Computational tools for the study of RNA processing and function

Amadís Pagès Pinós

DOCTORAL THESIS UPF / 2016

THESIS SUPERVISORS

Dr. Eduardo Eyras and Dr. Roderic Guigó

DCEXS

Acknowledgements

Tot i ser les primeres línies d'aquesta tesi, aquestes són en realitat unes línies de clausura. El punt i final a una etapa èpica, a un treball de més de quatre anys que no hagués pogut ser possible sense l'ajut, el recolzament i el suport de molts. Només espero no oblidar-me de ningú, i si és així, les meves més sinceres disculpes.

En primer lugar, quiero dar las gracias a mi supervisor Eduardo Eyras por haberme acogido en el grupo de investigación Regulatory Genomics y por haberme dado la oportunidad de desarrollar allí mi tesis doctoral. De manera muy sincera, gracias por tu apoyo en esta carrera a contrarreloj que han venido a ser los últimos meses del doctorado. També voldria donar les gràcies al meu supervisor Roderic Guigó pel seu suport i per obrir-me les portes del grup de recerca Computational Biology of RNA Processing.

Voldria donar les gràcies a totes les persones del PRBB amb les que he compartit grup, despatx, passadís, equips de volley, beer sessions, conferències, retreats, sopars, viatges, tertúlies i confidències. A l'Ivan, la Núria, el Salva i el Daniel (Teniente Poglayen!), amb qui hem iniciat un projecte que ara per ara ens ha portat a Cork, però que espero ens porti molt més lluny. A Alexandros, a ver si celebramos el final del doctorado en Sifnos o en el Canigó. Als més que il·lustres comensals del Porvenir, pilars fonamentals d'aquests anys de doctorat: Eneritz, Jaume, Inma, Núria, Steve i Mireya. A Nico, espero que nos reencontremos algun dia, ya sea en Barcelona o Bariloche. A Juanra, Gael y las pintas del Michael Collins. Thanks Babita for the late afternoon conversations about how to survive a Ph.D. thesis and how to get rich doing crazy businesses, y también gracias Juanlu, Héctor, JC y Janet por esta última etapa en el grupo y algún que otro karaoke. Gràcies als meus companys de despatx: a l'Aina (espero que conservis l'excel·lent gust musical), a l'Adriano i especialment al Max, amb qui he compartit moments memorables. Als Ravolleys a la Biolognesa i al Mix & Match. Al grup del Roderic, gràcies per acollir-me.

Gràcies als que feu que les coses funcionin dia rere dia brindant-nos sempre un tracte excel·lent: Alfons, Miguel, Carina, Martina i Romina.

Gràcies a tots els que, des de ja fa anys, m'heu fet viure cada minut com si fos l'últim. A l'eskuadron, als mestres i mestresses, i a tots amb els qui la Vall de Pineta ha forjat un vincle indestructible.

A la Neus, a qui tinc la immensa sort de poder-li agraïr tot, cada dia.

Finalment, vull donar les gràcies a la meva família. A ma germana, font inesgotable de complicitat, anècdotes i bons moments. I als meus pares, últims responsables de que estigui escrivint aquestes línies, a qui dedico aquesta tesi, i a qui vull expressar la més profunda, sentida i sincera gratitud.

Cork, 20 de juliol de 2016.

Resum

El processament de les cadenes d'àcids ribonucleics (ARN) és un mecanisme mol·lecular crucial gràcies al qual els precursors dels ARN missatgers es converteixen en ARN missatgers madurs. L'exemple més notable és l'anomenat empalmament, procés en el qual els introns són eliminats del precursor, i que sovint origina formes alternatives d'ARN missatgers madurs. Els ARN no codificants, o ARN que no tenen la capacitat de ser traduïts en proteïna, també estan sotmesos a diversos passos de processament, i alguns estudis estableixen una connexió entre aquest processament i la funció que exerceixen. Addicionalment, estudis recents assenyalen els ARN no codificants com a reguladors de l'empalmament alternatiu. Tanmateix, aquests mecanismes no es coneixen en profunditat. Aquest treball inclou el desenvolupament de tres novedoses propostes centrades en (i) l'anàlisi del processament de petits ARN no codificants, (ii) l'estudi de l'empalmament alternatiu i (iii) l'estudi dels processos cel·lulars que determinen les interaccions entre aquests dos.

Abstract

RNA processing is a crucial molecular mechanism by which precursor RNAs are converted into mature RNAs. The most notable processing step is splicing, in which introns are removed from precursor messenger RNAs, and that often gives birth to alternative forms of mature messenger RNAs. Non-coding RNAs, or RNAs that lack the capacity to be translated into a protein, also undergo extensive RNA processing steps during their biogenesis, and several studies establish a relation between the processing of non-coding RNAs and the function they exert. Moreover, recent studies point non-coding RNAs as regulators of alternative splicing, although the regulation mechanisms are not completely understood. The present work includes the development of three novel computational approaches focused on (i) the analysis of the processing of small non-coding RNAs, (ii) the study of alternative splicing and (iii) the study of the cellular processes that guide the interplay between both of them.

Preface

I was born at the same year Temple Smith and Michael Waterman published their seminal paper, entitled “Identification of common molecular subsequences”, in which they presented the Smith-Waterman algorithm to perform local sequence alignment. Today, more than three decades later, the coupling between computer science and biology have evolved to a point where biology constantly benefits from the research in computer science, and at the same time biology constantly poses new problems and challenges that act as driving forces for the advancement in the field of computer science. The disciplines that emerged at the interface of both fields are called Bioinformatics and Computational Biology. Although the scientific community seems to struggle to pinpoint the differences between them, both disciplines pivot over two different entities: the *tools* and the *data*. The *tools* are the pieces of software including all their components: the core algorithms, the mathematical models, the user interface or even the user manual. The *data* is whatever type of information that needs to be mined, analyzed or visualized in order to understand the biology that lies beneath it.

That said, I believe the present work is a paradigmatic example of a thesis in Bioinformatics or Computational biology. While some of the results presented in this dissertation could be broadly classified as basic research in genomics, the power engine that enabled them has been built upon the research efforts done in the field of computer science. The contribution of this thesis is then, double: thanks to the novel tools developed during the four years this thesis has taken, I have been able to analyze large amounts of diverse data and report a number of results that might contribute to a better understanding of open biological questions.

Table of contents

List of figures

List of tables

INTRODUCTION

1 The centrality of RNA to life

The elucidation of the DNA structure by James D. Watson and Francis Crick in 1953 (Francis and Crick, 1953), widely regarded as the starting point of modern molecular biology, ignited the fuse that led to unraveling the central role of RNA in life. The double helix structure, coupled to the finding by Frederick Sanger that proteins had a fixed sequence of amino acids (Sanger and Thompson, 1953), established a link between genes and proteins. However, a big enigma was still to be solved: how could the information stored in DNA be used to produce proteins in the proper amount, in the correct cell and in the right moment? At that time, knowledge about RNA was still scarce, being merely considered as a cell constituent, a nucleic acid with ribose instead of deoxyribose.

Shortly, the description of how messenger RNAs carry genetic information that directs protein synthesis and the findings related to the crucial roles of RNA in translation led to the deciphering of the genetic code. At that point, the consensus view was cogently summarized by Crick's declaration of "the central dogma of molecular biology" in 1958 (Crick, 1958), and later revised in 1970 (Crick, 1970), that stated that information flows from DNA to RNA to protein. Implicit in this paradigm was that the final functional product of a gene coded in the DNA was one protein with a specific sequence.

While the majority of the aforementioned studies were accomplished with bacteria and viruses, copying DNA into RNA did not suffice for genetic function in eukaryotic cells, suggesting the existence of some sort of still unknown mechanism of RNA processing. This gave birth to a wealth of relevant findings that culminated in the late seventies with the discovery of precursor messenger RNA (pre-mRNA) splicing. This mechanism, by which introns are removed from the pre-mRNA to create mature, functional RNA molecules, was first observed independently by Phillip A Sharp (Berget et al., 1977) and Richard J. Roberts (Chow et al., 1977), and opened the mind of the scientific community

towards the idea that regulation of mRNA might occur. A few years later, the finding that pure RNA molecules could perform chemical catalysis in many ways analogous to those of proteins, paved the way for RNA to be considered a centric, key player in determining cellular characteristics.

From that point on, the last four decades of biomolecular investigation on RNA shed even more light on how information is processed in living cells. Numerous co- and post-transcriptional RNA processing mechanisms such as alternative splicing, RNA editing, RNA interference or RNA-directed chromatin remodeling have been described, and an unprecedented number of non-coding RNAs (ncRNAs), RNA molecules that are not translated into protein - traditionally labeled as "junk DNA" (Gilbert, 1985) because of their apparent inability to make a meaningful contribution to the cell function - have been reported and characterized, specially thanks to the advent of the high-throughput sequencing technologies. In particular, the members of the ENCODE (ENcyclopedia of Dna Elements) project, a public research consortium launched by the US National Human Genome Research Institute back in 2003, recently reported that three quarters of the human genome is capable of being transcribed (Djebali et al., 2012), and claimed that it was possible to assign biochemical functions for 80% of the genome (ENCODE Project Consortium, 2012), an assertion that has been subject to intense scientific debate (Graur et al., 2013). In any case, genes are nowadays considered fuzzy transcription units capable of generating many products, including different proteins and non-coding RNAs.

In the light of all the discoveries made in the past few years, the "one gene one product" motto is challenged by the reality of pervasive transcription and extensive RNA processing, and there is no doubt that the classical conception of Crick's central dogma needs some reassessment, a task that had already been undertaken by some authors (Mattick, 2003; Shapiro, 2009). Figure 1-1 depicts a reformulation of the so-called central dogma of the molecular biology, and serves as the starting point and a declaration of intentions for this doctoral thesis centered in RNA, life's indispensable molecule.

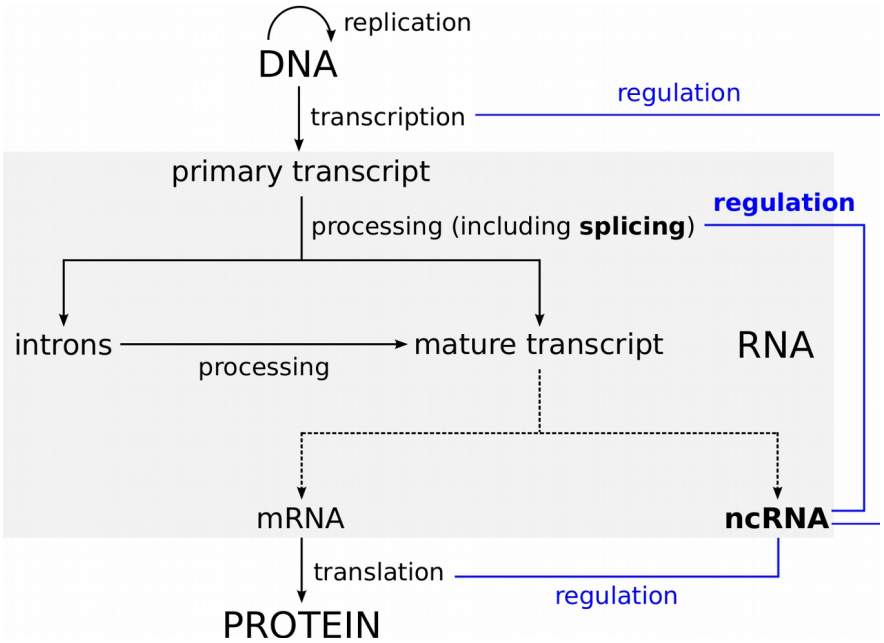


Figure 1-1. **A reformulation of the central dogma**. Flow of genetic information in higher eukaryotes. Dashed lines do not represent biochemical processes but rather subtypes of a specific product. Superimposed to the flow diagram, in blue, we show the network of regulatory functions carried out by non-coding RNAs. We give in boldface the main questions addressed in the present thesis.

2 Non-coding RNAs

Albeit the term non-coding RNA is nowadays used on a daily basis in the scientific literature, a bit of historical perspective is needed to fully understand the importance of non-coding RNAs in today's scientific research, and how its relevance has evolved through the history of molecular biology.

During the early 1950s, scientists observed in the cell cytoplasm for the first time the existence of a number of small granules similar in size. Those granules, that were named *ribosomes* by Robert George in a 1958 meeting (Darnell, 2011), were later found to be rich in RNA and the site of protein synthesis in *E. Coli* (McQuillen et al., 1959). The RNA contained in the ribosomes was naturally called ribosomal RNA (rRNA). Soon thereafter, a series of experiments performed by Paul C. Zamecnik and Mahlon B. Hoagland suggested that a completely different type of RNA molecule, later named transfer RNA (tRNA), was able to adapt to activated amino acids to enable protein synthesis in association with the ribosomes (Hoagland et al., 1957; Hoagland et al., 1958). After the discovery of messenger RNA (mRNA) in the early 1960s, it was clear that even though rRNAs and tRNAs play key roles in protein synthesis, neither of them were responsible of coding and carrying the genetic information stored in the DNA.

The 1960s also witnessed two other important discoveries concerning RNA. The first one was that in animal cells there was at least one molecular mechanism capable of generating functional cytoplasmic RNA products from nuclear primary transcripts, also named precursors. This precursor-to-product process was termed *RNA processing*, and was observed for the first time in rRNAs (Scherrer et al., 1963). Extensive RNA processing by the molecular machinery turned out to be a main characteristic of almost all the catalogued non-coding RNAs today. The second discovery concerned the observation by Harris Busch and colleagues of low molecular weight RNA subjects in the nuclei and nucleoli of mammalian cells that had a nucleotide composition different from the other known RNAs at that time (Muramatsu et al., 1966).

By the late 1970s, scientists discovered that genomic coding sequences (exons) were interspersed with non-coding regions (introns) that were present in the primary gene transcript, the pre-messenger RNA (pre-mRNA), but removed by a nuclear machinery, the spliceosome, to give rise to the mature RNA (mRNA). The quest to decipher the nature of the splicing mechanism and the RNA processing pathways shed some light on the unknown RNAs found by Busch's team: in the early 1980s, approximately 10 species of RNAs other than rRNA, tRNA and mRNAs were shown to be stable in mammalian cells. These were hypothesized to be involved in the processing of high molecular weight RNAs. Those ten species were divided into two groups, termed small nuclear RNAs (snRNAs) and Class III RNAs (Zieve, 1981). Some of the species in the snRNA group later emerged as key players in the pre-mRNA splicing and accepted as core components of the spliceosome (reviewed in Guthrie and Patterson, 1988), while other species of the same group were further characterized as modulators of the modification and processing of rRNA, and named small nucleolar RNAs (snoRNAs), because of their predominant localization in the cell nucleoli (reviewed in Maxwell and Fournier, 1995). Class III RNAs included the 7SL RNA, found to be a component of the signal recognition particle (SRP) (Walter and Blobel, 1981), a ribonucleoprotein complex in charge of directing the protein traffic within the cell; and YRNAs, a highly conserved RNA in mammals originally found in the serum of patients with the autoimmune disorder lupus erythematosus (Lerner et al., 1981).

During the 1980s, although terms like non-coding sequence or non-coding region were commonly used to refer to the untranslated and intronic regions of the pre-mRNA, researchers prefered the term structural RNAs (Walker, 1983) to denominate the major classes of characterized non-coding RNAs that were directly or indirectly required for mRNA processing and translation. One of the most prominent finding concerning RNA of this decade was that RNA itself was capable of enzymatic catalysis. Thomas Cech and colleagues found that the RNA structure formed by an intron of an rRNA precursor was sufficient to make the splicing reaction happen (Kruger et al., 1982), and Sidney Altman and his collaborators characterized the RNA subunit of the RNase P ribonucleoprotein complex as its catalytic center (Guerrier-Takada et al., 1983).

At the beginning of the 1990s, two sets of independent discoveries were determinant for the forthcoming revolution on non-coding RNAs. The first of these discoveries started the “miRNA revolution” (Morris and Mattick, 2014), and involved the description of two small (~22 nt) regulatory RNAs, found in 1993 (Lee et al., 1993) and 2000 (Reinhart et al., 2000), that are transcribed from the *C. Elegans* lin-4 and let-7 loci, respectively, and that regulate its developmental timing. Those two non-coding RNAs remained mere eccentricities until the elucidation of the RNA interference (RNAi) pathway (see Chapter 2.2.4), which led to the identification of those RNAs as micro RNAs (miRNAs) and to the discovery of many more like them.

The second of these discoveries took place in 1990. A gene was found in human to produce a cytoplasmic transcript that had no open reading frame and was not associated with the translation machinery, but that was transcribed by RNA polymerase II, spliced and polyadenylated (Brannan et al., 1990), which are characteristic features of mRNAs (refer to Chapter 3.1). Two years later, the *XIST* gene was reported to produce a human non-coding transcript containing at least 8 exons and totalling a length of 17kb, which was involved in the X inactivation process (Brown et al., 1992). This set of findings unearthed the existence of long, functional RNA molecules that, like pre-mRNAs, undergo splicing and maturation, but that did not code for protein.

The following decades (1990s till today) probably constitute the golden era of non-coding RNA research with the rising of many classes of regulatory RNAs such as Piwi-interacting RNAs (piRNAs), a large class of small RNAs expressed in animal cells linked to transcriptional gene silencing of retrotransposons and other genetic elements in germ line cells (Weick and Miska, 2014). Increasingly fast and accurate sequencing methods have enabled whole genomes and expressed RNA to be sequenced (refer to chapter 4.1). These methods, coupled to an exponentially growing catalogue of computational tools (refer to chapter 4.2), have enabled the discovery and characterization of an unprecedented number of non-coding RNAs. By the beginning of the 21st century, non-coding RNAs already attracted the attention of the overall scientific community. The advent of new sequencing technologies provided evidence that suggested that the vast majority (97-98%) of the

transcriptional output from the human genome did not code for protein (Mattick, 2001), pointing towards a yet-to-be identified key role for non-coding RNAs in eukaryotic complexity and evolution.

Non-coding RNA class	Description	Number of subjects
3prime_overlapping_ncrna	Short non-coding transcripts transcribed from the 3'UTR	33
antisense	Transcript that overlap the genomic span (i.e. exon or introns) of a protein-coding locus on the opposite strand	11194
bidirectional_promoter_lncrna	A non-coding locus that originates from within the promoter region of a protein-coding gene, with transcription proceeding in the opposite direction on the other strand.	5
lincRNA	Long, intergenic noncoding (linc) RNA that can be found in evolutionarily conserved, intergenic regions	13481
macro_lncRNA	Unspliced lncRNA that is several kb in size	1
miRNA	Micro RNA	4093
misc_RNA	Miscellanous RNAs (includes 7SL, 7SK, YRNA among others)	2312
Mt_rRNA	Mitochondrial rRNA	2
Mt_tRNA	Mitochondrial tRNA	22
non_coding	Transcript which is known from the literature to not be protein coding	3
processed_transcript	Transcript from a protein-coding gene that doesn't contain an ORF	26977
retained_intron	Alternatively spliced transcript believed to contain intronic sequence relative to other, coding, variants.	26704

ribozyme	Ribozyme	8
rRNA	Ribosomal RNA	544
scaRNA	Small Cajal Antobody-associated RNA	49
sense_intronic	Long non-coding transcript in introns of a coding gene that does not overlap any exons	978
sense_overlapping	Long non-coding transcript that contains a coding gene in its intron on the same strand	343
snoRNA	Small nucleolar RNA	961
snRNA	Small nuclear RNA	1896
sRNA	Small RNA regulator	20
vaultRNA	Vault RNA	1

Table 2-1. **Non-coding RNAs in the human genome**. Catalogue of the major classes of non-coding RNAs in the last version of Gencode, including a short description and the number of transcripts found in the human genome.

RNA sequencing have also revealed that the majority of the human genome appears to be transcribed, a phenomenon described as pervasive transcription. Moreover, new sequencing technologies allowed scientists to reveal a new plethora of non-coding RNAs associated to this pervasive transcription, such as promoter-associated short RNAs (paRNAs) and transcription initiation RNAs (tiRNAs) (Kapranov et al., 2007).

Today the term non-coding RNA is defined by *The Dictionary of Genomics, Transcriptomics and Proteomics* (Günter, 2015) as “Any ribonucleic acid that does not encode a protein and can therefore not be annotated by a search for open reading frames. Instead, ncRNAs are encoded by intergenic, intronic and promoter sequences”, and this definition applies to thousands of RNA molecules organized within tens of functionally characterized different families (Table 2-1). Not only non-coding RNAs play a role in virtually all the cellular processes, numerous studies have also linked abnormalities

in several small and long ncRNAs to a wide spectrum of diseases (reviewed in Esteller, 2011; Taft et al., 2010; and Cooper et al. 2009), yielding a potential wealth of new biomarkers and therapeutic targets.

But, are all the non-coding RNAs and their functions known? This question is tackled in in the present thesis. The reader can find in the Results section the description of the novel computational methods developed in this thesis work to identify an extended catalogue of small non-coding RNAs in human, and a predicted list of non-coding RNAs, short and long, with the capacity to regulate the alternative splicing of certain genes.

