein localization (Berkovits and Mayr 2015).

Initial studies that reported 3′ UTR shortening in dividing cells hypothesized that shortened 3′ UTRs harbor a reduced number of miRNA binding sites, the corresponding mRNAs being more stable and having an increased translation rate (Sandberg et al. 2008; Mayr and Bartel 2009). However, genome-wide measurements of mRNA and protein levels in dividing and resting cells revealed that systematic 3′ UTR shortening has a relatively minor impact on mRNA stability, translation, and protein output (Spies et al. 2013; Gruber et al. 2014). Instead, evidence has started to emerge that 3′ UTR shortening results in the loss of interaction with various RBPs, whose effects are not limited to mRNA stability and translation (Gupta et al. 2014) but reach as far as the transport of transmembrane proteins to the plasma membrane (Berkovits and Mayr 2015). The CD47 protein provides a striking example of 3′ UTR–dependent protein localization. However, the upstream signals and perhaps additional targets of this mechanism remain to be uncovered. Here we have demonstrated that HNRNPC can modulate polyadenylation of a large number of transcripts, leading to the inclusion or removal of U-rich elements. When these elements remain part of the 3′ UTRs, they can be subsequently bound by a variety of U-rich element binding proteins, including ELAVL1, which has been recently demonstrated to play a decisive role in the UDPL of *CD47* (Berkovits and Mayr 2015). Indeed, we found that the knock-down of *HNRNPC* promoted the expression of the long *CD47* 3′ UTR that is accompanied by an increased membrane localization of the CD47 protein. Although HNRNPC did not appear to target any particular class of transcripts, nearly one-quarter (>23%) of the HNRNPC-responsive transcripts encoded proteins that were annotated with the Gene Ontology category “integral component of membrane” (GO:0016021). Thus, our results provide an extended set of candidates for the recently discovered UDPL mechanism.

In conclusion, PolyAsite, available at http://www.polyasite.unibas.ch, is a large and extendable resource that supports investigations into the polyadenylation programs that operate during changes in cell physiology, during development, and in malignancies.

## Methods

### Uniform processing of publicly available 3′ end sequencing data sets

Publicly available 3′ end sequencing data sets were obtained from the NCBI GEO archive (www.ncbi.nlm.nih.gov/geo) and from NCBI SRA (www.ncbi.nlm.nih.gov/sra). To ensure uniform processing of 3′ end sequencing data generated by diverse 3′ end sequencing protocols, we developed the following computational pipeline (Supplemental Fig. 21). First, raw sequencing files were converted to FASTA format. For samples generated with protocols that leave a 5′ adapter sequence in the reads, we only retained the reads from which the specified adapter sequence could be trimmed. Next, we trimmed the 3′ adapter sequence, and when the protocol captured the reverse complement of the RNAs, we reverse complemented the reads. Reads were then mapped to the corresponding genome assembly (hg19 and mm10, respectively) and to mRNA and lincRNA-annotated transcripts (GENCODE v14 release for human [Harrow et al. 2012] and Ensembl annotation of mouse [Flicek et al. 2013], both obtained from UCSC [Meyer et al. 2013] in June 2013). The sequence alignment was done with segemehl with default parameters (Hoffmann et al. 2009). In cases where the sex of the organism from which the sample was prepared was female, mappings to the Y Chromosome were excluded from further analysis. For each read, we only kept the mappings with the highest score (smallest edit distance). Mappings overlapping splice junctions were only retained if they covered at least 5 nt on both sides of the junction and they had a higher score compared with any mapping of the same read to the genomic sequence. Based on the genome coordinates of individual exons and the mapping coordinates of reads within transcripts, next we converted read-to-transcript mapping coordinates into read-to-genome mapping coordinates. For generating a high-confidence set of pre-mRNA 3′ ends, we started from reads that consisted of no more than 80% of adenines and that mapped uniquely to the genome such that the last 3 nt of the read were perfectly aligned. Furthermore, we required that the 3′ end of the read was not an adenine and collapsed the 3′ ends of the sequencing reads into putative 3′ end processing sites. Finally, we filtered out those sites that showed one of the following patterns: one of the AAAA, AGAA, AAGA, or AAAG tetramers immediately downstream from the apparent cleavage site; or six consecutive or more than six adenines within the 10 nt downstream from the apparent cleavage site. We empirically found that these patterns were associated with many spurious poly(A) sites (for details on the entire pipeline, see Supplemental Fig. 21).

