Feature Review

CellPress

Alternative mRNA transcription, processing, and translation: insights from RNA sequencing

Eleonora de Klerk and Peter A.C. 't Hoen

Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

The human transcriptome comprises >80 000 protein-coding transcripts and the estimated number of proteins synthesized from these transcripts is in the range of 250 000 to 1 million. These transcripts and proteins are encoded by less than 20 000 genes, suggesting extensive regulation at the transcriptional, post-transcriptional, and translational level. Here we review how RNA sequencing (RNA-seq) technologies have increased our understanding of the mechanisms that give rise to alternative transcripts and their alternative translation. We highlight four different regulatory processes: alternative transcription initiation, alternative splicing, alternative polyadenylation, and alternative translation initiation. We discuss their transcriptome-wide distribution, their impact on protein expression, their biological relevance, and the possible molecular mechanisms leading to their alternative regulation. We conclude with a discussion of the coordination and the interdependence of these four regulatory layers.

Regulatory layers defining gene expression

The diversification of cellular and organismal functions observed in higher eukaryotes cannot be explained by the sheer number of genes but is mostly due to the expression of different transcripts and proteins from the same genes. Variation in the expression of coding genes is controlled at multiple levels, from transcription to RNA processing and translation. Alternative transcripts and proteins may arise from alternative transcription initiation, alternative splicing, alternative polyadenylation (APA), and alternative translation initiation. These co- and post-transcriptional regulatory mechanisms expand the genome's coding capacity modifying protein function, stability, localization, and expression levels. In this review, we discuss how high-throughput RNA-seq has helped us to understand these four regulatory processes. We describe their transcriptome-wide abundance in mammalian cells, their impact on protein expression, their biological relevance, and the molecular mechanisms underlying these processes. Finally, we highlight how the interdependence between transcription, RNA processing, and translation restricts the number of combinations of possible alternative transcripts and proteins.

Corresponding author: 't Hoen, P.A.C. (p.a.c.hoen@lumc.nl).

Keywords: gene expression; transcriptome; RNA sequencing; alternative polyadenylation; alternative splicing; translation.

0168-9525/

 http://dx.doi.org/10.1016/j.tig.2015.01.001

Initiation of transcription: alternative promoters

During the biogenesis of mRNAs, regulation of transcription initiation represents the first layer in the control of gene expression [1–4]. Alternative transcription initiation leads to the formation of transcripts differing in their first exon or in the length of the 5′ untranslated region (5′-UTR). The use of alternative first exons leads to transcripts with different open reading frames (ORFs) and diversifies the repertoire of encoded proteins giving rise to protein isoforms with alternative N termini [5] (Figure 1A). Alternatively, transcripts sharing the same coding region but a different 5′-UTR can be subject to differential translational regulation (Figure 1B) [6] through short upstream ORFs (uORFs) involved in translational control [7–9] or in the production of biologically relevant peptides [10–12].

The use of alternative promoters and transcription start sites (TSSs) in protein coding transcripts was established before the development of transcriptome-wide approaches, through studies based on a method called cap analysis of gene expression (CAGE) [13]. CAGE still represents the basic technology for the detection of TSSs. Recently, several high-throughput CAGE methods, such as DeepCAGE, have been developed [14]. These transcriptome-wide studies suggest that TSS use is highly tissue specific [4,15–18] and that the number of alternative TSSs differs by tissue type, with the hippocampus accounting for a larger number of TSSs than any other tissue [18,19]. To what extent alternative TSSs lead to alternative 5′ noncoding regions or translate into novel protein isoforms is virtually impossible to determine from DeepCAGE reads, which consist of 25 or 26 nucleotides. To assess the potential for novel ORFs arising from the use of alternative TSSs, it is essential to integrate DeepCAGE data with RNA-seq, ribosome profiling, and proteomics.

The FANTOM Consortium is leading most of the research in the field of promoters and TSSs. In their most recent TSS survey [4], which includes approximately 200 human primary cell types, 150 human tissues, and 250 human cancer cell lines, it was shown that on average there are four TSSs per gene, but the number of TSSs reported strictly relies on the filtering method used. An estimate of the transcriptome-wide distribution of alternative TSSs can indeed be complicated by the presence of CAGE peaks marking enhancer regions [4], 3′-UTRs

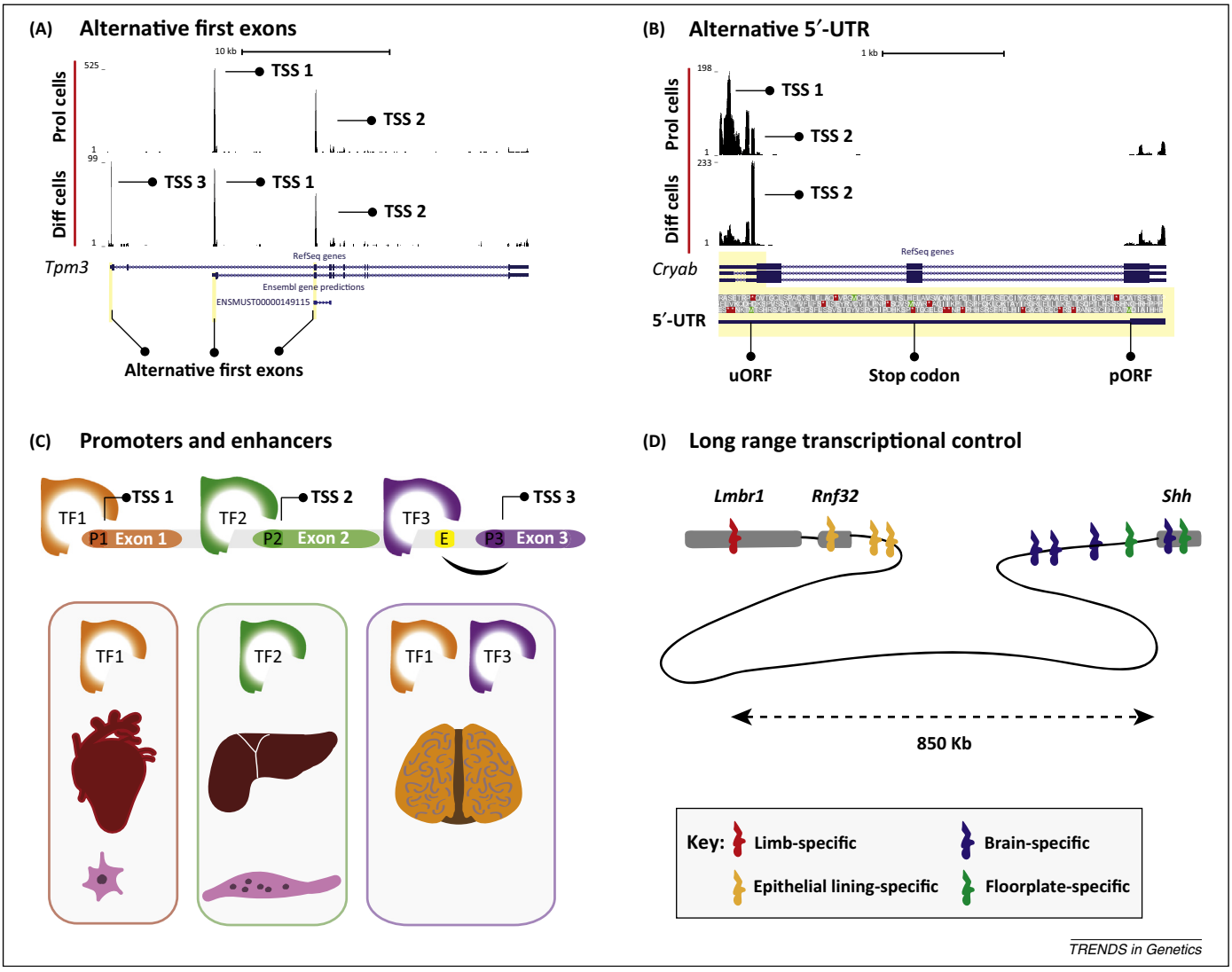

Figure 1. Alternative transcription initiation. (**A**) Data from a deep cap analysis of gene expression (DeepCAGE) experiment showing alternative transcription start sites (TSSs) used during muscle differentiation in proliferating myoblasts and differentiated myotubes [16]. In the *Tpm3* gene, different promoters lead to the formation of transcripts with different first exons. One alternative TSS (TSS3) is specifically used in differentiated cells. (**B**) In the *Cryab* gene, proliferating cells make use of an alternative TSS to extend their 5′ untranslated region (5′-UTR). The sequence of the 5′-UTR is shown below the reference track. The extension on the 5′-UTR leads to the transcription of a potential upstream open reading frame (uORF) starting at a canonical AUG codon and ending before the start codon of the primary ORF (pORF). (**C**) An illustrative example of cell- and tissue-specific alternative TSSs regulated by the binding of transcription factors (TFs) to promoters and enhancer regions. While TF1 and TF2 bind to promoters (P1, P2) surrounding the TSS, TF3 binds to a distal upstream sequence corresponding to an enhancer region (E), which enhances transcription from a third TSS (TSS3). Some TFs are present in multiple tissues (TF1) whereas others are tissue specific (TF2, TF3), and their transcription can also be regulated during cell differentiation (TF1 regulates transcription in undifferentiated cells and TF2 in differentiated cells). (**D**) Long-range transcriptional control mediated by enhancers. Transcriptional regulation of the *Shh* gene is tightly controlled during development by enhancer regions located up to 850 kb from the gene. Whereas some enhancers are located within the coding region of *Shh*, others are located in intergenic regions or within intronic regions of the *Lmbr1* and *Rnf32* genes. Genes are depicted as gray boxes. Known enhancer regions in the mouse are marked in different colors according to their tissue specificity.

