MECHANISMS CONTROLLING mRNA PROCESSING AND TRANSLATION:

Decoding the regulatory layers defining gene expression through RNA sequencing

Eleonora de Klerk

ISBN: 978-94-6182-593-3

Cover image: The journey of a PhD student
Cover design and layout: A.G. Termorshuizen and E. de Klerk
Printed by: Offpage

Publication of this book was financially supported, in part, by the Department of Human Genetics, LUMC.

MECHANISMS CONTROLLING mRNA PROCESSING AND TRANSLATION:

Decoding the regulatory layers defining gene expression through RNA sequencing

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 30 september 2015
klokke 11.15 uur

door

Eleonora de Klerk
geboren te Catania, Italië
in 1983

Promotor:	prof.dr. J.T. den Dunnen
Co-promotor:	dr. P.A.C. ’t Hoen
Promotiecommissie:	prof.dr. P. Devilee
	prof.dr. R. Agami[1]
	dr. A.G. Jochemsen
	dr. M. von Lindern[2]
	prof.dr. B. Wieringa[3]

[1] Nederlands Kanker Instituut (NKI), Amsterdam; Erasmus MC, Rotterdam
[2] Sanquin Research and Landsteiner Laboratory, AMC/UvA, Amsterdam
[3] Radboud UMC, Nijmegen

"Research is to see what everybody has seen,
and to think what nobody else has thought"

Albert Szent-Györgyi
Bioenergetics, 1957(*)

(*) Concise version, originally from
Arthur Schopenhauer
Parerga und Paralipomena, 1851

TABLE OF CONTENTS

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.

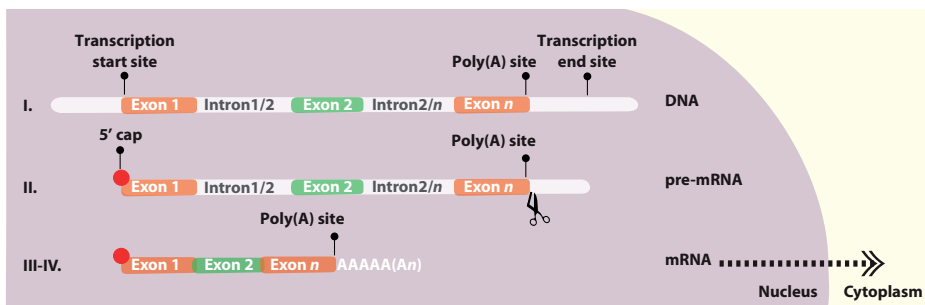


Figure 1. Biogenesis of an mRNA. Schematic representation of capping, splicing and polyadenylation.

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

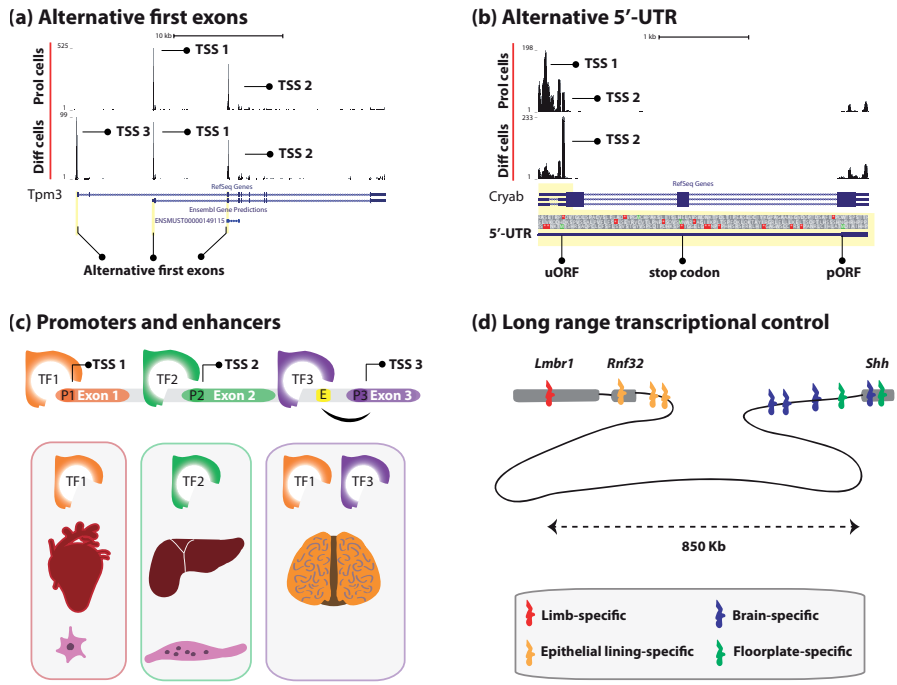


Figure 2. Alternative transcription initiation. (a) Data from a DeepCAGE experiment showing alternative transcription start sites (TSS) used during muscle differentiation in proliferating myoblasts and differentiated myotubes [65]. In the Tpm3 gene different promoters lead to the formation of transcripts 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

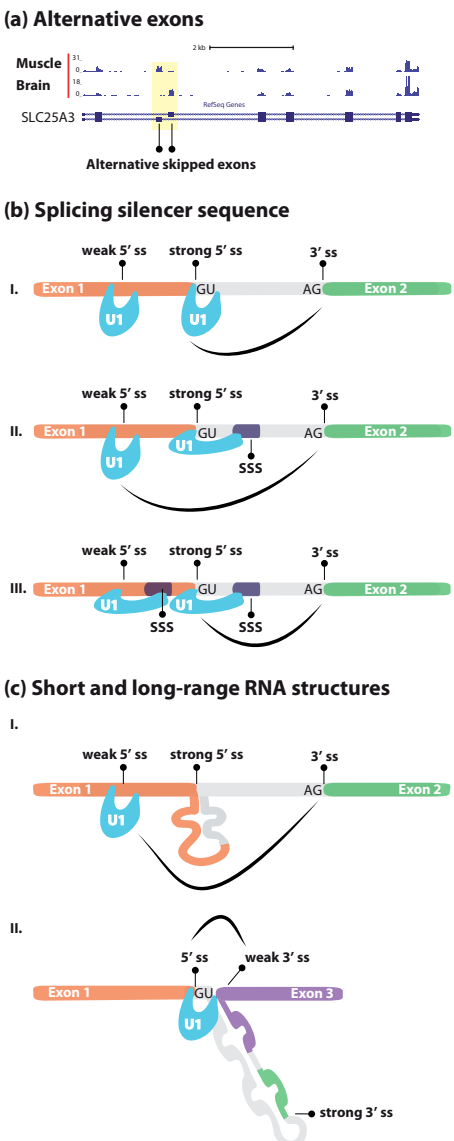


Figure 3. Alternative splicing. (a) Data from an RNAseq experiment showing tissue-specific alternative splicing [129]. The SLC25A3 gene is differentially spliced in brain and muscle tissues through exon skipping. **(b)** Alternative splicing regulated by silencer sequences. In (I) the U1 snRNP splicing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

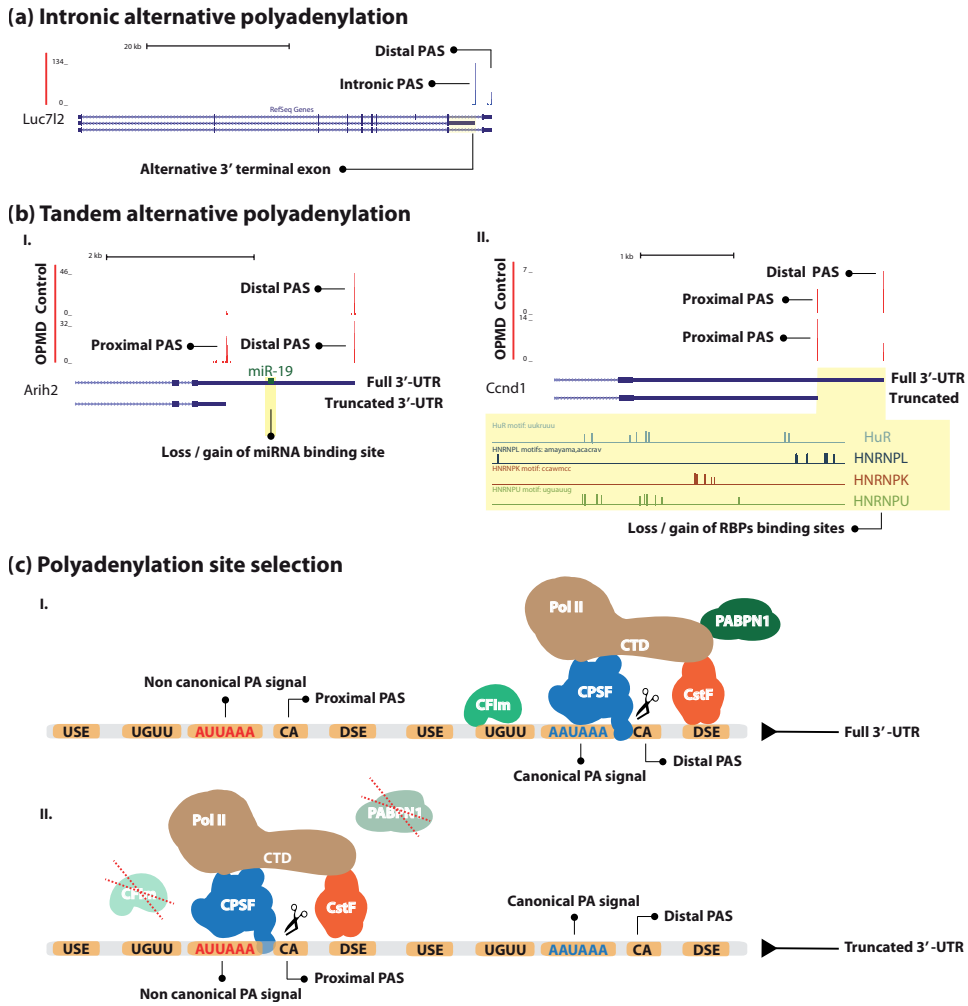


Figure 4. Alternative polyadenylation. (a) Data from a Poly(A)-seq experiment [this thesis, Chapter 2] showing alternative polyadenylation (APA) in the intron of Luc7l2 gene, leading to an intronic proximal PAS located in a different terminal exon, giving rise to transcript variants with different ORFs. **(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

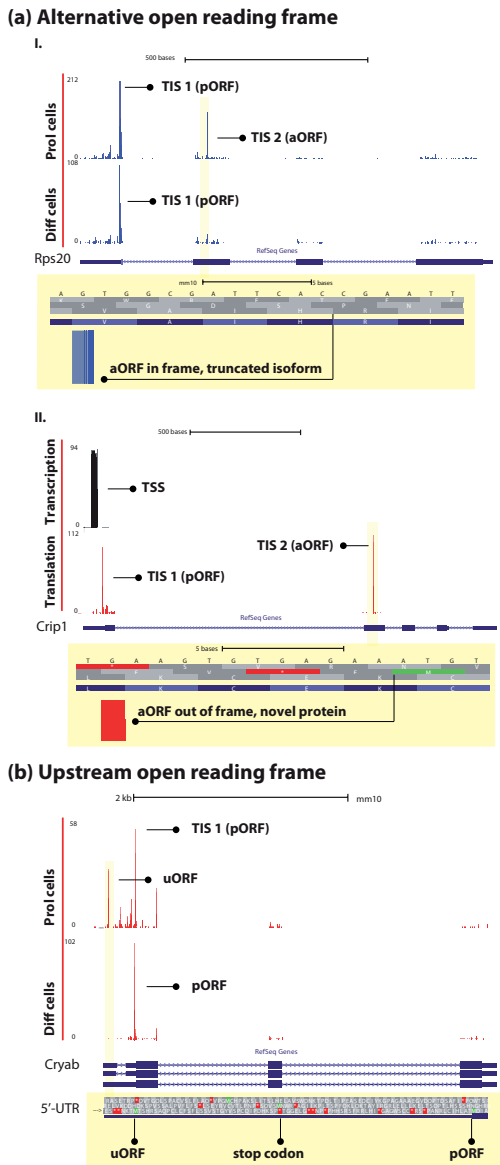


Figure 5. Alternative translation initiation. Alternative translation initiation sites (TISs) detected by ribosome profiling [this thesis, Chapter 4]. **(a)** Examples of alternative TISs leading to alternative open reading frames (aORFs) in frame (I) or out-of-frame (II) with the primary ORF. In the 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.

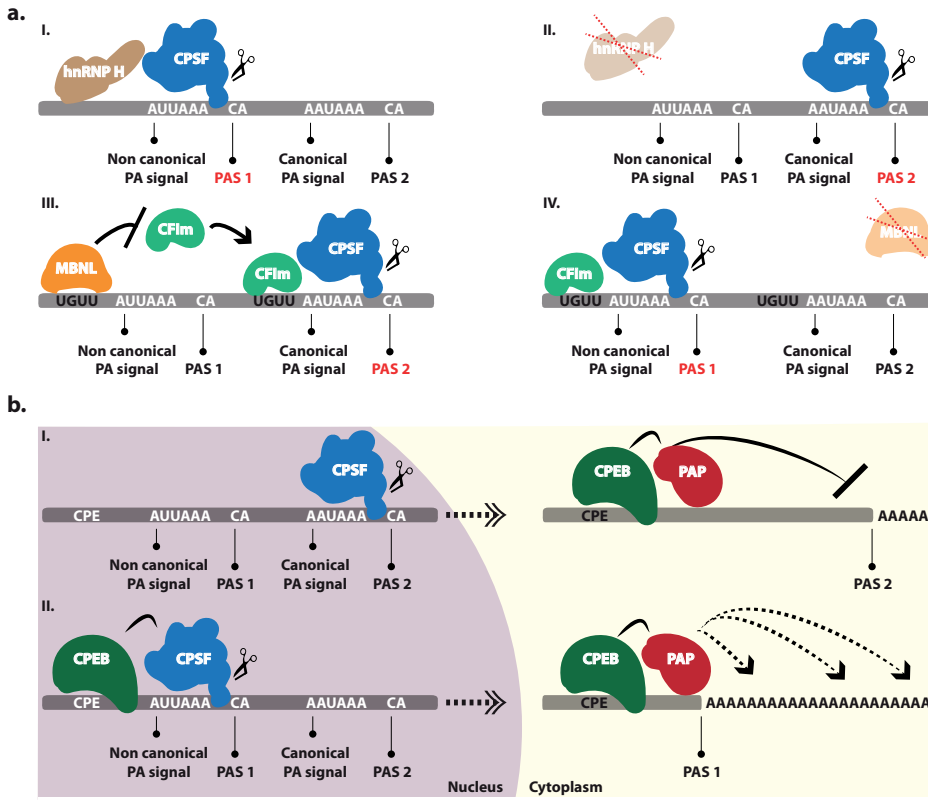


Figure 6. Coupled regulatory mechanisms. (a) Tandem alternative polyadenylation (APA) regulated by splicing factors. The RNA binding proteins hnRNP H and MBNL regulate APA in opposing ways. In the presence of hnRNP H (I), the Cleavage and Polyadenylation Specifity Factor (CPSF) binds weaker non-canonical 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.

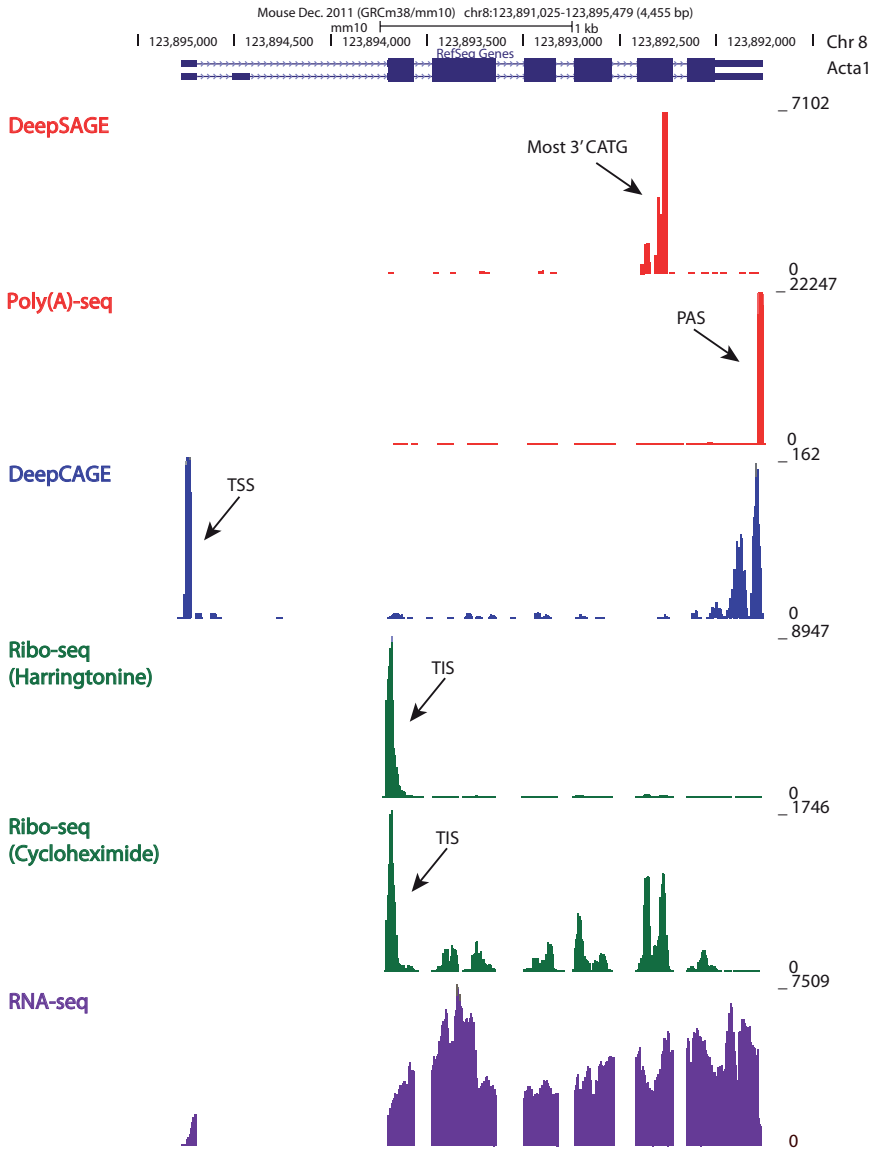


Figure 7. A screenshot from UCSC Genome Browser (http://genome.ucsc.edu) displaying the different regions sequenced by tag-based and shot-gun methods in Acta1 gene. The y-axis represents the coverage, corresponding to the number of reads mapping at each location. Six independent traces are shown. The 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.

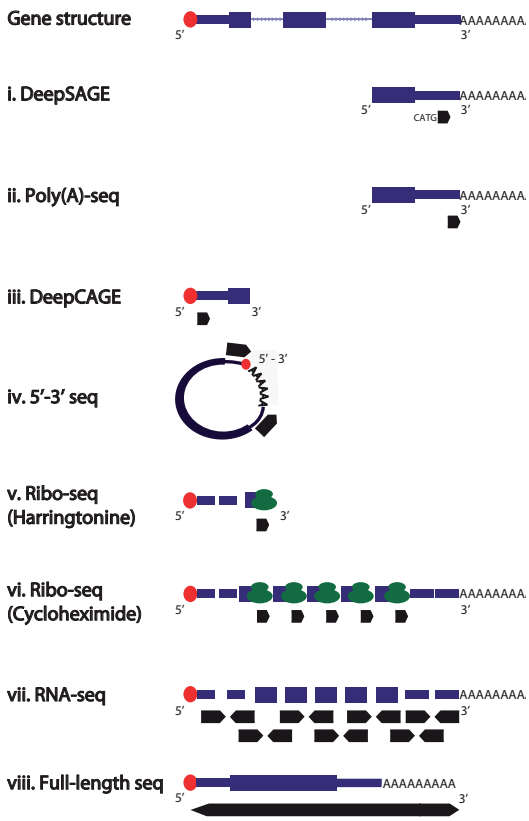


Figure 8. Schematic representation of sequencing reads generated by tag-based (i-iv), shot-gun (v-vii) or full-length (viii) sequencing. Thick black arrows indicate the sequenced reads. Paired-end reads are displayed by two opposite black arrows. Red circles indicate the 5′ cap structure. Ribosomes are 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

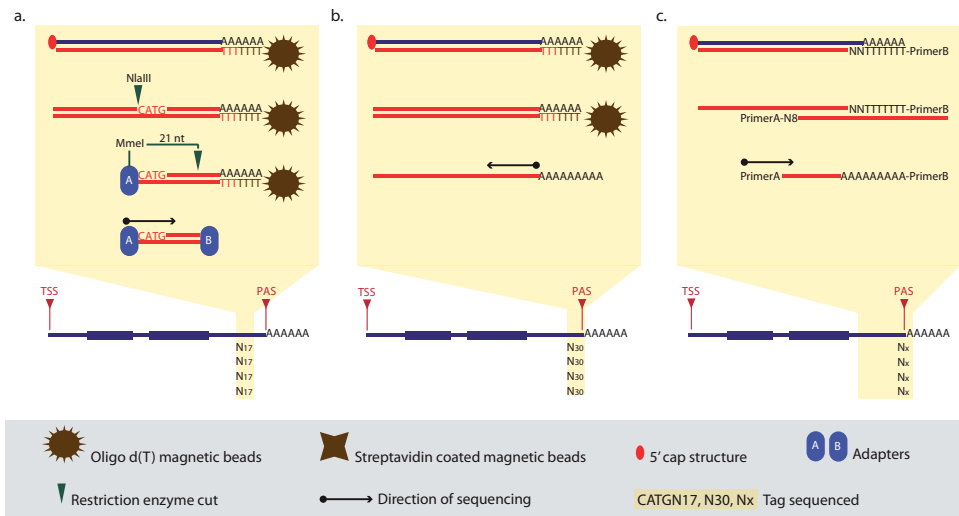


Figure 9. 3'end sequencing methods. (a) In DeepSAGE poly(A)+ RNAs are captured by oligo d(T) magnetic beads and reverse transcribed. cDNA is digested with NlaIII and adapter A is ligated. A second digestion with MmeI generates a 21-bp tag, and adaptor B is ligated to the 3' end. The construct is amplified 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.

CHAPTER 2

POLY(A) BINDING PROTEIN NUCLEAR 1 LEVELS AFFECT ALTERNATIVE POLYADENYLATION

Eleonora de Klerk, Andrea Venema,
S. Yahya Anvar, Jelle J. Goeman, OuHua Hu, Capucine Trollet,
George Dickson, Johan T. den Dunnen, Silvère M. van der Maarel,
Vered Raz and Peter A.C. 't Hoen.

Nucleic Acids Res. 2012 Oct; 40(18): 9089–9101.
doi: 10.1093/nar/gks655.

ABSTRACT

The choice for a polyadenylation site determines the length of the 3′-untranslated region (3′-UTRs) of an mRNA. Inclusion or exclusion of regulatory sequences in the 3′-UTR may ultimately affect gene expression levels. Poly(A) binding protein nuclear 1 (PABPN1) is involved in polyadenylation of pre-mRNAs. An alanine repeat expansion in PABPN1 (exp-PABPN1) causes oculopharyngeal muscular dystrophy (OPMD).

We hypothesized that previously observed disturbed gene expression patterns in OPMD muscles may have been the result of an effect of PABPN1 on alternative polyadenylation, influencing mRNA stability, localization and translation. A single molecule polyadenylation site sequencing method was developed to explore polyadenylation site usage on a genome-wide level in mice overexpressing exp-PABPN1. We identified 2012 transcripts with altered polyadenylation site usage. In the far majority, more proximal alternative polyadenylation sites were used, resulting in shorter 3′-UTRs. 3′-UTR shortening was generally associated with increased expression. Similar changes in polyadenylation site usage were observed after knockdown or overexpression of expanded but not wild-type PABPN1 in cultured myogenic cells.

Our data indicate that PABPN1 is important for polyadenylation site selection and that reduced availability of functional PABPN1 in OPMD muscles results in use of alternative polyadenylation sites, leading to large-scale deregulation of gene expression.

INTRODUCTION

Poly(A) binding protein nuclear 1 (PABPN1) is a ubiquitous protein involved in polyadenylation of pre-mRNAs (1–4). An expansion mutation in the polyalanine repeat in the N-terminus of PABPN1 causes oculopharyngeal muscular dystrophy (OPMD) (5). OPMD is an autosomal dominant, late-onset and progressive muscle disorder. The expanded PABPN1 (exp-PABPN1) accumulates in insoluble nuclear inclusions in affected muscles of OPMD patients (6). We have previously shown that the muscle mRNA expression profiles of OPMD patients and animal models are widely different from controls (7,8). However, it is not clear if this is directly related to the function of PABPN1 in polyadenylation.

Polyadenylation of mRNAs requires a range of multi-subunit protein complexes. The cleavage and polyadenylation specificity factor (CPSF), the cleavage stimulation factor (CstF)n and other proteins are involved in the endonucleolytic cleavage at the poly(A) cleavage site (polyadenylation site) preceding the addition of the poly(A) tail (9). Poly(A) polymerase (PAP), PABPN1 and CPSF are involved in the addition of the poly(A) tail itself (10,11). The assembly of the 3′-end processing machinery is directed by specific RNA sequences: the polyadenylation signal (consensus sequence AAUAAA, recognized by CPSF), the downstream sequence element (recognized by CstF (12,13)) and the upstream sequence element (14–16). So far, two major roles for PABPN1 in polyadenylation have been established. PABPN1 increases the processivity of PAP during the elongation of the tail (10,17), and it controls the length of the poly(A) tail to ~250 nucleotides (1–3). Polyadenylation at different positions in the mRNA increases the variety of transcripts (18). Polyadenylation sites within different exons or introns give rise to alternative 3′-terminal exons and transcripts coding for different protein isoforms (19). Polyadenylation sites located at different positions in the same 3′-untranslated region (3′-UTR) give rise to transcript variants that differ in the length of the 3′-UTR. Shortening or lengthening of the 3′-UTR may result in the loss or gain of regulatory elements, such as miRNA binding sites or binding sites for proteins that can stabilize or destabilize the transcript (20,21). This may affect mRNA stability and overall gene expression. Alternative polyadenylation is a common regulatory mechanism in various developmental and physiological processes such as the immune response (21–24), and it may also contribute to carcinogenesis (25).

To investigate the role of PABPN1 in alternative polyadenylation, we developed a single molecule sequencing approach for genome-wide detection of polyadenylation sites and studied alternative polyadenylation in A17.1 mice, which overexpress exp-PABPN1 in muscle (26). We further investigated the effects of mutation and modulation of PABPN1 expression levels on polyadenylation site selection in a myogenic cell model. We found that manipulation of PABPN1 expression levels lead to changes in polyadenylation site usage and that reduced PABPN1 levels lead to a general shortening of 3′-UTRs. We suggest an involvement of PABPN1 in polyadenylation site selection and a novel molecular mechanism for OPMD, where sequestering of exp-PABPN1 in insoluble inclusions interferes with normal polyadenylation and disrupts gene expression patterns.

MATERIALS AND METHODS

RNA isolation

Total RNA was extracted from quadriceps muscles of mice overexpressing the exp-PABPN1 (A17.1 mouse model) (26) and FBV mice using RNA Bee solution (Tel-Test, Bio-Connect) after homogenization of the tissue with glass beads (diameter: 1.0 mm) on the BeadBeater (BioSpec) according to the manufacturer's instructions. Quadriceps have been isolated from A17.1 and FVB mice aged 6 and 26 weeks (N=3 per group). RNA quality and concentration was determined on the Bioanalyzer (Agilent) with RNA 6000 Nano kit (RIN>8).

Sample preparation and polyadenylation site single molecule sequencing method

Poly(A)+ RNAs were isolated from 2 µg of total RNA using oligo(dT)25 magnetic beads (Invitrogen) according to the manufacturer's instructions. First strand cDNA synthesis (SuperScript III, Invitrogen) was performed on the beads primed by oligo(dT)25 following manufacturer's protocol. RNase H (Invitrogen) treatment and second strand synthesis were carried out at 16°C for 2.5 h. dsDNA was digested with NlaIII (New England Biolabs (NEB)) for 1 h at 37°C. During Poly(A)+ RNAs capture, first and second strand cDNA synthesis, and dsDNA digestion, RNA and DNA molecules were washed as described in the Tag Profiling Sample Prep Kit (Illumina), using respectively GEX binding and washing buffers, GEX cleaning solution and GEX buffer C and D. dsDNA was heat denatured at 95°C for 2 min in 100 µl of elution buffer (10 mM Tris–Cl, pH 8.5), and the eluted second strand cDNA was precipitated with 0.1 volumes of sodium acetate (3 M, ph 4.8–5.2), 1 µl of co-precipitant Pellet-Paint (EMD4Biosciences) and 2.5 volumes of ethanol 100% overnight at -20°C. Poly(A) tailing reaction and blocking reaction were performed using Terminal Transferase kit (NEB) with the following modifications. Poly(A) tailing reaction was carried out for 30 min at 42°C using 5 units of Terminal Transferase enzyme and 4 µl of 50 µM dATPs (Helicos PolyA tailing dATP) in presence of 2 µl of 2.5 mM CoCl2 and 0.2 µl of BSA (NEB). Blocking reaction was performed with 0.5 µl of 200 µM biotinylated ddATP (PerkinElmer) at 37°C for 1 h, followed by 20 min at 70°C. Biotinylated ddATPs are used to measure the concentration of the samples by biotin–streptavidin approach (OptyHyb assay, Helicos). Seventy-five microliters of each sample at a concentration of 200 pM was directly hybridized to the flow cell and ssDNA was sequenced on the HeliScope platform according to the manufacturer's instructions. A total of 12 samples were sequenced in individual lanes.

Data analysis

Microarray analysis

Differential expression analysis of A17.1 and wild-type mice was described previously (8). The expression profiles were generated on the Illumina Mouse Sentrix-6 v2 Beadchip platform. This platform generally contains one probe per transcript and for genes with multiple transcript variants, multiple probes may be present. Only genes with at least two probes were included. The number of genes with at least one significantly upregulated and at least one significantly downregulated probe was counted (false discovery rate (FDR)<0.05). We evaluated two control mouse datasets profiled on the exact same Beadchip platform using the same analysis procedure (normalization, differential expression and probe annotation) (27,28).

Sequencing analysis

During the sequencing process on the HeliScope platform some nucleotides can remain unlabelled and appear as deletions. Therefore, standard alignment software is not suitable for Helicos data. Preprocessing of raw reads, including filtering and alignment to the mouse genome (mm9), was performed using the basic pipeline (version 1.1.498.63) from the Helicos, described at http://open.helicosbio.com/helisphere_user_guide/ch04s07.html. We ran the alignment with a maximum of 100 possible mapping locations per read. A second filtering step after alignment was performed, filtering for the best location of each read and setting as threshold a minimum score of 4 and a minimum sequence length of 25 nucleotides. Downstream analysis was carried out using Custom Perl scripts to report the estimated number of tags in each region.

Polyadenylation site assignment and annotation

After alignment, we used the 5′-end of the reads to identify the position of the polyadenylation site and generated wiggle files using a Custom Perl script. Polyadenylation sites within a distance of 10 nt were clustered together and assembled into regions. Clustering of regions containing polyadenylation sites (Poly(A) clusters) was carried out in repeating cycles until every region was separate by a gap of at least 10 nucleotides. Regions were annotated based on the ENSEMBL GENE 63 (Sanger, UK) database, NCBI mouse genome build 37. Further statistical analysis focused only on reads mapping to the expanded 3′-UTR, up to 2 kb downstream of the annotated 3′-UTR. To this end, we used a Custom R script to filter out reads that were not mapping within the expanded 3′-UTR.

Differential expression analysis and functional annotation

The statistical programming language R (version R 2.12.0) was used for analysis of differential expression between A17.1 and control mice. The analysis was performed using the R Bioconductor package edgeR (29) (version 2.0.4). A negative binomial model was fitted and a common dispersion was estimated for all the tags prior testing procedures. Exact P-values were computed using the exact test and adjusted for multiple testing according to Benjamini and Hochberg method (30). Only reads mapping in the expanded 3′-UTR of a transcript were used for expression analysis. Poly(A) clusters containing just one read were filtered out prior any statistical analysis to reduce noise. KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis was performed using DAVID Functional Annotation Tool (31).

Statistical model to identify transcripts with usage of alternative polyadenylation sites

We fitted a logistic generalized linear mixed model to the counts for each polyadenylation site in a particular 3′-UTR using the R Bioconductor package lme4 (http://cran.r-project.org/web/packages/lme4/index.html). The counts nij for polyadenylation site j of mouse i were modeled as binomial with parameters Ni and pij. Here, Graphic, where k is the number of polyadenylation sites for a particular 3′-UTR, and Ni is the total number of reads for all polyadenylation sites for that 3′-UTR in mouse i. The log odds of the parameter pij was modeled using fixed effects for polyadenylation site and OPMD status, with their interaction, combined with a random intercept and polyadenylation site effect within a mouse. We tested for the presence of an OPMD effect with a chi-squared likelihood ratio test, using as the null hypothesis the same model, but with the OPMD effect and the OPMD-polyadenylation site

interaction set to zero. Modeling the fraction for each polyadenylation site within the total counts of all polyadenylation sites of the same transcript allows assessment of changes in relative, rather than absolute frequency of the polyadenylation site.

Sequence motif analysis

We used DREME (32) to identify enriched sequence motifs. From the list of transcripts showing changes in polyadenylation site usage, we expanded the sequence of every polyadenylation site up to 50 nucleotides upstream (accounting for the genomic strand on which the polyadenylation site was located), and grouped the sequences into two categories, one containing only the most distal (3′) polyadenylation site and the other containing all the other more proximal (5′) polyadenylation sites. All sequences were masked for repeats. We first ran DREME limiting the search to a maximum width of six bases for each motif, according to the length of known polyadenylation signals sequences with the full 3′-UTR sequences as background (18). Subsequently, a discriminative motif search was performed using the sequences upstream the distal polyadenylation site as negative background for the proximal polyadenylation site and vice versa.

Cell culture

Mouse myoblasts C2C12 were grown on collagen-coated plates in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal bovine serum, 1% glucose and 2% glutamax (Invitrogen). Differentiation was induced by serum deprivation for 7 days, by culturing in DMEM supplemented with 2% fetal bovine serum, 1% glucose and 2% glutamax. Cells were grown under 10% CO2.

Lentiviral transduction

Overexpression of the human wild-type PABPN1 and the exp-PABPN1 in C2C12 cells was achieved by transduction of lentiviruses expressing CFP-Ala10-PABPN1 and YFP-Ala16-PABPN1, which were generated from expression vectors previously described (33). For the downregulation of the endogenous Pabpn1, C2C12 were transduced with Short Hairpin (shRNA)-expressing lentiviruses. The shRNA lentiviral plasmids (pLKO.1-puro) were obtained from Sigma-Aldrich (TRCN0000102536 and TRCN0000000121). Lentiviral particles were produced as previously reported (34). C2C12 myoblasts were plated in either a 6-well (100000 cells/well, overexpression) or 12-well (35000 cells/well, knock-down) plate, and transduction was carried out in the presence of polybrene (8μg/ml) after an initial wash with EDTA (0.54mM EDTA in PBS). The titers of the lentiviruses were determined by the p24 elisa kit (Retro-Tek, ZeptoMetrix Corp.). A virus titer in the range of 35ng p24 per 105 cells was used for YFP-Ala16-PABPN1, a range of 35–70ng p24 per 105 cells for CFP-Ala10-PABPN1, and a titer of 80ng p24 per 105 cells was used for shRNA-TRCN0000102536 and shRNA-TRCN0000000121. Culture medium was replaced after 48h by differentiation medium and RNA was extracted after 7 days from C2C12 myotubes transduced with CFP-Ala10-PABPN1 and YFP-Ala16-PABPN1. CMV-GFP vector was used as negative control. The levels of CFP and YFP mRNA in transduced cells were checked by quantitative reverse transcriptase–polymerase chain reaction (qRT–PCR). C2C12 transduced with shRNA were treated with puromycin (40ng/ml) in fresh proliferating medium after 48h of transduction, and RNA was extracted after 24h from myoblasts. Experiments were performed in three independent wells and repeated on different days.

RNA isolation, reverse-transcription, qRT–PCR

RNA was extracted from C2C12 cells using the NucleoSpin RNA II kit (Macherey-Nagel) according to the manufacturer's instructions. cDNA was synthesized from 1µg of RNA using BioScript MMLV Reverse Transcriptase (Bioline) with 40ng of random hexamer and oligo(dT)18 primers following manufacturer's instructions. qRT–PCR was performed on the LightCycler 480 (Roche) using 2X SensiMix reagent (Bioline). Each measurement was performed in duplicates or triplicates. mRNA expression levels and PCR efficiency were determined using the LinRegPCR program (35) v.11.1 according to the described method (36). Relative expression levels of Pabpn1 were normalized to the geometric mean of two housekeeping genes, Gapdh and Hprt. Ratios between short and long variants were then calculated. Primers for qRT–PCR were designed using Primer3Plus program. Primers for poly(A) site switches validation in qRT–PCR were designed in the sequences proximal to the detected polyadenylation site of six candidate genes (Arih1, Atp1b1, Psmd14, Psme3, Tmod1 and Vldlr). Primers for Pabpn1 were designed to detect both the human exogenous and mouse endogenous mRNA. Primer sequences are listed in Supplementary Table S1.

RNA immunoprecipitation

RNA immunoprecipitation (RIP) was performed using C2C12 myoblasts extracts. Experiments were performed in 6-well plates 90% confluent. Cells were trypsinized (0.05% Trypsin–EDTA (Invitrogen)) and cell pellets were recovered by centrifugation (10min at 290 rcf). One-sixth of the cell pellet was used for total RNA extraction (input RNA) using NucleoSpin RNA II kit (Macherey-Nagel) according to the manufacturer's instructions. Five-sixth of the cell pellet were resuspended in 1ml lysis buffer (100mM KCl, 5mM MgCl2, 10mM HEPES (pH 7.0), 0.5% NP40, 1mM DTT, 80 U RNAse Inhibitor (Roche), Protease Inhibitor Cocktail (Roche)). The lysate was passed five times through a 29G needle and incubated for 10min on ice. The lysates was then clarified by centrifugation at 16000 rcf for 5min at 4°C and supernatant was recovered. Protein concentration was determined using Bradford assay. Immunoprecipitation of PABPN1 was conducted with 700µg of protein extract using the VHH-3F5 antibody (37) (overnight incubation at 4°C) and immunocomplexes were isolated with Protein A Sepharose beads (GE Healthcare) pre-coated with sperm-DNA. Following extensive washing with lysis buffer, RNA was isolated from the immunocomplexes using the NucleoSpin RNA II kit according to the manufacturer's instructions. To validate the RIP, a parallel immunoprecipitation was conducted with 150µg of protein extract for western blot analysis. The immunocomplexes were heat denatured (95°C for 5min) and resolved by sodium dodecyl sulphate–polyacrylamide gel electrophoresis on a 10% polyacrylamide gel followed by western blot. PABPN1 was detected with rabbit anti-PABPN1 (LSBio, 1:10000) using goat anti-rabbit as secondary antibody (IRDye800CW, Licor,1:10000) and Tubulin was used as loading control and negative control for immunoprecipitation (tubulin was detected with mouse monoclonal anti-tubulin, Sigma Aldrich, 1:2000, and goat anti-mouse IRDye680CW, Licor, 1:8000). Signals were visualized with the Odyssey Infrared Imaging System (LI-Cor Biosciences).

RESULTS

Microarray analysis show alternative polyadenylation events in A17.1 mice

In a previous microarray expression profile study comparing A17.1 mice with FVB parental mice (8), we noticed that probes in the same gene frequently detect discordant changes in expression. We determined that for ~8% of the genes which are represented by at least two probes on the microarray, opposite changes in expression levels were detected (**Table 1**). In contrast, two disease models and age-matched mouse datasets, where RNA was profiled on the same microarray platform, showed 10–20-fold fewer genes with probes showing opposite changes in expression direction (**Table 1**). This gave a first indication for frequent alternative polyadenylation events in A17.1 mice.

Mouse model	A17.1 PABPN1 overexpression	A17.1 PABPN1 overexpression	TNX [27] knockout	TBX3b [28] overexpression
Age (weeks)	6	26	32	8
Total number of deregulated genes	6012	4010	1111	1758
Number of deregulated genes with multiple probes	2819	1890	488	908
Number of deregulated genes with probes showing opposite direction in expression level	239	118	3	3
Percentage of deregulated genes with probes showing opposite direction in expression level	8.5	6.2	0.6	0.3

Table 1. Frequency of alternative polyadenylation events identified on microarrays

Polyadenylation site single molecule sequencing

The Illumina microarray platform contains only a limited number of genes where alternative transcripts from the same gene are interrogated by different probes and is therefore not suited for comprehensive analysis of alternative polyadenylation. To identify polyadenylation sites on a genome wide level in a largely unbiased way, we developed a polyadenylation site sequencing method based on the amplification-free HeliScope single molecule sequencer technology (**Figure 1**). To retain only the 3'-ends of the mRNAs and reduce internal priming events, poly(A)+ RNAs were captured on oligo(dT) beads, followed by first and second strand cDNA synthesis on the beads and a restriction enzyme digestion. Double stranded molecules were denatured, and the poly(A) stretch downstream of the polyadenylation site made the cDNA molecules directly amenable for sequencing on the HeliScope flow cell. Sequencing started directly after the poly(A) tail, and thus at the polyadenylation site. This method enables strand-specific mapping of alternative polyadenylation sites at single nucleotide resolution. In contrast to other recently developed methods for genome-wide assessment of alternative polyadenylation events (24,38), our method is amplification- and ligation-free, resulting in lower quantification bias.

Alternative polyadenylation sites in mouse transcripts

We used the single molecule sequencing method to investigate polyadenylation site usage in mouse muscles. Sequencing was performed on RNA isolated from quadriceps muscles of six individual A17.1 mice and six control mice (FVB). Mice of two different age groups were combined since there were no

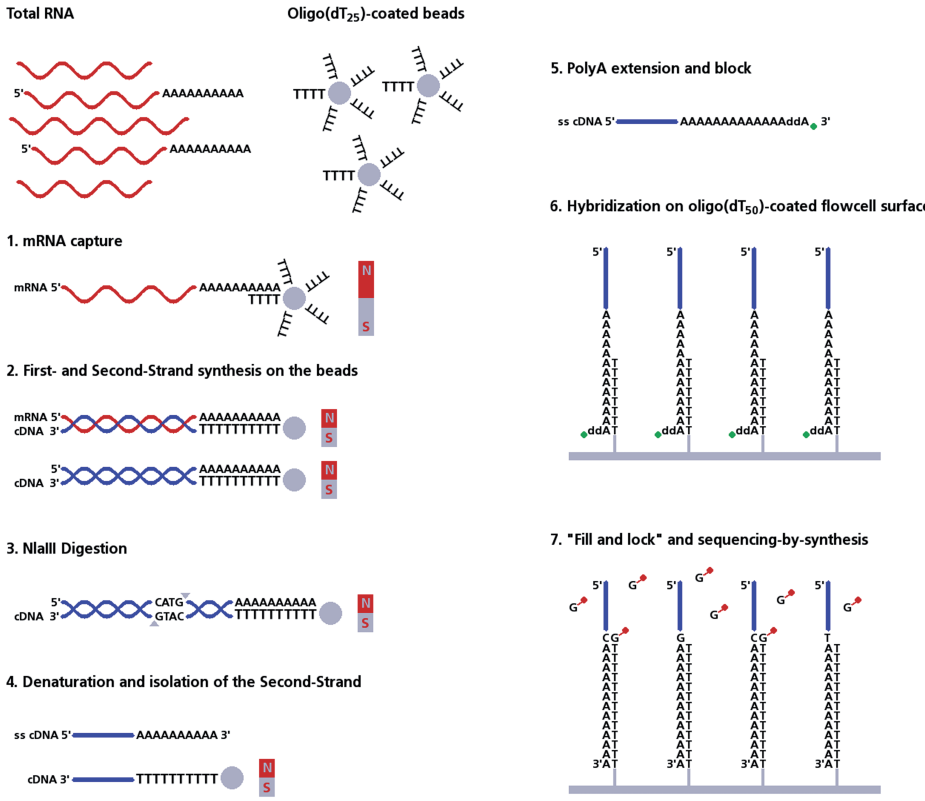


Figure 1. Single molecule polyadenylation site sequencing method. The procedure of the method is detailed in the results and material and methods sections. Red and grey boxes represent magnetic stands used to capture the oligo(dT) magnetic beads. Green anchors represent biotin labels, and red anchors represent fluorescent labels for nucleotide analogues.

significant differences in polyadenylation site usage between young (6 weeks) and adult (26 weeks) mice. On average, we obtained 24 million reads per sample. An average of ~4.8 million aligned reads per sample remained after filtering for low quality reads inherent to single molecule sequencing, and after applying a stringent threshold on the confidence with which the reads could be aligned to a position in the genome. Approximately 60% of those mapped to annotated 3′-UTRs or sequences up to 2 kb downstream of annotated 3′-UTRs (expanded 3′-UTR) and included many not yet annotated transcript ends. A summary of the sequencing results is shown in Supplementary Table S2. To assign the location of every polyadenylation site, we considered the biological heterogeneity of the cleavage site (18,39). Therefore, we clustered together polyadenylation sites located within a window of 10nt and continued with multiple rounds of clustering until all poly(A) clusters were separated by a gap of at least 10 nucleotides. The median width of the clusters of polyadenylation sites was 12 nucleotides, suggesting considerable variation in the exact position of the polyA site (**Figure 2A**). The width of the clusters is larger in case of mitochondrial RNAs, which may be polyadenylated at nearly any position in the transcript (Supplementary Figure S1) probably due to the polyadenylation-dependent degradation mechanism of mitochondrial RNAs (24,40). After removing noise by requiring a coverage

of minimum two reads per polyadenylation site, we detected a total of 32 820 polyadenylation sites. 28 853 polyadenylation sites (88%) located in the expanded 3′-UTR of 11 529 different transcripts. Our analysis showed that 56% of the detected transcripts have multiple polyadenylation sites (**Figure 2B**).

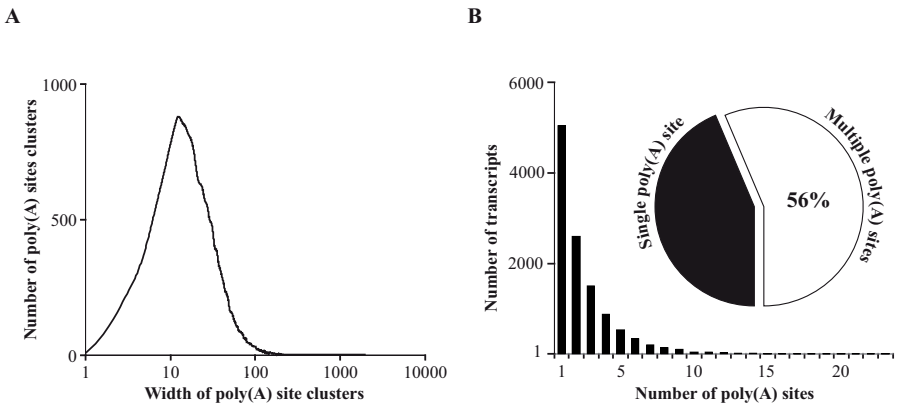


Figure 2. (A) Width of polyadenylation site clusters. The x-axis represents the width of the cluster on a logarithmic scale, the y-axis represents the number of polyadenylation site-clusters. **(B)** Bar graph showing the number of polyadenylation sites detected per transcript. Only polyadenylation sites mapping to the expanded 3'-UTR, and covered by at least two reads are shown. Pie chart shows the percentage of transcripts containing at least two polyadenylation sites.

Widespread changes in polyadenylation site usage in A17.1 mice

A17.1 mice showed large differences in the relative polyadenylation site usage compared with control mice. An illustrative example of a transcript showing a change in polyadenylation site usage is given in **Figure 3A**. We detected two major polyadenylation sites in the 3′-UTR of Psmd14, a distal site at the annotated 3′-end of the transcript and a more proximal site. The distal and the proximal polyadenylation sites, giving rise to variants with long or short 3′-UTRs, can be observed in both FVB and A17.1 mice, but at different levels. In FVB mice, the majority of the reads mapped to the distal polyadenylation site, while in A17.1 mice, the majority mapped to the proximal polyadenylation site. To analyse how widespread this type of switches in preferred polyadenylation site usage in A17.1 mice were, we evaluated the statistical significance of differences in the relative polyadenylation site usage using a generalized linear mixed model on the binomial distribution of the counts for each polyadenylation site in a transcript. From 11 529 transcripts detected with our polyadenylation site sequencing method, 6506 transcripts contained at least two polyadenylation sites. Out of those, 31% showed significant differences (FDR<0.05) in polyadenylation site usage between A17.1 and FVB mice (Supplementary Table S3). We observed a strong preference for the proximal polyadenylation site in A17.1 mice, because transcripts variants with proximal polyadenylation site were mainly upregulated and transcript variants with distal polyadenylation site were mainly downregulated in A17.1 mice (**Figure 3B**). To validate the changes in relative polyadenylation site usage observed by sequencing analysis, qRT–PCR was performed using primer pairs designed just upstream of the detected polyadenylation site. We tested a panel of six genes (Arih1, Atp1b1, Psmd14, Psme3, Tmod1 and Vldlr), identified from the transcriptome study in the mouse model (8) and from the cross-species study for OPMD (7). The ratio between the proximal PCR product (representing the shorter and longer isoforms) and the distal PCR product (representing only the level of the longer isoforms) was elevated for all six genes (**Figure 3C and D**), confirming the shortening of 3′-UTRs in A17.1 mice observed by

polyadenylation site sequencing.

