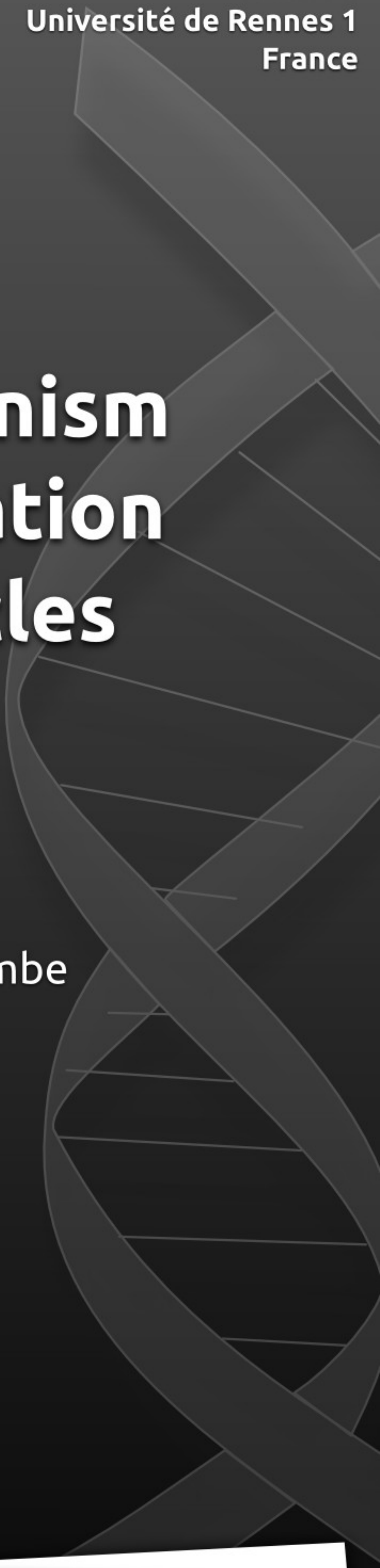
2010-2011

Master 2 Modélisation des systèmes biologiques
UFR SVE
Université de Rennes 1
France



Molecular Mechanism of Splicing regulation by hnRNP Particles

Isabelle Stévant

Supervised by
Kathi Zarnack and Nick Luscombe

Luscombe Group
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge
United Kingdom

UMR 6026
INTERACTIONS CELLULAIRES
ET MOLECULAIRES

Christian DELAMARCHE
Professeur à l'Université de Rennes 1, FRANCE
Equipe Structure et Dynamique des Macromolécules (SDM)

UNIVERSITÉ DE RENNES 1

ENGAGEMENT DE NON PLAGIAT

Je, soussigné(e) Isabelle Stévant
étudiant(e) en Master 2 Modélisation des systèmes biologiques
déclare être pleinement informé que le plagiat de documents ou d'une partie de document publiés sur toute forme de support, y comprit l'internet, constitue une violation des droits d'auteurs ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Date : 06/06/11

Signature :

Cet engagement devra être inséré en première page du document concerné.

Acknowlegements

I would like to thank Nick Luscombe and Kathi Zarnack to have accepted me for this internship after only two phone calls. It is really a pleasure to work with you! Thanks to Jernej Ule, Julian König and the Ule Group for all the experimental data you provided and for the interesting meetings where I felt sometimes so trivial face to your knowledge.

Thanks a lot to the Luscombe Group for the welcome and to have made this five months a pleasure: the coffee break with you every morning, your smile, your good mood and your help.

Thanks to Christophe Hitte and Catherine Belleannée for the references that surely helped a lot to make me be accepted at EBI. Thanks too to the Region Bretagne for the grant necessary to survive in a country where the rent is so expensive! *Breizh Atao!*

I thank Florence Cavalli (my French support in the group), Filipe Tavares-Cadete, Iñigo Martincorena and Sylvain Prigent for the review and the relevant comments on this thesis.

Thanks to Vincent Xue, my house mate, American student doing an intership at EBI too. You helped me a lot to put up this strange house and with you I have learnt to speak English fluently (even if I still have to improve). Thank you too to have spent two entire afternoons with me to correct my awful grammar.

Thanks to all the students and former students of the MSB Master for their interesting and entertaining discussions on IRC that helped me a lot to endure the distance between us.

Finally, thanks to you, reader. I hope that you are not allergic to molecular biology and you will read more that the half of the introduction!

Abbreviations

3’-OH : 3’-hydroxy

3’SS : 3’ Splice Site

5’SS : 5’ Splice Site

A3SS: Alternative 3’ Splice Site

A5SS: Alternative 5’ Splice Site

A : Adenine

AS : Alternative Splicing

ASPIRE : Analysis of SPlicing Isoform REciprocity

C : Cytosine

CE : Cassette Exon

DNA : DesoxyriboNucleic Acid

dT : desoxy-Thymine

EMBL- EBI : European Molecular Biology Laboratory - European Bioinformatics Institute

G : Guanine

hnRNP : heterogeneous nuclear RiboNucleoProtein

iCLIP : individual-nucleotide UV Cross-Linking and ImmunoPrecipitation

IGB : Integrated Genome Browser

IR: Intron Retention

MRC-LMB : Medical Research Council - Laboratory of Molecular Biology

mRNA : messenger RiboNucleic Acid

PTB : Polypyrimidine Tract-Binding protein

RBP : RNA-Binding Protein

RNA pol II : RNA polymerase II

RNA : RiboNucleic Acid

RNA-seq : RNA sequencing

SF : Splicing Facilitator

siRNA : small interfering RNA

snoRNA : small nucleolar RNA

snRNA : small nuclear RNA

snRNP : small nuclear ribonucleoprotein

T : Thymine

U2AF : U2 Auxiliary Factor

UK : United Kingdom

U : Uracil

UV : Ultra-Violet

Y : pYrimidine nucleotide (cytosine, uracil or thymine)

Contents

1 Introduction

In this study, we will highlight the role of the RNA-binding protein hnRNP C in the process of RNA splicing by combining several data types such as splice-junction microarrays, RNA-seq and iCLIP using computational methods. Before we proceed, we will explain the biological background and the different techniques used to generate the data.

1.1 Ribonucleic Acid

Ribonucleic Acid (RNA) is a macromolecule found in all known forms of life. It is one of the three major macromolecules, along side DNA and proteins. Like DNA, RNA is composed of an assembly of nucleotides. Each of these nucleotides consists of a ribose sugar, a phosphate group and a nucleobase. The four bases composing RNA are the purines, adenine (A) and guanine (G), and the pyrimidines, uracil (U) and cytosine (C) (Figure 1). RNA is synthesised by the RNA polymerase enzyme during a process called transcription. RNA polymerase (RNA pol II) uses DNA sequences as templates to assemble messenger RNA (mRNA) molecules by adding complementary nucleotides to the newly synthesised RNA chain.

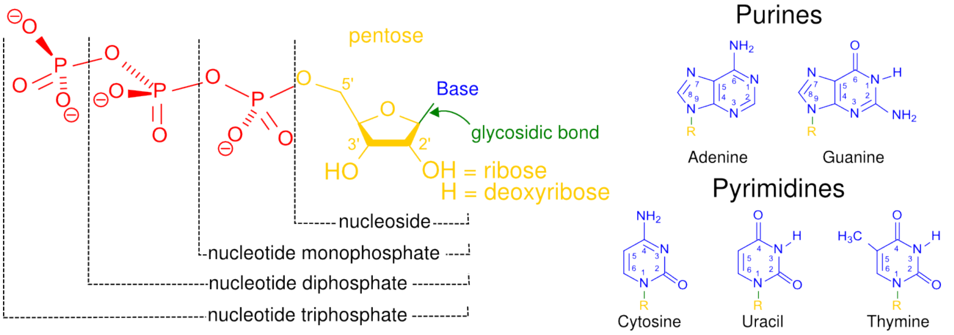


Figure 1 | Structural elements of the nucleotides.
Each RNA nucleotide is composed of a pentose (in yellow) in which carbon atoms are numbered 1' through 5'. A phosphate group is attached to the 3' position of one ribose and the 5' position of the next. This gives the RNA strand, a 5' to 3' orientation. The bases are defined in two groups, the purines and the pyrimidines. Uracil replace the thymine in the RNA. *Image taken from http://en.wikipedia.org/wiki/File:Nucleotides_1.svg*

The mRNA carries the necessary biological information to produce proteins. In eukaryotes, the precursor mRNA (pre-mRNA) is processed to form mature mRNA. Maturation of pre-mRNA consists of a series of important steps that take place at the same time as transcription. First, the nascent mRNA is capped by a modified nucleotide (7-methylguanosine) on the 5' side (5' capping). This cap has several roles: it protects the mRNA from degradation by ribonucleases (enzymes that catalyse the degradation of RNA), and later, it recruits the ribosomes that will translate the mRNA. After capping, the pre-mRNA is spliced to remove the non-coding regions called intron, that are specific to eukaryotes. The coding regions are named exons and often form only a small fraction of the genes. Splicing will be discussed in detail in the next section. When the full DNA sequence is transcribed, the mRNA is cleaved from DNA through the action of an endonuclease complex associated with RNA pol II. Then, a poly(A) tail is added to the 3' end of the mRNA by a polyadenylate polymerase. This poly(A) tail helps during the export of mRNA from the nucleus, protects the mRNA from enzymatic degradation in the cytoplasm and aids translation. After processing completes, the mature mRNA is translocated to the cytoplasm where it will be translated by ribosomes (Figure 2).

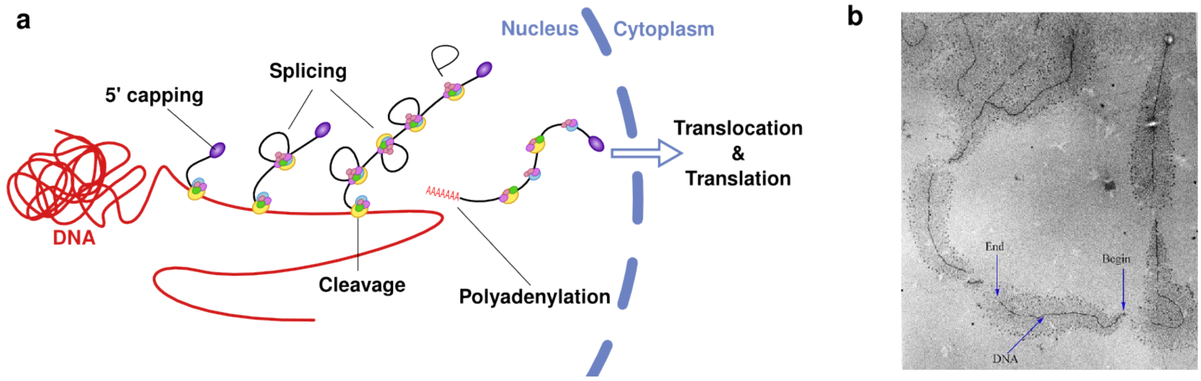


Figure 2 | RNA processing.
(**a**) RNA is synthesised by RNA pol II using DNA as a template. RNA is capped in its 5' side, spliced, cleaved from DNA and exported to the cytoplasm to be translated into proteins by the ribosomes. (**b**) Micrograph of gene transcription of ribosomal RNA illustrating the growing primary transcripts. "Begin" indicates the 5' end of the coding strand of DNA, where new RNA synthesis begins. "End" indicates the 3' end, where the primary transcripts are almost complete. *Image (**b**) taken from http://fr.wikipedia.org/wiki/Fichier:Transcription_label_en.jpg*

1.2 The RNA Splicing Process

Eukaryotic genes contain both intronic and exonic sequences. This was discovered in 1977 and came as a surprise because scientists were familiar only with bacterial genes, which consist of a continuous stretch of coding DNA[1, 2].

The basic molecular mechanism of splicing is shown in Figure 3. In the first step, a specific adenine nucleotide in a sequence called the branch point performs a nucleophilic attack on the 5' splice site which cuts the RNA. Thus, the 5' end of the intron becomes linked to the adenine nucleotide and creates a loop in the RNA molecule. The 3'-OH end of the preceding exon reacts with the start of the next exon, and the two exons join, releasing the intron in the shape of a *lariat* which is degraded. The joined exons form continuous coding sequence[3].

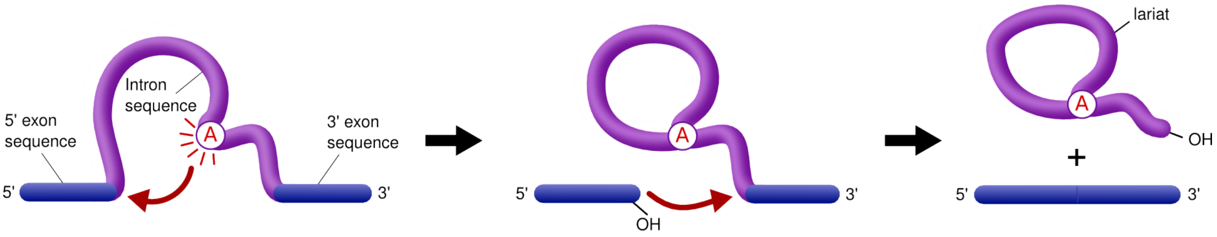


Figure 3 | The pre-mRNA splicing reaction.
A specific adenine branch point nucleotide performs a nucleophilic attack on the first nucleotide of the intron at the 5' splice site, forming a lariat. Then the 3'-OH of the intron is released and the two exons join (Alberts 2007).

