CHAPTER 1

REGULATORY LAYERS DEFINING GENE EXPRESSION

(1) Eleonora de Klerk and Peter A.C. 't Hoen.
(2) Eleonora de Klerk, Johan T. den Dunnen, Peter A.C. 't Hoen.

Partly published at
(1) Trends Genet. 2015 Mar; 31(3):128-139.
doi: 10.1016/j.tig.2015.01.001.

(2) Cell Mol Life Sci. 2014 Sep; 71(18):3537-51.
doi: 10.1007/s00018-014-1637-9.

The transcriptome can be described as the complete collection of RNA molecules expressed in a specific cell type or tissue at a given time. It includes coding RNAs (messenger RNA) and a multitude of non-coding RNAs (of which ribosomal RNA, transfer RNA, small nuclear RNA, small nucleolar RNA, microRNA, Piwi-interacting RNA, and long non-coding RNA are best characterized). RNA plays a central role in cell biology, where it not only serves as template for protein synthesis but also acts as a structural scaffold and as a regulatory molecule during post-transcriptional control of gene expression (David, 2012;Kung et al., 2013). 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. 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.

The first section of this chapter will elaborate on how high-throughput RNA sequencing technologies have increased our understanding of the mechanisms that give rise to alternative transcripts and their alternative translation, and it will highlight four different regulatory processes: alternative transcription initiation, alternative splicing, alternative polyadenylation, and alternative translation initiation. It will focus on their transcriptome-wide distribution, their impact on protein expression, their biological relevance, and the possible molecular mechanisms leading to their alternative regulation. Finally, it will address how the interdependence between transcription, RNA processing, and translation restricts the number of combinations of possible alternative transcripts and proteins. The second section of this chapter will focus on the major genome-wide RNA sequencing methods used to investigate specific aspects of gene expression and its regulation. Tag-based methods (for studying transcription, alternative initiation and polyadenylation events), shotgun methods (for detection of alternative splicing), full-length RNA sequencing (for the determination of complete transcript structures), and targeted methods (for studying the process of transcription and translation) will be presented.

1. Alternative mRNA transcription, processing and translation

The biogenesis of a messenger RNA (mRNA) is characterized by four major steps (**Figure 1**): transcription of long heterogeneous nuclear RNAs (hnRNAs, also known as nascent RNA or pre-mRNAs (Scherrer et al., 1963;Soeiro et al., 1968)), capping of its 5' end (Shatkin, 1976), splicing (consisting in the removal of noncoding intervening sequences [introns] and joining of expressed sequences [exons] (Gilbert, 1978)), and polyadenylation of the 3' end, which involves cleavage of the pre-mRNA and synthesis of a poly(A) tail (Manley et al., 1982). Once an mRNA is processed, it is transported to the cytoplasm where it serves as a template for protein synthesis during the process of translation, and lastly it is degraded. Capping, splicing and polyadenylation represent the most common co- and post-transcriptional mRNA processing events. Each of these processes influences the metabolism and therefore the future of the mRNA molecule.

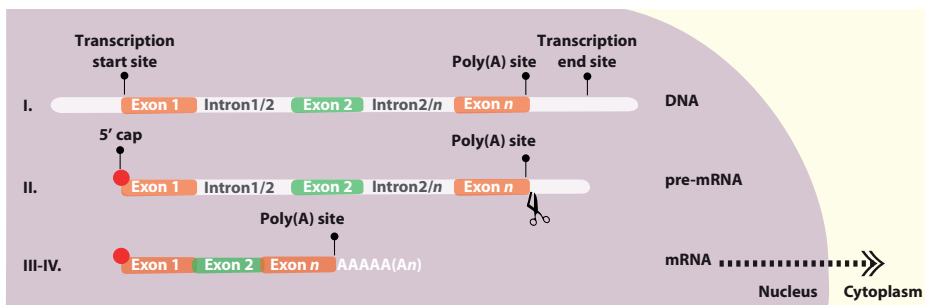

Figure 1. Biogenesis of an mRNA. Schematic representation of capping, splicing and polyadenylation.

The cap-structure consists of a 7-methylguanosine, which is linked to the first nucleotide of the mRNA and bound to cap-binding proteins. In the cytoplasm, the cap-structure is important for the initiation of translation, since the eukaryotic translation initiation factor eIF4A binds directly to the cap-structure (Sonenberg and Gingras, 1998).

Constitutive splicing occurs co- or post-transcriptionally, and is catalyzed by the spliceosome, a large RNA-protein complex. Whereas constitutive splicing is important to maintain a correct reading frame and therefore the coding potential of an mRNA, alternative splicing regulates whether a specific protein isoform is made, and its expression level. Furthermore, splicing has evolutionary implications, especially through recombination of exons which coincide with protein domains (Patthy, 1999).

Polyadenylation is a process required for nuclear export, stability of mature mRNA, and for its efficient translation, as mRNAs with short tails are generally subjected to degradation or stored to postpone their translation (Gorgoni and Gray, 2004).

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, 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.

1.1 Initiation of transcription: alternative promoters

During the biogenesis of mRNAs, regulation of transcription initiation represents the first layer in the control of gene expression (Djebali et al., 2012;Neph et al., 2012;Sanyal et al., 2012;The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014). 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 (Goossens et al., 2007) (**Figure 2a**). Alternatively, transcripts sharing the same coding region but a different 5′-UTR can be subject to differential translational regulation (**Figure 2b**) (Barbosa et al., 2013) through short upstream ORFs (uORFs) involved in translational control (Calvo et al., 2009;Fritsch et al., 2012;Yamashita et al., 2003) or in the production of biologically relevant peptides (Jorgensen and Dorantes-Acosta, 2012;Magny et al., 2013;Slavoff et al., 2013).

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) (Shiraki et al., 2003). CAGE still represents the basic technology for the detection of TSSs. Recently, several high-throughput CAGE methods, such as DeepCAGE, have been developed (**section 2.1.2, this Chapter**). These transcriptome-wide studies suggest that TSS use is highly tissue specific (de Hoon and Hayashizaki, 2008;Hestand et al., 2010;Suzuki et al., 2009;The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014;Valen et al., 2009) 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 (Gustincich et al., 2006;Valen et al., 2009). To what extent alternative TSSs lead to alternative 5′ non-coding 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 (The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014), 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, 3′-UTRs (Andersson et al., 2014;Kapranov et al., 2007), coding regions (a phenomenon called exon painting (Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project, 2009;Hestand et al., 2010;Otsuka et al., 2009), and promoter-associated short RNAs (PASRs) (Kapranov et al., 2007). 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 (The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014) 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.

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 2c**). Understanding the biological importance of

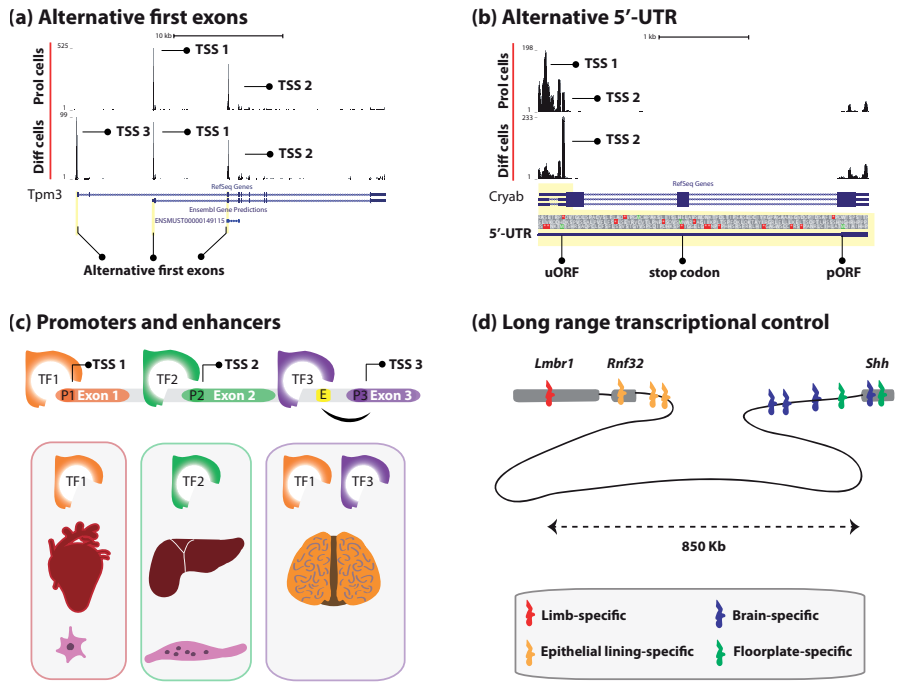

Figure 2. Alternative transcription initiation. (a) Data from a DeepCAGE experiment showing alternative transcription start sites (TSS) used during muscle differentiation in proliferating myoblasts and differentiated myotubes [65]. In the Tpm3 gene different promoters lead to the formation of transcripts 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′-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 open reading frame (pORF). **(c)** An illustrative example of cell- and tissue-specific alternative TSSs regulated by binding of transcription factors (TF) 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. The transcriptional regulation of the Shh gene is tightly controlled during development by enhancer regions located up to 850 kb away 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 mouse are marked in different colors, according to their tissue-specificity.

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 (Suzuki et al., 2009), 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 (Vitezic et al., 2010) 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 is critical for

(embryonic) development (Davis, Jr. and Schultz, 2000;Levanon and Groner, 2004;Steinthorsdottir et al., 2004) and cell differentiation (Pozner et al., 2007) and aberrations in alternative promoter and TSS use lead to various diseases including cancer (Agarwal et al., 1996;Pedersen et al., 2002), neuropsychiatric disorders (Tan et al., 2007), and developmental disorders (Hill and Lettice, 2013). 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 (Sanyal et al., 2012). An example of such a long-range interaction is Shh (Hill and Lettice, 2013), 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 2d**). These enhancers play a key role during development, as indicated by mutations in the limb-specific enhancer that lead to various skeletal limb abnormalities.

1.2 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 3a**). RNA-seq has the potential to elucidate the number, structure, and abundance of alternative transcripts and the molecular mechanisms responsible for their formation.

Of the regulatory mechanisms discussed in this chapter, alternative splicing is the most prevalent event, affecting approximately 95% of mammalian genes (Pan et al., 2008). 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%) (Sugnet et al., 2004).

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 to constitutive exons (Sorek et al., 2004b) and their length is divisible by three, thus maintaining the same reading frame when the alternative exon is skipped or included (Resch et al., 2004). 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 (Sorek et al., 2004a).

From analysis of the transcriptomes of 15 different human cell lines (Djebali et al., 2012), 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 (Gonzalez-Porta et al., 2013). According to the latest GENCODE release [version 20 (http://www.gencodegenes.org/stats.html)], there are almost 80,000 transcripts 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 (Harrow et al., 2012) 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

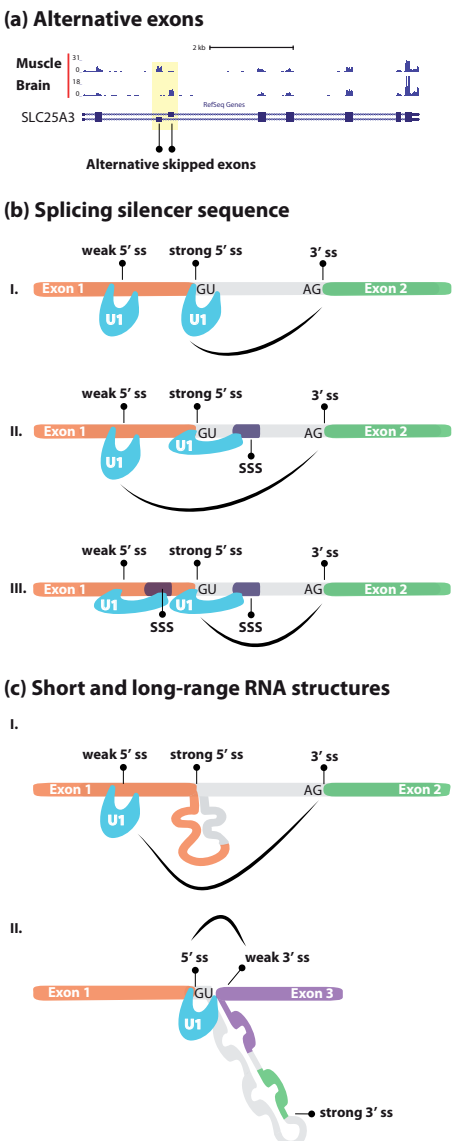

Figure 3. Alternative splicing. (a) Data from an RNAseq experiment showing tissue-specific alternative splicing [129]. The SLC25A3 gene is differentially spliced in brain and muscle tissues through exon skipping. **(b)** Alternative splicing regulated by silencer sequences. In (I) the U1 snRNP splicing 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 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 both weak and 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- (I) and long-range (II) 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).

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, and polypyrimidine tracts, 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 (Lebedeva et al., 2011;Licatalosi et al., 2008;Ule et al., 2003;Wang et al., 2012). Various crosslinking and RNA immunoprecipitation

techniques, followed by next-generation sequencing, have been developed to map RNA–protein interactions *in vivo* (**section 2.4, this Chapter**). 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 (Fu and Ares, Jr., 2014;Zhang et al., 2013). Moreover, many RBPs regulate the expression of other auxiliary factors. The differing cellular and temporal localization of RBPs (Ameur et al., 2011;Hao and Baltimore, 2013) may explain the different dynamics regulating alternative and constitutive splicing: whereas constitutive splicing mainly occurs cotranscriptionally, alternative splicing mainly occurs post-transcriptionally (Tilgner et al., 2012). For recent mechanistic models of splicing regulation through RBPs, see (Witten and Ule, 2011).

Alternative splicing can also be regulated in a manner totally independent of auxiliary splicing factors (Yu et al., 2008). Splicing silencer sequences regulate alternative splicing when competing 5′ splice sites are present in the same RNA molecule (**Figure 3b**). 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 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 (Wen et al., 2010).

Alternative splicing can also be regulated by RNA secondary structures (**Figure 3c**). Short-range RNA secondary structures can mask primary *cis* elements such as the acceptor and donor sites or the polypyrimidine tract (Pervouchine et al., 2012;Shepard and Hertel, 2008). This has 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 (Warf et al., 2009;Yuan et al., 2007). Long-range secondary structures bring distant splice sites into closer proximity, facilitating alternative splicing, and are associated with weak alternative 3′ splice sites (Pervouchine et al., 2012). 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 (Pervouchine et al., 2012). Recently developed high-throughput techniques combine nuclease digestion (Kertesz et al., 2010) or chemical probing (Lucks et al., 2011) 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 (Ding et al., 2014;Wan et al., 2014), by reporting the presence of strong secondary structures at 5′ splice sites that correlate with unspliced introns. 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 (Giudice et al., 2014;Kim et al., 2013) and cell, tissue (Pimentel et al., 2014), and species specificity (Gracheva et al., 2011). 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 (Buljan et al., 2012).

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 (Costa et al.,

2013;Pistoni et al., 2010).

1.3 3’ End maturation: alternative polyadenylation

Another step in mRNA processing is the process of polyadenylation (Danckwardt et al., 2008). The use of alternative polyadenylation (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 4a**) or in the length of their 3′-UTRs [tandem polyadenylation sites (PASs)] (**Figure 4b**). The impact of APA on the regulation of gene expression can be extended through effects on transcript localization (Andreassi and Riccio, 2009), stability, and translation efficiency (Fabian et al., 2010) 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 (**section 2.2.1, this Chapter**). 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 (Derti et al., 2012;Ozsolak et al., 2010;Shepard et al., 2011), 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 (Djebali et al., 2012). APA within the same last exon (tandem 3′-UTRs) is the most abundant type of APA (Shepard et al., 2011). Intronic APA events are reported less frequently and thousands of intronic PASs are usually suppressed (Yao et al., 2012). 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 (Ji et al., 2011) (**Chapter 2**). 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 (Ni et al., 2013). 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 (Hogg and Goff, 2010), or loss of AU-rich elements (AREs), which leads to ARE-directed mRNA degradation (Ji et al., 2011). However, there are many exceptions to the general rule, as proteins that bind to the 3′-UTR can also stabilize mRNAs (Gupta et al., 2014;Ray et al., 2013;Spies et al., 2013).

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) (Shepard et al., 2011) or poly(A)-binding protein nuclear 1 (PABPN1) lead to transcriptome-wide shortening of 3′-UTRs, corresponding to an increased preference for non-canonical polyadenylation signals (**Figure 4c**) (**Chapter 2**) (Jenal et al., 2012;Martin et al., 2012).