[4,20,21], coding regions (a phenomenon called exon painting [16,22,23]), and promoter-associated short RNAs (PASRs) [20]. Whereas exon painting may arise as a consequence of recapping of degradation products, many other CAGE peaks represent short capped transcripts whose functions remain largely unknown. A striking recent finding from this large TSS survey [4] is that most genes are regulated in a tissue-specific manner and only a small percentage can be considered to be truly housekeeping. The use of alternative tissue-specific TSSs seems to be regulated by the presence of enhancer regions more than by alternative core promoters. Half of all detected CpG island promoters and more than 90% of all promoters lacking both CpG islands and a TATA box exhibit cell type-restricted expression due to the presence of proximal enhancers [4].

The molecular mechanisms responsible for the choice of alternative promoters and TSSs can be divided into two categories: alteration of the chromatin state and regulation mediated by cell- and tissue-specific transcription factors (Figure 1C). Understanding the biological importance of alternative and tissue-specific TSSs requires learning how the choice of a specific TSS is made and which transcription factor and regulatory networks are involved. This can be achieved by making inferences on transcriptional networks. In a DeepCAGE time-course study on the differentiation of human monocytic leukemia cells [17], the authors predicted transcription factor binding sites around the

TSSs identified in each condition and subsequently built a network model of gene expression using motif activity response analysis. This provided important insights into the key regulators active in transcriptional control in distinct phases of differentiation. Similarly, another study [24] inferred transcriptional regulatory networks after the perturbation of specific transcription factors (PU.1, IRF8, MYB and SP1) in the same cells. This led to the discovery of target genes for each transcription factor and led to the identification of *de novo* binding site motifs.

Many studies focusing on single genes have shown that the choice of a specific TSS has critical roles during development [25–27] and cell differentiation [28] and aberrations in alternative promoter and TSS use lead to various diseases including cancer [29,30], neuropsychiatric disorders [31], and developmental disorders [32]. Whereas some disorders are caused by epigenetic changes or genetic aberrations in the promoter region, others are caused by genetic changes in distal elements affecting long-range transcriptional regulation. The ENCODE project has shown the presence of more than 1000 long-range interactions between TSSs and distal elements within a range of 120 kb [3]. An example of such a long-range interaction is *Shh* [32] (Figure 1D), a gene that is spatially and temporally regulated during development. To date, ten *Shh* enhancers have been identified, located within a region of 1 Mb in humans and 850 kb in mice (Figure 1D). These enhancers play a key role during development, as indicated by mutations in the limb-specific enhancer that lead to various skeletal limb abnormalities.

Splicing: alternative exons

During and after transcription, almost all mRNAs are spliced. Alternatively spliced transcripts result from the differential inclusion of subsets of exons (Figure 1A and Box 1). Of the regulatory mechanisms discussed in this review, alternative splicing is the most prevalent event, affecting approximately 95% of mammalian genes [33]. RNA-seq has the potential to elucidate the number, structure, and abundance of alternative transcripts and the molecular mechanisms responsible for their formation.

Box 1. Alternative splicing events

Five major alternative splicing events are distinguished: exon skipping (also called cassette exon), use of alternative acceptor and/or donor sites, intron retention, and mutually exclusive exons. Exon skipping appears to be the most common, occurring in ~38% of mouse and human genes, whereas intron retention is less common (~3%) [135]. How the spliceosome recognizes alternative exons and decides which exons to include remains not fully understood. Before the advent of RNA-seq, studies revealed some general characteristics in conserved alternative cassette exons: they tend to be smaller in size compared with constitutive exons [136] and their length is divisible by three, thus maintaining the same reading frame when the alternative exon is skipped or included [137]. Non-conserved cassette exons do not show these characteristics. In addition, alternative exons seem to contain weaker splice sites (the exon–intron junctions at the 5′ and 3′ ends of introns; i.e., donor and acceptor sites), although the other primary *cis*-acting elements used to define the intron (the branch site and the polypyrimidine tract located upstream of the acceptor site) are generally similar to those found in constitutive exons [138].

From analysis of the transcriptomes of 15 different human cell lines [1], it appears that up to 25 different transcripts can be produced from a single gene and that up to 12 alternative transcripts may be expressed in a particular cell. Alternative transcripts are not expressed at the same level, but one transcript is usually dominant [34]. According to the latest GENCODE release [version 20 (http://www.gencodegenes.org/stats.html)], there are almost 80 000 transcript variants encoded by about 20 000 protein-coding genes in humans – an average of four transcripts per gene. A previous GENCODE release (version 7) reported an average of six transcripts per gene, while RefSeq, the University of California, Santa Cruz (UCSC), and the Collaborative Consensus Coding Sequence (CCDS) project [35] report a much lower average. These discordances suggest that variations in the number of transcripts per gene reported are due to the different methods used to annotate RNA sequences, highlighting the current limitations in fully characterizing transcriptomes.

It remains challenging to predict which transcripts are present in a specific cell type. Splice site selection depends on multiple parameters including the presence of splicing regulators, the strength of splice sites, the structure of exon–intron junctions, and the process of transcription. So far, various molecular mechanisms have been shown to regulate alternative splicing.

Next to conserved *cis* elements such as the splice donor and acceptor sites, branch sites, polypyrimidine tracts, and a range of other sequence motifs are recognized by various auxiliary splicing factors. These auxiliary RNA-binding proteins (RBPs) are not part of the spliceosomal machinery but can enhance or suppress alternative splicing by interfering with it [36–39]. Various crosslinking and RNA immunoprecipitation techniques, followed by next-generation sequencing, have been developed to map RNA–protein interactions *in vivo* [14]. An early goal of these studies was the identification of RNA-binding sites. Many of these studies have shown that RBPs recognize short (~3–7 nt) degenerate motifs, have multiple RNA-binding domains, and display variable efficiency when multiple motifs cluster together [40,41]. Moreover, many RBPs regulate the expression of other auxiliary factors. The differing cellular and temporal localization of RBPs [42,43] may explain the different dynamics regulating alternative and constitutive splicing: whereas constitutive splicing mainly occurs cotranscriptionally, alternative splicing mainly occurs post-transcriptionally [44]. For recent mechanistic models of splicing regulation through RBPs, see [45]. Alternative splicing can also be regulated in a manner totally independent of auxiliary splicing factors [46]. Splicing silencer sequences regulate alternative splicing when competing 5′ splice sites are present in the same RNA molecule (Figure 2B). The competing 5′ splice sites are equally well recognized by the U1 small nuclear ribonucleoprotein (snRNP), but silencer sequences alter the configuration in which U1 binds to the 5′ splice sites, leading to silencing of the 5′ splice site. This can change the efficiency of a splice site: weak 5′ splice sites can be recognized and used instead of stronger 5′ splice sites. RNA-seq datasets can be used to computationally identify common and tissue-specific