This series of reactions is catalysed by a large complex of RNA-binding proteins and small nuclear

RNAs (snRNAs), called the spliceosome. Five snRNAs molecules are contained in the spliceosome (U1, U2, U4, U5 and U6), and each is complexed with at least seven protein subunits to form a small nuclear ribonucleoprotein (snRNP). Recent methods to purify spliceosomes coupled with mass spectrometry have revealed that the spliceosome may be composed of 300 distinct proteins, making at the most complex macromolecular machine in the eukaryotic cell[4].

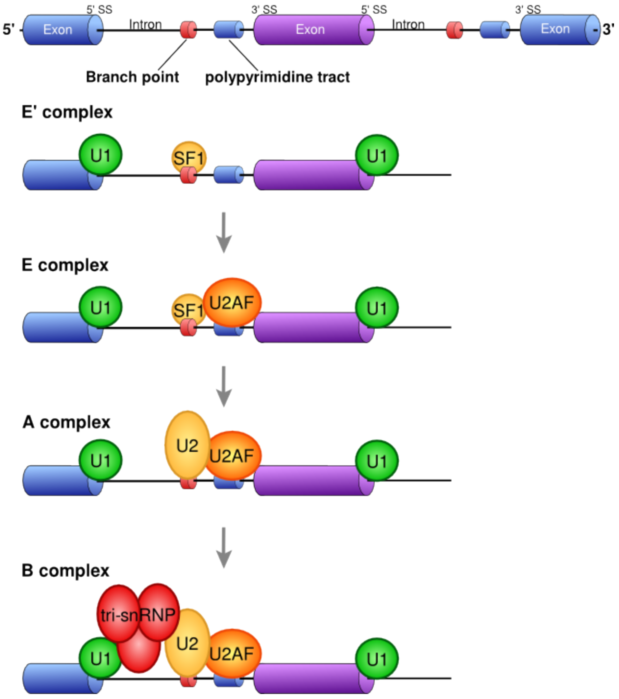


Figure 4 | Splicing and spliceosome assembly. The spliceosome assembly proceeds in several intermediates complexes that define the splice sites and catalyse the splicing reaction (Chen 2009).

Spliceosome assembly is described in Figure 4. It begins with the base-pairing of U1 snRNA to the 5' splice site (5'SS) and the binding of splicing factor 1 (SF1) to the branch point. This forms the E' complex. The E' complex can be converted into the E complex by the recruitment of the U2 auxiliary factor (U2AF) heterodimer (comprising U2AF^{65} and U2AF^{35}) to the polypyrimidine tract. The E complex is converted then into the pre-spliceosome A complex by the replacement of SF1 by U2 snRNP at the branch point. Further recruitment of the U4/U6-U5 tri-snRNP leads to the formation of the B complex, which contains all spliceosomal subunits that carry out pre-mRNA splicing. This is followed by extensive conformational changes and remodelling, including the loss of U1 and U4 snRNPs, ultimately resulting in the formation of the C complex, which is the catalytically active spliceosome[5].

1.3 Alternative Splicing

In many cases, the native transcripts of eukaryotic genes are spliced in more than one way to give rise to different isoforms. This allows the generation of a set of different proteins from one gene. Alternative splicing is a major contributor to protein diversity in metazoan organisms, thus in humans, about 95% of multi-exonic genes are alternatively spliced[6]. Different genes encode different number of isoforms: for instance some have two isoforms but others have thousands[7].

Alternative splicing takes many different forms. Exons can be included into mRNA or skipped (exons skipping or cassette exons). The positions of 5' and 3' splice sites can shift to make exons longer or shorter (alternative 5' or 3' sites). Introns that are normally excised can be retained in the mRNA (intron retention) (Figure 5). In addition to these changes in splicing, genes can also vary in transcriptional start site or polyadenylation site[8].

Although these mechanisms describe basic patterns of splicing, they do not reflect the full the complexity of splicing events. Exons are not randomly chosen for splicing; instead, splicing is regulated by

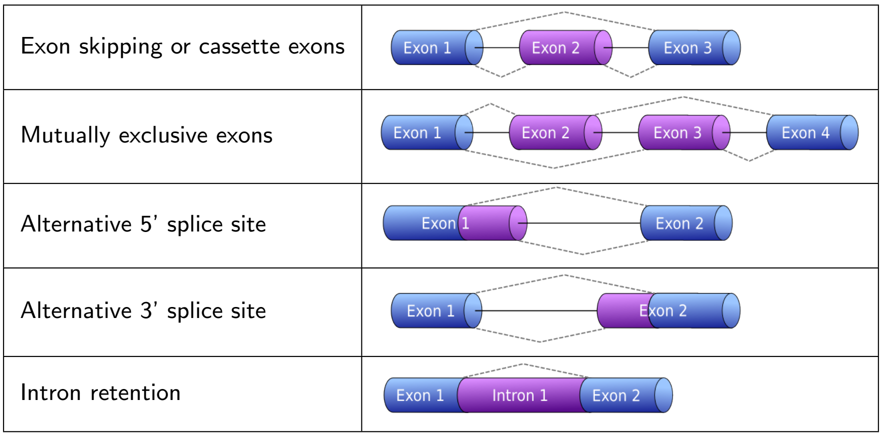


Figure 5 | The traditional classification of basic types of alternative splicing events.

trans-acting RNA-binding proteins (repressors and activators), which bind to their corresponding *cis*-activating regulatory sites (silencers and enhancers) on the mRNA. The secondary structure of mRNAs can also regulate splicing events.

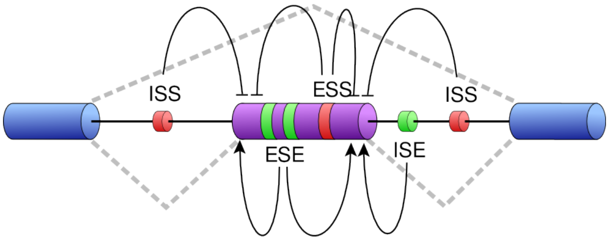


Figure 6 | Cis-regulatory elements.
ESE: exonic splicing enhanced; ESS: exonic splicing silenced; ISE: intronic splicing enhancer; ISS: intronic splicing silencer (Matlin 2005).

Cis-regulatory elements are conventionally classified according to their location and their ability to promote or inhibit inclusion of exons. Exonic splicing enhancers (ESEs) or silencers (ESSs) have an effect on the exon in which they reside, whereas intronic splicing enhancers (ISEs) or silencers (ISSs) influence adjacent splice sites or exons from an intronic location[9, 10] (Figure 6).

Splicing Activators Heterogeneous nuclear ribonucleoproteins (hnRNPs) in vertebrates and serine/arginine-rich proteins (SR proteins) in metazoans are involved in regulating and selecting splice sites[11] (Figure 7). Other proteins like TIA1 (binds to U-rich elements), NOVA1 and NOVA2 (bind to YCAY motifs), FOX1 and FOX2 (bind to UGCAUG motifs) and others also activate splicing events[5]. They often control alternative splicing by selecting the specific splice sites to be used.

Splicing Repressors Splicing repressors can sterically block the binding of splicing regulators. For example, the Hu/ELAV family of proteins inhibit U1 snRNP-binding by competing with the of TIA1-binding at the 5'SS of an exon of the *neurofibromatosis type 1* pre-mRNA[12]. FOX1 and FOX2 inhibit E complex formation by binding close to the ESE that is otherwise bound by the activators TRA2 and SRp55 in the *CALCA* pre-mRNA. Thus FOX1 and FOX2 prevent the recruitment of U2AF and prevent the spliceosome from assembling[13]. Finally, the polypyrimidine-tract binding protein (PTB) blocks the binding of U2AF^{65} [14].

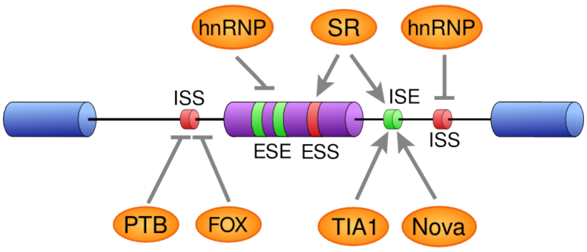


Figure 7 | Splicing auxiliary proteins and their binding sites (Keren 2010).

1.4 Alternative Splicing and Evolution: Alu Elements

Alternative splicing is a major mechanism generating transcriptomic and proteomic diversity. The diversity can be generated by several events. Alternative splicing produces protein isoforms with different molecular and biological functions, especially when alternative exons encode for specific protein domains. Comparative analysis of the human and mouse genomes has revealed that diversity in alternative splicing is often associated with recent creation and/or loss of exons.

One way to gain an exon is through the exonisation of transposable elements (TEs). About half of the human genome is derived from TEs, and these repeat-forming elements - particularly *Alu* elements - can become exonised. About 4% of human genes contain TE motifs in their coding regions that suggest exonisation events[15].

Exonisation of TEs is observed in 53% of 'orphan genes' (genes with a limited phylogenetic distribution, *i.e.* homologous genes that are retricted to closely related organisms) which shows the involvement of TE exonisation in species-specific adaptive processes. The formation of alternative exons from *Alu* elements permits new functions to be established without losing the original function of a protein[15].

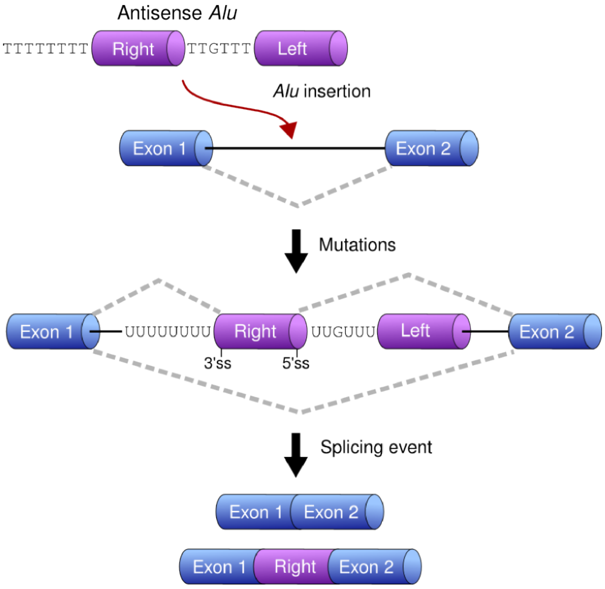


Figure 8 | Exonisation of Alu elements (Keren 2010).

Alu elements belong to the family of short interspersed elements (SINEs) and account for more that 10% of the human genome. A typical *Alu* is around 300 nucleotides (nt) long and contains two similar monomeric segments (the right arm and the left arm) joined by an A-rich linker and followed by a poly(A) tail-like region. *Alus* preferentially insert into the introns of primate genes[16] and is proceeded by retrotransposition.

85% of exonisation events occur in the antisense orientation. In such orientation, the consensus *Alu* sequence carries multiple sites that are similar to splice sites. For example, the poly(A) tract of the right arm in the antisense orientation creates a strong polypyrimidine tract (PPT), recognised as a specific binding site for U2AF^{65} to recruit the splicing machinery. To become an alternative exon, *Alu* elements require few supplementary mutations in the 3'SS and 5'SS[15, 16] (Figure 8).

1.5 Experimental Approaches

Genome-wide analysis of alternative splicing is an ongoing a challenge in biological research. By using high-throughput approaches, it is becoming easier to investigate splicing events. The methods to identify splicing events and the binding of RNA-binding proteins (RBPs) used in this study are splice-junction microarrays, RNA-seq and iCLIP (individual-nucleotide resolution UV cross-linking and immunoprecipitation).

1.5.1 Splice-Junction Microarrays

Before the development of the splice-junction microarrays, alternative splicing was studied on a case-by-case basis. In recent years, numerous studies have investigated global properties of alternative splicing using microarrays providing information about mRNA isoforms. We will describe briefly how to detect alternative transcripts with splice-junction microarrays.

The general principles of the Affymetrix microarrays are described in Figure 9. The mRNAs are extracted from cells, they are then reverse transcribed and transcribed again to incorporate nucleotides marked with biotin into the cRNA sequence. Then, the cRNA is fragmented and hybridised to the microarray. The microarray contains probes with a specific sequence that will bind the complementary cRNA (hybridisation). After the hybridisation, the microarray is washed to remove the cRNAthat ais not specifically bound. An anti-biotin antibody coupled with a fluorophore is used to reveal the hybridisation. The fluorescent signal is then detected and analysed, assuming that the intensity measured is proportionally related to the abundance of hybridised cRNA.

