Universidade de Lisboa
Faculdade de Medicina

GENOME WIDE MINING OF ALTERNATIVE SPLICING IN METAZOAN MODEL ORGANISMS

Inês Maria de Stoop Camões Teixeira Guerra Mollet

Tese orientada pela Prof. Doutora Maria Carmo-Fonseca
Instituto de Medicina Molecular, Faculdade de Medicina

Doutoramento em Ciências Biomédicas
Lisboa, 2008

INSTITUTO DE MEDICINA MOLECULAR

UNIDADE DE BIOLOGIA CELULAR

A impressão desta dissertação
foi aprovada pela Comissão Coordenadora
do Conselho Científico da Faculdade
de Medicina de Lisboa em reunião
de 17 Julho de 2008

As opiniões expressas nesta publicação são da exclusiva responsabilidade da sua autora

TO MY FAMILY

ACKNOWLEDGEMENTS

Funding

This project was supported by the Muscular Dystrophy Association (MDA3662), the European Commission (LSHG-CT-2005-518238, EURASNET), and Fundação para a Ciência e Tecnologia, Portugal (PTDC/SAU-GMG/69739/2006).

Professional acknowledgements

For professional support I am indebted first and foremost to my supervisor **Maria Carmo-Fonseca** who has successfully lead this project through many twists and turns, and given much visionary input; to **Juan Valcárcel** (CRG-Centre de Regulació Genómica, Barcelona, Spain) for many detailed discussions on the data and to members of his lab **Claudia Ben-Dov** who collaborated in revising the output data, **Britta Hartmann** who gave useful feedback on the Drosophila genome data, and **Josefin Lundgren** who validated alternative splicing patterns for genes involved in Myotonic Dystrophy; to **Samuel Aparicio** (BC Cancer Agency, Vancouver, British Columbia, Canada) for initial ideas critical to the success of this project, and much very productive advice; to **Daniel Felício Silva** (IMM - Instituto de Medicina Molecular, Lisbon, Portugal) with whom I collaborated in the design of the ExonMine web- interface and who has patiently complied with all my requests for changes; to **Ana Rita Grosso** (IMM - Instituto de Medicina Molecular, Lisbon, Portugal) who designed the five web-interface tools to aid in laboratory validation of data, was involved in creating the ExonMine interface and with whom I collaborate on microarray data analysis; to **Pedro Eleutério** (IMM - Instituto de Medicina Molecular, Lisbon, Portugal) for his invaluable support in maintaining the ExonMine server, receiving all the updates and keeping the data safe; to **Tito Santos Silva** (Escola de Engenharia da Universidade Católica Portuguesa) for advice on optimizing data formats for the ExonMine database and Web interface; to **Teresa Duarte Pacheco** (IMM - Instituto de Medicina Molecular, Lisbon, Portugal) with whom I collaborated in the first validation of alternative splicing regulatory sequence elements detected using ExonMine data; to **Margarida Gama Carvalho** (IMM - Instituto de Medicina Molecular, Lisbon, Portugal) for many useful discussions; to **Francisco Enguita** (IMM - Instituto de Medicina Molecular, Lisbon, Portugal) for all his knowledge on microRNAs and the successful validation of experiments

I asked him to perform on an alternative transcript carrying a microRNA; to **Anita Gomes** (IMM - Instituto de Medicina Molecular, Lisbon, Portugal) who performed validation experiments on genes involved in Myotonic Dystrophy; to **Joana Borlido** (Uro-Oncology Research Group, Department of Oncology, University of Cambridge, CR-UK Cambridge Research Institute) who validated ExonMine data for the U2AF35 family of splicing factors and for many stimulating discussions; to **Sandra Martins** (IMM - Instituto de Medicina Molecular, Lisbon, Portugal) who was involved in validation U2AF35 family splice variants and has given very useful feedback; to **Clarie Shovlin** (Imperial College London, UK) who has taken great interest in the ExonMine data and is currently collaborating in validation experiments; to **Nuno Morais** (Cancer Genomics Program, Department of Oncology University of Cambridge, Hutchison-MRC Research Centre, Cambridge, UK) who patiently gave me invaluable bioinformatics support at the start of the project; to **Bin Liu** for great informatics support with Linux; to **Juan José Lozano** (CRG-Centre de Regulació Genómica, Barcelona, Spain) and **Natalie Thorne** (Department of Oncology, University of Cambridge, Cancer Research UK Cambridge Research Institute, Cambridge, UK) for discussions on microarray analysis; to **James Brenton** (Department of Oncology, University of Cambridge, Cancer Research UK Cambridge Research Institute, Cambridge, UK) for much support on the workings of Linux; and to **José Rino**, **José Braga**, **Joana Desterro**, **Ana Teresa Maia** and **Ahmed** for many useful discussions and encouragement.

CONTENTS

PUBLICATIONS

Mollet, I.G., Ben-Dov, C., Felício-Silva, D., Grosso, A.R., Eleutério, P., Silva, T.S., and Carmo-Fonseca, M. (2008). **ExonMine - Unconstrained mining for exons and alternative splicing.** *BMC Bioinformatics.* Submitted
Contribution: Mollet, I.G., conceived the idea for the ExonMine database, wrote all the Perl scripts, generate the pipeline which produces the data, designed the MySQL database structure, did the statistical analysis of data including comparison with microarray data, is responsible for all updates and composed the manuscript.

Mollet, I., Barbosa-Morais, N.L., Andrade, J., and Carmo-Fonseca, M. (2006). **Diversity of human U2AF splicing factors.** *FEBS J.*, 273(21):4807-16.
Contribution: Mollet, I., performed the analysis of alternative splicing within the U2AF family of splicing factors based on data generated by ExonMine.

Pacheco, T.R., Coelho, M.B., Desterro, J.M., Mollet, I., and Carmo-Fonseca, M. (2006). **In vivo requirement of the small subunit of U2AF for recognition of a weak 3' splice site.** *Mol Cell Biol.*, 26(21):8183.
Contribution: Mollet, I., performed the bioinformatic search for candidate genes to validate the requirement of the small subunit of U2AF for recognition of a weak 3' splice site, based on data generated by ExonMine.

PREFÁCIO

Nesta dissertação são apresentados os resultados do trabalho de investigação desenvolvido no Instituto de Medicina Molecular na Faculdade de Medicina de Lisboa da Universidade de Lisboa, entre os anos de 2003 e 2008, sob a orientação da Professora Doutora Maria do Carmo-Fonseca.

Este trabalho foi desenvolvido para responder à escassez de dados consistentes sobre splicing alternativo e está dividido em duas partes. A primeira parte consiste no desenvolvimento de um algoritmo flexível e rápido que cria uma base de dados de splicing alternativo para qualquer organismo a partir de sequências de transcritos depositadas no GenBank, e que fornece informação num formato acessível não só para biólogos, para a análise de genes individuais, mas também para tecnologias que permitem a análise de milhares de genes em simultâneo.

O trabalho está organizado em quatro capítulos. O primeiro capítulo consiste numa introdução onde são abordados os vários aspectos relativos ao splicing, em particular ao splicing alternativo, e a outros mecanismos que resultam na variação de transcritos. É apresentado o caso particular de alterações no splicing alternativo em distrofia miotónica. A área de estudo, relativamente recente, de transcritos que não codificam proteínas, em particular os microRNAs, é em seguidas abordada. Por fim é feita uma breve introdução à tecnologia dos microarrays.

A segunda parte do trabalho descreve o método implementado pelo algoritmo desenvolvido para gerar uma base de dados de splicing alternativo para 13 modelos de organismos do reino animal.

A terceira parte apresenta um estudo estatístico dos dados produzidos, incluindo uma análise de exões até agora ignorados, e descreve e discute as várias aplicações de pequena e grande escala aos quais os dados produzidos pela base de dados foram sujeitos e que os validam. Por fim são tiradas conclusões sobre o sucesso do método e das várias aplicações efectuadas com os dados.

SUMÁRIO

Objectivos

O trabalho apresentado tem como objectivo o desenvolvimento de um algoritmo que gera uma base de dados de splicing alternativo e que detecta novos elementos envolvidos em splicing alternativo de maneira consistente e reprodutível. Este trabalho foi desenvolvido em duas fases. A primeira fase consistiu no desenvolvimento e optimização do algoritmo que cria uma base de dados de splicing alternativo para qualquer organismo a partir de informação de sequências depositados no GenBank [Mollet et al. 2008]. A segunda foi a aplicação dos dados gerados a análises de splicing alternativo de grande e pequena escala [Pacheco *et al.* 2006; Mollet *et al*. 2006].

Conceitos Teóricos

A quantidade fenomenal de sequências depositadas no GenBank todos os anos requer o desenvolvimento de algoritmos de triagem e condensação desta informação de modo a que esta seja útil ao biólogo. Só para o humano há mais de 8 milhões de sequências de transcritos depositadas. Um dos aspectos de enorme interesse é o do splicing alternativo que permite a um gene produzir variantes de uma proteína utilizando combinações alternativas de exões [Black 2003]. Para além da geração de variantes de proteínas, a geração de transcritos não codificantes é um mecanismo importante de regulação da expressão genética [Lewis *et al.* 2003; Blencowe 2006; Graveley 2001; Modrek e Lee 2002]. Estima-se que cerca de 73% dos cerca de 20,000 genes humanos sofram splicing alternativo [Johnson *et al.* 2003; Kan *et al*. 2001]. A informação útil retirada de bases de dados como o GenBank sobre splicing alternativo terá de ser útil não só para a análise de pequena escala, pelo biólogo que se dedica a aspectos específicos de alguns genes, mas também deverá ser útil para análises de grande escala, nomeadamente para análises por microarray. Os microarrays permitem monitorizar milhares genes em simultâneo num dado tecido. A comparação por microarray de dados em células de diferentes tecidos, ou de células saudáveis com células doentes, permite detectar os grupos de genes responsáveis pelas diferenças.

Desenvolvimento do algoritmo

O algoritmo foi desenvolvido na linguagem Perl interagindo com bases de dados construídas em MySQL [Mollet *et al.* 2008]. Os dados de partida consistem em tabelas de alinhamentos de transcritos (mRNA e ETSs) a um genoma pelo programa BLAT [Kent 2002], retiradas do UCSC Genome Browser [Karolchik *et al.* 2008]. A qualidade dos alinhamentos é analisada pelo programa que procede a uma triagem e à condensação de informação repetida. A informação condensada é utilizada para gerar tabelas com todo o splicing alternativo existente nos transcritos em relação a cada gene. O programa ainda determina a tradução dos transcritos, locais de poliadenilação e informação sobre expressão em diferentes tecidos ao nível de isoformas e exões individuais. O resultado consiste numa base de dados em MySQL com dez tabelas por genoma, optimizadas de modo a se poder fazer todo o tipo de pesquisas. A base de dados foi chamada ExonMine e, em colaboração com outros elementos do grupo, foi integrada num servidor no Instituto de Medicina Molecular com acesso ao público [http://www.imm.fm.ul.pt/exonmine/]. Os updates são gerados de quatro em quatro meses.

Estudo estatístico - Descoberta de novos elementos

O estudo estatístico dos dados gerados e uma comparação com outras bases de dados, revelou que a ExonMine detecta níveis mais altos de splicing alternativo do que qualquer outra base de dados. Uma análise dos dados revelou que se detectam mais exões novos, que não têm sequência codificante, no humano do que no rato. Sabe-se que transcritos não codificiantes estão envolvidos na regulação da expressão de genes e que elementos móveis (como novos exões) estão na base da evolução de genes e de genomas em geral. Este resultado identifica portanto potenciais elementos que poderão ser fundamentais para se decifrar a diferença em termos de regulação genética, entre o humano e outros organismos menos complexos.

Aplicações de grande escala

Procura de motivos envolvidos na regulação de splicing alternativo

A base de dados ExonMine foi desenhada de modo a se poderem facilmente fazer procuras de elementos regulatórios envolvidos em splicing alternativo (ou mesmo de qualquer elemento regulatório a nível de sequência primária). O sucesso deste objectivo foi

demonstrado num estudo em que, com base num pequeno algoritmo que questiona a base de dados ExonMine, se detectaram, no espaço de umas horas, combinações de elementos regulatórios que foram posteriormente validados experimentalmente e publicados [Pacheco et al. 2006] pela Doutora Teresa Pacheco.

Construção de um microarray para estudo da Distrofia Muscular

Em colaboração com o grupo do Doutor Juan Valcárcel (CRG-Centre de Regulació Genómica, Barcelona, Spain), a base de dados ExonMine foi utilizada para a construção de um microarray especificamente concebido para o estudo de alterações de splicing alternativo em distrofia muscular incluindo informação sobre cerca de 400 genes específicos de musculo e cerca de 400 factores de splicing.

Análise do microarray comercial ExonArray 1.0 ST e de tiling arrays da Affymetrix

Para aplicações de grande escala a base de dados está a ser utilizada para a análise de microarrays produzidos comercialmente pela Affymetrix com cerca de 5.5 milhões de sondas e que visam cobrir o genoma inteiro ao nível de um exão. Dados de 'tiling arrays' de Affymetrix que cobrem o genoma inteiro com uma resoluçao de 5 nucleótidos também foram comparados aos dados da ExonMine e são objecto de projectos em curso.

Aplicações de pequena escala

Análise de variação do splicing na família U2AF

A diversificação por splicing alternativo da importante família de factores de splicing U2AF foi investigada e publicada [Mollet et al. 2006]. Em todos os sete genes analisados foram detectados, com base no ExonMine, novas formas de splicing alternativo, com ou sem sequência codificante.

Análise de novos elementos na família U2AF35

Para a família de genes U2AF35 foi feita uma análise ainda mais aprofundada com o objectivo de validar as novas isoformas detectadas pelo ExonMine. Todas a isoformas, escolhidas para uma validação experimental, continham novos exões que se encontravam em humano mas não em rato, e foram confirmadas por RT-PCR por outros elementos do

grupo da Professora Doutora Maria Carmo-Fonseca. Para além da validação experimental foi ainda feita uma análise do potencial codificante das novas isoformas e da conservação dos novos exões noutros genomas. Em dois casos de isoformas contendo novos exões, mas sem sequência codificante, foi feito o estudo da estrutura secundária do mRNA e foi detectado potencial para expressão de microRNAs, um dos quais foi confirmado experimentalmente por outro elemento do grupo da Professora Doutora Maria Carmo-Fonseca.

Conclusões

Neste trabalho não só se desenvolveu e optimizou um método reprodutível e eficaz para gerar informação sobre splicing alternativo para qualquer organismo, como também se realizaram várias análises que mostram claramente que os objectivos pretendidos foram atingidos. Foram detectadas mais variações em splicing alternativo do que em qualquer outra base de dados, incluindo formas de splicing alternativo envolvendo novos exões não codificantes, até agora ignorados. Algumas das novas formas de splicing alternativo foram validadas por elementos do laboratório de Professora Doutora Maria Carmo-Fonseca, e foram aplicados os dados à análise de genes em grande escala, nomeadamente para análise de motivos envolvidos na regulação do splicing alternativo e na análise de microarrays.

Esta base de dados é além disso um projecto activo e com continuidade, tendo vários projectos dependentes dos seus dados em execução, nomeadamente: análise de microarrays independentes no laboratório de Professora Doutora Maria do Carmo-Fonseca e em colaboração com o laboratório do Doutor Juan Valcárcel (CRG-Centre de Regulació Genómica, Barcelona, Spain); análise de splicing alternativo de factores de splicing (projecto em execução); e análise de splicing alternativo em factores de coagulação pelo grupo de Claire Shovlin (Imperial College, London). A base de dados ExonMine será mantida no Instituto de Medicina Molecular da Faculdade de Medicina da Universidade de Lisboa.

OBJECTIVES

One of the difficulties in modern day genomics is that vast amounts of data are being generated at a breathtaking rate and biologists need to be able to tap into this information with a minimum of effort. This requires the biologist to acquire computational skills in order to extract information to meet his/her particular needs.

The objective of this doctorate was the development of an alternative splicing database that could be easily updated and which should be structured so as to serve both the biologist, focusing on a single gene, and the computational biologist, seeking to do large-scale analysis of alternative splicing data. The first phase of the work would consist in the development and optimization of a Perl program, which would generate the alternative splicing data from BLAT alignments of transcript sequence data deposited in GenBank. The second phase of the project would consist in validating the alternative splicing data generated in small and large-scale projects in collaboration with members of our lab and other labs. Large-scale analysis would consist in genome wide analysis of sequence elements involved in the regulation of alternative splicing; using the database to custom design microarrays; and for analysis of commercial microarrays aiming at measuring gene expression at the level of an exon. Small-scale projects would consist in investigating the protein-coding potential of new isoforms and the regulatory potential of non-protein-coding isoforms detected by the algorithm and in providing data for RT-PCR validation of these isoforms.

Since the data generated would be on a genome wide scale it was intended that this data should be shared with the scientific community in order to give it a maximum exposure for validation. With this objective in mind a web server was to be established in collaboration with other members of the group at the Institute of Molecular Medicine, with open access of the database to the public through a friendly interface.

ABSTRACT

Background

Mining current mRNA and EST databases for novel alternatively spliced isoforms is of paramount importance for shedding light on the way in which the maturation of RNA is used to regulate gene expression. Preliminary observations revealed a tendency for greater amounts of potentially non-protein-coding alternative transcripts in human genes than in orthologous genes from other organisms. However, many of these isoforms did not appear in recently published alternative splicing databases on account of constraints imposed in the selection of transcripts. This prompted us to develop a less constrained database with the aim of contributing to the identification of the full repertoire of splice variants in the transcriptome of different organisms. Given that mechanisms of control of gene expression involving non-protein-coding splice variants have been described in a variety of genes, this information may be crucial to deciphering more intricate layers of gene regulation in complex organisms brought about by alternative splicing.

Description

An algorithm was developed to cluster mRNA and EST BLAT alignments to annotated gene regions. Consensus splice sites were the main requirement imposed on the selection of transcripts. The method was applied to thirteen model organisms. The alternative splicing information generated has been incorporated into a database with clear graphical displays representing the splicing patterns and is available from the ExonMine website (http://www.imm.fm.ul.pt/exonmine). It incorporates information on constitutive exons, poly-A signals, open reading frames and translation, expression specificity of any exon or splicing pattern relative to biological source of mRNA/EST, alternative splicing events and respective exon and junction sequences for microarray probe design. The ExonMine interface also provides several tools to support laboratory validation of splicing patterns.

Conclusions

ExonMine detects a higher percentage of spliced genes and isoforms than currently available alternative splicing databases. The analysis reveals a marked increase, in complex organisms, of splice variants with either retained introns or incorporating novel exons with no apparent protein-coding potential. About 18% of unannotated exons detected in ExonMine were found expressed in primary human cells using tiling arrays. Validation of some of these results for the U2AF family of splicing factors was successfully performed in collaboration with members of the lab revealing primate specific transcripts and an alternatively spliced transcript carrying a microRNA. The database was also successfully used for genome wide analysis of sequence elements involved in the regulation of alternative splicing and for custom alternative splicing microarray design. Matching of ExonMine data to a commercial exon microarray platform covering the majority of human exons was also performed and will assist in large-scale analysis of alternative splicing data. The algorithm developed also provides for easy updatability, taking only 48 hours to generate data for the whole human genome and far less time for less complex organisms. In conclusion, ExonMine represents a new useful resource for future research on alternative splicing and gene regulation.

1. INTRODUCTION

The overwhelming amount of biological sequence information now available requires the creation of automated procedures for retrieving useful information from databases containing vast amounts of data. For the human genome alone over 8 million transcribed sequences have been deposited in GenBank [Benson *et al.* 2007], however only about 20,000 human genes have so far been identified. The fact that the nematode worm, *Caenorhabditis elegans,* measuring just one millimetre, contains about the same amount of genes as the human was initially perplexing [Hodgkin 2001]. However, it is becoming clear that, in more complex organisms, mechanism for generating a variety of transcripts from a single gene have evolved through alternative transcription initiation, alternative splicing and alternative polyadenylation. It has been estimated that, in the human genome, more than 73% of genes are alternatively spliced [Johnson *et al.* 2003; Kan *et al*. 2001]. Alternative splicing is therefore thought to play a dominant role in expanding the proteomic complexity encoded in a limited number of genes. Regulation of alternative splicing not only produces a dynamic proteome by modulating expression of functionally diverse protein isoforms but also regulates gene expression through on/off switches using premature termination codons and nonsense-mediated decay [Blencowe 2006; Graveley 2001; Modrek and Lee 2002; Black 2003]. Alternative splicing events occur in response to a wide range of stimuli which can result in changes at the level of the encoded protein, through the inclusion, exclusion or variability of particular protein domains [Pacheco *et al.* 2004; Schischmanoff *et al.* 1997]. Altered protein function can result in differential expression, activity, and localization of a gene product [Kwon *et al.* 1999]. Alternative splicing has been shown to play a vital role in a variety of physiological and developmental events, ranging from sexual differentiation [Lopez 1998; Elliott and Grellscheid 2006], to neuronal functions [Grabowski 2007; Schmucker and Flanagan 2004; Lipscombe 2005], neuronal development [Boutz *et al.* 2007b], immune functions [Lynch 2004], apoptosis in mammalian cells [Schwerk and Schulze-Osthoff 2005]; the cell cycle [Blencowe 2003]; and also in response to extracellular signals [Pelisch *et al*. 2005] and the environment [Pleiss *et al.* 2007; Srebrow *et al.* 2002]. Alternative splicing in untranslated regions i.e. that do not affect the reading frame, play an important role in regulation of gene expression [Hughes 2006]; they can, for example affect translation in the cytoplasm [Gebauer et al. 1998]. The study of alternative splicing is also important from a therapeutic point since it has been shown

to be associated with many diseased states [Wang and Cooper 2007] including cancer [Kalnina *et al*. 2005].

Although it has also become apparent that many of the alternatively spliced isoforms in a gene may not come to produce a protein, post transcriptional regulation through alternative splicing constitutes an important mechanism for the control of gene expression [Lewis *et al.* 2003] and lately much attention is being focused on non-protein-coding RNAs (ncRNAs) [Strausberg and Levy 2007; Prasanth and Spector 2007].

The work presented in this thesis consists in the creation of a program, which produces a database of alternative splicing patterns in genes from information deposited in GenBank. The program is applicable genome wide and its speed allows for easy updating of the database. The database is user friendly and informative and has been successfully applied to a variety of problems involving alternative splicing, including: microarray design, detection of control sequences, detection of microRNAs and cross-species comparison of alternative splicing. In addition to its application a statistical analysis of the output data revealed, particularly for higher organisms, a striking number of alternatively spliced exons which do not contain open reading frames in frame with known exons and which may play an important part in complex layers of gene regulation in higher organisms.

1.1. ALTERNATIVE SPLICING

The majority of genes in metazoans are composed of exons separated by intervening sequences referred to as introns. Figure 1.1 represents an overview of the steps which take place in the process of converting the information encoded in a gene into a protein. Splicing is normally thought to occur exclusively in the nucleus; a spliced messenger RNA is then transported to the cytoplasm where translation is expected to occur. A gene is encoded in a segment of DNA by combinations of four nucleotides two purines, adenosine (A) and guanine (G), and two pyrimidines, cytosine (C) and thymine (T). The gene is expressed via transcription into a pre-messenger RNA transcript (pre-mRNA), in which only the nucleotide base T is replaced by the nucleotide base uridine (U). In order to be translated into protein this RNA message must undergo a series of processing reactions, which include the removal of introns and joining of exons; and protection of the 5' and 3' ends of the RNA with a Cap structure and polyadenylation respectively.

The splicing of exons refers to the biochemical reactions which take place on a pre-mRNA molecule and result in the excision of an intervening sequence (intron) and concomitant splicing of the bordering exons [Sanford and Caceres 2004]. Splicing occurs in defined cellular loci [Moen *et al.* 1995; Misteli 2000]. Alternative splicing refers to this same mechanism taking place in the presence of additional splicing factors, which influence the choice of splice sites thereby producing alternative combinations of spliced exons. These combinations of exons normally occur not by shuffling but in a linear arrangement and can be systematized into seven basic types (Figure 1.2). One or more exons may be skipped (Figure 1.2A); exons, which can be skipped, are referred to as cassette exons. Cassette exons can be mutually exclusive (Figure 1.2B), i.e. never occur together in the same isoform. Exons may also be extended at either end through the use of alternative 5' splice sites (Figure 1.2C) or alternative 3' splice sites (Figure 1.2D). Strictly speaking, alternative first exons (Figure 1.2 E) are primarily an issue of the use of alternative transcription initiation sites (or alternative promoters) [Tsuritani *et al.* 2007], however, transcription factors can affect alternative splicing through promoter identity and occupation involving the recruitment of factors with dual functions in transcription and splicing and the control of RNA polymerase II elongation [Kornblihtt 2005; Pagani *et al.* 2003; Nogués *et al.* 2002; Kadener *et al.* 2002] furthermore, promoter specific polar effects on alternative splicing

have been observed [Fededa *et al.* 2005]. Alternative terminal exons will be associated with alternative polyadenylation signals, although alternative polyadenylation signals may in fact occur on the same terminal exon. And finally an intron can be retained in the final transcript (Figure 1.2 F). A transcript may however be subject to highly complex superimposed patterns of alternative splicing as is shown in (Figure 1.2H). Exons, which are always included in the final mRNA, are referred to as 'constitutive exons'.

Trans-splicing [Horiuchi and Aigaki 2006] (splicing between distinct pre-mRNAs) commonly occurs in clusters of genes transcribed polycistronically in unicellular organisms and in *Caenorhabditis elegans,* and has been observed in two *Drosophila melanogaster* genes; in mammals the physiological significance of the rare cases of *trans*-splicing reported *in vitro* is unclear.

Figure 1.1. The diagram represents a gene encoded in a segment of DNA, composed of 3 exons separated by two intervening sequences, or introns. The gene is expressed via transcription into a pre-messenger RNA transcript (pre-mRNA). In order to be translated into protein this pre-mRNA must undergo a series of processing reactions, which include the removal of introns and concomitant splicing of exons; and protection of the 5' and 3' ends of the RNA with a Cap structure and polyadenylation respectively.

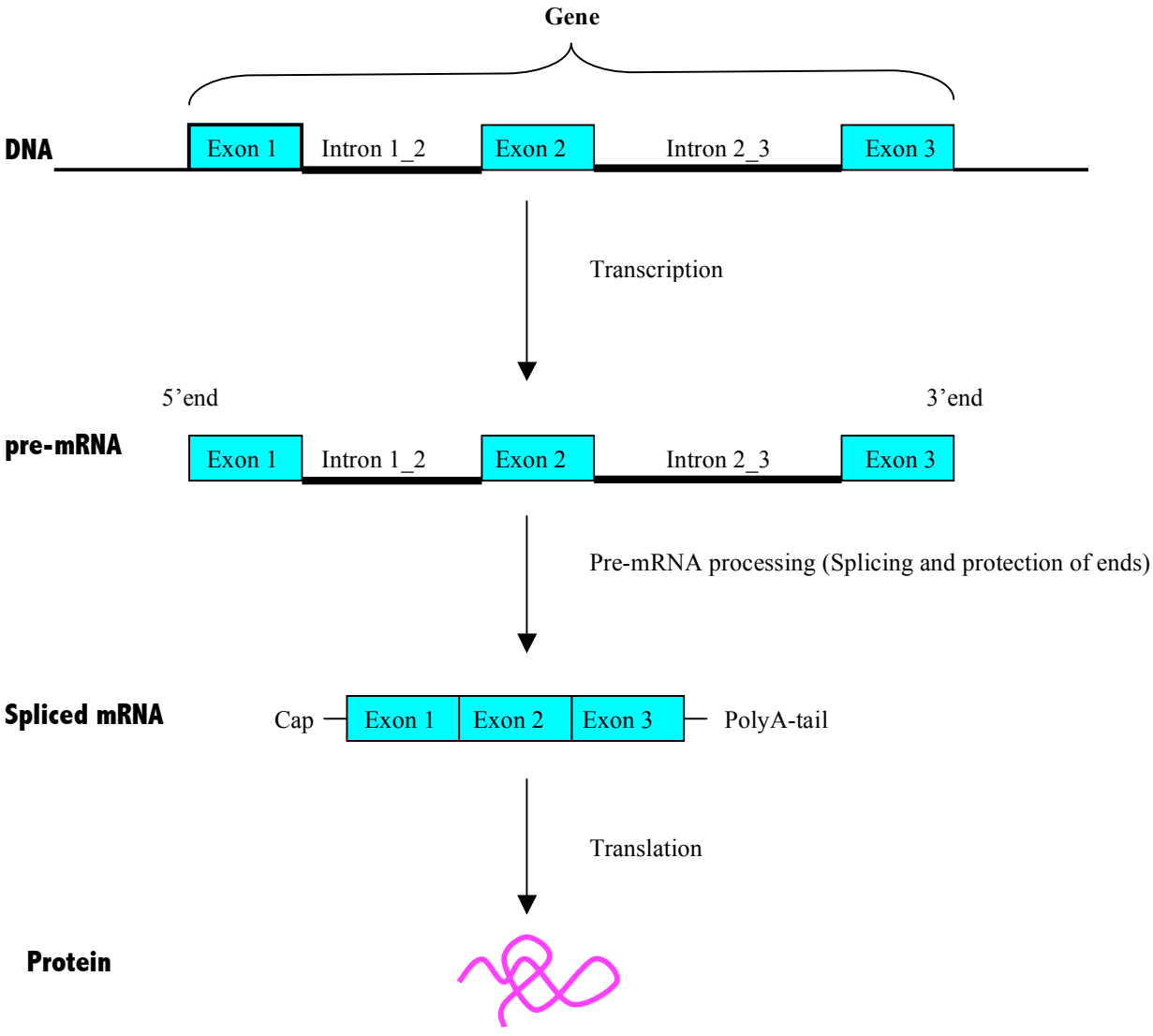
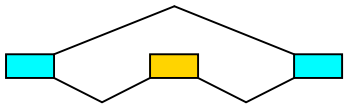
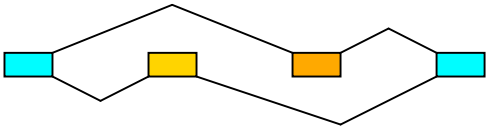


Figure 1.2. Alternative splicing can be systematised into six basic categories A-G. However, in reality, not only can there be several distinct alternatively spliced regions in a gene, but within a region there can be superimposed patterns of alternative splicing. H. An example of complex alternative splicing occurring in the 4.1R gene [Schischmanoff *et al.* 1997] where in addition to a basic alternative 3' splice site and cassette exons, multiple superimposed patterns of alternative splicing are observed in three distinct regions.

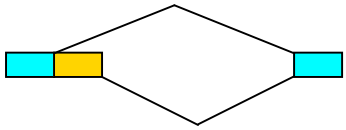
A – Cassette Exon



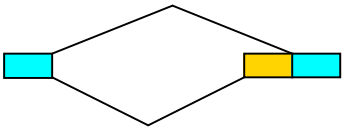
B – Mutual exclusion of cassette exons



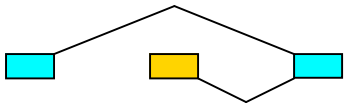
C – Alternative 5' splice site



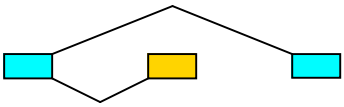
D – Alternative 3' splice site



E – Alternative First Exons

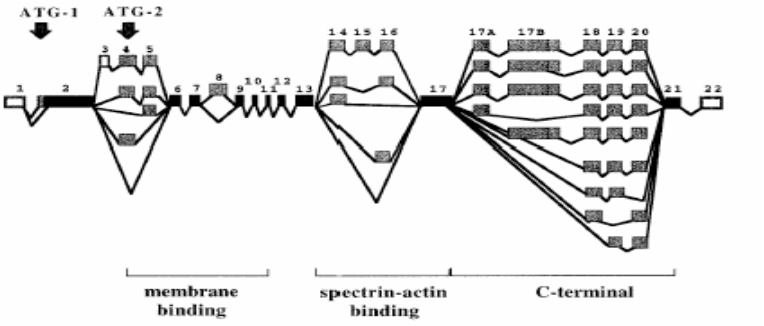


F – Alternative Terminal Exons



G – Intron Retention

H – Complex alternative splicing



1.1.1. THE SPLICEOSOME AND THE SPLICING REACTION

The spliceosome and main consensus sequences

The basic splicing reaction takes place on a macromolecular assembly called the spliceosome, which may be the most complex machine in a cell [Nilsen 2003]. The spliceosome is composed of five small nuclear ribonucleoprotein particles (snRNPs) U1, U2, U5, U4 and U6, and an array of non-snRNP splicing factors and splicing associated factors [Chiara *et al.* 1996; Jurica and Moore 2003; Zhou et al. 2002]. The ribonucleoprotein particles each contain a different small nuclear RNA (snRNA), a common RNP core structure composed of seven Sm proteins (E,F,G,D1,D2,D3, and B/B') [Kambach *et al.* 1999b; Raker *et al.* 1996; Stark *et al.* 2001; Urlaub *et al.* 2001; Walke *et al.* 2001] and additional proteins particular to a given snRNP particle [Will and Lührmann 2006; Burge *et al.* 1999; Kambach *et al.* 1999a; Will and Lührmann 2001]. The sequence of the spliceosomal snRNAs, in particular those regions engaging in base-pairing interactions, and to some extent the secondary structures, are highly conserved evolutionarily [Guthrie and Patterson 1988; Reddy and Busch 1988, Kiss 2004; Yu *et al.* 1999]. In addition to the snRNPs it is estimated that there may be as many as 300 splicing factors and splicing associated factors [Rappsilber *et al.* 2002; Jurica and Moore 2003] involved in the splicing reaction, many of which may be tissue specific or may participate in other cellular processes.

The snRNPs are recruited concomitantly to the pre-mRNA proceeding through a series of stabilization and destabilization reactions [Vilardell and Valcarcel 2007], remodelling multiple RNA-RNA (intra- and inter-snRNA , as well as snRNA – pre-mRNA) [Nilsen 1998], RNA-protein [Shen and Green 2007] and protein-protein interactions [Corsini *et al.* 2007], to generate a catalytically active RNP structure. This process of activation has been proposed to take place through an allosteric cascade of interactions [Brow 2002].

The operation of the spliceosome relies on specific sequences encoded in the pre-mRNA at the 5'splice site (5'ss), the 3' splice site (3'ss), the branch point sequence (BPS) which carries a conserver A nucleotide, and the polypirimidine tract (Py), all of which direct the initial assembly of the spliceosome. The nature of these sequences can influence the dynamics of the splicing reaction. The sequence of the Py-tract is highly variable [Singh et

al 1995], strong Py-tract being rich in uridines. The major class of introns, called U2-type introns have a branch signal consensus **yUnAy** [Gao *et al.* 2008] which is quite degenerate, where only the U and A seem to be consistently present, and is normally located at about 40 nucleotides from the 3' splice site but can also be found at a highly variable position relative to the 3'splice site [Gooding *et al.* 2006]; and have conserved 5' and 3' splice site sequences |**GU**R and Y**AG**| in more than 99.9% of cases, or very rarely |**GC** and Y**AG**|, where | represents the splice junction (R=A or G, and Y=U or C). A minor class of introns, called U12-type introns [Tarn and Steitz 1997], carry a very strongly conserved branch point sequence (BPS) UCCUUA**A**C (branch point underlined), equally strongly conserved 5' and 3' splice site sequences, either |**AU**AUCCUUU and Y**AC**| or more commonly |**GU**AUCCUUU and Y**AG**|, and typically lack a polypyrimidine tract. U12-type introns are spliced by a compositionally distinct spliceosome referred to as the minor spliceosome [Patel and Steitz 2003; Will and Luhrmann 2005], which has an early evolutionary origin [Russell *et al.* 2006].

The proportion of all nuclear introns that are spliced by the U12-type spliceosome is estimated to be in the order of one in a thousand in vertebrates and lower in other taxa [Patel et al 2002]. Both spliceosomes form similar RNA-RNA, RNA-protein interactions and exhibit similar protein compositions [Schneider *et al.* 2002; Will *et al.* 1999]. The U12-dependent spliceosome is formed from the U11, U12, U5, and U4atac/U6atac snRNPs, the U5 snRNA being common to both spliceosomes [Hall and Padgett 1996; Tarn and Steitz 1996a, 1996b]. Additional non-snRNP splicing factors are common to both spliceosomes [Shen and Green 2007]. The physiological significance of the existence of the two distinct spliceosomes is not clear and since genes containing U12-type introns also contain U2-type introns, it appears that the U2-type and U12-type spliceosomes can be active in the same region of the nucleus, however, splicing of transcripts with U12-type introns has been observed in anucleate platelets [Meshorer and Misteli 2005; Denis *et al.* 2005] and recent work indicates that the major and minor spliceosomes may belong to two spatially segregated pathways acting on distinct sets of genes; the minor spliceosome being preferentially active in the cytoplasm [Cáceres and Misteli 2007], rather than in the nucleus, and on genes involved in cell proliferation [König *et al.* 2007], providing a means for nucleus-independent key regulatory events in the control of gene expression. Indeed,

many splicing components cycle between the nucleus and the cytoplasm as part of their normal life cycle [Cáceres *et al.* 1998; Yong *et al.* 2004; Gama-Carvalho *et al.* 2006].