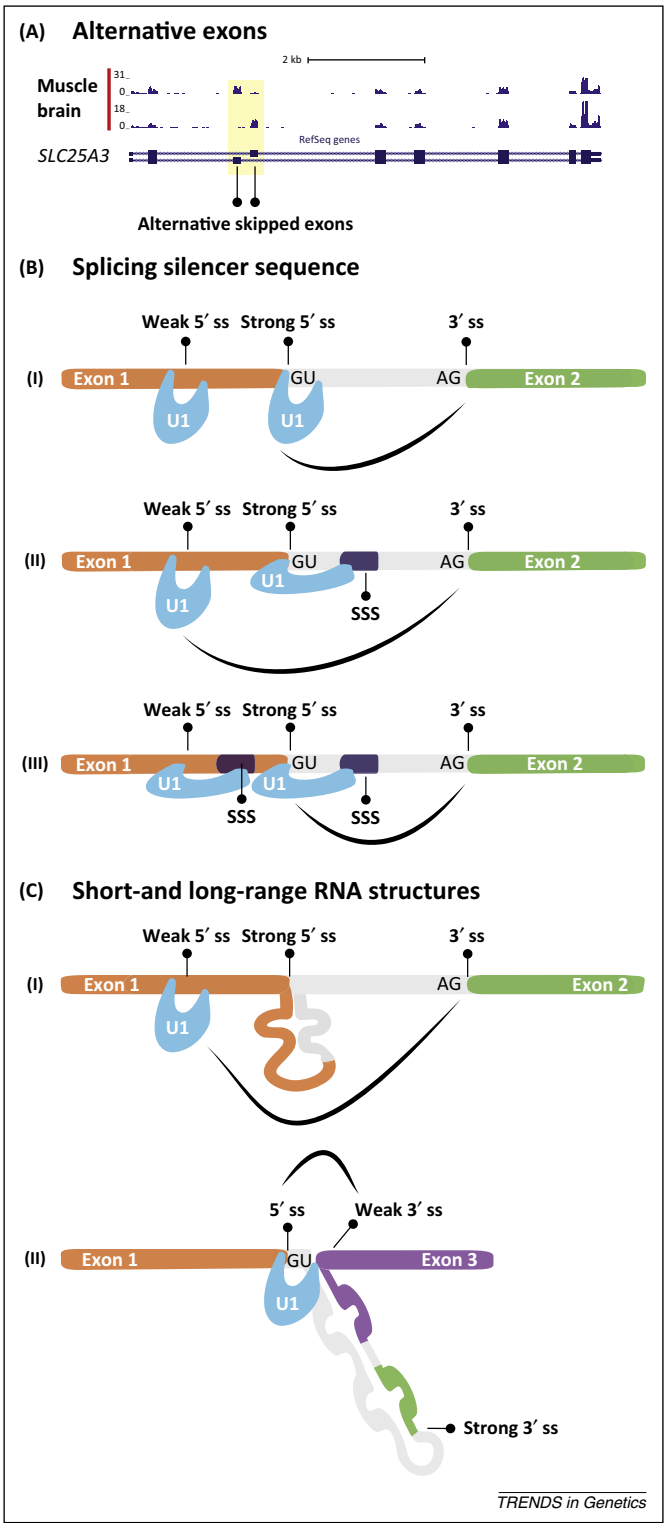

Figure 2. Alternative splicing. (**A**) Data from an RNA sequencing (RNA-seq) experiment showing tissue-specific alternative splicing [139]. The *SLC25A3* gene is differentially spliced in brain and muscle tissues through exon skipping. (**B**) Alternative splicing regulated by silencer sequences. In (I) the U1 small nuclear ribonucleoprotein (snRNP) splicing factor recognizes both strong and weak 5′ splice sites (5′ss) but splicing occurs only at the strong 5′ss. In (II) a splicing silencer sequence (sss) is located downstream of the strong 5′ss. U1 binds both the weak and the strong 5′ss, but the conformation in which it binds the strong 5′ss is suboptimal for splicing; therefore, only the weak 5′ss is used for splicing. In (III) the sss is located downstream of both the weak and the strong 5′ss. U1 binds both with suboptimal conformation, but only the strong 5′ss is used for splicing. (**C**) Alternative splicing regulated by RNA secondary structures. Example of short- and long-range RNA secondary structures. (I) The short-range RNA secondary structure masks a strong 5′ss, leading to the recognition of a weaker 5′ss located upstream. (II) The long-range RNA secondary structure brings together a strong 5′ss and a weak 3′ss, causing the loss of a complete exon (in green) and a region of the last exon (in purple).

splicing regulatory sequences. These studies have shown that the same sequence can act as an enhancer or a silencer in different tissues, but experimental validations of these predicted regulatory sequences are needed to confirm these observations [47].

Alternative splicing can also be regulated by RNA secondary structures (Figure 2C). Short-range RNA secondary structures can mask primary *cis* elements such as the acceptor and donor sites or the polypyrimidine tract [48,49]. They have been associated with alternative splicing at alternative 5′ splice sites. For example, the RBP MBNL1 forms a secondary structure upstream of exon 5 of human *TNNT2* and upstream of the fetal exon of mouse *Tnnt3*, blocking U2AF65 binding to the polypyrimidine tract [50,51]. Long-range secondary structures bring distant splice sites into closer proximity, facilitating alternative splicing, and are associated with weak alternative 3′ splice sites [49]. Computational studies based on RNA-seq datasets suggest that the splicing of thousands of mammalian genes is dependent on RNA structures, both short and long range [49]. Recently developed high-throughput techniques combine nuclease digestion [52] or chemical probing [53] with next-generation sequencing to provide transcriptome-wide RNA structural information. Two studies have recently shown a transcriptome-wide relationship between secondary structures and alternative splicing [54,55], by reporting the presence of strong secondary structures at 5′ splice sites that correlate with unspliced exons. The question that remains unsolved by RNA-seq studies is whether the plethora of transcript variants produced affect protein expression. This question has been recently addressed by studies using ribosome profiling, discussed further below. A general observation from transcriptome-wide studies is that alternative splicing is essential for development [56,57] and cell, tissue [58], and species specificity [59]. A plausible explanation of how alternative exons can confer such specificity is the inclusion or exclusion of binding motifs and post-translational modification sites, as shown in a study where the authors investigated the structural and functional properties of alternative exons [60].

Due to the widespread role of alternative splicing, it is unsurprising that errors in this process lead to various diseases, from neurodegenerative disorders to muscle dystrophies and cancer; we refer the reader to recent detailed reviews [61,62].

3′ End maturation: APA

Another step in mRNA processing is the process of polyadenylation [63]. The use of APA sites represents an extra regulatory layer during gene expression that results in the formation of transcripts differing in their 3′ ends. Transcripts arising from APA may differ in their coding region (if APA sites are located in a different exon or intron) (Figure 3A) or in the length of their 3′-UTRs [tandem polyadenylation sites (PASs)] (Figure 3B). The impact of APA on the regulation of gene expression can be extended