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 (Shepard et al., 2011;Smibert et al., 2012;Ulitsky et al., 2012). 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 (Hafez et al., 2013). A-rich sequences have also been reported upstream of cleavage sites for transcripts lacking canonical motifs (Nunes et al., 2010).

Numerous studies based on expressed sequence tags and microarrays have previously shown the biological relevance of APA (Tian et al., 2005;Yan and Marr, 2005). A study based on expressed sequence tags comprising 42 human tissues (Zhang et al., 2005) 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 (Ni et al., 2013). 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 (Ji et al., 2009). By contrast, shortening was observed during proliferation (Sandberg et al., 2008) and during reprogramming of somatic cells (Ji and Tian, 2009). 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 (Miura et al., 2013;Smibert et al., 2012;Ulitsky et al., 2012). Modulation of APA has also been widely observed during proliferation, differentiation, and development (Hoque et al., 2013;Li et al., 2012;Mangone et al., 2010;Shepard et al., 2011).

Widespread alteration of APA profiles has been observed in several diseases. Many studies have reported shortening of 3′-UTRs in cancer (Fu et al., 2011;Lin et al., 2012;Mayr and Bartel, 2009), linked to extensive upregulation and activation of oncogenes. More recently, altered APA profiles have been linked to muscle disorders such as myotonic dystrophy (Batra et al., 2014) and oculopharyngeal muscular dystrophy (**Chapter 2**).

1.4 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 (Kochetov, 2008), as translation may start at uORFs or at alternative ORFs (aORFs) (**Figure 5a, 5b**). 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 (Fritsch et al., 2012), whereas non-overlapping uORFs affect the translation of pORFs in various ways (Wethmar, 2014): 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 (Vanderperre et al., 2013). 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 [(Michel et al., 2012).

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 assessed via ribosome profiling (Ingolia et al., 2012). Deep sequencing of RNA fragments protected by ribosomes

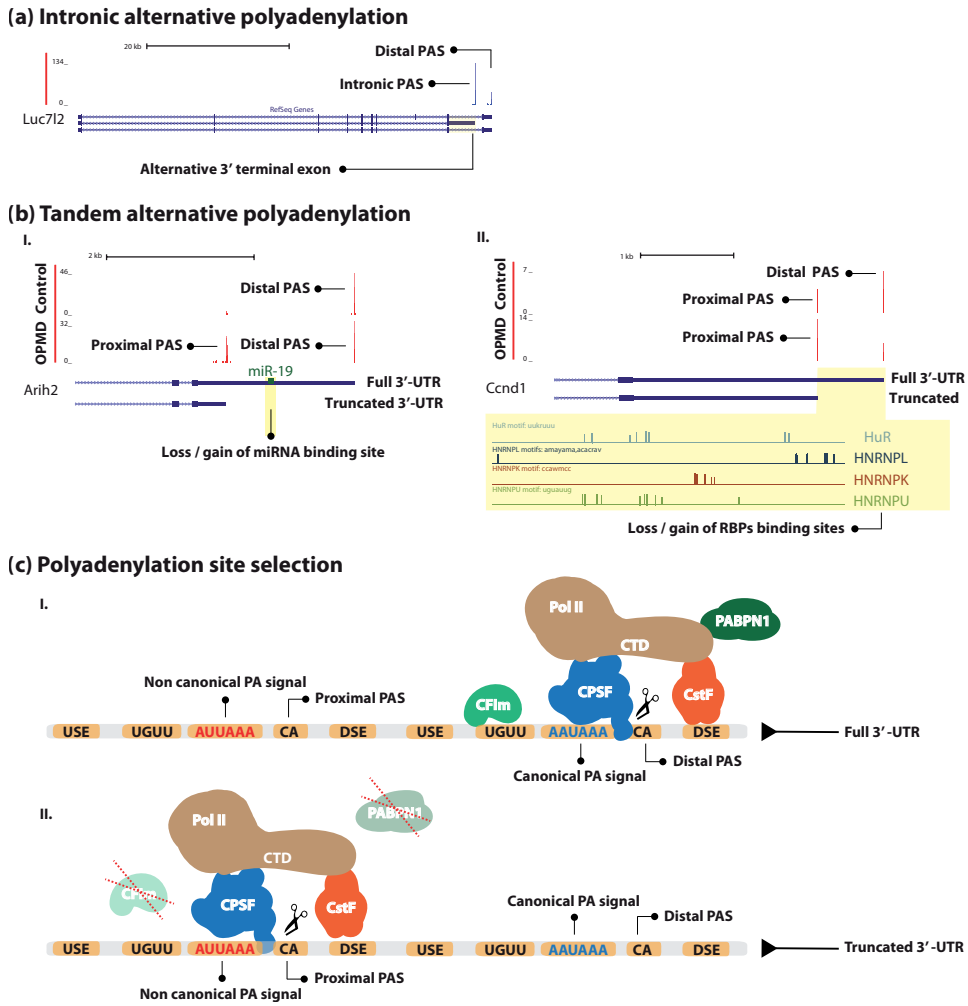

Figure 4. Alternative polyadenylation. (a) Data from a Poly(A)-seq experiment [this thesis, Chapter 2] showing alternative polyadenylation (APA) in the intron of Luc7l2 gene, leading to an intronic proximal PAS located in a different terminal exon, giving rise to transcript variants with different ORFs. **(b)** Two examples of tandem APA in muscle tissue from a mouse model for Oculopharyngeal muscle dystrophy (OPMD) [this thesis, Chapter 2]. In the Arih2 gene (I) both the distal and the proximal PASs can be used in the disease state. The recognition of a proximal PAS leads to shortening of the 3'-UTR and loss of a miRNA binding site, causing an increase in transcript levels. In the Ccnd1 gene (II) the 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 [this thesis, Chapter 2]. Common sequences in the 3'-UTR that regulate polyadenylation are the upstream sequence element (USE), the UGUU sequence recognized by the Cleavage Factor 1 (CFIm), the polyadenylation signal (PA) recognized by the Cleavage and Polyadenylation Specifity Factor (CPSF) and the downstream sequence element (DSE) recognized by the Cleavage Stimulation Factor (CstF). CPSF and CstF are brought to the RNA by the RNA polymerase II (Pol II), together with the Poly(A) Binding Protein Nuclear 1 (PABPN1), through its C-terminal domain (CTD). Generally, CPSF recognizes the canonical PA signal and cut at a distal polyadenylation site (PAS), at a CA dinucleotide (I). If PABPN1 or CFIm are present at a lower concentration, the CPSF recognizes non-canonical (weaker) PA signals (II) and cuts at proximal PASs, leading to the formation of transcripts with truncated 3'-UTRs.

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 (Calvo et al., 2009;Ingolia et al., 2011;Lee et al., 2012). 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 (Wasinger et al., 2013).

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 (Ingolia et al., 2011;Lee et al., 2012). These non-canonical 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 (Menschaert et al., 2013). 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 (Kozak, 2005), 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 (Lee et al., 2012).

Both aORFs and uORFs can give rise to ORFs with reading frames different from the pORFs, a phenomenon known as dual coding (Michel et al., 2012). 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) (Michel et al., 2012).

The extent to which mRNA levels explain differences in protein abundance is still debated. Although some studies have reported a poor correlation (Maier et al., 2009) – in the range of approximately 40% of protein levels explained by mRNA levels (Lundberg et al., 2010;Schwanhausser et al., 2011;Tian et al., 2004;Vogel et al., 2010) or even less than 20% (Ingolia et al., 2009) – others claim a much higher correlation of up to approximately 80% (Li et al., 2014). 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% (Ingolia et al., 2009;Wang et al., 2013b). Nevertheless, a study that compared changes

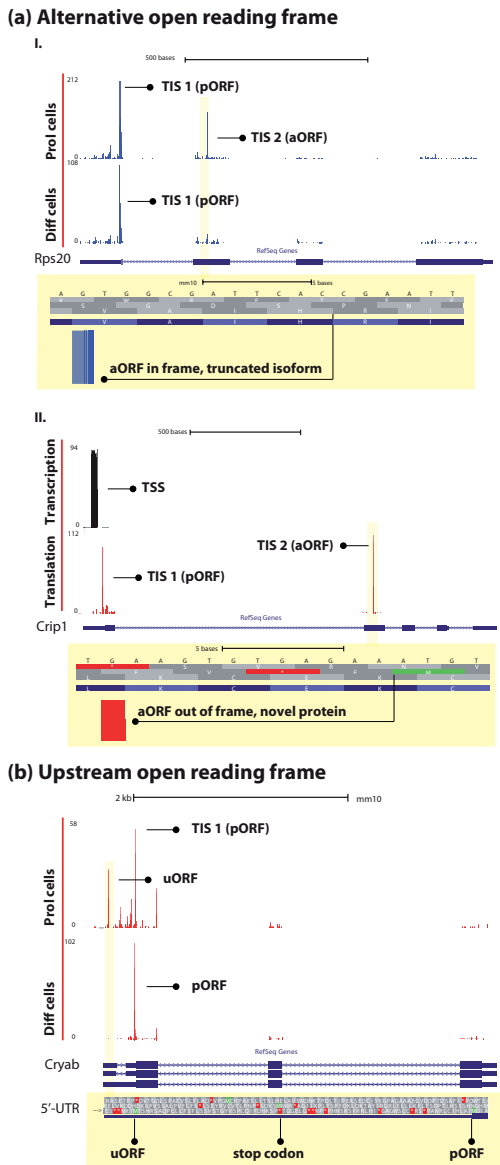

Figure 5. Alternative translation initiation. Alternative translation initiation sites (TISs) detected by ribosome profiling [this thesis, Chapter 4]. **(a)** Examples of alternative TISs leading to alternative open reading frames (aORFs) in frame (I) or out-of-frame (II) with the primary ORF. In the Rps20 gene (I) a switch in TIS usage occurs during cell differentiation. Proliferating cells use two TISs, one corresponding to the annotated start codon and the other corresponding to an alternative open reading frame, 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, deepCAGE) 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 to the pORF, indicating the presence of a dual coding region. **(b)** Examples of alternative TIS leading to an upstream open reading frame (uORF) in the Cryab gene. Proliferating cells use two TISs, one located in the 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. The extension of the 5'-UTR leads to the translation of an upstream open reading frame (uORF), with a canonical AUG codon and ending before the start codon of the primary open reading frame (pORF), negatively regulating translation.

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 (Tebaldi et al., 2012). 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.

1.5 Transcription, RNA processing, and translation: interdependent processes

The molecular machineries involved in transcription and RNA processing are spatiotemporally coupled. Co-transcriptional regulation of capping, splicing, and polyadenylation has been extensively described (Auboeuf et al., 2005;Bentley, 2014). 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 (Hsin and Manley, 2012). There is ample support of the coupling between transcription and splicing. Splicing predominantly occurs during transcription (Djebali et al., 2012;Tilgner et al., 2012), 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 (Brown et al., 2012). 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 may 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 (Dujardin et al., 2013) or proximal PASs (Hazelbaker et al., 2013;Pinto et al., 2011) and the choice of TSS is linked to a specific splicing pattern (Benson et al., 2012;Huang et al., 2009) or to the use of specific PASs (Huang et al., 2012;Ji et al., 2011;Nagaike et al., 2011).

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 (Martinson, 2011)) and that 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 (Berget, 1995). 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 (Shi et al., 2009). 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 (Lee and Tarn, 2014).

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 6a, 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 (Katz et al., 2010). An increase in proximal PAS use was also observed after alteration of Nova, a RBP involved in alternative splicing (Licatalosi et al., 2008).

High CLIP tag density surrounding proximal PASs has also been observed for the RBPs MBNL1 and MBNL2 (**Figure 6a, bottom**), which are known to regulate splicing (Wang et al., 2012), 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 (Batra et al., 2014).

Whether alternative splicing is also coupled to non-tandem APA remains 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 (Kaida et al., 2010), or enhanced in the case of suboptimal splicing (Tian et al., 2007). The coupling observed in this case represents kinetic competition between splicing and polyadenylation (Luo et al., 2013).

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 (Bava et al., 2013) (**Figure 6b**). Interestingly, CPEB1 can also regulate alternative splicing. CPEB1 prevents recruitment of the splicing factor U2AF65 to the 3′ splice 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.

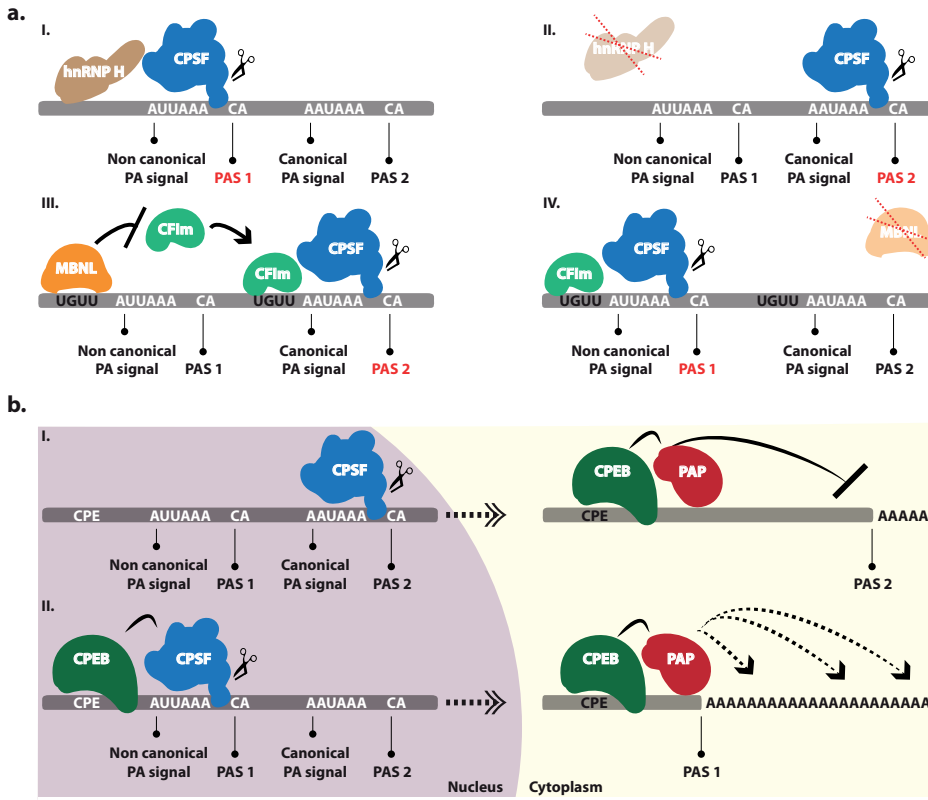

Figure 6. Coupled regulatory mechanisms. (a) Tandem alternative polyadenylation (APA) regulated by splicing factors. The RNA binding proteins hnRNP H and MBNL regulate APA in opposing ways. In the presence of hnRNP H (I), the Cleavage and Polyadenylation Specifity Factor (CPSF) binds weaker non-canonical polyadenylation (PA) signals and cuts at proximal poly(A) sites (PAS 1), leading to shortening of the 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 non-canonical PA signals, blocking the binding of the Cleavage Factor 1 (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 usage of 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 proximal PAS (PAS 1) and shortening of the 3'-UTR. **(b)** Coupling of APA and translation. In the nucleus, in the absence of the 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 non-canonical 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 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 poly(A) tails of transcripts bearing proximal PASs (PAS1) (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 longer due to the 3'-UTR lengthening in transcript bearing distal PAS (PAS 2).

2. RNA sequencing: from Tag-based profiling methods to resolving complete transcript structure

Numerous next-generation sequencing (NGS)-based RNA profiling methods are nowadays available to specifically investigate different levels of regulation. Whereas some RNA sequencing methods focus on a particular region of the transcript and are zooming in on specific RNA processing events, others provide a more comprehensive picture of the transcript, simultaneously characterizing different processing events (**Figure 7**). In this perspective, we can classify RNA sequencing methods into two categories: (1) tag-based methods, where only a short fragment (tag) at a defined position in each RNA molecule is sequenced, and (2) shotgun methods, where the molecule is divided and sequenced in multiple fragments and reconstruction of the original transcript is attempted through computational and statistical approaches (**Figure 8**). A completely different categorization is needed for RNA sequencing methods based on the PacBio sequencing platform. PacBio long-read sequencing provides full-length transcript sequencing, allowing an exact characterization of the structure of the transcript (Koren et al., 2013;Sharon et al., 2013). In this way, different RNA processing events can be simultaneously detected and specifically assigned to a certain transcript, without the ambiguity faced in all other shotgun methods developed for short-read sequencing platforms.