The splicing reaction

All nuclear pre-mRNA introns are removed by an identical two-step mechanism involving two consecutive trans-esterification reactions (Figure 1.3). In the first step, the 2'- OH group of the branch point adenosine sugar carries out a nucleophilic attack on the phosphate at the 5' splice site (Figure 1.3A), which results in the cleavage of the pre-mRNA at the 5' splice site, leaving a free 3'- hydroxyl group, and the ligation of the 5' end of the intron to the branch point adenosine. The second step consists in another nucleophilic attack, this time by the free 3'- hydroxyl group of the 5' exon on the phosphate at the 3'splice site (Figure 1.3B), leading to the ligation of the 5' and 3' exons, and concomitant release of the intron in the form of a lariat structure. It is notable that the splicing reaction, as we know it, can never take place on DNA, it can only take place on RNA because DNA does not possess 2'- OH groups on the sugar backbone.

Spliceosome assembly model

The initial identification of splice sites may take place across an intron or an exon and the choice of recognition mechanism is influenced by the size of both exons and introns [Sterner *et al.* 1996]. In mammals, where introns can be several orders of magnitude longer than exons, recognition of splice sites appears to take place across an exon, where these, being much shorter than introns, serve as a unit of recognition in a process called "exon definition" [Berget 1995; Reed 1996; Hoffman and Grabowski 1992; Robberson *et al.* 1990], even though the splice sites flanking an exon will not be spliced by the same spliceosome. The process of exon definition appears to involve interaction with the carboxy terminal domain of RNA polymerase II [Zeng and Berget 2000]. Although some observations suggest that the spliceosome exists as a large, preformed entity that engages the pre-mRNA as such [Stevens *et al.* 2002; Das *et al.* 2000; Maroney *et al.* 2000; Sergeant *et al.* 2007], studies performed on a variety of systems also show that the major spliceosome assembles in a stepwise manner by the ordered interaction of the U1 and U2 snRNPs and the U4/U6.U5 tri-snRNP with the pre-mRNA (Figure 1.4). This stepwise

model has recently been confirmed by studies of co-transcriptional spliceosome assembly coupled to 5' end processing and transcription [Görnemann *et al.* 2005]. The classic representation of spliceosome assembly is pictured to proceed across an intron [Burge *et al.* 1999; Will and Lührmann 2006; Brow 2002; Nilsen 2002; Staley and Guthrie 1998], from the set of five small nuclear ribonucleoproteins (snRNPs) and numerous accessory proteins, through five stages (E, A, B, B* and C) which correspond to the most stable assembly intermediates of the human spliceosome that have been detected biochemically. Spliceosomal proteins contribute to splice site recognition indirectly by stabilizing snRNA-pre-RNA base-pairing interactions; or directly by recognition of the 5' and 3' splice sites. Recent work has revealed a link between exon definition and an intron-defined spliceosome [Sharma *et al.* 2008].

In the early E, or commitment Complex (Figure 1.4), the U1 snRNP is recruited to the 5' splice site, with the U1 snRNA base pairing with the 5'splice site, an interaction which is stabilised by U1 specific proteins (U1-70K and U1-C) as well as members of the SR protein family [Will and Lührmann 1997; Rossi *et al.* 1996; Du and Rosbach 2001, 2002]. At this stage, the 3' splice site elements are bound cooperatively by a special set of proteins [Abovich and Rosbash 1997; Berglund *et al.* 1997, 1998a, 1998b]: SF1 binds at the branch-point; the 65-kDa subunit of the dimeric U2 auxiliary factor (U2AF) binds to the polypyrimidine tract and contacts the BPS via its RS domain [Gaur *et al.* 1995; Valcárcel *et al.* 1996; Kent *et al.* 2003; Shen and Green 2004; Sickmier *et al.* 2006]; and, in at least some cases, the 35-kDa subunit of U2AF binds directly to the AG dinucleotide at the intron/exon junction [Das *et al.* 2000; Merendino *et al.* 1999; Wu *et al.* 1999]. U2AF independent formation of this early complex has also been observed in mammals [Kent *et al*. 2005]. It is thought that, at this stage, a molecular bridge involving SR proteins is formed between the U1 snRNP at the 5' splice site and SF1 and U2AF bound to the BPS and 3' splice site [Reed 1996; Boukis *et al.* 2004]. This complex is joined by the U2 snRNP to form the A complex [Dönmez *et al.* 2004, 2007].

In the A Complex (Figure 1.4), or pre-spliceosome, the U2 snRNP apparently displaces SF1 and stably associates with the branch point sequence (BPS), with the U2 snRNA base pairing with the branch point sequence; an interaction which requires U2AF and is facilitated by other heteromeric splicing factors (SF3a and SF3b) which contact the pre-

mRNA in the vicinity of the branch site [Reed 1996]. The U2 snRNA forms an RNA duplex with a bulged adenosine [Berglund *et al.* 2001] at the branch point (Figure 1.5) positioning this adenosine to carry out the nucleophilic attack in the first catalytic step of splicing [Query *et al.* 1994]. The bulged branch point adenosine is directly contacted by a subunit of SF3b, namely p14/SF3b14a, which appears to remain associated at least through the first step of splicing [Query *et al.* 1996; Will *et al.* 2001]. This subunit is thought, in higher eukaryotes, to play a central role in the recognition and/or selection of the BPS and is likely present at the catalytic core of the spliceosome. However, during enhancer-dependent spliceosome assembly, RS domains of SR proteins also directly contact the BPS at the time of A complex assembly [Shen *et al.* 2004]. The nature of the molecular bridge formed between U1 and U2 snRNPs, bringing together the 5' and 3'splice sites, has been recently studied [Behzadnia *et al.* 2007] and it appears that it involves base pairing interactions between U1 snRNA and U2 snRNA. Non- snRNP DEAD-box protein Prp5 has also been implicated in bridging the U1 and U2 snRNPs in the major spliceosome [Xu *et al.* 2004]. This complex is joined by the U4/U5/U6 tri-snRNP to form the B complex.

The B Complex [Deckert *et al.* 2006] (Figure 1.4) carries the pre-assembled U4/U6.U5 tri-snRNP, which has displaced proteins U2AF65 and U2AF35. This complex undergoes a complicated rearrangement of RNA-RNA and RNA-protein interactions leading to the destabilization of the U1 and U4 snRNPs to form the catalytically activated B* Complex.

Finally the conversion into the catalytic C Complex takes place, in which the U1 snRNP interaction at the 5' splice site is replaced with the U6 snRNP and the U1 and U4 snRNPs are lost from the complex. In this complex the first of the two catalytic steps of splicing occurs (Figure 1.4), with the U5 protein Prp8 being implicated in properly positioning the 5' splice site within the spliceosome for this first reaction. After the first trans-esterification, U5 also contacts exon nucleotides just downstream from the 3' splice site and is thought not only to tether the 5' exon to the spliceosome, but also to align both exons for the second catalytic step, interactions that involve several other splicing factors [Aronova *et al.* 2007]. U6 snRNA is also involved in genetic interactions between the 5' and 3' splice site consensus sequences during this step [Collins and Guthrie 2001]. After the second catalytic step the spliceosome is thought to disassemble and the snRNPs are thought to take part in additional rounds of splicing.

This stepwise assembly model provides a clear picture of when individual snRNPs are stably recruited to the pre-mRNA, and is not necessarily contradictory with the existence of a preassembled spliceosome, as the intermediate stages of assembly which have been detected are likely to reflect different stabilization/destabilization states of the spliceosomal snRNPs and splicing factors with the pre-mRNA.

Many additional non-snRNP spliceosomal proteins are involved in major structural rearrangements in the RNA-RNA and RNA-protein and protein-protein networks prior and subsequent to the formation of a catalytically active spliceosome [Burge *et al.* 1999; Will and Lührmann 2006; Tycowski *et al.* 2006; Bartels *et al.* 2002, 2003; Kuhn *et al.* 2002; Tarn *et al.* 1993a,b; Makarova *et al.* 2004; Chen *et al.* 2002; Chan *et al.* 2003; Ajuh *et al.* 2000; Horowitz *et al.* 2002; Collins and Guthrie 1999; Graveley 2000]. At least some 300 proteins have been shown to be implicated in splicing [Jurica and Moore 2003; Chen *et al.* 2007; Barbosa-Morais *et al.* 2006; Gabut *et al.* 2008; Will and Luhrmann 2006]. These include members of the Serine/Arginine (SR) protein family, heterogeneous nuclear ribonucleoproteins (hnRNPs), and members of the DEXH/D-box family of RNA unwindases/RNPases [Teigelkamp *et al.* 1994; Staley and Guthrie 1998; Schwer and Meszaros 2000; Schwer 2001] and many others, but the precise spliceosomal targets of these proteins remain for the most part unclear. In addition, protein modifications, by phosphorylation, are thought to contribute to spliceosome dynamics at virtually every step of splicing, in particular the phosphorylation of SR proteins [Wang *et al.* 1999; Wang *et al.* 1998; Misteli and Spector 1999; Bollen and Beullens 2002; Lai *et al.* 2003; Soret and Tazi 2003; Tazi *et al.* 1992].

Figure 1.3. Details of the two trans-esterification reactions which result in the removal of the intron and the concomitant splicing of the flanking exons. **A**. In the first reaction the 2'-OH group of the Branch Point conserved adenosine (A) sugar displaces the 3'- OH group at the 5' splice site. **B**. In the second transesterification reaction the free 3'- OH group of the 5' exon displaces the 3'- OH group at the 3' splice site. (B stands for any nucleotide base A, C, U or G).

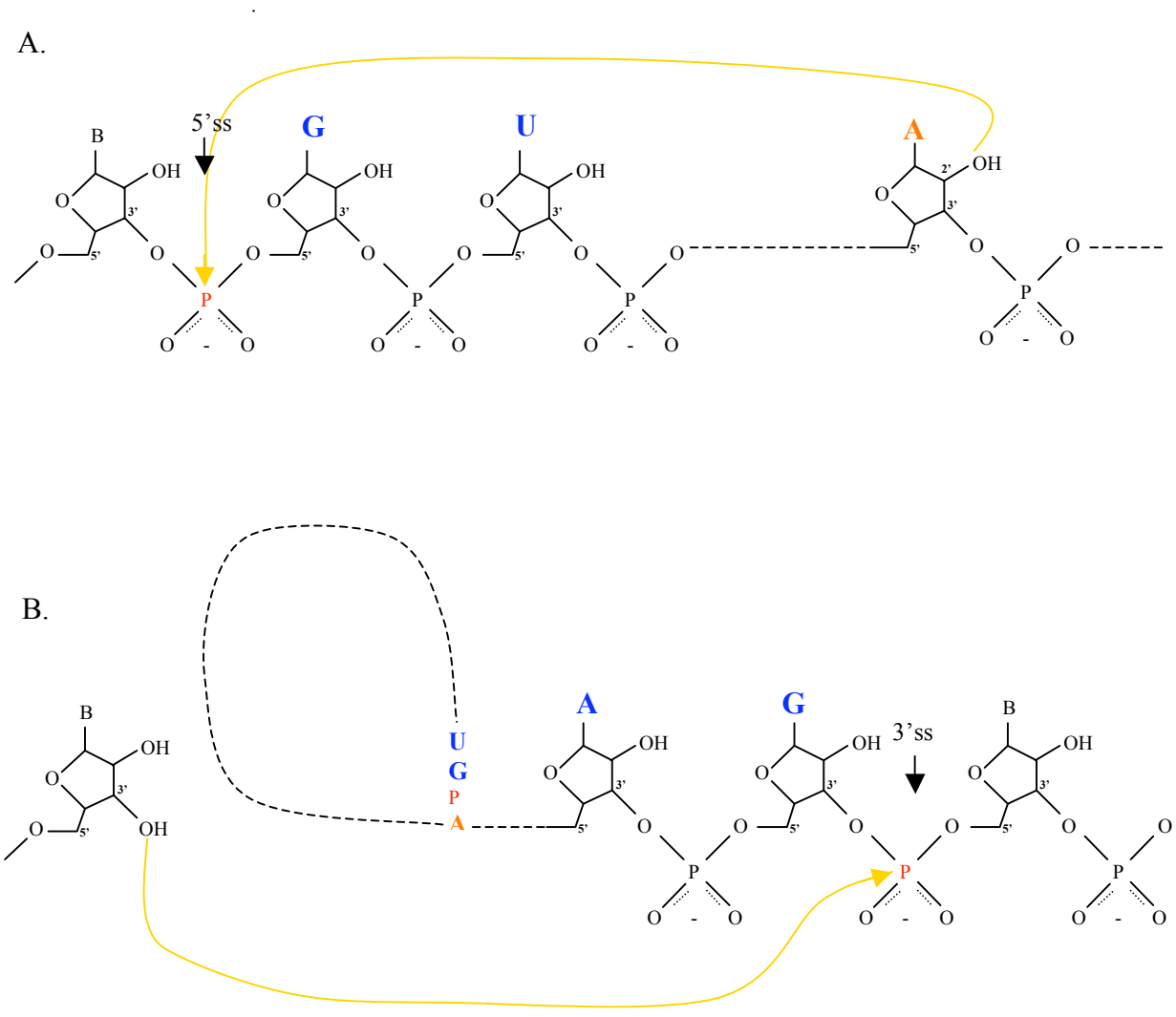


Figure 1.4. A. Layout of the main splicing signals in an intron and assembly of the main components of the spliceosome through the E, A, B, B* and C complexes. The 5' splice site (5'ss) is at the 5' end of the intron. The 3' splice site (3'ss) is at the 3' end of the intron. Branch Point Sequence (BPS) with conserved adenosine (**A**) nucleotide is expected to be found 10-20 nt or 18-40nt upstream of the 3'ss in U12 type and U2- type introns respectively. The polypyrimidine tract (Py) in U2-type introns lies between the BPS and the 3'ss but its position and composition are highly variable.

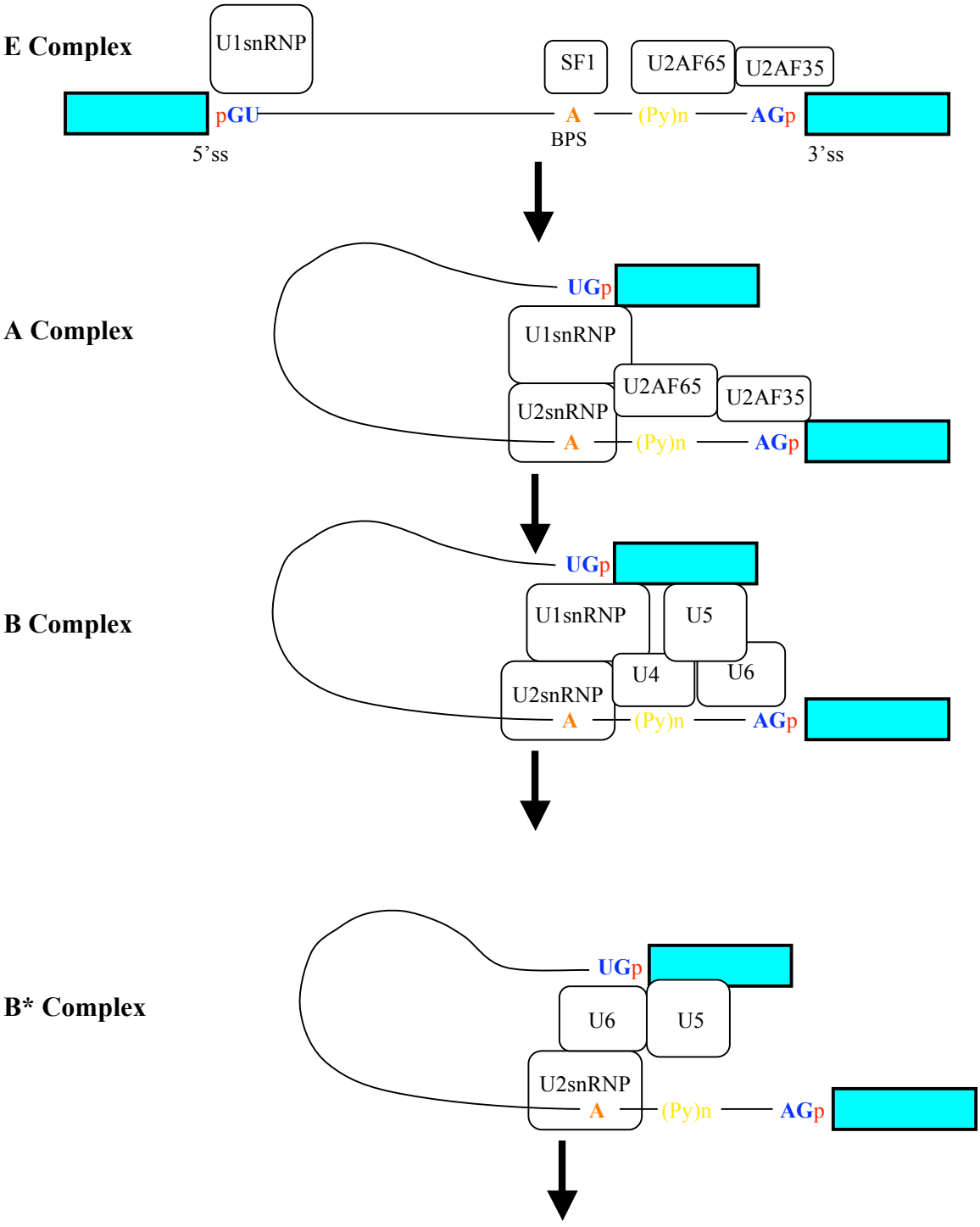


Figure 1.4 (cont.) Schematic representation of the two trans-esterification reactions which take place on the catalytic C complex (spliceosome components omitted for clarity)

C Complex reactions

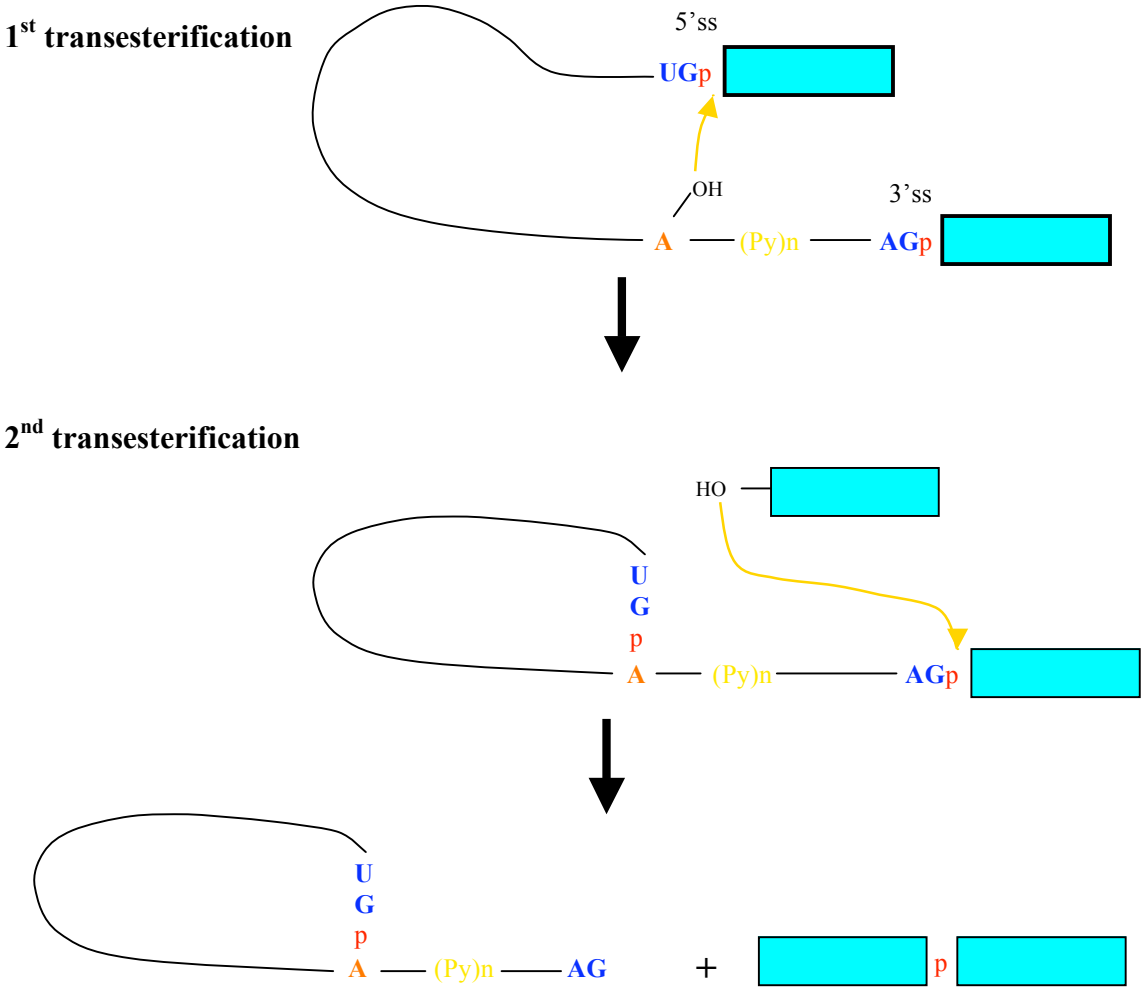
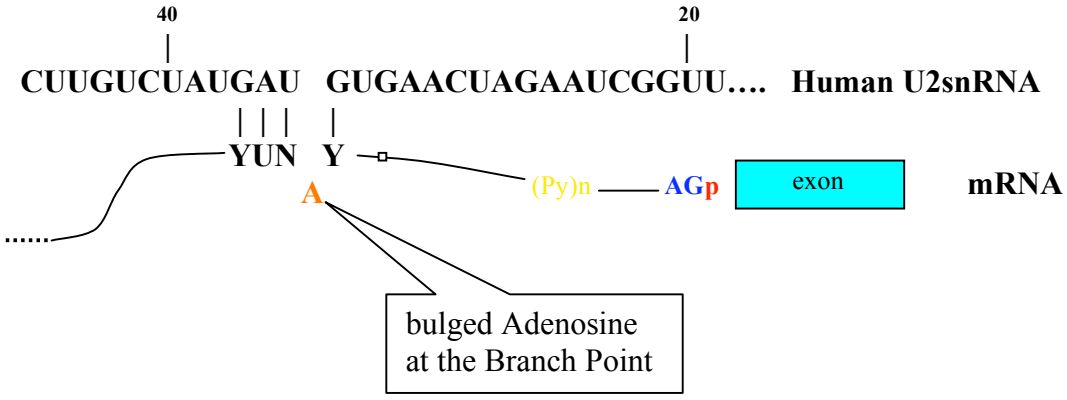


Figure 1.5. Base pairing (indicated by vertical lines) between a messenger RNA (mRNA) branch point sequence and the U2snRNA, producing the bulged Adenosine (A), which is involved in the first trans-esterification reaction in splicing.



RNA catalysis

Despite its absolute dependence on numerous proteins for catalytic activation, and the fact that it is predominantly made up of protein, the spliceosome appears to be an RNA enzyme [Nilsen 2000] (or ribozyme) with an active site or sites composed of RNA [Collins and Guthrie 2000; Villa *et al.* 2002; Valadkhan 2005]. Several intermolecular structures formed by the pre-mRNA and the U2, U5 and U6 snRNAs exhibit similarities with intramolecular structures formed by self-splicing-group II introns supporting the idea that pre-mRNA splicing is catalysed by RNA [Madhani and Guthrie 1992; Weiner 1993]. The first step of splicing has also been shown to be mediated by a metal ion [Sontheimer 2001], consistent with RNA-based catalysis [Sontheimer *et al.* 1997, Gordon *et al.* 2000; Yean *et al.* 2000]. Because splicing appears to by catalysed by RNA, a major job of the set of proteins present in the C complex will likely be to generate and/or maintain the RNA structures required for catalysis. In addition to facilitating the formation of critical RNA-RNA interactions, spliceosomal proteins are thought to be the driving force behind the numerous structural rearrangements required for the catalytic activity of this dynamic RNP machine. Splicing is indeed a very dynamic process, and some functionally important proteins are thought to act only transiently with the spliceosome at defined stages.

Integration of splicing with other cellular processes

Pre-mRNA splicing has been shown to be linked spatially and temporally to several other cellular processes [Hirose and Manley 2000; Bird *et al*. 2004, Bird *et al*. 2005; Aguilera 2005; Bentley 2005; Moore *et al.* 2006]. Changes in chromatin structure have been shown to affect splicing [Nogués *et al.* 2002; Sacco-Bubulya and Spector 2002; Batsché *et al.* 2006] perhaps because hyperacetylation of core histones facilitates the passage of the transcribing polymerase. Promoter structure [Cramer *et al.* 1997; Kornblihtt 2005], and recruitment of transcription factors [Kadener *et al.* 2001; Nogués *et al.* 2002] and coactivators [Auboeuf *et al.* 2004] can greatly affect alternative splicing. Interactions between polymerase II carboxy terminal domain (CTD) and splicing factors [de la Mata and Kornblihtt 2006], the rate of RNA polymerase elongation [Roberts *et al.* 1998; de la Mata 2003; Howe *et al.* 2003; Fededa *et al.* 2005], transcription [Kornblihtt *et al.* 2004;

Das *et al.* 2006; Hicks *et al.* 2006; Lynch 2006], RNA editing [Laurencikiene *et al.* 2006], capping [Görnemann *et al.* 2005], polyadenylation [Glover-Cutter *et al.* 2008] and mRNA export [Zhao *et al.* 2004, Reed and Hurt 2002] have also been shown to affect alternative splicing. Furthermore, on completion of the pre-mRNA splicing reaction, the spliceosome deposits a set of protein factors 20-24 nucleotides upstream of the exon-exon junction referred to as the exon junction complex (EJC) [Kataoka et al 2000; Le Hir *et al.* 2000a, 2000b]. The presence of these complexes will shape subsequent events in gene expression. These exon junction complexes have been shown to enhance translation [Nott *et al.* 2004] of an mRNA, although they are removed by the translation machinery in the pioneer round of translation in the cytoplasm (i.e. the first time the mRNA is used as a template for translation into protein). The EJC is also implicated in post-splicing events such as mRNA nuclear export and nonsense-mediated mRNA decay (NMD) [Le Hir *et al.* 2001a, 2001b; Maquat 2004] since they will remain on the mRNA if there is a premature stop codon, providing a signal to the NMD machinery, which will target mRNA for degradation. However, in the event of a premature stop codon, NMD can be inhibited by binding of poly(A)-binding protein [Silva *et al*. 2008]. The interplay of so many factors involved in pre-mRNA processing requires us to visualise the integration of these interactions over extended lengths of RNA sequences during the process of splicing [Cramer *et al.* 2001; Rigo *et al.* 2008; Bentley 2002; Maniatis and Reed 2002; Reed 2003; Mc Cracken *et al.* 1997].

Given these facts it is not surprising that many proteins related to other cellular processes have been detected in spliceosomal complexes [Chen *et al.* 2007]. For example, drosophila Sxl splicing regulator binding to one of its mature mRNA targets can also block translation in the cytoplasm [Bashaw and Baker 1997; Kelley *et al.* 1997; Gebauer *et al.* 1998].

However, as some of these proteins have also been documented to function in splicing, they may be multifunctional rather than being responsible for linking splicing to other gene expression machines. Candidates for mediating these types of interactions are proteins containing an arginine-serine (RS or SR) domain, which is present in many proteins important for splicing in vertebrates. RS domains are involved in both protein-protein and protein-RNA interactions [Manley and Tacke 1996; Tacke and Manley 1999; Sanford *et al.* 2005; Das *et al.* 2007]. Proteins containing RS domains have been shown to be critical in both early and late stages of spliceosome assembly in vertebrate cells [Manley and Tacke

1996; Valcárcel and Green 1996; Graveley and Maniatis 1998; Graveley *et al.* 1999; Hertel and Graveley 2005], and the heterogeneous distribution of such proteins in the vertebrate nucleus suggests an intimate relationship between splicing and nuclear structure [Xing *et al.* 1993]. SR proteins have also been linked to signal transduction pathways in which concerted regulation of alternative splicing and translation is observed in response to extracellular signals [Blaustein *et al.* 2005; Graveley 2005]. An example is the SF2/ASF SR protein, which shuttles between the nucleus and the cytoplasm where it is expected to perform distinct functions [Sanford *et al.* 2004].

1.1.2. SPLICING REGULATION AND SPLICING REGULATORY ELEMENTS

Changes in splice site choice are responsible for alternative splicing. It is thought that such changes occur mainly during the early stages of spliceosome assembly. Splice site choice can be determined at the stage of splice site recognition through exon or intron definition, but commitment to alternative splice site choice pairing may occur after initial splice site recognition and E complex formation [Lim and Hertel 2004]. There is also evidence of regulation of 3' splice site choice after the first catalytic step [Lallena *et al.* 2002].

Additional *cis*-regulatory sequence elements

The core human splice site motifs, 3' splice site, 5' splice site, branch point and polypyrimidine tract are too short and degenerate to reliably identify the splice junctions even in short introns [Lim and Burge 2001]. The higher level of degeneracy of the splice sites and the branch site in the more common U2- type introns, in relation to the U12 – type introns, suggests that the former may depend more critically on recognition of additional sequences to specify a splice junction [Berget 1995; Black 1995]. Branch point plasticity has also been shown to contribute to splicing regulation [Kol *et al.* 2005]. However, whether for U2-type introns or U12-type introns, it is by now fairly well established that, in metazoans, for a splice site to be reliably identified, additional mRNA *cis*-sequence elements are necessary [Black 2003]. These additional *cis*-acting pre-mRNA elements can be present on exons or introns. Elements within an exon, which promote or inhibit inclusion of that exon, are conventionally classified as exonic splicing enhancers (ESEs) or exonic splicing silencers (ESSs) [Blencowe 2000; Smith and Valcárcel 2000; Black 2003; Wang and Burge 2008]; many large scale and small scale computational and experimental approaches have been dedicated to identifying and compiling ESE and ESS [Chasin 2007; Tian and Kole 1995; Tian and Kole 2001; Liu *et al.* 1998, 2000; Fairbrother *et al.* 2002; Zhang and Chasin 2004; Cartegni *et al.* 2002; Wang *et al.* 2004; Wang *et al.* 2005; Goren *et al.* 2006]. Similarly, elements within an intron, which enhance or inhibit usage of the adjacent splice sites or exons, are conventionally classified as intronic splicing enhancers (ISEs) or intronic splicing silencers (ISSs), and a few have been identified [Ladd and Cooper 2002; McCullough and Berget 1997, 2000; Yeo *et al.* 2004; Hui *et al.* 2005; Hung *et al.* 2007; Brudno *et al.* 2001; Jin *et al.* 2003; Minovitsky *et al.* 2005; Nakahata and

Kawamoto 2005; Underwood *et al.* 2005; Jensen *et al.* 2000; Ule *et al.* 2003; Matlin *et al.* 2005; Wagner *et al.* 2005; Singh *et al.* 2006; Kashima *et al*. 2007].

These *cis*-acting pre-mRNA elements are typically short and diverse in sequence, most seem to bind regulatory proteins, but some can form RNA secondary structures that affect splice site recognition [Libri *et al.* 1991; Liu *et al*. 1995; Jacquenet *et al.* 2001; Buratti 2004; Hiller *et al*. 2007]. These elements modulate splicing of both constitutively and alternatively spliced exons [Garg 2007]; in fact it is thought that most, if not all exons, contain exonic splicing enhancers (ESEs) that stimulate splicing at the adjacent splice sites [Schaal and Maniatis 1999; Fairbrother *et al.* 2002; Cartegni *et al.* 2003]. Exons without these elements are generally not recognized and tend to be excluded from the mature mRNA [Wang and Burge 2008]. Silencers seem to play a more predominant role in alternative splicing [Fairbrother *et al.* 2002; Wang *et al.* 2004; Zhang and Chasin 2004] causing intronic sequence around alternatively spliced exons to be more conserved than that around constitutive exons [Sorek and Ast 2003]. These regulatory elements normally function by recruiting *trans*-acting splicing factors [Matlin *et al.* 2005; Chasin 2007]. There can be several copies of a particular regulatory element which may function additively by increasing the number of factors recruited to it [Huh and Hynes 1994; McCullough and Berget 1997; Chou *et al.* 2000; Wang *et al.* 2004; Zhang and Chasin 2004; Dominguez and Allain 2006]. Different regulatory elements may also function cooperatively to regulate alternative splicing [Han *et al.* 2005; Modafferi and Black 1999]. The context in which these *cis*-acting elements are and the relative concentration of the *trans*-acting splicing regulators will determine the outcome of the splicing reaction [Mabon and Misteli 2005]. Certain motifs can even function as silencer or enhancer depending on their location [McCullough and Berget 1997; Chen *et al.* 1999; Hui *et al.* 2005; Ule *et al.* 2006; Goren *et al.* 2006]. Although the issue of pre-mRNA secondary structure playing a role in the recognition of regulatory elements has been debated [D'Souza and Schellenberg 2002], it has been found to be essential in several cases [Hiller *et al.* 2007; Jensen *et al.* 2000; Warf and Berglund 2007; Donahue *et al.* 2006].

Many exonic splicing enhancers function by recruiting SR proteins [Lynch and Maniatis 1996; Graveley 2000]. SR proteins contain RNA recognition motifs (RRM) that bind to the RNA elements; and RS domains that facilitate protein-protein interactions and assembly of the spliceosome [Graveley and Maniatis 1998]; in some cases RS domains can also interact

with the branch point sequence and the 5' splice site.

Another large group of proteins, identified by their association with unspliced mRNA precursors (also called heterogeneous nuclear RNA, hnRNA), are commonly referred to as hnRNP proteins. These are not a single family of related proteins, rather they are identitified by their association with RNA in the nucleus. hnRNP proteins contain RNA binding domains that can bind to exonic splicing silencers and thereby act as repressors. They may also contain inhibitory glycine-rich motifs [Pozzoli and Sironi 2005]. They have been shown to function through a wide variety of mechanisms: by blocking interactions between spliceosomal snRNPs [Izquierdo *et al.* 2005; Sharma *et al.* 2005]; by binding on either side of an exon and looping it out; or by displacing snRNP binding [Zhu *et al.* 2001; Nasim *et al.* 2002].

Examples of alternative splicing mechanisms

Many examples of regulation of the splicing reaction have been summarized in a number of reviews [Black 2003; Konarska and Query 2005; Matlin *et al.* 2005; Blencowe 2006; House and Lynch 2008].

The simplest known mechanism for altering splice site choice is by direct competition with essential splicing factors for RNA binding sites. An example of this mechanism is illustrated by the splicing regulator Sxl in drosophila which can promote alternative 3'splice site choice is one target pre-mRNA [Crowder 1999; Handa 1999], by competing directly with U2AF for its uridine-rich binding site, which coincides with the Py tract, at the 3'ss.

To determine the consequences of the presence of a splicing factor, the location of its binding sites has to be taken into account. The same drosophila splicing regulator Sxl, for example, can also promote intron retention when its uridine-rich binding site is also present near the 5'ss of another target pre-mRNA [Förch *et al.* 2000; Förch *et al.* 2001], directly competing with two other essential splicing factors, TIA-1 and U1snRNP; or it can promote exon skipping when its uridine-rich binding sites are present on flanking introns [Lallena *et al.* 2002].

Exonic splicing enhancers can also be composed of repeats [Tian and Maniatis 1992; Lynch and Maniatis 1996; Hertel *et al.* 1996] onto which proteins bind cooperatively through protein-protein interactions of RS domains and protein-RNA interactions of RNA-binding

domains [Hertel *et al*. 1997; Smith and Valcarcel 2000; Wang and Manley 1997] forming a stabilizing multiprotein enhancer complex which leads to spliceosome assembly and splicing of the exon [Zuo and Maniatis 1996; Graveley *et al*. 2001; Graveley 2000]

Exon definition, being an early event during splicing, can lead to commitment of the exon to splicing, therefore this step is critical for splicing regulation and specificity. One example of control of alternative splicing at this level is found in the definition of Fas gene exon 6 which is inhibited by polypyrimidine tract binding protein (PTB/hnRNP I) binding to an exonic splicing enhancer sequence which inhibits exon definition complex formation causing skipping of Fas exon 6 [Izquierdo *et al.* 2005]. PTB has also been found to inhibit the spliceosome assembly across introns (intron definition) in repressing splicing of the c-src N1 exon [Sharma *et al*. 2005]. These two examples illustrate context dependence of *cis*-RNA sequence regulatory elements, i.e. the presence of the same *trans*-acting splicing factors can result in different alternative splicing patterns and mechanisms of splicing on different substrates depending on the location of the binding site on the RNA.

1.1.3. ADDITIONAL SOURCES OF VARIATION

Alternative transcription initiation

Alternative first exons (or alternative transcription initiation) are primarily an issue of use of alternative promoters. It has been estimated that, in human, more than 20% of genes have functional alternative promoters [Cooper *et al.* 2006; Landry *et al.* 2003] providing distinct regulation for alternate isoforms of the same gene.

RNA polymerase II is a multisubunit enzyme that catalyzes the synthesis of mRNA from the DNA template. RNA polymerase II promoters are generally viewed as composed of a core region and an extended region. The core region of a promoter is defined as the minimal stretch of contiguous DNA sequence, encompassing the transcription start site, that is sufficient to direct accurate initiation of transcription by the RNA polymerase II machinery, [Woychik and Hampsey 2002]. There seem to be two major types of core promoters - focused and dispersed [Juven-Gershon *et al.* 2008]. Focused promoters, which are more ancient and widespread throughout nature, will contain either a single transcription start site or a distinct cluster of start sites over several nucleotides [Smale and Kadonaga 2003]. Dispersed promoters, which, in vertebrates, are more common than focused promoters, contain several start sites over 50-100 nucleotides and are typically found in CpG islands. Core promoters contain a complex array of sequence motifs, refered to as TATA box, INR, DPE, MTE, BRE, DCE, and XCPE1 [Juven-Gershon *et al.* 2008; Smale and Kadonaga 2003; Maston *et al*. 2006] that specify different mechanisms of transcription and responses to enhancers. However, these motifs can occur in various combinations and not only have a high degree of degeneracy but there seems to be no universally required element within promoters necessary for promoter activity [Cooper *et al.* 2006].