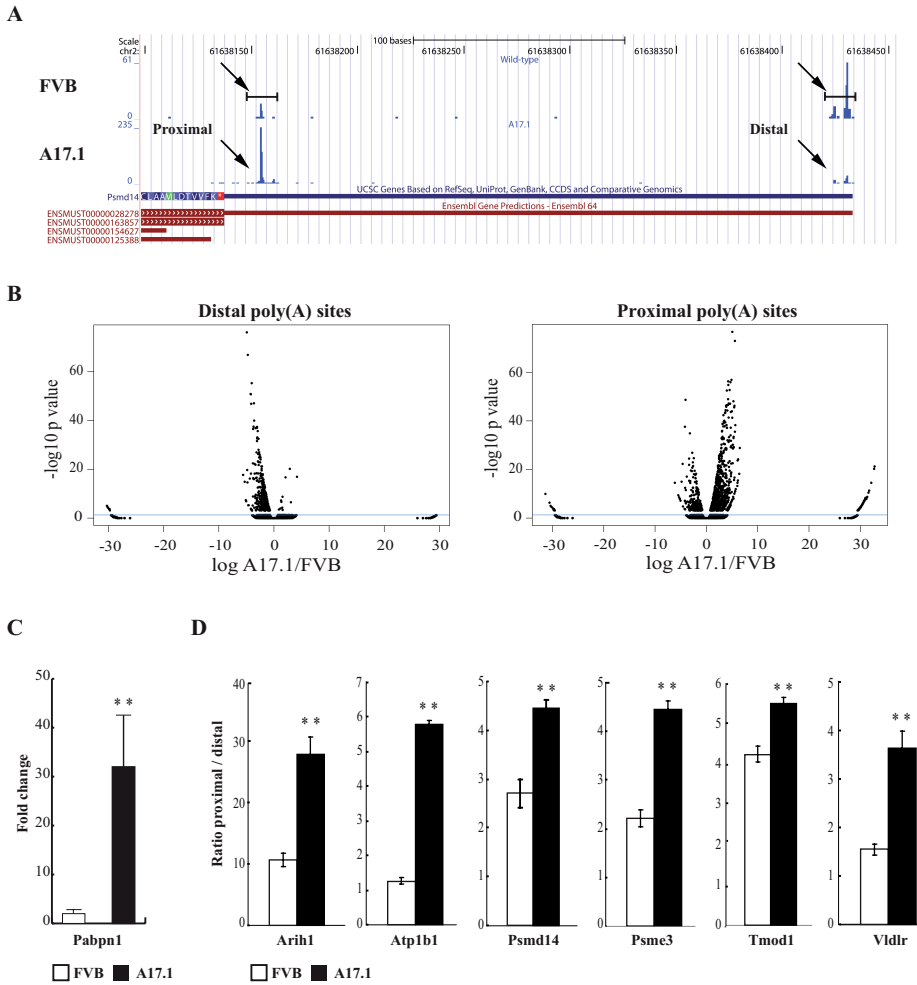


Figure 3. Altered polyadenylation site usage in A17.1 mice. (A) A screenshot of USCS Genome Browser displaying polyadenylation sites in Psdm14 gene. The y-axis represents the coverage of the peaks, corresponding to the number of reads mapping at each polyadenylation site. A control (FVB) and an A17.1 mouse are shown in independent traces. Below the coverage tracks, the four annotated transcripts are shown. The longest transcript (ENSMUST00000028278) contains a 3'-UTR of 297 nucleotides. There are two major polyadenylation sites in this region (indicated by arrows), a distal one (peak location chr2:61,638,431-61,638,432) at the annotated 3'-end of the transcript and a proximal one located 276 nucleotides upstream and just 23 nucleotides downstream of the stop codon (peak location chr2:61,638,155-61,638,156). **(B)** Volcano plots showing transcripts variants containing the distal polyadenylation site (left panel) or more proximal polyadenylation sites (right panel). The y-axis represents the -10log of the multiple testing adjusted p-values, while the x-axis represents the ratio of expression in A17.1 over wild-type mice on a logarithmic scale (base 2). The blue line represents a p-value threshold of 0.05. The outliers present in the graphs represent data points were the total counts in one of the two groups is zero. **(C)** Pabpn1 expression level in quadriceps muscles of A17.1 mice and FVB mice, as measured by qRT-PCR with primers measuring both endogenous and exogenous Pabpn1. **(D)** The ratio of proximal PCR over distal PCR products in A17.1 mice (black bars) and FVB mice (white bars), as measured by qRT-PCR. Values are means + standard deviation for n=6 mice per group (*P<0.05, **P<0.01).

Alteration of the length of the 3'-UTR results in disturbed gene expression patterns

A largely deregulated gene expression pattern in A17.1 mice has already been shown using microarray technology (8). Our polyadenylation site sequencing method can be used to identify differentially expressed transcripts, because every read uniquely identifies a transcript. To investigate the impact of changes in relative polyadenylation site usage on transcript levels, we performed a differential expression analysis using the R Bioconductor package edgeR (29) on the sum of all reads mapping to the expanded 3'-UTR of a transcript. With this procedure, we analysed the combined expression levels of short and long transcripts. From a total of 11 529 transcripts included in the analysis, 3441 were significantly deregulated (FDR<0.05) (Supplementary Table S4). Approximately 60% of the deregulated transcripts were upregulated. The proteasome and ubiquitin-mediated proteolysis pathways were the most significantly deregulated KEGG pathways (Supplementary Table S4), confirming our previous microarray-based results (7,8,41). To assess whether changes in polyadenylation site usage resulted in differences in transcript expression levels, we determined the overlap between deregulated transcripts in A17.1 mice and transcripts showing changes in relative polyadenylation site usage. Out of the 6506 transcripts with two or more polyadenylation sites, 2263 transcripts were differentially expressed between A17.1 and FVB mice (Supplementary Table S6), of which 1249 (55%) were upregulated. The overlap between transcripts with differential polyadenylation site usage and those which were upregulated is highly significant (**Figure 4A and B**). This suggests that shortening of 3'-UTRs generally results in the loss of negative regulatory elements, such as miRNA binding sites, and higher transcript stability.

Alternative polyadenylation sites used in A17.1 mice contain primarily non-canonical polyadenylation signals

To obtain further insight into the mechanism leading to a preferential use of proximal polyadenylation sites in the A17.1 mice, we performed a sequence motif analysis to examine the sequences 50 nucleotides upstream of the distal and proximal polyadenylation sites. Most sequences with distal polyadenylation site contained one of the two canonical polyadenylation signals. The frequency of canonical polyadenylation signals in proximal sequences was lower (**Table 2**, 43% vs. 83%). We then performed a discriminative motif analysis using DREME (32), contrasting the motifs in sequences located upstream of the distal or the proximal polyadenylation site directly. The use of a discriminative approach enables the identification of motifs, which are enriched in only one of the two subsets of sequences. Distal polyadenylation site were enriched for the two canonical polyadenylation signals (**Table 3**). Proximal polyadenylation sites showed very moderate enrichment for nine hexamers, mainly GC-rich. These results indicate that the proximal polyadenylation sites preferentially used in A17.1 mice are predominantly non-canonical and do not contain a strong consensus sequence. We also performed a motif analysis on the sequences 50 nucleotides downstream of the detected polyadenylation site, but did not find enriched motifs there.

PABPN1 levels affect 3'-end processing

To confirm that expression of the exp-PABPN1 alters polyadenylation site usage in muscle cells, we expressed the expanded human PABPN1 in C2C12 myoblasts by transduction of lentiviral particles containing YFP-Ala16-PABPN1. Myoblasts were then fused into myotubes. The overexpression level was assessed by measuring total (endogenous and exogenous) Pabpn1 mRNA levels by qRT–PCR.

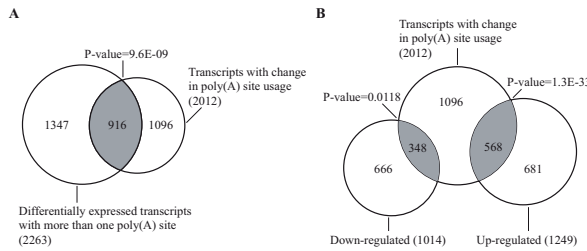


Figure 4. Switches in polyadenylation site induce changes in gene expression. (A) Overlap between deregulated transcripts and transcripts with changes in polyadenylation site usage. **(B)** The overlap is shown for downregulated and upregulated transcripts separately. Indicated P-values were calculated with Fisher's test.

Number of sequences (a)	**Total**	**With AATAAA**	**With ATTAAA**
Distal	1917	1212 (63%)	397 (20%)
Proximal	6508	1855 (28%)	975 (15%)

(a) A DREME motif analysis was performed on sequences 50 nt upstream the detected polyadenylation sites with the full 3'UTR sequence as background

Table 2. Canonical polyadenylation signals in sequences upstrea of polyadenylation

Discriminative analysis (a)	**Motif**	***E*- values**
Distal polyadenylation signals	AATAAA	3.60*E*−122
	ATTAAA	1.00*E*−03
	CCYTCY	7.10*E*−09
	CWGGYC	1.80*E*−06
	GARGAM	7.20*E*−05
	AGTGVC	1.20*E*−03
Proximal polyadenylation signals	GGTGMA	5.20*E*−03
	GCCCAC	1.10*E*−02
	GGGCTY	2.00*E*−02
	CYAGCA	3.60*E*−02
	GMACTA	3.70*E*−02

(a) A DREME descriminative motif analysis was performed to identify enriched motifs upstream of distal polyadenylation signals compared with proximal polyadenylation signals and enriched motifs upstream of proximal polyadenylation signals compared with distal polyadenylation signals.

Table 3. Discriminatif motif analysis in sequences upstream of poliadenylation sites

Pabpn1 was ~4-fold overexpressed (**Figure 5A**), much lower than in the A17.1 mice (**Figure 3C**). Overexpression at this level did not affect cell differentiation or fusion, as assessed by the expression levels of the myogenic markers Myog, Tnnc1, Myh7 and Myf5 (Supplementary Figure S2). We analysed proximal and distal polyadenylation site usage in the muscle cells with the same qRT–PCR assay as used for the A17.1 mice. The ratio between shorter and longer transcripts was significantly increased for five out of six tested genes (**Figure 5B**). Our in vitro data therefore confirm the effect of overexpression of exp-PABPN1on polyadenylation site usage observed in the A17.1 mouse model. The differences, however, were smaller than those found in the A17.1 mice, likely due to lower overexpression levels. We next addressed whether overexpression of wild-type PABPN1 would also affect alternative polyadenylation. C2C12 myoblasts were transduced with lentiviral particles expressing the CFP-Ala10-PABPN1 construct. We did not succeed in obtaining 4-fold overexpression level of wild-type PAPBN1. To be able to compare the effects of exp-PABPN1 and wild-type PABPN1, we overexpressed both forms 1.5-fold (**Figure 5C**). At this low level, overexpression of exp-PABPN1 caused significant changes in the length of 2 out of the 6 tested 3′-UTRs, whereas no significant increases were observed when transducing with wild-type PABPN1 (**Figure 5D** and Supplementary Figure S3A and B). Following these findings, we asked whether reduction of Pabpn1 expression

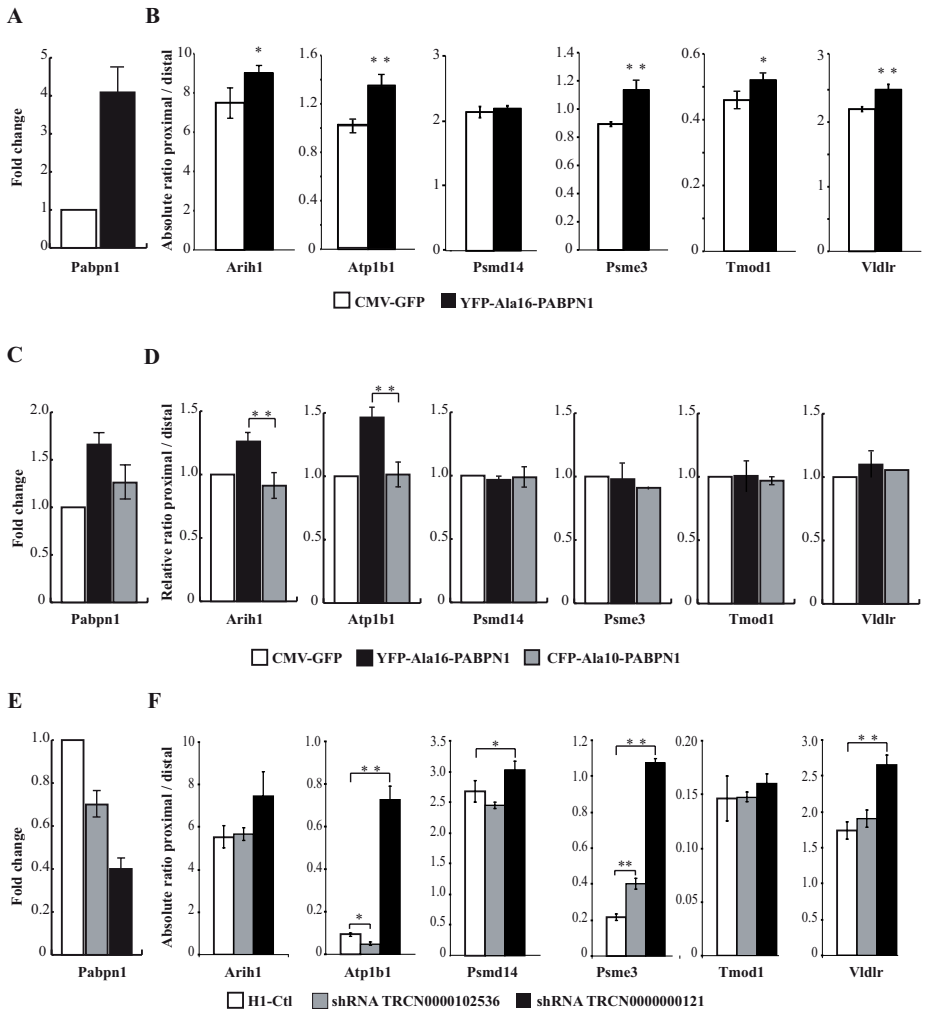


Figure 5. Modulation of PABPN1 expression levels induces changes in polyadenylation site usage in C2C12 cells. (A) Total (endogenous and exogenous) Pabpn1 mRNA levels in C2C12 myotubes transduced with CMV-GFP (white) and YFP-Ala16-PABPN1 (black). **(B)** Ratio of proximal:distal PCR products, representing a combination of short and long 3'-UTRs respectively, in CMV-GFP (white) and YFP-Ala16-PABPN1 (black) transduced myotubes. **(C)** Total Pabpn1 mRNA levels in C2C12 myotubes transduced with CMV-GFP (white), YFP-Ala16-PABPN1 (black) and CFP-Ala10-PABPN1 (grey). **(D)** Relative proximal:distal ratio between YFP-Ala16-PABPN1:CMV-GFP (black) and CFP-Ala10-PABPN1: CMV-GFP (grey), compared to the control CMV-GFP (white). (E) Pabpn1 mRNA levels in C2C12 myoblasts transduced with TRCN0000102536 (grey) and TRCN0000000121 (black) shRNAs targeting Pabpn1 and cells transduced with the control shRNA H1-Ctl (white). **(F)** Proximal:distal ratio for C2C12 myoblasts transduced with the two different sh-RNAs against Pabpn1 (grey, black) and the control shRNA H1-Ctl (white). Values are means + standard deviation for 3 different wells. All experiments were repeated multiple times with similar results.

would also affect polyadenylation site usage. PABPN1 knockdown by siRNAs is known to decrease myoblast differentiation in vitro (1). Thus, we downregulated endogenous Pabpn1 in C2C12 myoblasts by lentiviral transduction with shRNAs. We used two different constructs which downregulated the expression of Pabpn1 to 70% and 40% of control levels (**Figure 5E**). Interestingly, 40% downregulation of Pabpn1 did not result in consistent alterations in polyadenylation site usage, whereas 70%

downregulation resulted in shortening of the 3′-UTRs of 4 out of 6 genes tested (**Figure 5F**). These results suggest that polyadenylation site usage is affected by PABPN1 expression levels. To support the notion that PABPN1 enhances the use of distal polyadenylation sites, we performed RIP experiments to investigate the binding abundance of transcripts with short and long 3′-UTRs to PABPN1. We investigated the same panel of genes previously used in **Figure 5** and calculated the ratio between shorter and longer transcripts, comparing total RNA extracts and PABPN1 immunoprecipitates. Experiments were performed on C2C12 myoblasts. The proximal:distal ratio was significantly lower in PABPN1 immunoprecipitated-RNA compared with input RNA for five out of six genes (**Figure 6A and B**). This suggests that PABPN1 preferentially binds to transcripts with distal polyadenylation sites.

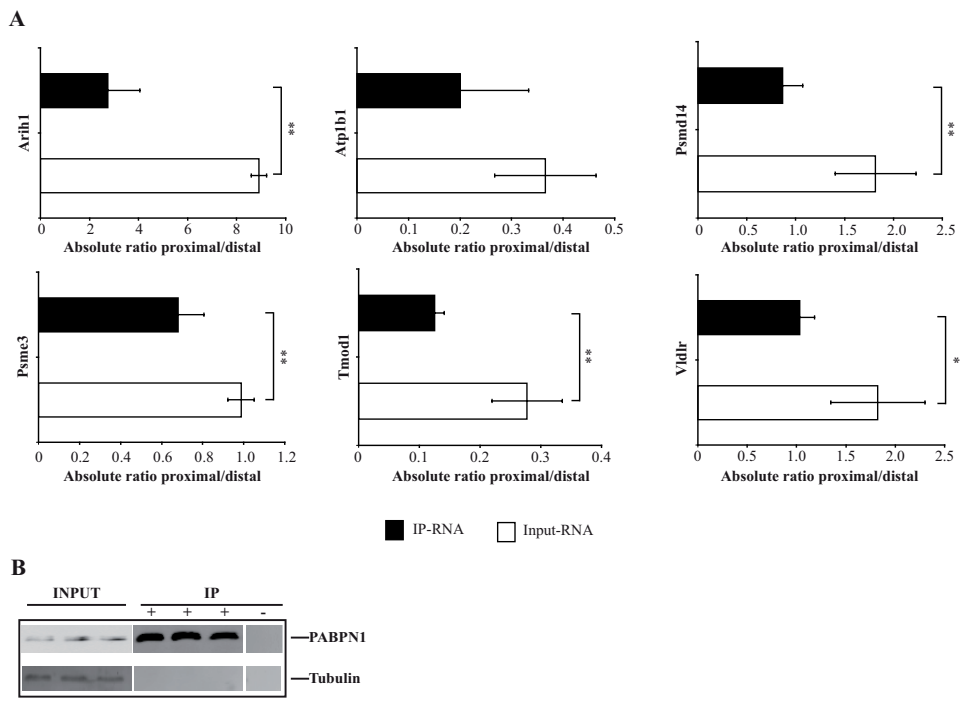


Figure 6. Transcripts with distal polyadenylation site are enriched in PABPN1 immunoprecipitated RNA. (A) The proximal:distal ratio for input RNA (white) and PABPN1-immunoprecipitated RNA (black) was determined by qRT-PCR using the same primer sets, as in Figure 5. Values are means + standard deviation for 3 independent experiments. **(B).** Western blot analysis shows levels of PABPN1 and tubulin, loading control, in the input. Levels of immunoprecipitated (IP) PABPN1 in C2C12 and tubulin, as negative control for immunoprecipitation, are shown. An IP without anti-PABPN1 antibody (-) was used as negative control.

DISCUSSION

Alternative polyadenylation is important to fine-tune gene expression levels, but it is currently unclear how the choice for alternative polyadenylation sites is regulated. The choice of polyadenylation site may depend on the proper orchestration of cleavage and polyadenylation factors (9,22,23,42–44), splicing (45–47) and general transcription factors (22). In this study, we provide evidence that PABPN1 regulates alternative polyadenylation in addition to its role in regulating poly(A) tail length (1–4). To study alternative polyadenylation events on a genome wide level, we developed the polyadenylation

site single molecule sequencing method. Our method is different from the recent SAPAS and PAS-Seq methods (24,38) because it is amplification and ligation free, allowing a more direct and therefore likely less biased, quantitative analysis of polyadenylation site usage. The noise in our data is low and internal priming events are rare as evident from the large majority of reads mapping to the 3′-UTR and having a canonical polyadenylation signal within a distance of 50 nucleotides. Moreover, our method has more power to detect alternative polyadenylation events than general RNA-seq technology (45), which usually comes with low coverage at the 3′-ends. A direct RNA sequencing method would also avoid a possible reverse transcription bias (48), but is not yet available for customers. This study detected many not yet annotated polyadenylation sites and provides an extensive catalogue of alternative polyadenylation events in mouse skeletal muscle. Our results showed more widespread alternative polyadenylation in the mouse transcriptome compared to previous studies (18,24). This indicates that the number of alternative polyadenylation events identified depends very much on the technology used and suggests that alternative polyadenylation frequencies are high in most eukaryotes (49). We investigated alternative polyadenylation events in A17.1 mice overexpressing exp-PABPN1. Our analysis showed a widespread change in relative polyadenylation site usage, with an increased use of proximal polyadenylation sites and thus a shortening of 3′-UTRs. Interestingly, we found that 45% of the transcripts showing changes in polyadenylation site usage were also deregulated. We found an increase in expression level of transcripts with shorter 3′-UTRs, which could be a result of increased stability of transcripts lacking certain miRNA binding sites or other destabilizing elements. Combined with a reduced potential for fine tuning of gene expression due to 3′-UTR shortening, this may contribute to the disturbed gene expression patterns observed in OPMD animal models and patient muscles (7,8). Recently, Jenal et al. (50) published similar findings using the amplification-dependent Illumina sequencing technology. We both show an increase in proximal polyadenylation site usage after overexpression of Exp-PABPN1 or knock down of endogenous Pabpn1. In addition to the results presented by Jenal et al., we directly compared the overexpression of wild-type and expPABPN1 in muscles cells. We demonstrated similar effects of the knockdown of endogenous Pabpn1 and the overexpression of the expanded but not with wild-type PABPN1 in muscle cells. These results are in line with a reduction in the availability of functional PAPBN1 in exp-PABPN1 expressing cells and OPMD muscle as a consequence of the higher aggregation potential of exp-PABPN1 compared with its wild-type counterpart, which will drain the nucleus from soluble PABPN1 (41). Importantly, OPMD muscles show myogenic defects (51) which are similar to knockdown of PABPN1 in mouse cells (1). The exact molecular mechanism by which alteration in PABPN1 expression affects site selection is still not fully elucidated. Jena et al. excluded an effect of PABPN1 on the stability of long versus short transcripts and suggested that PABPN1 binds proximal polyadenylation signals, masking those sites and protecting them from cleavage. However, Jenal et al. only considered binding of PABPN1 to proximal polyadenylation sites. Our RNA-immunoprecipitation experiments provide evidences for a preferential or stronger binding of PABPN1 to transcripts with distal polyadenylation sites. This suggests that PABPN1 enhances the 3′-end processing at the stronger, canonical sites but future studies with targeted mutagenesis should formally prove this. Moreover, Wahle (2) demonstrates a role for PABPN1 in conferring specificity to CPSF for canonical polyadenylation sites. Based on this, the observations in our paper, and the proven physical interaction between PABPN1 and RNA polymerase II (52), it may be postulated that PABPN1 and CPSF comigrate with the RNA polymerase II complex. PABPN1 may then subsequently restrict the site in the RNA where CPSF dissociates from the RNA polymerase II complex to distal, canonical polyadenylation sites. Reduced soluble PABPN1 levels may alter the stoichiometry of the RNA polymerase II/CPSF/PABPN1 complex, leading to premature

dissociation of CPSF at proximal, non-canonical poly(A) sites, resulting in general shortening of the 3′-UTRs of transcripts. An alternative mechanism that might be considered is that lower levels of PABPN1 directly or indirectly reduce the RNA Pol II-mediated transcriptional elongation rate, which has been shown to result in the preferred use of proximal poly(A) sites (53). In any case, these alternative polyadenylation events resulting from reduced PABPN1 levels may affect mRNA stability and partly explain the observed aberrant muscle gene expression patterns and muscle weakness in OPMD animal models and patients (7).

FUNDING

The Netherlands Organisation for Scientific Research [NWO investment grant]; Center for Medical Systems Biology within the framework of the Netherlands Genomics Initiative (NGI)/NWO, the Association Française contre les Myopathies [15,123]; European Commission [TRI-EX QLG2-CT-2001-01673 and POLYALA LSHM-CT-2005-018675]; Muscular Dystrophy Association [68015]. Funding for open access charge: Leiden University Medical Center.

ACKNOWLEDGEMENTS

The authors thank Henk Buermans (Leiden Genome Technology Center, The Netherlands) for the pre-processing of the sequencing data, Martijn Rabelink (Leiden University Medical Center, The Netherlands) for lentivirus production and Yavuz Ariyurek (Leiden Genome Technology Center, The Netherlands) for critical advices on the developed poly(A) site sequencing method. Muscle biopsies from A17.1 mice were kindly provided by Capuchine Trollet and George Dickson and were previously reported.

REFERENCES

1. Apponi LH, Leung SW, Williams KR, Valentini SR, Corbett AH, Pavlath GK. Loss of nuclear poly(A)-binding protein 1 causes defects in myogenesis and mRNA biogenesis. Hum. Mol. Genet. 2010;19:1058-1065.
2. Kuhn U, Gundel M, Knoth A, Kerwitz Y, Rudel S, Wahle E. Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. J. Biol. Chem. 2009;284:22803-22814.
3. Wahle E. Poly(A) tail length control is caused by termination of processive synthesis. J. Biol. Chem. 1995;270:2800-2808.
4. Benoit B, Mitou G, Chartier A, Temme C, Zaessinger S, Wahle E, Busseau I, Simonelig M. An essential cytoplasmic function for the nuclear poly(A) binding protein, PABP2, in poly(A) tail length control and early development in Drosophila. Dev. Cell 2005;9:511-522.
5. Brais B, Bouchard JP, Xie YG, Rochefort DL, Chretien N, Tome FM, Lafreniere RG, Rommens JM, Uyama E, Nohira O, et al. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nat. Genet. 1998;18:164-167.
6. Calado A, Tome FM, Brais B, Rouleau GA, Kuhn U, Wahle E, Carmo-Fonseca M. Nuclear inclusions in oculopharyngeal muscular dystrophy consist of poly(A) binding protein 2 aggregates which sequester poly(A) RNA. Hum. Mol. Genet. 2000;9:2321-2328.
7. Anvar SY, ‘t Hoen PA, Venema A, van der Sluijs B, van EB, Snoeck M, Vissing J, Trollet C, Dickson G, Chartier A,

et al. Deregulation of the ubiquitin-proteasome system is the predominant molecular pathology in OPMD animal models and patients. Skelet. Muscle 2011;1:15.

8. Trollet C, Anvar SY, Venema A, Hargreaves IP, Foster K, Vignaud A, Ferry A, Negroni E, Hourde C, Baraibar MA, et al. Molecular and phenotypic characterization of a mouse model of oculopharyngeal muscular dystrophy reveals severe muscular atrophy restricted to fast glycolytic fibres. Hum. Mol. Genet. 2010;19:2191-2207.
9. Shi Y, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, Yates JR III., Frank J, Manley JL. Molecular architecture of the human pre-mRNA 3′ processing complex. Mol. Cell 2009;33:365-376.
10. Wahle E. A novel poly(A)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation. Cell 1991;66:759-768.
11. Wahle E, Martin G, Schiltz E, Keller W. Isolation and expression of cDNA clones encoding mammalian poly(A) polymerase. EMBO J. 1991;10:4251-4257.
12. Keller W, Bienroth S, Lang KM, Christofori G. Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3′ processing signal AAUAAA. EMBO J. 1991;10:4241-4249.
13. MacDonald CC, Wilusz J, Shenk T. The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. Mol. Cell Biol. 1994;14:6647-6654.
14. Danckwardt S, Hentze MW, Kulozik AE. 3′ end mRNA processing: molecular mechanisms and implications for health and disease. EMBO J. 2008;27:482-498.
15. Legendre M, Gautheret D. Sequence determinants in human polyadenylation site selection. BMC Genomics 2003;4:7.
16. Moreira A, Takagaki Y, Brackenridge S, Wollerton M, Manley JL, Proudfoot NJ. The upstream sequence element of the C2 complement poly(A) signal activates mRNA 3′ end formation by two distinct mechanisms. Genes Dev. 1998;12:2522-2534.
17. Kerwitz Y, Kuhn U, Lilie H, Knoth A, Scheuermann T, Friedrich H, Schwarz E, Wahle E. Stimulation of poly(A) polymerase through a direct interaction with the nuclear poly(A) binding protein allosterically regulated by RNA. EMBO J. 2003;22:3705-3714.
18. Tian B, Hu J, Zhang H, Lutz CS. A large-scale analysis of mRNA polyadenylation of human and mouse genes. Nucleic Acids Res. 2005;33:201-212.
19. Tian B, Pan Z, Lee JY. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. Genome Res. 2007;17:156-165.
20. Barreau C, Paillard L, Osborne HB. AU-rich elements and associated factors: are there unifying principles? Nucleic Acids Res. 2005;33:7138-7150.
21. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. Science 2008;320:1643-1647.
22. Ji Z, Lee JY, Pan Z, Jiang B, Tian B. Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc. Natl Acad. Sci. USA 2009;106:7028-7033.
23. Ji Z, Tian B. Reprogramming of 3′ untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. PLoS One 2009;4:e8419.
24. Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA 2011;17:761-772.
25. Mayr C, Bartel DP. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 2009;138:673-684.
26. Davies JE, Wang L, Garcia-Oroz L, Cook LJ, Vacher C, O'Donovan DG, Rubinsztein DC. Doxycycline attenuates and delays toxicity of the oculopharyngeal muscular dystrophy mutation in transgenic mice. Nat. Med. 2005;11:672-677.
27. Voermans NC, Verrijp K, Eshuis L, Balemans MC, Egging D, Sterrenburg E, van Rooij IA, van der Laak JA, Schalkwijk J, van der Maarel SM, et al. Mild muscular features in tenascin-X knockout mice, a model of Ehlers-Danlos syndrome. Connect. Tissue Res. 2011;52:422-432.
28. Singh R, Hoogaars WM, Barnett P, Grieskamp T, Rana MS, Buermans H, Farin HF, Petry M, Heallen T, Martin JF, et al. Tbx2 and Tbx3 induce atrioventricular myocardial development and endocardial cushion formation. Cell Mol. Life Sci 2012;69:1377-1389.

29. Robinson MD, McCarthy DJ, Smyth GK. EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139-140.

30. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B (Methodol.) 1995;57:289-300.

31. Huang,da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 2009;4:44-57.

32. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics. 2011;27:1653-1659.

33. Raz V, Abraham T, van Zwet EW, Dirks RW, Tanke HJ, van der Maarel SM. Reversible aggregation of PABPN1 pre-inclusion structures. Nucleus 2011;2:208-218.

34. Raz V, Carlotti F, Vermolen BJ, van der Poel E, Sloos WC, Knaan-Shanzer S, de Vries AA, Hoeben RC, Young IT, Tanke HJ, et al. Changes in lamina structure are followed by spatial reorganization of heterochromatic regions in caspase-8-activated human mesenchymal stem cells. J. Cell Sci. 2006;119:4247-4256.

35. Ramakers C, Ruijter JM, Deprez RH, Moorman AF. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci. Lett. 2003;339:62-66.

36. Ruijter JM, Ramakers C, Hoogaars WM, Karlen Y, Bakker O, van den Hoff MJ, Moorman AF. Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. Nucleic Acids Res. 2009;37:e45.

37. Verheesen P, de Kluijver A, van Koningsbruggen S, de Brij M, de Haard HJ, van Ommen GJ, van der Maarel SM, Verrips CT. Prevention of oculopharyngeal muscular dystrophy-associated aggregation of nuclear polyA-binding protein with a single-domain intracellular antibody. Hum. Mol. Genet. 2006;15:105-111.

38. Fu Y, Sun Y, Li Y, Li J, Rao X, Chen C, Xu A. Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome Res. 2011;21:741-747.

39. Pauws E, van Kampen AH, van de Graaf SA, de Vijlder JJ, Ris-Stalpers C. Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis. Nucleic Acids Res. 2001;29:1690-1694.

40. Slomovic S, Laufer D, Geiger D, Schuster G. Polyadenylation and degradation of human mitochondrial RNA: the prokaryotic past leaves its mark. Mol. Cell. Biol. 2005;25:6427-6435.

41. Raz V, Routledge S, Venema A, Buijze H, van der Wal E, Anvar S, Straasheijm KR, Klooster R, Antoniou M, van der Maarel SM. Modeling oculopharyngeal muscular dystrophy in myotube cultures reveals reduced accumulation of soluble mutant PABPN1 protein. Am. J. Pathol. 2011;179:1988-2000.

42. Chuvpilo S, Zimmer M, Kerstan A, Glockner J, Avots A, Escher C, Fischer C, Inashkina I, Jankevics E, Berberich-Siebelt F, et al. Alternative polyadenylation events contribute to the induction of NF-ATc in effector T cells. Immunity 1999;10:261-269.

43. Shell SA, Hesse C, Morris SM Jr., Milcarek C. Elevated levels of the 64-kDa cleavage stimulatory factor (CstF-64) in lipopolysaccharide-stimulated macrophages influence gene expression and induce alternative poly(A) site selection. J. Biol. Chem. 2005;280:39950-39961.

44. Takagaki Y, Manley JL. Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation. Mol. Cell 1998;2:761-771.

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

46. Danckwardt S, Kaufmann I, Gentzel M, Foerstner KU, Gantzert AS, Gehring NH, Neu-Yilik G, Bork P, Keller W, Wilm M, et al. Splicing factors stimulate polyadenylation via USEs at non-canonical 3′ end formation signals. EMBO J. 2007;26:2658-2669.

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

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

49. Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V, et al. The landscape of C. elegans 3′UTRs. Science 2010;329:432-435.

50. Jenal M, Elkon R, Loayza-Puch F, van HG, Kuhn U, Menzies FM, Vrielink JA, Bos AJ, Drost J, Rooijers K, et al. The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. Cell

2012;149:538-553.

51. Perie S, Mamchaoui K, Mouly V, Blot S, Bouazza B, Thornell LE, St. Guily JL, Butler-Browne G. Premature proliferative arrest of cricopharyngeal myoblasts in oculo-pharyngeal muscular dystrophy: Therapeutic perspectives of autologous myoblast transplantation. Neuromuscul. Disord. 2006;16:770-781.

52. Bear DG, Fomproix N, Soop T, Bjorkroth B, Masich S, Daneholt B. Nuclear poly(A)-binding protein PABPN1 is associated with RNA polymerase II during transcription and accompanies the released transcript to the nuclear pore. Exp. Cell Res. 2003;286:332-344.

53. Pinto PA, Henriques T, Freitas MO, Martins T, Domingues RG, Wyrzykowska PS, Coelho PA, Carmo AM, Sunkel CE, Proudfoot NJ, et al. RNA polymerase II kinetics in polo polyadenylation signal selection. EMBO J. 2011;30:2431-2444.

SUPPORTING INFORMATION

Supplementary Tables 1-6 are available at NAR Online

Table S1. List of primer sequences.

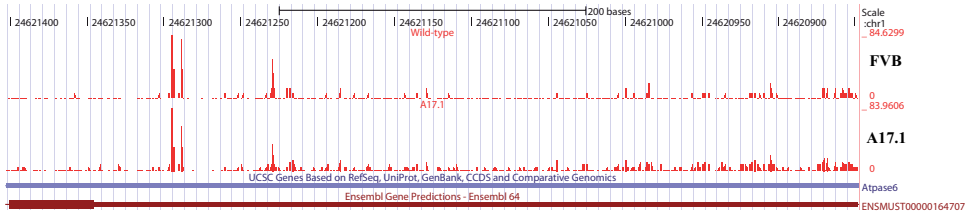
Table S2. Summary of polyadenylation site sequencing results.

Table S3. List of transcripts with change in polyadenylation site usage in A17.1 mice compared to FVB mice ((FDR) < 0.05).

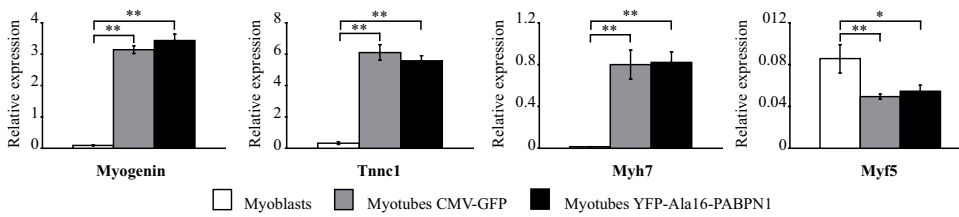
Table S4. List of differentially expressed transcripts in A17.1 mice compared to FVB mice ((FDR) < 0.05).

Table S5. KEGG pathway analysis for differentially expressed transcripts in A17.1 mice compared to FVB mice.

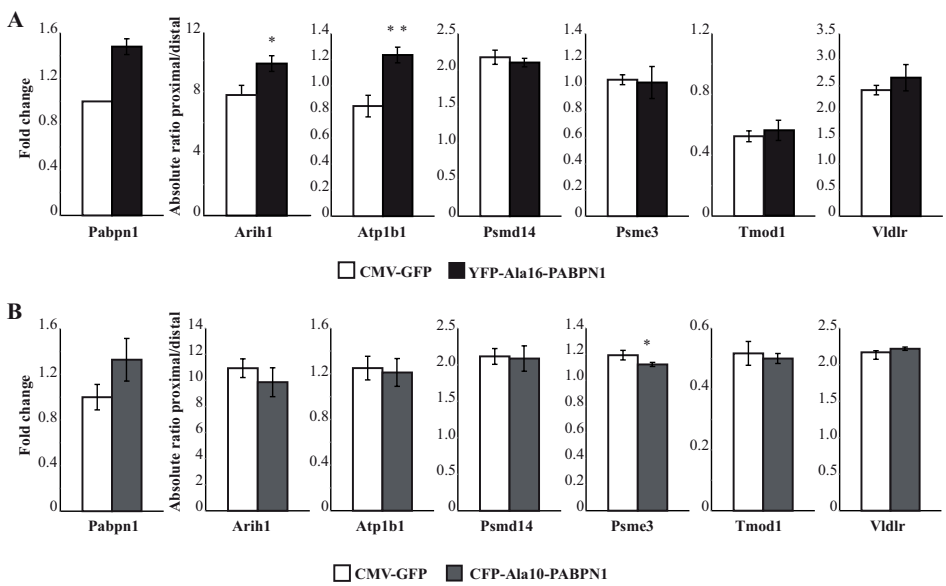
Table S6. List of differentially expressed transcripts in A17.1 mice compared to FVB mice ((FDR) < 0.05), containing at least two polyadenylation sites.



Supplementary Figure 1. Polyadenylation sites in mitochondrial transcripts.



Supplementary Figure 2. Effect of exp-PABPN1 overexpression on myotubes. Expression level of myogenic markers in myotubes transduced with YPF-Ala16-PABPN1 (black) compared to CMV-GFP control (grey) and myoblasts (white).



Supplementary Figure 3. Effects of low overexpression of wild-type and expanded PABPN1 on polyadenylation sites usage in C2C12 cells. Differently from Figure 5D, ratio of proximal and distal PCR products are calculated as absolute.

CHAPTER 3

DEEPSAGE REVEALS GENETIC VARIANTS ASSOCIATED WITH ALTERNATIVE POLYADENYLATION AND EXPRESSION OF CODING AND NON-CODING TRANSCRIPTS

Daria V. Zhernakova, Eleonora de Klerk, Harm-Jan Westra, Anastasios Mastrokolias, Shoaib Amini, Yavuz Ariyurek, Rick Jansen, Brenda W. Penninx, Jouke J. Hottenga, Gonneke Willemsen, Eco J. de Geus, Dorret I. Boomsma, Jan H. Veldink, Leonard H. van den Berg, Cisca Wijmenga, Johan T. den Dunnen, Gert-Jan B. van Ommen, Peter A.C. 't Hoen, Lude Franke.

PLoS Genet. 2013 Jun; 9(6): e1003594.
doi: 10.1371/journal.pgen.1003594.

ABSTRACT

Many disease-associated variants affect gene expression levels (expression quantitative trait loci, eQTLs) and expression profiling using next generation sequencing (NGS) technology is a powerful way to detect these eQTLs.

We analyzed 94 total blood samples from healthy volunteers with DeepSAGE to gain specific insight into how genetic variants affect the expression of genes and lengths of 3'-untranslated regions (3'-UTRs). We detected previously unknown cis-eQTL effects for GWAS hits in disease- and physiology-associated traits. Apart from cis-eQTLs that are typically easily identifiable using microarrays or RNA-sequencing, DeepSAGE also revealed many cis-eQTLs for antisense and other non-coding transcripts, often in genomic regions containing retrotransposon-derived elements. We also identified and confirmed SNPs that affect the usage of alternative polyadenylation sites, thereby potentially influencing the stability of messenger RNAs (mRNA). We then combined the power of RNA-sequencing with DeepSAGE by performing a meta-analysis of three datasets, leading to the identification of many more cis-eQTLs.

Our results indicate that DeepSAGE data is useful for eQTL mapping of known and unknown transcripts, and for identifying SNPs that affect alternative polyadenylation. Because of the inherent differences between DeepSAGE and RNA-sequencing, our complementary, integrative approach leads to greater insight into the molecular consequences of many disease-associated variants.

INTRODUCTION

Genome-wide association studies (GWAS) have associated genetic variants, such as single nucleotide polymorphisms (SNPs) and copy number variants (CNVs), with numerous diseases and complex traits. However, the mechanisms through which genetic variants affect disease phenotypes or physical traits often remain unclear. To gain insight into these mechanisms, we have combined genotype data with gene expression data by conducting expression quantitative trait locus (eQTL) mapping. Previously, the level of gene expression was primarily assessed using oligonucleotide microarrays, which was a powerful method to profile the transcriptome [1–6]. But recently, high-throughput next generation sequencing (NGS) has become available, which allows quantification of expression levels by counting mRNA fragments (RNA-seq) or sequence tags (including serial analysis of gene expression (SAGE), cap analysis of gene expression (CAGE), and massively parallel signature sequencing (MPSS)) [7].

To date, two NGS eQTL studies have been published [8,9], both of which used RNA-seq. Although RNA-seq is a versatile technique, the coverage in the ultimate 3'-end is usually lower due to the fragmentation and random hexamer priming steps involved in the sample preparation [10] (**Figure 1B**). DeepSAGE technology [11,12], however, concentrates on capturing information on the 3' end of transcripts. In DeepSAGE, enzymatic cDNA digestions generate one specific tag of 17 nucleotides at the most 3'-CATG sequence of each transcript (**Figure 1A**). The majority of these 21-mer tags ('CATG' + 17 nucleotides) can be uniquely mapped to the genome to identify the genes expressed.

There are several features of NGS-based expression quantification methods that are especially important for eQTL mapping. While oligonucleotide arrays target a predefined set of transcripts or exons, both RNA-seq and DeepSAGE are capable of detecting novel and unannotated transcripts. If such a novel gene later turns out to be cis-regulated by trait- or disease-associated SNPs, it can represent an interesting causal candidate gene for the trait or disease under investigation. RNA-seq is extremely versatile, as it can quantify the expression of alternative transcripts, which makes it possible to detect SNPs regulating the choice between alternative transcripts. DeepSAGE, however, is generally not suited to detecting alternative splicing because of the 3' bias of the tag locations [13]. Because only sequence data is generated for these short tags, the read depth per tag is generally much greater than with RNA-seq, permitting accurate quantification of these tags [11,14]. Thus, this 3' emphasis makes DeepSAGE suitable for transcript variants that differ in 3'-UTRs and also for detecting alternative polyadenylation events, a widespread phenomenon that generates variation in 3'-UTR length [15,16]. Shortening or lengthening of the 3'-UTR may result in the loss or gain of regulatory elements, such as miRNA binding sites or binding sites for proteins that can stabilize or destabilize the transcript [17,18]. Several SNPs that influence the choice for alternative polyadenylation sites have been detected by RNA-seq on a small number of individuals [19]. Here, we analyzed this phenomenon in more depth by performing cis-eQTL mapping on DeepSAGE data from total blood samples of 94 individuals.

RESULTS

DeepSAGE dataset

For cis-eQTL mapping, we used DeepSAGE sequencing of 21 bp tags (16 ± 7 million tags) from total blood samples from 94 healthy, unrelated individuals from the Netherlands Twin Register (NTR) and the Netherlands Study of Depression and Anxiety (NESDA) [20]. Sequence reads were mapped to the reference genome hg19 using Bowtie [21] and assigned to transcripts. We mapped 85 ± 5% of tags to the genome and found that 77 ± 9% of these mapped to exonic regions. Although 66 ± 18% of these reads mapped to hemoglobin-alpha or –beta (HBA1, HBA2, HBB) genes, we were left with sufficient sequencing depth to detect a total of 9,562 genes at a threshold of at least two tags per million.

Cis-eQTL mapping

Once reads had been mapped, we quantified the expression levels of sequenced tags and performed cis-eQTL mapping, evaluating only those combinations of SNPs and tags that were located within a genomic distance of 250 kb, while using a Spearman rank correlation test (tag-level false discovery rate (FDR) controlled at 0.05). We identified 540 unique cis-regulated tags. To subsequently increase the statistical power of eQTL detection, we used principal component analysis (PCA) to correct for technical and known and unknown biological confounders. The first principal components (PC) generally capture a high percentage of the expression variation, and these PCs mostly reflect technical, physiological and environmental variability. Removing this variation allows for the detection of more eQTLs [6,22,23]. In our data the first principal component significantly correlated with sample GC content, and principal components 7 and 11 correlated with various blood cell count parameters (for details see Text S1, Figures S1 and S2). When using the PC corrected data, we observed an almost two-fold increase in the number of significant cis-eQTLs (1,011 unique cis-regulated tags, corresponding to 896 unique cis-regulated genes at tag-level FDR < 0.05). The list of detected eQTLs is given in Table S1.

Comparison with microarray results

We then compared the DeepSAGE cis-eQTLs with cis-eQTLs that we had identified using the Affymetrix HG-U219 expression microarrays on the same 94 samples. In that analysis we detected cis-eQTLs for only 274 genes (FDR < 0.05), only a third of what we identified using DeepSAGE. We observed that this substantial difference could mostly be explained by the fact that the cis-eQTLs detected using Affymetrix microarrays nearly always reflected genes that are highly expressed in blood, whereas for DeepSAGE the detected cis-eQTL genes had expression levels that could be much lower (**Figure 2**). Although we only concentrated on tags that were expressed, there was no clear relationship between the mean tag level expression and the probability of showing a significant cis-eQTL. As such, DeepSAGE is much more capable of identifying cis-eQTLs for genes showing low expression than conventional microarrays. It was therefore not a surprise that only 39% of the identified DeepSAGE cis-eQTLs could also be significantly detected in the microarray-based dataset (with identical allelic direction) (Figure S3). Indeed, the cis-eQTLs that were not replicated in the microarray-based dataset generally had a much lower expression than the replicating cis-eQTLs (Wilcoxon Mann Whitney $P < 2 \times 10\text{-}3$). And vice versa, we could significantly replicate 75% of the detected Affymetrix cis-eQTLs with the same allelic direction in the DeepSAGE data (Figure S3), indicating that DeepSAGE shows overlapping results with array-based data. At the same time, this provides insight into the regulation of gene expression by SNPs at many more loci. We estimated the reduction that could be made in the sample size of

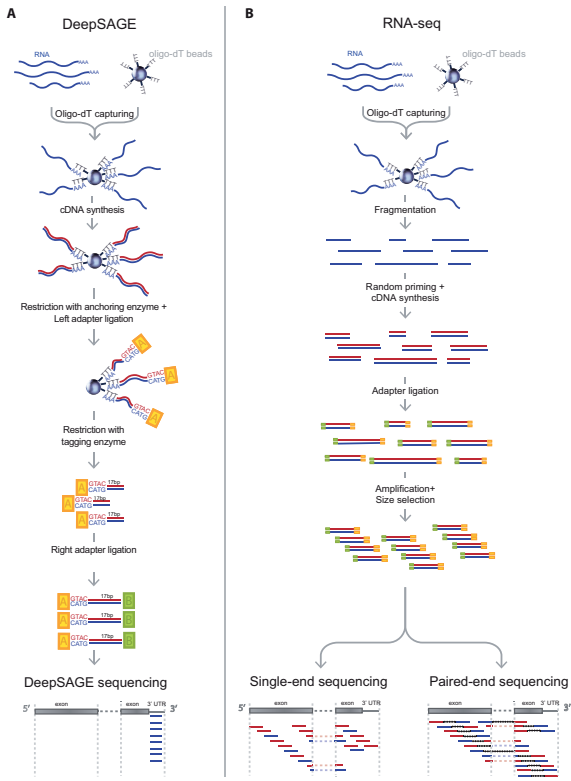


Figure 1. Comparison of typical DeepSAGE and RNA-seq data generation steps. A) DeepSAGE data preparation consists of the following basic steps: after RNA extraction the polyadenylated mRNA fraction is captured with oligo-dT beads. While RNA is still bound to the beads, double-stranded cDNA synthesis is performed. Next, cDNA is digested by NlaIII restriction enzyme (an anchoring enzyme), which cuts the DNA at CATG recognition sequences, leaving only the fragment with the most distal (3') CATG site associated with the beads. Subsequently, a GEX adapter is attached to the 5' end. This adapter contains a recognition sequence for the MmeI restriction enzyme that cuts the sequence 17 bp downstream of CATG site. After ligation of a second GEX adapter, fragments containing 21 bp tags (17 unknown nucleotides + CATG) are ready for sequencing. **B)** A typical protocol for RNA-seq data preparation has the following steps: after RNA extraction the polyadenylated mRNA fraction is captured with oligo-dT beads. Captured RNA is fragmented and for each fragment cDNA synthesis is performed using random hexamer primers. Sequencing adapters are then ligated to each fragment. This is followed by size selection of the DNA fragments and PCR amplification. Then one end of the fragment is sequenced (single-end sequencing) or both ends (paired-end sequencing).

