The handle http://hdl.handle.net/1887/35768 holds various files of this Leiden University dissertation.

Author: Klerk, Eleonora de

Title: Mechanisms controlling mRNA processing and translation: decoding the regulatory layers defining gene expression through RNA sequencing

Issue Date: 2015-09-30
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.
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 temporarily 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 temporarily 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
CHAPTER 6

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$^6$A) (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$^6$A 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$^6$A might represent a novel fundamental mechanism controlling protein expression.

The effects of m$^6$A 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$^6$A 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^6^A 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^6^A 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
CHAPTER 6

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
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


