TO WHAT END? - COMPUTATIONAL TOOLS TO UNCOVER REGULATORS OF PRE-MRNA POLYADENYLATION SITE SELECTION

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
Ralf Schmidt
aus
Deutschland

Basel, 2018

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Mihaela Zavolan, Prof. Dr. Elmar Wahle

Basel, den 27.03.2018

Prof. Dr. Martin Spiess
Dekan

ABSTRACT

In eukaryotic cells, remarkably orchestrated regulatory steps ensure the availability of proteins and non-coding RNAs at the right spot, at the right time. While many of these steps, such as splicing, have been well studied since decades, the choice of the mRNA 3' end, which leads to expression of one of the many possible primary transcripts from a single locus has been recognized as key mechanism of post-transcriptional gene regulation only in recent years.

Transitions between cell states have been found to be associated with specific patterns of change in poly(A) site usage, leading to coordinated changes in the length of 3' untranslated regions (3' UTRs). As 3' UTRs carry a plethora of *cis*-regulatory elements, their systematic shortening or lengthening has global effects on the responsivity of the transcriptome to regulation, which in turn affects essentially every aspect of RNA metabolism, including stability, transport and translation. However, the mechanisms underlying alternative polyadenylation (APA) under physiological or pathological conditions remain largely unknown. Likely, changes in poly(A) site choice are caused by changes in the availability of regulators that bind in the vicinity of poly(A) sites and impact their processing efficiency.

The projects summarized in this thesis were devoted to a better understanding of the regulation of APA. Integrative analysis of a large number of data sets allowed us to establish a comprehensive annotation of poly(A) sites in the human genome. Tools developed for the projects described here could then exploit this resource to quantify and model the changes of poly(A) site usage in different contexts. In particular, the application of PAQR to quantify 3' end processing from RNA-seq data and of KAPAC to relate the abundance of individual sequence motifs to changes in poly(A) site usage led to intriguing insights into the regulation of APA in cancer. For gliobastoma, a CU-dinucleotide repeat motif was most significantly associated with the observed 3' UTR shortening, an effect that is likely to be explained by the binding of PTBP1, a factor previously known for its role in splicing regulation.

Together with HNRNPC, another splicing factor that was implicated in the regulation of poly(A) site choice through analyses presented here, these results suggest an extensive coupling between splicing and 3' end processing. In particular, it appears that many regulators of both mRNA maturation steps exist and remain to be uncovered. Previous results from glioma cell lines indicated that PTBP1 levels directly affect proliferation and migration. Considering its role in splicing and 3' end processing, PTPB1 may emerge as an important regulator of gene expression with direct implications for tumor progression in glioblastoma. Potentially, PTBP1 can serve as therapeutic target or diagnostic marker in brain tumor.

In summary, the work of this thesis illustrates how the deployment of computational tools can condense the information contained in large-scale data sets into biologically relevant results, shedding light on novel aspects of mRNA 3' end processing in physiological and pathological conditions. The uncovered regulators may be amenable to targeting by small molecules, thereby restoring the RNA processing patterns specific to the healthy states.

ACKNOWLEDGEMENTS

This thesis marks the end of a decisive chapter of my life making it the right place to express my gratitude to several people that accompanied me during this time. First, I have to deeply thank Mihaela Zavolan for being an enormously supportive supervisor. Her incredible energy, at-all-times responsiveness and her readiness to promote the development of her lab members paved my way to become a mature scientist. Beyond numerous enlightening research lessons, she taught me what it means to combine passion with the ambition to achieve tangible results. Even though a lucky coincidence enabled my start in her lab rather than a well founded decision, I can not imagine a better place for my PhD studies.

Most of my time as a PhD student, I was in the fortunate position to be mentored by two Andreas Gruber. I am deeply grateful to *Andreas Senior* for his support during the first month in the lab when I often found myself amidst overwhelming challenges. He was a role model in his way to resolve problems and his conviction to be able to do so. In combination with his utterly pleasant personality, *Andreas Senior* was a perfect group member and co-worker. I am equally obliged to *Andreas Junior* for the memorable time we shared the same office. Copying his way to work felt always like a crash course in good practice for scientists. But even stronger I will remember our conversations and the level of familiarity we developed over the years. Apart from both Andreas', I am truly thankful to all members of our group who created an atmosphere that was inspiring and made me feel comfortable at all times. Especially our group meetings, I will always memorize as rewarding and enlightening discussions.

I'd like to specifically mention my PhD committee members Elmar Wahle, Helge Grosshans and Erik van Nimwegen. Before our first meeting I felt bothered by the obligation of this official procedure, but their diverse comments and impulses made their visits invaluable milestones during my PhD studies. Their sometimes disillusioning assessment of my projects secured me from ending up with too many loose ends and I highly appreciate their effort. Beside being a committee member, I particularly like to thank Erik for being a great teacher. His enthusiasm to convey fundamental concepts was truly inspiring and certainly strengthened my admiration for statistics.

Almost exclusively, my PhD projects depended on the high-performance computer facility of the University Basel. I am grateful to all members of the scicore team for their amazing support, no matter if hard- or software-wise. It was a constant relief to have a competent team next door that provided an exceptional environment which allowed myself to focus on research instead of solving administration and configuration issues.

No matter what I started or decided to do in my life, I always experienced an amazingly strong support and trust from my parents and my brother. I can not remember a single moment they would have been dubious about any of my plans and I have always derived inner strength from the security they were able to convey. Thinking back, it is impressive to realize that they always prioritized my development over their own. I am thankful to Holger and his decision

to study Bioinformatics. Without him I would not have dared to start it in the first place and profiting from his experiences has been a reliable aid more than once. I would like to thank Michel for sharing his way to shape PhD studies. His unpretentious and rational interpretation of what it means to be a PhD student helped me constantly to reframe my focus and beliefs. I never want to miss him as a friend, critic and discussion partner.

Last but not least I want to thank my wife, Danie, for being my best friend. By no means, my development over the last years would have been as fruitful without Danie's encouragement. She taught me to believe in myself and to take responsibility but she also critically questioned my attitude towards my environment and life in general. Danie built a fort if I was seeking shelter and tore it down again to make me feel free. Because of her I could make my PhD studies a lesson for life that I will be able to finish as a grown person.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The human body contains trillions of cells that share virtually the same genetic information. This shared genomic blueprint enables the formation of 200 distinct cell types [1]. It is a major challenge of modern biology to uncover the principles that guide the establishment of such a complex cellular landscape starting from a single cell and based on a unique genome. The transcriptional program has been recognized as a major determinant of a cell's function and state [2], distinct cell types having distinct transcriptomes. In eukaryotic cells, the set of expressed transcripts depends on many factors, beyond those that regulate transcription. In particular, gene expression is regulated at various co-transcriptional and post-transcriptional levels [3]. The best-characterized step of mRNA processing is certainly alternative splicing: co-transcriptional excision of introns is guided by various RNA-binding proteins (RBPs) which are expressed in a cell type-specific manner, leading to the cell type-specific expression of transcript isoforms, which differ in exon composition and function [4].

Similarly, much work in recent years has established that the processing of mRNA 3' ends also occurs at a variety of different sites, leading to transcript isoforms that sometimes only differ in their 3' untranslated regions (3' UTRs) [5, 6, 7]. Except for the mRNAs encoding replication-related histones [8], all mRNAs are matured through endonucleolytic cleavage at the 3' end and subsequent addition of a stretch of adenosines (the poly(A) tail) [9, 10, 11]. More than half of all human genes have multiple cleavage and polyadenylation (CPA) sites, which are used in a context-dependent manner, a phenomenon which was called alternative polyadenylation (APA) [7, 12, 13]. In the simplest case, APA only affects the 3' UTR of a protein-coding mRNA, so that some transcripts have short 3' UTRs while others have long, frequently much longer, 3' UTRs. The choice of the 3' UTR can have far-reaching implications for the fate of the mRNA, because 3' UTRs are hubs where *cis*-regulatory elements facilitate the binding

of microRNAs (miRNAs), long non-coding RNAs (lncRNAs) and RBPs, that in turn affect the stability, localization and translational efficiency of the mRNA [14, 15, 16, 17]. The relevance of 3' UTRs for gene regulation is underpinned by the observation that their median length has increased during evolution from around 140 nucleotides (nt) in worms to 1200 nt in humans [18, 19], paralleling the increase in overall genome size and organism complexity. Thus, 3' UTR length modulation through APA shapes the potential for the binding of interaction partners, with effects for distinct aspects of an mRNA's life cycle and functionality.

In recent years, APA has been recognized as a pervasive mechanism of regulating the expressed transcriptome [7, 20]. Strikingly, changes in poly(A) site usage were shown to be globally coordinated in relation to the cellular state [6, 21]. The observation that cancer cells, similar to other highly proliferative cells, express transcripts with systematically shortened 3' UTRs [22, 23] renewed the interest in APA, in particular in the context of human diseases.

The main focus of the work presented in this thesis was to unravel the regulation of poly(A) site choice. The following sections review the current state of the field.

1.1 Core elements guiding 3' end cleavage and polyadenylation

Cleavage and polyadenylation sites are defined by a specific configuration of *cis*-regulatory elements that surround the actual processing (cleavage) site (CS) and enable the correct assembly of the 3' end processing machinery (Fig. 1.1). It is thought that the 'strength' of poly(A) sites, which corresponds to the efficiency with which they are processed, is modulated by the various *cis*-elements in a combinatorial manner [24]. The most conserved element is a hexamer, typically AAUAAA, called poly(A) signal [25]. Genomic analyses uncovered seventeen other close variants that likely serve as poly(A) signals in both human and mouse, because they occur in a specific relationship to the CS, 21 nucleotides (nt) upstream (as will be discussed in Chapter 2). However, the poly(A) sites with an upstream AAUAAA signal are by far the most abundant and are processed with the highest efficiency [26, 27]. Poly(A) sites frequently have a degenerate U/GU-rich downstream sequence element (DSE), which is located within 30 nt—sometimes even further downstream—from the CS, and stimulates 3' end processing [28]. The cleavage of the pre-mRNA is carried out at a site between the poly(A) signal and the DSE, often after a CA-dinucleotide [29]. However, the exact position of the cleavage site can vary by several nucleotides depending on the individual context of *cis*-regulatory motifs [26, 30]. Finally, U-rich elements upstream of the poly(A) signal, often UGUA, aid in the recognition of the poly(A) site [31]. From the multiple poly(A) sites of a given gene, the most distal was generally found to have a more canonical composition in regulatory elements and to be most efficiently processed [11, 32, 33].

Beside linear sequence motifs, the RNA structure [34, 35], the conformation of chromatin [36] and the nucleosome positioning [37] were proposed to impose further constraints and

thereby influence the pre-mRNA cleavage. Overall, research from several decades suggests that poly(A) sites are defined by characteristic and conserved *cis*-regulatory elements, whose precise sequence and constellation determines the efficiency with which the site is processed [32, 38].

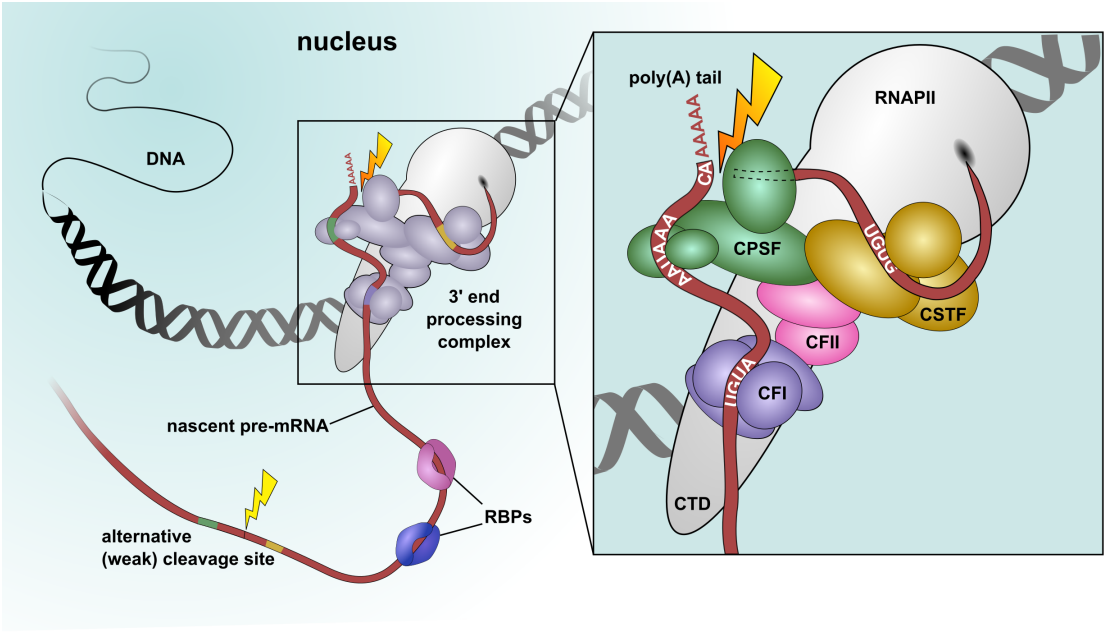


Figure 1.1: pre-mRNA maturation through cleavage and polyadenylation. The 3' end processing machinery assembles on the CTD of the RNAPII. An already transcribed proximal poly(A) site (yellow bolt) was ignored as cleavage site. The main 3' end processing subcomplexes along with their corresponding *cis* elements are shown in the frame on the right-hand side. The pre-mRNA part between the proximal and the processed distal poly(A) site contains multiple binding sites for interaction partners, shown is the binding of two RBPs. RNAPII - RNA polymerase II, CTD - C-terminal domain, RBP - RNA-binding protein, CPSF - Cleavage and Polyadenylation Specificity Factor, CSTF - Cleavage Stimulation Factor, CFI - Cleavage Factor I, CFII - Cleavage Factor II.

1.2 The 3' end processing machinery

In spite of being a relatively simple two-step reaction, pre-mRNA maturation through CPA involves a large and complex machinery [39]. Experiments using affinity purification uncovered over 85 proteins that participate in pre-mRNA cleavage and polyadenylation [40], of which about 20 form the core molecular machinery that selects and processes the poly(A) site (PAS) [41]. These further form four highly conserved multisubunit protein complexes (shown in the panel on the right-hand side of Figure 1.1). The cleavage and polyadenylation specific factor (CPSF), consisting of CPSF1 (CPSF160), CPSF2 (CPSF100), CPSF3 (CPSF73), CPSF4 (CPSF30),

WDR33 and FIP1L1, binds the poly(A) signal and cleaves the pre-mRNA. While binding of the poly(A) signal was attributed for a long time to CPSF160 [42], recent evidence demonstrates that PAS binding is due to WDR33 and CPSF4 [43, 44, 45, 46, 47]. CPSF3 contains the catalytic domain that provides the endonuclease activity [48]. CPSF cooperatively acts together with the heterotrimeric cleavage stimulation factor (CSTF) that contains the subunits CSTF1 (CstF50), CSTF2 (CstF64) or its paralogue CSTF2T (CstF64τ) and CSTF3 (CstF77). The binding of CSTF to the UG-rich DSE is achieved via a RNA recognition motif (RRM) domain of CSTF2. However, it is unclear how CSTF2 is able to bind rather unspecific degenerate UG-rich sequence motifs in the downstream region [49]. The cleavage factor I (CFI), a tetramer of two small NUDT21 (CFI 25) subunits and two large CPSF6 (CFI 68) and/or CPSF7 (CFI 59) subunits, binds to UGUA motifs upstream of the CS. The interaction of CFI with UGUA alone can already initiate CPA in the absence of a poly(A) signal [31]. The least well characterized complex of the core machinery, cleavage factor II (CFII), is composed of CLP1 and PCF11 and might be involved in the stabilization of the entire machinery [29]. In addition to the four protein complexes, the core machinery also contains the proteins symplekin (SYMPK), the poly(A) polymerase (which has three paralogues, PAPOLA, PAPOLB, PAPOLG) and the nuclear poly(A) binding protein (PABPN1) [50, 51, 52].

1.3 Consequences of alternative polyadenylation

Most human genes express multiple isoforms, that differ not only in their internal exons, but also in their 3' end [7, 53]. Depending on the location of poly(A) sites relative to the coding sequences, APA events can (1) alter the coding potential of transcripts, (2) change the terminal exon or (3) alter only the length of the 3' UTR [52]. The latter, most extensively covered in the present dissertation, are presumed to serve mostly in remodeling post-transcriptional regulatory interactions [54]. When the cleavage occurs at proximally located poly(A) sites, 3' UTRs are shortened and binding sites for factors like RBPs are not included in the transcripts which in turn affects various aspects of the mRNA life cycle.

Maybe the most extensively characterized examples involve changes in transcript stability, depending on whether the transcript has a short 3' UTR (due to processing at a proximal poly(A) site) or a long 3' UTR (when the processing occurs at a distal site). Short 3' UTR transcripts are slightly more stable than the corresponding long 3' UTR transcripts [55, 56]. Often, 3' UTRs contain target sites for microRNAs (miRNAs). miRNAs are non-coding RNAs which repress gene expression through mRNA destabilization [57]. The miRNA-mediated regulation is evaded by short 3' UTR isoforms [58], and this has been proposed as an explanation for the increase in stability of short 3' UTR isoforms [21]. Indeed, expression of short 3' UTR isoforms lacking miRNA binding sites has been associated with the upregulation of several oncogenes [22]. However, the generality of this pattern has been under debate, mainly because the expected

changes in transcript abundance that should parallel the APA shifts were not found in several studies (summarized in [59]). Moreover, in mouse fibroblasts a substantial fraction of APA events produced shorter 3' UTR isoforms that were less stable [55]. One putative mechanism to explain partially contradictory findings involves the activity level of miRNAs which is generally higher in cases of binding sites that are close to the 5' or 3' end of the 3' UTR [60]. APA at a more proximal site can thus strengthen the activity of a previously weak miRNA binding site rather than only depleting binding sites [61]. Additionally, a screen for putative regulatory elements revealed equally many activating and repressing elements in 3' UTRs indicating that 3' UTR shortening can also result in the loss of regulatory elements with a positive effect on expression [62].

Adenylate/uridylate-rich elements (ARE) are another class of *cis*-regulatory elements that influence mRNA stability and reside mainly in the 3' UTR. A variety of proteins (ARE-binding proteins or AREBPs) can bind to these elements. Most AREBPs have a destabilizing effect on the host mRNA [63], but examples of stabilization upon AREP binding are known [64]. Thus, loss of AREBP binding sites upon 3' UTR shortening should lead to increased gene expression, similar to the loss of miRNA-mediated repression [65, 66].

3' UTR isoforms are also exported from the nucleus with different efficiencies. Alu repeats, a family of short interspersed elements (SINEs), were suggested as driving force for this effect [67], as was incomplete splicing [68]. The longer isoforms tend to be more abundant in the nucleus, an observation that matches findings that relate long transcripts with nuclear retention in general [68, 69]. To date, it remains challenging to disentangle the effects of altered stability from those of changes in nuclear export. APA and the choice of 3' UTR also affect the movement of the mRNA within the cytoplasm. For example in neurons, the subcellular localization of transcripts can be 3' UTR dependent. The isoform of brain-derived neurotrophic factor (BDNF) with a short 3' UTR is retained in the soma, whereas the long isoform is localized in dendrites thereby facilitating an energy-efficient protein localization [70, 71]. Strikingly, it is not only the transcript localization that changes through APA; Berkovits and Mayr showed that the localization of the encoded protein can depend on the 3' UTR of the mRNA, a mechanism called 3' UTR-dependent protein localization, or UDPL [72]. Only the long isoform of CD47 enables 3' UTR mediated protein-protein interactions that lead to the shuttling of the protein to the plasma membrane. The protein from the short isoform remains in the endoplasmic reticulum (ER) despite having the very same amino acid sequence as the protein encoded by the long 3' UTR isoform.

Whether APA also affects the translation of the encoded protein has remained unclear, even though translation is known to be under the control of RBPs that specifically bind to 3' UTRs [73]. An initial study supported the notion that shorter transcripts are translated more efficiently, as they lack repressive 3' UTR elements [22]. The mean 3' UTR length of isoforms in the low polysome fraction was also found to be significantly longer than that of isoforms

isolated from the high polysomes [74]. In active neurons, translational repression is mediated through regulatory elements in extended 3' UTRs, which is again indicative for a direct link between translational capacity and 3' UTR length [75]. However, a few genome-wide studies did not find a global correlation between 3' UTR length and protein abundance [55, 56]. On the level of individual genes, the translational activity can differ significantly between short and long isoforms, although sometimes in an unexpected manner [76]. Hence, the relation of APA and translation is likely to be more complex than anticipated. The outlined examples demonstrate the multifaceted effect of APA on various aspects of mRNA life cycle, with direct implications on gene expression.

1.4 Regulators of poly(A) site usage

As the pervasiveness of APA as a means of regulating gene expression is becoming increasingly clear, it is important to understand what regulates APA itself. Obviously, factors of the core processing machinery are expected to influence the poly(A) site choice, their deregulation leading to global changes in the 3' UTR landscape. Indeed, the small interfering RNA-mediated knock-down of either CPSF6 or NUDT21, both subunits of the CFI complex, leads to drastic genome-wide 3' UTR shortening [33, 77]. Nuclear and cytoplasmic poly(A) binding proteins (PABPN1, PABPC1) influence poly(A) site usage in a similar fashion, depletion of either of these factors resulting in higher usage of proximal poly(A) sites [78, 79]. 3' UTR length is also changing upon deregulation of CSTF2, a component of the CSTF complex, but in the opposite direction: 3' UTR shortening is induced by upregulation of CSTF2. One of the first functionally relevant examples of APA, the immunoglobulin M (IgM) heavy chain switch from a membrane-bound to a secreted form, is regulated by CSTF2: higher CSTF2 levels induce the usage of a proximal poly(A) site, i.e. the expression of the short isoform, encoding the secreted form of the protein [80, 81]. However, CSTF2 downregulation has a limited effect; only the simultaneous downregulation of CSTF2 and its paralogue, CSTF2T, leads to a global shift towards more distal poly(A) site usage [49]. FIP1L1, a subunit of the CPSF complex, and the CFII component PCF11 also strengthen the usage of more proximal sites, at least in mouse [79].

Beyond core 3' end processing factors, poly(A) site choice is also sensitive to the abundance of splicing specific factors like the U1 small nuclear ribonucleoprotein (U1 snRNP, short U1) [82]. During cell activation, U1 becomes limited, which leads to premature cleavage at proximal sites that are no longer masked by U1 [83]. Other splicing factors that also modulate APA are muscleblind like splicing regulator (MBNL1/2) [84], the poly(C) binding protein (PCBP1) [85] or the neuron-specific NOVA alternative splicing regulator 2 (NOVA2) [86]. The heterogeneous ribonucleoprotein C (HNRNPC), another protein with a described role in splicing [87], is discussed in Chapter 2, as we have recently found that HNRNPC also masks poly(A) sites, preventing CPA. These results strongly suggest a tight coupling between splicing and 3' end

processing, an insight gained already through the multitude of reported interactions between components of both machineries [52]. Additionally, whether NOVA2 acts as a repressor or activator of splicing or CPA depends on the relative binding-position with respect to the splice or poly(A) site [86]. An interesting emerging paradigm is that the position-dependent activity for factors that regulate splicing as well as CPA is the same in both cases [88]. When NOVA2 binds close to a splice site or 3' end, it interferes with the assembly of the processing machinery and represses exon inclusion or poly(A) site usage, respectively. Similarly, PTBP1 represses exon inclusion and poly(A) site usage if its binding outcompetes the binding of factors from the canonical processing machinery [89, 90].

Apart from factors directly involved in either 3' end processing or splicing, other proteins like the cytoplasmic polyadenylation element binding protein 1 (CPEB1) have an effect on poly(A) site usage. CPEB1, known to regulate translation in the cytoplasm, shuttles to the nucleus and facilitates the usage of proximal poly(A) sites. Of note, the isoforms matured in this way are also under translational control by CPEB1 in the cytoplasm. Hence, CPEB1 elegantly couples the regulation of 3' end processing and translation [91].

All described factors are assumed to act via sequence-specific binding to *cis*-regulatory elements in the vicinity of poly(A) sites and the findings for NOVA2 even suggest an influence of the relative position of binding. Hence, one can think of 3' ends as combinations of sequence motifs that determine the usage of the poly(A) sites; the efficiency of usage directly depends on the prevalent motifs and their interaction partners. Chapter 3 follows this conceptual regime and tries to gain insights into the regulation of APA by identifying sequence motifs that best explain changes in 3' end processing between conditions.

A previously proposed, though still poorly studied mechanism for regulating APA is the so called "kinetic model" [92]. This model explicitly recognizes that RNAs are produced through polymerization, which leads to a situation in which 3' end processing sites emerge sequentially, and are not available all at the same time to the 3' end processing machinery. It is thus expected that upstream poly(A) sites have more time to undergo processing compared to the distal sites. Hence, the efficiency with which different poly(A) sites recruit the 3' end processing complex, coupled with the elongation rate of RNA polymerase II should govern the choice of poly(A) sites [92]. Indeed, mutant fruit flies with a slow polymerase show preferential usage of proximal poly(A) sites for several genes [93]. Another example of regulation of 3' end processing in line with the kinetic model involves DICER1 and EHMT2: both proteins influence the chromatin landscape around the proximal poly(A) site of the ETNK1 gene causing the slow-down of RNA polymerase II, which may be sufficient to facilitate the usage of the site [68].

Many modulators of poly(A) site choice have been identified already. Especially the effect and contribution of components of the core processing machinery has become increasingly clear. However, it remains elusive to which extent APA upon physiological changes of the cell state reflects changes in expression of core 3' end processing proteins or of other factors.

1.5 Physiological and disease-related changes in poly(A) site use

The switch from the membrane-bound to the secreted form of the IgM heavy chain during B cell activation [80] is one of the earliest described examples of APA in a physiological context. Strikingly, changes in 3' end processing during transitions from one cell state to another seem to occur on a global scale. Proliferating cells such as activated T cells and cancer cells have shorter 3' UTRs than their naïve or normal counterparts [21, 22]. Conversely, 3' UTRs undergo lengthening during cell differentiation, for example in neurons [6]. APA was also found to provide cell-type specificity to the expression of genes that are ubiquitously transcribed [19]. Genes that are expressed in multiple tissues tend to have alternative 3' ends and their 3' UTRs are generally longer relative to genes that only have a single poly(A) site. Together with an enrichment of *cis*-regulatory elements in the regions with differential inclusion into transcripts (those between proximal and distal poly(A) sites), these observations suggest that multiple poly(A) sites are related to more extensive post-transcriptional regulation of the corresponding genes [94, 95].

Extensive changes in 3' processing in cancers have been linked to tumorigenesis and tumor invasiveness [23, 96]. More specifically, expression of shorter 3' UTRs resulting from preferential usage of more proximal poly(A) sites is thought to allow transcripts to escape miRNA-dependent repression and to activate oncogenes [22]. In triple-negative breast cancer, elevated protein levels of the oncogenes JUN and NRAS are a direct consequence of 3' UTR shortening which protects the mRNAs from destabilization through PUM1 binding [96]. Generally, a disproportionate enrichment of *cis*-regulatory motifs in 3' UTRs of oncogenes and tumor suppressor genes makes them particularly suited to regulation by APA [96]. Overall, multiple studies put forth the concept that malignant transformation is coupled with 3' UTR shortening [22, 97, 98, 99]. However, some discrepancies were also reported. For example, although both MCF7 and MDA-MB-231 are breast cancer cell lines, only MCF7 showed the expected APA shift towards more proximal poly(A) site usage, whereas MDA-MB-231 exhibited rather 3' UTR lengthening relative to a mammary epithelial cell line [100]. Similarly, an analysis of 114 selected oncogenes and tumor suppressors across several cancer cohorts did not reveal a clear trend in APA, further arguing for a more complex pattern of APA in cancer [101]. Nevertheless, the patterns of poly(A) site use are cancer-specific and proved to be sufficiently informative to enable stratification of murine B cell leukemia samples into subgroups with different prognostic [102]. Additional prognostic power beyond clinical markers was also reported for selected APA events in a broader study of seven tumor types [98].

Besides cancer, APA has been associated with other disease as well. The poly(A) site usage was shown to be sensitive to PABPN1, a protein with a pivotal role in autosomal-dominant oculopharyngeal muscular dystrophy (OPMD) [78]. OPMD is caused by short triplet repeat expansion mutations in PABPN1. These lead to a 3' UTR shortening similar to that due to the knock-down of PABPN1, suggesting that the APA events induced through the mutated PABPN1

are relevant for the development of OPMD. Altered levels of NUDT21, a component of the core machinery subcomplex CFI, were even more directly linked with disease. Individuals with a copy number variation of NUDT21 have aberrant levels of MECP2, small changes of which cause neuropsychiatric diseases. Increased NUDT21 levels, e.g. due to gene duplication, result in an increased usage of the distal poly(A) site and expression of a longer MECP2 isoform, which is inefficiently translated [103]. This highlights the role of NUDT21 as a relevant factor for mental disability and neuropsychiatric disease. Similar to MECP2, a single APA event at the cyclin CCND1 gene is disease-relevant. A point mutation introduces a strong poly(A) signal (AAUAAA) that leads to premature CPA and the expression of a shorter but more stable CCND1 isoform. The resulting increase in overall CCND1 mRNA levels was correlated with high proliferation and shorter survival of strongly proliferative mantle cell lymphoma tumors [104].

Clearly, APA profiles are cell state specific and undergo dynamical changes in a context-dependent manner. As mentioned earlier, how these global changes are induced in physiological conditions is largely unknown. Our own analysis of numerous patient data sets obtained from The Cancer Genome Atlas (TCGA), as discussed in Chapter 3, not only finds global 3' UTR shortening but also determines potential physiological APA regulators. For example, our results implicate PTBP1, a known splicing regulator, as regulator of APA in glioblastoma (GBM).

1.6 Genome-wide methods to analyze 3' end usage

Over the last two decades, technological improvements have drastically changed the way 3' end processing is studied. After pioneering work done based on cloning [11], it was the analysis of expressed sequence tags (ESTs), generated with oligo-dT primers, that enabled large-scale studies of 3' end processing [12]. Clusters of EST sequences were thus exploited to identify alternative CPA sites contributing to a more accurate transcript annotation [105]. The first studies that reached the genome-wide scale were based on the microarray technology. Despite the shortcoming of a limited set of probes, which were not optimally designed to measure the relative usage of poly(A) sites, this technology enabled important insights, for example into the global increase of proximal poly(A) site usage during T cell activation [21, 106]. Ultimately, high-throughput sequencing (HTS) is the method of choice to date, enabling a quasi-comprehensive detection and quantification of poly(A) site usage. Once stable protocols for transcriptome sequencing (RNA-seq) were established, a variety of methods specifically devised to capture RNA 3' ends was developed [29]. Perhaps not unexpectedly, each specific method has its own bias, and the usage of poly(A) sites, determined with different methods, is not always consistent. To overcome these limitations and provide a complete annotation of poly(A) sites in the human and mouse genomes, we have recently undertook a study of hundreds of sequencing libraries. The work that entailed this analysis is presented in Chapter 2.

The vast majority of publicly available HTS data comes from standard RNA-seq experiments rather than from 3' end sequencing. To take advantage of this wealth of data for the study of mRNA 3' end processing, methods for quantifying relative poly(A) site usage directly from RNA-seq data were proposed. Their biggest challenge is to cope with the non-homogeneous read coverage along transcripts that can especially obscure sites of inefficient 3' end processing. The basic principle of all of the proposed methods is the segmentation of 3' UTRs (or terminal exons) into parts that end at individual 3' end processing sites. Looking from the perspective of an internal poly(A) site, there is one isoform that ends at this site and a longer isoform, that extends beyond this PAS. Because both share the region upstream of the PAS, reads mapping to this region are compatible with both isoforms. On the contrary, reads that map to the region downstream of the PAS are only compatible with the long isoform. Early approaches used the available transcript annotation to define the up- and downstream regions and quantified the usage of the poly(A) sites by comparing the read counts within the corresponding regions [88, 107]. Methods that did not rely on pre-existing gene annotations, inferring alternative 3' ends (i.e. alternative transcript isoforms) *de novo*, were also developed. KLEAT is one such approach, carrying out first a *de novo* transcript reconstruction and then estimating isoform abundances. The abundances then serve as the input to compute relative usages of isoforms that only differ in their 3' ends [108]. A more prevalent approach to identify proximal cleavage sites seeks to determine specific patterns of fluctuation in read density, an idea initially applied to probe intensity measurements from microarray data [109]. Since the upstream region coverage is a composite of reads from the short and the long isoform, a drop in read coverage together with its position and extent should be indicative for the location and the relative usage of the proximal poly(A) site. The PHMM tool uses a Hidden Markov Model [110] to implement this approach, while ChangePoint deploys a generalized likelihood ratio statistics [111], IsoSCM a Bayesian [112] and DaPars a regression model [23]. All of these methods have to apply rather high thresholds to ensure a good specificity despite the high heterogeneity of the coverage profile itself. Moreover, read density profiles from RNA-seq data are not well suited to identify poly(A) sites with single nucleotide resolution. To overcome these problems, we developed a method called PAQR, introduced in Chapter 3, that exploits the single nucleotide resolution of data from 3' end sequencing protocols in combination with read coverage fluctuations, to infer and quantify high-confidence APA events from RNA-seq data.

In the light of the enormous effort that has been taken to establish RNA-seq data repositories from large-scale studies of patients or entire populations, methods that infer APA from standard RNA-seq are highly desirable. Although their precision is still somewhat limited, they remain the most promising approach until the next revolution of full transcript sequencing is widely available.

1.7 The importance of RNA integrity

To infer 3' end usage based on RNA-seq read coverage profiles, the sequenced reads should reflect the underlying transcript isoform abundances as precisely as possible. In the optimal case, transcript bodies would be uniformly covered by RNA-seq reads so that transcript starts or ends would be visible directly by an increase or drop in coverage, respectively. However, this optimal coverage profile is hardly ever observed. Especially for sequence libraries prepared from clinical samples, the uniformity in read coverage can be compromised by the degradation of RNA that occurs in the time between sample collection and RNA extraction [113, 114]. Depending on the specimen collection and storage conditions, RNA decay occurs to variable extents, while the protocol used for library preparation can further amplify the resulting bias [115]. In data sets such as those from the TCGA, where library preparation included selection with oligo-dT, advanced RNA degradation resulted in severe 3' end bias in read coverage [116]. The most widely used metric to assess the degree of RNA degradation is called RNA integrity number (RIN) and relies mainly on the amount of 18S and 28S ribosomal RNAs which is evaluated based on the electropherogram of the isolated RNA [117]. To allow also post hoc assessments of the RNA integrity from the library of sequenced reads, different methods were proposed. In fact, all of them show a high correlation with the RIN scores for the set of test samples that was used in the corresponding publication [115, 116]. Such methods are particularly important in the context of large-scale data analyses of hundreds or thousands of samples with largely variable backgrounds, like discussed in Chapter 3 for the APA analysis of TCGA patient data. Unfortunately, it's not clear whether coverage bias can be corrected sufficiently at the computational level, to still be able to make use of the data and quantify transcript abundances or poly(A) site usage. However, at the minimum, samples with low RNA integrity can be excluded from further analysis which would otherwise distort the results. Despite a missing consensus at which RNA integrity score RNA decay renders a sample as unusable, the consideration of the metric at least ensures that samples with comparable RNA quality are analyzed. The importance of this quality control step was demonstrated for the analysis of gene expression: the estimated expression values vary in an RNA integrity dependent manner [114].

Decades of work from the scientific community revealed APA as a pivotal mechanism of post-transcriptional regulation of gene expression. Beyond identifying individual events of biologically relevant poly(A) site switching, evidence for systematic APA changes on a genome-wide scale has also emerged. To date, APA in different conditions, distinct cell types and states could be characterized. Furthermore, the regulatory role of various factors has been explored. Specific protocols to capture mRNA 3' ends were instrumental for these analyses. The integration of the data sets from these studies, presented in Chapter 2, enabled a more comprehensive

annotation of the genome-wide locations of CPA for mouse and human. The obtained atlas proved as a helpful resource when we expanded the scope of our research to include also the analysis of APA in samples for which only standard RNA-seq data are available, such as those from cancer patients. This work, discussed in Chapter 3, aims to uncover regulators of polyadenylation in human diseases, starting solely from the transcriptomics data that is available. The initial results are encouraging, as they provide evidence for the involvement of PTBP1 in APA in glioblastoma.

The work of this thesis contributed to the current understanding of alternative polyadenylation by delivering resources and tools which we demonstrated to be instrumental in the study of the regulation of poly(A) site choice in physiological and pathological conditions, at an unprecedented detail.

CHAPTER 2

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

This work was published in *Genome Research* in 2016 (see reference [118]).

2.1 Abstract

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.

2.2 Introduction

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 [119]. Recent studies have revealed systematic changes in 3' UTR lengths upon changes in cellular states, either those that are physiological [21, 83] or those during pathologies [23]. 3' UTR lengths are sensitive to the abundance of specific spliceosomal proteins [82], core pre-mRNA 3' end processing factors [33, 120], and polyadenylation factors [78]. 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., [58]), the choice of polyadenylation (poly(A)) sites has important regulatory consequences that reach up to the subcellular localization of the resulting protein [72]. Studies of presumed regulators of polyadenylation would greatly benefit from the general availability of comprehensive catalogs of poly(A) sites such as PolyA_DB [20, 121], which was introduced in 2005 and updated 2 years later.

Full-length cDNA sequencing offered a first glimpse on the pervasiveness of transcription across the genome and on the complexity of gene structures [122]. 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 [123]. By increasing the depth of coverage of transcription start sites and mRNA 3' ends, these protocols aimed to improve the quantification accuracy [6, 124, 125, 126]. 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 [6, 97], tissues [7, 127], developmental stages [128, 129], and cell differentiation stages [53], as well as following perturbations of specific RNA processing factors [33, 78, 85, 120, 130]. 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 [20], 439,390 [7], and 1,287,130 [97] sites have been reported in the human genome.

The current knowledge about sequence motifs that are relevant to cleavage and polyadenylation (for review, see [11]) goes back to studies conducted before next-generation sequencing technologies became broadly used [12, 27, 131]. 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) [43, 44] 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 [132] and is found at a large proportion of pre-mRNA cleavage sites identified in different studies [12, 133, 134]. 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 [27].

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 [135, 136], the polypyrimidine tract binding protein 1 [90, 136], FIP1L1, and CSTF2 [136], and by proteins that bind G-rich elements—cleavage stimulation factor CSTF2 [137] and HNRNPs F and H1 [138]—or C-rich elements—poly(rC)-binding protein 2 [85]. Some of these proteins are multifunctional splicing factors that appear to couple various steps in pre-mRNA processing, such as splicing, cleavage, and polyadenylation [139]. The sequence elements to which these regulators bind are also frequently multifunctional, enabling positive or negative regulation by different RBPs [137]. 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.

2.3 Results

2.3.1 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 [7, 20, 97] for the human genome. By comparing the 3' end processing sites from two recent genome-wide studies [7, 127], we found that a substantial proportion was unique to one or the other of the two studies (Supplemental Table A.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 A.2, A.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 [140]. 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.

2.3.2 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 [12]. 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 [12, 19, 33]. 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 2.5 Methods). For consistency with previous studies [12], 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 2.5 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. 2.1A), as reported previously [12, 27]. 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 [27] 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 [27] 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. 2.1A). 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.* [26]. 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. A.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 2.5 Methods section (Fig. 2.1B).

2.3.3 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

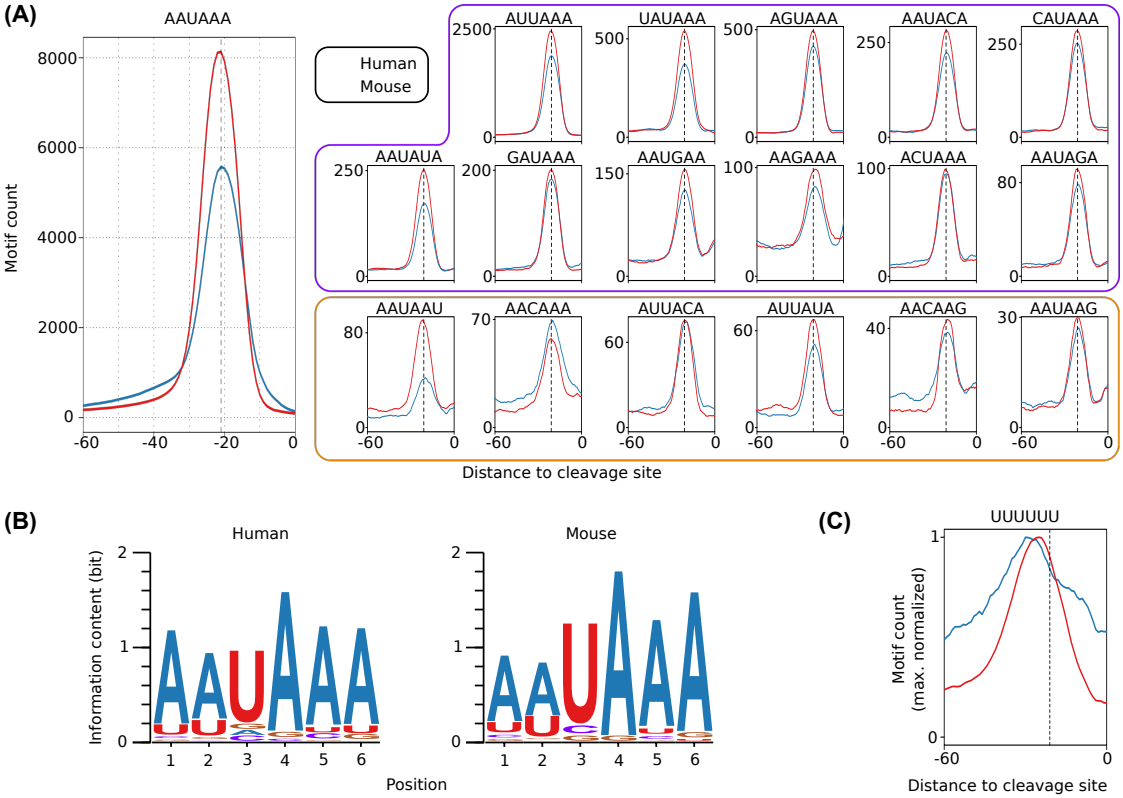


Figure 2.1: Hexamers with highly specific positioning upstream of human and mouse pre-mRNA 3' end cleavage sites. (A) The frequency profiles of the 18 hexamers that showed the positional preference expected for poly(A) signals in both human and mouse. The known poly(A) signal, AAUAAA, had the highest frequency of occurrence (left). Apart from the 12 signals previously identified (AAUAAA and motifs with the purple frame) [27], 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).

Fig. A.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 [6, 7, 56]. 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 2.5 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. A.3). Supplemental Figures A.4A and A.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 [12, 33, 141], 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. A.4B, A.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) [7, 127], only 61.82% [127] and 41.38% [7] 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 A.9 (human) and A.10 (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.

2.3.4 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 [31]. 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. 2.1C). 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 A.4, A.5). Although the A- and U-richness of pre-mRNA 3' end processing regions have been observed before [132], 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) [142, 143], and its knock-down causes a systematic increase in 3' UTR lengths [79, 143].

2.3.5 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 [144]. Of these, HNRNPC has been recently studied with crosslinking and immunoprecipitation (CLIP) and found to bind the majority of protein-coding genes [87], with high specificity for poly(U) tracts [87, 144, 145, 146, 147]. HNRNPC appears to nucleate the formation of ribonucleoprotein particles on nascent transcripts and to regulate pre-mRNA splicing [87, 145] and polyadenylation at Alu repeats [148]. 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 A.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 4,405 exons with a total of 22,698,094 mapped reads (Supplemental Table A.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 (CFIm)

[33, 120]. Out of the 5,152 poly(A) sites that showed consistent behavior across replicates, we found 1,402 poly(A) sites (27.2%) to increase in usage, 1,378 poly(A) sites (26.7%) to decrease in usage, and 2,372 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. 2.2A).

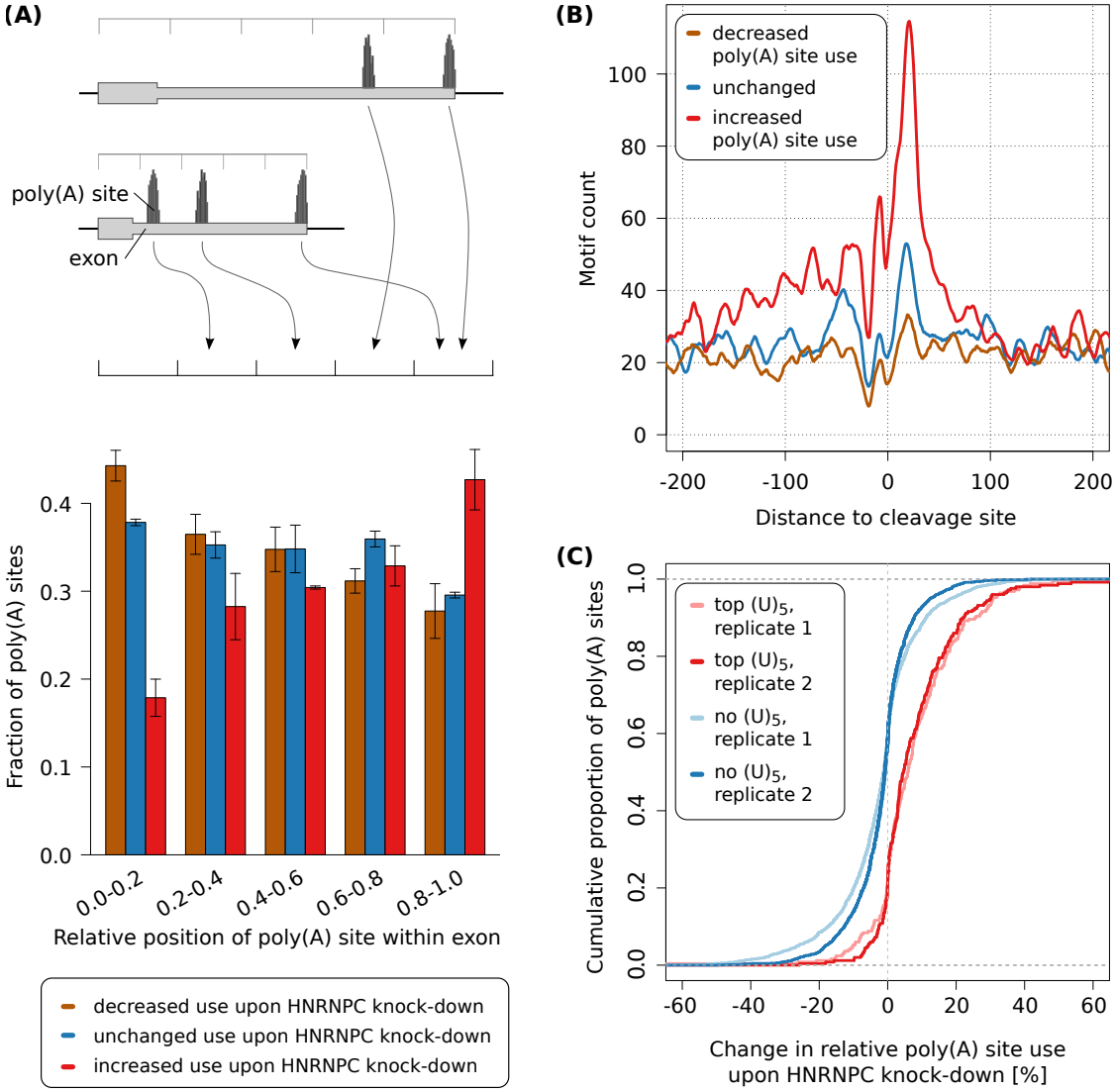


Figure 2.2: siRNA-mediated knock-down of HNRNPC leads to increased use of distal poly(A) sites. (A) Relative location of sites whose usage decreased (brown), did not change (blue) or increased (red) in response to HNRNPC knock-down within 3' UTRs. We identified the 1000 poly(A) sites whose usage increased 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.

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: 1.1×10^{-19}; replicate 2 p-value: 3.1×10^{-61}; one-sided Wilcoxon signed-rank test) (Supplemental Figs. A.7,A.8). It was noted before that distal poly(A) sites are predominantly used in HEK 293 cells [33]. 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 [87, 144, 147], 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. 2.2B). No such enrichment emerged from a similar analysis of untransfected versus si-Control transfected cells (Supplemental Fig. A.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. A.10). We found, consistent with the results shown in Figure 2.2B, 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. A.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: 2.4×10^{-36}; replicate 2 p-value: 1.9×10^{-42}; one-sided Mann-Whitney U test) (Fig. 2.2C). We therefore concluded that HNRNPC's binding in close proximity of 3' end processing sites likely masks them from cleavage and polyadenylation.