It is important to note that each of these methods capture RNA molecules in different ways, some rely on the presence of the 5′-cap or the poly(A) tail, others allow a full sampling of the transcriptome by capturing also non-capped and non-polyadenylated molecules. The transcripts detected by different techniques are therefore only partially overlapping. Another issue to consider is the transcript's orientation. While all tag-based methods are strand specific, meaning that they preserve information about the transcript's orientation, shotgun methods may be strand specific or not strand specific. Strand specificity is important to determine the exact gene expression levels in the presence of antisense transcription.

These advanced RNA sequencing methods and platforms generate a huge amount of data, giving us the possibility to understand the complexity of the transcriptome and its fine regulation. RNA sequencing methods have been adapted for the most common DNA sequencing platforms [HiSeq systems (Illumina), 454 Genome Sequencer FLX System [Roche], Applied Biosystems SOLiD (Life Technologies), IonTorrent (Life Technologies)]. These platforms require initial reverse transcription of RNA into cDNA. Conversely, the single molecule sequencer HeliScope (Helicos BioSciences) is able to use RNA as a template for sequencing (Ozsolak et al., 2009;Ozsolak et al., 2010) and a few studies have shown its potential (Geisberg et al., 2014;Graber et al., 2013;Moqtaderi et al., 2013;Sherstnev et al., 2012). A proof of principle for direct RNA sequencing on the PacBio RS platform has also been demonstrated (Pacific Bioscience). However, direct RNA sequencing technologies are currently not available to regular customers.

The sequencing platforms differ also in the number of reads generated, leading to a difference in sensitivity. While common short-read platforms can generate millions of reads (http://res.illumina.com/documents/products/appnotes/appnote_hiseq2500.pdf), allowing an accurate quantitative analysis of high and low abundant transcripts, PacBio currently yields ~50,000 long reads (http://files.pacb.com/pdf/PacBio_RS_II_Brochure.pdf), restricting the number of transcripts that can be detected, unless multiple runs are performed (Au et al., 2013;Sharon et al., 2013;Steijger et al., 2013).

To correctly interpret sequencing data and reach a full understanding of the hidden biological meaning in it, a parallel development of statistical and computational approaches is fundamental. Numerous algorithms have been developed to detect differentially expressed genes and spliced

variants. For an extensive comparison of some of the most commonly used methods, and for a general overview of the computational challenges, we refer to (Garber et al., 2011;Soneson and Delorenzi, 2013;Steijger et al., 2013). Moreover, dedicated algorithms to identify switches between polyadenylation (**Chapter 2**) (Katz et al., 2010) or transcription start sites (**Chapter 4**) (Balwierz et al., 2009;Frith et al., 2008) have been developed.

2.1 Tag-based methods

In tag-based methods, each transcript is represented by a unique tag. Initially, tag-based approaches were developed as a sequence-based method to measure transcript abundance and identify differentially expressed genes, assuming that the number of tags (counts) directly corresponds to the abundance of the mRNA molecules. The reduced complexity of the sample, obtained by sequencing a defined region, was essential to make the Sanger-based methods affordable. When NGS technology became available, the high number of reads that could be generated facilitated differential gene expression analysis. A transcript length bias in the quantification of gene expression levels, such as observed for shotgun methods (Gao et al., 2011;Zheng et al., 2011), is not encountered in tag-based methods. This makes tag-based method a potentially less biased approach when studying gene expression. Moreover, all tag-based methods are by definition strand specific.
Recently, an increased interest in the determination of transcripts' structure led to the development of numerous directed tag-based strategies which aim to precisely define 3′ and 5′ transcript ends. We will refer to them as 3′ end sequencing and 5′ end sequencing methods.

2.1.1 3'-End sequencing

3′ end sequencing methods specifically focus on the end of the transcript, allowing the detection of transcripts which differ in the 3′-terminal exon used or in the length of their 3′ untranslated region (3′-UTR). Different 3′ ends arise from alternative polyadenylation of pre-mRNAs (Danckwardt et al., 2008;Legendre and Gautheret, 2003;Shi et al., 2009).

A variety of 3′ end sequencing methods have been developed in the last years, from serial analysis of gene expression (SAGE)-like methods to more dedicated protocols, where the detection of the actual polyadenylation site used is even more precise. Here some of these methods are described, focusing the level of precision in which polyadenylation sites are determined.

DeepSAGE (Nielsen et al., 2006) represents the first high-throughput tag-based method developed to generate tags at the most 3′ end of a transcript. DGE ('t Hoen et al., 2008), Tag-Seq (Morrissy et al., 2009) and HT-SuperSAGE (Matsumura et al., 2010) are improved versions which have been adapted to different sequencing platforms. All these approaches are based on the SAGE method described by Velculescu et al. (Velculescu et al., 1995). Minor differences characterize these techniques, such as the length of the tag (21 or 25–26 nt), the restriction enzymes used to release the 3′ end of a transcript and generate a unique tag (NlaIII/MmeI or NlaIII/EcoP15I), and the sequencing platform used. Except for these minor differences, the steps necessary to generate a sequencing library are similar (**Figure 9a**).The first steps consist in capturing all polyadenylated transcripts and converting the RNA molecules into double-stranded cDNA molecules. The cDNA molecules are then cut at the most 3′ CATG by enzymatic digestion and ligated to a 5′ adapter, which introduces a recognition site for a specific restriction enzyme (MmeI/EcoP15I). A second digestion, downstream of the incorporated restriction site, produces a short fragment (tag of 21 or 25–26 nt) which is then ligated to a 3′ adapter. Both adapters make the cDNA tag suitable for amplification and high-throughput sequencing.

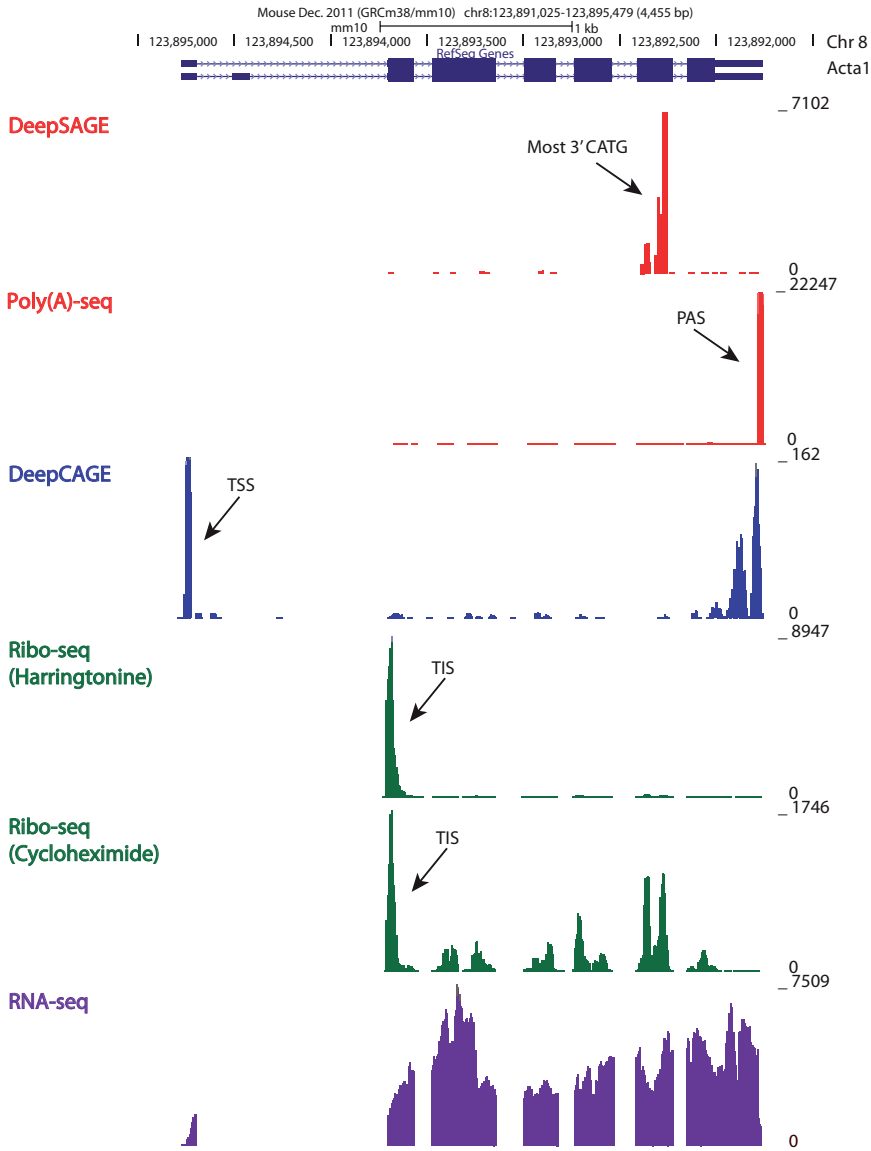

Figure 7. A screenshot from UCSC Genome Browser (http://genome.ucsc.edu) displaying the different regions sequenced by tag-based and shot-gun methods in Acta1 gene. The y-axis represents the coverage, corresponding to the number of reads mapping at each location. Six independent traces are shown. The top two traces (in red) show a peak at the most 3' CATG site and at the exact polyadenylation site (PAS, indicated by an arrow) detected by DeepSAGE and Poly(A)-seq, respectively. The third trace (in blue) shows a peak at the transcription start site (TSS, indicated by an arrow) detected by DeepCAGE. The fourth trace (in green) shows a peak at the translation start site (TIS, indicated by an arrow) detected by ribosome profiling based on harringtonine treatment. The fifth trace (also in green) shows a major peak at the detected translation start site (TIS, indicated by arrow) and a lower coverage at each translated exons, detected by ribosome profiling based on cycloheximide treatment. The last trace (in purple) shows a typical RNA-seq profile, where all exons and untranslated regions are detected. On top of the coverage tracks, the RefSeq gene track shows two transcript variants for Acta1, with exons shown as thick boxes, untranslated regions as thin boxes and introns as consecutive arrows.

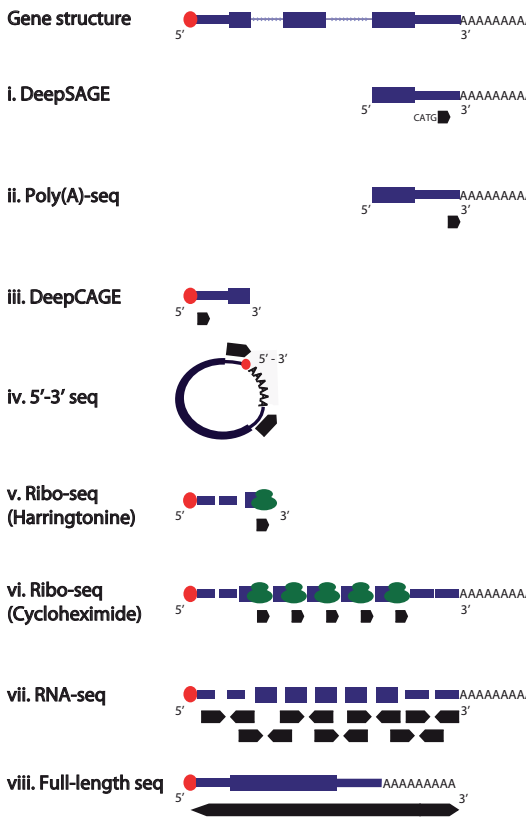

Figure 8. Schematic representation of sequencing reads generated by tag-based (i-iv), shot-gun (v-vii) or full-length (viii) sequencing. Thick black arrows indicate the sequenced reads. Paired-end reads are displayed by two opposite black arrows. Red circles indicate the 5′ cap structure. Ribosomes are displayed in green. The complete gene model is displayed on top, with exons shown as thick boxes, untranslated regions as thin boxes and introns as consecutive thin arrows.

Different studies have shown that SAGE-like methods are suitable to detect alternative polyadenylation events (**Chapter 3**) (‘t Hoen et al., 2008;Hestand et al., 2010;Ji et al., 2009;Nordlund et al., 2012). Nonetheless, the possibility to distinguish transcripts with different 3′ end relies on the presence of a restriction site in the sequence between the two alternative polyadenylation sites. All transcripts with alternative 3′ ends lacking restriction sites in between the polyadenylation sites are, therefore, missed. The same applies for transcripts which do not contain that specific restriction site. According to RefSeq human transcript database, ~1% of the transcripts lack an NlaIII recognition site, meaning that almost 1000 transcripts are not accessible to SAGE-like approaches (Unneberg et al., 2003). Another limitation of these methods is that they do not give information regarding the position of the polyadenylation site.

To overcome the limitations observed in all SAGE-like methods, several dedicated protocols have been developed to specifically characterize polyadenylation sites and quantify their relative usage genome wide (**Chapter 2**) (Beck et al., 2010;Derti et al., 2012;Fox-Walsh et al., 2011;Fu et al., 2011;Hoque et al., 2013;Jan et al., 2011;Jenal et al., 2012;Lin et al., 2012;Martin et al., 2012;Ozsolak et al., 2009;Ozsolak et al., 2010;Pelechano et al., 2012;Shepard et al., 2011;Wang et al., 2013a;Wilkening et al., 2013;Yoon et al., 2012) (**Figure 9b, 9c**). These methods do not rely on the presence of a specific restriction enzyme site and therefore detect all polyadenylation sites.

Limitations in the detection of the exact polyadenylation site location and biased quantifications

may arise due to various steps involved in the preparation of the sequencing library. Oligo(dT) priming, DNA or RNA ligase-mediated adapter ligation, reverse transcription and amplification represent the main sources of bias.

The available poly(A) site sequencing protocols may differ in the level of precision in which the polyadenylation site is determined, in the number of possible biasing steps introduced and in the number of false polyadenylation sites detected, mainly arising from internal priming events.

The main technical differences between the reviewed methods are summarized in **Table 1**. Internal priming events remain one of the limitations of all methods based on oligo(dT) priming (Derti et al., 2012;Elkon et al., 2012;Fox-Walsh et al., 2011;Fu et al., 2011;Martin et al., 2012;Shepard et al., 2011;Wilkening et al., 2013). Internal priming can occur due to priming of oligo(dT) on internal A-rich regions of the transcript, yielding artifacts which are difficult to distinguish from authentic polyadenylation sites.

Different approaches have been taken to minimize internal priming artifacts. In 3P-Seq (Elkon et al., 2012), ligation of a biotinylated double-stranded oligo (containing an overhanging stretch of Ts) to the end of the poly(A) tail is used to eliminate the chance of priming in internal poly(A) stretches. In another method, 3′READS (Hoque et al., 2013), discrimination of 3′ poly(A) tails from internal A-rich sequences is achieved by capturing fragmented RNA onto beads coated with a chimeric oligonucleotide consisting of thymidines (Ts) at the 5′ and uridines (Us) at the 3′ end (CU5T45). Subsequently, RNaseH digestion is used to release the molecules from the beads and to remove most of the As of the poly(A) tail. This method enriches for RNAs with longer A stretches. Wang et al. (Wang et al., 2013a) used a computational analysis to distinguish authentic polyadenylation sites from potential internal priming events based on the distinct pattern of nucleotide composition of the 3′ end region. This method is compatible with any 3′ end sequencing technology.

Next to differences in dealing with the internal priming issue, protocols display different degrees of resolution in the identification of the exact polyadenylation sites. If sequencing starts from the 5′ end of the library construct (Beck et al., 2010;Elkon et al., 2012;Fox-Walsh et al., 2011;Jenal et al., 2012), there is a chance that a fraction of reads will not reach the polyadenylation site. If sequencing starts at the very 3′ end of the library construct (Fu et al., 2011), including the stretch of As, other issues may arise, such as polymerase slippage or mispriming of the sequencing oligo, due to the presence of the homopolymeric stretch. The 3P-Seq approach described above (Jan et al., 2011) overcomes this last issue by digesting the poly(A) tail before incorporating the adapters necessary for amplification and sequencing. The PAS-Seq [46] approach avoids sequencing the poly(A) tail using a sequencing primer with an oligo(dT) extension at the 3′ end. Another method which avoids sequencing through the poly(A) tail is described by Wilkening et al. (Wilkening et al., 2013). In this method, named 3′T-fill, the poly(A) stretch is filled in with dTTPs before the sequencing reaction starts.