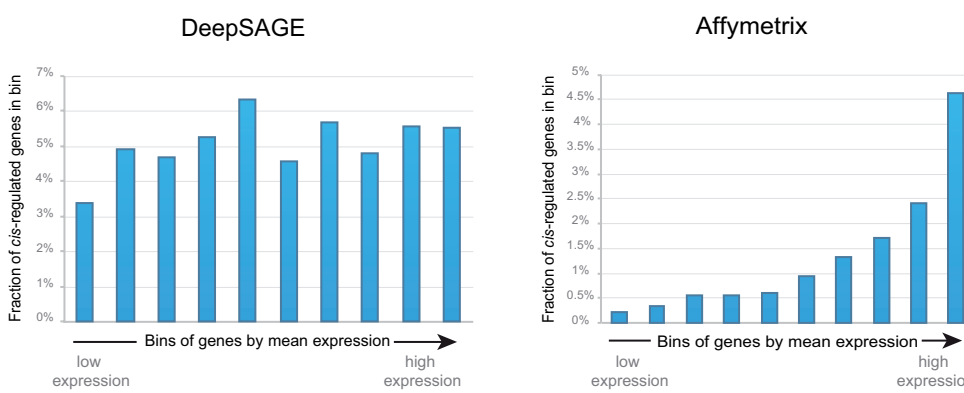


Figure 2. Fraction of cis-regulated genes in bins by mean gene expression levels for DeepSAGE and Affymetrix data. For each dataset, all genes were sorted by their mean gene expression levels, and divided into ten equal bins. The X-axis reflects these bins, which are sorted by increasing mean gene expression levels. The Y-axis reflects the fraction of cis-regulated genes that fall into each bin.

the sequencing-based dataset to get the same number of cis-regulated genes as in microarray-based data. We observed that the DeepSAGE sample size could be reduced by almost half (to 55 samples) to get the same number of significant cis-regulated genes as identified in the microarray analysis of the

94 samples. As such, these results clearly indicate that DeepSAGE has higher statistical power than microarrays.

Cis-eQTL effects on non-coding genes

While most microarray platforms interrogate mainly the protein-coding part of the transcriptome, NGS-based expression profiling will detect the majority of all expressed transcripts. Indeed, we detected eQTLs for known, but non-protein coding, genes: 8 antisense genes and 31 lincRNAs (**Figure 3**). We also expected to find a number of cis-eQTL effects on previously unknown transcripts. Of the 1,011 tags with a significant cis-eQTL effect, 230 did not map to known transcripts. Many of these tags map to retrotransposon-derived elements in the genome, which are known to be a source of novel exons [24]: 73 DeepSAGE tags with significant cis-eQTLs that did not map to annotated genes mapped to 72 unique LINE, SINE and LTR elements in the genome (**Table 1**).

New regulatory roles for disease- and trait-associated SNPs

We checked how many of our cis-acting SNPs were associated with complex traits or complex diseases ('trait-associated SNPs'), as published in the Catalog of Published Genome-Wide Association Studies. 104 of the 6,446 unique trait-associated SNPs were significant cis-eQTLs in our data (Table S2). We were interested to determine whether the DeepSAGE data had revealed cis-eQTL effects for trait-associated SNPs that had been missed when using conventional arrays on much larger cohorts. We therefore compared our results to a re-analysis of a large-scale, array-based cis-eQTL mapping that we had conducted in whole peripheral blood samples when using a much larger sample size of 1,469 (using Illumina oligonucleotide arrays [6]). We identified 13 trait-associated SNPs that did show a significant cis-eQTL effect in DeepSAGE eQTL mapping, but which did not show a cis-eQTL effect in the large, array-based, blood dataset (**Table 2**). This indicates that many trait-associated SNPs have regulatory effects that will, so far, likely have been missed using microarrays. While some of the tags map in the exons of annotated transcripts, we also found three cis-regulated tags in introns (sense direction), two tags antisense to the known transcripts, and two tags outside the annotated transcripts. These results indicate that several trait-associated SNPs affect the expression of previously unknown transcripts, adding functional relevance to SNPs and transcripts that are so far without annotation. Some newly discovered eQTLs provide novel insights into genome-wide association hits for diseases or physiological traits, e.g. SNP rs216345, which has been associated with bipolar disorder. While it is located just downstream of PRSS3, we now saw that it also affects the expression of UBE2R2. There are many links between the ubiquitin system and bipolar disorder reported in the literature (e.g. [25,26]), making UBE2R2 a more plausible candidate gene for bipolar disorder than PRSS3.

Genes with multiple SAGE tags and opposite allelic direction

In DeepSAGE, 21-bp-long cDNA fragments begin at the 'CATG' closest to the polyadenylation site (**Figure 1**). These individual 'tags' represent transcripts sharing the same polyadenylation site. If a SNP increases the abundance of one tag of a gene and decreases the abundance of another tag of the same gene, this indicates that the SNP is acting like a switch between transcripts with different 3'-UTRs or between alternative polyadenylation sites [19] (**Figure 4)**. Twelve genes with highly significant cis-eQTLs (p-value < 10-7) contained tags that were regulated in opposite directions (**Table 3**). Most of the tags regulated in opposite direction could be explained by switches in alternative

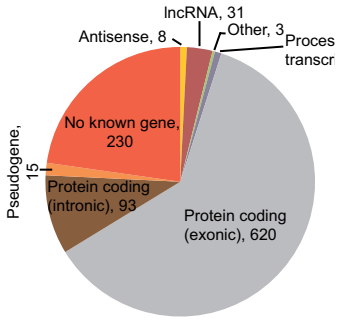


Figure 3. Mapping regions of cis-regulated tags. The gene biotypes and exon/intron locations of unique cis-regulated tags, according to Ensembl v.69 annotation, are shown. The numbers indicate the number of tags mapping in the genes of the corresponding type.

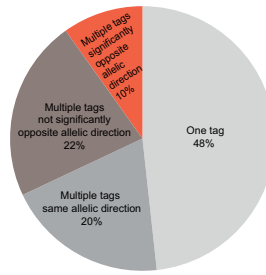


Figure 4. The number of cis-regulated tags per gene. The percentages of cis-regulated tags mapping into the same gene are indicated (781 genes overall). For nearly half of the genes (48%) only one tag shows an eQTL effect. If multiple tags map within the same gene, only one eQTL tag should pass the FDR < 0.05 significance threshold while the other tag could be less significant. For these eQTLs the allelic direction is shown: same allelic direction (multiple tags within the same gene are cis-regulated by a SNP in the same direction), significantly opposite allelic direction (multiple tags within the same gene are cis-regulated by a SNP but with opposite directions and the difference between the correlation coefficients is significant), or opposite allelic direction but not significant (if the difference between correlation coefficients is not significant).

Type of genomic region	Number of *cis*-regulated tags
LINE	32
SINE	14
LTR	17

Table 1. Number of cis-regulated tags mapping to different genomic regions in tag-wise DeepSAGE eQTL mapping.

polyadenylation sites, as the tags were observed in the same last exon. The effect on alternative polyadenylation in IRF5 has been found before [19,27] and was also validated in our cohort by RT-qPCR with primers in the proximal and distal parts of the 3'-UTR (**Figure 5**). As a further confirmation of the observed switches in using polyadenylation sites, we tested genotype-dependent alternative polyadenylation in two other RNA-seq datasets [8,9]. In these datasets, we confirmed the effect of two cis-regulating SNPs on THEM4 and F11R. However, we could not confirm the effect of other SNPs on targets validated experimentally, including IRF5. This shows the limitation of RNA-seq data in detecting alternative polyadenylation events, likely due to limited and unequal coverage of the 3'-UTR. For only two genes, OAS1 (also reported earlier [28]) and RP11-493L12.2, the observed opposite allelic effect originated from transcripts with different last exons, likely due to alternative splicing. As we hasve identified several SNPs that affect alternative polyadenylation, we subsequently used a more permissive strategy, which required that, for a given SNP, only one eQTL tag should pass the FDR < 0.05 significance threshold while the other tag could be less significant. However, for such SNP-tag pairs, we then tested whether the allelic directions were opposite and if the difference between correlation coefficients was significant. With a differential correlation significance p-value threshold of 10-7, we detected 41 unique genes showing regulation in opposite directions (Table S3). Of these, 23 (56%) showed opposite regulation of two tags in the same annotated 3'-UTR and a further 7 genes (17%) showed opposite regulation of tags in the same exons, both indicative of a switch in polyadenylation sites. Of these we picked HPS1, and validated a genotype-determined switch in preferred polyadenylation site usage by RT-qPCR analysis (**Figure 5**), indicating that the more permissive list also holds genuine changes in polyadenylation sites. The remaining 11 genes showed significant genotype-determined switches in expression of alternative transcripts not sharing the final exon. Thus, switches between shorter and longer 3'-UTRs occur more frequently than

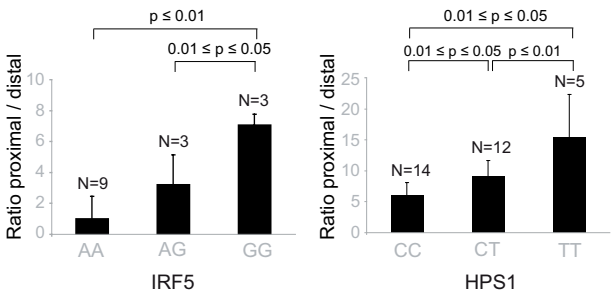


Figure 5. The choice of proximal/distal polyadenylation site in genes IRF5 and HPS1 depends on the genotypes of rs10488630 and rs11189600, respectively. The ratio between the abundance of transcripts with proximal and distal 3'-UTR RT-qPCR products in IRF5 (left) and HPS1 (right) depends on the genotypes of cis-regulating SNPs rs10488630 and rs11189600, respectively. N denotes the number of individuals included in the analysis. These results indicate allele-specific preference for use of a proximal and distal polyadenlyation site.

switches between transcripts with different 3'-UTRs. To check whether such results appeared by chance, we took an equal number of top hits from a permuted eQTL run (shuffling the phenotype labels of the expression data, thus breaking the relationship between genotype and expression, but retaining linkage disequilibrium (LD) structure and structure in the expression data) and performed the same analysis as above (assessing an equal number of top eQTLs from the permuted analysis as we had investigated in the real analysis). Using the differential correlation significance threshold of 10-7 and conducting this permutation analysis ten times, we did not find any SNP that affected two tags in the same gene in a significantly different way, indicating this method is robust. Since the eQTL SNPs are usually in strong LD with multiple SNPs, it is difficult to conclude whether a SNP is causal or which SNP is the likely causal variant. To identify the likely causal variant, we assessed whether any of these SNPs caused changes in polyadenylation site usage. A direct effect on alternative polyadenylation can be explained by a change in the polyadenylation site (corresponding to the cleavage site) or in the polyadenylation signal (a six-nucleotide motif located between 10–30 bases upstream of the cleavage site). We searched for likely causative SNPs in linkage disequilibrium with the polyA-QTL SNP ($R2 \geq 0.8$). We did not find any strong evidence for SNPs influencing the cleavage site and focused on cis-regulating SNPs located within polyadenylation signals. Considering the length and the motif of canonical and non-canonical polyadenylation signals [15], we performed a motif analysis in the sequence surrounding each cis-regulating SNP. We identified five SNPs that likely affect polyadenylation because there was a change in the polyadenylation signal (**Table 4**). As previously shown, rs10954213 causes the formation of a stronger polyadenylation signal in IRF5. Similar changes from non-canonical to stronger, canonical polyadenylation signals were observed for rs1062827 in F11R and rs6598 in GIMAP5. Moreover, rs12934747 creates a new canonical AATAAA polyadenylation signal in LPCAT2. The presence of this alternative polyadenylation signal at the beginning of the 3'-UTR leads to a decrease in transcripts containing the full length 3'-UTR, as observed by DeepSAGE (**Figure 6**). An opposite effect is observed for rs7063 in the ultimate 3'-end of the ERAP1 gene, where the SNP causes the disruption of the strong canonical motif, and results in the use of a more proximal polyadenylation signal. Unfortunately we were not able to identify likely causative SNPs for each of these eQTLs. This could have several reasons: we imposed strict thresholds ($R2 \geq 0.8$) on the LD between the detected cis-eQTLs and the putative causative SNPs; by imputing to the 1000 genomes dataset we may have missed causative SNPs unique to the Dutch population; and the list of experimentally validated polyadenylation sites is not exhaustive, because their detection depends on the expression level and cell type analyzed. Seven of the SNPs affecting polyadenylation are reported in the GWAS catalog as associated with diseases (Table S3), including rs2188962 and rs12521868,

which are associated with Crohn's disease. We found that these SNPs were associated with a switch in the polyadenlyation site of IRF1. This may reinforce previous evidence that IRF1 is the gene in the IBD5 locus responsible for its association with Crohn's disease [29]. IRF1 is a family member of the IRF5 gene. Thus, in the family of interferon regulatory factors, we found two members with genetic regulation of alternative polyadenylation sites, likely explaining susceptibility for Crohn's disease and systemic lupus erythematosus, respectively. Another example is rs3194051, located in the IL7R gene. This SNP was not found in the analysis described above since it affects the expression of a tag on the same strand, downstream of IL7R in a LINE element (**Table 2**). However, this tag may represent an alternative 3'-UTR for IL7R. The SNP is associated with ulcerative colitis and IL7R may be another example of a gene in the inflammatory response pathway demonstrating alternative polyadenylation.

Meta-analysis with RNA-seq data

To increase the statistical power to detect associations of SNPs with gene expression, we performed a first-of-its-kind eQTL mapping meta-analysis, combining DeepSAGE data with two published RNA-seq datasets. We used paired-end sequencing of mRNA derived from lymphoblastoid cell lines from HapMap individuals of European origin [8] and 35 and 46 bp single-end sequencing of mRNA derived from lymphoblastoid cell lines from HapMap individuals of Yoruba origin [9]. Sequence reads were mapped to the reference genome hg19 using Tophat [30] and assigned to transcripts. A consistently high percentage of reads (86-87% of aligned reads) mapped to exonic regions (**Table 5**). We first performed eQTL mapping separately in all three datasets (**Table 6**), summarizing expression on the transcript level to permit comparisons between the datasets. The numbers of cis-regulated genes detected in transcript-wise analysis was lower than in tag-wise analysis, possibly because we missed resolution on alternative splicing- and alternative polyadenylation events. Again, PC correction greatly improved the number of cis-eQTLs detected (**Table 6**). We applied PC correction to the individual datasets. As for the DeepSAGE analysis, the first PC correlated strongly with the mean GC-percentage in the two RNA-seq datasets (Figure S1). We then assessed the robustness of the identified cis-eQTLs: we checked whether those in one dataset could be significantly replicated in the other two datasets. We observed that in each of the RNA-seq datasets approximately one-third of cis-eQTLs could be replicated in the other dataset (Table S4). The overlap between RNA-seq and DeepSAGE was smaller, reflecting differences in the two technologies, in cell types and in populations. In each comparison, we observed a very high concordance in the allelic direction of cis-eQTLs that could be replicated in another dataset. We also looked at the replication of RNA-seq eQTLs in corresponding micro-array-based datasets. 80-88% of such eQTLs could be replicated in microarray data (Table S5). As we could cross-replicate many cis-eQTLs, we decided to conduct a meta-analysis to increase the statistical power. We calculated joint p-values using a weighted Z-score method. The number of cis-regulated genes then increased to 1,207 unique genes (**Table 6**) (a list of detected eQTLs is given in Table S6), indicating that a meta-analysis of different types of sequencing-based eQTL datasets reveals many more cis-regulated genes than the individual analyses. For our meta-analysis results we determined the number of disease- and trait-associated SNPs using the Catalog of Published Genome-Wide Association Studies in the same way as for the DeepSAGE dataset. 107 of the 6,446 unique trait-associated SNPs showed a significant cis-eQTL effect in the meta-analysis. The overlap with 104 trait-associated SNPs detected in tag-wise DeepSAGE eQTL mapping was 37, indicating that the DeepSAGE revealed other trait-associated cis-eQTLs than a meta-analysis on the level of whole transcripts. 21 of the 107 SNPs showed a significant cis-eQTL effect in the sequencing-based meta-analysis, but did not show a cis-eQTL effect in the large array-based blood dataset (**Table 7**).

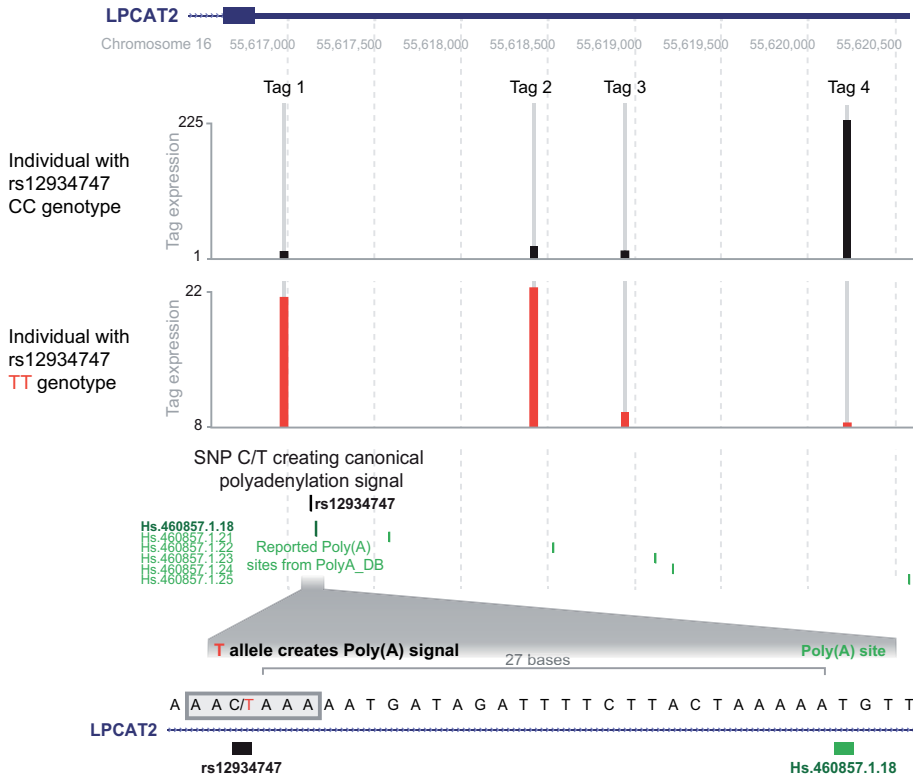


Figure 6. rs12934747*T creates a poly(A) signal in LPCAT2 and leads to alternative polyadenylation site usage. The y-axis represents the number of counts for the deepSAGE tags. Two samples with different genotypes for SNP rs12934747, CC (reference allele) and TT (alternative allele), are shown as different traces. Below the coverage tracks, the position of rs12934747 is shown, together with the position of all reported polyadenylation sites from polyA_DB. An enlargement of the region containing the SNP is shown below. rs12934747 is located at the beginning of the 3'-untranslated region (3'-UTR) of LPCAT2, 27 nucleotides upstream a reported and experimentally validated polyadenylation site. This SNP changes the sequence, creating a polyadenylation signal that leads to the usage of the reported polyadenylation site. The square block indicates the sequence of the polyadenylation signal created by rs12934747. The creation of this signal shortens the 3'-UTR, as indicated by the higher abundance of the proximal DeepSAGE just upstream of the polyadenylation signal, and the nearly absent distal DeepSAGE, in the sample with the TT genotype (both tags indicated by arrows). Tag 2 was filtered out because it was expressed in less than 90% of individuals. There is an additional tag 3 in-between the proximal and distal tags, which is not cis-regulated.

DISCUSSION

We have described the results from cis-eQTL mapping on DeepSAGE sequencing, a technique that is different from RNA-seq since it mainly targets the 3'-end of transcripts. We identified 1,011 unique cis-regulated tags (significant at tag-level FDR < 0.05). We performed eQTL mapping on the microarray expression data of the same samples and the number of detected cis-eQTLs was much smaller than in the DeepSAGE data, indicating the higher power of DeepSAGE in eQTL mapping. Moreover, for 220 of the cis-eQTLs SNPs detected by DeepSAGE we did not detect a significant cis-eQTL in a much larger microarray-based study in 1,469 whole peripheral blood samples [6]. 13 of these SNPs were reported as disease- or trait-associated in the GWAS catalog. We observed that the number of cis-

eQTLs detected in microarray data was higher in highly expressed genes, whereas for DeepSAGE the detected cis-eQTL genes had expression levels that could be much lower (**Figure 2**). This means that DeepSAGE is much better at identifying cis-eQTLs for genes showing low expression than conventional microarrays. This is because gene expression quantification using microarrays is more difficult as there is always a background signal present that needs to be accounted for. This is not the case for next-generation sequencing: although stochastic variation plays a major role in determining what RNA molecules will eventually be sequenced (especially for transcripts of low abundance), detection of such an RNA molecule is direct proof that it is being expressed. Clearly, DeepSAGE can capture events that are likely to be missed by RNA-seq and conventional microarrays. It is not surprising, due to the different emphasis of DeepSAGE, that we could only replicate 39% of the DeepSAGE cis-eQTLs in the microarray data with a consistent allelic direction (Figure S3). The limited overlap between DeepSAGE- and microarray-based eQTL studies may be partly explained by the fixed thresholds applied, the interrogation of different transcript variants, and by the smaller dynamic range of microarrays. In addition, we found that more highly expressed genes were more often replicated than lower expressed ones. Moreover, DeepSAGE allows for the detection of non-coding and novel transcripts not present on the microarrays. We showed that genetic variation affects the expression of a substantial number of lincRNAs and antisense genes, some of which have been linked to clinical traits. This suggests that clinical traits may be modified by expression of antisense transcripts or alternative 3'-UTR selection, which are not separately quantified in the microarray-based studies or in most RNA-seq, where standard protocols are still not strand-specific. We also noticed a relatively high proportion of eQTLs with DeepSAGE tags mapping in SINE, LINE and LTR elements. These transposable elements contribute to the evolution and inter-individual variation of the human genome and to the diversification of the transcriptome, the latter facilitated by their inherent potential to be transcribed and the presence of cryptic splice acceptor and donor sites [24,31,32]. Some of the DeepSAGE tags we identified may be located in entirely new transcripts, but the majority is likely to represent alternative exons or 3'-UTRs of known transcripts, in accordance with the observed preferential location in introns or near genes. Associations with transcripts and transcript variants not yet annotated may help to discover a function for these transcripts, as they are likely to play a role in the physiological and clinical traits associated with the SNP. Moreover, this will complement our knowledge of the pathways associated with these physiological and clinical traits.

In our study, we have demonstrated that genotype-dependent switches in the preference of alternative polyadenylation sites are common. One of these events has been well characterized: SNP rs10954213 creates an alternative polyadenylation site in IRF5, shortens the 3'-UTR, stabilizes the mRNA, and increases IRF5 expression, explaining its genetic association with systemic lupus erythematosus [19,27]. We have now discovered more examples where SNPs create or disrupt polyadenylation motifs. Amongst others, we identified a new, similar, genotype-dependent switch in preferred polyadenylation site for family member IRF1, with a probable link to Crohn's disease. Alternative polyadenylation associated with shortening of 3'-UTRs is a prominent event in the activation of immune cells [18]. Thus, genetically determined use of a proximal polyadenylation sites may predispose to inflammatory disorders such as Crohn's disease. The opposite correlations that we observed for most genes were slightly less pronounced than for IRF5. This indicates that mechanisms other than the creation or disruption of canonical polyadenylation motifs may also play a role. For example, SNPs in miRNA or protein-binding sites may specifically affect the stability of the transcript variant with the long 3'-UTR. We subsequently conducted a cis-eQTL meta-analysis on the heterogeneous types of data using methods extended from those we developed for microarray-

SNP name	Tag chr.	Tag position (midpoint)	In gene	Location in gene	Sense/antisense	Closest gene	Repeat masker annotation	Associated trait
rs6704644	2	234380527	*DGKD*	3′-UTR	sense	-	None	Bilirubin levels
rs9875589	3	14196086	*XPC*	intron	antisense	-	LINE L1MB3	Ovarian reserve
rs4580814	5	1050754	*SLC12A7*	3′-UTR	sense	-	None	Hematological and biochemical traits
rs3194051	5	35884591	*None*	-	-	*IL7R*	LINE L2C	Ulcerative colitis
rs4917014	7	50472441	*IKZF1*	3′-UTR	sense	-	None	Systemic lupus erythematosus
rs10092658	8	131017411	*FAM49B*	intron	sense	-	None	Protein quantitative trait loci
rs216345	9	33917317	*UBE2R2*	3′-UTR	sense	-	None	Bipolar disorder
rs12219125	10	20519590	*PLXDC2*	intron	sense	-	SINE AluJb	Diabetic retinopathy
rs7181230	15	40325714	*EIF2AK4*	intron	antisense	-	None	Dehydroepiandrosterone sulphate levels
rs4924410	15	40328035	*SRP14*	3′-UTR	sense	-	None	Ewing sarcoma
rs12594515	15	45995320	*None*	-	-	lincRNA RP11-718O11.1	LTR MLT1A	Waist circumference, weight
rs6504218	17	62400467	*PECAM1*	3′-UTR	antisense	-	None	Coronary heart disease
	17	62397000	*PECAM1*	3′-UTR	antisense	-	LINE L1ME4A	Coronary heart disease
rs4072910	19	8640274	*MYO1F*	intron	sense	-	LINE MER1B	Height

Table 2. Trait-associated SNPs affecting DeepSAGE tags of 94 peripheral blood samples, but not detected in an array-based eQTL dataset of 1,469 peripheral blood samples.

SNP Name	SNP Type	Allele Assessed	Probe Chr.	Probe Center	Overall Z-Score	HGNC Name	Annotation
rs12568757	G/A	G	1	150782318	4.404	*ARNT*	Alternative polyadenylation
			1	150782604	−4.314		
rs12566232	A/C	C	1	151846229	−6.859	*THEM4*	Alternative polyadenylation
			1	151846628	4.292		
rs1062826	G/C	C	1	160965239	−4.46	*F11R*	Alternative polyadenylation
			1	160966976	8.012		
rs13160562	G/A	A	5	96110323	−7.883	*ERAP1*	Alternative polyadenylation
			5	96111908	5.555		
rs3185733	A/C	A	5	112320282	4.027	*DCP2*	Alternative polyadenylation
			5	112356357	−4.46		
rs6948928	T/C	T	7	128589824	8.451	*IRF5*	Alternative polyadenylation
			7	128589265	−7.31		
rs2111903	G/C	C	12	47603121	5.143	*RP11-493L12.2*	Different exon
			12	47599911	−4.676		
rs841718	A/G	G	12	57489368	5.506	*STAT6*	Alternative polyadenylation
			12	57489809	−7.002		
rs2285934	G/T	T	12	113357275	4.931	*OAS1*	Different exon
			12	113355465	−5.191		
rs168822	C/T	T	16	55616984	7.37	*LPCAT2*	Alternative polyadenylation
			16	55620233	−4.664		
rs922446	T/C	T	16	56395733	−4.966	*AMFR*	Alternative polyadenylation
			16	56396100	4.392		
rs1674159	C/T	T	19	5915589	−7.103	*CAPS*	Alternative polyadenylation
			19	5916143	6.126		

*Only significant eQTLs with FDR<0.05 for both *cis*-regulated tags were used.

Table 3. Cis-regulating SNPs significantly* affecting multiple tags of the same gene in opposite directions.

Cis-regulating SNP	Causal SNP	R^2	SNP type	Gene	Reference sequence	Alternative polyA signal	Distance to polyA site (bp)	Effect on 3′-UTR length
Formation/activation of polyA signal								
rs6948928	rs10954213	0.76*	G/A	*IRF5*	AATGAA	AATAAA	15	Shortening
rs168822	rs12934747	0.87	C/T	*LPCAT2*	AACAAA	AATAAA	27	Shortening
rs1062826	rs1062827	0.99	G/A	*F11R*	AGTAAA	AATAAA	21	Shortening
rs759011	rs6598	1	G/A	*GIMAP5*	AATAGA	AATAAA	13	Shortening
Disruption of polyA signal								
rs13160562	rs7063	1	T/A	*ERAP1*	AATAAA	AAAAAA	23	Shortening

*This SNP was reported in [27] and is validated by our data.

Table 4. SNPs that likely affect polyadenylation due to a change in the polyadenylation signal.

based eQTL meta-analysis [6]. We identified 1,207 unique cis-regulated genes. This number is substantially higher than in each of the datasets separately and indicates that different protocols for digital gene expression generally deliver consistent results. Nevertheless, the overlap at a fixed FDR of 0.05 is rather small, in particular between DeepSAGE and RNA-seq data. While this is partly attributable to using a strong threshold, there are other important reasons: firstly, the RNA-seq and DeepSAGE technologies frequently interrogate different transcript variants. Secondly, the RNA-seq studies were done on lymphoblastoid cell lines (LCLs) while the DeepSAGE study was on total blood, and some cis-eQTLs may be tissue-specific [33,34]. Finally, the DeepSAGE technology is strand-specific but the RNA-seq technologies evaluated here are not: where DeepSAGE will evaluate the expression of sense and antisense transcripts separately, RNA-seq will sum them. These reasons could

Dataset	Sequencing type	Cell tissue type	Number of samples	Read length	Million reads per sample	Average % of mapped reads	Average % of mapped reads mapping to exons
Montgomery *et al.*	Paired-end RNA-seq	LCL	60	37 bp	9.5±3	56	87
Pickrell *et al.* Yale	Single-end RNA-Seq	LCL	72	35 bp	8.1±2.3	85	86
Pickrell *et al.* Argonne	Single-end RNA-Seq	LCL	72	46 bp	8.1±1.8	80	86
NTR-NESDA	DeepSAGE	Total blood	94	21 bp	16±7	85	88

Table 5. Description of RNA next generation sequencing datasets.

	Number of unique genes with cis-eQTLs	
	Without principal component correction	**With principal component correction**
Montgomery *et al.* (paired-*end RNA-seq*)	94	145
Pickrell *et al.* (*single-end RNA-seq*)	199	438
NTR-NESDA transcript-wise *(DeepSAGE)*	292	579
Meta-analysis	651	1,207

Table 6. Number of detected cis-eQTLs in transcript-wise analysis of three harmonized RNA NGS datasets.

SNP name	Chr.	Transcript position (midpoint)	*Cis*-regulated gene	Associated trait
rs1052501	3	41963564	*ULK4*	Multiple myeloma
rs347685	3	141782879	*TFDP2*	Chronic kidney disease
rs4580814	5	1081324	*SLC12A7*	Hematological and biochemical traits
rs4947339	6	28911984	*C6orf100*	Platelet aggregation
rs2517532	6	31024818	*HCG22*	Hypothyroidism
rs2844665	6	31024818	*HCG22*	Stevens-Johnson syndrome and toxic epidermal necrolysis (SJS-TEN)
rs6457327	6	31024818	*HCG22*	Follicular lymphoma
rs3130501	6	31324124	*HLA-B*	Stevens-Johnson syndrome and toxic epidermal necrolysis (SJS-TEN)
rs2858870	6	32434437	*HLA-DRB9*	Nodular sclerosis Hodgkin lymphoma
rs3129889	6	32434437	*HLA-DRB9*	Multiple sclerosis
rs3135388	6	32434437	*HLA-DRB9*	Multiple sclerosis
rs477515	6	32434437	*HLA-DRB9*	Inflammatory bowel disease
rs9271100	6	32524134	*HLA-DRB6*	Systemic lupus erythematosus
rs9273349	6	32632106	*HLA-DQB1*	Asthma
rs3807989	7	116183034	*CAV1*	PR interval
rs12680655	8	135604552	*ZFAT*	Height
rs4929923	11	8642408	*TRIM66*	Menarche (age at onset)
rs12785878	11	71161461	*RP11-660L16.2*	Vitamin D insufficiency
rs12580100	12	56436876	*RPS26*	Psoriasis
rs4924410	15	40329664	*SRP14*	Ewing sarcoma
rs7364180	22	42184613	*MEI1*	Alzheimer's disease biomarkers

Table 7. Trait-associated SNPs detected in the sequencing-based transcript-wise meta-analysis, but not detected in array-based eQTL dataset of 1,469 peripheral blood samples.

explain why the percentage of RNA-seq-derived eQTLs that can be replicated by DeepSAGE is higher than the percentage of DeepSAGE-derived eQTLs that can be replicated by RNA-seq. We conclude that DeepSAGE technology is useful to determine cis-eQTLs, as it is able to quantify the expression of novel transcripts, and to detect alternative polyadenylation effects and alternative 3'-UTR selection. It is complementary to other sequencing-based approaches, as they each reveal slightly different regulatory effects of genetic variants. Different sequencing-based eQTL analyses generally deliver consistent results, allowing for meta-analyses across different technologies. Future eQTL mapping studies based on DeepSAGE and other next generation sequencing strategies, using larger cohorts and different techniques, will likely reveal a more comprehensive picture of how far genetic variation affects the expression of protein-coding genes and non-coding RNAs.

MATERIALS AND METHODS

Ethics statement

The medical ethical committee of the VUMC, Amsterdam, the Netherlands, approved the collection and analysis of material blood, DNA and RNA from the 94 participants from the Netherlands Twin Registry (NTR) and the Netherlands Study of Depression and Anxiety (NESDA).

NTR-NESDA dataset

We analyzed 21 bp DeepSAGE data from total blood RNA of 94 unrelated individuals who participated in NTR or NESDA. RNA was isolated using PaxGene tubes [20,35,36]. DeepSAGE sample preparation protocols, and alignment approaches were described in [37]. One sample was run on one lane of the Illumina GAII instrument. Data are available in ArrayExpress under accession number E-MTAB-1181. The NTR-NESDA data was imputed using Beagle v3.1.0, with HapMap2 release 24 as a reference.

Tag mapping and expression estimation

Tags from DeepSAGE sequencing were aligned to the NCBI build 37 reference genome using Bowtie v. 0.12.7 allowing for a maximum of 1 mismatch and a maximum of 2 possible alignments (-n 1 -k 1 -m 2 --best --strata options). The expression values were both quantified on an individual tag and transcript level. For the tag-wise analysis, the total number of occurrences of each unique tag in each sample was counted. We only included tags that were present in >90% of samples. Tags with SNPs in the CATG recognition sequence (according to dbSNPv135) and the transcripts containing those tags were removed before eQTL analysis, since these SNPs can affect the position of the SAGE tag in the transcript. For the transcript-wise analysis, the tag counts for tags overlapping the exons of a transcript by at least half of the tag length were summed. Coordinates of LINE, SINE, LTR elements were derived from UCSC's RepeatMasker track (update: 2009-04-24).

GC content bias estimation

To calculate the GC content per individual for DeepSAGE data, GC frequencies for all mapped tags were summed after excluding the twenty most abundant tags, since their high abundance would give biased estimates.

Cis-eQTL mapping and correction for confounding effects through principal component analysis

Before eQTL mapping, transcript and tag expression values were quantile normalized. To perform cis-eQTL mapping, association of SNPs with the expression levels of tags or transcripts within a distance of 250 kb (as this is the average size of linkage regions) of the midpoint of the transcript or tag were tested with a non-parametric Spearman's rank correlation. Multiple testing correction was performed, controlling the false discovery rate (FDR) at 0.05. To determine the FDR, we created a null distribution by repeating the cis-eQTL analysis after permuting the sample labels 10 times [38]. We argue that gene expression levels from NGS-based datasets are, like micro-array based data, derived from genetic, technical and environmental effects. As such, compensating for these non-genetic effects would increase the power to detect cis-eQTL effects. To mitigate the effects of non-genetic sources of variability, we first log2 transformed the data and centered and scaled each tag, and subsequently applied PCA on the sample correlation matrix. We then used the first PCs as covariates, and re-did the non-parametric cis-eQTL mapping on the residual expression data (using the procedure described by [6]).

Validation of genotype-dependent alternative polyadenylation in RNA-seq datasets

The genomic coordinates of the 3'-UTR, obtained from Refseq Genes, were split into two separate regions (distal and proximal 3'-UTRs) according to the position of the DeepSAGE tags with opposite

directions, the position of LongSAGE tags from CGAP, and the position of reported and predicted polyadenylation sites from polyA_DB database. To calculate the coverage in proximal and distal regions in RNA-seq datasets, we created a coverage histogram from each .bam alignment file using coverageBed tool from BEDTools package (version 2.17.0) [39]. Subsequently, a custom Python script was used to convert the histogram in number of nucleotides mapped per region, normalized by the length of the region. The ratio between the number of counts in the proximal region and the distal region was then calculated.

qPCR validation of alternative polyadenylation

Expression of short and long variants of HPS1 and IRF5 was quantified by qRT-PCR, which was performed on a subset of RNA samples used for the DeepSAGE sequencing. cDNA was synthesized from 400 ng of total RNA using BioScript MMLV Reverse Transcriptase (Bioline) with 40 ng of random hexamer and oligo(dT)18 primers following manufacturer's instructions (for the list of primer sequences see Table S7). Primers specific to short or long variants of HPS1 were designed using Primer3Plus program, primers for IRF5 were designed as previously described [40]. qRT-PCR was performed on the LightCycler 480 (Roche) using 2X SensiMix reagent (Bioline). 45 cycles of two-step PCR were performed for HPS1, and 55 cycles of three-step PCR (95°C for 15 s, 48°C for 15 s, and 60°C for 40 s) for IRF5. Each measurement was performed in duplicates. PCR efficiency was determined using the LinRegPCR program [41] v.11.1 according to the described method [42]. Ratios between distal and proximal PCR products were then calculated and significance was tested performing a T-test.

Identifying causal SNPs affecting polyadenylation

We obtained all the proxy SNPs for all SNPs identified as cis-regulating the choice of polyadenylation site. To do this we used bi-allelic SNPs that pass QC from the 1000G European panel (v3.20101123) and took all SNPs that were in linkage disequilibrium with the query SNPs ($R2 \geq 0.8$, distance between SNPs within 250 kb). From this list of cis-regulating SNPs in linkage disequilibrium, we kept only SNPs, which were located in the cis-regulated genes. The filtering was performed by intersecting .bed files containing SNPs coordinates and coordinates of cis-regulated genes from RefSeq database, using table browser tool in UCSC genome browser and the overlap intervals tool in Galaxy (version 1.0.0). Intersection of SNPs with validated and predicted polyadenylation sites was performed using annotation in the PolyA-DB database (PolyA_DB 1 and PolyA_SVM) on UCSC (table browser tool). Detection of SNPs within polyadenylation signals was performed by extracting the strand specific sequence five nucleotide upstream and downstream each SNP (using table browser tool in UCSC) and performing a motif search using custom Perl script. Canonical and non-canonical polyA motifs searched were AATAAA, ATTAAA, TATAAA, AGTAAA, AAGAAA, AATATA, AATACA, CATAAA, GATAA, AATGAA, TTTAAA, ACTAAA, and AATAGA. For every SNP located in a putative polyadenylation signal motif, the distance to validated and predicted polyadenylation sites from PolyA-DB was calculated. Only motifs within a distance of 30 nucleotides from a polyadenylation site were considered true polyadenylation signals. Newly formed polyadenylation signals were detected by changing the reference allele of the SNP with the alternative allele, followed by the same polyadenylation signal motif search using custom Perl scripts. For the cis-regulated genes where the SNP is located within a true polyadenylation signal, we retrieved the coverage of every SAGE tag upstream and downstream the putative affected polyadenylation site and calculated the ratio between proximal and distal tags for the different genotypes to confirm the expected effects of polyadenylation site formation or

disruption.

RNA-seq datasets

For the meta-analysis we combined DeepSAGE data with two published RNA-seq datasets. The first dataset was 37bp paired-end RNA-sequencing data from HapMap individuals ([8], [ArrayExpress:E-MTAB-197]): RNA from lymphoblastoid cell lines of 60 HapMap CEPH individuals was sequenced on the Illumina GAII sequencer, while genotype data had already been generated within the HapMap project. The second dataset was single-end RNA-sequencing data from HapMap individuals [9, 43] [GEO:GSE19480 and at http://eqtl.uchicago.edu/RNA_Seq_data/]: RNA was sequenced from lymphoblastoid cell lines of 72 HapMap Yoruba individuals from Nigeria on the Illumina GAII platform in two sequencing centers: Yale (using 35bp reads) and Argonne (using 46bp reads). Since the Montgomery et al. paper used genotype data for some individuals that were not in the HapMap3 panel (NA0851, NA12004, NA12414 and NA12717), we imputed these individuals using Beagle v3.1.0, with HapMap2 release 24 as a reference.

RNA-seq read mapping

Reads from single- and paired-end RNA-sequencing were mapped to the human genome NCBI build 37 (reference annotation from Ensembl GRCh37.65) using Tophat v. 1.3.3 [30] – a splice-aware aligner that maps RNA-seq reads to the reference genome using Bowtie [21]. We used default settings (maximum 2 mismatches, 20 possible alignments per read) with a segment length value of 17bp. Reads that corresponded to the flag 1796 in the .bam alignment file (read unmapped, not primary alignment, read fail quality check, read is PCR or optical duplicate) were filtered out. The numbers of raw and mapped reads for each dataset are given in **Table 5**.

Read quantification

To estimate expression levels in RNA-seq data, reads that overlapped with exons from known transcripts (GRCh37.65) were quantified using the coverageBed method from BEDTools suite [39]. For transcript level quantification the read count C_s^{tr} for sample s for transcript tr was calculated as a sum of expression values over all exons contained in this transcript: $C_s^{tr} = 10^6 \cdot \sum_{e \in \{E_{tr}\}} n_e \cdot B_e$ where:

$\{E_{tr}\}$ set of all exons of transcript tr,
n_e number of reads overlapping exon e by not less than half of read's length,
B_e breadth of coverage for exon e (% of exon length covered by the reads mapping to that exon).

In case a read mapped to multiple transcripts, the read was counted for all transcripts, since the short reads are difficult to assign to a specific transcript. Multiplication by breadth of coverage was performed to help in distinguishing between different isoforms by assigning higher weight to exons fully covered by reads in contrast to alternative exons covered only partly. Because different methods have different capacity to identify alternative splicing events, we subsequently summarized our eQTL results to unique genes.

Meta-analysis

Meta-analysis was conducted by using a weighted Z-method, weighing each of the datasets by the square root of the number of samples per dataset [6].

Microarray datasets

We compared the results to corresponding microarray dataset eQTL mapping results. For each of the 94 individuals from NTR-NESDA study, Affymetrix HG-U219 expression data were generated at the Rutgers University Cell and DNA Repository (RUCDR, http://www.rucdr.org). NTR and NESDA samples were randomly assigned to plates with seven plates containing subjects from both studies to better inform array QC and study comparability. Gene expression data were required to pass standard Affymetrix QC metrics (Affymetrix expression console) before further analysis. Probe sets were removed when their mapping location was ambiguous or if their location intersected a polymorphic SNP (dropped if the probe oligonucleotide sequence did not map uniquely to hg19 or if the probe contained a polymorphic SNP based on HapMap3 [44] and 1000 Genomes [45] project data). Expression values were obtained using RMA normalization implemented in Affymetrix Power Tools (APT, v 1.12.0). MixupMapper revealed no sample mix-ups [46]. For RNA-seq data we used corresponding microarray datasets that were available for most of the individuals present in RNA-seq datasets. We used Illumina expression data provided by Stranger et al. [3] of the 72 HapMap YRI individuals (56 of which were also present in RNA-seq dataset from Pickrell et al.) and 60 HapMap CEU individuals provided by Montgomery et al. (58 of which were also present in RNA-seq dataset from Montgomery et al.). The same normalization procedure was performed as for the sequencing-based datasets: quantile normalization, and subsequent probe set centering to zero, z-score transformation, and scaling to a standard deviation of one.

Data access

The newly generated DeepSAGE data for NTR-NESDA dataset is available in ArrayExpress under accession number E-MTAB-1181 (ENA: ERP001544).

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: PACtH LF GJBvO JTdD. Performed the experiments: PACtH YA AM. Analyzed the data: DVZ EdK PACtH HJW SA. Contributed reagents/materials/analysis tools: PACtH HJW RJ BWP JJH EJdG BIB JHV LHvdB CW. Wrote the paper: DVZ EdK PACtH LF.

REFERENCES

1. Schadt EE, Monks SA, Drake TA, Lusisk AJ, Chek N, et al. (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422: 297–302.
2. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, et al. (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437: 1365–1369.
3. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848–853.

4. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. (2008) Genetics of gene expression and its effect on disease. Nature 452: 423–428.

5. Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. Nature Genetics 42: 295–302.

6. Fehrmann RSN, Jansen RC, Veldink JH, Westra H-J, Arends D, et al. (2011) Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet 7: e1002197.

7. Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 12: 87–98.

8. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773–777.

9. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768–772.

10. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63.

11. 't Hoen PAC, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RHAM, et al. (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.

12. Nielsen KL, Høgh AL, Emmersen J (2006) DeepSAGE-digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. Nucleic Acids Res 34: e133.

13. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, et al. (2002) Using the transcriptome to annotate the genome. Nat Biotechnol 20: 508–512.

14. Asmann YW, Klee EW, Thompson EA, Perez EA, Middha S, et al. (2009) 3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. BMC Genomics 10: 531.

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

16. Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, et al. (2012) A quantitative atlas of polyadenylation in five mammals. Genome Res 22: 1173–1183.

17. Barreau C, Paillard L, Osborne HB (2005) AU-rich elements and associated factors: are there unifying principles? Nucleic Acids Res 33: 7138–7150.

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

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

20. Maugeri N, Powell JE, 't Hoen PAC, de Geus EJC, Willemsen G, et al. (2011) LPAR1 and ITGA4 regulate peripheral blood monocyte counts. Hum Mutat 32: 873–876.

21. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3): R25.

22. Biswas S, Storey JD, Akey JM (2008) Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. BMC Bioinformatics 9: 244.

23. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3: 1724–1735.

24. Cordaux R, Batzer MA (2009) The impact of retrotransposons on human genome evolution. Nat Rev Genet 10: 691–703.

25. Ryan MM, Lockstone HE, Huffaker SJ, Wayland MT, Webster MJ, et al. (2006) Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. Mol Psychiatry 11: 965–978.

26. 26. Bousman CA, Chana G, Glatt SJ, Chandler SD, Lucero GR, et al. (2010) Preliminary evidence of ubiquitin proteasome system dysregulation in schizophrenia and bipolar disorder: convergent pathway analysis findings from two independent samples. Am J Med Genetics Part B, Neuropsychiatric genetics 153B: 494–

502.