### Clustering of closely spaced 3′ end sites into 3′ end processing regions

Putative 3′ end processing sites identified as described above were used to construct clusters to (1) identify poly(A) signals, (2) derive sample-specific cutoffs for the number of reads necessary to support a site, and (3) determine high-confidence 3′ end processing sites in the human and mouse genomes. In clustering putative 3′ end processing sites from multiple samples, as done for analyses 1 and 3, we first sorted the list of 3′ end sites by the number of supporting samples and then by the total normalized read count (read counts were normalized per sample as reads per million [RPM], and for each site a total RPM was obtained by summing these numbers over all samples). In contrast, to generate clusters of putative reads from individual samples (analysis 2), we only ranked genomic positions by RPM. Clusters were generated by traversing the sorted list from top to bottom and associating lower-ranking sites with a representative site of a higher rank, if the lower-ranked sites were located within a specific maximum distance upstream ($d_\mathrm{u}$) of, or downstream ($d_\mathrm{d}$) from, the representative site (Supplemental Fig. 22).

To determine a maximum distance between sites that seem to be under the same regulatory control, we applied the above-described clustering procedure for distances $d_\mathrm{u}$ and $d_\mathrm{d}$ varying between 0 and 25 nt and evaluated how increasing the cluster length affects the number of generated clusters that contain more than one site (Supplemental Fig. 23). Consistent with previous observations, we found that at a distance of 8 nt from the representative site, ∼40% of the putative 3′ end processing sites are part of multisite clusters; this proportion increases to 43% for a distance of 12 nt and reaches 47% at a distance of 25 nt. For consistency with previous studies, we used $d_\mathrm{u} = d_\mathrm{d} = 12$ nt (Tian et al. 2005; You et al. 2014). Only for the clustering of putative 3′ end processing sites in individual samples, we used a larger distance, $d_\mathrm{u} = d_\mathrm{d} = 25$, resulting in a more conservative set of clusters, with a maximum span of 51 nt.

### Identification of poly(A) signals

To obtain a set of high-confidence 3′ end processing sites from which to identify poly(A) signals, we filtered the preliminary 3′ end clusters, retaining only those that were supported by data from at least two protocols. For clusters with at least two putative sites, we took the center of the cluster as the representative cleavage site. Then, we constructed the positional frequency profile in the −60 to −5 nt region upstream of the representative cleavage sites for each of the 4096 possible hexamers (Supplemental Fig. 24A). We did not consider the 5 nt upstream of the putative cleavage sites to reduce the impact of artifacts originating from internal priming at poly(A) nucleotides, which are very close in sequence to the main poly(A) signal, AAUAAA (see below for details on "PAS priming sites"). Before fitting a specific functional form to the frequency profiles, we smoothed them, taking at each position the average frequency in a window of 11 nt centered on that position, and we subtracted a motif-specific "background" frequency which we defined as the median of the 10 smallest frequencies of the motif in the entire 55-nt window. To identify motifs that have a specific positional preference upstream of the cleavage site, we fitted a Gaussian density curve to the background-corrected frequency profile with the "nls" function in R (R Core Team 2014), assessing the quality of the fit by the $r^2$ value and by the height:width ratio of the fitted peak, where the width was defined as the standard deviation of the fitted Gaussian density (Supplemental Fig. 24A). Alternative poly(A) signals should have the same positional preference as the main signal, AAUAAA. However, when considering 60 nt upstream of the cleavage site, poly(A) signals can occur not only at −21 nt, which seems to be the preferred location of these signals, but also at other positions, particularly when the poly(A) signal is suboptimal and co-occurs with the main signal. Thus, we started from motifs that peaked in the region upstream of the cleavage site ($r^2 \geq 0.6$ for the fit to the Gaussian and a height:width ratio ≥5) but allow a permissive position of the peak, between −40 to −10 nt. Putative poly(A) signals were then determined according to the following iterative procedure (Supplemental Fig. 24B).

We sorted the set of putative signals by their strength. The strongest signal was considered to be the one with the lowest

*P*-value of the test that the peak frequency of the motif could have been generated by Poisson sampling from the background rate inferred as the mean motif frequency in the regions of 100 to 200 nt upstream of and downstream from the cleavage site. As expected, in both human and mouse data sets, the most significant hexamer was the canonical poly(A) signal AAUAAA. Before every iteration, we removed all sequences that contained the most significant signal of the previous iteration in the −60-nt window upstream of the cleavage sites and repeated the procedure on the remaining set of sequences. Signals with an $r^2$ value of the fit to a Gaussian ≥0.9 and a height:width ratio ≥4 were retained and the most significant added to the set of potential signals. The fitted Gaussian densities of almost all of the putative poly(A) signals recovered with this procedure had highly similar peak positions and standard deviations. Therefore, only signals that peaked at most 1 nt away from the most significant hexamer, AAUAAA, were retained in the final set of poly(A) signals. The only hexamers that did not satisfy this condition were the AAAAAA hexamer in the mouse and AAAAAA as well as UUAAAA in the human.