In addition to the core region there are other specific *cis*-acting DNA regulatory sequences forming the extended promoter upstream and downstream of the transcription start site. These sequences regulate RNA polymerase II transcription controlling spatial and temporal expression of the downstream gene. They include the proximal promoter, enhancers, silencers and boundary/insulator elements [Butler and Kadonaga 2002]. The proximal

promoter is the region in the immediate vicinity of the transcription start site (roughly from −250 to +250 nt). Enhancers and silencers can be located many kbp from the transcription start site and act either to activate or to repress transcription. Boundary/insulator elements appear to prevent the spreading of the activating effects of enhancers or the repressive effects of silencers or heterochromatin [Butler and Kadonaga 2002]. These elements are recognized by both sequence-specific and general transcriptional regulators during transcription initiation, and serve to integrate signals from multiple cellular pathways to deliver specific, highly regulated expression of a gene [Smale and Kadonaga 2003].

Although much effort has been dedicated to identifying and describing eukaryotic promoters [Butler and Kadonaga 2002; Kim *et al.* 2005a; Cooper *et al.* 2006; Trinklein *et al.* 2007] and many sequence elements have been identified in subsets of promoters, the identification of the true start sites for all human transcripts is far from complete [Cramer 2007].

Alternative polyadenylation

The formation of mature mRNAs in vertebrates involves processing of the 3' untranslated region through a coupled reaction involving endonucleolytic cleavage followed by poly(A) synthesis 10–30 nt downstream of an AAUAAA or AUUAAA signal sequence. Additional U-rich elements upstream and downstream of the hexamer can also be present [Colgan and Manley 1997; Legendre and Gautheret 2003] and contribute to the definition of the correct cleavage and polyadenylation site. Alternative cleavage site selection for synthesis of poly(A) has been experimentaly validated [Moucadel *et al.* 2007] and shown to vary in 2.8% of cases [Pauws *et al.* 2001] and is thought to depend on variation of polyadenylation signals. Although the AAUAAA signal is often considered to be present in 90% of the mRNAs (and the AUUAAA variant in the other 10%) [Colgan and Manley 1997], alternate signals are certainly present in a significant fraction of the 3' ends. In addition to the two signals mentioned above, 10 other patterns of variant polyadenylation signal usage have been identified in human [Beaudoing *et al.* 2000] (Figure 1.6), which could account for 14.9% of the putative mRNA 3' ends. All the observed signals are single-base variants of the canonical AAUAAA hexamer. This suggests a model where a unique polyadenylation machinery is tolerant to a limited level of mutation in its regular signal. Positions 3, 4, and 6 are highly conserved, while positions 1, 2, and 5 are tolerant to point mutations (Figure 1.6). Variant signals (including the common AUUAAA) are processed less efficiently than the canonical signal and could therefore be selected for regulatory purposes.

A number of studies reveal that at least 50% of genes in mammalian genomes contain several polyadenylation sites and mRNAs with different 3'UTR regions can be produced from a single gene [Beaudoing *et al.* 2000; Lee *et al.* 2007a; Tian et *al*. 2005]. The choice of poly(A) sites may influence the stability, translation efficiency, or localization of an mRNA in a tissue- or disease-specific manner [Edwalds-Gilbert *et al.* 1997; Graber *et al.* 1999; Jankovic *et al.* 1990; van Solinge *et al.* 1996].

Alternative poly(A) sites are commonly classified into tandem poly(A) sites located on the same 3'-exon, and sites located on different exons (including composite exons) formed by alternative splicing [Edwalds-Gilbert *et al.* 1997; Tian *et al*. 2005; Yan and Marr 2005]. Alternative 3' ends involving different 3' exons may affect the coding sequence and therefore have obvious functional consequences. However, the actual functional

consequence of tandem poly(A) sites, producing 3' ends that differ solely by the length of the 3' UTR, is still largely unknown. The mRNAs with multiple poly(A) sites tend to use non-canonical polyadenylation signals (including the common AUUAAA) more often than mRNAs with a single poly(A) site [Beaudoing *et al.* 2000]. Regulation of 3' end variation through alternative poly(A) site usage may be tissue-specific [Beaudoing and Gautheret 2001; Hu *et al.* 2005; Zhang *et al.* 2005]. Recently reports show that alternative polyadenylation can also be regulated by genomic imprinting [Wood *et al.* 2008].

Alternative polyadenylation patterns have been shown to be conserved in mammals [Ara *et al.* 2006]. Features such as the presence of multiple cleavage sites, distribution of poly (A) signal variants and nucleotide composition of flanking regions were reported to be similar in human and mouse [Tian *et al*. 2005]. All this evidence suggests that alternative polyadenylation is a widespread mechanism that contributes to transcript diversity in eukaryotes.

Figure 1.6 – The 12 putative human polyadenylation signals. The canonical signal AAUAAA is thought to be present on 90% of mRNAs. The other 11 variant signals observed are single-base variants of the canonical AAUAAA hexamer (single base variation highlighted in blue). Positions conserved in more than 90% of the variants are highlighted in yellow [Beaudoing et al. 2000]. Consensus sequence shown below (N = any nucleotide).

1. AAUAAA
2. AUUAAA
3. AGUAAA
4. UAUAAA
5. CAUAAA
6. GAUAAA
7. AAUAUA
8. AAUACA
9. AAUAGA
10. ACUAAA
11. AAGAAA
12. AAUGAA

NNUANA

Alternative translation initiation

The process of translation initiation of messenger RNAs in eukaryotes is conventionally thought to proceed through three steps: binding of a 40S ribosomal subunit-factor complex to the capped 5' end of the mRNA, followed by linear migration of the ribosome along the mRNA, with translation of the genetic code into protein beginning upon encounter of the ribosome with the first AUG codon [Kozak 1989]. However, the context of the AUG codon can affect the start of translation [Kozak 1987a; Fekete et al. 2007]. In mammals, the optimal consensus sequence, GCC**R**CC**AUG**G (R = purine), has been identified as the most efficient context for translation initiation [Kozak 1987b, 1989, 1995, 1997]. A less favorable context may cause the first AUG codon to be bypassed and the scanning mechanism to continue until an AUG codon in a more favorable context for translation initiation is found. AUG start codons are designated strong when position -3 is a purine base (**R** shown in bold above) and position +4 is a guanine base (**G** shown in bold above); or weak based on lack of conformity to the consensus sequence at positions -3 and +4 [Kozak 1987b, 1989, 1995, 1997]. This variation in translation initiation is referred to as reinitiation and context-dependent leaky scanning [Kozak 2001, Kozak 2002]. Bifunctional mRNAs [Kozak 1986] have also been found where a stop codon is followed closely by a start codon and the ribosome reinitiates translation of a second protein from the same mRNA. mRNA secondary structure at the 5' end has also been shown to affect translation initiation at AUG codons within an unfavourable context [Kozak 1990] including initiation at non-AUG codons [Kochetov 2007]. Another mechanism of translation has been discussed in the literature where it is proposed that the ribosome enters the mRNA not at the 5' capped end but at an internal ribosome entry site (IRES), however, some degree of controversy surrounds this proposed mechanism [Fujimura *et al*. 2008; Kozak 2005].

1.1.4. ABBERANT ALTERNATIVE SPLICING IN MYOTONIC DYSTROPHY – A CASE STUDY

Myotonic Dystrophy is a multisystemic disorder caused by microsatellite expansions of CTG (in DM1) and CCTG (in DM2) repeats in non translated regions of the DMPK and the ZNF9 genes respectively [Timchenko 1999]. Although the expansions are located on different chromosomes, there appears to be a common pathogenic mechanism involving the accumulation of transcripts into discrete nuclear RNA foci containing long tracts of CUG or CCUG repeats expressed from the expanded allele, and that these effects are independent of the DMPK or ZNF9 locus [Margolis *et al.* 2006]. The disease is mainly characterised by defects in skeletal muscle, but many other tissues and systems are involved such as cardiac muscle, brain, eye, and the endocrine system. Some of these defects have been traced to misregulation of alternative splicing of a number of genes [Kuyumcu-Martinez and Cooper 2006]. At least 20 alternative splicing events are disrupted in DM heart, skeletal muscle, or brain tissues [Cooper 2007]. The mechanism by which these expanded repeats alter the regulation of pre-mRNA alternative splicing is complex. A model of RNA-mediated pathogenesis has recently been propose [Kanadia *et al*. 2006]. Several aspects of this model have been studied in DM1 and are summarised in **Figure 1.7.**

Figure 1.7

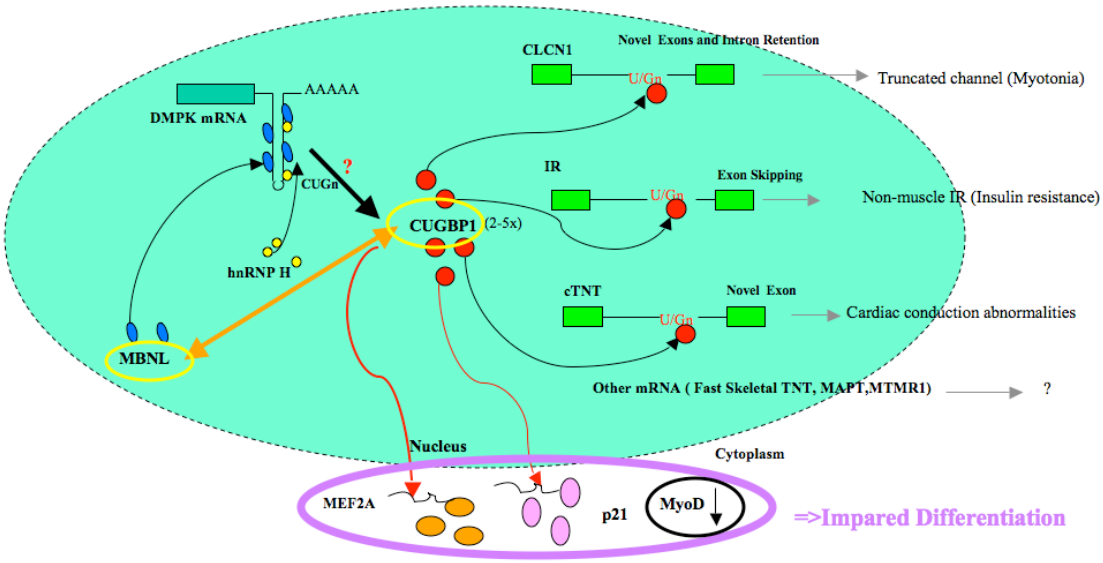


Figure 1.7 – Summary of some of the complex events which occur in Myotonic Dystrophy Type 1 (DM1) as a consequence of expanded CUG repeats in the DMPK gene. CUG repeats form a highly stable single-loop hairpin with an extended stem and mRNAs with expanded repeats accumulate in nuclear foci [Davis *et al.* 1997; Napierala and Krzyzosiak 1997; Kock and Leffert 1998]. These hairpins recruit double stranded CUG binding proteins such as MBNL [Mankodi *et al*. 2001; Miller *et al.* 2000] and hnRNP H [Kim *et al*. 2005b], both of which are involved in the regulation of splicing [Ho *et al.* 2004]. hnRNP H may also be responsible for the nuclear retention of the DMPK transcripts [Kim *et al.* 2005b]. Nuclear accumulation of CUG expanded repeats also results in a 2-5x overexpression of CUGBP1 activity and retention in the nucleus by some unknown mechanism [Timchenko *et al.* 2001]. However, CUGBP1 does not co-localize preferentially with the CUG expanded repeats, but rather binds preferentially to U/G rich sequences [Mori *et al*. 2008]. CUGBP1 is known to regulate at least three of the pre-mRNAs that are misregulated in DM striated muscle by binding to intronic U/G rich regions on these genes; these genes are chlorine channel 1(CLCN1), Insulin Receptor (IR), and cardiac Troponin T (cTNT). Aberrant alternative splicing in CLCN1 displays a novel exon and intron retentions which introduce a premature stop codon, the absence of

this channel being symptomatic of myotonia [Charlet-B *et al.* 2002; Mankodi *et al*. 2002; Lueck *et al.* 2007]. In the Insulin receptor an exon skipping event produces a non-muscle lower affinity isoform which results in insulin resistance [Savkur et al. 2001]. And in cardiac troponin T the introduction of a novel exon is associated with cardiac conduction abnormalities [Philips *et al.* 1998]. Other genes are also known to suffer aberrant alternative splicing in DM1 such as, microtubule-associated protein tau (MAPT) [Leroy et al. 2006] in both muscle and brain; myotubularin related protein 1 (MTMR1) [Buj-Bello et al. 2002] ; homeobox gene SIX5 [Sato et al. 2002]; and others, but the exact consequences of the mis-splicing are not known. Furthermore it has been demonstrated that inappropriate splicing patterns of cTNT and IR can be reproduced in normal cells by overexpression of CUGBP1 [Timchenko et al. 2004], which strongly suggests that aberrant regulation of alternative splicing of these genes in DM1 is, at least in part, a consequence of CUGBP1 overexpression. However although altered expression of genes post-transcriptionally regulated by CUGBP1 are likely to contribute to DM pathogenesis many other factors are simultaneously at play: cardiac troponin T is also a target for the splicing factor MBNL [Warf and Berglund 2007]; it has also been demonstrated that MBNL and CUGBP splicing factors are antagonistic regulators of alternative splicing [Ho et al. 2004] promoting opposite splicing patterns for cTNT and IR alternative exons; and that coordinated physical and functional interactions between hnRNP H, CUG-BP1 and MBNL1 dictate IR splicing in normal and DM1 myoblasts [Paul et al. 2006].

At the translational level, it has also been demonstrated in mice that the overexpression of CUGBP1 increases translation of proteins that are required for myogenesis such as MEF2A and p21 [Timchenko et al. 2004]. A similar induction of these proteins is observed in DM1 which diplays increased levels of CUGBP1. Overexpression of MEF2A and p21 may inhibit myogenesis and contribute to the deficiency in muscle development observed in DM1. A marked reduction of MyoD, a transcription factor required for the differentiation of myoblasts during muscle regeneration, is also observed in DM [Amack et al. 2002].

Interestingly in the case of chloride channel 1, ryanodine receptor 1, and sarcoplamic/endoplasmic reticulum Ca2+-ATPase 1 aberrant splicing in DM, the actually mRNA splice variants are not abnormal, but the normal developmental splicing pattern seems to be disrupted, resulting in expression of fetal protein isoforms that are inappropriate for adult tissues [Kimura et al 2005; Lin et al. 2006; Dhaenens et al. 2008;

Cooper 2007; Ranum and Cooper 2006]. A recent report also reveals the existence of MBNL1 fetal isoforms in the DM1 brain [Dhaenens et al. 2008]. The delayed differentiation in human myotonic dystrophy myoblasts affects fetal muscle development and compromises muscle regeneration in adults [Cooper 2007] . The case of disruption of alternative splicing in myotonic dystrophy may have revealed tampering with pathways of regulation of developmentally programmed alternative splicing [Pascual *et al.* 2006], involving splicing factors such as hnRNP-H, CUGBP1 and MBNL1. This hypothesis is supported by recent work which shows that MBNL1 participates in the post-natal remodelling of skeletal muscle by controlling a key set of developmentally regulated splicing switches and that sequestration of MBNL1 in DM and consequent failure to maintain these splicing transitions, has a pivotal role in muscle pathogenesis of this disease [Lin et al. 2006].

1.2. NON-PROTEIN-CODING RNA

Noncoding RNA (ncRNA) genes make transcripts that function directly as RNA, rather than encoding proteins [Eddy 1999, 2001]. The best known ncRNAs have complex three-dimensional RNA structures and play roles as catalytic or structural parts of RNA-protein machines; examples include transfer RNA, ribosomal RNA, and spliceosomal RNAs. Many other ncRNAs, especially many of the recently discovered ones, act in a relatively unsophisticated manner by base pairing to a target RNA. This binding can regulate gene expression directly (for instance, by sterically occluding a ribosome binding site), or provide RNA targeting specificity for a protein-based regulatory or modification mechanism: Examples include the micro RNAs (miRNAs) [Ambros 2001; Bartel 2004], E. coli translational regulatory RNAs [Wassarman *et al.* 1999], and small nucleolar RNAs (snoRNAs) [Eliceiri 1999; Kiss 2002; Kiss 2001]. In vertebrates snoRNAs are intimately involved in alternative splicing regulation [Bazeley et al. 2008]. Hundreds of small nucleolar RNAs are processed from pre-mRNA introns [Hirose *et al.* 2006; Kiss 2006]. Other RNA genes include snRNAs (small nuclear RNAs), scRNAs (small cytoplasmic RNAs), telomere RNAs, splice leaders and small regulatory RNAs. Some of these, such as the snoRNAs are hard to recognize in raw sequence [Hertel *et al.* 2008], and consequently their exact number is unknown. Others, such as the developmental timing regulators *lin-4* and *lin-7*, are known to perform important biological functions, but their discovery has depended entirely on genetic methods [Reinhart *et al.* 2000]. The number of known RNA genes in *C. elegans* is about 1000.

MicroRNAs

MicroRNAs constitute a large family of small, approximately 21-nucleotide-long, non-coding RNAs that have emerged as key post-transcriptional regulators of gene expression in metazoans and plants [He and Hannon 2004]. In mammals, microRNAs are predicted to control the activity of approximately 30% of all protein-coding genes, and have been shown to participate in the regulation of almost every cellular process investigated so far [Ambros 2003; Carrington and Ambros 2003; Ambros 2004; Farh *et al*. 2005; Maroney *et al.* 2006; Nilsen 2007; Wu and Belasco 2008] including cancer [Calin *et al.* 2006; Esquela-Kerscher and Slack 2006]. By base pairing to mRNAs, microRNAs mediate translational

repression or mRNA degradation [Filipowicz et al. 2008]. Some clustered microRNAs are likely to coordinately regulate target genes and specific regulatory relationships appear to be conserved between flies and mammals [Grün et al 2005].

Although biochemical methods [Zhang *et al.* 2007] have been more successful for predicting microRNAs, many bioinformatics tools have been developed to predict microRNAs and their targets [Eddy 2002; Rehmsmeier *et al.* 2004; Legendre *et al.* 2005; Ng Kwang Loong and Mishra 2007]. Although little is known about the transcriptional regulation of primary microRNAs (pri-miRNAs), and most metazoan miRNA genes do not have the classical signals for polyadenylylation [Ohler et al., 2004], evidence suggests that many are transcribed by RNA polymerase II [Bartel 2004] and are therefore capped; some however can be transcribed by RNA polymerase III [Bartel 2004; Borchert *et al.* 2006]. Many are located within introns of host genes, including protein-coding genes and non-coding genes and may therefore be regulated through the host gene promoters. Certain miRNAs are clustered in polycistronic transcripts suggesting that these mRNA are coordinately regulated [Lagos-Quintana et al. 2003]. The current model for the biogenesis of microRNAs is summarized in Figure 1.8 [He and Hannon 2004; Kim 2005].

Figure 1.8. (adapted from [He and Hanon 2004])

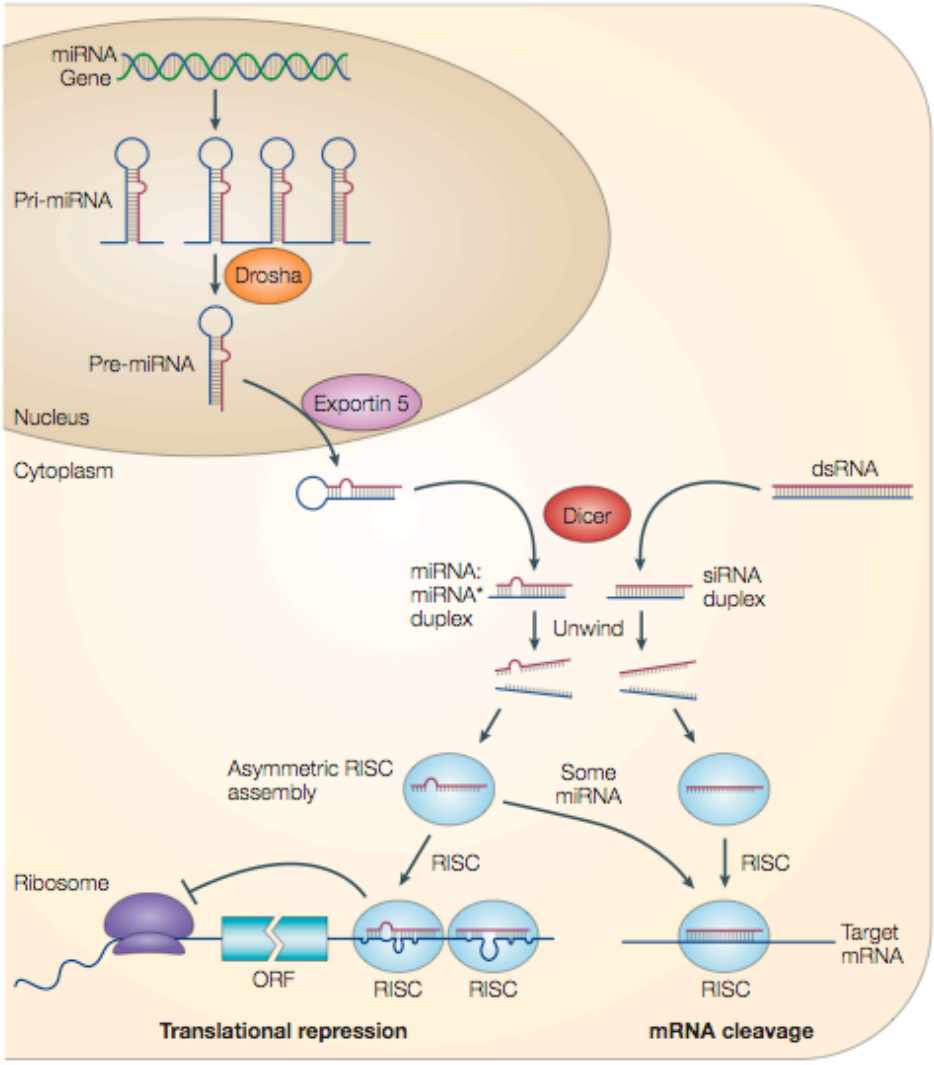


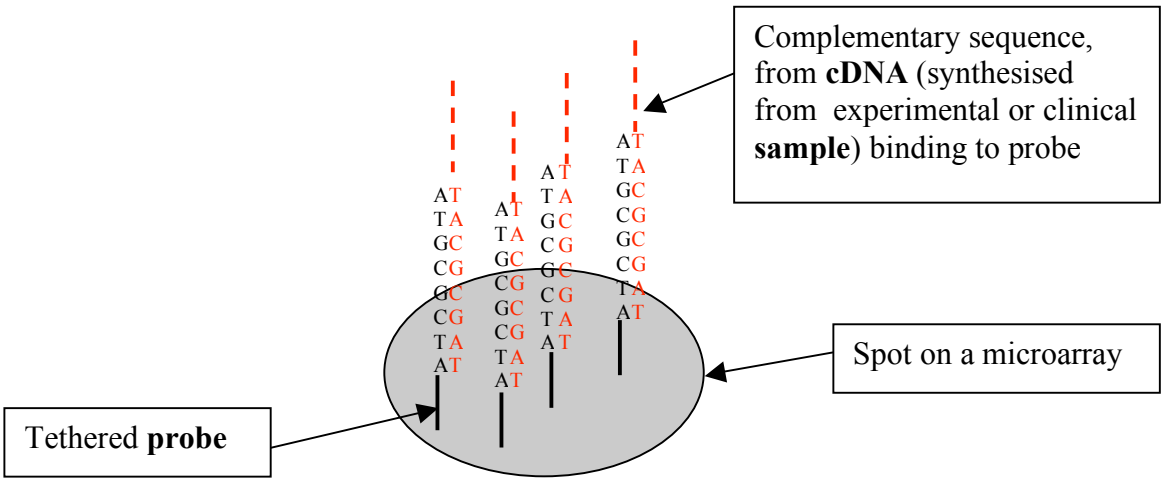
Figure 1.8. Current model for the biogenesis of microRNAs **(**Figure taken from [He and Hannon 2004]). Regardless of their diverse primary sequences and structures, pri-miRNAs are processed inside the nucleus, by the RNase III enzyme Drosha, into ~70-nucleotide pre-miRNAs that consist of an imperfect stem-loop structure [Lee et al. 2003]. Pre-miRNAs are transported to the cytoplasm by Exportin 5 [Lund *et al*. 2004]. In the cytoplasm these hairpin precursors are cleaved by Dicer [Hutvagner *et al.* 2001] into a small, imperfect double stranded RNA (dsRNA) duplex (miRNA: miRNA*) that contains both the mature miRNA strand and its complementary strand (miRNA*). Dicer also processes long dsRNA molecules into small interfering RNA (siRNA) duplexes. Cleavage by Dicer generates mature miRNAs that range from 21 to 25 nucleotides, the differences in size possibly

resulting from the presence of bulges and mismatches on the pre-miRNA stem. Only one strand of the miRNA:miRNA* duplex or the siRNA duplex is selectively assembled into the RNA-induced silencing complex (RISC), which subsequently acts on its target by translational repression or mRNA cleavage, depending, at least in part, on the level of complementarity between the small RNA and its target. This process is also known as post-transcriptional gene silencing (PTGS).

1.3. MICROARRAYS

Microarray technology is a widely used tool which allows the expression levels of thousands of genes to be simultaneously monitored [Allison *et al.* 2006; Schena *et al.* 1995; Schena *et al.* 1996]. The principle behind microarray analysis is the binding of complementary strands of nucleic acids. A microarray consist of a solid support, such as glass, plastic or silicon, onto which specific oligonucleotides are tethered in defined locations (or spots), by various methods. Each microscopic spot will contain several thousand copies of a unique oligonucleotide sequence referred to as a probe. These oligos are designed so as to be complementary to transcribed gene sequences as illustrated in Figure 1.9. Although transcribed genes are captured as RNA this molecule is less stable than DNA and therefore, for these experiments, it is common to transcribe it into cDNA which will contain both the sense and the antisence strand.

Figure 1.9. Complimentary binding of a tethered DNA probe to the sample cDNA



Probes can be designed either as the sense strand or the antisense strand. To ensure that binding of cDNA to a microarray is indeed specific, probes are designed to comply to several requirement such as: low potential for forming hairpins (binding to itself); absence of repeats; balanced nucleotide coverage; no secondary binding (binding to other genes); and good annealing temperature.

In order to visualise the bound cDNA the sample is labelled with a fluorophore, either Cy3, which has a fluorescence emission wavelength of 570 nm (green), or Cy5 with a fluorescence emission wavelength of 670 nm (red).

A one-colour (or single-channel, or single dye) microarray experiment will use cDNA from a single sample. A two-colour (or dual-channel) microarray experiment will simultaneously monitor two samples, labelled with different fluorophores, on the same microarray. Once the hybridization has taken place the results are viewed by scanning the microarray with a laser beam of a defined wavelength. The intensities (or relative intensities) of the fluorophore(s) are then used to identify up or down regulated genes.

Oligonucleotide microarrays will normally contain control probes designed to hybridize with RNA spike-ins (additional sequences added to the sample). The degree of hybridization between the spike-ins and the control probes is used to normalize the data.

A single-channel microarray will normally be analysed either by comparing genes within an array or to reference "normalizing" probes used to calibrate data across the entire array, or across multiple arrays. Although twice the amount of arrays are required to compare samples within an experiment, single channel arrays allow comparison of data from different experiments conducted at different times. A dual-channel microarray can be used to compare relative levels of expression in a diseased tissue versus a healthy tissue, for example. The advantage of using a dual channel is that only one microarray is needed.

One of the most popular companies producing commercial platforms is Affymetrix which has produced microarray platforms monitoring 3' end expression of genes. Recently Affymetrix has produced commercial arrays containing millions of probes which aim at covering the entire genome at the scale of an exon [W10; Gardina et al. 2006] rather than just monitoring the 3'end of a gene. One such array is the GeneChip® Human Exon 1.0 ST array containing approximately 5.5 million probes from a 1.4 millon probe set which can be used to interrogate 1 million known and predicted exons. This type of array is naturally inviting for the study of alternative splicing as the study of gene expression can be monitored at the level of individual isoforms and exons [Okoniewski 2007]

2. MATERIALS AND METHODS

2.1. CONSTRUCTION OF THE DATABASE

2.1.1. DATA SOURCES

Tables of BLAT alignments [Kent 2002] of RefSeq [Pruitt *et al.* 2007], mRNAs and spliced ESTs to a genome assembly were obtained from the UCSC Genome Browser [Karolchik *et al.* 2008; Karolchik *et al.* 2004; W6] and loaded into a MySQL database. The Perl programming language [Wall *et al.* 2000] and the Perl::DBI [W1] module were used to interface with the MySQL database. BLAT alignments of spliced ESTs taken from UCSC Genome Browser have a minimum of 96% base identity with the genomic sequence and will only carry introns smaller than 750,000 bases; in addition, such spliced ESTs retrieved from UCSC BLAT alignments must have at least one intron of minimum length 32 bases with a GT…AG consensus splice site. For the RefSeq alignments the RefFlat UCSC table was used to include a gene symbol associated with the alignment. The chromosome sequences for the genome assemblies of thirteen model organisms were obtained from the UCSC Genome Browser [W6]: human, *Homo sapiens*; chimp, *Pan troglodytes*; rhesus, *Macaca mulatta*; mouse, *Mus musculus*; rat, *Rattus norvegicus*; cow, *Bos taurus*; dog, *Canis lupus familiaris*; chicken, *Gallus gallus*; frog, *Xenopus tropicalis*; zebrafish, *Danio rerio*; fruit fly, *Drosophila melanogaster*; nematode worm, *Caenorhabditis elegans*; and sea squirt, *Ciona intestinalis*. Human ESTs originate from libraries generated from a variety of tissue sources (45% from normal tissue, 26% from tumor tissue or cell lines, 2% from disease tissue, and 27% with no tissue annotation).

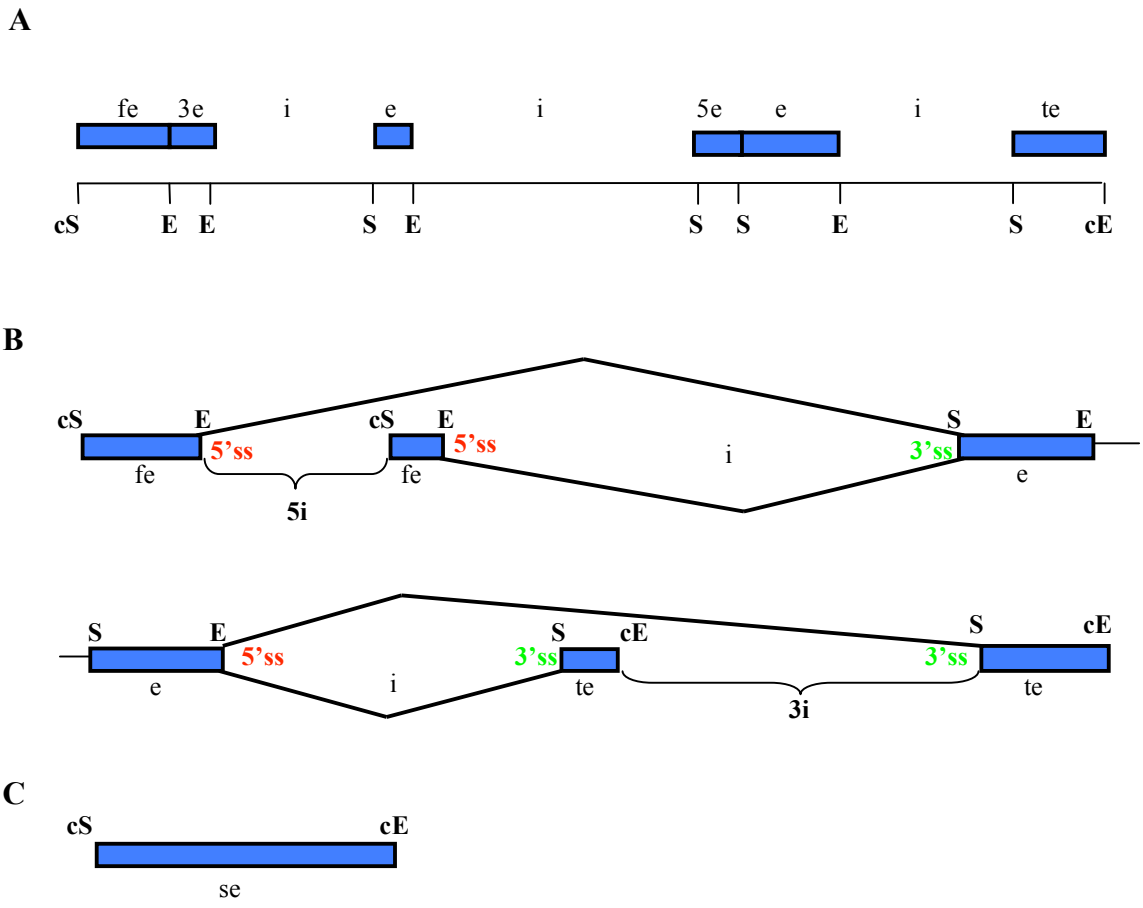
2.1.2. DATA PROCESSING

The genomic region of a gene was determined as that to which all RefSeq BLAT alignments associated with a given Entrez GeneID [Maglott et al. 2007] mapped, extended to any mRNA or spliced EST with at least one spliced junction in common with the RefSeq alignments. For unspliced RefSeq genes all mRNA and spliced EST alignments falling within the boundaries of the gene were considered. BLAT alignments consist of a series of

aligned blocks separated by gaps of genomic sequence. One would normally expect the aligned blocks to correspond to exons and the gaps in the alignment to correspond to introns, however, this is not always the case and the following steps are taken to filter out data not considered of sufficiently good quality to permit extraction of information on alternative splicing. To establish the minimum intron size we looked for the smallest RefSeq intron carrying the splice site consensus sequences GT_AG, GC_AG or AT_AC and found it to be 30 nt (SupplemetaryTable1.doc). Based on this result, a minimum intron size of 30nt was postulated. Two adjacent blocks separated by gaps smaller than 30nt were not considered to be reliable introns and were joined into a single block. These small gaps can be due to short repeats of variable length in the genomic sequence and the transcript or to short stretches of the transcript which the BLAT program did not succeed in aligning due to single nucleotide polymorphisms; their occurrence was mostly seen in 5' and 3' untranslated regions. A record of the sequence and position of all these 'filled in' regions was kept, including the transcript GenBank accession. Blocks in the BLAT alignment of a transcript mRNA or EST are potential exons. The smallest exon which we have detected in RefSeq annotated human genes is 6nt, in the RELN gene, but it is not mapped correctly to the genome assembly in UCSC BLAT alignments. A total of eight blocks with 7nt were found in RefSeq genes but of these only one mapped with consensus splice sites in UCSC BLAT alignments. A total of nine blocks with 8nt were found in RefSeq genes and, of these, seven mapped with consensus splice sites. There are more than 50 blocks of 9nt in RefSeq genes and these map with consensus splice sites in UCSC BLAT alignments, with few exceptions. A 9nt block was therefore chosen as a reliable lower threshold to impose on the size of blocks. Subsequently, the first two and last two nucleotides of each remaining gap in the alignment were analysed for splice site consensus sequences GT...AG, GC...AG, or AT...AC in the case of an mRNA source; and GT...AG or GC...AG in the case of an EST source. Where a splice site consensus did not exist, or a block was smaller than 9 nucleotides, an alignment was cut into fragments at the non-consensus junction or on either side of the small block respectively. Given that RefSeq sequences are constantly reviewed and are here used as reference sequences for a given gene, the program overrides this filtering procedure for RefSeq alignments and all of these are accepted regardless of the consensus sequence at the splice junction and the size of the aligned blocks. For the above set of filtered data the start and end of each alignment was corrected to the nearest known spliced junction or, for the terminal regions, to the longest corresponding block, according to the set of rules described in the Appendix A. The resulting blocks were considered to

correspond to exons and the intervening gaps to introns and each sequence segment was numbered and assigned a type as described in Figure 2.1. Thus, only a single EST or mRNA is required to support a new exon, provided that 96% of its sequence aligns to the genome assembly, that it contains at least one intron with consensus splice sites and has at least one splice site in common with a RefSeq annotated exon. No special requirement was imposed on first or last exons other than that these must have at least 9nt (the minimum block size required) and splice site consensus sequences. Any gaps smaller than 30 nt will have been patched up. If a gap greater than 30 nt between the first exon and the next exon does not have consensus splice sites the first exon is simply excluded.

Figure 2.1 - Nomenclature used to describe the organisation of the gene using the known starts and ends of all superimposed exons. The known starts and ends of all superimposed exons are used to partition the gene into segments. **A**. A cut start (cS) followed by an exon end (E) is considered a first exon (fe); two consecutive ends (E) is considered a 3' exon extension (3e); an end (E) followed by a start (S) is considered an intron; an exon start (S) followed by an exon end (E) is considered an exon (e); two consecutive starts (S) is considered a 5' exon extension (5e); and an exon start followed by a cut end is considered a terminal exon (te). **B**. Nomenclature used for alternative first exons and alternative terminal exons: an end (E) followed by a cut start (cS) is considered an intervening sequence which contains only a 5'ss (5i); and a cut end (cE) followed by a start (S) is considered an intervening sequence which contains only a 3'ss (3i). **C.** Nomenclature used for a gene containing a single exon: a cut start (cS) followed by a cut end (cE) is considered a single exon (se).