2.1 Classification of non-coding RNAs

The classification of the myriad of non-coding RNAs known to date into functionally well characterized classes is still an open problem. Although a general consensus has been reached as to separate them according their size, into small (< 200 nt) and long (the rest) non-coding RNAs, it is still a rather intricate task. Many of the aforementioned classes of RNAs, including the structural tRNAs, snRNAs snoRNAs, the regulatory miRNAs and the transcription associated tiRNAs and paRNAs, fall into the small RNA category. However, meanwhile these classes populating the small RNA world are well characterized, the panorama changes completely for the long non-coding RNAs, to the extent that multiple reviews have been written in the last few years attempting to propose a classification of long non-coding RNAs into different classes to facilitate the study of their functionalities (Laurent et al., 2015; Ma et al., 2013). In this thesis we will focus on the straightforward size-based classification of non-coding RNAs and on the major classes, miRNAs, snRNAs, snoRNAs and tRNAs.

2.2 Small non-coding RNAs

2.2.1 transfer RNAs

tRNAs consist of 75-95 nt molecules ubiquitous in all organisms, and are amongst the most abundant of all RNA molecules, constituting 4-10% of all cellular RNA (Kirchner and Ignatova, 2015). Their aforementioned role in protein synthesis is performed as follows: transfer RNAs are adaptor molecules capable of loading a specific amino acid at its 3'-end and read the mRNA three nucleotides at a time by base pairing, forming a codon (mRNA) - anticodon (tRNA) interaction that determines the position of amino acids in proteins. The interactions with both the mRNA and the ribosome are possible thanks to the cloverleaf structure characteristic to all tRNAs, first proposed by Robert Holley in 1965 (Holley et al., 1965), consisting of three loops where the second is the AC-loop containing the anticodon sequence and the third is the D-loop that binds to the ribosomes.

In addition to their role in protein synthesis, tRNAs have been reported to be implicated in many other cellular processes including cell wall synthesis and reverse transcription (reviewed in Giegé, 2008). Moreover, mutations in tRNA-encoding genes, or in genes encoding the enzymes responsible of processing, charging and modifying the tRNAs, have been linked to several diseases such as diabetes, ataxia, intellectual disability and cancer (Scheper et al., 2007; Torres et al., 2014).

Biogenesis of tRNAs

Eukaryotic tRNA molecules are transcribed by RNA polymerase III from tRNA-encoding genes as precursors that are heavily processed by a sequence of maturation events (Figure 2-1b). These events comprise leader removing and trailer trimming, splicing, addition of the 3'-terminal acceptor residues and a number of post-transcriptional modifications of multiple nucleoside residues. Correctly processed mature tRNAs undergo a receptor-mediated export process to the cytoplasm, where they can perform their function as adaptors in translation.

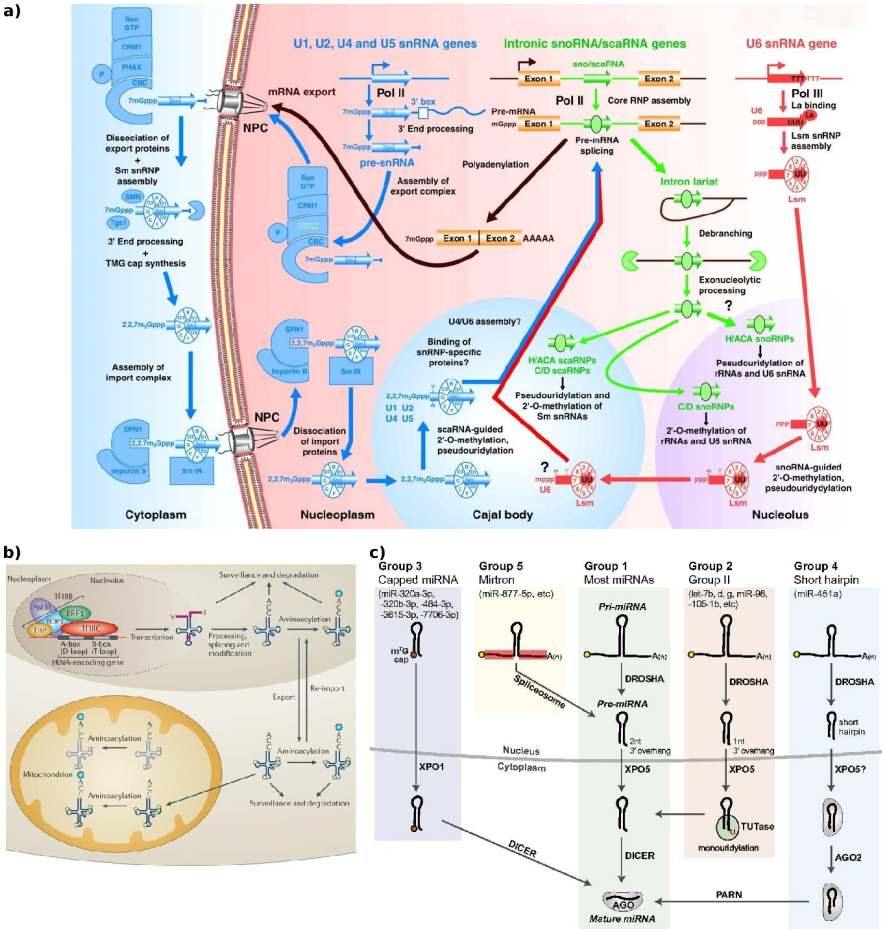


Figure 2-1. **Biogenesis of small RNAs**. Biogenesis pathways and effector machineries of (**a**) small nuclear RNAs from the major spliceosome and small nucleolar RNAs (adapted from Kiss, 2004), (**b**) transfer-RNAs (adapted from Kirchner and Ignatova, 2015) and (**c**) micro-RNAs, including canonical and non-canonical pathways (adapted from Kim et al., 2016). Note that in panel (**a**) U1, U2, U4 and U5 genes belong to the Sm class of snRNAs, while U6 snRNA gene belongs to the Sm-like class.

The number of tRNA-encoding genes varies significantly among organisms, and tends to increase with the complexity of the organism. Eukaryotic genomes might contain from 168 (*Schizosaccharomyces pombe*) to 12.292 (Danio Rerio) genes, being 610 the estimated number of human tRNA-encoding genes (Chan et al., 2009). Although these genes are dispersed throughout the nuclear and mitochondrial genomes, it has been reported that they

tend to cluster in the nucleolar portion of the linear nuclear genome, raising the possibility of a coordinated regulation of their transcription (Thompson et al., 2003).

2.2.2 small nuclear RNAs

Small nuclear RNAs comprise a large number of highly abundant, metabolically stable, non-polyadenylated RNAs localized in the nucleus, where they recruit a large set of proteins to form small nuclear ribonucleoprotein complexes (snRNPs) (Gesteland and Atkins, 1993). Those snRNPs interact to form a dynamic ribonucleoprotein machine called spliceosome, responsible of catalysing pre-mRNA splicing, including the recognition of the core splicing regulatory signals (refer to chapter 3.1). snRNAs can be subdivided in two major classes on the basis of sequence similarities and recruited protein cofactors: Sm and Sm-like snRNAs (reviewed in Matera et al., 2007).

Biogenesis of snRNAs

Functional mature snRNAs are 60-450 nt molecules born from a complex process that includes four major steps: transcription of a large snRNA precursor, processing of the precursor into mature-sized snRNA, introduction of covalent modifications of site-specific nucleotides and assembly with ribonucleoproteins (reviewed in Kiss, 2004). While adhering to those four steps, the two classes of snRNAs follow different biogenesis pathways (Figure 2-1a) linked to distinct subcellular compartments. In the case of Sm snRNAs, precursor transcription by a specialized form of RNA polymerase II is followed by 3’-end processing, addition of a 5’-cap structure and exportation from the nucleus for cytoplasmic maturation. In the cytosol, mature Sm snRNAs are assembled into snRNP complexes and transported back to the nucleoplasm for snRNP maturation in the Cajal Bodies. In contrast, biogenesis of Sm-like snRNAs is confined to the nucleus, where they undergo site-specific modifications and are assembled into snRNPs.

2.2.3 Small Nucleolar RNAs

Small nucleolar RNAs are a group of ~60-220 nt molecules typically localized to the nucleolus, where they play a canonical role in the posttranscriptional modification of rRNA and snRNA, upon binding to a core complement of proteins to form a small nucleolar ribonucleoprotein (snoRNP) complex (reviewed in Bratkovič and Rogelj, 2011) with catalytic activity. The structural features of snoRNAs include an antisense element that serve as a guide specifying the nucleotides in the target RNA to be modified, and a number conserved stretches of nucleotides, termed boxes, largely responsible for recruiting the core proteins that conform the snoRNPs (Maxwell and Fournier, 1995).

Based on the sequence motifs of these structural boxes and the proteins they recruit, snoRNAs are broadly divided into two classes, known as the C/D-box and H/ACA-box snoRNAs. C/D-box snoRNAs typically have two sets of internal boxes, named C (RUGAUGA where R is a purine) and D (CUGA), and direct 2'-O-ribose methylation of specific nucleotides in their target RNAs. H/ACA snoRNAs are identified by their characteristic H (ANANNA, where N is any nucleotide) and ACA (ACA) boxes, and guide pseudouridylation of specific nucleotides in their target RNAs. There is a third class of snoRNAs, the small cajal body-specific RNAs (scaRNAs) that have sequence, structural and functional features of either or both classes of snoRNAs, and are responsible for the 2'-O-ribose methylation and pseudouridylation of some snRNAs (Bachellerie et al., 2002).

In addition to directing RNA modifications, some snoRNAs have been detected to play intriguing roles outside the nucleolus, which include mRNA modification (Cavaillé et al., 2000), regulation of pre-mRNA splicing (refer to chapter 2.3), miRNA-like regulation of gene expression (refer to chapter 2.2.5) and control of chromatin accessibility (Schubert et al., 2012). Moreover, snoRNAs have been shown to be implicated in neurological disorders: the loss of the HBII-52 snoRNA cluster within the human 15q11q13 locus is implicated in the progress of the Prader-Willi and Angelman syndromes, possibly due to inappropriate splicing of the serotonin receptor mRNA (Nicholls and Knepper, 2001).

Biogenesis of snoRNAs

Both major snoRNA families are closely related by their unusual genomic organization and biogenesis pathways (Figure 2-1a). In vertebrates, snoRNAs are encoded within introns and are not independently transcribed by any RNA polymerase. Instead, snoRNAs are processed from the spliced pre-mRNA introns by exonucleolytic digestion of the debranched lariat. It appears that binding of the snoRNP proteins occurs on the host pre-mRNA and is facilitated by the splicing machinery bound to that pre-mRNA (reviewed in Kiss, 2004).

2.2.4 micro-RNAs

The most widely studied class of non-coding RNAs are probably micro-RNAs (miRNAs), a class of endogenous, 19-24 nt regulatory small RNAs that regulate gene expression at the post-transcriptional level by base-pairing with mRNAs (Holley and Topkara, 2011). miRNAs are bound to an RNA-induced silencing complex (RISC), which contains at its core an Argonaute (AGO) protein, and serve as the guide to specific sites in the 3-'UTR or, rarely, in the coding sequence, of target mRNAs, causing mRNA degradation or repressing mRNA translation (Hutzinger and Izaurralde, 2011). Up to 2.600 miRNAs have been catalogued so far in the human genome, many of which targeting hundreds of mRNAs (Griffiths-Jones et al., 2008), to the extent that it is estimated that human miRNAs regulate more than 60% of all human protein-coding genes (Friedman et al., 2009).

Most non-coding RNAs exert their function by forming RNA:RNA duplexes with their target molecules. These interactions have been well characterized for miRNAs, which require full complementarity in plants but show only partial complementarity between the miRNA and the target mRNA in animals. In particular, nucleotides 2 to 8 from the 5'-end of the miRNA, referred as the seed, and certain structures in the 3'-end are crucial for the binding to the mRNA (Bartel, 2009). While miRNAs have been shown to be implicated in a large proportion of biological pathways at the cell level, they appear to be particularly linked to differentiation and in deciding cell fate (Kloosterman and Plasterk, 2006). Moreover, miRNA dysregulation has deep implications in human disease, as

they have been linked to tumorigenesis (Esquela-Kerscher and Slack, 2006; Hammond, 2005; Croce, 2009), neurological (Schaefer et al., 2007; Williams et al., 2009; Haramati et al., 2010) and cardiovascular (Zhao et al., 2007; Cordes et al., 2009) disorders, and other conditions such as Chron's disease (Brest et al., 2011) or deafness (Lewis et al., 2009).

Biogenesis of miRNAs

In humans, the vast majority of canonical miRNAs are encoded in intergenic regions or in introns of coding and non-coding transcripts, although some miRNAs are exceptionally encoded by exonic regions (Kim et al., 2009). Recent studies also report the existence of a number of miRNAs with precursors located across exon-intron junctions, triggering a competitive interaction between the spliceosome and the miRNA processing machinery (Melamed et al., 2013).

In the canonical biogenesis pathway (Figure 2-1c), miRNAs are transcribed primarily by RNA polymerase II as part of longer primary miRNA (pri-miRNA) transcripts. The first step in pri-miRNA maturation is carried out by the Microprocessor complex, which contains the RNAse III enzyme Drosha, the RNA binding protein DGCR8 and multiple co-factors (Denli et al., 2004). The Microprocessor recognizes a secondary structure in the pri-miRNA and cleaves it, originating a precursor miRNA hairpin (pre-miRNA) that is exported to the cytoplasm by the Exportin5/RanGTP complex (Lund et al., 2004). In the cytoplasm, the pri-miRNA is cleaved by the RNAse Dicer to generate a mature miRNA duplex that is composed by the guide strand miRNA, that is incorporated into the RISC, and the passenger strand miRNA (or miRNA*), that is usually degraded (Kim, 2005).

Apart from this canonical miRNA biogenesis pathway, several alternative mechanisms have been described (Figure 2-1c) (Kim et al., 2016). A number of miRNAs encoded in introns can bypass Drosha-mediated processing and originate from intronic lariat debranching (Ruby et al., 2007). Another alternative mechanism is involved in the biogenesis of mir-451, which does not require Dicer, but instead involves the catalytic activity of AGO2 (Cheloufi et al., 2010).

2.2.5 small RNAs derived from structural non-coding RNAs

The quest to decipher the microRNAome, i.e. the complete set of microRNAs in an organism, of human and other species is intimately coupled with the application of deep sequencing technologies, specially small RNA-Seq, which has the ability to quantitatively measure the expression of small RNAs of typically 17 to 35 nucleotides long. The output of an small RNA-Seq study is a library of reads, which are short stretches of nucleotides that can be mapped back to the genome of interest. Chapters 4 and 5 of the present thesis address RNA-Seq technologies and their applications in greater detail.

While small RNA deep sequencing studies are traditionally directed to the study of microRNAs, several of these studies provided evidence that a substantial number of the sequenced small RNAs were derived from previously well-characterized longer RNAs, such as tRNAs (Cole et al., 2009; Lee et al., 2009; Haussecker et al., 2010; Liao et al., 2010; Pederson, 2010), snoRNAs (Taft et al., 2009; Scott et al., 2009; Ono et al., 2011; Saraiya and Wang, 2008; Scott et al., 2012; Ender et al., 2008), snRNAs (Burroughs et al., 2011; Li et al., 2012), rRNAs (Zywicki et al., 2012) or YRNAs (Nicolas et al., 2012), which were for a long time considered to be stable transcripts (Figure 2-2). Despite being expressed in a stable and consistent manner, the function of these non-coding RNA derived small RNAs remains elusive for the vast majority of them, a question addressed in the present thesis.

<u>snoRNA-derived RNAs</u>

snoRNA-derived RNAs (sdRNAs) are ~15-33 nt small RNAs mainly derived from the 5’ and 3’ ends of C/D-box snoRNAs, and from stems of internal hairpins of H/ACA-box snoRNAs. Such positional profiles are evolutionary conserved in vertebrates, invertebrates, plants and unicellular eukaryotes, although sdRNAs show different size distributions amongst these organisms (Taft et al., 2009).

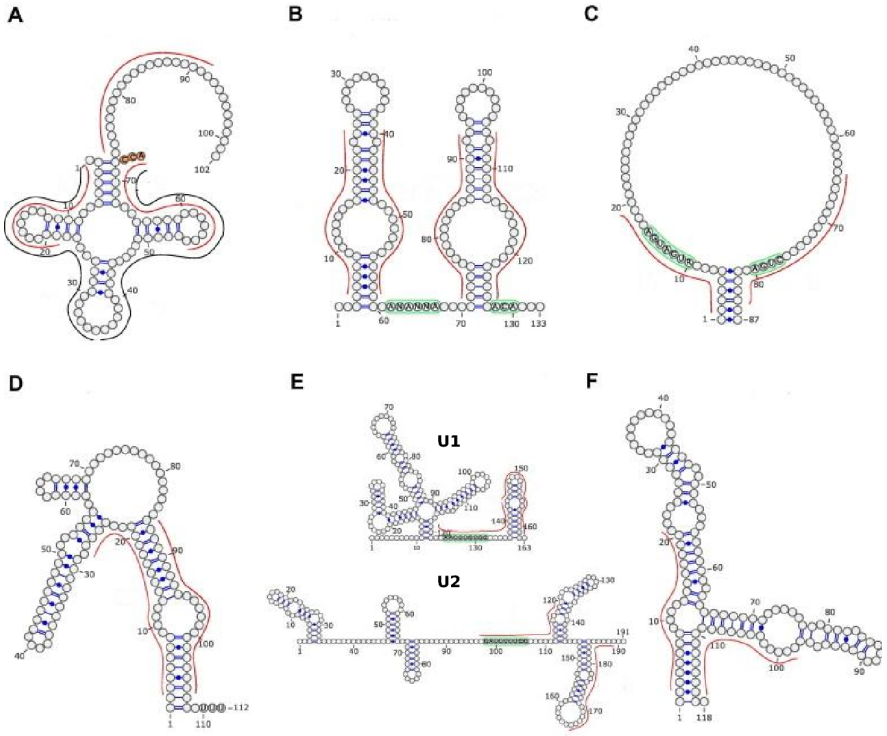


Figure 2-2. **Small RNAs derived from longer non-coding RNAs**. Schematic secondary structure and locations of derived small RNAs from longer non-coding RNAs, indicated as red and black lines for each non-coding RNA family: (**a**) tRNA, (**b**) H/ACA-box snoRNA, (**c**) C/D-box snoRNA, (**d**) YRNA, (**e**) U1 and U5 snRNAs and (**f**) 5S rRNA. Highlighted nucleotides belong to (**a**) 3'-termini adaptor, (**b**) H and ACA boxes, (**c**) C and D boxes and (**e**) Sm sites. Adapted from Chen and Heard, 2013.

Some sdRNAs have been shown to have miRNA-like processing features and activity. In human, a small RNA derived from the ACA45 scaRNA was shown to be produced in a Microprocessor-independent and Dicer-dependent manner to regulate the expression of its target, the cyclin-dependent kinase 11B (CDK11B) gene (Ender et al., 2008). While only a minor portion of the ACA45 scaRNA transcripts are exported to the cytosol for Dicer processing, the majority of them localize to the nucleolus where they fulfill its canonical function, a dual function that has been recently reported for further snoRNAs belonging to both C/D-box (Brameier et al., 2011) and H/ACA-box (Scott et al., 2009) families. But, while H/ACA-box- and scaRNA-derived small RNAs have the typical

size of Dicer products (~22nt), most of the C/D-box-derived small RNAs show a bimodal size distribution, between ~18 and ~27 nt, suggesting other functions and a biogenesis pathway mediated by nucleases other than Dicer (Martens-Uzunova et al., 2013). A potential function for this subset of sdRNAs was uncovered investigating the Prader-Willy syndrome-related HBII-52 family of snoRNAs, whose sdRNAs were shown to be involved in the regulation of alternative splicing (refer to Chapter 3).

tRNA-derived RNAs

Small RNAs derived from tRNAs can be separated in two major classes according to their size and the part of the tRNA from which they are derived: tRNA halves and small tRNA fragments (tRFs) (Sobala and Hutvagner, 2011). tRNA halves are 30-35 nt long and are produced by Angiogenin-mediated cleavage in or near the anticodon loop in response to induced stress. tRFs, of approximately 20 nt in length, are divided into three groups, according to their positions in the pre-tRNA: tRF-5, tRF-3 and tRF-1. tRFs that belong to the first two groups are conserved in mammals and plants and are specifically processed from the 5'- and 3'-ends of mature tRNAs respectively, probably via Dicer-dependent pathways (Cole et al., 2009). Members of the tRF-1 group, detected in different vertebrates, are produced from the 3'-pre-tRNA trailers during pre-tRNA processing by the nuclease RNase Z and are enriched in the cytosol, indicating the involvement of nuclear export pathways during their biogenesis (Lee et al., 2009). Albeit the exact roles of tRNA halves and tRFs are yet to be elucidated, accumulating evidence indicates that tRNA-derived small RNAs might have a role in the regulation of gene expression (Diebel et al., 2016).

2.3 Long non-coding RNAs

Long non-coding RNAs (lncRNAs), which are generally defined to be non-coding RNA molecules longer than 200 nt, have been in the spotlight of molecular biology researchers in the recent years due to their potential as components of an entire new layer of biological regulation. New studies uncovering novel lncRNAs and linking them to particular diseases or describing novel mechanisms of action appear on nearly a weekly basis. Surprisingly, despite this explosion of information, little is still known about how lncRNAs function, how many of them really exist or even whether most of them bear any biological significance (kung and Colognori, 2013).