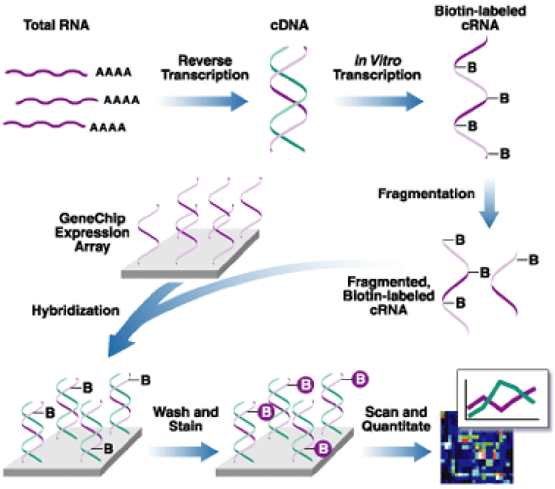


Figure 9 | Protocol of DNA microarrays.
Taken from affymetrix.com.

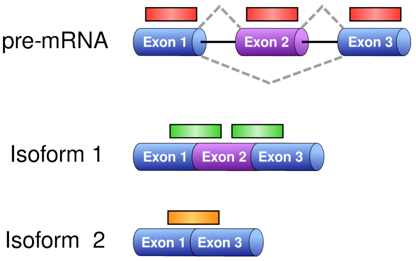


Figure 10 | Probes used to detect an exon skipping event.

In the case of splice-junction microarrays, the aim is to detect the different splice forms present in cells. The microarrays contain multiple probes for each alternative splicing events, as shown in Figure 10. Exon-body probes (red) are used to monitor the inclusion of each of the three exons and junction probes are used to monitor the two different types of junctions formed by the inclusion of alternative exon[17] (green and yellow).

Figure 11 shows example data for a casette exon. The blue bars represents the signal measured when the alternative exon is included into the mature mRNA (isoform 1), and the grey bars the case when the exon is skipped[17] (isoform 2).

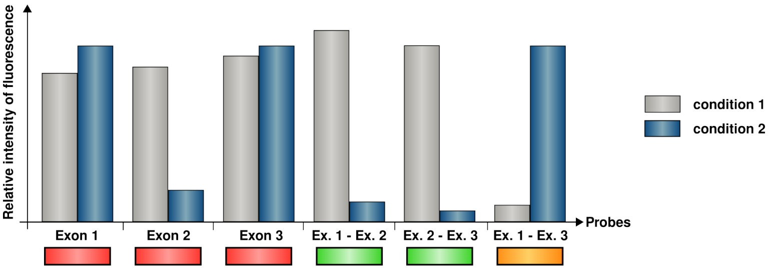


Figure 11 | Example of signals detected for a cassette exon in two cases. In blue, the cassette exon is included. In grey, the cassette exon is skipped.

1.5.2 RNA-seq

RNA sequencing (RNA-seq) is a recently developed technique using high-throughput sequencing to measure the transcriptome. mRNA is extracted from cells and converted into cDNA by reverse transcription with a poly(dT) primer binding the poly(A) tail. The cDNA is then fragmented into short reads and sequenced. The reads are mapped to a reference genome sequence. Reads overlapping exon bodies are mapped first. Then different methods can be used to discover reads spanning known or novel exon-exon junctions, and are invaluable to identify different isoforms[18]. Figure 12 depicts the mapping of the reads to a reference genome.

1.5.3 Individual-nucleotide Resolution UV Cross-linking and Immunoprecipitation (iCLIP)

iCLIP provides a robust methodtTo detect protein-RNA interactions in a context of an intact cell and determine the exact sequence recognised by the protein (Figure 13).

Cells are irradiated with ultra-violet (UV), leading to the formation of covalent bonds between protein and RNA. This method uses the natural photoreactivity of the nucleobases, especially pyrimidines, and of specific aminoacids, such as Cys, Lys, Phe, Trp and Tyr at 254 nm wavelength. The UV irradiation at 254 nm only cross-link proteins with nucleic acids, thus only the direct protein-RNA interactions are detected[19].

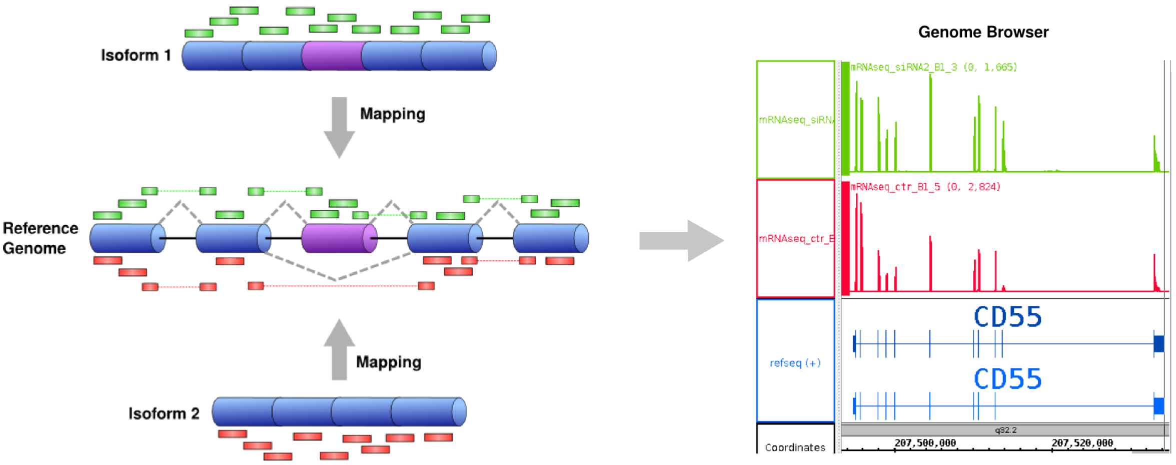


Figure 12 | RNA-seq mapping method.
The reads from two samples from two conditions are represented in red and green. They are mapped to a reference genome. Exon-exon junctions are detected by spanning reads that overlap both exons.

After UV irradiation, the covalently linked RNA is co-immunoprecipitated with the RNA-binding protein (RBP) with a specific antibody and a 3' adaptor is ligated to the RNA. The RBP is then digested by proteinase K which leaves a polypeptide at the cross-linking nucleotide. The mRNA is converted to cDNA by reverse transcription, using a primer comprising two cleavable adaptor regions and a three-nucleotide random barcode. The remaining polypeptide causes premature truncation of reverse transcription at the cross-link site, so that the last nucleotide added during the reverse transcription corresponds to the nucleotide directly downstream of the cross-link nucleotide.

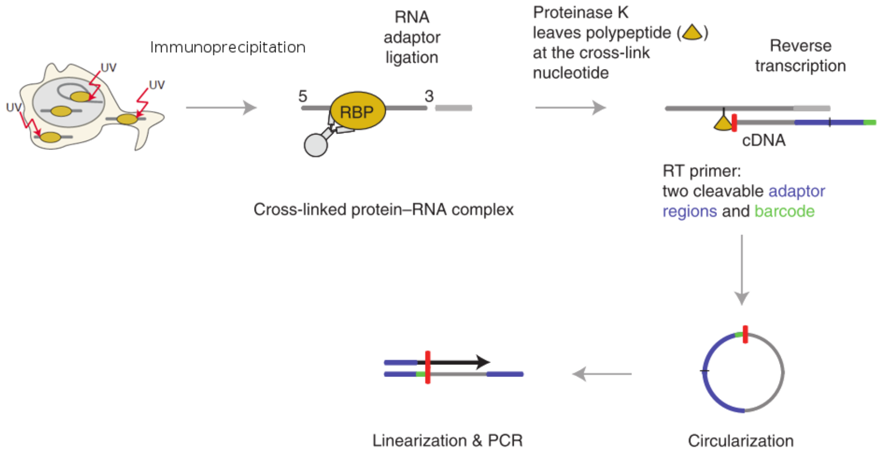

UV
UV
UV
Immunoprecipitation
RNA adaptor ligation
5
RBP
3
Cross-linked protein–RNA complex
Proteinase K leaves polypeptide ( ) at the cross-link nucleotide
Reverse transcription
cDNA
RT primer: two cleavable adaptor regions and barcode
Circularization
Linearization & PCR

Figure 13 | Schematic representation of the iCLIP protocol (König 2010).

The cDNA is circularised and linearised by cleaving the adaptor with a specific retriction enzyme. The cDNA is then amplified and sequenced using Illumina high-throughput sequencing. The reads are mapped to the human genome after removing the adaptors from both ends of the sequences. The reads presenting the same barcode and the same position are removed to prevent PCR amplification artefacts. We obtain the number of binding events for a single nucleotide by counting the cDNAs with a different barcode that identify the same cross-link nucleotide (cDNA count, Figure 14).

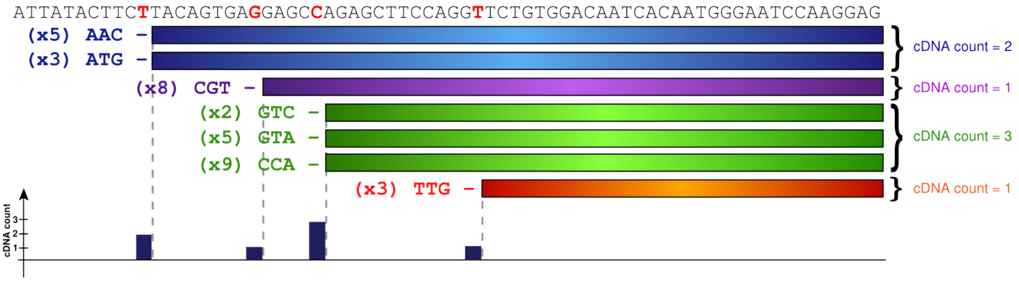


Figure 14 | cDNA count.
The reads (colored rectangles) are mapped to a referencd genome. The barcode (three random nucleotides on the left of the reads) allows to distinguish the PCR duplicates (number in brackets) from number of cross-linked mRNAs (cDNA count on the right) (König 2010).

1.6 Molecular Mechanism of mRNA Splicing Regulation by hnRNP Particles

The aim of this project is to gain understanding of the molecular mechanism of splicing regulation by hnRNP particles in human. We especially focus on hnRNP C, which appears to have a role in alternative splicing.

A first study was carried out in cooperation between three institutions: the MRC-LMB (Medical Research Council - Laboratory of Molecular Biology), the EMBL-EBI (European Molecular Biology Laboratory - European Bioinformatics Institute), both located in Cambridge (UK), and the Faculty of Computer and Information Science in Ljubljana, Slovenia[20]. The first publication described the novel iCLIP experiment and revealed the the involvement of hnRNP C in splicing regulation. The present work continues this project and aims to unravel the role of hnRNPC in the control of Alu exonisation and the interaction of hnRNPC with the component of the splicing machinery U2AF[65] . My work has focused so far on the later question and is being carried out at the Luscombe Group at EMBL-EBI, in collaboration with the Ule Group from the MRC-LMB.

1.6.1 hnRNP Particles

hnRNP C is a 304 amino-acid protein of the heterogeneous nuclear ribonucleoprotein family (hnRNP) and binds to poly-uridine (poly(U)) tracts on RNA. It is one of the most abundant proteins in the nucleus. hnRNP C1 and C2 are protein isoforms which derive from an alternative splicing event (hnRNP C2 is 13 amino acids longer than hnRNP C1). Each monomer contains three regions: an N-terminal RNA-recognition motif (RRM), a C-terminal auxiliary domain that is rich in acidic residues, and a motif that promotes the oligomerisation[21]. They form a tetramer composed of three hnRNP C1 and one hnRNP C2. Figure 15 shows the tetramer conformation by electron microscopy. The 3D structure of hnRNP C is not completely solved. The tetramer binds approximatively 235 nucleotides and is thought to pack RNA in a similar fashion to DNA within nucleosomes[22]. hnRNP C only present in the nucleus most of the time, but it can also be found in the cytoplasm during mitosis when the nuclear membrane is diseassembled[23].

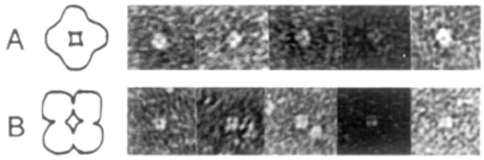


Figure 15 | Electron micrographs of negative staining of hnRNP C tetramer.
The most compact and common forms of the hnRNP C tetramer are shown (Rech 1995).

1.6.2 How can hnRNP C regulate splicing events?

Despite its abundance and the large number of studies, the exact role of hnRNP C is not clearly identified yet[24]. However, several studies have indicated of hnRNP C in alternative splicing events. The iCLIP experiment performed by Ule and colleges showed that hnRNP C promotes the exclusion of exons by binding the polypyrimidine tract at 3' splice sites. By contrast, exons included by the action of hnRNP C do not appear to have a specific binding site. The study proposed a model for splicing regulation[20] (Figure 16). As the average size of an exon is about 160 nt, when hnRNP C binds directly downstream an exon, the entire exon is wrapped aroud the tetramer and excluded from the mRNA. When hnRNP C wraps up an intron, it helps the splicing process by moving the two splice sites closer to each other. By using the iCLIP technic coupled with splice-junction microarrays and RNA-seq data, we will try to understand more precisely how hnRNP C can influence alternative splicing. As hnRNP C binds to poly(U) tracts like other well known proteins involved in the splicing process, there is a hypothesis that it prevents the binding of splicing regulators and as a consequence influences the choice of the exon to splice. Interesting candidates for competitive binding are PTB and U2AF[65] .