27. Cunninghame Graham DS, Manku H, Wagner S, Reid J, Timms K, et al. (2007) Association of IRF5 in UK SLE families identifies a variant involved in polyadenylation. Hum Mol Gen 16: 579–591.
28. Heap GA, Trynka G, Jansen RC, Bruinenberg M, Swertz MA, et al. (2009) Complex nature of SNP genotype effects on gene expression in primary human leucocytes. BMC Med Genomics 2: 1.
29. Huff CD, Witherspoon DJ, Zhang Y, Gatenbee C, Denson LA, et al. (2012) Crohn's disease and genetic hitchhiking at IBD5. Mol Biol Evol 29: 101–111.
30. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105–1111.
31. Belancio VP, Hedges DJ, Deininger P (2006) LINE-1 RNA splicing and influences on mammalian gene expression. Nucleic Acids Res 34: 1512–1521.
32. Kim D-S, Kim T-H, Huh J-W, Kim I-C, Kim S-W, et al. (2006) LINE FUSION GENES: a database of LINE expression in human genes. BMC Genomics 7: 139.
33. Fu J, Wolfs MGM, Deelen P, Westra H-J, Fehrmann RSN, et al. (2012) Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS Genet 8: e1002431.
34. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, et al. (2011) The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet 7: e1002003.
35. Willemsen G, De Geus EJC, Bartels M, Van Beijsterveldt CEMT, Brooks AI, et al. (2010) The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. Twin Res Hum Genet 13: 231–245.
36. Penninx BWJH, Beekman ATF, Smit JH, Zitman FG, Nolen WA, et al. (2008) The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. Int J Methods Psychiatr 17: 121–140. doi:10.1002/mpr.
37. Hestand MS, Klingenhoff A, Scherf M, Ariyurek Y, Ramos Y, et al. (2010) Tissue-specific transcript annotation and expression profiling with complementary next-generation sequencing technologies. Nucleic Acids Res 38: e165.
38. Breitling R, Li Y, Tesson BM, Fu J, Wu C, et al. (2008) Genetical genomics: spotlight on QTL hotspots. PLoS Genet 4: e1000232.
39. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.
40. Graham RR, Kyogoku C, Sigurdsson S, Vlasova IA, Davies LRL, et al. (2007) Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. Proc Nat Acad Sci U S A 104: 6758–6763.
41. Ramakers C, Ruijter JM, Deprez RHL, Moorman AF. (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci Lett 339: 62–66.
42. Ruijter JM, Ramakers C, Hoogaars WMH, Karlen Y, Bakker O, et al. (2009) Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. Nucl. Acids Res 37: e45.
43. Pickrell JK, Pai AA, Gilad Y, Pritchard JK (2010) Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genet 6: e1001236.
44. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58.
45. Durbin RM, Bentley DR, Chakravarti A, Clark AG, Collins FS, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
46. Westra H-J, Jansen RC, Fehrmann RSN, Te Meerman GJ, Van Heel D, et al. (2011) MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinformatics 27: 2104–2111.

SUPPORTING INFORMATION

Supplementary Tables 1-7 are available at PLoS Genetics Online

Table S1. List of detected eQTLs in tag-wise eQTL mapping.
Table S2. Trait-associated SNPs affecting the expression of DeepSAGE tags of 94 peripheral blood samples.
Table S3. List of candidate genes with alternative polyadenylation event detected using a permissive strategy.
Table S4. Replications between RNA-seq and DeepSAGE eQTLs.
Table S5. Replication of RNA-seq eQTLs in microarray-based datasets.
Table S6. List of detected eQTLs in the meta-analysis.
Table S7. Primer sequences for qPCR validation.
Text S1. Additional details on principal component analysis of DeepSAGE expression data.

Supplementary text

Additional details on principal component analysis of DeepSAGE expression data
To increase the statistical power of eQTL detection, we used principal component analysis (PCA) to correct for technical and biological confounders. We determined that using 15 PCs as covariates yielded the highest number of significant cis-eQTLs, reflecting an almost two-fold increase.

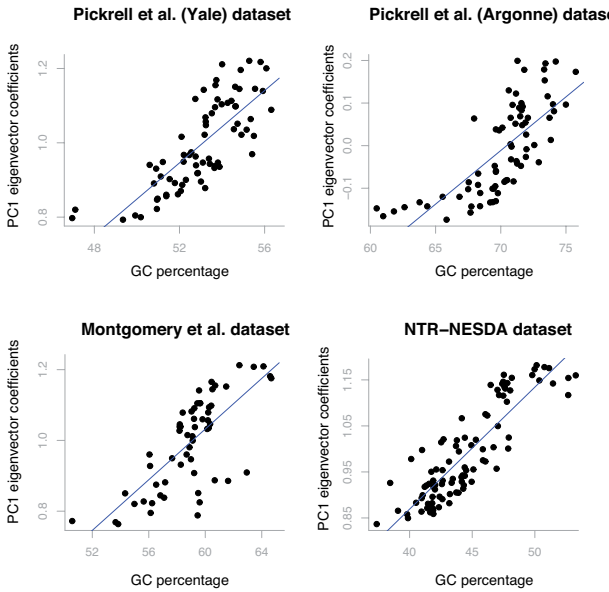
Although correction for the first principal components substantially increased the number of detectable cis-eQTLs, it remains somewhat elusive why this correction procedure is so effective. We therefore investigated which phenomena these components represent and investigated the correlation with various sample characteristics. The first principal component was highly significantly correlated with the percentage of GC in the reads of a sample (r2 = 0.76) (Figure S1). GC content is one of the most important sources of bias in RNA-seq data and strongly affects gene expression measurements [1,2]. Although various dedicated strategies have been proposed to overcome this bias (for a review see [3]) and more sophisticated algorithms to correct for technical and biological confounders exist such as PEER and PANAMA [4–6], this straightforward PCA-based method also efficiently corrects for GC content differences across samples.

Principal components seven and eleven correlated significantly with various blood cell count parameters, indicating that these PCs reflect differences in cell type compositions between samples (Figure S2). To further substantiate this latter point, we associated the top 100 genes that had the most extreme (highest and lowest) factor loadings on PC7 and PC11 with cell types reported in the literature, using the Anni software for text concept association [7] and observed that:

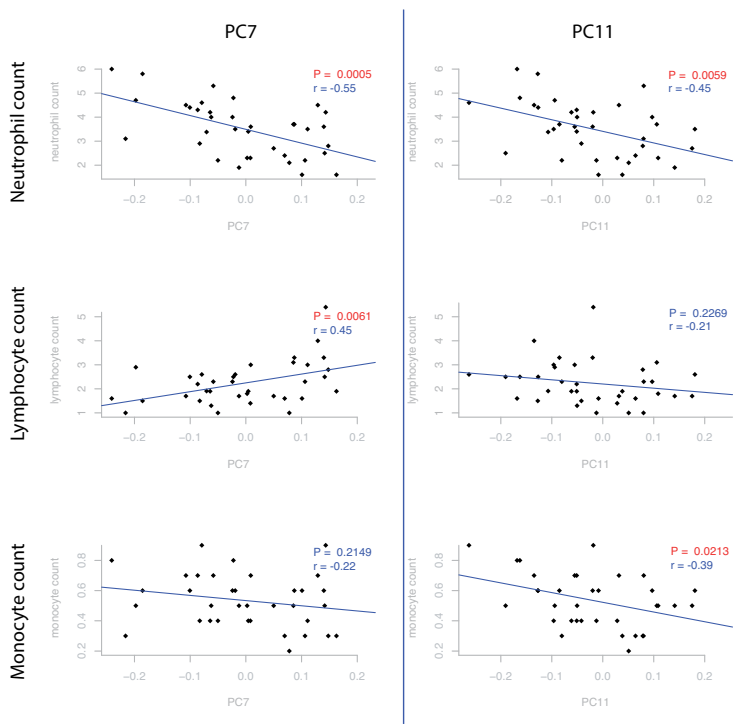
- Genes with highly positive factor loadings on PC7 are strongly associated with (and therefore likely expressed in) lymphocytes. This is in agreement with the positive correlation of PC7 with lymphocyte counts (Figure S2). Genes with the most negative factor loadings on PC7 are strongly associated with macrophages and neutrophils. This is in agreement with the negative correlation of PC7 with neutrophil counts (Figure S2).
- Genes with highly positive factor loadings on PC11 are strongly associated with different types of leukocytes, while genes with the most negative factor loadings on PC11 are strongly associated with erythrocytes.

References:

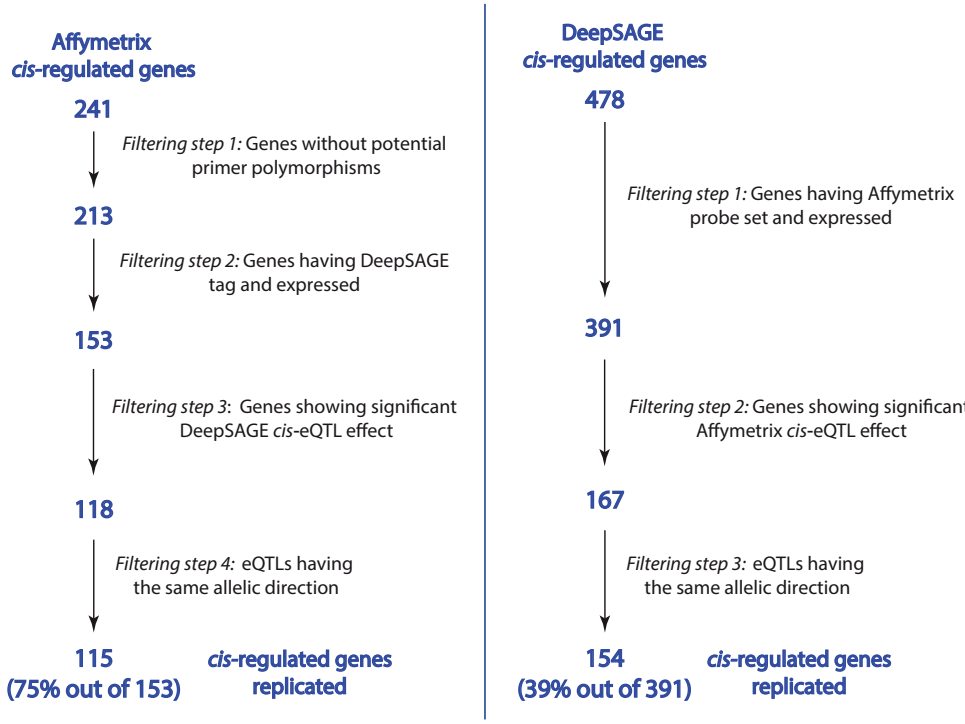
1. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768–772.
2. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36: e105.
3. Risso D, Schwartz K, Sherlock G, Dudoit S (2011) GC-Content Normalization for RNA-Seq Data. BMC Bioinformatics 12: 480.
4. Fusi N, Stegle O, Lawrence ND (2012) Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. PLoS Comput Biol 8: e1002330.
5. Parts L, Stegle O, Winn J, Durbin R (2011) Joint genetic analysis of gene expression data with inferred cellular phenotypes. PLoS Genet 7: e1001276.
6. Stegle O, Parts L, Piipari M, Winn J, Durbin R (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc 7: 500–507.
7. Jelier R, Schuemie MJ, Veldhoven A, Dorssers LCJ, Jenster G, et al. (2008) Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome Biol 9: R96.



Supplementary Figure 1. Correlation of GC content with principal component 1 (PC1) eigenvector coefficients for all the three datasets. Pearson correlation coefficient and corresponding p-value are shown in the plot.



Supplementary Figure 2. Blood cell counts in DeepSAGE data captured by the eigenvector coefficients on principal components PC7 (left) and PC11 (right). Experimentally determined blood cell counts at the time of RNA isolation were available for 36/94 samples. Blood cell counts are expressed as (number of cells)×109/L. Pearson correlation coefficients and corresponding p-values are shown in the plot.



Supplementary Figure 3. Replication of Affymetrix eQTLs in DeepSAGE dataset and DeepSAGE eQTLs in Affymetrix data. The numbers of unique cis-regulated genes is given after each filtering step.

CHAPTER 4

ASSESSING THE TRANSLATIONAL LANDSCAPE OF MYOGENIC DIFFERENTIATION BY RIBOSOME PROFILING

Eleonora de Klerk, Ivo F.A.C. Fokkema, Klaske A.M.H. Thiadens,
Jelle J. Goeman, Magnus Palmblad, Johan T. den Dunnen,
Marieke von Lindern, Peter A.C. 't Hoen.

Nucleic Acids Res. 2015 March.
doi: 10.1093/nar/gkv281.

ABSTRACT

The formation of skeletal muscles is associated with drastic changes in protein requirements known to be safeguarded by tight control of gene transcription and mRNA processing. The contribution of regulation of mRNA translation during myogenesis has not been studied so far.

We monitored translation during myogenic differentiation of C2C12 myoblasts, using a simplified protocol for ribosome footprint profiling. Comparison of ribosome footprints to total RNA showed that gene expression is mostly regulated at the transcriptional level. However, a subset of transcripts, enriched for mRNAs encoding for ribosomal proteins, was regulated at the level of translation. Enrichment was also found for specific pathways known to regulate muscle biology. We developed a dedicated pipeline to identify translation initiation sites (TISs) and discovered 5333 unannotated TISs, providing a catalog of upstream and alternative open reading frames used during myogenesis. We identified 298 transcripts with a significant switch in TIS usage during myogenesis, which was not explained by alternative promoter usage, as profiled by DeepCAGE. Also these transcripts were enriched for ribosomal protein genes. This study demonstrates that differential mRNA translation controls protein expression of specific subsets of genes during myogenesis.

Experimental protocols, analytical workflows, tools and data are available through public repositories (http://lumc.github.io/ribosomeprofiling- analysis-framework/).

INTRODUCTION

Myogenesis, the formation and maintenance of skeletal muscles, occurs during embryogenesis and muscle regeneration. During embryonic development, muscle progenitor cells are committed to the myogenic program and becomemyoblasts. Myoblasts fuse to formmultinucleatedmyotubes, which will give rise to muscle fibers. During muscle regeneration, the process is similar. Satellite cells are differentiated into myoblasts, which can fuse with existing myotubes to repair the adult muscle tissue (1).

The molecular mechanisms controlling myogenesis at the transcriptional level are well characterized. Several myogenic transcription factors, including MYF5, MYOD1, MYOG, MEF2 and MYF6, are expressed at different stages of myogenesis to tightly control the transcription of numerous muscle-specific genes encoding contractile proteins and to reorganize cell metabolism (2,3).

Less is known about the control of myogenesis at the level of mRNA translation. Several mechanisms enhance or repress translation through RNA binding proteins or miRNAs (4,5). The presence of translational enhancers able to interact with translation initiation complexes and increase protein synthesis have been reported also in the context of skeletalmuscle differentiation, where they target crucial differentiation factors (6). However, a genome wide overview of translational regulation, as it exists for transcription (7), is missing. Therefore we set out to investigate control of mRNA translation during myogenesis, with a focus on translation initiation.

Regulation at the translational level defines not only the abundance of a protein, but also the identity through the use of alternative translation initiation sites (TISs). Translation can initiate upstream or downstream of the primary open reading frame (pORF). TISs located in the 5' untranslated region (5'-UTR) of a transcript may give rise to upstream open reading frames (uORFs) or protein isoforms with extended N-termini (8). Translation of the uORF may have various consequences for the translation of the pORF: uORFs may repress translation, induce translation of protein isoforms truncated at their N-termini or even enhance translation of the pORF (9–19). TISs located in the coding region of the pORF may give rise to N-terminal truncated isoforms, with possibly different biological functions (20).

The complexity of the translatome is further increased by the presence of dual coding regions, nucleotide sequences that can be translated in more than one reading frame (21).

Recent studies based on ribosome footprint profiling have reported extensive regulation of protein expression at the translational level, in particular as a part of stress responses, but also under normal physiological conditions (8,22,23). Translational regulation is mostly exerted at the level of translation initiation, whereas translational elongation rates are more constant across conditions (22,24–27).

In mammals, between 50 and 65% of transcripts have been reported to contain at least two TISs (8,24,26), more than 50% of which are located upstream of the pORF. Nonetheless, to what extent gene expression is regulated at translational level is still being debated. A major role for translational regulation was hinted by studies that found a poor correlation between total mRNA and protein levels (20% (28) or no more than 40% (29–33)). However, other studies reported a much higher correlation (up to 80% (34)) and suggested that previously observed discrepancies between mRNA and protein levels were mainly of technical nature. Nevertheless, there is a role for translational regulation, at least for subsets of (functionally linked) proteins (35).

To explore how and to what extent myogenesis is regulated at translational level in mammalian skeletal muscles, we monitored translation at nucleotide resolution in a genome-wide high-throughput manner, using ribosome profiling on the murine C2C12 cell line, a model for skeletal muscle differentiation.

Ribosome profiling (25) is a method based on deep-sequencing of ribosome-protected mRNA

fragments that are recovered from mRNAs engaged by ribosomes, after digestion of non-protected regions of the mRNAs. Even though the ribosome profiling technique has been standardized and used in several studies (24,26,36,37), the isolation and sequencing of the ribosome footprints is laborious and the analysis represents a challenge due to short read length and noise surrounding genuine TISs.

We simplified the existing protocol and developed a data analysis pipeline to characterize translation initiation during differentiation of myoblasts into myotubes, to detect switches in the use of alternative TISs, as well as to quantify translation.

We further investigated the extent of translational control over transcriptional control by comparing ribosome profiling data with RNAseq, miRNA-seq and DeepCAGE data. miRNA-seq data was used to investigate the contribution of miRNAs in the regulation of gene expression at translational level. DeepCAGE data was used to identify transcription start sites (TSSs) and detect switches in the use of alternative promoters; this allowed us to discriminate between switches in TISs usage due to changes in the transcriptome and switches purely controlled during translation.

MATERIALS AND METHODS

Cell culture

Mouse myoblasts C2C12 were grown on collagen-coated plates in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 1% glucose and 2% glutamax (Invitrogen). Differentiation was induced by serum deprivation for 7 days, by culturing in DMEM supplemented with 2% FBS, 1% glucose and 2% glutamax.Cells were grown under 10% CO2.

Ribosome footprint profiling, DeepCAGE, RNAseq and miRNAseq sequencing libraries

Ribosome footprints libraries were prepared starting from 5 million C2C12 cells, seeded in 10 cm dishes. After 24 h in proliferation phase, myoblasts were treated with 100 g/ml cycloheximide (C7698-1G, Sigma) for 10 min or with 2 µg/ml harringtonin (sc-204771A, Santa Cruz Biotechnology) for 5 min followed by 10 min of cycloheximide treatment. Same treatment was performed in myotubes, after 7 days of serum deprivation.

After drug treatment at 37°C, dishes were transferred on ice and cells were washed with ice-cold phosphate buffered saline supplemented with 100 µg/ml cycloheximide. Cells lysis was performed using 1 ml ice-cold lysis buffer (1× salt buffer [10× solution contained 100 mM Tris, 120 mM MgCl2, 1.4MNaCl, pH 7.4], 0.5% IGEPAL) supplemented with RnaseOUT (500 U/ml, Invitrogen), dithiothreitol (DDT) (1.5 mM), cOmplete Protease Inhibitor Cocktail (40 µl of 25× stock,Roche) and cycloheximide (100 µg/ml). Nuclei were removed by centrifugation at 13 000 rpm in a FA-45-30-11 rotor (18 000 g) for 10 min. Supernatant was digested with RnaseI (1500U/ml, Ambion) for 30 min at room temperature. Digestion was blocked with SuperaseIN (600U/ml, Ambion) and lysate was layered on frozen sucrose gradients (7–46% sucrose) and separated by ultracentrifugation at 35 000 rpm in a SW 41 Ti rotor (210 000 g) for 3 h at 4°C.

Twelve fractions (750 µl each) were collected from the top and digested with proteinase K (0.15 mg/750 µl) for 30 min at 42 °C in the presence of 1% sodium dodecyl sulphate.

RNA was extracted by acid phenol (Ambion) purification followed by ethanol precipitation. For each sucrose gradient separation, an undigested lysate was used to monitor the polysome profile, determined on the Bioanalyzer (Agilent) with the RNA 6000 Nano kit. Fractions containing monosomes (corresponding to fractions nine and ten) were combined. Cytoplasmic rRNAs and mitochondrial

rRNAs were removed using Ribo-Zero Magnetic Gold Kit (Epicentre) according to manufacturer's instructions, with the following modifications: removal solution was incubated for 4 min at 68°C prior RNA addition, then mixed with RNA and kept at 37°C for 15 min; RNA hybridized to removal solution was incubated with magnetic beads at room temperature for 5 min, followed by 1 min at 50°C and 4 min at 37°C. Size selection of footprints with length 28–32 nt was performed on 15% TBE-urea gel (Invitrogen).

Footprints were dephosphorylated with T4 polynucleotide kinase (10U, NEB) and ligated to double stranded RNA adapters at both ends (SOLID Total RNASeq Kit, Ambion). RNA was reverse transcribed and amplified using indexed custom primers adapted for Illumina Hiseq 2000 (5'-AATGATACGGCGACCACCGATGGGCAGTCGGTGAT-3', 5'-GCGGAACCGGCATGTCGTC|index|AGCATAC GGCAGAAGACGAA-3'). Sequencing libraries were size selected for amplicons of 120 bps on 4–12% polyacrylamide gel electrophoresis gel (Novex TBE, Life Technologies).Atotal of twelve strand specific libraries were pooled and sequenced in one lane. Single end sequencing was performed on the Illumina Hiseq2000 for 50 cycles.

The complete protocol is available in the extended experimental procedures.

DeepCAGE libraries were prepared as described previously (38).

Strand specific RNAseq libraries were generated using the method described by Parkhomchuk et al. (39) with minor modifications. In short, mRNA was isolated from 500 ng total RNA using oligo-dT Dynabeads (Life Technologies) and fragmented to 150–200 nt in first strand buffer for 3 min at 94°C. First strand cDNA was generated using random primers. Second strand was generated using dUTP instead of dTTP to tag the second strand. Subsequent steps to generate the sequencing libraries were performed with the NebNext kit for Illumina sequencing with the following modifications: after adapter ligation to the dsDNA fragments, libraries were treated with USER enzyme (NEB M5505L) in order to digest the second strand derived fragments. Amplified libraries were pooled and sequenced in one single lane. Paired-end (2 × 100 bps) sequencing was performed on the Illumina HiSeq2000. miRNAseq libraries were prepared starting from purified small RNAs isolated with mirVana miRNA Isolation kit (Ambion) according to manufacturer's instructions. Sequencing libraries were prepared according to the method previously described (40) and single-end sequencing was performed on the Illumina Genome Analyzer II.

Protein isolation and western blot analysis

Protein isolation was performed starting from cell pellet recovered from 75 cm2 flasks. Cell pellet was resuspended in 500 µl of protein lysis buffer (50 mM HEPES, 50CmM NaCl, 10 mM ethylenediaminetetraacetic acid, 10 mM dithiothreitol (DTT), 0.1% 3-((3-Cholamidopropyl) dimethylammonium)-1- propanesulfonate (CHAPS), Complete Mini Protease inhibitor cocktail tablet (Roche)). Cell lysate was sonicated with ultrasound (5 s at amplitude 60 for three times) and incubated for 1 h at 4°C while rotating. Supernatant was recovered after centrifugation at 14 000 rpm in a FA-45–30–11 rotor (20 800 g) for 15 min at 4°C. Protein concentration was assessed using BCA Protein Assay kit (Pierce) according to manufacturer's instructions. Protein separation was performed on 18% Criterion TGX Gel (Bio-Rad) in 1× XT Tricine running buffer. A total of 30 µg of protein lysate were heat denatured in 2× Laemmli sample buffer (95°C for 5 min) prior loading. Proteins were transferred with Trans-Blot turbo transfer system (Bio-Rad) on a nitrocellulosemembrane (0.2 µmTrans blot turbo, Bio-Rad). The following primary antibodies were used: rabbit anti-RPL7 antibody (Bethyl, 1:2000), rabbit anti-RPS15 middle region (Aviva System Biology, 1:1000), rabbit anti-RPL34 (Abcam, 1:1000), anti-beta Actin (Abcam, 1:5000). RPL7, RPS15 and RPL34 were detected using goat anti-rabbit secondary

antibody (IRDye800CW, Licor,1:5000), b-Actin was detected using goat anti-mouse secondary antibody (IRDye680CW, Licor, 1:5000). Signals were visualized with the Odyssey Infrared Imaging Sistem (LI-Cor Biosciences).

Data analysis

Mapping of ribosome footprints, DeepCAGE and RNAseq reads.

Ribosome footprints reads were aligned to both transcriptome and genome references using a combined approach. Reads were first aligned to a transcriptome reference using Bowtie (41,42), with the following parameters: -k 1 -m20 -n 1, –best –strata –norc. An index transcriptome reference was built based on RefSeq RNA sequences (ftp:// ftp.ncbi.nlm.nih.gov/refseq/M musculus/mRNA Prot/ last modified 2013/05/08). Unmapped reads were then mapped to theGRCm38/mm10 genome reference using Bowtie with the following parameters: -k 1 -m 2 -n 1 –best –strata. For each SAM file, reads shorter than 25 nt were filtered out. SAM files were converted into a wiggle format, in which only the 5' end of each read was reported. For SAM files obtained from the transcriptome alignment, transcriptomic coordinates were converted into genomic coordinates and stored into a wiggle format. Wiggle files are available at http://gwips.ucc.ie/.

To retrieve corresponding genomic coordinates, we first mapped RefSeq RNA sequences (the same used to build the transcriptome reference) to the GRCm38/mm10 genome assembly using GMAP (43), with the following parameters: -f samse -n0. The corresponding genomic coordinateswere used to convert the transcriptomic coordinates of the mapped footprint reads. RefSeq RNA sequences which mapped to the genome with insertions and/or deletions (introns excluded) were not included when building the transcriptome reference.Wiggle files of each alignment containing the 5' ends of reads mapped were then merged.

DeepCAGE reads were trimmed to 27 nt and the first nucleotide at the 5' end was removed. Trimmed reads were aligned to the GRCm38/mm10 genome reference, with the following parameters: -m 10 -k 10 -n 2 –best –strata. For CAGE tags mapping to multiple genomic locations, we applied a weighting strategy, based on the number of CAGE tags within a 200 bp window around each candidate mapping location. A weight of 1.0 was assigned for uniquely mapped sequences, for multi-mapped tags weight varied from 0.0 to 1.0. Only tags with a weight equal or higher than 0.9 were kept (44).

Paired end RNAseq reads were aligned to RefSeq RNA sequences (ftp://ftp.ncbi.nlm.nih.gov/ refseq/M musculus/ mRNA Prot/) using Bowtie2 (45), with the following parameters: -N 1 –norc.

Triplet periodicity analysis in ribosome footprints.

Using a custom PHP script the merged wiggle files were converted to a format suitable for the Batch PositionConverter Interface in Mutalyzer (46) 2.0.beta-32 (https://mutalyzer.nl/ batch-jobs?job type=position-converter). These converted files with genomic coordinates were manually loaded into Mutalyzer to retrieve positions relative to the annotated TIS. The triplet periodicity pattern was analyzed by calculating the number of reads mapping in the first, second and third nucleotide of each codon for all detected transcripts. Positions were filtered out if they had coverage lower than 3 reads, or if they only mapped in intronic or intergenic regions (500 nt upstream or downstream from annotated coding regions) or if they only mapped to non-coding transcripts.

Since for samples treated with harringtonin the 5' end of a footprint was expected to be located −12 nt far from the TIS, positions located up to −15 nt were counted as positions in coding regions. To calculate the triplet periodicity, positions shared by overlapping transcripts were filtered as follows: if

a position was shared between a coding and an untranslated region (3' or 5'-UTR), the position was counted only for the coding region; if it was shared between 3'-UTR and 5'-UTR of two overlapping transcripts, the position closest to the coding region was reported, but only if the difference in the distance of the two positions relative to the coding region was larger than 100 nt, otherwise both positions were discarded.

Transcription start site assignment and annotation.

TSSs were assigned by summing the weighted number of CAGE tags at each genomic position. Weighted numbers were based on the MuMRescue software (44). Peaks located within a window of 20 bp were merged and reported whenever the coverage was at least 10 tags per million (tpm) in at least one experimental condition. BAMfiles were converted into BED files and annotated based on RefSeq collection, using intersectBed (BEDTools (47)). The RefSeq collection was modified by extending the 5'-UTR of each transcript with 500 nt. Peaks located more than 850 nt downstream the annotated TSS were not considered for further analysis. BAM files were converted into a wiggle format using custom scripts.

Translation initiation site assignment and annotation.

A dynamic local peak calling algorithm was developed to identify TISs in the ribosome footprint data from harringtonin treated cells. To discriminate between genuine initiation sites and noise, we evaluated the signal in the region surrounding each peak. Peaks were first filtered following the same procedures used for the triplet periodicity analysis, except for positions shared by overlapping transcripts, which were filtered as follows: if a position was shared between the 3'-UTR and a coding or 5'-UTR, the position was counted only for the coding region or the 5'-UTR. Peak calling was then performed after combining footprints from three independent biological replicates. Each position with a coverage of at least 20 reads was analysed and called if the following conditions were met: the peak had higher coverage compared to any peak located 3, 6, 9, 12 or 15 bases upstream; the peak showed a triplet periodicity pattern (the two nucleotides following the peak had a summed coverage 40% or lower than the total coverage of that codon); the five codons downstream should not contain a base with a coverage higher than that of the peak analysed; the five downstream codons, when having a coverage of at least 10% of the peak analysed, should show a triplet periodicity pattern.

Once a peak was called, the analysis continued at the next nucleotide, allowing the detection of TISs in different frames. For each gene, the TIS with the highest coverage was kept as reference, and any other TISs which had a coverage lower than 10% of the reference TIS was discarded for further analysis. For each called TIS, the coverage of that peak in each individual sample was reported. TISs were then classified into six categories: annotated TIS, 5'-UTR (or unannotated 5'-UTR) TIS, coding TIS, 3'-UTR TIS and multiple TIS. TISs mapping in position −12, −11 and −10 nt relative to the start codon were reported as annotated TISs. TISs mapping upstream of position −12 were annotated as 5'-UTR TISs (or unannotated 5'-UTR TISs if the TIS was not located in the 5'-UTR sequence present in the transcript's reference sequence), TISs located between position− 10 and the stop codonwere annotated as coding TISs, TISs located after the stop codonwere annotated as 3'-UTR TISs, TISs which fell in more than one category were annotated asmultiple (unless one of the categorieswas annotated TIS, which was then the only one reported). Peaks located 5 kb downstream of the annotated start codon (counted as transcript positions, not genome) were not considered for downstream analysis. The background noise in these regions was higher, likely because the ribosomes were not allowed

sufficient time to finish the translation of transcripts on which they had already engaged in the elongation phase at the start of the harringtonin treatment.
Wiggle files showing all the mapped footprints are available and visualized at http://gwips.ucc.ie/.

Differential expression analysis and functional annotation.

For RNAseq, CAGE and ribosome profiling data, custom scripts were used to quantify the number of mapped reads.

For miRNAseq data, the E-miR software package was used to map trimmed sequencing reads and quantify the number of mapped reads (40).

The statistical programming language R (version 2.15.1) was used for analysis of differential expression between myoblasts and myotubes. The analysis was performed using the R Bioconductor package edgeR (48) (version 3.0.8). A negative binomial model was fitted and GLM Tag-wise dispersion was estimated prior testing procedures. Exact P-values were computed using the exact test and adjusted for multiple testing according to Benjamini–Hochberg method (49). Differential expression analysis was performed at gene level after summing reads mapped to Refseq sequences (ftp://ftp.ncbi.nlm.nih. gov/refseq/M musculus/mRNA Prot/). KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis was performed using DAVID Functional Annotation Tool (50).

Statistical model to identify genes with alternative TIS or TSS usage.

We fitted the previously published logistic generalized linear mixed model (51) to the counts for each TIS (or TSS) using fixed effects for location, myotube, and their interaction, and a random intercept and location effect within cell culture. The model was fitted using R Bioconductor package lme4.0 (R version 2.15.1, http://cran.r-project.org/web/packages/lme4/index. html. Chi-squared likelihood ratio tests were used for testing the presence of location-myotube interactions, i.e. switches in TIS (or TSS) usage. Both a global chi-squared likelihood ratio test for the presence of any interaction and t-tests for individual effects per TIS were calculated.

Codon usage, uORFs and out-of-frame analysis.

For each TIS, the nucleotide sequence of the codon was reported based on RefSeq. For 5'-UTR TISs, sequences were reported up to the annotated TIS and translated into the corresponding amino acid sequence until the first stop codon or the annotated TIS. For TISs in unannotated 5'-UTR, the genomic sequences were retrieved from GRCm38/mm10 genome assembly using the genomic Refseq reference sequences). For any TIS located in the 5'-UTR and leading to a stop codon (upstream or downstream of the annotated start codon), the length of the amino acid sequence was calculated.The frame of coding TISs was defined by dividing the mRNA position (adjusted for the distance of the 5' end of the read relative to the actual TIS position) by 3.

IRES and 5' TOP analysis.

The 5'-UTR sequences of transcripts containing TISs in their 5'-UTR and transcripts containing TISs only in the annotated TIS and/or the coding region were retrieved from Refseq using UCSC Table Browser. Fasta files were uploaded into UTRScan (52) (http://itbtools.ba.itb.cnr.it/utrscan/help) and analysed for IRES and 5' TOP motifs. Enrichment was calculated by comparing the number of transcripts containing IRES and TIS in the 5'-UTR versus those containing IRES but no TISs in the 5'-

UTR. Transcripts containing IRES were then overlapped with transcripts containing uORFs.

In silico screening of alternative TISs.

Raw MS/MS proteomic datasets were retrieved from PRoteomics IDEntifications (PRIDE) database (accession numbers: PXD000328, PXD000022, PXD000065). Amino acid sequences for 24665 Mus Musculus proteins were retrieved from UniProt (http://www.uniprot.org/uniprot/?query=organism%3A%22mus+musculus%22+AND+reviewed%3Ayes+AND+keyword%3A1185&sort=score) in fasta format and used as background reference. A fasta file containing amino acid sequences of a set of candidate alternative and uORFs was created and merged with the UniProt reference file. For candidate ORFs containing non-canonical start codons, an alternative peptide sequence was included, where the first amino-acid was replaced with methionine.

The MS/MS analysis was performed using the Trans-Proteomic Pipeline v 4.6.3 (53). The raw MS/MS data were converted to mzXML and peptides identified by X!Tandem. The output files were then processed with PeptideProphet for spectrum-level validation and only spectra with probability greater than 0.90 were reported for manual inspection.

Accession codes and hyperlinks to public repositories.

Raw deep sequencing data from the C2C12 RNAseq, miRNAseq and ribosome footprint profiling are available for download at European Nucleotide Archive (http://www.ebi.ac.uk/ena) under accession number PRJEB7207.

Wiggle files of ribosome profiling data from cycloheximide and harringtonin experiments are available at http://gwips.ucc.ie/ and can be visualized as ‘Elongating Ribosomes’ and ‘Initiating Ribosomes’ tracks, respectively.

All analysis scripts together with a README file containing instructions for users are publicly available at GitHub: http://lumc.github.io/ribosome-profiling-analysis-framework/.

RESULTS

A simplified ribosome profiling protocol

The standard protocol for ribosome profiling (25) involves numerous steps, including the isolation of the protected ribosome footprints from the monosomes, obtained by RNAse digestion of cytosolic extracts and the conversion of the small single stranded RNA footprints into a double stranded DNA sequencing library. The conversion is usually accomplished by ligation of single stranded adapters to the 3' ends of the RNA footprints, followed by reverse transcription and circularization. The circular template is then used for polymerase chain reaction amplification (**Figure 1A**). Each of these stepsmay be subject to certain biases.

We simplified the existing method by converting the ribosome footprints into sequencing libraries with a standard small RNA sample preparation protocol, which avoids the multistep circularization procedure. Double stranded RNA adapters were ligated to the small RNA footprints, reverse transcribed and directly amplified for sequencing. This resulted in high quality ribosome footprints, as evidenced from the analyses described below. The complete protocol is available in the extended experimental procedures.

Analysis pipeline

High quality ribosome footprints are characterized by a distinct triplet periodicity pattern originating from the translocation of a ribosome from one codon to the next during translation elongation. In case of initiating ribosomes, the first nucleotide of each read is usually 12 nt upstream of the start of the codon that is being translated (25,54). These characteristics are commonly used as metrics for the quality of ribosome profiling data.

We developed a custom script to analyze the triplet periodicity pattern by converting the first position of the aligned reads to transcript coordinates and relating those coordinates to annotated TISs and the reading frame downstream of the annotated TIS. The script reports the number of reads mapping to the first, second and third nucleotide of each codon for all detected coding transcripts and the number of reads in each position relative to the annotated TIS.

The results of this procedure clearly show in our data that ~80% of reads mapped to the first nucleotide of each codon, as expected from previous studies (28,54,55) (**Figure 1B top, 1C,** Supplementary Table S1). For all samples, a major peak was observed at −12 nt from the annotated TIS, which is in accordance with previously reported data on the size and the position of the fragment protected by the ribosome (25,54). A higher percentage of 5' end reads mapping −12 nt from the annotated TIS was also observed for footprints generated by halting initiation of translation (harringtonin treated cells) compared to footprints generated by halting translation elongation (cycloheximide treated cells), as expected (**Figure 1D**).

The alignment of short (28–29 nt (Supplementary Figure S1)) ribosome footprints represents a challenge because footprints often span splice junctions, with short overhangs on either side of the junction. We calculated that 5421 murine transcripts (Supplementary Table S2) contain a TIS that is not mappable by standard genome alignment, because the TIS is located within a splice junction or it is located <15 nt upstream or downstream an exon–exon junction.

Common procedures to avoid loss of reads crossing exon–exon junctions use splicing-aware short-read alignment programs such as TopHat v1 (25,56). Alternatively, reads were mapped to the genome reference using a standard short read aligner, followed by the mapping of unaligned reads to known splice junctions using TopHat v1 (57). Both analyses are potentially flawed by the suboptimal

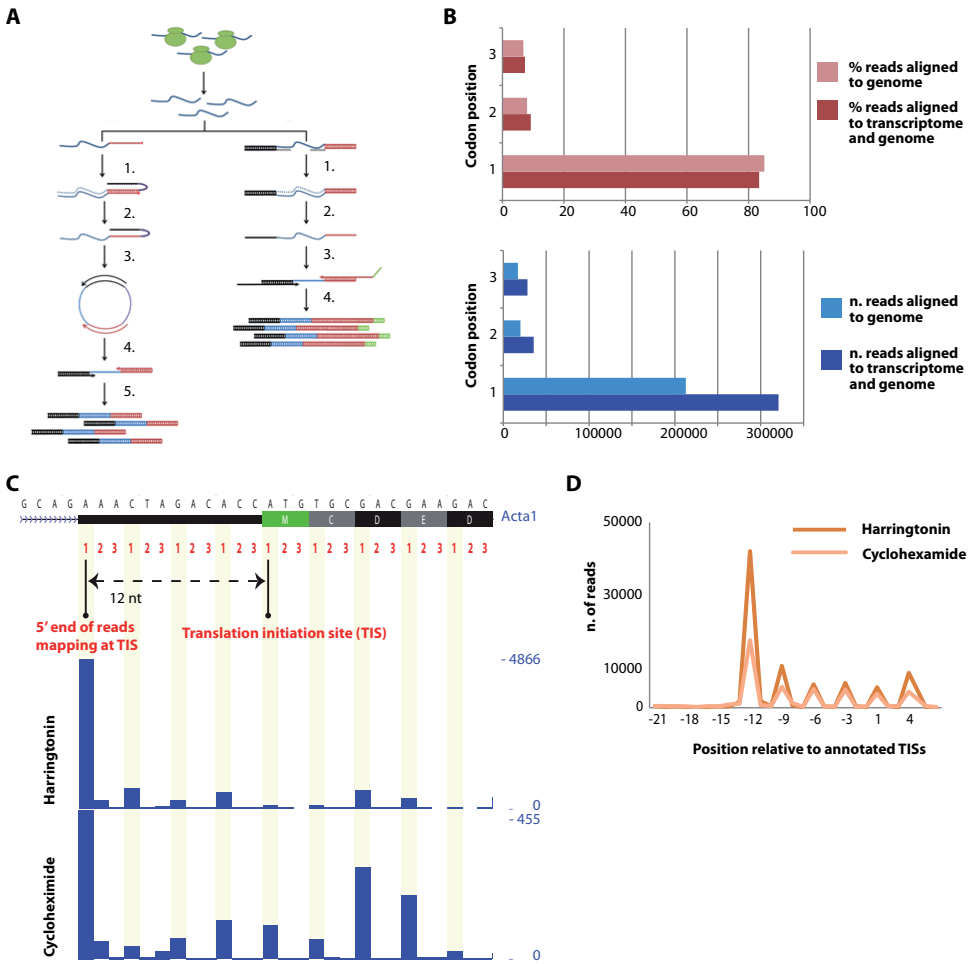


Figure 1. Generation and quality control of the ribosome profiling data sets. (A) An outline of our experimental procedures. The initial steps include the halting of ribosomes on the mRNA by harringtonin or cycloheximide, the treatment of cytosolic extracts with RNAse and the isolation of monosomes on sucrose gradients. These steps are identical to the original protocol developed by Ingolia et al. (23). The original protocol further includes the steps indicated in the left panel: single stranded adapters are ligated to the 3'ends of the RNA footprints (1), reverse transcribed (2, 3) and circularized (4) prior to amplification (5). The right panel shows our simplified sample preparation protocol. Double stranded RNA adaptors, with an overhang of six degenerated nucleotides, are ligated to the RNA footprints (1). Footprints are reverse transcribed (2, 3) and amplified for sequencing (4). **(B)** Percentage of reads mapped to the first, second and third position of each codon in all detected translated transcripts (top) and number of reads (bottom) mapped to the genome reference (light bars) or to combined transcriptome and genome reference (dark bars). **(C)** A screenshot of UCSC Genome browser displaying the triplet periodicity of the 5' ends of footprints mapped to Acta1 gene. Harringtonin and cycloheximide treated myoblasts are shown as independent traces. The y-axis represents the coverage of the highest peak. On top of the coverage tracks, the first, second and third nucleotide positions are shown for each codon for the first 27 nucleotides of the first exon. Arrows display the distance of the highest peak relative to the annotated start codon. **(D)** Number of reads mapped to the first 2 codons and up to 21 nucleotides upstream the start codon for harringtonin (dark yellow) and cycloheximide (light yellow) treated myoblasts.

performance of TopHat v1 on reads that are as short as 30 nt. Even though the upgraded TopHat2 performs better in the alignment of exon–exon junction reads that extend 10 nt or less into one of the exons (58), its performance has been optimized for long paired-end reads. Another problem in the genomic mapping of short RNA-derived reads is the presence of pseudogenes. The alternative of mapping exclusively to the transcriptome is also not ideal because it may miss hits in unannotated transcripts or in unannotated parts of transcripts, such as alternative first exons (59).

To overcome these limitations, we performed a combination of transcriptome and genome alignment. Footprint reads were aligned to a transcriptome reference, and only reads that did not map to the transcriptome were aligned to the genome (Supplementary Table S3). Mapping first to the transcriptome and then to the genome slightly reduced the number of reads mapping to pseudogenes (Supplementary Tables S4 and S5). The coordinates from the reads mapping to the transcriptomewere converted to genomic coordinates and then combined with the mappings from the genome alignment. The improvements obtained by the combined alignment can be appreciated by the recovery of ~30% of otherwise unmappable reads. These reads are likely genuine ribosome footprints as they show a triplet periodicity identical to the reads that do not span exon–exon junctions (**Figure 1B, bottom**).

A dynamic local peak calling algorithm was then developed to identify TISs in the ribosome footprint data from harringtonin treated cells. The developed algorithm evaluates the signal in the region surrounding each peak, takes into account the triplet periodicity in the nearby codons and is able to report start codons in different frames.Acomplete description is available in 'Materials and Methods' section.

Scripts used for the combined alignment, triplet periodicity analysis and peak calling are publicly available at http://lumc.github.io/ribosome-profiling-analysis-framework/.

Experimental setup

We performed ribosome profiling on undifferentiated myoblasts and differentiatedmyotubes fromthemurine C2C12 cell line, a well-characterized model for skeletal muscle differentiation (60). Ribosome footprints were recovered from initiating ribosomes and elongating ribosomes after halting translation with harringtonin or cycloheximide, respectively, analyzing three independent cultures for each condition.

Ribosome footprints derived from coding and non-coding genes

Footprints recovered after halting translation with harringtonin or cycloheximide mainly mapped to protein coding genes (**Figure 2A**, Supplementary Tables S4 and S5). Reads mapping to repetitive sequences, including contamination from ribosomal and transfer RNAs, are shown separately in Supplementary Table S3.

In addition, a relative high proportion of reads mapped to long intergenic non-coding RNAs (lincRNAs) (between 5 and 10% in average) and small RNAs (between 10 and 20% in average). To address the coding potential of lincRNAs in our dataset, we compared the read length distribution of footprints mapping to coding genes, non-coding genes (all genes with accession prefix 'NR ' in the RefSeq collection, including also lincRNAs) or only lincRNAs (**Figure 2B**). Footprints mapping to protein-coding genes were preferentially 29 nt long, whereas footprints from non-coding genes did not show this preference in length. The read-length distribution of footprints mapping only to lincRNAs was similar to the one of footprints mapping to any other noncoding genes. Nevertheless, in both cases a portion of reads was 27–30 nt long.

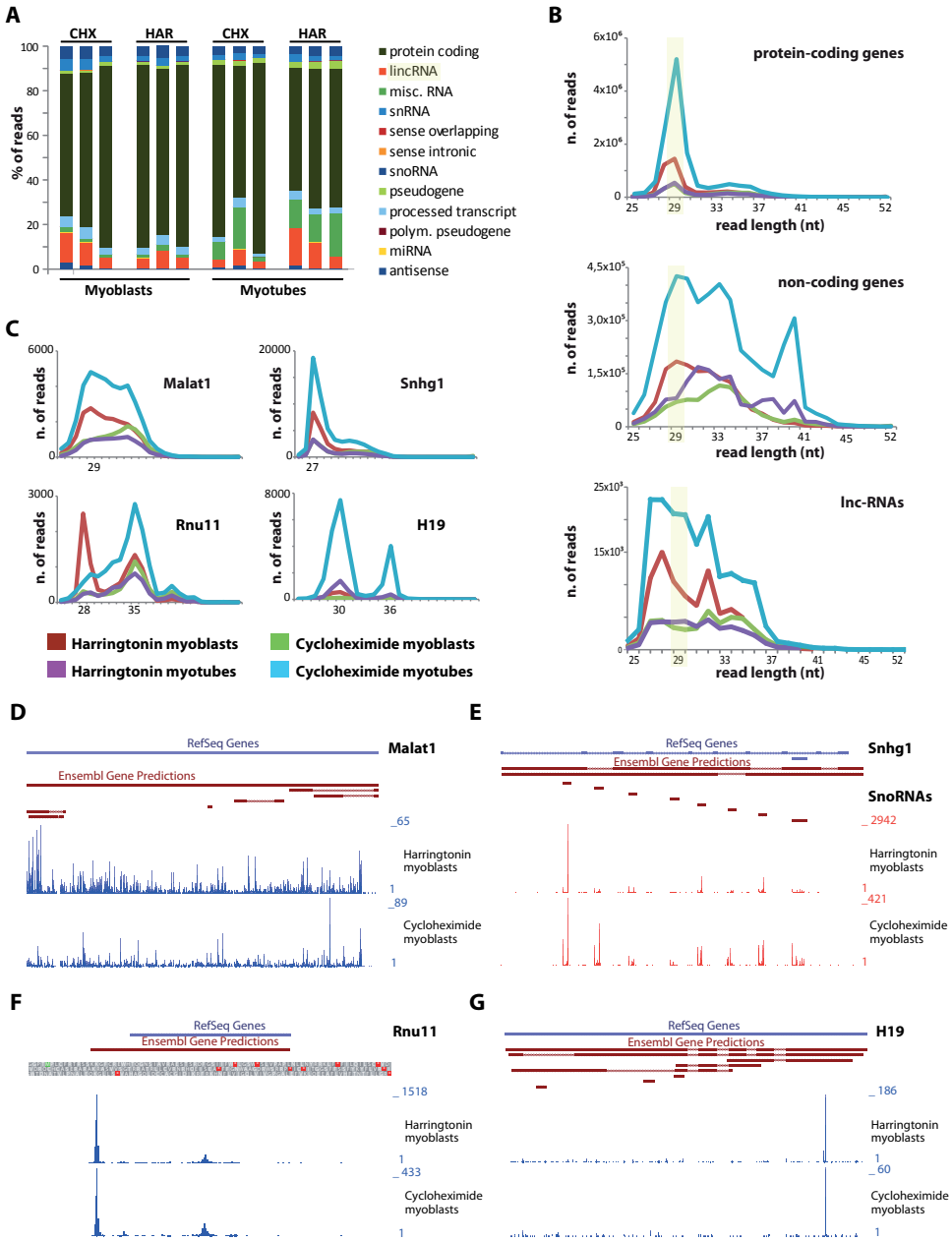


Figure 2. (A) Percentage of reads mapped to coding and non-coding genes in myoblasts and myotubes treated with harringtoning (Har) or cycloheximide (Chx). **(B)** Read-length distribution of footprints mapping to protein-coding genes (top), non-coding genes (including small and long non-coding genes) or only to lincRNAs (bottom). **(C)** Read-length distribution of footprints mapping to Malat1, Snhg1, Rnu11 and H19. **(D-G)** Coverage patterns for Malat1, Snhg1, Rnu11 and H19 in harringtonin (top traces) and cycloheximide (bottom traces) treated myoblasts.