### Treatment of putative 3′ end sites originating from internal priming

Priming within A-rich, transcript-internal regions rather than to the poly(A) tail is known to lead to many false-positive sites with most of the existing 3′ end sequencing protocols. We tried to identify and eliminate these cases as described above. An underappreciated source of false positives seems to be the annealing of the poly(T) primer in the region of the poly(A) signal itself, which is A-rich and close to the poly(A) site (Tian et al. 2005; Shi 2012). Indeed, a preliminary inspection of cleavage sites that seemed to lack poly(A) signals revealed that these sites were located on or in the immediate vicinity of a motif that could function as a poly(A) signal. To reduce the rate of false positives generated by this mechanism, we undertook an additional filtering procedure as follows (Supplemental Fig. 25). First, every 3′ end site that was located within a poly(A) signal or had a poly(A) signal starting within 5 nt downstream from the apparent cleavage site was marked initially as “PAS priming site.” Then, during the clustering procedure, each cluster that contained a “PAS priming site” was itself marked as putative internal priming candidate, and the most downstream position of the cluster was considered as the representative site for the cluster. Finally, internal priming candidate clusters were either (1) merged into a downstream cluster, if all annotated poly(A) signals of the downstream cluster were also annotated for the internal priming candidate, or (2) retained as valid poly(A) cluster when the distance between the representative site to the closest poly(A) signal upstream was at least 15 nt or (3) discarded, if neither condition (1) nor (2) was met.

### Generation of the comprehensive catalog of high-confidence poly(A) sites

#### *Annotating poly(A) signals*

The procedure outlined in the sections above yielded 18 signals that showed a positional preference similar to AAUAAA in both mouse and human. These signals were used to construct the catalog of 3′ end processing sites. We started again from all unique apparent cleavage sites from the 78 human and 110 mouse samples (Supplemental Tables 3, 4), amounting to 6,983,499 and 8,376,450 sites, respectively. For each of these sites, we annotated all occurrences of any of the 18 poly(A) signals within −60 to +5 nt relative to the apparent cleavage site.

#### *Identification of 3′ end processing clusters expressed above background in individual samples*

For each sample independently, we constructed clusters of 3′ end processing sites as described above. At this stage, we did not eliminate “PAS priming sites” but rather used a larger clustering distance, of $d_u = d_d = 25$, to ensure that “PAS priming sites” were captured as well. We kept track of whether any 3′ end processing site in each cluster had an annotated poly(A) signal or not. Next, we sorted the clusters by the total number of reads that they contained, and by traversing the sorted list from top (clusters with most reads) to bottom, we determined the read count *c* at which the percentage of clusters having at least one annotated poly(A) signal dropped below 90%. We then discarded all clusters with ≤ *c* read counts as not having sufficient experimental support (for outlines how to determine sample-specific cutoffs, Supplemental Fig. 26). This allowed for an efficient filtering of reads presumably representing background noise.

#### *Combining poly(A) site clusters from all samples into a comprehensive catalog of 3′ end processing sites*

By starting from the sites identified in at least one of the samples, we first normalized the read counts to the total number of reads in each sample to compute expression values as RPM and then merged all sites into a unique list that we sorted first by the number of protocols supporting each individual site and then by the total RPM across all samples that supported the site. These sites were clustered, and then internal priming candidates were eliminated as described above. Closely spaced clusters were merged (1) when they shared the same poly(A) signals or (2) when the length of the resulting cluster did not exceed 25 nt. The above procedure could result in poly(A) clusters that were still close to each other but with a combined length exceeding the maximum cluster size and that did not have any poly(A) signal annotated. To retain from these the most likely and distinct poly(A) sites, we merged clusters without poly(A) signals with an inter-cluster distance ≤12 nt and retained those whose total cluster span was ≤50 nt. A small fraction of the clusters had a span ≥50 nt, with some even wider than 100 nt. These clusters were not included in the atlas. Finally, the position with the highest number of supporting reads in each cluster was reported as the representative site of the cluster (Supplemental Fig. 27). The final set of clusters was saved in a BED-formatted file, with the number of supporting protocols as the cluster score. A cluster obtained support by a protocol if any of the reads in the clusters originated from that protocol. We used the protein-coding and lincRNA annotations from the UCSC GENCODE v19 Basic Set for human and the Ensembl mm10 transcript annotation from UCSC for mouse to annotate the following categories of clusters, listed here in the order of their priority (which we used to resolve annotation ambiguity):

TE—terminal exon,<br>
EX—any other exon except the terminal one,<br>
IN—any intron,<br>
DS—up to 1000 nt downstream from an annotated gene,<br>
AE—antisense to an annotated exon,<br>
AI—antisense to an annotated intron,<br>
AU—antisense and within 1000 nt upstream of an annotated gene, and<br>
IG—intergenic.