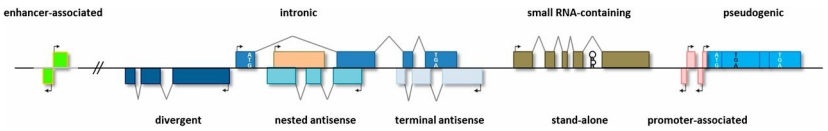


Figure 2-3. **Genomic contexts of long non-coding RNAs (lncRNAs)**. lncRNAs may be stand-alone transcription units or may be transcribed from enhancers, promoters or introns of protein-coding genes (shown with a start ATG and a stop TGA codon, in white); from pseudogenes (shown with a premature stop codon, in white); or antisense to other genes with varying degrees of overlap, from none (divergent), to partial (terminal), to complete (nested). lncRNAs may also host one or more small RNAs (black hairpin) within their transcription units. Adapted from Kung et al., 2013.

The classification of lncRNAs relies on the empirical attributes originally used to detect them, such as genomic context, transcript length, association with annotated protein-coding genes, association with other DNA elements of known function, sequence and structure conservation or association with subcellular structures (Laurent et al., 2015). While genomic context does not necessarily provide any information about their function, it serves as a convenient shorthand to organize these diverse species (Figure 2-3), and sometimes it is used for a “guilt by association” annotation of function by assigning the function of a nearby gene that correlates or anti-correlates in expression. Accordingly, lncRNAs have been grouped into six broad, non mutually exclusive categories: (1) stand-alone lncRNAs or long intergenic ncRNAs (lincRNAs), which are distinct transcription units located in intergenic regions,

transcribed by RNA Pol II, polyadenylated and spliced, usually with alternative isoforms (refer to chapters 3.1 and 3.3); (2) natural antisense transcripts (NATs), which arise from the opposite DNA strand of annotated transcription units (reviewed in Faghihi and Wahlestedt, 2009); (3) pseudogenes (Pink et al., 2011; Li et al., 2013); (4) long intronic ncRNAs, which are encoded in the introns of annotated genes (Rearick et al., 2011); (5) promoter-associated transcripts (PROMPTs), produced from the vicinity of transcription start sites in both sense and antisense direction, which are usually capped and polyadenylated (Seila et al., 2008); and (6) enhancer-associated lncRNAs (eRNAs), bidirectional transcripts that arise from enhancers (Kim et al., 2010), although it is not clear that those belong exclusively to the realm of lncRNAs, since eRNAs of 50 to 200 nt have been also reported (Natoli and Andrau, 2012).

The lack of an appropriate universal experimental approach to characterize lncRNAs function makes the identification of functional lncRNAs an arduous task, and despite the prevalence of long non-coding RNA genes in the eukaryotic genomes, only a small proportion have been examined for biological function. In fact, from the many thousands (~9.500) of lncRNAs annotated in the human genome (Harrow et al., 2012), no more than a couple of hundred of them are described to accomplish a specific biological function (Quek et al., 2014). At a cellular level, the best studied biological functions of lncRNAs include X chromosome inactivation (e.g. XIST) (Penny et al., 1996; Leung and Panning, 2014), imprinting (e.g. AIR) (Sleutels et al., 2002), control of development through the basal expression regulation of developmental genes (e.g. HOTAIR, HOTTIP and EVF2) (Rinn et al., 2007; Wang et al., 2011; Bond et al., 2009) and oncogenesis (e.g. MALAT1) (Gutschner et al., 2013). The blueprints of the underlying transcriptional and post-transcriptional mechanics that enable these lncRNAs to fulfill such complex functions can be constructed from the combinatorial usage of four archetypical molecular mechanisms: signals, decoys, guides and scaffolds (Wang et al., 2011). As signals, lncRNAs can faithfully reflect the combinatorial actions of transcription factors; as decoys, they can titrate transcription factors away from chromatin or work as sponges of regulatory miRNAs; as guides, lncRNAs can recruit chromatin-modifying enzymes to target genes, either in cis or in

trans; and as scaffolds, they can bring together multiple proteins to form ribonucleoprotein complexes.

Due to their prominent implication in the regulation of genes that are central to development and oncogenesis, it is not surprising that the dysregulation of lncRNAs is a primary feature of many complex human diseases (reviewed in Taft et al., 2010; Esteller, 2011). Indeed, lncRNAs have been shown to be implicated in diseases such as leukaemia (Calin et al., 2007), colon cancer (Pibouin et al., 2002), prostate cancer (Fu et al., 2006), breast cancer (Guffanti et al., 2009), psoriasis (Sonkoly et al., 2005) or Alzheimer (Faghihi et al., 2008), amongst many others.

2.3.1 Interplay between small RNAs and lncRNAs

Recent genome-wide studies suggest that a significant fraction of the annotated long non-coding RNA transcripts undergo post-transcriptional processing events not observed in mRNAs that yield small RNA products (reviewed in Quinn et al., 2016). One example is the lncRNA MALAT1, a highly abundant nuclear lncRNA expressed in many mammalian cell types, that has a tRNA-like structure which is cleaved by RNase P to give birth to the MALAT1-associated small cytoplasmic RNA (mascRNA), which is later capped and exported to the cytosol where it may exert its function (Wilusz et al., 2008). Another example are sno-lncRNAs, which consist of an intronic lncRNA flanked by two snoRNAs, thereby lacking a 5’-cap or a polyadenylated tail, that originate from introns encoding tandem snoRNAs (Yin et al., 2012).

3. Alternative splicing

Traditionally, splicing was defined as the mechanism by which introns are removed from precursor RNAs to create mature transcripts, which still stands as a valid but extremely simple definition. Rather than being considered as a mechanism solely devoted to intron removal, today we know that splicing is a complex mechanism of regulation modulated by specific factors or even by the transcription process itself, and that it is responsible for the RNA and protein diversity observed in higher eukaryotes due to its capacity to generate different alternatively spliced products. In this section I will give a brief description on alternative splicing regulation and its interplay with non-coding RNAs. I will start by outlining the underlying mechanics in the transcription and processing of eukaryotic protein-coding genes and some lncRNAs (refer to Chapter 3.2).

3.1 Eukaryotic transcription and processing

In eukaryotic organisms, the transfer of the genetic information contained in the DNA to the final mature RNA products involves a sequence of finely regulated biological reactions that befall in two stages: transcription and processing. While for a long time transcription and processing were thought to happen one after the other, now we're certain that both occur simultaneously and that each biological reaction step within them act as a sort of quality check for the next step (Orphanides and Reinberg, 2002).

According to the current vision (reviewed in Lee and Young, 200; Kornberg, 2007), eukaryotic transcription and processing starts with the formation of the so called preinitiation complex, which helps to position RNA polymerase II (RNAPII or Pol II) along with general transcription factors over gene promoters. Once bound to the DNA, the preinitiation complex separates the two DNA strands, and the template strand is oriented to the active site of the RNAPII. In this first phase the RNAPII enters into abortive cycles of synthesis and releases short RNA products before getting paused after 20 or 40 nucleotides from the transcription start site (LI et al., 1996). With the help of other cofactors, the Pol II gets released from most of the

bound transcription factors and escapes the promoter, entering the elongation step. Transcription elongation is a processive process in which the double stranded DNA is unzipped to make the template strand available to the Pol II for RNA synthesis. For every DNA base pair separated by the RNAPII, one RNA:DNA duplex is immediately formed. The two DNA strands then reunite at the trailing of the transcription complex while the single RNA strand emerges alone. Despite of its extremely processive nature, Pol II is an enzyme prone to transcriptional pausing and arrest for proofreading, which is regulated by the activity of several positive and negative elongation factors (Sims et al., 2004).

Most of the biological reactions involved in precursor RNA processing occur during transcription elongation. The nascent RNA is capped as soon as it exits the transcription complex, and further reactions catalyze the recruitment of the splicing machinery that catalyzes intron removal. The last processing step consists in the addition of a stretch of A nucleotides at the 3'-end of the nascent RNA, i.e. polyadenylation, and is coupled to the termination of transcription, which leads to the dissociation of the complete transcript and the release of the RNAPII from the DNA. Although splicing appears to be predominantly co-transcriptional in humans (Tilgner et al., 2012), it has been shown to happen in a post-transcriptional fashion as well (reviewed in Kornblihtt et al., 2004).

3.2 The splicing mechanism

The splicing reaction happens in the nucleus, is mediated by the spliceosome and leads to the definition of exons and introns in the precursor RNA, which are characterized by a number of splicing signals. Although the basic ability to splice introns is conserved throughout evolution, the splicing signals and their corresponding splicing factors have considerably evolved, uniquely shaping the splicing mechanisms of different organisms (Schwartz et al., 2008). Hence, here I will only discuss the features of the splicing mechanism in human.

3.2.1 Splicing signals

Splicing signals are regulatory regions in the pre-mRNA that are crucial for intron identification and removal, and present certain sequence particularities that allow their recognition by the splicing machinery. Splicing of introns is directed by four main splicing signals: the 5′ splice site (5′ss), the branch site (BS), the polypyrimidine tract (PP) and the 3′splice site (3′ss). There are two types of spliceosomes that are associated to two distinct types of introns with different sequence properties in eukaryotes: U2-dependent and U12-dependent introns (Figure 3-1) (Sharp and Burge, 1997). In human, while the vast majority of introns belong to the U2-dependent class, about 800 U12-dependent introns have been found within a similar number of genes that also contain introns from the U2-dependent class (Alioto, 2007).

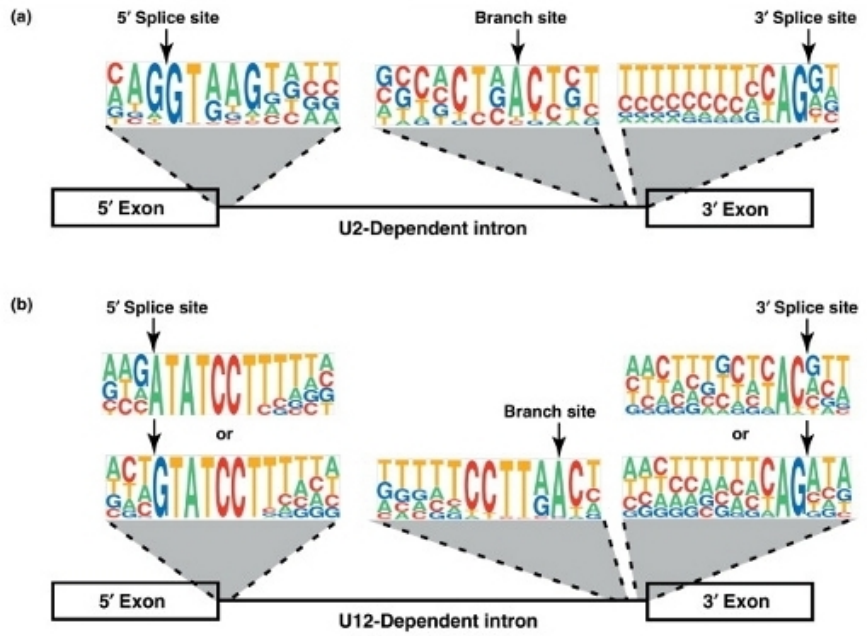


Figure 3-1. **Splicing consensus signals**. Splice sites consensus sequences for (**a**) U2-dependent introns and (**b**) U12-dependent introns. The boxes show graphical representations of the consensus sequences in which the size of each letter represents the frequency of that base at each position over all introns. Adapted from Padgett, 2012.

The 5’ss, also called donor site, delimits the exon/intron boundary at the 5’-end of the intron and has a highly conserved specific dinucleotide within a larger, less conserved region of about 9 nucleotides, that marks the beginning of the intron. While almost all of the human 5’ss in U2-dependent introns and several U12-dependent introns have a canonical GT dinucleotide, a number of U12-dependent introns hold a non-canonical GC or AT dinucleotide Padgett, 2012).

The branch site is characterized by an invariable A located within a highly degenerative region (Mercer et al., 2015) and is commonly located 21 to 34 nucleotides upstream of the 3’-end of the intron (Gao et al., 2008), although a considerable number of cases have been reported in which the BS is much more distant (Gooding et al., 2006; Corvelo et al., 2010).. The position at which the BS is located and the number of possible BSs have been reported as important features for the splicing mechanism, since increasing the distance between the BS and the 3’-end of the intron reduces its splicing efficiency (Cellini et al., 1986; Corvelo et al., 2010).

The polypyrimidine tract is a stretch of pyrimidines rich in uracil, usually 15 to 20 bp long, located downstream of the BS and often close to the 3’-end of the intron (Coolidge at al., 1997). However, in distant BSs the PPT remains located downstream of the BS (Corvelo et al., 2010). Finally, the 3’ss delimits the exon/intron boundary at the 3’-end of the intron and consists of an AG dinucleotide preferentially preceded by a T or a C (Padgett, 2012). U12-dependent introns that hold an AT dinucleotide in the 5’ss have an AC in this position. Despite the low information contained in the 3’ss signal, these 3 nucleotides seem to be necessary and sufficient for the splicing reaction in most introns (Wu et al., 1999).

3.2.2 The splicing reaction and the spliceosome

Biochemically, splicing consists of two cleavage-ligation reactions (Suzanne, 2008), which are transesterification reactions where one phosphodiester bond is exchanged for another. In the first reaction, the adenosine in the BS forms a phosphodiester bond with the guanosine in the 5’ss, releasing the 5’ exon from the intron and forming a lariat intermediate. In the second reaction, a phosphodiester bond is formed between the 3’-end of the released

exon and the guanine in the 3'ss, which results in the two exons being ligated together and in the release of the intron as a lariat structure.

Both cleavage-ligation reactions are catalyzed by the spliceosome, a large ribonucleoprotein assembly formed from five snRNPs (refer to Chapter 2.2.2) transiently associated to more than 700 non-snRNPs splicing factors (reviewed in Matera et al., 2014), which is dynamically assembled and disassembled at each round of splicing. The U2 or major spliceosome is formed by the U1, U2, U4, U5 and U6 snRNPs, which are responsible for the removal of U2-dependent introns from the precursor RNAs; whereas the U12 or minor spliceosome is formed by the U5, U11, U12, U4atac and U6atac snRNPs, which splices U12-dependent introns. From now on I will discuss the major spliceosome exclusively.

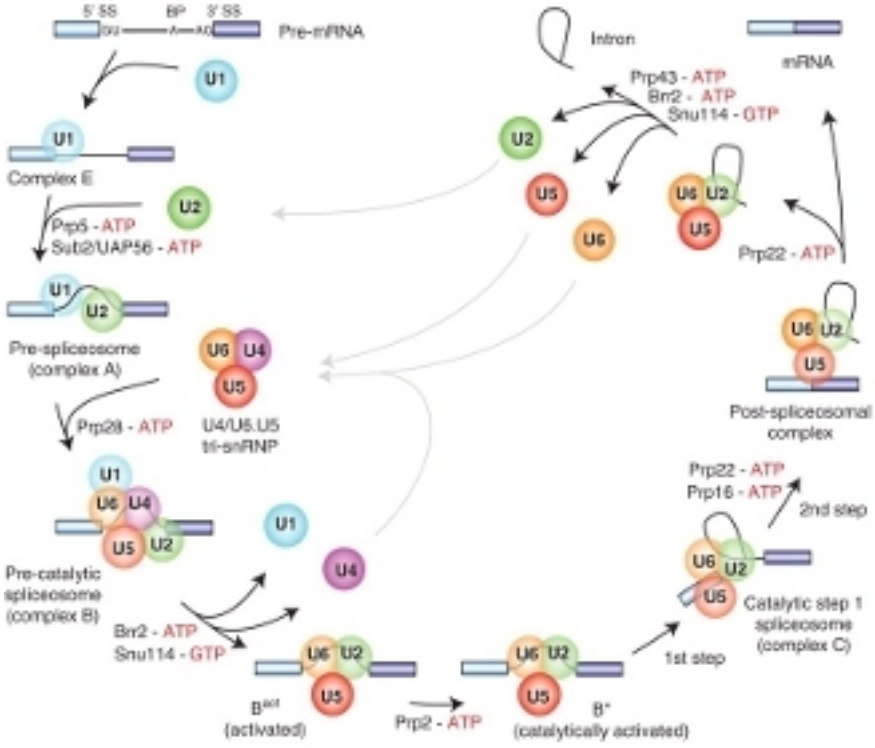


Figure 3-2. **Spliceosome cross-intron assembly and disassembly pathway**. For simplicity, only the ordered interactions of snRNPs (circles) are shown. Exons and introns are represented by boxes and lines, respectively. Adapted from Cindy and Lührmann, 2011.

The spliceosome assembly starts with the recognition of the 5'ss by the U1 snRNP through base-pairing with the sequence of nucleotides 3 to 10 of the U1 snRNA, followed by the binding of the U2 snRNP to the BS mediated by the sequence complementarity with the U2 snRNA that leaves the A unpaired. Recognition of the PPT and the 3'ss is carried on by the U2AF65 and U2AF35 splicing factors, respectively. Further recruitment of the U4/U5/U6 tri-snRNP forms the pre-catalytic spliceosome (B complex), which is converted to the catalytic step 1 spliceosome (C complex) after extensive conformational changes and remodelling, responsible of the two cleavage-ligation reactions. Figure 3-2 depicts the assembly and disassembly pathway by which the spliceosome performs the splicing of introns.

3.3 Alternative splicing

The term alternative splicing is used to define a regulated mechanism by which different forms of mature RNA are generated from the same precursor RNA. The first example of an alternatively spliced gene was described in 1981 for the mammalian gene encoding the thyroid hormone calcitonin, which was shown to produce two different transcripts containing 4 and 5 exons respectively (Leff and Rosenfeld, 1986). After calcitonin, many more examples of alternatively spliced genes were successively found, but these were considered to be the exception rather than the rule. By the time the Human Genome Project (HGP) was published in 2001, it was estimated that more than the 35% of human genes could undergo alternative splicing (Croft et al., 2000). At this point alternative splicing emerged as one of the most important mechanisms capable of generating a protein diversity that could potentially explain the complexity of a species whose genome encoded approximately 25000 genes, which is several orders of magnitude lower than other less complex organisms such as several types of plants. Today, thanks to the new generation of sequencing technologies, it is considered that more than the 95% of the human protein-coding genes undergo alternative splicing (Wang et al., 2010).

3.3.1 Types of alternative splicing events

According to the changes produced in the mature RNA there are five main types of alternative splicing events (Keren et al., 2010): exon skipping, alternative 3'ss selection, alternative 5'ss selection, intron retention and mutually exclusive exons (Figure 3-3). There are other mechanisms that can change exon composition in mature RNA, mainly alternative transcription start site selection, alternative polyadenylation site selection and trans-splicing events, the latter involving splicing reactions between two different precursor transcripts. These three forms of alternative splicing events are not covered in the present thesis, but the reader can refer to (Kimura et al., 2006; Tian et al., 2005) and (Labrador and Corcer, 2003) for excellent bibliography about them.

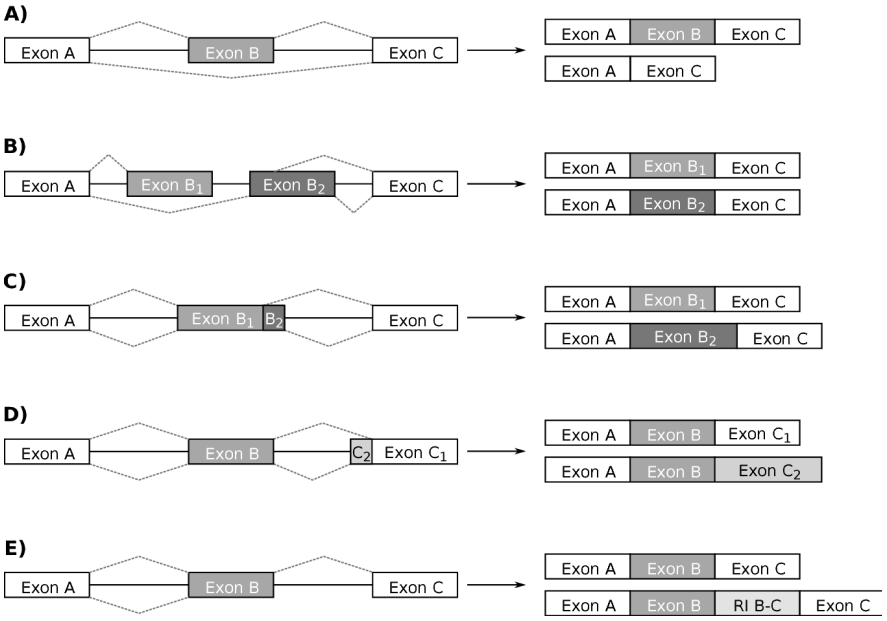


Figure 3-3. **Types of alternative splicing events**. (**a**) exon skipping, (**b**) mutually exclusive exons, (**c**) alternative 3'ss, (**d**) alternative 5'ss and (**e**) intron retention. We do not include alternative first and last exons in the figure, although these are generally considered as alternative splicing events too.