We identified highly covered lincRNAs with protected fragments of 27–36 nt (e.g. Malat1) or with a preference for reads 27–30 nt long (e.g. Snhg1, Rnu11, H19) (**Figure 2C**). Malat1 reads mapped along the full body of the transcript, it did not show a preferential peak at AUG codons nor other common non-AUG start codons and it lacked of a drop of coverage at any corresponding stop codon (**Figure 2D**). Snhg1 showed coverage in intronic regions transcribing for snoRNAs, as previously reported in the Gas5 transcript (**Figure 2E**) (61). The coverage in Rnu11 (**Figure 2F**) and H19 (**Figure 2G**) was restricted to one or two regions, and no difference was shown between the cycloheximide and harringtonin treatment.

Based on these observations, we suggest that the majority of footprints deriving from lincRNAs in our dataset do not have a coding potential.

Subsets of mRNAs primarily regulated at translational level during myoblasts differentiation

To investigate the impact of translational regulation inmyogenesis, ribosome profiling data were compared to regular RNAseq data on total poly(A)+ RNA. The numbers of genes detected by ribosome profiling and RNAseq were similar (Supplementary Figure S2). Switches in the abundance of known markers of myogenesis were observed in both the RNAseq and the ribosome profiling data, as exemplified by the upregulation of the myogenic markers Myog, Tnnc1 and Myh7, and the downregulation of Myf5 (Supplementary Table S6).

Differential expression between myoblasts and myotubes was analysed at the gene level and the calculated log fold changes were compared between ribosome-bound RNAs (Supplementary Tables S7 and S8) and total RNA (Supplementary Table S9). Overall we observed a positive correlation between total and ribosome-bound RNAs (r = 0.71 and 0.65 for cycloheximide and harringtonin footprints, respectively, Pearson correlation) (**Figure 3**). However, the fold change observed in ribosome-bound RNA is generally lower than the fold change in total RNA, as demonstrated by the slope of the regression line (0.46 for cycloheximide footprints [95% confidence interval: 0.457–0.474] and 0.42 for harringtonin footprints [95% confidence interval: 0.413–0.431]). This is indicative for a general dampening effect of translational regulation.

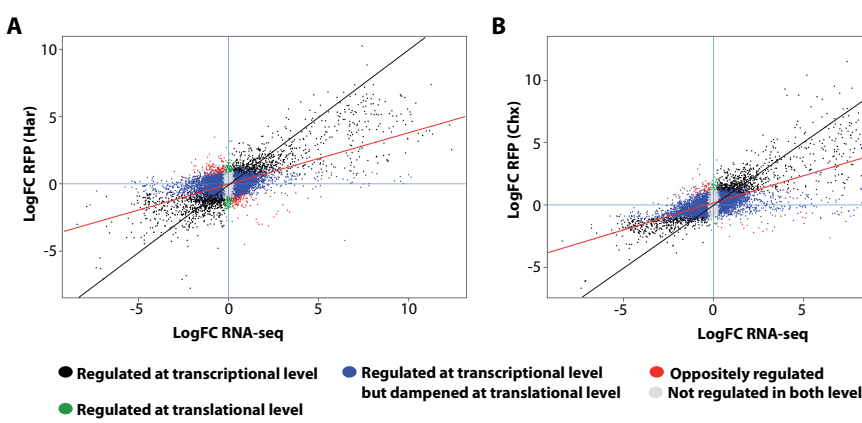


Figure 3. Changes in transcription and translation during myogenesis. Scatterplot showing differences in total RNA (x-axis) and ribosome-associated RNA (y-axis) from harringtonin **(A)** or cycloheximide **(B)** treated myoblasts and myotubes. Each data point represents the log-transformed fold change between myotubes and myoblasts. The red line indicates the slope, whereas the black line indicates the diagonal.

A subset of genes showed discrepant total and ribosome-bound RNA levels. In harringtonin-treated C2C12 (**Figure 3A**), 5680 genes showed significant changes between myoblasts and myotubes (P-value < 0.05) in total RNA but not in ribosome boundRNAlevels, indicative of a dampening effect of translation on transcription-induced changes.Atotal of 431 genes were regulated exclusively at translational level but not at the transcriptional level. Finally, 544 genes were regulated in opposite direction, meaning that they were upregulated at transcriptional level but showed lower translational efficiencies or vice versa. In cycloheximide-treated cells (**Figure 3B**), a similar trend was observed,

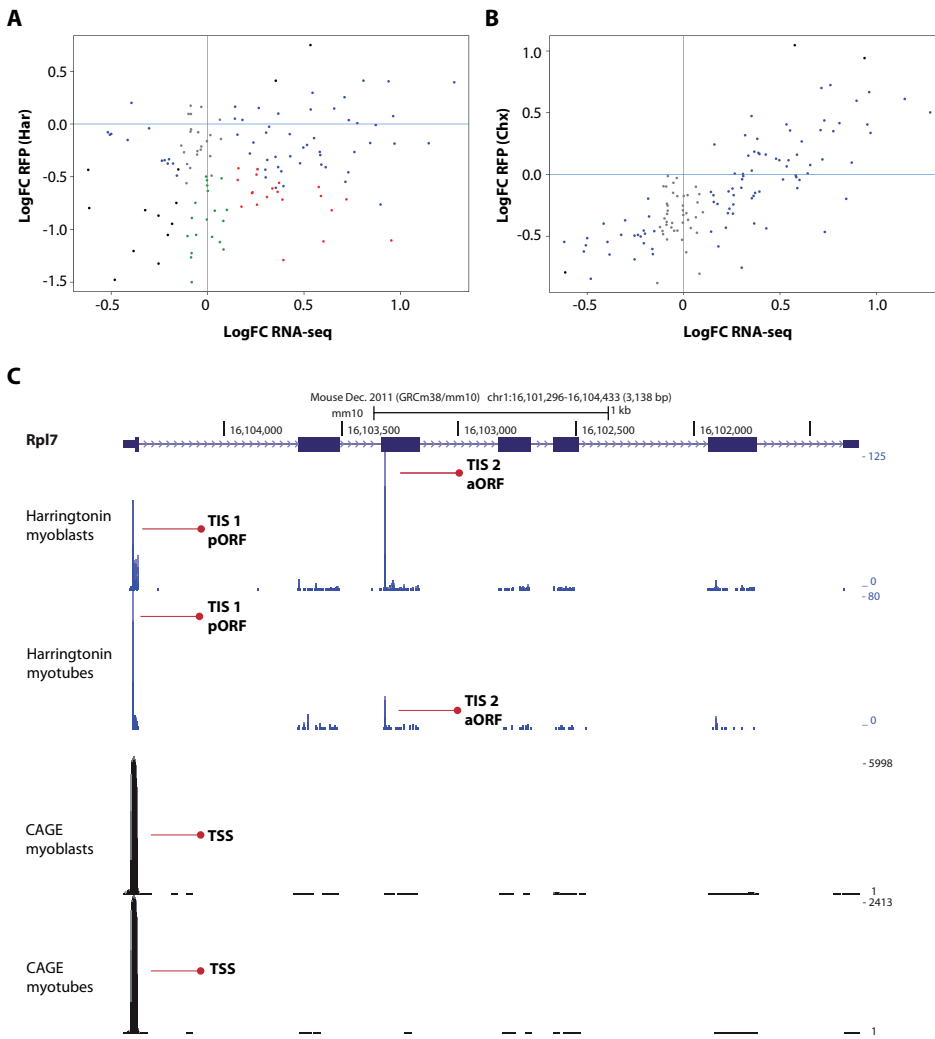


Figure 4. Translational regulation of ribosomal protein genes. Scatterplots show differences in levels of transcribed and translated ribosomal protein genes. Total RNA (x-axis) and ribosome-associated RNA (y-axis) from harringtonin **(A)** or cycloheximide **(B)** treated myoblasts and myotubes are shown for 145 genes belonging to the ribosome KEGG pathway. **(C)** A screenshot of UCSC Genome browser displaying alternative translation start sites (TISs, first and second traces) and transcription start sites (TSSs, third and fourth traces) in myoblasts and myotubes of the ribosomal protein gene Rpl7. TSSs and TISs leading to the translation of the primary open reading frame (TIS1) or predicted alternative open reading frame (TIS2) are indicated by red arrows.

even though the number of genes that reached a statistical significance (P-value < 0.05) was lower (6902 genes were regulated during transcription but dampened at translational level, 66 genes were regulated only at translational level and 73 showed antidirectional changes).

We next addressed the contribution of miRNAs on the regulation of gene expression at the level of translation.We found 105 miRNAs differentially expressed between myoblasts and myotubes (Supplementary Table S10) in our miRNAseq data, of which 66 were upregulated. We then focused on the effect of nine well-characterized myomiRs (mir-206, mir-1a, mir-22, mir-27b, mir-133a, mir-155, mir- 29c, mir-675 and mir-181a-5p) and compared the calculated log fold changes of experimentally validated targets between ribosome-bound RNAs (cycloheximide treatment) and totalRNA.

For 8 out of 9 analyzed myomiRrs, the correlation in fold change of their targets was not significantly different from the general correlation (Supplementary Table S11, Figure S3). These data suggest that miRNA regulation does not contribute strongly to the observed translational regulation (**Figure 3**).

We continued our comparison of transcriptome and translatome by performing a pathway enrichment analysis (Supplementary Table S12) on the subsets of genes showing discordant regulation.

mRNAs coding for ribosomal proteins displayed the highest enrichment in the subset of genes showing opposite regulation between transcription and translation (these genes were downregulated at translational level but upregulated at transcriptional level, Pvalue 2.2 × 10−7), and in the subset of genes downregulated only at translational level (P-value 3.7 × 10−13). mRNAs involved in the proteasome pathway showed a moderate enrichment in the subset of oppositely regulated genes (P-value 3.5 × 10−5), followed by mRNAs involved in focal adhesion (P-value 3.3 × 10−4), regulation of actin cytoskeleton (P-value 7.7 × 10−4) and calcium signaling (Pvalue 2 × 10−2).

To determine whether the discordant regulation was affecting the full pathway or only a subset, we compared log fold changes of all genes belonging to each enriched pathway.

The correlation observed between RNAseq data and ribosome profiling data for all genes that are part of the calcium signaling pathway was high (r = 0.84 for both cycloheximide and harringtonin footprints, Pearson correlation), suggesting that only a subset of calcium signaling genes is differentially translated. Similar high correlations were observed for all the other pathways, except for ribosomal protein genes. A poor correlation was found between RNAseq data and ribosome profiling data for ribosomal genes, when comparing RNAseq and harringtonin footprints (r = 0.27, Pearson correlation, P-value = 0.0018) (**Figure 4A**). The comparison between RNAseq and cycloheximide footprints, however, showed a positive correlation (r = 0.79, Pearson correlation, P-value < 2.2e-16) (**Figure 4B**). The discrepancy between ribosome footprints of initiating and elongating ribosomes suggested that not all initiating ribisomes were leading to translation of the ORF and/or that ribosome stalling was affecting the counts for elongating ribosomes. We therefore focused on the characterization of translation initiation.

Characterization of translation initiation in myogenesis

Data from harringtonin-treated cells were used to identify TISs used in myoblasts and myotubes. After mapping and filtering procedures, and combining the reads from the triplicate experiments, 3,052,146 and 976,468 reads were used to assign TISs in myoblasts and myotubes, respectively. The above described dynamic local peak calling algorithm was used to discriminate between noise and genuine initiation sites in the surrounding region of each peak.

We detected a total of 6,823 TISs in myoblasts (Supplementary Table S13) and 2,371 TISs in myotubes (Supplementary Table S14), corresponding to 4,106 and 1,561 coding genes, respectively. Our analysis showed that ~45% of the detected genes in myoblasts had two or more TISs, whereas in

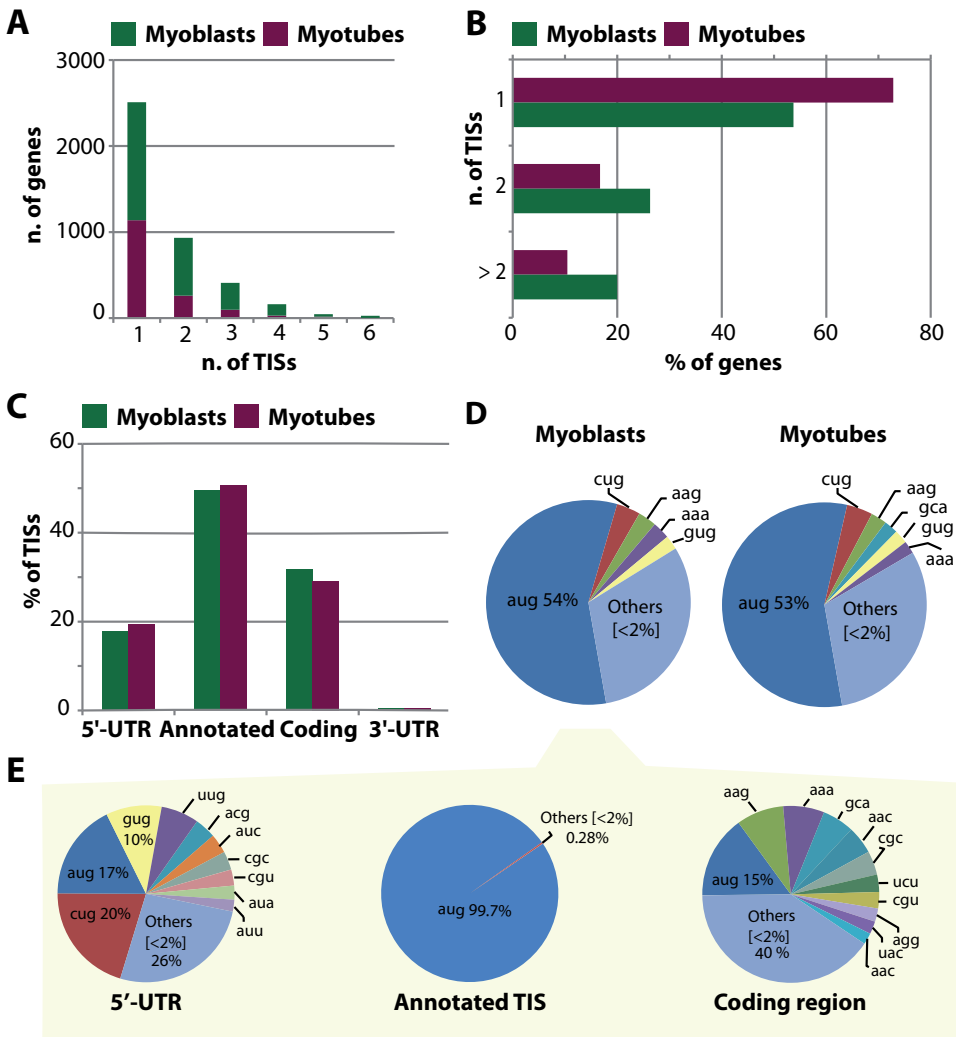


Figure 5. Alternative translation initiation sites used during myogenesis. Bar graph shows **(A)** the number of TISs per gene in myoblasts (green bars) and myotubes (purple bars), **(B)** the percentage of genes containing one, two or more than two TISs, **(C)** the percentage of TISs located in the 5'-UTRs, in the annotated start codons, in the coding regions or in the 3'-UTRs. Multiple indicates TISs mapping to more than one of the listed categories. **(D)** Pie chart shows the percentage of reads containing AUG and non-AUG codons at all detected TISs for myobalsts (left) and myotubes (right) treated with harringtonin. **(E)** Distribution of reads with AUG and non-AUG codons at detected TISs located in the 5'-UTRs, in the annotated start codons and in the coding regions. Distribution is shown only for myoblasts.

myotubes the percentage was slightly lower (~30%) (**Figures 5A and 5B**). The number of genes with more than six TISs was only ~0.5 and 0.6% in myoblasts and myotubes, respectively.

Approximately 50% of the footprints coincided with annotated start codons (**Figure 5C**), whereas ~20% mapped in the 5'-UTRs (of which 6.5% in unannotated 5'-UTRs, <500 nt upstream of the annotated start codon). A considerable amount of footprints (~30%) were found within coding regions, ~5% of which led to in-frame ORFs, hinting at alternative start codons for protein isoforms with truncated N-termini for 107 genes in myoblasts and 50 genes in myotubes. No general shift in the

localization of TISs was observed during myogenic differentiation (**Figure 5C**).

Around 55% of the footprints in the detected TISs contained the canonical AUG codon (**Figure 5D**). Notably, footprints of TISs located in the 5'-UTRs were enriched for alternative codons, primarily CUG and GUG, in accordance with the notion that uORFs frequently use weaker, non-canonical start codons (**Figure 5E**) (24,26,28,60). Footprints of TISs located in the unannotated 5'-UTRs were also mainly mapping to the non-canonical codons CUG and GUG (Supplementary Tables S15 and S16), except in myoblasts where the percentage of footprints with a canonical AUG codon was higher (32 against 16% in myotubes). This discrepancy mainly originated from footprints mapping to two TISs, corresponding to the highly expressed splicing factor Sf3b6 and mitochondrial gene Prelid1 (Supplementary Figure S4). These two detected TISs were followed by a stop codon upstream of the pORFs, according to the genomic sequence, but no TISs were detected at the annotated start codons, which may be due to the sort distance between uORF and annotated TIS. Since the unannotated 5'-UTR sequence may contain intronic sequences, it is impossible to determine whether these TISs represent uORFs or genuine start codons from wrongly annotated genes.

To distinguish between uORFs and alternative extended N-termini, we focused on TISs located in the annotated 5'-UTRs, and we classified them based on their reading frame in relation to the pORF and the presence of stop codons.

60% of the detected TISs located in the 5'-UTRs were leading to stop codons prior the start of the pORFs (corresponding to 1,274 TISs and 380 TISs in myoblasts andmyotubes, respectively) (**Figure 6A**). The length of these uORFs ranged from1 to more 100 amino acids (**Figure 6C**), but the majority (~85%) were between 1–30 amino acids (50% was shorter than 10 amino acids). The remaining 40% of the TISs located in 5'-UTRs were not leading to stop codons prior the start of the pORF, but ~72% of these uORFs was in a different reading frame than the pORFs, leading to overlapping uORFs, whereas the remaining 28% was in-frame with the pORF, suggesting the presence of isoforms with extended N-termini (**Figure 6B**). The length of the overlapping uORFs was longer than the one of the nonoverlapping uORFs, reaching up to 400 amino acids and with only ~40% being shorter than 30 amino acids (**Figure 6D**). We then investigated whether the usage of TISs in the 5'-UTRs sequences was associated with the presence of known regulatory elements, such as Internal Ribosome Entry Sites (IRESs) and Terminal Oligopyrimidine Tracts (5' TOP).

A significant enrichment of predicted IRES was found in transcripts with TISs in the 5'-UTRs, compared to transcripts forwhichwe detected TISs only in the annotated start codon and in the coding region. 36% of the transcripts containing TISs in the 5'-UTR had IRESs (Supplementary Table S17), whereas the percentage dropped to 24 for transcript without TISs in their 5'-UTRs in myoblasts (27 against 20% in myotubes, respectively).

No significant enrichment was found for predicted 5' TOPs (Supplementary Table S18), and overall the percentage of transcripts with a TIS in the5' UTR and the presence of a predicted 5' TOP was lower compared to the percentage of transcripts containing predicted IRES (~4% for both myoblasts and myotubes). These results suggest that for these genes uORFs do not play an important role in the regulation of mRNAs starting with a 5' TOP in myogenesis, whereas they may favor the use of IRESs in a subset of genes.

Alternative translation initiation independent of alternative promoter usage

Differences in TIS usage during skeletal muscle differentiation could derive from regulation at the transcriptional level, due to alternative promoter usage. Alternatively, a switch in TIS usage may occur

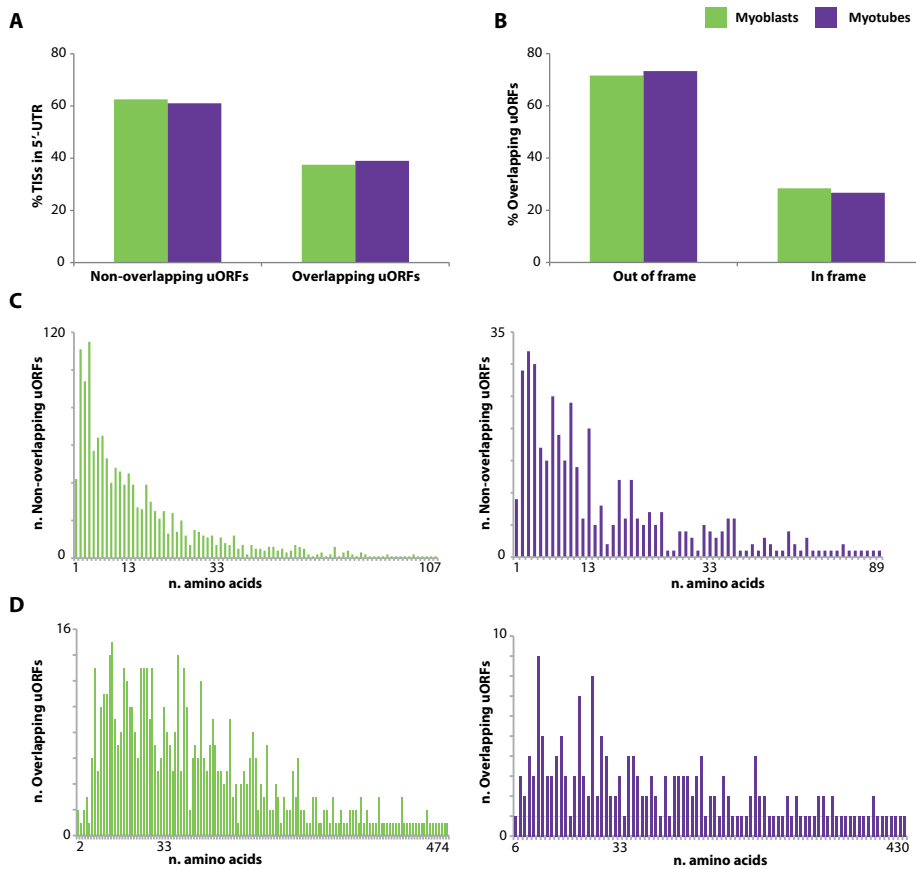


Figure 6. Potential uORFs used during myogenesis. (A) Percentage of TIS located in the 5'-UTRs leading to a stop codon before the annotated start codon of the primary open reading frame (non-overlapping uORFs) or overlapping the primary open reading frame. **(B)** Percentage of TISs in-frame and out-of-frame with the overlapped primary ORF. **(C)** Length distribution of non-overlapping and **(D)** overlapping uORFs in myoblasts (green) and myotubes (purple).

in the same transcript and arise from regulation at the translational level, because of differential recognition of upstream or alternative ORFs due to altered activity of translation initiation factors or RNA binding proteins. Regulation at the translational level can also occur through different efficiency in the translation of transcript variants. An example of a gene with a combination of both scenarios is Tpm3, a cytoskeletal protein involved in the calcium dependent regulation of muscle contraction. Two different TISs were detected in Tpm3: one TIS arising from a shorter transcript was predominantly used in myoblasts, another TIS arising from a longer transcript with alternative first exons was predominantly used in myotubes (**Figure 7A**). This results in the formation of proteins with two distinct N-termini, a longer isoform of 285 aa (UniProt P21107–1, also known as skeletal muscle isoform) and a shorter isoform of 248 aa (P21107-2, also known as cytoskeletal isoform). In addition to the nature of the transcribed protein, the efficiency of translation seems to be tightly controlled. As a measure for translational efficiency and to assess the effects of changes in TSSs, we analysed DeepCAGE data to detect 5'-ends of transcripts.

DeepCAGE data for the same gene in the same cells showed three different TSSs. The most distal

(3') TSS does not appear to code for a protein. The other two code for the short (cytoplasmic) and long (skeletal muscle) transcript variants and were transcribed at similar levels in myotubes (**Figure 7B**). However, the short variant was not translated in myotubes, but only in myoblasts (**Figure 7A**). Interestingly, tropomyosin proteins have already been shown to be regulated at translational level in slow-twitch and fast-twitch muscles (62).

To investigate the extent of translational regulation during myogenesis, we assessed the statistical significance of TIS switches for all genes with more than one TIS. From 4219 genes for which we could identify TISs, 1729 genes contained at least two TISs. Out of those, 312 genes (18%) (Supplementary Table S19) showed a significant difference (P-value < 0.05, after multiple testing correction) in alternative TIS usage between myoblasts and myotubes. To account for changes derived fromregulation at transcriptional level, we performed the same analysis to detect changes in TSS usage as detected by DeepCAGE. Out of 6426 detected genes, 635 genes contained two or more TSSs, and 28% (180) of those showed a significant change (P-value < 0.05, after multiple testing correction) in TSS usage betweenmyoblasts andmyotubes (Supplementary Table S20).

The overlap between genes with both changes in TISs and TSSs usage was small (**Figure 8A**), indicating that the majority of switches in TIS are occurring in transcriptswith the same start site. Even transcripts with a switch in both TIS and TSS usage appeared to be at least partly regulated at the translational level.

Cryab is an example of such a transcript. Two major TSSs were detected inmyoblasts, whereas in myotubes only one of the two TSS was detected (**Figure 8B**). Ribosome footprints from myoblasts showed (i) a TIS in the 5'-UTR, which represents an uORF with an harringtonin peak corresponding to an AUG start codon in a Kozak consensus sequence, and cycloheximide footprints on the entire 35 amino acids uORF, in addition to (ii) a TIS representing the pORF (**Figure 8C**). In myotubes, however, only the TIS corresponding to the annotated start codon was detected. Ribosome profiling footprints of cycloheximide treated cells showed a significant upregulation of Cryab in myotubes compared to myoblasts (Supplementary Table S8). This indicates a negative effect of the uORF on translation in myoblasts. The short distance between the uORF and the pORF, plus the relatively long uORF (35 amino acids) suggest that translation re-initiation in myoblasts is impaired. In agreement with our finding, a previous study has shown upregulation of Cryab at protein level in myotubes (63).

To identify which other genes are likely subjected to translational regulation by expression of uORFs, we selected genes with a significant switch between the annotated TIS and a TIS in the5'-UTR region, as evident from the interaction P-value of relative TIS usage and differentiation status (Supplementary Table S21). This led to the identification of 27 genes containing uORFs regulated during differentiation. Many of these genes, including Cryab (63,64), Vim (65), Spp1 (66), Eno3 (67,68), Pgam (69), Agl (70), Tmbim6 (71), Asb8 (72) and Cs (73), are known to be involved in the development, regeneration and/or homeostasis of skeletal muscles in humans (**Table 1**). Moreover Eno3 (74,75) and Spp1 (76) have been recently reported as biomarkers for Duchenne muscular dystrophy, where their protein expression levels changes in Duchenne patients through molecular mechanisms not yet fully understood.

KEGG pathway analysis on the complete set of genes with changes in alternative TIS usage showed moderate enrichment of only two pathways, ribosomal proteins genes and genes involved in the calcium signaling pathway (Supplementary Table S22), pathways that were also enriched in a comparison of transcriptome versus translatome (Supplementary Table S12). None of the 27 genes with switches involving uORFs was listed in the set of genes belonging to these two pathways, indicating that the observed switches identified in ribosomal protein genes and calcium signaling genes were mainly occurring between an annotated TIS and a TIS in the coding region, or between

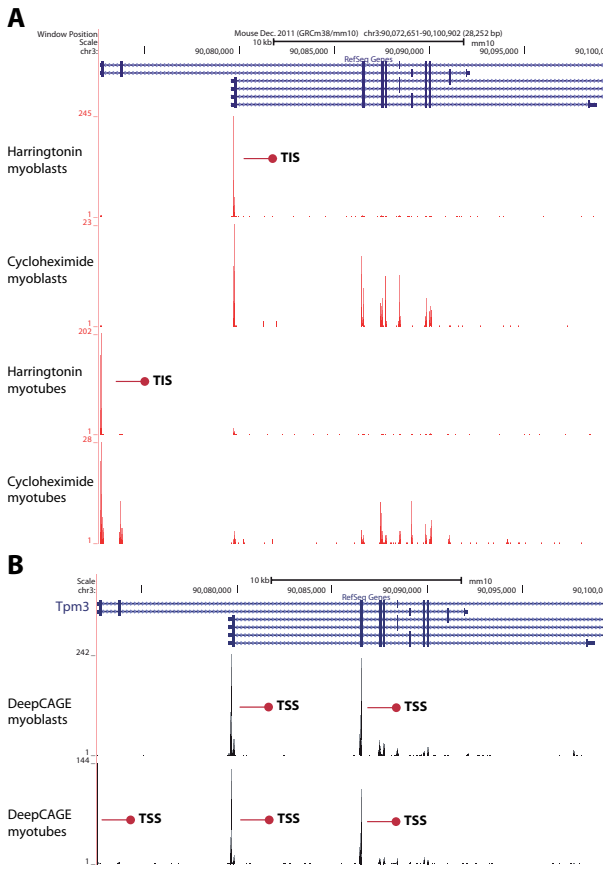


Figure 7. Alternative transcription and translation in Tpm3. On top of the coverage tracks, the six annotated transcripts are shown. **(A)** Two translation initiation sites (TISs, indicated by red arrows) arise from two transcript variants and are differentially used in myoblasts (first trace, harringtonin) and myotubes (third trace, harringtonin). Footprints in the complete open reading frame are shown for myoblasts (second trace, cycloheximide) and myotubes (fourth trace, cycloheximide). **(B)** Two transcription start sites (TSSs, indicated by red arrows) are used in myoblasts (top trace) and three transcription start sites are used in myotubes (bottom trace).

different TISs in the coding region. Another plausible explanation is the imperfect annotation of KEGG pathway, as seen for Tmbim6, a gene involved in calcium signaling (76) but not listed in the calcium signaling pathway (mmu4020).

Switches between alternative TISs which were observed in ribosomal genes were indeed mainly occurring between the annotated TISs and alternative TISs located downstream in the coding region. These switches were not due to alternative TSS usage as generally only one TSS was found. An example of such a ribosomal protein gene is Rpl7 (**Figure 4C**). The TIS detected in the coding region may lead to translation of a shorter novel protein, containing a methionine at its start codon and translated in a different frame compared to the pORF. This aORF may therefore represent a dual coding region. Rpl7 was downregulated at total RNA level, but the downregulation did not reach significanctly in both ribosome profiling datasets.

We determined RPL7 protein level by western blot analysis, and detected similar protein expression levels between myoblasts and myotubes (Supplementary Figure S5), suggesting that the alternative out-of-frame TIS does not affect the translation of the pORF and that the downregulation observed at total RNA level is buffered during translation. Western blot was also performed for two other ribosomal proteins, RPL34 and RPS15, where the detected alternative TISs in the coding regions were in-frame with the pORFs. Rpl34 was shown to be significantly upregulated at total RNA level, whereas

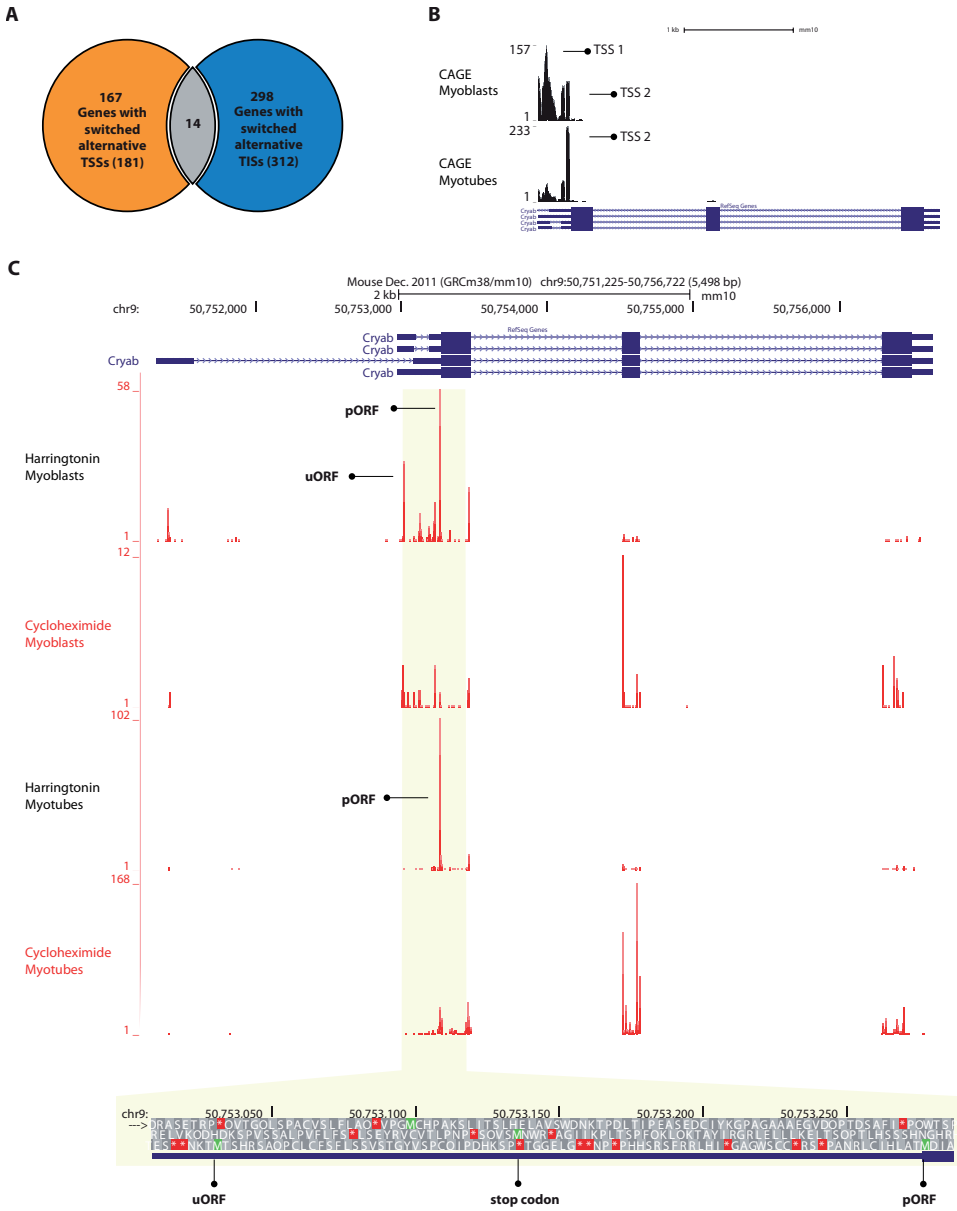


Figure 8. Switches in TISs usage during myogenesis. (A) Venn diagram showing the overlap between genes with significant switch in TSSs usage and genes with significant switch in TISs usage. The total number of genes with significant switches in TSS or TIS usage is reported between parentheses. **(B)** Coverage pattern and position of alternative TSSs used in Cryab gene, for myoblasts (top trace) and myotubes (bottom trace). TSSs are indicated with black arrows. **(C)** A screenshot of UCSC Genome browser displaying translation start sites (harringtonin, first and third traces) and ribosome footprints for the whole coding region (cycloheximide, second and fourth traces) in Cryab. TISs located in the annotated start codon (pORF) and in the 5'-UTR (uORF) are indicated with black arrows. Below the coverage tracks, the nucleotide sequence of the genomic region between the uORF and the pORF is shown. The stop codon in-frame with the uORF is indicated.

Gene symbol	Biological and/or clinical relevance	TIS position (5'-end)	Start codon	Length (aa)	uORF peptide sequence	Type of uORF
Cryab	AlphaB-crystallin modulates myogenesis by altering MyoD levels. CRYAB levels increase during differentiation, leading to an increase of MYOD levels. Loss of CryAB alters the capability of satellite cells to regenerate skeletal muscle.	chr9:50753019	AUG	35	MTSHRSAQPCLCFSFLSSVSTGY VSPCQIPDHKSP	Not overlapping
		chr9:50753228	UUG	47	LTSQPTLHSSSHNGHRHPPPLDPA PLLPLPLPKPPLRPVLRRAPVGV	Overlapping
Spp1	Osteopontin is a target of MyoD and Myf5 and a biomarker for Duchenne muscular dystrophy.	chr5:104435126	CAU	20	HPCLGLQSSAAGILGGNQPR	Not overlapping
		chr5:104435168	GGA	6	GGNQPR	Not overlapping
Eno3	The beta-subunit of the glycolytic enzyme enolase is upregulated at transcriptional level during differentiation of myoblasts. Mutations in this gene have been associated to muscle beta-enolase deficiency, which leads to glycogen storage disease. Eno3 is a biomarker for Duchenne muscular Dystrophy.	chr11:70657801	UCU	64	SSSLRDQLSTLAHSHLLWCSHGH AKNLRPGNPGLQGQPHGGGGPA HSQGSIPSSCAQWSFHGYL	Overlapping
Pgam	Phosphoglycerate mutase is regulated at transcriptional level during myogenesis and dysfunction of Pgam leads to metabolic myopathy.	chr19:41911995	UCG	21	SAILSCCCILCPSPWLPTSWC	Overlapping
Agl	Glycogen debranching enzyme is involved in glycogen storage disorders (Cori's and Lafora's disease).	chr3:116807384	GUU	19	VRILQKPKWNTVSRFEFYY	Overlapping
Tmbim6	The BAX inhibitor motif containing 6 gene modulates calcium homeostasis in the endoplasmatic reticulum.	chr15:99399869	CUG	10	LNRLWSHEYI	Overlapping
		chr15:99393038	UGU	4	CPVL	Not overlapping
Asb8	Ankyrin repeat and SOCS box gene 8 is expressed predominantly in skeletal muscle (i). A member of the same family (Asb15) regulates skeletal muscle growth by stimulating protein synthesis and regulating differentiation of muscle cells.	chr15:98145607	UUG	7	LEHVNTL	Not overlapping
Cs	Citrate synthase is a mitochondrial enzyme regulated during myogenesis, when mitochondrial content rapidly increases.	chr10:128337852	CUG	1	L	Not overlapping
Vim	Vimentin is expressed during the starting phase of differentiation and decreases during development progression in C2C12.	chr2:13574376	UUG	45	LQFFQPQQASPPSKPCLPGLCPR PPTAGCSVAPAHRAGPAPTGAM	Overlapping

Table 1. Candidate uORFs differentially used during myogenesis, in genes with biological and/or clinical relevance in muscle biology.

it was downregulated in harringtonin footprints (the downregulation did not reach significance in cycloheximide footprints). Rps15 was not differentially expressed at total RNA levels but, similarly to Rpl34, harringtonin footprints showed a downregulation in myotubes (which did not reach significance in cycloheximide footprints). For both RPL34 and RPS15 no significant change was detected at protein level with western blot analysis, neither the presence of truncated isoforms (Supplementary Figure S5).

We attempted to validate the presence of dually decoded regions, N-terminally truncated or extended isoforms and small ORFs derived from uORFs, in a genome wide scale, by screening publicly available raw LCMS/MS proteomic datasets, including two C2C12-specific datasets (77,78) and a HiRIEF LC-MS/MS deep proteome dataset from N2A mouse cell (79). None of the novel candidate peptides passed our stringent spectrum-level validation, consistent with an extremely low abundance of these peptides or detection of ribosome stalling (see Discussion section).

DISCUSSION

Gene expression programs control tissue development and regeneration. Whereas regulation of gene expression at transcriptional level is extensively studied at genome-wide level, control of mRNA translation has mostly been studied on individual genes. Polyribosomal mRNAs profiling has been used in the past to obtain a global overview of translation efficiency. However, the novel approach of ribosome footprint profiling enables translatome analysis at the same level as transcriptome analysis. Nonetheless, the existing protocol for ribosome profiling is laborious and, to date, there are no dedicated pipelines for the analysis of the short ribosome footprints.

Here we describe a simplified protocol for ribosome profiling and a novel data analysis pipeline, which includes a combined mapping procedure for short reads, the analysis of the triplet periodicity and a dynamic peak calling algorithm to detect annotated and/or novel TISs, including aORFs and uORFs in frame or out-of-frame compared to the annotated ORF. We used our simplified protocol and

custom pipeline to investigate the extent of translational control during the formation of mammalian skeletal muscles, based on the analysis of the translatome, promoterome and transcriptome of proliferating myoblasts and differentiated myotubes in the murine C2C12 cell model. We integrated ribosome profiling data, DeepCAGE data, RNAseq and miRNAseq data to assess the contribution of translational regulation to the changes in protein expression during myogenic differentiation.

Detection of TISs

To investigate the impact of alternative translation initiation, we used our custom dynamic peak calling algorithm to detect and quantify alternative TIS usage during differentiation in harringtonin treated myoblasts and myotubes.

Our algorithm detected 5,333 not yet annotated TISs, providing an extensive catalog of alternative TISs leading to uORFs, aORF and potentially dual coding regions, specifically used during myogenesis.

We report only a high confidence set of TISs. Not all peaks called from the harringtonin footprints may represent genuine TISs. False positive peaks may arise in the distal part of the coding regions, when the harringtonin treatment is too short for elongating ribosomes to finish the translation of the C-terminal part of the protein. For this reason we developed a dynamic peak calling algorithm which considers not only the triplet periodicity pattern, but also the coverage and the relative position of each candidate TIS.

Alternative TISs detected in the 5'-UTRs (corresponding to ~20% of mapped reads) showed a codon distribution similar to previously reported studies (24,25,55), with CUG and GUG codons being the most abundant non-AUG codons, whereas 50% of the footprints mapped to annotated start codons. Overall, these findings give confidence in our data. Likely, many more TIS are used during myogenesis, but they were not abundant enough to be detected in our experiments.

We detected a lower number of TISs in myotubes compared to myoblasts, which may relate to lower numbers of footprints prior to peak calling. Nevertheless, it does not exclude the possibility that the lower percentage of alternative TISs in myotubes reflects a true biological phenomenon, considering that differentiated cells become more specialized and therefore require a smaller protein repertoire.

We may also have lost alternative TIS due to our stringent thresholds: alternative TIS were only called when their abundancewas at least 10% of the full length isoform,where previous reports demonstrated that N-terminally truncated protein isoforms present at only 5% of the full length isoform can exert biologically significant effects (10,80). However, we preferred to not decrease this threshold and avoid false positives.

In our study ~30% of the reads mapped within the coding regions of pORFs. Only ~6% of the TISs located in coding regions were in-frame with their pORFs (~4% in case of TISs detected in myotubes), representing potential protein isoforms with truncated N-termini. We were not able to confirm the presence of alternative truncated protein isoforms for RPL34 and RPS15 at western blot level. An explanation could be pausing of ribosomes during the harringtonine treatment, or leaky scanning of the pORF TIS that results in recognition of a downstream start codon yielding to an instable alternative isoforms. The primary ribosomal proteins are stable and accumulate in the cell, whereas the isoform does not accumulate.Regulation of protein stability is another control mechanisms determining protein abundance, which cannot be addressed by ribosome profiling.

The remaining TIS located in the coding regions where outof- frame TISs. A portion of it may represent potential dual coding regions. Previous studies have detected dual coding regions in genes involved in fundamental cellular processes (21), such as translation (Eif4a2), cell cycle (Cdkn2a) and

protein degradation (Ube2e2). Many translation initiation factors, including Eif1, Eif4a2, Eif4e2, Eif4a1, Eif2s1 and Eif5 showed a switch in TIS usage during myoblasts differentiation.

We did not observe dual coding in Eif4a2 in our data, but we did detect two TISs in Eif1, one representing the annotated start codon and the other representing an out-of-frame aORF with an AUG start. Nevertheless, for the majority of the alternative out-of-frame TISs, we currently lack further evidence. Our attempt to validate dual coding regions, in-frame aORF and small peptides derived from uORFs, based on publicly available mass spectrometry data, present several limitations, even if the proteomic data used is of high quality and acquired using state-of-the art instrumentation and methodology. An untargeted proteomic approach is not ideal due to dynamic range limitations and difficulties in detecting and quantifying low-abundant proteins among a diverse pool (81). A recent study showed that ribosome profiling data could be used to improve identification of novel N-termini isoforms and translated upstream ORF from proteomic data (82). However, only a small number of translated uORFs and N-terminal extensions was validated. We therefore conclude that the lack of consistency between ribosome profiling data and mass spectrometry data does not invalidate our findings, but positive validation of these translated uORFs and aORFs on protein level may require enrichment of peptides by anti-peptide antibodies raised against a number of predicted and synthesized peptides.

In this study we restricted the detection of TISs in coding transcripts. Nevertheless, a percentage of footprints derived from non-coding transcripts. LincRNAs bound to ribosomes have been observed in previous ribosome profiling (24) and polysome profiling studies (83). Whether they lead to active translation is still debated, with some studies showing no coding potential (84,85) and others suggesting that translation occurs in portions of lincRNAs (61). The fragment-size of the protected footprint is one of the parameters commonly used to distinguish true ribosome footprints from RNA fragments derived from transcripts protected by other complexes that may co-sediment with ribosomes or fragments derived from stable RNA secondary structure. Our read-length distribution analysis showed that lincRNAs did not always display a preference for one specific read-length, as protein coding genes did, and for those which showed a preferential peak surrounding 30 nt, we did not observe characteristic signatures of translation, not even restricted to portions of the lincRNAs.

Cellular processes controlled by selective mRNA translation in myogenesis

During differentiation of myoblasts into multinucleated myotubes, protein synthesis generally correlated with mRNA levels for the majority of the genes. Genes with lower correlations are likely regulated at the level of mRNA translation. The latter were strongly enriched for genes encoding for ribosomal proteins, whereas a modest enrichment for genes involved in protein degradation, focal adhesions, regulation of actin cytoskeleton and calcium signaling was also observed. The ribosomal protein genes and the calcium signaling pathway were also enriched in the set of genes showing alternative TIS usage, but the enriched genes were different, indicating that these pathways are mainly regulated at translational level not only by different translation initiation but also through other mechanisms.

A previous study showed that the production of three ribosomal proteins (S16, L18 and L32) is regulated both at the level of transcription and translation during myoblast differentiation (86). The authors showed a decrease in transcription and a decrease in translation efficiency by measuring mRNA bound to polysomes. In line with their study, S16, L18 and L32 showed a significantly lower number of harringtonin footprints in myotubes, whereas the decrease in cycloheximide footprints did not reach statistical significance.

A general downregulation was observed for the majority of the ribosomal protein genes both at transcriptional and translational level. Despite a positive correlation between RNAseq and cycloheximide footprints (r = 0.79, Pearson correlation), we found a poor correlation between RNAseq data and harringtonin footprints (r = 0.27, Pearson correlation), much lower than the correlation for all genes (r = 0.65). This discrepancy observed may be explained in different ways, one of which could be ribosome stalling, a known limitation in ribosome profiling data (87). If elongating ribosomes are stalled, this may lead to accumulation of footprints, which might be detected as alternative TISs in harringtonin data. The same applies to cycloheximide footprints, where ribosomal pauses might interfere with a correct quantification of translation. However, even if the peaks and footprints do not always reflect the production of novel short peptides or protein isoforms, we observed significant changes in ribosome footprints at those sites during myogenesis. These changes are highly reproducible between replicates, they are cell specific and tightly controlled during differentiation and therefore they likely represent a regulatory mechanism with relevance for muscle differentiation.

The mechanisms regulating alternative TISs usage in myogenesis remain to be investigated. Previous studies have shown that proteins involved in the translation machinery are autoregulated (35) and their synthesis is mainly controlled at the level of translation (88). These mRNAs are mainly characterized by the presence of structural motifs, such as the 5' TOP. The mTOR signaling pathway is known to regulate translation of TOP mRNAs. Serum removal could represent a downregulating stimulus for the mTOR pathway, possibly leading to mTOR-pathway inactivation and mTOR-dependent translation repression. The protocol for C2C12 differentiation is based on serum reduction (from 10 to 2% FBS) but our data does not show evidence of amajor contribution of themTOR signaling pathway toward the control of TOP mRNAs translation during myogenic differentiation, as we do not observe any enrichment of transcripts bearing a 5' TOP and affected by a switch in IS usage. Other studies have previously shown that the inhibition of mTOR can have different outcomes, from a major effect to little or no effect on TOP mRNA translation, depending on the cellular context (88). The ribosomal protein genes and translation factors which showed a switch in IS usage did not contain a 5' TOP, therefore we suggest that a different mechanism is used.

Next to genes involved in the translational apparatus, we found that many of the genes showing a switch in TIS usage are known to play a role in muscle development, maintenance and regeneration. Cryab (63,64), Spp1 (66), Tmbim6 (71) and Cs (73) have been previously shown to be regulated at transcriptional level during myogenic differentiation. no3 (67,68), Pgam (69) and Agl (70) have been related to metabolic myopathies, whereas Eno3 (74) and Spp1 (75) have been recently reported as biomarkers formuscular dystrophies. We showed that these genes are regulated at translational level by switches of alternative TIS usage between uORFs and pORFs during differentiation. Due to themany regulatory potential of uORFs, a full understanding of the translational control of these genesmay be relevant for clinical purposes.

The contribution of mRNA translation in myogenesis

Even though we observed a general positive correlation between transcription and translation, suggesting that most of the regulation occurs at transcriptional level, we also observed a dampening effect of translational regulation. The causes of this dampening effect remain to be elucidated.