#### *Supplemental atlas versions*

To provide more details on different aspects of the inferred poly(A) site clusters, additional versions of the human and mouse atlas with extended information were generated. For human, we established a version that annotated one of the above categories to every poly(A) site cluster based on the UCSC GENCODE v19 Comprehensive Set annotation (not limited to protein-coding and lincRNA-encoding genes). Moreover, for mouse and human, a version with additional information about the tissues/cell types in which each poly(A) site was identified was constructed. All versions are publicly available and online at http://www.polyasite.unibas.ch.

#### *Sequence logos of the identified poly(A) signals*

The procedure described above was used again, this time to construct a version of the human and mouse poly(A) site atlases that incorporated the entire set of 22 organism-specific poly(A) signals, not just the 18 signals that were shared between species. Frequencies of all annotated poly(A) signals (possibly more than one per poly(A) cluster) across all identified clusters were calculated for the human and mouse catalog independently. FASTA files with poly(A) signals, including their multiplicities in the data, were used with the Weblogo program (Crooks et al. 2004) version 3.3, with default settings, to generate the sequence logos for human and mouse, respectively.

#### *Hexamer enrichment in upstream regions of 3′ end clusters*

We calculated the significance (*P*-value) of enrichment of each hexamer in the set of 3′ end clusters (and their 60 nt upstream regions) of our human and mouse atlas relative to what would be expected by chance, assuming the mononucleotide frequencies of the sequences and a binomial distribution of motif counts.

#### *Annotation of poly(A) sites with respect to categories of genomic regions*

We used the genomic coordinates of the protein-coding genes and lincRNAs from the UCSC GENCODE v19 Basic Set (human) and the Ensembl mm10 (mouse) annotations to annotate our and previously published sets of poly(A) sites with respect to genomic regions with which they overlap. A poly(A) site was assigned to an annotated feature if at least one of its genomic coordinates overlapped with the genomic coordinates of the feature.

PolyAsite: For every poly(A) cluster annotated in our catalog, the entire region of the cluster was used to test for an overlap with annotated genomic features.

PolyA-seq: Processed, tissue-specific data were downloaded as a BED file (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30198). Poly(A) sites from nine and five different samples were downloaded for human and mouse, respectively (Derti et al. 2012). Mouse genome coordinates were converted to the coordinates of the Ensembl mm10 annotation through LiftOver (Hinrichs et al. 2006). The genomic coordinates of all poly(A) sites (one position per poly(A) site) were intersected with the annotation features.

APASdb: Processed, tissue-specific data for human poly(A) sites were downloaded from http://mosas.sysu.edu.cn/utr/download_datasets.php. This included poly(A) sites from 22 human tissues (You et al. 2014). The genomic coordinates of all poly(A) sites (one position per poly(A) site) were intersected with the annotation features.

### Analysis of 3′ end libraries from *HNRNPC* knock-down experiments

#### *Sequencing of A–seq2 libraries and quantification of relative poly(A) site usage*

We considered all high-confidence A-seq2 (Gruber et al. 2014) reads that mapped to a unique position in the human genome (hg19) and that had 5′ ends that were located in a cluster supported by two or more protocols. For our A-seq2 protocol, high-confidence reads are defined as sequencing reads that do not contain more than two ambiguous bases (N), have a maximum A-content of 80%, and the last nucleotide is not an adenine. By using our atlas of poly(A) sites that was constructed considering the 18 conserved poly(A) signals, we calculated the relative usage of poly(A) sites. We considered in our analysis all exons that had multiple poly(A) clusters expressed at >3.0 RPM in one or more samples. There were 12,136 such clusters. We considered as “consistently” changing poly(A) sites those that had a change of at least 5% in the same direction in both replicates. We considered as “consistently” unchanged poly(A) sites those whose mean change and standard deviation across replicates were <2%.

#### *Determination of ELAVL1 binding sites that are affected by APA events taking place upon HNRNPC knock–down*

Determination of 3′ UTR regions that respond to *HNRNPC* knock-down: To identify putative HNRNPC regulated regions, we have selected exons that had exactly two poly(A) sites, one of which showing an increase in relative usage by at least 5% upon *HNRNPC* knock-down and harboring a putative HNRNPC binding site ($(U)_5$) within a region of −200 to 100 nt relative to the cleavage site. We considered as unchanged regions exons with exactly two poly(A) sites, both of which changing <5% upon *HNRNPC* knock-down.