A more direct approach is described in **Chapter 2**. This method, based on the HeliScope single molecule sequencer technology, allows to start sequencing directly after the 5′ end of the poly(A) tail, thus at the exact polyadenylation site. Molecules are directly hybridized, through their poly(A) tail, to a flow cell containing oligo(dT) probes. The poly(A) stretch downstream of each polyadenylation site makes the second-strand cDNA molecules directly amenable for sequencing, with the advantage that the first nucleotide on the 5′ end of each sequenced molecule represents the poly(A) addition site. An even less biased approach is described by Ozsolak et al. (Ozsolak et al., 2009;Ozsolak et al., 2010), and is based on direct RNA sequencing (DRS). All poly(A)-containing RNAs are sequenced starting from the polyadenylation site, without reverse transcription, right after one single enzymatic reaction consisting in the addition of dideoxy terminators at the end of the poly(A) tail. This is done to prevent

extension at the 3′ end of mRNAs which are not perfectly hybridized to the poly(T) stretch of the flow cell surface. Accurate detection of polyadenylation sites can also be achieved on the PacBio-RS single molecule sequencing platform. Here, transcripts are converted into a circular double-stranded DNA template capped by hairpin loops at both 3′ and 5′ ends (Travers et al., 2010). Since the full-length cDNA molecule is incorporated in a circular template, the poly(A) tail will be present, allowing the detection of the exact position of the polyadenylation site and the length on the poly(A) tail.

Methods relying on enzymatic ligation of adapter sequences to RNA molecules (such as A-Seq (Martin et al., 2012), 3P-Seq (Jan et al., 2011) and 3′READS (Hoque et al., 2013)), are known to be non-random, compromising quantification (Hafner et al., 2011;Zhuang et al., 2012). Ligation steps may be avoided using the template switch reverse transcription approach. Methods such as PAS-Seq (Shepard et al., 2011), SAPAS (Fu et al., 2011) and PolyA-seq (Derti et al., 2012), use this approach to incorporate known sequences at both ends of cDNA molecules during first-strand synthesis. Despite this, other artifacts may be introduced, e.g., through a process called strand invasion (Tang et al., 2013).

2.1.2 5′ End sequencing

5′ end sequencing methods can be considered as a mirror approach of the 3′ end sequencing methods, as they generate tags at the 5′ end of a transcript. 5′ end sequencing methods have been developed to specifically identify transcription start sites (TSS) and (proximal) promoters. The knowledge of the exact position of a transcription start site can also be used to investigate promoter usage and to identify transcription factor binding sites in these promoters (Vitezic et al., 2010).

The detection of the exact transcription start sites is highly important since alternative transcription start sites can lead to the formation of protein isoforms with totally different biological functions. Alternatively, shorter or longer 5′-UTRs may influence the efficiency of protein translation (Barbosa et al., 2013;Morris and Geballe, 2000).

The number of 5′ end sequencing methods available is restricted compared to the number of 3′ end sequencing approaches. A possible reason might be that the first method published, named DeepCAGE (de Hoon and Hayashizaki, 2008;Suzuki et al., 2009;Valen et al., 2009), already efficiently detected 5′ ends of transcripts, with a high level of precision. Whereas SAGE-like methods are restricted to the use of restriction enzymes and therefore to the presence and location of restriction sites, CAGE-like methods are based on the 5′ cap structure of a transcript, and can theoretically detect all capped 5′ ends of mRNA molecules. On the other hand, these methods are not suitable for non-capped transcripts.

DeepCAGE represents an improved NGS version of the previously published CAGE protocols (Kodzius et al., 2006;Shiraki et al., 2003). This technique makes use of the cap trapper method (Carninci et al., 1996) to capture the 5′-cap structure of RNA molecules. Trapped RNAs are converted to cDNAs, and an adapter is ligated to the 3′ end of the cDNAs. The adapter is used to introduce a recognition site for a specific restriction enzyme (Mme1 or EcoP15I), which is able to cut 21 or 25–27 nt downstream, generating the tag desired. After synthesis of the second cDNA strand, the double-stranded cDNA fragment is ligated to a second adapter, necessary for amplification before sequencing. DeepCAGE libraries have been analyzed on common DNA-based sequencing platforms (Illumina, 454) but also on the Helicos single molecule sequencer (Kanamori-Katayama et al., 2011;The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014). The Helicos-based DeepCAGE method (called Heliscope-CAGE) is a simplified method which consists of only three main steps: first-strand cDNA synthesis, 5′-cap trapping and poly(A) tailing of the 3′ ends. Heliscope-CAGE has the advantage to avoid second-strand synthesis, amplification, ligation, and digestion, reducing possible quantification

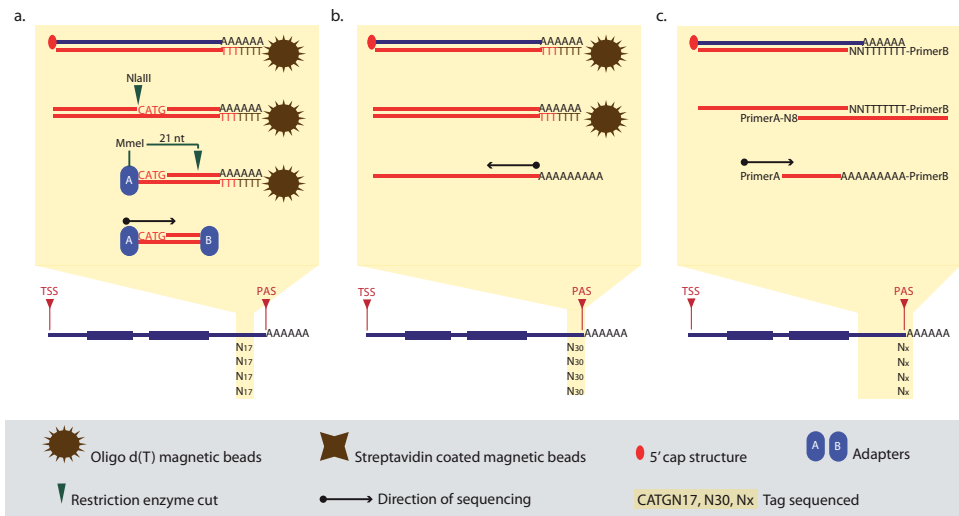

Figure 9. 3'end sequencing methods. (a) In DeepSAGE poly(A)+ RNAs are captured by oligo d(T) magnetic beads and reverse transcribed. cDNA is digested with NlaIII and adapter A is ligated. A second digestion with MmeI generates a 21-bp tag, and adaptor B is ligated to the 3' end. The construct is amplified and sequenced from adapter A. **(b)** In HeliScope-based Poly(A)seq poly(A)+ RNAs are captured by oligo d(T) magnetic beads and reverse transcribed. Second strand cDNA molecules are hybridized to the Helicos flowcell and sequenced starting precisely at the polyadenylation site. **(c)** In MAPS first and second strand synthesis are carried out using oligo d(T) linked to primer B and random primers linked to primer A, respectively. The construct is amplified and sequenced starting from the 5' end.

	PAS-Seq	SAPAS	PolyA-seq	A-seq	MAPS	3'Seq	3P-Seq	3'READS	3'T-fill	de Klerk et al.	Ozsolak et al.
Reverse transcription	▲	▲	▲	▲	▲	▲	▲	▲	▲	▲	
oligo(dT)-based	▲	▲	▲	▲	▲	▲			▲		
DNA ligase-mediated adapter ligation						▲			▲		
RNA ligase-mediated adapter ligation				▲			▲	▲			
Sequencing starts next or at PAS	▲ (*)		▲ (*)	▲ (*)			▲ (*)	▲ (*)	▲ (**)	▲ (**)	▲ (**)
Sequencing starts at poly(A) tail		▲									
Sequencing starts at 5' end					▲	▲					

(*) Sequencing starts next to PAS
(**) Sequencing starts at exact PAS

Table 1. Polyadenylation sites (PASs) sequencing protocols

bias that might arise from each of these steps. Molecules can be hybridized to the flow cell and sequencing can start directly after filling up the poly(A) tail.

Both DeepCAGE and HeliscopeCAGE are based on the cap-trapper method. A different approach is described by Salimullah et al. (Salimullah et al., 2011) in their protocol named NanoCAGE, initially developed by Plessy et al. (Plessy et al., 2010). NanoCAGE uses the template-switching method for reverse transcription. Compared to cap-trapper-based methods, an advantage of this approach is the low amount of starting material (~50 ng instead of ~5 µg) required and the possibility to sequence not only a single tag at the transcription start site, but also a second tag in a downstream exon. The position of the second tag is random, since it depends on the position of the random primer used during second-strand synthesis. Paired-end sequencing of NanoCAGE libraries will therefore provide extra information on the structure of the transcript compared to DeepCAGE methods. The same approach is used in the method called CAGEscan (Plessy et al., 2010). The limitation of NanoCAGE and

CAGEscan lies in the possible artifacts introduced by template switching (Tang et al., 2013).

All CAGE-like methods discussed so far are limited in their ability to correctly detect alternative transcription start sites, due to a phenomenon called 'exon painting' (Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project, 2009;Hestand et al., 2010;Kanamori-Katayama et al., 2011). The term 'exon painting' is used to indicate the presence of multiple CAGE peaks in exonic regions, next to the expected CAGE peak at the 5′ end of the transcript. This phenomenon is not caused by a technical artifact, but more likely arises from recapping of processed transcripts (Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project, 2009). To limit the number of false alternative transcription start sites detected, only TSS in intergenic regions are considered (Hestand et al., 2010).

2.1.3 5′ and 3′ End sequencing

The detection of alternative transcription start sites and alternative polyadenylation sites by tag-based methods, which focus on the 5′ and 3′ end of a transcript, respectively, is a proven method to characterize transcript structure. Nevertheless, the full information about transcript structure is missing. To overcome this limitation, tag-based methods able to detect the co-occurrence of a specific transcription start site and a polyadenylation site has been developed.

Methods able to determine both ends are called RNA-PET (Ruan and Ruan, 2012) and TIF-Seq (Pelechano et al., 2013). RNA-PET is a paired-end tag approach, where detection of both 3′ and 5′ ends occurs through paired-end sequencing. The initial step consists of capturing the 5′-cap structure by cap-trapper and synthesizing full-length cDNA. The double-stranded cDNA molecules are ligated to specific adapters which allow the formation of a circular template and the introduction of two restriction sites for EcoP15I. The restriction sites are inversely oriented, allowing the double cleavage of the PET construct, yielding a fragment of 27 nt from both the 3′ and the 5′ ends.

In TIF-Seq full mRNAs are first ligated to a single-strand oligo by oligo-capping. Then mRNAs are converted to cDNAs by reverse transcription and amplified using biotinylated primers. The double-stranded cDNA molecules are circularized through an intramolecular ligation, and fragmented by sonication. Fragments containing both 3′ and 5′ ends are captured by streptavidin-coated beads and ligated to adapters for amplification and paired-end sequencing.

An advantage of both paired-end tag approaches is the ability to detect fusion transcripts. On the other hand, generation of full-length cDNAs from long transcripts still represents a technical limitation for any 5′3′-sequencing method.

2.2 Shotgun methods

The advantage of a shotgun, sequence-it-all method, over a tag-based method, is the ability to quantify the expression level of each exon within a transcript, estimate their percent inclusion level and detect (differential) alternative splicing events. However, it is difficult to identify the exact 3′ and 5′ ends of transcripts due to various technical biases (such as random hexamer priming or oligo dT priming) leading to underrepresentation of sequences near 5′ and 3′ ends (Hansen et al., 2010;Roberts et al., 2011).

The term RNA-seq is used to indicate any RNA sequencing method based on a shotgun approach. Numerous protocols have been published so far, which have many steps in common: fragmentation (which can occur at RNA level or cDNA level, where RNA fragmentation appears to introduce less bias (Mortazavi et al., 2008)), conversion of the RNA into cDNA (performed by oligo dT or random primers),

	Mortazavi *et al.*	Lister *et al.*	He *et al.*	Parkhomchuk *et al.*
RNA fragmentation	▲	▲		
cDNA fragmentation			▲	▲
RNA ligase-mediated adapter ligation		▲		
Random hexamers priming	▲		▲	▲
Oligo(dT) priming				▲
Adapter priming		▲		
Bisulfite treatment			▲	
Deoxy-UTP incorporation in dsDNA				▲
Strand-specific		▲	▲	▲

Table 2. RNA-seq protocols

second-strand synthesis, ligation of adapter sequences at the 3′ and 5′ ends (at RNA or DNA level) and final amplification. RNA-seq can focus only on polyadenylated RNA molecules (mainly mRNAs but also some lncRNAs, snoRNAs, pseudogenes and histones (Kari et al., 2013;Lemay et al., 2010;Zheng et al., 2007)) if poly(A)+ RNAs are selected prior to fragmentation, or may also include non-polyadenylated RNAs if no selection is performed. In the latter case, ribosomal RNA (more than 80% of the total RNA pool (Lodish H et al., 2000)) needs to be depleted prior to fragmentation. It is, therefore, clear that differences in capturing of the mRNA part of the transcriptome lead to a partial overlap in the type of detected transcripts. Moreover, different protocols may affect the abundance and the distribution of the sequenced reads (Griebel et al., 2012). This makes it difficult to compare results from experiments with different library preparation protocols.

Whereas all tag-based methods are by definition strand specific, the first RNA-seq methods were not strand specific (Mortazavi et al., 2008), as the orientation of the molecule was lost during random-primed cDNA synthesis. In the last years, numerous strand-specific RNA-seq protocols have been developed (**Table 2**) (Armour et al., 2009;He et al., 2008;Lister et al., 2008;Parkhomchuk et al., 2009;Schaefer et al., 2009). Maintaining strand information is important given the widespread occurrence of antisense transcripts with a, likely regulatory, biological function.

Strand-specific methods can be classified into two categories: (1) RNA-seq methods based on ligation of two different adaptors in a known orientation relative to the 5′ and 3′ ends, and (2) RNA-seq methods based on chemical modification of the RNA, either by bisulfite treatment or by the incorporation of dUTPs during the second-strand cDNA synthesis. In both cases, the non-modified strand is degraded enzymatically. According to a comparative study published by Levin et al. (Levin et al., 2010), where 13 different protocols have been analyzed based on their strand specificity, the coverage along all exons and the accuracy in quantification, the dUTP approach was the best performing protocol. Nevertheless, in all strand-specific RNA-seq protocols a fraction of antisense reads will be generated, for example when RNA molecules fold back on themselves. Depending on the protocol, the percentage of antisense reads from sense transcripts amounts to 1–12% (Levin et al., 2010). Therefore, additional analytical approaches are required to discriminate naturally occurring antisense transcripts from artifacts.

Shotgun sequencing methods have the potential to identify alternative splicing events. Algorithms deriving transcript structure from short reads mostly use a combination of coverage patterns and

exon–exon spanning reads, and read pair information. To be able to detect alternative spliced variants, a certain coverage is necessary. Therefore, low expressed genes will give less information than highly expressed genes, unless a large number of reads are generated. A discussion of these algorithms falls outside the scope of this thesis, but the reader can refer to (Alamancos et al., 2014;Steijger et al., 2013).

2.3 Full-length sequencing

One of the main limitations of all short-read shotgun methods is the inability to directly characterize the structure of a transcript and/or to discriminate different alleles. Additional computational and statistical approaches are required to reconstruct the transcript, and the short fragment sizes limit the reconstruction to local regions of the transcripts.

The PacBio system is the only available platform potentially able to produce reads with a length up to ~30 kb. However, the limitation faced at the moment is the production of full-length double-stranded cDNAs (Sharon et al., 2013).

Different approaches are used to create full-length cDNAs suitable for full-length transcript sequencing. One of the possible approaches is based on template switching, consisting in the addition of a non-templated poly-cytosine tail to the 3′ end of the first-strand cDNA molecule through the terminal transferase activity of the MMLV reverse transcriptase. The addition of a poly-(C) tail allows the hybridization of an adapter with a poly(G) tail if the first-strand cDNA synthesis has reached the 5′ end of the transcript. A disadvantage of this approach is that degraded mRNAs containing a poly(A) tail will also be converted into cDNAs, simply due to the fact that cDNA synthesis starts at the poly(A) tail. Distinction between full-length transcripts and partially degraded transcripts will therefore be impossible.

A different approach based on the isolation of properly 5′-capped RNA molecules is also extensively used. It is based on first-strand cDNA synthesis starting at the poly(A) tail, followed by digestion of unconverted RNAs and capture of the 5′-cap. Only molecules where the cDNA synthesis has reached the 5′ cap will be used for second-strand synthesis.