In exon skipping, an internal exon named cassette exon is spliced out of the transcript together with its flanking introns as a single

intron. Exon skipping accounts for nearly 40% of alternative splicing events in higher eukaryotes, but is extremely rare in lower eukaryotes (Alekseyenko et al., 2007). Alternative 3'ss and 5'ss selection events occur when two or more splice sites are alternatively recognized at one end of an exon, and account for 18.4% and 7.9% of all alternative splicing events in higher eukaryotes, respectively. Intron retention events, consisting of an intron remaining unspliced and integrated within the mature RNA, are one of the rarest alternative splicing event in vertebrates and invertebrates, accounting for less than 5% of known events. Finally, in the rare event of mutually exclusive exons, the splicing of two adjacent exons is coordinated, when one is included the other is skipped, and the other way around.

3.3.2 Regulation of alternative splicing

Splicing signals in the junctions of introns and alternatively spliced exons show a general tendency to deviate from the consensus sequences (Figure 3-1), resulting in a lowered affinity to the spliceosome, which leads to a reduced recognition (Stamm et al., 2005). The recognition of these degenerate splicing signals, commonly called weak splicing signals, is modulated by the presence of additional sequence elements located in the exon and nearby introns: exon splicing enhancers (ESE) and intron splicing enhancers (ISE), which enhance the recognition of weak splicing signals, and exon splicing silencers (ESS), and intron splicing silencers (ISS), which prevent it (Chasin, 2008). These enhancing and silencing functions are mediated by numerous regulatory proteins that bind to these sequence elements and interact with the splicing machinery to favor or disfavor the choice of weak splicing sites (Cáceres and Kornblihtt, 2002). These proteins are called splicing factors and include the SR and hnRNP families of proteins, and along with the sequence elements act in a combinatorial fashion in a way such that the balance of the competing enhancers and silencers determines the final splicing outcome.

The strength of the splicing sites and the interplay between the regulatory sequence elements and the splicing factors are not the only modulators of alternative splicing. The secondary structures adopted by the precursor RNA can occlude the splicing signals and prevent their recognition by the spliceosome components, or

shorten the distance between them and facilitate their recognition (Hiller et al., 2007; Shepard and Hertel, 2008; Warf et al., 2010). The coupling between transcription and splicing has also deep implications in the regulation of alternative splicing: changes affecting the RNAPII elongation rate (Kornblihtt, 2004), histone modifications (Spies, 2009; Kolasinska-Zwierz et al., 2009) or chromatin organization (Schwartz et al., 2009; Nahkuri et al., 2009) have also been shown to influence the recognition of splicing sites.

3.4 Regulation of alternative splicing by non-coding RNAs

Non-coding RNAs and components of their biogenesis pathways have recently emerged as important regulators of alternative splicing, and a few mechanistic models - that can be loosely divided into indirect and direct regulation models - have been proposed in the past few years.

One of the first evidences of the role of ncRNAs in the regulation of alternative splicing came from miRNAs, which were shown to indirectly regulate the splicing outcome by repressing the expression of key splicing factors (Makeyev et al., 2007; Kalsotra et al., 2010; Boutz et al. 2007). Similarly, the long non-coding RNA MALAT1 was shown to regulate alternative splicing of a set of pre-mRNAs by modulating the levels of SR proteins (Tripathi et al., 2010). More complex indirect mechanisms of alternative splicing regulation involve some members of the Argonaute family of proteins, especially AGO1 and AGO2, which beyond their roles as effectors of the miRNA-directed gene silencing pathway (refer to chapter 2.2.4), have been shown to moonlight to act as alternative splicing regulators by interacting with the spliceosome and by triggering changes in the chromatin organization (Alló et al., 2009; Ameyar-Zazoua et al., 2012).

The first evidence of direct regulation involved the C/D-box snoRNA HBII-52 (refer to chapter 2.2.3), which was shown to regulate the alternative splicing of the serotonin 2C receptor pre-mRNA (Kishore et al., 2006). The proposed model consists of HBII-52 being processed into smaller fragments that bind to a silencer element of the exon Vb in the serotonin 2C receptor thanks

to an 18 nucleotides conserved region of perfect complementarity, thus promoting the inclusion of exon Vb (Kishore et al., 2010; Kishore and Stamm, 2006). In a recent study in which I participated, we found that the C/D-box snoRNA SNORD27 regulates the alternative splicing of the transcription factor E2F7 pre-mRNA through direct RNA:RNA interaction, likely by competing with the U1 snRNP (Falaleeva et al., 2016).

In the present thesis I address the question of whether there are more non-coding RNAs, long or small, with the capacity to regulate alternative splicing by direct RNA:RNA interaction with the pre-mRNA, using a novel, unbiased, genome-wide approach.

4. Computational biology of RNA processing

The coupling between the traditional field of molecular biology and the emerging fields of bioinformatics and computational biology has become so intimate in the recent years that today it is almost impossible to conceive the study of RNA biology, at least in a genome-wide manner, by relying exclusively on purely wet-lab techniques. Although bioinformatics and computational biology are considered to be distinct fields, in terms of the principles they apply and the purposes they pursue, their boundaries are not clearly defined and there is undoubtedly a significant overlap between them, hence in the present thesis I will refer indistinctly to bioinformatics and computational biology as fields within the life sciences devoted to the development and application of computational tools and approaches to facilitate the interpretation of biological data, including those to acquire, store, organize, archive, analyze, or visualize such data.

In this chapter I will review the state of the art of the computational methods designed for the study of non-coding RNAs and alternative splicing, with an special emphasis on those that rely on the analysis of data generated by RNA-Seq sequencing experiments. An introduction to this type of the so called high-throughput sequencing technologies is then necessary prior to start dissecting how RNA processing is studied with the aid of computational tools.

4.1 RNA Sequencing

RNA sequencing (RNA-Seq) is a high-throughput sequencing method that parallelizes the sequencing process, delivering single base resolution, almost noise-free data. It is based on the sequencing of short fragments of cDNA obtained through the fragmentation of the cDNA molecules converted from a rRNA-depleted RNA sample called library. cDNA fragmentation is followed by ligation of 5’ and 3’ adapters, PCR amplification and sequencing, resulting in the generation of millions of short reads that can be mapped to a reference genome (Figure 4-1). The number of reads sequenced in each experiment is proportional to the original number of molecules

in the library, allowing direct quantification of RNA expression (Wang et al., 2009).

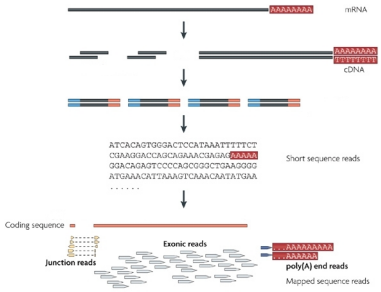


Figure 4-1. **Overview of the RNA-Seq technology**. Flow of an RNA-Seq experiment: long RNAs are converted to cDNA fragments, sequencing adaptors are added to the fragments and a short read is obtained for each cDNA. The sequencing reads are finally mapped to a reference genome or transcriptome. Adapted from Wang et al., 2009.

The number and length of the reads obtained in an RNA-Seq experiment depends largely on the sequencing platform used. On the other hand, a collection of reads may come in two different fashions: single-ended or paired-ended. While in single-end sequencing the cDNA fragments are sequenced from only one end, in paired-end sequencing the same fragments are sequenced from both ends, resulting in collections of reads better suited for downstream computational analysis at the expense of a higher monetary cost (Ozsolak and Milos, 2011). Although the landscape of RNA sequencing technologies is rapidly changing, and new sequencing platforms emerge with the promise of delivering faster, cheaper and more reliable sequencing methods, the Illumina short-read sequencing technology stands as a standard, and the choice for the vast majority of experiments published to date. Hence, from now on, whenever I refer to the reads obtained from an RNA-Seq experiment, the reader could assume that those have been obtained by means of the Illumina short read sequencing technology.

Prior to the amplification and sequencing steps, libraries can be built to capture RNA molecules that fulfil certain conditions, such as having a specific size range or having a polyadenylated tail. In small RNA-Seq (sRNA-Seq), designed to sequence miRNAs and other small RNAs, the cellular RNA is selected based on the desired size range (e.g. 15 to 30 nucleotides) and is then transformed to cDNA without being fragmented. After adapter ligation, the cDNA

molecules are sequenced from their 5'-ends, resulting in a collection of millions of single-end reads (Morin et al., 2008).

4.1.1 RNA-Seq read mapping

Read mapping, also called read alignment, is the process of aligning the collection of short reads coming from a high-throughput sequencing experiment to either a reference genome or transcriptome, and constitutes one of the most basic tasks in RNA-Seq analysis. However, the read mapping problem poses significant computational challenges, mainly because the number of reads per experiment can easily reach several hundreds of millions, and genome sizes are in the order of thousands of millions of base pairs (Garber et al., 2011). These two facts conspire to turn traditional alignment tools like BLAST (Altschul et al., 1990) or those based on dynamic programming algorithms (e.g. Smith-Waterman algorithm) (Smith and Waterman, 1981) to impracticable solutions to the read mapping problem.

To address the problem of large input size, both in number of reads and size of the references, two main algorithmic ideas have been applied to the read mapping problem: filtering and indexing (Reinert et al., 2015). Filtering methods exclude large regions of the reference where no approximate match with the read can be found. This is usually accomplished by identifying short regions in the reference (k-mers) that are discarded from the mapping process if they do not share a small piece of the read (seed) without errors. In addition to seeding filters, there are filters based on shared q-gram counts (Burkhardt et al., 1999) and on the pigeonhole lemma (Baeza-Yates and Navarro, 1999). Indexing strategies consist on the preprocessing of the reference sequence into string indices in a way such that the mapping of a short read does not require scanning the whole reference, drastically reducing the time of conducting queries at the expense of a larger memory consumption. Popular indexes currently used are the suffix array (Manber and Meyers, 1993), the enhanced suffix array (Abouelhoda et al., 2004) and the FM-index (Ferragina and Manzini, 2000).

Regardless of whether they incorporate filtering, indices or both, mappers (i.e. the tools that perform read alignment) can be divided into two broad groups: unspliced mappers, which align reads to a

reference without allowing any large gaps; and spliced mappers, which allow the presence of large gaps in the alignment, hence enabling the possibility to map reads derived from exon-exon junctions to a reference genome (reviewed in Engström et al., 2013 and in Alamancos et al. 2014).

Method	**Mapping to**	**Isoform quantification**	**Reference**
SAMMate	Genome	RPKM/FPKM	Xu et al. 2011
IsoformEx	Genome	RPKM	Kim et al. 2011
MISO	Genome	Isoform PSI	Katz et al. 2010
Alexa-Seq	Genome	Isoform expression level	Griffith et al. 2010
SOLAS	Genome	RPKM	Richard et al. 2010
Erange	Genome	RPKM	Mortazavi et al. 2008
rSeq	Genome	RPKM	Jiang et al. 2009
rQuant	Genome	RPKM	Bonhert et al 2009
FluxCapacitor	Genome	Isoform PSI	Montgomery et al. 2010
IQSeq	Genome	RPKM	Du et al. 2012
Cufflinks	Genome	FPKM	Trapnell et al. 2010
Casper	Genome	Isoform PSI	Rossell et al. 2012
CEM	Genome	Isoform expression level	Li et al 2012
IsoLasso	Genome	RPKM	Li et al 2011
IsoInfer	Genome	RPKM	Feng et al 2012
SLIDE	Genome	RPKM	Li et al 2011
RABT	Genome	RPKM	Roberts et al 2011
DRUT	Genome	FPKM	Mangui et al 20012

iReckon	Genome	RPKM	Mezlini et al 2013
RSEM	Transcriptome	TPM	Li et al 2011
IsoEM	Transcriptome	RPKM	Nicolae et al 2011
NEUMA	Transcriptome	FVKM	Lee et al 2011
BitSeq	Transcriptome	Isoform expression level	Glaus et al. 2012
MMSEQ	Transcriptome	Isoform expression level	Turro et al. 2011
eXpress	Transcriptome	FPKM	Roberts et al. 2013

Table 4-1. **Transcript quantification methods**. This table includes methods that use read mapping to reference genome or transcriptome to quantify annotated transcripts. Adapted from Alamancos et al. 2011.

4.2 Methods for the study of alternative splicing

4.2.1 Methods for transcriptome reconstruction and quantification

Expression quantification has long been an important application for the study of RNA populations in different tissues and conditions. When using RNA-Seq, the number of reads per nucleotide assigned to a given transcript serves as a surrogate for the original number of molecules of the same transcript in the sequenced sample. Once the reads have been assigned to a transcription unit, read counts need to be properly normalized to extract meaningful expression estimates due to two main sources of systematic variability inherent to the RNA-Seq technology: RNA fragmentation causes longer transcripts to generate more reads compared to shorter transcripts, and the number of fragments mapped across samples fluctuate due to the variability in the number of reads produced in each run (Marioni et al., 2008).

Three broad approaches have been proposed for estimating the set of transcripts in RNA samples using RNA-Seq data (Janes et al., 2015). The first one, and the simplest, consists on assuming that the transcripts in a sample are a subset of transcripts from a curated

annotation, such as RefSeq (Pruitt et al., 2012). In this approach, reads are aligned to a reference genome or to a transcriptome derived from the annotation, and statistical models are used to estimate expression of the annotated transcripts. The second is a more ambitious strategy which involves the alignment of reads to a reference genome and the use of these alignments to infer the transcript structure and expression. The last and most challenging approach is to assemble reads into transcript structures without the aid of a genome reference. From now on I will refer only to methods that follow the first approach.

One of the first methods for transcript quantification, which also provided the foundations for the computational analysis of quantitative transcriptome sequencing, was Erange (Mortazavi et al., 2008), where reads mapped to the exons and known junctions were distributed in isoforms, the expression levels of which were calculated in terms of Reads per Kilobase per Million Reads (RPKM). However, the uncertainty in the assignment of reads shared by two or more isoforms was not appropriately modeled. Since then, dozens of tools that overcome this drawback by implementing isoform disambiguation methods have been published. Table 4-1 summarizes the main isoform quantification tools published to date along with a brief description of the underlying methods implemented in each of them. Some of these tools report isoform quantifications in terms of isoform expression level values such as RPKM, FPKM (Trapnell et al., 2010) or TPM (Li et al., 2010), and other in terms of a relative expression value. All the tools listed in Table 2 use spliced or unspliced mappers to align the reads to a reference genome or transcriptome.

On top of these methods, a number of recently published alignment-free tools have revolutionized the area of computational isoform quantification due to their ability to complete transcriptome quantifications in a matter of minutes, without compromising the accuracy of the estimations. The first published alignment-free tool was Sailfish (Patro et al., 2014), which is based on the definition of k-mers that identify transcript sequences from a given transcriptome. Sailfish bypasses the read-mapping step and directly estimates transcript coverage by counting the k-mers occurring in reads. Inspired by Sailfish, several other tools soon followed similar principles or introduced mild modifications to improve either the

speed or the accuracy. RNA-Skim (Zhang et al., 2014), Salmon (Patro et al., 2015) and, more recently, Kallisto (Bray et al., 2015), are the pioneering examples of this new generation of isoform quantification tools (reviewed in Dapas, 2016).

4.2.2 Methods for the quantification of alternative splicing events

The majority of tools for the computational quantification of alternative splicing events are strongly dependent on the mapping of RNA sequencing reads to the genome, and often rely on the existing annotation to guide the prediction of the events. Like in the previous section, I will focus only on methods that are genome and annotation dependent or transcriptome dependent.

First reports using RNA-Seq to quantify splicing were based on the analysis of junctions built from known gene annotations (Wang et al., 2008), where splicing events were quantified in terms of exon inclusion ratios. The most popular metric of exon inclusion ratio is the ‘percent spliced in’ (PSI - Ψ), defined as the ratio of inclusion reads to inclusion plus exclusion reads, where inclusion reads align to candidate alternative exons and its junctions and exclusion reads align to flanking constitutive exons and its junctions (Figure 9). Based on this approach, various tools have been developed recently, which differ on whether the reads mapping on exons are used for the exon inclusion ratio calculations, whether they provide the mapping step or not, or whether a statistical model is used for estimating the exon inclusion ratios. MMES (Wang et al., 2010), SpliceTrap (Wu et al., 2011), RUM (Grant et al., 2011), SpliceSeq (Ryan et al., 2012), MISO (Katz et al., 2010), Alexa-Seq (Griffith et al., 2010) and SOLAS (Richard et al., 2010) figure amongst the most widely used exon and junction count tools for alternative splicing events quantification. Nonetheless, all of them rely on the mapping of reads to the genome and/or to the annotation to build the different events.

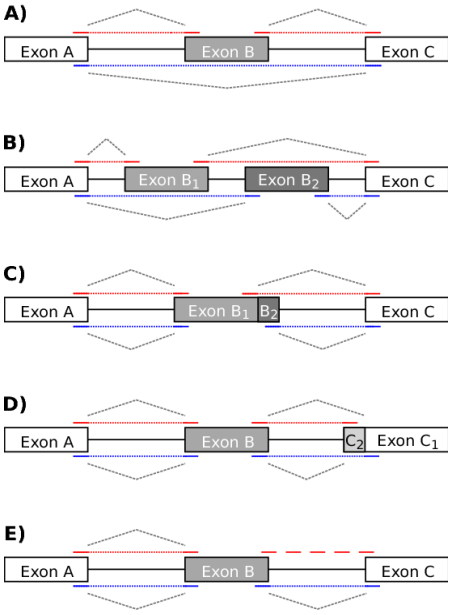


Figure 4-2. **Exon inclusion levels**. When using junction reads, PSI values are calculated as the ratio of the number of reads supporting the inclusion (read) to the number of reads supporting the inclusion and the exclusion (blue) of an alternative exon (**a**, **b**), alternative exon extension (**c**, **d**) or retained intron (**e**). Dashed lines represent the gaps of the reads over the exon-exon junctions. Solid lines represent the body of reads describing exonic regions.

An alternative approach to calculate exon inclusion levels was first used in a study conducted in cancerous breast tissues (Venables et al., 2008). In this study, researchers identified 600 different genes that had an exon skipping event, resulting in the transcription of two different isoforms, one including the exon (longer isoform) and the other skipping the exon (shorter isoform). The concentration of each of the isoforms was assessed by RT-PCR, and for each exon a PSI value was calculated as the ratio of the concentration of the longer isoform to the sum of concentrations of the longer and shorter isoforms.

In this thesis I present SUPPA, a novel method for alternative splicing event quantification that exploits the transcript abundances for estimating exon inclusion levels by taking advantage of the state of the art alignment-free isoform quantification methods (refer to chapter 4.2.1), thus delivering accurate exon inclusion levels at an unforeseen speed and enabling the systematic splicing analyses of large datasets with limited computational resources.

4.3 Methods to study non-coding RNAs

Understanding the function of non-coding RNAs in the age of high-throughput experiments has been largely possible due to the emergence of new computational approaches for the structural analysis, discovery of RNA:RNA interactions and annotation of genomic data (Washietl et al., 2012). Although a close connection exists between structure and function for many noncoding RNAs, structural analysis is not a central theme in this thesis. Thus this chapter will not review any of the methods devoted to this purpose. The reader can refer to (Backofen et al., 2014) for excellent literature on this topic.

To exert their functions, ncRNAs interact with a wide spectrum of biological molecules, including mRNAs and other non-coding RNAs. For example, by direct RNA:RNA base-pairing interactions, miRNAs and snoRNAs regulate the expression and the alternative splicing of their target mRNAs respectively (refer to chapters 2.2.4 and 3.4). Predicting RNA:RNA interactions can thus elucidate RNA interaction partners and potential novel functional mechanisms. Accordingly, a plethora of methods have been developed for the identification of non-coding RNA targets. While most of these methods focus on the prediction of miRNA targets (reviewed in Reyes and Ficarra, 2012; Watanabe et al., 2007), a number of tools have been designed for the prediction of ncRNA targets without assuming any specific constraints in the binding patterns (reviewed in Lai and Meyer, 2015). These tools include IntaRNA (Wright et al., 2014), GUUGle (Gerlach and Giegerich, 2006), RactIP (Kato et al., 2010), RNAup (Mückstein et al., 2006) and LncTar (Li et al., 2015), the latter specifically designed for predicting RNA targets of long non-coding RNAs. In general these methods establish limits in the search space of the RNA target to speed up the search and require the user to input parameters which are difficult to determine in advance, such as the region of interaction between both RNAs in the case of RNAup. In this thesis I describe STSCAN, a sequence complementarity based algorithm for target prediction that exhaustively scans for binding sites in the target molecule in linear time, thus enabling genome-wide searches in a reasonable time, and requires minimal parameter inputs from the user.