2.3.6 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 2.3A, 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.

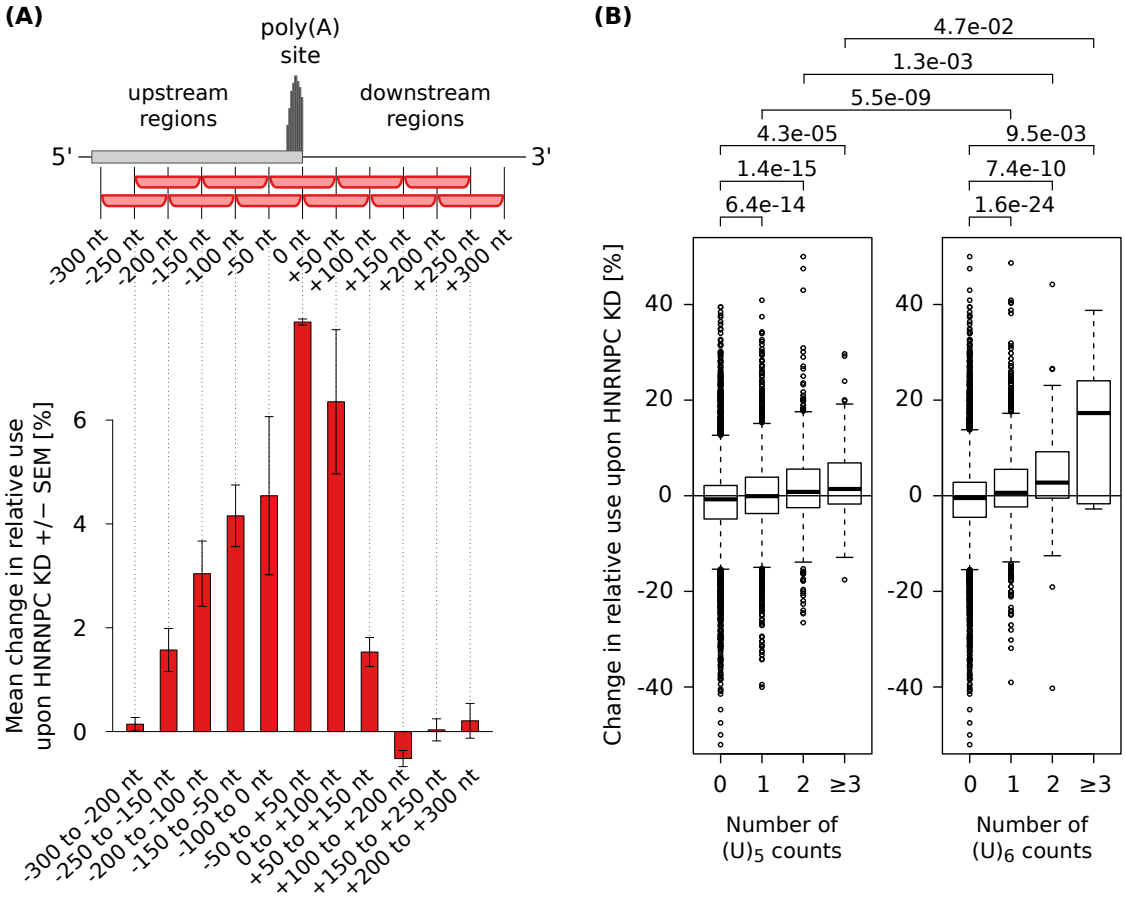


Figure 2.3: The length, number, and location of poly(U) tracts with respect to poly(A) sites influence the change in poly(A) site use upon HNRNPC knock-down. (A) Mean change in the use of sites containing the highest number of $(U)_5$ motifs within 100-nt-long regions located at specific distances from 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.

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. 2.3A).

Although the minimal RNA recognition motif of HNRNPC consists of five consecutive uridines [144, 146, 147], longer uridine tracts are bound with higher affinity [87, 145, 146]. 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. 2.3B).

2.3.7 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. A.12, A.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 [149]. 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) [150] 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. 2.4A).

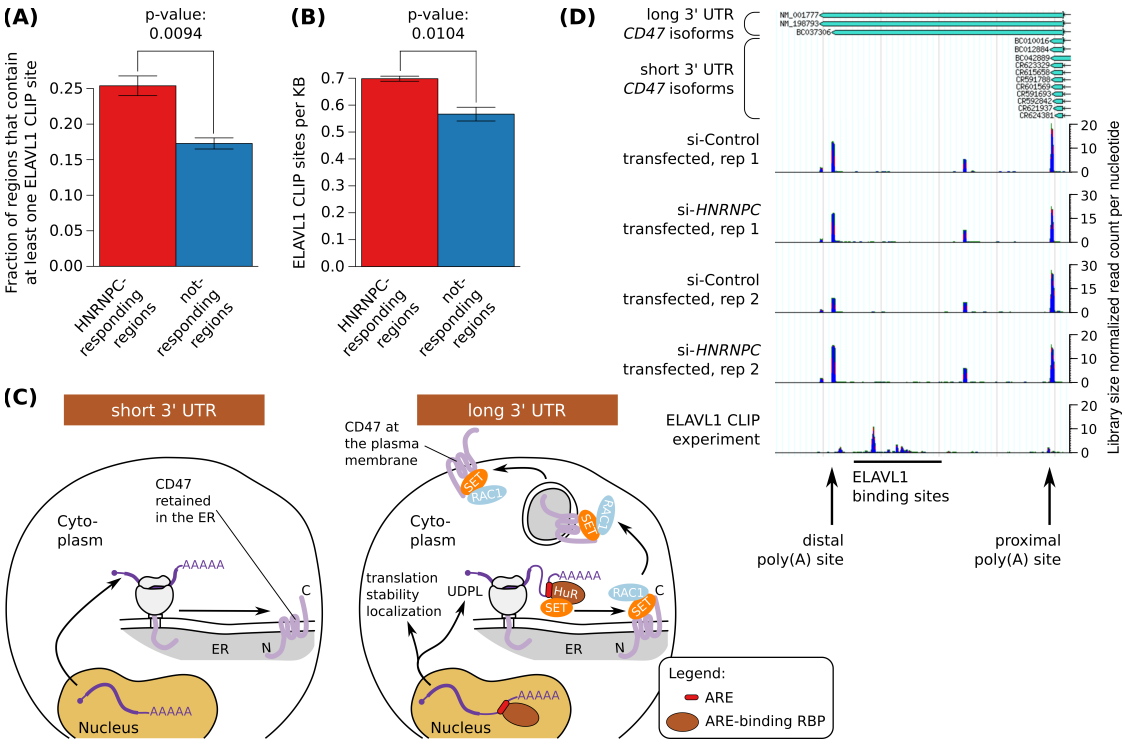


Figure 2.4: HNRNPC-responsive 3' UTRs are enriched in ELAVL1 binding sites. (A) Fraction of HNRNPC-responding and not-responding 3' UTR regions that contain one or more ELAVL1 CLIP sites. The p-value of the one-sided t-test is shown. **(B)** Density of ELAVL1 CLIP sites per kilobase (kb) in the 3' UTR 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 [72]. **(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.

Moreover, the density of ELAVL1 binding sites and not only their absolute number was enriched across these 3' UTR regions (Fig. 2.4B). 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 [72] 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. 2.4C) [72, 151]. 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. 2.4D). 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. A.14) but are the consequence of increased usage of the distal poly(A) site upon knock-down of HNRNPC (Fig. 2.4D). To find out whether HNRNPC can act as an upstream regulator of UDPL, we quantified 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. 2.5A; Supplemental Fig. A.15).

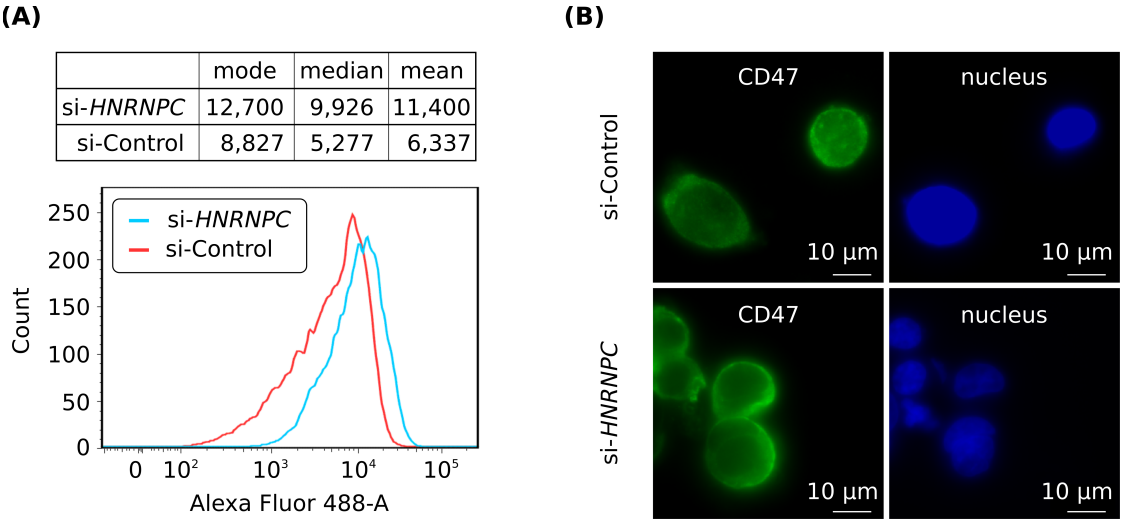


Figure 2.5: The knock-down of HNRNPC affects CD47 protein localization. **(A)** Indirect immunophenotyping of membrane-associated CD47 in HEK 293 cells that were treated either with an si-HNRNPC (blue) or with si-Control (red) siRNA. Mean, median, and mode of the Alexa Fluor 488 intensities computed 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.

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. A.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. 2.5B). Overall, our results suggest that HNRNPC can function as an upstream regulator of UDPL.

2.3.8 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 [152] 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 1.4×10^{-30} and 5.1×10^{-29}) (Fig. 2.6A).

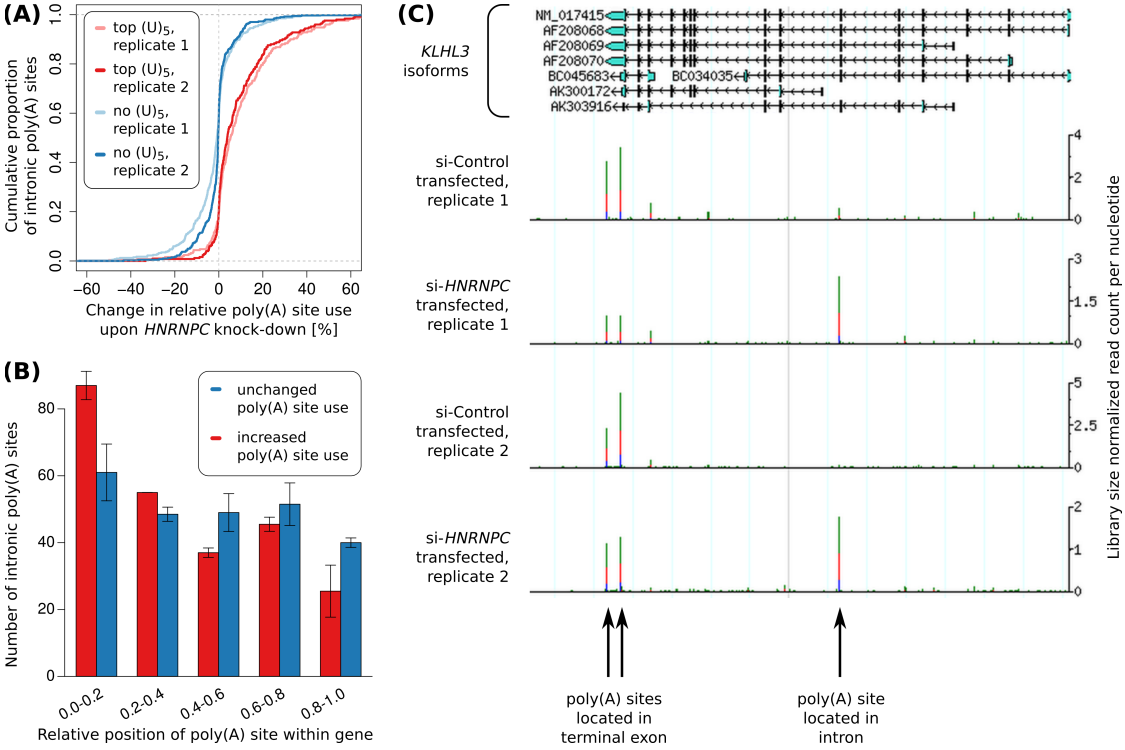


Figure 2.6: HNRNPC knock-down leads to increased usage of intronic poly(A) sites. (A) The change in the relative use of intronic poly(A) sites that did not contain any $(U)_5$ within ±200 nt and of the top 250 intronic poly(A) sites according to the number of $(U)_5$ motifs within ±50 nt around the cleavage 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.

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. A.17). Importantly, only the intronic sites that responded to HNRNPC knock-down were strongly enriched in $(U)_5$ tracts immediately downstream from the poly(A) site (Supplemental Fig. A.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 2.6B, 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 2.6C shows the example of the Kelch Like family member 3 (KLHL3) gene, which harbors one of the most derepressed intronic poly(A) sites.

2.4 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., [13]). 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 [153, 154, 155], and short 3' UTRs are generated in proliferating and malignant cells [20, 21, 98]. The key regulators of these polyadenylation programs are unknown. Reduced expression of the U1 snRNP [83] or of the mammalian cleavage factor I (CFIm) components NUDT21 and CPSF6 [33, 120] 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 [23, 83]. 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 [78], 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 [98]; and the retinoblastoma binding protein 6, whose reduced expression results in reduced transcript levels and increased use of distal poly(A) sites [156].

Many experimental protocols to capture transcript 3' ends and enable studies of the dynamics of polyadenylation have been developed (for review, see [157]), and consequently, a few databases of 3' end processing sites are available [7, 20, 127]. 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 [7, 127] 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. [7] and You et al. [127], 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. [7] and You et al. [127] (29.57% and 24.51%, respectively) (Supplemental Fig. A.19). The lower coverage of lincRNAs is probably due to their lower expression in comparison with protein-coding genes [158] and to the fact that some of them are bimorphic, appearing in both the poly(A)$^+$ and poly(A)$^-$ fraction [159], 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. A.20; for detailed numbers, see Supplemental Tables A.7, A.8).

The 3' end processing sites reported by other studies [7, 127] 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 [132], 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 [144, 146, 147] and has a variety of functions including pre-mRNA splicing [87] and mRNA transport [160], also modulates the processing of pre-mRNA 3' ends. HNRNPC has originally been identified as a component of the HNRNP core particle [161, 162] and found to form stable tetramers that bind to nascent RNAs [163]. Systematic evolution of ligands by exponential enrichment (SELEX) experiments have shown that HNRNPC particles bind to uninterrupted tracts of five or more uridines [164], and studies employing CLIP indicated that longer tracts are bound with higher affinity [87]. 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" [82, 83]. 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 [165]. Interestingly, the competition between HNRNPC and U2AF2 appears to regulate exonization of Alu elements [145] and, furthermore, impacts polyadenylation at Alu exons [148]. These studies have emphasized the complex cross-talk between regulators that come into play during RNA splicing and polyadenylation [11]. 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 [130] up to protein localization [72].

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 [21, 22]. 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 [55, 56]. 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 [54] but reach as far as the transport of transmembrane proteins to the plasma membrane [72]. 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 [72]. 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.

2.5 Methods

2.5.1 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. A.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 [152] and Ensembl annotation of mouse [166], both obtained from UCSC [167] in June 2013). The sequence alignment was done with segemehl with default parameters [168]. 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. A.21).

2.5.2 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_u) of, or downstream (d_d) from, the representative site (Supplemental Fig. A.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_u and d_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. A.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 $d_u = d_d = 12$ nt [12, 127]. Only for the clustering of putative 3' end processing sites in individual samples, we used a larger distance, $d_u = d_d = 25$, resulting in a more conservative set of clusters, with a maximum span of 51 nt.

2.5.3 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. A.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 [169], 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. A.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. A.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.

2.5.4 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 [12, 165]. 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. A.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.

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

2.5.5.1 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 A.2, A.3), 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.

2.5.5.2 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 $\leq c$ read counts as not having sufficient experimental support (for outlines how to determine sample-specific cutoffs, Supplemental Fig. A.26). This allowed for an efficient filtering of reads presumably representing background noise.

2.5.5.3 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. A.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,
- EX: any other exon except the terminal one,
- IN: any intron,
- DS: up to 1000 nt downstream from an annotated gene,
- AE: antisense to an annotated exon,
- AI: antisense to an annotated intron,
- AU: antisense and within 1000 nt upstream of an annotated gene, and
- IG: intergenic

2.5.5.4 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`.

2.5.5.5 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 [170] version 3.3, with default settings, to generate the sequence logos for human and mouse, respectively.

2.5.5.6 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.

2.5.5.7 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 [7]. Mouse

genome coordinates were converted to the coordinates of the Ensembl mm10 annotation through LiftOver [171]. 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 [127]. The genomic coordinates of all poly(A) sites (one position per poly(A) site) were intersected with the annotation features.

2.5.6 Analysis of 3' end libraries from HNRNPC knock-down experiments

2.5.6.1 Sequencing of A-seq2 libraries and quantification of relative poly(A) site usage

We considered all high-confidence A-seq2 [56] 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%.

2.5.6.2 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 [150], 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 [172, 173] 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 [174]. 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 [175].

2.5.6.3 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, 2,454 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.

2.5.7 Experiments

2.5.7.1 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 AGGUAGUGUAUCGCCUUGTT) and si-HNRNPC1/2 (sc-35577 from Santa Cruz Biotechnologies), both applied at 20 nM in 2.5 mL DMEM on six-well plates.

2.5.7.2 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.

2.5.7.3 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.

2.5.7.4 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.

2.5.7.5 PAR-CLIP and A-seq2 libraries

A-seq2 libraries were generated as previously described [56] and sequenced on an Illumina HiSeq 2500 sequencer. The HNRNPC PAR-CLIP was performed as previously described [33] 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.

2.5.8 HNRNPC PAR-CLIP analysis

The raw data were mapped using CLIPZ [150]. 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 A.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.

2.5.9 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 [147] 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 [176].

2.5.9.1 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.

2.6 Authors information

2.6.1 List of authors

The following authors have contributed to the work discussed in Chapter 2:

1. Andreas Johannes Gruber[1] (Abbr.: AJG),
2. Ralf Schmidt[1] (Abbr.: RS),
3. Andreas R. Gruber[1] (Abbr.: ARG),
4. Georges Martin[1] (Abbr.: GM),
5. Souvik Ghosh[1] (Abbr.: SG),
6. Manuel Belmadani[1] (Abbr.: MB),
7. Walter Keller[1] (Abbr.: WK) &
8. Mihaela Zavolan[1] (Abbr.: MZ)

[1] Biozentrum, University of Basel, Klingelberstrasse 50-70, CH-4056 Basel, Switzerland

2.6.2 Author contributions

The order of authors in the previous subsection (2.6.1) reflects the authors' contributions, with the first two authors (AJG and RS) being responsible for all major aspects of the data analysis. The last two authors are principal investigators and thus their listing follows the opposite ranking.

In detail, using the abbreviations specified in subsection 2.6.1: MZ, AJG, ARG, and WK designed the project. ARG, RS, and AJG collected data sets and created the catalog of poly(A) sites. MB developed the PolyAsite web interface. RS and AJG identified poly(A) signals. GM performed the HNRNPC PAR-CLIP and A-seq2 experiments. AJG and RS analyzed the data. GM and SG performed the experiments. MZ supervised the project. AJG, MZ, RS, and ARG wrote the manuscript.

2.7 Acknowledgments

We thank Erik van Nimwegen for his input on data analysis, Beatrice Dimitriades for technical assistance, and Josef Pasulka for suggestions on the analysis.

2.8 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.

2.9 Supplementary materials

Supplementary materials can be found in Appendix A.

2.10 Supplementary materials

The work discussed in this Chapter was supported by the Swiss National Science Foundation grant 31003A-143977 to Walter Keller and by the Swiss National Science Foundation NCCR project "RNA & Disease" (51NF40_141735).

CHAPTER 3

DISCOVERY OF PHYSIOLOGICAL AND CANCER-RELATED REGULATORS OF 3' UTR PROCESSING WITH KAPAC

This work was published by *Genome Biology* in 2018 (see reference [177]).

3.1 Abstract

3' Untranslated regions (3' UTRs) length is regulated in relation to cellular state. To uncover key regulators of poly(A) site use in specific conditions, we have developed PAQR, a method for quantifying poly(A) site use from RNA sequencing data and KAPAC, an approach that infers activities of oligomeric sequence motifs on poly(A) site choice. Application of PAQR and KAPAC to RNA sequencing data from normal and tumor tissue samples uncovers motifs that can explain changes in cleavage and polyadenylation in specific cancers. In particular, our analysis points to polypyrimidine tract binding protein 1 as a regulator of poly(A) site choice in glioblastoma.

3.2 Background

The 3' ends of most eukaryotic mRNAs are generated through endonucleolytic cleavage and polyadenylation (CPA) [11, 51, 178]. These steps are carried out in mammalian cells by a 3' end processing complex composed of the cleavage and polyadenylation specificity factor (which includes the proteins CPSF1 (also known as CPSF160), CPSF2 (CPSF100), CPSF3 (CPSF73), CPSF4 (CPSF30), FIP1L1, and WDR33), the mammalian cleavage factor I (CFIm, a tetramer of two small, NUDT21 (CFIm 25) subunits, and two large subunits, of CPSF7 (CFIm 59) and/or CPSF6 (CFIm 68)), the cleavage factor II (composed of CLP1 and PCF11), the cleavage stimula-

tion factor (CstF; a trimer of CSTF1 (CstF50), CSTF2 (Cstf64) and CSTF3 (CstF77)), symplekin (SYMPK), the poly(A) polymerase (PAPOLA, PAPOLB, PAPOLG), and the nuclear poly(A) binding protein (PABPN1) [51, 52]. Crosslinking and immunoprecipitation (CLIP) revealed the distribution of core 3' end processing factor binding sites in pre-mRNAs [33] and the minimal polyadenylation specificity factor that recognizes the polyadenylation signal, consisting of the CPSF1, CPSF4, FIP1L1, and WDR33 proteins, has been identified [43, 44].

Most genes have multiple poly(A) sites (PAS), which are differentially processed across cell types [19], likely due to cell type-specific interactions with RBPs. The length of 3' UTRs is most strongly dependent on the mammalian cleavage factor I (CFIm), which promotes the use of distal poly(A) sites [23, 33, 77, 79, 120]. Reduced expression of CFIm 25 has been linked to 3' UTR shortening, cell proliferation and oncogene expression in glioblastoma cell lines [23], while increased levels of CFIm 25 due to gene duplication have been linked to intellectual disability [103]. The CSTF2 component of the CstF subcomplex also contributes to the selection of poly(A) sites [33, 49], but in contrast to CFIm, depletion of CSTF2 leads to increased use of distal poly(A) sites (dPAS), especially when the paralogous CSTF2T is also depleted [49]. PCF11 and FIP1L1 proteins similarly promote the use of proximal poly(A) sites (pPAS) [79].

Many splicing factors modulate 3' end processing. Most strikingly, the U1 small nuclear ribonucleoprotein (snRNP) promotes transcription, masking poly(A) sites whose processing would lead to premature CPA, through a "telescripting" mechanism [82, 83]. The U2AF65 spliceosomal protein interacts with CFIm [179] and competes directly with the heterogeneous nucleoprotein C (HNRNPC) for binding to uridine (U)-rich elements, regulating the splicing and thereby exonization of Alu elements [145]. HNRNPC represses CPA at poly(A) sites where U-rich sequence motifs occur [118]. Other splicing factors that have been linked to poly(A) site selection are the neuron-specific NOVA1 protein [86], the nuclear and cytoplasmic poly(A) binding proteins [78, 79], the heterogeneous ribonucleoprotein K (HNRNPK) [180], and the poly(C) binding protein (PCBP1) [85]. However, the mechanisms remain poorly understood. An emerging paradigm is that position-dependent interactions of pre-mRNAs with RBPs influence poly(A) site selection, as well as splicing [87]. By combining mapping of RBP binding sites with measurements of isoform expression, Ule and colleagues started to construct "RNA maps" relating the position of *cis*-acting elements to the processing of individual exons [181]. However, whether the impact of a regulator can be inferred solely from RNA sequencing data obtained from samples with different expression levels of various regulators is not known.

To address this problem, we have developed KAPAC (for **k**-mer **a**ctivity on **p**oly**a**denylation site **c**hoice), a method that infers position-dependent activities of sequence motifs on 3' end processing from changes in poly(A) site usage between conditions. By analogy with RNA maps, and to emphasize the fact that our approach does not use information about RBP binding to RNA targets, we summarize the activities of individual motifs inferred by KAPAC from different regions relative to poly(A) sites as "impact maps". As 3' end sequencing remains

relatively uncommon, we have also developed PAQR, a method for **p**oly**a**denylation site usage **q**uantification from **R**NA sequencing data, which allows us to evaluate 3' end processing in data sets such as those from The Cancer Genome Atlas (TCGA) Research Network [182]. We demonstrate that KAPAC identifies binding motifs and position-dependent activities of regulators of CPA from RNA-seq data obtained upon the knock-down of these RBPs, and in particular, that CFIm promotes CPA at poly(A) sites located ~ 50 to 100 nucleotides (nt) downstream of the CFIm binding motifs. KAPAC analysis of TCGA data reveals pyrimidine-rich elements associated with the use of poly(A) sites in cancer and implicates the polypyrimidine tract-binding protein 1 (PTBP1) in the regulation of 3' end processing in glioblastoma.

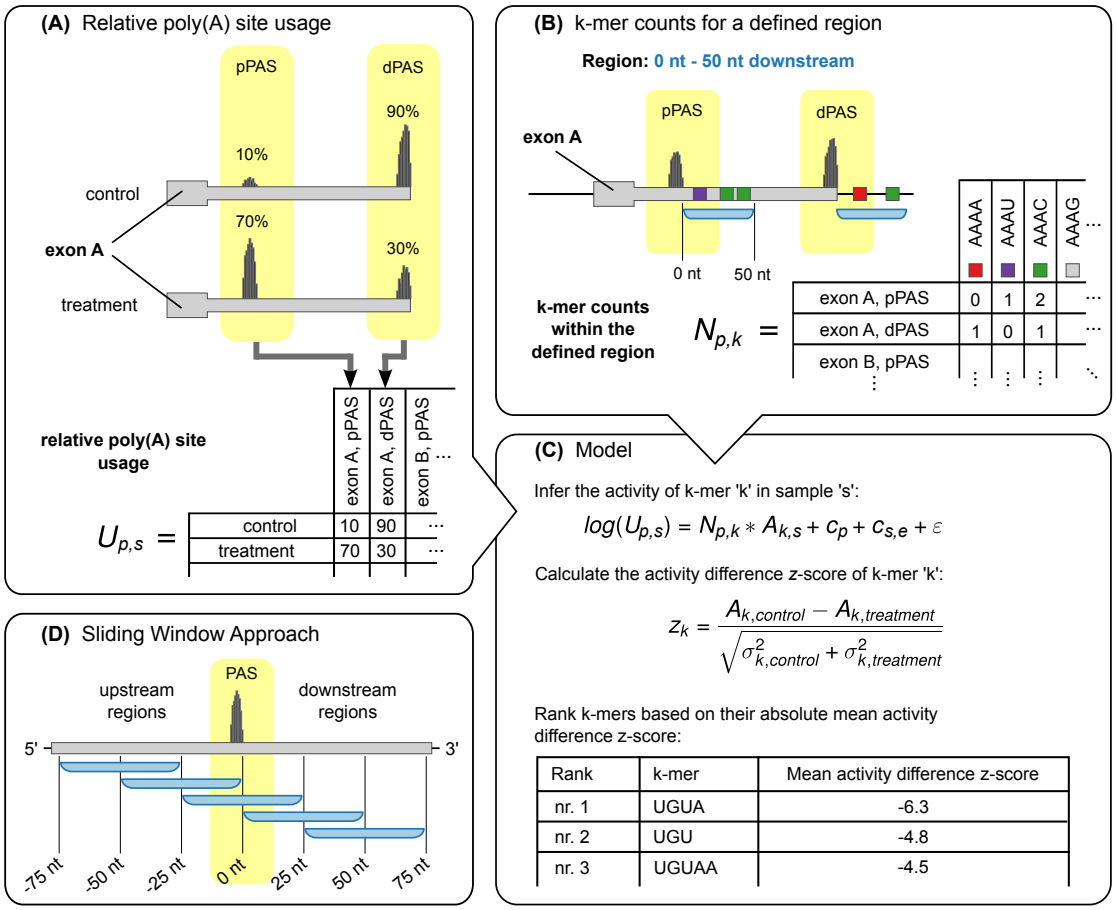
3.3 Results

3.3.1 Inferring sequence motifs active on PAS selection with KAPAC

As binding specificities of RBPs have only recently been started to be determined *in vivo* in high-throughput [183], we developed an unbiased approach, evaluating the activity of all possible sequences of length k (k-mers, with k in the range of RBP-binding site length, 3–6 nt [184]) on PAS usage. Briefly, we first compute the relative use of each PAS p among the P poly(A) sites ($P > 1$) in a given terminal exon across all samples s, as

$$U_{p,s} = \frac{R_{p,s}}{\sum_{p'=1}^{P} R_{p',s}}, \tag{3.1}$$

where $R_{p,s}$ is the number of reads observed for poly(A) site p in sample s (Figure 3.1A). KAPAC aims to explain the observed changes in relative poly(A) site usage $U_{p,s}$ in terms of the activity of a k-mer k within a sample s and the excess counts (over the background expected based on the mononucleotide frequencies; see section B.2.2.1) $N_{p,k}$ of the k-mer within a region located at a specific distance relative to the poly(A) site p (Figure 3.1B, C). Running KAPAC for regions located at various relative distances with respect to the PAS (Figure 3.1D) allows the identification of the most significantly active k-mers as well as their location.



Rank	k-mer	Mean activity difference z-score
nr. 1	UGUA	-6.3
nr. 2	UGU	-4.8
nr. 3	UGUAA	-4.5

Figure 3.1: Schematic outline of the KAPAC approach. (A) Tabulation of the relative usage of poly(A) sites in different experimental conditions (here, control and treatment). **(B)** Tabulation of k-mer counts for regions (blue) located at a defined distance with respect to poly(A) sites p. **(C)** Based on the usage of poly(A) sites relative to the mean across samples and the counts of k-mers k in windows located at specific distances from the poly(A) sites p, KAPAC infers activities $A_{k,s}$ of k-mers in samples s. $c_{s,e}$ is the mean relative usage of poly(A) sites from exon e in sample s, c_p is the mean $\log_2$-relative usage of poly(A) site p across samples, and ϵ is the residual error. KAPAC ranks k-mers based on the absolute z-score of the mean activity difference in two conditions (here, in control relative to treatment). **(D)** Fitting the KAPAC model for windows located at specific distances relative to poly(A) sites, position-dependent activities of sequence motifs on poly(A) site use are inferred.

3.3.2 KAPAC uncovers expected position-specific activities of RBPs on pre-mRNA 3' end processing

To evaluate KAPAC we first analyzed PAS usage data obtained by 3' end sequencing upon perturbation of known RBP regulators of CPA. Consistent with the initial study of poly(C) binding protein 1 (PCBP1) role in CPA [85], as well as with the density of its CCC—$(C)_3$—

binding element around PAS that do and PAS that do not respond to PCBP1 knock-down (Figure 3.2A), KAPAC revealed that $(C)_3$ motifs strongly activate the processing of poly(A) sites located 25–100 nt downstream (Figures 3.2B, C; Supplementary Table B.1).

As in a previous study we found that the multi-functional HNRNPC modulates 3' end processing (see also Figure 3.2D), we also applied KAPAC to 3' end sequencing data obtained upon the knock-down of this protein. Indeed, we found that $(U)_n$ sequences ($n = 3$–5 nt) have a strongly repressive activity on poly(A) site choice, which, reminiscent of HNRNPC's effect on exon inclusion [145], extends to a broad window, from approximately −200 nt upstream to about 50 nt downstream of poly(A) sites (Figure 3.2E, F, Supplementary Table B.1). In contrast to the density of $(U)_5$ motifs, which peaks immediately downstream of poly(A) sites, KAPAC inferred an equally high repressive activity of $(U)_5$ motifs located upstream of the poly(A) site.

These results demonstrate that being provided only with estimates of poly(A) site expression in different conditions, KAPAC uncovers both the sequence specificity of the RBP whose expression was perturbed in the experiment, and the position-dependent, activating, or repressing activity of the RBP on poly(A) site choice.

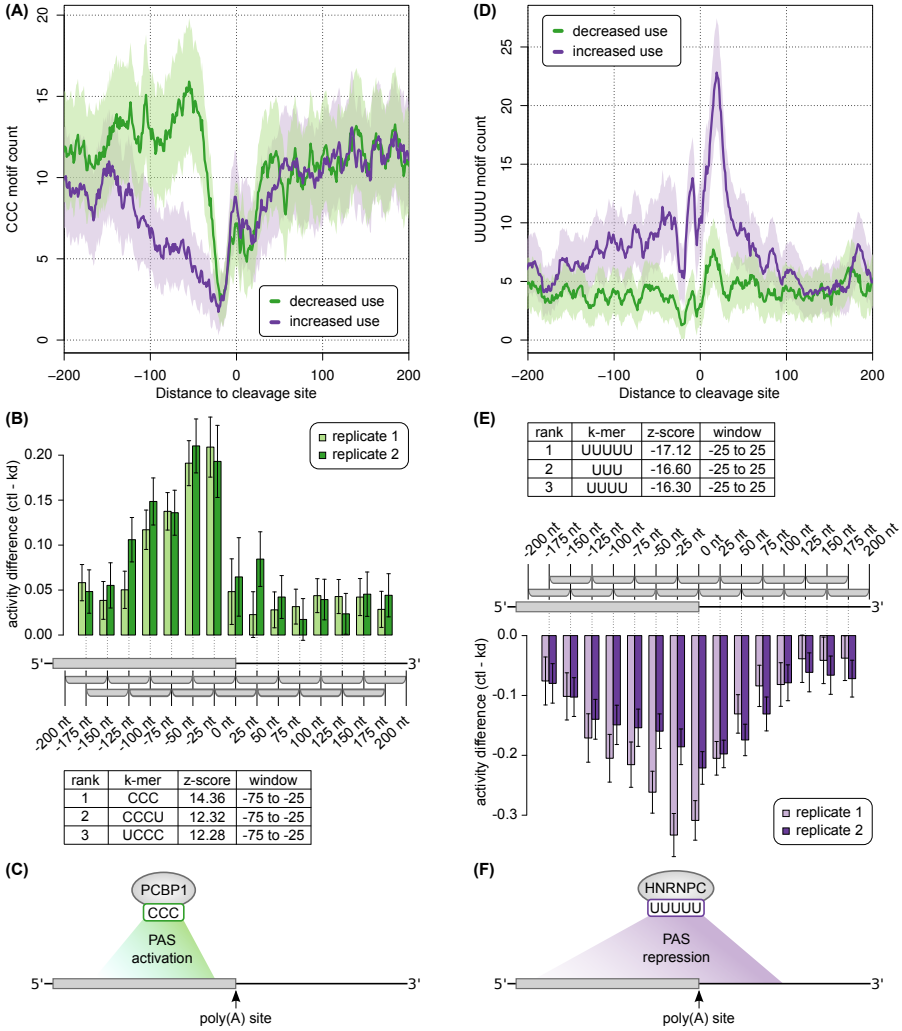


Figure 3.2: KAPAC accurately uncovers the activity of known regulators of poly(A) site choice. (A) Smoothened (±5 nt) density of non-overlapping $(C)_3$ motifs in the vicinity of poly(A) sites that are consistently processed (increased or decreased use) in two PCBP1 knock-down experiments from which 3' end sequencing data are available [85]. *Shaded areas* indicate standard deviations based on binomial sampling. **(B)** Difference of $(C)_3$ motif activity inferred by KAPAC in the two replicates of control (Ctrl) versus PCBP1 knock-down (KD) experiments (number of PAS n = 3737). The positive differences indicate that $(C)_3$ motifs are associated with increased PAS use in control samples. The table shows the three most significant motifs, with the z-score and position of the window from which they were inferred. **(C)** Model of the KAPAC-inferred impact of PCBP1 on CPA. **(D)** Smoothened (±5 nt) density of non-overlapping $(U)_5$ tracts in the vicinity of sites that are consistently processed (increased or decreased use) in two HNRNPC knock-down experiments [147]. **(E)** Difference of $(U)_5$ motif activity inferred by KAPAC in the two replicates of control (Ctrl) versus HNRNPC knock-down (KD) experiments (n = 4703). The negative differences indicate that $(U)_5$ motifs are associated with decreased PAS use in the control samples. The table with the three most significant motifs is also shown, as in **(B)**. **(F)** Model of the KAPAC-inferred impact of HNRNPC on CPA.

3.3.3 The PAQR method to estimate relative PAS use from RNA-seq data

As 3' end sequencing data remain relatively uncommon, we sought to quantify poly(A) site use from RNA sequencing data. The drop in coverage downstream of proximal PAS has been interpreted as evidence of PAS processing, generalized by the DaPars method to identify changes in 3' end processing genome-wide [23]. However, DaPars (with default settings) reported only eight targets from the RNA-seq data obtained upon the knock-down of HNRNPC [147], and they did not include the previously validated HNRNPC target CD47 [118], whose distal PAS shows increased use upon HNRNPC knock-down (Figure 3.3A). Furthermore, DaPars quantifications of relative PAS use in replicate samples had limited reproducibility (Supplementary Figures B.1, B.2), as did the motif activities inferred by KAPAC based on these estimates (Figure 3.3B, Supplementary Figure B.1). These results prompted us to develop PAQR, a method to quantify PAS use from RNA-seq data (Figure 3.3C). PAQR uses read coverage profiles to progressively segment 3' UTRs at annotated poly(A) sites. At each step, it infers the breakpoint that decreases most the squared deviation from the mean coverage of a 3' UTR segment when dividing the segment in two regions with distinct mean coverage (Figure 3.3C and 3.6) relative to considering it as a single segment with one mean coverage. A key aspect of PAQR is that it only attempts to segment the 3' UTRs at experimentally identified poly(A) sites, from an extensive catalog that was recently constructed [118]. Using the HNRNPC knock-down data set that was obtained independently [147] for benchmarking, we found that the PAQR-based quantification of PAS use led to much more reproducible HNRNPC binding motif activity and more significant difference of mean z-scores between conditions (-22.92 with PAQR-based quantification vs. -10.19 with DaPars quantification; Figure 3.3B, D, Supplementary Figure B.2). These results indicate that PAQR more accurately and reproducibly quantifies poly(A) site use from RNA-seq data.

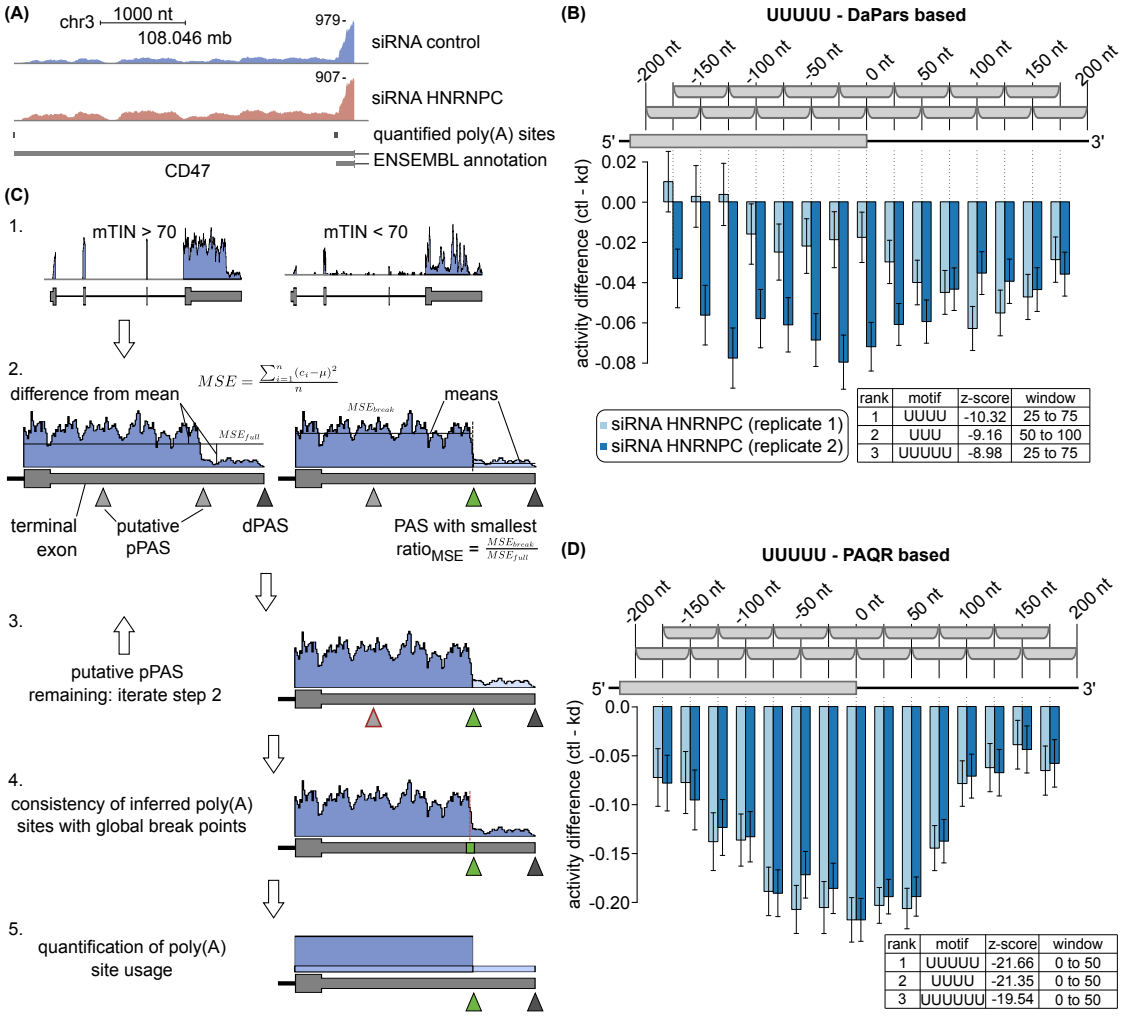


Figure 3.3: Overview on PAQR. (A) Read coverage profile of the CD47 terminal exon, whose processing is affected by the knock-down of HNRNPC [118]. **(B)** KAPAC-inferred position-dependent activities of the $(U)_5$ motif based on DaPars-based estimates of relative PAS use (number of PAS n = 13388) in the same data set as in **(A)**. **(C)** Sketch of PAQR: 1) Samples with highly biased read coverage along transcripts (low mTIN score), presumably affected by RNA degradation, are identified and excluded from the analysis. 2) Usage of proximal PAS (pPAS) in a sample is determined based on the expected drop in coverage downstream of the used PAS (ratio of the mean squared deviation from mean coverage (MSE) in the full region compared to two distinct regions, split by the poly(A) site). 3) Step 2 is repeated iteratively for subregions bounded by already determined PAS. 4) The consistency between PAS called as used and the global best break points in corresponding regions is evaluated and in case of discrepancy, terminal exons are discarded from the analysis. 5) Relative PAS use is calculated from the average read coverage of individual 3' UTR segments, each corresponding to the terminal region of an isoform that ends at a used poly(A) site. **(D)** Similar HNRNPC activity on PAS use is inferred by KAPAC from estimates of PAS use generated either by PAQR from RNA sequencing data (n = 3599), or measured directly by 3' end sequencing (Figure 3.2E).

3.3.4 KAPAC reveals a position-dependent activity of CFIm binding on cleavage and polyadenylation

As KAPAC allows us to infer position-dependent effects of RBP binding on 3' end processing, we next sought to unravel the mechanism of CFIm, the 3' end processing factor with a relatively large impact on 3' UTR length [33, 77, 79, 120]. We thus depleted either the CFIm 25 or the CFIm 68 component of the CFIm complex by siRNA-mediated knock-down in HeLa cells, and carried out RNA 3' end sequencing. As expected, CFIm depletion led to marked and reproducible 3' UTR shortening (Figure 3.4A, see 3.6 for details). We found that the UGUA CFIm binding motif occurred with high frequency upstream of the distal poly(A) sites whose usage decreased upon CFIm knock-down, whereas it was rare in the vicinity of all other types of PAS (Figure 3.4B). These results indicate that CFIm promotes the processing of poly(A) sites that are located distally in 3' UTRs and are strongly enriched in CFIm binding motifs in a broad region upstream of the poly(A) signal. KAPAC analysis supported this conclusion, further uncovering UGUA as the second most predictive motif for the changes in poly(A) site use in these experiments, after the canonical poly(A) signal AAUAAA (Figure 3.4C, Supplementary Table B.1), which is also enriched at distal PAS [33]. Interestingly, the activity profile further suggests that UGUA motifs located downstream of PAS may repress processing of these sites, leading to an apparent decreased motif activity when CFIm expression is high.

We repeated these analyses on RNA-seq data obtained independently from HeLa cells depleted of CFIm 25 [23], obtaining a similar activity profile (Figure 3.4D, Supplementary Table B.2), including the apparent negative activity of sites that are located downstream on PAS processing. These results demonstrate that CFIm binds upstream of distal PAS to promote their usage, whereas binding of CFIm downstream of PAS may, in fact, inhibit processing of poly(A) sites.

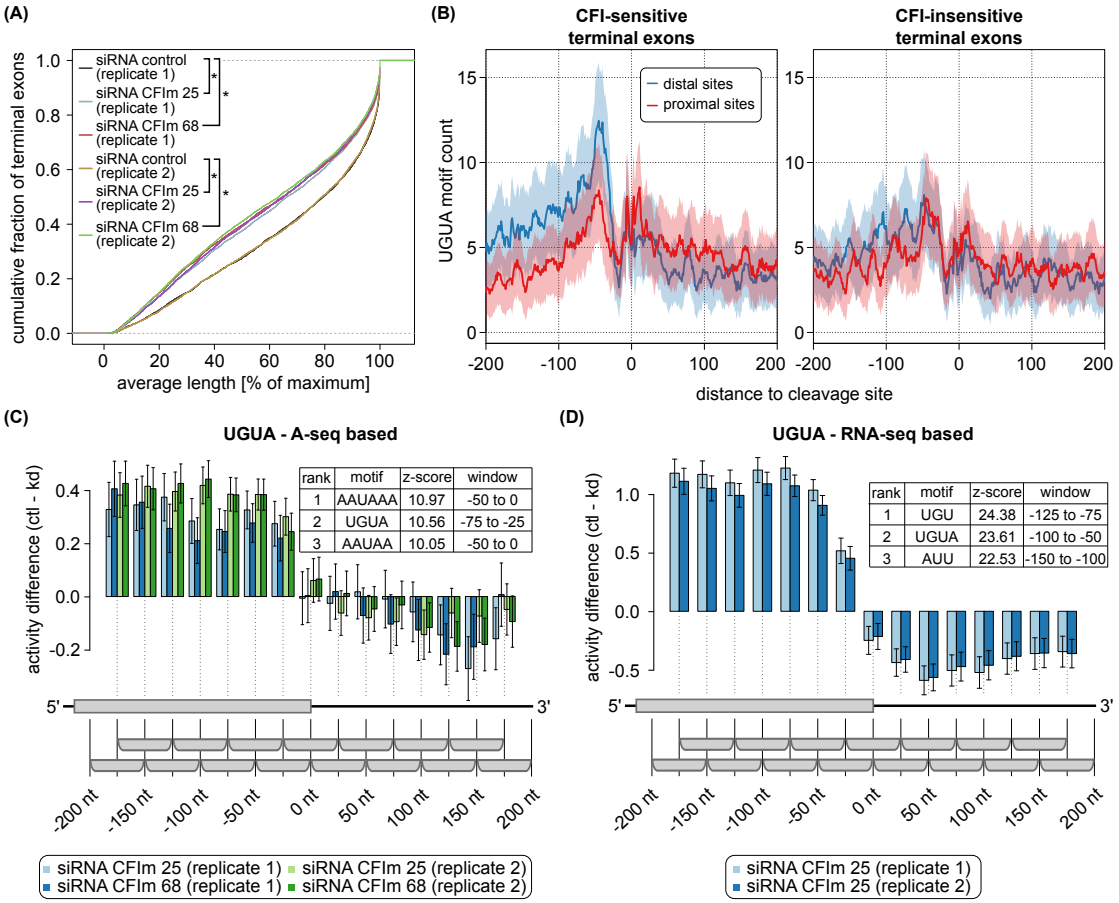


Figure 3.4: Position-dependent activation of pre-mRNA processing by CFIm. (A) The distributions of average terminal exon lengths (see 3.6) computed from 5123 multi-PAS terminal exons quantified in CFIm 25, CFIm 68 knock-down, and control samples indicate significant shortening of 3' UTRs upon CFIm depletion (*asterisks* indicate two-sided Wilcoxon signed-rank test p-value < 0.0001). **(B)** Smoothened (±5 nt) UGUA motif density around PAS of terminal exons with exactly two quantified poly(A) sites, grouped according to the log fold change of the proximal/distal ratio (p/d ratio) upon CFIm knock-down. The left panel shows the UGUA motif frequency around the proximal and distal PAS of the 750 exons with the largest change in p/d ratio, while the right panel shows similar profiles for the 750 exons with the smallest change in p/d ratio. **(C)** KAPAC analysis of CFIm knock-down and control samples uncovers the poly(A) signal and UGUA motif as most significantly associated with changes in PAS usage ($n = 3727$). **(D)** UGUA motif activity is similar when the PAS quantification is done by PAQR from RNA sequencing data of CFIm 25 knock-down and control cells ($n = 4287$) [23].