Representative set of splicing patterns

To obtain a representative set of isoforms, the minimal set of longest transcripts covering all the collected spliced junctions is retrieved by pattern matching of the various combinations of numbered sequence segments. Given that a gene can have several thousand ESTs mapping to it, using a representative set of splicing patterns can reduce the list of all splicing patterns collected by several orders of magnitude. A reference sequence, chosen as the RefSeq transcript with the highest version number and the most recent update, is used to number novel exons in relation to it. All the exons in this condensed set of isoforms are numbered in relation to a reference sequence as illustrated in Figure 2.5 B.

Alternative Splicing Events

The representative set of splicing patterns is built into a directed graph object using the Perl Graph module [W2] which is used to retrieve a description of all alternative splicing events using a generic description of an event as described in Figure 2.2. A region containing mutually exclusive exons is described using two or more events (Figure 2.3). Different isoforms, however, may carry different combinations of alternative splicing events and the alternative splicing events must be viewed in combination with the representative set of isoforms. The description of the gene as a succession of segments, as shown in Figure 2.1, allows for junctions to be described as a string of characters referred to as the "splice string" (Figure 2.4). This string of characters allows us to have additional information, on what segments are being skipped, when a region between two exons contains more than two splicing patterns (Figure 2.4).

Figure 2.2. Generic description of an alternative splicing event. C1, C2 indicate the constitutive exons and A1, A2 are the alternative sequences, either exons or exon extensions. Each junction C1,C2 skipping more than a simple intron or containing an intron retention is considered as a region of alternative splicing. In the Graph object, A1 is a successor of C1 and A2 as a predecessor of C2. Several options are possible for A1 and A2: for skipping of a single exon, an alternative 3' or 5' splice site or an intron retention, A1 will be equal to A2 (A1=A2); for alternative first exons there will be no A1; for alternative terminal exons there will be no A2; and for other more complex patterns of splicing A1 and A2 can be different exons or 3' or 5' exons extensions (A1≠A2).

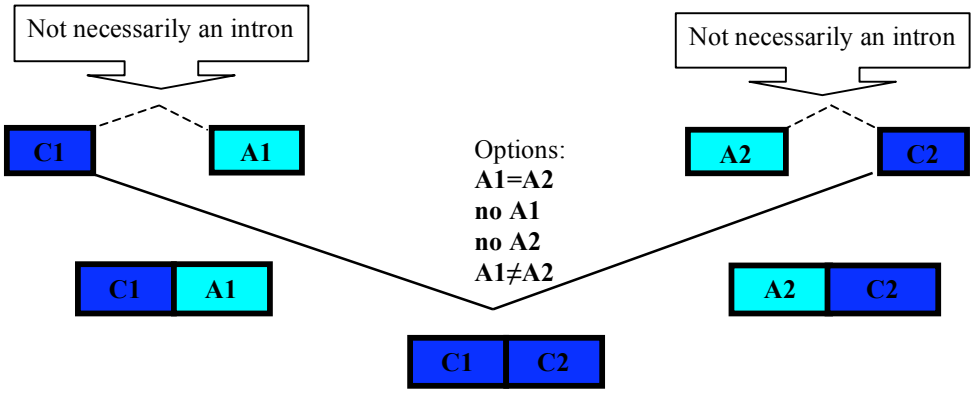


Figure 2.3. – Mutual exclusion events are described by two or more events of the type "no A1" and "no A2" (see generic description of an event in Figure 2.2).

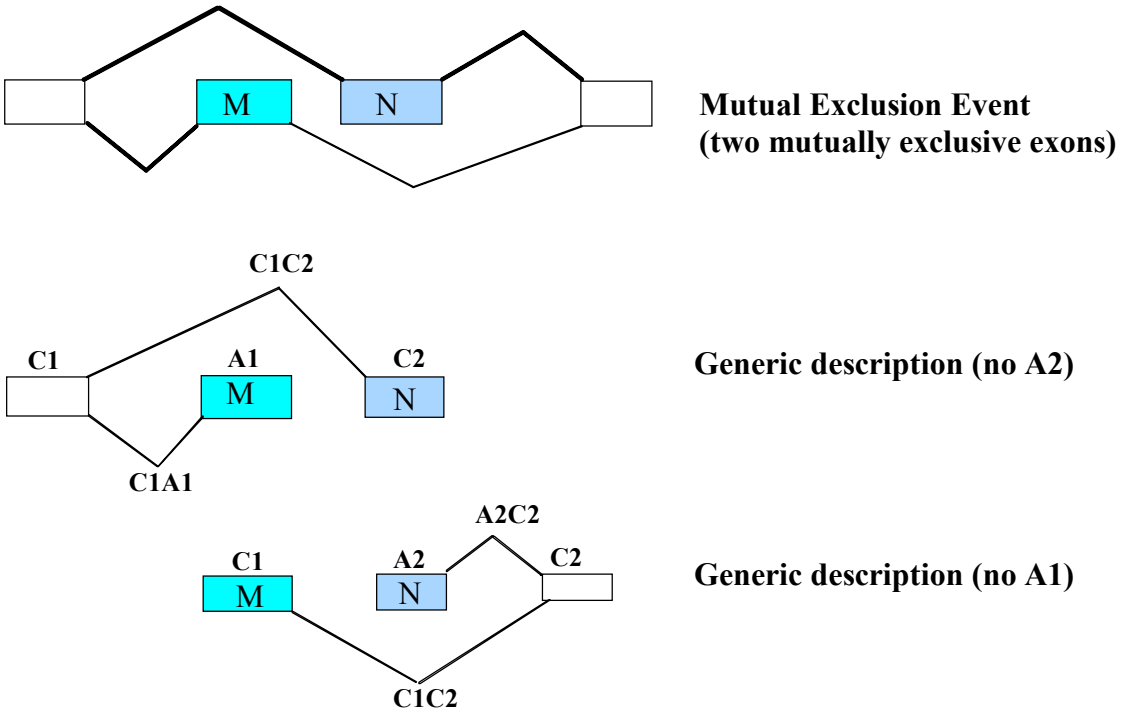
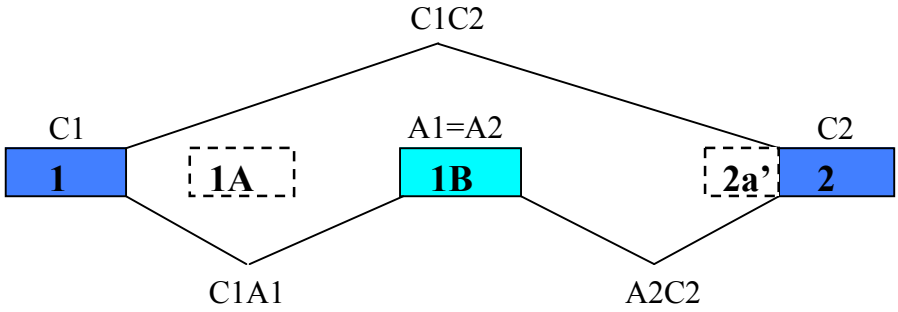


Figure 2.4 – Description of complex alternative splicing. In this figure a hypothetical complex alternative splicing event is presented, where more than two splicing patterns are possible between exons 1 and 2. Boxes indicate exons and the spaces between the boxes indicate intervening sequences, or introns. The generic event shown in Figure 2.2 allows for an alternative splicing event to be described by comparing two splicing patterns: e.g. splicing pattern (1,1B,2) and splicing pattern (1,2) resulting in the event description "Cassette exon 1B". On the other hand the description of the gene as a succession of segments, as shown in Figure 2.1, allows for junctions such as that between exons 1 and 2 to be described as a string of characters (the splice string) in which the spliced segment type characters are in uppercase and the skipped segment type characters are in lowercase. In this case the splice string reveals that in the junction (1,2) there are two cassette exons and a 5' extension of exon 2 being skipped.



Event description: Cassette exon 1B

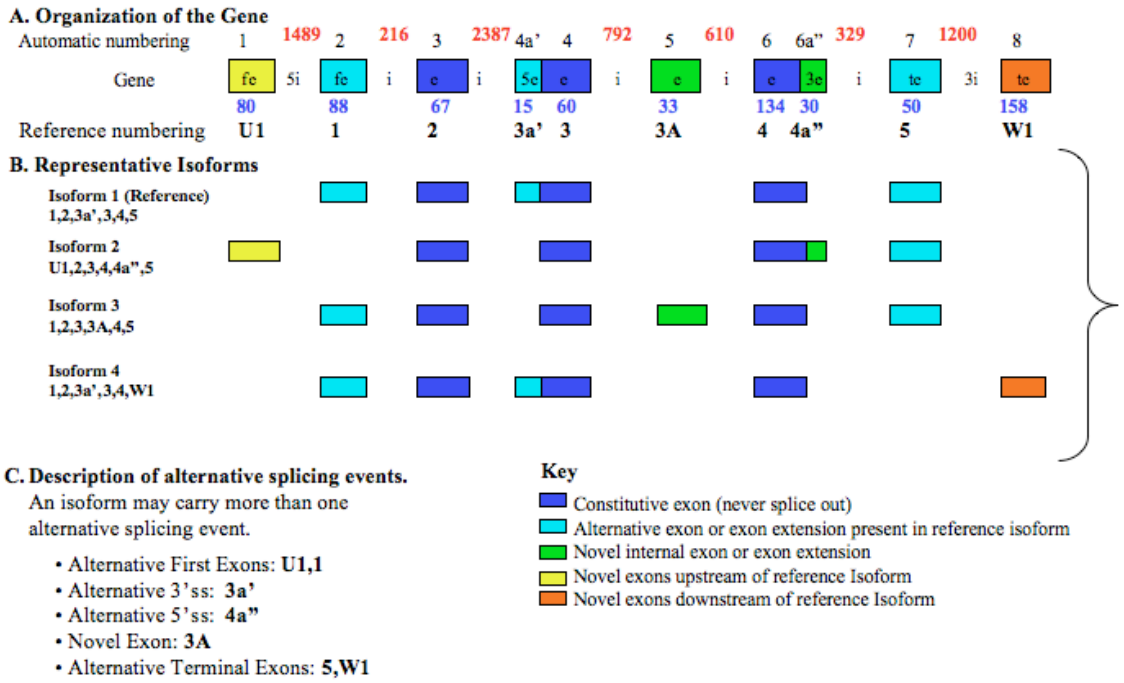
Splice string of junction C1C2: E,i,e,i,e,i,5e,E

Constitutive exons

Given the fragmented nature of ESTs, and some mRNAs, used here for mining alternative splicing, determination of constitutive exons is not straightforward. Our approach for selecting constitutive exons is done by choosing those which are not detected as being spliced out in the representative set of isoforms; first or terminal exons will only be marked as constitutive if the gene contains a single first or single terminal exon respectively. Exons C1 and C2 in Figure 2.2 are only constitutive in relation to the description of a particular event; they will not be marked as constitutive if they are spliced out in another event. This approach can result in exons upstream and downstream of the reference sequence being marked as constitutive, however these can easily be distinguished by the reference numbering system (Figure 2.5).

In Figure 2.5 we display a hypothetical gene to illustrate the re-numbering of exons in relation to the reference isoform; and a representative set of isoforms with novel, upstream, downstream, alternative and constitutive exons colour coded. This representation was used as a basis for an automatically generated web graphical display. The description of alternative splicing events is provided by comparing pairs of isoforms, however different isoforms may carry different combinations of alternative splicing events and the alternative splicing events must be viewed in combination with the representative set of isoforms.

Figure 2.5 - **A.** Organization of a hypothetical gene. GenBank accessions for all the representative set of isoforms are normally given. Boxes represent exons; exon lengths in nucleotides are indicated below each box (blue); intron lengths are indicated in nucleotides above in the gaps between the boxes (red). **B.** Set of determined representative isoforms. A reference sequence is chosen from this set and exons can be renumbered in relation to this reference sequence. In the reference numbering system upstream exons are preceeded by the letter U and downstream exons by the letter W. Novel exons take the number of the previous exon succeeded by a capital letter. **C.** Literal description of alternative splicing events.



Polyadenylation sites

Alternative polyadenylation has been shown to be associated to alternative splicing [Edwalds-Gilbert *et al.* 1997]. Polyadelylation sites and their positions in a transcript, based on variants of the canonical AAUAAA motif [Beaudoing *et al.* 2000] are determined for the whole transcript of each representative isoform.

Tissue expression

To determine tissue expression we use the gbCdnaInfo table available from the UCSC Table Browser [Karolchik *et al*. 2004], which associates each accession with various data retrieved from the original GenBank record such as: tissue, developmental stage or sex. Every mRNA and EST collected for each gene in our database is associated with the sequential numbering of exons as exemplified in Figure 2.5. Using simple pattern matching the user can determine all the accessions which contain a particular exon, junction or splicing pattern and thereby find the corresponding tissue expression information.

Open reading frames

For information on open reading frames a table is provided with translation of each representative isoform starting on the first, second and third nucleotide and then from each AUG start codon to the first in-frame stop codon. This allows the user to select individual putative proteins from this table.

Orthologs

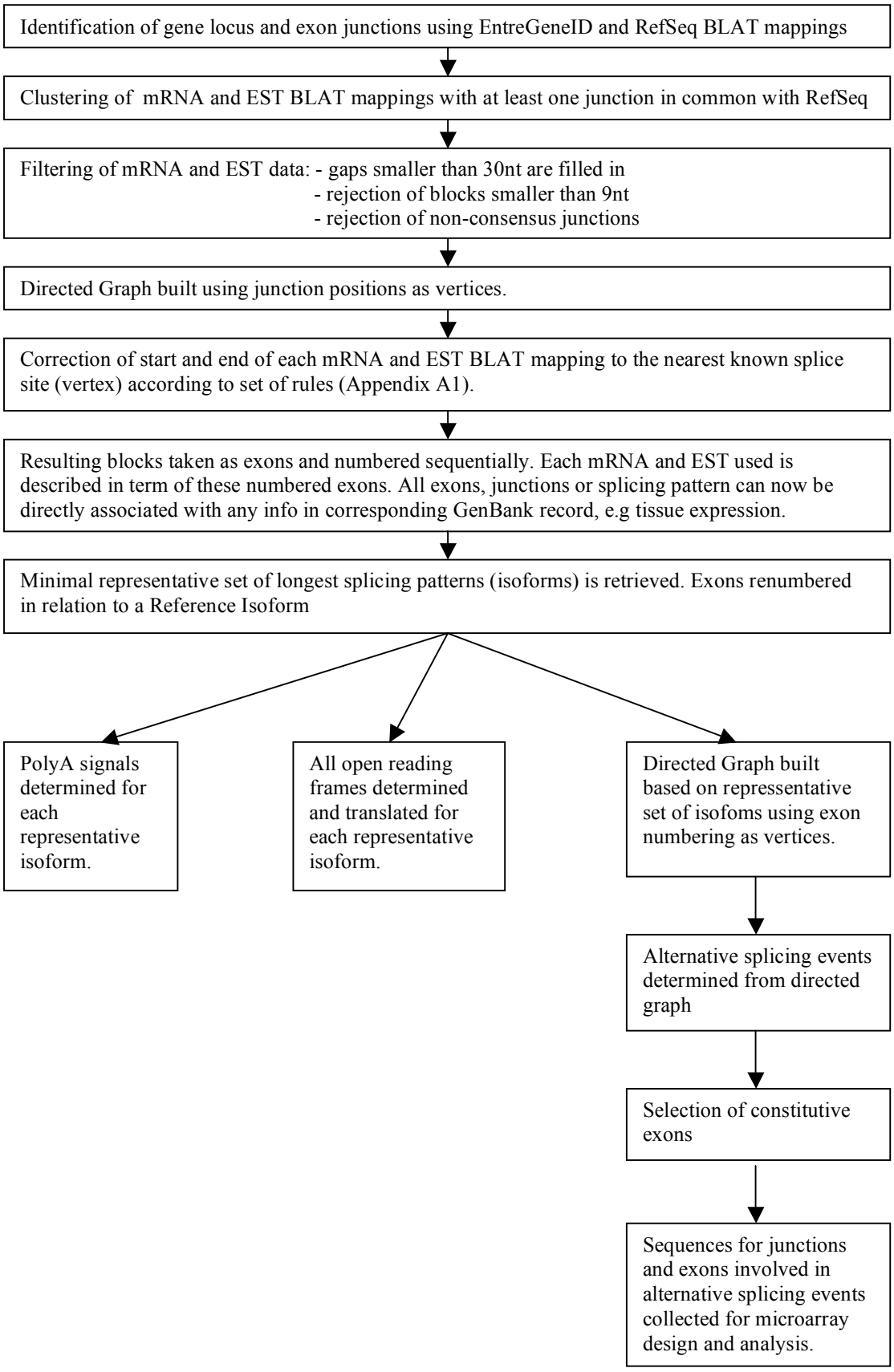
Information on orthologs of human genes is also provided. These are obtained by selecting BLAT alignments of RefSeq accessions from other organisms to a particular human gene locus. These alignments are available from the UCSC Genome Browser.

Data for Microarray experiments

ExonMine also provides data for microarray experiments in the form of tables with all junctions (50 nucleotides centered on the junction) and exons involved in alternative

splicing events. The information on constitutive exons can also be used for designing or analysing alternative splicing microarrays. All information on exons and junctions can be referenced back to the original alternative splicing events. This type of information can be used either for custom design of alternative splicing microarrays or for relating commercial exon array probes to alternative splicing events. Figure 2.6 summarises the steps followed in processing the data for a given organism.

Figure 2.6 Summary of steps followed in the algorithm to produce ExonMine data for a give organism



3. RESULTS AND DISCUSSION

3.1. THE EXONMINE DATABASE

3.1.1. STATISTICS

All information collected is stored in a MySQL database structure. Summary of the input data used to determine alternative splicing in 13 model organisms and resulting number of genes and isoforms detected is described in Table 3.1. Here we show the number of alternatively spliced genes measured as the percentage of genes in which more than one isoform was detected, this includes intron retentions and transcripts that begin or end in an intronic region. For the human genome, for example , 3976715 spliced ESTs were loaded into the program and 3851376 (97%) of these were found to share at least one splice site with a RefSeq exon and to have at least one intron with consensus splice sites - these are referred to as captured or clustered ESTs. A direct comparison of the values in Table 3.1 can only be done for human, mouse, drosophila and C. elegans, for which we are assuming an equivalent coverage of alternative splicing events detected using ESTs (i.e., more than 95% of spliced ESTs were clustered). For these genomes, the increase in complexity of the organisms clearly accompanies the increase in splicing and the increase in alternative splicing, as the number of exons per gene increases and as both the number of genes with more than one isoform and the number of isoforms per gene rise. ExonMine detected twice as many splicing patterns in human than in mouse, with a comparable percentage of ESTs clustered, 97% and 96% respectively.

For organisms where more than 95% of spliced ESTs were clustered, we further compared the number of splicing patterns, alternative exons and open reading frames (Table 3.2). Data from chimp was not included because ESTs and mRNAs used are mainly of human origin. The proportional increase observed in different types of splicing patterns, relative to *C. elegans,* is indicated in Table 3.2. We see that the human data contains almost twice the amount of splicing patterns as the mouse data with an equivalent percentage of these

containing intron retentions (35-36%). The high level of intron retentions detected is most likely due to incompletely spliced transcripts. Similar levels of intron retentions have been detected in another study of EST transcripts to monitor alternative splicing [Zan et al. 2002]. A lower percentage of intron retention is observed in the fruit fly (18%) and nematode (11%). The proportional increase observed in the number of splicing patterns with alternative exons is significant: 30 fold for the human, 13 fold for the mouse and 3 fold for the fruit fly. Finally, Table 3.2 reveals that human and mouse have significantly more splicing patterns where the transcript contains a new exon (i.e. one not annotated by RefSeq), greater than 50nt, with no open reading frame in-frame with known exons (i.e. exons annotated by RefSeq). The proportional increase observed in the latter, relative to *C. elegans,* is dramatic: 305 fold for the human, 128 fold for the mouse but only 5 fold for the fruit fly (the data used to obtain these values is available on request from the ExonMine team). This contrasts greatly with the proportional increase observed in the total number of splicing patterns with alternative exons. These results are consistent with our premise that more complex layers of gene regulation involving putative non-protein-coding transcripts containing new exons are likely to exist in general in higher organisms and in particular in humans. About 4% (human) and 1.3%(mouse) of these have open reading frames (data not shown) which could potentially code for more than 150 amino acid; the possibility of some of these transcripts coding for small proteins could also be explored.

For human and mouse Table 3.3 shows that the majority of alternative exons captured by ExonMine have not been annotated by RefSeq (92% and 94% respectively) with a greater percentage of these being new first exons revealing an increased detection of putative alternative promoters. In the case of *D. melanogaster* and *C. elegans* two thirds of the exons captured have been annotated by RefSeq. Of the resulting 24345 new first exons (45% of 54515, Table 3.3) detected in the human the April 2008 update of ExonMine, 89% are greater than 50nt.

Notwithstanding the fact that the transcript data used by ExonMine to determine alternative splicing in chimp is mainly of human origin, the complexity of splicing obtained is very close to that of human; the chimp would therefore naturally provide a better model organism than the mouse for extrapolation to the human.

A manual analysis of the data obtained revealed that alternative splicing involving sequences shorter than 20 nucleotides should not be taken for granted. This is due to limitations of the BLAT program which does not always succeed in aligning junctions containing single nucleotide polymorphisms or exons shorter than 20 nucleotides, producing alignments that will appear as alternative splicing but are not. Normally these are expected to produce non-consensus splice sites and will be filtered out, but chance may produce a consensus splice site in the alignment. We tried to filter out such alignments but found that doing so could sometimes remove another valid alternative splicing event in the offending transcript and therefore decided instead to include them with this warning.

Table 3.1 - Summary of input and output data

Input data used to determine alternative splicing in each organism and resulting output number of genes and splicing patterns detected (April 2008 update). For *P. Troglodytes* both human and chimp mRNAs and ESTs were used. Genome assemblies may contain sequenced regions which have not yet been definitively assigned to the complete chromosome but are still used in the BLAT mapping process; this can result in a gene being assigned both to a chromosome and an unassigned region; on the other hand there may be more than one copy of a gene in the genome, this is particularly evident in *G. gallus*, *X. tropicalis* and *D. rerio* as the number of genes detected is greater than the number of RefSeq sequences used.

Organism	Genome assembly	Number of RefSeq transcripts	Number of clustered mRNAs	Number of clustered spliced ESTs	% Spliced ESTs captured	Number of Genes	% Genes with more than one splicing pattern detected	Average number of core exons per gene #	Average number of splicing patterns per gene
H. sapiens	hg18	25196	132059	3851376	97	19431	88	14	10.5
P. troglodytes	panTro2	25656	130388	3722877	97	19750	86	14	9.9
M. mulatta	rheMac2	466	1031	2603	8	466	23	7	1.5
M. musculus	mm9	21216	129099	1834467	96	19393	79	11	5.6
R. norvegicus	rn4	14345	16121	272192	82	14260	56	9	2.5
C. familiaris	canFam2	873	1109	17352	10	884	39	10	1.8
B. taurus	bosTau3	9496	13057	654778	78	10052	71	9	3.7
G. gallus	galGal3	4161	7149	125947	46	4267	71	11	3.3
X. tropicalis	xenTro2	6837	10588	504015	70	7764	64	9	3.0
D. rerio	Zv7	12659	18194	697710	81	13647	60	9	2.7
C. intestinalis	ci2	688	1125	41947	10	736	51	9	2.2
D. melanogaster	dm3	20714	16676	265263	98	14556	37	5	1.9
C. elegans	ce4	23121	2279	245458	98	20320	20	6	1.3

The number of exons is counted taking into account only the constitutive part of exons and not exon extentions.

Table 3.2 - Splicing patterns with intron retentions or with new exons but no open reading frames

The number of splicing patterns where the transcript contains a 'New Exon' greater than 50nt but no open reading frame (orf) in-frame with known exons was determined, excluding splicing patterns with intron retentions and splicing to more than one exon upstream or downstream of the reference sequence. Percentages (in brackets) are in relation to the total number of splicing patterns detected in that organism. The proportional increase in the number of splicing patterns relative to *C. elegans* is indicated in square brackets.

Organism	Genes	Total no. of Splicing Patterns [value/value for C. elegans]	Splicing Patterns with Intron Retentions (% Total Splicing Patterns) [value/value for C. elegans]	Splicing Patterns with Alternative Exons* (% Total Splicing Patterns) [value/value for C. elegans]	No. of splicing patterns where transcript contains New Exon** >50nt with no orf in-frame with known exons (% Total Splicing Patterns) [value/value for C. elegans]
H. sapiens	19431	204556 [8]	74649 (36%) [26]	118468 (58%) [30]	30475 (15%) [305]
M. musculus	19393	109303 [4]	38632 (35%) [13]	53210 (49%) [13]	12805 (12%) [128]
D. melanogaster	14556	27391 [1]	4979 (18%) [2]	10071 (37%) [3]	526 (2%) [5]
C. elegans	20320	26624 [1]	2894 (11%) [1]	3953 (15%) [1]	100 (<0.5%) [1]

* 'Alternative Exons' here refers to exons which do not appear in the reference sequence

** 'New Exon' here refers to an exon which has not been annotated by RefSeq

Table 3.3 - Alternative exons determined by ExonMine compared to those annotated by RefSeq

Alternative exons here are counted as those which are spliced out in at least one splicing event and do not appear in the reference sequence. First exons (F) are those for which no 3' splice site was detected. Terminal exons (T) are those for which no 5' splice site was detected. Internal exons (I) have both a 3' splice site and 5' splice site. First exons and terminal exons are counted as alternative if the gene contains more than one first exon or more than one terminal exon respectively.

Organism	Total Alternative Exons in ExonMine	Alternative Exons Annotated by RefSeq	Alternative Exons Annotated by RefSeq			Alternative Exons not annotated by RefSeq	Alternative Exons not annotated by RefSeq		
			% F	% I	% T		% F	% I	% T
H. sapiens	59451	4936 (8%)	26	52	22	54515 (92%)	45	35	20
M. musculus	27457	1741 (6%)	33	47	20	25716 (94%)	52	28	19
D. melanogaster	4408	2991 (68%)	54	37	9	1417 (32%)	53	19	28
C. elegans	1477	1043 (71%)	44	44	12	434 (29%)	54	21	25

3.1.2. COMPARISON WITH OTHER ALTERNATIVE SPLICING DATABASES

Comparison of ExonMine data for the human and mouse with some recently published alternative splicing databases (Table 3.4) clearly shows that ExonMine is detecting a greater percentage of alternative splicing in protein coding genes: in human ExonMine detects 88%, Fast DB [de la Grange *et al.* 2007] coming second with 78%. A detailed comparison of our results for selected human genes with other database revealed the most accurate and complete to be Fast DB which provides alternative splicing information based on mRNAs and ESTs for the human and mRNAs for the mouse. The increase in alternative splicing detected by ExonMine in relation to Fast DB most likely results from the filtering process imposed on the transcripts which is less stringent in ExonMine. Fast DB transcripts are initially selected by blasting each exon defined by EnsEMBL [W3] against mRNA and EST alignments to the genome, whereas ExonMine relies on BLAT alignments and requires a transcript to have only one splice site in common with RefSeq transcripts. Global percent identity of alignment of transcripts in Fast DB is 98% versus 96% in ExonMine. In Fast DB at least 95% of the transcript has to be aligned, it must cover 10% of the genomic region, and the ratio of exon and intron lengths in a given transcript cannot exceed three times the average ratio of all defined exon and intron lengths; whereas ExonMine cuts alignments into fragments at non-consensus junctions and on either side of exons smaller than nine nucleotides and recovers any spliced fragments. As a result of the data selection criteria ExonMine succeeds in capturing more than three times the number of ESTs and mRNAs than Fast DB (3,983,435 vs. 1,154,554), and almost twice as many as ASTD [Stamm *et al.* 2006] and ECgene [Lee *et al.* 2007b] (2,018,294 and 2,189,150 respectively). In terms of the number of exons detected, ASTD presents a higher number than ExonMine (325,692 vs. 311,488) because ASTD exons include all variations of an exon by 3' and 5' extensions.

Data from ECgene presents a low detection of alternative splicing (43%) and the high number of genes detected (49,546 in the human and 43,932 in the mouse) include spliced genes with no apparent protein coding potential. ExonMine, however, captured the great majority of human spliced ESTs (97%), which suggests a near complete coverage of genes

by human RefSeq transcripts and that the overwhelming majority of spliced transcripts reside in protein coding loci. ExonMine does, on the other hand, detect a high number of non-protein coding spliced isoforms incorporating new exons in relation to the known RefSeq annotated exons. This may explain the discrepancy between the number of genes detected by ECgene and ExonMine in that the latter may be capturing the transcripts with no apparent protein coding potential as being spliced to exons in known protein coding genes.

Table 3.4 - Comparison of ExonMine results for human and mouse with other alternative splicing databases: FastDB [de la Grange *et al.* 2007], ASAP [Kim *et al.* 2007b], ASTD [Stamm *et al.* 2006] and ECgene [Lee *et al.* 2007b].

	Human					Mouse				
Database	**Genes***	**% Genes with AS**	**mRNAs and ESTs used****	**Exons*****	**Internal Exons#**	**Genes***	**% Genes with AS**	**mRNAs and ESTs used****	**Exons*****	**Internal Exons#**
ExonMine	19,431	88%	3,983,435	311,488	197,572	19,393	79%	1,963,566	240,854	164,607
FastDB	16,053	78%	1,154,554	201,245	-	13,913	54%	78,170	157,920	-
ASAP II	22,220	53%	-	-	129,981	16,404	53%	-	-	105,260
ASTD	16,715	68%	2,018,294	325,692	-	16,491	57%	1,229,339	275,612	-
ECgene	49,546	43%	2,189,150	-	-	43,932	40%	1,416,309	-	-

* In the case of ECgene this count includes spliced non-coding genes many of which may be artefacts but one should not rule out the possibility of non-protein coding spliced genes.

**In the case of ExonMine and ECgene this count includes mRNAs and spliced ESTs; for ASAP II and FastDB this count includes all mRNAs and ESTs.

*** In the case of ExonMine this count includes core exons and exon extensions.

In the case of ExonMine this count includes only core exons.

3.1.3. CONSIDERATIONS ON TRANSCRIPTION INITIATION AND TERMINATION

As a result of the procedure followed for clustering mRNAs and ESTs (Appendix A), the start and end of transcripts as reported in the original GenBank records have been altered. These corrections are normally only of a few nucleotides, and were applied to deal with the fragmented nature of the input data. However, if the start or end of the mRNA or EST is indeed the start or end of a transcript, the correction applied moves the start or end of the transcript further upstream or downstream respectively. In a small number of case, where splicing to upstream or downstream exons is observed, a first or terminal exon of a representative splicing pattern may come described as an internal exon (type 'e'). In the case of transcription initiation this means that the first exon of a representative splicing pattern may not be the actual site of transcription initiation but will include upstream promoter sequences. In the case of alternative polyadenylation all polyadenylation signals, and variants of the canonical signal, are reported for each representative isoform. However, the user is given access to all the raw accession data described, using the same exon numbering as that used in the representative isoforms, so that all the variations in the start and end of clustered transcripts can be checked.

3.1.4. WEB INTERFACE

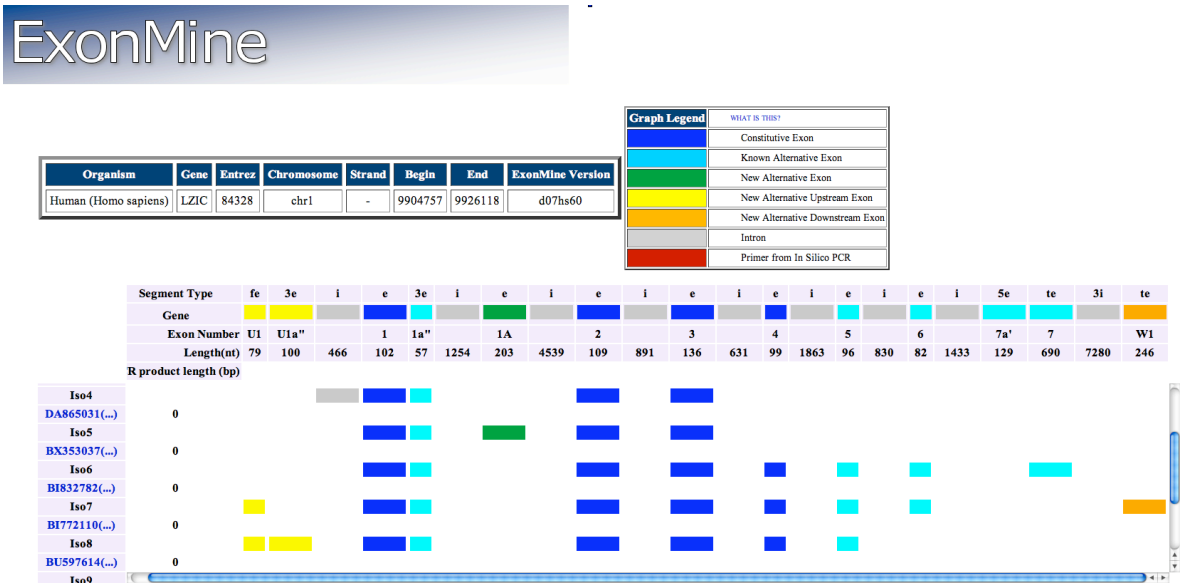
A web interface for public access to the ExonMine data was developed in collaboration with Daniel Felício Silva and Ana Rita Grosso at the Institute of Molecular Medicine. It is available at [http://www.imm.fm.ul.pt/exonmine] and is supported by clear graphical displays of the gene and transcripts.

To query ExonMine, the user selects an organism and enters the Gene Symbol ID, GenBank accession or Entrez Gene ID. ExonMine output provides information in both graphical and downloadable format. The graphical display is divided in two main parts: one with gene organization in terms of exons, exon extensions, introns, and their respective length; and the other with the splicing patterns (Figure 3.1). Exons are colour coded according to whether they are constitutive, known alternative exons, or new upstream, internal or downstream alternative exons (i.e. new exons are not annotated by RefSeq). The mRNA and/or EST accessions corresponding to each complete splicing pattern are given. For each gene we also provide the orthologue prediction in other organisms. Exon, intron and isoform sequences can be downloaded in fasta format. Tab delimited tables can be downloaded with information relative to: exons, introns and isoforms (including presence of poly-A signals, based on variants of the canonical AATAAA motif [Beaudoing *et al*. 2000]) for consideration of alternative polyadenylation [Ara et al. 2006]; start codons with neighboring context [Kozak 1987a], stop codons, and translation of all open reading frames allowing the user to explore possible variation through reinitiation and context-dependent leaky scanning [Kozak 2002], bifunctional mRNAs [Kozak 1986] and the controversial internal ribosome entry sites (IRES) [Kozak 2005]. The number and origin of ESTs and mRNAs supporting any exon, junction or splicing pattern can be obtained from the 'Downloads Options' by entering the exon number(s) in 'Tissue distribution in Gene given the transcript, junction or exon'. Tissue expression relative to source of mRNA and/or EST is based on the gbCdnaInfo Table available from the UCSC Table Browser. Junction sequences involved in alternative splicing events for microarray probe design (50 nucleotide sequences centered on the junction) are also available.

The web interface also provides five tools to aid in laboratory validation of individual splicing patterns. These tools were developed by Ana Rita Grosso using R language and packages *RMySQL* and *seqinr* [W5]. An "*In silico* PCR" tool allows the user to visualize a graphical representation of primer location and amplicon length on the relevant splicing pattern for a given gene. The user can also verify if the primers cross-amplify transcripts from other genes using the second tool. The third tool allows the search for isoform-specific sequences for primer design given a gene symbol and an isoform. Finally, with an input transcript sequence, the user can search for matching isoforms in ExonMine or decompose the input transcript sequence into ExonMine exons.

Figure 3.1 - Snapshot of the ExonMine graphical display

A gene is displayed using a succession of adjacent boxes representing exons, exon extension and introns in their genomic order. Constitutive and alternative exons are colour-coded and the length of each sequence is indicated in nucleotides. The genomic assembly sequence of each segment can be viewed by clicking on a box. Splicing patterns are displayed below the gene by the relevant boxes. GenBank accessions for each complete splicing pattern are given.



3.2. EXONMINE APPLIED TO GENOME WIDE ANALYSIS

3.2.1. MINING OF ALTERNATIVE SPLICING SEQUENCE REGULATORY ELEMENTS

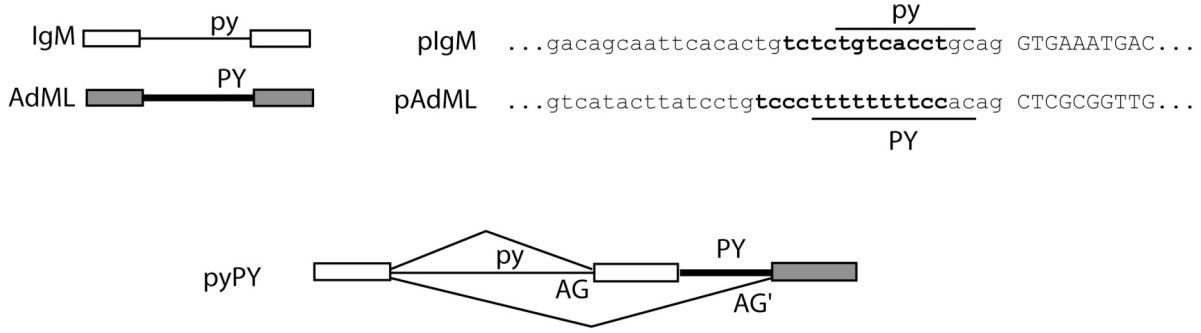
The usefulness of the ExonMine data produced, in searching for control sequences involved in alternative splicing, was demonstrated in a data mining procedure carried out to support experiments performed *in vitro*, on a construct, which revealed the requirement of the small subunit of U2AF for the recognition of a weak 3' splice site [Pacheco *et al.* 2006].