Another major challenge in understanding the function of ncRNAs is to find, characterize and annotate them in complete genomes. Dozens of tools have been designed to fulfill this purpose, either for specific families of ncRNAs (miRNAs, tRNAs, rRNAs, snoRNAs) (Kang and Friedländer, 2015; Lowe and Eddy, 1997; Lagesen et al., 2007; Hertel et al., 2008) or in a generic fashion. A common approach for the generic classification of non-coding RNAs is the searching of sequence and structural homology in specialized databases or genomics data (Nawrocki and Eddy, 2013). Another interesting approach is the use of read profiles, i.e. the distinct coverage patterns formed by the reads coming from an RNA-Seq experiment when mapped to the reference genome (Pundhir et al., 2015). A read profile essentially represents the positional arrangement of reads along a specific region in the genome, and can be regarded as the footprint of the processing steps through which this transcribed genomic region has undergone. In this direction, a number of methods capable of detecting new members of known RNA families have been published in the past few years, mainly deepBlockAlign (Langerberger et al., 2012) and more recently BlockClust (Videm et al., 2014), which operate under the assumption that ncRNAs showing similar read profiles would likely belong to the same ncRNA family. These methods show low to moderate accuracy when predicting certain known classes of ncRNAs, are often slow and require large amounts of computer memory.

In this thesis I describe SeRPeNT, a fast and memory efficient profile-based tool for the discovery and annotation of small non-coding RNAs. SeRPeNT not only annotates novel sRNAs that belong to known classes with higher accuracy than previous methods, but also detects potential novel ncRNAs families.

OBJECTIVES

The general objectives of this thesis are:

General Objective 1 (GO1)
The development of novel, efficient computational methods to study how RNAs are processed in living cells

General Objective 2 (GO2)
To elucidate the mechanisms of noncoding RNA mediated regulation of alternative splicing

These two general objectives can be materialized into four different concrete objectives:

Concrete Objective 1 (CO1)
To develop a method for the identification of small RNAs from size-selected RNA-seq experiments data

Concrete Objective 2 (CO2)
To develop a method for the fast quantification of alternative splicing events oriented to the analysis of large datasets

Concrete Objective 3 (CO3)
To develop a method to find binding sites between pairs of RNAs.

Concrete Objective 4 (CO4)
To find a set of small RNAs with the ability to regulate the alternative splicing of a number of mRNAs or long non-coding RNAs

RESULTS

5 The discovery potential of RNA processing profiles

Pagès A, Dotu I, Pallarès-Albanell J, Martí E, Guigó R, Eyras E. The discovery potential of RNA processing profiles. Nucleic Acids Res. 2018 Feb 16;46(3):e15–e15. DOI: 10.1093/nar/gkx1115

6 Leveraging transcript quantification for fast computation of alternative splicing profiles

Alamancos GP, Pagès A, Trincado JL, Bellora N, Eyras E. Leveraging transcript quantification for fast computation of alternative splicing profiles. RNA. 2015 Sep;21(9):1521–31. DOI: 10.1261/rna.051557.115

7 STSCAN: prediction of sRNA:pre-mRNA interactions with the potential to regulate alternative splicing

Pagès A, Falaleeva M, Guigó R, Stamm S, Eyras E.

STSCAN: prediction of sRNA:pre-mRNA interactions with the potential to regilate alternative splicing.

Manuscript in preparation.

STSCAN: prediction of sRNA:pre-mRNA interactions with the potential to regulate alternative splicing

Amadís Pagès[1,2], Marina Falaleeva[3], Roderic Guigó[1,2], Stefan Stamm[3], Eduardo Eyras[1,4] *

[1]Universitat Pompeu Fabra (UPF), E08003 Barcelona, Spain.
[2]Centre for Genomic Regulation (CRG), E08003 Barcelona, Spain.
[3]Department of Molecular and Cellular Biochemistry, University of Kentucky, Lexington, KY 40536
[4]Catalan Institution for Research and Advanced Studies, E08010 Barcelona, Spain.

*To whom correspondence should be addressed

Abstract

Alternative splicing is a key molecular mechanism in the processing of the precursor RNAs and contributes to the increase of transcriptome complexity in most eukaryotes. Although alternative splicing is mainly regulated by the combinatorial control of multiple protein factors that enhance or repress the recognition of splice-sites by the spliceosome, recent evidence points to alternative mechanisms of regulation involving the base-pairing of small non-coding RNAs to the nascent pre-mRNA to affect its splicing. Here we test whether there is a general mechanism of splicing regulation by the base pairing of non-coding RNAs to pre-mRNAs. We describe a new method, STSCAN, based on finite state machine theory, for the fast and exhaustive identification of putative RNA:RNA binding sites, which we test on a set of experimentally validated targets. Furthermore, we describe non-coding RNAs that function as potential regulators of alternative splicing. Our analysis provides evidence for a new layer of splicing regulation by the direct interaction of non-coding RNAs to the pre-mRNA.

Introduction

Alternative splicing is a key molecular mechanism by which the same precursor RNA is processed into distinct mature RNAs (Black 2003). Alternative splicing affects nearly all the human protein-coding genes as well as a significant number of long non-coding RNAs, and bears major importance in the proper regulation of many cellular processes (Wang et al. 2008). Hence, its dysregulation is linked to multiple diseases, including cancer (Wang & Cooper 2007). Alternative splicing is mainly regulated through the transient activity of ribonucleoprotein complexes that bind to regulatory sequences located on and around exons, which act either by enhancing or repressing the recognition of nearby splice sites (Black 2003). Alternative splicing can also be modified through the transfection in cells and animal models of a variety of molecules. Synthesized antisense oligonucleotides can modulate the splicing of events in a very specific way to recapitulate different cellular phenotypes (Bechara et al., 2013; Sebestyén et al., 2016). Similarly, designed small-interfering RNAs (siRNAs) can be directed to specific gene loci to trigger changes in the local chromatin that indirectly affect the splicing of a gene (Alló et al., 2009; Alló et al., 2014). Furthermore, RNA sequences engineered to target the branch-point of a specific gene have been shown to induce splicing changes in the target gene (Semenov et al., 2012). These results raise the question of whether similar mechanisms take place endogenously in human cells. Recent evidence indicates that small nucleolar RNAs (snoRNAs) could act as potential regulators of alternative splicing through direct RNA:RNA interactions with the pre-mRNA. snoRNAs are a highly abundant type of short noncoding RNA that assemble into large ribonucleoprotein complexes in order to direct chemical modifications of rRNAs and snRNAs with which the snoRNAs share a region of almost perfect complementarity. Although many snoRNAs identified so far through sequence and secondary structure studies are known to carry out this canonical function, many remain of unknown function, also called orphan, since they lack complementarity with known rRNAs or snRNAs. Notably, one of such orphan snoRNAs was implicated in alternative splicing regulation. The snoRNA SNORD115 (HBII-52) binds to a region of perfect complementarity of the 5-Hydroxytryptamine (Serotonin) Receptor 2C (HTR2C) pre-

mRNA and promotes the inclusion of an alternative exon, likely by blocking an intronic splicing silencer located nearby the alternative exon (Kishore et al. 2010). Similarly, the snoRNA SNORD27 (U27), which is known to guide the 2'-O-ribose methylation of the 18s rRNA, also binds to a 29 nt complementary region, including two G:U base pairs and four mismatches, of the transcription factor E2F7 pre-mRNA (Falaleeva et al. 2016). In this case, the complementarity region encompasses seven of the nine bases of the 5' splice site of an alternative exon, thus masking its recognition and promoting exon exclusion (Falaleeva et al. 2016).

Despite these evidences, it is still unknown whether there are other non-coding RNAs with the potential to regulate alternative splicing by direct RNA:RNA interaction. Moreover, the mechanisms of action of these RNAs over their target precursor RNAs are still to be elucidated. With the aim of identifying general mechanisms of non-coding RNA mediated regulation of alternative splicing through direct RNA:RNA interactions with the pre-mRNA, we developed STSCAN, a novel algorithm based on finite state machine theory, for the fast and exhaustive identification of putative RNA:RNA binding sites. Using RNA sequencing on cells with the knockdown of SNORD116 and controls, we applied STSCAN to identify possible regulatory modes of splicing by the interaction of SNORD116 with pre-mRNAs.

Methods

Identification and scoring of binding sites

To identify regions of sequence complementarity between two RNA molecules, typically a sRNA (the query sequence, of length *m*) and a pre-mRNA (the target sequence, of length *n*), we developed STSCAN, an algorithm based on finite state machine theory that exhaustively finds all the longest possible regions of sequence complementarity (the binding sites) between the query and target sequences that include a subregion of perfect complementarity (the seed). STSCAN takes as parameters the minimum length of the seed (*sml*), the minimum length of the target site (*bml*) and the maximum number of mismatches (*mnm*); and reports the longest possible binding sites that fulfill the following criteria: (1) length of

the binding site >= *bml*, (2) length of the seed >= *sml* and (3) number of mismatches in the binding site <= *mnm*. G:U is not considered as base pair of perfect complementarity for the seed, neither a mismatch of the overall predicted binding site. The STSCAN algorithm proceeds through 5 steps:

1. The query sequence is reverse complemented.
2. A finite state machine is built from the query sequence.
3. The finite state machine is used for seed recognition on the target sequence. Seeds shorter than *sml* are discarded.
4. Binding sites are extended from the seeds until the maximum number of mismatches is exceeded, discarding those binding sites that are shorter than *bml*.

STSCAN performs the first operation in linear $O(m)$ time, the second operation in quadratic $O(m^2)$ time, and the two last operations in linear $O(n)$ time, enabling the identification of target sites at genome scale since typically the target sequence is much larger than the query sequence ($m << n$). More details of the finite state machine building algorithm are provided in the Supplementary Methods.

The score of each binding site is calculated using the Nearest Neighbor Database (NNDB) (Turner and Mathews. 2010), which contains experimentally obtained parameters for the prediction of free energy changes of different structural elements. We adopted the scoring transformation scheme from RIsearch (Wenzel et al. 2012) and multiplied by -100 the original scores in NNDB to build the STSCAN scoring matrix.

Quantification of alternative splicing events

We quantified the alternative splicing events in two knockdown experiments of the snoRNA SNORD116 and their respective controls (Falaleeva et al. In preparation). We first downloaded the RefSeq annotation (Release 75) from UCSC and used the *eventGenerator* operation of the *SUPPA* software (Alamancos et al. 2015) to obtain the set of alternative splicing events. We then downloaded the sequence of the transcripts annotated in RefSeq and used Salmon (Patro et al. 2015) to quantify their expression levels for each of the RNA-seq knockdown and control samples. Finally,

we used *SUPPA*'s *psiPerEvent* operation to quantify the annotated alternative splicing events from the isoform expression values in each of the samples. Quantification of alternative splicing events was reported as 'percent spliced in' (PSI or Ψ) values, which represent an estimate of the relative abundance of a particular alternative splicing event (i.e. an alternative exon, an alternative splice site or a retained intron).

Identification of exons regulated by SNORD116

To build the set of exons regulated by the snoRNA SNORD116 we selected those exons that did not change splicing between replicates but that substantially changed splicing between control and knockdown samples. Let C_1 and C_2 be the two control replicates, and KD_1 and KD_2 the two SNORD116 knockdown replicates, and let $\Psi(C_1, n)$, $\Psi(C_2, n)$, $\Psi(KD_1, n)$ and $\Psi(KD_2, n)$ be the PSI values assigned to the alternative splicing event n in each of the samples, we defined the event n as regulated by SNORD116 if it fulfilled the following conditions:

1. $|\Psi(C_1, n) - \Psi(C_2, n)| < 0.05$
2. $|\Psi(KD_1, n) - \Psi(KD_2, n)| < 0.05$
3. $\forall i,j \in \{1,2\}: |\Psi(C_i, n) - \Psi(KD_j, n)| > 0.25$

Identification of exons not regulated by SNORD116

To build the set of exons that are not regulated by the snoRNA SNORD116 we selected those exons that did not change splicing between replicates neither between control and knockdown samples. Let C_1 and C_2 be the two control replicates, and KD_1 and KD_2 the two SNORD116 knockdown replicates, and let $\Psi(C_1, n)$, $\Psi(C_2, n)$, $\Psi(KD_1, n)$ and $\Psi(KD_2, n)$ be the PSI values assigned to the alternative splicing event n in each of the samples, we defined the event n as not regulated by SNORD116 if it fulfilled the following conditions:

4. $|\Psi(C_1, n) - \Psi(C_2, n)| < 0.05$
4. $|\Psi(KD_1, n) - \Psi(KD_2, n)| < 0.05$
4. $\forall i,j \in \{1,2\}: |\Psi(C_i, n) - \Psi(KD_j, n)| < 0.05$

Enrichment value calculation

Let *r* be a regulated exon, $U = (u_1, \ldots, u_m)$ the set of non-regulated exons and let *S(e, x)* be the score of the nucleotide that is *x* bases upstream (*-500 <= x <= -1*) or downstream (*1 <= x <= 500*) of an exon *e*.

Let's now define *S(e, x)* as the maximum score amongst the targets that overlap nucleotide *x* in exon *e*. We then calculate the enrichment value *Z(r, x)* of nucleotide *x* in the regulated exon *r* as a z-score calculated as follows:

$$Z(r,x) = \frac{S(r,x) - \mu_{U,x}}{\sigma_{U,x}}$$

where:

$$\mu_{Ux} = \frac{1}{m}\sum_{i=1}^{m} S(U_i, x) \quad \text{and} \quad \sigma_{Ux} = \frac{1}{m}\sum_{i=1}^{m} \left(S(U_i, x) - \mu_{U,x}\right)^2$$

Results and discussion

STSCAN discriminates putative regulatory target sites

To validate the accuracy of STSCAN, we used an experimentally confirmed dataset of 54 fungal snoRNA:rRNA interactions for 59 snoRNAs and 2 rRNAs, and 109 bacterial sRNA:mRNA interactions for 27 sRNAs and 90 mRNAs (Lai and Meyer 2015). STSCAN was run on each pair of fungal RNAs (snoRNA:rRNA) and each pair of bacterial RNAs (sRNA:mRNA) that had an experimentally validated interaction, imposing a minimum seed length of 5 bases, a minimum target length of 10 bases and a maximum of 2 mismatches.

The number of distinct interacting sites obtained for each pair of query and target sequence greatly depends on the length of both of them (Supplementary Figure 1), thus it may contribute to an increased number of false positive binding sites. For the test data we

obtained a total of 4,337 target sites for the fungal snoRNA:rRNA dataset and 5,550 target sites for the bacterial sRNA:mRNA dataset. From these, we recovered 93 and 51 experimentally validated putative interactions for the two sets, which accounts for the 85% and 86% of validated interactions respectively (Supplementary Data 1). Notably, STSCAN scoring system tends to rank higher the experimentally validated sites amongst the rest of predicted sites (Figure 1 and Supplementary Data 1). This discriminative power is higher for the snoRNA:rRNA interactions in the fungal dataset, possibly due to these interactions having in general small bulges but almost no internal loops.

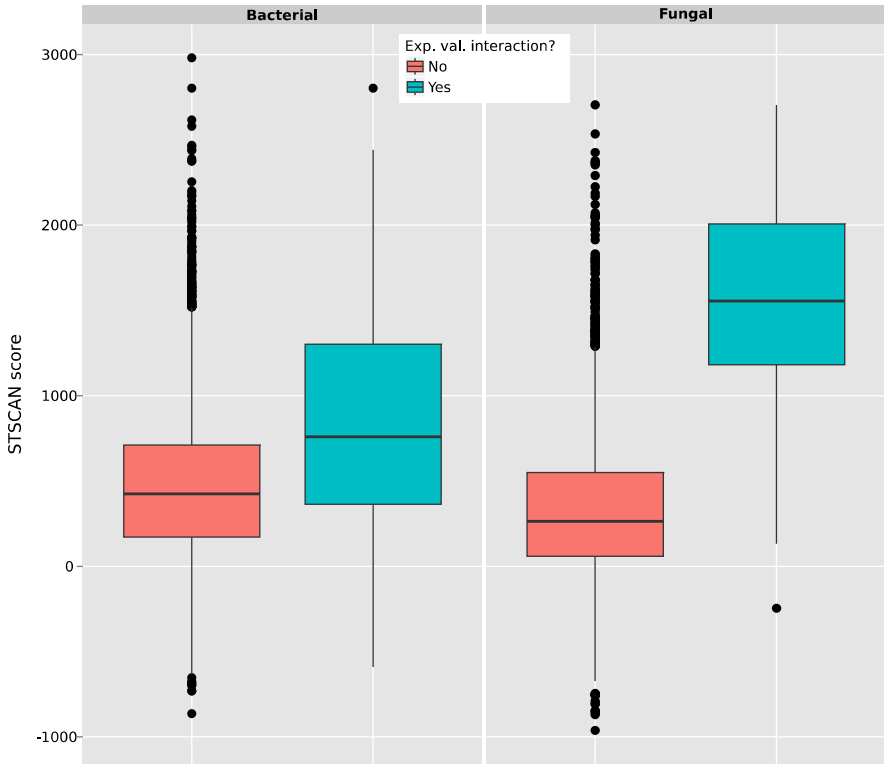


Figure 1. Distribution of the scores (y-axis) for the interactions predicted with STSCAN for the bacterial (left panel) (Mann-Whitney U-test p-value = $9.04810e^{-15}$) and fungal (right panel) (Mann-Whitney U-test p-value = $2.210e^{-16}$) datasets comparing the experimentally validated interactions (blue) and other non-validated interactions (red).

Potentially splicing regulatory interaction sites are located nearby splicing signals

We used knockdown experiments of the snoRNA SNORD116 compared with controls to search for potential interactions of SNORD116 with pre-mRNAs that may impact splicing regulation. Using stringent conditions of reproducibility across replicates and for the splicing change between conditions (Methods), we identified 6 differentially spliced exon skipping events, 5 of which showed an increase in the inclusion of the cassette exon upon the depletion of SNORD116 (Table 1). The remaining event showed a decrease in the inclusion of the cassette exon in the absence of SNORD116 (Table 1). This suggests that SNORD116 might be promoting the inclusion of this target regulated exon, and possibly other exons, while promoting the exclusion of other of its target regulated exons. We also identified 755 exons that did not change splicing upon knockdown of SNORD116 (Methods).

Host gene	Chromosome	Coordinates	Strand	Effect of SNORD116
MKI67	chr10	129913192-129914271	-	Promotion of exon skipping
STOX2	chr4	184930311-184932576	+	Promotion of exon skipping
AKAP13	chr15	86201768-86201821	+	Promotion of exon skipping
SON	chr21	34921782-34927697	+	Promotion of exon skipping
BBX	chr3	107491475-107492483	+	Promotion of exon skipping
SLC37A2	chr11	124956100-124956156	+	Promotion of exon inclusion

Table 1. List of cassette exons regulated by the snoRNA SNORD116, including the host gene, the exon genomic coordinates and the regulatory effect that SNORD116 would exert on the exon as derived from the knockdown experiment.

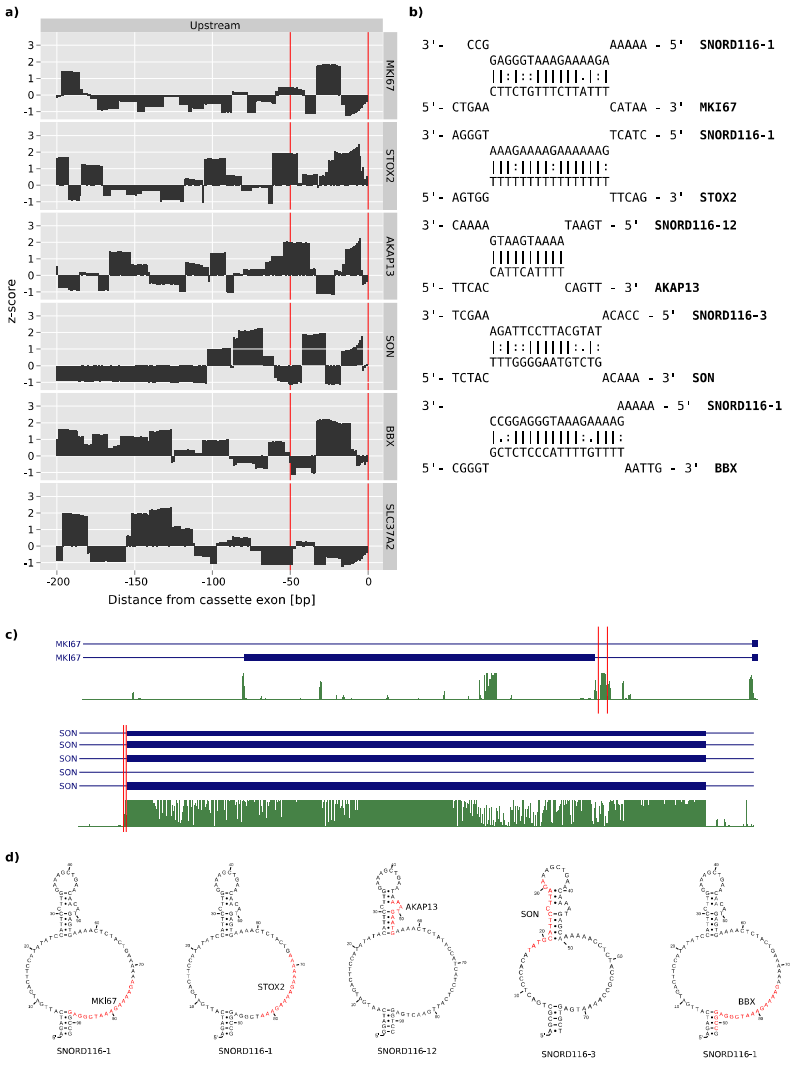


Figure 2. (**a**) Score enrichment of SNORD116 interaction sites in the flanking region upstream of the cassette exon and (**b**) depiction of the interaction sites in the region immediately upstream of the exon's 3'ss, indicated between red lines in (a) (**c**) Conservation track from the UCSC genome browser showing conservation of the interaction sites (between red lines). (**d**) Regions in SNORD116 for the interaction sites in Figure 2b. Secondary structure and sequence are from Ensembl release 75.