3.3.5 KAPAC implicates the pyrimidine tract binding proteins in 3' end processing in glioblastoma

We then asked whether KAPAC can uncover a role of CFIm 25 in 3' UTR shortening in glioblastoma (GBM), as has been previously suggested [23]. We found that while 3' UTRs are indeed

markedly shortened in these tumors (Figure 3.5A), UGUA was not among the 20 motifs that most significantly explained the change in PAS usage in these samples. This may not be unexpected because, in fact, once a certain threshold of RNA integrity is met, normal and tumor samples have CFIm expression in the same range (Supplementary Figure B.3).

Rather, KAPAC revealed that variants of the CU dinucleotide repeat, located from ~ 25 nt upstream to ~ 75 nt downstream of PAS, are most significantly associated with the change in PAS usage in tumors compared to normal samples (Figure 3.5B, Supplementary Table B.3). Among the many proteins that can bind polypyrimidine motifs, the mRNA level of the pyrimidine tract binding protein 1 (PTBP1) was strongly anti-correlated with the median average length of terminal exons in this set of samples (Figure 3.5C). This suggested that PTBP1 masks the distally-located, CU repeat-containing PAS, which are processed only when PTBP1 expression is low, as it is in normal cells. Of the 203 sites where the CU repeat motif was predicted to be active, 181 were located most distally in the corresponding terminal exons. The PTBP1 crosslinking and immunoprecipitation data recently generated by the ENCODE consortium [185] confirmed the enriched binding of the protein downstream of CU-containing, KAPAC-predicted target PAS (Figure 3.5D), whose relative usage decreases in tumor compared to control samples (Supplementary Figure B.4). Furthermore, the enrichment of PTBP1-eCLIP reads was highest for the highest scoring PTBP1 targets (Figure 3.5E). A similar pattern of PTBP1-eCLIP reads was obtained when the 200 PAS with the strongest decrease in relative usage were considered instead of KAPAC-predicted targets. In contrast, no obvious enrichment was observed for the 200 distal PAS with the least change in usage in glioblastoma compared to normal tissue (Supplementary Figure B.5). Strikingly, KAPAC analysis of mRNA sequencing data obtained upon the double knock-down of PTBP1 and PTBP2 in HEK 293 cells [186] confirmed this hypothesized effect of PTBP1 on 3' end processing (Figure 3.5F). These results implicate PTBP1 rather than CFIm 25 in the regulation of PAS use in glioblastoma.

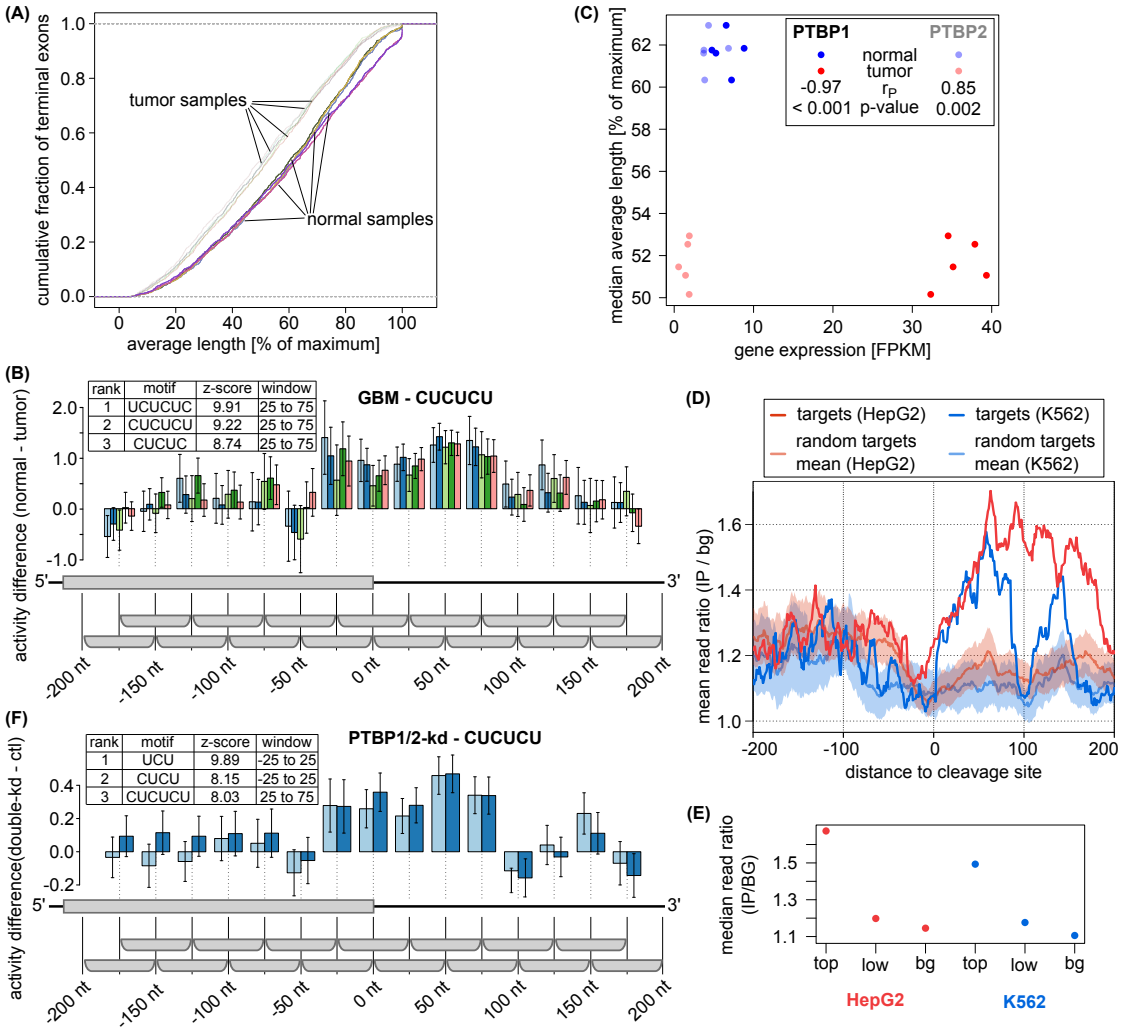


Figure 3.5: Regulation of PAS choice in glioblastoma samples from TCGA. (A) Cumulative distributions of weighted average length of 1172 terminal exons inferred by applying PAQR to five normal and five tumor samples (see 3.6 for the selection of these samples) show that terminal exons are significantly shortened in tumors. **(B)** Activity profile of CUCUCU, the second most significant motif associated with 3' end processing changes in glioblastoma (number of PAS used in the inference $n = 2119$). The presence of the motif in a window from -25 to $+75$ relative to PAS is associated with increased processing of the site in normal tissue samples. (Continued on next page)

3.3.6 A novel U-rich motif is associated with 3' end processing in prostate cancer

Cancer cells, particularly from squamous cell and adenocarcinoma of the lung, express transcripts with shortened 3' UTRs (Figure 3.6A, Supplementary Table B.4). The negative correlation between the mRNA level expression of CSTF2 and the 3' UTR length (Figure 3.6B) led to the

Figure 3.5: (C) Expression of PTBP1 in the 10 samples from **(A)** is strongly anti-correlated (dark colored points; Pearson's r (r_P) = −0.97, p-value < 0.0001) with the median average length of terminal exons in these samples. In contrast, the expression of PTBP2 changes little in tumors compared to normal tissue samples, and has a positive correlation with terminal exon length (light colored points; r_p = 0.85, p-value = 0.002). **(D)** Position-dependent PTBP1 binding inferred from two eCLIP studies (in HepG2 (thick red line) and K562 (thick blue line) cell lines) by the ENCODE consortium is significantly enriched downstream of the 203 PAS predicted to be regulated by the CU-repeat motifs. We selected 1000 similar-sized sets of poly(A) sites with the same positional preference (distally-located) as the targets of the CU motif and the density of PTBP1 eCLIP reads was computed as described in 3.6. The mean and standard deviation of position-dependent read density ratios from these randomized data sets are also shown. **(E)** The median ratio of PTBP1-IP to background eCLIP reads over nucleotides 0 to 100 nt downstream of the PAS (position-wise ratios computed as in **(D)**), for the top 102 ("top") and bottom 101 ("low") predicted PTBP1 targets as well as for the background set ("bg") of distal PAS. **(F)** Activity profile of the same CUCUCU motif in the PTBP1/2 double knock-down (where the motif ranked third) compared to control samples (two biological replicates from HEK cells, number of PAS n = 2493).

suggestion that overexpression of this 3' end processing factor plays a role in lung cancer [98]. Applying KAPAC to 56 matching normal-tumor paired, lung adenocarcinoma samples, we did not find any motifs strongly associated with PAS use changes in this cancer. In particular, we did not recover G/U-rich motifs, as would be expected if CSTF2 were responsible for these changes [98]. This was not due to functional compensation by the paralogous CSTF2T, as the expression of CSTF2T was uncorrelated with the 3' UTR length (Figure 3.6C). Rather, the CSTF2-specific GU repeat motif had highly variable activity between patients and between poly(A) sites, which did not exhibit a peak immediately downstream of the PAS (Figure 3.6D), where CSTF2 is known to bind [33]. Thus, as in glioblastoma, PAS selection in lung adenocarcinoma likely involves factors other than core 3' end processing components.

Exploration of other cancer types for which many paired tumor-normal tissue samples were available revealed that U-rich motifs are more generally significantly associated with changes in PAS use in these conditions (Supplementary Table B.3). Most striking was the association of the presence of poly(U) and AUU motifs with an increased PAS use in colon and prostate cancer, respectively (Figure 3.6E, F). These results indicate that KAPAC can help identify regulators of 3' end processing in complex tissues environments such as tumors.

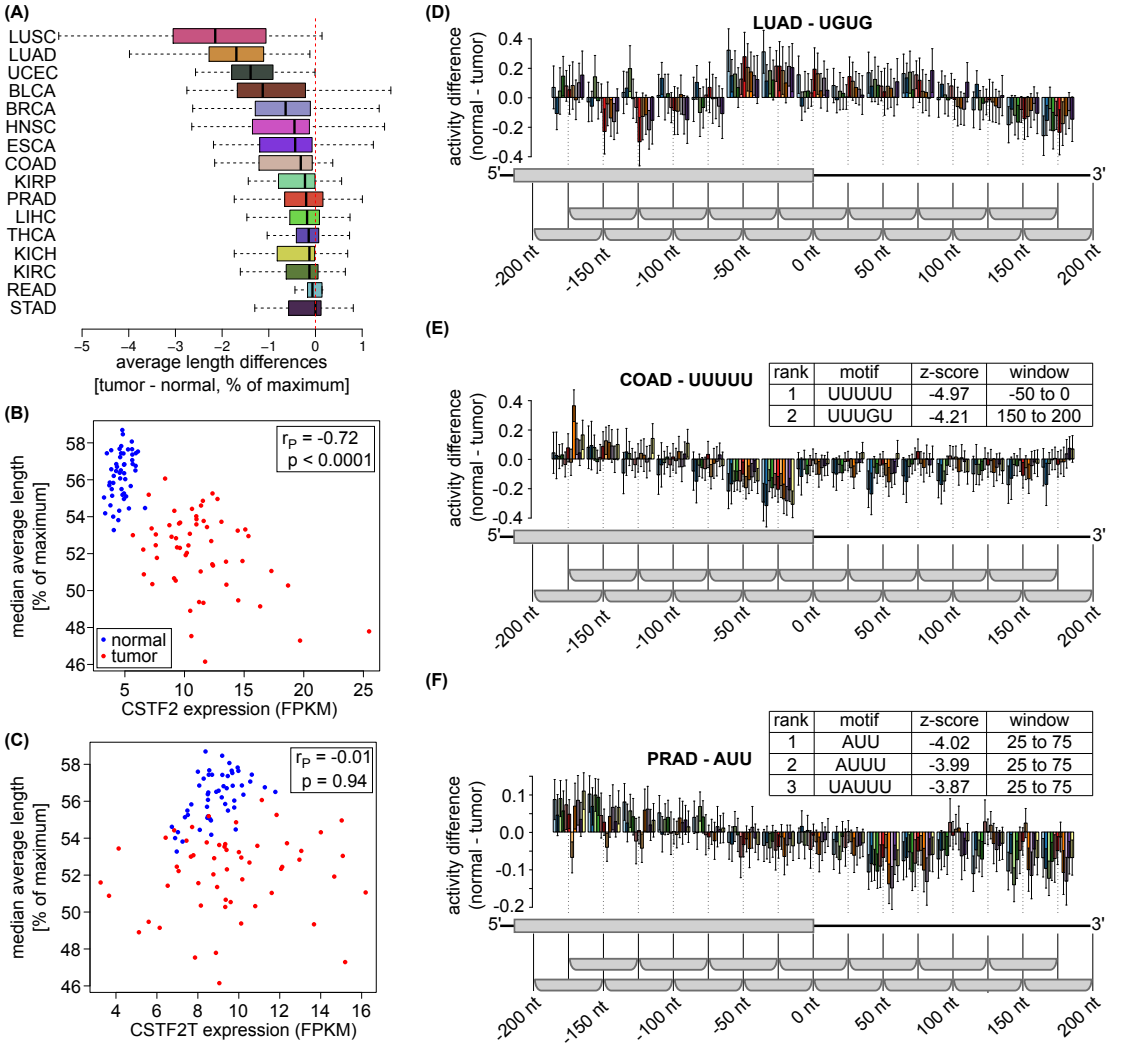


Figure 3.6: Analysis of TCGA data sets. **(A)** For TCGA data sets with at least five matching normal-tumor pairs with high RNA integrity (mTIN > 70), the distributions of patient-wise medians of tumor-normal tissue differences in average terminal exon lengths are shown. Except the adenocarcinoma of the stomach (STAD), the median is negative for all cancers, indicating global shortening of 3' UTRs in tumors. **(B)** Among 56 matching lung adenocarcinoma (LUAD)-normal tissue pairs (from 51 patients) where global shortening of terminal exons was observed, the CSTF2 expression (in fragments per kilobase per million (FPKM)) was negatively correlated ($r_P = -0.72$, p-value = 2.5e−18) with the median of average exon length. **(C)** For the same samples as in **(B)**, no significant correlation ($r_P = -0.01$, p-value = 0.89) between the expression of CSTF2T and the median of average exon length was observed. **(D)** Activity profile of the UGUG CSTF2-binding motif inferred from matched LUAD tumor-normal tissue sample pairs ($n = 1054$). For visibility, 10 randomly selected sample pairs are shown instead of all 56. **(E,F)** Activity profiles of UUUUU and AUU, the motifs most significantly associated by KAPAC with changes in PAS use in colon adenocarcinoma (COAD, number of PAS $n = 1294$) **(E)** and prostate adenocarcinoma (PRAD, number of PAS $n = 1835$) **(F)**, respectively (11 tumor-normal tissue sample pairs in both studies).

3.4 Discussion

Sequencing of RNA 3' ends has uncovered a complex pattern of PAS and 3' UTR usage across cell types and conditions, and particularly that the length of 3' UTRs increases upon cell differentiation and decreases upon proliferation [21, 154]. However, the responsible regulators remain to be identified.

The knock-down of most 3' end processing factors leads to short 3' UTRs [79]. Paradoxically, similar 3' UTR shortening is also observed in cancers, in spite of a positive correlation between expression of 3' end processing factors and the proliferative index of cells [51]. This may suggest that 3' end processing factors are not responsible for 3' UTR processing in cancers, and that other regulators remain to be discovered. However, the possibility remains that 3' end processing factors, although highly expressed, do not match the increased demand for processing in proliferating cells. Although reduced levels of CFIm 25 have been linked to 3' UTR shortening and increased tumorigenicity of glioblastoma cells [23], once we applied a threshold on the RNA integrity in the samples to be analyzed, CFIm 25 expression was similar between tumors and normal tissue samples (Supplementary Figure B.3). Thus, it seems that an apparent low expression of CFIm 25 is associated with stronger 3' end bias in read coverage and partial RNA degradation (Supplementary Figure B.6). Consistently, our KAPAC analysis of samples with high RNA integrity did not uncover the CFIm 25-specific UGUA motif as significantly explaining the PAS usage changes in glioblastoma compared to normal brain tissue. Of note, in the study of Masamha et al. [23] only 60 genes had significantly shortened 3' UTRs in glioblastoma relative to normal brain, and only 24 of these underwent significant 3' UTR shortening upon CFIm 25 knock-down in HeLa cells, in spite of 1453 genes being affected by the CFIm 25 knock-down. However, applying KAPAC to five normal and five glioblastoma tumor samples which showed most separable distributions of terminal exon lengths, we uncovered a pyrimidine motif, likely bound by PTBP1, as most significantly associated with changes in PAS use in these tumors. Our findings are supported by previous observations that PTBP1 acts antagonistically to CSTF2, repressing PAS usage [90], and that increased PTBP1 expression, as we observed in glioblastoma tumors, promotes proliferation and migration in glioblastoma cell lines [187]. Our analysis demonstrates that, *de novo*, unbiased motif analysis of tumor data sets with high RNA integrity can reveal specific regulators of PAS usage.

In spite of mounting evidence for the role of CFIm in the regulation of polyadenylation at alternative PAS in terminal exons, its mechanism has remained somewhat unclear. "Canonical" PAS, containing consensus signals for many of the 3' end processing factors including CFIm, tend to be located distally in 3' UTRs [33]. If core 3' end processing factors bind to specific PAS and select them for processing, reducing the concentration of 3' end processing factors should increase the stringency of PAS selection. Yet the siRNA-mediated knock-down of CFIm leads to increased processing at proximal sites, and not to preferential processing of the "high-affinity", distal PAS. Here we have found that CFIm indeed promotes the usage of distal PAS to which

it binds, while CFIm binding motifs are depleted at both the proximal and the distal PAS of terminal exons whose processing is insensitive to the level of CFIm. Therefore, the decreased processing of distal PAS upon CFIm knock-down is not explained by a decreased "affinity" of these sites. A model that remains compatible with the observed pattern of 3' end processing is the so-called "kinetic" model, whereby reducing the rate of processing at a distal, canonical site when the regulator is limiting, leaves sufficient time for the processing of a suboptimal proximal site [92]. Kinetic aspects of pre-mRNA processing have started to be investigated in cell lines that express slow and fast-transcribing RNA polymerase II (RNAPII) [188]. Analyzing RNA-seq data from these cells, we found that terminal exons that respond to CFIm knock-down in our data underwent more pronounced shortening in cells expressing the slow polymerase (Supplementary Figure B.7), in agreement with the kinetic model. Nevertheless, this effect was also apparent for exons in which proximal and distal poly(A) sites were located far apart; it was not limited to CFIm targets. Furthermore, the changes in 3' UTR length in a sample from the fast RNAPII-expressing cell line were surprisingly similar to the changes we observed for the slow polymerase. Thus, current data do not provide unequivocal support to the kinetic model underlying the relative increase in processing of proximal PAS upon CFIm knock-down.

Generalized linear models have been widely used to uncover transcriptional regulators that implement gene expression programs in specific cell types [189, 190]. Similar approaches have not been applied to 3' end processing, possibly because the genome-wide mapping of 3' end processing sites has been lagging behind the mapping of transcription start sites. Here we demonstrate that the modeling of PAS usage in terms of motifs in the vicinity of PAS can reveal global regulators, while the reconstructed position-dependent activity of their corresponding motifs provides insights into their mechanisms. Interestingly, some of the proteins that we touched upon in our study are splicing factors. This underscores a general coupling between splicing and polyadenylation that has been long surmised (e.g. [179]), and for which evidence has started to emerge [191]. Interestingly, the activities of splicing factors on poly(A) site choice paralleled the activities of these factors on splice site selection. Specifically, we found that both HNRNPC, which functions as an "RNA nucleosome" in packing RNA and masking decoy splice sites [87], and PTBP1, which has repressive activity on exon inclusion [192], repress the processing of the PAS to which they bind. This unexpected concordance in activities suggests that other splicing factors simultaneously modulating 3' end processing are to be uncovered. Splicing is strongly perturbed in cancers [193], and the role of splicing factors in the extensive change of the polyadenylation landscape remains to be defined.

Sequencing of RNA 3' ends has greatly facilitated the study of 3' end processing dynamics. However, such data remain relatively uncommon, and many large-scale projects have already generated a wealth of RNA sequencing data that could, in principle, be mined to uncover regulators of CPA. We found a previously proposed method for inferring the relative use of alternative PAS from RNA-seq data, DaPars [23], to have limited reproducibility, possibly because

biases in read coverage along RNAs are difficult to model. To overcome these limitations, we developed PAQR, which makes use of a large catalog of PAS to segment the 3' UTRs and infer the relative use of PAS from RNA-seq data. We show that PAQR enables a more reproducible as well as accurate inference of motif activities in PAS choice compared to DaPars. PAQR strongly broadens the domain of applicability of KAPAC to include RNA sequencing data sets that have been obtained in a wide range of systems, as we have illustrated in our study of TCGA data. As single-cell transcriptome analyses currently employ protocols designed to capture RNA 3' ends, it will be especially interesting to apply our methods to single-cell sequencing data.

3.5 Conclusions

In this study, we developed PAQR, a robust computational method for inferring relative poly(A) site use in terminal exons from RNA sequencing data, and KAPAC, an approach to infer sequence motifs that are associated with the processing of poly(A) sites in specific samples. We demonstrate that these methods help uncover regulators of polyadenylation in cancers and also shed light on their mechanism of action. Our study further underscores the importance of assessing the quality of samples used for high-throughput analyses, as this can have substantial impact on the estimates of gene expression.

3.6 Methods

3.6.1 Datasets

3.6.1.1 A-seq2 samples

3' End sequencing data from HeLa cells that were treated with either a control siRNA or siRNAs targeting the CFIm 25 and the CFIm 68 transcripts were generated as follows. HeLa cells were cultured in DMEM (# D5671, Sigma Aldrich) supplemented with L Glutamine (#25030081, ThermoFisher Scientific) and 10% fetal bovine serum (#7524, Sigma-Aldrich). For siRNA treatment, cells were seeded in six well polystyrene-coated microplates and cultured to reach a confluence of ~ 50%. Subsequently, the cells were separately transfected with 150 picomoles of siRNA, either control (sense strand sequence 5' AGG UAG UGU AAU CGC CUU GTT 3'), or directed against CFIm 25 (sense strand sequence 5' GGU CAU UGA CGA UUG CAU UTT 3') or against CFIm 68 (sense strand sequence 5' GAC CGA GAU UAC AUG GAU ATT 3'), with Lipofectamine RNAiMAX reagent (#13778030, ThermoFisher Scientific). All siRNAs were obtained from Microsynth AG and had dTdT overhangs. The cells were incubated with the siRNA Lipofectamine RNAiMax mix for at least 48 h before cells were lysed. Cell lysis and polyadenylated RNA selection was performed according to the manufacturer's protocol (Dynabead™ mRNA DIRECT™ Purification Kit, #61011, Thermo Scientific). Polyadenylated RNA was subsequently processed and libraries were prepared for sequencing on the Illumina HiSeq 2500 platform as

described earlier [118]. Sequencing files were processed according to Martin et al. [194] but without using the random 4-mer at the start of the sequence to remove duplicates. A-seq2 3' end processing data from control and si-HNRNPC-treated cells was obtained from a prior study [118].

3.6.1.2 3' end sequencing data pertaining to PCBP1

3' End sequencing data from control and si-PCPB1-treated cells were downloaded from SRA (accession: SRP022151) and converted to fastq format. Reverse complemented and duplicate-collapsed reads were then mapped to the human genome with segemehl version 0.1.7 [168]. We did not use STAR for these data set because these libraries, generated by DRS (direct RNA sequencing) had a high fraction of short reads that STAR did not map. From uniquely mapped reads for which at least the last 4 nucleotides at the 3' end perfectly matched to the reference, the first position downstream of the 3' end of the alignment was considered as cleavage site and used for quantification of PAS use.

3.6.1.3 RNA-seq data from The Cancer Genome Atlas

BAM files for matching normal and tumor RNA-seq samples (the number which is listed in Supplementary Table B.5) were obtained from the Genomic Data Commons (GDC) Data Portal [195] along with gene expression values counted with HTSeq and reported in fragments per kilobase per million (FPKM).

3.6.1.4 Other RNA-seq data sets

Publicly available raw sequencing data were obtained from NCBI's Gene Expression Omnibus (GEO) [196] for the studies of CFIm 25 knock-down in HeLa cells [23] (accession number GSE42420), HNRNPC knock-down in HEK293 cells [147] (GSE56010), PTBP1/2 knock-down in HEK293 cells [186] (GSE69656) and for HEK293 cells expressing mutated versions of POLR2A that have overall different rates of RNAPII transcription elongation [188] (GSE63375).

3.6.1.5 PTBP1 CLIP data

PTBP1-eCLIP data generated by the ENCODE consortium [185] was obtained from the ENCODE Data Coordination Center [197] (accession numbers for the IP and control samples from K562 cells ENCSR981WKN and ENCSR445FZX, and from HepG2 cells ENCSR384KAN and ENCSR438NCK).

3.6.2 Processing of the sequencing data

Raw reads obtained from RNA-seq experiments were mapped according to the RNA-seq pipeline for long RNAs provided by the ENCODE Data Coordinating Center [198] using the

GENCODE version 24 human gene annotation. Raw reads from the study conducted by Gueroussov et al. [186] were additionally subjected to 3' adapter trimming with cutadapt, version 1.14 [199] prior to mapping. Raw reads from eCLIP experiments carried out by the ENCODE consortium for the PTBP1 were first trimmed with cutadapt version 1.9.1 [199], at both the 5' and 3' ends to remove adapters. A second round of trimming guaranteed that no double ligation events were further processed. The reads were then mapped to the genome with STAR, version 2.5.2a [176]. Detection and collapsing of PCR duplicates was done with a custom python script similar to that described by Van Nostrand et al. [183]. BAM files corresponding to biological replicates were then merged.

3.6.3 PAQR

3.6.3.1 Inputs

PAQR requires an alignment file in BAM-format and a file with all poly(A) sites mapped on the genome, in BED-format. The assessment of RNA integrity (see below) also requires the transcript annotation of the genome, in BED12-format.

3.6.3.2 Poly(A) sites

PAQR quantifies the relative use of poly(A) sites in individual terminal exons. We started from the entire set of poly(A) sites in the PolyAsite resource [118], but this set can be exchanged or updated, and should be provided as a BED-file to the tool. We converted the coordinates of the poly(A) sites to the latest human genome assembly version, GRCh38, with liftOver [171]. Terminal exons with more than one poly(A) site (terminal exons with tandem poly(A) sites, TETPS) and not overlapping with other annotated transcripts on the same strand were identified based on version 24 of the GENCODE [152] annotation of the genome. When analyzing RNA-seq data that were generated with an unstranded protocol, PAQR does not quantify poly(A) site usage in terminal exons that overlap with annotated transcripts on the opposite strand.

3.6.3.3 Quantification of PAS usage

The main steps of the PAQR analysis are as follows: first, the quality of the input RNA sequencing data is assessed, to exclude samples with evidence of excessive RNA degradation. Samples that satisfy a minimum quality threshold are then processed to quantify the read coverage per base across all TETPS and poly(A) sites with sufficient evidence of being processed are identified. These are called "used" poly(A) sites (or uPAS). Finally, the relative use of the uPAS is calculated.

3.6.3.4 Assessment of sample integrity

The integrity of RNA samples is usually assessed based on a fragment analyzer profile [117]. Alternatively, a post hoc method, applicable to all RNA sequencing data sets, quantifies the uniformity of read coverage along transcript bodies in terms of a "transcript integrity number" (TIN) [116]. We implemented this approach in PAQR, calculating TIN values for all transcripts containing TETPS. For the analysis of TCGA samples and of RNA-seq samples from cells with different RNAPII transcription speeds, we only processed samples with a median TIN value of at least 70, as recommended in the initial publication [116].

3.6.3.5 RNA-seq read coverage profiles

For each sample, nucleotide-wise read coverage profiles along all TETPS were calculated based on read-to-genome alignments (obtained as described above). In processing paired-end sequencing data, PAQR ensured unique counting of reads where the two mates overlap. When the data were generated with an unstranded protocol, all reads that mapped to the locus of a specific TETPS were assumed to originate from that exon. The locus of each TETPS was extended by 200 nt at the 3' end, to ensure inclusion of the most distal poly(A) sites (see below). To accurately quantify the usage of the most proximal PAS, when poly(A) sites were located within 250 nt of the start of the terminal exon, the coverage profile was first extended upstream of the PAS based on the reads that mapped to the upstream exon(s). Specifically, from the spliced reads, PAQR identified the upstream exon with most spliced reads into the TETPS and computed its coverage. When the spliced reads that covered the 5' end of the TETPS provided evidence for multiple splice events, the most supported exons located even further upstream were also included (Supplementary Figure B.8).

3.6.3.6 Identification of the most distal poly(A) sites

From the read coverage profiles, PAQR attempted to identify the poly(A) sites that show evidence of processing in individual samples as follows. First, to circumvent the issue of incomplete or incorrect annotations of PAS in transcript databases, PAQR identified the most distal PAS in each terminal exon that had evidence of being used in the samples of interest. Thus, alignment files were concatenated to compute a joint read coverage profile from all samples of the study. Then, the distal PAS was identified as the 3'-most PAS in the TETPS for which: 1) the mean coverage in the 200-nt region downstream of the PAS was lower than the mean coverage in a region twice the read length (to improve the estimation of coverage, as it tends to decrease towards the poly(A) site) upstream of the poly(A) site; and 2) the mean coverage in the 200-nt region downstream of the PAS was at most 10% of the mean coverage from the region at the exon start (the region within one read length from the exon start) (Supplementary Figure B.9). For samples from TCGA, where read length varied, we have used the maximum read length

in the data for each cancer type. After the distal PAS was identified, PAQR considered for the relative quantification of PAS usage only those TETPS with at least one additional PAS internal to the TETPS and with a mean raw read coverage computed over the region between the exon start and distal PAS of more than five.

3.6.3.7 Identification of used poly(A) sites

PAQR infers the uPAS recursively, at each step identifying the PAS that allows the best segmentation of a particular genomic region into upstream and downstream regions of distinct coverage across all replicates of a given condition (Figure 3.3C). Initially, the genomic region is the entire TETPS, and at subsequent steps genomic regions are defined by previous segmentation steps. Given a genomic region and annotated PAS within it, every PAS is evaluated as follows. The mean squared error (MSE) in read coverage relative to the mean is calculated separately for the segments upstream (MSE_u) and downstream (MSE_d) of each PAS for which the mean coverage in the downstream region is lower than the mean coverage in the upstream region. A minimum length of 100 nt is required for each segment, otherwise the candidate PAS is not considered further. The sum of MSE in the upstream and downstream segments is compared with the MSE computed for the entire region (MSE_t). If $(MSE_u + MSE_d) / MSE_t \leq 0.5$ (see also below), the PAS is considered "candidate used" in the corresponding sample. When the data set contains at least two replicates for a given condition, PAQR further enforces the consistency of uPAS selection in replicate samples by requiring that the PAS is considered used in at least two of the replicates and, furthermore, for all PAS with evidence of being used in a current genomic region, the one with the smallest median MSE ratio computed over samples that support the usage of the site is chosen in a given step of the segmentation. The segmentation continues until no more PAS have sufficient evidence of being used. If the data consist of a single sample, the segmentation is done based on the smallest MSE at each step.

To further minimize incorrect segmentations due to PAS that are used in the samples of interest but are not part of the input set, an additional check is carried out for each TETPS in each sample, to ensure that applying the segmentation procedure considering all positions in the TETPS rather than the annotated PAS recovers positions that fall within at most 200 nt upstream of the uPAS identified in previous steps for each individual sample (Supplementary Figure B.10). If this is not the case, the data for the TETPS from the corresponding sample are excluded from further analysis.

3.6.3.8 Treatment of closely spaced poly(A) sites

Occasionally, distinct PAS occur very close to each other. While 3' end sequencing may allow their independent quantification, the RNA-seq data do not have the resolution to distinguish between closely spaced PAS. Therefore, in the steps described above, closely spaced (within

200 nt of each other) PAS are handled first, to identify one site of the cluster that provides the best segmentation point. Only this site is then compared with the more distantly spaced PAS.

3.6.3.9 Relative usage and library size normalized expression calculation

Once used poly(A) sites have been identified, library size-normalized expression levels and relative usage within individual terminal exons are calculated. Taking a single exon in a single sample, the following steps are performed: the mean coverage of the longest 3' UTR is inferred from the region starting at the most distal poly(A) site and extending upstream up to the next poly(A) site or to the exon start. Mean coverage values are similarly calculated in regions between consecutive poly(A) sites and then the coverage of an individual 3' UTR is determined by subtracting from the mean coverage in the terminal region of that 3' UTR the mean coverage in the immediately downstream region. As some of the poly(A) sites are not identified in all samples, their usage in the samples with insufficient evidence is calculated as for all other sites, but setting the usage to 0 in cases in which the upstream coverage in the specific sample was lower than the downstream coverage. The resulting values are taken as raw estimates of usage of individual poly(A) sites and usage relative to the total from poly(A) sites in a given terminal exon are obtained.

To obtain library size normalized expression counts, raw expression values from all quantified sites of a given sample are summed. Each raw count is divided by the summed counts (i.e., the library size) and multiplied by 10^6, resulting in expression estimates as reads per million (RPM).

3.6.3.10 PAQR modules

PAQR is composed of 3 modules: 1) a script to infer transcript integrity values based on the method described in a previous study [116]—the script builds on the published software which is distributed as part of the Python RSeQC package version 2.6.4 [200]; 2) a script to create the coverage profiles for all considered terminal exons—this script relies on the HTSeq package version 0.6.1 [201]; and 3) a script to obtain the relative usage together with the estimated expression of poly(A) sites with sufficient evidence of usage. All scripts, intermediate steps, and analysis of the TCGA data sets were executed as workflows created with snakemake version 3.13.0 [202].

3.6.4 KAPAC

KAPAC, standing for k-mer activity on polyadenylation site choice, aims to identify k-mers that can explain the change in PAS usage observed across samples. For this, we model the relative change in PAS usage within terminal exons (with respect to the mean across samples) as a linear function of the occurrence of a specific k-mer and the unknown "activity" of this

k-mer. Note that by modeling the relative usage of PAS within individual terminal exons we will capture only the changes that are due to alternative polyadenylation and not those that are due to overall changes in transcription rate or to alternative splicing. We are considering k-mers of a length from 3 to 6 nt in order to match the expected length of RBP binding sites [184].

KAPAC attempts to explain the change in the relative use of a given PAS in terms of the motifs (k-mers) that occur in its vicinity, each occurrence of a k-mer contributing a multiplicative constant to the site use. Thus, we write the number of reads observed from PAS i in sample s as

$$R_{i,s} = \alpha * \exp(N_{i,k} * A_{k,s}), \tag{3.2}$$

where $N_{i,k}$ is the count of k-mer k around PAS i, $A_{k,s}$ is the activity of the k-mer in sample s, which determines how much the k-mer contributes to the PAS use, and α is the overall level of transcription at the corresponding locus. Then, for poly(A) sites in the same terminal exon we can write their base 2 logarithm relative use $\log(U_{i,s})$ as a function of the number of k-mer counts found in a defined window at a specific distance from the site i and the activity of these k-mers:

$$\log(U_{i,s}) = N_{i,k} * A_{k,s} - \log\Big(\sum_{p=1}^{P} \exp(N_{p,k} * A_{k,s})\Big) \tag{3.3}$$

(see B.2.2.2 for a detailed derivation). By fitting the relative use of poly(A) sites to the observed number of motifs around them, we can obtain the activities $A_{k,s}$ for each k-mer k in each sample s and calculate mean activity difference z-scores across treatment versus control pairs of samples (see Figure 3.1 and B.2).

3.6.4.1 Parameters used for KAPAC analysis of 3' end sequencing data

We considered terminal exons with multiple poly(A) sites within protein coding transcripts (GRCh38, GENCODE version 24) whose expression, inferred as previously described [118], was at least 1 RPM in at least one of the investigated samples. To ensure that the position-dependent motif activities could be correctly assigned, exons containing expressed PAS that were closer than 400 nt from another PAS were excluded from the analysis, as we applied KAPAC to regions ±200 nt around poly(A) sites. We randomized the associations of changes in poly(A) site use with k-mer counts 100 times in order to calculate p-values for mean activity difference z-scores (see B.2).

3.6.4.2 Parameters used for KAPAC analysis of RNA-seq data

All KAPAC analyses for RNA-seq data sets considered terminal exons with at least two PAS of any transcripts from the GENCODE version 24 annotation of the human genome. Filtering of the closely-spaced PAS, activity inference and randomization tests were done similar to the processing of 3' end sequencing libraries. No RPM cutoff was applied as the used PAS

are already determined by PAQR. In the case of TCGA data analysis, mean activity difference z-scores were inferred based on comparisons of tumor versus normal tissue. For the KAPAC analysis of PTBP1/2 knock-down in HEK293 cells, double knock-down samples were considered as control and the actual control samples as treatment, since this comparison corresponds directly to that in the GBM analysis (see also Figure 3.5C and Supplementary Figure B.11).

3.6.5 Average terminal exon length

An average terminal exon length can be calculated over all transcripts expressing a variant of that terminal exon as

$$\hat{l} = \sum_{p=1}^{P} f_p l_p, \tag{3.4}$$

where f_p is the relative frequency of use of PAS p in the terminal exon and l_p is the length of the terminal exon when PAS p is used for CPA. To compare terminal exons with different maximum lengths, we further normalize the average exon length to the maximum and express this normalized value percentually. Thus, when the most distal site is exclusively used the average terminal exon length is 100, while when a very proximal site is used exclusively, the average terminal exon length will be close to 0 (Supplementary Figure B.12).

3.6.6 Average length difference

The difference in average length of a terminal exon between two samples is obtained by subtracting the average length inferred from one sample from the average length inferred from the second sample. 3' UTR shortening is reflected in negative average length differences, while 3' UTR lengthening will lead to positive differences.

3.6.7 Definition of the best MSE ratio threshold

Two studies of HNRNPC yielded 3' end sequencing [118] and RNA sequencing [147] data of control and si-HNRNPC-treated cells. We used these data to define a PAQR parameter (the threshold MSE ratio) such as to maximize the reproducibility of the results from the two studies. MSE ratio values ranging from 0.2 to 1.0 were tested (Supplementary Figure B.13). Relative use of PAS was calculated based on the A-seq2 data sets as described before [118]. The RNA-seq data were processed to infer PAS use with different MSE cutoffs, and we then calculated average terminal exon lengths for individual exons in individual samples and also differences in average exon lengths between samples. For the comparison of the RNA-seq based PAS quantifications with those from A-seq2, we considered both the overall number of terminal exons quantified in replicate data sets as well as the correlation of average length differences. As shown in Supplementary Figure B.13 stringent (low) cutoff in MSE leads to few exons being quantified with high reproducibility, but the number of quantified exons has a peak relative to the MSE. At a threshold of 0.5 on MSE we are able to quantify the largest number of exons

with relatively good reproducibility, and we therefore applied this value for all our subsequent applications of PAQR.

3.6.8 Selection of normal-tumor sample pairs for analysis of 3' UTR shortening

For the analysis of motifs associated with 3' UTR length changes in cancers, we computed the distribution of 3' UTR length differences in matched tumor-normal samples. We carried out hierarchical clustering of vectors of 3' UTR length changes for each cancer type separately (using Manhattan distance and complete linkage). We then identified the subcluster in which the median change in 3' UTR length was negative for all samples and that also contained the sample where the median change over all transcripts was smallest over all samples. Samples from these clusters were further analyzed with KAPAC.

3.6.9 Selection of normal-tumor pairs from GBM data

From the six normal tissue sample that had a median transcript integrity number > 70, five had similar average exon length distributions (all of them being among the samples with the highest median average length). We used these five normal tissue samples and selected five primary tumor samples with similarly high TIN and the lowest median average exon length. We then generated random pairs of normal-tumor tissue samples and analyzed them similarly to paired samples from other cancers.

3.6.10 eCLIP data analysis

We predicted targets of the CU-repeat motif as described in B.2 and obtained a total of 203 targets. We either used the entire set or divided the set into the top half and bottom half of targets. For each poly(A) site from a given set, the read coverage profiles of the 400 nt region centered on the poly(A) site were constructed from both the protein-specific immunoprecipitation (IP) experiment and the related size-matched control. At every position, we computed the ratio of the library size normalized read coverage (RPM) in the IP and in the background sample (using a pseudo-count of 0.1 RPM) and then average these ratios position-wise across all poly(A) sites from a given set, considering any poly(A) site with at least a single read support in either of both experiments. For comparison, we carried out the same analysis for 1000 random sets of poly(A) sites with the same size as the real set, and then inferred the mean and standard deviation of the mean read ratios at each position.

3.6.11 Motif profiles

Motif profiles were generated by extracting the genomic sequences (from the GRCh38 version of the human genome assembly) around poly(A) sites from a given set, scanning these sequences and tabulating the start positions where the motif occurred. The range of motif

occurrence variation at a given position was calculated as the standard deviation of the mean, assuming a binomial distribution with the probability of success given by the empirical frequency (smoothened over 7 nucleotides centered on the position of interest) and the number of trials given by the number of poly(A) sites in the set.

3.6.12 Selection of CFIm-sensitive and insensitive terminal exons

For terminal exons with exactly two quantified poly(A) sites that were expressed with at least 3 RPM in all samples (1776 terminal exons) we calculated the proximal/distal ratio. Next, we calculated the average (between replicates) $\log_{10}$ fold change (in knock-down relative to control) in proximal/distal ratio. The 750 terminal exons with the largest average $\log_{10}$ fold change in the CFIm 25 and CFIm 68 knock-down experiments were selected as CFIm sensitive, while the 750 with an average $\log_{10}$ fold change closest to zero were considered insensitive.

3.7 List of abbreviations

TGCA cancer cohort abbreviations used in the previous chapter correspond to the following full names:

BCLA: Bladder Urothelial Carcinoma
BRCA: Breast Invasive Carcinoma
COAD: Colon Adenocarcinoma
ESCA: Esophageal Carcinoma
GBM: Glioblastoma Multiforme
HNSC: Head and Neck Squamous Cell Carcinoma
KICH: Kidney Chromophobe
KIRC: Kidney Renal Clear Cell Carcinoma
KIRP: Kidney Renal Papillary Cell Carcinoma
LIHC: Liver Hepatocellular Carcinoma
LUAD: Lung Adenocarcinoma
LUSC: Lung Squamous Cell Carcinoma
PRAD: Prostate Adenocarcinoma
READ: Rectum Adenocarcinoma
STAD: Stomach Adenocarcinoma
THCA: Thyroid Carcinoma
UCEC: Uterine Corpus Endometrial Carcinoma

3.8 Declarations

3.8.1 Availability of data and materials

3' End sequencing data from HeLa cells treated with control siRNA or siRNAs directed against CFIm 25 and CFIm 68 and generated with the A-seq2 protocol [194] have been submitted to the NCBI Sequence Read Archive (SRA) [203] and are available under accession number SRP115462. A-seq2 data pertaining to HNRNPC were obtained from SRA under accession number SRP065825. Direct RNA sequencing data from the PCBP1 study of Ji et al. [85] were obtained from SRA with accession number SRP022151. RNA sequencing data from the studies involving CFIm 25 knock-down [23], HNRNPC knock-down [147], PTBP1/2 knock-down [186] and RNAPII with altered elongation rate [188] were obtained from GEO [196], with accession numbers GSE42420, GSE56010, GSE69656, and GSE63375, respectively. Data from the eCLIP study of PTBP1 were obtained from the ENCODE Data Coordination Center [197], having the following accession numbers: ENCSR981WKN, ENCSR445FZX, ENCSR384KAN and ENCSR438NCK. TCGA data (sample sets listed in Supplementary Tables B.4 and B.5) were obtained from the GDC Portal [195], following permission. The source code of PAQR and KAPAC is available from `https://github.com/zavolanlab/PAQR_KAPAC.git`. The snakemake pipeline to execute PAQR and KAPAC as we have done in the manuscript, with input data pertaining to HNRNPC as an example is available from `https://doi.org/10.5281/zenodo.1147433`. Both are distributed under the terms of the GNU General Public License as published by the Free Software Foundation which permits the free redistribution and/or modification of the code.

3.8.2 Acknowledgements

We are grateful to the specimen donors and to the research groups that were part of the TCGA research network for making these data available. We would like to thank Florian Geier for fruitful discussions and sharing R code for regression models. Also, we would like to thank the sciCORE team for their maintenance of the HPC facility at the University Basel and John Baumgartner for his R implementation of Iwanthue (`https://github.com/johnbaums/hues/blob/master/R/iwanthue.R`).

3.8.3 List of authors

The following authors have contributed to the work discussed in Chapter 3:

1. Andreas Johannes Gruber[1] (Abbr.: AJG),
2. Ralf Schmidt[1] (Abbr.: RS),
3. Souvik Ghosh[1] (Abbr.: SG),
4. Georges Martin[1] (Abbr.: GM),
5. Andreas R. Gruber[1] (Abbr.: ARG),

6. Erik von Nimwegen[1] (Abbr.: EvN) &
7. Mihaela Zavolan[1] (Abbr.: MZ)

[1] Biozentrum, University of Basel, Klingelberstrasse 50-70, CH-4056 Basel, Switzerland

3.8.4 Authors' contributions

The order of authors in the previous subsection (3.8.3) reflects the authors' contributions, with the first two authors (AJG and RS) contributing equally to this work. The last two authors are principal investigators and thus their listing follows the opposite ranking.

AJG developed KAPAC, RS developed PAQR, SG and GM generated the 3' end sequencing data in HeLa cells, ARG contributed to the analysis of 3' end sequencing data sets with KAPAC and EvN contributed to the KAPAC model. AJG and RS analyzed the 3' end and RNA sequencing data sets. RS analyzed the TCGA data sets. Mihaela Zavolan contributed to model development and analyses. AJG, RS, MZ wrote the manuscript with help from all authors.

3.8.5 Ethics approval and consent to participate

Authorization to use RNA-seq data from patient samples, which is obtained by the TCGA Research Network, has been granted.

3.8.6 Funding

This work was supported by Swiss National Science Foundation grant #31003A_170216 to MZ and by the project #51NF40_141735 (National Center for Competence in Research "RNA & Disease").

3.8.7 Competing interests

The authors declare that they have no competing interests.

3.8.8 Additional files

Supplementary materials can be found in Appendix B.