The construct (Figure 3.2) consisted of a proximal weak 3' splice site followed by a distal strong 3' splice site with the alternate exon sequence carrying a purine rich exonic splicing enhancer. The requirement of U2AF35 for splicing to the proximal weak 3' splice site had been demonstrated by Teresa Pacheco and an analysis of alternative 3' splice sites in existing genes was required to extrapolate this mechanism of alternate 3' splice site selection to known genes.

In purine rich exonic splicing enhancers the presence of the motif GAA has been shown to be essential [Aznarez *et al.* 2003; Liu *et al.* 1998; Yeakley *et al*. 1996; Tacke *et al.* 1998]. Following this lead, the alternate region of the exon was considered to house a candidate exonic splicing enhancer if it contained at least one GAA motif. The 3' splice site was considered 'strong' if it contained at least four consecutive T bases within the 30 nucleotides upstream of the splice site; and 'weak' if it contained no more than two consecutive T bases within that region. The alternate exon sequence was also restricted to the range 20-350 nucleotides (Figure 3.3 A). The search for candidates was performed using a Perl script which queries the ExonMine database taking only 1-2 minutes to scan about 270 thousand splicing events in about 19 thousand genes. In addition to returning the region of the polypyrimidine tracts and the putative ESE region surrounding a GAA motif, the script also returns all tissue information relative to the two alternative splicing patterns. This search resulted in 70 potential candidates (Appendix B) from which eight (Figure 3.3 B) were randomly selected for validation.

Figure 3.2 – **A.** Schematic representation of the minigene construct pyPY consisting of the 5' end of Immunoglobulin M (IgM) containing a weak polypyrimidine tract (py) and a purine rich exonic splicing enhancer (ESE); and the 3' end of adenovirus major late promoter (AdML) harbouring the second exon and 70 nucleotides upstream containing a strong polypyrimidine tract. **B.** Nucleotide sequence of the construct showing the weak (low in consecutive T nucleotides) and strong (high in consecutive T nucleotides) polypyrimidine tracts (boxed) and the purine rich exonic splicing enhancer (underlined) downstream of the proximal 3' splice site. Constitutive exon sequences are highlighted in green and the alternative exon sequence is highlighted in yellow; AG dinucleotides of the proximal and distal 3' splice sites are highlighted in red.

A



B

AGGGAGGTGAATGAGGAGGCTTTGAGAACCTGTGGACCACTGCCTCCACCTTCATCGTCCT
CTTCCTCCTGAGCCTCTTCTACAGCACCACCGTCACCCTGTTCAAGGTAGTATGGTTGTGG
GGCTGAGGACACAGGGCTGGGACAGGGAGTCACCAGTCCTCACTGCCTCTACCTCTACTCC
CTACAAGTGGACAGCAATTCACACTGTCTCTGTCACCTGCAGGTGAAATGACTCTCAGCAT
GGAAGGACAGCAGAGACCAAGAGATTCAAGCTTGCTGCACGTCTAGGGCGCAGTAGTCCAG
GGTTTCCTTGATGATGTCATACTTATCCTGTCCCTTTTTTTTCCACAGCTCGCGGTTGAGG
ACAAACTCTTCGCGGTCTTTCCAGTGGGGGC

Figure 3.3 – A. Schematic representation summarising the requirements imposed in the searching the ExonMine database for candidate alternative 3' splice sites containing a proximal weak 3' splice site followed by a distal strong 3' splice site with the alternative 3' exon extension (3e) harbouring an exonic splicing enhancer. **B**. Sequences of eight candidate endogenous genes selected for further analysis. Intron sequences are in lowercase letters; exon sequences are in capital letters. The number of nucleotides between the two alternative 3' sites (AltExSeq) is indicated.

A

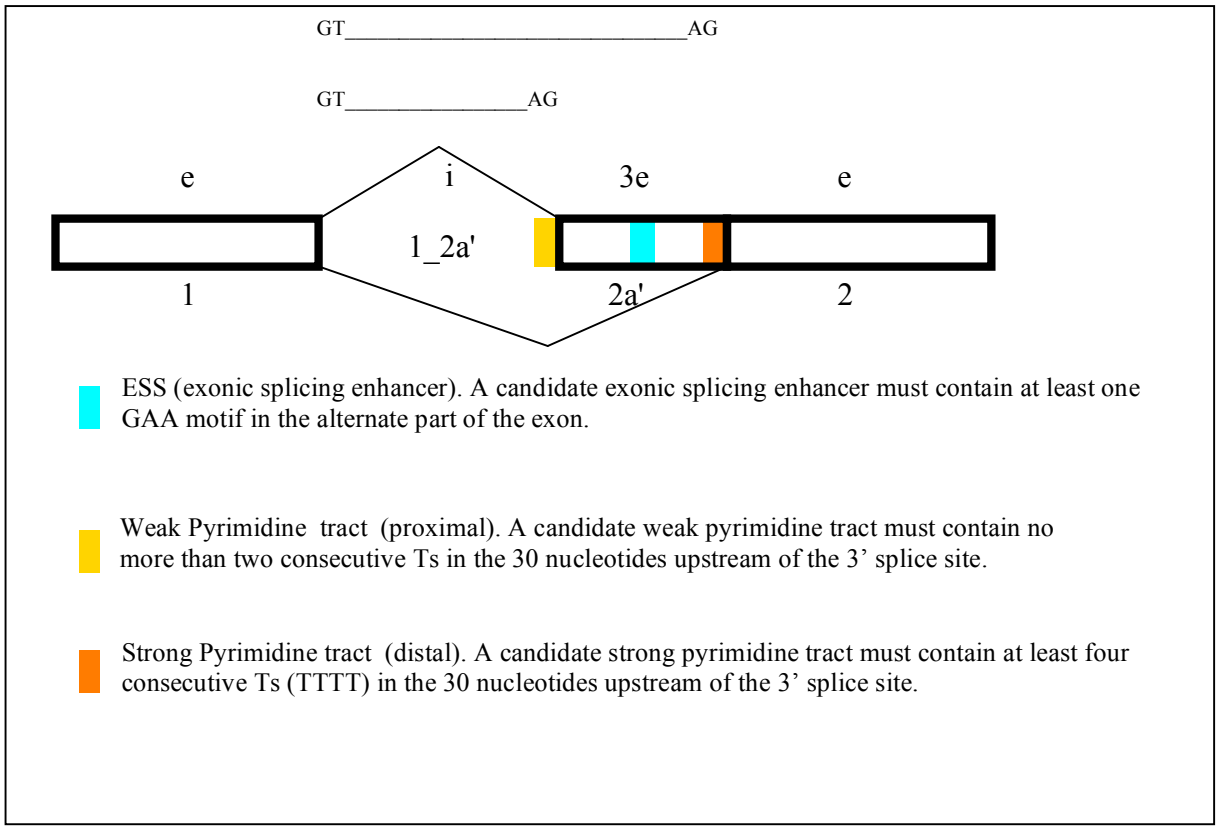


Figure 3.3 (cont.)

B

Gene		3' Splice Site Sequence (5'- 3')	AltExSeq (nt)
CUEDC1	Prox	tagtaacttctcccaaaactgtctctccag GAACCAGTGACCAGGAAAGACCATT	72
	Dist	attcctcccatccctttttcccctaccag ATGCCAGCAGCTCTTCCTTGCCAGA	
EIF3S7	Prox	gcctgaaagttgtctctcctgctccaccag CTGGTGACTGAAGGAAATACCTGCC	168
	Dist	catttctcttacatttttgtgtctttcag GAAGAAACTTAAACCAGTGATGTGG	
MBNL3	Prox	ctagtactagtaccatctctccccccacag GCCCTGCAGCCTGGTACACTGCAAC	105
	Dist	tttaatcccactgttttccactgccaacag GCTCTGACTAACCTGCAGCTCCCAC	
PPP1R13L	Prox	ctgatgtctgaagatcctgtctctccccag CTGTTCCCCAGGGTGAAGCCTCAAA	302
	Dist	acaggaaaccacttccttttgccaaatcag ATCCCGTCCAAAGTGCCTCCCATGC	
PTK9	Prox	agcccggtccctgctgccgtcacgttgcag ATTGATCCGAAATTCGGGGGAGCGA	117
	Dist	cgcccaactttcccccttttggtcacgtag GGAAGAGCTGGGACAGTGGAGGCGA	
SRRM2	Prox	gggagggaatgcagatgggagttgggggag GGGAGGATACAGTTCAGGATACCCC	248
	Dist	aggcattttggttttttaaaatctgtacag CAAGAGCAACTTTTTCTGTCAAATA	
TCP1	Prox	ttgcacatgttgaatatccggcacactcag ACCAGTGAATGTTTTCTCAACAGGT	23
	Dist	cactcagaccagtgaatgttttctcaacag GTTATTATTGCAGCAGAACTCCTAA	
ZNF124	Prox	attcttcatacgtcattcatcccttcatag GCACATCATATCTCATTCTGGAAAC	186
	Dist	aatatgtgagaaagtttttaatattcccag TTCATTTCAGATACATCAGAGAAAT	

Results, performed by Teresa Pacheco on HeLa cells at the Cell Biology Unit of the Institute of Molecular Medicine (Lisbon) and published [Pacheco *et al.* 2006], confirmed the existence of the splicing patterns found in the ExonMine database for four of the eight randomly selected candidates: CUEDC1, EIF3S7 (also called EIF3D), MBNL3, and PTK9 (also called TWF1). For CUEDC1 and EIF3S7 the major isoform detected in non-treated cells corresponded to splicing to the strong distal 3' splice site and RNAi depletion of U2AF35 displayed reduced splicing to the weak 3' splice site, supporting the requirement of the small subunit of U2AF for splicing to the weak 3' splice site. For MBNL3 and PTK9 the major isoform detected in non-treated cells corresponded to splicing to the weak proximal weak 3' splice site and RNAi depletion of U2AF35 did not significantly reduce splicing to the weak 3' splice site. The latter result suggest that predominant splicing to a weak 3' splice site depends on additional elements that stimulate splicing to the weak 3' splice site in a U2AF-independent manner.

The results published in [Pacheco et al. 2006] were dependent on an ExonMine database version based on the human genome assembly hg17 and data downloaded form the UCSC Genome Browser in June 2006. When the same analysis was performed on the December 2007 update of ExonMine, which is based on the human genome assembly hg18, one of the eight selected candidates, SRRM2 (for which, incidentally, the predicted splicing patterns were not detected in untreated HeLa cells), no longer appeared in the list of candidate genes. An analysis of the two ESTs found to display the alternative splicing pattern in the December 2007 version of ExonMine revealed BLAT mappings to gene SRRM2 which did not display the previously detected splicing pattern. This incident confirms the necessity of maintaining stable sets of data for reproducibility of results while simultaneously providing for updates at relatively frequent intervals.

3.2.2. CUSTOM DESIGN OF AN ALTERNATIVE SPLICING MICROARRAY

In collaboration with Juan Valcarcel's Lab at the 'Centre de Regulació Genòmica' in Barcelona, Spain, the ExonMine Database was used to produce a custom designed microarray platform for the study of alternative splicing in genes expressed in heart and skeletal muscle. The main objective of this project is to find new targuets of misregulation of alternative splicing in Myotonic Dystrophy and to identify variations in splicing factor concentrations and splicing patterns.

Given that dystrophic muscle seems to display foetal alternative splicing patterns, this project offers the opportunity to verify to what extent the patterns of alternative splicing vary between genes in myoblasts and those in dystrophic muscle. It may be that in Myotonic Dystrophy a mechanism of control of alternative splicing of adult versus foetal isoforms is being tampered with. This study may therefore produce insight not just into misregulation of alternative splicing in myotonic dystrophy but also into the workings of developmentally regulated mechanisms of alternative splicing.

A list of some 10,000 genes expressed in human muscle was obtained from microarray data deposited at NCBI in the Gene Expression Omnibus [W4] repository (Series id GSE740) [Johnson *et al.* 2003]. This collection of genes was compared with GNF Expression Atlas 1 Human Data on Affy U95 Chips avalable on the UCSC Genome Browser for UCSC annotated genes (Known Genes). A selection of 2000 genes with matching high values of expression in heart and skeletal muscle in both the above mentioned chips was made (Supplementary Data File 1). Alternative splicing data produced by ExonMine was analysed for these genes and used to custom design of probes for a microarray

Alternative splicing in splicing factors themselves is known to occur. MBNL genes, for example, have multiple splice variants, and different protein isoforms may have different binding affinities, binding specificities and splicing activities. Therefore, alterations in splicing factors at the level of expression and splicing in dystrophic muscle will also be monitored. A list of 420 alternative splicing factors and splicing associated factors was

compiled from various sources of published [Jurica and Moore 2003] and unpublished data supplied by various collaborators.

The selection of alternative splicing events to include in the microarray was made using the following criteria:

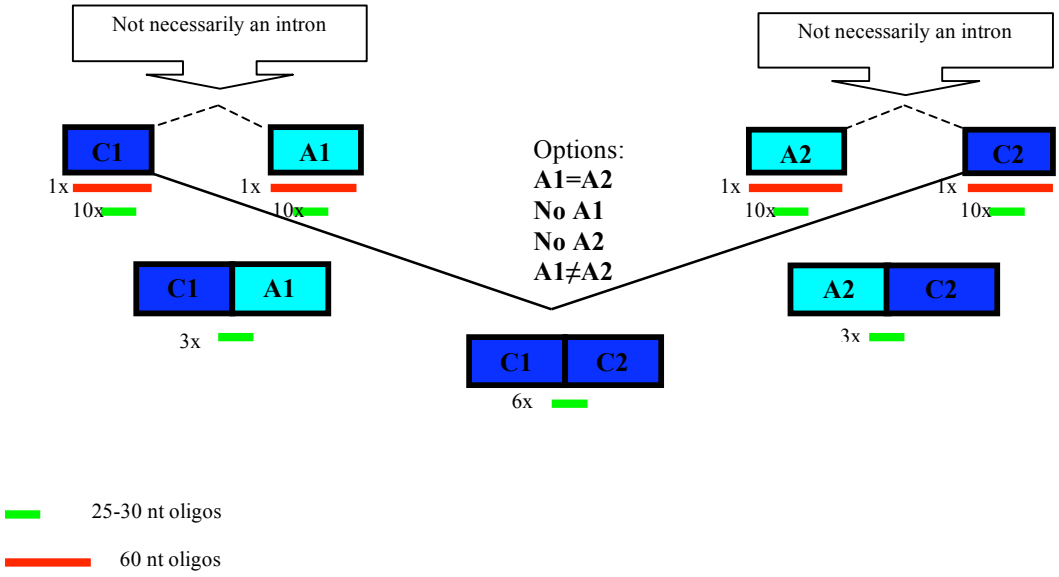
1. Exclusion of all intron retention events;
2. Exclusion of all alternative 3' and 5' splice sites involving exon segments shorter than 36 nucleotides;
3. Exclusion of events involving exons upstream or downstream of the reference sequence;
4. Only events described in relation to the reference isoform were included;
5. Only GT_AG or GC_AG consensus splice sites allowed.

The organization of data in ExonMine allows these selection criteria to be made in a single MySQL selection statement:

```
mysql> SELECT
SplID,SeqC1,SeqA1,SeqA2,SeqC2,JunctionOlig_C1A1,SpliceStringOlig_C1A1,SSConsen
susSeq_C1A1,JunctionOlig_A2C2,SpliceStringOlig_A2C2,SSConsensusSeq_A2C2,Juncti
onOlig_C1C2,SpliceStringOlig_C1C2,SSConsensusSeq_C1C2 FROM SplicingEvents
WHERE ((GeneID = '$GeneID') AND (RefEventDescription NOT REGEXP 'Intron') AND
(SSConsensusSeq_C1A1 REGEXP 'G[CT]..._..AG' OR SSConsensusSeq_C1A1 = '_' OR
SSConsensusSeq_C1A1 = 'null') AND (SSConsensusSeq_A1C2 REGEXP 'G[CT]..._..AG'
OR SSConsensusSeq_A2C2 = '_' OR SSConsensusSeq_A2C2 = 'null') AND
(SSConsensusSeq_C1C2 REGEXP 'G[CT]..._..AG' OR SSConsensusSeq_C1C2 = '_' OR
SSConsensusSeq_C1C2 = 'null') AND (SpecialID NOT REGEXP 'N') AND (LengthSeqC1
>= $MinExonLength OR LengthSeqC1 = 'null') AND (LengthSeqA1 >= $MinExonLength
OR LengthSeqA1 = 'null') AND (LengthSeqA2 >= $MinExonLength OR LengthSeqA2 =
'null') AND (LengthSeqC2 >= $MinExonLength OR LengthSeqC2 = 'null') AND
NOT((SeqC1 REGEXP '[Uu]' AND SeqA1 REGEXP '[Uu]' AND SeqA2 REGEXP '[Uu]'
AND SeqC2 REGEXP '[Uu]') OR (SeqC1 REGEXP '[Wu]' AND SeqA1 REGEXP '[Wu]'
AND SeqA2 REGEXP '[Wu]' AND SeqC2 REGEXP '[Wu]')) AND ((RepAcc_C1A1
REGEXP 'Reference' AND RepAcc_A2C2 REGEXP 'Reference') OR RepAcc_C1C2
REGEXP 'Reference')  );
```

For the selected alternative splicing events, exon and junction sequences corresponding to exons C1, C2, A1 and A2 (Figure 3.4) were sent for oligo design to Juan Valcarcel's team at the Centre de Regulació Genòmica' in Barcelona, Spain. For the junctions, three sequences per junction were sent each of 30 nucleotides: one centered on the junction, and the other two each with a one nucleotide shift upstream and downstream. Probes were designed using Array Designer 4 (PremierBiosoft commercial software) by Claudia Bendóv. Information on constitutive exons generated by the ExonMine database was used to add probes for an additional constitutive exon per gene.

Figure 3.4 . Probe design for alternative splicing event.



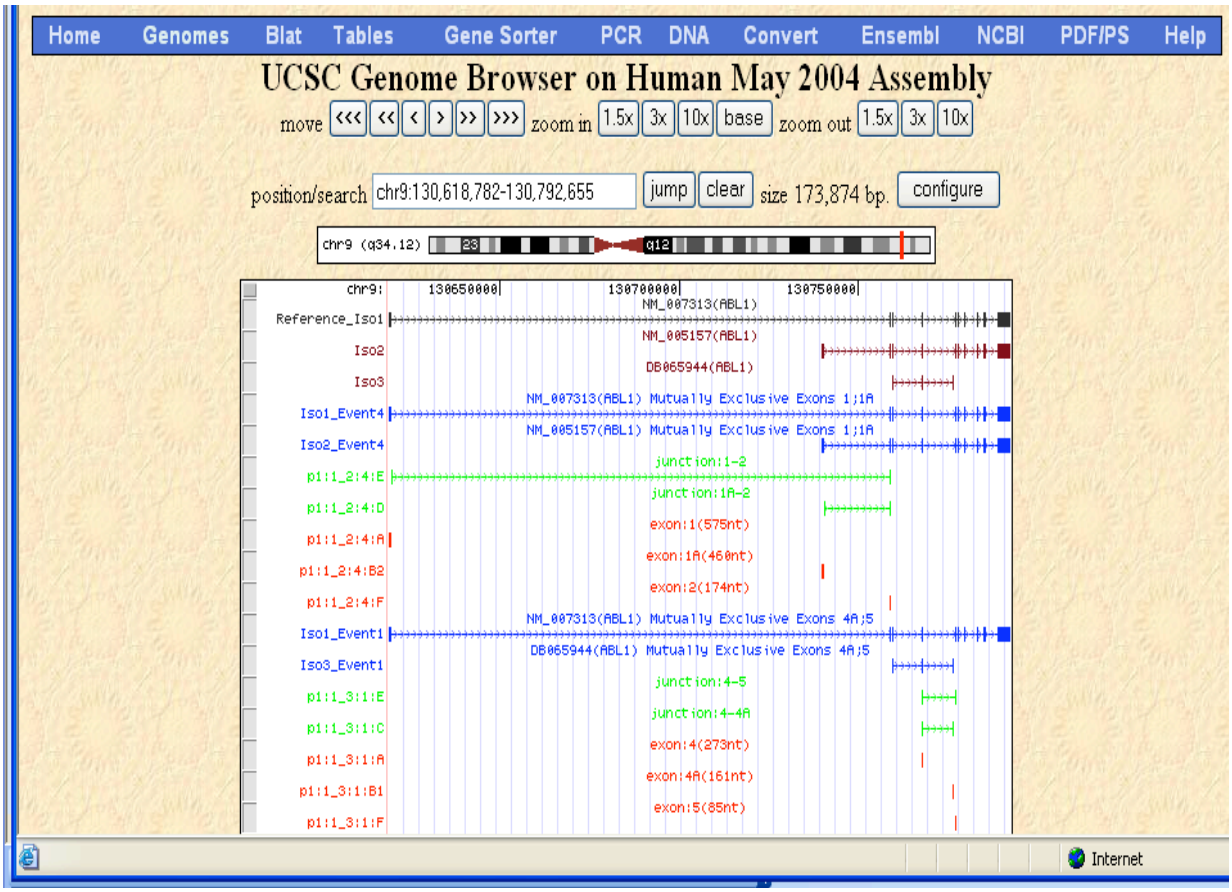
Given the high number of alternative splicing events detected by ExonMine and due to space constraints on the microarray chip, from the list of 2000 muscle genes a subset of 443 genes. A full description of all alternative splicing events detected in the 443 muscle genes and the 420 splicing factors and splicing associated factors is provided in Supplementary Data File 2 and 3 respectively.

Agilent microarrays with Inkjet-Deposited Presynthesized Oligos with two sets of 22,000 spots on a single chip, were ordered: one 22,000 spot array for the muscle specific events and the other for the splicing factors.

Visualization of the events on the Genome Browser

A Perl program was written to produce BED format files from the ExonMine Database which allow the representative isoforms and alternative splicing events to be visualised on the UCSC Genome Browser. Figure 3.5 shows the result of such a display on the UCSC Genome Browser for the gene ABL1.

Figure 3.5. Visualization of alternative splicing events on the UCSC Genome Browser



3.2.3. MATCHING DATABASE EVENTS TO AFFYMETRIX HUMAN EXON ARRAY

Recently Affymetrix has produced commercial arrays containing probes which aim at covering the entire genome at the scale of an exon. One such array is the GeneChip® Human Exon 1.0 ST array containing approximately 5.5 million probes from a 1.4 million probe set which can be used to interrogate 1 million known and predicted exons. This type of array has been extensively used for the study of alternative splicing as gene expression can be monitored at the level of individual isoforms and exons [53].

The set 1.4 million probes from the GeneChip® Human Exon 1.0 ST obtained from the UCSC Genome Browser for Human genome assembly hg18, was matched against the ExonMine database for human April 2008 update. Table 5 presents a summary statistics of the results obtained. Probe selection regions (PSRs) were found to match 19017 genes in ExonMine, out of a total of 19431 genes. Of the total of 1409312 PSRs, from the GeneChip® Human Exon 1.0 ST, 732157 match ExonMine genes: half of these match exons and the other half introns, 210995 PSRs match constitutive exons, 141320 PSRs match alternative exons or exon extensions and 108024 PSRs match junctions at the boundaries of an exon and an exon extensions or an intron.

An exon sequence may have several probes matching to it; out of a total of 311488 exons or exon extensions in ExonMine 252937 are covered by the GeneChip® Human Exon 1.0 ST probes. Table 5 also shows details of the number of matches to constitutive and alternative internal exons (e), first exons (fe), terminal exons (te), single exon genes (se), and alternative 3' (3e) and 5' (5e) exon extensions. Every alternative exon in ExonMine is linked to at least one splicing pattern and one alternative splicing event. The information on constitutive exons in the ExonMine database could be used to determine the level of expression of alternative exons by comparing the expression of probes matching each alternative exon in the gene against the average expresssion of the constitutive exons in a gene within an array, using a formula of the type shown below:

$$M_C = \frac{1}{N_C} \sum_{k=1}^{N_C} C_k$$

Average expression of probes on constitutive exons from a given gene

$$M_A = \frac{1}{N_A} \sum_{j=1}^{N_A} A_j$$

Average expression of probes on alternative exon, exon extention or intron from a given gene

$$V_A = \frac{M_A}{Mc}$$

Relative expression of probes on alternative exon, exon extention or intron from a given gene

where,

- N_C is the number of probes matching constitutive exons in a gene,
- C_k is the expression level of each probe matching a constitutive exon,
- N_A is the number of probes matching a particular alternative exon, exon extension or intron (A),
- A_j is the expression level of each probe matching an alternative exon, exon extention or intron, and
- V_A is the relative expression of the alternative exon versus the level of expression of the constitutive exons in a given gene. A value of V_A close to 1 would be interpreted as the alternative region being present.

Furthermore, the information on constitutive exons will make comparison of global gene expression more accurate by allowing a priori exclusion of probes matching alternative exons from the analysis. This data is currently being used in Carmo-Fonseca's Lab to analyse data from experiments performed on this platform.

Detailed analysis of probes in Affymetrix GeneChip® Human Exon 1.0 ST

Each one of the 1,409,312 probes selection regions (PSR) of the Affymetrix GeneChip® Human Exon 1.0 ST corresponds to a region on an exon, an exon extension or an intron which ranges from 24-9552nt. For each region one to four 25nt probes are designed giving a total of 5.5 million probes on the array. An example of this design and the relatonship with the ExonMine Data alignment is given in Appendix C.

Table 3.5 – Summary statistics of alignment of Affymetrix GeneChip® Human Exon 1.0 ST array probe selection regions to the ExonMine data

This analysis was performed using the affyHuEx1 table from UCSC Genome Browser and the ExonMine database for the human (April 2008 update).

Description	Quantity
Total affyHuEx1 probes	1,409,312
AffyHuEx1 probes that match ExonMine genes	732,157
AffyHuEx1 probes that match ExonMine constitutive exons (type e, fe or te)	210,995
AffyHuEx1 probes that match ExonMine alternative exons (type e,3e,5e,fe or te)	141,320
AffyHuEx1 probes that match junctions	108,024
No. of genes in ExonMine	19,431
No. of genes from ExonMine that match affiHuEx1 probes	19,017
Exon sequences in ExonMine (of type e,3e,5e,fe,te or se)	311,488
Exon sequences in ExonMine that match affiHuEx1 probes	252,937
AffyHuEx1 probes that match ExonMine exon sequences of type 'e'	206,228
ExonMine exon sequences of type 'e' that match affyHuEx1 probes	178,353
ExonMine constitutive exon sequences of type 'e' that match affyHuEx1 probes	131,217
ExonMine alternative exon sequences of type 'e' that match affyHuEx1 probes	47,136
AffyHuEx1 probes that match ExonMine exon sequences of type 'fe'	49,264
ExonMine exon sequences of type 'fe' that match affyHuEx1 probes	30,814
ExonMine constitutive exon sequences of type 'fe' that match affyHuEx1 probes	6,941
ExonMine alternative exon sequences of type 'fe' that match affyHuEx1 probes	23,873
AffyHuEx1 probes that match ExonMine exon sequences of type 'te'	71,152
ExonMine exon sequences of type 'te' that match affyHuEx1 probes	25,518
ExonMine constitutive exon sequences of type 'te' that match affyHuEx1 probes	10,873
ExonMine alternative exon sequences of type 'te' that match affyHuEx1 probes	14,645
AffyHuEx1 probes that match ExonMine exon sequences of type 'se'	5,734
ExonMine exon sequences of type 'se' that match affyHuEx1 probes	1,127
AffyHuEx1 probes that match ExonMine exon sequences of type '3e'	10,565
ExonMine exon sequences of type '3e' that match affyHuEx1 probes	8,815
AffyHuEx1 probes that match ExonMine exon sequences of type '5e'	10,321
ExonMine exon sequences of type '5e' that match affyHuEx1 probes	8,310

3.2.4. EXONMINE DATA COMPARED TO HUMAN TRANSCRIPTOME PROFILED BY TILING ARRAYS

Results produced by ExonMine were compared to steady-state polyadenylated RNA profiles obtained by using tiling arrays over the entire nonrepetitive portion of the human genome [Kapranov et al. 2007]. Table 3.6 presents the correlation found between ExonMine and the tiling array data for primary human dermal fibroblasts. A total of 585342 transcribed fragments detect transcripts greater than 200nt. Of these, 393428 fragments (67%) match 17672 protein-coding gene loci in ExonMine. About 38% of all known exons and 18% of new non-coding exons detected in ExonMine were found transcribed in primary fibroblasts. In a random selection of 50 of the new exons identified in ExonMine and verified in the tiling array data, none were found in Fast DB. Taken together these results strengthen the view that the human transcriptome contains even greater diversity of splice variants than previously suspected. Noteworthy, 69% of new non-coding exons identified in ExonMine are first exons, and the tiling arrays reveal that a significant fraction of protein-coding genes produce transcripts that overlap the promoter and the first exon and intron regions but do not include most of the other exons [Kapranov et al. 2007]. 53% of introns in ExonMine are detected in transcripts, although only 23% of these are retained in ExonMine data. Unraveling the potential biological function of such novel RNA species will clearly be the focus of future research.

Table 3.6 - Matching transcribed fragments from Affymetrix tiling array data from a primary cell line of Human Dermal Fibroblasts (HDF) to ExonMine Exons and Introns.

The counting of exons, known exons and new exons includes the constitutive part of an exon and 3' or 5' extentions. In the case of non-coding Exons (ncExons) only the constitutive part of the exon is counted. A transcribed fragment matching the junction of an exon and an intron is counted in both the Exon count and the Intron count. Transfrag (transcribed fragment), ncExons (non-coding exons), Known Exons (exons annotated by RefSeq), New Exons (exons not annotated by RefSeq), New Retained Introns (retained introns not annotated by RefSeq). Tiling array data from GEO Series accession number GSE-7576 [Kapranov et al. 2007].

Description	Quantity
Total number of long transcribed fragments	585,342
Genes in ExonMine matching transfrag	17,672
Transfrag matching Genes in ExonMine	393,428
Total Genes in ExonMine	19,431
Exons in ExonMine matching transfrag	103,017
Transfrag matching Exons in ExonMine	151,802
Total Exons in ExonMine	311,488
Known Exons in ExonMine matching transfrag	88,273
Transfrag matching Known Exons in ExonMine	140,871
Total Known Exons in ExonMine	230,492
New Exons in ExonMine matching transfrag	14,744
Transfrag matching New Exons in ExonMine	16,580
Total New Exons in ExonMine	80,996
ncExons in ExonMine matching transfrag	4,260
Transfrag matching ncExons in ExonMine	5,352
Total ncExons in ExonMine	23,727
Introns in ExonMine matching transfrag	133,459
Transfrag matching Introns in ExonMine	298,321
Total Introns in ExonMine	251,486
Retained Introns in ExonMine matching transfrag	31,369
Transfrag matching Retained Introns in ExonMine	62,799
Total Retained Introns in ExonMine	55,266
New Retained Introns in ExonMine matching transfrag	28,299
Transfrag matching New Retained Introns in ExonMine	52,592
Total New Retained Introns in ExonMine	49,450

3.3. ANALYSIS OF EXONMINE DATA FOR INDIVIDUAL GENES

3.3.1. ALTERNATIVE SPLICING AND DIVERSITY OF HUMAN U2AF PROTEINS

Our laboratory has previously reported that human transcripts encoding $U2AF^{35}$ can be alternatively spliced giving rise to three different mRNA isoforms called $U2AF^{35}$a, $U2AF^{35}$b, and $U2AF^{35}$c [Pacheco *et al.* 2004]. This discovery raised the question of whether additional U2AF genes produce alternatively spliced mRNAs. To date, only two CAPERβ mRNAs and four CAPERα mRNAs were detected in several human tissues by Northern blotting [Dowhan *et al.* 2005], and a splicing variant of PUF60/FIR was identified in colorectal cancers [Matsushita *et al.* 2006]. This scarcity of data prompted the use bioinformatics search strategies to review alternative splicing of U2AF genes. An analysis of alternative splicing information obtained from the ExonMine database for the family of $U2AF^{35}$ and $U2AF^{65}$ splicing factors was performed and published [Mollet *et al.* 2006].

The $U2AF^{35}$ family of proteins is composed of four genes: $U2AF^{35}$ (U2AF1), $U2AF^{35}$-RS1 (ZRSR1), $U2AF^{35}$-RS2 (ZRSR2) and $U2AF^{26}$ (U2AF1L4) [Shepard *et al.* 2002]. In the $U2AF^{65}$ family there are also four genes: $U2AF^{65}$ (U2AF2), CAPERα (RBM39), CAPERβ (RBM23), and PUF60 (SIAHBP1). The gene symbol in parenthesis corresponds to the official gene symbol used by Entrez Gene [Maglott *et al.* 2007] and is the one that appears in the ExonMine datababase.

What follows is a detailed discussion of the of the results obtained for each gene, which is subsequently summarised in Table 3.7:

For the gene that encodes $U2AF^{65}$ (U2AF2) we found two RefSeq mRNAs (NM_001012478.1, NM_007279.2) that reveal one alternative 5'ss at exon 10 which extends the exon by 12 nt (Appendix D, Table II.A, Figure II.A). Another mRNA (CR609498.1) is a 3'end fragment with a novel intron of 668nt removed, indicated in the figure by 12A*, and with no stop codon detected. There is also EST evidence for two novel

exons 1A (CD624005.1, CR982513.1) and 5A (CR982513.1, BI909492.1, CA488904.1), both of which contain an in-frame stop codon.

For the gene that encodes PUF60 (SIAHBP1) there are two RefSeq sequences NM_014281.3 and NM_078480.1 the latter including the novel exon 4A (Appendix D, Table II.B, Figure II.B). For each of these isoforms the RefSeq database reports a coding sequence running from exon 1 to 11, but the SwissProt database reports a coding sequence running from exon 3 to 11. Two other mRNAs BC009734.1 and BC011265.1 which are identical to the two RefSeq sequences except for the alternate first exon 1A, do indeed have a start codon only on exon 3 and the resulting proteins would therefore correspond to the ones reported in SwissProt for the RefSeq isoforms. Two EST fragments (BI915396.1, AL522753.3) show an alternative 3'ss at exon 2, which removes only three nucleotides, and cause no frame-shift; EST BX397429.2 also uses the alternative 3'ss on exon 2 but the alternative first exon 1A moves the start codon to exon 3, and therefore this splicing pattern does not affect the coding sequence. Skipping of exon 2 appears in ESTs AL514886.3 and BX384203.2 removing 87nt without causing a frame-shift. Another alternative first exon 2A comes up in mRNA AK055941.1 with a start codon on that same exon producing no frame-shift. EST BQ421738.1 uses exons 1A and 1B with a new start codon on exon 1B which again does not produce a frame-shift. Substitution of exon 1B by 1C in EST BM558085.1 moves the start codon to exon 3. Another combination is found in EST BQ956878.1 which uses exons 1 and 1B with the start codon now moving to 1B. A novel exon 1D in EST BG115238.1 also contains a start codon on the same exon and produces no frame-shift. The fragment EST BE393389.1 has an apparent first exon 4B with a start codon on exon 5 which would produce a coding sequence covering all RNA binding domains. And finally we found an odd splicing pattern in fragment EST BU170641.1 which shows sequence 6a", which previously functioned as an intron, replacing 7a', leaving the latter to function as an intron instead; this substitution adds 6nt to the mRNA and does not produce a frameshift. It is interesting to note that none of the alternative splicing patterns found removed the RNA binding domains.