Translation can be regulated by many different mechanisms. Here we specifically focused on the alternative use of start codons. Our study showed that 312 genes were subjected to switches in alternative TIS usage during differentiation. Although we showed that the presence of a myotube-specific promoter in Tpm3 resulted in an alternative TIS, we found that the majority of the switches

detected at translational level was independent from transcription. Switches in TIS usage mostly occurred in genes with a single promoter, thus the transcription of genes from distinct promoters, and the translation initiation from distinct start codons, seem to be two complementary mechanisms to control gene and protein expression in myogenesis.

Moreover, we showed that alternative promoters may also lead to recognition of regulatory uORFs located in the 5'-UTR, as shown for Cryab. Therefore, alternative TSS can be used to regulate protein levels. Nevertheless the detection of alternative TSSs may be challenging when TSSs are characterized by a broad peak (59). The DeepCAGE technology and the subsequent clustering procedure may not have the resolution to identify SSs which are in close proximity, leading to the incorporation of alternative TSSs into one single TSS. Our analysis might therefore underestimate the number of alternative TSSs which are in very close proximity and therefore overestimate the number of switches in TIS usage exclusively dependent on the translational control. It remains to be investigated to which extent this phenomenon may alter our results.

A considerable amount of footprint mapped in the 5'-UTRs. Even though it is difficult to predict the effect of an uORF based on the length, many reports suggest that short uORFs are regulatory, whereas long uORFs and out-offrame uORFs overlapping the pORF primarily inhibit protein synthesis (89,90). We showed that the majority of the non-overlapping uORFs were between 1–30 amino acids long, whereas the majority of the overlapping uORFs were longer than 30 amino acids, suggesting a likely stronger regulatory potential.

We further investigated the contribution of miRNAs in the regulation of translation, focusing onwell-characterized myomiRs. For all experimentally validated targets we did not observe any major effect on translation inhibition. The amount of mRNAs targets present at transcriptome level and the amount of mRNAs targets translated reflected the general dampening effect observed for all other non-target genes, indicating that the myimiRs do not primarily affect the translational control of their target mRNAs.

In conclusion, our results demonstrate that translation initiation represent a layer of regulation of protein expression in myogenesis for specific subsets of functionally correlated genes.

ACCESSION NUMBER

European Nucleotide Archive PRJEB7207

ACKNOWLEDGEMENT

We thank the Leiden Genome Technology Center (LGTC) for providing RNAseq libraries and sequencing, Henk Buermans for the quantitative analysis of the miRNAseq data using E-miR software, LUMC's Sequencing Analysis Support Core (SASC) for support on data submission to the European Nucleotide Archive (ENA),Martijn Vermaat and Jeroen Laros for support on Mutalyzer, Pietro Spitali for discussions on biomarkers for Duchenne muscular dystrophy.

FUNDING

Landsteiner Foundation for Blood Transfusion Research (LSBR) (in part). Funding for open access charge: LUMC. Conflict of interest statement. None declared.

REFERENCES

1. Bentzinger,C.F., Wang,Y.X. and Rudnicki,M.A. (2012) Building muscle: molecular regulation of myogenesis. Cold Spring Harb. Perspect. Biol., 4.
2. Buckingham,M. and Rigby,P.W. (2014) Gene regulatory networks and transcriptional mechanisms that control myogenesis. Dev. Cell, 28, 225-238.
3. Moyes,C.D., Mathieu-Costello,O.A., Tsuchiya,N., Filburn,C. and Hansford,R.G. (1997) Mitochondrial biogenesis during cellular differentiation. Am. J. Physiol, 272, C1345-C1351.
4. David,R. (2012) Small RNAs: miRNAs' strict schedule. Nat. Rev. Genet., 13, 378.
5. Pimentel,J. and Boccaccio,G.L. (2014) Translation and silencing in RNA granules: a tale of sand grains. Front Mol. Neurosci., 7, 68.
6. Polesskaya,A., Cuvellier,S., Naguibneva,I., Duquet,A., Moss,E.G. and Harel-Bellan,A. (2007) Lin-28 binds IGF-2 mRNA and participates in skeletal myogenesis by increasing translation efficiency. Genes Dev., 21, 1125-1138.
7. Sterrenburg,E., Turk,R., 't Hoen,P.A., van Deutekom,J.C., Boer,J.M., van Ommen,G.J. and den Dunnen,J.T. (2004) Large-scale gene expression analysis of human skeletal myoblast differentiation. Neuromuscul. Disord., 14, 507-518.
8. Fritsch,C., Herrmann,A., Nothnagel,M., Szafranski,K., Huse,K., Schumann,F., Schreiber,S., Platzer,M., Krawczak,M., Hampe,J. et al. (2012) Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Res., 22, 2208-2218.
9. Calkhoven,C.F., Bouwman,P.R., Snippe,L. and Ab,G. (1994) Translation start site multiplicity of the CCAAT/enhancer binding protein alpha mRNA is dictated by a small 5' open reading frame. Nucleic Acids Res., 22, 5540-5547.
10. Calkhoven,C.F., Muller,C. and Leutz,A. (2000) Translational control of C/EBPalpha and C/EBPbeta isoform expression. Genes Dev., 14, 1920-1932.
11. Cao,J. and Geballe,A.P. (1995) Translational inhibition by a human cytomegalovirus upstream open reading frame despite inefficient utilization of its AUG codon. J. Virol., 69, 1030-1036.
12. Grant,C.M. and Hinnebusch,A.G. (1994) Effect of sequence context at stop codons on efficiency of reinitiation in GCN4 translational control. Mol. Cell Biol., 14, 606-618.
13. Hill,J.R. and Morris,D.R. (1993) Cell-specific translational regulation of S-adenosylmethionine decarboxylase mRNA. Dependence on translation and coding capacity of the cis-acting upstream open reading frame. J. Biol. Chem., 268, 726-731.
14. Landers,J.E., Cassel,S.L. and George,D.L. (1997) Translational enhancement of mdm2 oncogene expression in human tumor cells containing a stabilized wild-type p53 protein. Cancer Res., 57, 3562-3568.
15. Lincoln,A.J., Monczak,Y., Williams,S.C. and Johnson,P.F. (1998) Inhibition of CCAAT/enhancer-binding protein alpha and beta translation by upstream open reading frames. J. Biol. Chem., 273, 9552-9560.
16. Mize,G.J., Ruan,H., Low,J.J. and Morris,D.R. (1998) The inhibitory upstream open reading frame from mammalian S-adenosylmethionine decarboxylase mRNA has a strict sequence specificity in critical positions. J. Biol. Chem., 273, 32500-32505.
17. Raney,A., Baron,A.C., Mize,G.J., Law,G.L. and Morris,D.R. (2000) In vitro translation of the upstream open reading frame in the mammalian mRNA encoding S-adenosylmethionine decarboxylase. J. Biol. Chem., 275, 24444-24450.
18. Ruan,H., Shantz,L.M., Pegg,A.E. and Morris,D.R. (1996) The upstream open reading frame of the mRNA encoding S-adenosylmethionine decarboxylase is a polyamine-responsive translational control element. J. Biol. Chem., 271, 29576-29582.
19. Schleiss,M.R., Degnin,C.R. and Geballe,A.P. (1991) Translational control of human cytomegalovirus gp48 expression. J. Virol., 65, 6782-6789.
20. Vanderperre,B., Lucier,J.F., Bissonnette,C., Motard,J., Tremblay,G., Vanderperre,S., Wisztorski,M., Salzet,M., Boisvert,F.M. and Roucou,X. (2013) Direct detection of alternative open reading frames translation products in human significantly expands the proteome. PLoS. One., 8, e70698.
21. Michel,A.M., Choudhury,K.R., Firth,A.E., Ingolia,N.T., Atkins,J.F. and Baranov,P.V. (2012) Observation of

dually decoded regions of the human genome using ribosome profiling data. Genome Res., 22, 2219-2229.

22. Bazzini,A.A., Lee,M.T. and Giraldez,A.J. (2012) Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science, 336, 233-237.
23. Hsieh,A.C., Liu,Y., Edlind,M.P., Ingolia,N.T., Janes,M.R., Sher,A., Shi,E.Y., Stumpf,C.R., Christensen,C., Bonham,M.J. et al. (2012) The translational landscape of mTOR signalling steers cancer initiation and metastasis. Nature, 485, 55-61.
24. Ingolia,N.T., Lareau,L.F. and Weissman,J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell, 147, 789-802.
25. Ingolia,N.T., Brar,G.A., Rouskin,S., McGeachy,A.M. and Weissman,J.S. (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc., 7, 1534-1550.
26. Lee,S., Liu,B., Lee,S., Huang,S.X., Shen,B. and Qian,S.B. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc. Natl. Acad. Sci. U. S. A, 109, E2424-E2432.
27. Uemura,S., Aitken,C.E., Korlach,J., Flusberg,B.A., Turner,S.W. and Puglisi,J.D. (2010) Real-time tRNA transit on single translating ribosomes at codon resolution. Nature, 464, 1012-1017.
28. Ingolia,N.T., Ghaemmaghami,S., Newman,J.R. and Weissman,J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science, 324, 218-223.
29. Lundberg,E., Fagerberg,L., Klevebring,D., Matic,I., Geiger,T., Cox,J., Algenas,C., Lundeberg,J., Mann,M. and Uhlen,M. (2010) Defining the transcriptome and proteome in three functionally different human cell lines. Mol. Syst. Biol., 6, 450.
30. Maier,T., Guell,M. and Serrano,L. (2009) Correlation of mRNA and protein in complex biological samples. FEBS Lett., 583, 3966-3973.
31. Schwanhausser,B., Busse,D., Li,N., Dittmar,G., Schuchhardt,J., Wolf,J., Chen,W. and Selbach,M. (2011) Global quantification of mammalian gene expression control. Nature, 473, 337-342.
32. Tian,Q., Stepaniants,S.B., Mao,M., Weng,L., Feetham,M.C., Doyle,M.J., Yi,E.C., Dai,H., Thorsson,V., Eng,J. et al. (2004) Integrated genomic and proteomic analyses of gene expression in Mammalian cells. Mol. Cell Proteomics., 3, 960-969.
33. Vogel,C., Abreu,R.S., Ko,D., Le,S.Y., Shapiro,B.A., Burns,S.C., Sandhu,D., Boutz,D.R., Marcotte,E.M. and Penalva,L.O. (2010) Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol. Syst. Biol., 6, 400.
34. Li,J.J., Bickel,P.J. and Biggin,M.D. (2014) System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ., 2, e270.
35. Tebaldi,T., Re,A., Viero,G., Pegoretti,I., Passerini,A., Blanzieri,E. and Quattrone,A. (2012) Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells. BMC. Genomics, 13, 220.
36. Gonzalez,C., Sims,J.S., Hornstein,N., Mela,A., Garcia,F., Lei,L., Gass,D.A., Amendolara,B., Bruce,J.N., Canoll,P. et al. (2014) Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. J. Neurosci., 34, 10924-10936.
37. Oh,E., Becker,A.H., Sandikci,A., Huber,D., Chaba,R., Gloge,F., Nichols,R.J., Typas,A., Gross,C.A., Kramer,G. et al. (2011) Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. Cell, 147, 1295-1308.
38. Hestand,M.S., Klingenhoff,A., Scherf,M., Ariyurek,Y., Ramos,Y., van,W.W., Suzuki,M., Werner,T., van Ommen,G.J., den Dunnen,J.T. et al. (2010) Tissue-specific transcript annotation and expression profiling with complementary next-generation sequencing technologies. Nucleic Acids Res., 38, e165.
39. Parkhomchuk,D., Borodina,T., Amstislavskiy,V., Banaru,M., Hallen,L., Krobitsch,S., Lehrach,H. and Soldatov,A. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res., 37, e123.
40. Buermans,H.P., Ariyurek,Y., van,O.G., den Dunnen,J.T. and 't Hoen,P.A. (2010) New methods for next generation sequencing based microRNA expression profiling. BMC. Genomics, 11, 716.
41. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., 10, R25.

42. Langmead,B. (2010) Aligning short sequencing reads with Bowtie. Curr. Protoc. Bioinformatics., Chapter 11, Unit.

43. Wu,T.D. and Watanabe,C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics., 21, 1859-1875.

44. Hashimoto,T., de Hoon,M.J., Grimmond,S.M., Daub,C.O., Hayashizaki,Y. and Faulkner,G.J. (2009) Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite. Bioinformatics., 25, 2613-2614.

45. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods, 9, 357-359.

46. Wildeman,M., van,O.E., den Dunnen,J.T. and Taschner,P.E. (2008) Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum. Mutat., 29, 6-13.

47. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics., 26, 841-842.

48. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics., 26, 139-140.

49. Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal. Statistical. Society. Series. B (Methodological.), 57, 289-300.

50. Huang,d.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc., 4, 44-57.

51. de Klerk,E., Venema,A., Anvar,S.Y., Goeman,J.J., Hu,O., den Dunnen,J.T., van der Maarel,S.M., Raz,V. and 't Hoen,P.A. (2012) Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation. Nucleic Acids Res..

52. Grillo,G., Turi,A., Licciulli,F., Mignone,F., Liuni,S., Banfi,S., Gennarino,V.A., Horner,D.S., Pavesi,G., Picardi,E. et al. (2010) UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res., 38, D75-D80.

53. Deutsch,E.W., Mendoza,L., Shteynberg,D., Farrah,T., Lam,H., Tasman,N., Sun,Z., Nilsson,E., Pratt,B., Prazen,B. et al. (2010) A guided tour of the Trans-Proteomic Pipeline. Proteomics., 10, 1150-1159.

54. Guo,H., Ingolia,N.T., Weissman,J.S. and Bartel,D.P. (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature, 466, 835-840.

55. Ingolia,N.T. (2010) Genome-wide translational profiling by ribosome footprinting. Methods Enzymol., 470, 119-142.

56. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics., 25, 1105-1111.

57. Olshen,A.B., Hsieh,A.C., Stumpf,C.R., Olshen,R.A., Ruggero,D. and Taylor,B.S. (2013) Assessing gene-level translational control from ribosome profiling. Bioinformatics., 29, 2995-3002.

58. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol., 14, R36.

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

60. Puri,P.L., Wu,Z., Zhang,P., Wood,L.D., Bhakta,K.S., Han,J., Feramisco,J.R., Karin,M. and Wang,J.Y. (2000) Induction of terminal differentiation by constitutive activation of p38 MAP kinase in human rhabdomyosarcoma cells. Genes Dev., 14, 574-584.

61. Ingolia,N.T., Brar,G.A., Stern-Ginossar,N., Harris,M.S., Talhouarne,G.J., Jackson,S.E., Wills,M.R. and Weissman,J.S. (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep., 8, 1365-1379.

62. Pieples,K. and Wieczorek,D.F. (2000) Tropomyosin 3 increases striated muscle isoform diversity. Biochemistry, 39, 8291-8297.

63. Singh,B.N., Rao,K.S. and Rao,C. (2010) Ubiquitin-proteasome-mediated degradation and synthesis of MyoD is modulated by alphaB-crystallin, a small heat shock protein, during muscle differentiation. Biochim. Biophys. Acta, 1803, 288-299.

64. Neppl,R.L., Kataoka,M. and Wang,D.Z. (2014) Crystallin-alphaB regulates skeletal muscle homeostasis via modulation of argonaute2 activity. J. Biol. Chem., 289, 17240-17248.

65. Casadei,L., Vallorani,L., Gioacchini,A.M., Guescini,M., Burattini,S., D'Emilio,A., Biagiotti,L., Falcieri,E. and Stocchi,V. (2009) Proteomics-based investigation in C2C12 myoblast differentiation. Eur. J. Histochem., 53, 261-268.

66. Ishibashi,J., Perry,R.L., Asakura,A. and Rudnicki,M.A. (2005) MyoD induces myogenic differentiation through cooperation of its NH2- and COOH-terminal regions. J. Cell Biol., 171, 471-482.

67. Comi,G.P., Fortunato,F., Lucchiari,S., Bordoni,A., Prelle,A., Jann,S., Keller,A., Ciscato,P., Galbiati,S., Chiveri,L. et al. (2001) Beta-enolase deficiency, a new metabolic myopathy of distal glycolysis. Ann. Neurol., 50, 202-207.

68. Lamande,N., Brosset,S., Lucas,M., Keller,A., Rouzeau,J.D., Johnson,T.R., Gros,F., Ilan,J. and Lazar,M. (1995) Transcriptional up-regulation of the mouse gene for the muscle-specific subunit of enolase during terminal differentiation of myogenic cells. Mol. Reprod. Dev., 41, 306-313.

69. Castella-Escola,J., Urena,J., Alterio,J., Carreras,J., Martelly,I. and Climent,F. (1990) Expression of phosphoglycerate mutase mRNA in differentiating rat satellite cell cultures. FEBS Lett., 268, 24-26.

70. Cheng,A., Zhang,M., Gentry,M.S., Worby,C.A., Dixon,J.E. and Saltiel,A.R. (2007) A role for AGL ubiquitination in the glycogen storage disorders of Lafora and Cori's disease. Genes Dev., 21, 2399-2409.

71. Bultynck,G., Kiviluoto,S., Henke,N., Ivanova,H., Schneider,L., Rybalchenko,V., Luyten,T., Nuyts,K., De,B.W., Bezprozvanny,I. et al. (2012) The C terminus of Bax inhibitor-1 forms a Ca2+-permeable channel pore. J. Biol. Chem., 287, 2544-2557.

72. Liu,Y., Li,J., Zhang,F., Qin,W., Yao,G., He,X., Xue,P., Ge,C., Wan,D. and Gu,J. (2003) Molecular cloning and characterization of the human ASB-8 gene encoding a novel member of ankyrin repeat and SOCS box containing protein family. Biochem. Biophys. Res. Commun., 300, 972-979.

73. Kraft,C.S., LeMoine,C.M., Lyons,C.N., Michaud,D., Mueller,C.R. and Moyes,C.D. (2006) Control of mitochondrial biogenesis during myogenesis. Am. J. Physiol Cell Physiol, 290, C1119-C1127.

74. Ayoglu,B., Chaouch,A., Lochmuller,H., Politano,L., Bertini,E., Spitali,P., Hiller,M., Niks,E.H., Gualandi,F., Ponten,F. et al. (2014) Affinity proteomics within rare diseases: a BIO-NMD study for blood biomarkers of muscular dystrophies. EMBO Mol. Med., 6, 918-936.

75. Piva,L., Gavassini,B.F., Bello,L., Fanin,M., Soraru,G., Barp,A., Ermani,M., Angelini,C., Hoffman,E.P. and Pegoraro,E. (2012) TGFBR2 but not SPP1 genotype modulates osteopontin expression in Duchenne muscular dystrophy muscle. J. Pathol., 228, 251-259.

76. Xu,C., Xu,W., Palmer,A.E. and Reed,J.C. (2008) BI-1 regulates endoplasmic reticulum Ca2+ homeostasis downstream of Bcl-2 family proteins. J. Biol. Chem., 283, 11477-11484.

77. Forterre,A., Jalabert,A., Berger,E., Baudet,M., Chikh,K., Errazuriz,E., De,L.J., Chanon,S., Weiss-Gayet,M., Hesse,A.M. et al. (2014) Proteomic analysis of C2C12 myoblast and myotube exosome-like vesicles: a new paradigm for myoblast-myotube cross talk? PLoS. One., 9, e84153.

78. Kristensen,A.R., Gsponer,J. and Foster,L.J. (2013) Protein synthesis rate is the predominant regulator of protein expression during differentiation. Mol. Syst. Biol., 9, 689.

79. Branca,R.M., Orre,L.M., Johansson,H.J., Granholm,V., Huss,M., Perez-Bercoff,A., Forshed,J., Kall,L. and Lehtio,J. (2014) HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nat. Methods, 11, 59-62.

80. Wethmar,K., Begay,V., Smink,J.J., Zaragoza,K., Wiesenthal,V., Dorken,B., Calkhoven,C.F. and Leutz,A. (2010) C/EBPbetaDeltauORF mice--a genetic model for uORF-mediated translational control in mammals. Genes Dev., 24, 15-20.

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

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

83. van,H.S., van,I.M., Jacobi,J., Boymans,S., Essers,P.B., de,B.E., Hao,W., MacInnes,A.W., Cuppen,E. and

Simonis,M. (2014) Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. Genome Biol., 15, R6.

84. Banfai,B., Jia,H., Khatun,J., Wood,E., Risk,B., Gundling,W.E., Jr., Kundaje,A., Gunawardena,H.P., Yu,Y., Xie,L. et al. (2012) Long noncoding RNAs are rarely translated in two human cell lines. Genome Res., 22, 1646-1657.
85. Guttman,M., Russell,P., Ingolia,N.T., Weissman,J.S. and Lander,E.S. (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell, 154, 240-251.
86. Agrawal,M.G. and Bowman,L.H. (1987) Transcriptional and translational regulation of ribosomal protein formation during mouse myoblast differentiation. J. Biol. Chem., 262, 4868-4875.
87. Ingolia,N.T. (2014) Ribosome profiling: new views of translation, from single codons to genome scale. Nat. Rev. Genet., 15, 205-213.
88. Meyuhas,O. (2000) Synthesis of the translational apparatus is regulated at the translational level. Eur. J. Biochem., 267, 6321-6330.
89. Kozak,M. (2001) Constraints on reinitiation of translation in mammals. Nucleic Acids Res., 29, 5226-5232.
90. Somers,J., Poyry,T. and Willis,A.E. (2013) A perspective on mammalian upstream open reading frame function. Int. J. Biochem. Cell Biol., 45, 1690-1700.

SUPPORTING INFORMATION

Supplementary Tables 1-22 and supplementary Text are available at NAR Online

Table S1. Triplet periodicity. Number, percentage and median of reads mapping to the first, second and third nucleotide of a codon, and percentage of reads mapping 12 nucleotides upstream of annotated translation start sites (TISs).

Table S2. List of mouse Refseq transcripts with TIS located within a splice junction or located less than 15 nt upstream or downstream an exon-exon junction. Distance relative to the 3' and 5' ends are reported.

Table S3. Alignment statistics. Number and percentage of reads mapped to the transcriptome reference or to the genome reference after transcriptome alignment. Number and percentage of reads mapped to the repeat mask.

Table S4. Number and percentages of ribosome profiling reads from harringtonin-treated C2C12 mapped to annotated biotypes, after genome alignment or combined alignment.

Table S5. Number and percentages of ribosome profiling reads from cycloheximide-treated C2C12 mapped to annotated biotypes, after genome alignment or combined alignment.

Table S6. Myogenic markers. Gene expression levels of Myog, Tnnc1, Myh7, Myf5 in RNAseq data and ribosome profiling data.

Table S7. Differentially expressed genes in ribosome profiling data (harringtonin, footprints of initiating ribosomes).

Table S8. Differentially expressed genes in ribosome profiling data (cycloheximide, footprints of elongating ribosomes).

Table S9. Differentially expressed genes in RNAseq data.

Table S10. Differentially expressed miRNAs in miRNAseq data.

Table S11. MyomiRs analysis. Estimated coefficients and confidence intervals for experimentally validated targets of nine myomiRs.

Table S12. KEGG pathway analysis on subsets of genes differentially regulated during transcription

and translation.

Table S13. List of TISs detected in myoblasts.

Table S14. List of TISs detected in myotubes.

Table S15. Codon distribution. Number of TISs and read counts per motif per category detected in myoblasts.

Table S16. Codon distribution. Number of TISs and read counts per motif per category detected in myotubes.

Table S17. Internal Ribosome Entry Sites. Predicted IRES in transcripts with TISs in their 5'-UTRs, for myoblasts (top list) and myotubes (bottom list).

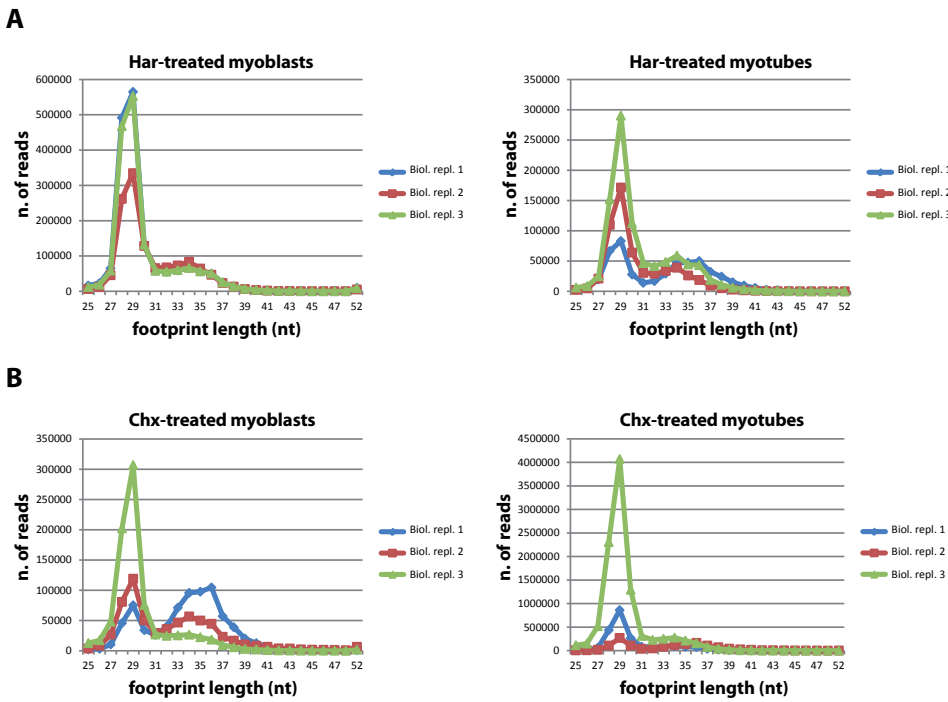
Table S18. Terminal Oligopyrimidine Tract. Predicted 5'TOPs in transcripts with TISs in their 5'-UTRs, for myoblasts (top list) and myotubes (bottom list).

Table S19. List of genes with alternative TIS usage during myogenesis

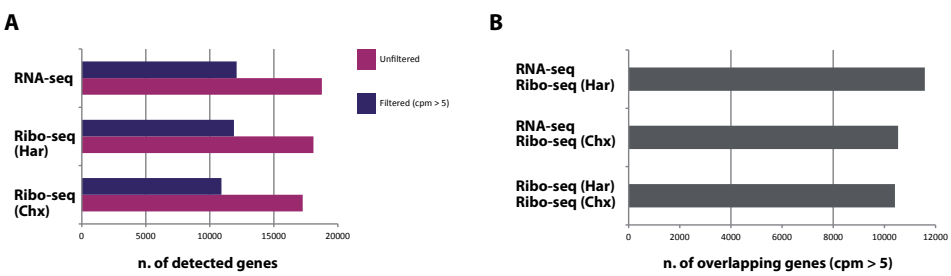
Table S20. List of genes with alternative TSS usage during myogenesis.

Table S21. List of genes with alternative TIS usage between myoblasts and myotubes and interaction p value of relative TIS usage.

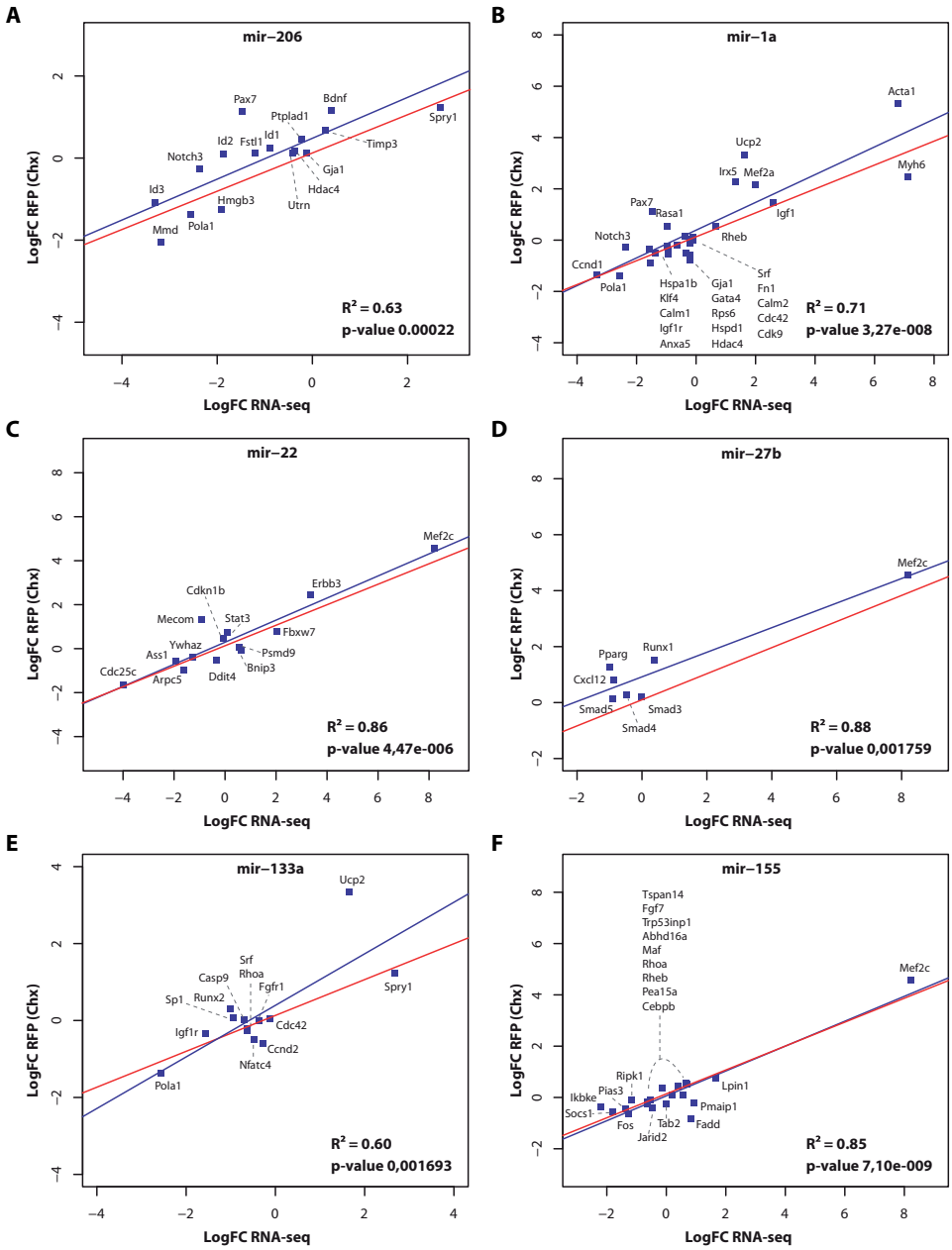
Table S22. KEGG pathway analysis on genes with changes in alternative TIS usage between myoblasts and myotubes.



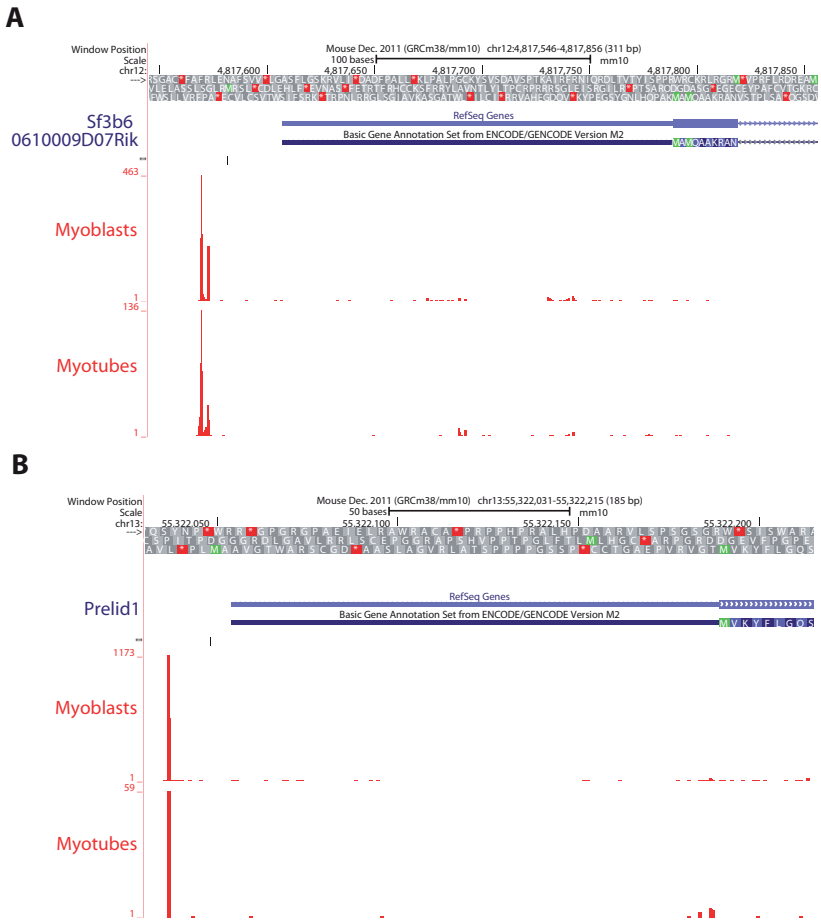
Supplementary Figure 1. Read length distribution of ribosome footprints.



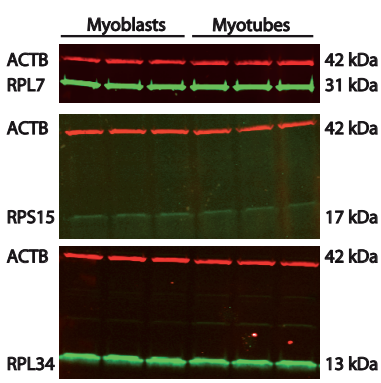
Supplementary Figure 2. Number of genes detected in RNAseq and Ribosome profiling datasets, before and after filtering step (cpm>5), and their overlap.



Supplementary Figure 3. Correlation between RNAseq and Ribosome profiling (CHX) data for experimentally validated myomiRs targets.



Supplementary Figure 4. TISs detected in (A) Sf3b6 and (B) Prelid1.



Supplementary Figure 5. Detection of RPL7, RPS15 and RPL34 by Western-blot analysis.

CHAPTER 5

FULL-LENGTH mRNA SEQUENCING UNCOVERS A WIDESPREAD COUPLING BETWEEN TRANSCRIPTION AND mRNA PROCESSING

Seyed Yahya Anvar, Eleonora de Klerk,
Martijn Vermaat, Johan T. den Dunnen, Stephen W. Turner,
Peter A.C. 't Hoen.

Manuscript Submitted. 2015.

ABSTRACT

Deciphering the interdependency between transcriptional and posttranscriptional regulatory events acting on the same RNA molecule is key in understanding the regulation of gene expression. The analysis of 7.4 million single-molecule long sequencing reads representing full-length mRNA molecules in MCF-7 human breast cancer cells provides the first comprehensive view of the degree of coordination between alternative transcription initiation, splicing and polyadenylation. In MCF-7 cells, an unforeseen amount of genes undergo vigorous and interdependent preferential selection during transcription and mRNA processing, which occur across the entire mRNA molecules. In particular, alternative polyadenylation sites that are coupled with alternative splicing events are depleted for known polyadenylation signals and enriched for MBNL binding motifs, supporting a dual role of MBNL proteins in regulating splicing and polyadenylation.

Our findings demonstrates that our understanding of transcriptome complexity is far from complete and provides a framework to reveal largely unresolved mechanisms that coordinate transcription and mRNA processing.

INTRODUCTION

The formation of a mature messenger RNA (mRNA) is a multi-step process. In higher eukaryotes, variations in each of these steps, e.g. selection of alternative transcription initiation, alternative exons, and alternative polyadenylation site, change the nature of the mature transcript. Tight regulation and coordination of these processes ensures the production of a set of cell-, tissue- and condition-specific transcript variants to meet variable cellular protein requirements (1-4). The co-transcriptional nature of mRNA processing suggests the presence of yet largely unresolved mechanisms that couple transcription with 5' end capping, splicing, and 3' end formation (reviewed in 5). Thus, resolving full transcript structures and accurate quantification of the abundance of alternative transcripts are important steps towards the delineation of these mechanisms.

RNA sequencing (RNA-Seq) has become a central technology for deciphering the global RNA expression patterns. However, reconstruction and expression level estimation of alternative transcripts using standard RNA-Seq experiments is limited and prone to error due to relatively short read length (typically up to 150 nucleotides) and required amplification steps of second-generation sequencing technologies (6, 7). It is apparent that single-molecule long reads that capture the entire RNA molecule can offer a better understanding of the rich patterns of alternative transcription and mRNA processing events and gene expression in human transcriptome and, hence, the underlying biology.

Despite a number of studies that have pursued long read sequencing to connect different exons or even capture entire transcripts with a rather limited sequencing depth (6, 8-14), the coupling between transcription and mRNA processing has not been extensively studied. Here, we investigate the global pattern of coupling between transcription, splicing and polyadenylation in MCF-7 human breast cancer cell line, which is deeply sequenced using the single-molecule real-time Pacific Biosciences RSII sequencing platform.

We show that transcription and mRNA processing are tightly coupled and that such interdependencies can be found across the entire RNA molecule and across large intra-molecular distances. We demonstrate that transcript identification and understanding of coupling between processes that are involved in the formation of these transcripts is far from complete, even in well-characterized human cell lines such as MCF-7. This study provides an in-depth view of the true complexity of the transcriptome and, for the first time, shows the tight and global interdependency between alternative transcription, splicing and polyadenylation.

RESULTS

Detection of transcript variants and the associated interdependencies between alternative exons

To investigate the genome-wide coupling of transcription and mRNA processing events, full-length mRNAs from MCF-7 human breast cancer cells were sequenced on 119 SMRT cells using Pacific Biosciences RSII platform (Supplementary Table 1). Prior to sequencing, parts of the sequencing library were size selected to allow for capturing rare and longer transcripts. The sequencing depth of our data, consisting of 7.4 million long reads, is equivalent to 70.3 million Illumina paired-end reads. Thus, this data enables reliable quantification of transcript abundance in MCF-7 transcriptome.

Transcript structures were defined by applying the isoform-level clustering algorithm (ICE) on full-length reads, capturing the entire mRNA molecule (containing both 5' and 3' primer sequences). Transcript sequences were further polished using both full-length and partial reads (**Figure 1A**). Our

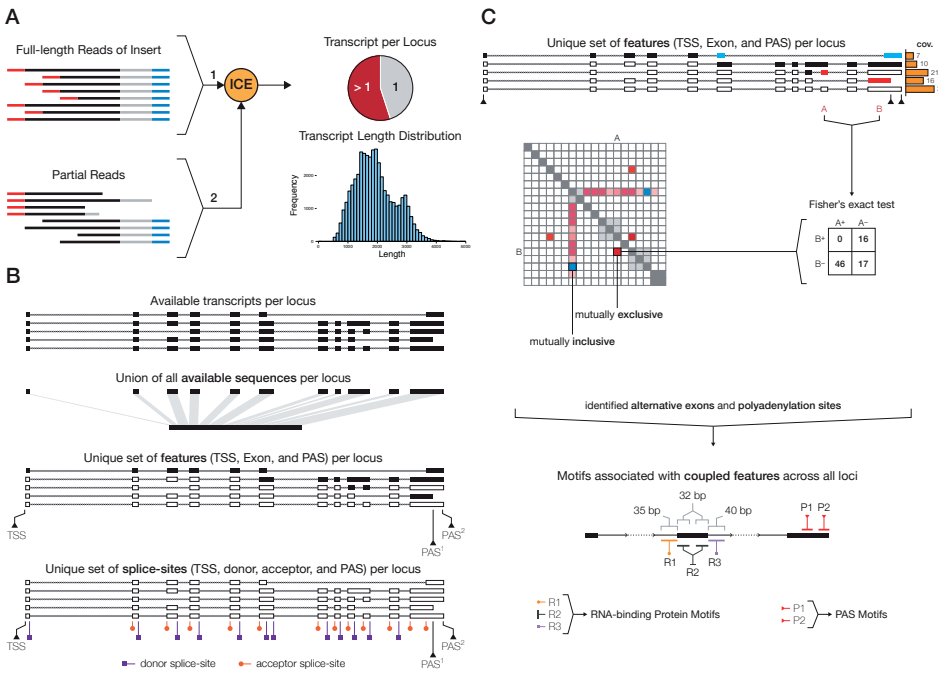


Figure 1. Schematic overview of the approach to characterize the interdependencies between mRNA transcription and processing events. A) Identified full-length reads of inserts are clustered into different transcript structures using the ICE algorithm and further polished using the partial reads. The number of unique transcript structures per locus and the distribution of transcript lengths are assessed. **B)** Based on the available transcripts per locus, the available sequence and unique set of features and splice-sites are identified. The available sequence is the union of all exonic sequences that are observed at each locus. Features are defined as a unique set of transcription start-sites (TSS), alternative exons, and alternative polyadenylation sites (PAS). The unique set of splice sites consists of unique donor and acceptor splice-sites as well as all alternative TSSs and PASs. **C)** The survey of coupling events is done by performing all possible pairwise tests between unique features of expressed and detected genes. The sum of the coverage of all transcripts that support the inclusion or exclusion of each pair is used in a contingency table to perform a Fisher's exact test for statistical significance. The odd ratio (OR) is used to differentiate between mutually inclusive (positive log-transformed OR) and exclusive (negative log-transformed OR) coupling events. Next, for all alternative exons that show significant linkage, a motif search is performed to assess the enrichment of specific RNA-binding protein motifs. For all alternative exons, the 35bp intronic sequences upstream of the acceptor site are defined as R1 domain (depicted in orange), the 32bp exonic sequences downstream of the acceptor site and upstream of the donor site are defined as R2 domain (depicted in dark grey), and 40bp intronic sequences downstream of the donor site are defined as R3 domain (depicted in purple). The 35bp sequence upstream of each PAS (depicted in red) is searched for the presence of canonical and non-canonical poly(A) signals.

analysis pipeline could precisely determine the position of polyadenylation sites (presence of poly(A) tail in the sequence) and intron-exon boundaries, as evident from the presence of the canonical GU motif in 94.9% of donor splice sites and the canonical AG motif in 96.6% of acceptor splice sites. From the 14,385 genes with detectable expression, 49% produced multiple transcript structures (Supplementary Figure 1). A total of 93 candidate fusion genes were identified based on the inter-chromosomal or distant intra-chromosomal split-alignment of transcripts to the human reference genome (Supplementary Table 2). In addition, 42% of identified transcripts in MCF-7 are potentially novel in comparison with the GENCODE annotation (Supplementary Table 3).

To detect and characterize the dependency between transcription and mRNA processing events, we designed the following analysis strategy (**Figure 1**). For each gene, the union of all exonic sequences

was considered as the available sequence and the union of all unique transcription start sites (TSSs), exons (defined as having distinct donor and acceptor splice sites), and polyadenylation sites (PASs) was used as a set of available features (**Figure 1B**). Mutual inclusivity or exclusivity of all possible pairs of features was assessed based on the number of reads that support the inclusion or exclusion of each pair of features. Subsequently, we applied a Fisher's exact test to evaluate statistical significance of the interdependency between a pair of features (**Figure 1C**; also see Methods).

General properties of coupling in human MCF-7 transcriptome

The MCF-7 transcriptome data consist of 14,385 genes containing 1,724,400 combinations of features (TSSs, exons, and PASs). The majority of combinations represent exon-exon pairs as many loci contain only a single TSS or PAS whereas most loci are multi-exonic (Supplementary Figure 2). Since the test is only applicable to genes with multiple transcripts, only 7,008 genes and 1,090,077 pairs of features (TSSs, exons, and PASs) were included in the statistical evaluation. Twenty percent of all feature pairs were significantly coupled (p-value < 4.6e-08, after Bonferroni correction for multiple testing). Generally, we observed large effect sizes for coupled features with the majority (65%) to be mutually inclusive, meaning that features were predominantly present in the same transcripts (Supplementary Figure 3,4). Remarkably, we observed coupling between mRNA features in nearly half of all genes that were evaluated (3,426 out of 7,008; **Figure 2A**; Supplementary Figure 5). We found a substantial amount of interdependencies between all types of features (**Figure 2B**). Of the 3,426 genes with at least one coupling event, 1,212 (35%) showed interdependencies between all classes of features: alternative TSS linked to alternative exons, alternative exon to alternative exon linkage, alternative PAS linked to alternative exons, and alternative TSSs to alternative PASs. Thus, the deep sequencing of full-length mRNAs provided a first image of the large degree of coordination in the usage of alternative TSSs, exons and PASs, mostly restricting the number of produced transcripts given the substantial amount of combinatorial possibilities.

Only 18% of the significantly coupled transcription and mRNA processing features were cataloged in the Ensembl Alternative Splicing Events set, version 75 (**Figure 2C**). These features were almost uniformly distributed across the different categories of alternative transcription or mRNA processing events. The majority of features (75%) that could not be attributed to any of the known categories represented interdependencies between alternatively spliced exons.

The length of individual transcripts was not associated with the likelihood of a significant coupling event in that transcript (Supplementary Figure 6). However, significant coupling events were enriched in genes with larger exonic sequence lengths (**Figure 2D,E**), giving rise to a larger repertoire of possible transcripts and requiring more extensive regulation of the synthesis for transcripts containing different subset of features.

We also examined the effect of the relative position in the gene and the distance between features on the observed degree of coupling. As expected, most TSSs were located at the most 5'-end of genes. The TSSs coupled or not coupled to alternative mRNA processing events showed a similar distribution over the gene (**Figure 3A** and Supplementary Figure 7). Interdependence between alternative TSSs was observed across the entire gene (**Figure 3B;** Supplementary Figure 8). However, alternative TSSs were preferentially coupled to alternative splicing events in relatively close proximity to the TSSs, near the 5'-end **(Figure 3B**). Nevertheless, examples of the coupling of alternative TSS and alternative exon usage across large distances, and spanning multiple exons were also frequently observed (**Figure 3C;** ITGB4). More evidence for interactions across the entire length of genes comes from the significant coupling between TSS and PAS (**Figure 3B,C**; NCAPD2).

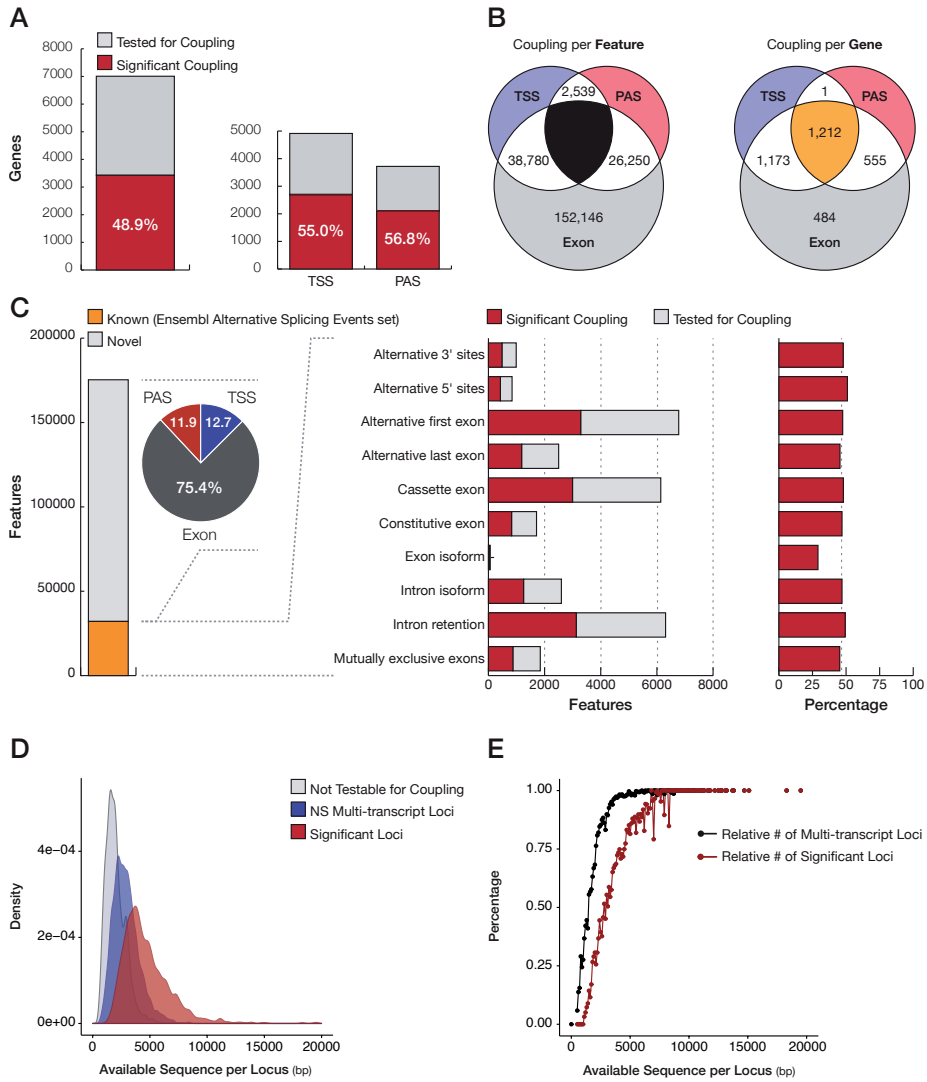


Figure 2. Alternative transcription, splicing, and polyadenylation is highly interdependent. A) The bar and pie charts illustrate the number and proportion of genes that show significant coupling. **B)** Venn diagrams show the number of coupled features based on the type of processes that are tested as well as the number of genes that show significant coupling between different processes. **C)** The number and proportion of known alternative features (TSS, exon, PAS) that are located in genomic regions, associated with Ensembl annotated alternative splicing events. **D)** The distribution of the length of available sequences per locus for loci with only one transcript (not tested; grey), multi-transcript loci with no significant coupling (blue), and all loci that show at least one significant coupling event (red). **E)** The relative number of loci with multiple transcripts (black) and the relative number of multi-transcript loci with significant coupling were plotted against the length of the available sequence. 100bp bins were used to group examined loci by length.