CHAPTER 4

DISCUSSION

Alternative polyadenylation (APA) has been emerged as an important mechanism for regulating gene expression in higher eukaryotes [7, 12, 128, 141]. The characterization of individual APA events elucidated the fundamental role of this mechanism in diverse cellular processes: From the early example of the protein isoform switch of the IgM heavy chain during B cell activation [80] to the formation of paraspeckles which depends on a short isoform of the long non-coding RNA NEAT1 [180] to the 3' UTR dependent protein localization of CD47 which is shuttled to the plasma membrane only upon the translation of its long 3' UTR isoform [72]. Apparently, the impact of APA exceeds beyond the regulation of the RNA metabolism and modulates a wide range of cellular functions. Yet, it was the detection of intriguing dynamics and systematic alterations of 3' end processing during cell state transitions that generated a new impetus in the research of APA. Proliferating cells including cancer show a preferential use of proximal poly(A) sites (PAS) while differentiating cells systematically express long isoforms matured at more distal PAS' [21, 22, 53, 154, 204, 205]. These insights unveiled APA as regulatory step that putatively impacts the cellular state globally. Even though relatively little is known about the underlying mechanisms that are responsible for APA in physiological contexts, pioneering studies shed light on potential key regulators, best known being the CFI complex which is a part of the core processing machinery. CFI knock-down has been shown by our laboratory and others to cause global shifts of poly(A) site usage leading to genome-wide 3' UTR shortening [23, 33]. Subsequent results indicated far reaching consequences for APA events causes by CFI: cell differentiation and reprogramming are sensitive to the levels NUDT21, a component of the CFI complex [206]; NUDT21 levels were also implicated in the activation of oncogenes in liver cancer [207] and in tumor growth of glioblastoma [23]; copy-number variations of NUDT21 were proposed as reason for mental disability [103]. Such findings promoted the tempting

idea, that the cellular state can be modulated in a directed and specific manner by the targeted manipulation of individual regulators of alternative polyadenylation.

The work of this thesis was devoted to a better understanding of the processes and regulators that modulate APA. In recent years, various protocols to specifically sequence the 3' ends of polyadenylated mRNAs have been developed and applied in different biological contexts (a summary of considered protocols can be found in Chapter A). The integrative analysis of the corresponding data sets as presented in Chapter 2 resulted in a comprehensive atlas of poly(A) sites for the human and the mouse genome. This resource provides experimentally supported genomic sites of 3' end processing with a single-nucleotide resolution, which opened the door for analyses of biological aspects of APA.

The large number of individual 3'ends inferred from a wide range of conditions could be exploited for the characterization of sequence motifs important for 3' end processing in unprecedented detail. The focus was on the poly(A) signal which is considered to be the core sequence element for PAS recognition [11, 46]. The applied approach, developed during these PhD studies, revealed additional conserved hexameric motifs that likely function in cleavage and polyadenylation. An interesting question that emerges from these findings in combination with recent insights into structural aspects of the binding of the CPSF complex to AAUAAA [45, 47] is how CPSF can bind a wide array of eighteen or more sequence motifs to guide the 3' end processing reaction. It has been known for a long time that the presence of variant poly(A) signals leads to less efficient cleavage [26]. This can be partially explained through results on the structure of the CPSF-RNA complex that revealed a decreased binding affinity of CPSF4 and WDR33 to AAGAAA (personal communication with Clerici et al.). A mutation in the poly(A) signal of the HBA2 (α2-globin) that changes the canonical signal to AAUAAg causes α-thalassaemia [208] while the AAcAAA mutation in HBB (β-globin) leads to β-thalassaemia [209], both suggesting a complete disruption of the 3' end processing at those sites. In contrast, our analysis revealed the recurring availability of both signals upstream of processed 3' ends with the same positional preference as the canonical AAUAAA, clearly suggesting cleavage and polyadenylation at these PAS. Maybe, processing of these sites can be rescued by other auxiliary motifs in the up- and downstream regions [31] whereas the same AACAAA or AAUAAG signal, acquired through a mutation, can render a PAS non-functional in other sequence contexts. Apparently, input from multiple factors and sequence elements determine the functionality and processing efficiency of poly(A) sites. The extended set of poly(A) signals and their corresponding PAS provides valuable information about possible sequence motif contexts of *bona fide* 3' ends.

In a first application, the poly(A) site atlas facilitated the identification of HNRNPC as a repressor of PAS usage. The heterotetramer had been implicated before mainly in the regulation of splicing [87]. Here, it was demonstrated that HNRNPC levels affect 3' end processing globally and the 3' UTR isoform of the CD47 transcript in particular. The corresponding CD47 protein

was shown to undergo 3' UTR dependent protein localization: only the protein translated from the long transcript isoform is shuttled to the cell membrane whereas the protein from the short form remains in the endoplasmatic reticulum [72]. Thus, HNRNPC acts as an upstream regulator of this mechanism which makes the functionality of CD47 as cell surface protein directly dependent on HNRPC levels. While this relation was described here in the context of HNRNPC knock-down, it might be relevant also in physiological conditions: CD47 was reported to be upregulated in cancer cells to inhibit phagocytosis [210], a function that requires CD47 to reside in the plasma membrane, hence, that can be modulated by HNRNPC levels.

The results of this initial study brought the physiological contexts of APA into focus. Consequently, the follow-up project aimed to better characterize the mechanisms responsible for the dynamic APA changes in cancer. The constantly decreasing costs for high-throughput RNA sequencing (RNA-seq) made this technology widely accessible and prompted collaborative initiatives like The Cancer Genome Atlas (TCGA) to establish large-scale repositories of data from matching normal and tumor tissue samples, with the number of cases per cancer type being in the range of 100-1000 [182]. However, the RNA-seq data did not allow an immediate inference of PAS usage. Thus, to obtain accurate estimates, PAQR (**p**oly**a**denylation site **q**uantification from **R**NA-seq) was developed, a method that quantifies 3' end processing events from RNA-seq data based on the distribution of sequencing reads along terminal exons. Although the principle approach of PAQR to detect drops in read coverage that are indicative of 3' end processing was used before [98, 110, 111, 112], the nonuniform coverage profile leads to many false positive hits and inaccurate usage estimates when utilized for the *de novo* PAS detection. PAQR instead relies on the poly(A) site atlas and identifies genomic positions for which a read coverage drop was in concordance with an annotated poly(A) site. This approach made PAQR less susceptible to false positives and simultaneously gives it single nucleotide precision in PAS usage quantification.

The processing of the TCGA data sets revealed an unexpected variability in read coverage of the terminal exons across samples. In several cases the read distribution was very skewed with an enrichment at the 3' end of exons indicating advanced RNA degradation [115]. Sample collection in clinical settings is often accomplished under conditions that are not optimal to preserve the RNA. RNA degradation has been recognized as an important confounder in the analysis of clinical samples [114, 116].

The analysis of quality-controlled samples from the TCGA provided further corroboration for global 3' UTR shortening of tumor compared to normal tissue across a large number of cancer types [22, 98]. The application of a computational approach to infer sequence motifs whose abundance is significantly associated with PAS usage changes (called KAPAC for ***k***-mer **a**ctivity on **p**oly**a**denylation site **c**hoice) was intended to unravel novel aspects of APA regulation in cancer. Unexpectedly, KAPAC and further analyses revealed the polypyrimidine tract binding protein 1 (PTBP1) as putative regulator of APA in GBM. Previously, mainly factors of the core 3'

end processing machinery were proposed to regulate APA in cancer [23, 98, 99] whereas PTBP1 does not belong to this complex. Nevertheless, PTBP1 joins the rank of factors with observed activity in splicing and 3' end processing [90, 186, 192]. Moreover, its inferred repressive effect on PAS usage upon binding matches the well-studied role in splicing to primarily prevent exon inclusion [192]. PTBP1 is ubiquitously expressed across many although not all human tissues and belongs to the subfamily of heterogeneous nuclear ribonucleoproteins that were reported to regulate various aspects of mRNA metabolism including pre-mRNA processing or mRNA export [211]. During neurogenesis, skipping of exon 9 leads to a reduced PTBP1 activity on splicing regulation resulting in an alternative splicing program [186]. Moreover, PTBP1 levels were found to correlate with clinical features of Parkinson's disease and with the degree of transformation of mammary epithelial cells indicating a potential role of PTBP1 for disease [212, 213]. Evidently, PTPB1 is of physiological relevance and its regulatory activity affects the cellular state.

Interestingly, elevated PTBP1 levels were associated with advanced proliferation and migration of glioma cell lines [187]. Similarly, another study linked 3' UTR shortening in glioma cell lines with anchorage-dependent growth and cellular invasion [23]. One interesting interpretation of both results is to attribute a direct role of alternative 3' end processing to cancer progression and tumorigenesis. It will be of particular interest to clarify if APA events are driving certain tumor characteristics or if they concomitantly emerge during cellular transformation. If PTBP1 can directly impact tumorigenesis or tumor growth, it might be an potential therapeutic target: Its modulation will have global consequences for the gene expression of cells due to its involvement into splicing and 3' end processing.

Additionally to novel insights into the regulation of APA in GBM, the presented approach revealed a potential role for uridine-rich motifs in the control mechanisms of poly(A) site choice for several cancers. Further analyses will be required to associate the identified motifs with the factors by which they are bound and that exert the regulation of PAS usage. Promising advances in this direction entail large-scale studies of RNA-protein interactions with crosslinking and immunoprecipitation (CLIP) and similar approaches [183, 214]. CLIP consists in ultraviolet light-induced crosslinking of RNA binding proteins (RBPs) to RNAs, precipitation of the RBP of interest with a specific antibody, and sequencing of RBP-bound RNA fragments. This allows one to infer positional information about the sites of interaction between RBP and RNA that can be used subsequently to select enriched sequence motifs in these regions of binding. Such short sequence elements can then be considered as specific binding sites for the corresponding RBP [183]. Another method, called RNA Bind-n-Seq, incubates a purified RBP that contains a streptavidin binding peptide tag with a pool of randomized RNAs and selects RNA bound to the protein via magnetic beads. Finally, the short RNA fragments are sequenced and allow a more direct investigation of binding motifs [215]. Both studies, CLIP and RNA Bind-n-Seq potentially define the binding motifs for a near to complete set of RBPs. These can then be intersected

with motifs revealed by our approach to draw conclusions about possible regulators.

However, not for all analyzed cancer types a clear and recurring motif was determined. This might be due to the fact that PAS usage changes can not be sufficiently described through the abundance of individual motifs. Another reason might be the intra-cancer heterogeneity of the surveyed samples. Earlier reports indicated that PAS usage patterns are specific to individual subtypes or that these patterns are even sufficiently characteristic for the classification of cancer subtypes [96, 102]. Hence, combining all samples from a single cancer type in one analysis might reduce the capacity to capture existing signals when the variability in 3' end processing patterns between distinct subtypes does not allow the inference of dominant whole-cancer motif elements. Potentially, better results can be obtained through the prior selection of a subgroup of samples which might be even identified by clustering the samples based on poly(A) site usage, and separate analysis of individual subtypes.

The patterns of 3' end usage may also change depending on the tumor grade. Earlier results already indicated that the stratification of B cell leukemia in mice adds prognostic power [102]. Hence, the identification of APA events that correlate with the progression of tumors may be exploited as diagnostic biomarkers, support the classification of cancer subtypes or even serve as therapeutic targets. Especially for the last case, the comparison of APA patterns across cancer types might help to reveal recurring and biologically relevant APA events that can be used for the design of therapeutic agents suited for the treatment of different cancer types.

With HNRNPC and PTBP1, two factors that were before described as splicing regulators [87, 186, 192] were implicated in the regulation of APA in the presented projects. These findings strongly support earlier reports of the involvement of splicing factors like U1 small nuclear ribonucleoprotein (U1 snRNP, or simply U1) in the regulation of 3' end processing [82] and emphasize the extensive integration of both processes. Multiple interactions of core components of the splicing and 3' end processing machineries are known (for an overview see [52]) which for example promote the coordinated processing of the 3' terminal intron and CPA [216, 217, 218]. The results of this thesis prompt a coupling of both regulatory mechanisms beyond the interaction of factors directly involved in the RNA processing. Instead, they suggest that splicing and 3' end processing are regulated through the same factors and maybe even through similar mechanisms. Further support for this theory comes for example from a study that examined the binding specificity and the impact of binding for the neuron-specific regulator NOVA2 [86]. It will require further effort to unravel the full scope of interdependence between both mechanisms of post-transcriptional gene expression regulation.

Another feature that is shared between HNRNPC and PTBP1 is their repressive effect on poly(A) site usage. Importantly, the presumed mechanism to regulate poly(A) site choice for both factors differs in a crucial detail from the proposed model for U1: While U1 masks proxi-

mal poly(A) sites to prevent premature cleavage during transcription [82], our analyses indicate that HNRNPC and PTBP1 also repress the usage of distal poly(A) sites. From a mechanistic point of view, this raises the question how the cell ensures the correct 3' end processing when the proximal poly(A) site was already transcribed and skipped from processing but the distal poly(A) site is unavailable due to the binding of HNRNPC or PTPB1. It is completely unclear whether the maturation of transcripts in such situation fails and they are decayed. Alternatively, one can conjecture specific "connection" between the distal and a more proximal poly(A) site that allows a delayed processing of the proximal site in such situation. Testing the validity of both models is not easy because it requires detailed information about the processing events on the level of single molecules. However, if available the method would be similarly suited to examine the "kinetic model" which basically states that different PAS compete for processing during transcription [92].

The prevalence of alternative polyadenylation in diverse physiological contexts provides testimony for its relevance in modulating post-transcriptional gene expression regulation. Despite major advances to characterize the PAS usage patterns and their changes in various conditions, surprisingly little is known about the fundamental mechanisms that guide such changes. The development and application of computational approaches in the course of the presented projects provided novel biological insights and contributed to the discovery of regulators of APA. With the prospective technological improvement of the current RNA-seq methods, the read coverage profiles will become more accurate which will enhance the ability to quantify poly(A) site usage based on such data. Moreover, the next revolution of sequencing has already started and in the medium term single-molecule RNA sequencing will be available. This technique will ease the identification and quantification of poly(A) site usage drastically and will give much more accurate insights into the context dependent changes in the transciptome. These novel sequencing methods are equally connected with the expectation to enable the direct inference of RNA modifications. Hence, the technological progress will drive the search for a different class of APA regulators other than *trans*-acting factors that bind to sequence motifs. Already today, the methylation of adenosines (m^6A), the most abundant modification of mRNAs [219], was associated with 3' end processing: the knock-down of m^6A writers was shown to cause APA [220]. Presumably, also other RNA modifications are coupled with the post-transcriptional regulation of gene expression. Another factor with elusive effect on the regulation of APA is the secondary structure. Maybe, sequencing full transcripts will simultaneously allow to acquire information on the secondary structure of the message. This would enable the analysis of the effect of structure elements on the processing of poly(A) sites. In the prospect of upcoming developments, the presented results can only be considered as a starting point for the analysis of alternative polyadenylation. However, the obtained understanding of the data and their biological interpretation will equally support

the interpretation of future results. Also, the large repertoire of available data sets will remain a rich resource and may be re-evaluated in the light of novel insight that were obtained on limited but more accurate measurements.

Especially the results on the regulation of PAS usage in GBM motivates future work on cancer-related APA events in our laboratory and perhaps others. Altered poly(A) site usage in cancer provides a highly relevant context for studying 3' UTR based regulation and may allow one to manipulate the behavior of cancer cells. Possibly, restoring the poly(A) site usage pattern of tumor cells to match those of its normal counterparts would support a reconversion of the cellular state.

APPENDIX A

SUPPLEMENTARY MATERIAL TO CHAPTER 2

A.1 3' end sequencing protocols

A.1.1 2P-Seq

In the 2P-Seq protocol, reverse transcription is accomplished by an anchored oligo(dT) primer. The products of reverse transcription and PCR amplification are expected to have 20 As preceding the 3' adapter. Libraries are sequenced in anti-sense direction with a custom primer. Reads should be reverse complemented [55, 130].

A.1.2 3'-Seq

In the 3'-Seq protocol of Mayr and colleagues, reverse transcription is accomplished by an anchored oligo(dT) primer. The products of reverse transcription and PCR amplification are expected to have 17 As preceding the 3' adapter. Libraries are sequenced in sense direction requiring removal of the 3' adapter sequence and preceding As to pinpoint the 3' end [19].

A.1.3 3P-Seq

In the 3P-seq protocol, a biotinylated adapter is ligated to the end of the poly(A) tail via splint-ligation. After partial digestion, poly(A) regions are captured with streptavidin and reverse transcription is carried out only with dTTP. Most of the poly(A) tail is then removed through RNase H digestion. Adapter ligation, reverse transcription and PCR amplification follow before the library is sequenced in anti-sense direction. Consequently, pinpointing the 3' end requires the reads to be reverse complemented [18, 58].

A.1.4 3'READS

3' region extraction and deep sequencing (3'READS) is a protocol that utilizes a special primer (45 thymidines followed by 5 uridines) to capture poly(A) containing RNA fragments. RNase H digestion releases transcripts 3' ends from the most of the poly(A) tail. Subsequently, the fragments are subjected to adapter ligation, reverse transcription, and PCR amplification before they are sequenced in anti-sense direction. The cleavage site is inferred as the first non-A of the 3' end of the read's reverse complement [53, 79].

A.1.5 A-seq

In the A-seq protocol, reverse transcription is accomplished by an anchored oligo-dT primer. The products of reverse transcription and PCR amplification are expected to have six As preceding the 3' adapter. Libraries are sequenced in sense direction requiring removal of the 3' adapter sequence and preceding As to pinpoint the 3' end [120].

A.1.6 A-seq (version 2)

The second version of the A-seq protocol has the following changes: (1) The steps of the protocol are conducted such that the generation of adapter dimers is minimized. (2) Libraries are sequenced in anti-sense direction and the mRNA cleavage site is inferred as the first nucleotide after a stretch of 4 random nucleotides and 3 Ts [56].

A.1.7 DRS

In the direct RNA sequencing (DRS) protocol, 3' ends of transcripts are hybridized to poly(dT)-coated flow cell surfaces where antisense strand synthesis is initiated. This has the advantage that no prior reverse transcription or cDNA amplification is needed [49, 85, 143, 221].

A.1.8 PAS-seq

In the PAS-Seq protocol, reverse transcription is accomplished with an anchored oligo-dT primer. The products of reverse transcription and PCR amplification are expected to have 20 As preceding the 3' adapter. Libraries are sequenced in anti-sense direction with a custom primer requiring the reverse complement of the reads to pinpoint the 3' end [6].

A.1.9 PolyA-seq

Library preparation for the PolyA-seq protocol includes the following steps: (1) Reverse transcription, primed with an oligo-dT sequence, (2) second strand synthesis with random hexamers linked to a second PCR primer, and (3) PCR amplification. Sequencing is accomplished in

anti-sense orientation with a primer ending in 10 Ts and the resulting reads need to be reverse complemented to pinpoint the pre-mRNA cleavage site [7, 84].

A.1.10 SAPAS

In the SAPAS protocol, reverse transcription is accomplished by an anchored oligo-dT primer. The products of reverse transcription and PCR amplification are expected to have the sequence AAAAAAGAAAAAAGAAAAAA preceding the 3' adapter. Libraries are sequenced in anti-sense direction with a regular primer requiring to trim 20 nucleotides from the 5' end of reads and to reverse complement reads to pinpoint the 3' end [100, 127].

A.2 Supplementary Figures

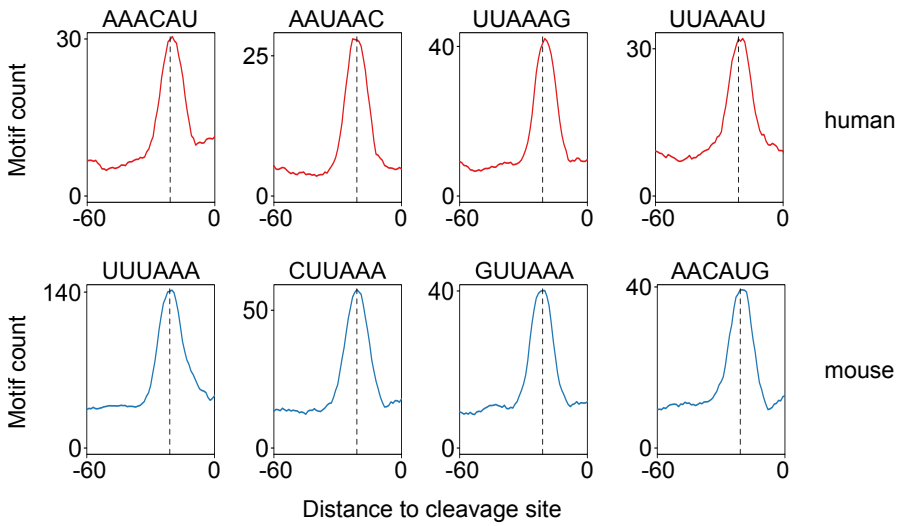


Figure A.1: Frequency profiles of the poly(A) signals that have been identified only in human (red) or mouse (blue).

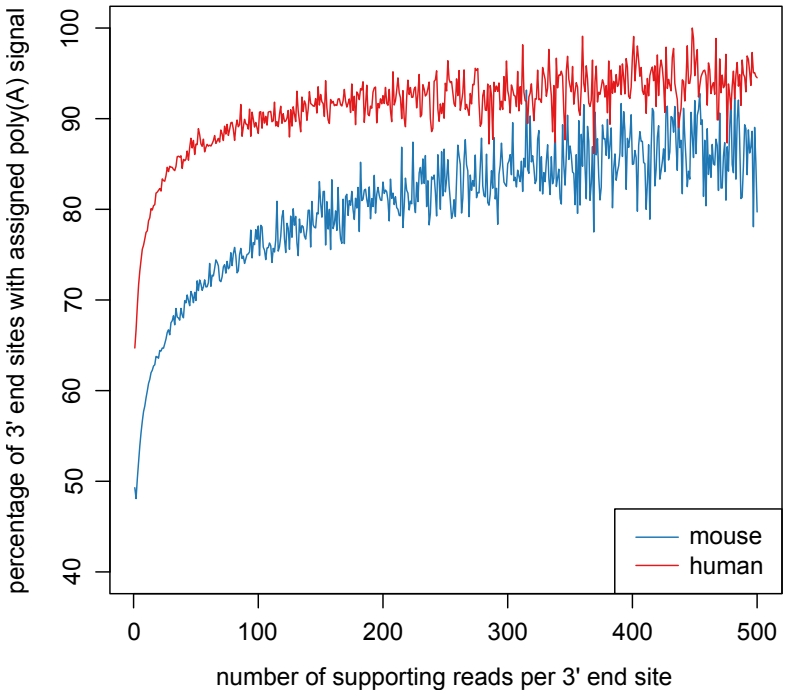


Figure A.2: Fraction of PAS' with poly(A) signal. Fraction of the putative 3' end sites with an assigned poly(A) signal in their upstream region (60 to 10 nucleotides upstream) as a function of the number of supporting reads per site (summed reads over all considered samples).

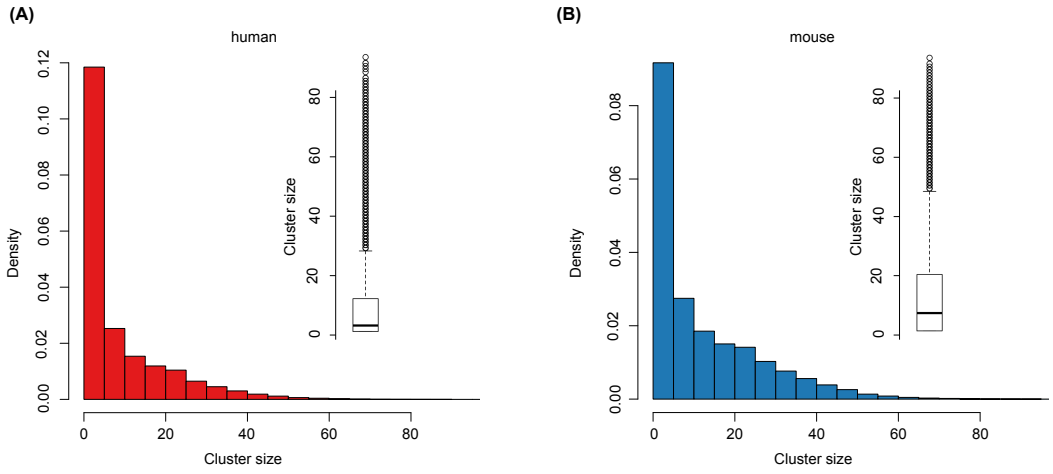


Figure A.3: Distribution of cluster sizes. **(A)** human catalog **(B)** mouse catalog. The large majority of clusters has a short span (less than 20 nt) in both human and mouse.

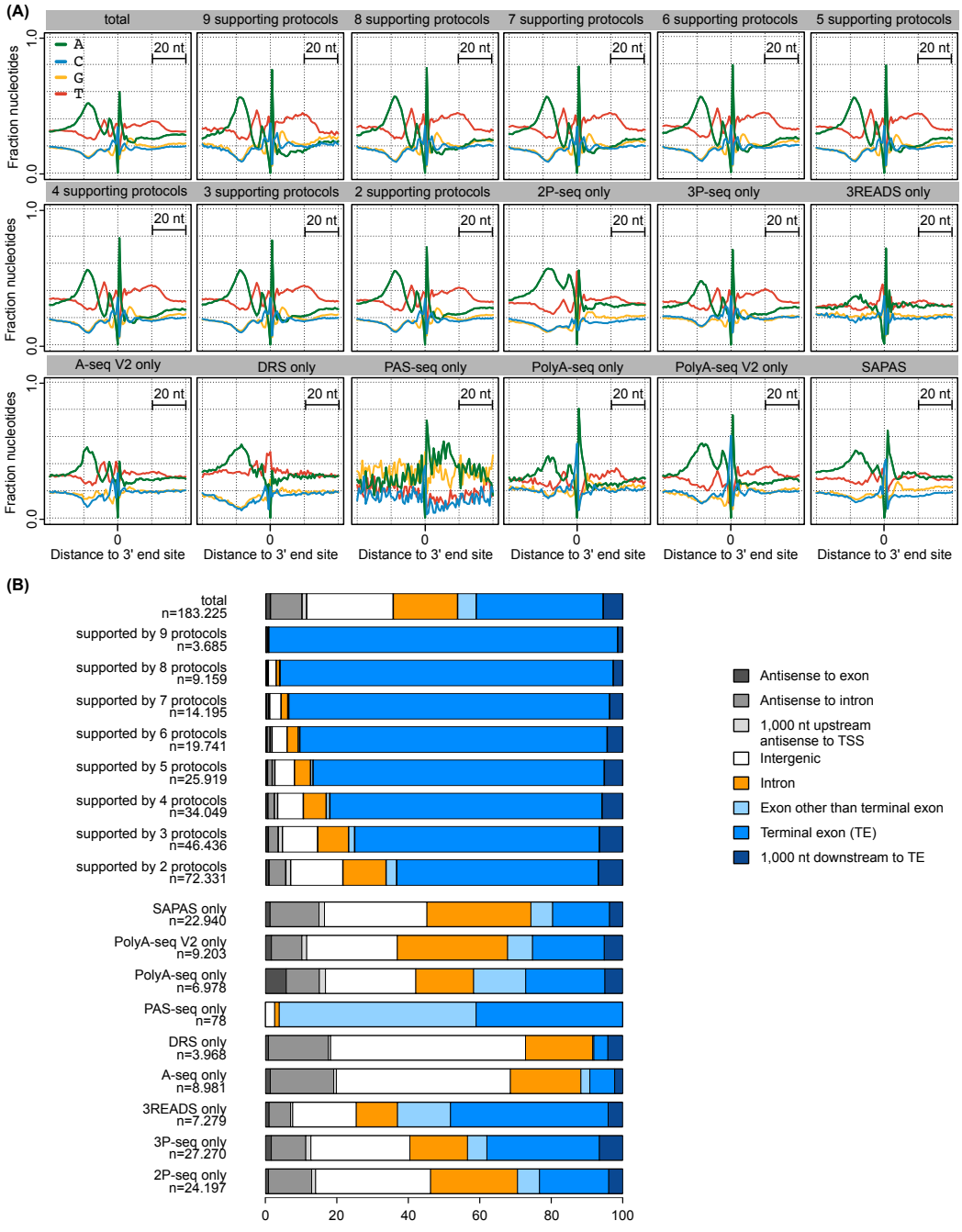


Figure A.4: Characteristics of mouse poly(A) clusters. (A) Nucleotide composition around cleavage sites supported by the indicated number of protocols or the name of the protocol for clusters that had a single protocol support. **(B)** Annotation of clusters supported by various types of protocols (n - number of poly(A) clusters in the indicated category).

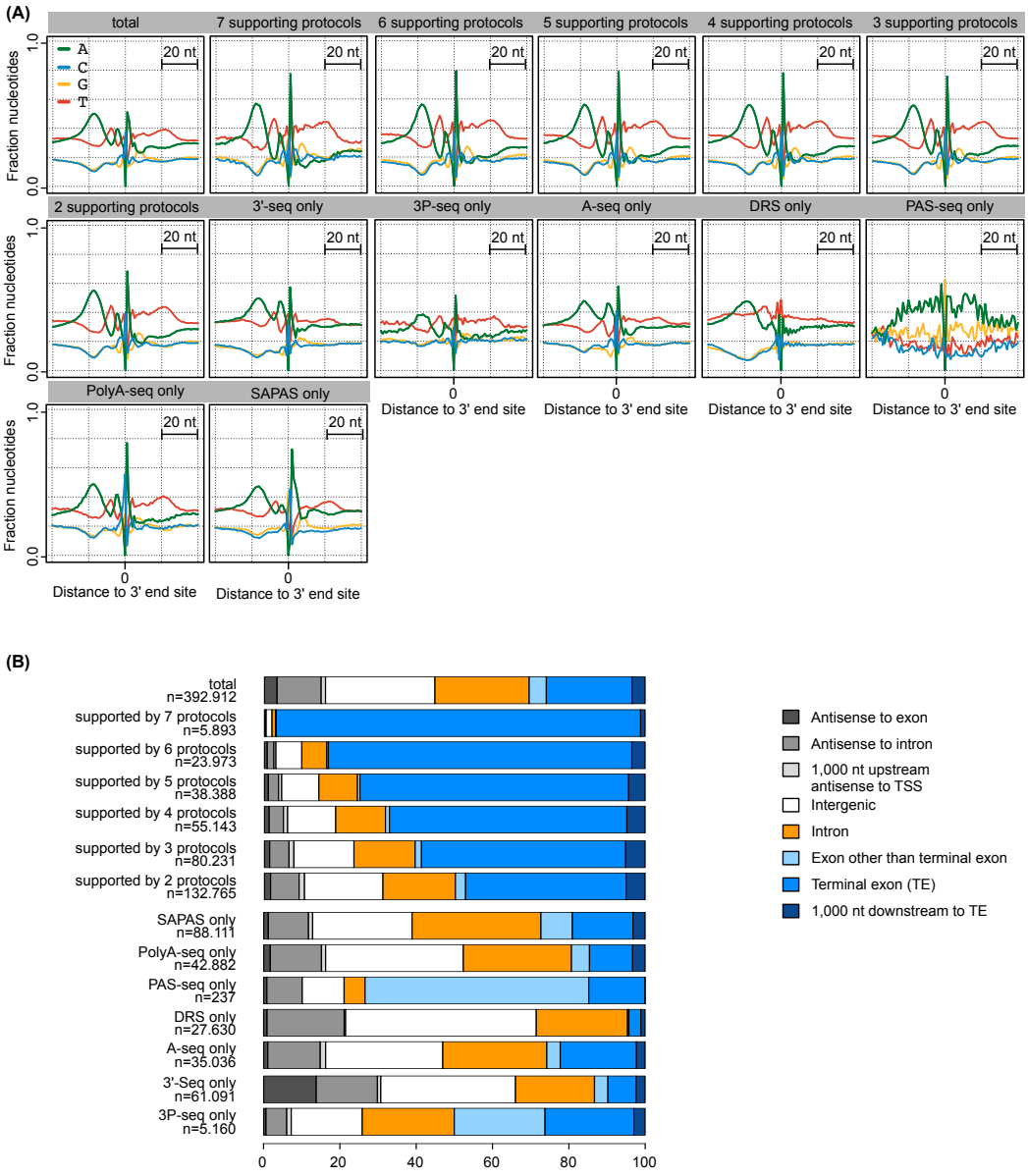


Figure A.5: Characteristics of human poly(A) clusters. (A) Nucleotide composition around cleavage sites supported by the indicated number of protocols or the name of the protocol for clusters that had a single protocol support. **(B)** Annotation of clusters supported by various types of protocols (n - number of poly(A) clusters in the indicated category).

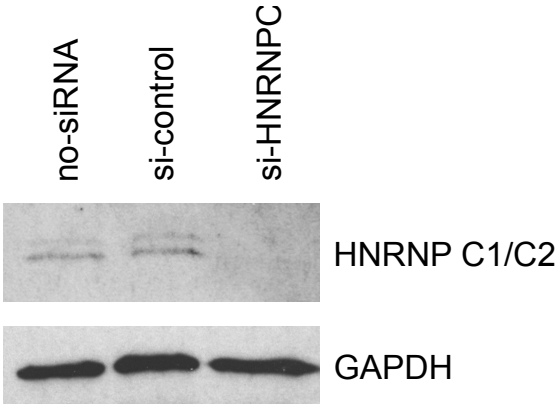


Figure A.6: Expression of HNRNPC. Western blot showing the expression levels of HNRNP C1/C2 and GAPDH in cells that were either untreated, or treated with either a control siRNA or with si-HNRNPC (50 picomoles siRNA per well of a 6-well plate).

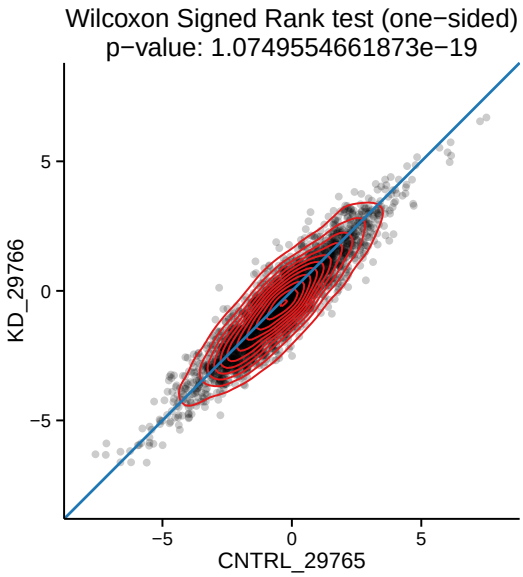


Figure A.7: Scatterplot of the proximal distal ratio in control and knock-down of replicate 1. Contour plot of the proximal-to-distal poly(A) site usage ratios in si-HNRNPC transfected versus si-Control transfected HEK 293 cells in replicate 1. For each plot only exons having exactly two expressed poly(A) sites were considered (2607 exons in total). The proximal-to-distal ratio is significantly higher in cells treated with the control siRNA indicating that on average 3' UTRs tend to be elongated, rather than shortened, upon knockdown of HNRNPC.

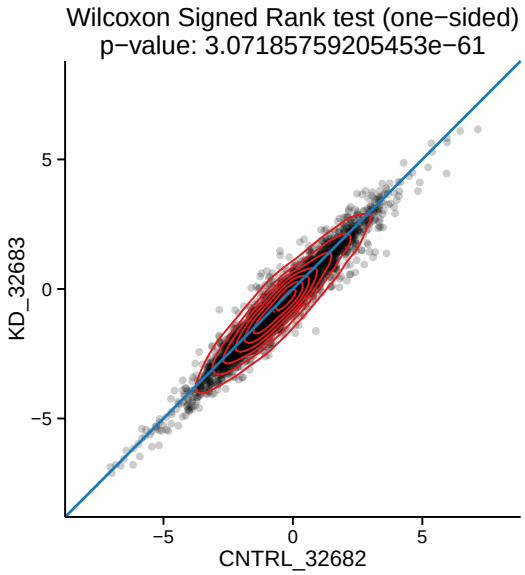


Figure A.8: Scatterplot of the proximal distal ratio in control and knock-down of replicate 2. Contour plot of the proximal-to-distal poly(A) site usage ratios in si-HNRNPC transfected versus si-Control transfected HEK 293 cells in replicate 2. For each plot only exons having exactly two expressed poly(A) sites were considered (2607 exons in total). The proximal-to-distal ratio is significantly higher in cells treated with the control siRNA indicating that on average 3' UTRs tend to be elongated, rather than shortened, upon knockdown of HNRNPC.

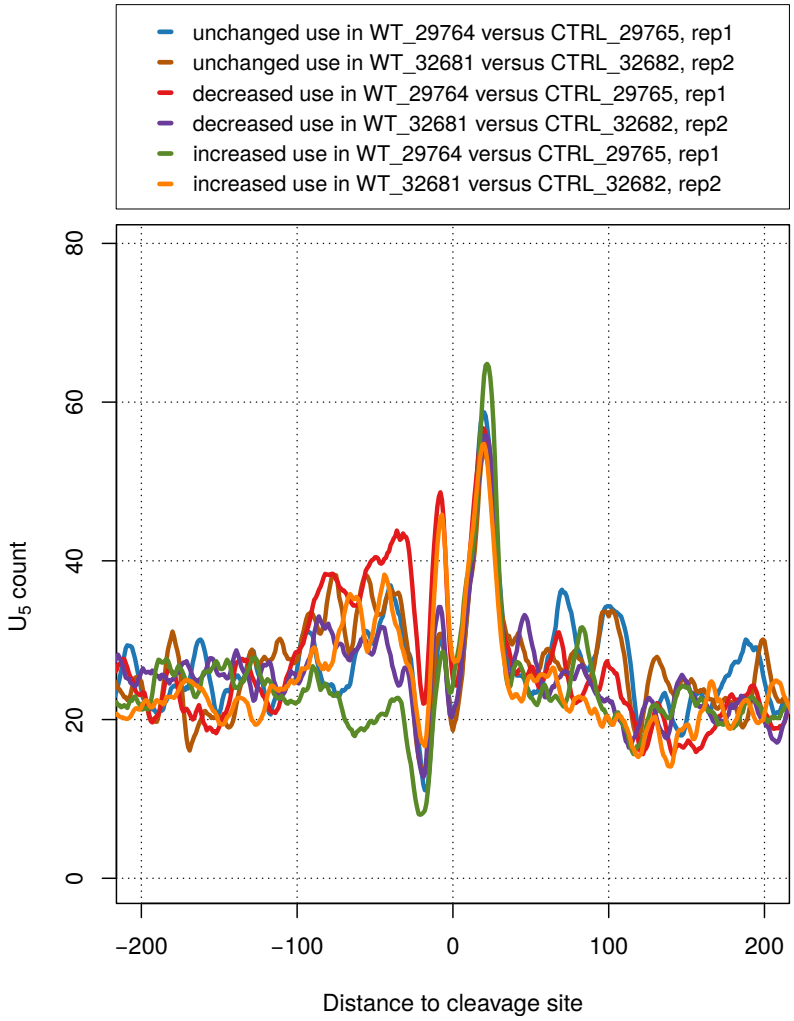


Figure A.9: $(U)_5$ motif count profiles. Smoothened (±5 nt) density of non-overlapping $(U)_5$ tracts in the vicinity of sites with a consistent behavior (increased, unchanged, decreased use) in untransfected (wild type, WT) compared to the si-Control transfected (CTRL) HEK 293 cells.

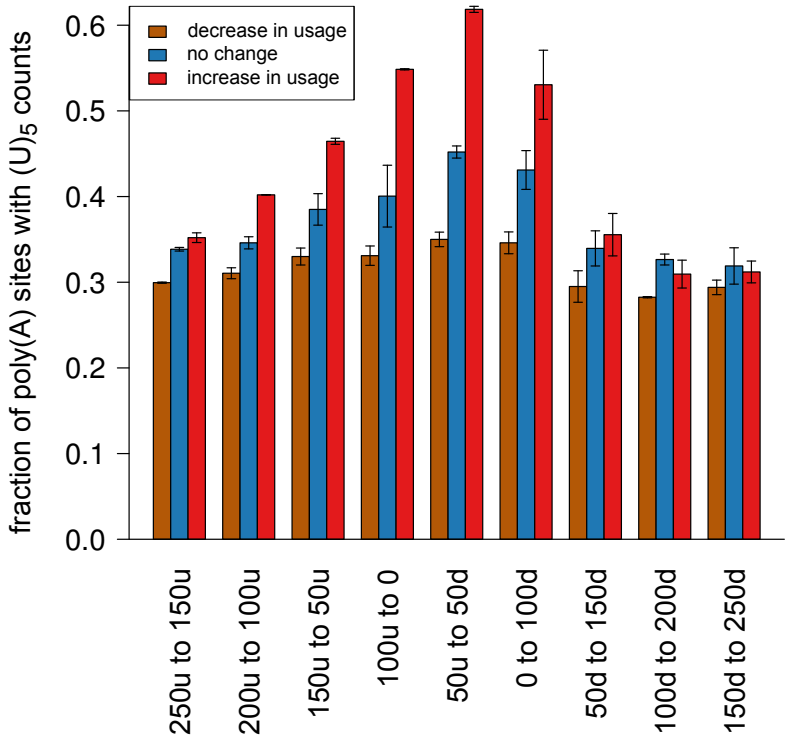


Figure A.10: Abundance of the $(U)_5$ motif around different sets of poly(A) sites. Relationship between the $(U)_5$ content around poly(A) sites and their behavior upon HNRNPC knock-down. 1000 poly(A) sites that increased most, decreased most or changed least (and reproducibly, between the two replicate experiments) in usage upon HNRNPC knock-down were extracted, and the fractions of each of these types of sites that had at least one occurrence of the $(U)_5$ motif at the indicated distance from the poly(A) site were calculated. "u" and "d" indicate upstream and downstream of poly(A) sites and the numbers indicate the boundaries (in nt) of the windows relative to poly(A) sites.

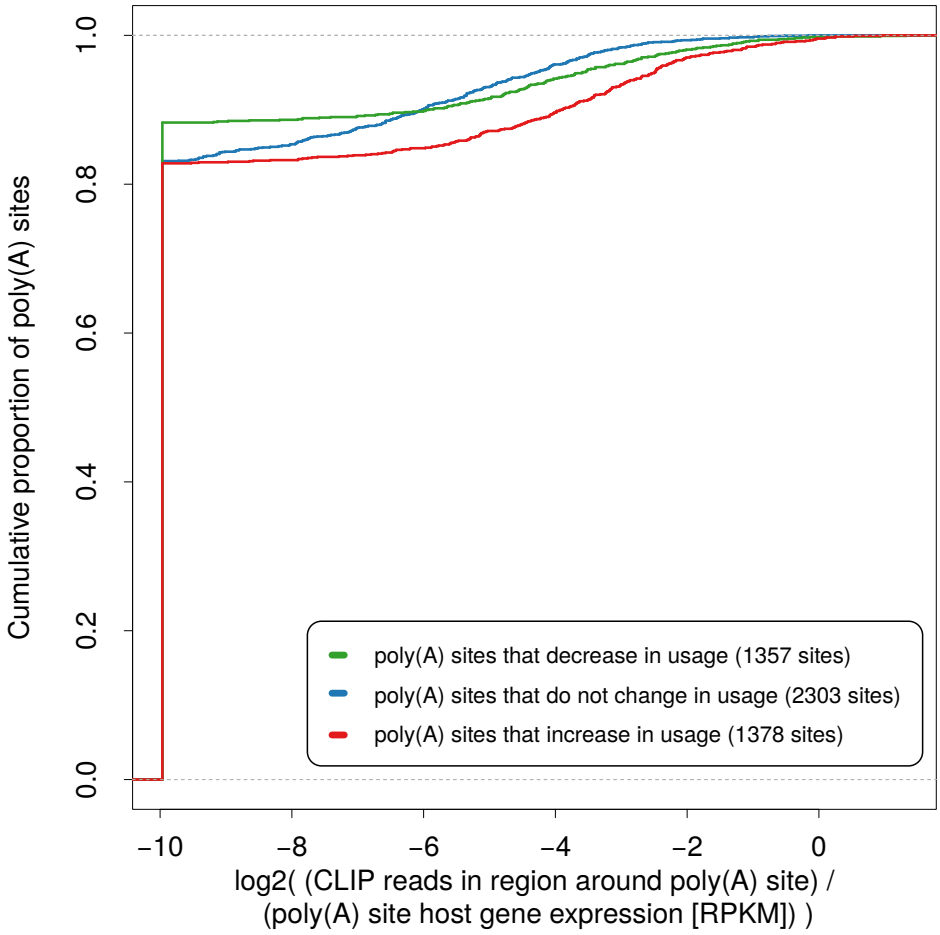


Figure A.11: HNRNPC CLIP reads around different sets of poly(A) sites. Number of HNRNPC CLIP reads that intersect with a region of ±50 nucleotides around poly(A) sites belonging to different categories (consistently decreased/unchanged/increased poly(A) site usage upon HNRNPC knock-down). The number of HNRNPC CLIP reads was normalized by the expression ([RPKM]) of each poly(A) site's host gene. Poly(A) sites that increase in usage have a significantly higher CLIP read support compared to poly(A) sites that do not change in usage upon HNRNPC knock-down (p-value <0.0007, two-sided Kolmogorov-Smirnov test).

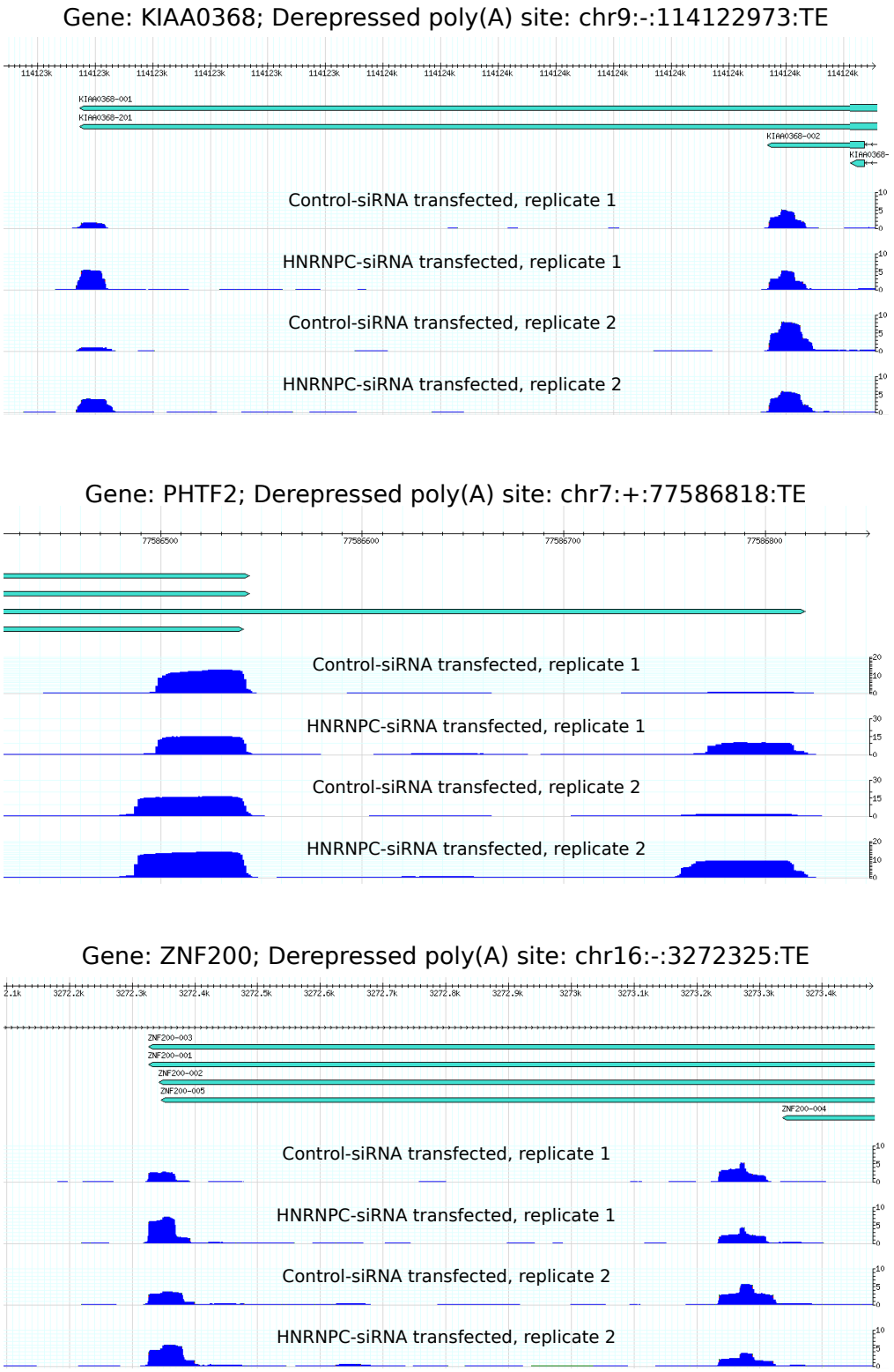


Figure A.12: Examples of genes with 3' UTR lengthening upon HNRNPC knock-down. Browser shots of A-Seq2 read densities within 3' UTRs with *distal poly(A) sites that are derepressed upon knock-down of HNRNPC*. The y-axis shows library size normalized read counts per nucleotide.