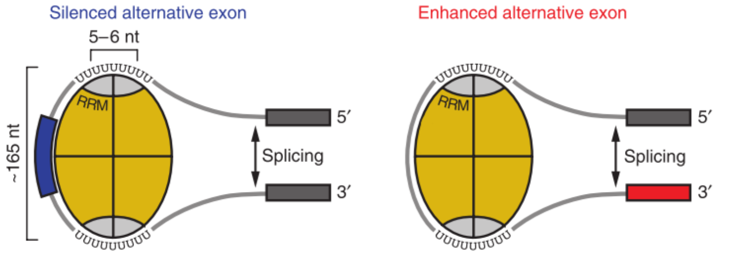


Figure 16 | A model of the hnRNP C tetramer binding at silenced and enhanced alternative exons.
The tetramer is represented in yellow. The grey correspond to the RNA-recognition motif, binding 6-5 nucleotides poly(U) tracts. hnRNP C wraps up arround 165 nt. The left schematic depict how hnRNP C silences the blue exon, and the right part how it enhances the inclusion of the red exon (König 2010).

The polypyrimidine tract-binding protein (PTB) is known as a splicing repressor and also as a splicing activator according to the *cis*-active site it binds. PTB can bind upstream of the 3' splicing site of an exon and prevent the binding of U2 auxiliary factor (U2AF) which is essential for the recruitment of the spliceosome. In addition, PTB can bind at intronic splicing enhancers (ISE), and promote exon inclusion[25, 26].

U2AF is composed of a large and a small subunit: U2AF^{65} and U2AF^{35}. The large subunit U2AF^{65} is well known to interact with polypyrimidine tracts upstream of the 3' splice site. It appears early in the assembly of the spliceosome and is essential for defining the 3' splice site[27]. When the polypyrimidine tract is not accessible, U2AF^{65} cannot recognise the splice site and the exon is excluded from the mRNA.

In this study, we will assess whether hnRNP C affects the regulatory factors PTB and U2AF^{65} .

2 Material and Methods

The study was conducted in two steps. In the first step members of the Ule Group performed the wet-lab experiments at the MRC-LMB. In the second step, the data analysis was performed by the members of the Luscombe Group, including myself.

2.1 Experiments and Available Data

All experiments were performed with HeLa cells, which is one of the oldest and most commonly used human cell lines.

2.1.1 Knockdown of hnRNP C

To understand how hnRNP C influences the splicing regulation, we have to see how splicing is affected when hnRNP C is absent from cells. For that, expression of hnRNP C is silenced using a small interfering RNAs (siRNAs), which are a class of double-stranded RNA molecules that are naturally found in the cell. Their most notable role is their involvement in the silencing of exogenous RNA *e.g.* from viruses. Double-stranded siRNAs are recognised within the cytoplasm by a protein complex called RNA-induced silencing complex (RISC). RISC becomes active by liberating one strand of the siRNA and using it as a template to recognise complementary mRNA. When it binds a complementary strand, it activates RNase which degrades the bound mRNA.[28]. This process was discovered by Andrew Z. Fire and Craig C. Mello who received the Nobel price in 2006 for this discovery. Since 2006, RNA interference is commonly used to silence the expression of a gene by synthetising siRNAs that bind specifically to the target mRNA.

Table 1 | siRNA sequences used to knockdown hnRNP C.

siRNA	Sequence
siRNA1	GCUUUGCCUUCGUUCAGUAUGUUAA
siRNA2	AAGCAGUAGAGAUGAAGAAUGAUAA

Table 1 represents the two siRNAs that were used to knockdown hnRNP C and Figure 17 shows where they bind to the hnRNP C transcript.

By silencing hnRNP C we can analyse the splicing events that are influenced by this protein. If an exon appears to be skipped more frequently in the knockdown, it means that inclusion of this exon is enhanced by hnRNP C. By contrast, if an exon is included more in the knockdown, it can be considered to be silenced by hnRNP C.

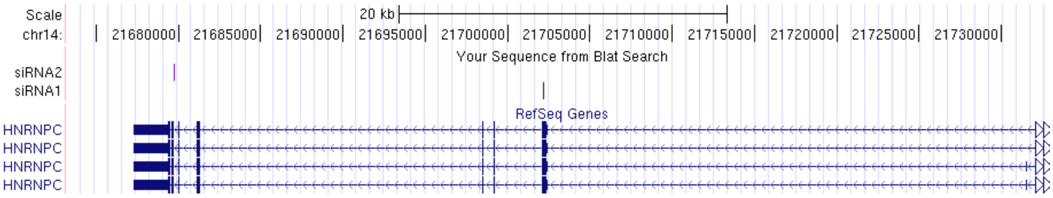


Figure 17 | Genomic location of the siRNAs used for the knockdown of hnRNP C. Screenshot from the UCSC Genome Browser.

2.1.2 Splice-Junction Microarray

The data below was acquired from the König 2010 study[20] (see the supplementary methods for more details). The labeled mRNA from hnRNP C knockdown and control HeLa cells was hybridised on a non-commercial human exon-junction microarray. The microarray produced by Affymetrix, monitors 260,488 exon-exon junctions of 315,137 exons. The analysis was done with the ASPIRE3 algorithm (Analysis of SPlicing Isoform REciprocity, version 3). ASPIRE3 predicts splicing changes from reciprocal sets of microarray probes and recognises inclusion or skipping of an alternative exon.

Table 2 | Preview of the splice-junction microarray data.

rowID	hg19 in	hg19 skip	strand hg19	annotated exons	ΔT	ΔI rank
1	chr1:321033-321264	chr1:320939-322038	+	CE	1.11	-0.05
2	chr1:322039-322228	chr1:320654-324288	+	CE	1.11	0.11
3	chr1:776581-778969	chr1:764485-783034	+	CE	0.82	0.03
4	chr1:783035-783186	chr1:764485-787307	+	CE	0.82	-0.12
5	chr1:784865-784982	chr1:783187-787307	+	CE	0.82	0.09
6	chr1:787308-787490	chr1:764485-788051	+	CE	0.82	-0.07
7	chr1:788052-788146	chr1:787491-788771	+	CE	0.82	-0.19

An extract of the results from the ASPIRE3 analysis is presented Table 2. The data structure includes the position of an alternative exon and its flanking constitutive exons. The "hg19 in" column corresponds to the chromosome number followed by the position of the first nucleotide of the preceding intron, and the last nucleotide of the proceding intron, according to the human reference genome. The "hg19 skip" column gives the positions of the first and last nucleotides of the alternative exon. The "annotated exons" column corresponds to the type of alternative exon (CE: Cassette Exon, A3SS: Alternative 3' splice site, A5SS: Alternative 5' splice site, IR: Intron Retention). ΔT represents the fold change in transcript abundance, and ΔI rank is the change in relative exon abundance corrected by a modified t-test.

ΔI rank is used to detect the significant exon splicing changes between the wild type and the knockdown of hnRNP C. The alternative exons with ΔI rank $\leq$ -1 are considered to be significantly more skipped in the wild type (defined as silenced by hnRNP C), and those with ΔI rank $\geq$ 1 are considered to be more included in the wild type but skipped in the knockdown (defined as enhanced by hnRNP C).

ASPIRE3 was able to monitor 53,624 alternative splicing events (AS). In the first part of study, we focus only on cassette exons annotated in Ensembl. From the 19,027 cassette exons in the dataset, 871 were significantly changed (Table 3).

Table 3 | Splice-junction microarray exons regulated by hnRNP C.

Total AS	CE	Enhanced CE	Silenced CE
53,624	19,027	450	421

2.1.3 RNA-seq

The RNA-seq was performed on poly(A)+ mRNA (*i.e.* a poly(dT) oligonucleotide that is complementary to the poly(A) tail is used as a primer during the reverse transcription) using Illumina GAII (Genome Analyser II) sequencing system. The reads were aligned to the human genome version hg19 using TopHat[29]. TopHat is a splicing-aware alignment program that maps reads that overlap exon-exon-junctions seperated by introns. By splitting the reads, it is able to detect splice junctions. The alternative exons were annotated using Ensembl annotation, and the novel exons detected by the RNA-seq were predicted using Cufflinks[30]. The RNA-seq experiments upon knockdown of hnRNP C with siRNA1 and siRNA2 are referred to as RNA-seq 1 and RNA-seq 2, respectively, in the following analysis.

Table 4 | Preview of the RNA-seq data.

space	start	end	strand	reg.type_siRNA1
chr1	169798404	169798590	+	enhanced
chr6	41057318	41057449	+	enhanced
chr6	41057943	41058048	+	enhanced
chrX	64749474	64749758	-	enhanced
chr12	9098825	9099001	-	enhanced
chr8	17417837	17418042	+	enhanced
chr3	129286335	129286447	-	enhanced

Table 4 shows the data from annotated cassette exons that are differentially spliced in the knockdown of hnRNP C by siRNA1 that are used for the first part of the study. The coordinates are taken from the Ensembl annotation. "Enhanced" means that the exon is downregulated in the knockdown compared to the wild type, *i.e.* its inclusion is enhanced by hnRNP C under normal condition. By contrast, a "silenced" exon is upregulated in the knowkdown. The number of regulated casette exons detected upon knockdown with siRNA1 and siRNA2 is presented Table 5.

Table 5 | RNA-seq exons regulated by hnRNP C in the two knockdowns.

siRNA	Total CE	Enhanced CE	Silenced CE
siRNA1	5,368	4,085	1,283
siRNA2	488	309	179

2.1.4 iCLIP Experiment

iCLIP experiments give the position of RNA-protein cross-link events at a single nucleotide resolution. Table 6 gives an example of iCLIP data on hnRNP C. The columns "Chr" and "Xnt" give the chromosome and the position of the cross-link nucleotide respectively. The count corresponds to the number of cDNAs detected at this position, which is also the number of binding events on this nucleotide.

Table 6 | Preview of the iCLIP data.

Chr	Xnt	strand	count
chr10	100000151	+	3
chr10	100000225	+	1
chr10	100000226	+	5
chr10	100000244	+	2
chr10	100000263	+	7
chr10	100000350	+	1
chr10	100000381	+	10

2.2 Computational Methods

2.2.1 RNA splicing map

The RNA splicing map is an integrative approach to study splicing regulation. By combining genome-wide protein-RNA interaction maps (iCLIP) with the results of splicing profiling (splice-junction microarray or RNA-seq), we are able to determine the position-dependent regulatory effects of an RNA-binding protein[31]. This representation was first used to study the position-splicing regulation by Nova[32], and later for hnRNP C[20]. This previous study combined hnRNP C-dependent splicing events detected by microarray analyses with hnRNP C positioning determined experimentally from iCLIP.

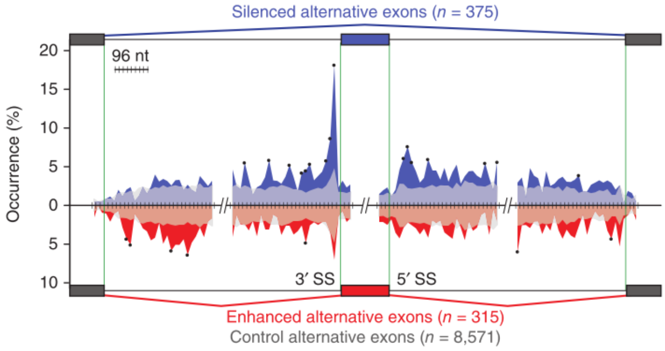


Figure 18 | Initial RNA splicing map of hnRNP C.
Results from König 2010 study.

Figure 18 shows the RNA splicing map that was obtained using a set of regulated exons from the splice-junction microarray previously described. In my work, I produced new RNA splicing maps using several new iCLIP data sets and new splicing profiles from RNA-seq.

I wrote the scripts in R using the Bioconductor packages (in particular GenomicRanges[1]). Bioconductor is an open source software project that provides tools for the analysis of high-throughput genomic data. GenomicRanges provides an efficient tool to manipulate interval positions to find overlap among data sets.

The general principle of the RNA splicing map is shown Figure 19. Exons are separated according to their splicing profile *i.e.* enhanced, silenced or non-regulated (control). A window of 450 nucleotides (nt) is defined around each end of an exon: 400 nt into the intron and 50 nt into the exon (grey rectangles).