The gene that encodes CAPERα (RBM39) is the longest of the genes we are analysing, covering about 17 exons (Appendix D, Table II.C, Figure II.C). There are five RefSeq variants for this gene, the longest isoforms, NM_184234.1 and NM_004902.2 run from

exon 2 through to exon 17 differing only in sequence 13b' which codes for 6aa. Two other isoforms NM_184241.1 and NM_184244.1 contain a premature stop codon in exon 2A and another, NM_184237.1, in exon 3C; in these three cases a start codon on exon 7 is used and produces two protein sequences, differing only in the sequence 13b', but which exclude the RS domain and include all the known RNA binding domains. The mRNA AL833168.1 includes exon 13A which introduces a frame shift resulting in a stop codon on exon 14; this would produce an isoform which excludes the UHM domain and the Poly-Ala stretch. mRNA BC107886.1 contains a known premature stop on exon 2A and two potential reading frames from exon 7 to 9A and from exon 10 to 17. Four ESTs (BM468718.1, BE816688.1, DA115481.1 and DB027200.1) include exon 2A with a premature stop at that exon 2A with no other in-frame start codon except on exon 7 for EST BM468718.1; the inclusion of exon 3C, skipping of exon 5, inclusion of exon 3B and exclusion of 4a', in the four first ESTs respectively has apparently no effect. The effect of splicing of exon 3A to exon 3B in EST AV691154.1 cannot be determined because this fragment has no known exons. The fragment EST AL711019.1 includes exon 3B with a premature stop codon on that exon. The inclusion of exon 2B in EST CA419145.1 also introduces a stop codon on that exon, and the start codon on exon 2B catches exon 3 in the known reading frame but encounters another stop codon on exon 3B. A fragment EST DA372839.1 shows exon 2A splicing to exon 2B, but here again there is a premature stop on exon 2B. Novel intron 1A* present in ESTs BP352717.1, DB027200.1 and DB264131.1 is in the 5'UTR; the first one of these has a putative alternative start on exon 2B after a premature stop on exon 2A; the second fragment has a premature stop on exon 2A and no other in-frame strat codon; and the third with the known start codon on exon 2 would produce a known coding sequence. A novel first exon 12A in EST DA109669.1 contains several start and stop codons, the longest orf would code for only 34aa. Alternate first exon 1A in ESTs DB150523.1, BG764840.1, DA922841.1 contains a premature stop codon on the same exon; in the second case the known start on exon 2 produces a known coding sequence; but in the third case there is another premature stop at exon 2A. The EST DA109669.1 showing a novel exon 12 A contains stop codons in all reading frames on exon 12A. Likely candidates of alternative splicing resulting in variation at the coding level come from the following examples: extension of exon 5 by 24 nt (sequence 5a') in EST BX483043.1 produces no frame-shift; evidence of skipping of exon 5, which does not produce a frame-shift, comes from ESTs BQ893325.1 and CR995560.1; skipping of exon 10 in EST BQ954122.1 would remove a known RNA binding domain containing an RNP1 motif; inclusion of novel

exons 11A and 11B in EST BM983358.1 or just 1B in EST BE933146.1 produces a putative in-frame orf. EST BI117009.1 shows a further extention 13a' of 35nt to the sequences 13b',13 which would introduce a stop codon on this very extension excluding the UHM domain; exclusion of sequence 6a' in EST BU075848.1 removes only 3nt coding for a serine residue; inclusion of sequence 2a" of 18 nt in EST DB023865.1 does not produce a frame-shift.

The gene that encodes CAPERβ (RBM23) contains about 13 exons (Appendix D, Table II.D, Figure II.D). Here there is one RefSeq sequence NM_018107.3 which has a coding sequence from exon 2 to exon 13, two other mRNAs (CR595426.1, BX161440.1) differ only in the skipping of exon 5 and the inclusion of exon 3B respectively; skipping of exon 5 will affect the RNA binding domain RRM1 and inclusion of exon 3B will add RS amino acids to the RS domain already known to be coded for by exon 4. One other complete mRNA sequence is known, AL834198.1, which includes exon 9A and subsequent intron retention; exon 9A contains a stop codon affecting the coding sequence at the RNA binding domain RRM2 and excludes the poly-Alanine stretch further downstream. EST BM712633.1 also shows inclusion of exon 9A but without the subsequent intron retention. Two ESTs, DA675412.1 and BG033916.1, give evidence of inclusion of exon 3B with exclusion of exon 5, which produces no frame-shift and is therefore a likely candidate for a novel splicing pattern. Six other ESTs with various splicing patterns at the 5'end of the gene were found that result in premature stop codons: DA821789.1 at exon 3A; DB164369.1, BM464794.1 and DA145418.1 at exon 3B; and BI823680.1 and DB166416.1 at exon 4. One EST fragment (DA117163.1) includes a novel exon 1C which carries a putative novel start codon. Another putative novel start codon is found in ESTs DA311282.1, BQ707907.1 and BQ071908.1 on exon 1Ea'. Other combinations of exons in the 5' UTR are found on ESTs AA324737.1, DA856230.1, CD692273.1 and BP356833.1 all of which have a start codon on exon 2. Splicing of exon 1 to 1A in EST AA633094.1 would equally be a 5'UTR pattern. There is also evidence of a novel intron in EST BX388764.2 which would remove the poly-Alanine stretch. Exclusion of exon 2 which contains the known start codon occurs in several ESTs: in EST BI915247.1 this produces a stop codon in exon 3, and in ESTs DA299707.1, DA026292.1, CN483101.1, CX165727.1 and mRNA BC106012.1 the stop codon occurs on exon 6; the only other in-frame start codon is on exon 6 for the ESTs and exon 8 for the mRNA. The EST 3'end fragment

BE269289.1 reveals exclusion of sequence 8a' which would produce a stop codon on exon 9.

For the gene that encodes U2AF^{35} (U2AF1) there are three RefSeq mRNAs that correspond to the previously identified isoforms U2AF^{35}a (NM_006758), U2AF^{35}b (NM_001025203), and U2AF^{35}c (NM_001025204), as described in Pacheco et al 2004. NM_006758 and NM_001025203 reveal alternate inclusion of exons 2A and 3 (Appendix D, Table II.E, Figure II.E). Consequently, the protein isoform U2AF^{35}b differs from U2AF^{35}a in only 7 amino acid residues located at the atypical RNA recognition motif (UHM) involved in dimerization with U2AF^{65}. However, biochemical experiments indicate that this substitution does not abolish the ability of U2AF^{35}b to interact with U2AF^{65}. U2AF^{35}a is 9- to 18-fold more abundant than U2AF^{35}b, with distinct tissue-specific patterns of expression [Pacheco et al 2004]. RefSeq mRNA NM_001025204 (corresponding to U2AF^{35}c) includes both exons 2A and 3 resulting in a premature stop codon in exon 3. Accordingly, experimental evidence indicates that the resulting mRNA is targeted to nonsense-mediated decay [Pacheco et al 2004]. However, an in-frame start codon in exon 4 would give rise to a coding sequence in the C-Terminal covering the RNA binding domain, a Zn-finger, the RS domain and the poly-Gly repeat. Additionally, EST BG612658 shows an alternate exon 2B in place of exons 2A and 3; this novel exon introduces a premature stop codon, however, a start codon exist on exon 2 with orf to exon 8, exons 3-8 being in-frame - this would produce a protein with a novel N-Terminal. In EST BE736536 there is an apparent novel first exon 1A containing several start codons; all, however, have a corresponding stop codon on the same exon.

For the gene that encodes U2AF^{26} (U2AF1L4) there are two RefSeq sequences, NM_144987.2 and NM_001040425.1, the latter differing only from the former by the inclusion of sequence 4a" extending exon 4 by 58nt at its 3'end (Appendix D, Table II.F, Figure II.F). This produces a change of frame which codes for a protein containing a Zn-finger which is not present in isoform NM_144987.2. An EST (BE856544.1) reveals two novel exon 2A and 2B which do not produce a frame-shift.

Inclusion of novel sequences 2Ab',2A in ESTs BM696851.1, BM970675.1 and AW274826.1 introduces a stop codon in exon 3, followed closely by a putative start codon on the same exon with a known stop codon on exon 6; the first two ESTs produce a known

coding sequence; but the third EST in this set skips both sequences 4a" and 6a'. In the case of EST DB127360.1 skiping of sequence 2a' introduces a premature stop codon in exon 3, and as in the case of EST AW274826.1 there is skipping of both exons 4a" and 6a' and a coding sequence going from exon 3 to 6.

In ESTs BU628789.1 and AA455588.1 inclusion of exon 2A produces a premature stop codon on exon 2A followed closely by a start codon on exon 3 with open reading frame all the way to exon 6. Other patterns of splicing at the 5'end revealed in ESTs BI770029.1, BC010865.1, BG481735.1 and W51842.1 produce premature stop codons in sequences 2Aa', 2Ab' or 2A. Inclusion of novel exons 5A and 5B in EST BU608847.1 would produces a frame shift with a stop codon on exon 6a'. Finally, two more ESTs, DB338076.1 and BF821614.1, include sequence 5Ba' in addition to 5A and 5B which introduces a stop codon at this new sequence 5Ba'.

For the gene that encodes U2AF^{35}-RS2 (ZRSR2) there is only one complete RefSeq sequence, NM_005089.2 (Appendix D, Table II.G, Figure II.G). Although there is mRNA and EST evidence (BC065719.1, DA173194.1, DA383795.1, CN289520.1, BE619312.1) for several patterns of splicing at the 5' end, all contain premature stop codons. Two novel first exons, 1A and 2A appear in ESTs DA261525.1 and CA425173.1 respectively; the latter also shows skipping of exon 6, however neither contains an open reading frame which includes known exons of this gene.

In conclusion, our analysis revealed that, with a single exception, all genes coding for U2AF proteins can be alternatively spliced. The exception is the U2AF^{35}-RS1 gene, which is devoid of introns. Many alternatively spliced mRNA isoforms are predicted to contain premature stop codons and are therefore expected to be targeted for degradation, thus contributing to regulate level of expression of the encoded protein [Lareau et al. 2004]. Additionally, we found evidence for several alternatively spliced mRNAs that could generate functional protein isoforms (see Table 3.7). Unraveling the roles played by these putative new human proteins awaits further experimentation.

Table 3.7. Predicted number of mRNA isoforms generated by alternative splicing of the U2AF family of genes. An alternatively spliced mRNA isoform was considered confirmed if its corresponding protein sequence is annotated in RefSeq or SwissProt databases. A splicing pattern observed in an mRNA or EST was predicted to produce a premature coding sequence termination if it contained an inframe stop codon within an internal exon. For the predicted patterns of splicing there is redundance in the number of accessions shown due to the fragmented nature of ESTs and some mRNAs. The information in this table is a summary of the more extended description presented in Appendix D.

Protein (Gene Symbol)	Confirmed mRNA isoforms (Accessions)	Predicted splicing patterns producing a premature stop codon (Accessions)	Predicted splicing patterns of candidates for putative novel protein (Accessions)
U2AF^{65} (U2AF2)	**2** (NM_007279.2, NM_001012478.1)	**2** (CD624005.1, CR982513.1, CA488904.1)	**2** (CR609498.1, BI909492.1)
PUF60 (SIAHBP1)	**4** (NM_014281.3, NM_078480.1, BC009734.1, BC011265.1)	**0**	**10** (BI915396.1, AL522753.3, AL514886.3, BX384203.2, AK055941.1, BQ421738.1, BQ956878.1, BG115238.1, BE393389.1, BU170641.1)
CAPERα (RNPC2)	**5** (NM_184234.1, NM_004902.2, NM_184241.1, NM_184244.1, NM_184237.1)	**5** (NM_184241.1, NM_184244.1, NM_184237.1, BC107886.1, BM468718.1, BE816688.1, DA115481.1, AL711019.1, CA419145.1, DA372839.1, BP352717.1, DB027200.1, DB150523.1, BG764840.1, DA922841.1, AW993266.1, AL513896.3)	**10** (BC107886.1, AL833168.1, BP352717.1, BX483043.1, BQ893325.1, CR995560.1, BQ954122.1, BE933146.1, BM983358.1, BU075848.1, DB023865.1)
CAPERβ (RBM23)	**4** (NM_018107.3, CR595426.1, BX161440.1, AL834198.1)	**10** (DA821789.1, DB164369.1, BM464794.1, DA145418.1, BI823680.1, DB166416.1, AA633094.1, BI915247.1, DA299707.1, DA026292.1, CN483101.1, CX165727.1, BC106012.1)	**8** (DA675412.1, BG033916.1, DA117163.1, DA311282.1, BQ707907.1, BQ071908.1, BX388764.2, BI915247.1, DA299707.1, DA026292.1, CN483101.1, CX165727.1,BC106012.1)
U2AF^{35} (U2AF1)	**3** (NM_006758.2, NM_001025203.1, NM_001025204.1)	**2** (NM_001025204.1, BE736536.1)	**1** (BG612658.1)
U2AF^{26} (U2AF1L4)	**2** (NM_144987.2, NM_001040425.1)	**6** (BM696851.1, BM970675.1, AW274826.1, DB127360.1, BU628789.1, AA455588.1, BI770029.1, BC010865.1, BG481735.1, W51842.1)	**6** (BE856544.1, BM696851.1, BM970675.1, AW274826.1, DB127360.1, BU628789.1, AA455588.1, BU608847.1, DB338076.1, BF821614.1)
U2AF^{35}R2 (U2AF1L2)	**1** (NM_005089.2)	**6** (BC065719.1, DA173194.1, DA383795.1, CN289520.1, BE619312.1, DA261525.1, CA425173.1)	**0**

3.3.2. VALIDATION OF ALTERNATIVE SPLICING IN U2AF35 FAMILY OF SPLICING FACTORS

The results of the previous analysis of the U2AF family of proteins revealed new putative protein-coding transcripts and non-protein-coding transcripts in the human. As a follow-up of the bioinformatic analysis, and in order to weed out possible artifacts, a more detailed analysis of a number of new transcripts and a study of the extent of conservation of alternative splicing patterns across six model organisms (*H. sapiens*, *M. musculus*, *D. rerio, G. gallus, D. melanogaster and C. elegans*) was performed for the three members of the family of genes of the small subunit of U2AF, namely $U2AF^{35}$(U2AF1), $U2AF^{26}$(U2AF1L4) and $U2AF^{35}$-RS2(ZRSR2). This family of proteins have in common a central UHM domian (U2AF homology motif) flanked by zinc fingers and a terminal Arginine/Serine rich domain (RS domain) except in the case of $U2AF^{26}$ where the RS domain is missing. Each UHM [Kielkopf *et al.* 2004] domain includes: an eight-residue RNA binding motif RNP1 (ribonucleoprotein 1) with consensus [RK]-G-[FY]-[GA]-[FY]-[ILV]-X-[FY]; a six amino acid RNA binding motif RNP2 (ribonucleoprotein 2) with consensus [ILV]-[FY]-[ILV]-X-N-L; and a protein interaction motif of the type RXF, where X is any amino acid.

Of these three genes *C. elegans* possesses only an orthologue of $U2AF^{35}$, called Y116A8C.35, with no alternative splicing; this gene contains four exons with an intronic division which does not match that of vertebrates. In *D. melanogaster* we find an orthologue for $U2AF^{35}$ called U2af38 with only one intron, corresponding to the first intron found in vertebrates, but no alternative splicing was detected. This would indicate that Arthropods and Nematodes diverged prior to the division of this gene into exons (see phylogeny of animal model organisms in Appendix E).

D. melanogaster also has an orthologue of $U2AF^{35}$-RS2 composed of 5 exons. This gene has only a 26% sequence similarity to the human ortholog and the position of the introns does not correspond to that found in vertebrates. Two isoforms are known (accession NM_135013.2 and NM_175962.1) and no other alternative splicing was detected by ExonMine.

For the four vertebrates studied the division into exons of the orthologues of these three genes is conserved and the layout of exons in these genes is presented in Figure 3.6. The work that follows consists in a more detailed study and validation of selected alternatively spliced transcripts which are present in human but not in the other vertebrate model organisms considered. This work was done in collaboration with two other members of Carmo Fonseca's lab, Joana Borlido and Sandra Martins.

Figure 3.6. – Diagram showing the layout of the exons determined by ExonMine for vertebrate orthologues of **A**. U2AF^{35}(U2AF1), **B**. U2AF^{26} (U2AF1L4) and **C**. U2AF^{35}-RS2 (ZRSR2): *H. sapiens* (H), *M. musculus* (M), *D. rerio* (Z), *G. gallus* (G). Boxes represent exons and these have been numbered above the boxes. Exon lengths in nucleotides are indicated below each box. Intron lengths are indicated in nucleotides above each diagram in the gaps between the boxes. Known protein domains, motifs are shown: the UHM domain composed of motifs RNP1, RNP2 and RWF; zinc finger motifs of type CCCH (Zn-f), Arginine /Serine rich domain (RS); and poly glycine stretch (Poly-Gly). The RNA binding domain has a secondary structure containing alpha-helices (α) and beta-sheets (β) with the pattern and conventional numbering β_1 α_A β_2 β_3 α_B β_4. The exons which have been described as belonging to a protein coding gene by RefSeq or SwissProt have been coloured grey. Boxes with diagonal pattern indicate exons not annotated by RefSeq which were confirmed by RT-PCR. The size of the boxes, representing exons, and the intervals between then, representing intervening sequences, are not to scale. For Chicken and Zebrafish U2AF^{35}-RS2 (ZRSR2) no RefSeq gene was available and the results are based on mRNA and EST sequences clustered by ExonMine using gene loci defined by GenScan genes chr1_25.193 (chicken) and chr11_3.36/ chr11_3.27 (duplicated gene in zebrafish). In the case of the zebrafish duplicated gene U2AF^{35}-RS2(ZRSR2) the exons in the region between exons 4 and 7 could not be determined by ExonMine with sufficient precision from the existing EST and mRNA data; the size of exons determined for this duplicated gene are identical and where differences exist in intron sizes these have been indicated for chr11_3.36/ chr11_3.27 respectively.

*Duplicated gene with slight variations at the intron level. The region between exons 4 and 7 is not covered with sufficient confidence by mRNA and EST information for the ExonMine program to detect exons 5 and 6, but a manual analysis of the sequence between exons 4 and 7 revealed the existence of sequences identical to the missing exons 5 and 6.

Figure 3.6 A - U2AF^{35}(U2AF1) orthologs

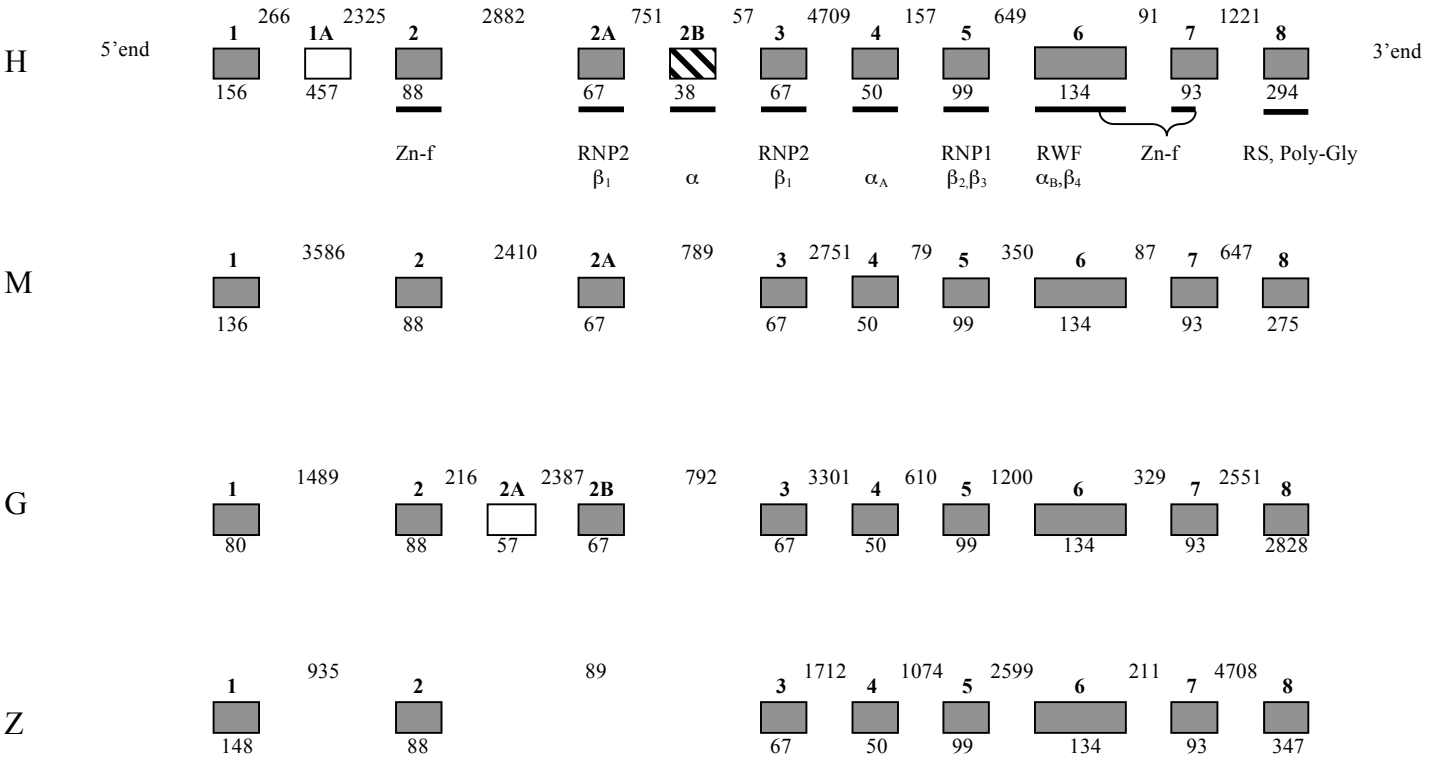


Figure 3.6 - B. U2AF^{26}(U2AF1L4) orthologs

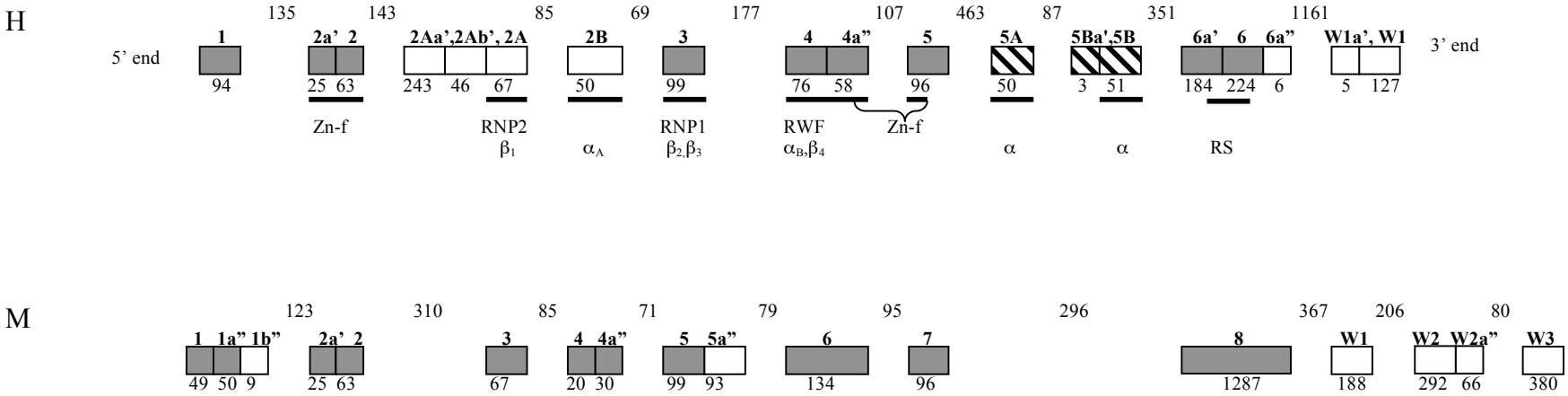
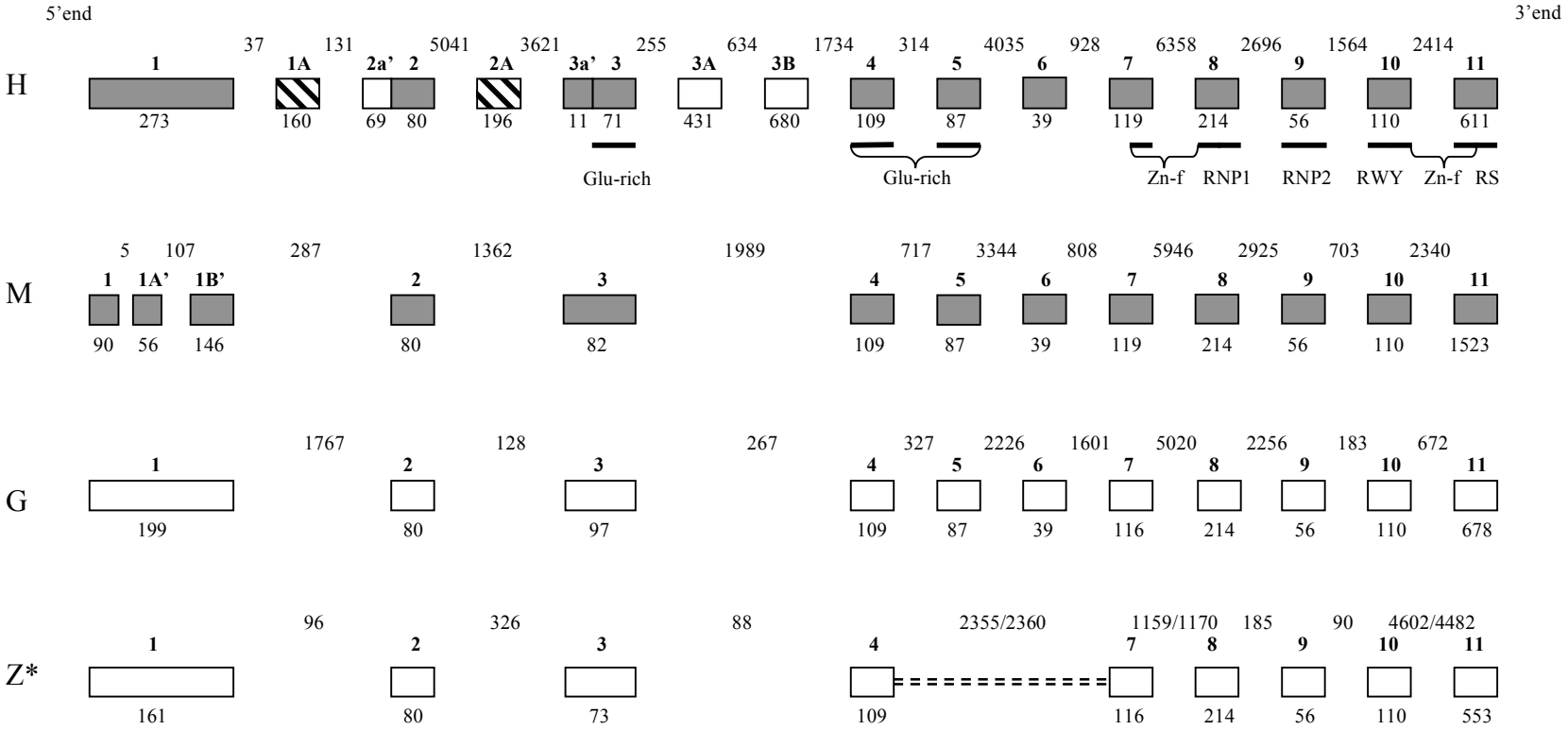


Figure 3.6 - C. U2AF^{35}-RS2 (ZRSR2) orthologs



Validation of novel U2AF^{35} transcripts

In U2AF^{35} there are two transcripts containing exons not present in mouse, chicken or zebrafish. These are: an EST containing exon 2B, which contains exon 1 through to exon 8 (1,2,2B,4,5,6,7,8), with GenBank accession BG612658.1; and another EST containing a novel first exon 1A, which only goes as far as exon 4 (1A,2,3,4), with Genbank accession BE736536.1.

The former transcript (BG612658.1), containing exon 2B, was experimentaly validated by members of Carmo Fonseca's Lab (Figure 3.11). The open reading frame spans exons 2 to 8, but known exon 2 is in a new reading frame. Figure 3.7.A shows a secondary structure prediction of this novel region which is predicted to fold into a helical motif covering 13 amino acids. This transcript is therefore expected to generate a new U2AF^{35} isoform with a new N-terminal motif.

The latter transcript (BE736536.1), containing novel first exon 1A, was found to have no open reading frames in frame with known exons. Exon 1A is also rather large, covering 437 nucleotides, (that is about half the size of the full U2AF^{35} transcript) and the transcript reveals splicing of exon 1A to known exons 2, 3 and 4. This kind of exon falls into the large category of novel non-coding exons (ncExons) examined in the previous section (see Table 3.2). There are 6 start codons on this exon with corresponding premature stop codons on the same exon. It is therefore not likely to produce a novel isoform of U2AF^{35}. However, this exon contains two predicted overlaping open reading frames (orfs) with upstream start codons only 4 nucleotides apart , which would code for 93 and 92 amino acid respectively (Figure 2.7. B). A secondary structure analysis of the translated orfs reveals one with three alpha helices and the other with one extended alpha helix covering 34 amino acids. Exon 1A was found to be conserved in two other primates (chimp and rhesus) but not on mouse. Figure 3.7.C shows the extent of conservation of the translation of the first open reading frame. This analysis indicates that this exon may be unique to primates. Although this transcript may not produce a novel U2AF^{35} isoform it may be involved in a regulatory pathway which is unique to primates, or it may code for a small protein unique to primates as there is a poly(A) signal hexamer (AATAAA) on exon 2.

Figure 3.7. A. Secondary structure prediction of amino acid sequences obtained from open reading frame on the U2AF^{35} transcript containing the novel exon 2B (BG612658). Spliced exons validated by RT-PCR are underlined (see Figure 3.11). The open reading frame spans exons 2 to 8 (highlighted in bold green), but known exon 2 is in a new reading frame. **B.** Secondary structure prediction of amino acid sequences obtained from open reading frame on the U2AF35 transcript containing the novel first exon 1A (BE736536). The sequence shown is extended upstream to the first methionine residues found. Two open reading frames exists that cover exon 1A: the first contains three short α-helices and the second one extended α-helix covering 34 amino acids. Amino acid secondary structure predictions were made with Scratch Protein Predictor SSpro8 [Pollastri *et al.* 2002; Cheng *et al.* 2005; W7] and coloured and annotated manually; H=alpha-helix, E=extended strand, T=Beta turn, S=Bend region, C=random coil. **C.** Conservation in primates (human, chimp and rhesus) of the translation of the 1st orf derived from human exon 1A of U2AF^{35} present in EST BE736536. The alignment was performed using the program MULTALIN [Corpet 1988**]**.

A. Variant d (BG612658.1, EST) Exons: 1,**2,2B,4,5,6,7,8**

```
(exon2)METGALGCTINRRLAS(exon2B)TESQLKHLPEAAG(exon4).....known orf
(exon2)CCCHCEEEEHHHHHHH(exon2B)HHHHHHTCTTTCC(exon4).....known orf
```

B. (BE736536.1, EST) Exons: **1A**,2,3,4

1st orf

```
MNGVSFRGAAAELSPALG(exon1A)SGLADRPTSGPPSRRLPEAASSRAESCLEHPGLLDVDTFVPTFGCECPVEQALVPASPVFSRLSN
CCCEEEHHHHHHHHCHHTT(exon1A)TTCCCSCCCSCCCCCCCHHHHHHHHHHHHCTTCCCCCEECCCECCCCCHHEEECCCCHHHHHHHT

LGS(STOP)
TCC(STOP)
```

2nd orf

```
MGSLFGAPRPSFLRPWG(exon1A)PDWLIARRPGLRVADSRKPPRAALKVAWSIRAFWTWTRLFQLSAVNVLSSKLWYQPRQYSVAY
CCCETCCCCCCGCCCTC(exon1A)CHEHHEECTTCEEECTSSCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHECCCCEEEEE

LT(STOP)
EC(STOP)
```

C.

```
                    10        20        30        40        50        60
                     |         |         |         |         |         |
Human          MNGVSFRGAAAELSPALGSGLADRPTSGPPSRRLPEAASSRAESCLEHPGLLDVDTFVPT
Chimp          MNGASFRGAAAELSPALGSGLADRPTSGPPSRRLPEAASSRAESCLEHPGLLDVDTFAPT
Rhesus         MNGASFRGAAAELSPALRPGLADRPTSGPPSRQLPEAASSDAESCLEHPGLFDVDTFTPT
Consensus      MNGaSFRGAAAELSPALgsGLADRPTSGPPSRrLPEAASSrAESCLEHPGLlDVDTF PT
Prim.cons.     MNGASFRGAAAELSPALGSGLADRPTSGPPSRRLPEAASSRAESCLEHPGLLDVDTF3PT

                    70        80
                     |         |
Humanx0        FGCECPVEQALVPASPVFSRLSNLGS
Chimpx1        FGCECPVEQALVPASPVFSRLSNLGS
Rhesus         LSCECPVSNKLWYQPCQYSVTYLT--
Consensus      fgCECPVe#aLvpaspv%Srlsnlgs
Prim.cons.     FGCECPVEQALVPASPVFSRLSNLGS
```

Validation of novel U2AF^{26} transcripts

The U2AF^{26} gene may be unique to mammals as it has been found in rat, cow and pig as well mouse and human but there is no evidence of it in the genomes of birds, amphibians or fish [Mollet *et al.* 2006]; it is thought to have occurred as a duplication of U2AF^{35} when the mammalian lineage was already established. However, in human, the annotated gene for U2AF^{26} lacks exons 3 and 4, which are present in the mouse ortholog. Exon 3 contains the RNP2 motif and exon 4 contains a large part of an alpha helix A, both these motifs forming part of the canonical UHM motif in the mouse orthologue. The crystal structure of the U2AF^{65}- U2AF^{35} heterodimer has revealed that, when phosphorylated, this helix can present a highly negatively charged face which could provide for interaction with other splicing factors [Kielkopf *et al.* 2001]. The absence of these two exons in the known human isoform is therefore certain to affect RNA-protein interactions. ExonMine does however detect these "missing exons" shown in Figure 3.6B as exons 2A and 2B for the human gene. The length of exons 2A and 2B in the human is identical to the mouse U2AF26 ortholog and the U2AF35 homolog and the translated sequence contains the expected protein motifs as indicated on Figure 3.6B and Figure 3.8. The transcripts detected by ExonMine, however, are either incomplete (Appendix F) or include splicing to 5' extentions to exon 2A which are not seen in any other ortholog or homolog, and it would be important to validate alternative splicing to these exons.

Figure 3.8 – Variation in RNP2 motif in U2AF35 and U2AF26 (RNP2 motif highlighted in Red)

```
human U2AF35 (exon 3)     TIALLNIYRNPQNSSQSADGLR
                          ||   ||||||||| | |||
human U2AF35 (exon 2B)    TILIQNIYRNPQNSAQTADGSH
                          ||   |||||||| ||||||||
human U2AF26 (exon 2A)    TIVLLNLYRNPQNTAQTADGSH
                          ||||||||||||||||||||||
mouse U2AF26 (exon 3)     TIVLLNLYRNPQNTAQTADGSH
```

In addition to this difference between mouse and human U2AF26, in the human, the RefSeq annotation describes a 3' extention of exon 4 (4a") which changes the reading frame of the 3' end of the transcript which will now code for the expected second zinc

finger and teminal RS domain not present in the translation of the transcript skipping 4a" (see Appendix F for details).

Two new exons, 5A and 5B, not found in mouse, were detected in human. Exon 5B has an alternative 5' extention of 3 nucleotides which is indicated as 5Ba'. Two splicing patterns containing these exons were confirmed by RT-PCR (variants 3 and 4 in Figure 3.11). ExonMine tables containing information on open reading frames show that variant 3 has its first start codon at position 29 with a good neighboring Kozak consensus **A**GU**AUGG** but a stop at position 47. However there is a second start codon at position 46 with a neighbouring Kozak consensus **G**AG**AUGA,** and an open reading frame extending all the way to exon 6a' (Figure 3.10 A). The predicted secondary structure of the translation of this reading frame reveals an extended alpha helix covering the novel exons 5A and 5B. A search for similar protein sequences with Mpsrch [http://www.ebi.ac.uk/MPsrch/] reveals an interesting match to serine threonine-protein kinase for exon 5A (Figure 3.9). Variant 4 (Figure 3.11) contains the extention 5Ba' which introduces a stop codon just before exon 5B (Figure 3.10B), otherwise the translation is as for variant 3. The putative proteins are very short and may not be real, however, validation by RT-PCR confirms that these splicing patterns exist and these isoforms should be considered, if not for novel protein fuction, for involvement in the regulation of the expression of this gene.

Figure 3.9 – Translation of U2AF26 Exon 5A (highlighted cian) matches a Serine/threonine kinase. Search made with Mpsrch [W9].

```
ID   NEK5_HUMAN              Reviewed;         708 AA.
DE   Serine/threonine-protein kinase Nek5 (EC 2.7.11.1) (NimA-related
DE   protein kinase 5).

  DB  1;  Score     115;  Match 81.2%;  QryMatch 39.8%;  Pred. No. 2.36e-05;
  Matches    13;  Conservative    3;  Mismatches   0;  Indels   0;  Gaps   0;

          ***.*.***.******
Db    534 WSAMARSWLTATSASQ 549
Qy      1 wsakaqswlsatsasq 16
```

Figure 3.10 - Secondary structure prediction of amino acid sequences obtained from open reading frame on the U2AF[26] transcript containing the novel exons 5A, 5Ba'and 5B. Exons included in open reading frame are highlighted in bold green. Spliced exons validaded by RT-PCR are underlined (see Figure 3.10). **A**. Variant 3 (BU608847.1). **B.** Variant 4 (DB338076.1). Amino acid secondary structure predictions were made with Scratch Protein Predictor SSpro8 [Pollastri *et al.* 2002; Cheng *et al.* 2005; W7] and coloured and annotated manually; H=α-helix, E=extended strand, T=Beta turn, S=Bend region, C=random coil.