Minor improvements in cDNA length have been observed in recent template switch-based methods like Smart-seq2 (Picelli et al., 2013), where the majority of the cDNA molecules reach a read length of 2 kb.

Independently from which approach is used to generate full-length cDNAs, for PacBio sequencing these are converted into a SMRTbell library (Travers et al., 2010), consisting of double-stranded cDNA molecules capped by two harpin adapters on both side. The hairpin adapters are used to convert the linear double-stranded cDNAs into circular cDNA molecules, which due to this structure and long-read lengths will be sequenced multiple times by the same polymerase. Fragmentation and amplification steps are not performed, with the advantage that any possible technical artifact commonly faced in most of the current methods is avoided.

Taking into account the actual limitations observed in full-length cDNA preparation, full-length sequencing on PacBio still represents a unique approach to interrogate full transcript structure on a single molecule level (**Chapter 5**). Unfortunately, the number of reads offered by the PacBio technology is limited, and full characterization of a transcriptome requires performing of many runs (Au et al., 2013;Sharon et al., 2013) and is costly.

2.4 Immunoprecipitation-based methods

Whereas previous methods usually reflect steady-state RNA levels, there are also dedicated methods available to monitor active transcription. A first approach is the immunoprecipitation of genomic DNA bound by RNA Polymerase II (Sun et al., 2011). Depending on the antibody used, only transcription initiation complexes are immunoprecipitated or also actively transcribed DNA. Alternatively, nascent RNA molecules can be sequenced by NET-seq (Churchman and Weissman, 2011) (native elongating transcript sequencing). In this approach, the ternary complex formed by the RNA pol II, DNA and RNA is immunoprecipitated. Crosslinking can be avoided due to the stable ternary complex.

RNA immunoprecipitation-based methods are also used to understand how protein–RNA complexes interactions regulate gene expression at transcriptional and post-transcriptional level. Various targeted approaches have been developed to investigate the interaction between RNA-binding proteins and their target RNA molecules (**Table 3**).

HITS-CLIP (Licatalosi et al., 2008) and CLIP-seq (Yeo et al., 2009) represent the first high-throughput methods developed to generate genome-wide RNA–protein interaction maps. Both methods are based on the crosslinking-immunopurification (CLIP) strategy (Jensen and Darnell, 2008;Ule et al., 2003), which relies on the principle that ultraviolet light causes the formation of a covalent bound between RNAs and proteins in direct contact. Cells or tissues can be irradiated in vivo, and after cell lysis the crosslinked RNA–protein complexes can be purified by immunoprecipitation using specific antibodies. To be able to map each binding site, RNA is digested up to a length of ~50 nt, reverse transcribed after RNA adapter ligation, and amplified prior sequencing. In the traditional CLIP method the resolution is low, since the mapped binding sites correspond to the total length of the fragmented co-purified RNAs. Another limitation is represented by the low efficiency of crosslinking using UV light at a wavelength of 254 nm. Different approaches, such as PAR-CLIP (Hafner et al., 2010b;Hafner et al., 2010a) and iCLIP (Konig et al., 2010), have been developed to more precisely map the exact binding sites at nucleotide resolution and to increase the efficiency of the crosslinking.

PAR-CLIP (Hafner et al., 2010b;Hafner et al., 2010a) (photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation) is based on the incorporation of photoreactive ribonucleoside analogs (4-thiouridine or 6-thioguanosine) into newly synthesized RNAs. The use of ribonucleoside analogs leads to two advantages: they allow crosslinking with UV light at 365 nm (more efficient than the crosslinking at 254 nm), and they lead to a base transition during reverse transcription (thymidine to cytidine or guanosine to adenosine when using 4-thiouridine or 6-thioguanosine, respectively) which can be used to exactly define the crosslink site at nucleotide resolution.

HITS-CLIP, CLIP-seq and PAR-CLIP face the problem of truncated cDNAs generated during reverse transcription. Reverse transcription can stop due to the presence of undigested peptides which are still crosslinked to the RNA molecules. Truncated cDNAs are usually lost because they cannot be amplified, due to the missing 5′ adapter primer.

iCLIP (Konig et al., 2010) makes use of partial peptide digestion to appositely create truncated cDNA molecules, which can be converted into circular cDNA molecules. The crosslink position can be exactly defined since it corresponds to one nucleotide upstream of the truncation site.

Any of the CLIP methods mentioned above require numerous enzymatic steps which can bias the detection of true binding sites (from RNA and protein digestion, to RNA ligase-mediated adapter ligation, reverse transcription and amplification). Moreover, even though a crosslinking at 365 nm is generally considered more efficient, the efficiency of a crosslink might differ from protein to protein (Kishore et al., 2011). Most of the CLIP-based studies performed so far focus on splicing factors (Konig et al., 2010;Licatalosi et al., 2008;Yeo et al., 2009).

	NET-seq	HITS-CLIP	CLIP-seq	PAR-CLIP	iCLIP
Cross-link UV 254 nm		▲	▲		▲
Cross-link UV 365 nm				▲	
RNA ligase-mediated adapter ligation	▲	▲	▲	▲	▲
Reverse transcription	▲	▲	▲	▲	▲
Photoreactive ribonucleoside analogs				▲	
Identification of precise cross-linked site				▲	▲

Table 3. Immunoprecipitation-based methods

2.5 Ribosome profiling

All methods discussed so far focus on measuring the abundance and characterizing the structure of a transcript, or defining its interaction with RNA-binding proteins. The information derived is therefore restricted to the composition of the transcriptome. However, transcript levels are not necessarily a good approximation of protein levels because the process of translation is also highly controlled, probably to the same extent as transcription or splicing (Plotkin, 2010). Ribosome-associated mRNA levels are a better proxy for protein levels than total mRNA levels (Ingolia et al., 2009).

Ribosome profiling (also called Ribo-seq) (Ingolia et al., 2009;Ingolia, 2010;Ingolia et al., 2012) has been developed to study the process of translation and its efficiency. This method is also often combined with RNA-seq to define untranslated RNAs (e.g., lncRNAs), whether all alternative transcripts are actively translated and to study the extent of regulation at the level of transcription and translation (**Chapter 4**).

Ribosome profiling is a shotgun method based on deep sequencing of ribosome-protected mRNA fragments, which allow to determine which transcript is actively translated at a specific moment in the cell, the rate of translation, the reading frame used and thereby the exact protein product. The technique is based on the observation that ribosomes bound to mRNA molecules protect ~28 nt fragments from nuclease digestion (ribosome footprints). After halting translation, ribosome-bound mRNAs are digested and the ribosome:mRNA complexes (monosomes) are recovered by ultracentrifugation on sucrose gradients or by size-exclusion chromatography. The short protected fragments are released from the monosomes, and converted into a cDNA library, which can be amplified and sequenced. Different variants of the original protocol have been developed to study translational control at different levels. Using drugs arresting ribosome initiation complexes, such as harringtonine or lactimidomycin, it is possible to detect alternative translation start sites or regulatory upstream open reading frames. By inhibiting ribosome translocation with cycloheximide or by thermal freezing, it is possible to quantify the level of translation, to identify the translational reading frame, potential reading frame switches, and to investigate ribosome pausing.

It has been shown that some of the methods commonly used to halt translation may lead to artifacts. Cycloheximide is known to cause a profound accumulation of ribosomes at the translation initiation codon, due do the fact that translation can still initiate while elongation is already blocked (Ingolia et al., 2009). Harringtonine, on the contrary, might fail in halting the ribosomes at the start codon (Lee et al., 2012). No disadvantages have been observed so far when halting translation using lactimidomycin, which currently seems to be the method of choice (Lee et al., 2012).

2.6 From bulk transcriptome to single cell

Large required amounts of input material represent an obstacle when studying rare and heterogeneous cell populations, micro-dissected tissues, subcellular fractions or simply when there is a limited accessible quantity of RNA from patients. Therefore, some RNA profiling methods are limited to bulk transcriptome analysis of large numbers of cells or pieces of tissues.

The targeted approaches, such as the immunoprecipitation-based methods and the ribosome profiling method, require the highest amount of input material, in the range of millions of cells. The suggested amount of RNA for a PAR-CliP experiment ranges between 100 and 400 million cells (Hafner et al., 2010a), but iCLIP experiments can be performed in <10 million cells (Konig et al., 2010), and the same applies for ribosome profiling experiments (Ingolia et al., 2012). None of these approaches has been so far optimized to analyze transcriptome from single cells or from a small population of cells.

PacBio long-read sequencing also requires a high amount of input RNA, in the range of hundreds of thousands of cells. Successful full-length libraries have been generated starting from ~10 μg of total RNA (Sharon et al., 2013) or ~1 μg of poly(A)+ RNA (Au et al., 2013).

Tag-based and shotgun methods have been extensively improved with regards to the amount of starting material. While the older DeepCAGE approach required ~50 μg of total RNA (Valen et al., 2009), the single molecule HeliScopeCAGE method requires only ~5 μg of total RNA (Kanamori-Katayama et al., 2011) and the nanoCAGE approach has been optimized to be used with an amount of total RNA ranging from 10 ng to 1 μg (even though the most reliable results are obtained when using at least 50 ng of total RNA) (Plessy et al., 2010). This allows investigating 5′ ends of transcripts from a small population of cells.

The 3′ end sequencing methods generally require low amounts of input RNA. Even though some poly(A) sequencing methods requires between 10 and 50 μg of total RNA (Fu et al., 2011;Jan et al., 2011;Martin et al., 2012) or between 0.5 and 1 μg of poly(A)+ RNA (Jenal et al., 2012;Shepard et al., 2011), others, such as 3Seq (Wang et al., 2013a), the Helicos-based poly(A) seq (**Chapter 2**), PolyA-seq (Derti et al., 2012) and MAPS (Fox-Walsh et al., 2011), require only between 0.5 and 3 μg of total RNA. The fact that there are no single-cell studies based on poly(A) sequencing does not imply their unfeasibility, given the fact that the sample preparation for some of these methods partially resemble the one for RNA-seq libraries.

RNA-seq remains at the moment the only method which has been used for whole-transcriptome single-cell sequencing.

One of the main challenges in single-cell RNA-seq is the ability to distinguish between biological variation and technical variation, which suffers from biases introduced during cDNA synthesis and amplification. Next to the ambiguity in the quantification, when the starting amount is lowered to single-cell level, it also becomes difficult to detect lowly expressed transcripts (Ramskold et al., 2012). Recently, numerous RNA-seq methods specific for single-cell transcriptome sequencing have been developed to decrease technical variation (Islam et al., 2014;Ramskold et al., 2012), together with statistical methods to distinguish the true biological variability (Brennecke et al., 2013). A comparison of commercially available kits showed that single-cell RNA sequencing can detect the same transcriptome complexity observed with standard RNA-seq on millions of cells (Wu et al., 2014). The advantage of single-cell RNA sequencing over standard RNA-seq on a bulk of cells relies in the possibility to detect expression differences which could be overlooked when looking at a heterogeneous population of cells, such as allele-specific expression (Deng et al., 2014). Even though studies have shown the possibility to detect splicing events (Ramskold et al., 2012), alternative 3′ or 5′ ends (Islam et al., 2011;Tang et al., 2009;Tang et al., 2010), SNPs and mutations (Ramskold et al.,

2012), in single-cell analysis further improvements are still needed to decrease the technical variation introduced during sample preparation, and to be able to obtain high-coverage transcriptomes. For bioinformatics tools specific for single-cell analysis (out of the scope of this thesis), the reader can refer to (Ning et al., 2014).

3. Outline and scope of this thesis

The main objective of the research in this thesis was to investigate regulatory mechanisms of gene expression, based on a diverse set of high-throughput RNA sequencing technologies. The first part of this thesis (**Chapter 1**) elaborated on how high-throughput RNA sequencing technologies have increased our understanding of the mechanisms that give rise to alternative transcripts and their alternative translation, and described the major RNA sequencing methods used to investigate specific aspects of gene expression.

In **Chapter 2** and **Chapter 3**, the process of alternative polyadenylation is investigated. **Chapter 2** describes the role of alternative polyadenylation in the context of oculopharyngeal muscular dystrophy (OPMD), by demonstrating transcriptome-wide shortening of 3' ends of mRNAs in OPMD. This study led to the proposition of a new role for the Poly(A) binding protein nuclear 1 (PABPN1) in polyadenylation site selection. **Chapter 3** shows the application of *cis*-eQTL (expression quantitative trait loci) analysis based on DeepSAGE data to identify single nucleotide polymorphisms affecting the usage of alternative polyadenylation sites, by disrupting or forming polyadenylation signal sequences.

In **Chapter 4** mechanisms controlling protein translation are investigated in the context of skeletal muscles. This chapter shows the application of the ribosome footprint profiling method to investigate the regulation of mRNA translation in skeletal muscle cells during myogenic differentiation.

Chapter 5 shows the application of full length mRNA sequencing to investigate interdependences between alternative regulatory events in gene expression, such as the coupling between alternative transcription, alternative splicing and alternative polyadenylation.

Finally, a general discussion in **Chapter 6** present limitations in the current high-throughput RNA sequencing technologies and outlines other regulatory mechanisms which have not been addressed in **Chapter 1**. The chapter ends with an overview of promising RNA-based diagnostic and theraupetic approaches

REFERENCES

1. 't Hoen,P.A., Y.Ariyurek, H.H.Thygesen, E.Vreugdenhil, R.H.Vossen, R.X.de Menezes, J.M.Boer, G.J.van Ommen, and J.T.den Dunnen. 2008. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res. 36: e141.
2. 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.
3. Agarwal,V.R., S.E.Bulun, M.Leitch, R.Rohrich, and E.R.Simpson. 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.
4. Alamancos,G.P., E.Agirre, and E.Eyras. 2014. Methods to study splicing from high-throughput RNA sequencing data. Methods Mol. Biol. 1126: 357-397.
5. Ameur,A., A.Zaghlool, J.Halvardson, A.Wetterbom, U.Gyllensten, L.Cavelier, and L.Feuk. 2011. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat. Struct. Mol. Biol. 18: 1435-1440.
6. Andersson,R., C.Gebhard, I.Miguel-Escalada, I.Hoof, J.Bornholdt, M.Boyd, Y.Chen, X.Zhao, C.Schmidl,

T.Suzuki, E.Ntini, E.Arner, E.Valen, K.Li, L.Schwarzfischer, D.Glatz, J.Raithel, B.Lilje, N.Rapin, F.O.Bagger, M.Jorgensen, P.R.Andersen, N.Bertin, O.Rackham, A.M.Burroughs, J.K.Baillie, Y.Ishizu, Y.Shimizu, E.Furuhata, S.Maeda, Y.Negishi, C.J.Mungall, T.F.Meehan, T.Lassmann, M.Itoh, H.Kawaji, N.Kondo, J.Kawai, A.Lennartsson, C.O.Daub, P.Heutink, D.A.Hume, T.H.Jensen, H.Suzuki, Y.Hayashizaki, F.Muller, A.R.Forrest, P.Carninci, M.Rehli, and A.Sandelin. 2014. An atlas of active enhancers across human cell types and tissues. Nature 507: 455-461.

7. Andreassi,C. and A.Riccio. 2009. To localize or not to localize: mRNA fate is in 3'UTR ends. Trends Cell Biol. 19: 465-474.

8. Armour,C.D., J.C.Castle, R.Chen, T.Babak, P.Loerch, S.Jackson, J.K.Shah, J.Dey, C.A.Rohl, J.M.Johnson, and C.K.Raymond. 2009. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. Nat. Methods 6: 647-649.

9. Au,K.F., V.Sebastiano, P.T.Afshar, J.D.Durruthy, L.Lee, B.A.Williams, B.H.van, E.E.Schadt, R.A.Reijo-Pera, J.G.Underwood, and W.H.Wong. 2013. Characterization of the human ESC transcriptome by hybrid sequencing. Proc. Natl. Acad. Sci. U. S. A 110: E4821-E4830.

10. Auboeuf,D., D.H.Dowhan, M.Dutertre, N.Martin, S.M.Berget, and B.W.O'Malley. 2005. A subset of nuclear receptor coregulators act as coupling proteins during synthesis and maturation of RNA transcripts. Mol. Cell Biol. 25: 5307-5316.

11. Balwierz,P.J., P.Carninci, C.O.Daub, J.Kawai, Y.Hayashizaki, B.W.Van, C.Beisel, and N.E.van. 2009. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. Genome Biol. 10: R79.