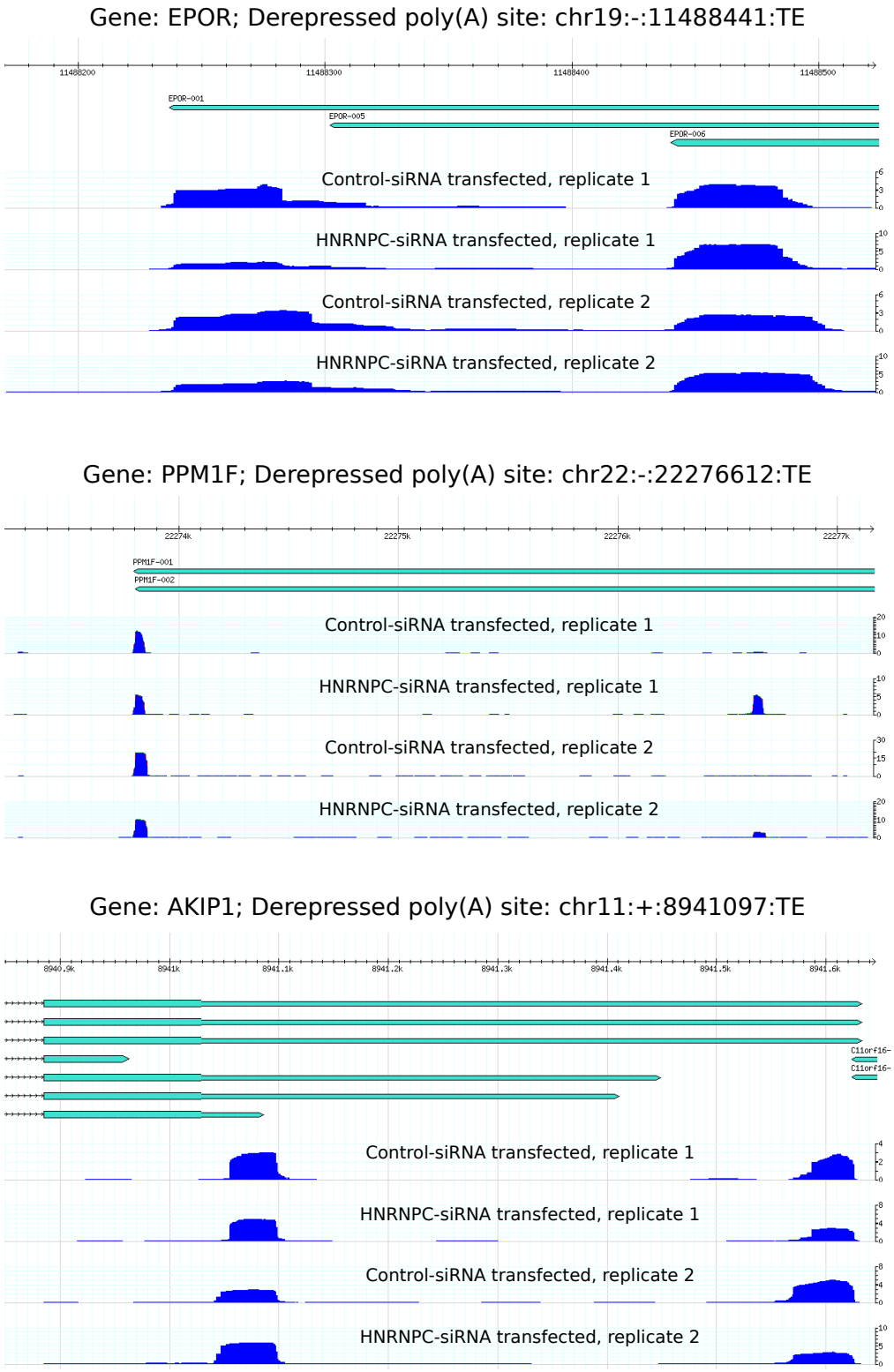
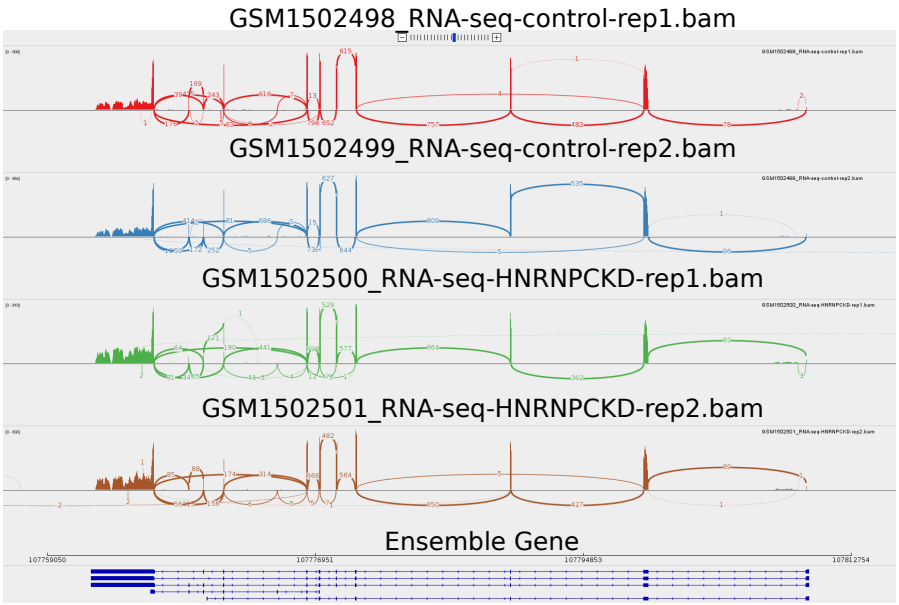


Figure A.13: Examples of genes with 3' UTR shortening upon HNRNPC knock-down. Browser shots of A-Seq2 read densities within 3' UTRs with *proximal poly(A) sites that are derepressed upon knock-down of HNRNPC*. The y-axis shows library size normalized read counts per nucleotide.

(A) Sashimi plots of the CD47 locus as derived from mRNA-Seq data
region: chr3:107756068-107815808 (human genome version hg19)



(B)
Sashimi plots of the CD47 3'UTR locus as derived from mRNA-Seq data
region: chr3:107756992-107766867 (human genome version hg19)

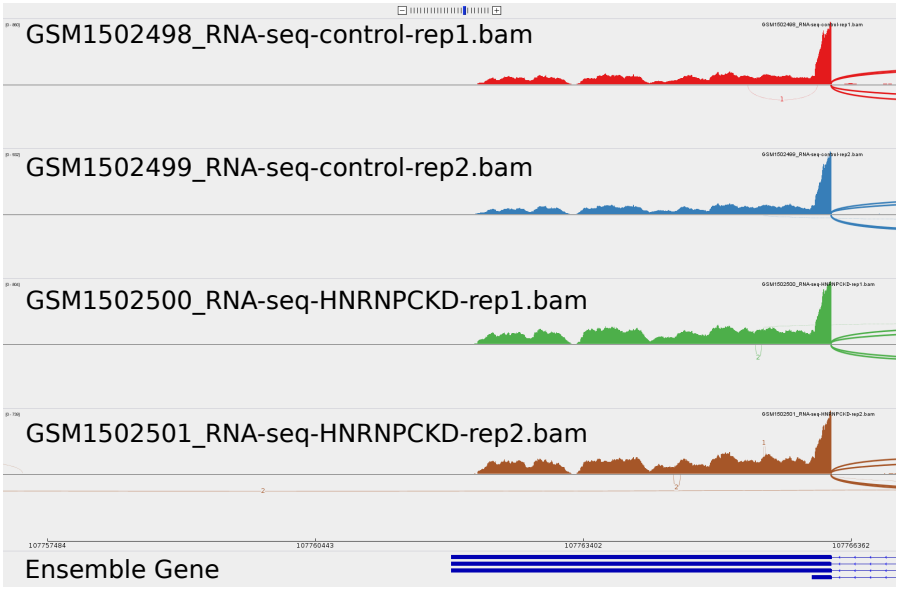


Figure A.14: Read coverage profiles from RNA-seq for the CD47 gene. "Sashimi" plots constructed from previously published (see [147]) mRNA-Seq data (2 replicates of 2 experiments) obtained from HEK 293 cells that have been transfected with si-Control or si-HNRNPC, respectively. After adaptor removal, paired-end reads were mapped applying the STAR aligner with default settings [176]. The mappings were visualized (Sashimi plots) using the Integrative Genomics Viewer (IGV) software [222].

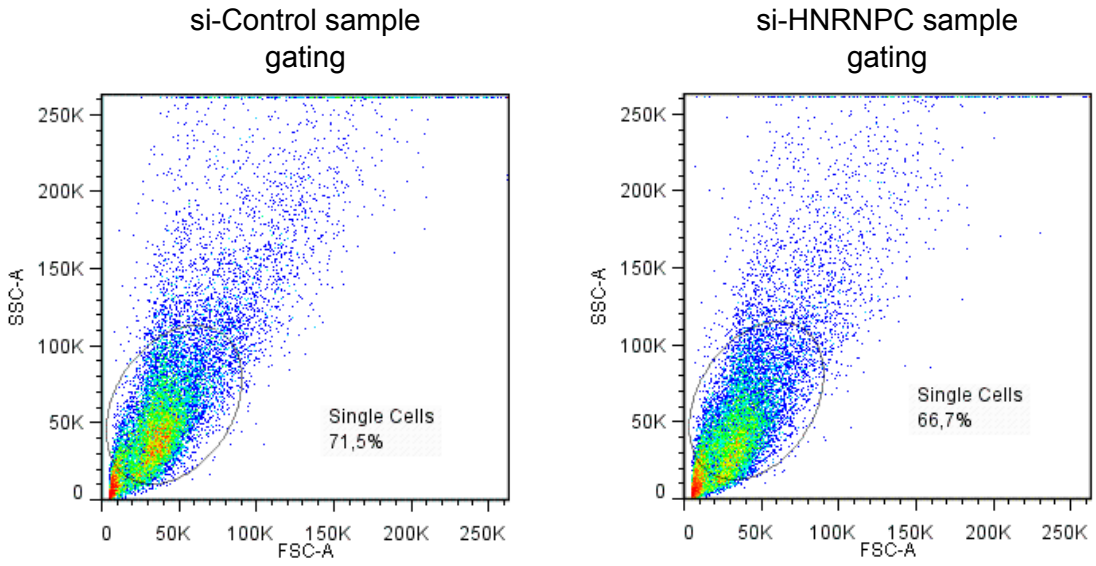


Figure A.15: FACS gating of HEK293 cells. For indirect immunophenotyping of membrane CD47 levels in HEK 293 cells that were either treated with a control siRNA (left panel) or with si-HNRNPC (right panel) a minimum of 10000 gated events was considered. The gate is indicated.

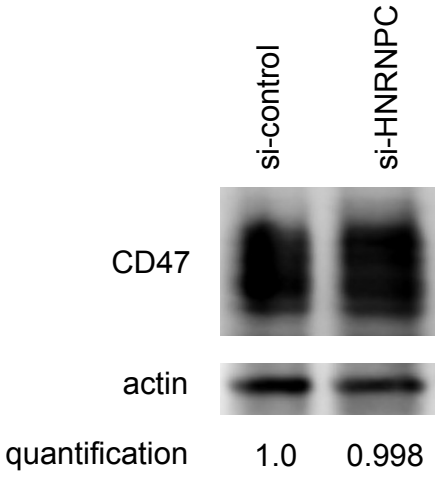


Figure A.16: Western blots of CD47 and Actin proteins. Cells were treated with either a control siRNA or with si-HNRNPC for 72 hrs. Signals were quantified with the ImageJ software and relative CD47 levels are reported with respect to Actin and control siRNA = 1.0.

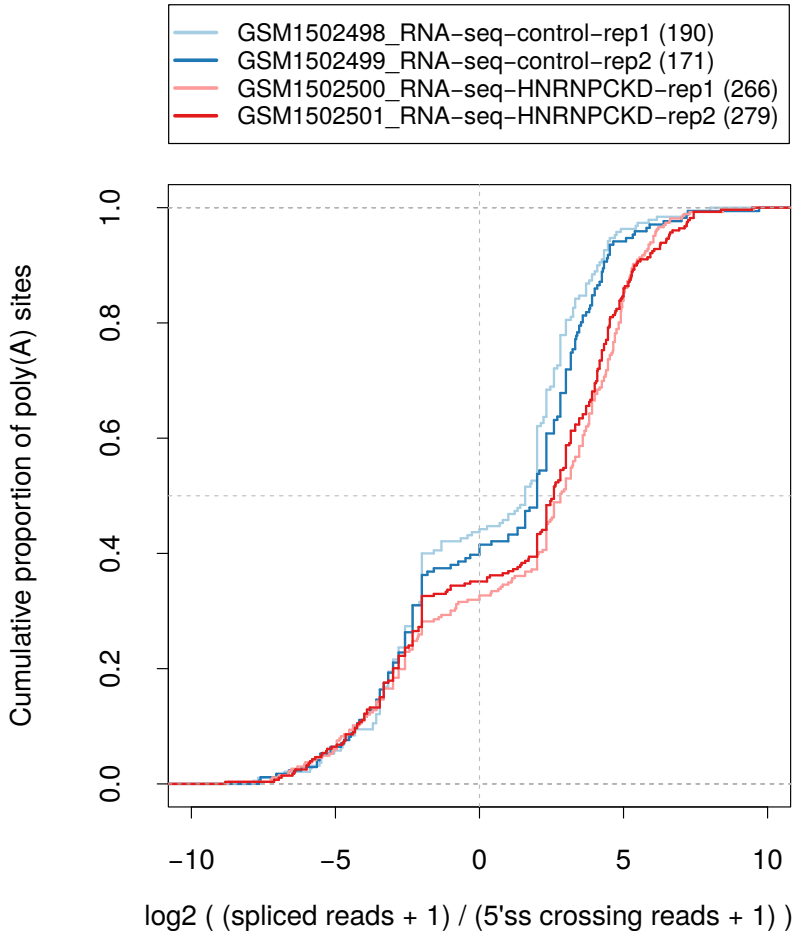


Figure A.17: Distribution of the ratios of splice-reads by non-splice reads at 5' splice sites. Cumulative distribution functions of the log2 ratios of spliced reads to reads that map beyond the 5' splice site (5'ss) of the closest, upstream located exon of each consistenly derepressed, intronic poly(A) site. Intronic poly(A) sites are associated predominantly with the emergence of new exons relative to the extension of internal exons, in both si-Control and si-HNRNPC transfected cells. The HNRNPC knock-down causes a further significant shift towards novel terminal exons created by splicing rather than by internal exon extension (replicate 1 p-value: 4.0e-06, replicate 2 p-value: 8.6e-03, two-sided Mann-Whitney U test). The numbers shown in the legend (written in brackets) indicate the number of intronic poly(A) sites that were used to construct this plot (for more details, see the Methods section).

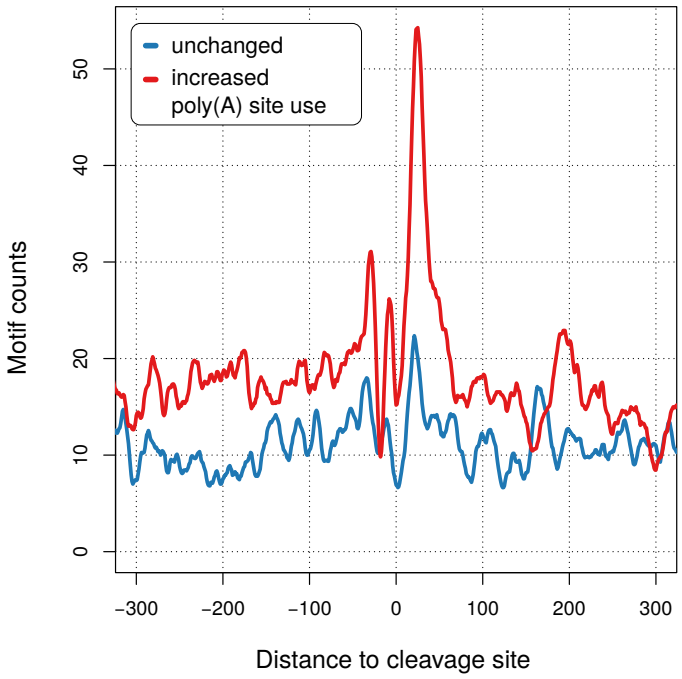


Figure A.18: Abundance of $(U)_5$ motif around different sets of intronic poly(A) sites. Smoothened (±5 nt) density of non-overlapping $(U)_5$ tracts in the vicinity of intronic poly(A) sites with a consistent behavior (increased or unchanged use) in the two HNRNPC knock-down A-seq2 experiments.

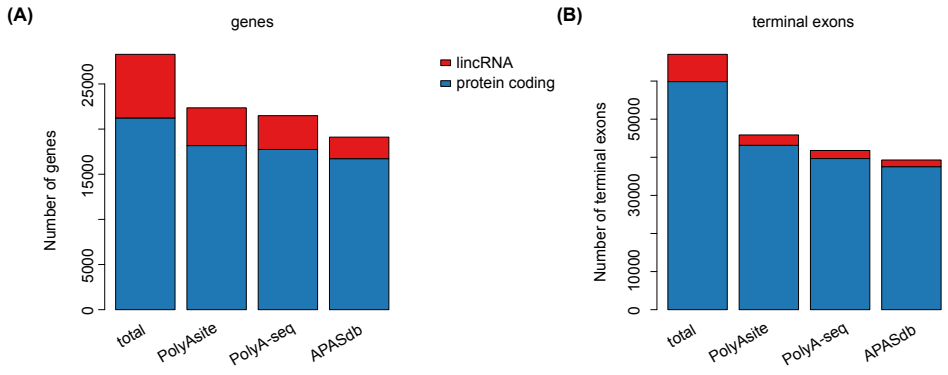


Figure A.19: Number of annotated features (based on the UCSC Basic Table of the GENCODE v19 human (hg19) annotation) that are covered by sites from different atlases. (A) Coverage of genes by sites from PolyAsite (present manuscript), PolyA-seq [7] and APASdb [127]. A gene was considered covered if the genomic position of at least one poly(A) site was within the genomic range of the gene. **(B)** Same as **(A)** but for the terminal exons from the annotation.

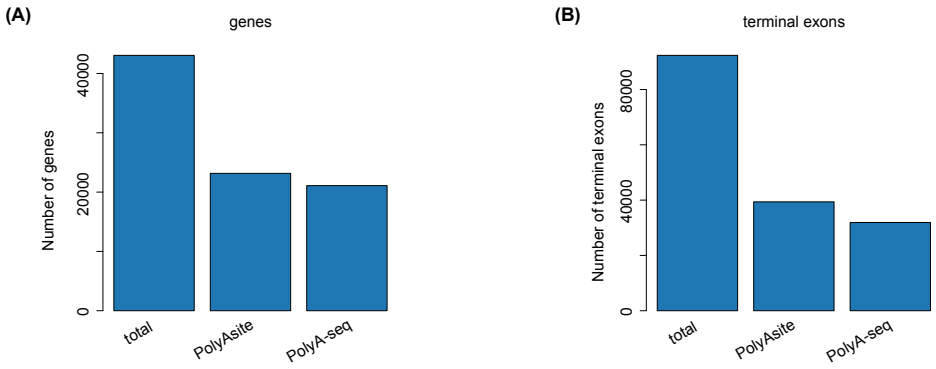


Figure A.20: Number of annotated features (based on the ENSEMBL mouse (mm10) annotation from UCSC) that are covered by sites from different atlases. **(A)** Coverage of genes by sites from PolyAsite and PolyA-seq [7]. A gene was considered to be covered if the genomic position of at least one poly(A) site was within the genomic range of the gene. **(B)** Same as **(A)** but for the terminal exons from the annotation.

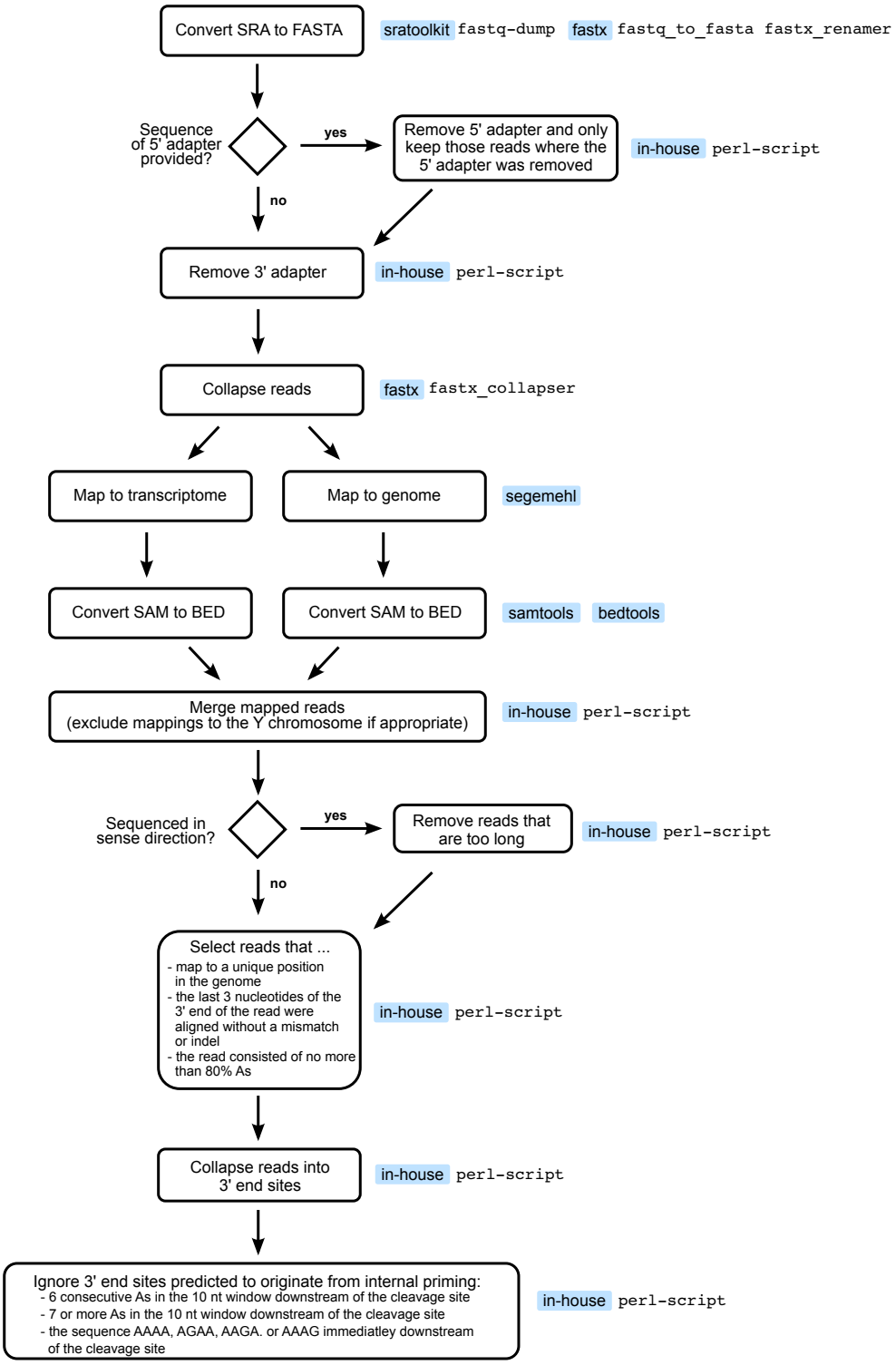


Figure A.21: Outline of the computational pipeline for processing 3' end sequencing data.

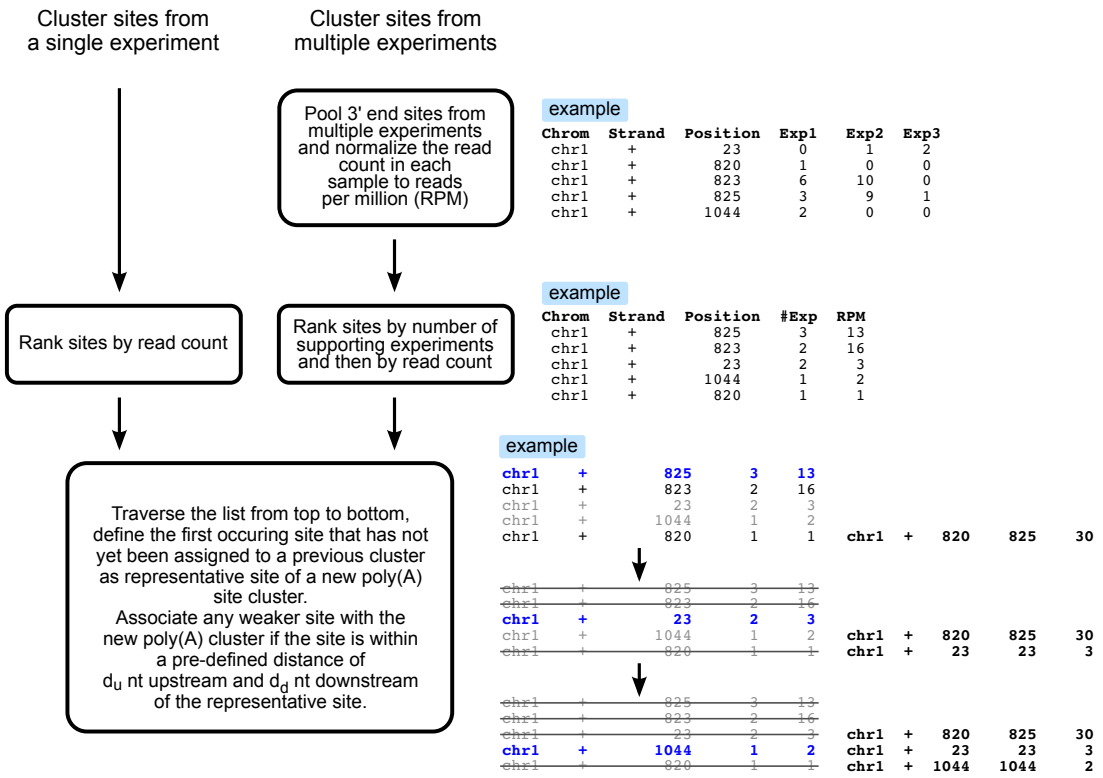


Figure A.22: Outline of the computational pipeline for clustering closely spaced 3' end sites into 3' end processing regions. A toy example data set is used to illustrate the procedure.

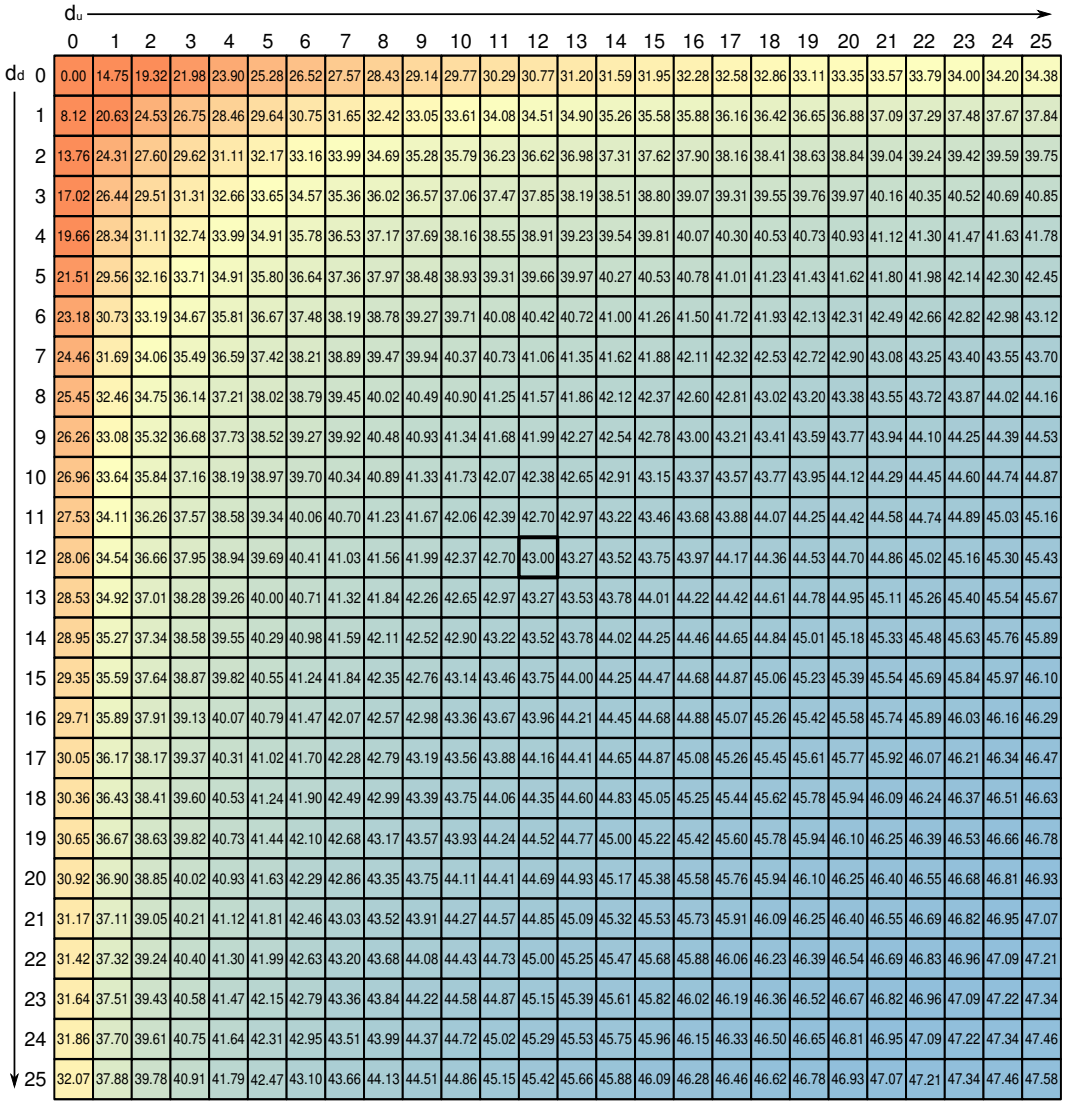


Figure A.23: Analysis results for the inference of distance parameters to associate individual 3' ends into poly(A) site clusters. Evaluation of the distance parameters for clustering closely spaced, putative 3' end processing sites. d_u and d_d refer to the distance upstream and downstream of the representative site, respectively. Values in the plot denote the percentage of 3' end processing sites that were part of a multi-site cluster when a particular set of distance parameters was applied to cluster individual sites. While initially there is a steep increase in the proportion of reads in clusters, a plateau is soon reached. Distances $d_u = 12$ and $d_d = 12$ were chosen in this study.

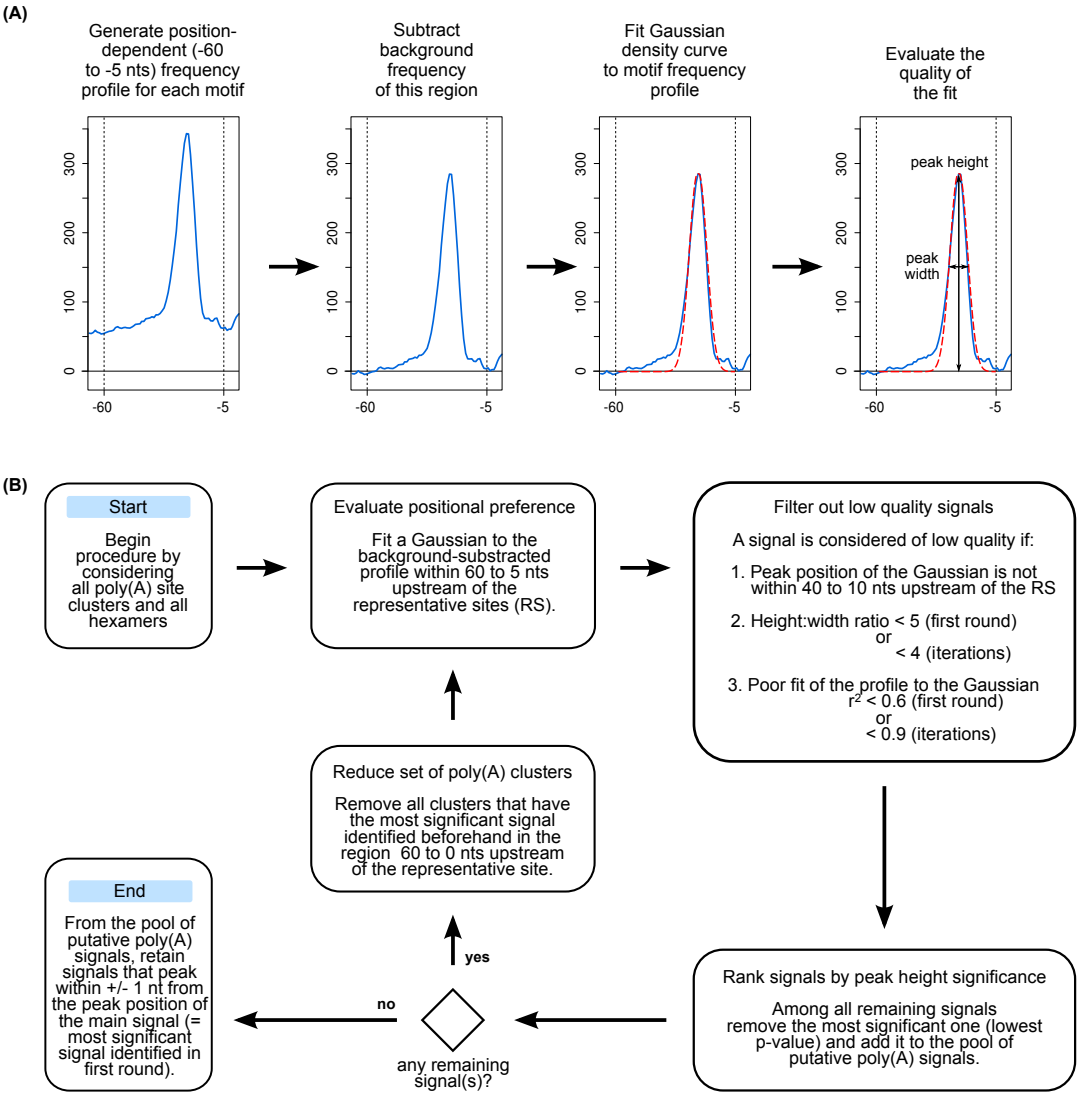


Figure A.24: Outline of the computational procedure that we used to identify poly(A) signals from poly(A) site clusters obtained from high-throughput sequencing of pre-mRNA 3' ends.

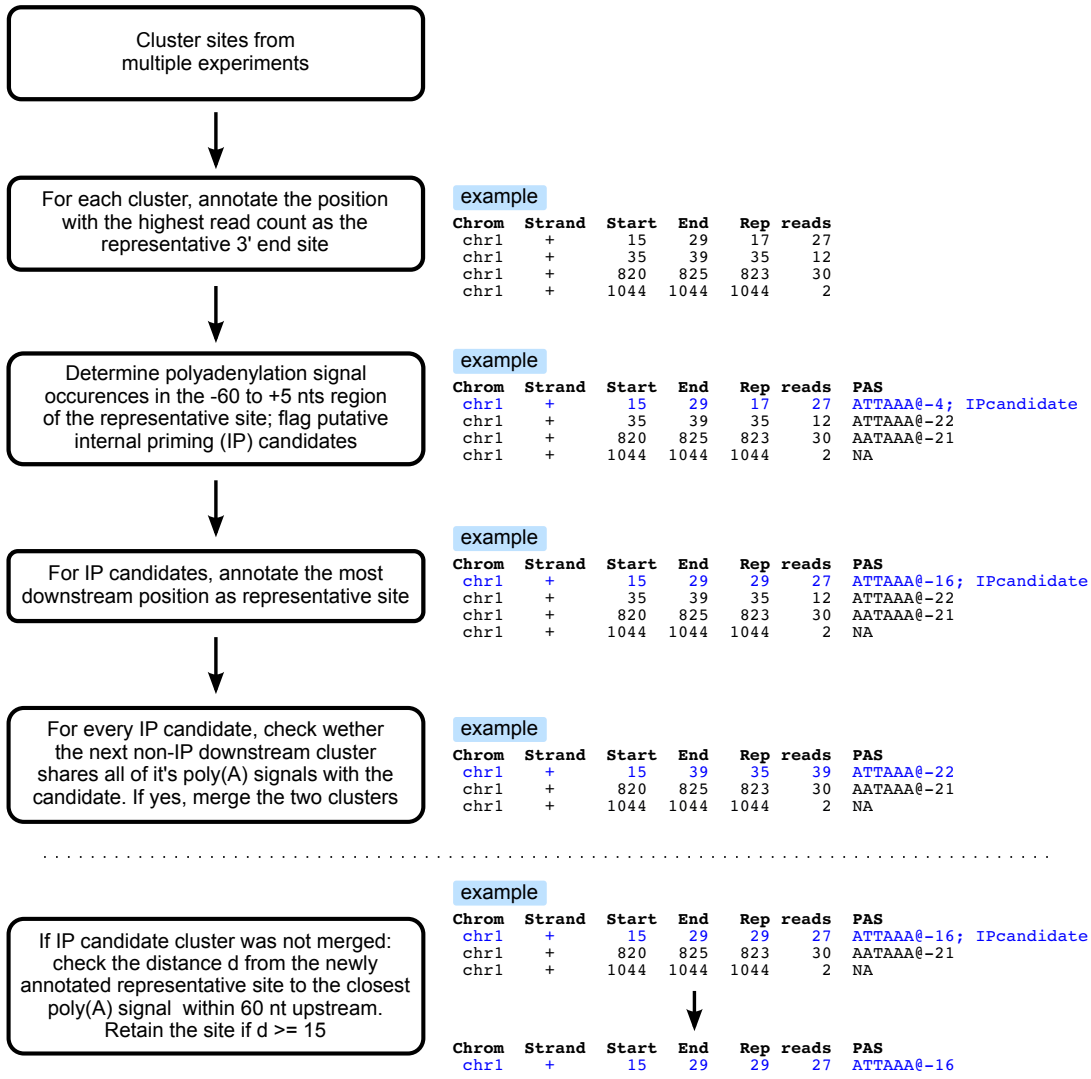


Figure A.25: Outline of the strategy to evaluate poly(A) clusters potentially originating from internal priming.

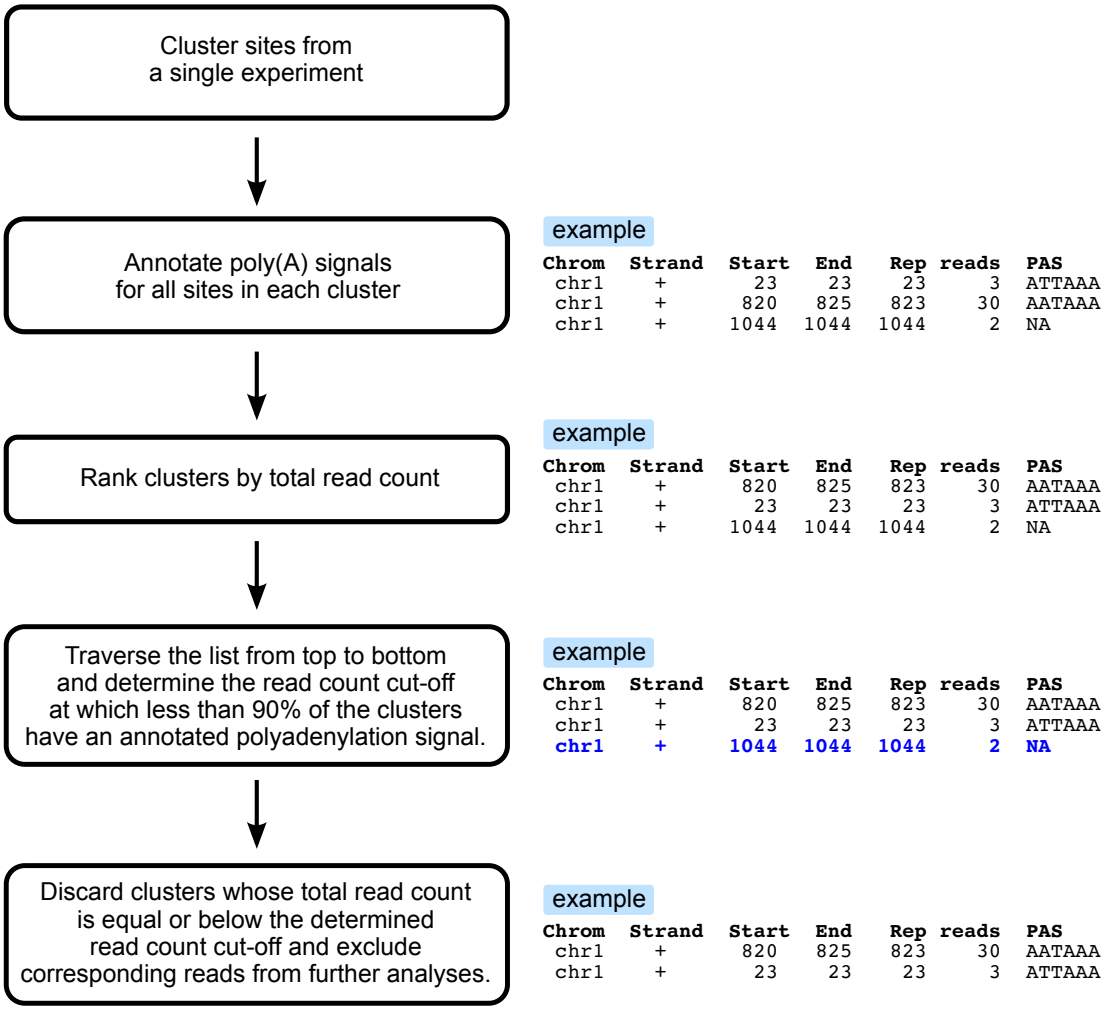


Figure A.26: Outline of the procedure that we used to filter out clusters that do not have sufficient experimental support (sample-specific cut-off of read counts).

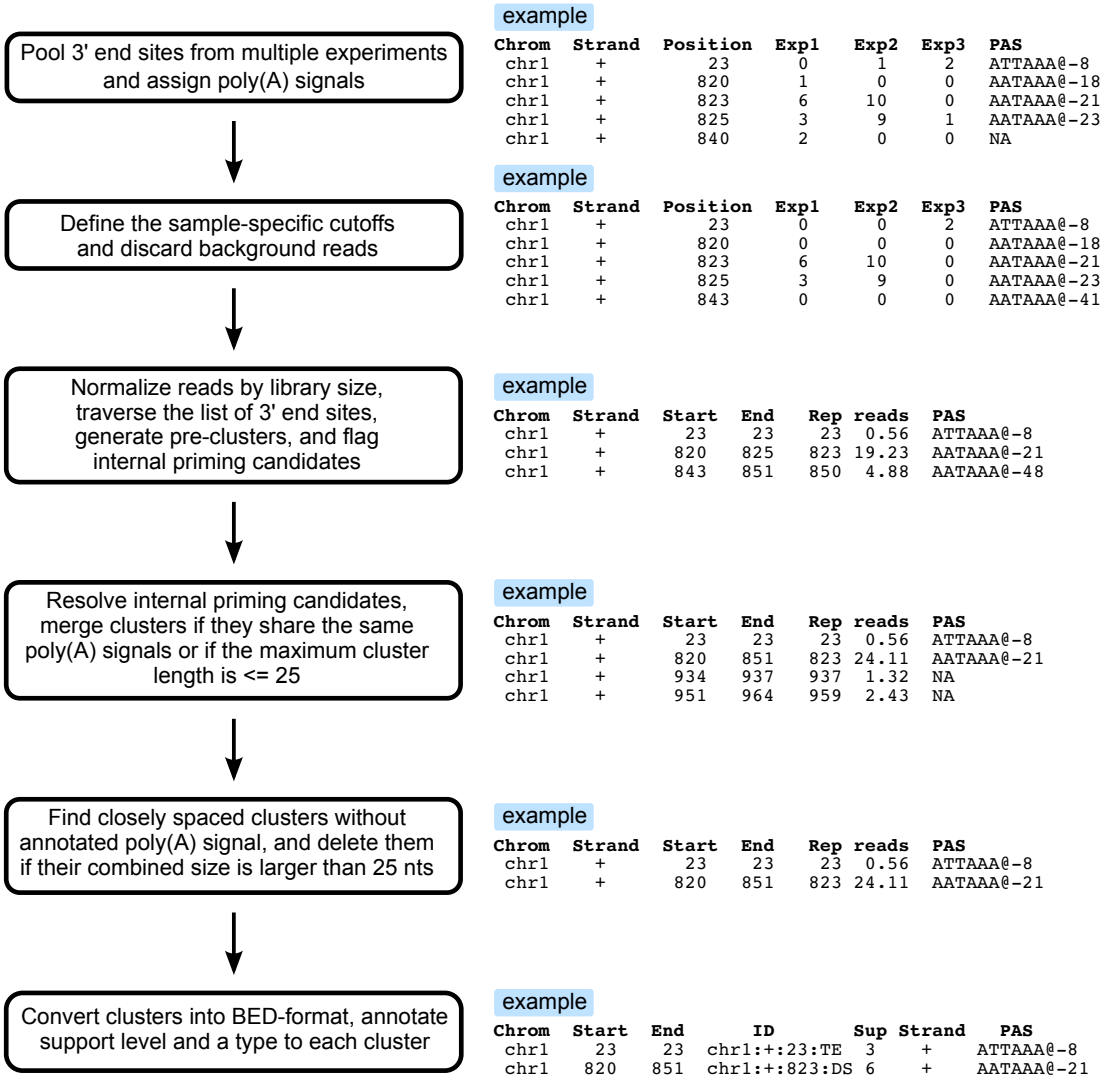


Figure A.27: Outline of the computational procedure that we used to combine 3' end processing sites from multiple experiments into a comprehensive catalog of 3' end processing clusters.

A.3 Supplementary Tables

Table A.1: Number of poly(A) sites in different studies. Comparison of poly(A) sites that were reported by Derti et al. [7] and You et al. [127] for different human tissues. Both of these studies reported only one genomic position per poly(A) site cluster. To be more permissive in evaluating the overlap of these data sets, we first extended the poly(A) sites from these data sets by 25 nt up- and downstream. A poly(A) site from one study was considered to overlap if there was at least one cluster in the other data set such that both clusters overlapped each other by at least one nucleotide. For each tissue we report both the number of poly(A) site clusters that overlapped as well as those that were unique to a specific data set. In parentheses, the average number of reported reads for the underlying poly(A) sites of the corresponding set of clusters is indicated.