[1] http://www.bioconductor.org/packages/release/bioc/html/GenomicRanges.html

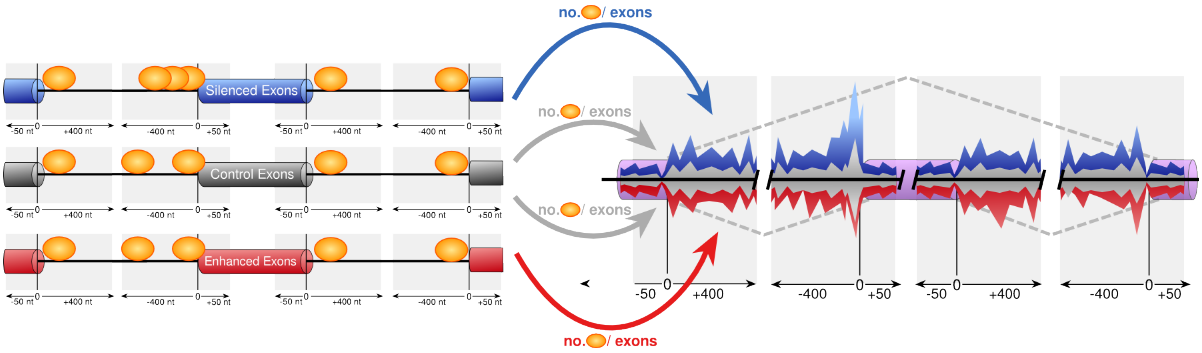


Figure 19 | Schematic of the RNA splicing map.
Exons are separated according to how a protein regulates their inclusion (splicing profile). Exons are either enhanced (red), silenced (blue) or non-regulated (grey). A 450-nucleotide window is defined around each end of an exon (grey rectangles). From each exon, the number of protein binding events (orange ellipse) is summed, averaged and plotted on the RNA splicing map.

If exons are shorter than 100 nt or introns are shorter than 800 nt, a smaller window is used. In each window, I calculated the relative positions of the RBP binding events. The RBP binding events are summarised by summing the cDNA count for each position and averaged by the total number of regulated exons. To smooth the map, the number of binding events per position is averaged every 15 nt. The three splicing profiles are then plotted in one graph. The silenced exons are in blue, the enhanced in red and shown as negative values, and the non-regulated exons are in grey, plotted twice as positive and negative values.

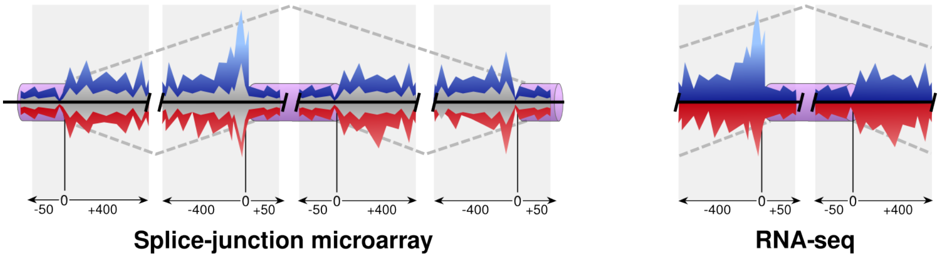
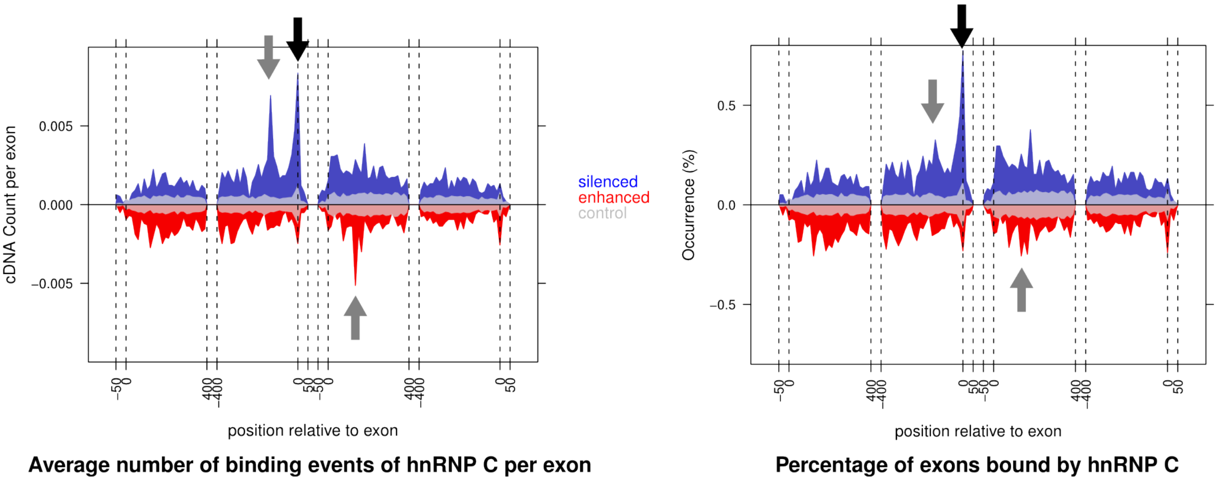


Figure 20 | The RNA splicing map design for the splice-junction microarray and the RNA-seq data.
The RNA splicing map design for splice-junction microarray data shows the cassette exons, the flanking exons, and the control set of non-regulated exons. RNA-seq data focus on the cassette exons thus the RNA splicing map shows only the regulated exons.

The splice-junction microarray provides the positions of the cassette exons and their flanking exons. This information comes from the probe design used for the microarray. The RNA-seq data only provides the position of the annotated cassette exons. For the moment, we do not have a set of control exons from the RNA-seq. Therefore, the RNA splicing maps drawn with the regulated exons from RNA-seq are smaller and do not have any control, but the RNA splicing maps from the microarray can be used as a reference. A comparison of the RNA splicing maps from the two techniques are presented in Figure 20.

For each method of splicing profiling, the script produces two RNA splicing maps. One provides the average number of binding events per exon (cDNA count per exon), whereas the other provides the percentage of exons bound by the protein per nucleotide (occurrence). iCLIP data are not normalised to the transcript abundance. When a mRNA is highly expressed, more binding sites are present for the protein, which can generate higher cDNA counts for the same exon. As a result, the map which counts the average number of binding events can contain a high peak resulting from only one exon, which does not reflect a general position-dependent binding of the protein. By comparing the the two maps, we can discriminate global position-dependent bindings from artefacts.



Average number of binding events of hnRNP C per exon

Percentage of exons bound by hnRNP C

Figure 21 | RNA splicing maps of hnRNP C.
The RNA splicing maps for hnRNP C generated with the published iCLIP data are plotted with the regulated exons from the splice-junction microarray. The RNA splicing map that counts the binding events (on the left) presents two high peaks, one on the silenced exons and the other on the enhanced exons (marked by grey arrows), whereas the occurrence graph (on the right) does not have them. These peaks on the left correspond to two non-coding RNA highly expressed in the nucleus.

Figure 21 shows a typical example of an artefact caused by two highly expressed non-coding RNA. The RNA splicing map that counts the binding events, presents two peaks upstream of the 3' SS of the silenced cassette exons, and one peak downstream of the 5' SS of the enhanced exons (marked by grey arrows). The two peaks correspond to two highly abundant snoRNAs bound by hnRNP C (small nucleolar RNA, SNORD17 and SNORD12B). These two non-coding RNAs have been removed from the set of regulated exons for all remaining results. On the RNA map counting the number of exons, only one peak on the silenced exons is conserved (marked by black arrows).

3 Results

3.1 Genome Browser Preview

The first way to explore the data is to transform the files into bed format and import them in a Genome Browser. The Integrated Genome Browser[33] (IGB) is a tool used to plot RNA-seq results with iCLIP data on the genome.

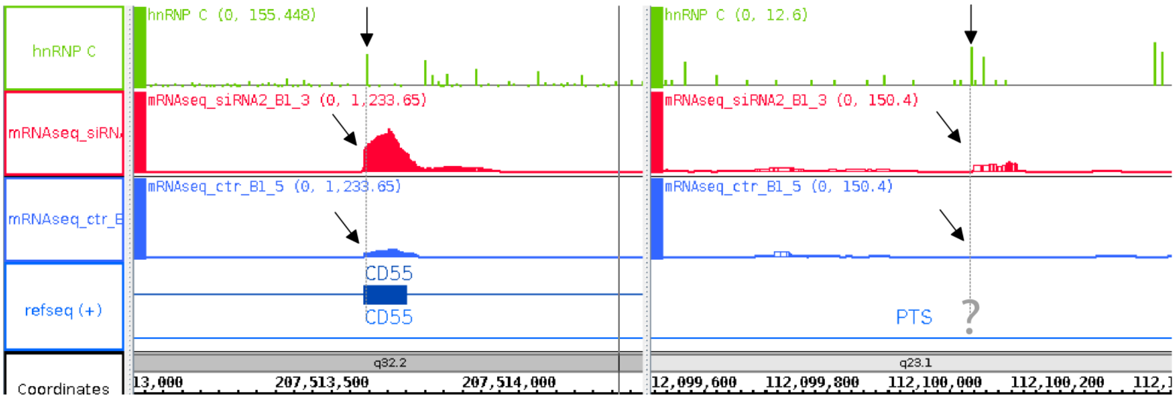


Figure 22 | The genomic location of hnRNP C cross-link nucleotides on two genes.
In this diagram, green track plots the number of hnRNP C binding events from iCLIP, red track plots the RNA-seq reads from the knockdown of the hnRNP C protein, and blue track plots reads from the wild type. The arrows point to the binding sites of hnRNP C. In the knockdown, we see the inclusion of exons that are mostly excluded in the wild type. On the left, we see expression of cassette exon from *CD55* gene, but on the right, we see the expression of an unannotated exon (marked with ”?”).

Figure 22 shows two splicing events linked with the hnRNP C binding position. The number of hnRNP C binding events are represented in the green track, RNA-seq reads from the wild type in the blue track and RNA-seq reads from the knockdown of hnRNP C by siRNA2 in the red track. On the left, the *CD55* gene contains an annotated cassette exon (blue box) that is mostly silenced in the wild type but highly expressed in the knockdown. Upstream of this exon, we can see a peak of hnRNP C binding on the green track. On the right, we also see hnRNP C binding events on the *PTS* gene, however, here there is expression in the knockdown but there is no annotated exon.

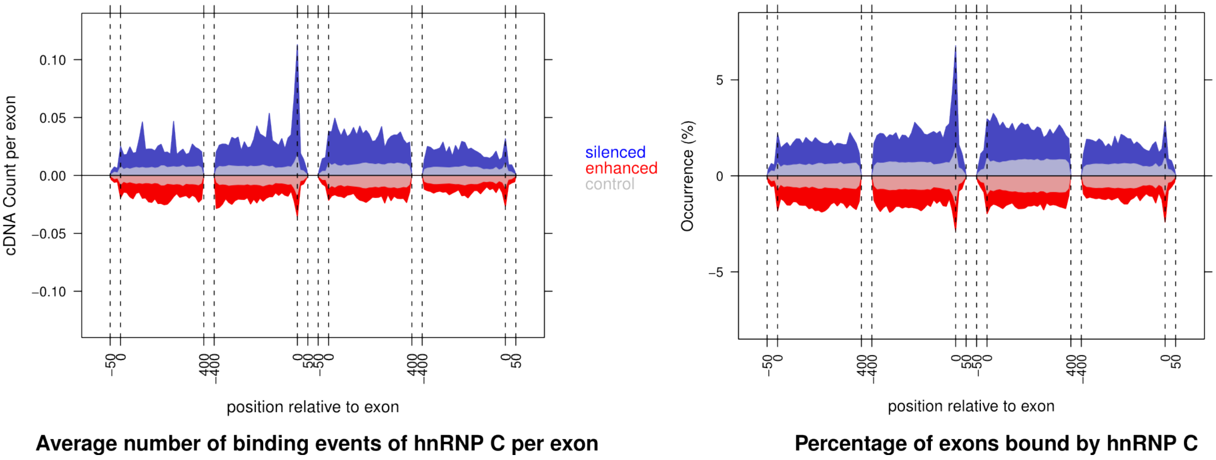
Looking at the exons one by one is an exhausting task. The RNA splicing map is an appropriate tool to summarise the effect of hnRNP C on the splicing regulation.

3.2 RNA splicing map on the Sets of Regulated Exons

3.2.1 hnRNP C RNA splicing map

Several iCLIP experiment for hnRNP C have been performed to calibrate the protocol. The following RNA splicing maps use the largest iCLIP dataset. This dataset differs from the previously published iCLIP experiment where it uses a more specific antibody that increases the quantity of extracted cross-linked pre-mRNA.