12. Barbosa,C., I.Peixeiro, and L.Romao. 2013. Gene expression regulation by upstream open reading frames and human disease. PLoS. Genet. 9: e1003529.

13. Batra,R., K.Charizanis, M.Manchanda, A.Mohan, M.Li, D.J.Finn, M.Goodwin, C.Zhang, K.Sobczak, C.A.Thornton, and M.S.Swanson. 2014. Loss of MBNL Leads to Disruption of Developmentally Regulated Alternative Polyadenylation in RNA-Mediated Disease. Mol. Cell.

14. Bava,F.A., C.Eliscovich, P.G.Ferreira, B.Minana, C.Ben-Dov, R.Guigo, J.Valcarcel, and R.Mendez. 2013. CPEB1 coordinates alternative 3'-UTR formation with translational regulation. Nature 495: 121-125.

15. Beck,A.H., 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. 2010. 3'-end sequencing for expression quantification (3SEQ) from archival tumor samples. PLoS ONE 5: e8768.

16. Benson,M.J., T.Aijo, X.Chang, J.Gagnon, U.J.Pape, V.Anantharaman, L.Aravind, J.P.Pursiheimo, S.Oberdoerffer, X.S.Liu, R.Lahesmaa, H.Lahdesmaki, and A.Rao. 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.

17. Bentley,D.L. 2014. Coupling mRNA processing with transcription in time and space. Nat. Rev. Genet. 15: 163-175.

18. Berget,S.M. 1995. Exon recognition in vertebrate splicing. J. Biol. Chem. 270: 2411-2414.

19. Boutet,S.C., T.H.Cheung, N.L.Quach, L.Liu, S.L.Prescott, A.Edalati, K.Iori, and T.A.Rando. 2012. Alternative polyadenylation mediates microRNA regulation of muscle stem cell function. Cell Stem Cell 10: 327-336.

20. Brennecke,P., S.Anders, J.K.Kim, A.A.Kolodziejczyk, X.Zhang, V.Proserpio, B.Baying, V.Benes, S.A.Teichmann, J.C.Marioni, and M.G.Heisler. 2013. Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods 10: 1093-1095.

21. Brown,S.J., P.Stoilov, and Y.Xing. 2012. Chromatin and epigenetic regulation of pre-mRNA processing. Hum. Mol. Genet. 21: R90-R96.

22. Buljan,M., G.Chalancon, S.Eustermann, G.P.Wagner, M.Fuxreiter, A.Bateman, and M.M.Babu. 2012. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol. Cell 46: 871-883.

23. Calvo,S.E., D.J.Pagliarini, and V.K.Mootha. 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.

24. Carninci,P., C.Kvam, A.Kitamura, T.Ohsumi, Y.Okazaki, M.Itoh, M.Kamiya, K.Shibata, N.Sasaki, M.Izawa, M.Muramatsu, Y.Hayashizaki, and C.Schneider. 1996. High-efficiency full-length cDNA cloning by biotinylated

CAP trapper. Genomics 37: 327-336.

25. Churchman,L.S. and J.S.Weissman. 2011. Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature 469: 368-373.
26. Costa,V., M.Aprile, R.Esposito, and A.Ciccodicola. 2013. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. Eur. J. Hum. Genet. 21: 134-142.
27. Danckwardt,S., M.W.Hentze, and A.E.Kulozik. 2008. 3' end mRNA processing: molecular mechanisms and implications for health and disease. EMBO J. 27: 482-498.
28. David,R. 2012. Small RNAs: miRNAs' strict schedule. Nat. Rev. Genet. 13: 378.
29. Davis,W., Jr. and R.M.Schultz. 2000. Developmental change in TATA-box utilization during preimplantation mouse development. Dev. Biol. 218: 275-283.
30. de Hoon,M. and Y.Hayashizaki. 2008. Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. Biotechniques 44: 627-8, 630, 632.
31. Deng,Q., D.Ramskold, B.Reinius, and R.Sandberg. 2014. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343: 193-196.
32. Derti,A., P.Garrett-Engele, K.D.Macisaac, R.C.Stevens, S.Sriram, R.Chen, C.A.Rohl, J.M.Johnson, and T.Babak. 2012. A quantitative atlas of polyadenylation in five mammals. Genome Res. 22: 1173-1183.
33. Ding,Y., Y.Tang, C.K.Kwok, Y.Zhang, P.C.Bevilacqua, and S.M.Assmann. 2014. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 505: 696-700.
34. Djebali,S., C.A.Davis, A.Merkel, A.Dobin, T.Lassmann, A.Mortazavi, A.Tanzer, J.Lagarde, W.Lin, F.Schlesinger, C.Xue, G.K.Marinov, J.Khatun, B.A.Williams, C.Zaleski, J.Rozowsky, M.Roder, F.Kokocinski, R.F.Abdelhamid, T.Alioto, I.Antoshechkin, M.T.Baer, N.S.Bar, P.Batut, K.Bell, I.Bell, S.Chakrabortty, X.Chen, J.Chrast, J.Curado, T.Derrien, J.Drenkow, E.Dumais, J.Dumais, R.Duttagupta, E.Falconnet, M.Fastuca, K.Fejes-Toth, P.Ferreira, S.Foissac, M.J.Fullwood, H.Gao, D.Gonzalez, A.Gordon, H.Gunawardena, C.Howald, S.Jha, R.Johnson, P.Kapranov, B.King, C.Kingswood, O.J.Luo, E.Park, K.Persaud, J.B.Preall, P.Ribeca, B.Risk, D.Robyr, M.Sammeth, L.Schaffer, L.H.See, A.Shahab, J.Skancke, A.M.Suzuki, H.Takahashi, H.Tilgner, D.Trout, N.Walters, H.Wang, J.Wrobel, Y.Yu, X.Ruan, Y.Hayashizaki, J.Harrow, M.Gerstein, T.Hubbard, A.Reymond, S.E.Antonarakis, G.Hannon, M.C.Giddings, Y.Ruan, B.Wold, P.Carninci, R.Guigo, and T.R.Gingeras. 2012. Landscape of transcription in human cells. Nature 489: 101-108.
35. Dujardin,G., C.Lafaille, E.Petrillo, V.Buggiano, L.I.Gomez Acuna, A.Fiszbein, M.A.Godoy Herz, M.N.Nieto, M.J.Munoz, M.Allo, I.E.Schor, and A.R.Kornblihtt. 2013. Transcriptional elongation and alternative splicing. Biochim. Biophys. Acta 1829: 134-140.
36. Elkon,R., J.Drost, H.G.van, M.Jenal, M.Schrier, J.A.Vrielink, and R.Agami. 2012. E2F mediates enhanced alternative polyadenylation in proliferation. Genome Biol. 13: R59.
37. Fabian,M.R., N.Sonenberg, and W.Filipowicz. 2010. Regulation of mRNA translation and stability by microRNAs. Annu. Rev. Biochem. 79: 351-379.
38. Fox-Walsh,K., J.Davis-Turak, Y.Zhou, H.Li, and X.D.Fu. 2011. A multiplex RNA-seq strategy to profile poly(A+) RNA: application to analysis of transcription response and 3' end formation. Genomics 98: 266-271.
39. Frith,M.C., E.Valen, A.Krogh, Y.Hayashizaki, P.Carninci, and A.Sandelin. 2008. A code for transcription initiation in mammalian genomes. Genome Res. 18: 1-12.
40. Fritsch,C., A.Herrmann, M.Nothnagel, K.Szafranski, K.Huse, F.Schumann, S.Schreiber, M.Platzer, M.Krawczak, J.Hampe, and M.Brosch. 2012. Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Res. 22: 2208-2218.
41. Fu,X.D. and M.Ares, Jr. 2014. Context-dependent control of alternative splicing by RNA-binding proteins. Nat. Rev. Genet.
42. Fu,Y., Y.Sun, Y.Li, J.Li, X.Rao, C.Chen, and A.Xu. 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.
43. Gao,L., Z.Fang, K.Zhang, D.Zhi, and X.Cui. 2011. Length bias correction for RNA-seq data in gene set analyses. Bioinformatics. 27: 662-669.
44. Garber,M., M.G.Grabherr, M.Guttman, and C.Trapnell. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods 8: 469-477.
45. Geisberg,J.V., Z.Moqtaderi, X.Fan, F.Ozsolak, and K.Struhl. 2014. Global analysis of mRNA isoform half-lives

reveals stabilizing and destabilizing elements in yeast. Cell 156: 812-824.

46. Gilbert,W. 1978. Why genes in pieces? Nature 271: 501.

47. Giudice,J., Z.Xia, E.T.Wang, M.A.Scavuzzo, A.J.Ward, A.Kalsotra, W.Wang, X.H.Wehrens, C.B.Burge, W.Li, and T.A.Cooper. 2014. Alternative splicing regulates vesicular trafficking genes in cardiomyocytes during postnatal heart development. Nat. Commun. 5: 3603.

48. Gonzalez-Porta,M., A.Frankish, J.Rung, J.Harrow, and A.Brazma. 2013. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. 14: R70.

49. Goossens,S., B.Janssens, G.Vanpoucke, R.R.De, H.J.van, and R.F.van. 2007. Truncated isoform of mouse alphaT-catenin is testis-restricted in expression and function. FASEB J. 21: 647-655.

50. Gorgoni,B. and N.K.Gray. 2004. The roles of cytoplasmic poly(A)-binding proteins in regulating gene expression: a developmental perspective. Brief. Funct. Genomic. Proteomic. 3: 125-141.

51. Graber,J.H., F.I.Nazeer, P.C.Yeh, J.N.Kuehner, S.Borikar, D.Hoskinson, and C.L.Moore. 2013. DNA damage induces targeted, genome-wide variation of poly(A) sites in budding yeast. Genome Res. 23: 1690-1703.

52. Gracheva,E.O., J.F.Cordero-Morales, J.A.Gonzalez-Carcacia, N.T.Ingolia, C.Manno, C.I.Aranguren, J.S.Weissman, and D.Julius. 2011. Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats. Nature 476: 88-91.

53. Griebel,T., B.Zacher, P.Ribeca, E.Raineri, V.Lacroix, R.Guigo, and M.Sammeth. 2012. Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic Acids Res. 40: 10073-10083.

54. Gupta,I., S.Clauder-Munster, B.Klaus, A.I.Jarvelin, R.S.Aiyar, V.Benes, S.Wilkening, W.Huber, V.Pelechano, and L.M.Steinmetz. 2014. Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions. Mol. Syst. Biol. 10: 719.

55. Gustincich,S., A.Sandelin, C.Plessy, S.Katayama, R.Simone, D.Lazarevic, Y.Hayashizaki, and P.Carninci. 2006. The complexity of the mammalian transcriptome. J. Physiol 575: 321-332.

56. Hafez,D., T.Ni, S.Mukherjee, J.Zhu, and U.Ohler. 2013. Genome-wide identification and predictive modeling of tissue-specific alternative polyadenylation. Bioinformatics. 29: i108-i116.

57. Hafner,M., M.Landthaler, L.Burger, M.Khorshid, J.Hausser, P.Berninger, A.Rothballer, M.Ascano, A.C.Jungkamp, M.Munschauer, A.Ulrich, G.S.Wardle, S.Dewell, M.Zavolan, and T.Tuschl. 2010a. PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins. J. Vis. Exp.

58. Hafner,M., M.Landthaler, L.Burger, M.Khorshid, J.Hausser, P.Berninger, A.Rothballer, M.Ascano, Jr., A.C.Jungkamp, M.Munschauer, A.Ulrich, G.S.Wardle, S.Dewell, M.Zavolan, and T.Tuschl. 2010b. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141: 129-141.

59. Hafner,M., N.Renwick, M.Brown, A.Mihailovic, D.Holoch, C.Lin, J.T.Pena, J.D.Nusbaum, P.Morozov, J.Ludwig, T.Ojo, S.Luo, G.Schroth, and T.Tuschl. 2011. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA. 17: 1697-1712.

60. Hansen,K.D., S.E.Brenner, and S.Dudoit. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 38: e131.

61. Hao,S. and D.Baltimore. 2013. RNA splicing regulates the temporal order of TNF-induced gene expression. Proc. Natl. Acad. Sci. U. S. A 110: 11934-11939.

62. Harrow,J., A.Frankish, J.M.Gonzalez, E.Tapanari, M.Diekhans, F.Kokocinski, B.L.Aken, D.Barrell, A.Zadissa, S.Searle, I.Barnes, A.Bignell, V.Boychenko, T.Hunt, M.Kay, G.Mukherjee, J.Rajan, G.Despacio-Reyes, G.Saunders, C.Steward, R.Harte, M.Lin, C.Howald, A.Tanzer, T.Derrien, J.Chrast, N.Walters, S.Balasubramanian, B.Pei, M.Tress, J.M.Rodriguez, I.Ezkurdia, B.J.van, M.Brent, D.Haussler, M.Kellis, A.Valencia, A.Reymond, M.Gerstein, R.Guigo, and T.J.Hubbard. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22: 1760-1774.

63. Hazelbaker,D.Z., S.Marquardt, W.Wlotzka, and S.Buratowski. 2013. Kinetic competition between RNA Polymerase II and Sen1-dependent transcription termination. Mol. Cell 49: 55-66.

64. He,Y., B.Vogelstein, V.E.Velculescu, N.Papadopoulos, and K.W.Kinzler. 2008. The antisense transcriptomes of human cells. Science 322: 1855-1857.

65. Hestand,M.S., A.Klingenhoff, M.Scherf, Y.Ariyurek, Y.Ramos, W.W.van, M.Suzuki, T.Werner, G.J.van Ommen, J.T.den Dunnen, M.Harbers, and P.A.'t Hoen. 2010. Tissue-specific transcript annotation and expression

profiling with complementary next-generation sequencing technologies. Nucleic Acids Res. 38: e165.

66. Hill,R.E. and L.A.Lettice. 2013. Alterations to the remote control of Shh gene expression cause congenital abnormalities. Philos. Trans. R. Soc. Lond B Biol. Sci. 368: 20120357.
67. Hogg,J.R. and S.P.Goff. 2010. Upf1 senses 3'UTR length to potentiate mRNA decay. Cell 143: 379-389.
68. Hoque,M., Z.Ji, D.Zheng, W.Luo, W.Li, B.You, J.Y.Park, G.Yehia, and B.Tian. 2013. Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. Nat. Methods 10: 133-139.
69. Hsin,J.P. and J.L.Manley. 2012. The RNA polymerase II CTD coordinates transcription and RNA processing. Genes Dev. 26: 2119-2137.
70. Huang,d.W., B.T.Sherman, and R.A.Lempicki. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4: 44-57.
71. Huang,Y., W.Li, X.Yao, Q.J.Lin, J.W.Yin, Y.Liang, M.Heiner, B.Tian, J.Hui, and G.Wang. 2012. Mediator complex regulates alternative mRNA processing via the MED23 subunit. Mol. Cell 45: 459-469.
72. Ingolia,N.T. 2010. Genome-wide translational profiling by ribosome footprinting. Methods Enzymol. 470: 119-142.
73. Ingolia,N.T., G.A.Brar, S.Rouskin, A.M.McGeachy, and J.S.Weissman. 2012. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc. 7: 1534-1550.
74. Ingolia,N.T., S.Ghaemmaghami, J.R.Newman, and J.S.Weissman. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324: 218-223.
75. Ingolia,N.T., L.F.Lareau, and J.S.Weissman. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147: 789-802.
76. Islam,S., U.Kjallquist, A.Moliner, P.Zajac, J.B.Fan, P.Lonnerberg, and S.Linnarsson. 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 21: 1160-1167.
77. Islam,S., A.Zeisel, S.Joost, M.G.La, P.Zajac, M.Kasper, P.Lonnerberg, and S.Linnarsson. 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods 11: 163-166.
78. Jan,C.H., R.C.Friedman, J.G.Ruby, and D.P.Bartel. 2011. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. Nature 469: 97-101.
79. Jenal,M., R.Elkon, F.Loayza-Puch, H.G.van, U.Kuhn, F.M.Menzies, J.A.Vrielink, A.J.Bos, J.Drost, K.Rooijers, D.C.Rubinsztein, and R.Agami. 2012. The poly(a)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. Cell 149: 538-553.
80. Jensen,K.B. and R.B.Darnell. 2008. CLIP: crosslinking and immunoprecipitation of in vivo RNA targets of RNA-binding proteins. Methods Mol. Biol. 488: 85-98.
81. Ji,Z., J.Y.Lee, Z.Pan, B.Jiang, and B.Tian. 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.
82. Ji,Z., W.Luo, W.Li, M.Hoque, Z.Pan, Y.Zhao, and B.Tian. 2011. Transcriptional activity regulates alternative cleavage and polyadenylation. Mol. Syst. Biol. 7: 534.
83. Ji,Z. and B.Tian. 2009. Reprogramming of 3UTR Untranslated Regions of mRNAs by Alternative Polyadenylation in Generation of Pluripotent Stem Cells from Different Cell Types. PLoS ONE 4: e8419.
84. Jorgensen,R.A. and A.E.Dorantes-Acosta. 2012. Conserved Peptide Upstream Open Reading Frames are Associated with Regulatory Genes in Angiosperms. Front Plant Sci. 3: 191.
85. Kaida,D., M.G.Berg, I.Younis, M.Kasim, L.N.Singh, L.Wan, and G.Dreyfuss. 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature 468: 664-668.
86. Kanamori-Katayama,M., M.Itoh, H.Kawaji, T.Lassmann, S.Katayama, M.Kojima, N.Bertin, A.Kaiho, N.Ninomiya, C.O.Daub, P.Carninci, A.R.Forrest, and Y.Hayashizaki. 2011. Unamplified cap analysis of gene expression on a single-molecule sequencer. Genome Res. 21: 1150-1159.
87. Kapranov,P., J.Cheng, S.Dike, D.A.Nix, R.Duttagupta, A.T.Willingham, P.F.Stadler, J.Hertel, J.Hackermuller, I.L.Hofacker, I.Bell, E.Cheung, J.Drenkow, E.Dumais, S.Patel, G.Helt, M.Ganesh, S.Ghosh, A.Piccolboni, V.Sementchenko, H.Tammana, and T.R.Gingeras. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316: 1484-1488.