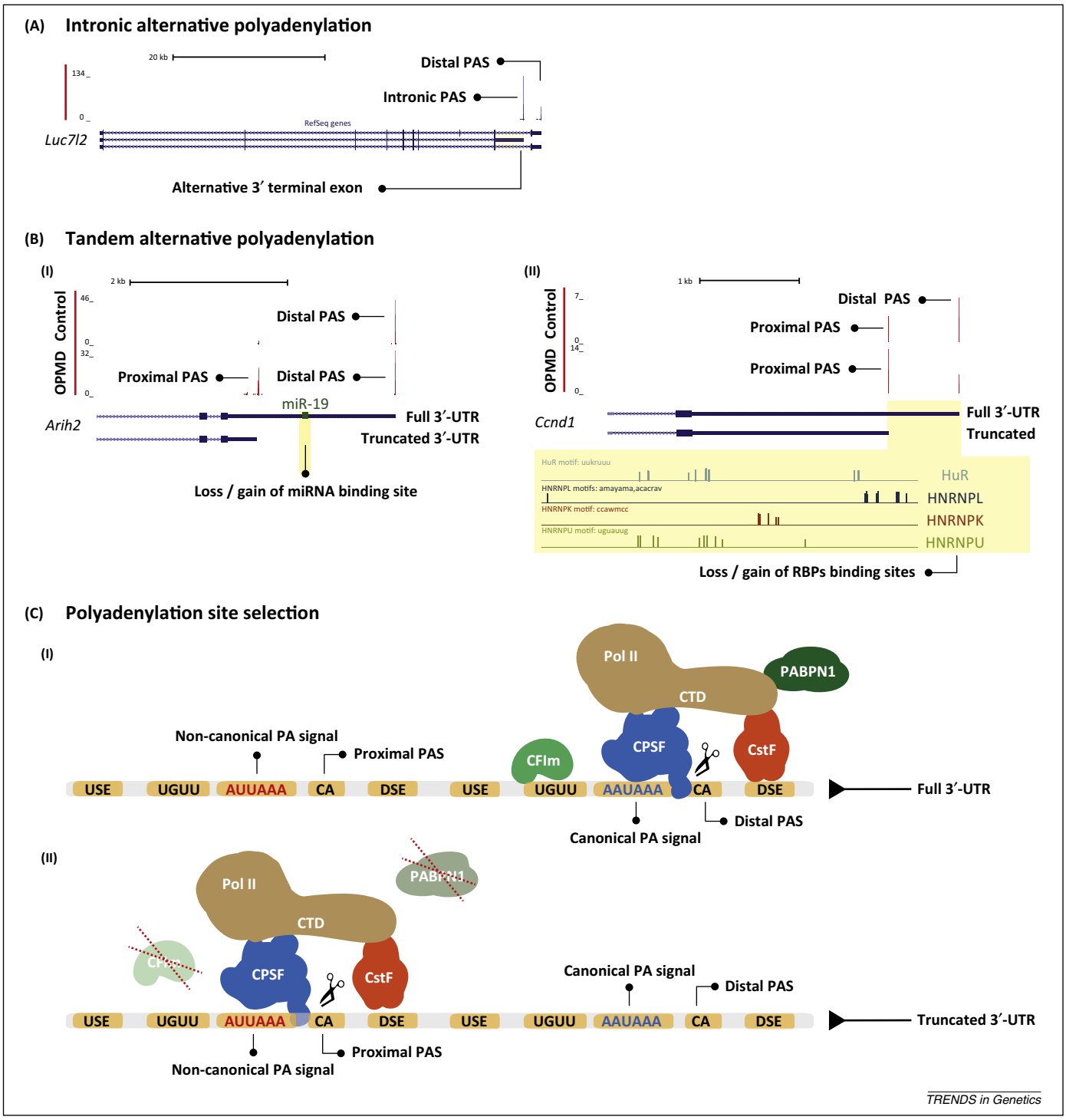

Figure 3. Alternative polyadenylation (APA). (**A**) Data from a poly(A)-sequencing experiment showing APA in the intron of the *Luc7l2* gene [71], leading to an intronic proximal polyadenylation site (PAS) located in a different terminal exon giving rise to transcript variants with different open reading frames (ORFs). (**B**) Two examples of tandem APA in muscle tissue from a mouse model for oculopharyngeal muscle dystrophy (OPMD) [71]. In the *Arih2* gene (I), both the distal and the proximal PASs can be used in the disease state. Recognition of a proximal PAS leads to shortening of the 3′ untranslated region (3′-UTR) and loss of a miRNA binding site, causing an increase in transcript levels. In the *Ccnd1* gene (II), shortening of the 3′-UTR leads to the loss of many recognition sites for RNA-binding proteins (RBPs) that stabilize the transcript. Loss of stability leads to a decrease in transcript level. (**C**) Model mechanisms regulating tandem APA. Common sequences in the 3′-UTR that regulate polyadenylation are the upstream sequence element (USE), the UGUU sequence recognized by cleavage factor I (CFIm), the polyadenylation (PA) signal recognized by cleavage and polyadenylation specific factor (CPSF), and the downstream sequence element (DSE) recognized by cleavage stimulation factor (CstF). CPSF and CstF are brought to the RNA by RNA polymerase II (Pol II), together with poly(A)-binding protein nuclear 1 (PABPN1), through its C-terminal domain (CTD). Generally, CPSF recognizes the canonical PA signal and cuts at a distal PAS, at a CA dinucleotide (I). If PABPN1 or CFIm is present at a lower concentration, the CPSF recognizes noncanonical (weaker) PA signals (II) and cuts at proximal PASs, leading to the formation of transcripts with truncated 3′-UTRs.

through effects on transcript localization [64], stability, and translation efficiency [65] and on the nature of the encoded protein. Numerous RNA-seq methods have contributed to our understanding of APA, ranging from RNA-seq studies able to detect overall changes in polyadenylation, to serial analysis of gene expression (SAGE)-based methods able to specifically quantify and characterize the 3′ ends of transcripts, to a series of dedicated protocols for the

accurate detection and quantification of PASs [14]. These transcriptome-wide studies have deepened our understanding of APA, providing information on newly discovered PASs, elucidating the impact of APA on gene expression, and discovering new APA regulatory mechanisms.

Although the number of alternative PASs detected differs greatly between studies [66–68], these studies contribute to the notion of the ubiquity of APA events, which involve approximately 70% of human genes. According to a study conducted on 15 human cell lines, there are on average two PASs per gene [1]. APA within the same last exon (tandem 3′-UTRs) is the most abundant type of APA [68]. Intronic APA events are reported less frequently and thousands of intronic PASs are usually suppressed [69]. APA is generally linked to changes in gene expression levels and, ultimately, to protein abundance. Studies have shown an inverse correlation between 3′-UTR length and protein expression levels [70,71]. Some human tissues (such as brain, testis, lung, and breast) are enriched for highly abundant transcripts with short 3′-UTRs, whereas others (such as heart and skeletal muscle) contain many low-abundance transcripts with long 3′-UTRs [72]. Increased expression of transcripts with shortened 3′-UTRs can be explained by loss of miRNA target sequences, loss of UPF1-binding sites, which leads to RNA decay [73], or loss of AU-rich elements (AREs), which leads to ARE-directed mRNA degradation [71]. However, there are many exceptions to the general rule, as proteins that bind to the 3′-UTR can also stabilize mRNAs [74–76].

Transcriptome-wide studies have been undertaken to elucidate the dynamics of APA regulation. In general, disruption of the polyadenylation machinery leads to loss of fidelity in the choice of PAS and shortening of the 3′-UTRs. There are numerous 3′ processing factors involved in polyadenylation; nevertheless, changes in the expression levels of a single specific factor are sufficient to influence the choice of PAS. For example, decreased levels of cleavage factor I (CFIm) 68 or poly(A)-binding protein nuclear 1 (PABPN1) lead to transcriptome-wide shortening of 3′-UTRs, corresponding to an increased preference for noncanonical polyadenylation signals (Figure 3C) [70,77,78].

Many recent transcriptome-wide studies have confirmed that distal PASs generally have a strong canonical signal motif [A(A/U)UAAA], whereas proximal PASs diverge from the canonical sequence [68,79–81]. Interestingly, tissue-specific regulated PASs can be depleted of the canonical motif. For example, APA in brain seems to be regulated by an A-rich motif starting just downstream of the PAS [82]. A-rich sequences have also been reported upstream of cleavage sites for transcripts lacking canonical motifs [83].

Numerous studies based on expressed sequence tags and microarrays have previously shown the biological relevance of APA (Box 2) [84,85]. APA profiles are tissue specific and appear to be tightly regulated during development and cell differentiation. Most of the findings achieved by recent transcriptome-wide approaches confirm at a larger scale what was previously observed. The tissue specificity of APA and the correlation between tissue and 3′-UTR length seem to be highly conserved between different species and APA profiles from different species are similar for the same tissues [80,81,86]. Modulation of APA has also been widely observed during proliferation, differentiation, and development [68,87–89].

Widespread alteration of APA profiles has been observed in several diseases. Many studies have reported shortening of 3′-UTRs in cancer [90–92], linked to extensive upregulation and activation of oncogenes. However, shortening of 3′-UTRs poorly correlates with breast, lung, and colorectal cancer prognosis [93,94], suggesting that the relationship between APA and cancer is not straightforward. More recently, altered APA profiles have been linked to muscle disorders such as myotonic dystrophy [95] and oculopharyngeal muscular dystrophy [70].

Box 2. The biological relevance of APA

A study based on expressed sequence tags comprising 42 human tissues [140] showed that certain tissues preferentially produce mRNAs of a certain length. Brain, pancreatic islet, ear, bone marrow, and uterus showed a preference for distal PASs, leading to longer 3′-UTRs. Retina, placenta, ovary, and blood showed a preference for proximal PASs. This classification might change when considering the levels at which these mRNAs are expressed. Although most of the transcripts detected in the brain contain distal PASs, the transcripts that are highly abundant generally show a preference for proximal PASs and have short 3′-UTRs [72]. Other studies showed that the choice between a distal and a proximal PAS was modulated during differentiation and development. Progressive lengthening of 3′-UTRs was shown for most of the transcripts during cell differentiation and during embryonic development [141]. By contrast, shortening was observed during proliferation [142] and during reprogramming of somatic cells [143].

From mRNA to protein: alternative translation initiation

In addition to the regulation of transcription and processing, the translation of transcripts is also tightly regulated. Regulation of translation defines not only the abundance of a protein but also its amino acid composition through the use of different start codons [96], as translation may start at uORFs or at alternative ORFs (aORFs) (Box 3 and Figure 4).

In the past, changes in protein synthesis were measured exclusively based on proteomic approaches or estimated based on total mRNA levels. More recently, they have been

Box 3. Alternative translation initiation

uORFs are located in the 5′-UTR of a transcript. Depending on the presence or absence of stop codons and their coding frame, a uORF can overlap with the pORF or not. Overlapping and in-frame uORFs lead to N-terminal extended protein isoforms [8], whereas non-overlapping uORFs affect the translation of pORFs in various ways [144]: they can block the translation of the pORFs, reducing protein production; they can promote reinitiation of translation at downstream start codons; or they can enhance translation of the main pORFs. aORFs are located downstream of the annotated start codon. In-frame aORFs give rise to N-terminal truncated isoforms [145]. uORFs and aORFs can also be out of frame with respect to the pORFs and lead to the production of different peptides. The sequences translated in more than one reading frame are called dual coding regions [103]. We also note that uORFs and aORFs are not the only events that increase the diversity of the translated mRNAs and affect protein production. The genetic code can be read in alternative ways, leading to frameshifting, hopping, stop codon read-through, recoding, and codon reassignment [146,147], topics beyond the scope of this review.