Figure 23 demonstrates that my script is able to generate similar results to the published RNA



Average number of binding events of hnRNP C per exon

Percentage of exons bound by hnRNP C

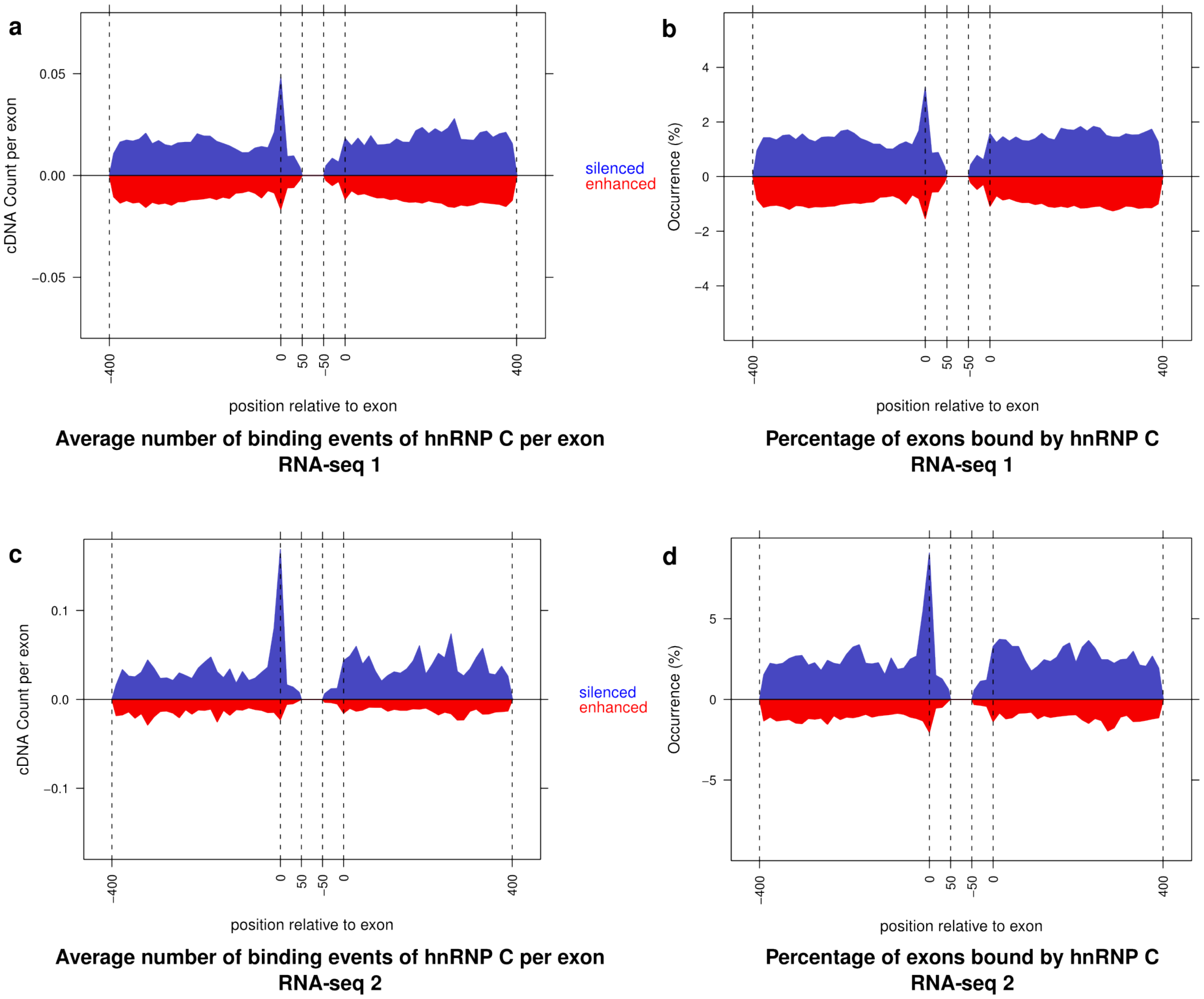
Figure 23 | Microarray-based RNA splicing map of hnRNP C.

splicing map. The new iCLIP experiment produced thirteen times more data than the one presented in the first publication and the global effect of hnRNP C on the exon silencing is still visible. hnRNP C highly binds to the polypyriminine tract upstream of the 3’ SS of the silenced cassette exons, whereas exons enhanced by hnRNP C do not present a position-dependent binding of the protein. The maps do not show strong artefacts, and the percentage of bound exons in the area of the high peak is ten times higher than in the previous study.

The following RNA splicing maps are based on the regulated annotated exons detected by the two RNA-seq experiments (Figure 24). As previously explained, these maps neither show the flanking exons nor a control set of non-regulated exons.

We can see that even through the RNA-seq 1 experiment has detected more regulated exons, the RNA splicing map reveals that this set of exons might contain a lot of false positives. Indeed, the binding profile of hnRNP C is closer to the profile on the non-regulated exons shown in grey in Figure 23. The peak upstream of the silenced cassette exons is still visible, but less notable than in the RNA-seq 2-based map. The set of regulated exons detected with the RNA-seq 2 experiment is smaller, but seems to be more specific to show the position-dependent regulation by hnRNP C. As the maps plotting the number of bound exons do not show any noticeable artefacts, and the two maps produce identical profiles, we will not show both maps in the following sections.

In conclusion, the two different maps based on the microarray and the RNA-seq 2 are able to depict the position-dependent regulation by hnRNP C and support the regulation model described in Figure 16. The dataset provided by RNA-seq 1 is less reliable and less informative than the microarray and the RNA-seq 2 datasets.



Average number of binding events of hnRNP C per exon RNA-seq 1

Percentage of exons bound by hnRNP C RNA-seq 1

Average number of binding events of hnRNP C per exon RNA-seq 2

Percentage of exons bound by hnRNP C RNA-seq 2

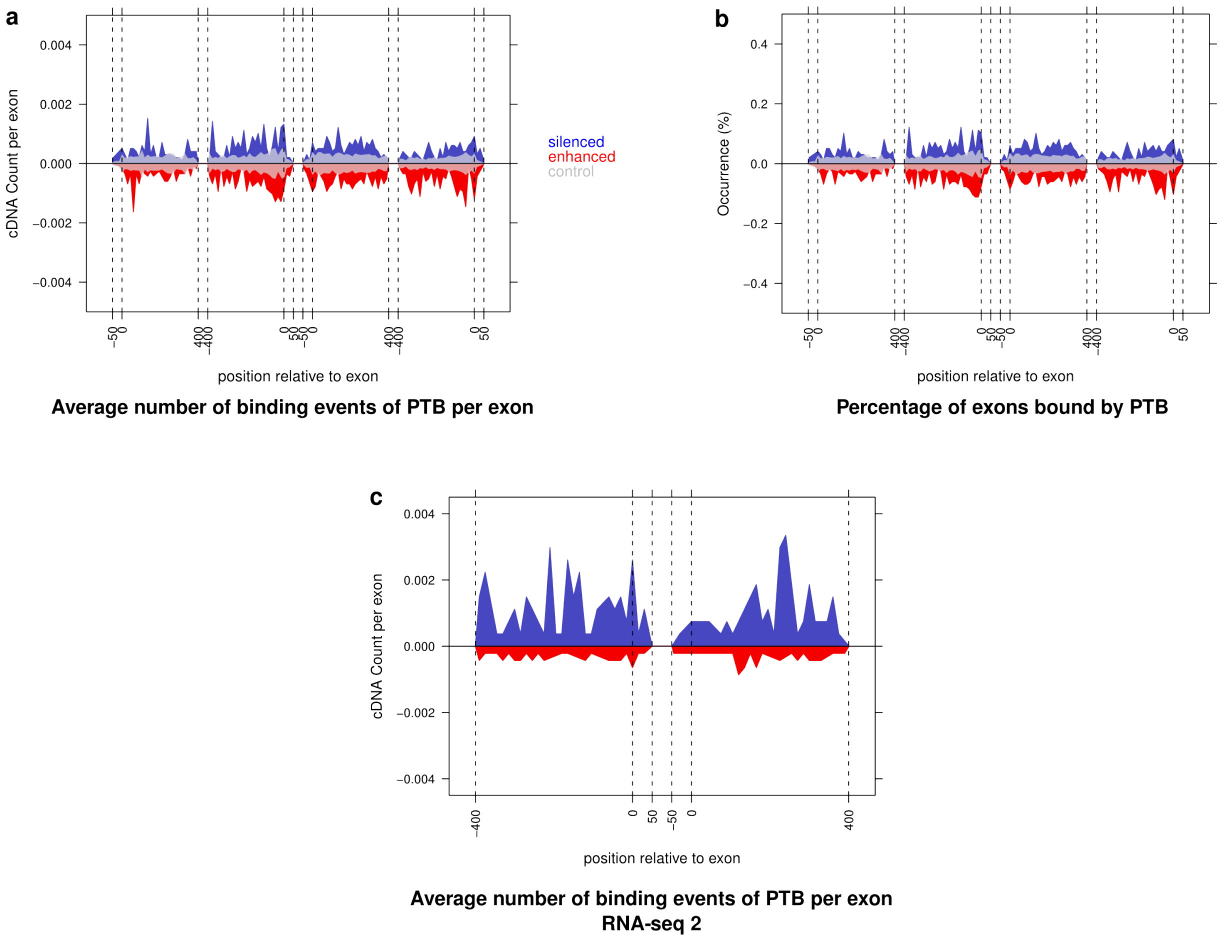
Figure 24 | hnRNP C RNA splicing map with the RNA-seq exon sets.

3.2.2 PTB RNA splicing map

Having defined two reliable sets of exons regulated by hnRNPC and knowing the position-dependent regulation profile for hnRNP C, we can now investigate important proteins related to the hnRNP C function. The first protein we test is PTB. This protein is known to bind polypyrimidine tracts and regulate the exon inclusion.

By plotting the binding location of PTB on the exons regulated by hnRNP C, we want to see if PTB plays a role in hnRNP C-mediated regulation. PTB has a position-dependent effect on splicing (upstream of the exon to silence it, or downstream of the exon to help its inclusion). If PTB regulates the same exons as hnRNP C, we expect to see a specific position on the RNA splicing maps.

Figure 25 shows PTB binding events on the two sets of regulated exons. The scales demonstrate that PTB binds much less than hnRNP C (0.15% maximum of bound exons, compared to about 5% for hnRNP C). The microarray-based maps (Figure 25 a and b) do not show a specific binding location of PTB on the regulated exons compared to the control. The RNA-seq 2-based map provides similar results. Indeed, it seems that PTB has more significant binding events on the silenced exons, but the



Average number of binding events of PTB per exon

Percentage of exons bound by PTB

Average number of binding events of PTB per exon RNA-seq 2

Figure 25 | PTB RNA splicing maps.

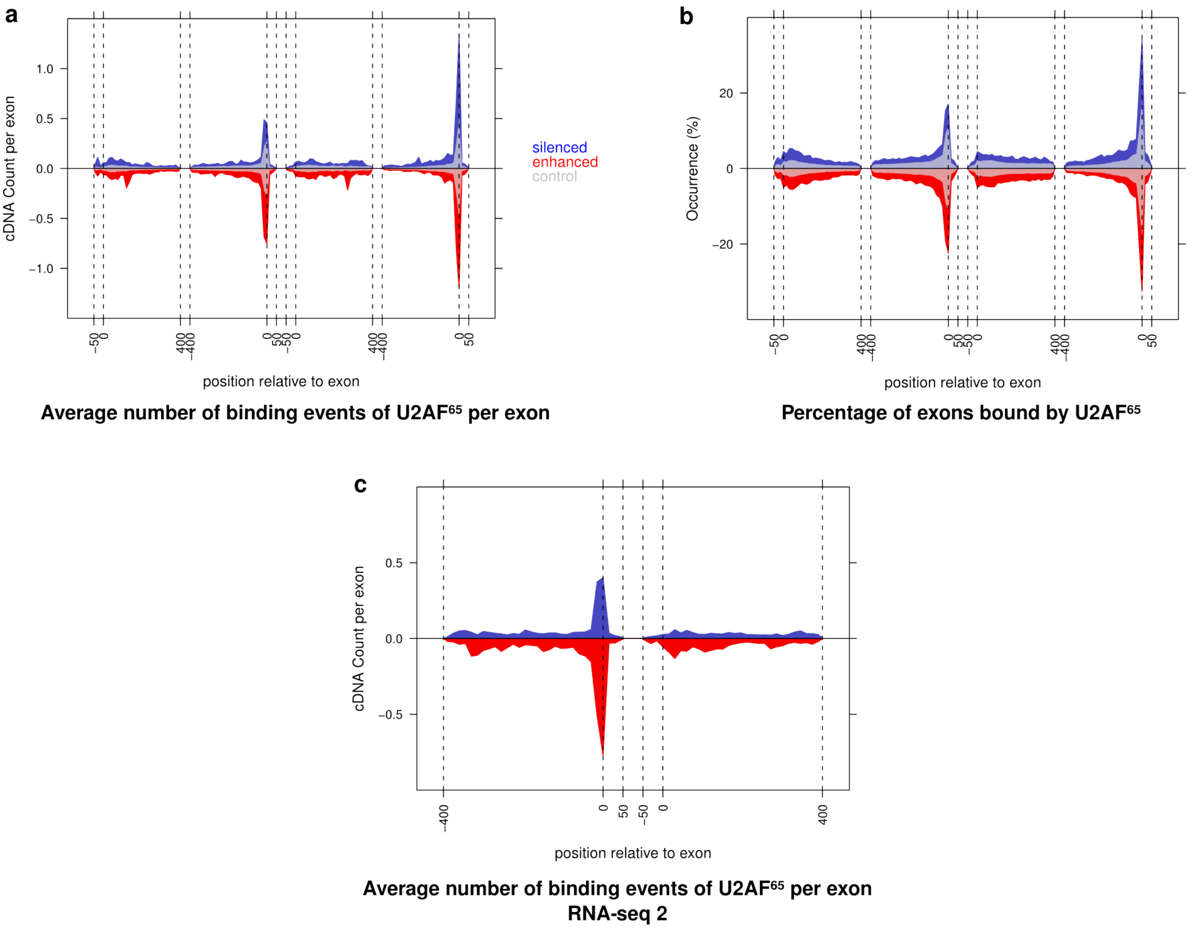
signal is so weak that an increase in one position can drastically change the proportions in the graph. Thus, the slight increase of binding events in Figure 25c is not significant.