ELAVL1 binding site extraction from PAR-CLIP: We used data from a previously published ELAVL1 CLIP experiment (Kishore et al. 2011), Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) database accession GSM714641. Enriched binding sites were determined by applying the mRNA site extraction tool available on CLIPZ (Khorshid et al. 2011; Jaskiewicz et al. 2012) using the mRNA-seq samples with GEO accessions GSM714684 and GSM714685 as background. CLIP sites with an enrichment score ≥5.0 were translated into genome coordinates (hg19) using GMAP (Wu and Watanabe 2005). To identify ELAVL1 CLIP sites located within transcript regions that are included/excluded through APA, we intersected the set of enriched ELAVL1 CLIP sites with genomic regions enclosed by tandem poly(A) sites (located on the same exon) using BEDTools (Quinlan and Hall 2010).

### Determination of intronic poly(A) sites

To make sure that we can capture premature cleavage and polyadenylation events that might occur spontaneously upon knock-down of *HNRNPC* and are therefore observable in the *HNRNPC* knock-down samples only, for each sample we created clusters as described above, using conserved poly(A) signals only. By analogy to tandem poly(A) sites within exons, we calculated the relative usage of clusters within genes by considering all genes having multiple poly(A) clusters that were expressed at >3.0 RPM in one or more samples. There were 22,498 such clusters, 2454 of which were annotated to be intronic. Finally, we determined the set of sites that showed a consistent change upon *HNRNPC* knock-down as described above.

### Cell culture and RNAi

HEK 293 cells (Flp-In-293, from Life Technologies) were grown in Dulbecco’s modified Eagle’s medium (DMEM; from Sigma) supplemented with 2 mM L-glutamine (Gibco) and 10% heat-inactivated fetal calf serum (Gibco). Transfections of siRNA were carried out using Lipofectamine RNAiMAX (Life Technologies) following the manufacturer’s protocol. The following siRNAs were used:

negative-control from Microsynth (sense strand AGGUAGUGUA UCGCCUUGTT) and si-*HNRNPC*1/2 (sc-35577 from Santa Cruz Biotechnologies), both applied at 20 nM in 2.5 mL DMEM on six-well plates.

### Western blotting

Cells were lysed in 1 × RIPA buffer, and protein concentration was quantified using BCA reagent (Thermo Scientific). A stipulated amount of the sample (usually 10 µg) was then used for SDS gel separation and transferred to ECL membrane (Protran, GE Healthcare) for further analysis. Membranes were blocked in 5% skim milk (Migros) in TN-Tween (20 mM Tris-Cl at pH 7.5, 150 mM NaCl, 0.05% Tween-20). The following antibodies were used for Western blots: Actin, sc-1615 from Santa Cruz Biotechnology; hnRNP C1/C2 (N-16), sc-10037 from Santa Cruz Biotechnology (used at 1:1000 dilution); CD47, AF-4670 from R&D Systems (used at 1:200 dilution). HRP-conjugated secondary antibodies were applied at 1:2000 dilution. After signal activation with ECL Western blotting detection reagent (GE Healthcare), imaging of Western blots was performed on an Azure c600 system. Signal quantification was done with ImageJ software.

### Immunofluorescence

For the immunofluorescence analysis, HEK 293 cells were transfected with either control siRNA or siRNAs targeting *HNRNPC* as described under Cell Culture and RNAi, 48 h post transfection cells were fixed with 4% paraformaldehyde for 30 min, permeabilized, and blocked with PBS containing 1% BSA and 0.1% Triton X-100 for 30 min. Primary anti-CD47 antibody (sc-59079 from Santa Cruz Biotechnology) was incubated for 2 h at room temperature at a dilution of 1:100 in the same buffer. To visualize CD47 in cells, secondary antibody conjugated with Alexa Fluor 488 was applied, while the nucleus was labeled with Hoechst dye. Imaging was performed with a Nikon Ti-E inverted microscope adapted with a LWD condenser (WD 30mm; NA 0.52), Lumencor SpectraX light engine for fluorescence excitation LED transmitted light source. Cells were visualized with a CFI Plan Apochromat DM 60× lambda oil (NA 1.4) objective, and images were captured with a Hamatsu Orca-Flash 4.0 CMOS camera. Image analysis and edge detection was performed with NIKON NIS Elements software version 4.0. All images were subsequently adjusted uniformly and cropped using Adobe Photoshop CS5.

### FACS analysis

FACS analyses of siRNA transfected cells were performed similar to immunofluorescence studies (see above) except that cells were not permeabilized prior to the treatment with antibody against CD47 (sc-59079 from Santa Cruz Biotechnology). Analysis of Alexa Fluor 488 signal and counts was carried out on a BD FACS Canto II instrument, and data were analyzed with the FLOWJO software. An equal pool of siRNA samples from each transfection set was mixed for the IgG control staining to rule out nonspecific signals.