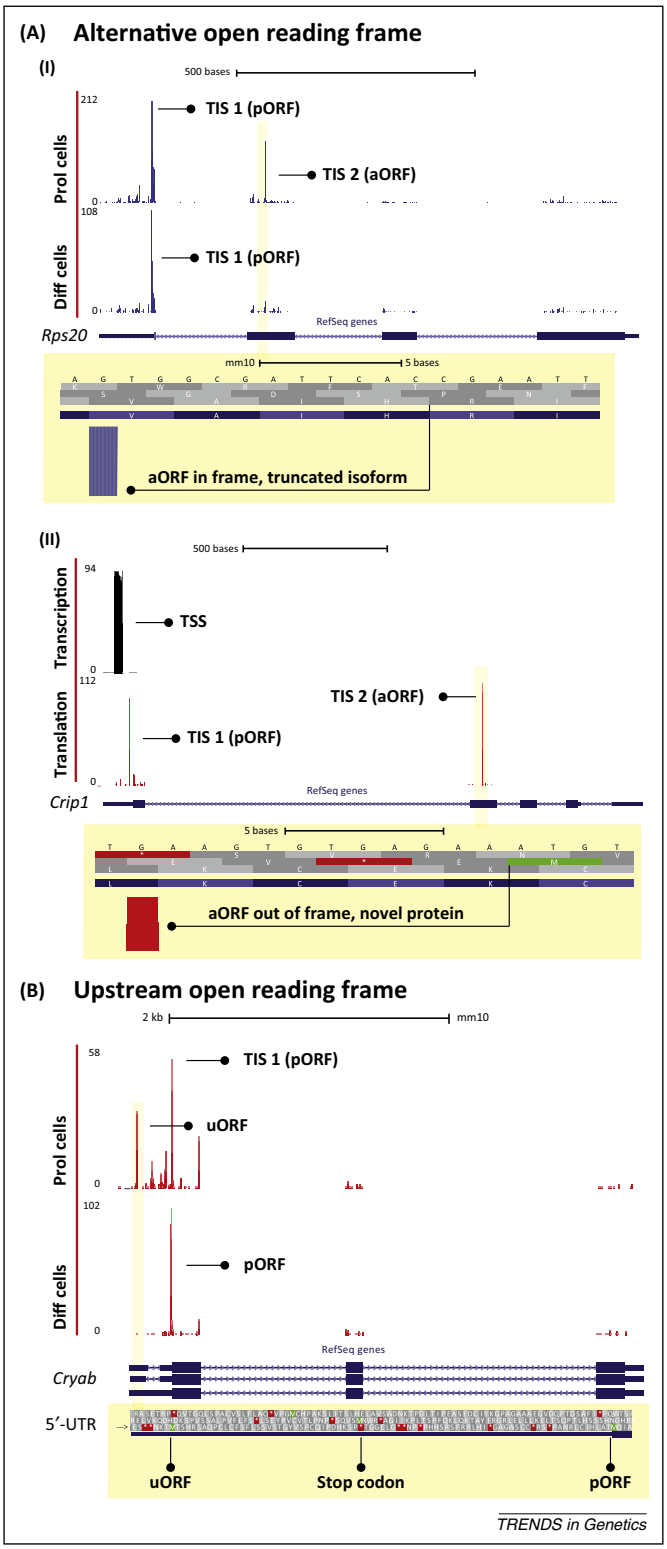

Figure 4. Alternative translation initiation. Alternative translation initiation sites (TISs) detected by ribosome profiling (http://www.ebi.ac.uk/ena/data/view/PRJEB7207). (**A**) Examples of alternative TISs leading to alternative open reading frames (aORFs) in frame (I) or out of frame (II) with the primary ORF (pORF). In the *Rps20* gene (I), a switch in TIS use occurs during cell differentiation. Proliferating cells use two TISs, one corresponding to the annotated start codon and the other corresponding to an aORF, the latter of which leads to a truncated protein isoform. The alternative TIS is shown in the highlighted box. The top part (gray) shows the three possible frames and the blue bar shows the frame of the pORF. Because ribosome profiling peaks are usually displayed using only the 5′ end of each mapped read, the black line indicates the actual TIS location of the aORF, located 12 bp downstream of the mapped peak. In the *Crip1* gene (II), only one transcription start site (TSS) is present (top track, deep cap analysis of gene expression (DeepCAGE) [16]) but two different TISs are used (bottom track, ribosome profiling), one corresponding to the annotated start codon and one located downstream of the annotated start codon, leading to an aORF. The alternative TIS is shown in the highlighted box. The alternative TIS corresponds to an AUG start codon that is out of frame compared with the pORF, indicating the presence of a dual coding region. (**B**) Examples of alternative TISs leading to an upstream ORF (uORF) in the *Cryab* gene. Proliferating cells use two TISs, one located in the 5′ untranslated region (5′-UTR) and one corresponding to the annotated start codon. The sequence of the 5′-UTR incorporated by the alternative TIS is shown below the reference track. Extension of the 5′-UTR leads to the translation of an uORF, with a canonical AUG codon and ending before the start codon of the pORF, negatively regulating translation.

assessed via ribosome profiling [97]. Deep sequencing of RNA fragments protected by ribosomes determines the position of the ribosomes on the RNA molecule at nucleotide resolution, allowing exact characterization of the translation initiation site (TIS) and quantification of levels of translation. Ribosome profiling studies in combination with RNA-seq have assessed the extent of alternative translation initiation, provided insights into the regulatory mechanisms of this process, and shed light on how it impacts gene expression.

A common finding of many recent ribosome profiling studies is the widespread use of alternative TISs. Initiation of translation at alternative TISs may be caused by various forms of stress but is also observed under normal physiological conditions. Between 50% and 65% of transcripts contains more than one TIS [7,98,99]. Most of the detected TISs are located upstream of the annotated start codons (50–60%), leading to potential uORFs. A minority are located downstream of the annotated start codons (~20%) and lead to N-terminally truncated proteins or out-of-frame ORFs. However, some ribosome profiling peaks detected as alternative TISs may represent cases of ribosomal stalling. To distinguish these from genuine TISs, proteomic data are essential. These are often difficult to obtain because the peptides are usually short and unstable. Moreover, the study of the proteome in a high-throughput fashion presents certain technical limitations, especially for low-abundance proteins, which are difficult to detect among a diverse pool of proteins [100].

Insights into the mechanisms regulating the choice of an uORF or aORF over a primary ORF are starting to emerge. Initiation of translation at near-cognate codons and non-AUG codons, previously reported for a small number of mRNAs, appears to be common, as approximately 50% of translation is initiated at noncanonical codons [98,99]. These noncanonical start codons are enriched in uORFs. By contrast, TISs located downstream of annotated TISs comprise mainly AUG codons. The use of near-cognate and non-AUG start codons has been confirmed by mass spectrometry [101]. Interestingly, these codons are recoded to regular methionines, as all of the produced proteins seem to contain an N-terminal methionine.

Recent studies support the leaky scanning theory [102], according to which the choice of a downstream TIS depends on the strength of the Kozak consensus sequence. It was shown on a transcriptome-wide scale that initiation at downstream TISs usually occurs when the Kozak sequence in the annotated start codon is suboptimal. A similar mechanism applies for initiation at uORFs. uORFs are translated in parallel to their downstream primary ORFs (pORFs) if the start codon used in the uORF is a non-AUG,

but translation of pORFs is usually repressed if the uORFs contain an AUG start codon and a strong Kozak sequence [99].

Both aORFs and uORFs can give rise to ORFs with reading frames different from the pORFs, a phenomenon known as dual coding [103]. The triplet periodicity observed in ribosome profiling data enables the detection of dually decoded regions. Although the extent of dual coding observed in the human genome in ribosome profiling studies is only approximately 1%, it has been suggested that this might be an underestimate due to technical and analytical limitations (low coverage and the assumption that the two frames must be translated at the same rate) [103].

The extent to which mRNA levels explain differences in protein abundance is still debated. Although some studies have reported a poor correlation [104] – in the range of approximately 40% of protein levels explained by mRNA levels [105–108] or even less than 20% [109] – others claim a much higher correlation of up to approximately 80% [110]. Ribosome-associated RNA levels seem to be a good proxy for protein levels, as the correlations between mRNA and protein observed are between 60% and 90% [109,111]. Nevertheless, a study that compared changes at mRNA levels and ribosome-bound mRNAs showed profound uncoupling between transcription and translation in several different experiments after treatments with extracellular stimuli or during cell and tissue differentiation [112]. Therefore, it remains unclear whether regulation at the translational level has a major influence on global protein abundance or whether it is restricted to a subset of genes.

Transcription, RNA processing, and translation: interdependent processes

The molecular machineries involved in transcription and RNA processing are spatiotemporally coupled. Several reviews have extensively described cotranscriptional regulation of capping, splicing, and polyadenylation [113,114]. RNA polymerase II (Pol II) is an important player in the regulation of this coupling, as its C terminus recruits proteins involved in capping, splicing, and polyadenylation [115]. There is ample support of the coupling between transcription and splicing. Splicing predominantly occurs during transcription [1,44], as indicated by the following three observations: many introns are already spliced in chromatin-associated RNAs; there is enrichment of spliceosomal small nuclear RNAs in chromatin-associated RNAs; and exons that are spliced are enriched for epigenetic chromatin marks [116]. Nevertheless, splicing events at the 3′ end of a transcript might occur post-transcriptionally, giving a general 5′–3′ trend in splicing completion.

Transcription and splicing are coupled not simply in space and time but are also jointly responsible for the formation of alternative transcripts. The interdependence of different RNA-processing events restricts the number of combinations of alternative TSSs, exons, and PASs. Splicing and polyadenylation might be influenced not only by the transcription elongation rate but also by transcription initiation: a lower elongation rate is linked to slower splicing and polyadenylation and therefore to an increased chance of recognizing alternative exons [117] or proximal PASs [118,119] and the choice of TSS is linked to a specific splicing pattern [120,121] or to the use of specific PASs [71,122,123].