	PolyA-seq clusters overlapping with APASdb clusters	**APASdb clusters overlapping with PolyA-seq clusters**	**PolyA-seq unique clusters**	**APASdb unique clusters**
brain	31,356 (58.47)	30,856 (90.04)	57,754 (19.25)	23,827 (10.83)
kidney	23,793 (104.27)	23,090 (121.53)	71,152 (29.39)	12,006 (19.78)
liver	25,923 (175.45)	25,152 (116.98)	62,317 (16.23)	10,741 (7.26)
muscle	21,910 (151.16)	21,227 (123.36)	90,888 (17.03)	10,743 (37.56)
testes	34,810 (117.72)	34,057 (66.84)	80,258 (11.61)	34,860 (18.47)

Table A.2: Overview of the samples used to build the genome-wide catalog of 3' end processing site in human.

series ID	**sample ID**	**protocol**	**tissue/cell type**	**gender**	**publication**
GSE40859	GSM1003590	"DRS"	"HeLa"	F	[49]
GSE40859	GSM1003591	"DRS"	"HeLa"	F	[49]
GSE40859	GSM1003592	"DRS"	"HeLa"	F	[49]
SRP025988	SRX388391	"DRS"	"HeLa"	F	[143]
SRP022151	SRX275752	"DRS"	"K562"	F	[85]
SRP022151	SRX275753	"DRS"	"K562"	F	[85]
SRP022151	SRX275806	"DRS"	"K562"	F	[85]
SRP022151	SRX275827	"DRS"	"K562"	F	[85]
SRP003483	SRX026582	"SAPAS"	"MDA-MB-231"	F	[100]
SRP003483	SRX026583	"SAPAS"	"MCF-10A"	F	[100]
SRP003483	SRX026584	"SAPAS"	"MCF-7"	F	[100]
GSE25450	GSM624686	"PAS-Seq"	"HeLa"	F	[6]
GSE30198	GSM747470	"PolyA-seq"	"Brain"	NA	[7]
GSE30198	GSM747471	"PolyA-seq"	"Kidney"	NA	[7]
GSE30198	GSM747472	"PolyA-seq"	"Liver"	NA	[7]
GSE30198	GSM747473	"PolyA-seq"	"MAQC Brain"	NA	[7]
GSE30198	GSM747474	"PolyA-seq"	"MAQC Brain"	NA	[7]
GSE30198	GSM747475	"PolyA-seq"	"MAQC UHR"	NA	[7]
GSE30198	GSM747476	"PolyA-seq"	"MAQC UHR"	NA	[7]

Continued on next page

Table A.2 – continued from previous page

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE30198	GSM747477	"PolyA-seq"	"Muscle"	NA	[7]
GSE30198	GSM747479	"PolyA-seq"	"Testis"	NA	[7]
GSE30198	GSM747480	"PolyA-seq"	"UHR"	NA	[7]
GSE37037	GSM909242	"A-seq"	"HEK293"	F	[33]
GSE37037	GSM909243	"A-seq"	"HEK293"	F	[33]
GSE37037	GSM909244	"A-seq"	"HEK293"	F	[33]
GSE37037	GSM909245	"A-seq"	"HEK293"	F	[33]
GSE40137	GSM986133	"A-seq"	"HEK293"	F	[120]
GSE40137	GSM986134	"A-seq"	"HEK293"	F	[120]
GSE40137	GSM986135	"A-seq"	"HEK293"	F	[120]
GSE40137	GSM986136	"A-seq"	"HEK293"	F	[120]
GSE40137	GSM986137	"A-seq"	"HEK293"	F	[120]
GSE40137	GSM986138	"A-seq"	"HEK293"	F	[120]
SRP029953	SRX351949	"3'-Seq"	"native B cells"	NA	[19]
SRP029953	SRX351950	"3'-Seq"	"native B cells"	NA	[19]
SRP029953	SRX351952	"3'-Seq"	"brain"	NA	[19]
SRP029953	SRX351953	"3'-Seq"	"breast"	F	[19]
SRP029953	SRX359328	"3'-Seq"	"embryonic stem cells (H9)"	F	[19]
SRP029953	SRX359329	"3'-Seq"	"ovary"	F	[19]
SRP029953	SRX359330	"3'-Seq"	"skeletal muscle"	NA	[19]
SRP029953	SRX359331	"3'-Seq"	"testis"	NA	[19]
SRP029953	SRX359332	"3'-Seq"	"MCF10A"	F	[19]
SRP029953	SRX359333	"3'-Seq"	"MCF10A"	F	[19]
SRP029953	SRX359334	"3'-Seq"	"MCF7"	F	[19]
SRP029953	SRX359335	"3'-Seq"	"HeLa"	F	[19]
SRP029953	SRX359336	"3'-Seq"	"HEK293"	F	[19]
SRP029953	SRX359337	"3'-Seq"	"NTERA2"	M	[19]
SRP029953	SRX359339	"3'-Seq"	"B-LCL cells"	NA	[19]
SRP029953	SRX359340	"3'-Seq"	"MCF10A"	F	[19]
SRP029953	SRX359341	"3'-Seq"	"MCF10A"	F	[19]
GSE52527	GSM1268942	"3P-Seq"	"HeLa"	F	[58]
GSE52527	GSM1268943	"3P-Seq"	"HEK293"	F	[58]
GSE52527	GSM1268944	"3P-Seq"	"Huh7"	NA	[58]
GSE52527	GSM1268945	"3P-Seq"	"IMR90"	F	[58]
GSE56657	GSM1366428	"DRS"	"neuroendocrine tumor"	F	[221]
GSE56657	GSM1366429	"DRS"	"neuroendocrine tumor"	M	[221]
GSE56657	GSM1366430	"DRS"	"Pituitary"	M	[221]
SRP041182	SRX517334	"SAPAS"	"testis"	M	[127]
SRP041182	SRX517333	"SAPAS"	"ovary"	F	[127]
SRP041182	SRX517332	"SAPAS"	"skeletal muscle"	M	[127]
SRP041182	SRX517331	"SAPAS"	"adipose"	M	[127]
SRP041182	SRX517330	"SAPAS"	"thymus"	M	[127]
SRP041182	SRX517329	"SAPAS"	"small intestine"	M	[127]
SRP041182	SRX517328	"SAPAS"	"pancreas"	F	[127]
SRP041182	SRX517327	"SAPAS"	"liver"	M	[127]
SRP041182	SRX517326	"SAPAS"	"prostate"	M	[127]
SRP041182	SRX517325	"SAPAS"	"breast"	F	[127]
SRP041182	SRX517324	"SAPAS"	"bladder"	F	[127]
SRP041182	SRX517323	"SAPAS"	"uterus"	F	[127]
SRP041182	SRX517322	"SAPAS"	"lung"	M	[127]
SRP041182	SRX517321	"SAPAS"	"placenta"	F	[127]

Continued on next page

Table A.2 – continued from previous page

series ID	sample ID	protocol	tissue/cell type	gender	publication
SRP041182	SRX517320	"SAPAS"	"lymph node"	M	[127]
SRP041182	SRX517319	"SAPAS"	"heart"	M	[127]
SRP041182	SRX517318	"SAPAS"	"cervix"	F	[127]
SRP041182	SRX517317	"SAPAS"	"kidney"	M	[127]
SRP041182	SRX517316	"SAPAS"	"stomach"	M	[127]
SRP041182	SRX517315	"SAPAS"	"spleen"	M	[127]
SRP041182	SRX517314	"SAPAS"	"thyroid"	F	[127]
SRP041182	SRX517313	"SAPAS"	"brain"	F	[127]

Table A.3: Overview of the samples used to build the genome-wide catalog of 3' end processing site in mouse.

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE30198	GSM747481	"PolyA-seq"	"Brain"	NA	[7]
GSE30198	GSM747482	"PolyA-seq"	"Kidney"	NA	[7]
GSE30198	GSM747483	"PolyA-seq"	"Liver"	NA	[7]
GSE30198	GSM747484	"PolyA-seq"	"Muscle"	NA	[7]
GSE30198	GSM747485	"PolyA-seq"	"Testis"	NA	[7]
GSE54950	GSM1327166	"A-seq V2"	"T cells"	NA	[56]
GSE54950	GSM1327167	"A-seq V2"	"T cells"	NA	[56]
GSE54950	GSM1327168	"A-seq V2"	"T cells"	NA	[56]
GSE54950	GSM1327169	"A-seq V2"	"T cells"	NA	[56]
GSE46433	GSM1130096	"2P-Seq"	"embryonic stem cells"	NA	[130]
GSE46433	GSM1130097	"2P-Seq"	"embryonic stem cells"	NA	[130]
GSE46433	GSM1130098	"2P-Seq"	"embryonic stem cells"	NA	[130]
GSE46433	GSM1130099	"2P-Seq"	"embryonic stem cells"	NA	[130]
GSE46433	GSM1130100	"2P-Seq"	"embryonic stem cells"	NA	[130]
GSE46433	GSM1130101	"2P-Seq"	"embryonic stem cells"	NA	[130]
SRP025988	SRX304982	"DRS"	"embryonic stem cell line E14Tg2a"	M	[143]
SRP025988	SRX304983	"DRS"	"embryonic stem cell line E14Tg2a"	M	[143]
GSE44698	GSM1089085	"2P-Seq"	"3T3"	NA	[55]
GSE44698	GSM1089086	"2P-Seq"	"3T3"	NA	[55]
GSE44698	GSM1089087	"2P-Seq"	"3T3"	NA	[55]
GSE44698	GSM1089088	"2P-Seq"	"3T3"	NA	[55]
GSE44698	GSM1089089	"2P-Seq"	"3T3"	NA	[55]
GSE44698	GSM1089090	"2P-Seq"	"3T3"	NA	[55]
GSE44698	GSM1089091	"2P-Seq"	"3T3"	NA	[55]
GSE44698	GSM1089092	"2P-Seq"	"3T3"	NA	[55]
GSE44698	GSM1089093	"2P-Seq"	"3T3"	NA	[55]
GSE44698	GSM1089094	"2P-Seq"	"3T3"	NA	[55]
GSE44698	GSM1089095	"2P-Seq"	"3T3"	NA	[55]
GSE44698	GSM1089096	"2P-Seq"	"3T3"	NA	[55]
GSE52528	GSM1268946	"3P-seq"	"heart"	NA	[58]
GSE52528	GSM1268947	"3P-seq"	"muscle"	NA	[58]
GSE52528	GSM1268948	"3P-seq"	"liver"	NA	[58]
GSE52528	GSM1268949	"3P-seq"	"lung"	NA	[58]
GSE52528	GSM1268950	"3P-seq"	"wat"	NA	[58]
GSE52528	GSM1268951	"3P-seq"	"kidney"	NA	[58]
GSE52528	GSM1268952	"3P-seq"	"heart"	NA	[58]
GSE52528	GSM1268953	"3P-seq"	"muscle"	NA	[58]
GSE52528	GSM1268954	"3P-seq"	"liver"	NA	[58]

Continued on next page

Table A.3 – continued from previous page

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE52528	GSM1268955	"3P-seq"	"lung"	NA	[58]
GSE52528	GSM1268956	"3P-seq"	"wat"	NA	[58]
GSE52528	GSM1268957	"3P-seq"	"kidney"	NA	[58]
GSE52528	GSM1268958	"3P-seq"	"embryonic stem cells"	NA	[58]
GSE25450	GSM624687	"PAS-Seq"	"ES"	NA	[6]
GSE60487	GSM1480973	"PolyA-seq V2"	"MEF"	NA	[84]
GSE60487	GSM1480974	"PolyA-seq V2"	"MEF"	NA	[84]
GSE60487	GSM1480975	"PolyA-seq V2"	"MEF"	NA	[84]
GSE60487	GSM1480976	"PolyA-seq V2"	"MEF"	NA	[84]
GSE60487	GSM1480977	"PolyA-seq V2"	"MEF"	NA	[84]
GSE60487	GSM1480978	"PolyA-seq V2"	"MEF"	NA	[84]
GSE60487	GSM1480979	"PolyA-seq V2"	"MEF"	NA	[84]
GSE60487	GSM1480980	"PolyA-seq V2"	"MEF"	NA	[84]
GSE62001	GSM1518105	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518106	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518107	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518108	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518109	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518110	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518111	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518112	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518113	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518082	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518089	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518090	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518102	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518103	"3READS"	"NA"	NA	[79]
GSE62001	GSM1586365	"3READS"	"NA"	NA	[79]
GSE62001	GSM1586366	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518096	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518097	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518098	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518072	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518073	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518074	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518075	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518076	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518077	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518078	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518079	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518080	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518081	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518083	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518084	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518085	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518086	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518087	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518088	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518091	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518092	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518093	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518094	"3READS"	"NA"	NA	[79]

Continued on next page

Table A.3 – continued from previous page

series ID	sample ID	protocol	tissue/cell type	gender	publication
GSE62001	GSM1518095	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518099	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518101	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518104	"3READS"	"NA"	NA	[79]
GSE62001	GSM1586367	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518071	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518114	"3READS"	"NA"	NA	[79]
GSE62001	GSM1586368	"3READS"	"NA"	NA	[79]
GSE62001	GSM1518100	"3READS"	"NA"	NA	[79]
GSE62001	GSM1586363	"3READS"	"NA"	NA	[79]
GSE62001	GSM1586364	"3READS"	"NA"	NA	[79]
SRP039327	SRX480169	"SAPAS"	"thymus"	NA	[127]
SRP039327	SRX480179	"SAPAS"	"thymus"	NA	[127]
SRP039327	SRX480205	"SAPAS"	"thymus"	NA	[127]
SRP039327	SRX480212	"SAPAS"	"thymus"	NA	[127]
SRP039327	SRX480221	"SAPAS"	"thymus"	NA	[127]
SRP039327	SRX480227	"SAPAS"	"thymus"	NA	[127]
SRP039327	SRX480229	"SAPAS"	"thymus"	NA	[127]
SRP039327	SRX480250	"SAPAS"	"thymus"	NA	[127]
SRP039327	SRX480287	"SAPAS"	"thymus"	NA	[127]

Table A.4: Hexamer enrichment upstream of human poly(A) sites. The 100 most significantly enriched hexamers (binomial test relative to what is expected given the mononucleotide composition of the region from -60 to 0 nt relative to poly(A) site) in the human poly(A) site catalog.

hexamer	-log p-value
AATAAA	122788.1
AAATAA	42670.49
AAAAAA	33960.3
ATAAAA	33379.19
TAAAAA	24249.76
AAAATA	21755.03
AAAAAT	19162.31
TTAAAA	16451.96
ATAAAT	14493.43
AAAAAG	14079.72
TTTTTT	13455.43
ATTAAA	12302.28
TAAAAT	11913.92
GCCTGG	11751.91
ATAAAG	11628.45
CCTGGG	11165.77
TTTTCT	10964.83
TGTTTT	10879.94
CCAGCC	10729.18
AAAATG	9002.596
CAGCCT	8279.236
CTTTTT	8043.175

Continued on next page

Table A.4 – continued from previous page

hexamer	-log p-value
AGAAAA	7959.7
TTTCTT	7707.476
CTGGGC	7594.283
AAAGAA	7535.008
AAGAAA	7519.484
AAATGT	7297.44
GAAAAA	7156.527
AGCCTG	7106.297
TTTAAA	7019.924
TTTTTC	6929.253
TTTTGT	6754.398
CCTCCC	6622.351
TTGTTT	6515.799
TTCTTT	6484.465
TTTTAA	6444.964
TTTCTG	6351.61
CAATAA	6137.289
TAAATG	5913.602
TTTTTG	5750.779
AAAAAC	5741.94
TAAATA	5719.061
TCTTTT	5691.07
ATTTTT	5690.314
CTCCAG	5609.213
CAAAAA	5564.294
TTTGTT	5252.513
TTTTTA	5163.368
CTGTCT	5128.945
TGTGTG	5124.415
AAAACA	5094.2
CCCAGC	5042.282
TTCTGT	5016.795
CTCTGT	4984.282
ATAAAC	4984.15
CTCCCC	4866.824
TATTTT	4738.292
AAAAGA	4679.872
TTTCCT	4662.104
CTGCTG	4550.984
TTTTCC	4286.656
CCTGGC	4259.37
CCTGCC	4236.644
CTGCCT	4207.258
CTGTTT	4086.569

Continued on next page

Table A.4 – continued from previous page

hexamer	-log p-value
CCCTCC	4082.152
GGAAAA	4078.892
ACAGAG	4074.031
CTGTGT	4001.796
TCTGTG	3969.594
GTTTTT	3911.444
CCCAGG	3869.135
TGTCTC	3865.269
GCCTCC	3851.923
TGCTTT	3843.789
TGCCTG	3713.514
CTTCCC	3708.302
CCCCAG	3686.223
TAATAA	3629.887
TTTCTC	3577.619
TGGAAA	3574.17
TAAAAG	3557.743
TGCTGT	3532.84
TTTATT	3526.132
CCCCCA	3524.531
TCCAGC	3520.258
GAATAA	3458.727
GCTGTG	3405.909
TCTCTG	3392.311
CCACTG	3378.823
CCTCTG	3304.089
TTTCCC	3297.584
GGGAGG	3271.045
CATTTT	3270.061
TTCCTG	3266.088
CTGCCC	3236.691
CTTTCT	3230.07
CAGAGC	3226.857
CTGTGG	3207.589

Table A.5: Hexamer enrichment upstream of murine poly(A) sites. The 100 most significantly enriched hexamers (binomial test relative to what is expected given the mononucleotide composition of the region from -60 to 0 nt relative to poly(A) site) in the mouse poly(A) site catalog.

hexamer	-log p-value
AATAAA	78344.66
AAAAAA	33032.07
AAATAA	28932.12
ATAAAT	17302.62
ATAAAA	14803.36

Continued on next page

Table A.5 – continued from previous page

hexamer	-log p-value
TAAAAA	12938.72
TTAAAA	10366.85
TAAATA	10122.15
AAAAAG	8097.119
ATTAAA	7668.254
CAGTGT	6974.536
ATAAAG	6855.813
AAAATA	6839.607
ACAGTG	6185.573
CTGCCT	5763.978
TGTTTT	5692.668
TGTCTG	5583.763
CCTCCC	5520.302
TTTAAA	5008.553
GTGTAC	4968.018
GTGTGT	4958.019
GACAGC	4933.256
TAAAAT	4914.887
AAAAAT	4852.199
CCTCTG	4693.22
TAATAA	4460.155
CTTCTG	4436.615
TGTGTG	4411.729
CTGAAG	4159.753
TGTACT	4135.415
TTGTTT	3858.373
TTTTGT	3721.03
ATAAAC	3683.916
CCTGCC	3667.125
GTGTCT	3663.924
TTTTCT	3652.31
TGCCTC	3617.359
CTACAG	3575.848
AAAGAA	3570.49
GCTACA	3527.289
TTCTGG	3512.262
CTGTCT	3499.525
TTTGTT	3488.113
CTCCCC	3386.621
AGACAG	3353.467
TCTGAA	3231.828
ACAGCT	3161.227
CTGGTG	3148.898
AAATCT	3076.442

Continued on next page

Table A.5 – continued from previous page

hexamer	-log p-value
TCTGCC	3032.614
AAATGT	3023.56
CTGTGT	2979.327
CTCTGC	2974.548
AGTGTA	2935.839
CAATAA	2867.629
TTTCCT	2843.454
GGTGTG	2836.151
TGTGTC	2810.496
CCTGTC	2803.988
TTTTTT	2748.095
CCCTGT	2719.253
TGAAGA	2718.407
CTTCCT	2690.973
AAGAAA	2651.799
AAAAGA	2636.556
CCCTCC	2573.799
CTGCTG	2560.113
TTTCTT	2559.386
GCTGGG	2522.802
AAAAAC	2519.491
TCTCTG	2486.791
TCTGTG	2482.156
TTTCTG	2480.577
AAACCC	2460.335
AGCTAC	2456.855
TTTTAA	2438.885
TGCTGG	2436.94
CCTGGG	2436.371
GTCTGA	2414.336
TGCTGT	2412.297
CTCTGT	2361.324
TTCTGT	2360.056
GTGCTG	2358.721
AAAATG	2341.729
CAGCTA	2295.836
CCCTCT	2275.77
TACAGT	2265.152
TGTCTC	2255.793
TAAATG	2252.428
CTCCTG	2230.726
TTCTTT	2206.821
AAAACA	2176.917
CTGGGA	2176.094

Continued on next page

Table A.5 – continued from previous page

hexamer	-log p-value
TGCCTG	2171.784
CTCTTC	2161.823
GCCTCC	2150.538
GCTGTG	2141.131
TAAATC	2138.624
ACCCTG	2131.258
CCTGTG	2111.563

Table A.6: Summary statistics of 3' end sequencing libraries (A-Seq2 protocol [56]) for control-siRNA and HNRNPC-siRNA transfected HEK 293 cells.

	control-siRNA replicate 1 (ID: 29765)	**HNRNPC-siRNA replicate 1 (ID: 29766)**	**control-siRNA replicate 2 (ID: 32682)**	**HNRNPC-siRNA replicate 2 (ID: 32683)**
Number of reads sequenced	55,274,416	47,917,208	68,650,218	78,065,144
considered high-confidence reads that mapped to a unique position in the genome	6,836,446	9,265,965	13,818,252	15,319,388
Number of reads assigned to tandem poly(A) site clusters having >1 protocol support	2,991,716	4,115,507	6,989,361	8,601,510
Number of reads assigned to sample-specific clusters	2,976,577	4,107,667	6,893,361	8,529,512

Table A.7: Overview of the number and the proportion of features annotated in the human genome that are covered by poly(A) sites from different atlases.

		total	**PolyAsite**		**PolyA-seq**		**APASdb**	
			covered sites	percentage	covered sites	percentage	covered sites	percentage
genes	protein coding	21,232	18,139	85.43 %	17,742	83.56 %	16,724	78.77 %
	lincRNA	7,048	4,160	59.02 %	3,745	53.14 %	2,387	33.87 %

Continued on next page

Table A.7 – continued from previous page

		total	**PolyAsite**		**PolyA-seq**		**APASdb**	
			covered sites	percentage	covered sites	percentage	covered sites	percentage
terminal exons	protein coding	59,869	42,579	71.12 %	39,670	66.26 %	37,533	62.69 %
	lincRNA	7,153	2,689	37.59 %	2,115	29.57 %	1,753	24.51 %

Table A.8: Overview of the number and the proportion of features annotated in the mouse genome that are covered by poly(A) sites from different atlases.

	total	**PolyAsite**		**PolyA-seq**	
		covered sites	percentage	covered sites	percentage
genes	43,054	22,988	53.39 %	21,088	48.98 %
terminal exons	92,351	38,529	41.72 %	31,903	34.55 %

A.4 Supplementary Data

Please request the data from the author or access it online at:
`http://genome.cshlp.org/content/suppl/2016/07/04/gr.202432.115.DC1/Supplementary_Data_S1.bed`

Table A.9: Supplemental Data Human

Please request the data from the author or access it online at:
`http://genome.cshlp.org/content/suppl/2016/07/04/gr.202432.115.DC1/Supplementary_Data_S2.bed`

Table A.10: Supplemental Data Mouse

APPENDIX B

SUPPLEMENTARY MATERIAL TO CHAPTER 3

B.1 Supplementary Figures

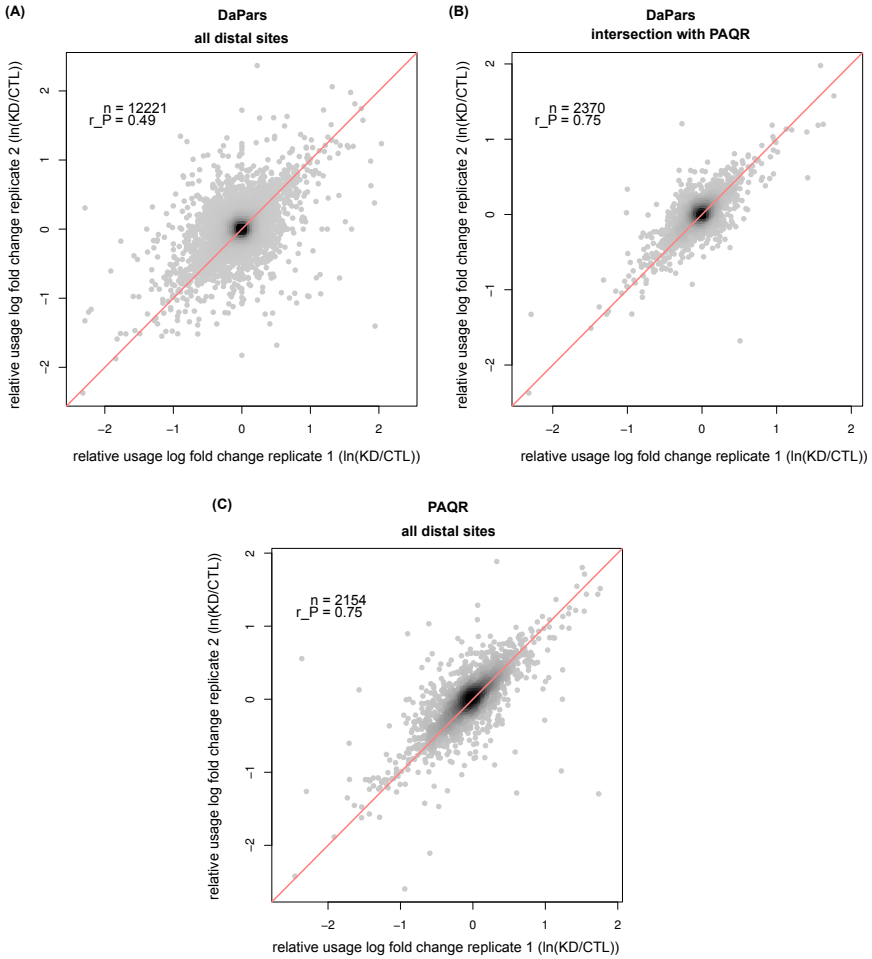


Figure B.1: DaPars estimates usage of many putative PAS at the cost of accuracy. **(A)** Reproducibility of PAS usage changes in replicates of HNRNPC knock-down compared to control samples computed based on DaPars PAS usage estimates. **(B)** The reproducibility strongly increases when only sites that are also quantified by PAQR are considered (number of sites differs slightly, as sometimes, multiple DaPars sites are closely spaced and considered as one site by PAQR. **(C)** Reproducibility of PAS usage changes in replicates of HNRNPC knock-down compared to control samples computed based on PAQR PAS usage estimates.

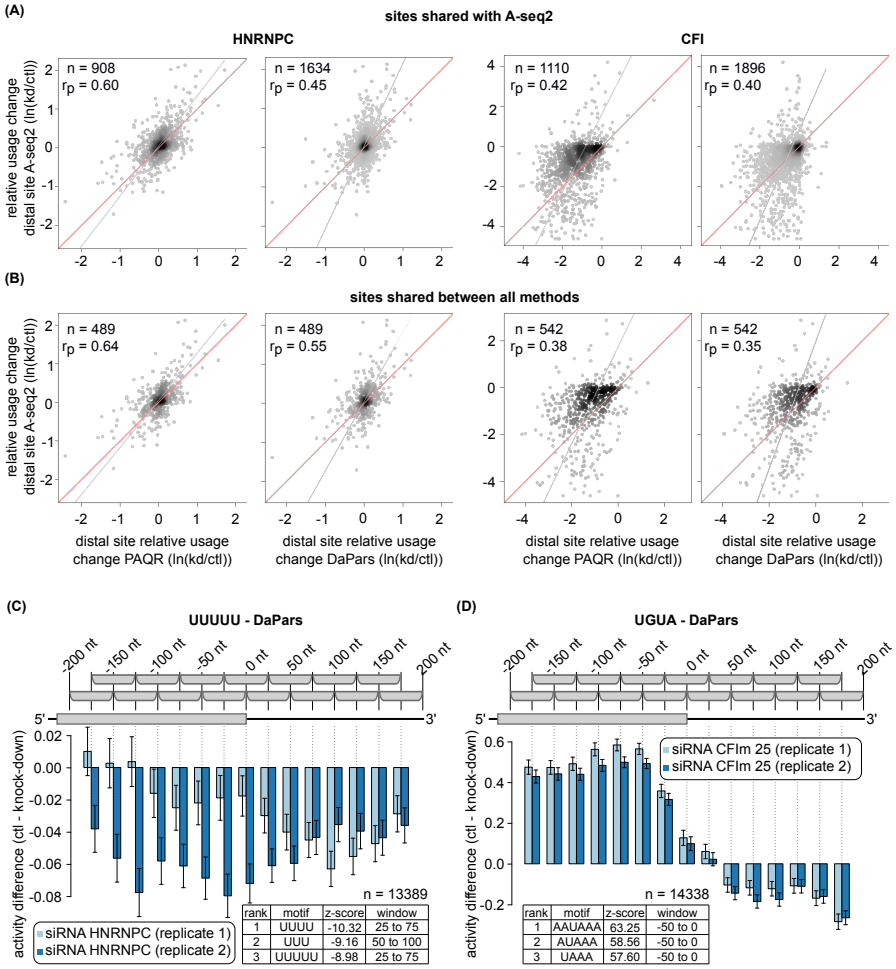


Figure B.2: Accuracy of estimates of relative PAS usage by DaPars and PAQR. (A) Log-fold changes in the relative usage of the distal poly(A) sites in the HNRNPC knock-down experiment [147] and a CFIm 25 knock-down experiment [23] were calculated with DaPars and PAQR. The number of quantified sites that overlapped with the set of distal PAS quantified by A-seq2 differed between the methods. The Pearson correlation coefficients were always positive and significant but were larger for the PAQR-based quantification. **(B)** This is also the case when focusing on the PAS quantified by all three methods. **(C)** PAQR-based quantification of PAS usage yields more significant and reproducible KAPAC-inferred motif activities compared to the DaPars-based quantification. Shown is the profile of the UUUUU motif which was ranked highest based on the A-seq2 quantifications, and third based on DaPars quantifications. **(D)** KAPAC-inferred motif activity profile for UGUA, the binding motif of CFIm, which was ranked 13th, in the KAPAC analysis based on the DaPars PAS usage quantifications of control and CFIm 25 knock-down samples [23]. Note also that with DaPars-based quantification the motif activity remain positive also in the region downstream of PAS, which was not the case when KAPAC used A-seq2-based quantification.

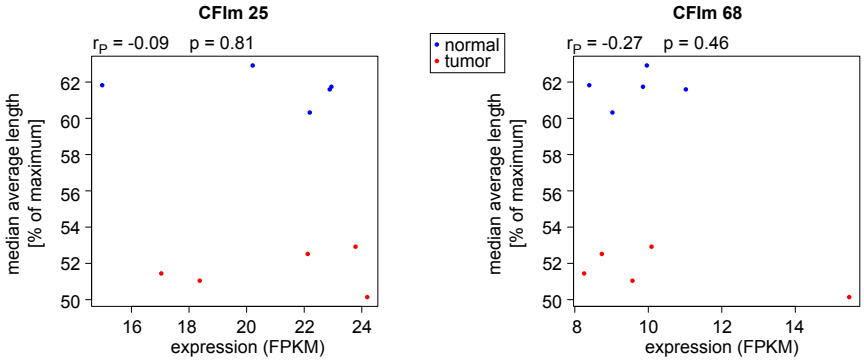


Figure B.3: CFIm 25 and 68 expression estimates (in FPKM) and corresponding average exon lengths of the ten selected GBM samples.

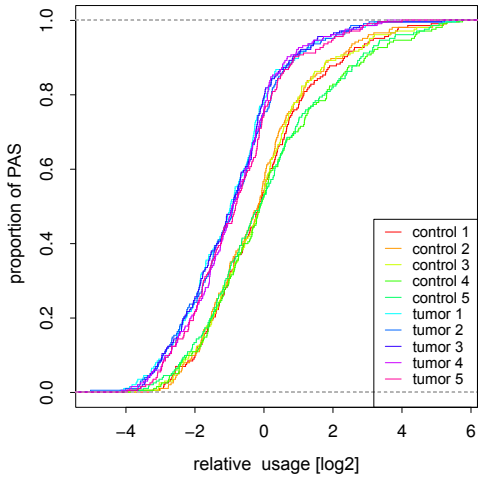


Figure B.4: Distribution of relative usages of poly(A) sites for GBM and normal brian tissue samples. Analysis of PTBP1 activity in CPA in glioblastoma. Cumulative densitity functions for the relative usage of the 203 PAS inferred by KAPAC to be targets of the PTBP1-binding UCUC motif in glioblastoma. The CDFs are shifted to the left in tumors, indicating decreased relative usage of the sites carrying the PTBP1-binding motif when the regulator has high expression (see Figure 3.5 in the main text).

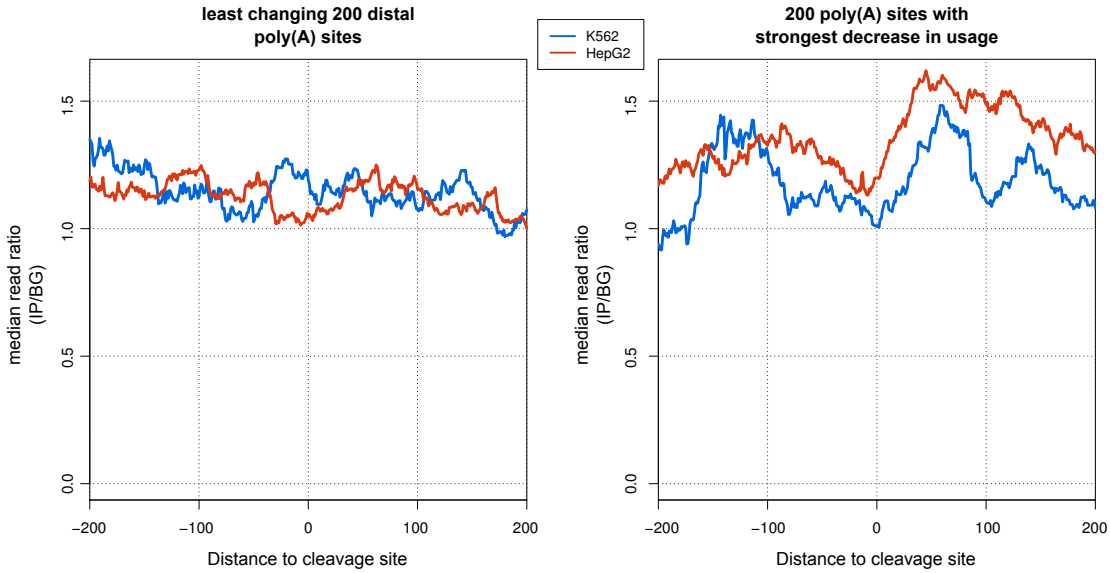


Figure B.5: Position-dependent densities of PTBP1-eCLIP reads observed in two studies (in the HepG2 (thick red line) and K562 (thick blue line) cell lines). (A) Median per-nucleotide ratio of eCLIP read densities from the foreground (PTBP1-IP) and the background (size matched control) samples in regions around 200 distal poly(A) sites that changed least between GBM and normal tissue samples (mean change between selected random pairs of tumor-normal samples). **(B)** Same as (A) but for the 200 poly(A) sites with the strongest mean decrease in usage in GBM tumors compared to normal brain samples.

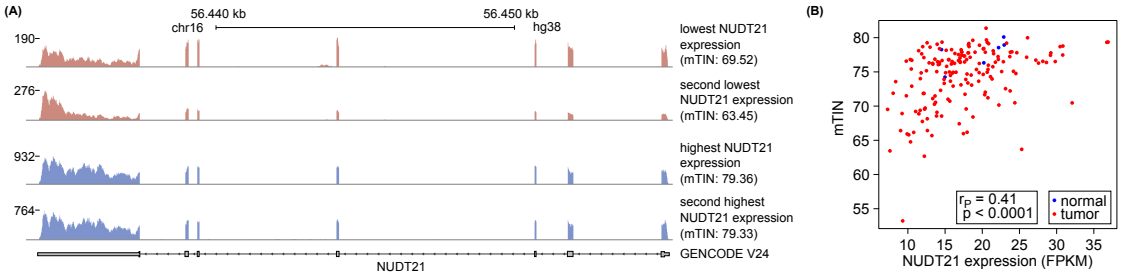


Figure B.6: Partial RNA degradation can lead to apparent reduction in gene expression level. **(A)** Screen shot of the NUDT21 (CFIm 25) locus coverage in four different RNA-seq tumor samples from the TCGA glioblastoma cohort, shown in the IGV [222] browser. The samples were selected based on their NUDT21 expression according to the FPKM values reported by the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/). The top two profiles (in red) are from samples with the lowest estimated NUDT21 expression (FPKM), whereas the two profiles at the bottom correspond to samples withe highest estimated NUDT21 expression. The median transcript integrity numbers (mTIN) computed over the entire transcriptome for the corresponding samples are also shown on the right side of the profiles. **(B)** Scatter plot of NUDT21 gene expression estimates (FPKM) obtained the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/) and mTIN values [116], indicative of RNA degradation in the samples, shows the positive correlation between apparent expression level and RNA integrity. Normal tissue samples are in blue, tumor samples in red.

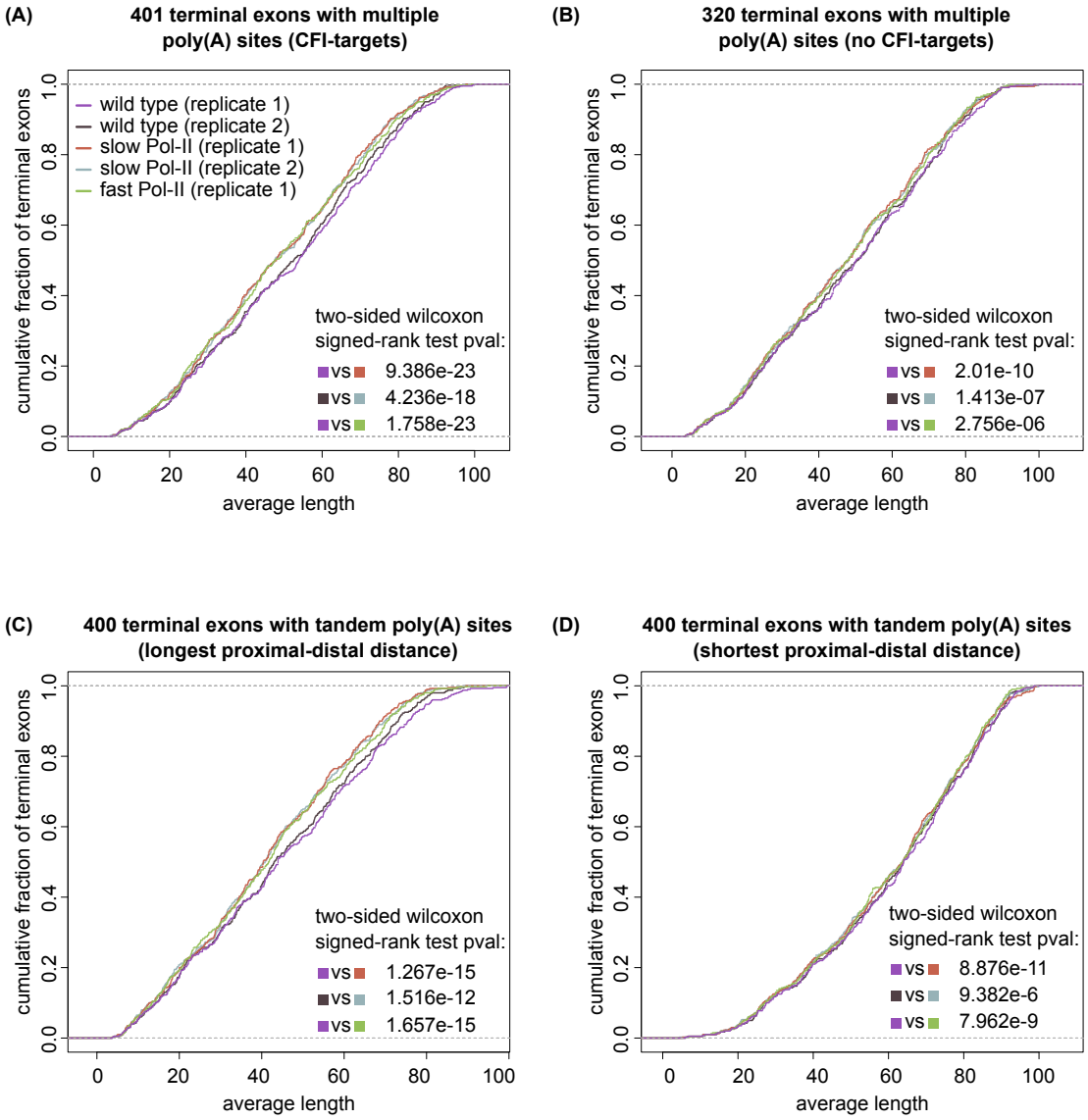


Figure B.7: Cumulative distribution functions of average terminal exon length computed across the entire transcriptome from RNA-seq data sets obtained from cells expressing RNA polymerase II (RNAPII) mutants that effect the transcription elongation rate [188]. The color scheme is preserved throughout the figure. **(A)** Distributions of average length for 401 CFIm-responsive terminal exons. **(B)** Distributions of average length for a set of control exons, that did not show a large and consistent change in length upon CFIm knock-down. **(C)** Cumulative density functions of average length of 400 terminal exons with the largest distance between proximal and distal poly(A) site (among all the terminal exons that we quantified). **(D)** Complementary to (C), this panel contains the average length distributions for 400 terminal exons with the smallest distance between the proximal and the distal site.

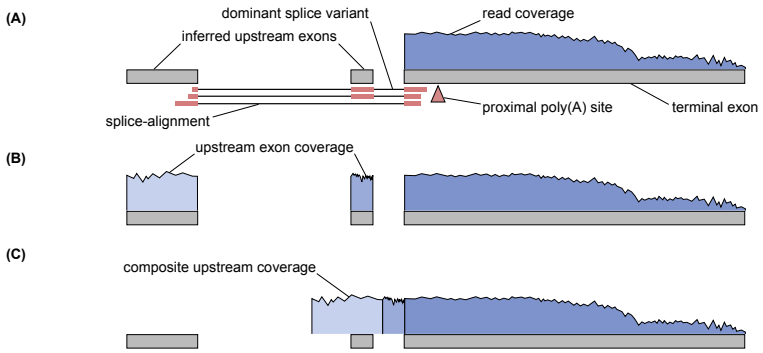


Figure B.8: Schematic representation of the procedure used to compute the coverage upstream of the poly(A) site in situations when the poly(A) site is close to the start of the terminal exon. (A) Based on all reads that align with splicing into the terminal exon of interest, the upstream exon(s) of the major splice variant is(are) inferred. The procedure is repeated until a sufficiently long upstream exonic region is reconstructed **(B)** The read coverage profile is calculated for all inferred upstream exons. **(C)** The read coverage profile of the terminal exon is then extended upstream by the coverage profile(s) of the inferred upstream exon(s).

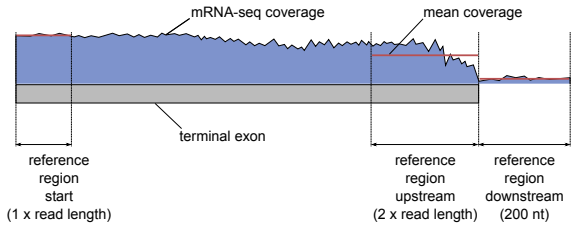


Figure B.9: Schematic representation of the features used to define the most distal poly(A) site in a terminal exon.

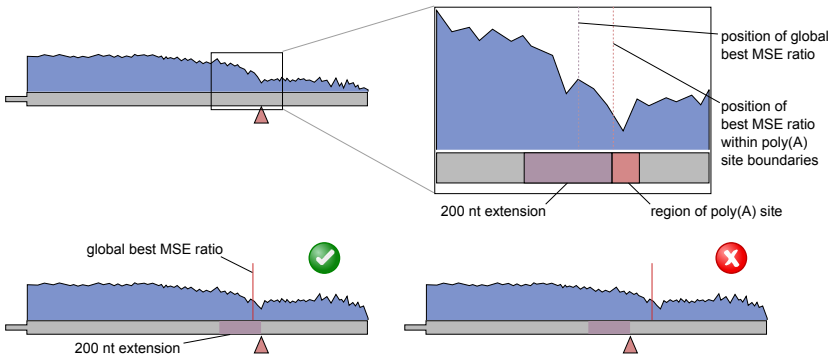


Figure B.10: Schematic representation of the final consistency check of PAQR. In a final pass of PAQR, the segmentation procedure is applied to every position within the terminal exon, identifying segments with evidence of distinct coverage. If the best segmentation point falls outside of 200 nt-long regions ending at the used poly(A) sites defined as described in the main manuscript, the exon is discarded from the analysis (bottom right panel).

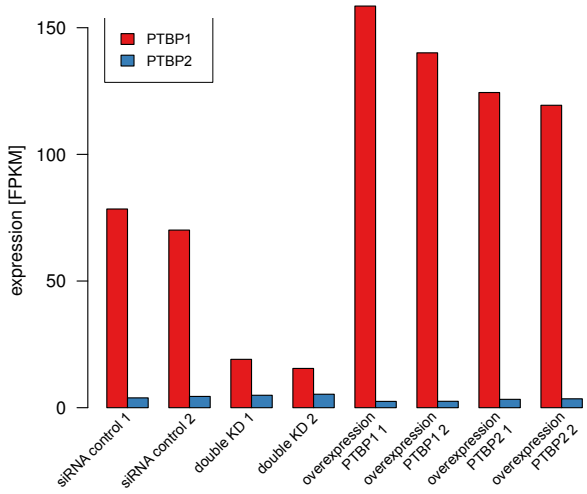


Figure B.11: Expression levels of PTBP1/2 mRNAs in HEK293 cells. PTBP1 is shown in red and PTBP2 in blue, the HEK 293 cells were treated as indicated by the x-axis labels (data from [186]). These samples were used to infer PTBP1/2 activity in polyadenylation in a human cell line (see Figure 3.6 from main text).

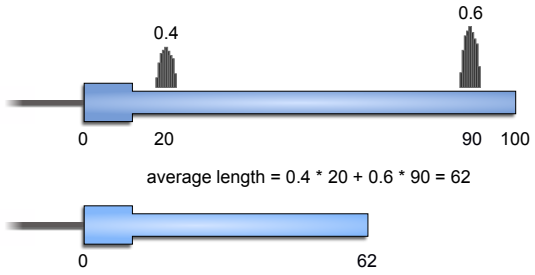


Figure B.12: Schematic illustration of the calculation of average terminal exon length.

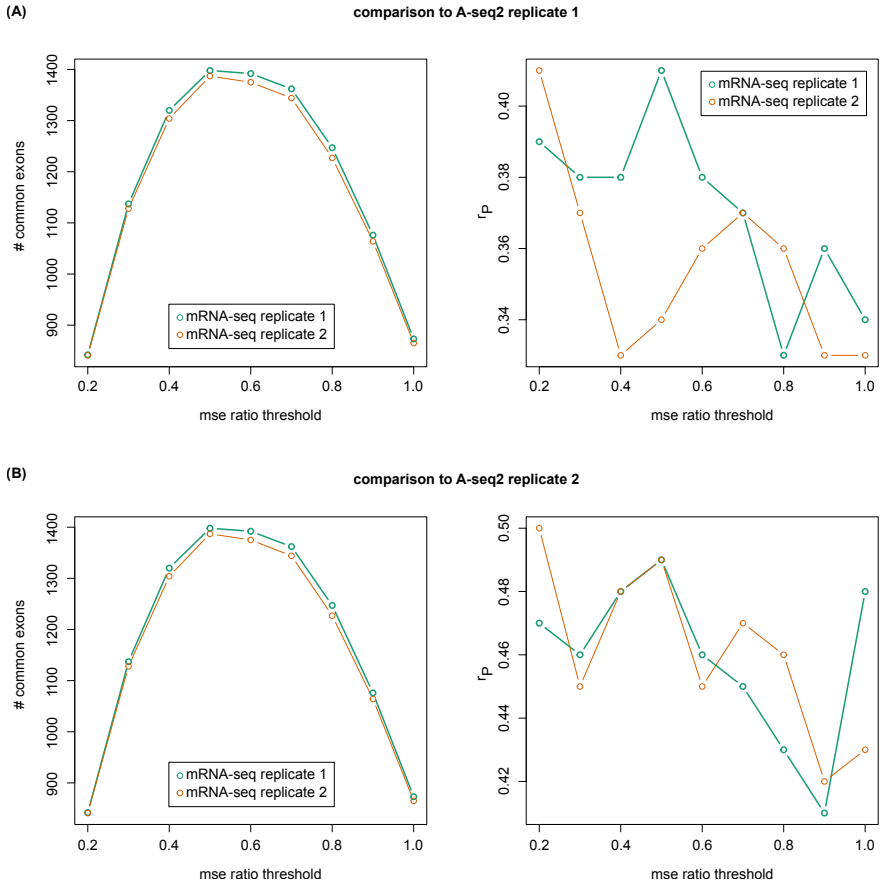


Figure B.13: Comparisons of poly(A) site quantifications based on A-seq2 (SRP065825) and on RNA-seq (GSE56010) data sets. Quantifications based on RNA-seq data were done for different mean squared error (mse) ratios used to define processed poly(A) sites. (A) Comparisons of the A-seq2 quantifications based on samples SRX1436120 and SRX1436124 (siRNA control and si-HNRNPC replicate 1) with the RNA-seq quantifications based on GSM1502498 and GSM1502500 (replicate 1) as well as GSM1502499 and GSM1502501 (replicate 2). The left panel shows the number of terminal exons with more than one poly(A) sites for which the same set of poly(A) sites was quantified in the A-seq2 samples as well as the RNA-seq samples. For the right panel, differences in average length (between control and HNRNPC knock-down) computed for individual exons either based on A-seq2 or RNA-seq data were correlated. Shown are Pearson's correlation coefficients (r_P). **(B)** Same as shown for (A) except that the quantification of terminal exon lengths from RNA-seq data was compared with the A-seq2-based quantification (samples SRX1436128 and SRX1436130 (replicate 2)).