We ran STSCAN, with the same parameter configuration used before and predicted 11.020 interaction sites between 28 copies of the SNORD116 and the ± 200 bp intronic regions flanking the exons of the regulated events. We also ran STSCAN between the SNORD116 copies and the exon flanking regions of the non-regulated events, obtaining 530.601 interaction sites. We then aligned the predicted interaction sites on the flanking regions of the target events and calculated nucleotide-based enrichment values of the scores of the interaction sites for the regulated events respect to the non-regulated events (Methods). We observed that all the events that show an increase in the exon inclusion upon depletion of SNORD116 have at least one target enriched in the region comprised within the 50 base pairs immediately upstream of the acceptor site (Figures 2a and 2b) (Supplementary Figure 2). It is known that this region usually contains the branch point (Corvelo et al., 2010), one of the three obligatory signals required for pre-mRNA splicing. Overall, these results suggest that SNORD116 might be binding to the region immediately upstream of its target cassette exons, inhibiting the recognition of the cassette exon, thus leading to the promotion of its skipping. The lack of targets in this region for the exon that shows a decrease in the inclusion upon depletion of SNORD116 (SLC37A2) suggests that the snoRNA might be binding to another regulatory region nearby such as a splicing silencer, thereby enhancing the recognition of the cassette exon. We also observed that some of the interaction sites were located in genomic regions that are highly conserved among vertebrates (Figure 2c), which points to the existence of an evolutionary constraint favoring the persistence of these regions and hence providing an extra layer of evidence to our hypothesis.

We aligned the interaction sites nearby the branch points of the five events downregulated by SNORD116 (i.e. that increase inclusion upon depletion of SNORD116) to its correspondent SNORD116 copy (Figure 2d). We observed that SNORD116-1 targets events in genes MKI67, STOX2 and BBX, and that all three targets span approximately the same region of the snoRNA, close to the stem in the 3’-end. Moreover, this region of the C/D-box snoRNAs is known to yield snoRNA-derived small RNAs (sdRNAs) (Chen and Heard, 2103), which agrees with previous findings reported in (Kishore et al., 2010) stating that the snoRNA HBII-52 is processed into smaller RNAs and that those smaller products are the effectors

of the regulation of alternative splicing. We also observed that the interaction sites in the snoRNAs SNORD116-12 and SNORD116-3 are located between nucleotides 47 to 54 and between nucleotides 17 to 31 respectively, regions close to the center of the snoRNA that have been also reported to yield sdRNAs in C/D-box snoRNAs (Taft et al., 2009).

Conclusions

We have developed STSCAN, a novel computational algorithm that finds RNA:RNA interaction sites in an unbiased and exhaustive manner. Requiring minimal input from the user (seed length, site length and number of mismatches), STSCAN reports all the possible interaction sites of the query RNA on the target RNA that fulfills the user requirements. We applied STSCAN to two datasets of experimentally validated RNA:RNA interactions in bacteria and fungi and showed that STSCAN discriminates the putative interaction sites in both of them.

We also applied STSCAN to a set of exons regulated by the snoRNA SNORD116. Our results indicate that a frequent mechanism of regulation of splicing by SNORD116 is based on the occlusion of possibly the branch-point, poly-pyrimidine tract or other splicing regulatory signals nearby the splice-sites (Kishore et al., 2010; Falaleeva et al., 2016; Corvelo et al., 2010). Our results also indicate a possible direct binding of the small non-coding RNAs on other regions potentially harboring other splicing regulatory signals, in particular intronic silencers.

References

Alamancos, G. P. et al. (2015) Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA* **21**, 1521-1531.

Alló, M. et al. (2009) Control of alternative splicing through siRNA-mediated transcriptional gene silencing. *Nature structural & molecular biology* **16**, 717-724.

Alló, M. et al. (2014) Argonaute-1 binds transcriptional enhancers and controls constitutive and alternative splicing in human cells. *Proceedings of the National Academy of Sciences* **111**, 15622-15629.

Bechara, E. G. et al. (2013) RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. *Molecular cell* **52**, 720-733.

Black, D. L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**, 291-336.

Chen, C. J. and Heard, E. (2013) Small RNAs derived from structural non-coding RNAs. *Methods* **63**, 76-84.

Falaleeva, M. et al. (2016) Dual function of C/D box small nucleolar RNAs in rRNA modification and alternative pre-mRNA splicing. *PNAS*, **advance access**.

Kishore, S. et al. (2010) The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. *Human molecular genetics* **19**, 1153-1164.

Lai, D., and Meyer, M. (2015) A comprehensive comparison of general RNA–RNA interaction prediction methods. *Nucleic acids research*, **advance access**.

Patro, R. et al. (2015) Accurate, fast, and model-aware transcript expression quantification with Salmon. *bioRxiv*, http://dx.doi.org/10.1101/021592.

Sebestyén, E. et al. (2016) Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Research*, **advance access**.

Semenov, D.V. et al. (2012) Analogues of artificial human box C/D small nucleolar RNA as regulators of alternative splicing of a pre-mRNA target. *Acta Naturae* **4**, 2012.

Taft, R. et al. (2009) Small RNAs derived from snoRNAs. *Rna* **15**, 1233-1240.

Turner, D.H., and Mathews, D.H. (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic acids research* **38**, D280-D282.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* **27**, 470-6.

Wang, G. S. and Cooper, T. A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics* **8**, 749-761.

Wenzel, A. et al. (2012) RIsearch: fast RNA–RNA interaction search using a simplified nearest-neighbor energy model. *Bioinformatics* **28**, 2738-2746.

Supplementary Methods

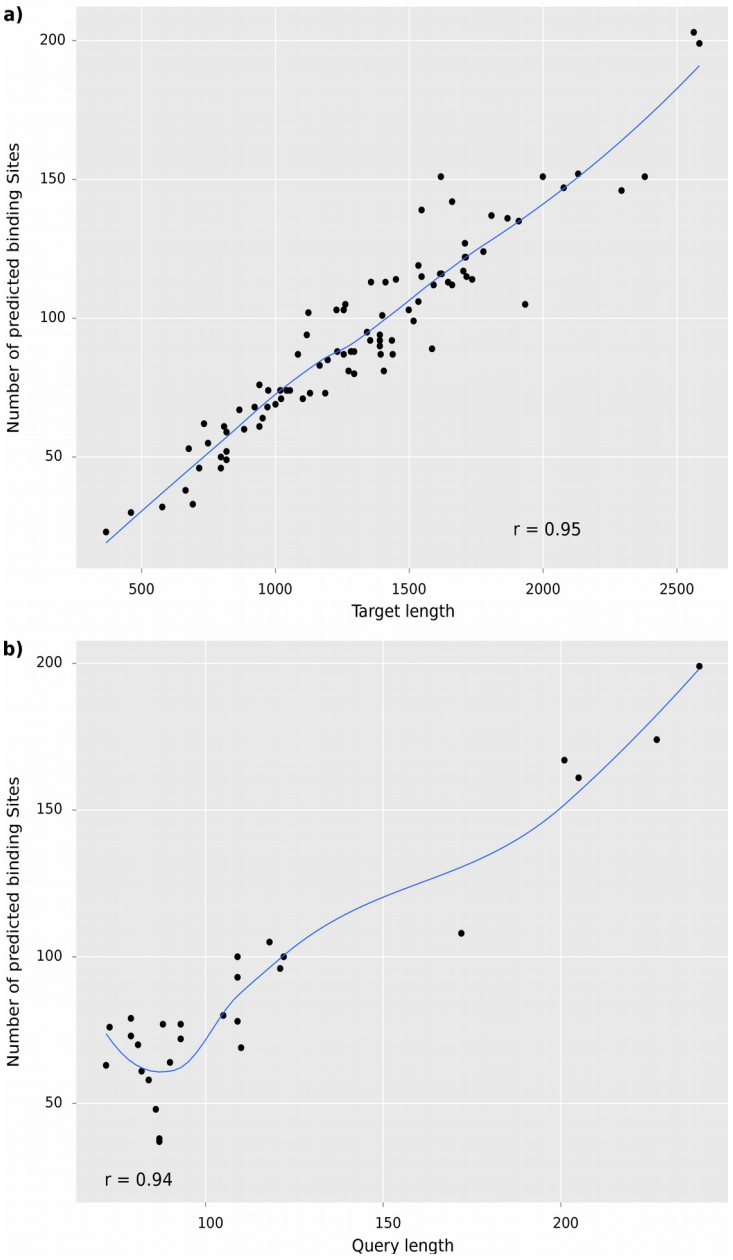
Finite state machine building algorithm

Given a sequence S of length n, the finite state machine (FSM) that recognizes all the subsequences of S is defined as a 4-tuple (Σ, Q, δ, q_0) consisting of:

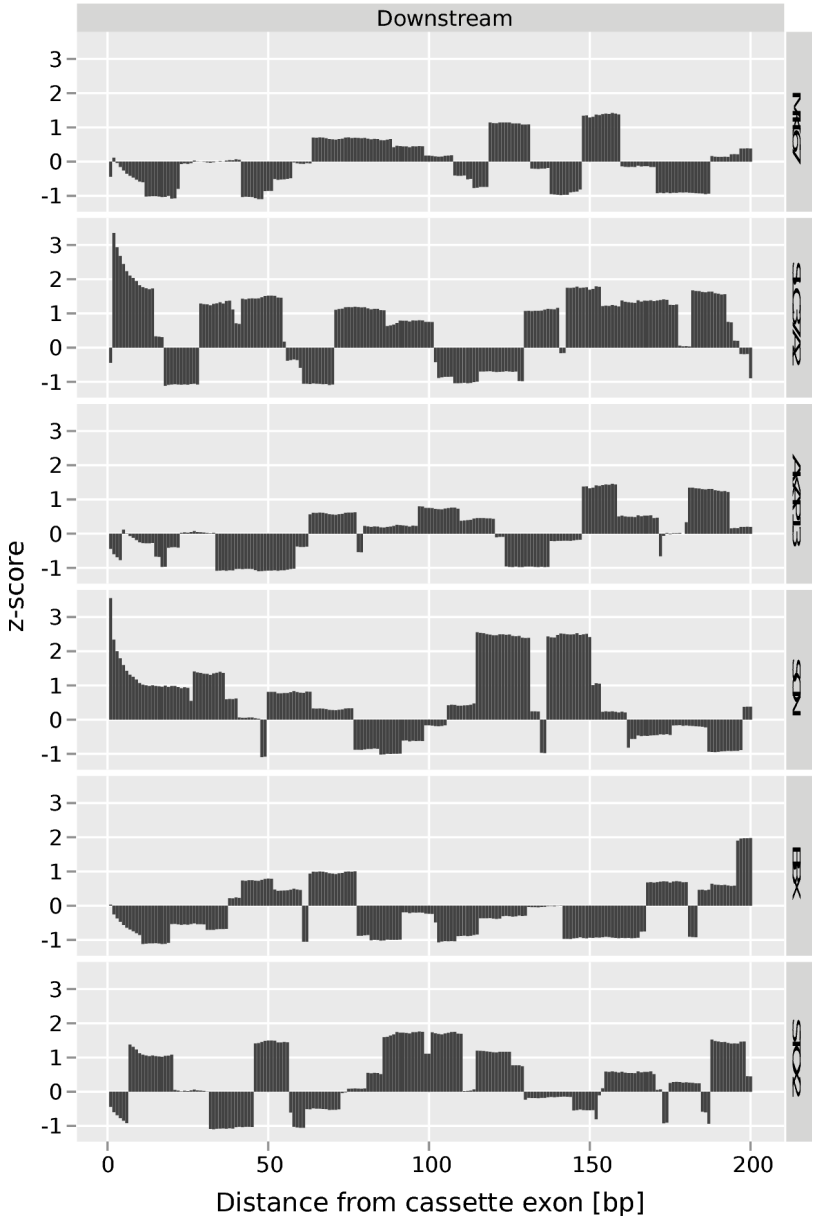
- A finite set of input symbols $\Sigma = \{“A”, “C”, “G”, “T”\}$
- A finite set of states (Q)
- A transition function ($\delta: Q \times \Sigma \rightarrow Q$)
- An initial or start state ($q_o \in Q$)

The FSM is built in a way such that the initial state q_0 represents the empty sequence, and each other state in Q represents a unique subsequence of S. For each state $q \in Q$ and each symbol $s \in \Sigma$, the transition $\delta(q, s)$ is defined as the state $p \in Q$ that represents the longest suffix of $q+s$. The operator + denotes the concatenation operator. If $p = q+s$ then the transition is called a forward transition, otherwise the transition is called a backward transition. Supplementary Figure 3 illustrates an example of FSM that recognizes all the subsequences of the sequence ATGTC.

Supplementary Figures

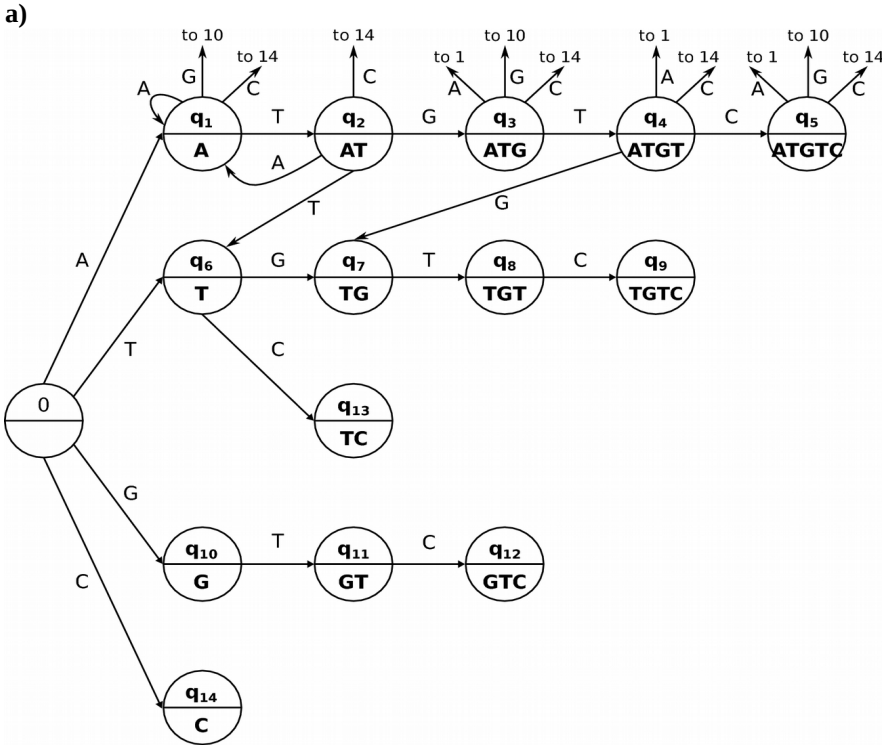


Supplementary Figure 1. Correlations between the length of the targets and the number of predicted biding sites (**a**), and the length of the queries and the number of predicted binding sites (**b**) for the bacterial dataset.



Supplementary Figure 2. Score enrichment of SNORD116 interaction sites in the flanking region downstream of the cassette exon.

a)



b)

	q_1	**q_2**	**q_3**	**q_4**	**q_5**	**q_6**	**q_7**	**q_8**	**q_9**	**q_{10}**	**q_{11}**	**q_{12}**	**q_{13}**
A	q_1	q_1	q_1	q_1	q_1	q_1	q_1	q_1	q_1	q_1	q_1	q_1	q_1
C	q_{14}	q_{14}	q_{14}	q_5	q_{14}	q_{13}	q_{14}	q_8	q_{14}	q_{14}	q_{12}	q_{14}	q_{14}
G	q_{10}	q_3	q_{10}	q_7	q_{10}	q_7	q_{10}	q_7	q_{10}	q_{10}	q_7	q_{10}	q_{10}
T	q_6	q_2	q_{11}	q_6	q_6	q_6	q_{11}	q_6	q_{11}	q_6	q_6	q_6	q_6

Supplementary Figure 3. The FSM for the sequence ATGTC, including all the states, all the forward transitions and the backward transitions for states q_1 to q_5 (**a**), and a matrix representing all the possible transitions of the transition function (**b**).

Supplementary Data

Supplementary Data 1. Experimentally validated putative interactions for the fungal and bacterial datasets recovered by STSCAN. Included in the attached CD.

DISCUSSION

8 General discussion

Each of the research articles in the Results section of the present thesis includes its own discussion of the corresponding results. Hence, this section does not attempt to go over the same points but to provide a general, end to end overview of the development of this Ph.D. thesis and to outline its contributions and a list of future lines of research that it might furnish (Figure 10).

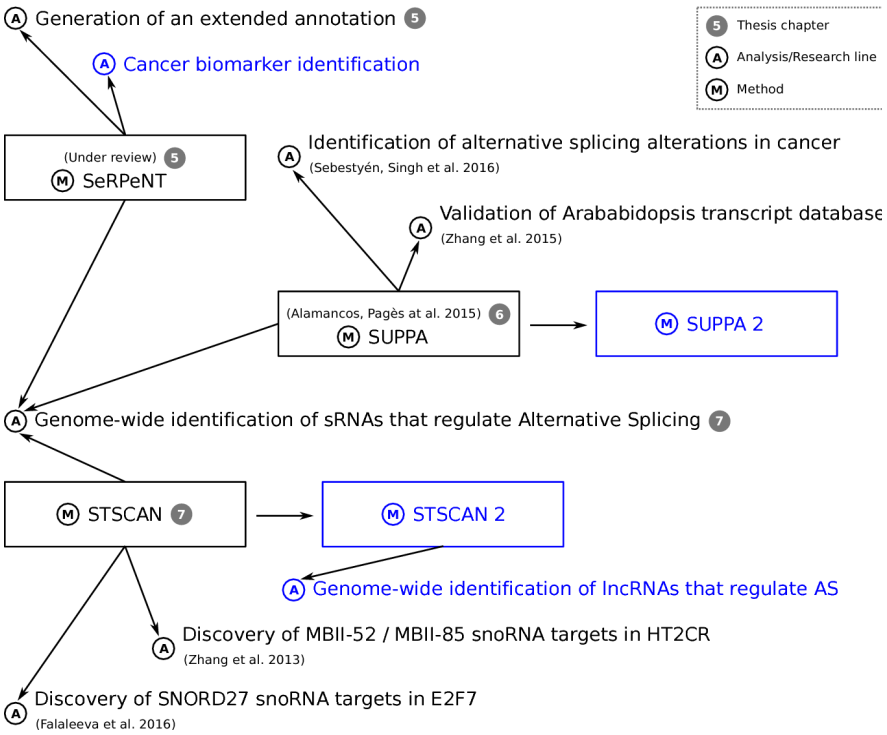


Figure 10. **Dissection of the Ph.D. project**. Schematic of the development of the Ph.D. project, including methods implemented, analysis performed and future lines of research (in blue).

In its inception, this thesis was envisioned as a four-year project focused on the development of computational tools to enhance the study of how RNA is processed in living cells, with the ultimate

goal of deciphering the mechanisms of alternative splicing regulation mediated by small non-coding RNAs. Nonetheless, the first tool developed within this framework was somehow unrelated to RNA processing: it was a pipeline for analyzing genomic data, which integrated three tools: a method to mine data from different sources, a Biomart (Smedley et al. 2015) powered database to store this data and make it publicly available, and a Weka (Hall et al. 2009) based machine learning platform that enabled the pipeline's users to manipulate this data in order to find meaningful relations and gain insight into its underlying biology. Although this pipeline was implemented to run on the IT infrastructure of the laboratory where it was developed and therefore it was never publicly released, it served as the basis for two different published studies. In the first study, histone modification data was mined to build a chromatin code of gene regulation (Althammer et al. 2012). In the second, the pipeline was used to identify changes in chromatin signals that were associated with splicing regulation (Agirre et al. 2015). This last study reported that a considerable number of alternative splicing events could have a chromatin-dependent regulation involving the binding of CTCF, AGO1 and HP1α proteins nearby regulated exons. The other tools developed within the Ph.D. project are the core of the Results section in the present thesis: SeRPeNT, SUPPA and STSCAN.

Chapter 5 describes SeRPeNT (Small Rna ProfiliNg Toolkit), a method that exploits the processing patterns of small RNAs to identify new members of known small RNAs families and to uncover and characterize potential new small RNA classes. Although the idea of using processing patterns for the classification of small non-coding RNAs is not new (Langenberger et al. 2012) (Videm et al. 2014), SeRPeNT exploits this idea at an unprecedented scale. SeRPeNT not only performs better than these preceding methods in terms of accuracy and time/memory efficiency, it also uses reproducibility across replicates and implements a tool to identify small RNAs that show differential processing between cellular conditions. In the same chapter the authors report how SeRPeNT was used to build a catalogue of more than 800 novel sRNAs, including new snoRNAs and tRNAs that resemble microRNAs, and show that a large fraction of those novel sRNAs (and other previously annotated sRNAs) undergo extensive differential processing between different cell compartments. At the

time of submission of this thesis, SeRPeNT was already submitted for peer review.