### PAR-CLIP and A-seq2 libraries

A-seq2 libraries were generated as previously described (Gruber et al. 2014) and sequenced on an Illumina HiSeq 2500 sequencer. The HNRNPC PAR-CLIP was performed as previously described (Martin et al. 2012) with a modification consisting of preblocking of the Dynabeads–Protein A (Life Technologies), resulting in reduced background and higher efficiency of library generation. To this end, Dynabeads were washed three times with PN8 buffer (PBS buffer with 0.01% NP-40), and incubated in 0.5 mL of PN8-preblock (1 mM EDTA, 0.1% BSA from Sigma [A9647], and 0.1 mg/mL heparin from Sigma [H3393], in PN8 buffer) for 1 h on a rotating wheel. The preblock solution was removed and replaced by the antibody in 0.2 mL preblock solution and rotated for 2–4 h. We used the goat polyclonal antibody sc-10037 against HNRNPC (Santa Cruz Biotechnology). The 5′ adapter was GTTCAGAGTTCTACAGTCCGACGATC and the 3′ adapter was TGGAATTCTCGGGTGCCAAGG.

### HNRNPC PAR-CLIP analysis

The raw data were mapped using CLIPZ (Khorshid et al. 2011). For each poly(A) site, the uniquely mapping reads that overlapped with a region of ±50 nt around the cleavage site were counted and normalized (divided) by the expression level (RPKM) of the poly(A) sites host gene using the mRNA-seq samples with GEO accession GSM714684. For Supplemental Figure 11, normalized CLIP read counts of poly(A) sites belonging to different categories of consistently behaving poly(A) sites across replicates, as defined above, were used.

### Analysis of mRNA-seq libraries from *HNRNPC* knock-down experiments

Publicly available libraries of *HNRNPC* knock-down and control experiments (two replicates) that have been published recently (Liu et al. 2015) were downloaded from the sequence read archive (SRA) database of the National Center for Biotechnology Information (accession numbers SRX699496/GSM1502498, SRX699497/GSM1502499, SRX699498/GSM1502500, and SRX699499/GSM1502501). After adapter removal, the FASTQ file containing the reads sequenced in sense direction was mapped using the STAR aligner with default settings (Dobin et al. 2013).

#### *Evaluation of novel exon vs. extended internal exon contribution to intronic poly(A) sites*

First we identified all poly(A) sites that were located in introns according to gene structures reflected in the GENCODE v19 (human) transcript set and that were putative HNRNPC targets. That is, they were consistently derepressed upon knock-down of *HNRNPC* (see above) and contained putative HNRNPC-binding $(U)_5$ motifs within −200 to +100 nt around their cleavage site. For each of these intronic sites, we determined the closest upstream exon, here referred to as u-exon. To find out whether this type of poly(A) sites represented the 3′ ends of novel terminal exons or of extended versions of the u-exon, we calculated the ratio $R = \frac{S+1}{C+1}$, where $C$ is the number of reads that map over the 3′ end of the u-exon (extending by at least 10 nt in the downstream region), and $S$ is the number of reads that map across a splice boundary, the 5′ splice site (ss) being within ±3 nt of the 3′ end of the u-exon and the 3′ end of the read mapping upstream of the intronic poly(A) site. The $C$ type of reads provide evidence for the extension of the u-exon, whereas the $S$ type of reads provide evidence for a novel terminal exon. In order to prevent artifacts that may result from poorly expressed transcripts, we required the u-exon to intersect with at least 10 reads within a sample, and we only included regions for which we had at least three reads of either $C$ or $S$ type (or both). We used a pseudo-count of one for both read types.

## Data access

The sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under accession number SRP065825.

## Acknowledgments

We thank Erik van Nimwegen for his input on data analysis, Béatrice Dimitriades for technical assistance, and Josef Pasulka for suggestions on the analysis. This work was supported by the Swiss National Science Foundation grant 31003A-143977 to W. K. and by the Swiss National Science Foundation NCCR project "RNA & Disease" (51NF40_141735).

*Author contributions:* M.Z., A.J.G., A.R.G., and W.K. designed the project. A.R.G., R.S., and A.J.G. collected data sets and created the catalog of poly(A) sites. M.B. developed the PolyAsite web interface. R.S. and A.J.G. identified poly(A) signals. G.M. performed the HNRNPC PAR-CLIP and A-seq2 experiments. A.J.G. analyzed the data with help from R.S. G.M. and S.G. performed the experiments on CD47. M.Z. supervised the project. A.J.G., M.Z., R.S., and A.R.G. wrote the manuscript. All authors read and approved the final manuscript.

## References

Alkan SA, Martincic K, Milcarek C. 2006. The hnRNPs F and H2 bind to similar sequences to influence gene expression. *Biochem J* **393:** 361–371.