A. Variant 3 (BU608847.1)

Exons: 2Ab',2A,**3,4,5,5A,5B,6a'**,6 **(129 amino acids)**

```
(3)MNVCDNLGDHLVGNVYVK(4)FRREEDGERAVAELSNRWFNGQAVHG(5)NVPEVASATSCICGPFPRTSRGSSMGGDPGAG
(3)CCEEHHHHHHEEEEEEEE(4)EEECTTHHHHHHHHTTCEETSCEEEE(5)EECCCCCHHHHECCCCCTCCCCCCCTCCTTCC

(5A)WSAKAQSWLSATSASQR(5B)WGFTTLASLVSNSGPQV(6a')TPEVPYWPPSPREEPSVFP(STOP)
(5A)HHHHHHHHHHHHHHHHH(5B)HTHHHHHHHHHTTSCCC(6a')CCCCCCCCCCCCCCCCCCC(STOP)
```

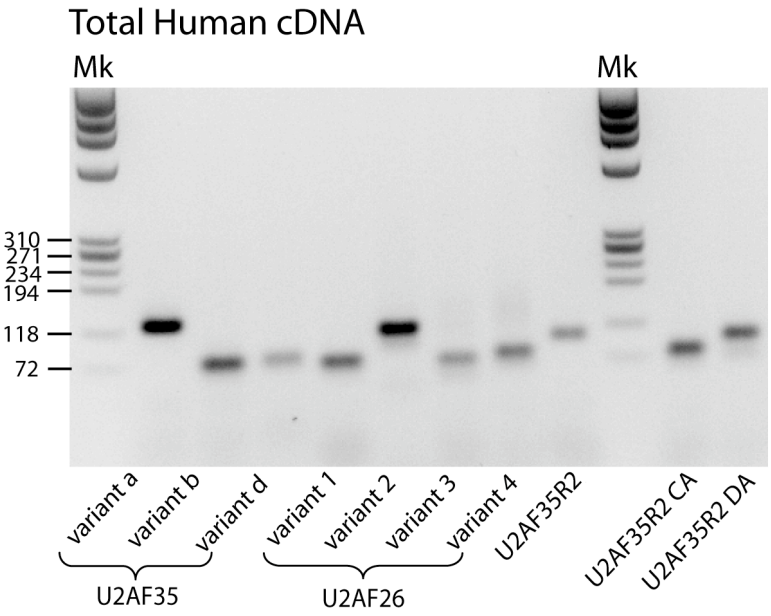
B. Variant 4 (DB338076.1)

Exons: 2A,**3,4,5,5A,5Ba',**5B,6 **(92 amino acids)**

```
(3)MNVCDNLGDHLVGNVYVK(4)FRREEDGERAVAELSNRWFNGQAVHG(5)NVPEVASATSCICGPFPRTSRGSSMGGDPGAG
(3)CCEEHHHHHHEEEEEEEE(4)EEECTTHHHHHHHHTTCEETSCEEEE(5)EECCCCCHHHHECCCCCTCCCCCCCTCCTTCC

(5A)WSAKAQSWLSATSASQ(5Ba')(STOP)
(5A)HHHHHHHHHHHHHCCC(5Ba')(STOP)
```

Figure 3.11. **A**. Total human cDNA of known and predicted U2AF^{35}(U2AF1), U2AF^{26}(U2AF1L4) and U2AF^{35}-RS2(ZRSR2) alternatively spliced mRNA isoforms validated by quantitative real time reverse transcriptase-PCR. Variants (RT-PCR covers bold underlined exons): U2AF^{35} variant a **(**NM_006758.2 has exons 1,**2,3,4,5**,6,7,8**),** U2AF^{35} variant b **(**NM_001025203.1 has exons 1,2,**2A,4,5**,6,7,8**)** and U2AF^{35} variant d **(**BG612658.1 has exons 1,2,**2B,4,5**,6,7,8**);** U2AF^{26} variant 1 (NM_001040425.1 has exons 1,2a',2,3,4,**4a",5**,6a',6), U2AF^{26} variant 2 (NM_144987.2 has exons 1,2a',2,3,**4,5,6a'**,6), U2AF^{26} variant 3 (BU608847.1 has exons 2Ab',2A,3,4,5,**5A,5B,6a'**,6) and U2AF^{26} variant 4 (DB338076.1 has exons 2A,3,4,5,**5A,5Ba',5B**,6); U2AF^{35}R2 (NM_005089.2 has exons **1,2,3a'**,3,4,5,6,7,8,9,10,11), U2AF^{35}R2 CA (DA261525.1 has exons **1A**,2,3,4,5,7), U2AF^{35}R2 DA (CA425173.1 has exons **2A**,3a',3,4). The experimental validation was performed by Joana Borlido, methods detailed in Appendix G.



Validation of novel U2AF^{35}-RS2 transcripts

For U2AF^{35}-RS2 the expression of two exons, of 160 and 196 nucleotides, which do not have any open reading frames in frame with known exons, were validated by RT-PCR by Joano Borlido (variants indicated as DA and CA in Figure 3.11). These were found in transcripts DA261525.1 with exons (**1A**,2,3,4,5,7), and CA425173.1 with exons (**2A**,3a',3,4), both of which contain a putative poly(A) signal on exon 4 (Figure 3.14 and Figure 3.15). ExonMine data on open reading frames shows that the first two open reading frames on these transcripts contain premature stop codons. The putative peptides would have 16 and 55 amino acids for the former, and 12 and 35 amino acids for the latter, which is indeed very short and are not likely to be real.

Using BLAT we find that Exon 1A maps well to the Chimp, Orangutan and Rhesus genome assemblies on orthologous gene U2AF35-RS2 (Figure 3.12), but does not map to the orthologous gene on Mouse, or other mammal model organisms. Similarly Exon 2A maps to Orangutan and Rhesus ortholog of U2AF35-RS2 (Figure 3.13) but a match was not found on the chimp ortholog although matches were found on other genes. This is an indication that this exon may be a mobile element [Kazazian 2004]. Mobile elements are transposable sequences which have shaped the evolution of genes and genomes of both plants and animals. A match for exon 2A was not found on ortholog of U2AF35-RS2 in Mouse or other mammal model organisms.

Figure 3.12 – BLAT alignment of Exon1A from Human, to **Chimp** and **Rhesus** ortholog of U2AF35-RS2 gene. Matches shown in blue.

Exon 1A (DA261525.1) mapping to **Chimp** Genomic chrX on gene U2AF35-RS2 :

```
ccgagaaacc aaggtaagct ccgtacgggg agatgagcag gcgagcgggc  15773473
CGATCGTGGG CATGGAGGAT GTGGTGGCAG GAAGAAAAGA GAAGCTGGGC  15773523
CAAGCTAAGT GAGCCACCGC GGCAGCTCCA TTCCTTCCTG GTGCGCGCCC  15773573
TGGTGCCCTC CCCGGACTTG CACGGGCGGC TGTGGCTCAG GCAAGCGAGG  15773623
AGAGGATCAG gttggcggat ggagaagaac tgtctgggag tgggaggatg  15773673
```

Exon 1A (DA261525.1) mapping to **Orangutan** Genomic chrX on gene U2AF35-RS2 :

```
cagagaaacc aaggtaagct ccgtacgggg agatgagcag gcaagtgggc  15760311
CGATCGTGGG CATGGAGGAT GTGGTGGCAG GAAGAAAAGA GAAGCTGGGC  15760361
CAAGCTAAGT GAGCCACCGC GGCAGCTCCA TTCCTTCCTG GTGCGCGCcC  15760411
TGGTGCCCTC CCCGGACTTG CACGGGCGGC TGTGGgCTCA GGCAAGCGAG  15760461
GAGAGGATCA Ggttggcgga tggagaagaa ctgtctggga gtgggaggat  15760511
```

Exon 1A (DA261525.1) mapping to **Rhesus** Genomic chrX on gene U2AF35-RS2:

```
cagagaaacc aaggtaagct ccgtacgggg agatgagcag gcaagcgggc  13550981
CGATCGTGGG CATGGAGGAT GTGGTGGCAG cAAGAAAAGg GAAGCTaGGC  13551031
CAAGCTAAGT GAGCCACCGC GGCAGCTCgA TTCCTTCCTG GTGCGCGCCC  13551081
TGGTcCCCTC CCCGGACTTt CACGGGCGGC TGTGGgCTCA GGCAAGCGAG  13551131

GAGAGGATCA Ggttggcgga tggagaagaa ctgccgtctg gaagtcggag  13551181
```

Figure 3.13 – BLAT alignment of Exon 2A, from Human, to **Orangutan** and **Rhesus** ortholog of U2AF35-RS2 gene. Matches shown in blue.

Exon 2A (CA425173.1) mapping to **Orangutan** Genomic chrX on gene U2AF35-RS2 :

```
tcaatgattt tctttttgt tttgttttgt tttgtttttt ttgagatgga  15765809
GTCTTGCTCT GTCGCCCAGG CTGGAGTGCA GTGGCATGAA CTCGGCTCAC  15765859
TGCAACCTCC GCCTaCCGGG TTcAAGCAAT TCTCCTGCCT CAGCCTCCAG  15765909
AGTAGcTGGG ATTACAGACA CGCACCACCG CGCCCaGCCA TCATGACTTT  15765959
TCTCTGCTTC TTGAGAGCAC TTCCcGCATC aCTAGTCGCA CTTTGTgtga  15766009
gtctcatgat gttattcaag gtttaccatt ttgtagtaaa cacaatgaaa  15766059
```

Exon 2A (CA425173.1) mapping to **Rhesus** Genomic chrX on gene U2AF35-RS2 :

```
ctcaatgatt ttcttttctt tctttctttt attttatttt ttgagacaga  13556085
GTCTTGCTCT GTtGCCCAGG CTGGAGTGCA GTGGCATGAA CTCGGCTCAC  13556135
TGCAACCTCC GCCTCCCaGG TTcAAGCAgT TCTCCTGCCT CAGCCTCCct  13556185
AGTAGcTGGa ATTACAGgCg tGCACCACac tGCCCGGtCA TCATGACTTT  13556235
TCTCTGCTTC TTGAGAGCAC TTCCAGCATC aCTAGTtgca ttttgtgtga  13556285
gtctcatgat gttattcaag gtttaccatt ttgtagtaaa cacgatgaaa  13556335
```

To explore the RNA sequence of these exons, a secondary structure analysis was peformed using the sfold program [W8]. Exon 1A is predicted to form a secondary structure containing a hairpin which could code for a microRNA (Figure 3.14). Based on this hypothesis, Francisco Enguita from Carmo-Fonseca's Lab attempted to validate this microRNA and preliminary experimental results showed that this exon may in fact be coding for a microRNA.

The transcript carrying exon 2A is also predicted to form a stable hairpin, but in this case the region which folds into a hairpin covers a spliced junction (Figure 3.15). This hairpin was not validated experimentally as a microRNA coding sequence.

The analysis of novel exon 1A and 2A in only preliminary but it opens a new avenue to explore the possibility that other novel exons in known protein coding genes, which do not contain open reading frames covering other known exons in the gene, may be coding for microRNAs, and that some may be unique to the primate lineage.

Figure 3.14 – A. U2AF35-RS2 spliced exons **1A**,2,3,4,5,7 (DA261525.1) and putative Poly(A) signal **AATAAA** on exon 4. Sequence forming a stable hairpin highlighted in red covers nucleotides 14 to 102 of the transcript. **B.** RNA secondary structure, of nucleotides 1 to 160nt of transcript, produced by sfold program [W8]. Stable hairpin circled in red.

A.

CGATCGTGGGCATGGAGGATGTGGTGGCAGGAAGAAAAGAGAAGCTGGGC
CAAGCTAAGTGAGCCACCGCGGCAGCTCCATTCCTTCCTGGTGCGCGCTC
TGGTGCCCTCCCCGGACTTGCACGGGCGGCTGTGGCTCAGGCAAGCGAGG
AGAGGATCAGCCACAAAAAGTACAGGGCCGCCCTGAAGAAGGAGAAACGA
AAGAAACGTCGGCAGGAACTTGCTCGACTGAGAGACTCAGAAGGAGGAAG
AGGAGGACACTTTTATTGAAGAACAACAACTAGAAGAAGAGAAGCTATTG
GAAAGAGAGAGGCAAAGATTACATGAGGAGTGGTTGCTAAGAGAGCAGAA
GGCACAAGAAGAATTCAGAATAAAGAAGGAAAAGGAAGAGGCGGCTAAAA
AACGGCAAGAAGAACAAGAGAGAAAGTTAAAGGAACAATGGGAAGAACAG
CAGAGGAAAGAGAGAGAAGAGGAGGAGCAGAAACGACAGGAGAAGAAAGA
AAAAGAGTTGGAAAATGGTACCACATGGCAAAACCCAGAACCACCCGTGG
ATTTCAGAGTAATGGAGAAGGATCGAGCTAATTGTCCCTTCTACAGTAAA
ACAGGAGCTTGCAGATTTGGAGATAG

B. 14-102nt (putative microRNA hairpin)

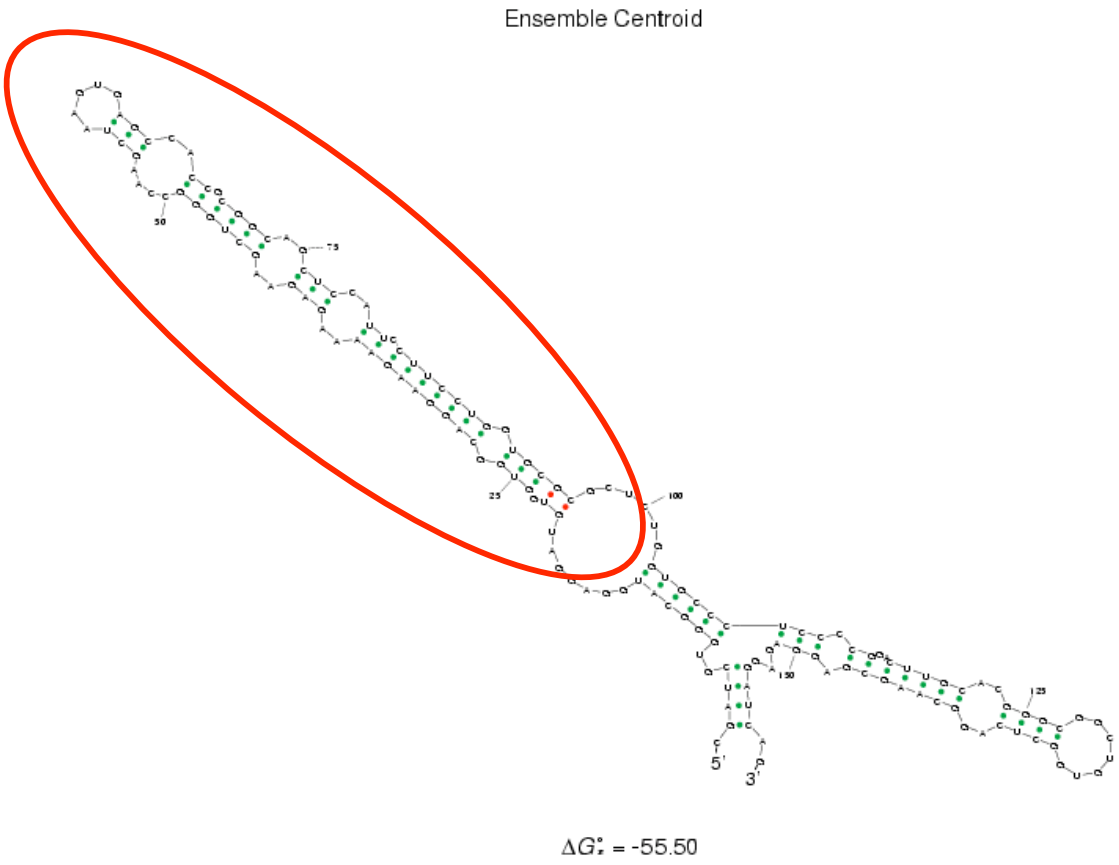
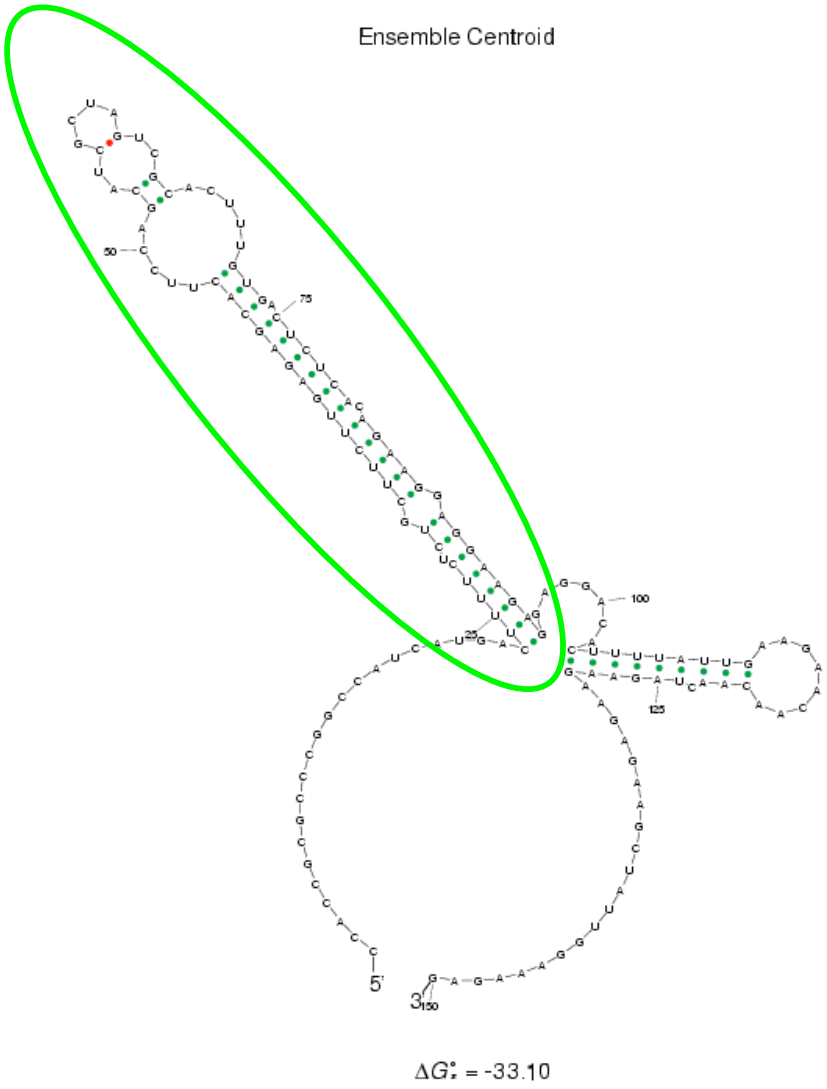


Figure 3.15 – **A.** U2AF35-RS2 spliced exons **2A**,**3a'**,**3**,**4** (CA425173.1, EST) and putative Poly(A) signal **AATAAA** on exon 4. Sequence forming a stable hairpin highlighted in green covers nucleotides 145 to 221 of the transcript. **B.** RNA secondary structure nucleotides 125 to 225nt of transcript produced by sfold program [W8]. Stable hairpin circled in green.

A.

GTCTTGCTCTGTCGCCCAGGCTGGAGTGCAGTGGCATGAACTCGGCTCAC
TGCAACCTCCGCCTCCCGGGTTGAAGCAATTCTCCTGCCTCAGCCTCCAG
AGTAGTTGGGATTACAGACACGCACCACCGCGCCCGGCCATCATGACTTT
TCTCTGCTTCTTGAGAGCACTTCCAGCATCGCTAGTCGCACTTTGTGACT
CTCACAGAAGGAGGAAGAGGAGGACACTTTTATTGAAGAACAACAACTAG
AAGAAGAGAAGCTATTGGAAAGAGAGAGGCAAAGATTACATGAGGAGTGG
TTGCTAAGAGAGCAGAAGGCACAAGAAGAATTCAGAATAAAGAAGGAAAA
GGAAGAGGCGGCTAAAAAACGGCAAGAAGAACAAGAG

B. 145-221nt (putative microRNA hairpin)



4. CONCLUSIONS

Novelty

By imposing fewer constraints on the selection of transcripts, ExonMine produces a high level of clustering of spliced EST data to a given gene. The method succeeds in capturing a greater number of transcripts and detecting a higher percentage of spliced genes and isoforms when compared with other recently published alternative splicing databases. ExonMine provides a minimal set of isoforms representing all splicing patterns found but also provides the user with all the raw data for each gene. The discrimination of all exons, present in all input GenBank transcripts used in the clustering, allows the user to backtrack to the original experimental conditions including tissue expression of each transcript. A statistical analysis of our output data also revealed a marked increase, in complex organisms, of splice variants incorporating intron retentions or novel exons with no apparent protein-coding potential. A significant number of these match transcription data from tiling arrays. Several lines of evidence indicate that post-transcriptional regulation through alternative splicing involving non-protein coding transcripts may constitute an important mechanism for the control of gene expression [Lewis *et al.* 2005; Hughes 2006]. Mechanisms of this nature include regulated unproductive splicing and translation (RUST) [Lewis *et al.* 2005] and nonsense mediated decay (NMD) [Lareau *et al.* 2007; Grellscheid and Smith 2006]. Such splice variants may form an integral part of regulatory pathways built into more intricate layers of gene regulation in humans. Deciphering complex pathways of gene regulation in higher organisms involving non-protein-coding splicing patterns will expand our knowledge of the regulatory role of alternatively spliced untranslated regions.

Validation

Validation of data produced by the ExonMine program has confirmed the utility of this resource. The ease with which mining of control sequences, associated with alternative splicing events, was performed to validate experimental models [Pacheco *et al.* 2006] reveals the versatility of ExonMine. Cross species analysis of ExonMine data for the U2AF family of alternative splicing factors has given insight into the evolution of functions brought about by alternative splicing in this set of genes [Mollet *et al.* 2007]. A selection of novel isoforms detected in ExonMine for the U2AF35 family of splicing factors was

validated experimentally by members of Carmo-Fonseca's lab. A transcript from U2AF35, including a novel exon which does not have an open reading frame in-frame with known exons, was found to be conserved in primates but not in mouse. Another non-coding transcript in U2AF35_RS2, which was validated by members of Carmo-Fonseca's lab, was found to contain a potential microRNA hairpin. A recent study done in plants suggest that the number of microRNAs in a system is a crucial factor in determining the complexity of that system as it is associated with more complex expression patterns in the target genes [Takuno and Innan 2008], surely an identical case could be made for animal systems. Our results show that the new information mined by ExonMine is not spurious and seems to confirm our premise that more complex layers of gene regulation, involving apparently non-protein-coding transcripts, exist in humans.

Potential for large-scale analysis

ExonMine data has been used for custom design of an alternative splicing array designed to monitor alterations in alternative splicing in myotonic dystrophy. On the other hand, the good correlation obtained between ExonMine data and both Affymetrix tiling array data and the commercial exon microarray platform, which covers the majority of exons in human, could allow large scale analysis of alternative splicing to be performed. The MySQL structure of the data has also provided a resource for easy large-scale analysis of regulatory sequences elements involved in alternative splicing [Pacheco *et al.* 2006].

Plasticity of algorithm

The algorithm has currently been applied using RefSeq and EntrezGeneIDs as a means to identify the locus and exons of a particular gene, and used a particular level of filtering of EST and mRNA data in order to weed out determination of false alternative splicing events which occur due to incorrect or incomplete BLAT mapping of transcripts to the genome assembly. This algorithm does however present a high level of plasticity and could be applied to any set of gene predictions. The level of filtering of mRNAs and ESTs is also open to easy manipulation at several levels: minimum exon or intron size allowed; number of bases in the cDNA unmapped by the BLAT program due to single nucleotide polymorphism; number of n bases in the mRNA and EST allowed; and the splice site consensus sequences allowed. The algorithm can also be applied to gene predictions when a RefSeq annotation is not available and this was successfully done in an analysis of chicken

and zebrafish orthologs of human U2AF^{35}-RS2 - this provides a means for manoeuvring into new territory.

Cross Species Genomics

Certain mechanisms of gene expression display a high degree of conservation across all organisms. For example, mRNA translation by the ribosome follows a set of rules that is essentially invariant in all known organisms. On the other hand, although many elements of pre-mRNA processing are conserved, the rules governing RNA splicing clearly differ between different groups of eukaryotes. In addition, the rules for splicing appear to be significantly more complex than those for translation, involving presence of multiple degenerate motifs occurring with appropriate spacing in the transcript. Information on alternative splicing in orthologous genes (i.e. identical genes found in different organisms) is a fundamental tool not only for establishing the common ground in regulation of gene expression through alternative splicing across several species but more importantly for singling out what distinguishes more complex organisms from less complex ones and thereby broadening the road to deciphering the intricate layers of gene regulation in human. The cross species analysis and validation of alternative splicing patterns detected by ExonMine in the U2AF35 family of splicing factors revealed elements which seem to be unique to primates.

Updatability

We stress that the objective of ExonMine is not to give a near perfect description of alternative splicing but rather to mine as much alternative splicing information as possible from the vast amounts of transcript sequence deposited in GenBank. Given the rate at which novel sequence data is deposited in GenBank on a daily basis, the issue of easy updatability is a great asset (48 hours for the human data and less time for less complex genomes). On the other hand it is important to provide stable sets of data which can be referenced in the process of validation or ongoing work and a four month gap between updates has been settled on, allowing nonetheless continued access to previous updates.

Future work

The objectives of this project have been attained. The datasets made available through ExonMine are part of an actively growing project which will involve continuous inclusion of novel information relating to alternative splicing, such as splicing control sequences, tissue-specific splicing events and disease-associated splicing defects. For model organisms which have not been extensively covered by RefSeq, or have no such gene annotation, future work will include the application of the ExonMine algorithm by defining gene loci using gene predictions rather than RefSeq transcripts, and ESTs from closely related organisms. Further work is currently being undertaken in several independent projects at the level of microarray analysis and at the level of validation of individual alternative splicing events. It is hoped that this resource will continue to support work in validation and functional analysis of alternative splicing information both on the scale of a single gene and by using genome-wide technologies.

5. APPENDICES

Appendix A - *Rules followed to adjust the ends of transcript BLAT alignments*

Figure I. Rules followed to adjust the ends of transcript BLAT alignments

This figure describes in detail the adjustments made to transcript BLAT alignments taking into account information from all superimposed alignments of mRNAs and ESTs, that is, the rules followed for adjusting the start (cS – cut Start) and the end (cE – cut End) of each alignment to the nearest known start of an exon (S) or end of an exon (E). **Figure I. A.** For a cS lying between a known exon start S and a known exon end E on the same exon the cS is extended to the position of S. When no isoform contains both S and E the program looks at the other cE and cS in the region between S and E: if the first cS is downstream of the last cE then cS is extended to the position of the first cS (Case 1); if the first cS is upstream of the last cE (Case 2) or there are no cE between S and E (Case 3) then cS is extended to the position of S. **B**. For a cS lying between a known exon end E and a known exon start S, the cS is aligned to the position of S if |S-cS|<9nt or to position E otherwise. **C.** For a cS lying between two known exon starts S_1 and S_2 the cS is extended upstream to the position S_1. **D**. For a cS lying between two known exon ends E_1 and E_2: if E_2 has a known start S then cS is extended upstream to the position E_1 otherwise cS is extended to the furthest upstream cS lying between E_1 and E_2 (first cS). **E.** For a cE lying between an S and an E on the same exon the cE is extended to the position of E. When no isoform contains both S and E the program looks at the other cE and cS in the region between S and E: if the first cS is downstream of the last cE, then, cE is extended to the position of the last cE (Case 1); if the first cS is upstream of the last cE (Case 2) or there are no cS in the region between S and E (Case 3) then cE is extended to the position of E. **F**. For a cE lying between an E and an S the cE is aligned to the position of E if |cE-E|<9nt or to position S otherwise. **G**. For a cE lying between two known exon ends E_1 and E_2 the cE is aligned with E_2. **H**. For a cE lying between two known exon starts S_1 and S_2: if S_1 has an known end E then cE in extended to the position of S_2 otherwise cE is extended to the furthest downstream cE lying between S_1 and S_2 (last cE).

Figure I.

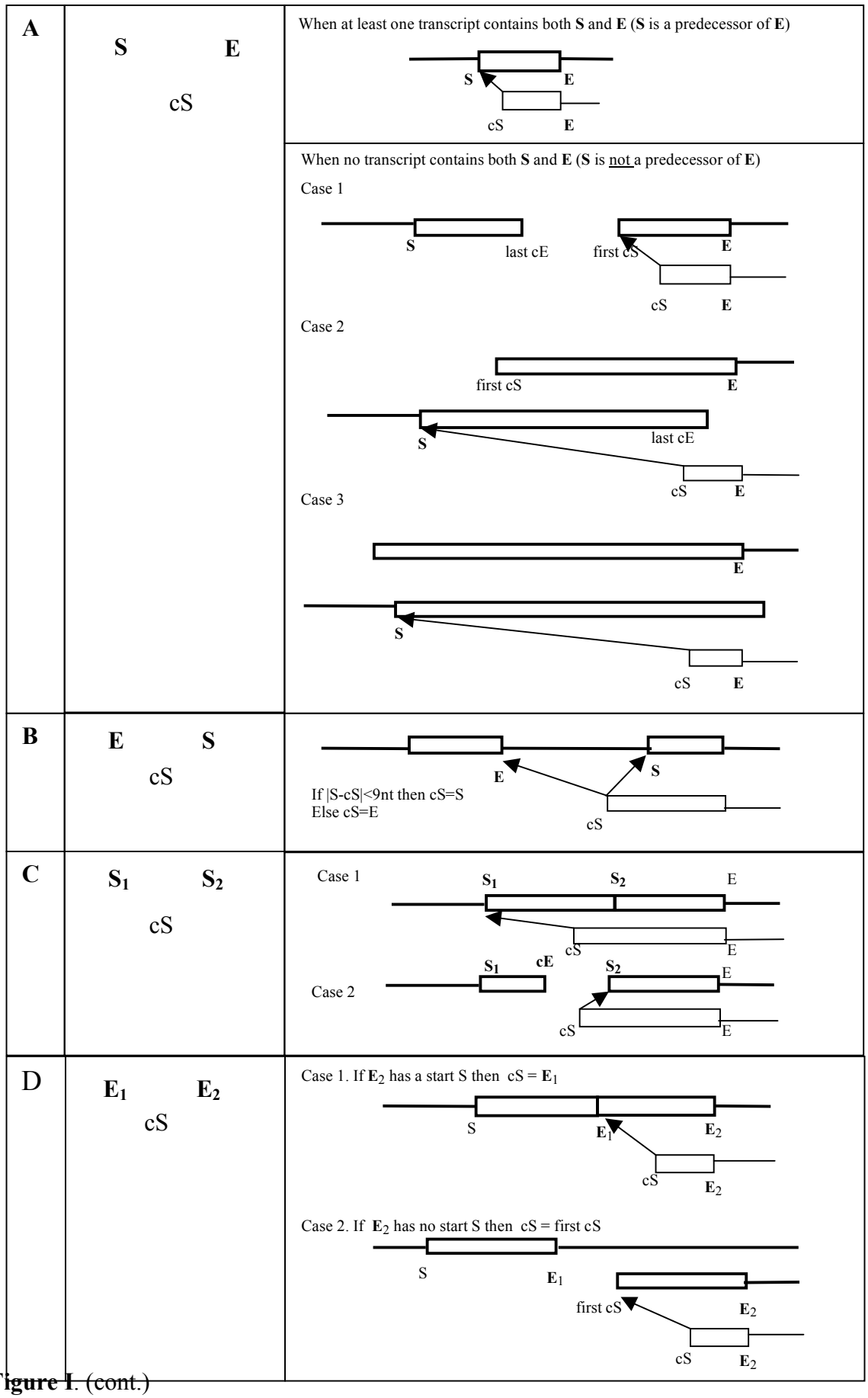
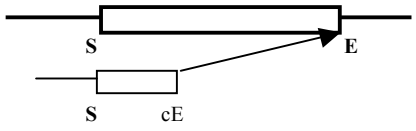
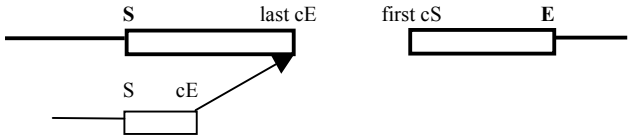
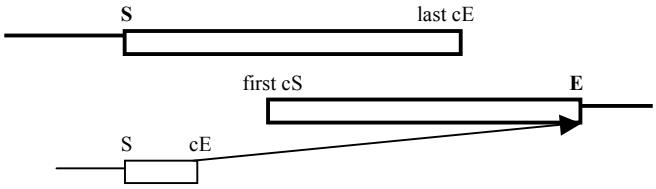
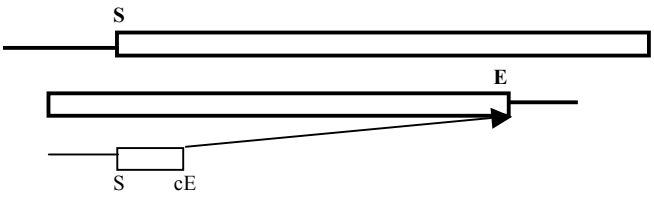
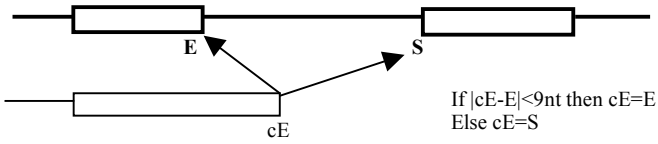


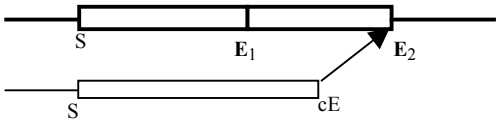
Figure I. (cont.)

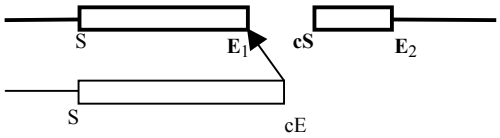
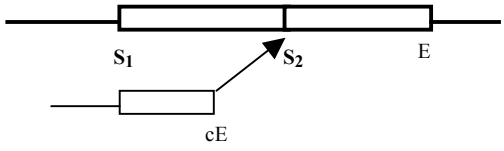
<table>
<tr>
<td>E</td>
<td>S E
cE</td>
<td>When at least one transcript contains both S and E (E is a successor of S)
 

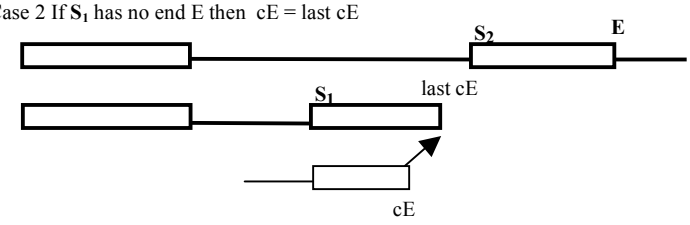
When no transcript contains both S and E (E is <u>not</u> a successor of S)
Case 1
 

Case 2
 

Case 3
 
</td>
</tr>
<tr>
<td>F</td>
<td>E S
cE</td>
<td> 

If |cE-E|<9nt then cE=E
Else cE=S</td>
</tr>
<tr>
<td>G</td>
<td>E₁ E₂
cE</td>
<td>Case 1
 

Case 2
 
</td>
</tr>
<tr>
<td>H</td>
<td>S₁ S₂
cE</td>
<td>Case 1. If S₁ has end E then cE = S₂
 

Case 2 If S₁ has no end E then cE = last cE
 
</td>
</tr>
</table>

Appendix B - *List of candidates for validation*

Table I. Candidate genes containing alternative 3' splice sites (3'ss) carrying a weak proximal polypyrimidine tract and a strong distal polypyrimidine tract and with the alternative exon extention containing a putative exonic splicing enhancer.