In addition to links between transcription and mRNA processing, alternative splicing and APA also appear to be interdependent. Twenty years ago, it was shown that splicing of the last intron requires definition of the last exon (at least in mammals [124]) and this occurs through the cooperation of splicing and polyadenylation factors that interact across the last exon, leading to mutual enhancement of both splicing and polyadenylation [125]. The snRNPs U1 and U2 and the U2 auxiliary factor 65 kDa subunit (U2AF65), all spliceosome components, are also part of the human pre-mRNA 3′ processing complex [126]. These spliceosome components directly interact with cleavage and polyadenylation specific factor (CPSF) and with CFIm. Splicing factors can also play a role in premature cleavage and polyadenylation, as shown by the spliceosomal factor TRAP150 [127].

Recent transcriptome-wide studies further support the links between splicing and polyadenylation. Alteration of the splicing factor hnRNP H has been shown to have widespread effects on tandem APA, with increased 3′-UTR shortening in the presence of hnRNP H and lengthening in its absence (Figure 5A, top). Changes in APA were accompanied by changes in alternative splicing. A direct link between hnRNP H and the choice of a specific PAS was shown by crosslinking immunoprecipitation sequencing (CLIP-seq) analysis, by the presence of a higher CLIP tag density next to the proximal PAS [128]. An increase in proximal PAS use was also observed after alteration of Nova, a RBP involved in alternative splicing [36].

High CLIP tag density surrounding proximal PASs has also been observed for the RBPs MBNL1 and MBNL2 (Figure 5A, bottom), which are known to regulate splicing [38], and a direct link between MBNL proteins and APA was recently explained by the competition of MBNL with CFIm68, a component of the polyadenylation machinery [95].

Whether alternative splicing is also coupled to non-tandem APA remain unclear. A few studies have specifically investigated the interdependency between intronic polyadenylation and splicing. Cryptic intronic PASs are mainly located in large introns with weak 5′ splice sites. This suggests that intronic polyadenylation can be inhibited if there are splicing enhancers that recognize the 5′ splice site, as shown for U1 [129], or enhanced in the case of suboptimal splicing [130]. The coupling observed in this case represents kinetic competition between splicing and polyadenylation [131].

Finally, coupling is not restricted to processes connected in space and time. Interdependency has also been shown between processes occurring in different subcellular compartments; for example, between APA and translation. Cytoplasmic polyadenylation element-binding protein 1 (CPEB1), which shuttles between the nucleus and the cytoplasm, has been shown to play a dual role in APA and translation [132] (Figure 5B). Interestingly, CPEB1 can also regulate alternative splicing. CPEB1 prevents recruitment of the splicing factor U2AF65 to the 3′ splice

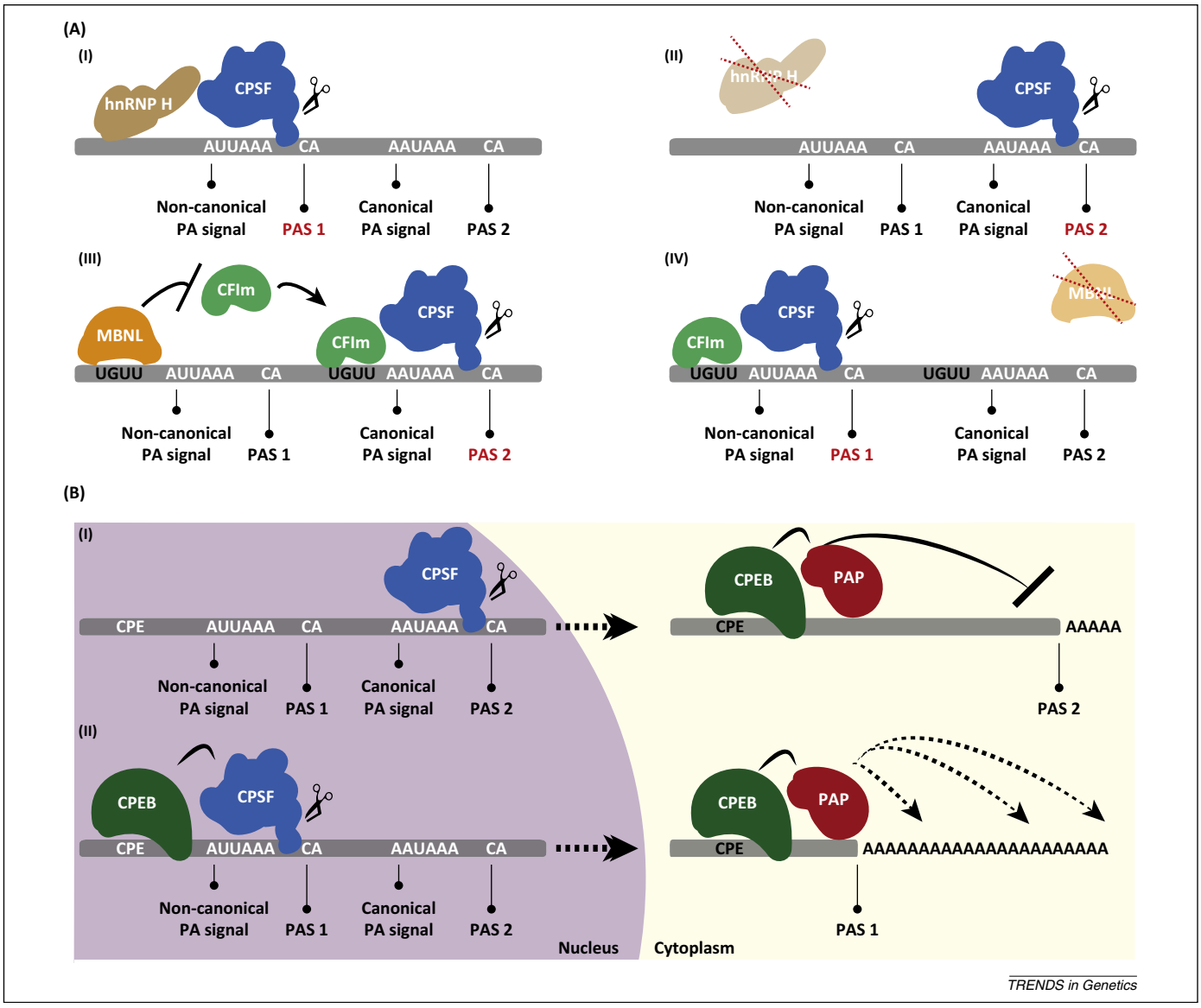

Figure 5. Coupled regulatory mechanisms. (**A**) Tandem alternative polyadenylation (APA) regulated by splicing factors. The RNA-binding proteins hnRNP H and MBNL regulate APA in opposing ways. In the presence of hnRNP H (I), cleavage and polyadenylation specific factor (CPSF) binds weaker noncanonical polyadenylation (PA) signals and cuts at the proximal polyadenylation site (PAS 1) leading to shortening of the 3′ untranslated region (3′-UTR), while in its absence (II) only the canonical PA signal is recognized and cleavage occurs in the distal PAS (PAS 2). (III) MBNL masks the region upstream of weak noncanonical PA signals, blocking the binding of cleavage factor I (CFIm). This leads to binding of CFIm to a more distal UGUU sequence, followed by binding of CPSF to the distal canonical PA signal and use of the distal PAS (PAS 2). In the absence of MBNL (IV), CFIm can bind proximal UGUU regions and bring the CPSF to weaker PA signals, causing cleavage at the proximal PAS (PAS 1) and shortening of the 3′-UTR. (**B**) Coupling of APA and translation. In the nucleus, in the absence of cytoplasmic polyadenylation element-binding protein 1 (CPEB1) (I), CPSF binds the canonical PA signal and cleaves the RNA at a distal PAS (PAS 2). In the presence of CPEB1 (II), CPEB1 binds the cytoplasmic polyadenylation element (CPE) located upstream of weak noncanonical PA signals. CPEB1 directly interacts with CPSF, bringing it to regions proximal to the weak PA signal. This leads to their recognition by CPSF and cleavage at the proximal PAS (PAS 1). When CBEP1 shuttles to the cytoplasm, it again binds to the CPE, but this time to promote lengthening of the poly(A) tail by poly(A) polymerase (PAP), which results in increased translation efficiency. Lengthening of the poly(A) tails of transcripts bearing proximal PASs (PAS 1) (II) is enhanced by the fact that the CPE, PAP, and the polyadenylation site are in close proximity, whereas this enhancement is disrupted when the distance is greater due to the 3′-UTR lengthening in transcripts bearing a distal PAS (PAS 2).

site, but simultaneously recruits the polyadenylation machinery. The RBP CPEB1 is an example of a master regulator that affects three layers of gene expression: splicing, polyadenylation, and translation.

Concluding remarks

RNA-seq technologies are elucidating the mechanisms that expand the genome's coding capacity and are quickly redefining the concept of gene expression regulation.

Although there is a continuing increase in the number of transcripts identified, and in the understanding of the molecular mechanisms that coordinate their formation during transcription and mRNA processing, we still face technical limitations due to the short read length of next-generation sequencing data and reliance on statistical and computational approaches to reconstruct transcript structure. This represents an obstacle when trying to link different events occurring in the same RNA molecule. The only way to specifically determine the exact transcript structure for each detected RNA molecule is by sequencing full-length RNAs, an option that is currently becoming more feasible [133,134] and that is opening a new era in the field of RNA-seq.