In conclusion, I find that PTB binds the exons regulated by hnRNP C in a same way as the non-regulated exons. Thus, PTB is not involved in the splicing regulation by hnRNP C in a position-specific manner.

3.2.3 U2AF^{65} RNA splicing maps

U2AF^{65} is a protein that is directly involved in the splicing process. It recognises a polypyrimidine tract upstream of the exon and defines the 3' splice site of the intron. Like PTB, we want to see how U2AF^{65} binds on the exons regulated by hnRNP C. The binding of U2AF^{65} recruits of the spliceosome and thus causes inclusion of the exon. If hnRNP C has a direct effect on the splicing regulation, we expect to see a different profile of binding events of U2AF^{65} among the exons enhanced and silenced by hnRNP C. U2AF^{65} should bind less upstream of the silenced exons compared to the enhanced, as they are not included into the mature mRNA.

Figure 26 shows the binding profile of U2AF^{65} on the two different sets of exons regulated by hnRNP C. As expected, U2AF^{65} has a high affinity for the polypyrimidine tracts upstream of the exons. The



Average number of binding events of U2AF⁶⁵ per exon

Percentage of exons bound by U2AF⁶⁵

Average number of binding events of U2AF⁶⁵ per exon
RNA-seq 2

Figure 26 | U2AF^{65} RNA splicing maps.

first notable observation from Figure 26a is that U2AF^{65} has two times more binding events on the flanking exons than on the cassette exons. We can explain this by the fact that the flanking exons are predominantly constitutive exons, so they are always included into the mature mRNA. As a consequence they are always bound by U2AF^{65} , whereas the cassette exons are not always included, thus the average number of binding events of U2AF^{65} reflects this difference of inclusion. The second notable observation is the difference of the binding profiles on the enhanced and silenced exons. U2AF^{65} binds more on the enhanced exons than on the silenced exons.

Here we plot the U2AF^{65} binding events detected in wild type cells on hnRNP C regulated exons. If hnRNP C prevents the binding of U2AF^{65} , the binding profile of U2AF^{65} should change in cells which do not express hnRNP C.

In order to test this hypothesis, two iCLIP experiments were performed to detect U2AF^{65} binding events on two cell lines where hnRNP C was silenced using siRNA1 or siRNA2. The RNA splicing maps based on RNA-seq 2 are presented in Figure 27. The U2AF^{65} iCLIP experiment in the wild type produced slightly more reads than in the knowkdown. The number of binding events in the wild type and the knockdown were normalised to compare the two maps.

The comparison shows that the binding profile of U2AF^{65} in the knockdown of hnRNP C is completely

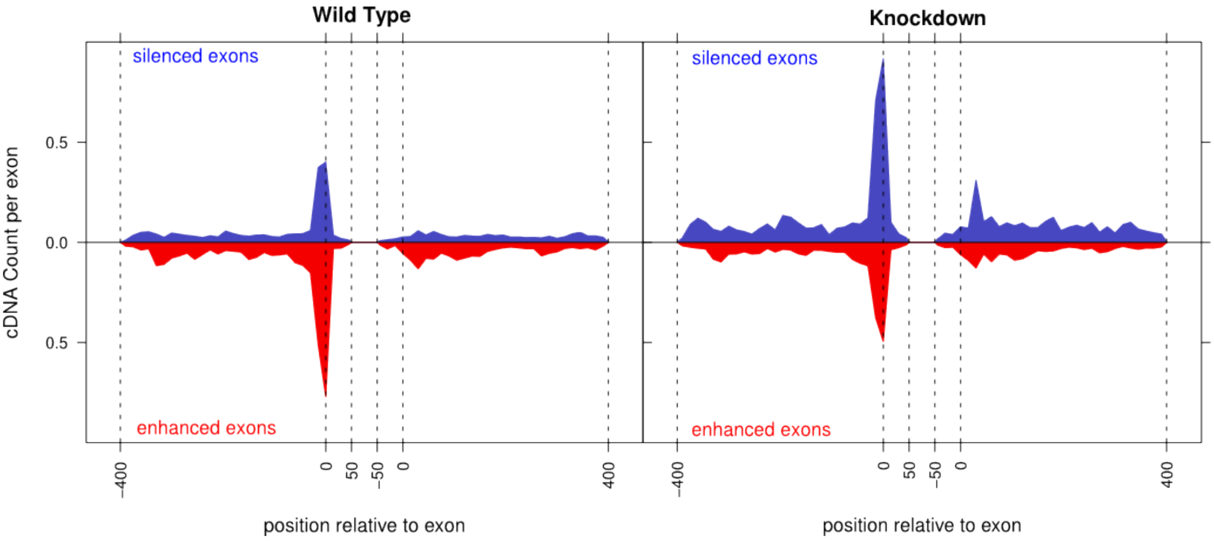


Figure 27 | RNA splicing maps of U2AF^{65} binding events in the wild type and upon the knockdown of hnRNP C with siRNA2.

inverted. In the wild type, U2AF^{65} binds two times more on the enhanced exons than on the silenced. By contrast, in the knockdown, we see that U2AF^{65} binds more on the silenced exons than on the enhanced. These results show good evidence for the competition we suggested.

hnRNP C seems to have a direct effect on the splicing decision by controlling the binding of U2AF^{65} (recall that U2AF^{65} is necessary for the recruitment of the spliceosome). Our results show that hnRNP C regulates the inclusion of exons by competing with U2AF^{65} for the polypyrimidine tract. This information enriches our understanding of the model of splicing regulation by hnRNP C presented in Figure 16.

3.3 Analysis on Alu elements

Now that we have a model of the splicing regulation by hnRNP C, we can focus on a specific set of exons that show specialised regulation by hnRNP C. We have introduced the role of alternative splicing in evolution, and illustrated it with the example of the exonisation of the *Alu* elements. In the next sections, we will begin to look at how hnRNP C is involved in the silencing of the *Alu* elements.

The previous analysis described above uses annotated exons from Ensembl. The remaining part of this paper describes analysis using unanotated exons differentially expressed from the RNA-seq. The RNA-seq data from the wild type and the knockdown of hnRNP C was analysed using Cufflinks to reveal the unannotated exons. By comparing these exons with the *Alu* dedicated database using RepeatMasker, we extracted the hypothetical exonisation events of *Alu* elements.

3.3.1 U2AF^{65} on Alu exons

To confirm the model of splicing regulation by hnRNP C on the *Alu* exons, I performed the same analysis on the new set of *Alu* exons from RNA-seq 2. Figure 28a shows the binding profile of hnRNP C on the *Alu* exons detected by RNA-seq 2. hnRNP C binds to these exons with exactly the same profile as the previously analysed exons. In the wild type, U2AF^{65} binds more on the enhanced *Alu* exons, and much less on the silenced. In the knockdown of hnRNP C, both enhanced and silenced exons are bound by U2AF^{65} (Figure 28c), but we can see that there is a much stronger increase detected on the silenced exons. Our results show that hnRNP C has a clear effect on the exclusion of the *Alu* exons, suggesting that hnRNP C silences the *Alu* exons which prevents the emergence of new isoforms in human.

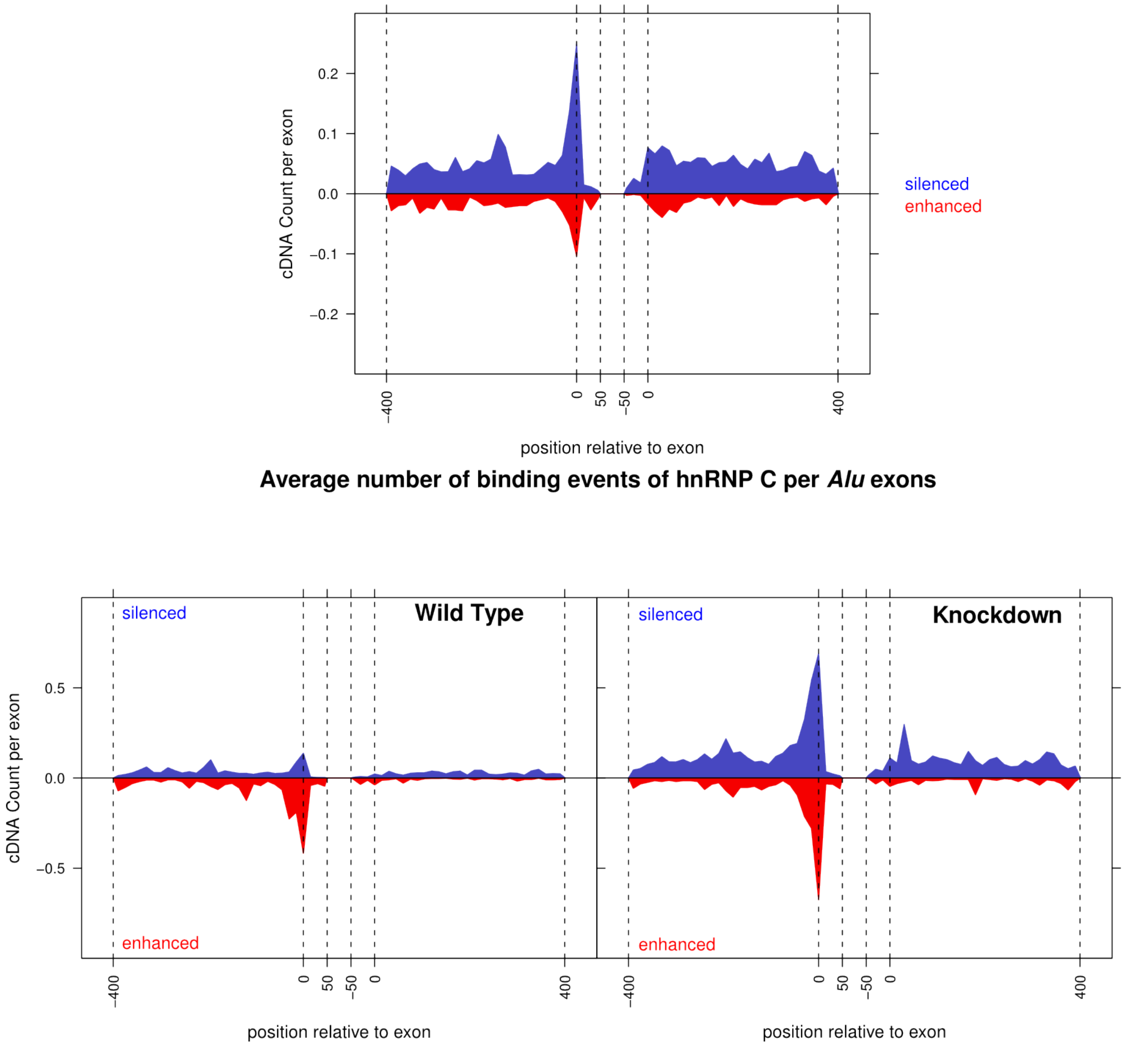


Figure 28 | RNA splicing maps of hnRNP C and U2AF^{65} binding events in the wild type and upon the knockdown with siRNA2.

3.3.2 Heatmap

Another way to see the competition between hnRNP C and U2AF^{65} on the *Alu* exons is to generate a Heatmap. A Heatmap is a graphical representation of data in a two-dimentional table where the values are represented as colors. Here, I generate a Heatmap that represents the number of binding events of hnRNP C and U2AF^{65} in the wild type and the knockdown in a window of 80 nucleotides upstream of the 3' SS (Figure 29). Each line represents one regulated *Alu* exon, and the gradient color encode the number of binding events (the darker the blue, the more binding occurs).