Similarly, coupling events linked to alternative PAS usage were found across the entire gene (Supplementary Figure 8; **Figure 3D**). In concordance with published literature (15-17), alternative PAS usage was preferentially coupled to nearby alternative exons (**Figure 3E**). Nevertheless, a substantial

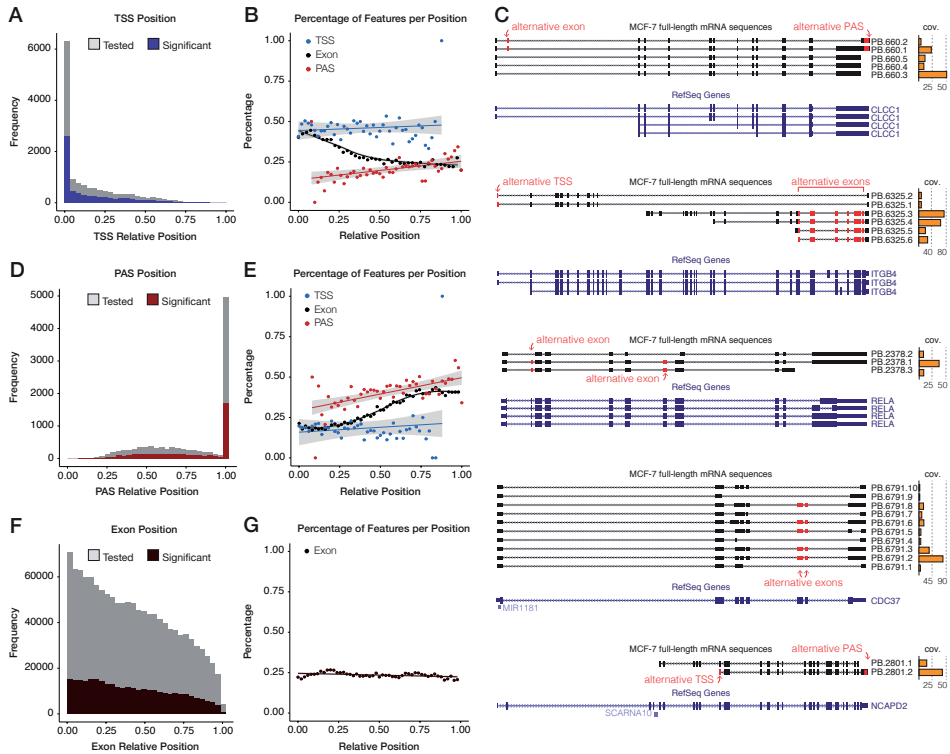


Figure 3. The interdependence of transcription and mRNA processing events can range across large distances. **A)** Histogram of the relative positions of transcription start sites (TSS) with (blue) and without (grey) significant coupling to mRNA processing events. Relative positions are calculated based on the length of the total exonic sequence at each locus. **B)** The fraction of significantly coupled TSSs (blue) to alternative exons (black) and PASs (red) is plotted at each relative position. **C)** Examples of genes that show evidence of coupling between transcription and mRNA processing events across the entire length of the gene. CLCC1 gene shows an example of long-range mutual inclusivity of the alternative second exon and the usage of the most distal PAS (depicted in red). ITGB4 gene presents an example of coupling between TSS and mutually exclusion of a cassette of alternative exons. RELA transcripts support the mutual inclusion of non-consecutive alternative exons. CDC37 gene provides evidence of mutual inclusion of consecutive exons during alternative splicing. NCAPD2 gene shows an example of coupling transcription with alternative polyadenylation. The number of supporting reads for each transcript is shown in orange bars. RefSeq annotated transcript structures are presented. **D)** Histogram of the relative positions of polyadenylation sites (PAS) with (red) and without (grey) significant coupling to alternative transcription and splicing events. **E)** The fraction of significantly coupled PASs (red) to alternative TSSs (blue) and exons (black) is plotted at each relative position. **F)** Histogram of the relative positions of alternative exons with (brown) and without (grey) significant coupling to other exons. **G)** The fraction of significantly coupled exons to other exons is plotted at each relative position. For plots depicting the percentage of linked features per position, the bin size of 0.02 was used.

proportion of PASs was coupled to alternative exons in more 5'-regions of genes (see CLCC1, **Figure 3C**). The proportion of alternative exons was higher at the 5'-end; however, dependencies between multiple alternative splicing events were uniformly observed across the entire gene (**Figure 3F, 3G**). In spite of this uniform distribution of exon-exon coupling events and the presence of distant coupling events (**Figure 3C**, RELA), the majority of independent alternative splicing events was between nearby or neighboring exons (Supplementary Figure 9; **Figure 3C,** CDC37).

We performed pathway enrichment analysis to analyze whether the coupling between alternative

transcription and mRNA processing events was associated with the molecular function of the proteins encoded by the transcripts. A number of pathways associated with mRNA processing and protein degradation such as spliceosome, proteasome, and ubiquitin mediated proteolysis, were enriched in transcripts demonstrating significant coupling (Supplementary Table 4).

Poly(A) signal usage for coupled polyadenylation sites

The majority of the alternative PASs in MCF-7 cells was found in different exons. From 3,719 genes that contain alternative PASs, we identified only 200 tandem PASs in the same 3' UTR. From these, only 56 loci (28%) included PASs that were significantly coupled with alternative exons. The low number of tandem 3' UTRs, in both coupled and uncoupled PAS-exon pairs (3.2% and 2.8%, respectively), has been previously explained by a general shortening of 3' UTRs in MCF-7 cell line (18). Thus, the majority of coupling events between alternative PASs and inclusion or exclusion of alternative exons are due to the use of exonic and intronic PASs, leading to the formation of new 3' UTRs.

To assess whether certain poly(A) signals are preferentially associated with alternative transcription and splicing, we searched for canonical (AATAAA and ATTAAA) and eleven known non-canonical poly(A) signals in the 35bp sequences upstream of the identified PASs. Canonical poly(A) signals could be found in the 35bp sequences upstream of 51.5% of all PASs (**Figure 4A**; Supplementary Figure 10, 11). This percentage is lower than what is generally reported and is most likely due to a global shortening of the 3' UTRs in MCF-7 cell line (18). Interestingly, the proportion of PASs that could be associated with canonical poly(A) signals was significantly lower (40.7%) for those that were coupled with TSSs or alternative exons. PASs that were linked with TSSs showed an even lower proportion of canonical poly(A) signals (34.7%). This decrease was not accompanied by an increase in known non-canonical poly(A) signals, but was mainly due to the use of PASs for which no known poly(A) signal could be identified (**Figure 4B**). This suggests that a novel poly(A) signal and alternative mechanisms may be involved in transcription- or splicing-coupled polyadenylation in MCF-7 cells. Thus, we screened for enriched motifs in the 35bp sequences upstream of PASs that are not associated with known poly(A) signals. While we did not observe any enrichment for PASs that were coupled with alternative TSSs, the ones that were coupled with alternative splicing were enriched for ASCCTG and GYGACA motifs. Interestingly, the ASCCTG motif could be associated with the binding site of muscleblind-like (MBNL) protein family, known to play a dual role in the regulation of splicing and polyadenylation (19, 20). Each MBNL isoform can bind to slightly different motifs (20) and a few motifs have been associated with MBNL proteins (20-22). Although all three MBNL proteins are expressed in MCF-7 cells, the enrichment of de novo identified ASCCTG and the recently reported CWGCMWKS motifs that can be recognized by MBNL3 protein20 were more prominent. Notably, previously identified binding motifs for MBNL1 (CTSCYB21 and RSCWTGSK20) and MBNL2 (TGCYTSYY20) were also enriched in sequences upstream of the PASs without a known poly(A) signal (**Table 1**). However, these motifs were not found to be associated with PASs that were coupled with alternative TSSs or alternative exons. Together, these results support an important role of MBNL proteins in the coupling between alternative polyadenylation and alternative splicing.

Identification of binding motifs for RNA-binding proteins potentially involved in coupling

We investigated the potential involvement of RNA-binding proteins (RBPs) in the coordination of alternative transcription and mRNA processing events by enrichment analysis of their binding motifs in

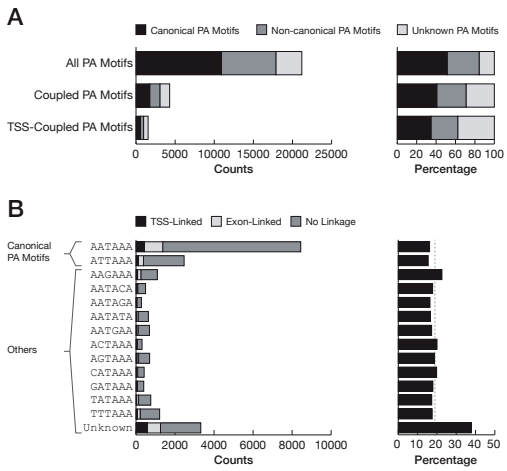


Figure 4. Alternative transcription start sites and exons are significantly associated with non-canonical poly(A) signals. A) Bar charts of the number and relative proportion of PASs that are associated with canonical or non-canonical poly(A) signals for all PASs, PAS with significant coupling, and TSS-linked PASs. **B)** The number and relative proportion of poly(A) signals for TSS-linked, exon-linked, or non-significantly coupled PASs.

Motif s	Source	Total	Random Set	P-value *	Coupled PAS	Not Coupled PAS	P-value §
ASCCTG	DREME	2232	102	0	710 (26.0%)	1522 (20.5%)	4.6E-09
CTSCYB	Masuda, 2012 [21]	970	650	5.8E-17	237 (8.7%)	733 (9.9%)	9.7E-01
YGCY	Purcell, 2012 [22]	2499	3485	1.0E-00	611 (22.3%)	1888 (25.5%)	9.9E-01
RSCWTGSK	Batra, 2014 [20]– MBNL1	195	99	9.1E-09	59 (2.2%)	136 (1.8%)	1.7E-01
TGCYTSYY	Batra, 2014 [20]– MBNL2	92	51	3.7E-04	16 (0.6%)	76 (1.0%)	9.9E-01
CWGCMWKS	Batra, 2014 [20]– MBNL3	1894	129	0	590 (21.6%)	1304 (17.6%)	3.9E-06
	Total PASs	10146	10146	–	2736	7410	–

* The enrichment of binding motifs in sequences upstream of PASs without a known poly(A) signal were calculated by Fisher's exact test (one-sided). A randomly generated set was used as a background for enrichment analysis.

§ PASs without significant coupling were used as the background set to identify a binding site that is enriched in the coupled PAS s without a known poly(A) signal.

Table 1. Enrichment of MBNL binding site motifs in sequences upstream of alternative PAS with unknown poly(A) signal that are coupled with alternative TSS or alternative exons.

coupled versus non-coupled exons. We screened three genomic regions relative to donor and acceptor splice-sites of coupled exons for enriched sequence motifs (**Figure 1C**; also see Methods): the 35bp intronic sequences upstream of the acceptor site (R1), the 32bp exonic sequences downstream of the acceptor sites and upstream of the donor sites (R2), and the 40bp intronic sequences downstream of the donor sites (R3).

For exons linked to alternative TSSs, the sequences from the R1 and R3 domains (upstream of the acceptor and downstream of the donor splice-sites, respectively) were both enriched for motifs (**Table 2**) that can be recognized by the splicing modulator RBM14 protein, known to play a dual role in regulating transcription and splicing (23, 24). In addition, the R2 sequences were enriched for binding sites for PPRC1, RBM8A, and TRA2 proteins. RBM8A has been shown to couple pre- and post-mRNA splicing events (25, 26) and TRA2 is also associated with regulating pre-mRNA splicing (27, 28). R3 regions immediately downstream of exons that are linked to alternative PASs were enriched for an A-rich motif, which can be recognized by a number of poly(A) binding proteins (**Table 2**), suggesting a competitive binding to these R3 sequences and genuine poly(A) tails.

Discussion

Short-read RNA sequencing has become central in assessing the global RNA expression patterns. However, as a result of the complexity of human transcriptome, these approaches disappoint in precise reconstruction and reliable expression estimation of transcript variants (6, 7, 29), owing to the short length of sequencing reads. In contrast, single-molecule long-read sequencing provides a unique opportunity to reveal the true complexity of the transcriptome as it can determine the full structure of individual transcripts by single-pass and full-length sequencing.

Here, we have analyzed the deepest and longest transcriptome data so far to better understand the extent of interdependencies between transcription and mRNA processing. Notably, full-length mRNA sequencing and de novo identification of high-quality sequence of transcript variants uncovered an unprecedented amount of potentially novel transcripts. The majority of alternative mRNA processing events could not be attributed to those that are cataloged in the latest Ensembl Alternative Splicing Events database. Our findings not only unravel a higher level of alternative transcription, splicing and polyadenylation in MCF-7 transcriptome than previously appreciated, but also provide valuable information on the preferential selection and interdependency between these processes.

We showed that transcription initiation, splicing and 3' end formation are tightly coupled in nearly 50% of genes with multiple transcripts and such interdependencies can be found across the entire length of the mRNA molecule. Notably, we report an unforeseen and unprecedented number of genes that undergo a vigorous preferential selection during transcription and mRNA processing as the choice of transcription initiation subsequently influences both alternative splicing of exons and the usage of alternative poly(A) site. These genes were enriched in mRNA processing and protein degradation pathway that may be in line with the previously observed auto-regulation of mRNA processing factors (30) and feedback loops between protein degradation and mRNA synthesis.

Ample evidence points at the critical role for RNA Pol II in the coordination between transcription and mRNA processing (reviewed in 5, 31-33). It has been shown that RNA Pol II initiation, pausing, and elongation rate can influence alternative splicing and polyadenylation of transcripts (34-37). Moreover, the C-terminal domain of RNA Pol II likely acts as a scaffold for regulatory factors that are involved in

TSS -coupled exons						
Domain	**Motif**	**Length Restriction**	**E-value**	**Fold Enrichment**	**RBP Motif**	**RBP**
R1	SGCGSGC	7 nt.	4.2E-02	1.44	RNCMPT00113	RBM14
R2	BCGCG	5 nt.	2.1E-02	1.18	RNCMPT00045	PPRC1
					RNCMPT00056	RBM8A
	GAWGARG	5 nt., 7 nt.	1.8E-02	1.16	RNCMPT00078	TRA2
R3	CGCSG	-	6.7E-09	1.35	RNCMPT00052	RBM14
Exon -Exon Coupling						
Domain	**Motif**	**Length Restriction**	**E-value**	**Fold Enrichment**	**RBP Motif**	**RBP**
R1	-	-	-	-	-	-
R2	RAAGAAG	7 nt.	1.8E-02	1.15	RNCMPT00078	TRA2
R3	-	-	-	-	-	-
PAS -coupled exons						
Domain	**Motif**	**Length Restriction**	**E-value**	**Fold Enrichment**	**RBP Motif**	**RBP**
R1	-	-	-	-	-	-
R2	-	-	-	-	-	-
R3	AAAARH	-	3.3E-56	1.33	RNCMPT00043	PABPC4
	AAAAAAVB	7 nt.	3.4E-55	1.64	RNCMPT00043	PABPC4

Table 2. The RNA-binding protein motifs associated with alternative exons that are coupled to TSS, other alternative exons, or PAS.

splicing and polyadenylation (reviewed in 33). Concordantly, we found an enrichment of coupling events in larger genes that seem to undergo a more extensive regulation during mRNA synthesis. However, the exact mechanisms by which the coordination is achieved remain largely unclear.

From previous studies it became clear that polyadenylation couples with splicing machinery to influence the removal or inclusion of the last intron (15, 38, 39). We now show that (i) the interdependencies between splicing and polyadenylation are not necessarily restricted to the final introns, (ii) that they can also involve introns that are far from the poly(A) site and (iii) that the coupling between splicing and alternative polyadenylation is not restricted to tandem 3' UTRs. The exact mechanisms by which these coupling events are achieved fall beyond the scope of this study. Previously, it has been shown that spliceosome components are also part of the human pre-mRNA 3'-end processing complex (40). Moreover, it is likely that there are RNA-binding proteins with a dual role in alternative splicing and polyadenylation in order to coordinate mRNA processing events. hnRNP H17, CstF6439, MBNL1 and ELAV1 (HuR) (19, 41-43) are a few examples of such proteins. We found MBNL binding motifs enriched in the sequences upstream of polyadenylation sites coupled with alternatively spliced exons. Interestingly, these regions often lacked canonical or non-canonical poly(A) signals. This suggests that MBNL proteins mark alternative poly(A) sites and play a dual and possibly coordinating role in splicing and polyadenylation. This is in line with previous studies in MBNL1-deficient cells where both splicing and polyadenylation were shown to be disrupted (19, 20).

Based on the reported sequence preference of MBNL proteins (20), MBNL3 is the most likely candidate of the MBNL family responsible for the coordination between alternative splicing and polyadenylation of transcripts in MCF-7 cells. However, it is not clear to what extent these findings can be extrapolated to other cell lines and cell types. In MCF-7 cells, the balance between alternative poly(A) site usage is shifted to more proximal poly(A) sites (18, 44). The absence of binding sites for regulatory proteins and miRNAs can enhance the tumorigenic activity of MCF-7 cells by allowing transcripts to escape from inhibition (18). Our findings mostly relate to the use of alternative polyadenylation by utilizing different 3' UTRs and not tandem polyadenylation sites that are in the same 3' UTR region. It is not clear whether MBNL-mediated polyadenylation, coupled with transcription initiation and splicing, is achieved through direct recruitment of RNA processing machinery or via alteration of secondary structure and formation of RNA molecules that, in turn, affect the choice for poly(A) site usage. Our analysis also identified a few more candidates with dual roles in mRNA processing, notably RBM14 (23, 24), RBM8A (25, 26, 45) and TRA22 (7, 28), which warrant further investigations by performing additional functional assays.

This study demonstrates that our understanding of transcript structures and coordinating mechanisms that regulate transcription and mRNA processing is far from complete, even in well-characterized human cell lines such as MCF-7. Single-molecule full-length RNA sequencing of other human tissues and cell-lines can provide a comprehensive view of the true complexity of the human transcriptome. Moreover, although it has been shown that single-nucleotide variants can alter the inclusion of exons in transcripts (9), it is of interest to identify variants that can affect allele-specific coupling between transcription and mRNA processing. Together, these can offer a better understanding of the mechanisms that control transcription and mRNA processing.

Methods

RNA sample preparation, library preparation, and sequencing

The methodologies and experimental settings for RNA preparation, cDNA synthesis, library preparation,

and sequencing are described at: http://blog.pacificbiosciences.com/2013/12/data-release-human-mcf-7-transcriptome.html.

Annotation of transcripts using isoform-level clustering algorithm (ICE)

The identification, polishing, and annotation of transcripts were previously carried out using the ICE algorithm and made public by Pacific Biosciences. To find transcript clusters, ICE performs a pairwise alignment and reiterative assignment of full-length reads to clusters based on likelihood. This process is followed by consensus calling and further polishing of the sequence to reduce the redundancy and increase the overall accuracy of sequences for identified transcript variants. For further information on the methodology and experimental settings visit: https://github.com/PacificBiosciences/cDNA_primer/wiki.

Comparison to the GENCODE annotation

We used GENCODE annotated transcripts (version 19) as reference to compare with the identified transcripts in the human MCF-7 transcriptome data. The comparison was carried out using cuffcompare from the Cufflinks suite (46).

Definition of transcription start site, polyadenylation site, and donor and acceptor splice sites

In this study, by processing the GFF file that contains the annotation of all identified transcripts and exon/intron boundaries (defined by the genomic position and strand on the hg19 reference sequence), a list of all transcription and mRNA processing events is produced. Transcription start sites (TSSs) are defined as the first genomic position of each transcript structure. Polyadenylation sites (PASs) are defined as the last genomic position of each transcript. The most upstream and downstream genomic positions of exons were used to define donor and acceptor splice-sites, respectively. However, for the first exon only the donor site is described as the first position is defined as transcription start site. Likewise, the last exon does not contain a donor splice site as the position is defined as polyadenylation site. If multiple transcripts share the same feature, then only one copy is kept in the unique set of features at each locus. Furthermore, the union of all unique exons is defined as the available sequence at each locus. This is also illustrated in **Figure 1B**.

Alignment and quantification of supporting reads for each transcript

The number of reads aligned to each transcript was used as the supporting evidence for each transcript structure. To identify the number of supporting reads, the polished sequences of all unique transcripts were used as a reference for the unique alignment of raw reads using BLASR47. Other parameters were set default and according to the Pacific Biosciences guidelines.

Statistical analysis

After defining unique features (transcription start sites, exons, and polyadenylation sites) and identifying the number of supporting reads for transcripts at each locus, all possible pairwise comparisons between features were made. To do this, the sum of all reads that support the presence of the two selected features in all observed transcripts is reported in a two-by-two contingency table. The table describes the number of times two features are observed in the same transcript

or exclusively, as well as the sum of reads that are mapped to transcripts that do not support the presence of either features (**Figure 1C**). A significant linkage between two features is assessed using the Fisher's exact test. The mutual inclusivity or exclusivity of coupled features are defined using their log-transformed odd-ratio. All p-values are adjusted using Bonferroni multiple testing correction. Many aspects of this analysis were carried out in Python and R.

Pathway analysis

This analysis was performed on a subset of genes that contain at least one coupling event and separated based on the type of coupling between features: TSS-exon, TSS-PAS, exon-exon, and exon-PAS. A list of all genes that could be detected in this study and subsequently annotated using GENCODE v19 (10,673 ENSEMBL gene IDs) was used as a background. Prior to the analysis, official gene symbols were converted to DAVID IDs. KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis was performed using DAVID Functional Annotation Tool (48).

Annotation of alternative exons

The genomic region of significantly coupled transcription start sites, alternative exons, polyadenylation sites are compared to the Ensembl Alternative Splicing Events annotation to characterize regions that have already been associated with one of the following ten classes:

1) Cassette exon
2) Intron retention
3) Mutually exclusive exons
4) Constitutive exon
5) Exon isoform
6) Intron isoform
7) Alternative 3' site
8) Alternative 5' site
9) Alternative first exon
10) Alternative last exon

To assess the enrichment of different categories of alternative splicing events in the Ensembl annotation, all the transcription start sites, exons, and polyadenylation sites that are present in the MCF-7 transcriptome data were also attributed to this annotation to serve as a background quantification.

Sequence motif analysis relative to polyadenylation sites

For each detected locus, we reported the last nucleotide as polyadenylation site. Each genomic location was converted into a BED format. Strand specific genomic sequences located up to 35 nucleotides upstream each unique polyadenylation site were extracted, in a FASTA format, using UCSC Table Browser (GRCh37/hg19). FASTA files were parsed using a custom bash script to count the number of sequences containing specific 6-mer motifs: one of the two canonical polyadenylation signals AATAAA and ATTAAA, or one of the eleven non-canonical polyadenylation signals (AAGAAA, AATACA, AATAGA, AATATA, AATGAA, ACTAAA, AGTAAA, CATAAA, GATAAA, TATAAA, TTTAAA). Subsequently, the same 6-mer motifs were counted for each unique PAS significantly coupled to TSSs or exons and for each unique PAS that did not show a significant coupling.

For PASs that could not be attributed to known poly(A) signals, we ran DREME (49) (v. 4.9.1) to

identify enriched motifs. Randomly shuffled set of sequences was generated from the original sequences of the examined PASs and used as a background set. In addition, the sequences of known recognition motifs for MBNL proteins (20-22) were counted for each set using a custom script. Subsequently, the enrichment of each motif was assessed by Fisher's exact test.

Tandem 3' UTR analysis

This analysis was performed to identify loci that contain tandem 3' UTRs (loci that contain more multiple PASs located in the same last exon). Custom scripts were used to identify loci that contain at least two PASs that share the same coordinates of the last exon start. The number of loci with tandem 3' UTRs was calculated for those in which PAS was significantly coupled to alternative exons and for those that did not show any significant interdependencies between alternative exons and the PAS usage.

Sequence motif analysis relative to acceptor and donor sites

For each detected locus, we reported the first and last nucleotide of each exon as acceptor splice site and donor splice site, respectively. Each unique genomic position was converted into a BED format and the strand specific sequences of 2 nucleotides length were extracted using UCSC Table Browser (GRCh37/hg19) for both acceptor and donor splice sites. A custom bash script was used to count the number of dinucleotide sequences containing 'GT' and/or 'AG'.

RNA binding motif analysis

We used MEME suite tools to identify enriched sequence motifs present in exons significantly coupled with TSSs, PASs or other alternative exons. For each unique exon, three regions were considered: R1 (containing up to 35 nucleotides upstream the acceptor splice site), R2 (containing 32 nucleotides downstream the acceptor splice site and 32 nucleotides upstream the donor splice site), and R3 (containing up to 40 nucleotides downstream the donor splice site). R1, R2 and R3 regions were obtained by extracting strand specific FASTA sequences using UCSC Table Browser (GRCh37/hg19).

We locally ran DREME (49) (v. 4.9.1) for each region separately, and performed a motif search using a negative background (R1, R2 and R3 regions from exons that were not significantly coupled). We ran DREME in two modes, one without any limitation for the motifs' width, and one with limiting the search to a minimum width of 5 or 7 nucleotides. In each case, a maximum of 10 motifs with E-values < 0.05 was reported. The remaining parameters were kept as default. We then compared each motif found by DREME against the human RNA-binding motifs database CISBP-RNA using TOMTOM Motif Comparison tool (50). We ran the analysis by setting the Pearson correlation coefficient as comparison function and considered only matches with a minimum false discovery rate (q-values) < 0.05.

REFERENCES

1. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature genetics 40, 1413-1415 (2008).
2. Barash, Y. et al. Deciphering the splicing code. Nature 465, 53-59 (2010).
3. Wang, E.T. et al. Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470-476 (2008).
4. Auboeuf, D. et al. A subset of nuclear receptor coregulators act as coupling proteins during synthesis and maturation of RNA transcripts. Molecular and cellular biology 25, 5307-5316 (2005).
5. Bentley, D.L. Coupling mRNA processing with transcription in time and space. Nature reviews. Genetics 15, 163-175 (2014).
6. Tilgner, H. et al. Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. G3 3, 387-397 (2013).
7. Steijger, T. et al. Assessment of transcript reconstruction methods for RNA-seq. Nature methods 10, 1177-1184 (2013).
8. Koren, S. et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature biotechnology 30, 693-700 (2012).
9. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M.P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. Proceedings of the National Academy of Sciences of the United States of America 111, 9869-9874 (2014).
10. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. Nature biotechnology 31, 1009-1014 (2013).
11. Au, K.F. et al. Characterization of the human ESC transcriptome by hybrid sequencing. Proceedings of the National Academy of Sciences of the United States of America 110, E4821-4830 (2013).
12. Thomas, S., Underwood, J.G., Tseng, E., Holloway, A.K. & Bench To Basinet Cv, D.C.I.S. Long-read sequencing of chicken transcripts and identification of new transcript isoforms. PloS one 9, e94650 (2014).
13. Treutlein, B., Gokce, O., Quake, S.R. & Sudhof, T.C. Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing. Proceedings of the National Academy of Sciences of the United States of America 111, E1291-1299 (2014).
14. Schreiner, D. et al. Targeted Combinatorial Alternative Splicing Generates Brain Region-Specific Repertoires of Neurexins. Neuron (2014).
15. Berget, S.M. Exon recognition in vertebrate splicing. The Journal of biological chemistry 270, 2411-2414 (1995).
16. Martinson, H.G. An active role for splicing in 3'-end formation. Wiley interdisciplinary reviews. RNA 2, 459-470 (2011).
17. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods 7, 1009-1015 (2010).
18. Fu, Y. et al. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome Res 21, 741-747 (2011).
19. Wang, E.T. et al. Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. Cell 150, 710-724 (2012).
20. Batra, R. et al. Loss of MBNL Leads to Disruption of Developmentally Regulated Alternative Polyadenylation in RNA-Mediated Disease. Molecular cell (2014).
21. Masuda, A. et al. CUGBP1 and MBNL1 preferentially bind to 3' UTRs and facilitate mRNA decay. Scientific reports 2, 209 (2012).
22. Purcell, J., Oddo, J.C., Wang, E.T. & Berglund, J.A. Combinatorial mutagenesis of MBNL1 zinc fingers elucidates distinct classes of regulatory events. Molecular and cellular biology 32, 4155-4167 (2012).
23. Auboeuf, D. et al. CoAA, a nuclear receptor coactivator protein at the interface of transcriptional coactivation and RNA splicing. Molecular and cellular biology 24, 442-453 (2004).
24. Kang, Y.K. et al. Dual roles for coactivator activator and its counterbalancing isoform coactivator modulator

in human kidney cell tumorigenesis. Cancer research 68, 7887-7896 (2008).

25. Kataoka, N. et al. Specific Y14 domains mediate its nucleo-cytoplasmic shuttling and association with spliced mRNA. Scientific reports 1, 92 (2011).
26. Kataoka, N. et al. Pre-mRNA splicing imprints mRNA in the nucleus with a novel RNA-binding protein that persists in the cytoplasm. Molecular cell 6, 673-682 (2000).
27. Dauwalder, B., Amaya-Manzanares, F. & Mattox, W. A human homologue of the Drosophila sex determination factor transformer-2 has conserved splicing regulatory functions. Proceedings of the National Academy of Sciences of the United States of America 93, 9004-9009 (1996).
28. Tacke, R., Tohyama, M., Ogawa, S. & Manley, J.L. Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing. Cell 93, 139-148 (1998).
29. Engstrom, P.G. et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nature methods 10, 1185-1191 (2013).
30. Lewis, B.P., Green, R.E. & Brenner, S.E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proceedings of the National Academy of Sciences of the United States of America 100, 189-192 (2003).
31. Kornblihtt, A.R. et al. Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nature reviews. Molecular cell biology 14, 153-165 (2013).
32. Schor, I.E., Gomez Acuna, L.I. & Kornblihtt, A.R. Coupling between transcription and alternative splicing. Cancer treatment and research 158, 1-24 (2013).
33. Hsin, J.P. & Manley, J.L. The RNA polymerase II CTD coordinates transcription and RNA processing. Genes & development 26, 2119-2137 (2012).
34. Danko, C.G. et al. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. Molecular cell 50, 212-222 (2013).
35. de la Mata, M. et al. A slow RNA polymerase II affects alternative splicing in vivo. Molecular cell 12, 525-532 (2003).
36. Nogues, G., Kadener, S., Cramer, P., Bentley, D. & Kornblihtt, A.R. Transcriptional activators differ in their abilities to control alternative splicing. The Journal of biological chemistry 277, 43110-43114 (2002).
37. Pinto, P.A. et al. RNA polymerase II kinetics in polo polyadenylation signal selection. The EMBO journal 30, 2431-2444 (2011).
38. Cooke, C., Hans, H. & Alwine, J.C. Utilization of splicing elements and polyadenylation signal elements in the coupling of polyadenylation and last-intron removal. Molecular and cellular biology 19, 4971-4979 (1999).
39. Movassat, M., Crabb, T., Busch, A., Shi, Y. & Hertel, K. Coupling between alternative polyadenylation and alternative splicing is limited to terminal introns (560.2). The FASEB Journal 28 (2014).
40. Shi, Y. et al. Molecular architecture of the human pre-mRNA 3' processing complex. Molecular cell 33, 365-376 (2009).
41. Lebedeva, S. et al. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. Molecular cell 43, 340-352 (2011).
42. Mukherjee, N. et al. Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. Molecular cell 43, 327-339 (2011).
43. Barnhart, M.D., Moon, S.L., Emch, A.W., Wilusz, C.J. & Wilusz, J. Changes in cellular mRNA stability, splicing, and polyadenylation through HuR protein sequestration by a cytoplasmic RNA virus. Cell reports 5, 909-917 (2013).
44. Mayr, C. & Bartel, D.P. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 138, 673-684 (2009).
45. Ishigaki, Y. et al. Depletion of RNA-binding protein RBM8A (Y14) causes cell cycle deficiency and apoptosis in human cells. Experimental biology and medicine 238, 889-897 (2013).
46. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology 28, 511-515 (2010).
47. Chaisson, M.J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC bioinformatics 13, 238 (2012).

48. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4, 44-57 (2009).

49. Bailey, T.L. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics 27, 1653-1659 (2011).

50. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. & Noble, W.S. Quantifying similarity between motifs. Genome biology 8, R24 (2007).

SUPPORTING INFORMATION

Supplementary Figures and Tables are available upon request.

S**upplementary Text is available online at** http://nbviewer.ipython.org/urls/git.lumc.nl/mcf7/full-length-rna-coupling/raw/master/2015%20-%20Coupling%20between%20transcription%20and%20mRNA%20processing%20events.ipynb#section%200

CHAPTER 6

GENERAL DISCUSSION

(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.

1. Current limitations in the RNA-sequencing field

The expression of coding RNA molecules is a complex process regulated not only at transcriptional and post-transcriptional level, but also during and after translation. To fully characterize this process on a genome-wide scale and at a nucleotide level, numerous high-throughput RNA profiling sequencing methods have been developed (**Chapter 1, section 2**). The use of a combination of these approaches focusing at transcriptional, post-transcriptional and translational level is helping to comprehensively characterize gene expression regulation.

RNA-seq technologies are elucidating the mechanisms that expand the genome's coding capacity and are quickly redefining the concept of gene expression regulation. Although there is a continuing increase in the number of transcripts identified, and in the understanding of the molecular mechanisms that coordinate their formation during transcription and mRNA processing, we still face technical limitations due to the short read length of next-generation sequencing data and reliance on statistical and computational approaches to reconstruct transcript structure. This represents an obstacle when trying to link different events occurring in the same RNA molecule.

The determination of the actual structure of a transcript cannot be achieved without capturing different processing and regulatory events occurring in the same transcript. Capturing these events by combining different complementary methods comes with limitations, due to the uncertainty associated with transcript reconstruction. The only way to specifically determine the exact transcript structure for each detected RNA molecule is the sequencing of full-length RNAs.

From a technological point of view, it is already possible to sequence full-length cDNA molecules on the PacBio RS sequencing platform (Pacific Bioscience). This option is currently becoming more feasible (Au et al., 2013;Sharon et al., 2013) and is opening a new era in the field of RNA-seq.

Full-length transcript sequencing helps defining any coupling between the different layers of regulation of gene expression (**Chapter 5**) and leads to a better understanding of the complexity of the transcriptome and its expression, even though future improvements in the production of cDNA molecules are still required to fully investigate the exact structure of each transcript variant. cDNA generation per se may preclude the determination of long transcripts, as only minor improvements in cDNA length have been observed in recent cDNA synthesis methods available, and the majority of the cDNA molecules produced reach a read length of ~2kb (**Chapter 1, section 2.3**). Improvements are also necessary in the PacBio RS sequencing platform, which current yield does not allow an accurate quantitative analysis of high and low abundant transcripts.

Direct use of RNA as a template for sequencing will further reduce biases introduced in the sample preparation procedure. Since a proof of principle for direct RNA sequencing on the PacBio RS platform has already been demonstrated (**Chapter 1, section 2**), it is expected that this option will become available in the near future.

2. Additional regulatory mechanism shaping gene and protein expression

The final outcome of gene expression cannot be fully characterized without considering the full set of regulatory mechanisms. Alternative transcription initiation (**Chapter 1, section 1.1; Chapter 4**), alternative splicing (**Chapter 1, section 1.2**), alternative polyadenylation (**Chapter 1, section 1.3; Chapter 2; Chapter 3**), and alternative translation initiation (**Chapter 1, section 1.4; Chapter 4**) represent only a portion of the known mechanisms which affect gene and protein expression in eukaryotes. Many more processes need to be considered when trying to elucidate the underlying regulatory mechanisms which determine protein levels, thus leading to specific phenotypes.

Regulatory mechanisms arising from transcription, RNA processing and translation

Regulation of gene expression starts at DNA level through epigenetic marks, such as DNA methylation and histone proteins modifications. Epigenetic marks shape the chromatin structure influencing its accessibility, leading to silencing or activation of specific DNA regions. Changes in the epigenome can be re-established, after clearance of the existing marks, or inherited. Inheritance can occur during mitosis, but also during meiosis, a phenomenon known as transgenerational epigenetic inheritance (Daxinger and Whitelaw, 2010). Some epigenetic marks can be influenced by the environment, therefore environmental event in one generation can affect the phenotype in subsequent generations.

Once a gene is transcribed, its structure can be influenced not only during the initiation of transcription (**Chapter 1, section 1.1**), but also during the elongation and termination processes. The speed of transcription elongation and termination can affect alternative splicing and polyadenylation (**Chapter 1, section 1.5**), with consequent impact on mRNA stability, localization and function.

Processed mRNAs are then transported to the cytoplasm, prior their translation. The processes of mRNA transport and mRNA localization can be tightly regulated to ensure when and where to translate an mRNA, a phenomenon called spatially controlled translation. This control is performed through the interaction with RNA binding proteins (RBPs), which localize the mRNAs but can also repress its translation in a reversible way (Rodriguez et al., 2008).

mRNA molecules are indeed never bared molecules, but molecules packed with RBPs to form messenger ribonucleoprotein (mRNP) complexes. Examples of mRNP complexes are the polysomes, the RNA particles and RNA granules, the stress granules, and the processing bodies (P-bodies). Whereas polysomes, in the majority of the cases, represent sites of active translation (with the exception of ribosome stalling events, see further), RNA particles and RNA granules represent two transport complexes which are sites of translation repression. mRNAs packed in these transport complexes are protected from degradation and temporary translationally repressed, to allow their transport in specific cellular regions and their local translation. The only difference between RNA particles and RNA granules is the absence or presence of ribosomes, respectively: RNA granules contain polysome-associated mRNAs whose translation is temporary repressed, whereas mRNAs contained in RNA particles are not yet engaged by the translational apparatus.

Stress granules represent also sites of temporary translation repression, with the exception that the mRNAs are not transported in different cellular regions, but are temporary protected from degradation during cellular stress. On the contrary, P-bodies are mainly defined as sites of degradation for translationally repressed mRNAs, even though some mRNAs can leave the P-bodies and re-associate with the translational apparatus.

The process of translation itself is controlled at multiple levels. Part of the regulation occurs during

the initiation process (**Chapter 1, section 1.4; Chapter 4**), and part occurs during the elongation, even though correlation between gene length and translation efficiency, or between codon usage and translation efficiency, remains a controversial subject. According to some studies (Ingolia et al., 2009;Ingolia et al., 2011), the speed of translation is independent of the length of the transcript, the abundance of the transcript and the codon usage, whereas others affirm that shorter genes are more efficiently translated (Arava et al., 2003) and that translation elongation speed seem to be affected by codons within the ORF, local mRNA folding, and amino acids charges. The latter leads to the theory that the speed of translation is not similar between transcripts (Dana and Tuller, 2012), and that codon usage is one of the causes leading to poor correlation between protein and mRNA levels (Olivares-Hernandez et al., 2011).

Pauses during elongation can also regulate synthesis, folding and localization of a protein (Darnell et al., 2011;Mariappan et al., 2010;Zhang et al., 2009). These pauses, known as ribosome stalling (**Chapter 4**), represent a mechanism which can regulate the speed of elongation in order to maintain protein homeostasis (Liu et al., 2013), and is a major component of the cellular stress response (Shalgi et al., 2013). Ribosome stalling can also lead to a complete block of translation, when ribosomes permanently stop moving during the elongation process, and eventually lead to degradation, an event which commonly occurs when polysomes associate with the MicroRNA-loaded RISC (miRISC) complex (Houseley and Tollervey, 2009).

In addition to the regulation of translation initiation and elongation, the genetic code can be read in alternative ways, leading to frameshifting, hopping, stop codon read-through and recoding (Atkins JF, 2010).

Frameshifting is caused by insertions or deletions in the coding region of a DNA sequence. When the number of nucleotides added or removed is not divisible by three, the reading frame is changed, leading to the translation of a complete different protein. This can lead to the premature inclusion of stop-codons, which will ultimately bring to degradation through NMD.

Many different human diseases are caused by indel mutations leading to frameshifting (Iannuzzi et al., 1991;Chung et al., 2011;Truong et al., 2010;Myerowitz, 1997). Interestingly, these alternative ways of translating an mRNA may also be used to restore protein translation. The codon read-through mechanism has been often used as therapeutic approach in diseases caused by premature termination codons, through the use of drugs that induce the ribosome to bypass the premature stop codon (Bidou et al., 2012).

The last regulatory control in the life of an mRNA is represented by degradation. mRNA degradation allows regulated turnover, and occurs when a mRNA is not needed in the cell anymore. Degradation also occurs if an mRNA is defective, such as misprocessed or misfolded. Defective mRNAs are recognized through a mechanism known as mRNA surveillance. Different mRNA surveillance pathways (Houseley and Tollervey, 2009) are known, as degradation of an mRNA can occur through endonucleases that cut the mRNA internally, or through exonucleases that degrade the mRNA from the 5’ end or the 3’ end.

The most observed degradation pathway is the nonsense mediated decay (NMD). The NMD is activated after the first round of translation and leads to the degradation of mRNAs containing premature stop codons, preventing the formation of truncated proteins (Kervestin and Jacobson, 2012). This mechanism is usually generated by defective alternative splicing, representing therefore a surveillance mechanism.

The coupling between alternative splicing and NMD is also used as an autoregulatory negative feedback loop by many splicing factors. Splicing factors can bind their own transcripts and appositely program a defective splicing, leading to the inclusion of alternative exons containing premature

stop codons. This autoregulatory negative feedback loop has been observed in many SR and hnRNP proteins (Lareau et al., 2007;Ni et al., 2007;Saltzman et al., 2008) as common self-limiting mechanism, through which splicing factors regulate its own splicing and production of its own protein.

These feedback loops can consist of complex interplays between different regulatory layers. An example is the autoregulation of the splicing factor TDP-43 (Avendano-Vazquez et al., 2012), which involves interplay between transcription, splicing and polyadenylation. In the presence of high levels of TDP-43, an alternative spliced and polyadenylated transcript is formed. The switch in splicing and APA pattern is autoregulated by the binding of the TDP-43 on its own 3'-UTR, and lead to the formation of a transcript which is retained in the nucleus, thus leading to a decrease of available protein. The control of gene and protein expression by negative feedback loops is observed not only for splicing factors, but also for translation factors (Betney et al., 2010;Betney et al., 2012). An example of such negative feedback is the autoregulatory repression of the eukaryotic translation initiation factor 1 (eIF1), upon its overexpression (Ivanov et al., 2010).

Regulatory mechanisms arising from changes in the nucleotide sequence of an mRNA

Next to regulatory mechanisms arising from transcription, RNA processing and translation, other regulatory mechanisms have been described, which are caused by post-transcriptional changes in the nucleotide sequence of the mRNA, which do not reflect changes at DNA level. To date, more than hundred different RNA chemical modifications have been reported (Machnicka et al., 2013), but the function of most of them remains unknown. Nonetheless, for some of them, fundamental biological aspects been discovered.

An example of chemical modification which is known to affect gene expression is RNA editing. The most common type of RNA editing involves deamination of adenosine (A) to create inosine (I) (Nishikura, 2010). The result is that splicing and translational machineries recognize inosine as guanosine. A-to-I RNA editing occurs mainly within Alu repetitive elements, or within introns and UTRs, whereas only a small percentage occurs in coding sequences (Park et al., 2012;Daniel et al., 2014;Levanon et al., 2005). Even though the frequency of an A-to-I editing event is low, the effects reported so far are numerous, from alteration of the amino acid sequence and RNA folding, through changes in the coding sequence of the translated exons, to alternative splicing (Farajollahi and Maas, 2010) through creation or disruption of splice sites.

Altered editing has been linked to human disorders, such as amyotrophic lateral sclerosis, epilepsy, and brain tumors (Maas et al., 2006;Paz et al., 2007;Kawahara et al., 2004).

The list of chemical modifications that regulate gene expression has been recently enlarged, after the discovery that methylation of internal adenosines (m^6A) (Jia et al., 2011), the most prevalent internal chemical modification of all higher eukaryotes, is a reversible mechanism, which resembles DNA methylation.

Similarly to DNA methylation, and unlike A-to-I RNA editing, m^6A does not alter the coding capacity of a transcript, therefore it does not lead to proteins with different amino acid sequences. Due to its reversible nature, m^6A might represent a novel fundamental mechanism controlling protein expression.

The effects of m^6A on biochemical, physiological and developmental processes are still poorly understood. mRNAs are methylated at internal adenosines by the methyltransferase complex (including METTL3, METTL14 (Liu et al., 2014) and WTAP (Ping et al., 2014)) and they are dynamically demethylated by two different enzymes, FTO (Jia et al., 2011) and ALKBH5 (Zheng et al., 2013). m^6A is the most common internal mRNA modification, affecting more than 7000 human genes (Dominissini

et al., 2012;Meyer et al., 2012), and it is conserved amongst eukaryotes, from yeast to humans (Rottman et al., 1976;Schwartz et al., 2013). Deletion, over-expression, or mutations in components of the methyltransferase complex or the demethylases appear to have dramatic effects in mouse and human, ranging from developmental defects, postnatal retardation, malformations to obesity (Boissel et al., 2009;Church et al., 2010;Dina et al., 2007;Fischer et al., 2009;Frayling et al., 2007;Rottman et al., 1976;Scuteri et al., 2007). However, a direct link of these diseases with RNA methylation still needs to be established.

Pioneering studies are suggesting broad biological roles at cellular level, including a possible interplay between RNA methylation and splicing (Dominissini et al., 2012), nuclear export (Fustin et al., 2013), and mRNA stability (Wang et al., 2014), with an emerging role for m^6A as negative regulator of gene expression. Whereas methylation at long internal exons seems to be associated with alternative splicing, methylation in the 3'-UTRs affects binding of the YTHDF2 and ELAV1 proteins (Dominissini et al., 2012), both influencing mRNA stability. YTHDF2 is able to partially re-localize its target mRNAs from translating ribosomes to cytoplasmic foci (P-bodies), with possible negative effect on gene expression (Wang et al., 2014).

We currently lack knowledge of the molecular mechanisms through which m^6A affects gene expression, and we do not understand why certain adenosines get methylated and others not.

3. Connecting fundamental research in the RNA field to clinical care

The recent findings in the RNA field and the understanding of alternative modes that regulate gene expression at transcriptional, post-transcriptional and translational level, represent a wealth of information useful to elucidate disease-related regulatory events and inspire new diagnostic and therapeutic approaches.

Currently, RNA-based analysis is being used in diagnostic mainly for gene expression-based patient stratification. Breast cancer arrays are an example of such application. An increase or decrease in mRNA levels could be caused by the presence of a variant which activates NMD, aberrant splicing, aberrant polyadenylation or aberrant translation. The gene expression-based patient stratification method currently used might be improved if the effect of a disease-causing variant is predicted, and the mechanism leading to disease is more specifically targeted and treated. The increased knowledge achieved to date allows more refined applications, both for diagnostic, prognostic and therapeutic purposes, which will lead towards personalized medicine.

This final section will discuss some of the applications and approaches currently in development. The first part will show an example of how alternative regulatory events could be used for diagnostic and prognostic purposes, whereas the second part will highlight how alternative regulatory mechanisms could be used as targets for personalized medicine.

Signatures from alternative regulatory events can be used as molecular biomarkers for diagnostic and prognostic purposes.

Currently, an example of such application is the use of APA profiles as potential molecular biomarker for cancer diagnostic. Widespread alteration of APA profile has been observed in many different cancer types, where shortening of 3'-UTRs has been linked to extensive upregulation and activation of oncogenes (**Chapter 1, section 1.3**). Lymphoma tumor subtypes with various survival characteristics can be distinguished based on their APA profile, even when the tumors are histologically identical (Singh et al., 2009). Prostate cancers can be stratified into subtypes with different risk of relapse based on APA profile (Li et al., 2014). APA profiles can also be used as molecular biomarker with prognostic potential for breast and lung cancer (Lembo et al., 2012) and to monitor progression of colorectal cancer (Morris et al., 2012). Shorter 3'-UTRs from specific mRNAs seem to correlate with tumor aggressiveness and poor prognosis in breast and lung cancer, therefore APA profile may be used to stratify patients in different risk classes (Lembo et al., 2012).