SUPPA (Alamancos, Pagès et al. 2015) was conceived as a software oriented to exploit the high speed of the state-of-the-art isoform quantification algorithms, such as Sailfish (Patro et al. 2014), to quantify alternative splicing events by estimating their inclusion values at an unforeseen speed (refer to chapter 4.2.2). In Chapter 6, we show that SUPPA is capable of analyzing both synthetic and experimental datasets with comparable or even higher accuracy than other similar methods (Shen et al. 2012) (Katz et al. 2010) but with a significant improvement in terms of speed, a fact that bears special relevance when the datasets to analyze are composed of a large number of samples and the available computational resources are limited. This feature of SUPPA enabled the study of alternative splicing alterations in several tumor types using data from the Cancer Genome Atlas (TCGA) project (Sebestyén, Singh et al. 2016). TCGA is a gigantic dataset that integrates immense amounts of sequencing data from matched normal and tumor tissues from 11.000 patients. In this study, we analyzed more than 30.000 alternative splicing events in 11 different solid tumor types with data from more 4000 samples and found that several of them showed enrichment of differentially spliced events in driver genes. Additionally, we found that several cancer hallmarks (Liberzon et al. 2015) are enriched in differentially spliced events but not in differentially expressed genes, suggesting that alternative splicing contributes to cancer development independently of expression alterations. Besides this study, the approach introduced in SUPPA for the quantification of alternative splicing events was used to validate a database of non-redundant transcripts in Arabidopsis Thaliana (Zhang et al. 2015). In this publication the authors used SUPPA on the database of transcripts to validate the quantification of splicing ratios from RNA-seq by high-resolution reverse transcription polymerase chain reaction (HR-RT-PCR).

The last tool developed during this Ph.D. project was STSCAN, which is described in Chapter 7. STSCAN is a tool developed to find potential interaction sites between a number of RNA query sequences and a number of RNA target sequences. It finds all the occurrences of a particular interaction pattern between the query and targets, allowing G:U wobbles and a limited number of

mismatches or bulges, and scores the interactions according the free energy contributed by the matching dinucleotides. Due to memory limitations of the finite state machine underlying STSCAN, the length of the query sequences is limited to 500nt. Accordingly, STSCAN is particularly useful for the discovery of interactions between small RNAs and any other type of RNA molecule (mRNA, lncRNAs…), but requires some modifications to be used for longer RNAs. Results in the same chapter demonstrate that STSCAN is capable of discriminating putative binding sites of bacterial sRNAs in their target mRNA molecules, and putative binding sites of fungal snoRNAs in their target rRNAs. Moreover, STSCAN was used to elucidate new mechanisms of regulation of the alternative splicing mediated by snoRNAs in human. In this context, STSCAN was used in two different studies in collaboration with Prof. Stefan Stamm from the University of Kentucky to find targets of human snoRNAs in genes for which the alternative splicing was regulated by these snoRNAs. The first study reported the binding sites of snoRNAs MBII-52 (SNORD115) and MBII-85 (SNORD116) in the human serotonin receptor 2c pre-mRNA (*HTR2C*) (Zhang et al. 2013). The second study used STSCAN to predict binding sites of the C/D-box snoRNA SNORD27 on the transcription factor *E2F7* pre-mRNA (Falaleeva et al. 2016). This study is particularly striking since, meanwhile SNORD115 and SNORD116 were orphan snoRNAs without known function, SNORD27 is the first snoRNA characterized to have an additional function, in this case splicing regulation, besides its canonical function of modifying its target 18S ribosomal RNA. We are currently expanding this collaboration to characterize the pre-mRNA and mRNA targets of SNORD116 and further describe its role in regulating post-transcriptional RNA processing.

Overall, the methods developed in the present thesis served as a driving force for several studies that achieved significant conclusions on important biological questions such as the role of alternative splicing in cancer or the mechanism by which sRNAs regulate alternative splicing of pre-mRNAs in human. But the work presented in this Ph.D. thesis hold other contributions in the form of the novelty of the algorithms implemented in the tools; a type of contribution, in my opinion, often overlooked in the biomedical sciences. SeRPeNT, for example, introduces a variant of the time warping dynamic algorithm (Kruskal and Lieberman 1999) to

calculate the normalized cross-correlation between two processing pattern profiles, and an enhancement of a density-based clustering algorithm proposed by Rodriguez and Laio in 2014 (Rodriguez and Laio. 2014). STSCAN also implements a novel algorithm based on a finite state machine built from the query sequence to find seeds on the target sequence in linear time, with an additional extension step that runs also in linear time.

In research, the completion of a project often leads to the inception of a new one. This way, new small pieces of knowledge are arranged above the existing ones in a subtle equilibrium to expand the human knowledge step by step. Let me then finish this discussion by outlining three research lines that the completion of this Ph.D. project could originate. The first one would be to extend SUPPA with a new module for the calculation of differential alternative splicing events between two or more conditions. The second one would be to use SeRPeNT to identify differentially processed sRNAs between tumor and normal tissues using the TCGA data, with the objective of finding potential new biomarkers for up to 33 different cancer types. Another research line would to apply STSCAN to all annotated long non-coding RNAs to find potential targets in pre-mRNAs and mRNAs, and relate the presence of targets with the expression and splicing correlations of queries and targets across multiple conditions. This would highlight potential new functions for many of the yet-to-be characterized long non-coding RNAs.

CONCLUSIONS

The main contributions of the work presented in each of the chapters of the Results section in the present thesis can be summarized as follows:

Chapter 5

- SeRPeNT is an efficient and accurate computational method for the discovery and characterization of small RNAs that outperforms similar methods in terms of accuracy, speed and memory management.

- We discovered 671 new members from the known major small RNA classes (snoRNA, snRNA, tRNA and miRNA) and 131 members from new potential small RNA classes.

- A significant proportion of small RNAs show pervasive differential processing among cellular compartments, especially tRNAs that are prominently processed in the cytosol.

- Processing patterns can be used to assign function to small RNAs irrespectively of their sequence and secondary structure.

Chapter 6

- SUPPA provides a method for leveraging fast transcript quantification for efficient and accurate alternative splicing analysis for large number of samples.

- SUPPA is comparable to other similar methods in terms of accuracy but outperforms them in terms of speed and memory management.

Chapter 7

- STSCAN is a computational method that finds RNA:RNA interaction sites in an unbiased and exhaustive manner, with high accuracy and speed.
- snoRNAs might be regulating alternative splicing of their target pre-mRNAs by direct binding on the upstream region of the alternative splicing event, occluding the branch point of the alternative exon and hindering its recognition.

REFERENCES

Abouelhoda, Mohamed Ibrahim, Stefan Kurtz, and Enno Ohlebusch. "Replacing suffix trees with enhanced suffix arrays." Journal of Discrete Algorithms 2.1 (2004): 53-86.

Agirre, Eneritz et al. "A chromatin code for alternative splicing involving a putative association between CTCF and HP1α proteins." *BMC biology* 13.1 (2015): 1.

Alamancos, Gael P et al. "Leveraging transcript quantification for fast computation of alternative splicing profiles." *RNA* 21.9 (2015): 1521-1531.

Alekseyenko, Alexander V, Namshin Kim, and Christopher J Lee. "Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes." Rna 13.5 (2007): 661-670.

Alioto, Tyler S. "U12DB: a database of orthologous U12-type spliceosomal introns." Nucleic acids research 35.suppl 1 (2007): D110-D115.

Alló, Mariano et al. "Control of alternative splicing through siRNA-mediated transcriptional gene silencing." Nature structural & molecular biology 16.7 (2009): 717-724.

Althammer, Sonja, and Eduardo Eyras. "Predictive models of gene regulation from high-throughput epigenomics data." *Comparative and functional genomics* 2012 (2012).

Altschul, Stephen F et al. "Basic local alignment search tool." Journal of molecular biology 215.3 (1990): 403-410.

Ameyar-Zazoua, Maya et al. "Argonaute proteins couple chromatin silencing to alternative splicing." Nature structural & molecular biology 19.10 (2012): 998-1004.

Bachellerie, Jean-Pierre, Jérôme Cavaillé, and Alexander Hüttenhofer. "The expanding snoRNA world." Biochimie 84.8 (2002): 775-790.

Backofen, Rolf et al. "Bioinformatics of prokaryotic RNAs." RNA biology 11.5 (2014): 470-483.

Baeza-Yates, Ricardo, and Gonzalo Navarro. "Faster approximate string matching." Algorithmica 23.2 (1999): 127-158.

Bartel, David P. "MicroRNAs: target recognition and regulatory functions." Cell 136.2 (2009): 215-233.

Berget, Susan M, Claire Moore, and Phillip A Sharp. "Spliced segments at the 5'terminus of adenovirus 2 late mRNA." Proceedings of the National Academy of Sciences 74.8 (1977): 3171-3175.

Bond, Allison M et al. "Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry." Nature neuroscience 12.8 (2009): 1020-1027.

Boutz, Paul L et al. "MicroRNAs regulate the expression of the alternative splicing factor nPTB during muscle development." Genes & development 21.1 (2007): 71-84.

Brameier, Markus et al. "Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs." Nucleic acids research 39.2 (2011): 675-686.

Brannan, Camilynn I et al. "The product of the H19 gene may function as an RNA." Molecular and cellular biology 10.1 (1990): 28-36.

Bratkovič, Tomaž, and Boris Rogelj. "Biology and applications of small nucleolar RNAs." Cellular and Molecular Life Sciences 68.23 (2011): 3843-3851.

Bray, Nicolas et al. "Near-optimal RNA-Seq quantification." arXiv preprint arXiv:1505.02710 (2015).

Brest, Patrick et al. "A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-

dependent xenophagy in Crohn's disease." Nature genetics 43.3 (2011): 242-245.

Brown, Carolyn J et al. "The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus." Cell 71.3 (1992): 527-542.

Burkhardt, Stefan et al. "q-gram based database searching using a suffix array (QUASAR)." Proceedings of the third annual international conference on Computational molecular biology 1 Apr. 1999: 77-83.

Burroughs, Alexander Maxwell et al. "Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin." RNA biology 8.1 (2011): 158-177.

Cáceres, Javier F, and Alberto R Kornblihtt. "Alternative splicing: multiple control mechanisms and involvement in human disease." TRENDS in Genetics 18.4 (2002): 186-193.

Calin, George A et al. "Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas." Cancer cell 12.3 (2007): 215-229.

Cavaillé, Jérôme et al. "Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization." Proceedings of the National Academy of Sciences 97.26 (2000): 14311-14316.

Cellini, Alessandra, Eduard Felder, and John J Rossi. "Yeast pre-messenger RNA splicing efficiency depends on critical spacing requirements between the branch point and 3'splice site." The EMBO journal 5.5 (1986): 1023.

Chan, Patricia P, and Todd M Lowe. "GtRNAdb: a database of transfer RNA genes detected in genomic sequence." Nucleic acids research 37.suppl 1 (2009): D93-D97.

Chasin, Lawrence A. "Searching for splicing motifs." Advances in experimental medicine and biology 623 (2008): 85.

Cheloufi, Sihem et al. "A dicer-independent miRNA biogenesis pathway that requires Ago catalysis." Nature 465.7298 (2010): 584-589.

Chen, Chong-Jian, and Edith Heard. "Small RNAs derived from structural non-coding RNAs." Methods 63.1 (2013): 76-84.

Chow, Louise T et al. "An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA." Cell 12.1 (1977): 1-8.

Clancy, Suzanne. "RNA splicing: introns, exons and spliceosome." Nature Education 1.1 (2008): 31.

Cole, Christian et al. "Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs." Rna 15.12 (2009): 2147-2160.

Coolidge, Candace J, Raymond J Seely, and James G Patton. "Functional analysis of the polypyrimidine tract in pre-mRNA splicing." Nucleic acids research 25.4 (1997): 888-896.

Cooper, Thomas A, Lili Wan, and Gideon Dreyfuss. "RNA and disease." Cell 136.4 (2009): 777-793.

Cordes, Kimberly R et al. "miR-145 and miR-143 regulate smooth muscle cell fate and plasticity." Nature 460.7256 (2009): 705-710.

Corvelo, André et al. "Genome-wide association between branch point properties and alternative splicing." PLoS Comput Biol 6.11 (2010): e1001016.

Crick, Francis. "Central dogma of molecular biology." Nature 227.5258 (1970): 561-563.

Crick, Francis H. "On protein synthesis." Symposia of the Society for Experimental Biology 1958: 138.

Croce, Carlo M. "Causes and consequences of microRNA dysregulation in cancer." Nature reviews genetics 10.10 (2009): 704-714.

Croft, Larry et al. "ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome." Nature genetics 24.4 (2000): 340-341.

Dapas, Matthew et al. "Comparative evaluation of isoform-level gene expression estimation algorithms for RNA-seq and exon-array platforms." Briefings in bioinformatics (2016): bbw016.

Darnell, James E. RNA: life's indispensable molecule. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2011.

Denli, Ahmet M et al. "Processing of primary microRNAs by the Microprocessor complex." Nature 432.7014 (2004): 231-235.

Diebel, Kevin W et al. "Beyond the Ribosome: Extra-translational Functions of tRNA Fragments." Biomarker insights 11.Suppl 1 (2016): 1.

Djebali, Sarah et al. "Landscape of transcription in human cells." Nature 489.7414 (2012): 101-108.

ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome." Nature 489.7414 (2012): 57-74.

Ender, Christine et al. "A human snoRNA with microRNA-like functions." Molecular cell 32.4 (2008): 519-528.

Engström, Pär G et al. "Systematic evaluation of spliced alignment programs for RNA-seq data." Nature methods 10.12 (2013): 1185-1191.

Esquela-Kerscher, Aurora, and Frank J Slack. "Oncomirs—microRNAs with a role in cancer." Nature Reviews Cancer 6.4 (2006): 259-269.

Esteller, Manel. "Non-coding RNAs in human disease." Nature Reviews Genetics 12.12 (2011): 861-874.

Faghihi, Mohammad Ali, and Claes Wahlestedt. "Regulatory roles of natural antisense transcripts." Nature reviews Molecular cell biology 10.9 (2009): 637-643.

Faghihi, Mohammad Ali et al. "Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase." Nature medicine 14.7 (2008): 723-730.

Falaleeva, Marina et al. "Dual function of C/D box small nucleolar RNAs in rRNA modification and alternative pre-mRNA splicing." *Proceedings of the National Academy of Sciences* 113.12 (2016): E1625-E1634.

Ferragina, Paolo, and Giovanni Manzini. "Opportunistic data structures with applications." Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on 2000: 390-398.

Friedman, Robin C et al. "Most mammalian mRNAs are conserved targets of microRNAs." Genome research 19.1 (2009): 92-105.

Fu, Xiaoqin et al. "Regulation of apoptosis by a prostate-specific and prostate cancer-associated noncoding gene, PCGEM1." DNA and cell biology 25.3 (2006): 135-141.

Gao, Kaiping et al. "Human branch point consensus sequence is yUnAy." Nucleic acids research 36.7 (2008): 2257-2267.

Garber, Manuel et al. "Computational methods for transcriptome annotation and quantification using RNA-seq." Nature methods 8.6 (2011): 469-477.

Gerlach, Wolfgang, and Robert Giegerich. "GUUGle: a utility for fast exact matching under RNA complementary rules including G–U base pairing." Bioinformatics 22.6 (2006): 762-764.

Gesteland, Raymond, and J Atkins. "The {RNA} World." (1993).

Giegé, Richard. "Toward a more complete view of tRNA biology." Nature structural & molecular biology 15.10 (2008): 1007-1014.

Gilbert, Walter. "Genes-in-pieces revisited." Science 228.4701 (1985): 823-824.

Gooding, Clare et al. "A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones." Genome Biol 7.1 (2006): R1.

Grant, Gregory R et al. "Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)." Bioinformatics 27.18 (2011): 2518-2528.

Graur, Dan et al. "On the immortality of television sets:“function” in the human genome according to the evolution-free gospel of ENCODE." Genome biology and evolution 5.3 (2013): 578-590.

Griffith, Malachi et al. "Alternative expression analysis by RNA sequencing." Nature methods 7.10 (2010): 843-847.

Griffiths-Jones, Sam et al. "miRBase: tools for microRNA genomics." Nucleic acids research 36.suppl 1 (2008): D154-D158.

Guerrier-Takada, Cecilia et al. "The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme." Cell 35.3 (1983): 849-857.

Guffanti, Alessandro et al. "A transcriptional sketch of a primary human breast cancer by 454 deep sequencing." BMC genomics 10.1 (2009): 1.

Guthrie, C, and B Patterson. "Spliceosomal snRNAs." Annual review of genetics 22.1 (1988): 387-419.

Gutschner, Tony et al. "The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells." Cancer research 73.3 (2013): 1180-1189.

Hall, Mark et al. "The WEKA data mining software: an update." *ACM SIGKDD explorations newsletter* 11.1 (2009): 10-18.

Hammond, Scott M. "MicroRNAs as tumor suppressors." Nature genetics 39.5 (2005): 582-583.

Haramati, Sharon et al. "miRNA malfunction causes spinal motor neuron disease." Proceedings of the National Academy of Sciences 107.29 (2010): 13111-13116.

Harrow, Jennifer et al. "GENCODE: the reference human genome annotation for The ENCODE Project." Genome research 22.9 (2012): 1760-1774.

Haussecker, Dirk et al. "Human tRNA-derived small RNAs in the global regulation of RNA silencing." Rna 16.4 (2010): 673-695.

Hertel, Jana, Ivo L Hofacker, and Peter F Stadler. "SnoReport: computational identification of snoRNAs with unknown targets." Bioinformatics 24.2 (2008): 158-164.

Hiller, Michael et al. "Pre-mRNA secondary structures influence exon recognition." PLoS Genet 3.11 (2007): e204.

Hoagland, Mahlon B et al. "A soluble ribonucleic acid intermediate in protein synthesis." J Biol Chem 231.1 (1958): 241-257.

Hoagland, Mahlon B, Paul C Zamecnik, and Mary L Stephenson. "Intermediate reactions in protein biosynthesis." Biochimica et biophysica acta 24 (1957): 215-216.

Holley, Christopher L, and Veli K Topkara. "An introduction to small non-coding RNAs: miRNA and snoRNA." Cardiovascular drugs and therapy 25.2 (2011): 151-159.

Holley, Robert W et al. "Structure of a ribonucleic acid." Science 147.3664 (1965): 1462-1465.

Huntzinger, Eric, and Elisa Izaurralde. "Gene silencing by microRNAs: contributions of translational repression and mRNA decay." Nature Reviews Genetics 12.2 (2011): 99-110.

Jänes, Jürgen et al. "A comparative study of RNA-seq analysis strategies." Briefings in bioinformatics 16.6 (2015): 932-940.

Kahl, Günter. The dictionary of genomics, transcriptomics and proteomics. John Wiley & Sons, 2015.

Kalsotra, Auinash et al. "MicroRNAs coordinate an alternative splicing network during mouse postnatal heart development." Genes & development 24.7 (2010): 653-658.

Kang, Wenjing, and Marc R Friedländer. "Computational prediction of miRNA genes from small RNA sequencing data." Frontiers in bioengineering and biotechnology 3 (2015).

Kapranov, Philipp et al. "RNA maps reveal new RNA classes and a possible function for pervasive transcription." Science 316.5830 (2007): 1484-1488.

Kato, Yuki et al. "RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming." Bioinformatics 26.18 (2010): i460-i466.

Katz, Yarden et al. "Analysis and design of RNA sequencing experiments for identifying isoform regulation." Nature methods 7.12 (2010): 1009-1015.

Keren, Hadas, Galit Lev-Maor, and Gil Ast. "Alternative splicing and evolution: diversification, exon definition and function." Nature Reviews Genetics 11.5 (2010): 345-355.

Kim, Tae-Kyung et al. "Widespread transcription at neuronal activity-regulated enhancers." Nature 465.7295 (2010): 182-187.

Kimura, Kouichi et al. "Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes." Genome research 16.1 (2006): 55-65.

Kim, V Narry, Jinju Han, and Mikiko C Siomi. "Biogenesis of small RNAs in animals." Nature reviews Molecular cell biology 10.2 (2009): 126-139.

Kim, V Narry. "MicroRNA biogenesis: coordinated cropping and dicing." Nature reviews Molecular cell biology 6.5 (2005): 376-385.

Kim, Young-Kook, Boseon Kim, and V Narry Kim. "Re-evaluation of the roles of DROSHA, Exportin 5, and DICER in microRNA biogenesis." Proceedings of the National Academy of Sciences 113.13 (2016): E1881-E1889.

Kirchner, Sebastian, and Zoya Ignatova. "Emerging roles of tRNA in adaptive translation, signalling dynamics and disease." Nature Reviews Genetics 16.2 (2015): 98-112.

Kishore, S, and S Stamm. "Regulation of alternative splicing by snoRNAs." Cold Spring Harbor symposia on quantitative biology 1 Jan. 2006: 329-334.

Kishore, Shivendra, and Stefan Stamm. "The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C." Science 311.5758 (2006): 230-232.