88. Kari,V., O.Karpiuk, B.Tieg, M.Kriegs, E.Dikomey, H.Krebber, Y.Begus-Nahrmann, and S.A.Johnsen. 2013. A subset of histone H2B genes produces polyadenylated mRNAs under a variety of cellular conditions. PLoS. One. 8: e63745.

89. Katz,Y., E.T.Wang, E.M.Airoldi, and C.B.Burge. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods 7: 1009-1015.

90. Kertesz,M., Y.Wan, E.Mazor, J.L.Rinn, R.C.Nutter, H.Y.Chang, and E.Segal. 2010. Genome-wide measurement of RNA secondary structure in yeast. Nature 467: 103-107.

91. Kim,K.K., J.Nam, Y.S.Mukouyama, and S.Kawamoto. 2013. Rbfox3-regulated alternative splicing of Numb promotes neuronal differentiation during development. J. Cell Biol. 200: 443-458.

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

93. Kochetov,A.V. 2008. Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. Bioessays 30: 683-691.

94. Kodzius,R., M.Kojima, H.Nishiyori, M.Nakamura, S.Fukuda, M.Tagami, D.Sasaki, K.Imamura, C.Kai, M.Harbers, Y.Hayashizaki, and P.Carninci. 2006. CAGE: cap analysis of gene expression. Nat. Methods 3: 211-222.

95. Konig,J., K.Zarnack, G.Rot, T.Curk, M.Kayikci, B.Zupan, D.J.Turner, N.M.Luscombe, and J.Ule. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. 17: 909-915.

96. Koren,S., G.P.Harhay, T.P.Smith, J.L.Bono, D.M.Harhay, S.D.McVey, D.Radune, N.H.Bergman, and A.M.Phillippy. 2013. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol. 14: R101.

97. Kozak,M. 2005. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene 361: 13-37.

98. Kung,J.T., D.Colognori, and J.T.Lee. 2013. Long noncoding RNAs: past, present, and future. Genetics 193: 651-669.

99. Lebedeva,S., M.Jens, K.Theil, B.Schwanhausser, M.Selbach, M.Landthaler, and N.Rajewsky. 2011. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. Mol. Cell 43: 340-352.

100. Lee,K.M. and W.Y.Tarn. 2014. TRAP150 activates splicing in composite terminal exons. Nucleic Acids Res.

101. Lee,S., B.Liu, S.Lee, S.X.Huang, B.Shen, and S.B.Qian. 2012. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc. Natl. Acad. Sci. U. S. A 109: E2424-E2432.

102. Legendre,M. and D.Gautheret. 2003. Sequence determinants in human polyadenylation site selection. BMC. Genomics 4: 7.

103. Lemay,J.F., A.D'Amours, C.Lemieux, D.H.Lackner, V.G.St-Sauveur, J.Bahler, and F.Bachand. 2010. The nuclear poly(A)-binding protein interacts with the exosome to promote synthesis of noncoding small nucleolar RNAs. Mol. Cell 37: 34-45.

104. Levanon,D. and Y.Groner. 2004. Structure and regulated expression of mammalian RUNX genes. Oncogene 23: 4211-4219.

105. Levin,J.Z., M.Yassour, X.Adiconis, C.Nusbaum, D.A.Thompson, N.Friedman, A.Gnirke, and A.Regev. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat. Methods 7: 709-715.

106. Li,J.J., P.J.Bickel, and M.D.Biggin. 2014. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ. 2: e270.

107. Li,Y., 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. 2012. Dynamic landscape of tandem 3' UTRs during zebrafish development. Genome Res. 22: 1899-1906.

108. Licatalosi,D.D., 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. 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature 456: 464-469.

109. Lin,Y., 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. 2012. An in-depth map of polyadenylation sites in cancer. Nucleic Acids Res. 40: 8460-8471.

110. Lister,R., R.C.O'Malley, J.Tonti-Filippini, B.D.Gregory, C.C.Berry, A.H.Millar, and J.R.Ecker. 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133: 523-536.

111. Lodish H, Berk A, and Zipursky SL. 2000. Processing of rRNA and tRNA. in Molecular Cell Biology (ed. W.H.Freeman), New York.

112. Lucks,J.B., S.A.Mortimer, C.Trapnell, S.Luo, S.Aviran, G.P.Schroth, L.Pachter, J.A.Doudna, and A.P.Arkin. 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.

113. Lundberg,E., L.Fagerberg, D.Klevebring, I.Matic, T.Geiger, J.Cox, C.Algenas, J.Lundeberg, M.Mann, and M.Uhlen. 2010. Defining the transcriptome and proteome in three functionally different human cell lines. Mol. Syst. Biol. 6: 450.

114. Luo,W., Z.Ji, Z.Pan, B.You, M.Hoque, W.Li, S.I.Gunderson, and B.Tian. 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.

115. Magny,E.G., J.I.Pueyo, F.M.Pearl, M.A.Cespedes, J.E.Niven, S.A.Bishop, and J.P.Couso. 2013. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. Science 341: 1116-1120.

116. Maier,T., M.Guell, and L.Serrano. 2009. Correlation of mRNA and protein in complex biological samples. FEBS Lett. 583: 3966-3973.

117. Mangone,M., A.P.Manoharan, D.Thierry-Mieg, J.Thierry-Mieg, T.Han, S.D.Mackowiak, E.Mis, C.Zegar, M.R.Gutwein, V.Khivansara, O.Attie, K.Chen, K.Salehi-Ashtiani, M.Vidal, T.T.Harkins, P.Bouffard, Y.Suzuki, S.Sugano, Y.Kohara, N.Rajewsky, F.Piano, K.C.Gunsalus, and J.K.Kim. 2010. The landscape of C. elegans 3'UTRs. Science 329: 432-435.

118. Manley,J.L., P.A.Sharp, and M.L.Gefter. 1982. Rna synthesis in isolated nuclei processing of adenovirus serotype 2 late messenger rna precursors. J. Mol. Biol. 159: 581-599.

119. Martin,G., A.R.Gruber, W.Keller, and M.Zavolan. 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.

120. Martinson,H.G. 2011. An active role for splicing in 3'-end formation. Wiley. Interdiscip. Rev. RNA. 2: 459-470.

121. Matsumura,H., K.Yoshida, S.Luo, E.Kimura, T.Fujibe, Z.Albertyn, R.A.Barrero, D.H.Kruger, G.Kahl, G.P.Schroth, and R.Terauchi. 2010. High-throughput SuperSAGE for digital gene expression analysis of multiple samples using next generation sequencing. PLoS ONE 5: e12010.

122. Mayr,C. and D.P.Bartel. 2009. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 138: 673-684.

123. Menschaert,G., C.W.Van, T.Notelaers, A.Koch, J.Crappe, K.Gevaert, and D.P.Van. 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.

124. Michel,A.M., K.R.Choudhury, A.E.Firth, N.T.Ingolia, J.F.Atkins, and P.V.Baranov. 2012. Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res. 22: 2219-2229.

125. Miura,P., S.Shenker, C.Andreu-Agullo, J.O.Westholm, and E.C.Lai. 2013. Widespread and extensive lengthening of 3' UTRs in the mammalian brain. Genome Res. 23: 812-825.

126. Moqtaderi,Z., J.V.Geisberg, Y.Jin, X.Fan, and K.Struhl. 2013. Species-specific factors mediate extensive heterogeneity of mRNA 3' ends in yeasts. Proc. Natl. Acad. Sci. U. S. A 110: 11073-11078.

127. Morris,D.R. and A.P.Geballe. 2000. Upstream open reading frames as regulators of mRNA translation. Mol. Cell Biol. 20: 8635-8642.

128. Morrissy,A.S., R.D.Morin, A.Delaney, T.Zeng, H.McDonald, S.Jones, Y.Zhao, M.Hirst, and M.A.Marra. 2009. Next-generation tag sequencing for cancer gene expression profiling. Genome Res. 19: 1825-1835.

129. Mortazavi,A., B.A.Williams, K.McCue, L.Schaeffer, and B.Wold. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5: 621-628.

130. Nagaike,T., C.Logan, I.Hotta, O.Rozenblatt-Rosen, M.Meyerson, and J.L.Manley. 2011. Transcriptional activators enhance polyadenylation of mRNA precursors. Mol. Cell 41: 409-418.

131. Neph,S., J.Vierstra, A.B.Stergachis, A.P.Reynolds, E.Haugen, B.Vernot, R.E.Thurman, S.John, R.Sandstrom, A.K.Johnson, M.T.Maurano, R.Humbert, E.Rynes, H.Wang, S.Vong, K.Lee, D.Bates, M.Diegel, V.Roach, D.Dunn, J.Neri, A.Schafer, R.S.Hansen, T.Kutyavin, E.Giste, M.Weaver, T.Canfield, P.Sabo, M.Zhang, G.Balasundaram, R.Byron, M.J.MacCoss, J.M.Akey, M.A.Bender, M.Groudine, R.Kaul, and J.A.Stamatoyannopoulos. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489: 83-90.

132. Ni,T., Y.Yang, D.Hafez, W.Yang, K.Kiesewetter, Y.Wakabayashi, U.Ohler, W.Peng, and J.Zhu. 2013. Distinct polyadenylation landscapes of diverse human tissues revealed by a modified PA-seq strategy. BMC. Genomics 14: 615.

133. Nielsen,K.L., A.L.Hogh, and J.Emmersen. 2006. DeepSAGE--digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. Nucleic Acids Res. 34: e133.

134. Ning,L., G.Liu, G.Li, Y.Hou, Y.Tong, and J.He. 2014. Current Challenges in the Bioinformatics of Single Cell Genomics. Front Oncol. 4: 7.

135. Nordlund,J., A.Kiialainen, O.Karlberg, E.C.Berglund, H.Goransson-Kultima, M.Sonderkaer, K.L.Nielsen, M.G.Gustafsson, M.Behrendtz, E.Forestier, M.Perkkio, S.Soderhall, G.Lonnerholm, and A.C.Syvanen. 2012. Digital gene expression profiling of primary acute lymphoblastic leukemia cells. Leukemia 26: 1218-1227.

136. Nunes,N.M., W.Li, B.Tian, and A.Furger. 2010. A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence. EMBO J. 29: 1523-1536.

137. Otsuka,Y., N.L.Kedersha, and D.R.Schoenberg. 2009. Identification of a cytoplasmic complex that adds a cap onto 5'-monophosphate RNA. Mol. Cell Biol. 29: 2155-2167.

138. Ozsolak,F., P.Kapranov, S.Foissac, S.W.Kim, E.Fishilevich, A.P.Monaghan, B.John, and P.M.Milos. 2010. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell 143: 1018-1029.

139. Ozsolak,F., 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. 2009. Direct RNA sequencing. Nature 461: 814-818.

140. Pan,Q., O.Shai, L.J.Lee, B.J.Frey, and B.J.Blencowe. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. 40: 1413-1415.

141. Parkhomchuk,D., T.Borodina, V.Amstislavskiy, M.Banaru, L.Hallen, S.Krobitsch, H.Lehrach, and A.Soldatov. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 37: e123.

142. Patthy,L. 1999. Genome evolution and the evolution of exon-shuffling--a review. Gene 238: 103-114.

143. Pedersen,I.S., P.Dervan, A.McGoldrick, M.Harrison, F.Ponchel, V.Speirs, J.D.Isaacs, T.Gorey, and A.McCann. 2002. Promoter switch: a novel mechanism causing biallelic PEG1/MEST expression in invasive breast cancer. Hum. Mol. Genet. 11: 1449-1453.

144. Pelechano,V., W.Wei, and L.M.Steinmetz. 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. Nature 497: 127-131.

145. Pelechano,V., S.Wilkening, A.I.Jarvelin, M.M.Tekkedil, and L.M.Steinmetz. 2012. Genome-wide polyadenylation site mapping. Methods Enzymol. 513: 271-296.

146. Pervouchine,D.D., E.E.Khrameeva, M.Y.Pichugina, O.V.Nikolaienko, M.S.Gelfand, P.M.Rubtsov, and A.A.Mironov. 2012. Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. RNA. 18: 1-15.

147. Picelli,S., A.K.Bjorklund, O.R.Faridani, S.Sagasser, G.Winberg, and R.Sandberg. 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods 10: 1096-1098.

148. Pimentel,H., M.Parra, S.Gee, D.Ghanem, X.An, J.Li, N.Mohandas, L.Pachter, and J.G.Conboy. 2014. A dynamic alternative splicing program regulates gene expression during terminal erythropoiesis. Nucleic Acids Res. 42: 4031-4042.

149. Pinto,P.A., 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. 2011. RNA polymerase II kinetics in polo polyadenylation signal selection. EMBO J. 30: 2431-2444.

150. Pistoni,M., C.Ghigna, and D.Gabellini. 2010. Alternative splicing and muscular dystrophy. RNA. Biol. 7: 441-452.

151. Plessy,C., N.Bertin, H.Takahashi, R.Simone, M.Salimullah, T.Lassmann, M.Vitezic, J.Severin, S.Olivarius,

D.Lazarevic, N.Hornig, V.Orlando, I.Bell, H.Gao, J.Dumais, P.Kapranov, H.Wang, C.A.Davis, T.R.Gingeras, J.Kawai, C.O.Daub, Y.Hayashizaki, S.Gustincich, and P.Carninci. 2010. Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. Nat. Methods 7: 528-534.

152. Plotkin,J.B. 2010. Transcriptional regulation is only half the story. Mol. Syst. Biol. 6: 406.

153. Pozner,A., J.Lotem, C.Xiao, D.Goldenberg, O.Brenner, V.Negreanu, D.Levanon, and Y.Groner. 2007. Developmentally regulated promoter-switch transcriptionally controls Runx1 function during embryonic hematopoiesis. BMC. Dev. Biol. 7: 84.

154. Ramskold,D., S.Luo, Y.C.Wang, R.Li, Q.Deng, O.R.Faridani, G.A.Daniels, I.Khrebtukova, J.F.Loring, L.C.Laurent, G.P.Schroth, and R.Sandberg. 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol. 30: 777-782.

155. Ray,D., H.Kazan, K.B.Cook, M.T.Weirauch, H.S.Najafabadi, X.Li, S.Gueroussov, M.Albu, H.Zheng, A.Yang, H.Na, M.Irimia, L.H.Matzat, R.K.Dale, S.A.Smith, C.A.Yarosh, S.M.Kelly, B.Nabet, D.Mecenas, W.Li, R.S.Laishram, M.Qiao, H.D.Lipshitz, F.Piano, A.H.Corbett, R.P.Carstens, B.J.Frey, R.A.Anderson, K.W.Lynch, L.O.Penalva, E.P.Lei, A.G.Fraser, B.J.Blencowe, Q.D.Morris, and T.R.Hughes. 2013. A compendium of RNA-binding motifs for decoding gene regulation. Nature 499: 172-177.