B.2 Supplementary methods

B.2.1 Inference of poly(A) site usage from mRNA sequencing data

The drop in coverage by RNA sequencing reads within a 3' UTR has been interpreted as evidence for an internal poly(A) site [111]. However, the read coverage per position along the 3'

UTR varies quite widely, which makes it difficult to distinguish small drops in coverage that are due to the use of proximal poly(A) sites to fluctuations that occur for unknown reasons. In contrast, the read coverage generally ends abruptly in the region downstream of the gene, which makes the distal sites easier to detect. Nevertheless, methods that exploit drops in the average coverage of 3' UTR segments to infer proximal poly(A) site usage have been developed. In particular, the DaPars method [23] has been used in multiple studies that attempted to infer regulators of polyadenylation. Attempting to quantify poly(A) site usage in two systems in which both 3' and RNA sequencing data were available, we found that DaPars quantifies a large number of putative poly(A) sites internal to 3' UTRs, of which only a very small number are deemed to have signficantly different expression changes when a specific regulator of 3' end processing is depleted by siRNA-mediated knock-down (8 of 24021 for the HNRNPC knock-down and 1265 of 26520 for the CFIm knock-down). This prompted us to develop the alternative, PAQR method, which differs from DaPars in two main aspects. First, PAQR uses an extensive database of poly(A) sites that have been determined experimentally, and does not define poly(A) sites based on the coverage profile from RNA-seq. Aside from presumably avoiding some false positive sites, this approach has the advantage that it allows us to infer position-dependent motif activities. If the location of poly(A) sites had some uncertainty, this would propagate to the location of the active motifs. Second, PAQR only aims to identify terminal exons in which there is significant usage of more than one poly(A) site. Thus, terminal exons in which internal poly(A) sites are used only to a small extent, which does not stand out of the fluctuations in coverage along the 3' UTR, are reported by PAQR as "single poly(A) site" exons.

As shown in Supplementary Figure B.2, when we compare the DaPars-based and PAQR-based quantifications of distal poly(A) site usage changes in individual exons from RNA sequencing data, with the usage changes determined by 3' end sequencing, we obtain systematically higher correlations for PAQR. This is not only due to DaPars reporting more exons, because the correlations remain higher for the set of exons that are quantified by all three methods (3' end sequencing, DaPars-based and PAQR-based quantification of RNA seq data). Consistent with the improved quantification of usage by PAQR, the absolute values as well as the significance of the motif activities that we infer with KAPAC, are higher when we use PAQR-based quantifications of poly(A) site usage than when we use DaPars quantifications (B.2C and D).

B.2.2 K-mer Activity on Polyadenylation Site Choice (KAPAC)

KAPAC, standing for **k**-mer **a**ctivity on **p**oly**a**denylation site **c**hoice, aims to identify sequence motifs (of length $k = 3-6$ nucleotides (nt), hence k-mers) that can explain changes in poly(A) site (PAS) use across conditions (e.g. in samples in which the expression of a potential regulator has been perturbed). It models the change in the use of alternative PAS within a transcript as a

linear function of the occurrence of specific k-mers and the unknown regulatory impact (also called "activity") of these k-mers.

B.2.2.1 Determination of k-mer counts within defined regions relative to poly(A) sites

For a defined window relative to the poly(A) site we count the occurrences of a specific k-mer k. As we aim to identify general regulators of poly(A) site use, we discard k-mers that are found in less than 1% of all poly(A) sites. As we seek to identify factors that act in a tissue/condition-specific manner on a subset of transcripts, we calculate the number of k-mers that are found in 'excess' of what is expected to be found per chance within a sequence window (region) of interest based on the (mono)nucleotide composition of regions around poly(A) sites. Considering a region of ±200 nt around poly(A) sites from the PolyAsite atlas (human genome hg38) [118] we have obtained the following base (B) frequencies f_B:

- Adenine: f_A=0.2973
- Cytosine: f_C=0.1935
- Guanine: f_G=0.2007
- Uracil: f_U=0.3084

Using the determined base frequencies f_B we then calculated the probability to find a k-mer k_L of length L at a specific position as follows:

$$P(k_L) = \prod_{l=1}^{L} f_{B,l} \tag{B.1}$$

whereat $f_{B,l}$ is the frequency of base B observed at position l. Using equation (B.1) we then calculated how many counts of k-mer k_L are expected to be found in a region r_W of length W given the observed base frequencies (see above):

$$N_k^e = P(k_L) * (W - L) \tag{B.2}$$

Finally, for a specific region (of length L) relative to a poly(A) site i we calculate the number of "excess" counts $N_{k,i}$ by subtracting the number of expected counts of k-mer k (N_k^e) from the number of counts observed within the region $N_{k,i}^o$:

$$N_{k,i} = f^+(N_{k,i}^o - N_k^e) = \max(0, N_{k,i}^o - N_k^e) \tag{B.3}$$

We use the "excess" counts determined for k-mer k within a defined region relative to poly(A) site i to explain the relative usage of the site observed within a sample s (see below).

B.2.2.2 Derivation of k-mer activities from genome-wide changes in poly(A) site use

We use a simple linear model that tries to explain PAS use as a function of the number of occurrences of each k-mer within a defined region in close proximity to the cleavage site and

the activity of the k-mer within this region. Specifically, we define the relative use of a poly(A) site i from a terminal exon with P poly(A) sites in sample s, $U_{i,s}$ as

$$U_{i,s} = \frac{R_{i,s}}{\sum_{p=1}^{P} R_{p,s}}, \tag{B.4}$$

$R_{i,s}$ being the number of reads from the poly(A) site i in sample s. We relate $R_{i,s}$ to the transcription rate from the corresponding locus through the parameter α, the number $N_{i,k}$ of occurrences of k-mer k within a specific region relative to the poly(A) site i (see section B.2.2.1, equation B.3) and the activity $A_{k,s}$ of k-mer k within sample s:

$$R_{i,s} = \alpha * \exp(N_{k,i} * A_{k,s}). \tag{B.5}$$

Combining equations B.4 and B.5 gives:

$$U_{i,s} = \frac{\exp(N_{k,i} * A_{k,s})}{\sum_{p=1}^{P} \exp(N_{k,p} * A_{k,s})}, \tag{B.6}$$

or, in log-space, the relative use of poly(A) site i in sample s can be written as:

$$\begin{aligned} \log(U_{i,s}) &= \log\left(\frac{\exp(N_{k,i} * A_{k,s})}{\sum_{p=1}^{P} \exp(N_{k,p} * A_{k,s})}\right) \\ &= \log\Big(\exp(N_{k,i} * A_{k,s})\Big) - \log\Big(\sum_{p=1}^{P} \exp(N_{k,p} * A_{k,s})\Big) \\ &= N_{k,i} * A_{k,s} - \log\Big(\sum_{p=1}^{P} \exp(N_{k,p} * A_{k,s})\Big) \end{aligned} \tag{B.7}$$

We can define the mean of the log of the relative use $\langle \log(U_{t,s}) \rangle$ of a poly(A) site from terminal exon t with P_t poly(A) sites, in sample s as:

$$
\begin{aligned}
\langle \log(U_{t,s}) \rangle &= \frac{\sum_{i=1}^{P_t} \log(U_{i,s})}{P_t} \\
&= \frac{\sum_{i=1}^{P_t} \left(N_{k,i} * A_{k,s} - \log\left(\sum_{p=1}^{P_t} \exp(N_{k,p} * A_{k,s}) \right) \right)}{P_t} \\
&= \frac{\sum_{i=1}^{P_t} \left(N_{k,i} * A_{k,s} \right) - P_t * \log\left(\sum_{p=1}^{P_t} \exp(N_{k,p} * A_{k,s}) \right)}{P_t} \\
&= \frac{\sum_{i=1}^{P_t} \left(N_{k,i} * A_{k,s} \right)}{P_t} - \log\left(\sum_{p=1}^{P_t} \exp(N_{k,p} * A_{k,s}) \right) \\
&= \frac{\left(\sum_{p=1}^{P_t} N_{k,p} \right)}{P_t} * A_{k,s} - \log\left(\sum_{p=1}^{P_t} \exp(N_{k,p} * A_{k,s}) \right) \\
&= \langle N_k \rangle_t * A_{k,s} - \log\left(\sum_{p=1}^{P_t} \exp(N_{k,p} * A_{k,s}) \right)
\end{aligned}
\tag{B.8}
$$

where $\langle N_k \rangle_t$ is the mean count of k-mer k across the poly(A) sites of terminal exon t. We can obtain the per "terminal-exon-centered" log relative use δ_i of poly(A) site i in sample s by combining equations (B.7) and (B.8):

$$
\begin{aligned}
\delta_{i,s} &= \log(U_{i,s}) - \langle \log(U_{t,s}) \rangle \\
&= N_{k,i} * A_{k,s} - \log\left(\sum_{p=1}^{P_t} \exp(N_{k,i} * A_{k,s}) \right) \\
&\quad - \langle N_k \rangle_t * A_{k,s} + \log\left(\sum_{p=1}^{P_t} \exp(N_{k,i} * A_{k,s}) \right) \\
&= N_{k,i} * A_{k,s} - \langle N_k \rangle_t * A_{k,s} \\
&= (N_{k,i} - \langle N_k \rangle_t) * A_{k,s} \\
&= \bar{N}_{k,i} * A_{k,s}
\end{aligned}
\tag{B.9}
$$

where $\bar{N}_{k,i}$ are the per "terminal-exon-centered" site counts of k-mer k at poly(A) site i from terminal exon t.

As we are interested in finding k-mers that can explain changes in poly(A) site use between samples (e.g. control vs. knock-down) we finally calculate the change of the per "terminal-exon-centered" log relative use $\delta_{i,s}$ relative to the mean use across samples:

$$\begin{aligned}\bar{\delta}_{i,s} = \delta_{i,s} - \langle\delta_i\rangle &= (N_{i,k} - \langle N_k\rangle_t) * A_{k,s} - (N_{i,k} - \langle N_k\rangle_t) * \langle A_k\rangle \\ &= (N_{i,k} - \langle N_k\rangle_t) * (A_{k,s} - \langle A_k\rangle) \\ &= (N_{i,k} - \langle N_k\rangle_t) * \bar{A}_{k,s}\end{aligned} \tag{B.10}$$

whereas $\bar{A}_{k,s}$ is the activity of each k-mer k relative to the mean activity across samples $\langle A_k\rangle$ and $\langle\delta_i\rangle$ is the mean relative use of poly(A) site i across samples which is defined as:

$$\langle\delta_i\rangle = (N_{i,k} - \langle N_k\rangle_t) * \langle A_k\rangle \tag{B.11}$$

with $\langle A_k\rangle$ being the mean activity of motif k across samples.

Substituting the per "terminal-exon-centered" counts into equation (B.10) we find that the relative use of poly(A) site i in sample s should satisfy

$$\bar{\delta}_{i,s} = \bar{N}_{k,i} * \bar{A}_{k,s} + \epsilon \tag{B.12}$$

which allows us to obtain a fitted activity $\tilde{A}_{k,s}$ and a corresponding error $\tilde{\sigma}_{k,s}$ for each k-mer k in sample s, using a standard least-squares approach to solve the simple linear regression model (equation (B.12)).

B.2.2.3 Ranking of k-mers

For each pair of treatment-control samples tcp (or tumor versus (matched) normal tissue) we calculate for each k-mer k an activity difference z-score:

$$z_{tcp,k} = \frac{\tilde{A}_{k,control} - \tilde{A}_{k,treatment}}{\sqrt{\tilde{\sigma}^2_{k,control} + \tilde{\sigma}^2_{k,treatment}}} \tag{B.13}$$

We then combine the data from multiple replicates (or multiple patients) by calculating a mean activity difference z-score (Z_k) considering all treatment-control pairs:

$$Z_k = \frac{\sum_{tcp}^{TCP} z_{tcp,k}}{TCP} \tag{B.14}$$

where TCP is the number of treatment versus control pairs of samples (tcp). KAPAC ranks k-mers by their absolute mean activity difference z-scores.

B.2.2.4 Determination of significant mean activity difference z-scores

Given that real sequences have compositional biases that are difficult to model, we evaluate the significance of an inferred mean activity difference z-score (Z_k) for k-mer k (see equation (B.14)) using a randomization approach. Namely, we randomize the associations of changes in poly(A) site use with k-mer counts, by randomizing the expression values of individual poly(A) sites across the genome and then calculating the relative use of poly(A) sites according to equation (3.1). We fit the model, repeating the procedure (e.g. 100 times). Then, for each k-mer k, we calculate the p-value of the real mean activity difference z-score (Z_k), assuming a Gaussian distribution of the score for the k-mer, with mean and variance determined from the randomized runs. KAPAC reports the obtained p-value, the Bonferroni adjusted p-value (taking into account the total number of considered k-mers) as well as the p-value obtained by conducting a Shapiro-Wilk normality test on the mean activity difference z-scores from the randomized runs.

B.2.3 K-mer rankings and activity plots presented in Figures 3.2–3.6 of the main manuscript

As 3' end processing factors generally bind at defined distances with respect to the processing site, we have performed the KAPAC analysis independently for regions located at specific distances from the poly(A) sites. We have used windows of 50 nt, sliding by 25 nt at a time ('Sliding Window Approach'). A similar analysis could be implemented for windows extending to defined distances upstream or downstream of the PAS ('Extending Window Approach'). To identify the most active k-mers across all regions, we used for each k-mer the highest absolute mean activity difference z-score (Z_k, see equation (B.14)) across all regions, as a ranking criterion. k-mers with a Bonferroni adjusted p-value ≥ 0.05 were not considered (e.g. Supplementary Table B.3).

B.2.4 Prediction of "targets" of the CU-rich repeat motif used for the PTBP1-eCLIP data analysis

We evaluate the "quality" of a target, containing a specific motif inferred to be active in samples of interest, by comparing the log likelihoods of the model that includes the counts $k_{i,s}$ of the motif k in the putative target region p and sample s and a model that does not. The difference in likelihoods gives us a measure of how important the motif is for predicting the observed change in the use of the respective poly(A) site.

We predict the relative use $\tilde{\delta}^N_{i,s}$ of a poly(A) site i in sample s from the inferred k-mer activities:

$$\tilde{\delta}^N_{i,s} = \bar{N}_{k,i} * \tilde{A}_{k,s} \tag{B.15}$$

We then calculate the $\chi^{2,N}_{i,s}$ statistic for all the poly(A) sites (P_t) in the corresponding terminal exon:

$$\chi^{2,N}_{i,s} = \sum_{t}^{P_t} (\delta_{t,s} - \tilde{\delta}^{N}_{t,s})^2 \tag{B.16}$$

where $\delta_{t,s}$ is the measured per "exon-centered" log relative use of poly(A) site t in sample s (see equation B.9).

Next, we calculate the predicted expression using the model in which we have set the counts of k-mer k at poly(A) site i ($N_{i,k}$) to zero and then re-centered the k-mer counts for all poly(A) sites in the corresponding exon thereby obtaining a new k-mer count matrix N'. We use N' to predict the relative use $\tilde{\delta}^{N'}_{i,s}$ of a poly(A) site i in sample s using the inferred k-mer activities:

$$\tilde{\delta}^{N'}_{i,s} = \bar{N'}_{k,i} * \tilde{A}_{k,s} \tag{B.17}$$

Afterwards we calculate the $\chi^{2,N'}_{i,s}$ of a poly(A) site i in sample s using the new k-mer count matrix N' by summing over all poly(A) sites that are located in the exon (containing P_t expressed poly(A) sites, including i) as follows:

$$\chi^{2,N'}_{i,s} = \sum_{t}^{P_t} (\delta_{t,s} - \tilde{\delta}^{N'}_{t,s})^2 \tag{B.18}$$

where $\delta_{t,s}$ is the measured per "exon-centered" log relative use (as in equation B.16 above).

We use the normalized log likelihood ratio as score $S_{i,k}$ for k-mer k targeting poly(A) site i:

$$S_{i,k} = \frac{\sum_{s}^{S} (\chi^{2,N'}_{i,s} - \chi^{2,N}_{i,s})}{\langle \chi^2 \rangle * P_t} \tag{B.19}$$

whereas P_t is the number of expressed poly(A) sites in the exon that contains poly(A) site i and the average squared-deviation per sample/poly(A) site combination $\langle \chi^2 \rangle$ is defined as:

$$\langle \chi^2 \rangle = \frac{\sum_{s}^{S} \sum_{i}^{I} (\delta_{i,s} - \tilde{\delta}^{N}_{i,s})^2}{S * I} \tag{B.20}$$

with S being the total number of samples and I being the total number of expressed poly(A) sites.

B.2.5 Processing of RNA-seq data from the study of RNAPII elongation rate

Raw reads were obtained from GEO (accession number: GSE63375) and processed according to the RNA-seq pipeline for long RNAs provided by the ENCODE Data Coordinating Center [198] using the GENCODE version 24 human gene annotation. Based on the obtained bam files, the median TIN score (mTIN) according to Wang et al. [116] was calculated based on all transcripts with a terminal exon containing more than one poly(A) site. The obtained values were:

sample	mTIN
wild type replicate 1	78.384943
wild type replicate 2	79.868301
slow RNAPII replicate 1	77.286761
slow RNAPII replicate 2	77.233262
fast RNAPII replicate 1	79.240207
fast RNAPII repicate 2	52.603342

All samples with a mTIN score of at least 70 were further processed, i.e. replicate two of the fast RNAPII sample was not considered.

B.2.6 Definition of CFI targets for the analysis of RNAPII elongation rate

Only terminal exons with at least two poly(A) sites that were quantified in both studies CFIm 25/CFIm 68 knock-down in HeLa cells and RNAPII speed mutations in HEK293 cells were considered. For all samples from the CFIm study, the average length differences (knock-down/mutant versus control) were obtained and the exons were stratified as follows: exons with a consistent length change in all comparisons were used as targets whereas those with an inconsistent change in any of the comparisons were marked as non-target.

B.2.7 Expression estimation for PTBP1 and PTBP2

FPKM values were calculated by strictly following the GDC workflow for TCGA RNA-seq sample processing (`https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline`). After obtaining a raw read quantification of genes with HTseq-count [201], FPKM values were calculated by using the length of a composite exon model per gene and all reads that map to protein-coding genes as library size.

B.2.8 Selection of distal PAS

For the comparison of DaPars [23] with A-seq2, the 3' ends of the exons reported by DaPars were intersected with the poly(A) site clusters quantified by A-seq2. Poly(A) site clusters that overlapped with a 3' end quantified by DaPars were considered to be expressed in both studies. Since PAQR uses annotated PAS as input, the same PAS must have been identified as distal site for the comparison of samples from PAQR and A-seq2.

B.2.9 Selection of subsets of poly(A) sites for the analysis of PTBP1-eCLIP read enrichment

From exons with at least two quantified poly(A) sites we selected the 200 that had the strongest mean decrease in relative usage across all GBM tumor-normal pairs. As control set, we used the 200 distal poly(A) sites whose absolute mean relative change in usage was smallest.

B.3 Supplementary Tables

Table B.1: Overview on KAPAC results for the 3' end sequencing data of the indicated study. Shown are the 20 most significant k-mers, or all if there are less than 20 significant ones (Bonferroni corrected p-value < 0.05) with their mean activity difference z-score (between normal and tumor samples), their adjusted p-value and the window for which the activity difference z-score was inferred.

study	rank	k-mer	z-score difference	adjusted p-value	window
PCBP1 [85]	nr 1	CCC	14.3619386027758	3.92210e-29	u75to25.d0to0
	nr 2	CCCU	12.3234804203847	1.35376e-22	u75to25.d0to0
	nr 3	UCCC	12.2883102355394	4.62113e-23	u75to25.d0to0
	nr 4	CCU	11.2584739909402	8.03581e-20	u100to50.d0to0
	nr 5	UCC	10.3078840034045	9.40719e-16	u100to50.d0to0
	nr 6	GCCC	9.26200680989259	7.40421e-10	u75to25.d0to0
	nr 7	UCCCU	8.95753524332782	6.62195e-12	u75to25.d0to0
	nr 8	CCCC	8.60555616896579	5.53819e-06	u75to25.d0to0
	nr 9	CCCUU	8.44098566788859	2.74658e-07	u100to50.d0to0
	nr 10	CCCA	7.93988407765	2.47397e-05	u125to75.d0to0
	nr 11	CCUU	7.47584995939994	3.68167e-07	u100to50.d0to0
	nr 12	UCCCC	7.37898773800494	1.23901e-03	u75to25.d0to0
	nr 13	GCC	7.15590255944904	6.66673e-04	u100to50.d0to0
	nr 14	CUCC	7.06136031907776	1.13564e-03	u100to50.d0to0
	nr 15	CUUCC	7.04848749984694	5.63203e-04	u100to50.d0to0
	nr 16	CUCCC	6.97920980000055	1.25769e-03	u100to50.d0to0
	nr 17	UGCCC	6.97713370646291	2.43523e-06	u75to25.d0to0
	nr 18	UUCCC	6.78133271365627	2.14076e-04	u75to25.d0to0
	nr 19	CCCCU	6.66612158690429	2.24487e-03	u75to25.d0to0
	nr 20	UCCCUU	6.65186476207128	1.61041e-05	u75to25.d0to0
HNRNPC [118]	nr 1	UUUUU	-17.1226127126567	3.13308e-43	u25to0.d0to25
	nr 2	UUU	-16.5958380846563	7.05715e-41	u25to0.d0to25
	nr 3	UUUU	-16.3028956832816	4.92083e-38	u25to0.d0to25
	nr 4	UUUUUU	-13.0554616188556	4.11552e-24	u25to0.d0to25
	nr 5	UUUG	-11.7071099513396	6.27981e-15	u50to0.d0to0
	nr 6	UUUUG	-11.1561957571474	3.77578e-14	u50to0.d0to0
	nr 7	CAG	10.627744491379	1.38687e-15	u50to0.d0to0
	nr 8	UUUUUG	-10.2783217400175	1.23036e-13	u50to0.d0to0
	nr 9	AUUUUU	-10.1232256769064	4.77752e-18	u25to0.d0to25
	nr 10	CUC	10.0776142444971	3.43206e-11	u75to25.d0to0
	nr 11	UUG	-9.97742310106386	1.53611e-10	u75to25.d0to0
	nr 12	AUUU	-9.66082399472801	1.10721e-10	u50to0.d0to0
	nr 13	GUUU	-9.5733404808427	1.22066e-11	u25to0.d0to25
	nr 14	UAUU	-9.51544130944631	4.81927e-15	u50to0.d0to0
	nr 15	UUGU	-9.43882410226739	1.50859e-12	u50to0.d0to0
	nr 16	UUUUUA	-9.37333750745204	1.73277e-09	u50to0.d0to0
	nr 17	UUUGU	-9.2243373238644	3.35869e-10	u50to0.d0to0
	nr 18	AUUUU	-9.20476745475456	5.42214e-08	u50to0.d0to0
	nr 19	UGU	-9.11478153686116	4.99393e-09	u75to25.d0to0
	nr 20	CUUUUU	-9.10287686716165	3.52874e-13	u25to0.d0to25
CFIm	nr 1	AAUAAA	10.9656674102611	4.64404e-18	u50to0.d0to0
	nr 2	UGUA	10.5588960353076	2.00776e-15	u75to25.d0to0
	nr 3	AAUAA	10.054856313557	1.39373e-14	u50to0.d0to0
	nr 4	UGU	10.0267388803685	1.84259e-14	u150to100.d0to0

Continued on next page

TableB.1 – continued from previous page

study	rank	k-mer	z-score difference	adjusted p-value	window
	nr 5	AAUA	9.70222909873153	3.59413e-13	u50to0.d0to0
	nr 6	UAA	9.44033417374034	2.03604e-17	u50to0.d0to0
	nr 7	UAU	9.05991932725031	5.35025e-14	u100to50.d0to0
	nr 8	AAU	8.91980457038989	1.17062e-11	u50to0.d0to0
	nr 9	UAAA	8.90689935325978	2.04303e-12	u50to0.d0to0
	nr 10	UUGUA	8.86183077056082	3.07792e-14	u75to25.d0to0
	nr 11	AUAAA	8.56029151784637	9.77397e-11	u50to0.d0to0
	nr 12	UUGU	8.36396835293178	3.03781e-11	u150to100.d0to0
	nr 13	GUA	8.26380253803631	4.16344e-09	u75to25.d0to0
	nr 14	UUUA	8.23518199723387	2.24869e-08	u100to50.d0to0
	nr 15	GCC	-7.81567064349458	2.43275e-09	u100to50.d0to0
	nr 16	UUA	7.75266447353091	2.67136e-09	u100to50.d0to0
	nr 17	AAAU	7.72642936923094	7.20373e-09	u50to0.d0to0
	nr 18	UGUAA	7.64644569977783	2.41648e-08	u75to25.d0to0
	nr 19	AUAA	7.57449370048538	1.95289e-08	u50to0.d0to0
	nr 20	UUU	7.56573439493848	4.93710e-11	u150to100.d0to0

Table B.2: Overview on KAPAC results for the standard RNA-seq data of the indicated study. Shown are the 20 most significant k-mers, or all if there are less than 20 significant ones (Bonferroni corrected p-value < 0.05) with their mean activity difference z-score (between normal and tumor samples), their adjusted p-value and the window for which the activity difference z-score was inferred.

study	rank	k-mer	z-score difference	adjusted p-value	window
HNRNPC [147]	nr 1	UUUUU	-21.66427672982	2.47991e-23	u0to0.d0to50
	nr 2	UUUU	-21.3480701671228	3.68403e-29	u0to0.d0to50
	nr 3	UUUUUU	-19.5362176459321	1.88716e-18	u0to0.d0to50
	nr 4	UUU	-17.672200613449	2.53044e-16	u0to0.d0to50
	nr 5	AUUUUU	-12.4335263034769	9.15171e-07	u25to0.d0to25
	nr 6	CUUUUU	-12.0173198238539	5.40989e-06	u0to0.d0to50
	nr 7	AUUUU	-11.9391554672639	1.96168e-08	u25to0.d0to25
	nr 8	UUUUUG	-11.1279217133331	8.46027e-06	u0to0.d25to75
	nr 9	ACUGCA	-10.4544299166344	2.85868e-05	u0to0.d75to125
	nr 10	UUUUUA	-10.3723578596203	8.66192e-06	u25to0.d0to25
	nr 11	GCCUCC	-9.96981626596616	2.17326e-05	u0to0.d100to150
	nr 12	GCCUC	-9.87551667175939	1.62913e-04	u0to0.d100to150
	nr 13	UUUUG	-9.65914100131002	6.08779e-04	u0to0.d0to50
	nr 14	CUUUU	-9.60735872053696	1.47678e-03	u0to0.d0to50
	nr 15	GCAACC	-9.54646549107147	3.03805e-05	u0to0.d75to125
	nr 16	UUUUUC	-9.44597128693347	2.65561e-03	u0to0.d25to75
	nr 17	GUUUUU	-9.43523677200803	9.70228e-04	u75to25.d0to0
	nr 18	CACUGC	-9.40557479481014	2.03784e-03	u0to0.d75to125
	nr 19	GCCAUU	-9.25103729882374	6.03798e-04	u0to0.d100to150
	nr 20	GCUCAC	-8.95404320524543	2.68805e-03	u0to0.d75to125
CFIm 25 [23]	nr 1	UGU	24.3829655172491	9.18903e-28	u125to75.d0to0
	nr 2	UGUA	23.6103952635293	4.48714e-29	u100to50.d0to0
	nr 3	AUU	22.5276877058983	3.92576e-45	u150to100.d0to0
	nr 4	AAUAAA	22.1747258931642	1.70306e-29	u50to0.d0to0
	nr 5	UAU	21.6483480370493	1.63728e-26	u100to50.d0to0
	nr 6	UUGU	21.4616586590102	6.84804e-40	u150to100.d0to0

Continued on next page

TableB.2 – continued from previous page

study	rank	k-mer	z-score difference	adjusted p-value	window
	nr 7	UUA	19.9991395007045	3.66046e-30	u150to100.d0to0
	nr 8	AAUAA	19.9668688810485	4.81553e-21	u50to0.d0to0
	nr 9	UAUU	19.9580615697656	9.55971e-30	u150to100.d0to0
	nr 10	AUUU	19.0887468345882	4.89942e-23	u150to100.d0to0
	nr 11	UUGUA	18.3972829494701	9.38891e-20	u100to50.d0to0
	nr 12	UGUAU	18.314537459356	6.23009e-19	u125to75.d0to0
	nr 13	GUA	18.2355799427803	3.93916e-19	u100to50.d0to0
	nr 14	UAA	18.1279696821075	1.78465e-15	u200to150.d0to0
	nr 15	UUUA	18.117483027468	8.87175e-23	u175to125.d0to0
	nr 16	GGA	-18.0412579969026	1.60360e-22	u150to100.d0to0
	nr 17	CAG	-17.8522835736229	4.38540e-18	u150to100.d0to0
	nr 18	AUUGU	17.7561679662	3.11635e-18	u125to75.d0to0
	nr 19	AUAAA	17.4863596662839	7.61925e-16	u50to0.d0to0
	nr 20	UUAU	17.42803060558	1.78107e-20	u150to100.d0to0
PTBP1/2 [186]	nr 1	UCU	9.88770416432942	5.36039e-06	u25to0.d0to25
	nr 2	CUCU	8.15341786865457	1.98386e-03	u25to0.d0to25
	nr 3	CUCUCU	8.03270724839405	1.01339e-03	u0to0.d25to75
	nr 4	UUCU	7.85922178238784	3.30486e-03	u25to0.d0to25
	nr 5	UUCUC	7.51573593114661	1.49623e-02	u50to0.d0to0
	nr 6	UUGUGU	7.5029478372414	2.65028e-02	u0to0.d25to75
	nr 7	UCUUC	7.47146651882119	1.26587e-02	u50to0.d0to0
	nr 8	GACUA	7.37672158832967	1.72687e-02	u0to0.d75to125
	nr 9	UCUCCU	7.26369202083234	3.75263e-02	u50to0.d0to0
	nr 10	CUCUUC	7.16401311199207	4.13901e-02	u50to0.d0to0
	nr 11	CUC	7.04876822494446	3.88378e-02	u50to0.d0to0
	nr 12	UGGUGA	7.04005435369631	4.72596e-02	u0to0.d100to150
	nr 13	AGGACU	6.98761283234312	4.12172e-02	u0to0.d75to125
	nr 14	UUUAUA	-6.92047578472529	4.13303e-02	u25to0.d0to25

Table B.3: Overview on KAPAC results for the different cancer types if they are considered in the main manuscript. Shown are the 20 most significant k-mers, or all if there are less than 20 significant ones (Bonferroni corrected p-value < 0.05) with their mean activity difference z-score (between normal and tumor samples), their adjusted p-value and the window for which the activity difference z-score was inferred.

cancer cohort	rank	k-mer	z-score difference	adjusted p-value	window
COAD	nr 1	UUUUU	-4.96937708701905	3.74967e-02	u50to0.d0to0
	nr 2	UUUGU	-4.21481585142658	1.73656e-02	u0to0.d150to200
LUAD	nr 1	AGCUUG	4.40759145275076	7.68629e-03	u75to25.d0to0
	nr 2	CCUUCC	-4.23280169144492	8.94116e-03	u0to0.d0to50
	nr 3	UAUU	4.22326523375785	5.05276e-04	u125to75.d0to0
	nr 4	GAAG	-4.04537916018301	4.83960e-02	u200to150.d0to0
	nr 5	UAU	4.02434402961917	4.09451e-03	u125to75.d0to0
	nr 6	GUAUGA	3.88754196033867	2.36437e-02	u50to0.d0to0
	nr 7	CCUUC	-3.87652295059531	3.79403e-02	u0to0.d0to50
	nr 8	GUAU	3.84571509896243	1.80215e-02	u125to75.d0to0
	nr 9	CCCAG	-3.6454244635503	3.29534e-02	u150to100.d0to0
	nr 10	CCCCUU	3.473394114303	4.78457e-02	u0to0.d50to100
	nr 11	UGUAUU	3.40476817035171	4.85372e-03	u125to75.d0to0

Continued on next page

TableB.3 – continued from previous page

cancer cohort	rank	k-mer	z-score difference	adjusted p-value	window
PRAD	nr 1	AUU	-4.01827015983043	2.46805e-03	u0to0.d25to75
	nr 2	AUUU	-3.99168739737411	1.26721e-02	u0to0.d25to75
	nr 3	UAUUU	-3.88652328998694	1.32727e-03	u0to0.d25to75
	nr 4	UAUU	-3.53765426692504	4.47438e-02	u0to0.d25to75
GBM[1]	nr 1	UCUCUC	9.90740192572668	1.68536e-06	u0to0.d25to75
	nr 2	CUCUCU	9.22208774710919	3.06294e-05	u0to0.d25to75
	nr 3	CUCUC	8.74301074486753	7.44746e-06	u0to0.d25to75
	nr 4	UCUC	8.66858168373634	4.66253e-06	u0to0.d25to75
	nr 5	UCUCU	8.35363357910616	8.08023e-04	u0to0.d25to75
	nr 6	CUC	8.07444814776459	8.29237e-04	u0to0.d25to75
	nr 7	UUAU	-7.56922933209396	4.70394e-03	u0to0.d150to200
	nr 8	CUCU	7.48320885412783	1.79984e-03	u0to0.d25to75
	nr 9	AAUAU	-7.41968885048703	3.63770e-03	u0to0.d75to125
	nr 10	AUUU	-7.39880077048514	3.89400e-03	u0to0.d50to100
	nr 11	UGUGUG	7.12420875874659	2.80547e-02	u0to0.d50to100
	nr 12	AAUCCC	6.91920737044926	1.99104e-02	u0to0.d125to175
	nr 13	AUU	-6.86277219404571	2.72602e-02	u0to0.d50to100
	nr 14	AAAUAU	-6.81585214892369	1.98374e-02	u0to0.d75to125
	nr 15	CAGGC	6.7400778189719	1.44156e-02	u0to0.d150to200
	nr 16	GGC	6.6874316792108	3.34506e-02	u0to0.d125to175
	nr 17	CAUCUU	-6.65269394378098	3.90724e-03	u175to125.d0to0
	nr 18	CUGAC	-6.64458751628684	3.93609e-04	u175to125.d0to0
	nr 19	UUA	-6.63359494853029	1.86693e-02	u0to0.d125to175
	nr 20	CCCAGC	6.29440474138774	2.49740e-02	u0to0.d125to175

Table B.4: Overview on the number of considered normal-tumor comparisons and on the number of quantified terminal exons with at least two poly(A) sites per cancer cohort.

cancer cohort	number of normal-tumor comparisons	number of overall quantified exons
BLCA	15	2046
BRCA	93	2863
COAD	32	2470
ESCA	8	2338
HNSC	37	2779
KICH	19	2370
KIRC	63	2587
KIRP	24	2453
LIHC	13	1941
LUAD	58	2619

Continued on next page

[1] The given results were inferred based on randomly assigned pairs of normal and tumor samples, not on matching pairs of samples from individual patients.

TableB.4 – continued from previous page

cancer cohort	number of normal-tumor comparisons	number of overall quantified exons
LUSC	46	2879
PRAD	43	2450
READ	6	1784
STAD	22	3208
THCA	37	2219
UTEC	9	1916

Table B.5: Overall number of processed TCGA samples per cancer type (obtained from https://portal.gdc.cancer.gov/).

cancer cohort	number of samples
BLCA	40
BRCA	229
CESC	6
CHOL	18
COAD	87
ESCA	16
GBM	169
HNSC	86
KICH	46
KIRC	144
KIRP	62
LIHC	100
LUAD	125
LUSC	98
PAAD	8
PCPG	6
PRAD	106
READ	18
SARC	4
STAD	54
THYM	4

Continued on next page

TableB.5 – continued from previous page

cancer cohort	**number of samples**
THCA	116
UCEC	46

APPENDIX C

LIST OF PUBLICATIONS

As part of the projects presented here, the following articles were published or accepted for publication.

Gruber Andreas J., Schmidt Ralf, Gruber Andreas R., Martin Georges, Ghosh Souvik, Belmadani Manuel, Keller Walter, Zavolan Mihaela. 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. *Genome Res* **26**: 1145–1159.

Martin Georges, Schmidt Ralf, Gruber Andreas J., Ghosh Souvik, Keller Walter, Zavolan Mihaela. 2017. 3' End Sequencing Library Preparation with A-seq2. *J Vis Exp* e56129.

Gruber Andreas J., Schmidt Ralf, Ghosh Souvik, Martin Georges, Gruber Andreas R., van Nimwegen Erik, Mihaela Zavolan. 2018. Discovery of physiological and cancer-related regulators of 3' UTR processing with KAPAC. *Genome Biol* **19**: 44.

Schmidt Ralf, Ghosh Souvik, Zavolan Mihaela. accepted 2018. The 3' UTR landscape in cancer. *Encyclopedia of Life Sciences*

BIBLIOGRAPHY

[1] L. A. Boyer, T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch, and R. A. Young. "Core transcriptional regulatory circuitry in human embryonic stem cells". *Cell* 122.6 (Sept. 2005), pp. 947–956 (cit. on p. 1).

[2] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". *Cell* 161.5 (May 2015), pp. 1202–1214 (cit. on p. 1).

[3] E. Evguenieva-Hackenberg and G. Klug. "New aspects of RNA processing in prokaryotes". *Curr. Opin. Microbiol.* 14.5 (Oct. 2011), pp. 587–592 (cit. on p. 1).

[4] B. J. Blencowe. "Alternative splicing: new insights from global analyses". *Cell* 126.1 (July 2006), pp. 37–47 (cit. on p. 1).

[5] H. Zhang, J. Y. Lee, and B. Tian. "Biased alternative polyadenylation in human tissues". *Genome Biol.* 6.12 (Nov. 2005), R100 (cit. on p. 1).

[6] P. J. Shepard, E.-A. Choi, J. Lu, L. A. Flanagan, K. J. Hertel, and Y. Shi. "Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq". *RNA* 17.4 (Apr. 2011), pp. 761–772 (cit. on pp. 1, 2, 8, 14, 18, 82, 106, 109).

[7] A. Derti, P. Garrett-Engele, K. D. Macisaac, R. C. Stevens, S. Sriram, R. Chen, C. A. Rohl, J. M. Johnson, and T. Babak. "A quantitative atlas of polyadenylation in five mammals". *Genome Res.* 22.6 (June 2012), pp. 1173–1183 (cit. on pp. 1, 2, 4, 14, 15, 18, 19, 29, 30, 37, 73, 83, 97, 98, 106–108).

[8] W. F. Marzluff, E. J. Wagner, and R. J. Duronio. "Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail". *Nat. Rev. Genet.* 9.11 (Nov. 2008), pp. 843–854 (cit. on p. 1).

[9] D. F. Colgan and J. L. Manley. "Mechanism and regulation of mRNA polyadenylation". *Genes Dev.* 11.21 (Nov. 1997), pp. 2755–2766 (cit. on p. 1).

[10] P. Richard and J. L. Manley. "Transcription termination by nuclear RNA polymerases". *Genes Dev.* 23.11 (June 2009), pp. 1247–1269 (cit. on p. 1).

[11] N. J. Proudfoot. “Ending the message: poly(A) signals then and now”. *Genes Dev.* 25.17 (Sept. 2011), pp. 1770–1782 (cit. on pp. 1, 2, 9, 14, 31, 45, 74).

[12] B. Tian, J. Hu, H. Zhang, and C. S. Lutz. “A large-scale analysis of mRNA polyadenylation of human and mouse genes”. *Nucleic Acids Res.* 33.1 (Jan. 2005), pp. 201–212 (cit. on pp. 1, 9, 14–16, 19, 33, 35, 73).

[13] R. Davis and Y. Shi. “The polyadenylation code: a unified model for the regulation of mRNA alternative polyadenylation”. *J. Zhejiang Univ. Sci. B* 15.5 (May 2014), pp. 429–437 (cit. on pp. 1, 29).

[14] L. Duret, F. Dorkeld, and C. Gautier. “Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression”. *Nucleic Acids Res.* 21.10 (May 1993), pp. 2315–2322 (cit. on p. 2).

[15] S. Kuersten and E. B. Goodwin. “The power of the 3’ UTR: translational control and development”. *Nat. Rev. Genet.* 4.8 (Aug. 2003), pp. 626–637 (cit. on p. 2).

[16] A. Jambhekar and J. L. Derisi. “*Cis*-acting determinants of asymmetric, cytoplasmic RNA transport”. *RNA* 13.5 (May 2007), pp. 625–642 (cit. on p. 2).

[17] A. G. Lau, H. A. Irier, J. Gu, D. Tian, L. Ku, G. Liu, M. Xia, B. Fritsch, J. Q. Zheng, R. Dingledine, B. Xu, B. Lu, and Y. Feng. “Distinct 3’UTRs differentially regulate activity-dependent translation of brain-derived neurotrophic factor (BDNF)”. *Proc. Natl. Acad. Sci. USA.* 107.36 (Sept. 2010), pp. 15945–15950 (cit. on p. 2).

[18] C. H. Jan, R. C. Friedman, J. G. Ruby, and D. P. Bartel. “Formation, regulation and evolution of Caenorhabditis elegans 3’UTRs”. *Nature* 469.7328 (Jan. 2011), pp. 97–101 (cit. on pp. 2, 81).

[19] S. Lianoglou, V. Garg, J. L. Yang, C. S. Leslie, and C. Mayr. “Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression”. *Genes Dev.* 27.21 (Nov. 2013), pp. 2380–2396 (cit. on pp. 2, 8, 16, 46, 81, 107).

[20] J. Y. Lee, I. Yeh, J. Y. Park, and B. Tian. “PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes”. *Nucleic Acids Res.* 35.Database issue (Jan. 2007), pp. D165–8 (cit. on pp. 2, 14, 15, 29).

[21] R. Sandberg, J. R. Neilson, A. Sarma, P. A. Sharp, and C. B. Burge. “Proliferating cells express mRNAs with shortened 3’ untranslated regions and fewer microRNA target sites”. *Science* 320.5883 (June 2008), pp. 1643–1647 (cit. on pp. 2, 4, 8, 9, 14, 29, 31, 59, 73).

[22] C. Mayr and D. P. Bartel. “Widespread shortening of 3’UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells”. *Cell* 138.4 (Aug. 2009), pp. 673–684 (cit. on pp. 2, 4, 5, 8, 31, 73, 75).

[23] C. P. Masamha, Z. Xia, J. Yang, T. R. Albrecht, M. Li, A.-B. Shyu, W. Li, and E. J. Wagner. "CFIm25 links alternative polyadenylation to glioblastoma tumour suppression". *Nature* 510.7505 (June 2014), pp. 412–416 (cit. on pp. 2, 8, 10, 14, 29, 46, 51, 53, 54, 59, 60, 62, 71, 73, 76, 121, 129, 136, 139).

[24] Y. Cheng, R. M. Miura, and B. Tian. "Prediction of mRNA polyadenylation sites by support vector machine". *Bioinformatics* 22.19 (Oct. 2006), pp. 2320–2325 (cit. on p. 2).

[25] N. J. Proudfoot. "Sequence analysis of the 3' non-coding regions of rabbit alpha- and beta-globin messenger RNAs". *J. Mol. Biol.* 107.4 (Nov. 1976), pp. 491–525 (cit. on p. 2).

[26] M. D. Sheets, S. C. Ogg, and M. P. Wickens. "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.* 18.19 (Oct. 1990), pp. 5799–5805 (cit. on pp. 2, 17, 74).

[27] E. Beaudoing, S. Freier, J. R. Wyatt, J. M. Claverie, and D. Gautheret. "Patterns of variant polyadenylation signal usage in human genes". *Genome Res.* 10.7 (July 2000), pp. 1001–1010 (cit. on pp. 2, 14–18).

[28] E. Wahle and U. Rüegsegger. "3'-End processing of pre-mRNA in eukaryotes". *FEMS Microbiol. Rev.* 23.3 (June 1999), pp. 277–295 (cit. on p. 2).

[29] J. Neve, R. Patel, Z. Wang, A. Louey, and A. M. Furger. "Cleavage and polyadenylation: Ending the message expands gene regulation". *RNA Biol.* 14.7 (July 2017), pp. 865–890 (cit. on pp. 2, 4, 9).

[30] F. Zhang, R. M. Denome, and C. N. Cole. "Fine-structure analysis of the processing and polyadenylation region of the herpes simplex virus type 1 thymidine kinase gene by using linker scanning, internal deletion, and insertion mutations". *Mol. Cell. Biol.* 6.12 (Dec. 1986), pp. 4611–4623 (cit. on p. 2).

[31] K. Venkataraman, K. M. Brown, and G. M. Gilmartin. "Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition". *Genes Dev.* 19.11 (June 2005), pp. 1315–1327 (cit. on pp. 2, 4, 19, 74).

[32] M. Legendre and D. Gautheret. "Sequence determinants in human polyadenylation site selection". *BMC Genomics* 4.1 (Feb. 2003), p. 7 (cit. on pp. 2, 3).

[33] G. Martin, A. R. Gruber, W. Keller, and M. Zavolan. "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.* 1.6 (June 2012), pp. 753–763 (cit. on pp. 2, 6, 14, 16, 19, 21, 23, 29, 40, 46, 53, 57, 59, 73, 107).

[34] B. R. Graveley, E. S. Fleming, and G. M. Gilmartin. "RNA structure is a critical determinant of poly(A) site recognition by cleavage and polyadenylation specificity factor". *Mol. Cell. Biol.* 16.9 (Sept. 1996), pp. 4942–4951 (cit. on p. 2).

[35] M. Khaladkar, J. Liu, D. Wen, J. T. L. Wang, and B. Tian. “Mining small RNA structure elements in untranslated regions of human and mouse mRNAs using structure-based alignment”. *BMC Genomics* 9 (Apr. 2008), p. 189 (cit. on p. 2).

[36] W. Li, J. Y. Park, D. Zheng, M. Hoque, G. Yehia, and B. Tian. “Alternative cleavage and polyadenylation in spermatogenesis connects chromatin regulation with post-transcriptional control”. *BMC Biol.* 14 (Jan. 2016), p. 6 (cit. on p. 2).

[37] H. Huang, H. Liu, and X. Sun. “Nucleosome distribution near the 3’ ends of genes in the human genome”. *Biosci. Biotechnol. Biochem.* 77.10 (Oct. 2013), pp. 2051–2055 (cit. on p. 2).

[38] J. Hu, C. S. Lutz, J. Wilusz, and B. Tian. “Bioinformatic identification of candidate *cis*-regulatory elements involved in human mRNA polyadenylation”. *RNA* 11.10 (Oct. 2005), pp. 1485–1493 (cit. on p. 3).

[39] D. C. Di Giammartino, K. Nishida, and J. L. Manley. “Mechanisms and consequences of alternative polyadenylation”. *Mol. Cell* 43.6 (Sept. 2011), pp. 853–866 (cit. on p. 3).

[40] Y. Shi, D. C. Di Giammartino, D. Taylor, A. Sarkeshik, W. J. Rice, J. R. Yates 3rd, J. Frank, and J. L. Manley. “Molecular architecture of the human pre-mRNA 3’ processing complex”. *Mol. Cell* 33.3 (Feb. 2009), pp. 365–376 (cit. on p. 3).

[41] C. R. Mandel, Y. Bai, and L. Tong. “Protein factors in pre-mRNA 3’-end processing”. *Cell. Mol. Life Sci.* 65.7-8 (Apr. 2008), pp. 1099–1122 (cit. on p. 3).

[42] W. Keller, S. Bienroth, K. M. Lang, and G. Christofori. “Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3’ processing signal AAUAAA”. *EMBO J.* 10.13 (Dec. 1991), pp. 4241–4249 (cit. on p. 4).

[43] L. Schönemann, U. Kühn, G. Martin, P. Schäfer, A. R. Gruber, W. Keller, M. Zavolan, and E. Wahle. “Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33”. *Genes Dev.* 28.21 (Nov. 2014), pp. 2381–2393 (cit. on pp. 4, 14, 46).

[44] S. L. Chan, I. Huppertz, C. Yao, L. Weng, J. J. Moresco, J. R. Yates 3rd, J. Ule, J. L. Manley, and Y. Shi. “CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3’ processing”. *Genes Dev.* 28.21 (Nov. 2014), pp. 2370–2380 (cit. on pp. 4, 14, 46).

[45] Y. Sun, Y. Zhang, K. Hamilton, J. L. Manley, Y. Shi, T. Walz, and L. Tong. “Molecular basis for the recognition of the human AAUAAA polyadenylation signal”. *Proc. Natl. Acad. Sci. USA.* 115.7 (Feb. 2018), E1419–E1428 (cit. on pp. 4, 74).

[46] M. Clerici, M. Faini, R. Aebersold, and M. Jinek. “Structural insights into the assembly and polyA signal recognition mechanism of the human CPSF complex”. *Elife* 6 (Dec. 2017) (cit. on pp. 4, 74).

[47] M. Clerici, M. Faini, L. M. Muckenfuss, R. Aebersold, and M. Jinek. “Structural basis of AAUAAA polyadenylation signal recognition by the human CPSF complex”. *Nat. Struct. Mol. Biol.* 25.2 (Feb. 2018), pp. 135–138 (cit. on pp. 4, 74).