Kishore, Shivendra et al. "The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing." Human molecular genetics 19.7 (2010): 1153-1164.

Kiss, Tamás. "Biogenesis of small nuclear RNPs." Journal of cell science 117.25 (2004): 5949-5951.

Kloosterman, Wigard P, and Ronald HA Plasterk. "The diverse functions of microRNAs in animal development and disease." Developmental cell 11.4 (2006): 441-450.

Kolasinska-Zwierz, Paulina et al. "Differential chromatin marking of introns and expressed exons by H3K36me3." Nature genetics 41.3 (2009): 376-381.

Kornberg, Roger D. "The molecular basis of eukaryotic transcription." Proceedings of the National Academy of Sciences 104.32 (2007): 12955-12961.

Kornblihtt, Alberto R et al. "Multiple links between transcription and splicing." Rna 10.10 (2004): 1489-1498.

Kruger, Kelly et al. "Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena." cell 31.1 (1982): 147-157.

Kruskal, J. and Liberman, M. The symmetric time-warping problem: from continuous to discrete. In Sankoff, D. (eds.), *Time Warps, String Edits, and Macromolecules: The theory and Practice of Sequence Comparison*. CSLI Publications, Stanford, pp. 125-161.

Kung, Johnny TY, David Colognori, and Jeannie T Lee. "Long noncoding RNAs: past, present, and future." Genetics 193.3 (2013): 651-669.

Labrador, Mariano, and Victor G Corces. "Extensive exon reshuffling over evolutionary time coupled to trans-splicing in Drosophila." Genome research 13.10 (2003): 2220-2228.

Lagesen, Karin et al. "RNAmmer: consistent and rapid annotation of ribosomal RNA genes." Nucleic acids research 35.9 (2007): 3100-3108.

Lai, Daniel, and Irmtraud M Meyer. "A comprehensive comparison of general RNA–RNA interaction prediction methods." Nucleic acids research (2015): gkv1477.

Langenberger, David et al. "deepBlockAlign: a tool for aligning RNA-seq profiles of read block patterns." Bioinformatics 28.1 (2012): 17-24.

Laurent, Georges St, Claes Wahlestedt, and Philipp Kapranov. "The Landscape of long noncoding RNA classification." Trends in Genetics 31.5 (2015): 239-251.

Lee, Rosalind C, Rhonda L Feinbaum, and Victor Ambros. "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14." Cell 75.5 (1993): 843-854.

Lee, Tong Ihn, and Richard A Young. "Transcription of eukaryotic protein-coding genes." Annual review of genetics 34.1 (2000): 77-137.

Lee, Yong Sun et al. "A novel class of small RNAs: tRNA-derived RNA fragments (tRFs)." Genes & development 23.22 (2009): 2639-2649.

Leff, Stuart E, and Michael G Rosenfeld. "Complex transcriptional units: diversity in gene expression by alternative RNA processing." Annual review of biochemistry 55.1 (1986): 1091-1117.

Lerner, Michael R et al. "Two novel classes of small ribonucleoproteins detected by antibodies associated with lupus erythematosus." Science 211.4480 (1981): 400-402.

Leung, Karen N, and Barbara Panning. "X-inactivation: Xist RNA uses chromosome contacts to coat the X." Current Biology 24.2 (2014): R80-R82.

Lewis, Morag A et al. "An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice." Nature genetics 41.5 (2009): 614-618.

Liao, Jian-You et al. "Deep sequencing of human nuclear and cytoplasmic small RNAs reveals an unexpectedly complex subcellular distribution of miRNAs and tRNA 3′ trailers." PLos one 5.5 (2010): e10563.

Li, Baiyong et al. "Analyses of promoter-proximal pausing by RNA polymerase II on the hsp70 heat shock gene promoter in a Drosophila nuclear extract." Molecular and cellular biology 16.10 (1996): 5433-5443.

Li, Bo et al. "RNA-Seq gene expression estimation with read mapping uncertainty." Bioinformatics 26.4 (2010): 493-500.

Li, Jianwei et al. "LncTar: a tool for predicting the RNA targets of long noncoding RNAs." Briefings in bioinformatics 16.5 (2015): 806-812.

Li, Wen, Wei Yang, and Xiu-Jie Wang. "Pseudogenes: pseudo or real functional elements?." Journal of Genetics and Genomics 40.4 (2013): 171-177.

Li, Zhihua et al. "Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs." Nucleic acids research (2012): gks307.

Lowe, Todd M, and Sean R Eddy. "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." Nucleic acids research 25.5 (1997): 955-964.

Lund, Elsebet et al. "Nuclear export of microRNA precursors." Science 303.5654 (2004): 95-98.

Makeyev, Eugene V et al. "The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing." Molecular cell 27.3 (2007): 435-448.

Ma, Lina, Vladimir B Bajic, and Zhang Zhang. "On the classification of long non-coding RNAs." RNA biology 10.6 (2013): 924-933.

Manber, Udi, and Gene Myers. "Suffix arrays: a new method for on-line string searches." siam Journal on Computing 22.5 (1993): 935-948.

Marioni, John C et al. "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." Genome research 18.9 (2008): 1509-1517.

Martens-Uzunova, Elena S, Michael Olvedy, and Guido Jenster. "Beyond microRNA–novel RNAs derived from small non-coding RNA and their implication in cancer." Cancer letters 340.2 (2013): 201-211.

Matera, A Gregory, and Zefeng Wang. "A day in the life of the spliceosome." Nature reviews Molecular cell biology 15.2 (2014): 108-121.

Matera, A Gregory, Rebecca M Terns, and Michael P Terns. "Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs." Nature reviews Molecular cell biology 8.3 (2007): 209-220.

Mattick, John S. "Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms." Bioessays 25.10 (2003): 930-939.

Mattick, John S. "Non-coding RNAs: the architects of eukaryotic complexity." EMBO reports 2.11 (2001): 986-991.

Maxwell, Earl S, and MJ Fournier. "The small nucleolar RNAs." Annual review of biochemistry 64.1 (1995): 897-934.

McQuillen, Kenneth, Richard B Roberts, and Roy J Britten. "Synthesis of nascent protein by ribosomes in Escherichia coli." Proceedings of the National Academy of Sciences 45.9 (1959): 1437-1447.

Melamed, Ze'ev et al. "Alternative splicing regulates biogenesis of miRNAs located across exon-intron junctions." Molecular cell 50.6 (2013): 869-881.

Mercer, Tim R et al. "Genome-wide discovery of human splicing branchpoints." Genome research 25.2 (2015): 290-303.

Morin, Ryan D et al. "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells." Genome research 18.4 (2008): 610-621.

Morris, Kevin V, and John S Mattick. "The rise of regulatory RNA." Nature reviews. Genetics 15.6 (2014): 423.

Mortazavi, Ali et al. "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nature methods 5.7 (2008): 621-628.

Mückstein, Ulrike et al. "Thermodynamics of RNA–RNA binding." Bioinformatics 22.10 (2006): 1177-1182.

Nahkuri, Satu, Ryan J Taft, and John S Mattick. "Nucleosomes are preferentially positioned at exons in somatic and sperm cells." Cell cycle 8.20 (2009): 3420-3424.

Natoli, Gioacchino, and Jean-Christophe Andrau. "Noncoding transcription at enhancers: general principles and functional models." Annual review of genetics 46 (2012): 1-19.

Nawrocki, Eric P, and Sean R Eddy. "Infernal 1.1: 100-fold faster RNA homology searches." Bioinformatics 29.22 (2013): 2933-2935.

Nicholls, Robert D, and Jessica L Knepper. "Genome organization, function, and imprinting in Prader-Willi and Angelman syndromes." Annual review of genomics and human genetics 2.1 (2001): 153-175.

Nicolas, Francisco Esteban et al. "Biogenesis of Y RNA-derived small RNAs is independent of the microRNA pathway." FEBS letters 586.8 (2012): 1226-1230.

Ono, Motoharu et al. "Identification of human miRNA precursors that resemble box C/D snoRNAs." Nucleic acids research 39.9 (2011): 3879-3891.

Orphanides, George, and Danny Reinberg. "A unified theory of gene expression." Cell 108.4 (2002): 439-451.

Ozsolak, Fatih, and Patrice M Milos. "RNA sequencing: advances, challenges and opportunities." Nature reviews genetics 12.2 (2011): 87-98.

Padgett, Richard A. "New connections between splicing and human disease." Trends in Genetics 28.4 (2012): 147-154.

Patro, Rob, Geet Duggal, and Carl Kingsford. "Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment." bioRxiv (2015): 021592.

Patro, Rob, Stephen M Mount, and Carl Kingsford. "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms." Nature biotechnology 32.5 (2014): 462-464.

Pederson, Thoru. "Regulatory RNAs derived from transfer RNA?." Rna 16.10 (2010): 1865-1869.

Penny, Graeme D et al. "Requirement for Xist in X chromosome inactivation." Nature 379.6561 (1996): 131-137.

Pibouin, Laurence et al. "Cloning of the mRNA of overexpression in colon carcinoma-1: a sequence overexpressed in a subset of colon carcinomas." Cancer genetics and cytogenetics 133.1 (2002): 55-60.

Pink, Ryan Charles et al. "Pseudogenes: pseudo-functional or key regulators in health and disease?." Rna 17.5 (2011): 792-798.

Pruitt, Kim D et al. "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy." Nucleic acids research 40.D1 (2012): D130-D135.

Pundhir, Sachin, Panayiota Poirazi, and Jan Gorodkin. "Emerging applications of read profiles towards the functional annotation of the genome." Frontiers in genetics 6 (2015).

Quek, Xiu Cheng et al. "lncRNAdb v2. 0: expanding the reference database for functional long noncoding RNAs." Nucleic acids research (2014): gku988.

Quinn, Jeffrey J, and Howard Y Chang. "Unique features of long non-coding RNA biogenesis and function." Nature Reviews Genetics 17.1 (2016): 47-62.

Rearick, David et al. "Critical association of ncRNA with introns." Nucleic acids research 39.6 (2011): 2357-2366.

Reinert K. et al. “Alignment of Next-Generation Sequencing Reads”. Annu Rev Genomics Hum Genet. 16 (2015):133-51.

Reinhart, Brenda J et al. "The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans." nature 403.6772 (2000): 901-906.

Reyes, Paula H, and Elisa Ficarra. "One decade of development and evolution of microRNA target prediction algorithms." Genomics, proteomics & bioinformatics 10.5 (2012): 254-263.

Richard, Hugues et al. "Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments." Nucleic acids research 38.10 (2010): e112-e112.

Rinn, John L et al. "Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs." Cell 129.7 (2007): 1311-1323.

Ruby, J Graham, Calvin H Jan, and David P Bartel. "Intronic microRNA precursors that bypass Drosha processing." Nature 448.7149 (2007): 83-86.

Ryan, Michael C et al. "SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts." Bioinformatics 28.18 (2012): 2385-2387.

Sanger, Frederick, and EOP Thompson. "The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates." Biochemical Journal 53.3 (1953): 353.

Saraiya, Ashesh A, and Ching C Wang. "snoRNA, a novel precursor of microRNA in Giardia lamblia." PLoS Pathog 4.11 (2008): e1000224.

Schaefer, Anne et al. "Cerebellar neurodegeneration in the absence of microRNAs." The Journal of experimental medicine 204.7 (2007): 1553-1558.

Scheper, Gert C, Marjo S van der Knaap, and Christopher G Proud. "Translation matters: protein synthesis defects in inherited disease." Nature Reviews Genetics 8.9 (2007): 711-723.

Scherrer, Klaus, Harriet Latham, and James E Darnell. "Demonstration of an unstable RNA and of a precursor to ribosomal RNA in HeLa cells." Proceedings of the National Academy of Sciences 49.2 (1963): 240-248.

Schubert, Thomas et al. "Df31 protein and snoRNAs maintain accessible higher-order structures of chromatin." Molecular cell 48.3 (2012): 434-444.

Schwartz, Schraga, Eran Meshorer, and Gil Ast. "Chromatin organization marks exon-intron structure." Nature structural & molecular biology 16.9 (2009): 990-995.

Schwartz, Schraga et al. "Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes." Genome research 18.1 (2008): 88-103.

Scott, Michelle S et al. "Human box C/D snoRNA processing conservation across multiple cell types." Nucleic acids research 40.8 (2012): 3676-3688.

Scott, Michelle S et al. "Human miRNA precursors with box H/ACA snoRNA features." PLoS Comput Biol 5.9 (2009): e1000507.

Sebestyén, E., Singh, B. et al. “Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks.” bioRxiv (2015): doi: http://dx.doi.org/10.1101/023010.

Seila, Amy C et al. "Divergent transcription from active promoters." Science 322.5909 (2008): 1849-1851.

Shapiro, James A. "Revisiting the central dogma in the 21st century." Annals of the New York Academy of Sciences 1178.1 (2009): 6-28.

Sharp, Phillip A, and Christopher B Burge. "Classification of introns: U2-type or U12-type." Cell 91.7 (1997): 875-879.

Shepard, Peter J, and Klemens J Hertel. "Conserved RNA secondary structures promote alternative splicing." Rna 14.8 (2008): 1463-1469.

Shen, Shihao et al. "MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data." *Nucleic acids research* (2012): gkr1291.

Sims, Robert J, Rimma Belotserkovskaya, and Danny Reinberg. "Elongation by RNA polymerase II: the short and long of it." Genes & development 18.20 (2004): 2437-2468.

Sleutels, Frank, Ronald Zwart, and Denise P Barlow. "The non-coding Air RNA is required for silencing autosomal imprinted genes." Nature 415.6873 (2002): 810-813.

Smedley D et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 2015 Jul 1;43(W1):W589-98.

Smith, Temple F, and Michael S Waterman. "Identification of common molecular subsequences." Journal of molecular biology 147.1 (1981): 195-197.

Sobala, Andrew, and Gyorgy Hutvagner. "Transfer RNA-derived fragments: origins, processing, and functions." Wiley Interdisciplinary Reviews: RNA 2.6 (2011): 853-862.

Sonkoly, Eniko et al. "Identification and characterization of a novel, psoriasis susceptibility-related noncoding RNA gene, PRINS." Journal of Biological Chemistry 280.25 (2005): 24159-24167.

Spies, Noah et al. "Biased chromatin signatures around polyadenylation sites and exons." Molecular cell 36.2 (2009): 245-254.

Stamm, S. et al. “Function of alternative Splicing”. Gene 344 (2005): 1-20.

Taft, Ryan J et al. "Non-coding RNAs: regulators of disease." The Journal of pathology 220.2 (2010): 126-139.

Taft, Ryan J et al. "Small RNAs derived from snoRNAs." Rna 15.7 (2009): 1233-1240.

Thompson, Martin et al. "Nucleolar clustering of dispersed tRNA genes." Science 302.5649 (2003): 1399-1401.

Tian, Bin et al. "A large-scale analysis of mRNA polyadenylation of human and mouse genes." Nucleic acids research 33.1 (2005): 201-212.

Tilgner, Hagen et al. "Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs." Genome research 22.9 (2012): 1616-1625.

Torres, Adrian Gabriel, Eduard Batlle, and Lluis Ribas de Pouplana. "Role of tRNA modifications in human diseases." Trends in molecular medicine 20.6 (2014): 306-314.

Trapnell, Cole et al. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nature biotechnology 28.5 (2010): 511-515.

Tripathi, Vidisha et al. "The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation." Molecular cell 39.6 (2010): 925-938.

Venables, Julian P et al. "Identification of alternative splicing markers for breast cancer." Cancer Research 68.22 (2008): 9525-9531.

Videm, Pavankumar et al. "BlockClust: efficient clustering and classification of non-coding RNAs from short read RNA-seq profiles." Bioinformatics 30.12 (2014): i274-i282.

Walker, RT. "Mycoplasma evolution: a review of the use of ribosomal and transfer RNA nucleotide sequences in the determination of phylogenetic relationships." The Yale journal of biology and medicine 56.5-6 (1983): 367.

Walter, P, and G Blobel. "7SL small cytoplasmic RNA is an integral component of the signal recognition particle." Nature 299 (1982): 691-698.

Wang, Eric T et al. "Alternative isoform regulation in human tissue transcriptomes." Nature 456.7221 (2008): 470-476.

Wang, Kevin C, and Howard Y Chang. "Molecular mechanisms of long noncoding RNAs." Molecular cell 43.6 (2011): 904-914.

Wang, Kevin C et al. "A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression." Nature 472.7341 (2011): 120-124.

Wang, Liguo et al. "A statistical method for the detection of alternative splicing using RNA-seq." PloS one 5.1 (2010): e8529.

Wang, Zhong, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics." Nature Reviews Genetics 10.1 (2009): 57-63.

Warf, M Bryan, and J Andrew Berglund. "Role of RNA structure in regulating pre-mRNA splicing." Trends in biochemical sciences 35.3 (2010): 169-178.

Washietl, Stefan et al. "Computational analysis of noncoding RNAs." Wiley Interdisciplinary Reviews: RNA 3.6 (2012): 759-778.

Watanabe, Yuka, Masaru Tomita, and Akio Kanai. "Computational methods for microRNA target prediction." Methods in enzymology 427 (2007): 65-86.

Watson, James D, and Francis HC Crick. "Molecular structure of nucleic acids." Nature 171.4356 (1953): 737-738.

Weick, Eva-Maria, and Eric A Miska. "piRNAs: from biogenesis to function." Development 141.18 (2014): 3458-3471.

Will, Cindy L, and Reinhard Lührmann. "Spliceosome structure and function." Cold Spring Harbor perspectives in biology 3.7 (2011): a003707.

Williams, Andrew H et al. "MicroRNA-206 delays ALS progression and promotes regeneration of neuromuscular synapses in mice." Science 326.5959 (2009): 1549-1554.

Wilusz, Jeremy E, Susan M Freier, and David L Spector. "3′ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA." Cell 135.5 (2008): 919-932.

Wright, Patrick R et al. "CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains." Nucleic acids research 42.W1 (2014): W119-W123.

Wu, Jie et al. "SpliceTrap: a method to quantify alternative splicing under single cellular conditions." Bioinformatics 27.21 (2011): 3010-3016.

Wu, Shaoping et al. "Functional recognition of the 3′ splice site AG by the splicing factor U2AF35." Nature 402.6763 (1999): 832-835.

Yin, Qing-Fei et al. "Long noncoding RNAs with snoRNA ends." Molecular cell 48.2 (2012): 219-230.

Zhang, Zhaiyi et al. "The 5′ untranslated region of the serotonin receptor 2C pre-mRNA generates miRNAs and is expressed in non-neuronal cells." *Experimental brain research* 230.4 (2013): 387-394.

Zhang, Zhaojun, and Wei Wang. "RNA-Skim: a rapid method for RNA-Seq quantification at transcript level." Bioinformatics 30.12 (2014): i283-i292.

Zhang, Runxuan et al. "AtRTD–a comprehensive reference transcript dataset resource for accurate quantification of transcript-specific expression in Arabidopsis thaliana." *New Phytologist* 208.1 (2015): 96-101.

Zhao, Yong et al. "Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2." Cell 129.2 (2007): 303-317.

Zieve, Gary W. "Two groups of small stable RNAs." Cell 25.2 (1981): 296-297.

Zywicki, Marek, Kamilla Bakowska-Zywicka, and Norbert Polacek. "Revealing stable processing products from ribosome-associated small RNAs by deep-sequencing data analysis." Nucleic acids research 40.9 (2012): 4013-4024.

ANNEXES

Annex A. List of publications

Althammer, S., Pagès, A. and Eyras, E. "Predictive models of gene regulation from high-throughput epigenomics data." *Comparative and functional genomics* doi: 10.1155/2012/284786 (2012).

Zhang, Z., Falaleeva, M., Agranat-Tamir, L., Pagès, A., Eyras, E., Sperling, R. and Stamm, S. "The 5′ untranslated region of the serotonin receptor 2C pre-mRNA generates miRNAs and is expressed in non-neuronal cells." *Experimental brain research* 230.4 (2013): 387-394.

Agirre, E., Bellora, N., Alló, M., Pagès, A., Bertucci, P., Kornblihtt, A.R. and Eyras, E. "A chromatin code for alternative splicing involving a putative association between CTCF and HP1α proteins." *BMC biology* 13:31 (2015).

González-Vallinas, J., Pagès, A., Singh, B. and Eyras, E. "A semi-supervised approach uncovers thousands of intragenic enhancers differentially activated in human cells." *BMC genomics* 16:523 (2015).

Alamancos, G. P., Pagès, A., Trincado, J.L., Bellora, N. and Eyras, E. "Leveraging transcript quantification for fast computation of alternative splicing profiles." *RNA* 21.9 (2015): 1521-1531.

Falaleeva, M., Pagès, A., Matuszek, Z., Hidmi, S., Agranat-Tamir, L., Korotkov, K., Nevo, Y., Eyras, E., Sperling, R. and Stamm, S. "Dual function of C/D box small nucleolar RNAs in rRNA modification and alternative pre-mRNA splicing." *Proceedings of the National Academy of Sciences* 113.12 (2016): E1625-E1634.

Sebestyén, E., Singh, B., Miñana, B., Pagès, A., Mateo, F., Pujana, M.A., Valcárcel, J and Eyras, E. "Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks." *Genome Res.* (2016): advance access.