Nevertheless, the use of APA profile as potential molecular biomarker for cancer diagnostic, prognostic, and treatment comes with some limitations: APA profiles observed in cancer cell lines do not always overlap with what is observed in cancers from patients, suggesting that cancer cells might not be the best environment to study APA changes in cancer (Lembo et al., 2012); cancer cells do not always associated with 3'-UTRs shortening, but lengthening has also been observed, for example in MB231 breast cancer cell line (Fu et al., 2011), where APA profile is opposite to what is observed in MCF7 breast cancer cell line; 3'-UTR shortening is not a specific cancer signature.

Considering that transcriptome-wide alterations of APA profile have been observed in different contexts, both physiological (**Chapter 1, section 1.3**) and disease-related (**Chapter 2, Chapter 3**), it is essential to exclude possible alternative causes of APA before an APA-based diagnosis is established. Precautions need to be taken also when comparing APA profiles in the presence of an age-effect. Even though there are no studies describing widespread changes in APA during aging in human, and age

effect on the length of the 3'-UTRs has been observed in C. elegans, where the length of the 3'UTRs inversely correlates with the age of the animal (Mangone et al., 2010). The PABPN1 protein seems also to decrease during aging in human skeletal muscles (Anvar et al., 2013). This suggests a possible interplay between APA and aging, which need to be considered prior a APA-based diagnosis.

Alternative regulatory mechanisms can be used as targets for personalized medicine.
Many therapeutic approaches that entered clinical trials aim to control gene expression at the pre-mRNA level. These methods try to modulate mRNA production to interfere with processes leading to diseases.

Recent proof-of-concept studies have shown how artificial modulation of APA events can be used as therapeutic approach (**Figure 1a**). The choice for a specific polyadenylation site can be manipulated in order to (i) activate polyadenylation sites which are normally not used or (ii) inhibit correct polyadenylation, leading to degradation of the transcript variant.

The first case (a) has been applied to genes potentially coding for transcript variants whose localization strictly depends on the activation or suppression of intronic polyadenylation sites. Pre-mRNAs of different receptor tyrosine kinases and the vascular endothelial growth factor receptor 2 (VEGFR2) have been recently targeted with a novel antisense-based strategy, consisting in the inhibition of U1 small ribonucleoprotein particle, which normally suppresses intronic polyadenylation (Vorlova et al., 2011). Antisense oligonucleotides (AONs) are used to target the 5' splice site and inhibit binding of U1. In absence of splicing, intronic polyadenylation occurs, leading to the formation of transcript variants lacking trans-membrane domains. In the absence of these domains, the protein becomes anti-tumorigenic. In the second case (b), a method known as U1 small nuclear interference (U1i) is used. Different oncogenes have been targeted so far with this approach (pim-1 kinase, metabotropic glutamate receptor 1 and B-cell lymphoma 2), resulting in reduced tumor growth (Goraczniak et al., 2013;Weirauch et al., 2013). U1i makes use of artificial U1 adapters, consisting of oligonucleotides able to bind the terminal exon of a target pre-mRNA, and the U1 snRNA, recruiting the snRNP complex. The snRNP complex competes with the polyadenylation machinery, blocking correct polyadenylation, and leading to degradation of the pre-mRNA.

In cases where the disease is caused by erroneous activation of alternative polyadenylation sites, antisense-based strategies can be used to avoid the recognition of the alternative polyadenylation sites and reconstitute correct polyadenylation at the canonical polyadenylation site (Raz et al., 2014). This strategy may be used to target genomic variants that regulate gene expression levels by affecting the usage of alternative polyadenylation sites (**Chapter 3**). Variants localized within existing or newly created polyadenylation signals might influence the expression levels of single transcript variants leading to diseases such as islet autoimmunity in type I diabetes (Shin et al., 2007), mantle cell lymphoma (Wiestner et al., 2007), and systemic lupus erythematosus (Graham et al., 2007). In **Chapter 3**, novel causative SNPs affecting alternative polyadenylation by changes in the polyadenylation signal have been reported, seven of which have been also are reported in the GWAS catalog as associated with diseases. These loci might represent candidate therapeutic targets. In vitro studies on gastric cancer metastasis (Lai et al., 2015) have already shown that mRNAs with altered APA could represent novel targets for metastasis prevention.

These kind of targeted therapies are difficult to apply when APA changes occur transcriptome-wide. In **Chapter 2** we showed widespread 3'-UTR shortening in skeletal muscles of mice expressing a mutant form of the Poly(A) binding protein nuclear 1 (PABPN1), and proposed a novel role for the PABPN1 protein in poly(A) site selection. Due to the widespread effects, a therapeutic alternative

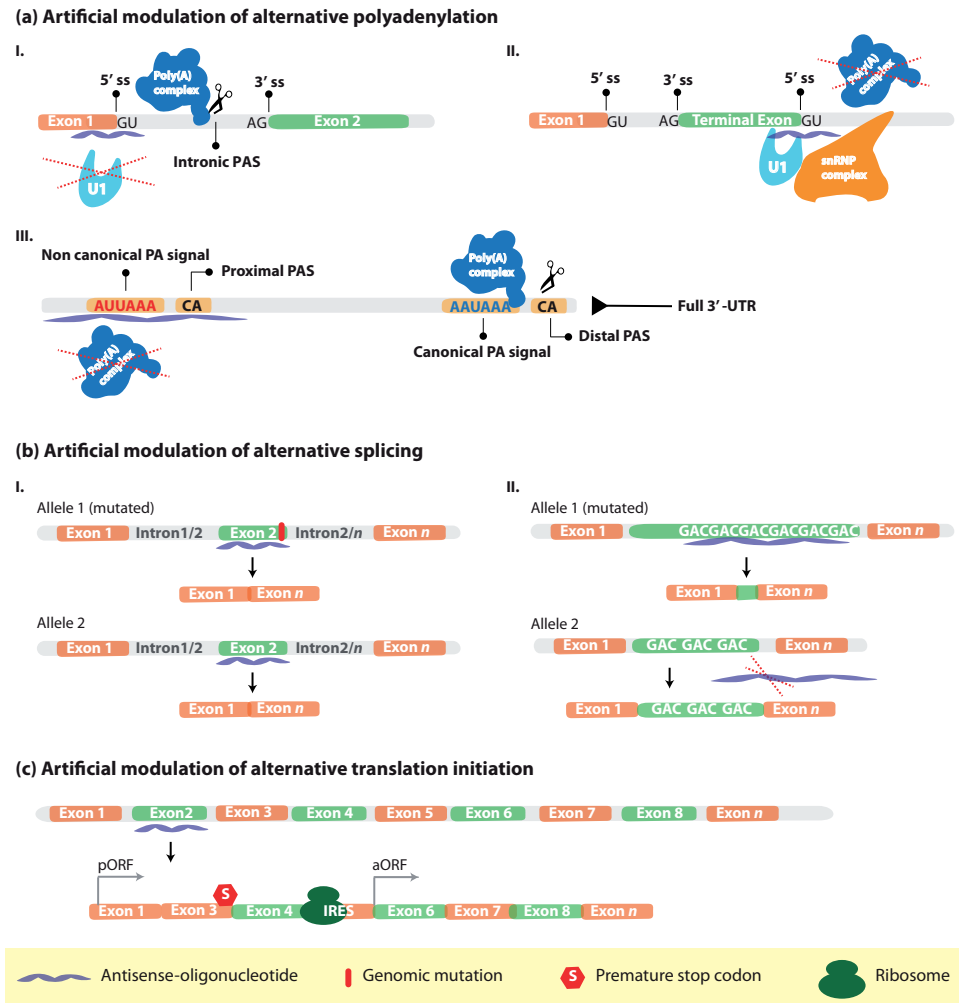


Figure 1. Schematic overview of RNA-based therapeutic approaches currently in development. **(a)** Modulation of APA through the use of an AON (i) which masks the 5′ splice site, inhibiting correct splicing and leading to intronic PAS, or (ii) through the use of an oligonucleotide which binds the terminal exon and U1, recruiting the snRNP complex, and causing a block of correct polyadenylation, leading to degradation of the pre-mRNA, or (iii) through the use on an OAN which masks non-canonical polyadenylation signals, to restore polyadenylation at canonical sites (or viceversa). **(b)** Modulation of splicing through the use of (i) an AON targeting an exon in a non-allele specific approach (the AON will target both alleles) or (ii) through the use of an AON targeting an expansion mutation within an exon in an allele-specific approach (the AON will preferentially bind to the exon containing an equal amount of repeats). **(c)** Modulation of translation initiation, through AON-mediated alternative splicing in the DMD gene. The skipping of exon 2 leads to a premature stop codon, which pushes the translation machinery to recognize an IRES and start translation from exon 6.

would be to target the mutated protein to modulate the activity of the polyadenylation machinery itself, instead of targeting the affected transcripts. A way to target the mutated protein is by using antisense-based strategies to modulate alternative splicing (Spitali and Aartsma-Rus, 2012).

Artificial modulation of alternative splicing through antisense mediated exon skipping (**Figure 1b**) represent a promising therapeutic tool through which targeted exon are hidden from the

splicing machinery and not included in the mRNA. This strategy aims to restore protein function in monogenetic disorders where a gene is affected by mutations that lead to truncated non-functional proteins, such as Duchenne muscular dystrophy (DMD) (Aartsma-Rus et al., 2004). A similar antisense-based approach has been tested also to modify protein toxicity in polyglutamine disorders, such as Spinocerebellar ataxia type 3 (SCA3), where the protein toxicity is reduced by removing the toxic polyglutamine repeat from the ataxin-3 protein (Evers et al., 2013).

Since the mutant PABPN1 is caused by an expansion mutation in the polyalanine repeat in the N-terminus of the protein, a similar approach could be used to skip the repeat and restore a reading frame that would code for a functional truncated protein. The advantage of this method, over a common exon skipping approach, is that only the mutated mRNA is targeted, whereas the functional allele produces the endogenous protein. This allele specificity is missing in commonly exon skipping approach, where both alleles are targeted and affected by the therapy.

Antisense oligonucleotide-based strategies can also be used to artificially modulate translation. Antisense oligonucleotides can be used to block the translation initiation complex, and lead to natural degradation of the targeted mRNAs. Ideally, uORFs and aORFs used in a physiological (**Chapter 4**) and/or disease context could therefore also represent a target for antisense-based strategies, to reduce protein production or allow the translation of truncated functional isoforms.

Next to modulating mRNA production, protein expression can also be modulated with similar approaches (**Figure 1c**). Artificial modulation of alternative translation initiation can therefore also be used to interfere with disease mechanisms. Wein et al. (Wein et al., 2014) have shown that, by inducing an out-of-frame exon skipping, it is possible to generate a premature stop codon which leads to the activation of an internal ribosome entry site (IRES) driving the expression of an aORF. This therapeutic approach was shown to produce truncated but functional dystrophin and correct muscle injury in DMD mice. Interestingly, activation of the IRES can also be achieved by glucocorticoids treatment, which represent a standard treatment in DMD patients (Manzur et al., 2008), even though the molecular mechanism is not clear.

Even though the approaches discussed here are promising, there are some limitations faced in the use of antisense oligonucleotides to interfere with RNA processing machineries and/or the translational apparatus. The most important limiting factors include their poor cellular uptake, possible off-target effects and toxicity (Kole et al., 2012).

To increase the therapeutic effect of these targeting approaches, a possible option might be to combine antisense-based strategies with transcript-therapy.

The term transcript-therapy refers to the use of chemically modified mRNAs (Kormann et al., 2011) to produce functional proteins that would act as endogenous proteins. The transcript-therapy represents an alternative to DNA-based gene-therapy, with some important advantages. The introduction of synthetic genes into the genome, through the use of viruses, has been associated with increased risk of leukemia, and strong immune responses. Chemically modified mRNAs, such as those carrying an anti-reverse cap analog nucleotide and pseudo-uridine or methyl-cytidine substitutions, do not show any of these side effects (Warren et al., 2010). These modifications decrease the binding of the mRNAs to toll-like receptors, avoiding therefore the activation of the innate immune system. Another advantage brought by these chemical modifications is the increased stability of the mRNAs (compared to non-modified mRNAs).

Proof-of-concept studies have shown the potential of transcript-therapy in different contexts: from restoration of lung function in mice affected by lethal congenital lung defects due to the lack

of surfactant protein B (Kormann et al., 2011), to increased cardiomyocyte survival after myocardial infarction (Huang et al., 2015).

Despite the current challenges discussed above, the targeting of regulatory processes involved in the production of mRNAs as therapeutic approach represents a promising path towards personalized medicine.

REFERENCES

1. Aartsma-Rus,A., A.A.Janson, W.E.Kaman, M.Bremmer-Bout, G.J.van Ommen, J.T.den Dunnen, and J.C.van Deutekom. 2004. Antisense-induced multiexon skipping for Duchenne muscular dystrophy makes more sense. Am. J. Hum. Genet. 74: 83-92.
2. Anvar,S.Y., Y.Raz, N.Verway, B.van der Sluijs, A.Venema, J.J.Goeman, J.Vissing, S.M.van der Maarel, P.A.'t Hoen, B.G.van Engelen, and V.Raz. 2013. A decline in PABPN1 induces progressive muscle weakness in oculopharyngeal muscle dystrophy and in muscle aging. Aging (Albany. NY) 5: 412-426.
3. Arava,Y., Y.Wang, J.D.Storey, C.L.Liu, P.O.Brown, and D.Herschlag. 2003. Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S. A 100: 3889-3894.
4. Atkins JF,R.F.G. Recoding: Expansion of Decoding Rules Enriches Gene Expression. Springer, New York.
5. 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.
6. Avendano-Vazquez,S.E., A.Dhir, S.Bembich, E.Buratti, N.Proudfoot, and F.E.Baralle. 2012. Autoregulation of TDP-43 mRNA levels involves interplay between transcription, splicing, and alternative polyA site selection. Genes Dev. 26: 1679-1684.
7. Betney,R., S.E.de, J.Krishnan, and I.Stansfield. 2010. Autoregulatory systems controlling translation factor expression: thermostat-like control of translational accuracy. RNA. 16: 655-663.
8. Betney,R., S.E.de, C.Mertens, Y.Knox, J.Krishnan, and I.Stansfield. 2012. Regulation of release factor expression using a translational negative feedback loop: a systems analysis. RNA. 18: 2320-2334.
9. Bidou,L., V.Allamand, J.P.Rousset, and O.Namy. 2012. Sense from nonsense: therapies for premature stop codon diseases. Trends Mol. Med. 18: 679-688.
10. Boissel,S., O.Reish, K.Proulx, H.Kawagoe-Takaki, B.Sedgwick, G.S.Yeo, D.Meyre, C.Golzio, F.Molinari, N.Kadhom, H.C.Etchevers, V.Saudek, I.S.Farooqi, P.Froguel, T.Lindahl, S.O'Rahilly, A.Munnich, and L.Colleaux. 2009. Loss-of-function mutation in the dioxygenase-encoding FTO gene causes severe growth retardation and multiple malformations. Am. J. Hum. Genet. 85: 106-111.
11. Buenrostro,J.D., P.G.Giresi, L.C.Zaba, H.Y.Chang, and W.J.Greenleaf. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10: 1213-1218.
12. Chung,W.K., C.Kitner, and B.J.Maron. 2011. Novel frameshift mutation in Troponin C (TNNC1) associated with hypertrophic cardiomyopathy and sudden death. Cardiol. Young. 21: 345-348.
13. Church,C., L.Moir, F.McMurray, C.Girard, G.T.Banks, L.Teboul, S.Wells, J.C.Bruning, P.M.Nolan, F.M.Ashcroft, and R.D.Cox. 2010. Overexpression of Fto leads to increased food intake and results in obesity. Nat. Genet. 42: 1086-1092.
14. Dana,A. and T.Tuller. 2012. Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. PLoS. Comput. Biol. 8: e1002755.
15. Daniel,C., G.Silberberg, M.Behm, and M.Ohman. 2014. Alu elements shape the primate transcriptome by cis-regulation of RNA editing. Genome Biol. 15: R28.
16. Darnell,J.C., S.J.Van Driesche, C.Zhang, K.Y.Hung, A.Mele, C.E.Fraser, E.F.Stone, C.Chen, J.J.Fak, S.W.Chi, D.D.Licatalosi, J.D.Richter, and R.B.Darnell. 2011. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. Cell 146: 247-261.
17. Daxinger,L. and E.Whitelaw. 2010. Transgenerational epigenetic inheritance: more questions than answers. Genome Res. 20: 1623-1628.
18. Dina,C., D.Meyre, S.Gallina, E.Durand, A.Korner, P.Jacobson, L.M.Carlsson, W.Kiess, V.Vatin, C.Lecoeur, J.Delplanque, E.Vaillant, F.Pattou, J.Ruiz, J.Weill, C.Levy-Marchal, F.Horber, N.Potoczna, S.Hercberg, S.C.Le, P.Bougneres, P.Kovacs, M.Marre, B.Balkau, S.Cauchi, J.C.Chevre, and P.Froguel. 2007. Variation in FTO contributes to childhood obesity and severe adult obesity. Nat. Genet. 39: 724-726.
19. Dominissini,D., S.Moshitch-Moshkovitz, S.Schwartz, M.Salmon-Divon, L.Ungar, S.Osenberg, K.Cesarkas, J.Jacob-Hirsch, N.Amariglio, M.Kupiec, R.Sorek, and G.Rechavi. 2012. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature 485: 201-206.

20. Evers,M.M., H.D.Tran, I.Zalachoras, B.A.Pepers, O.C.Meijer, J.T.den Dunnen, G.J.van Ommen, A.Aartsma-Rus, and W.M.van Roon-Mom. 2013. Ataxin-3 protein modification as a treatment strategy for spinocerebellar ataxia type 3: removal of the CAG containing exon. Neurobiol. Dis. 58: 49-56.

21. Farajollahi,S. and S.Maas. 2010. Molecular diversity through RNA editing: a balancing act. Trends Genet. 26: 221-230.

22. Fischer,J., L.Koch, C.Emmerling, J.Vierkotten, T.Peters, J.C.Bruning, and U.Ruther. 2009. Inactivation of the Fto gene protects from obesity. Nature 458: 894-898.

23. Frayling,T.M., N.J.Timpson, M.N.Weedon, E.Zeggini, R.M.Freathy, C.M.Lindgren, J.R.Perry, K.S.Elliott, H.Lango, N.W.Rayner, B.Shields, L.W.Harries, J.C.Barrett, S.Ellard, C.J.Groves, B.Knight, A.M.Patch, A.R.Ness, S.Ebrahim, D.A.Lawlor, S.M.Ring, Y.Ben-Shlomo, M.R.Jarvelin, U.Sovio, A.J.Bennett, D.Melzer, L.Ferrucci, R.J.Loos, I.Barroso, N.J.Wareham, F.Karpe, K.R.Owen, L.R.Cardon, M.Walker, G.A.Hitman, C.N.Palmer, A.S.Doney, A.D.Morris, G.D.Smith, A.T.Hattersley, and M.I.McCarthy. 2007. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316: 889-894.

24. 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.

25. Fustin,J.M., M.Doi, Y.Yamaguchi, H.Hida, S.Nishimura, M.Yoshida, T.Isagawa, M.S.Morioka, H.Kakeya, I.Manabe, and H.Okamura. 2013. RNA-methylation-dependent RNA processing controls the speed of the circadian clock. Cell 155: 793-806.

26. Goraczniak,R., B.A.Wall, M.A.Behlke, K.A.Lennox, E.S.Ho, N.H.Zaphiros, C.Jakubowski, N.R.Patel, S.Zhao, C.Magaway, S.A.Subbie, Y.L.Jenny, S.Lacava, K.R.Reuhl, S.Chen, and S.I.Gunderson. 2013. U1 Adaptor Oligonucleotides Targeting BCL2 and GRM1 Suppress Growth of Human Melanoma Xenografts In Vivo. Mol. Ther. Nucleic Acids 2: e92.

27. Graham,R.R., C.Kyogoku, S.Sigurdsson, I.A.Vlasova, L.R.Davies, E.C.Baechler, R.M.Plenge, T.Koeuth, W.A.Ortmann, G.Hom, J.W.Bauer, C.Gillett, N.Burtt, D.S.Cunninghame Graham, R.Onofrio, M.Petri, I.Gunnarsson, E.Svenungsson, L.Ronnblom, G.Nordmark, P.K.Gregersen, K.Moser, P.M.Gaffney, L.A.Criswell, T.J.Vyse, A.C.Syvanen, P.R.Bohjanen, M.J.Daly, T.W.Behrens, and D.Altshuler. 2007. Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. Proc. Natl. Acad. Sci. U. S. A 104: 6758-6763.

28. Houseley,J. and D.Tollervey. 2009. The many pathways of RNA degradation. Cell 136: 763-776.

29. Huang,C.L., A.L.Leblond, E.C.Turner, A.H.Kumar, K.Martin, D.Whelan, D.M.O'Sullivan, and N.M.Caplice. 2015. Synthetic Chemically Modified mRNA-Based Delivery of Cytoprotective Factor Promotes Early Cardiomyocyte Survival Post-Acute Myocardial Infarction. Mol. Pharm. 12: 991-996.

30. Iannuzzi,M.C., R.C.Stern, F.S.Collins, C.T.Hon, N.Hidaka, T.Strong, L.Becker, M.L.Drumm, M.B.White, B.Gerrard, and . 1991. Two frameshift mutations in the cystic fibrosis gene. Am. J. Hum. Genet. 48: 227-231.

31. 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.

32. 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.

33. Ivanov,I.P., G.Loughran, M.S.Sachs, and J.F.Atkins. 2010. Initiation context modulates autoregulation of eukaryotic translation initiation factor 1 (eIF1). Proc. Natl. Acad. Sci. U. S. A 107: 18056-18060.

34. Jia,G., Y.Fu, X.Zhao, Q.Dai, G.Zheng, Y.Yang, C.Yi, T.Lindahl, T.Pan, Y.G.Yang, and C.He. 2011. N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nat. Chem. Biol. 7: 885-887.

35. Kajiyama,K., M.Okada-Hatakeyama, Y.Hayashizaki, H.Kawaji, and H.Suzuki. 2013. Capturing drug responses by quantitative promoter activity profiling. CPT. Pharmacometrics. Syst. Pharmacol. 2: e77.

36. Kawahara,Y., K.Ito, H.Sun, H.Aizawa, I.Kanazawa, and S.Kwak. 2004. Glutamate receptors: RNA editing and death of motor neurons. Nature 427: 801.

37. Kervestin,S. and A.Jacobson. 2012. NMD: a multifaceted response to premature translational termination. Nat. Rev. Mol. Cell Biol. 13: 700-712.

38. Kole,R., A.R.Krainer, and S.Altman. 2012. RNA therapeutics: beyond RNA interference and antisense

oligonucleotides. Nat. Rev. Drug Discov. 11: 125-140.

39. Kormann,M.S., G.Hasenpusch, M.K.Aneja, G.Nica, A.W.Flemmer, S.Herber-Jonat, M.Huppmann, L.E.Mays, M.Illenyi, A.Schams, M.Griese, I.Bittmann, R.Handgretinger, D.Hartl, J.Rosenecker, and C.Rudolph. 2011. Expression of therapeutic proteins after delivery of chemically modified mRNA in mice. Nat. Biotechnol. 29: 154-157.

40. Lai,D.P., S.Tan, Y.N.Kang, J.Wu, H.S.Ooi, J.Chen, T.T.Shen, Y.Qi, X.Zhang, Y.Guo, T.Zhu, B.Liu, Z.Shao, and X.Zhao. 2015. Genome-wide profiling of polyadenylation sites reveals a link between selective polyadenylation and cancer metastasis. Hum. Mol. Genet.

41. Lareau,L.F., A.N.Brooks, D.A.Soergel, Q.Meng, and S.E.Brenner. 2007. The coupling of alternative splicing and nonsense-mediated mRNA decay. Adv. Exp. Med. Biol. 623: 190-211.

42. Lembo,A., C.F.Di, and P.Provero. 2012. Shortening of 3'UTRs correlates with poor prognosis in breast and lung cancer. PLoS. One. 7: e31129.

43. Levanon,K., E.Eisenberg, G.Rechavi, and E.Y.Levanon. 2005. Letter from the editor: Adenosine-to-inosine RNA editing in Alu repeats in the human genome. EMBO Rep. 6: 831-835.

44. Li,L., D.Wang, M.Xue, X.Mi, Y.Liang, and P.Wang. 2014. 3'UTR shortening identifies high-risk cancers with targeted dysregulation of the ceRNA network. Sci. Rep. 4: 5406.

45. Liu,B., Y.Han, and S.B.Qian. 2013. Cotranslational response to proteotoxic stress by elongation pausing of ribosomes. Mol. Cell 49: 453-463.

46. Liu,J., Y.Yue, D.Han, X.Wang, Y.Fu, L.Zhang, G.Jia, M.Yu, Z.Lu, X.Deng, Q.Dai, W.Chen, and C.He. 2014. A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. Nat. Chem. Biol. 10: 93-95.

47. Maas,S., Y.Kawahara, K.M.Tamburro, and K.Nishikura. 2006. A-to-I RNA editing and human disease. RNA. Biol. 3: 1-9.

48. Machnicka,M.A., K.Milanowska, O.O.Osman, E.Purta, M.Kurkowska, A.Olchowik, W.Januszewski, S.Kalinowski, S.Dunin-Horkawicz, K.M.Rother, M.Helm, J.M.Bujnicki, and H.Grosjean. 2013. MODOMICS: a database of RNA modification pathways--2013 update. Nucleic Acids Res. 41: D262-D267.

49. 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.

50. Manzur,A.Y., T.Kuntzer, M.Pike, and A.Swan. 2008. Glucocorticoid corticosteroids for Duchenne muscular dystrophy. Cochrane. Database. Syst. Rev. CD003725.

51. Mariappan,M., X.Li, S.Stefanovic, A.Sharma, A.Mateja, R.J.Keenan, and R.S.Hegde. 2010. A ribosome-associating factor chaperones tail-anchored membrane proteins. Nature 466: 1120-1124.

52. Meyer,K.D., Y.Saletore, P.Zumbo, O.Elemento, C.E.Mason, and S.R.Jaffrey. 2012. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. Cell 149: 1635-1646.

53. Morris,A.R., A.Bos, B.Diosdado, K.Rooijers, R.Elkon, A.S.Bolijn, B.Carvalho, G.A.Meijer, and R.Agami. 2012. Alternative cleavage and polyadenylation during colorectal cancer development. Clin. Cancer Res. 18: 5256-5266.

54. Myerowitz,R. 1997. Tay-Sachs disease-causing mutations and neutral polymorphisms in the Hex A gene. Hum. Mutat. 9: 195-208.

55. Ni,J.Z., L.Grate, J.P.Donohue, C.Preston, N.Nobida, G.O'Brien, L.Shiue, T.A.Clark, J.E.Blume, and M.Ares, Jr. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev. 21: 708-718.

56. Nishikura,K. 2010. Functions and regulation of RNA editing by ADAR deaminases. Annu. Rev. Biochem. 79: 321-349.

57. Olivares-Hernandez,R., S.Bordel, and J.Nielsen. 2011. Codon usage variability determines the correlation between proteome and transcriptome fold changes. BMC. Syst. Biol. 5: 33.

58. Park,E., B.Williams, B.J.Wold, and A.Mortazavi. 2012. RNA editing in the human ENCODE RNA-seq data. Genome Res. 22: 1626-1633.

59. Paz,N., E.Y.Levanon, N.Amariglio, A.B.Heimberger, Z.Ram, S.Constantini, Z.S.Barbash, K.Adamsky, M.Safran,

A.Hirschberg, M.Krupsky, I.Ben-Dov, S.Cazacu, T.Mikkelsen, C.Brodie, E.Eisenberg, and G.Rechavi. 2007. Altered adenosine-to-inosine RNA editing in human cancer. Genome Res. 17: 1586-1595.

60. Ping,X.L., B.F.Sun, L.Wang, W.Xiao, X.Yang, W.J.Wang, S.Adhikari, Y.Shi, Y.Lv, Y.S.Chen, X.Zhao, A.Li, Y.Yang, U.Dahal, X.M.Lou, X.Liu, J.Huang, W.P.Yuan, X.F.Zhu, T.Cheng, Y.L.Zhao, X.Wang, J.M.Rendtlew Danielsen, F.Liu, and Y.G.Yang. 2014. Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. Cell Res. 24: 177-189.

61. Raz,V., H.Buijze, Y.Raz, N.Verwey, S.Y.Anvar, A.Aartsma-Rus, and S.M.van der Maarel. 2014. A novel feed-forward loop between ARIH2 E3-ligase and PABPN1 regulates aging-associated muscle degeneration. Am. J. Pathol. 184: 1119-1131.

62. Rodriguez,A.J., K.Czaplinski, J.S.Condeelis, and R.H.Singer. 2008. Mechanisms and cellular roles of local protein synthesis in mammalian cells. Curr. Opin. Cell Biol. 20: 144-149.

63. Rottman,F.M., R.C.Desrosiers, and K.Friderici. 1976. Nucleotide methylation patterns in eukaryotic mRNA. Prog. Nucleic Acid Res. Mol. Biol. 19: 21-38.

64. Saltzman,A.L., Y.K.Kim, Q.Pan, M.M.Fagnani, L.E.Maquat, and B.J.Blencowe. 2008. Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. Mol. Cell Biol. 28: 4320-4330.

65. Schwartz,S., S.D.Agarwala, M.R.Mumbach, M.Jovanovic, P.Mertins, A.Shishkin, Y.Tabach, T.S.Mikkelsen, R.Satija, G.Ruvkun, S.A.Carr, E.S.Lander, G.R.Fink, and A.Regev. 2013. High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. Cell 155: 1409-1421.

66. Scuteri,A., S.Sanna, W.M.Chen, M.Uda, G.Albai, J.Strait, S.Najjar, R.Nagaraja, M.Orru, G.Usala, M.Dei, S.Lai, A.Maschio, F.Busonero, A.Mulas, G.B.Ehret, A.A.Fink, A.B.Weder, R.S.Cooper, P.Galan, A.Chakravarti, D.Schlessinger, A.Cao, E.Lakatta, and G.R.Abecasis. 2007. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. PLoS. Genet. 3: e115.

67. Shalgi,R., J.A.Hurt, I.Krykbaeva, M.Taipale, S.Lindquist, and C.B.Burge. 2013. Widespread regulation of translation by elongation pausing in heat shock. Mol. Cell 49: 439-452.

68. 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.

69. Shin,J.H., M.Janer, B.McNeney, S.Blay, K.Deutsch, C.B.Sanjeevi, I.Kockum, A.Lernmark, J.Graham, H.Arnqvist, E.Bjorck, J.Eriksson, L.Nystrom, L.O.Ohlson, B.Schersten, J.Ostman, M.Aili, L.E.Baath, E.Carlsson, H.Edenwall, G.Forsander, B.W.Granstrom, I.Gustavsson, R.Hanas, L.Hellenberg, H.Hellgren, E.Holmberg, H.Hornell, S.A.Ivarsson, C.Johansson, G.Jonsell, K.Kockum, B.Lindblad, A.Lindh, J.Ludvigsson, U.Myrdal, J.Neiderud, K.Segnestam, S.Sjoblad, L.Skogsberg, L.Stromberg, U.Stahle, B.Thalme, K.Tullus, T.Tuvemo, M.Wallensteen, O.Westphal, and J.Aman. 2007. IA-2 autoantibodies in incident type I diabetes patients are associated with a polyadenylation signal polymorphism in GIMAP5. Genes Immun. 8: 503-512.

70. Singh,P., T.L.Alley, S.M.Wright, S.Kamdar, W.Schott, R.Y.Wilpan, K.D.Mills, and J.H.Graber. 2009. Global changes in processing of mRNA 3' untranslated regions characterize clinically distinct cancer subtypes. Cancer Res. 69: 9422-9430.

71. Spitali,P. and A.Aartsma-Rus. 2012. Splice modulating therapies for human disease. Cell 148: 1085-1088.

72. Truong,H.T., T.Dudding, C.L.Blanchard, and S.H.Elsea. 2010. Frameshift mutation hotspot identified in Smith-Magenis syndrome: case report and review of literature. BMC. Med. Genet. 11: 142.

73. Vorlova,S., G.Rocco, C.V.Lefave, F.M.Jodelka, K.Hess, M.L.Hastings, E.Henke, and L.Cartegni. 2011. Induction of antagonistic soluble decoy receptor tyrosine kinases by intronic polyA activation. Mol. Cell 43: 927-939.

74. Wang,X., Z.Lu, A.Gomez, G.C.Hon, Y.Yue, D.Han, Y.Fu, M.Parisien, Q.Dai, G.Jia, B.Ren, T.Pan, and C.He. 2014. N6-methyladenosine-dependent regulation of messenger RNA stability. Nature 505: 117-120.

75. Warren,L., P.D.Manos, T.Ahfeldt, Y.H.Loh, H.Li, F.Lau, W.Ebina, P.K.Mandal, Z.D.Smith, A.Meissner, G.Q.Daley, A.S.Brack, J.J.Collins, C.Cowan, T.M.Schlaeger, and D.J.Rossi. 2010. Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. Cell Stem Cell 7: 618-630.

76. Wein,N., A.Vulin, M.S.Falzarano, C.A.Szigyarto, B.Maiti, A.Findlay, K.N.Heller, M.Uhlen, B.Bakthavachalu, S.Messina, G.Vita, C.Passarelli, F.Gualandi, S.D.Wilton, L.R.Rodino-Klapac, L.Yang, D.M.Dunn, D.R.Schoenberg, R.B.Weiss, M.T.Howard, A.Ferlini, and K.M.Flanigan. 2014. Translation from a DMD exon 5 IRES results in a functional dystrophin isoform that attenuates dystrophinopathy in humans and mice. Nat. Med. 20: 992-

1000.

77. Weirauch,U., A.Grunweller, L.Cuellar, R.K.Hartmann, and A.Aigner. 2013. U1 adaptors for the therapeutic knockdown of the oncogene pim-1 kinase in glioblastoma. Nucleic Acid Ther. 23: 264-272.

78. Wiestner,A., M.Tehrani, M.Chiorazzi, G.Wright, F.Gibellini, K.Nakayama, H.Liu, A.Rosenwald, H.K.Muller-Hermelink, G.Ott, W.C.Chan, T.C.Greiner, D.D.Weisenburger, J.Vose, J.O.Armitage, R.D.Gascoyne, J.M.Connors, E.Campo, E.Montserrat, F.Bosch, E.B.Smeland, S.Kvaloy, H.Holte, J.Delabie, R.I.Fisher, T.M.Grogan, T.P.Miller, W.H.Wilson, E.S.Jaffe, and L.M.Staudt. 2007. Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. Blood 109: 4599-4606.

79. Zhang,G., M.Hubalewska, and Z.Ignatova. 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat. Struct. Mol. Biol. 16: 274-280.

80. Zheng,G., J.A.Dahl, Y.Niu, P.Fedorcsak, C.M.Huang, C.J.Li, C.B.Vagbo, Y.Shi, W.L.Wang, S.H.Song, Z.Lu, R.P.Bosmans, Q.Dai, Y.J.Hao, X.Yang, W.M.Zhao, W.M.Tong, X.J.Wang, F.Bogdan, K.Furu, Y.Fu, G.Jia, X.Zhao, J.Liu, H.E.Krokan, A.Klungland, Y.G.Yang, and C.He. 2013. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. Mol. Cell 49: 18-29.

SUMMARY

SAMENVATTING (SUMMARY IN DUTCH)

CURRICULUM VITAE

LIST OF PUBLICATIONS

SUMMARY

In 1941, geneticists G.W. Beadle and E.L. Tatum formulated the ground-breaking hypothesis “one gene, one enzyme”, which led to their Nobel Prize in Physiology or Medicine in 1958. Today’s researchers have reformulated Beadle and Tatum’s hypothesis, as the link between DNA and proteins has been proven to be much more complex: even though the human genome contains less than 20,000 genes, these genes encode for more than 80,000 protein-coding messenger RNAs (mRNAs, intermediate molecules between DNA and proteins), which have been estimated to give rise to hundreds of thousands of proteins. One major remaining challenge in cellular biology is to understand the mechanisms regulating the diversity of mRNAs and proteins expressed from a single gene.

The work described in this thesis focuses on the mechanisms that give rise to alternative mRNAs and their alternative translation into proteins. Each of the described studies has been based on a specific set of high-throughput RNA sequencing technologies. Together these provide a comprehensive view of these alternative regulatory mechanisms. An overview of the available RNA sequencing methods, together with an introduction to different regulatory layers which define the expression of a gene, are presented in Chapter 1. This Chapter describes the processes of alternative transcription, alternative mRNA processing and alternative translation, focusing on what we have learnt from RNA sequencing studies.

Our work in Chapter 2 and Chapter 3 investigates the process of alternative polyadenylation, which is one of the steps during mRNA processing, and results in the inclusion or exclusion of sequences that affect the stability of an mRNA or the nature of the protein isoform formed.

Chapter 2 shows the role of alternative polyadenylation in the context of oculopharyngeal muscular dystrophy (OPMD), an autosomal dominant and progressive muscle disorder caused by mutation in the PABPN1 gene. In this study, we identified and quantified the usage of alternative polyadenylation sites in affected skeletal muscles using a novel high-throughput single-molecule poly(A)-site sequencing method. We demonstrated transcriptome-wide shortening of mRNAs in OPMD and propose a novel role for the PABPN1 protein in poly(A) site selection.

Chapter 3 describes genetic variants associated with alternative polyadenylation. In this study we used RNAseq and DeepSAGE to identify genetic variants affecting the usage of alternative polyadenylation sites, by disrupting or forming polyadenylation signal sequences. We confirmed the known genotype-dependent alternative polyadenylation in the gene IRF5 (explaining its genetic association with systemic lupus erythematosus), and we reported novel causative variants affecting alternative polyadenylation by changes in the polyadenylation signal, seven of which had been reported as associated with diseases.

Chapter 4 focuses on mechanisms controlling protein synthesis (translation) during skeletal muscle differentiation, highlighting changes in the use of alternative translation initiation sites. This chapter demonstrates that skeletal muscle differentiation is not only regulated at the level of mRNA transcription and processing, but that also mRNA translation is tightly controlled for specific subsets of functionally correlated genes and contributes to the diversity of proteins required for skeletal muscle function.

In Chapter 5 we investigated the interdependence between alternative regulatory events in gene expression. In this study, based on single-molecule full-length RNA sequencing, we demonstrated coordination and interdependence between alternative transcription initiation, alternative splicing, and alternative polyadenylation in nearly half of the detected genes, and suggested a coordinating role for RNA binding proteins from the muscle blind family (MBNL) in the regulation of splicing and

polyadenylation.

The alternative regulatory mechanisms described in Chapter 1 and investigated in this thesis represent only a portion of all the mechanisms affecting gene and protein expression. Additional regulatory mechanisms are shortly discussed in Chapter 6, to give a more comprehensive picture of the complexity of the process of gene expression. Finally, Chapter 6 connects fundamental research in the RNA field with clinical care, describing new diagnostic and therapeutic approaches that are based on the alternative modes regulating gene expression at transcriptional, post-transcriptional and translational level.

SAMENVATTING

In 1941 formuleerden de genetici G.W. Beadle en E.L. Tatum de volgende revolutionaire hypothese: "één gen correspondeert met één enzym". Hiervoor ontvingen zij in 1958 de Nobelprijs voor de Fysiologie of Geneeskunde. Hedendaagse onderzoekers herformuleerden de hypothese van Beadle en Tatum, omdat het verband tussen DNA en eiwitten een stuk ingewikkelder is gebleken dan aanvankelijk werd aangenomen: hoewel het humane genoom minder dan 20.000 genen bevat, coderen deze voor meer dan 80.000 boodschapper-RNA's of messenger-RNA's (mRNA's, intermediairs tussen DNA en eiwit). Deze mRNA's worden op hun beurt vertaald in honderdduizenden verschillende eiwitten. Een belangrijke nog openstaande uitdaging in de cellulaire biologie is om de mechanismen te begrijpen die zorgen dat een grote verscheidenheid aan mRNA's en eiwitten geproduceerd kan worden uit één enkel gen.

Het in dit proefschrift beschreven onderzoek richt zich op de mechanismen die ten grondslag liggen aan de vorming van alternatieve mRNAs en hun vertaling in verschillende eiwitten. Elk van de uitgevoerde onderzoeken is gebaseerd op een specifieke set van high-throughput RNA sequencing-technologieën. Gezamenlijk geven deze technieken een gedetailleerd beeld van de diversiteit aan geproduceerde mRNA's en eiwitten. Een overzicht van de beschikbare RNA sequencing-methoden en een introductie tot de verschillende niveaus waarop de genexpressie wordt gereguleerd, worden gegeven in Hoofdstuk 1. Dit hoofdstuk beschrijft de volgende processen: alternatieve transcriptie, alternatieve mRNA-verwerking en alternatieve translatie. De nadruk ligt op wat we hebben geleerd uit studies van RNA-sequencing.

In de studies beschreven in Hoofdstuk 2 en Hoofdstuk 3 is het proces van polyadenylering onderzocht, een van de stappen in de verwerking van mRNA. Alternatieve polyadenylering leidt tot de opname of uitsluiting van bepaalde sequenties in het mRNA die de stabiliteit van het mRNA of de aard van het gevormde eiwit beïnvloeden.

Hoofdstuk 2 beschrijft de rol van alternatieve polyadenylering in relatie tot oculopharyngeale spierdystrofie (OPMD), een autosomaal dominante en progressieve spierziekte veroorzaakt door mutatie van het PABPN1 gen. In deze studie hebben wij de polyadenyleringsplaatsen in kaart gebracht en het gebruik van deze plaatsen gekwantificeerd in zieke en gezonde spieren. Hierbij hebben wij een nieuwe high-throughput technologie gebruikt die polyadenyleringsplaatsen kan sequencen op het niveau van individuele moleculen. We hebben laten zien dat er een transcriptoombrede verkorting van mRNA's plaatsvindt in OPMD. We stellen een nieuwe rol voor PABPN1 in de selectie van polyadenyleringsplaatsen voor.

Hoofdstuk 3 beschrijft genetische varianten die geassocieerd zijn met alternatieve polyadenylering. In deze studie hebben wij gebruik gemaakt van RNAseq en DeepSAGE technologieën om de genetische varianten die het gebruik van alternatieve polyadenyleringsplaatsen beïnvloeden te identificeren. Deze genetische varianten verstoren of vormen een signaalsequentie die van belang is voor polyadenylering. We hebben een bekend effect in het IRF5 gen, dat genetisch in verband gebracht wordt met de ziekte systemic lupus erythematosus, bevestigd en daarnaast zeven vergelijkbare genetische varianten geïdentificeerd die eveneens in verband gebracht worden met ziekten.

Hoofdstuk 4 gaat in op de mechanismen die de eiwitsynthese (translatie) gedurende skeletspierdifferentiatie reguleren en besteedt speciale aandacht aan het gebruik van alternatieve translatie-initiatieplaatsen. In dit hoofdstuk tonen we aan dat skeletspierdifferentiatie niet alleen gereguleerd wordt op het niveau van transcriptie en mRNA-verwerking, maar dat hierbij ook de mRNA-translatie strak wordt gereguleerd. Dit is het meest duidelijk te zien in specifieke, functioneel

coherente genen die betrokken zijn bij het translatieproces. Regulering op het niveau van translatie draagt ook in belangrijke mate bij aan de vorming van de diversiteit aan eiwitten die nodig is voor het correct functioneren van de spier.

In hoofdstuk 5 hebben we de afhankelijkheid tussen verschillende regelmechanismen onderzocht. In deze studie, gebaseerd op de sequencing van individuele, intacte mRNA-moleculen, hebben we aangetoond dat er coördinatie en onderlinge afhankelijkheid is tussen alternatieve transcriptie-initiatie, alternatieve splicing en alternatieve polyadenylering. Dit treedt op in tenminste de helft van alle gedetecteerde genen. De resultaten van deze studie suggereren een rol voor RNA-bindende eiwitten uit de muscle blind (MBNL) familie in de coördinatie van splicing en polyadenylering.

De alternatieve regelmechanismen beschreven in Hoofdstuk 1 en bestudeerd in dit proefschrift bestrijken slechts een gedeelte van alle mechanismen die de gen- en eiwitexpressie beïnvloeden. De overige regelmechanismen worden kort bediscussieerd in Hoofdstuk 6, om zo een completer beeld te schetsen van de complexiteit van het proces van genexpressie. Tenslotte verbindt Hoofdstuk 6 fundamenteel onderzoek in het RNA-veld met klinische zorg en gaat in op nieuwe diagnostische en therapeutische toepassingen die gebaseerd zijn op de alternatieve mechanismen die de genexpressie reguleren op het niveau van transcriptie, mRNA-verwerking en translatie.

CURRICULUM VITAE

Eleonora de Klerk was born in Catania, Italy, on July 29th, 1983. She attended Enrico Boggio Lera Scientific-Linguistic Lyceum, and graduated cum laude in the summer of 2002. After finishing her high school, she started a Bachelor program in Biological Sciences at the University of Catania, and graduated cum laude in 2006. As part of her Bachelor study, she spent 6 months at the Human Genetics Laboratory of the Vittorio Emanuele Hospital in Catania, under the supervision of dr. Angela Ragusa. Her project focused on the clinical application of quantitative fluorescence PCR for rapid prenatal detection of common chromosome aneuploidies, based on short tandem repeats analysis.

In 2006 she began a Master of Cellular and Molecular Biology at the same univeristy. During her Master, she spent 9 months in the Molecular Biology Laboratory of the Chemistry Department at the University of Catania, under the supervision of prof. Vito De Pinto and dr. Angela Messina. Her internship focused on porin ion channels located on the outer mitochondrial membrane, and the functional role of the N-terminal segment of these proteins in the open and closed state of the channel. In summer 2009 she obtained her Master cum laude, and received a short-term scholarship for visiting other European laboratories.

In september 2009 she moved to the Netherlands and, with her scholarship, visited the Leiden Genome Technology Center (LGTC), a sequencing facility of the Human Genetics Department at the Leiden University Medical Center (LUMC). In January 2010 she began to work as research analyst at the LGTC, focusing on Next Generation Sequencing.

She began her PhD in November 2010, at the Human Genetics Department in LUMC, under the supervision of prof. Johan den Dunnen and dr. Peter-Bram 't Hoen. Since the start of her PhD, she investigated regulatory mechanisms of gene expression based on a diverse set of high-throughput RNA sequencing technologies, and the results of her work are presented in this thesis.

Her work was presented at many international conferences, including RNA Society, European Molecular Biology Organization (EMBO), European Neuromuscular International Centre (ENMC), American Society and International Conference of Human Genetics (ASHG/ICHG), and European Society of Human Genetics (ESHG), where she was candidate for Young Investigator Awards in 2012.

From December 2014 until July 2015, while finishing her PhD thesis, she joined the group of prof. Silvère van der Maarel and Lucia Clemens-Daxinger.

During her PhD, she has been actively involved in education with supervision of Bachelor and Master students, and teaching (Basic RNA-seq data analysis course, Frontiers of Science course for Master students Biomedical Sciences, Next Generation Sequencing course Avans Hogeschool, PhD student course in Molecular Neurobiology). She also enjoyed organizing scientific events for fellow PhD students, such as the organization of the MGC-PhD Workshop 2013 (Medish Genetisch Centrum Zuid-West Nederland) in Luxembourg.

From fall 2015, Eleonora will start her training as postdoctoral fellow in dr. Michael McManus's laboratory at the University of California San Francisco (UCSF).

LIST OF PUBLICATIONS

de Klerk E., Fokkema I.F.A.C., Thiadens K.A.M.H., Goeman J.J., Palmblad M., den Dunnen J.T., von Lindern M., ’t Hoen A.C. Assessing the translational landscape of myogenic differentiation by ribosome profiling. Nucleic Acids Research. 2015

de Klerk E., 't Hoen P.A. Alternative mRNA transcription, processing and translation: insights from RNA-sequencing. Trends in Genetics. 2015

de Klerk E., den Dunnen J.T. and 't Hoen P.A. RNA sequencing: from tag-based profiling to resolving complete transcript structure. Cellular and Molecular Life Sciences. 2014

Zhernakova D.V., **de Klerk E.**, Westra H.J., Mastrokolias A., Amini S., Ariyurek Y., Jansen R., Penninx B.W., Hottenga J.J., Willemsen G., de Geus E.J., Boomsma D.I., Veldink J.H., van den Berg L.H., Wijmenga C., den Dunnen J.T., van Ommen G.J., 't Hoen P.A., Franke L. DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. PLoS Genetics. 2013

de Klerk E., Venema A., Anvar S.Y., Goeman J.J., Hu O., den Dunnen J.T., van der Maarel S.M., Raz V. and ’t Hoen P.A. Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation. Nucleic Acids Research. 2012

Anvar S.Y., **de Klerk E.**, Vermaat M., den Dunnen J.T., Turner S.W., 't Hoen P.A. Full-length mRNA sequencing uncovers a widespread coupling between transcription and mRNA processing. (submitted)

Thiadens K.A.M.H., **de Klerk E.**, Fokkema I.F.A.C., ’t Hoen A.C., von Lindern M. Ribosome Profiling uncovers the role of uORFs in translational control of gene expression during erythroblast differentiation (in preparation)