[48] C. R. Mandel, S. Kaneko, H. Zhang, D. Gebauer, V. Vethantham, J. L. Manley, and L. Tong. “Polyadenylation factor CPSF-73 is the pre-mRNA 3’-end-processing endonuclease”. *Nature* 444.7121 (Dec. 2006), pp. 953–956 (cit. on p. 4).

[49] C. Yao, J. Biesinger, J. Wan, L. Weng, Y. Xing, X. Xie, and Y. Shi. “Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation”. *Proc. Natl. Acad. Sci. USA.* 109.46 (Nov. 2012), pp. 18773–18778 (cit. on pp. 4, 6, 46, 82, 106).

[50] R. S. Laishram. “Poly(A) polymerase (PAP) diversity in gene expression–star-PAP vs canonical PAP”. *FEBS Lett.* 588.14 (June 2014), pp. 2185–2197 (cit. on p. 4).

[51] A. R. Gruber, G. Martin, W. Keller, and M. Zavolan. “Means to an end: mechanisms of alternative polyadenylation of messenger RNA precursors”. *Wiley Interdiscip. Rev. RNA* 5.2 (Mar. 2014), pp. 183–196 (cit. on pp. 4, 45, 46, 59).

[52] B. Tian and J. L. Manley. “Alternative polyadenylation of mRNA precursors”. *Nat. Rev. Mol. Cell Biol.* 18.1 (Jan. 2017), pp. 18–30 (cit. on pp. 4, 7, 46, 77).

[53] M. Hoque, Z. Ji, D. Zheng, W. Luo, W. Li, B. You, J. Y. Park, G. Yehia, and B. Tian. “Analysis of alternative cleavage and polyadenylation by 3’ region extraction and deep sequencing”. *Nat. Methods* 10.2 (Feb. 2013), pp. 133–139 (cit. on pp. 4, 14, 73, 82).

[54] I. Gupta, S. Clauder-Münster, B. Klaus, A. I. Järvelin, R. S. Aiyar, V. Benes, S. Wilkening, W. Huber, V. Pelechano, and L. M. Steinmetz. “Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions”. *Mol. Syst. Biol.* 10 (Feb. 2014), p. 719 (cit. on pp. 4, 31).

[55] N. Spies, C. B. Burge, and D. P. Bartel. “3’ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts”. *Genome Res.* 23.12 (Dec. 2013), pp. 2078–2090 (cit. on pp. 4–6, 31, 81, 108).

[56] A. R. Gruber, G. Martin, P. Müller, A. Schmidt, A. J. Gruber, R. Gumienny, N. Mittal, R. Jayachandran, J. Pieters, W. Keller, E. van Nimwegen, and M. Zavolan. “Global 3’ UTR shortening has a limited effect on protein abundance in proliferating T cells”. *Nat. Commun.* 5 (Nov. 2014), p. 5465 (cit. on pp. 4, 6, 18, 31, 38, 40, 82, 108, 115).

[57] H. Guo, N. T. Ingolia, J. S. Weissman, and D. P. Bartel. “Mammalian microRNAs predominantly act to decrease target mRNA levels”. *Nature* 466.7308 (Aug. 2010), pp. 835–840 (cit. on p. 4).

[58] J.-W. Nam, O. S. Rissland, D. Koppstein, C. Abreu-Goodger, C. H. Jan, V. Agarwal, M. A. Yildirim, A. Rodriguez, and D. P. Bartel. "Global analyses of the effect of different cellular contexts on microRNA targeting". *Mol. Cell* 53.6 (Mar. 2014), pp. 1031–1043 (cit. on pp. 4, 14, 81, 107–109).

[59] J. Neve and A. Furger. "Alternative polyadenylation: less than meets the eye?" *Biochem. Soc. Trans.* 42.4 (Aug. 2014), pp. 1190–1195 (cit. on p. 5).

[60] D. Gaidatzis, E. van Nimwegen, J. Hausser, and M. Zavolan. "Inference of miRNA targets using evolutionary conservation and pathway analysis". *BMC Bioinformatics* 8 (Mar. 2007), p. 69 (cit. on p. 5).

[61] Y. Hoffman, D. R. Bublik, A. P. Ugalde, R. Elkon, T. Biniashvili, R. Agami, M. Oren, and Y. Pilpel. "3'UTR Shortening Potentiates MicroRNA-Based Repression of Pro-differentiation Genes in Proliferating Human Cells". *PLoS Genet.* 12.2 (Feb. 2016), e1005879 (cit. on p. 5).

[62] P. Oikonomou, H. Goodarzi, and S. Tavazoie. "Systematic identification of regulatory elements in conserved 3' UTRs of human transcripts". *Cell Rep.* 7.1 (Apr. 2014), pp. 281–292 (cit. on p. 5).

[63] C. Y. Chen and A. B. Shyu. "AU-rich elements: characterization and importance in mRNA degradation". *Trends Biochem. Sci.* 20.11 (Nov. 1995), pp. 465–470 (cit. on p. 5).

[64] X. C. Fan and J. A. Steitz. "Overexpression of HuR, a nuclear-cytoplasmic shuttling protein, increases the in vivo stability of ARE-containing mRNAs". *EMBO J.* 17.12 (June 1998), pp. 3448–3460 (cit. on p. 5).

[65] M. E. Lazarov, M. M. Martin, B. M. Willardson, and T. S. Elton. "Human phosducin-like protein (hPhLP) messenger RNA stability is regulated by *cis*-acting instability elements present in the 3'-untranslated region". *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* 1446.3 (Sept. 1999), pp. 253–264 (cit. on p. 5).

[66] Y. Audic and R. S. Hartley. "Post-transcriptional regulation in cancer". *Biol. Cell* 96.7 (Sept. 2004), pp. 479–498 (cit. on p. 5).

[67] L.-L. Chen, J. N. DeCerbo, and G. G. Carmichael. "Alu element-mediated gene silencing". *EMBO J.* 27.12 (June 2008), pp. 1694–1705 (cit. on p. 5).

[68] J. Neve, K. Burger, W. Li, M. Hoque, R. Patel, B. Tian, M. Gullerova, and A. Furger. "Subcellular RNA profiling links splicing and nuclear DICER1 to alternative cleavage and polyadenylation". *Genome Res.* 26.1 (Jan. 2016), pp. 24–35 (cit. on pp. 5, 7).

[69] Y. Lubelsky and I. Ulitsky. "Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells". *Nature* (Jan. 2018) (cit. on p. 5).

[70] J. J. An, K. Gharami, G.-Y. Liao, N. H. Woo, A. G. Lau, F. Vanevski, E. R. Torre, K. R. Jones, Y. Feng, B. Lu, and B. Xu. “Distinct role of long 3’ UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons”. *Cell* 134.1 (July 2008), pp. 175–187 (cit. on p. 5).

[71] A. R. Buxbaum, G. Haimovich, and R. H. Singer. “In the right place at the right time: visualizing and understanding mRNA localization”. *Nat. Rev. Mol. Cell Biol.* 16.2 (Feb. 2015), pp. 95–109 (cit. on p. 5).

[72] B. D. Berkovits and C. Mayr. “Alternative 3’ UTRs act as scaffolds to regulate membrane protein localization”. *Nature* 522.7556 (June 2015), pp. 363–367 (cit. on pp. 5, 14, 26, 31, 32, 73, 75).

[73] E. Szostak and F. Gebauer. “Translational control by 3’-UTR-binding proteins”. *Brief. Funct. Genomics* 12.1 (Jan. 2013), pp. 58–65 (cit. on p. 5).

[74] S. N. Floor and J. A. Doudna. “Tunable protein synthesis by transcript isoforms in human cells”. *Elife* 5 (Jan. 2016) (cit. on p. 6).

[75] J. D. Blair, D. Hockemeyer, J. A. Doudna, H. S. Bateup, and S. N. Floor. “Widespread Translational Remodeling during Human Neuronal Differentiation”. *Cell Rep.* 21.7 (Nov. 2017), pp. 2005–2016 (cit. on p. 6).

[76] C. S. Lutz. “Alternative polyadenylation: a twist on mRNA 3’ end formation”. *ACS Chem. Biol.* 3.10 (Oct. 2008), pp. 609–617 (cit. on p. 6).

[77] T. Kubo, T. Wada, Y. Yamaguchi, A. Shimizu, and H. Handa. “Knock-down of 25 kDa subunit of cleavage factor Im in Hela cells alters alternative polyadenylation within 3’-UTRs”. *Nucleic Acids Res.* 34.21 (Nov. 2006), pp. 6264–6271 (cit. on pp. 6, 46, 53).

[78] M. Jenal, R. Elkon, F. Loayza-Puch, G. van Haaften, U. Kühn, F. M. Menzies, J. A. F. Oude Vrielink, A. J. Bos, J. Drost, K. Rooijers, D. C. Rubinsztein, and R. Agami. “The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites”. *Cell* 149.3 (Apr. 2012), pp. 538–553 (cit. on pp. 6, 8, 14, 29, 46).

[79] W. Li, B. You, M. Hoque, D. Zheng, W. Luo, Z. Ji, J. Y. Park, S. I. Gunderson, A. Kalsotra, J. L. Manley, and B. Tian. “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.* 11.4 (Apr. 2015), e1005166 (cit. on pp. 6, 20, 46, 53, 59, 82, 109, 110).

[80] F. W. Alt, A. L. Bothwell, M. Knapp, E. Siden, E. Mather, M. Koshland, and D. Baltimore. “Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3’ ends”. *Cell* 20.2 (June 1980), pp. 293–301 (cit. on pp. 6, 8, 73).

[81] Y. Takagaki, R. L. Seipelt, M. L. Peterson, and J. L. Manley. “The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation”. *Cell* 87.5 (Nov. 1996), pp. 941–952 (cit. on p. 6).

[82] D. Kaida, M. G. Berg, I. Younis, M. Kasim, L. N. Singh, L. Wan, and G. Dreyfuss. “U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation”. *Nature* 468.7324 (Dec. 2010), pp. 664–668 (cit. on pp. 6, 14, 31, 46, 77, 78).

[83] M. G. Berg, L. N. Singh, I. Younis, Q. Liu, A. M. Pinto, D. Kaida, Z. Zhang, S. Cho, S. Sherrill-Mix, L. Wan, and G. Dreyfuss. “U1 snRNP determines mRNA length and regulates isoform expression”. *Cell* 150.1 (July 2012), pp. 53–64 (cit. on pp. 6, 14, 29, 31, 46).

[84] R. Batra, K. Charizanis, M. Manchanda, A. Mohan, M. Li, D. J. Finn, M. Goodwin, C. Zhang, K. Sobczak, C. A. Thornton, and M. S. Swanson. “Loss of MBNL leads to disruption of developmentally regulated alternative polyadenylation in RNA-mediated disease”. *Mol. Cell* 56.2 (Oct. 2014), pp. 311–322 (cit. on pp. 6, 83, 109).

[85] X. Ji, J. Wan, M. Vishnu, Y. Xing, and S. A. Liebhaber. “αCP Poly(C) binding proteins act as global regulators of alternative polyadenylation”. *Mol. Cell. Biol.* 33.13 (July 2013), pp. 2560–2573 (cit. on pp. 6, 14, 15, 46, 48, 50, 71, 82, 106, 138).

[86] D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, J. C. Darnell, and R. B. Darnell. “HITS-CLIP yields genome-wide insights into brain alternative RNA processing”. *Nature* 456.7221 (Nov. 2008), pp. 464–469 (cit. on pp. 6, 7, 46, 77).

[87] J. König, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D. J. Turner, N. M. Luscombe, and J. Ule. “iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution”. *Nat. Struct. Mol. Biol.* 17.7 (July 2010), pp. 909–915 (cit. on pp. 6, 20, 23, 25, 31, 46, 60, 74, 77).

[88] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. “Alternative isoform regulation in human tissue transcriptomes”. *Nature* 456.7221 (Nov. 2008), pp. 470–476 (cit. on pp. 7, 10).

[89] E. J. Wagner and M. A. Garcia-Blanco. “Polypyrimidine tract binding protein antagonizes exon definition”. *Mol. Cell. Biol.* 21.10 (May 2001), pp. 3281–3288 (cit. on p. 7).

[90] P. Castelo-Branco, A. Furger, M. Wollerton, C. Smith, A. Moreira, and N. Proudfoot. “Polypyrimidine tract binding protein modulates efficiency of polyadenylation”. *Mol. Cell. Biol.* 24.10 (May 2004), pp. 4174–4183 (cit. on pp. 7, 15, 59, 76).

[91] F.-A. Bava, C. Eliscovich, P. G. Ferreira, B. Miñana, C. Ben-Dov, R. Guigó, J. Valcárcel, and R. Méndez. “CPEB1 coordinates alternative 3’-UTR formation with translational regulation”. *Nature* 495.7439 (Mar. 2013), pp. 121–125 (cit. on p. 7).

[92] D. L. Bentley. “Coupling mRNA processing with transcription in time and space”. *Nat. Rev. Genet.* 15.3 (Mar. 2014), pp. 163–175 (cit. on pp. 7, 60, 78).

[93] P. A. B. Pinto, T. Henriques, M. O. Freitas, T. Martins, R. G. Domingues, P. S. Wyrzykowska, P. A. Coelho, A. M. Carmo, C. E. Sunkel, N. J. Proudfoot, and A. Moreira. “RNA polymerase II kinetics in polo polyadenylation signal selection”. *EMBO J.* 30.12 (May 2011), pp. 2431–2444 (cit. on p. 7).

[94] V. Pelechano, W. Wei, and L. M. Steinmetz. “Extensive transcriptional heterogeneity revealed by isoform profiling”. *Nature* 497.7447 (May 2013), pp. 127–131 (cit. on p. 8).

[95] C. Mayr. “Evolution and Biological Roles of Alternative 3’UTRs”. *Trends Cell Biol.* 26.3 (Mar. 2016), pp. 227–237 (cit. on p. 8).

[96] W. O. Miles, A. Lembo, A. Volorio, E. Brachtel, B. Tian, D. Sgroi, P. Provero, and N. Dyson. “Alternative Polyadenylation in Triple-Negative Breast Tumors Allows NRAS and c-JUN to Bypass PUMILIO Posttranscriptional Regulation”. *Cancer Res.* 76.24 (Dec. 2016), pp. 7231–7241 (cit. on pp. 8, 77).

[97] Y. Lin, Z. Li, F. Ozsolak, S. W. Kim, G. Arango-Argoty, T. T. Liu, S. A. Tenenbaum, T. Bailey, A. P. Monaghan, P. M. Milos, and B. John. “An in-depth map of polyadenylation sites in cancer”. *Nucleic Acids Res.* 40.17 (Sept. 2012), pp. 8460–8471 (cit. on pp. 8, 14, 15).

[98] Z. Xia, L. A. Donehower, T. A. Cooper, J. R. Neilson, D. A. Wheeler, E. J. Wagner, and W. Li. “Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3’-UTR landscape across seven tumour types”. *Nat. Commun.* 5 (Nov. 2014), p. 5274 (cit. on pp. 8, 29, 57, 75, 76).

[99] Y. Xiang, Y. Ye, Y. Lou, Y. Yang, C. Cai, Z. Zhang, T. Mills, N.-Y. Chen, Y. Kim, F. Muge Ozguc, L. Diao, H. Karmouty-Quintana, Y. Xia, R. E. Kellems, Z. Chen, M. R. Blackburn, S.-H. Yoo, A.-B. Shyu, G. B. Mills, and L. Han. “Comprehensive Characterization of Alternative Polyadenylation in Human Cancer”. *J. Natl. Cancer Inst.* (Nov. 2017) (cit. on pp. 8, 76).

[100] Y. Fu, Y. Sun, Y. Li, J. Li, X. Rao, C. Chen, and A. Xu. “Differential genome-wide profiling of tandem 3’ UTRs among human breast cancer and normal cells by high-throughput sequencing”. *Genome Res.* 21.5 (May 2011), pp. 741–747 (cit. on pp. 8, 83, 106).

[101] Z. Xue, R. L. Warren, E. A. Gibb, D. MacMillan, J. Wong, R. Chiu, S. A. Hammond, C. A. Ennis, A. Hahn, S. Reynolds, and I. Birol. “Pan-cancer analysis reveals complex tumor-specific alternative polyadenylation”. July 2017 (cit. on p. 8).

[102] P. Singh, T. L. Alley, S. M. Wright, S. Kamdar, W. Schott, R. Y. Wilpan, K. D. Mills, and J. H. Graber. “Global changes in processing of mRNA 3’ untranslated regions characterize clinically distinct cancer subtypes”. *Cancer Res.* 69.24 (Dec. 2009), pp. 9422–9430 (cit. on pp. 8, 77).

[103] V. A. Gennarino, C. E. Alcott, C.-A. Chen, A. Chaudhury, M. A. Gillentine, J. A. Rosenfeld, S. Parikh, J. W. Wheless, E. R. Roeder, D. D. G. Horovitz, E. K. Roney, J. L. Smith, S. W. Cheung, W. Li, J. R. Neilson, C. P. Schaaf, and H. Y. Zoghbi. “NUDT21-spanning CNVs lead to neuropsychiatric disease and altered MeCP2 abundance via alternative polyadenylation”. *Elife* 4 (Aug. 2015) (cit. on pp. 9, 46, 73).

[104] A. Wiestner et al. “Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival”. *Blood* 109.11 (June 2007), pp. 4599–4606 (cit. on p. 9).

[105] D. Gautheret, O. Poirot, F. Lopez, S. Audic, and J. M. Claverie. “Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering”. *Genome Res.* 8.5 (May 1998), pp. 524–530 (cit. on p. 9).

[106] S. W. Flavell, T.-K. Kim, J. M. Gray, D. A. Harmin, M. Hemberg, E. J. Hong, E. Markenscoff-Papadimitriou, D. M. Bear, and M. E. Greenberg. “Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection”. *Neuron* 60.6 (Dec. 2008), pp. 1022–1038 (cit. on p. 9).

[107] Z. Ji, W. Luo, W. Li, M. Hoque, Z. Pan, Y. Zhao, and B. Tian. “Transcriptional activity regulates alternative cleavage and polyadenylation”. *Mol. Syst. Biol.* 7 (Sept. 2011), p. 534 (cit. on p. 10).

[108] I. Birol, A. Raymond, R. Chiu, K. M. Nip, S. D. Jackman, M. Kreitzman, T. R. Docking, C. A. Ennis, A. G. Robertson, and A. Karsan. “Kleat: cleavage site analysis of transcriptomes”. *Pac. Symp. Biocomput.* (2015), pp. 347–358 (cit. on p. 10).

[109] C. Erdman and J. W. Emerson. “A fast Bayesian change point analysis for the segmentation of microarray data”. *Bioinformatics* 24.19 (Oct. 2008), pp. 2143–2148 (cit. on p. 10).

[110] J. Lu and P. R. Bushel. “Dynamic expression of 3’ UTRs revealed by Poisson hidden Markov modeling of RNA-Seq: implications in gene expression profiling”. *Gene* 527.2 (Sept. 2013), pp. 616–623 (cit. on pp. 10, 75).

[111] W. Wang, Z. Wei, and H. Li. “A change-point model for identifying 3’UTR switching by next-generation RNA sequencing”. *Bioinformatics* 30.15 (Aug. 2014), pp. 2162–2170 (cit. on pp. 10, 75, 128).

[112] S. Shenker, P. Miura, P. Sanfilippo, and E. C. Lai. “IsoSCM: improved and alternative 3’ UTR annotation using multiple change-point inference”. *RNA* 21.1 (Jan. 2015), pp. 14–27 (cit. on pp. 10, 75).

[113] J. Lee, A. Hever, D. Willhite, A. Zlotnik, and P. Hevezi. “Effects of RNA degradation on gene expression analysis of human postmortem tissues”. *FASEB J.* 19.10 (Aug. 2005), pp. 1356–1358 (cit. on p. 11).

[114] I. Gallego Romero, A. A. Pai, J. Tung, and Y. Gilad. “RNA-seq: impact of RNA degradation on transcript quantification”. *BMC Biol.* 12 (May 2014), p. 42 (cit. on pp. 11, 75).

[115] H. Feng, X. Zhang, and C. Zhang. “mRIN for direct assessment of genome-wide and gene-specific mRNA integrity from large-scale RNA-sequencing data”. *Nat. Commun.* 6 (Aug. 2015), p. 7816 (cit. on pp. 11, 75).

[116] L. Wang, J. Nie, H. Sicotte, Y. Li, J. E. Eckel-Passow, S. Dasari, P. T. Vedell, P. Barman, L. Wang, R. Weinshiboum, J. Jen, H. Huang, M. Kohli, and J.-P. A. Kocher. “Measure transcript integrity using RNA-seq data”. *BMC Bioinformatics* 17 (Feb. 2016), p. 58 (cit. on pp. 11, 64, 66, 75, 124, 135).

[117] A. Schroeder, O. Mueller, S. Stocker, R. Salowsky, M. Leiber, M. Gassmann, S. Lightfoot, W. Menzel, M. Granzow, and T. Ragg. “The RIN: an RNA integrity number for assigning integrity values to RNA measurements”. *BMC Mol. Biol.* 7 (Jan. 2006), p. 3 (cit. on pp. 11, 64).

[118] A. J. Gruber, R. Schmidt, A. R. Gruber, G. Martin, S. Ghosh, M. Belmadani, W. Keller, and M. Zavolan. “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”. *Genome Res.* 26.8 (Aug. 2016), pp. 1145–1159 (cit. on pp. 13, 46, 51, 52, 62, 63, 67, 68, 130, 138).

[119] A. O. Subtelny, S. W. Eichhorn, G. R. Chen, H. Sive, and D. P. Bartel. “Poly(A)-tail profiling reveals an embryonic switch in translational control”. *Nature* 508.7494 (Apr. 2014), pp. 66–71 (cit. on p. 14).

[120] A. R. Gruber, G. Martin, W. Keller, and M. Zavolan. “Cleavage factor Im is a key regulator of 3’ UTR length”. *RNA Biol.* 9.12 (Dec. 2012), pp. 1405–1412 (cit. on pp. 14, 21, 29, 46, 53, 82, 107).

[121] H. Zhang, J. Hu, M. Recce, and B. Tian. “PolyA_DB: a database for mammalian mRNA polyadenylation”. *Nucleic Acids Res.* 33.Database issue (Jan. 2005), pp. D116–20 (cit. on p. 14).

[122] J. Kawai et al. “Functional annotation of a full-length mouse cDNA collection”. *Nature* 409.6821 (Feb. 2001), pp. 685–690 (cit. on p. 14).

[123] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge. “Analysis and design of RNA sequencing experiments for identifying isoform regulation”. *Nat. Methods* 7.12 (Dec. 2010), pp. 1009–1015 (cit. on p. 14).

[124] M. de Hoon and Y. Hayashizaki. “Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference”. *Biotechniques* 44.5 (Apr. 2008), pp. 627–8, 630, 632 (cit. on p. 14).

[125] F. Ozsolak, A. R. Platt, D. R. Jones, J. G. Reifenberger, L. E. Sass, P. McInerney, J. F. Thompson, J. Bowers, M. Jarosz, and P. M. Milos. “Direct RNA sequencing”. *Nature* 461.7265 (Oct. 2009), pp. 814–818 (cit. on p. 14).

[126] A. H. Beck, Z. Weng, D. M. Witten, S. Zhu, J. W. Foley, P. Lacroute, C. L. Smith, R. Tibshirani, M. van de Rijn, A. Sidow, and R. B. West. “3’-end sequencing for expression quantification (3SEQ) from archival tumor samples”. *PLoS One* 5.1 (Jan. 2010), e8768 (cit. on p. 14).

[127] L. You, J. Wu, Y. Feng, Y. Fu, Y. Guo, L. Long, H. Zhang, Y. Luan, P. Tian, L. Chen, G. Huang, S. Huang, Y. Li, J. Li, C. Chen, Y. Zhang, S. Chen, and A. Xu. “APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals”. *Nucleic Acids Res.* 43.Database issue (Jan. 2015), pp. D59–67 (cit. on pp. 14, 15, 19, 29, 30, 33, 38, 83, 97, 106–108, 110).

[128] I. Ulitsky, A. Shkumatava, C. H. Jan, A. O. Subtelny, D. Koppstein, G. W. Bell, H. Sive, and D. P. Bartel. “Extensive alternative polyadenylation during zebrafish development”. *Genome Res.* 22.10 (Oct. 2012), pp. 2054–2066 (cit. on pp. 14, 73).

[129] Y. Li, Y. Sun, Y. Fu, M. Li, G. Huang, C. Zhang, J. Liang, S. Huang, G. Shen, S. Yuan, L. Chen, S. Chen, and A. Xu. “Dynamic landscape of tandem 3’ UTRs during zebrafish development”. *Genome Res.* 22.10 (Oct. 2012), pp. 1899–1906 (cit. on p. 14).

[130] A. E. Almada, X. Wu, A. J. Kriz, C. B. Burge, and P. A. Sharp. “Promoter directionality is controlled by U1 snRNP and polyadenylation signals”. *Nature* 499.7458 (July 2013), pp. 360–363 (cit. on pp. 14, 31, 81, 108).

[131] N. J. Proudfoot and G. G. Brownlee. “3’ non-coding region sequences in eukaryotic messenger RNA”. *Nature* 263.5574 (Sept. 1976), pp. 211–214 (cit. on p. 14).

[132] B. Tian and J. H. Graber. “Signals for pre-mRNA cleavage and polyadenylation”. *Wiley Interdiscip. Rev. RNA* 3.3 (May 2012), pp. 385–396 (cit. on pp. 15, 20, 30).

[133] J. H. Graber, C. R. Cantor, S. C. Mohr, and T. F. Smith. “In silico detection of control signals: mRNA 3’-end-processing sequences in diverse species”. *Proc. Natl. Acad. Sci. USA.* 96.24 (Nov. 1999), pp. 14055–14060 (cit. on p. 15).

[134] C. C. MacDonald and J.-L. Redondo. “Reexamining the polyadenylation signal: were we wrong about AAUAAA?” *Mol. Cell. Endocrinol.* 190.1-2 (Apr. 2002), pp. 1–8 (cit. on p. 15).

[135] J. Wilusz, D. I. Feig, and T. Shenk. “The C proteins of heterogeneous nuclear ribonucleoprotein complexes interact with RNA sequences downstream of polyadenylation cleavage sites”. *Mol. Cell. Biol.* 8.10 (Oct. 1988), pp. 4477–4483 (cit. on p. 15).

[136] X. Zhao, D. Oberg, M. Rush, J. Fay, H. Lambkin, and S. Schwartz. “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.* 79.7 (Apr. 2005), pp. 4270–4288 (cit. on p. 15).

[137] S. A. Alkan, K. Martincic, and C. Milcarek. “The hnRNPs F and H2 bind to similar sequences to influence gene expression”. *Biochem. J* 393.Pt 1 (Jan. 2006), pp. 361–371 (cit. on p. 15).

[138] G. K. Arhin, M. Boots, P. S. Bagga, C. Milcarek, and J. Wilusz. “Downstream sequence elements with different affinities for the hnRNP H/H’ protein influence the processing efficiency of mammalian polyadenylation signals”. *Nucleic Acids Res.* 30.8 (Apr. 2002), pp. 1842–1850 (cit. on p. 15).

[139] S. Millevoi, A. Decorsière, C. Loulergue, J. Iacovoni, S. Bernat, M. Antoniou, and S. Vagner. “A physical and functional link between splicing factors promotes pre-mRNA 3’ end processing”. *Nucleic Acids Res.* 37.14 (Aug. 2009), pp. 4672–4683 (cit. on p. 15).

[140] FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. “A promoter-level mammalian expression atlas”. *Nature* 507.7493 (Mar. 2014), pp. 462–470 (cit. on p. 16).

[141] F. Ozsolak, P. Kapranov, S. Foissac, S. W. Kim, E. Fishilevich, A. P. Monaghan, B. John, and P. M. Milos. “Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation”. *Cell* 143.6 (Dec. 2010), pp. 1018–1029 (cit. on pp. 19, 73).

[142] I. Kaufmann, G. Martin, A. Friedlein, H. Langen, and W. Keller. “Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase”. *EMBO J.* 23.3 (Feb. 2004), pp. 616–626 (cit. on p. 20).

[143] B. Lackford, C. Yao, G. M. Charles, L. Weng, X. Zheng, E.-A. Choi, X. Xie, J. Wan, Y. Xing, J. M. Freudenberg, P. Yang, R. Jothi, G. Hu, and Y. Shi. “Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal”. *EMBO J.* 33.8 (Apr. 2014), pp. 878–889 (cit. on pp. 20, 82, 106, 108).

[144] D. Ray et al. “A compendium of RNA-binding motifs for decoding gene regulation”. *Nature* 499.7457 (July 2013), pp. 172–177 (cit. on pp. 20, 23, 25, 31).

[145] K. Zarnack, J. König, M. Tajnik, I. Martincorena, S. Eustermann, I. Stévant, A. Reyes, S. Anders, N. M. Luscombe, and J. Ule. “Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements”. *Cell* 152.3 (Jan. 2013), pp. 453–466 (cit. on pp. 20, 25, 31, 46, 49).

[146] Z. Cieniková, F. F. Damberger, J. Hall, F. H.-T. Allain, and C. Maris. “Structural and mechanistic insights into poly(uridine) tract recognition by the hnRNP C RNA recognition motif”. *J. Am. Chem. Soc.* 136.41 (Oct. 2014), pp. 14536–14544 (cit. on pp. 20, 25, 31).

[147] N. Liu, Q. Dai, G. Zheng, C. He, M. Parisien, and T. Pan. “N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions”. *Nature* 518.7540 (Feb. 2015), pp. 560–564 (cit. on pp. 20, 23, 25, 31, 41, 50, 51, 62, 68, 71, 94, 121, 139).

[148] M. Tajnik, A. Vigilante, S. Braun, H. Hänel, N. M. Luscombe, J. Ule, K. Zarnack, and J. König. “Intergenic Alu exonisation facilitates the evolution of tissue-specific transcript ends”. *Nucleic Acids Res.* 43.21 (Dec. 2015), pp. 10492–10505 (cit. on pp. 20, 31).

[149] C. M. Brennan and J. A. Steitz. “HuR and mRNA stability”. *Cell. Mol. Life Sci.* 58.2 (Feb. 2001), pp. 266–277 (cit. on p. 25).

[150] S. Kishore, L. Jaskiewicz, L. Burger, J. Hausser, M. Khorshid, and M. Zavolan. “A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins”. *Nat. Methods* 8.7 (May 2011), pp. 559–564 (cit. on pp. 25, 38, 41).

[151] J. P. ten Klooster, I. v. Leeuwen, N. Scheres, E. C. Anthony, and P. L. Hordijk. “Rac1-induced cell migration requires membrane recruitment of the nuclear oncogene SET”. *EMBO J.* 26.2 (Jan. 2007), pp. 336–345 (cit. on p. 26).

[152] J. Harrow et al. “GENCODE: the reference human genome annotation for The ENCODE Project”. *Genome Res.* 22.9 (Sept. 2012), pp. 1760–1774 (cit. on pp. 28, 32, 63).

[153] Z. Ji and B. Tian. “Reprogramming of 3’ untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types”. *PLoS One* 4.12 (Dec. 2009), e8419 (cit. on p. 29).

[154] Z. Ji, J. Y. Lee, Z. Pan, B. Jiang, and B. Tian. “Progressive lengthening of 3’ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development”. *Proc. Natl. Acad. Sci. USA.* 106.17 (Apr. 2009), pp. 7028–7033 (cit. on pp. 29, 59, 73).

[155] P. Miura, S. Shenker, C. Andreu-Agullo, J. O. Westholm, and E. C. Lai. “Widespread and extensive lengthening of 3’ UTRs in the mammalian brain”. *Genome Res.* 23.5 (May 2013), pp. 812–825 (cit. on p. 29).

[156] D. C. Di Giammartino, W. Li, K. Ogami, J. J. Yashinskie, M. Hoque, B. Tian, and J. L. Manley. “RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3’ UTRs”. *Genes Dev.* 28.20 (Oct. 2014), pp. 2248–2260 (cit. on p. 29).

[157] E. de Klerk, J. T. den Dunnen, and P. A. C. ’t Hoen. “RNA sequencing: from tag-based profiling to resolving complete transcript structure”. *Cell. Mol. Life Sci.* 71.18 (Sept. 2014), pp. 3537–3551 (cit. on p. 29).

[158] Z. Wu, X. Liu, L. Liu, H. Deng, J. Zhang, Q. Xu, B. Cen, and A. Ji. “Regulation of lncRNA expression”. *Cell. Mol. Biol. Lett.* 19.4 (Dec. 2014), pp. 561–575 (cit. on p. 30).

[159] M. J. Hangauer, I. W. Vaughn, and M. T. McManus. “Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs”. *PLoS Genet.* 9.6 (June 2013), e1003569 (cit. on p. 30).

[160] A. McCloskey, I. Taniguchi, K. Shinmyozu, and M. Ohno. “hnRNP C tetramer measures RNA length to classify RNA polymerase II transcripts for export”. *Science* 335.6076 (Mar. 2012), pp. 1643–1646 (cit. on p. 31).

[161] A. L. Beyer, M. E. Christensen, B. W. Walker, and W. M. LeStourgeon. “Identification and characterization of the packaging proteins of core 40S hnRNP particles”. *Cell* 11.1 (May 1977), pp. 127–138 (cit. on p. 31).

[162] Y. D. Choi and G. Dreyfuss. “Isolation of the heterogeneous nuclear RNA-ribonucleoprotein complex (hnRNP): a unique supramolecular assembly”. *Proc. Natl. Acad. Sci. USA.* 81.23 (Dec. 1984), pp. 7471–7475 (cit. on p. 31).

[163] S. R. Whitson, W. M. LeStourgeon, and A. M. Krezel. “Solution structure of the symmetric coiled coil tetramer formed by the oligomerization domain of hnRNP C: implications for biological function”. *J. Mol. Biol.* 350.2 (July 2005), pp. 319–337 (cit. on p. 31).

[164] M. Görlach, C. G. Burd, and G. Dreyfuss. “The determinants of RNA-binding specificity of the heterogeneous nuclear ribonucleoprotein C proteins”. *J. Biol. Chem.* 269.37 (Sept. 1994), pp. 23074–23078 (cit. on p. 31).

[165] Y. Shi. “Alternative polyadenylation: new insights from global analyses”. *RNA* 18.12 (Dec. 2012), pp. 2105–2117 (cit. on pp. 31, 35).

[166] P. Flicek et al. “Ensembl 2013”. *Nucleic Acids Res.* 41.Database issue (Jan. 2013), pp. D48–55 (cit. on p. 32).

[167] L. R. Meyer et al. “The UCSC Genome Browser database: extensions and updates 2013”. *Nucleic Acids Res.* 41.Database issue (Jan. 2013), pp. D64–9 (cit. on p. 32).

[168] S. Hoffmann, C. Otto, S. Kurtz, C. M. Sharma, P. Khaitovich, J. Vogel, P. F. Stadler, and J. Hackermüller. “Fast mapping of short sequences with mismatches, insertions and deletions using index structures”. *PLoS Comput. Biol.* 5.9 (Sept. 2009), e1000502 (cit. on pp. 32, 62).

[169] R Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. 2014 (cit. on p. 34).

[170] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. “WebLogo: a sequence logo generator”. *Genome Res.* 14.6 (June 2004), pp. 1188–1190 (cit. on p. 37).

[171] A. S. Hinrichs et al. “The UCSC Genome Browser Database: update 2006”. *Nucleic Acids Res.* 34.Database issue (Jan. 2006), pp. D590–8 (cit. on pp. 38, 63).

[172] M. Khorshid, C. Rodak, and M. Zavolan. “CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins”. *Nucleic Acids Res.* 39.Database issue (Jan. 2011), pp. D245–52 (cit. on p. 38).

[173] L. Jaskiewicz, B. Bilen, J. Hausser, and M. Zavolan. “Argonaute CLIP–a method to identify in vivo targets of miRNAs”. *Methods* 58.2 (Oct. 2012), pp. 106–112 (cit. on p. 38).

[174] T. D. Wu and C. K. Watanabe. “GMAP: a genomic mapping and alignment program for mRNA and EST sequences”. *Bioinformatics* 21.9 (May 2005), pp. 1859–1875 (cit. on p. 38).

[175] A. R. Quinlan and I. M. Hall. “BEDTools: a flexible suite of utilities for comparing genomic features”. *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842 (cit. on p. 39).

[176] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. “STAR: ultrafast universal RNA-seq aligner”. *Bioinformatics* 29.1 (Jan. 2013), pp. 15–21 (cit. on pp. 41, 63, 94).

[177] A. J. Gruber, R. Schmidt, S. Ghosh, G. Martin, A. R. Gruber, E. van Nimwegen, and M. Zavolan. “Discovery of physiological and cancer-related regulators of 3’ UTR processing with KAPAC”. *Genome Biol.* 19.1 (Mar. 2018), p. 44 (cit. on p. 45).

[178] E. Wahle and U. Rüegsegger. “3’-End processing of pre-mRNA in eukaryotes”. *FEMS Microbiol. Rev.* 23.3 (June 1999), pp. 277–295 (cit. on p. 45).

[179] S. Millevoi, C. Loulergue, S. Dettwiler, S. Z. Karaa, W. Keller, M. Antoniou, and S. Vagner. “An interaction between U2AF 65 and CF I(m) links the splicing and 3’ end processing machineries”. *EMBO J.* 25.20 (Oct. 2006), pp. 4854–4864 (cit. on pp. 46, 60).

[180] T. Naganuma, S. Nakagawa, A. Tanigawa, Y. F. Sasaki, N. Goshima, and T. Hirose. “Alternative 3’-end processing of long noncoding RNA initiates construction of nuclear paraspeckles”. *EMBO J.* 31.20 (Oct. 2012), pp. 4020–4034 (cit. on pp. 46, 73).

[181] J. Ule, G. Stefani, A. Mele, M. Ruggiu, X. Wang, B. Taneri, T. Gaasterland, B. J. Blencowe, and R. B. Darnell. “An RNA map predicting Nova-dependent splicing regulation”. *Nature* 444.7119 (Nov. 2006), pp. 580–586 (cit. on p. 46).

[182] *The Cancer Genome Atlas*. `http://cancergenome.nih.gov`. Accessed: 2017-8-13 (cit. on pp. 47, 75).

[183] E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo. “Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)”. *Nat. Methods* 13.6 (June 2016), pp. 508–514 (cit. on pp. 47, 63, 76).

[184] B. M. Lunde, C. Moore, and G. Varani. “RNA-binding proteins: modular design for efficient function”. *Nat. Rev. Mol. Cell Biol.* 8.6 (June 2007), pp. 479–490 (cit. on pp. 47, 67).

[185] ENCODE Project Consortium. “An integrated encyclopedia of DNA elements in the human genome”. *Nature* 489.7414 (Sept. 2012), pp. 57–74 (cit. on pp. 55, 62).

[186] S. Gueroussov, T. Gonatopoulos-Pournatzis, M. Irimia, B. Raj, Z.-Y. Lin, A.-C. Gingras, and B. J. Blencowe. “An alternative splicing event amplifies evolutionary differences between vertebrates”. *Science* 349.6250 (Aug. 2015), pp. 868–873 (cit. on pp. 55, 62, 63, 71, 76, 77, 127, 140).

[187] H. C. Cheung, T. Hai, W. Zhu, K. A. Baggerly, S. Tsavachidis, R. Krahe, and G. J. Cote. “Splicing factors PTBP1 and PTBP2 promote proliferation and migration of glioma cell lines”. *Brain* 132.Pt 8 (Aug. 2009), pp. 2277–2288 (cit. on pp. 59, 76).

[188] N. Fong, H. Kim, Y. Zhou, X. Ji, J. Qiu, T. Saldi, K. Diener, K. Jones, X.-D. Fu, and D. L. Bentley. “Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate”. *Genes Dev.* 28.23 (Dec. 2014), pp. 2663–2676 (cit. on pp. 60, 62, 71, 125).

[189] B. C. Foat, A. V. Morozov, and H. J. Bussemaker. “Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE”. *Bioinformatics* 22.14 (July 2006), e141–9 (cit. on p. 60).

[190] P. J. Balwierz, M. Pachkov, P. Arnold, A. J. Gruber, M. Zavolan, and E. van Nimwegen. “ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs”. *Genome Res.* 24.5 (May 2014), pp. 869–884 (cit. on p. 60).

[191] G. Rot, Z. Wang, I. Huppertz, M. Modic, T. Lenče, M. Hallegger, N. Haberman, T. Curk, C. von Mering, and J. Ule. “High-Resolution RNA Maps Suggest Common Principles of Splicing and Polyadenylation Regulation by TDP-43”. *Cell Rep.* 19.5 (May 2017), pp. 1056–1067 (cit. on p. 60).

[192] N. Keppetipola, S. Sharma, Q. Li, and D. L. Black. “Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2”. *Crit. Rev. Biochem. Mol. Biol.* 47.4 (July 2012), pp. 360–378 (cit. on pp. 60, 76, 77).

[193] J. Chen and W. A. Weiss. “Alternative splicing in cancer: implications for biology and therapy”. *Oncogene* 34.1 (Jan. 2015), pp. 1–14 (cit. on p. 60).

[194] G. Martin, R. Schmidt, A. J. Gruber, S. Ghosh, W. Keller, and M. Zavolan. “3’ End Sequencing Library Preparation with A-seq2”. *J. Vis. Exp.* 128 (Oct. 2017), e56129 (cit. on pp. 62, 71).

[195] *Genomic Data Commons Data Portal*. `https://portal.gdc.cancer.gov`. Accessed: 2017-9-1 (cit. on pp. 62, 71).

[196] *NCBI Gene Expression Omnibus*. `https://www.ncbi.nlm.nih.gov/geo/`. Accessed: 2017-9-1 (cit. on pp. 62, 71).

[197] *ENCODE Portal*. `https://www.encodeproject.org/`. Accessed: 2017-9-1 (cit. on pp. 62, 71).

[198] *ENCODE RNA-seq pipeline*. `https://github.com/ENCODE-DCC/long-rna-seq-pipeline/blob/master/dnanexus/align-star-pe/resources/usr/bin/lrna_align_star_pe.sh`. Accessed: 2017-9-10 (cit. on pp. 62, 135).

[199] M. Martin. “Cutadapt removes adapter sequences from high-throughput sequencing reads”. *EMBnet.journal* 17.1 (May 2011), pp. 10–12 (cit. on p. 63).

[200] L. Wang, S. Wang, and W. Li. “RSeQC: quality control of RNA-seq experiments”. *Bioinformatics* 28.16 (Aug. 2012), pp. 2184–2185 (cit. on p. 66).

[201] S. Anders, P. T. Pyl, and W. Huber. “HTSeq–a Python framework to work with high-throughput sequencing data”. *Bioinformatics* 31.2 (Jan. 2015), pp. 166–169 (cit. on pp. 66, 136).

[202] J. Köster and S. Rahmann. “Snakemake–a scalable bioinformatics workflow engine”. *Bioinformatics* 28.19 (Oct. 2012), pp. 2520–2522 (cit. on p. 66).

[203] *NCBI Sequence Read Archive*. `https://www.ncbi.nlm.nih.gov/sra/`. Accessed: 2017-9-1 (cit. on p. 71).

[204] Z. Ji and B. Tian. “Reprogramming of 3’ untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types”. *PLoS One* 4.12 (Dec. 2009), e8419 (cit. on p. 73).

[205] V. Hilgers, M. W. Perry, D. Hendrix, A. Stark, M. Levine, and B. Haley. “Neural-specific elongation of 3’ UTRs during Drosophila development”. *Proc. Natl. Acad. Sci. USA.* 108.38 (Sept. 2011), pp. 15864–15869 (cit. on p. 73).

[206] J. Brumbaugh, B. Di Stefano, X. Wang, M. Borkent, E. Forouzmand, K. J. Clowers, F. Ji, B. A. Schwarz, M. Kalocsay, S. J. Elledge, Y. Chen, R. I. Sadreyev, S. P. Gygi, G. Hu, Y. Shi, and K. Hochedlinger. "Nudt21 Controls Cell Fate by Connecting Alternative Polyadenylation to Chromatin Signaling". *Cell* 172.1-2 (Jan. 2018), 106–120.e21 (cit. on p. 73).

[207] M. Sun, J. Ding, D. Li, G. Yang, Z. Cheng, and Q. Zhu. "NUDT21 regulates 3'-UTR length and microRNA-mediated gene silencing in hepatocellular carcinoma". *Cancer Lett.* 410 (Dec. 2017), pp. 158–168 (cit. on p. 73).

[208] D. R. Higgs, S. E. Goodbourn, J. Lamb, J. B. Clegg, D. J. Weatherall, and N. J. Proudfoot. "Alpha-thalassaemia caused by a polyadenylation signal mutation". *Nature* 306.5941 (1983), pp. 398–400 (cit. on p. 74).

[209] S. H. Orkin, T. C. Cheng, S. E. Antonarakis, and H. H. Kazazian Jr. "Thalassemia due to a mutation in the cleavage-polyadenylation signal of the human beta-globin gene". *EMBO J.* 4.2 (Feb. 1985), pp. 453–456 (cit. on p. 74).

[210] M. P. Chao, I. L. Weissman, and R. Majeti. "The CD47-SIRPα pathway in cancer immune evasion and potential therapeutic implications". *Curr. Opin. Immunol.* 24.2 (Apr. 2012), pp. 225–232 (cit. on p. 75).

[211] F. Weighardt, G. Biamonti, and S. Riva. "The roles of heterogeneous nuclear ribonucleoproteins (hnRNP) in RNA metabolism". *Bioessays* 18.9 (Sept. 1996), pp. 747–756 (cit. on p. 76).

[212] X. He, A. D. Arslan, T.-T. Ho, C. Yuan, M. R. Stampfer, and W. T. Beck. "Involvement of polypyrimidine tract-binding protein (PTBP1) in maintaining breast cancer cell growth and malignant properties". *Oncogenesis* 3 (Jan. 2014), e84 (cit. on p. 76).

[213] J. A. Santiago and J. A. Potashkin. "Blood Biomarkers Associated with Cognitive Decline in Early Stage and Drug-Naive Parkinson's Disease Patients". *PLoS One* 10.11 (Nov. 2015), e0142582 (cit. on p. 76).

[214] D. Dominguez, P. Freese, M. S. Alexis, A. Su, M. Hochman, et al. "Sequence, Structure and Context Preferences of Human RNA Binding Proteins". *bioRxiv* (2017) (cit. on p. 76).

[215] N. Lambert, A. Robertson, M. Jangi, S. McGeary, P. A. Sharp, and C. B. Burge. "RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins". *Mol. Cell* 54.5 (June 2014), pp. 887–900 (cit. on p. 76).

[216] M. Niwa, S. D. Rose, and S. M. Berget. "In vitro polyadenylation is stimulated by the presence of an upstream intron". *Genes Dev.* 4.9 (Sept. 1990), pp. 1552–1559 (cit. on p. 77).

[217] S. Millevoi, F. Geraghty, B. Idowu, J. L. Y. Tam, M. Antoniou, and S. Vagner. “A novel function for the U2AF 65 splicing factor in promoting pre-mRNA 3’-end processing”. *EMBO Rep.* 3.9 (Sept. 2002), pp. 869–874 (cit. on p. 77).

[218] S. Danckwardt, M. W. Hentze, and A. E. Kulozik. “3’ end mRNA processing: molecular mechanisms and implications for health and disease”. *EMBO J.* 27.3 (Feb. 2008), pp. 482–498 (cit. on p. 77).

[219] B. S. Zhao, I. A. Roundtree, and C. He. “Post-transcriptional gene regulation by mRNA modifications”. *Nat. Rev. Mol. Cell Biol.* 18.1 (Jan. 2017), pp. 31–42 (cit. on p. 78).

[220] S. Ke, E. A. Alemu, C. Mertens, E. C. Gantman, J. J. Fak, A. Mele, B. Haripal, I. Zucker-Scharff, M. J. Moore, C. Y. Park, C. B. Vågbø, A. Kusśnierczyk, A. Klungland, J. E. Darnell Jr, and R. B. Darnell. “A majority of m6A residues are in the last exons, allowing the potential for 3’ UTR regulation”. *Genes Dev.* 29.19 (Oct. 2015), pp. 2037–2053 (cit. on p. 78).

[221] A. Rehfeld, M. Plass, K. Døssing, U. Knigge, A. Kjær, A. Krogh, and L. Friis-Hansen. “Alternative polyadenylation of tumor suppressor genes in small intestinal neuroendocrine tumors”. *Front. Endocrinol.* 5 (Apr. 2014), p. 46 (cit. on pp. 82, 107).

[222] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. “Integrative genomics viewer”. *Nat. Biotechnol.* 29.1 (Jan. 2011), pp. 24–26 (cit. on pp. 94, 124).