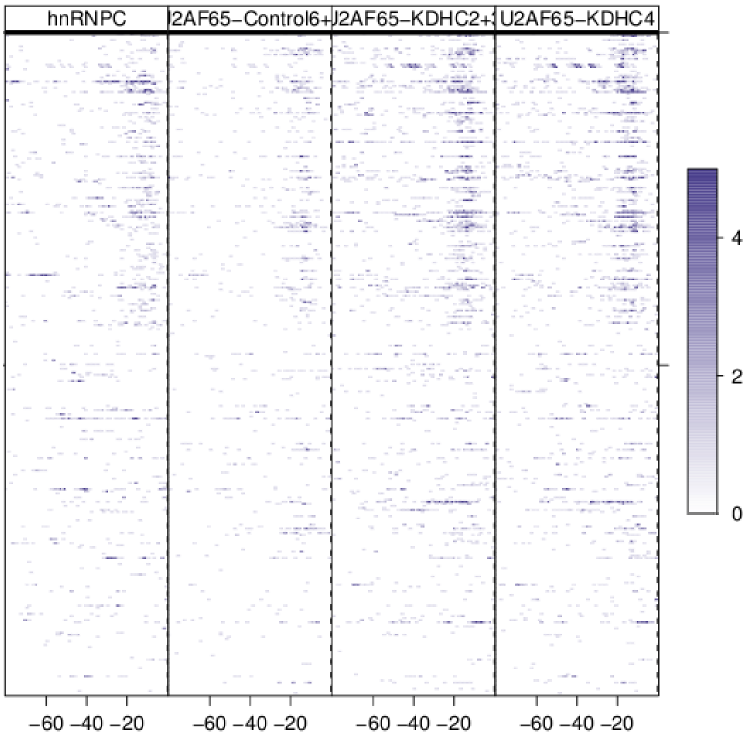


Figure 29 | Heatmap of the binding events of hnRNP C and U2AF^{65} in the wild type and the knockdown of hnRNP C on Alu exons.
Each line of the heatmap corresponds to the 80 nucleotides upstream an Alu exon. The blue spots are the binding events of the proteins. The gradient correspond to the cDNA count.

With this figure, we can directly visualise the competition between the two proteins. The first and the second columns represent hnRNP C and U2AF^{65} binding in the wild type, whereas the third and the fourth column show U2AF^{65} binding in the two knockdown experiments.

The script used to generate this figure could not be completed. In the future, the sequences will be clustered to better visualise the concurrent pattern between hnRNP C and U2AF^{65} binding. However, even with this incomplete heatmap we can see that the sequences are ordered, and that the top half of the Heatmap reveals competition. In the wild type, we see that U2AF^{65} binds less at positions where hnRNP C highly binds, whereas in the knockdown, U2AF^{65} binds much more at the same location than hnRNP C.

With the Heatmap, we can see that hnRNP C regulate a large fraction of the *Alu* exons. This reinforces the idea that a competition between hnRNP C and U2AF^{65} for the polypyrimidine tract exists upstream of the exons.

4 Discussion and Conclusion

Alternative splicing is a highly regulated process that increases the proteomic diversity. It plays an important role in cellular differentiation, in evolution and in disease. The principal factors of splicing regulation are RNA-binding proteins (RNPs). By combining genome-wide protein-RNA interaction maps with the analysis of splicing profiles, we are now able to determine the position-dependent regulatory effect of an RBP.

The aim of this study was to gain on understanding of the molecular mechanism of splicing regulation by hnRNP C in human. A previous study revealed the role of this RNA-binding protein in splicing regulation[20], but we continued to investigate hnRNP C's' role in affecting splicing decision. We used an integrative approach to study splicing regulation by hnRNP C by combining the data from several techniques: iCLIP, splice-junction microarray and RNA-seq.

We have demonstrated that hnRNP C has a high affinity for the polypyrimidine tracts upstream of the 3' splice site of the cassette exons. hnRNP C does not compete with the polypyrimidine tract-binding protein (PTB) to regulate splicing. However, we have shown that hnRNP C and U2AF^{65} share the same binding location. Our results show that hnRNP C is likely to compete with U2AF^{65} to regulate the splicing decision of cassette exons. A model of the competition is shown in Figure 30.

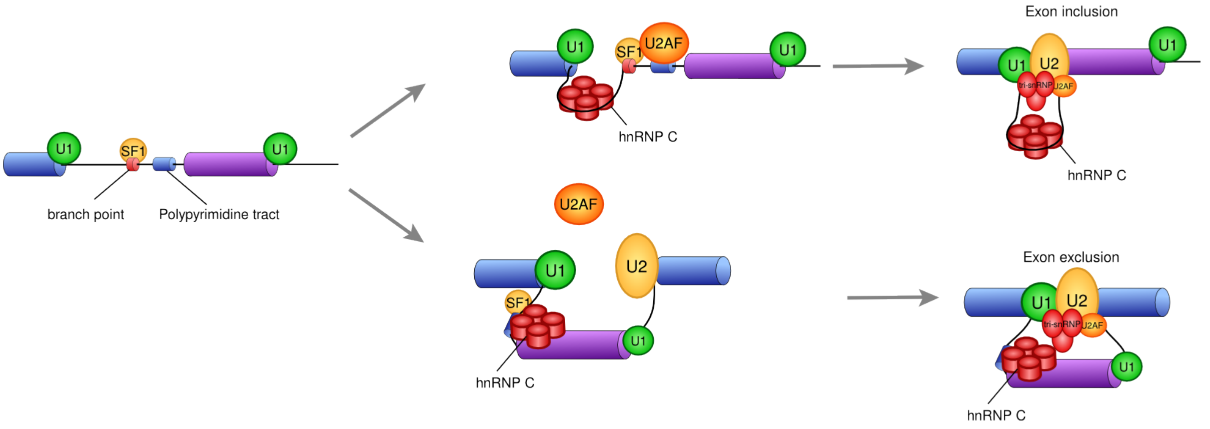


Figure 30 | Model of the splicing regulation by hnRNP C

We propose that by binding the intron in a non-position-dependent manner, hnRNP C packs the mRNA and moves the splice sites closer to facilitate splicing[20]. When hnRNP C binds the polypyrimidine tract upstream of the 3' splice site, it prevents the binding of the large subunit of U2 auxiliary factor that is necessary to recruit the spliceosome. As a consequence, the U1 snRNA molecule of the spliceosome interacts with U2 binding at the following exon, and thus ultimately induces the skipping of the cassette exon, as depicted in Figure 30.

hnRNP C might not regulate all the cassette exons, but the model of the regulation seems to apply for a particular type of exons involved in evolution. The *Alu* transposable elements are known for their ability to be exonised and to induce new isoforms. hnRNP C appears to have a high affinity for those exons and to silence their erroneous inclusion. The next part of the project will focus on the study of the regulation of the *Alu* elements by hnRNP C and its involvement in human evolution.

References

1. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 3171–5 (1977).

2. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**, 1–8 (1977).

3. Alberts, B. *et al. Molecular Biology of the Cell in Cell 5th*, vol. 54 (Garland Science, Taylor & Francis Group, USA, 2008), 5 edn.

4. Nilsen, T. W. The spliceosome: the most complex macromolecular machine in the cell? *BioEssays : news and reviews in molecular, cellular and developmental biology* **25**, 1147–9 (2003).

5. Chen, M. & Manley, J. L. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature reviews. Molecular cell biology* **10**, 741–54 (2009).

6. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics* **40**, 1413–5 (2008).

7. Black, D. L. Protein Diversity from Alternative Splicing. *Cell* **103**, 367–370 (2000).

8. Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annual review of biochemistry* **72**, 291–336 (2003).

9. Wang, Z. & Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA (New York, N.Y.)* **14**, 802–13 (2008).

10. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: towards a cellular code. *Nature reviews. Molecular cell biology* **6**, 386–98 (2005).

11. Barbosa-Morais, N. L., Carmo-Fonseca, M. & Aparício, S. Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome research* **16**, 66–77 (2006).

12. Zhu, H., Hinman, M. N., Hasman, R. A., Mehta, P. & Lou, H. Regulation of neuron-specific alternative splicing of neurofibromatosis type 1 pre-mRNA. *Molecular and cellular biology* **28**, 1240–51 (2008).

13. Zhou, H.-L. & Lou, H. Repression of prespliceosome complex formation at two distinct steps by Fox-1/Fox-2 proteins. *Molecular and cellular biology* **28**, 5507–16 (2008).

14. Saulière, J., Sureau, A., Expert-Bezançon, A. & Marie, J. The polypyrimidine tract binding protein (PTB) represses splicing of exon 6B from the beta-tropomyosin pre-mRNA by directly interfering with the binding of the U2AF65 subunit. *Molecular and cellular biology* **26**, 8755–69 (2006).

15. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nature reviews. Genetics* **11**, 345–55 (2010).

16. Häsler, J., Samuelsson, T. & Strub, K. Useful 'junk': Alu RNAs in the human transcriptome. *Cellular and molecular life sciences : CMLS* **64**, 1793–800 (2007).

17. Shai, O., Morris, Q. D., Blencowe, B. J. & Frey, B. J. Inferring global levels of alternative splicing isoforms using a generative model of microarray data. *Bioinformatics (Oxford, England)* **22**, 606–13 (2006).

18. Ozsolak, F. & Milos, P. M. RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics* **12**, 87–98 (2010).

19. Ule, J., Jensen, K., Mele, A. & Darnell, R. B. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods (San Diego, Calif.)* **37**, 376–86 (2005).

20. König, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology* **17**, 909–916 (2010).

21. Han, S. P., Tang, Y. H. & Smith, R. Functional diversity of the hnRNPs: past, present and perspectives. *The Biochemical journal* **430**, 379–392 (2010).

22. Rech, J. E., LeStourgeon, W. M. & Flicker, P. F. Ultrastructural morphology of the hnRNP C protein tetramer. *Journal of structural biology* **114**, 77–83 (1995).

23. Schepens, B. *et al.* A role for hnRNP C1/C2 and Unr in internal initiation of translation during mitosis. *The EMBO journal* **26**, 158–69 (2007).

24. Hossain, M. N., Fuji, M., Miki, K., Endoh, M. & Ayusawa, D. Downregulation of hnRNP C1/C2 by siRNA sensitizes HeLa cells to various stresses. *Molecular and cellular biochemistry* **296**, 151–7 (2007).

25. Llorian, M. *et al.* Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nature structural & molecular biology* **17**, 1114–1123 (2010).

26. Warf, M. B., Diegel, J. V., von Hippel, P. H. & Berglund, J. A. The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9203–8 (2009).

27. Kent, O. A., Reayi, A., Foong, L., Chilibeck, K. A. & MacMillan, A. M. Structuring of the 3' splice site by U2AF65. *The Journal of biological chemistry* **278**, 50572–7 (2003).

28. Timmons, L., Tabara, H., Mello, C. C. & Fire, A. Z. Inducible systemic RNA silencing in Caenorhabditis elegans. *Molecular biology of the cell* **14**, 2972–83 (2003).

29. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)* **25**, 1105–11 (2009).

30. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, 511–515 (2010).

31. Witten, J. T. & Ule, J. Understanding splicing regulation through RNA splicing maps. *Trends in genetics : TIG* **27**, 97–89 (2011).

32. Ule, J. *et al.* An RNA map predicting Nova-dependent splicing regulation. *Nature* **444**, 580–6 (2006).

33. Nicol, J. W., Helt, G. A., Blanchard, S. G., Raja, A. & Loraine, A. E. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics (Oxford, England)* **25**, 2730–1 (2009).

Master 2 Modélisation des systèmes biologiques, UFR SVE, Université de Rennes 1, France 2010-2011

Molecular Mechanism of Splicing regulation by hnRNP Particles

Isabelle Stévant

Abstract

Alternative splicing is a highly regulated process increasing proteomic diversity. It plays an important role in cellular differentiation, in evolution and disease. The principal factors of splicing regulation are RNA-binding proteins (RNPs).

In this study, we highlight the role of the RNA-binding protein hnRNP C in the process of RNA splicing by using several data types such as splice-junction microarrays, RNA-seq and iCLIP (individual-nucleotide-resolution UV cross-linking and immunoprecipitation) using computational methods.

We used an integrative approach that combines the data from different techniques to produce an efficient graph able to depict the position-dependent regulation of hnRNP C.

The results shows that hnRNP C and $U2AF^{65}$ compete for the same binding site. $U2AF^{65}$ is a necessary protein to recruit the spliceosome, which is the protein complex that catalyse the splicing reaction. hnRNP C silcences exons by preventing the binding of $U2AF^{65}$.

We now begin to focus on how hnRNP C can regulate the expression of the Alu transposable elements, which are involved in evolution.

Keywords : RNA-binding proteins, alternative splicing, spliceosome, iCLIP, RNA-seq

Résumé

L'épissage alternatif est un processus fortement régulé participant à la diversité du protéome. L'épissage joue un rôle important dans la différenciation cellulaire, l'évolution ou encore dans des maladies. Le principal facteur de régulation de l'épissage sont les protéines de liaison à l'ARN.

Dans cette étude, nous révélons le rôle de la protéine de liaison hnRNP C dans le processus d'épissage en étudiant les données obtenues par diverses techniques telles que les splice-junction microarrays, RNA-seq et iCLIP (individual-nucleotide-resolution UV cross-linking and immunoprecipitation).

Nous avons utilisé une approche intégrative combinant les données issues de ces techniques pour produire un graphique capable de décrire la régulation de l'épissage par hnRNP C.

Les résultats montrent que hnRNP C est en compétition avec $U2AF^{65}$ pour le même site de liaison. $U2AF^{65}$ est une protéine nécessaire au recrutement du spliceosome, le complexe protéique catalysant la réaction d'épissage. hnRNP C exclue un exon de l'ARN messager en empêchant $U2AF^{65}$ de s'y lier.

Nous travaillons maintenant à comprendre comment hnRNP C régule l'expression des éléments transposables Alu, connus pour leur impliquation dans l'évolution.

Mots clés : protéines de liaison à l'ARN, épissage alternatif, spliceosome, iCLIP, RNA-seq