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

Arhin GK, Boots M, Bagga PS, Milcarek C, Wilusz J. 2002. Downstream sequence elements with different affinities for the hnRNP H/H′ protein influence the processing efficiency of mammalian polyadenylation signals. *Nucleic Acids Res* **30:** 1842–1850.

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

Beck AH, Weng Z, Witten DM, Zhu S, Foley JW, Lacroute P, Smith CL, Tibshirani R, van de Rijn M, Sidow A, et al. 2010. 3′-end sequencing for expression quantification (3SEQ) from archival tumor samples. *PLoS One* **5:** e8768.

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

Berkovits BD, Mayr C. 2015. Alternative 3′ UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522:** 363–367.

Beyer AL, Christensen ME, Walker BW, LeStourgeon WM. 1977. Identification and characterization of the packaging proteins of core 40S hnRNP particles. *Cell* **11:** 127–138.

Brennan CM, Steitz JA. 2001. HuR and mRNA stability. *Cell Mol Life Sci* **58:** 266–277.

Castelo-Branco P, Furger A, Wollerton M, Smith C, Moreira A, Proudfoot N. 2004. Polypyrimidine tract binding protein modulates efficiency of polyadenylation. *Mol Cell Biol* **24:** 4174–4183.

Chan SL, Huppertz I, Yao C, Weng L, Moresco JJ, Yates JR, Ule J, Manley JL, Shi Y. 2014. CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3′ processing. *Genes Dev* **28:** 2370–2380.

Choi YD, Dreyfuss G. 1984. Isolation of the heterogeneous nuclear RNA–ribonucleoprotein complex (hnRNP): a unique supramolecular assembly. *Proc Natl Acad Sci* **81:** 7471–7475.

Cieniková Z, Damberger FF, Hall J, Allain FH-T, Maris C. 2014. Structural and mechanistic insights into poly(uridine) tract recognition by the hnRNP C RNA recognition motif. *J Am Chem Soc* **136:** 14536–14544.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14:** 1188–1190.

Davis R, Shi Y. 2014. The polyadenylation code: a unified model for the regulation of mRNA alternative polyadenylation. *J Zhejiang Univ Sci B* **15:** 429–437.

de Hoon M, Hayashizaki Y. 2008. Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *Biotechniques* **44:** 627–632.

de Klerk E, den Dunnen JT, ′t Hoen PAC. 2014. RNA sequencing: from tag-based profiling to resolving complete transcript structure. *Cell Mol Life Sci* **71:** 3537–3551.

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

Di Giammartino DC, Li W, Ogami K, Yashinskie JJ, Hoque M, Tian B, Manley JL. 2014. RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3′ UTRs. *Genes Dev* **28:** 2248–2260.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21.

FANTOM Consortium and the RIKEN PMI and CLST (DGT). 2014. A promoter-level mammalian expression atlas. *Nature* **507:** 462–470.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013. Ensembl 2013. *Nucleic Acids Res* **41:** D48–D55.

Görlach M, Burd CG, Dreyfuss G. 1994. The determinants of RNA-binding specificity of the heterogeneous nuclear ribonucleoprotein C proteins. *J Biol Chem* **269:** 23074–23078.

Graber JH, Cantor CR, Mohr SC, Smith TF. 1999. *In silico* detection of control signals: mRNA 3′-end-processing sequences in diverse species. *Proc Natl Acad Sci* **96:** 14055–14060.

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

Gruber AR, Martin G, Müller P, Schmidt A, Gruber AJ, Gumienny R, Mittal N, Jayachandran R, Pieters J, Keller W, et al. 2014. Global 3′ UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nat Commun* **5:** 5465.

Gupta I, Clauder-Münster S, Klaus B, Järvelin AI, Aiyar RS, Benes V, Wilkening S, Huber W, Pelechano V, Steinmetz LM. 2014. Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA–protein interactions. *Mol Syst Biol* **10:** 719.

Hangauer MJ, Vaughn IW, McManus MT. 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet* **9:** e1003569.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760–1774.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34:** D590–D598.

Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermüller J. 2009. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol* **5:** e1000502.

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

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

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. 2012. The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell* **149:** 538–553.

Ji Z, Tian B. 2009. Reprogramming of 3′ untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PLoS One* **4:** e8419.

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

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

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

Katz Y, Wang ET, Airoldi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7:** 1009–1015.

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

Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, et al. 2001. Functional annotation of a full-length mouse cDNA collection. *Nature* **409:** 685–690.

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

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

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

Lackford B, Yao C, Charles GM, Weng L, Zheng X, Choi E-A, Xie X, Wan J, Xing Y, Freudenberg JM, et al. 2014. Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *EMBO J* **33:** 878–889.