156. Resch,A., Y.Xing, A.Alekseyenko, B.Modrek, and C.Lee. 2004. Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. Nucleic Acids Res. 32: 1261-1269.

157. Roberts,A., C.Trapnell, J.Donaghey, J.L.Rinn, and L.Pachter. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 12: R22.

158. Ruan,X. and Y.Ruan. 2012. Genome wide full-length transcript analysis using 5' and 3' paired-end-tag next generation sequencing (RNA-PET). Methods Mol. Biol. 809: 535-562.

159. Salimullah,M., M.Sakai, C.Plessy, and P.Carninci. 2011. NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes. Cold Spring Harb. Protoc. 2011: db.

160. Sandberg,R., J.R.Neilson, A.Sarma, P.A.Sharp, and C.B.Burge. 2008. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science 320: 1643-1647.

161. Sanyal,A., B.R.Lajoie, G.Jain, and J.Dekker. 2012. The long-range interaction landscape of gene promoters. Nature 489: 109-113.

162. Schaefer,M., T.Pollex, K.Hanna, and F.Lyko. 2009. RNA cytosine methylation analysis by bisulfite sequencing. Nucleic Acids Res. 37: e12.

163. SCHERRER,K., H.LATHAM, and J.E.DARNELL. 1963. Demonstration of an unstable RNA and of a precursor to ribosomal RNA in HeLa cells. Proc. Natl. Acad. Sci. U. S. A 49: 240-248.

164. Schwanhausser,B., D.Busse, N.Li, G.Dittmar, J.Schuchhardt, J.Wolf, W.Chen, and M.Selbach. 2011. Global quantification of mammalian gene expression control. Nature 473: 337-342.

165. Sharon,D., H.Tilgner, F.Grubert, and M.Snyder. 2013. A single-molecule long-read survey of the human transcriptome. Nat. Biotechnol. 31: 1009-1014.

166. Shatkin,A.J. 1976. Capping of eucaryotic mRNAs. Cell 9: 645-653.

167. Shepard,P.J., E.A.Choi, J.Lu, L.A.Flanagan, K.J.Hertel, and Y.Shi. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA. 17: 761-772.

168. Shepard,P.J. and K.J.Hertel. 2008. Conserved RNA secondary structures promote alternative splicing. RNA. 14: 1463-1469.

169. Sherstnev,A., C.Duc, C.Cole, V.Zacharaki, C.Hornyik, F.Ozsolak, P.M.Milos, G.J.Barton, and G.G.Simpson. 2012. Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. Nat. Struct. Mol. Biol. 19: 845-852.

170. Shi,Y., D.C.Di Giammartino, D.Taylor, A.Sarkeshik, W.J.Rice, J.R.Yates, III, J.Frank, and J.L.Manley. 2009. Molecular architecture of the human pre-mRNA 3' processing complex. Mol. Cell 33: 365-376.

171. Shiraki,T., S.Kondo, S.Katayama, K.Waki, T.Kasukawa, H.Kawaji, R.Kodzius, A.Watahiki, M.Nakamura, T.Arakawa, S.Fukuda, D.Sasaki, A.Podhajska, M.Harbers, J.Kawai, P.Carninci, and Y.Hayashizaki. 2003. 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.

172. Slavoff,S.A., A.J.Mitchell, A.G.Schwaid, M.N.Cabili, J.Ma, J.Z.Levin, A.D.Karger, B.A.Budnik, J.L.Rinn, and

A.Saghatelian. 2013. Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat. Chem. Biol. 9: 59-64.

173. Smibert,P., P.Miura, J.O.Westholm, S.Shenker, G.May, M.O.Duff, D.Zhang, B.D.Eads, J.Carlson, J.B.Brown, R.C.Eisman, J.Andrews, T.Kaufman, P.Cherbas, S.E.Celniker, B.R.Graveley, and E.C.Lai. 2012. Global patterns of tissue-specific alternative polyadenylation in Drosophila. Cell Rep. 1: 277-289.

174. Soeiro,R., M.H.Vaughan, J.R.Warner, and J.E.Darnell, Jr. 1968. The turnover of nuclear DNA-like RNA in HeLa cells. J. Cell Biol. 39: 112-118.

175. Sonenberg,N. and A.C.Gingras. 1998. The mRNA 5' cap-binding protein eIF4E and control of cell growth. Curr. Opin. Cell Biol. 10: 268-275.

176. Soneson,C. and M.Delorenzi. 2013. A comparison of methods for differential expression analysis of RNA-seq data. BMC. Bioinformatics. 14: 91.

177. Sorek,R., G.Lev-Maor, M.Reznik, T.Dagan, F.Belinky, D.Graur, and G.Ast. 2004a. Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. Mol. Cell 14: 221-231.

178. Sorek,R., R.Shamir, and G.Ast. 2004b. How prevalent is functional alternative splicing in the human genome? Trends Genet. 20: 68-71.

179. Spies,N., C.B.Burge, and D.P.Bartel. 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.

180. Steijger,T., J.F.Abril, P.G.Engstrom, F.Kokocinski, J.F.Abril, M.Akerman, T.Alioto, G.Ambrosini, S.E.Antonarakis, J.Behr, P.Bertone, R.Bohnert, P.Bucher, N.Cloonan, T.Derrien, S.Djebali, J.Du, S.Dudoit, P.G.Engstrom, M.Gerstein, T.R.Gingeras, D.Gonzalez, S.M.Grimmond, R.Guigo, L.Habegger, J.Harrow, T.J.Hubbard, C.Iseli, G.Jean, A.Kahles, F.Kokocinski, J.Lagarde, J.Leng, G.Lefebvre, S.Lewis, A.Mortazavi, P.Niermann, G.Ratsch, A.Reymond, P.Ribeca, H.Richard, J.Rougemont, J.Rozowsky, M.Sammeth, A.Sboner, M.H.Schulz, S.M.Searle, N.D.Solorzano, V.Solovyev, M.Stanke, T.Steijger, B.J.Stevenson, H.Stockinger, A.Valsesia, D.Weese, S.White, B.J.Wold, J.Wu, T.D.Wu, G.Zeller, D.Zerbino, M.Q.Zhang, T.J.Hubbard, R.Guigo, J.Harrow, and P.Bertone. 2013. Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods 10: 1177-1184.

181. Steinthorsdottir,V., H.Stefansson, S.Ghosh, B.Birgisdottir, S.Bjornsdottir, A.C.Fasquel, O.Olafsson, K.Stefansson, and J.R.Gulcher. 2004. Multiple novel transcription initiation sites for NRG1. Gene 342: 97-105.

182. Sugnet,C.W., W.J.Kent, M.Ares, Jr., and D.Haussler. 2004. Transcriptome and genome conservation of alternative splicing events in humans and mice. Pac. Symp. Biocomput. 66-77.

183. Sun,H., J.Wu, P.Wickramasinghe, S.Pal, R.Gupta, A.Bhattacharyya, F.J.Agosto-Perez, L.C.Showe, T.H.Huang, and R.V.Davuluri. 2011. Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. Nucleic Acids Res. 39: 190-201.

184. Suzuki,H., The FANTOM Consortium & Riken Omics Science Center. 2009. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat. Genet. 41: 553-562.

185. Tan,W., Y.Wang, B.Gold, J.Chen, M.Dean, P.J.Harrison, D.R.Weinberger, and A.J.Law. 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.

186. Tang,D.T., C.Plessy, M.Salimullah, A.M.Suzuki, R.Calligaris, S.Gustincich, and P.Carninci. 2013. Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching. Nucleic Acids Res. 41: e44.

187. Tang,F., C.Barbacioru, S.Bao, C.Lee, E.Nordman, X.Wang, K.Lao, and M.A.Surani. 2010. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. Cell Stem Cell 6: 468-478.

188. Tang,F., C.Barbacioru, Y.Wang, E.Nordman, C.Lee, N.Xu, X.Wang, J.Bodeau, B.B.Tuch, A.Siddiqui, K.Lao, and M.A.Surani. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods 6: 377-382.

189. Tebaldi,T., A.Re, G.Viero, I.Pegoretti, A.Passerini, E.Blanzieri, and A.Quattrone. 2012. Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells. BMC. Genomics 13: 220.

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

191. Tian,B., J.Hu, H.Zhang, and C.S.Lutz. 2005. A large-scale analysis of mRNA polyadenylation of human and

mouse genes. Nucleic Acids Res. 33: 201-212.

192. Tian,B., Z.Pan, and J.Y.Lee. 2007. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. Genome Res. 17: 156-165.

193. Tian,Q., S.B.Stepaniants, M.Mao, L.Weng, M.C.Feetham, M.J.Doyle, E.C.Yi, H.Dai, V.Thorsson, J.Eng, D.Goodlett, J.P.Berger, B.Gunter, P.S.Linseley, R.B.Stoughton, R.Aebersold, S.J.Collins, W.A.Hanlon, and L.E.Hood. 2004. Integrated genomic and proteomic analyses of gene expression in Mammalian cells. Mol. Cell Proteomics. 3: 960-969.

194. Tilgner,H., D.G.Knowles, R.Johnson, C.A.Davis, S.Chakrabortty, S.Djebali, J.Curado, M.Snyder, T.R.Gingeras, and R.Guigo. 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.

195. Travers,K.J., C.S.Chin, D.R.Rank, J.S.Eid, and S.W.Turner. 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. 38: e159.

196. Ule,J., K.B.Jensen, M.Ruggiu, A.Mele, A.Ule, and R.B.Darnell. 2003. CLIP identifies Nova-regulated RNA networks in the brain. Science 302: 1212-1215.

197. Ulitsky,I., A.Shkumatava, C.H.Jan, A.O.Subtelny, D.Koppstein, G.W.Bell, H.Sive, and D.P.Bartel. 2012. Extensive alternative polyadenylation during zebrafish development. Genome Res. 22: 2054-2066.

198. Unneberg,P., A.Wennborg, and M.Larsson. 2003. Transcript identification by analysis of short sequence tags--influence of tag length, restriction site and transcript database. Nucleic Acids Res. 31: 2217-2226.

199. Valen,E., G.Pascarella, A.Chalk, N.Maeda, M.Kojima, C.Kawazu, M.Murata, H.Nishiyori, D.Lazarevic, D.Motti, T.T.Marstrand, M.H.Tang, X.Zhao, A.Krogh, O.Winther, T.Arakawa, J.Kawai, C.Wells, C.Daub, M.Harbers, Y.Hayashizaki, S.Gustincich, A.Sandelin, and P.Carninci. 2009. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. Genome Res. 19: 255-265.

200. Vanderperre,B., J.F.Lucier, C.Bissonnette, J.Motard, G.Tremblay, S.Vanderperre, M.Wisztorski, M.Salzet, F.M.Boisvert, and X.Roucou. 2013. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. PLoS. One. 8: e70698.

201. Velculescu,V.E., L.Zhang, B.Vogelstein, and K.W.Kinzler. 1995. Serial analysis of gene expression. Science 270: 484-487.

202. Vitezic,M., T.Lassmann, A.R.Forrest, M.Suzuki, Y.Tomaru, J.Kawai, P.Carninci, H.Suzuki, Y.Hayashizaki, and C.O.Daub. 2010. Building promoter aware transcriptional regulatory networks using siRNA perturbation and deepCAGE. Nucleic Acids Res. 38: 8141-8148.

203. Vogel,C., R.S.Abreu, D.Ko, S.Y.Le, B.A.Shapiro, S.C.Burns, D.Sandhu, D.R.Boutz, E.M.Marcotte, and L.O.Penalva. 2010. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol. Syst. Biol. 6: 400.

204. Wan,Y., K.Qu, Q.C.Zhang, R.A.Flynn, O.Manor, Z.Ouyang, J.Zhang, R.C.Spitale, M.P.Snyder, E.Segal, and H.Y.Chang. 2014. Landscape and variation of RNA secondary structure across the human transcriptome. Nature 505: 706-709.

205. Wang,E.T., N.A.Cody, S.Jog, M.Biancolella, T.T.Wang, D.J.Treacy, S.Luo, G.P.Schroth, D.E.Housman, S.Reddy, E.Lecuyer, and C.B.Burge. 2012. Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. Cell 150: 710-724.

206. Wang,L., R.D.Dowell, and R.Yi. 2013a. Genome-wide maps of polyadenylation reveal dynamic mRNA 3'-end formation in mammalian cell lineages. RNA. 19: 413-425.

207. Wang,T., Y.Cui, J.Jin, J.Guo, G.Wang, X.Yin, Q.Y.He, and G.Zhang. 2013b. Translating mRNAs strongly correlate to proteins in a multivariate manner and their translation ratios are phenotype specific. Nucleic Acids Res. 41: 4743-4754.

208. Warf,M.B., J.V.Diegel, P.H.von Hippel, and J.A.Berglund. 2009. The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. Proc. Natl. Acad. Sci. U. S. A 106: 9203-9208.

209. Wasinger,V.C., M.Zeng, and Y.Yau. 2013. Current status and advances in quantitative proteomic mass spectrometry. Int. J. Proteomics. 2013: 180605.

210. Wen,J., A.Chiba, and X.Cai. 2010. Computational identification of tissue-specific alternative splicing elements in mouse genes from RNA-Seq. Nucleic Acids Res. 38: 7895-7907.

211. Wethmar,K. 2014. The regulatory potential of upstream open reading frames in eukaryotic gene expression.

Wiley. Interdiscip. Rev. RNA.

212. Wilkening,S., V.Pelechano, A.I.Jarvelin, M.M.Tekkedil, S.Anders, V.Benes, and L.M.Steinmetz. 2013. An efficient method for genome-wide polyadenylation site mapping and RNA quantification. Nucleic Acids Res. 41: e65.

213. Witten,J.T. and J.Ule. 2011. Understanding splicing regulation through RNA splicing maps. Trends Genet. 27: 89-97.

214. Wu,A.R., N.F.Neff, T.Kalisky, P.Dalerba, B.Treutlein, M.E.Rothenberg, F.M.Mburu, G.L.Mantalas, S.Sim, M.F.Clarke, and S.R.Quake. 2014. Quantitative assessment of single-cell RNA-sequencing methods. Nat. Methods 11: 41-46.

215. Yamashita,R., Y.Suzuki, K.Nakai, and S.Sugano. 2003. Small open reading frames in 5' untranslated regions of mRnas. C. R. Biol. 326: 987-991.

216. Yan,J. and T.G.Marr. 2005. Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. Genome Res. 15: 369-375.

217. Yao,C., J.Biesinger, J.Wan, L.Weng, Y.Xing, X.Xie, and Y.Shi. 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.

218. Yeo,G.W., N.G.Coufal, T.Y.Liang, G.E.Peng, X.D.Fu, and F.H.Gage. 2009. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat. Struct. Mol. Biol. 16: 130-137.

219. Yoon,O.K., T.Y.Hsu, J.H.Im, and R.B.Brem. 2012. Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. PLoS Genet. 8: e1002882.

220. Yu,Y., P.A.Maroney, J.A.Denker, X.H.Zhang, O.Dybkov, R.Luhrmann, E.Jankowsky, L.A.Chasin, and T.W.Nilsen. 2008. Dynamic regulation of alternative splicing by silencers that modulate 5' splice site competition. Cell 135: 1224-1236.

221. Yuan,Y., S.A.Compton, K.Sobczak, M.G.Stenberg, C.A.Thornton, J.D.Griffith, and M.S.Swanson. 2007. Muscleblind-like 1 interacts with RNA hairpins in splicing target and pathogenic RNAs. Nucleic Acids Res. 35: 5474-5486.

222. Zhang,C., K.Y.Lee, M.S.Swanson, and R.B.Darnell. 2013. Prediction of clustered RNA-binding protein motif sites in the mammalian genome. Nucleic Acids Res. 41: 6793-6807.

223. Zhang,H., J.Y.Lee, and B.Tian. 2005. Biased alternative polyadenylation in human tissues. Genome Biol. 6: R100.

224. Zheng,D., A.Frankish, R.Baertsch, P.Kapranov, A.Reymond, S.W.Choo, Y.Lu, F.Denoeud, S.E.Antonarakis, M.Snyder, Y.Ruan, C.L.Wei, T.R.Gingeras, R.Guigo, J.Harrow, and M.B.Gerstein. 2007. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res. 17: 839-851.

225. Zheng,W., L.M.Chung, and H.Zhao. 2011. Bias detection and correction in RNA-Sequencing data. BMC. Bioinformatics. 12: 290.

226. Zhuang,F., R.T.Fuchs, Z.Sun, Y.Zheng, and G.B.Robb. 2012. Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. Nucleic Acids Res. 40: e54.