References

1 Djebali, S. *et al.* (2012) Landscape of transcription in human cells. *Nature* 489, 101–108
2 Neph, S. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489, 83–90
3 Sanyal, A. *et al.* (2012) The long-range interaction landscape of gene promoters. *Nature* 489, 109–113
4 FANTOM Consortium and RIKEN PMI and CLST (DGT) (2014) A promoter-level mammalian expression atlas. *Nature* 507, 462–470
5 Goossens, S. *et al.* (2007) Truncated isoform of mouse αT-catenin is testis-restricted in expression and function. *FASEB J.* 21, 647–655
6 Barbosa, C. *et al.* (2013) Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet.* 9, e1003529
7 Calvo, S.E. *et al.* (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7507–7512
8 Fritsch, C. *et al.* (2012) Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res.* 22, 2208–2218
9 Yamashita, R. *et al.* (2003) Small open reading frames in 5′ untranslated regions of mRNAs. *C. R. Biol.* 326, 987–991
10 Slavoff, S.A. *et al.* (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* 9, 59–64
11 Magny, E.G. *et al.* (2013) Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* 341, 1116–1120
12 Jorgensen, R.A. and Dorantes-Acosta, A.E. (2012) Conserved peptide upstream open reading frames are associated with regulatory genes in angiosperms. *Front. Plant Sci.* 3, 191
13 Shiraki, T. *et al.* (2002) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15776–15781
14 de Klerk, E. *et al.* (2014) RNA sequencing: from tag-based profiling to resolving complete transcript structure. *Cell. Mol. Life Sci.* 71, 3537–3551
15 de Hoon, M. and Hayashizaki, Y. (2008) Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *Biotechniques* 44, 627–628 630, 632
16 Hestand, M.S. *et al.* (2010) Tissue-specific transcript annotation and expression profiling with complementary next-generation sequencing technologies. *Nucleic Acids Res.* 38, e165
17 Suzuki, H. *et al.* (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.* 41, 553–562
18 Valen, E. *et al.* (2009) Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res.* 19, 255–265
19 Gustincich, S. *et al.* (2006) The complexity of the mammalian transcriptome. *J. Physiol.* 575, 321–332
20 Kapranov, P. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488
21 Andersson, R. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461
22 Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project (2009) Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* 457, 1028–1032
23 Otsuka, Y. *et al.* (2009) Identification of a cytoplasmic complex that adds a cap onto 5′-monophosphate RNA. *Mol. Cell. Biol.* 29, 2155–2167
24 Vitezic, M. *et al.* (2010) Building promoter aware transcriptional regulatory networks using siRNA perturbation and DeepCAGE. *Nucleic Acids Res.* 38, 8141–8148
25 Levanon, D. and Groner, Y. (2004) Structure and regulated expression of mammalian RUNX genes. *Oncogene* 23, 4211–4219
26 Steinthorsdottir, V. *et al.* (2004) Multiple novel transcription initiation sites for NRG1. *Gene* 342, 97–105
27 Davis, W., Jr and Schultz, R.M. (2000) Developmental change in TATA-box utilization during preimplantation mouse development. *Dev. Biol.* 218, 275–283
28 Pozner, A. *et al.* (2007) Developmentally regulated promoter-switch transcriptionally controls Runx1 function during embryonic hematopoiesis. *BMC Dev. Biol.* 7, 84
29 Pedersen, I.S. *et al.* (2002) Promoter switch: a novel mechanism causing biallelic PEG1/MEST expression in invasive breast cancer. *Hum. Mol. Genet.* 11, 1449–1453
30 Agarwal, V.R. *et al.* (1996) Use of alternative promoters to express the aromatase cytochrome P450 (CYP19) gene in breast adipose tissues of cancer-free and breast cancer patients. *J. Clin. Endocrinol. Metab.* 81, 3843–3849
31 Tan, W. *et al.* (2007) Molecular cloning of a brain-specific, developmentally regulated neuregulin 1 (NRG1) isoform and identification of a functional promoter variant associated with schizophrenia. *J. Biol. Chem.* 282, 24343–24351
32 Hill, R.E. and Lettice, L.A. (2013) Alterations to the remote control of *Shh* gene expression cause congenital abnormalities. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 368, 20120357
33 Pan, Q. *et al.* (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415
34 Gonzalez-Porta, M. *et al.* (2013) Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.* 14, R70
35 Harrow, J. *et al.* (2012) GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Res.* 22, 1760–1774
36 Licatalosi, D.D. *et al.* (2008) HITS–CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456, 464–469
37 Ule, J. *et al.* (2003) CLIP identifies Nova-regulated RNA networks in the brain. *Science* 302, 1212–1215
38 Wang, E.T. *et al.* (2012) Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell* 150, 710–724
39 Lebedeva, S. *et al.* (2011) Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol. Cell* 43, 340–352
40 Fu, X.D. and Ares, M., Jr (2014) Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.* 15, 689–701
41 Zhang, C. *et al.* (2013) Prediction of clustered RNA-binding protein motif sites in the mammalian genome. *Nucleic Acids Res.* 41, 6793–6807
42 Ameur, A. *et al.* (2011) Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* 18, 1435–1440
43 Hao, S. and Baltimore, D. (2013) RNA splicing regulates the temporal order of TNF-induced gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 110, 11934–11939
44 Tilgner, H. *et al.* (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* 22, 1616–1625
45 Witten, J.T. and Ule, J. (2011) Understanding splicing regulation through RNA splicing maps. *Trends Genet.* 27, 89–97
46 Yu, Y. *et al.* (2008) Dynamic regulation of alternative splicing by silencers that modulate 5′ splice site competition. *Cell* 135, 1224–1236
47 Wen, J. *et al.* (2010) Computational identification of tissue-specific alternative splicing elements in mouse genes from RNA-seq. *Nucleic Acids Res.* 38, 7895–7907
48 Shepard, P.J. and Hertel, K.J. (2008) Conserved RNA secondary structures promote alternative splicing. *RNA* 14, 1463–1469
49 Pervouchine, D.D. *et al.* (2012) Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. *RNA* 18, 1–15
50 Warf, M.B. *et al.* (2009) The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9203–9208
51 Yuan, Y. *et al.* (2007) Muscleblind-like 1 interacts with RNA hairpins in splicing target and pathogenic RNAs. *Nucleic Acids Res.* 35, 5474–5486
52 Kertesz, M. *et al.* (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature* 467, 103–107
53 Lucks, J.B. *et al.* (2011) Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-seq). *Proc. Natl. Acad. Sci. U.S.A.* 108, 11063–11068
54 Wan, Y. *et al.* (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 505, 706–709
55 Ding, Y. *et al.* (2014) *In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* 505, 696–700