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

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

Li W, You B, Hoque M, Zheng D, Luo W, Ji Z, Park JY, Gunderson SI, Kalsotra A, Manley JL, et al. 2015. Systematic profiling of poly(A)+ transcripts modulated by core 3′ end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. *PLoS Genet* **11:** e1005166.

Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27:** 2380–2396.

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

Liu N, Dai Q, Zheng G, He C, Parisien M, Pan T. 2015. $N^6$-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature* **518:** 560–564.

MacDonald CC, Redondo J-L. 2002. Reexamining the polyadenylation signal: Were we wrong about AAUAAA? *Mol Cell Endocrinol* **190:** 1–8.

Martin G, Gruber AR, Keller W, Zavolan M. 2012. 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* **1:** 753–763.

Masamha CP, Xia Z, Yang J, Albrecht TR, Li M, Shyu A-B, Li W, Wagner EJ. 2014. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* **510:** 412–416.

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

McCloskey A, Taniguchi I, Shinmyozu K, Ohno M. 2012. hnRNP C tetramer measures RNA length to classify RNA polymerase II transcripts for export. *Science* **335:** 1643–1646.

Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, et al. 2013. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* **41:** D64–D69.

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

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

Nam J-W, Rissland OS, Koppstein D, Abreu-Goodger C, Jan CH, Agarwal V, Yildirim MA, Rodriguez A, Bartel DP. 2014. Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol Cell* **53:** 1031–1043.

Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM. 2009. Direct RNA sequencing. *Nature* **461:** 814–818.

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

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

Proudfoot NJ, Brownlee GG. 1976. 3′ non-coding region sequences in eukaryotic messenger RNA. *Nature* **263:** 211–214.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842.

R Core Team. 2014. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499:** 172–177.

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* **320:** 1643–1647.

Schönemann L, Kühn U, Martin G, Schäfer P, Gruber AR, Keller W, Zavolan M, Wahle E. 2014. Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. *Genes Dev* **28:** 2381–2393.

Sheets MD, Ogg SC, Wickens MP. 1990. 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* **18:** 5799–5805.

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

Shi Y. 2012. Alternative polyadenylation: new insights from global analyses. *RNA* **18:** 2105–2117.

Spies N, Burge CB, Bartel DP. 2013. 3′ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res* **23:** 2078–2090.

Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. 2014. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508:** 66–71.

Tajnik M, Vigilante A, Braun S, Hänel H, Luscombe NM, Ule J, Zarnack K, König J. 2015. Intergenic *Alu* exonisation facilitates the evolution of tissue-specific transcript ends. *Nucleic Acids Res* **43:** 10492–10505.

ten Klooster JP, Leeuwen Iv, Scheres N, Anthony EC, Hordijk PL. 2007. Rac1-induced cell migration requires membrane recruitment of the nuclear oncogene SET. *EMBO J* **26:** 336–345.

Tian B, Graber JH. 2012. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip Rev RNA* **3:** 385–396.

Tian B, Hu J, Zhang H, Lutz CS. 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* **33:** 201–212.

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

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

Whitson SR, LeStourgeon WM, Krezel AM. 2005. Solution structure of the symmetric coiled coil tetramer formed by the oligomerization domain of hnRNP C: implications for biological function. *J Mol Biol* **350:** 319–337.

Wilusz J, Feig DI, Shenk T. 1988. The C proteins of heterogeneous nuclear ribonucleoprotein complexes interact with RNA sequences downstream of polyadenylation cleavage sites. *Mol Cell Biol* **8:** 4477–4483.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21:** 1859–1875.

Wu Z, Liu X, Liu L, Deng H, Zhang J, Xu Q, Cen B, Ji A. 2014. Regulation of lncRNA expression. *Cell Mol Biol Lett* **19:** 561–575.

Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, Li W. 2014. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nat Commun* **5:** 5274.

You L, Wu J, Feng Y, Fu Y, Guo Y, Long L, Zhang H, Luan Y, Tian P, Chen L, et al. 2014. APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. *Nucleic Acids Res* **43:** D59–D67.

Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, Reyes A, Anders S, Luscombe NM, Ule J. 2013. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of *Alu* elements. *Cell* **152:** 453–466.

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

Zhao X, Oberg D, Rush M, Fay J, Lambkin H, Schwartz S. 2005. A 57-nucleotide upstream early polyadenylation element in human papillomavirus type 16 interacts with hFip1, CstF-64, hnRNP C1/C2, and polypyrimidine tract binding protein. *J Virol* **79:** 4270–4288.

*Received November 25, 2015; accepted in revised form May 31, 2016.*

# A comprehensive analysis of 3′ end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation

Andreas J. Gruber, Ralf Schmidt, Andreas R. Gruber, et al.

*Genome Res.* 2016 26: 1145-1159 originally published online July 5, 2016
Access the most recent version at doi:10.1101/gr.202432.115