56 Giudice, J. *et al.* (2014) Alternative splicing regulates vesicular trafficking genes in cardiomyocytes during postnatal heart development. *Nat. Commun.* 5, 3603
57 Kim, K.K. *et al.* (2013) Rbfox3-regulated alternative splicing of Numb promotes neuronal differentiation during development. *J. Cell Biol.* 200, 443–458
58 Pimentel, H. *et al.* (2014) A dynamic alternative splicing program regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res.* 42, 4031–4042
59 Gracheva, E.O. *et al.* (2011) Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats. *Nature* 476, 88–91
60 Buljan, M. *et al.* (2012) Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol. Cell* 46, 871–883
61 Costa, V. *et al.* (2013) RNA-seq and human complex diseases: recent accomplishments and future perspectives. *Eur. J. Hum. Genet.* 21, 134–142
62 Pistoni, M. *et al.* (2010) Alternative splicing and muscular dystrophy. *RNA Biol.* 7, 441–452
63 Danckwardt, S. *et al.* (2008) 3′ End mRNA processing: molecular mechanisms and implications for health and disease. *EMBO J.* 27, 482–498
64 Andreassi, C. and Riccio, A. (2009) To localize or not to localize: mRNA fate is in 3′UTR ends. *Trends Cell Biol.* 19, 465–474
65 Fabian, M.R. *et al.* (2010) Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* 79, 351–379
66 Derti, A. *et al.* (2012) A quantitative atlas of polyadenylation in five mammals. *Genome Res.* 22, 1173–1183
67 Ozsolak, F. *et al.* (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* 143, 1018–1029
68 Shepard, P.J. *et al.* (2011) Complex and dynamic landscape of RNA polyadenylation revealed by PAS-seq. *RNA* 17, 761–772
69 Yao, C. *et al.* (2012) Transcriptome-wide analyses of CstF64–RNA interactions in global regulation of mRNA alternative polyadenylation. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18773–18778
70 de Klerk, E. *et al.* (2012) Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation. *Nucleic Acids Res.* 40, 9089–9101
71 Ji, Z. *et al.* (2011) Transcriptional activity regulates alternative cleavage and polyadenylation. *Mol. Syst. Biol.* 7, 534
72 Ni, T. *et al.* (2013) Distinct polyadenylation landscapes of diverse human tissues revealed by a modified PA-seq strategy. *BMC Genomics* 14, 615
73 Hogg, J.R. and Goff, S.P. (2010) Upf1 senses 3′UTR length to potentiate mRNA decay. *Cell* 143, 379–389
74 Ray, D. *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499, 172–177
75 Gupta, I. *et al.* (2014) Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA–protein interactions. *Mol. Syst. Biol.* 10, 719
76 Spies, N. *et al.* (2013) 3′ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.* 23, 2078–2090
77 Jenal, M. *et al.* (2012) The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell* 149, 538–553
78 Martin, G. *et al.* (2012) Genome-wide analysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor I in the regulation of 3′ UTR length. *Cell Rep.* 1, 753–763
79 Jan, C.H. *et al.* (2011) Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. *Nature* 469, 97–101
80 Smibert, P. *et al.* (2012) Global patterns of tissue-specific alternative polyadenylation in *Drosophila*. *Cell Rep.* 1, 277–289
81 Ulitsky, I. *et al.* (2012) Extensive alternative polyadenylation during zebrafish development. *Genome Res.* 22, 2054–2066
82 Hafez, D. *et al.* (2013) Genome-wide identification and predictive modeling of tissue-specific alternative polyadenylation. *Bioinformatics* 29, i108–i116
83 Nunes, N.M. *et al.* (2010) A functional human poly(A) site requires only a potent DSE and an A-rich upstream sequence. *EMBO J.* 29, 1523–1536
84 Tian, B. *et al.* (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* 33, 201–212
85 Yan, J. and Marr, T.G. (2005) Computational analysis of 3′-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Res.* 15, 369–375
86 Miura, P. *et al.* (2013) Widespread and extensive lengthening of 3′ UTRs in the mammalian brain. *Genome Res.* 23, 812–825
87 Hoque, M. *et al.* (2013) Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nat. Methods* 10, 133–139
88 Li, Y. *et al.* (2012) Dynamic landscape of tandem 3′ UTRs during zebrafish development. *Genome Res.* 22, 1899–1906
89 Mangone, M. *et al.* (2010) The landscape of *C. elegans* 3′UTRs. *Science* 329, 432–435
90 Fu, Y. *et al.* (2011) Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res.* 21, 741–747
91 Lin, Y. *et al.* (2012) An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res.* 40, 8460–8471
92 Mayr, C. and Bartel, D.P. (2009) Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* 138, 673–684
93 Lembo, A. *et al.* (2012) Shortening of 3′UTRs correlates with poor prognosis in breast and lung cancer. *PLoS ONE* 7, e31129
94 Morris, A.R. *et al.* (2012) Alternative cleavage and polyadenylation during colorectal cancer development. *Clin. Cancer Res.* 18, 5256–5266
95 Batra, R. *et al.* (2014) Loss of MBNL Leads to disruption of developmentally regulated alternative polyadenylation in RNA-mediated disease. *Mol. Cell* 56, 311–322
96 Kochetov, A.V. (2008) Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. *Bioessays* 30, 683–691
97 Ingolia, N.T. *et al.* (2012) The ribosome profiling strategy for monitoring translation *in vivo* by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* 7, 1534–1550
98 Ingolia, N.T. *et al.* (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147, 789–802
99 Lee, S. *et al.* (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.* 109, E2424–E2432
100 Wasinger, V.C. *et al.* (2013) Current status and advances in quantitative proteomic mass spectrometry. *Int. J. Proteomics* 2013, 180605
101 Menschaert, G. *et al.* (2013) Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol. Cell. Proteomics* 12, 1780–1790
102 Kozak, M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* 361, 13–37
103 Michel, A.M. *et al.* (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res.* 22, 2219–2229
104 Maier, T. *et al.* (2009) Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* 583, 3966–3973
105 Lundberg, E. *et al.* (2010) Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* 6, 450
106 Schwanhausser, B. *et al.* (2011) Global quantification of mammalian gene expression control. *Nature* 473, 337–342
107 Tian, Q. *et al.* (2004) Integrated genomic and proteomic analyses of gene expression in mammalian cells. *Mol. Cell. Proteomics* 3, 960–969
108 Vogel, C. *et al.* (2010) Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* 6, 400
109 Ingolia, N.T. *et al.* (2009) Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218–223
110 Li., J.J. *et al.* (2014) System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* 2, e270
111 Wang, T. *et al.* (2013) Translating mRNAs strongly correlate to proteins in a multivariate manner and their translation ratios are phenotype specific. *Nucleic Acids Res.* 41, 4743–4754

112 Tebaldi, T. *et al.* (2012) Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells. *BMC Genomics* 13, 220
113 Auboeuf, D. *et al.* (2005) A subset of nuclear receptor coregulators act as coupling proteins during synthesis and maturation of RNA transcripts. *Mol. Cell. Biol.* 25, 5307–5316
114 Bentley, D.L. (2014) Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* 15, 163–175
115 Hsin, J.P. and Manley, J.L. (2012) The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev.* 26, 2119–2137
116 Brown, S.J. *et al.* (2012) Chromatin and epigenetic regulation of pre-mRNA processing. *Hum. Mol. Genet.* 21, R90–R96
117 Dujardin, G. *et al.* (2013) Transcriptional elongation and alternative splicing. *Biochim. Biophys. Acta* 1829, 134–140
118 Hazelbaker, D.Z. *et al.* (2013) Kinetic competition between RNA polymerase II and Sen1-dependent transcription termination. *Mol. Cell* 49, 55–66
119 Pinto, P.A. *et al.* (2011) RNA polymerase II kinetics in polo polyadenylation signal selection. *EMBO J.* 30, 2431–2444
120 Benson, M.J. *et al.* (2012) Heterogeneous nuclear ribonucleoprotein L-like (hnRNPLL) and elongation factor, RNA polymerase II, 2 (ELL2) are regulators of mRNA processing in plasma cells. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16252–16257
121 Huang, D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57
122 Huang, Y. *et al.* (2012) Mediator complex regulates alternative mRNA processing via the MED23 subunit. *Mol. Cell* 45, 459–469
123 Nagaike, T. *et al.* (2011) Transcriptional activators enhance polyadenylation of mRNA precursors. *Mol. Cell* 41, 409–418
124 Martinson, H.G. (2011) An active role for splicing in 3′-end formation. *Wiley Interdiscip. Rev. RNA* 2, 459–470
125 Berget, S.M. (1995) Exon recognition in vertebrate splicing. *J. Biol. Chem.* 270, 2411–2414
126 Shi, Y. *et al.* (2009) Molecular architecture of the human pre-mRNA 3′ processing complex. *Mol. Cell* 33, 365–376
127 Lee, K.M. and Tarn, W.Y. (2014) TRAP150 activates splicing in composite terminal exons. *Nucleic Acids Res.* Published online October 17, 2014. (http://dx.doi.org/10.1093/nar/gku963)
128 Katz, Y. *et al.* (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* 7, 1009–1015
129 Kaida, D. *et al.* (2010) U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* 468, 664–668
130 Tian, B. *et al.* (2007) Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res.* 17, 156–165
131 Luo, W. *et al.* (2013) The conserved intronic cleavage and polyadenylation site of CstF-77 gene imparts control of 3′ end processing activity through feedback autoregulation and by U1 snRNP. *PLoS Genet.* 9, e1003613
132 Bava, F.A. *et al.* (2013) CPEB1 coordinates alternative 3′-UTR formation with translational regulation. *Nature* 495, 121–125
133 Au, K.F. *et al.* (2013) Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 110, E4821–E4830
134 Sharon, D. *et al.* (2013) A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31, 1009–1014
135 Sugnet, C.W. *et al.* (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.* 2004, 66–77
136 Sorek, R. *et al.* (2004) How prevalent is functional alternative splicing in the human genome? *Trends Genet.* 20, 68–71
137 Resch, A. *et al.* (2004) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res.* 32, 1261–1269
138 Sorek, R. *et al.* (2014) Minimal conditions for exonization of intronic sequences: 5′ splice site formation in Alu exons. *Mol. Cell* 14, 221–231
139 Mortazavi, A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* 5, 621–628
140 Zhang, H. *et al.* (2005) Biased alternative polyadenylation in human tissues. *Genome Biol.* 6, R100
141 Ji, Z. *et al.* (2009) Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci. U.S.A.* 106, 7028–7033
142 Sandberg, R. *et al.* (2008) Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* 320, 1643–1647
143 Ji, Z. and Tian, B. (2009) Reprogramming of 3′UTR untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PLoS ONE* 4, e8419
144 Wethmar, K. (2014) The regulatory potential of upstream open reading frames in eukaryotic gene expression. *Wiley Interdiscip. Rev. RNA* 5, 765–778
145 Vanderperre, B. *et al.* (2013) Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS ONE* 8, e70698
146 Atkins, J.F. and Raymond, F., eds (2010) *Recoding: Expansion of Decoding Rules Enriches Gene Expression*, Springer
147 Atkins, J.F. and Baranov, P.V. (2010) The distinction between recoding and codon reassignment. *Genetics* 185, 1535–1536