Gene Symbol	Chr	Strand	Position 5'ss	Position proximal 3'ss	Position distal 3'ss	Length Alternate Sequence	Accession for Isoform with proximal 3'ss	Accession for Isoform with distal 3'ss
ADA	chr20	-	42682354	42682000	42681902	98	BX414292	NM_000022
downstream of APOC1	chr19	+	50122130	50122707	50122892	185	AV645786	DA638554
ARHGEF19	chr1	-	16276560	16274375	16274341	34	NM_153213	AL137736
ARMCX4	chrX	+	100593144	100594216	100594365	149	NM_152583	CR595436
ATXN7L2	chr1	+	109741881	109742182	109742278	96	AK090460	NM_153340
BCL2L14	chr12	+	12135050	12138712	12138865	153	NM_030766	NM_138723
BCNP1	chr19	+	17504197	17505334	17505396	62	DA951184	NM_173544
BIRC4BP	chr17	+	6599694	6600034	6600085	51	DA430788	BX649188
BMP4	chr14	-	53489744	53488697	53488644	53	NM_001202	BE877424
C18orf8	chr18	+	19364259	19364349	19364473	124	NM_013326	AL526305
C22orf18	chr22	-	40666417	40665893	40665808	85	BM045658	NM_001002876
CAD	chr2	+	27373751	27373850	27373937	87	NM_004341	BG821112
CUEDC1	chr17	-	53298834	53295717	53295645	72	BI335053	NM_017949
DMTF1	chr7	+	86459947	86462006	86462059	53	AB209230	NM_021145
DNAJB13	chr11	+	73354944	73356909	73357038	129	BU859152	NM_153614
ECE1	chr1	-	21307649	21305622	21305460	162	BQ942755	NM_001397
ECM2	chr9	-	92359524	92357040	92356974	66	NM_001393	BC105958
EIF3S7	chr22	-	35232050	35231715	35231547	166	AI302828	NM_003753
FAM24B	chr10	-	124605151	124600056	124600026	30	NM_152644	BU602099
FBXO2	chr1	-	11648354	11647848	11647750	98	DA054667	BI545439
NAT11	chr11	+	63469983	63470773	63470999	226	DA731067	NM_024771
TTLL13	chr15	+	88601958	88602274	88602419	145	NM_001029964	BC052201
FRMD1	chr6	-	168286603	168286365	168286067	298	AK125963	NM_024919
GANC	chr15	+	40390336	40393372	40393570	198	BM541471	AK074330
HDAC7A	chr12	-	46478158	46477549	46477477	72	NM_015401	CR983547
KIAA0100	chr17	-	23984694	23984546	23984474	72	NM_014680	BU158839
KIAA0913	chr10	+	75227369	75227545	75227700	155	NM_015037	AL133662
LAD1	chr1	-	198083859	198083590	198083524	66	BG980318	NM_005558
LEPREL2	chr12	+	6811339	6812854	6813017	163	BI756620	NM_014262
LOC153684	chr5	+	43079426	43080252	43080302	50	BC101382	AK002146
LOC155060	chr7	+	148422515	148423170	148423360	190	AK098351	BC041610
LOC203427	chrX	+	118322546	118325914	118326035	121	AL590525	NM_145305
LOC348180	chr16	+	87305604	87305970	87306183	213	BC063512	NM_001012759
LOC374395	chr11	+	62311575	62313071	62313235	164	NM_199337	BU197210

Table I (continued)

Gene Symbol	Chr	Strand	Position 5'ss	Position proximal 3'ss	Position distal 3'ss	Length Alternate Sequence	Accession for Isoform with proximal 3'ss	Accession for Isoform with distal 3'ss
Upstream of LPIN1	chr2	+	11772216	11772915	11773022	107	CR593993	AK000417
MBNL3	chrX	-	131250410	131246374	131246269	105	NM_018388	NM_133486
MET	chr7	+	115992625	115993342	115993396	54	J02958	NM_000245
LIMD2	chr17	-	59130120	59130030	59129910	120	NM_030576	BE793757
MGC42174	chr2	+	233026089	233026356	233026429	73	CR598716	AL834174
NFRKB	chr11	-	129260540	129259987	129259951	36	NM_006165	BC063280
TTLL13	chr15	+	88601958	88602274	88602419	145	NM_001029964	BC052201
PAX4	chr7	-	126845950	126845853	126845681	172	BC107149	NM_006193
PBEF1	chr7	-	105519938	105519378	105519328	50	DA295479	DA338663
PEA15	chr1	+	156988385	156989891	156989996	105	DA736151	AK095879
PEG3	chr19	-	62029568	62027921	62027694	227	NM_006210	BP219205
PON3	chr7	-	94646136	94641490	94641451	39	AV660101	NM_000940
PPP1R13L	chr19	-	50577625	50575326	50575024	302	NM_006663	BM716707
PTK9	chr12	-	42486291	42486072	42485955	117	BP354520	BC022344
PYGO2	chr1	-	151747023	151746661	151746575	86	DA739729	NM_138300
RAG2	chr11	-	36576228	36573393	36573321	72	DB137123	DB143931
RBPSUHL	chr20	+	43376619	43378241	43378336	95	NM_014276	BI964057
RIMS3	chr1	-	40800159	40782607	40782527	80	NM_014747	BC003103
RNF31	chr14	+	23689937	23690180	23690348	168	NM_017999	AK122711
SEC13L1	chr3	-	10320710	10318058	10317831	227	NM_030673	CR623402
SERPINA6	chr14	-	93859357	93850925	93850757	168	BU935177	NM_001756
SGIP1	chr1	+	66873197	66881977	66882199	222	BM666838	BM689588
SLCO1C1	chr12	+	20740008	20740950	20741032	82	DA345251	DA142363
SLCO4A1	chr20	+	60769707	60769809	60769955	143	NM_016354	CR616855
SMOC2	chr6	+	168762908	168766322	168766355	33	NM_022138	BC047583
ZNF343	chr20	-	2429302	2422690	2422648	42	NM_024325	CD702168
SPATA1	chr1	+	84684122	84684360	84684581	221	AW204186,	BM055177
SPP2	chr2	+	234760667	234767489	234767669	180	NM_006944	AV681865
SRRM2	chr16	+	2761051	2761123	2761371	248	AI376062	AA703497
TCP1	chr6	-	160179186	160177463	160177440	23	BU508560	NM_001008897
TM2D2	chr8	-	38972889	38972470	38972369	101	NM_031940	DA114767
TOR3A	chr1	+	175783179	175783511	175783732	221	BX374780	NM_022371
ZNF124	chr1	-	243661191	243649210	243649156	54	BG257604	NM_003431
ZNF124	chr1	-	243648349	243646746	243646550	186	BC099661	NM_003431
ZNF177	chr19	+	9351815	9352265	9352344	79	U37251	BP260204
ZNF300	chr5	-	150263619	150263138	150262937	201	DA739710	NM_052860

Appendix C - *Matching AffyHumanExon1.0 ST probes to ExonMine exons*

Entry from table matching a Affymetrix GeneChip® Human Exon 1.0 ST Probe Selection Region (PSR). In the case shown below, the region where the probes are designed is 84nt long (segment highlighted in grey below), this is the PSR and the ID of this region is 3231391.

ProbeSetID	chr	strand	Start	End	PSRLength	score	GeneID	SegmentID	SegmentType	Constitutive
3231391	chr10	+	170747	170830	84	900	9976	163579	3e	A

Entry from Table **Exon** (**ExonID=**163579) matching the above Affymetrix GeneChip® Human Exon 1.0 ST Probe (**ProbeSetID =** 3231391)

ExonID	GeneID	ExonNum	RefExonNum	Constitutive	ExonType	Known	Length	chr	strand	Start	End
163579	9976	2a"	1a"	A	3e	Known	106	chr10	+	170727	170832

Affymetrix designs one to four 25nt probes in the this region. The probes for Probe Selection Region (PSR) 3231391 of Human Exon 1.0 ST (design date Dec 2004) the probes are

Probe
TCCACACATGCACTCCGTATCCACA
TGTTGGAGCATATTCCTCTGTCTCC
CCTCTATCCGAAATTTGTTGGAGCA
TCCCCACGTCCTGTCTTCGGCGCCT

Here the probe sequence is the antisense strand

Reverse Complement sequences for ProbeSetID=3231391

TGTGGATACGGAGTGCATGTGTGGA
GGAGACAGAGGAATATGCTCCAACA
TGCTCCAACAAATTTCGGATAGAGG
AGGCGCCGAAGACAGGACGTGGGGA

Exon Sequence for ExonID=163579

GTACGTAGCCTTCACGTGTGTGTGGATACGGAGTGCATGTGTGGAGACAGAGGAATATGCTCCAACAAATTTCGGATAGAGGCGCCGAAGACAGGACGTGGGGAAG

TGTGGATACGGAGTGCATGTGTGGA TGCTCCAACAAATTTCGGATAGAGG

GGAGACAGAGGAATATGCTCCAACA AGGCGCCGAAGACAGGACGTGGGGA

Appendix D - *Full set of data supporting Table 3.6*

Figure II - Predicted mRNA isoforms for genes encoding the U2AF family of proteins. For each gene, a table indicates GenBank accessions and corresponding exons according to BLAT mappings; chromosome positions are for genome assembly hg17. The gene names are followed by official gene symbol in parenthesis used by Entrez Gene. Confirmed mRNA isoforms are highlighted in yellow; the count of predicted splicing patterns producing a premature stop codon are highlighted in cian; the count of predicted splicing patterns of candidates for putative novel protein are highlighted in magenta. Transcripts carrying extensive intron retentions with no open reading frames were excluded. Coding sequence is indicated as 'complete' if a complete known protein is reported in RefSeq or SwissProt databases; or 'putative' for other mRNAs or ESTs where an in-frame coding sequence was found. Unknown open reading frames are also indicated. The exon on which the Start or Stop codon is found is indicated with the respective genomic position. A diagram of the exon structure of each gene is depicted below the corresponding table. Numbers below the exons represent exon length in nucleotides, and numbers above represent the intron length in nucleotides; junction consensus other than GT_AG is indicated; also indicated are the exons which code for known protein domains.

Abreviations: cds – coding sequence, aa – amino acids, mRNA – messenger RNA, EST – expressed sequence tag, RRM – RNA recognition motif, Zn-finger – zinc finger domain

Position of first nucleotide of Start codon.

Position of first nucleotide after Stop codon

Table II.A

Protein (Gene symbol) Chr position Strand	Order of exons in isoform (orf in bold)		Accession	Status of cds (length of cds in aa), protein Accession	Start codon Exon (chr postion)#	Stop codon Exon (chr position)##
U2AF[65] (U2AF2) chr19: 60857188-60877934 '+' strand	**1,2,3,4,5,6,7,8,9,10,10a",11,12,12A***,12A'		NM_007279.2 (RefSeq)	complete (475aa), NP_009210.1, P26368	1 (60858282)	12A* (60877246)
	1,2,3,4,5,6,7,8,9,10,11,12,12A*,12A'		NM_001012478.1 (RefSeq)	complete (471aa), NP_001012496.1, Q96HC5	1 (60858282)	12A* (60877246)
	9,10,11,12,12A'	1	CR609498.1 (mRNA)	putative fragment (191aa)	no start	no stop
	1,1A,2,3,4,5,6	1	CD624005.1 (EST)	premature stop (52aa)	1 (60858282)	1A (60862149)
	1,1A,2,3,4,5,5A	(1)	CR982513.1 (EST)	premature stop (52aa)	1 (60858282)	1A (60862149)
	5A,**6,7,8,9**	2	BI909492.1 (EST)	putative fragment (151aa)	6 (60865685)	no stop
	3,4,5,5A,6,7	2	CA488904.1 (EST)	premature stop (100aa)	no start	5A (60864951)

Figure II.A

U2AF65 (U2AF2)

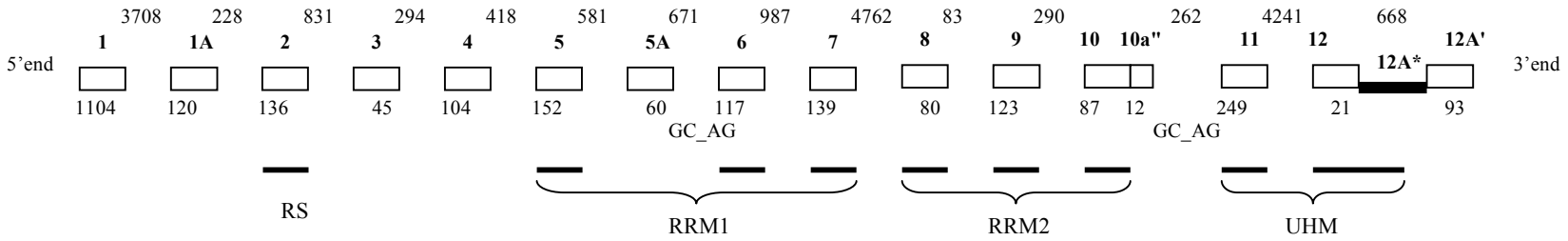


Table II.B

Protein (Gene symbol) Chr position Strand	Order of exons in isoform (orf in bold)	Accession	Status of cds (length of cds in aa), protein Accession	Start codon Exon (chr postion)#	Stop codon Exon (chr position)##
PUF60 (SIAHBP1) chr8:144,970,535-144,983,525 '-' strand	**1,2a',2,3,4,5,6,7a',7,8,9,10,11**	NM_014281.3 (RefSeq)	complete (542aa), NP_055096.2	1 (144983461)	11 (144970677)
	1,2a',2,**3,4,5,6,7a',7,8,9,10,11**	idem	complete (499aa), Q96D94	3 (144976053)	11 (144970677)
	1,2a',2,3,4,4A,5,6,7a',7,8,9,10,11	NM_078480.1 (RefSeq)	complete (559aa), NP_510965.1	1 (144983461)	11 (144970677)
	1,2a',2,**3,4,4A,5,6,7a',7,8,9,10,11**	idem	complete (516aa), Q969E7	3 (144976053)	11 (144970677)
	1A,2a',2,**3,4,5,6,7a',7,8,9,10,11**	BC009734.1 (mRNA)	complete (499aa), Q96D94	3 (144976053)	11 (144970677)
	1A,2a',2,**3,4,4A,5,6,7a',7,8,9,10,11**	BC011265.1 (mRNA)	complete (516aa), Q969E7	3 (144976053)	11 (144970677)
	1,2,3,4,5,6,7a',7,8 1	BI915396.1 (EST)	putative fragment (318aa)	1 (144983461)	no stop
	1,2,3,4,4A,5,6,7a',7,8 2	AL522753.3 (EST)	putative fragment (335aa)	1 (144983461)	no stop
	1A,2,**3,4,5,6,7a',7,8** **known cds**	BX397429.2 (EST)	putative fragment (276aa)	3 (144976053)	no stop
	1,3,4,5,6,7a',7,8,9 3	AL514886.3 (EST)	putative fragment (335aa)	1 (144983461)	no stop
	1,3,4,4A,5,6,7a',7,8 4	BX384203.2 (EST)	putative fragment (307aa)	1 (144983461)	no stop
	2A,3,4,4A,5,6,7a',7,8,9,10 5	AK055941.1 (mRNA)	putative fragment (449aa)	2A (144976721)	no stop
	1A,**1B,2a',2,3,4,5,6,7a',7** 6	BQ421738.1 (EST)	putative fragment (292aa)	1B (144981677)	no stop
	1A,1C,2a',2,**3,4,4A,5,6,7a',7 known cds**	BM558085.1 (EST)	putative fragment (229aa)	3 (144976053)	no stop
	1,**1B,2a',2,3,4,4A,5,6** 7	BQ956878.1 (EST)	putative fragment (238aa)	1B (144981677)	no stop
	1D,2a',2,3,4,4A 8	BG115238.1 (EST)	putative fragment (141aa)	1D (144979054)	no stop
	4B,**5,6,7a'** 9	BE393389.1 (EST)	putative fragment (119aa)	5 (144972692)	no stop
	5,6,6a'',7,8,9 10	BU170641.1 (EST)	putative fragment (270aa)	5 (144972692)	no stop

Figure II.B PUF60 (SIAHBP1)

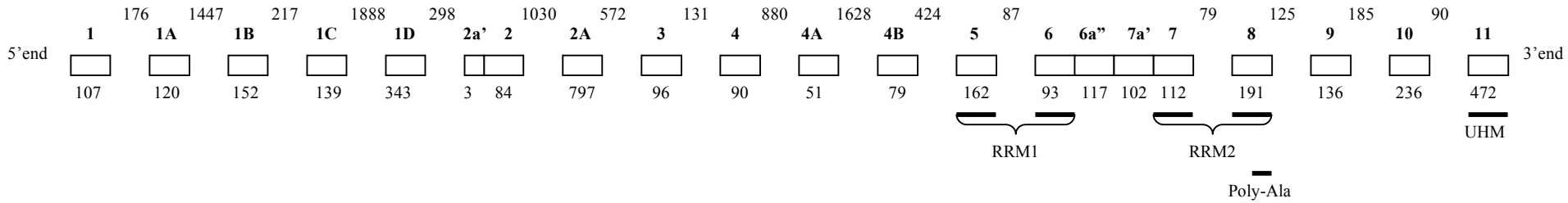


Table II.C

Protein (Gene symbol) Chr position Strand	Order of exons in isoform (orf in bold)		Accession	Status of cds (length of cds in aa), protein Accession	Start codon Exon (chr postion)#	Stop codon Exon (chr position)##
CAPERα (RNPC2) chr20:33,754,944-33,793,607 '-' strand	1,1A*,1A',**2,3,4a',4,5,6a',6,7,8,9,10,11,12,13b',13,14,15,16,17**		NM_184234.1 (RefSeq)	complete (530aa), NP_909122.1, Q14498-1	2 (33792210)	17 (33755816)
	1,1A*,1A',**2,3,4a',4,5,6a',6,7,8,9,10,11,12,13,14,15,16,17**		NM_004902.2 (RefSeq)	complete (524aa), NP_004893.1, Q14498-2	2 (33792210)	17 (33755816)
	1,1A*,1A',**2,2A**,3,4a',4,5,6a',6,7,8,9,10,11,12,13,14,15,16,17	1	NM_184241.1 (RefSeq)	premature stop (40aa)	2 (33792210)	2A (33791861)
	1,1A*,1A',2,2A,3,4a',4,5,6a',6,**7,8,9,10,11,12,13,14,15,16,17**		idem	complete (367aa), NP_909129.1, Q6N037	7 (33776436)	17 (33755816)
	1,1A*,1A',**2,2A**,3,4a',4,5,6a',6,7,8,9,10,11,12,13b',13,14,15,16,17	(1)	NM_184244.1 (RefSeq)	premature stop (40aa)	2 (33792210)	2A (33791861)
	1,1A*,1A',2,2A,3,4a',4,5,6a',6,**7,8,9,10,11,12,13b',13,14,15,16,17**		idem	complete (373aa), NP_909132.1, Q5QP20	7 (33776436)	17 (33755816)
	1,1A*,1A',**2,3,3C**,4a',4,5,6a',6,7,8,9,10,11,12,13b',13,14,15,16,17	2	NM_184237.1 (RefSeq)	premature stop (47aa)	2 (33792210)	3C (33785550)
	1,1A*,1A',2,3,3C,4a',4,5,6a',6,**7,8,9,10,11,12,13b',13,14,15,16,17**		idem	complete (373aa), NP_909125.1	7 (33776436)	17 (33755816)
	1A',**2,3,4a',4,5,6a',6,7,8,9,10,11,12,13,13A,14**,15,16,17	1	AL833168.1 (mRNA)	putative (423aa)	2 (33792210)	14 (33758525)
	1A',**2,2A**,3,4a',4,5,6a',6,7,8,9,9A,10,11,12,13b',13,14,15,16,17	(1)	BC107886.1 (mRNA)	premature stop (40aa)	2 (33792210)	2A (33791861)
	1A',2,2A,3,4a',4,5,6a',6,**7,8,9,9A**,10,11,12,13b',13,14,15,16,17	2	idem	putative (132aa)	7 (33776436)	9A (33771893)
	1A',2,2A,3,4a',4,5,6a',6,7,8,9,9A,**10,11,12,13b',13,14,15,16,17**		idem	putative (249aa)	10 (33768123)	17 (33755816)
	1A',**2,2A**,3,3C,4a',4,5,6a',6,7,8	(1)	BM468718.1 (EST)	premature stop (40aa)	2 (33792210)	2A (33791861)
	2,2A,3,4a',4,6	(1)	BE816688.1 (EST)	premature stop (40aa)	2 (33792210)	2A (33791861)
	1,1A*,1A',**2,2A**,3,3B	(1)	DA115481.1 (EST)	premature stop (40aa)	2 (33792210)	2A (33791861)
	3A,3B		AV691154.1 (EST)	unknown (77aa)	3A (33787895)	3B (33786230)
	1A',**2,3,3B**	3	AL711019.1 (EST)	premature stop (95aa)	2 (33792210)	3B (33786230)
	1A',**2,2B**,3,3B	4	CA419145.1 (EST)	premature stop (62aa)	2 (33792210)	2B (33790745)
	1A',2,**2B,3,3B**		idem	unknown (90aa)	2B (33790764)	3B (33786230)

Table II.C (cont.)

Protein (Gene symbol) Chr position Strand	Order of exons in isoform (orf in bold)	Accession	Status of cds (length of cds in aa), protein Accession	Start codon Exon (chr postion)#	Stop codon Exon (chr position)##
(cont.) CAPERα (RNPC2) chr20:33,754,944-33,793,607 '-' strand	1,1A*,1A',**2,2A**,2B (1)	DA372839.1 (EST)	premature stop (40aa)	2 (33792210)	2A (33791861)
	1,1A',**2,2A**,2B,3 (1)	BP352717.1 (EST)	premature stop (40aa)	2 (33792210)	2A (33791861)
	1,1A',2,2A,**2B,3** 3	idem	putative alternative start	2B (33790764)	no stop
	1,1A',**2,2A**,3,4a',4,5 (1)	DB027200.1 (EST)	premature stop (40aa)	2 (33792210)	2A (33791861)
	1,1A',**2,3,4a',4,5,6a',6**	DB264131.1 (EST)	known cds (NM_184234.1, NM_004902.2)	2 (33792210)	no stop
	3,4a',4,5a',5,6a',6,7 4	BX483043.1 (EST)	putative in-frame fragment (169aa)	no start	no stop
	1A',**2,3,4a',4,6a',6,7,8** 5	BQ893325.1 (EST)	putative fragment (207aa)	2 (33792210)	no stop
	4,6a',6,7,8,9,10,11,12 (5)	CR995560.1 (EST)	putative in-frame fragment (334aa)	no start	no stop
	6,7,8,9,11,12 6	BQ954122.1 (EST)	putative in-frame fragment (247aa)	no start	no stop
	11B,12,13,14,15 7	BE933146.1 (EST)	putative in-frame fragment (137aa)	no start	no stop
	7,8,9,10,11,11A,11B 8	BM983358.1 (EST)	putative in-frame fragment (255aa)	no start	no stop
	1A,1Aa",2,2a",2a"_2A,2A 5	DB150523.1 (EST)	premature stop (23aa)	1A (33792947)	1A (33792875)
	1A,2,3,4a',4,5,6a',6,7 (5)	BG764840.1 (EST)	premature stop (23aa)	1A (33792947)	1A (33792875)
	1A,**2,3,4a',4,5,6a',6,7**	idem	known cds (178aa)	2 (33792210)	no stop
	1A,2,2A,3,4a',4,5 (5)	DA922841.1 (EST)	premature stop (23aa)	1A (33792947)	1A (33792875)
	1A,**2,2A**,3,4a',4,5 (1)	idem	premature stop (40aa)	2 (33792210)	2A (33791861)
	12,13a',13b',13,14,15,16,17	BI117009.1 (EST)	putative fragment	no start	13a' (33760610)
	1A',**2,3,4a',4,5,6,7,8** 9	BU075848.1 (EST)	putative fragment (288aa)	2 (33792210)	no stop
	2A,3,4a',4,5,6,7 (1)?	AW993266.1 (EST)	premature stop	no start	2A (33791861)
	1,1A*,1A',**2,2a",3,4a',4** 10	DB023865.1 (EST)	putative fragment (104aa)	2 (33792210)	no stop
	12A,13	DA109669.1 (EST)	unknown (34aa)	12A (33761820)	12A (33761715)
	1A',**2,2A**,3,4,5,6a',6 (1)	AL513896.3 (EST)	premature stop (40aa)	2 (33792210)	2A (33791861)

Figure II.C

CAPERα (RNPC2)

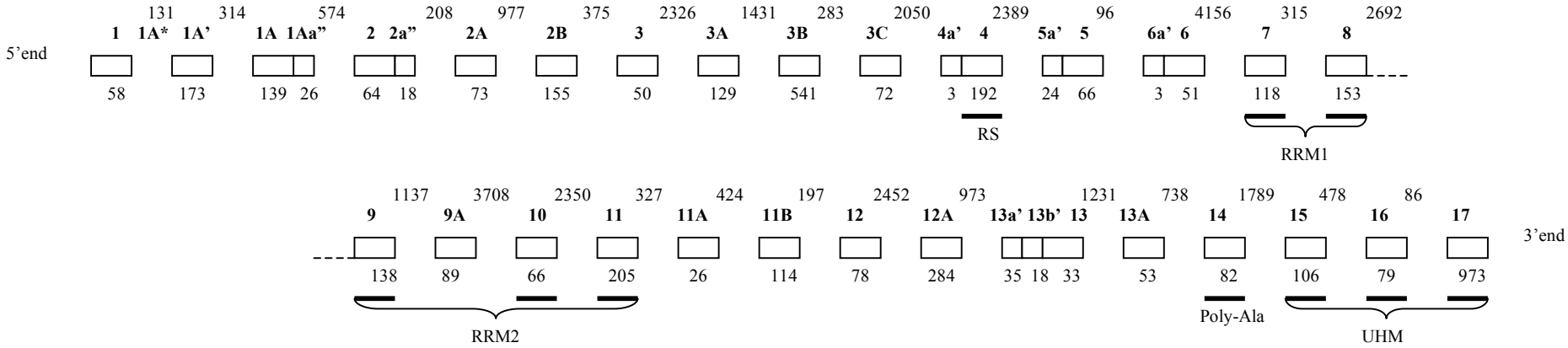


Table II.D

Protein (Gene symbol) Chr position Strand	Order of exons in isoform (orf in bold)		Accession	Status of cds (length of cds in aa), protein Accession	Start codon Exon (chr postion)#	Stop codon Exon (chr position)##
CAPERβ (RBM23) chr14:22,439,816-22,458,214 '-' strand	1,**2,3,4,5,6,7a',7,7a",8a',8,9,10,11,11A*,11A',12,13**		NM_018107.3 (RefSeq)	complete (424aa), NP_060577.2; (423aa), Q86U06-2	2 (22450442)	13 (22440778)
	1,**2,3,4,6,7a',7,7a",8a',8,9,10,11,11A*,11A',12,13**		CR595426.1 (mRNA)	complete (405aa), Q86U06-4	2 (22450442)	13 (22440778)
	1,**2,3,3B,4,5,6,7a',7,7a",8a',8,9,10,11,11A*,11A',12,13**		BX161440.1 (mRNA)	complete (439aa), Q86U06-1	2 (22450442)	13 (22440778)
	1,**2,3,3B,4,5,6,7a',7,7a",8a',8,9,9A**,9A_10,10,11,11A*,11A',12,13		AL834198.1 (mRNA)	complete (419aa), Q86U06-3	2 (22450442)	9A (22443209)
	1,**2,3,3B,4,6**	1	DA675412.1 (EST)	putative fragment (173aa)	2 (22450442)	no stop
	3,3B,4,6,7a',7,7a",8a',8	(1)	BG033916.1 (EST)	putative in-frame fragment (248aa)	no start	no stop
	1,**2,3,3A**	1	DA821789.1 (EST)	premature stop (91aa)	2 (22450442)	3A (22447692)
	1,**2,3,3Ba',3B**,4	2	DB164369.1 (EST)	premature stop (73aa)	2 (22450442)	3B (22447405)
	1,**2,3B**,4,5,6,7a',7,7a"	3	BM464794.1 (EST)	premature stop (29aa)	2 (22450442)	3B (22447405)
	1,**2,3B**,4,6,7a'	(3)	DA145418.1 (EST)	premature stop (29aa)	2 (22450442)	3B (22447405)
	1,**2,4**,6,7a',7,7a",8a',8	4	BI823680.1 (EST)	premature stop (23aa)	2 (22450442)	4 (22445411)
	1,**2,4**,5,6,7a',7,7a",8a',8	4	DB166416.1 (EST)	premature stop (23aa)	2 (22450442)	4 (22445411)
	1,**1C,2,3,4**	2	DA117163.1 (EST)	putative fragment (135aa)	1C (22456120)	no stop
	1,**1Ea',1E,2,3,3B,4**	3	DA311282.1 (EST)	putative fragment (177aa)	1Ea' (22452894)	no stop
	1Ea',1E,2,3,4,5,6,7a',7,7a",8a',8	4	BQ707907.1 (EST)	putative fragment (316aa)	1Ea' (22452894)	no stop
	1,**1Ea',1E,2,3,4,5,6**	(4)	BQ071908.1 (EST)	putative fragment (219aa)	1Ea' (22452894)	no stop
	1B,**2,3**		AA324737.1 (EST)	putative fragment, known cds (59aa)	2 (22450442)	no stop
	1,1D,1E,**2,3,3B**		DA856230.1 (EST)	putative fragment, known cds(75aa)	2 (22450442)	no stop

Table II.D (cont.)

Protein (Gene symbol) Chr position Strand	Order of exons in isoform	Accession	Status of cds (length of cds in aa), protein Accession	Start codon Exon (chr postion)#	Stop codon Exon (chr position)##
(cont.) CAPERβ (RBM23) chr14:22,439,816-22,458,214 '-' strand	1,1E,**2,3**	CD692273.1 (EST)	putative fragment, known cds(59aa)	2 (22450442)	no stop
	1,1G,**2,3,4**	BP356833.1 (EST)	putative fragment, known cds(117aa)	2 (22450442)	no stop
	11,11A' 5	BX388764.2 (EST)	putative fragment (19aa)	no start	no stop
	1,1A 5	AA633094.1 (EST)	premature stop (76aa)	1 (22458231)	1A (22456804)
	1,3,4,6,7a',7,7a",8a',8 6	BI915247.1 (EST)	premature stop (62aa)	1 (22458231)	3 (22448639)
	1,3,4,**6,7a',7,7a",8a',8** 6	idem	putative fragment, new start, known cds (62aa)	6 (22444447)	no stop
	1,3B,4,5,6,7a',7 7	DA299707.1 (EST)	unknown orf (159aa)	1 (22458231)	6 (22444482)
	1,3B,4,5,**6,7a',7** (6)	idem	putative fragment, new start, known cds (57aa)	6 (22444447)	no stop
	1,4,6,7a',7,7a" 8	DA026292.1 (EST)	unknown orf (125aa)	1 (22458231)	6 (22444482)
	1,4,**6,7a',7,7a"** (6)	idem	putative fragment, new start, known cds (63aa)	6 (22444447)	no stop
	1,4,5,6,7a',7,8a',8 9	CN483101.1 (EST)	unknown orf (143aa)	1 (22458231)	6 (22444482)
	1,4,5,**6,7a',7,8a',8** 7	idem	putative fragment, new start (103aa)	6 (22444447)	no stop
	1,4,5,6,7a',7,7a",8a',8,9 (9)	CX165727.1 (EST)	unknown orf (143aa)	1 (22458231)	6 (22444482)
	1,4,5,**6,7a',7,7a",8a',8,9** (6)	idem	putative fragment, new start, known cds (140aa)	6 (22444447)	no stop
	1,3B,4,6,7,7a",8a',8,9,10,11,11A*,11A',12,13 10	BC106012.1 (mRNA)	unknown orf (141aa)	1 (22458231)	6 (22444482)
	1,3B,4,6,7,7a",8a',**8,9,10,11,11A*,11A',12,13** 8	idem	putative fragment, new start, known cds (189aa)	8 (22444037)	13 (22440778)
	7a",8,9,10,11,11A*,11A',12,13	BE269289.1 (EST)	putative fragment (59aa)	no start	9 (22443356)
	7a",**8,9,10,11,11A*,11A',12,13**	idem	putative fragment, new start, known cds (189aa)	8 (22444037)	13 (22440778)
	9A,10,11,11A*,11A',12	BM712633.1 (EST)	putative fragment	no start	9A (22443209)

Figure II.D

CAPERβ (RBM23)

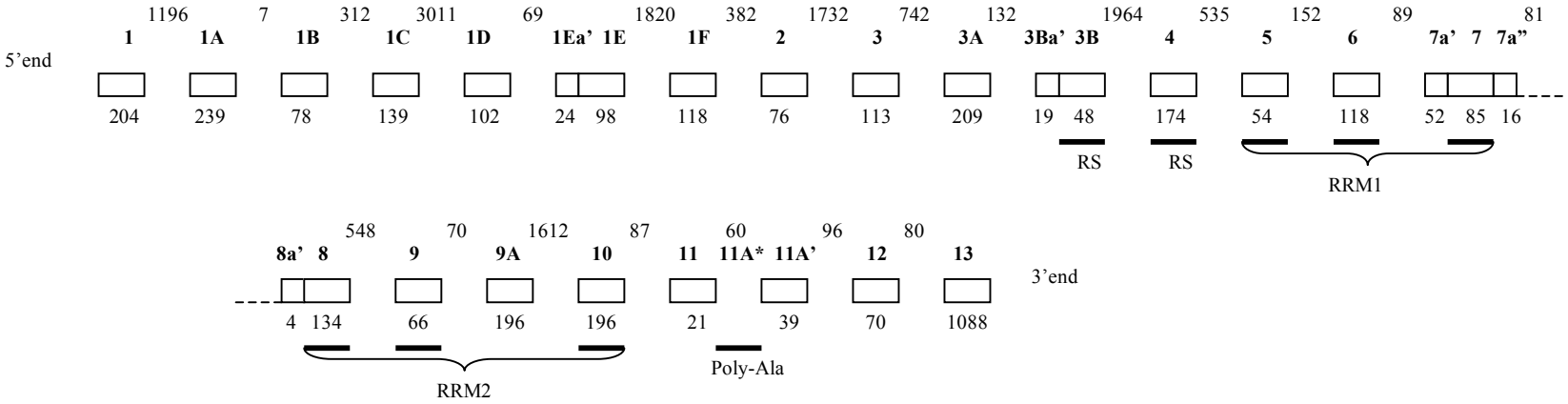


Table II.E

Protein(Gene symbol) Chr position Strand	Order of exons in isoform (orf in bold)	Accession	Status of cds (length of cds in aa), protein Accession	Start codon Exon (chr postion)#	Stop codon Exon (chr position)##
U2AF[35] (U2AF1) chr21: 43386096-43400798 '-' strand	**1,2,3,4,5,6,7,8**	NM_006758.2 (RefSeq)	complete (240aa), NP_006749.1 , Q01081	1 (43400673)	8 (43386280)
	1,2,2A,4,5,6,7,8	NM_001025203.1 (RefSeq)	complete (240aa), Q69YM7	1 (43400673)	8 (43386280)
	1,2,2A,3,4,5,6,7,8 1	NM_001025204.1 (RefSeq)	premature stop	1 (43400673)	3 (43393669)
	1,2,2A,3,**4,5,6,7,8**	idem	complete (167aa), Q71RF1	4 (43388902)	8 (43386280)
	1,**2,2B,4,5,6,7,8** 1	BG612658.1 (EST)	putative (202aa)	2 (43397540)	8 (43386280)
	1A,2,3,4 2	BE736536.1 (EST)	premature stop	1A (43400225)	1A (43400168)

Figure II.E

U2AF35 (U2AF1)

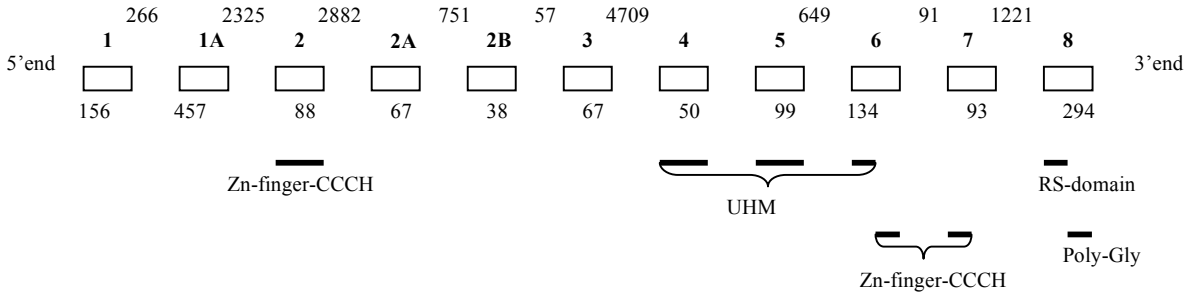


Table II.F

Protein (Gene symbol) Chr position Strand	Order of exons in isoform (coding exons in bold)	Accession	Status of cds (length of cds in aa), protein Accession	Start codon Exon (chr postion)#	Stop codon Exon (chr position)##
U2AF^{26} (U2AF1L4) chr19:40925185-40928187 '-' strand	**1,2a',2,3,4,5,6a',6**	NM_144987.2(RefSeq)	complete (202aa), NP_659424.2, Q8WU68	1 (40928132)	6 (40925338)
	1,2a',2,3,4,4a",5,6a',6	NM_001040425.1 (RefSeq)	complete (181aa), NP_001035515.1 , Q56UU3	1 (40928132)	6a' (40925459)
	2A,2B,3,4,5,6a',6 1	BE856544.1 (EST)	putative fragment (197aa)	no start	6 (40925338)
	1,2a',2,2Ab',2A,3,4,5,6a' 1	BM696851.1 (EST)	premature stop (97aa)	1 (40928132)	3 (40927113)
	1,2a',2,2Ab',2A,**3,4,5,6a'** 2	idem	putative fragment (129aa)	3 (40927117)	no stop
	2,2Ab',2A,3,4,5,6a',6 (1)	BM970675.1 (EST)	premature stop	no start	3 (40927113)
	2,2Ab',2A,**3,4,5,6a',6** (2)	idem	putative, known cds (143aa)	3 (40927117)	6 (40925338)
	2,2Ab',2A,3,4,5,6 (1)	AW274826.1 (EST)	premature stop	no start	3 (40927113)
	2,2Ab',2A,**3,4,5,6** 3	idem	putative (80aa)	3 (40927117)	6 (40925343)
	1,2,3,4,5,6 2	DB127360.1 (EST)	premature stop (51aa)	1 (40928132)	3 (40927113)
	1,2,**3,4,5,6** (3)	idem	putative (80aa)	3 (40927117)	6 (40925343)
	2,2A,3,4,4a",5,6a',6 3	BU628789.1 (EST)	premature stop	no start	2A (40927364)
	2,2A,**3,4,4a",5,6a',6** 4	idem	putative (122aa)	3 (40927117)	6a' (40925459)
	2,2A,3,4,5,6a' (3)	AA455588.1 (EST)	premature stop	no start	2A (40927364)
	2,2A,**3,4,5,6a'** (2)	idem	putative fragment (129aa)	3 (40927117)	no stop
	1,2,2Aa',2Ab',2A,2A_2B 4	BI770029.1 (EST)	premature stop (95aa)	1 (40928132)	2Aa' (40927541)
	1,2,2Ab',2A,3,3_4,4,5,6a',6 5	BC010865.1 (mRNA)	premature stop (39aa)	1 (40928132)	2Ab' (40927466)
	1,2Ab',2A,3 6	BG481735.1 (EST)	premature stop (18aa)	1 (40928132)	2Ab' (40927466)
	1,2,2A,3,3_4,4,4a" (3)?	W51842.1 (EST)	premature stop (37aa)	1 (40928132)	2A (40927426)
	2Ab',2A,**3,4,5,5A,5B,6a',6** 5	BU608847.1 (EST)	putative (129aa)	3 (40927117)	6a' (40925481)
	2A,**3,4,5,5A,5Ba'**,5B,6 6	DB338076.1 (EST)	putative (92aa)	3 (40927117)	5Ba' (40925946)
	4,5,5A,5Ba',5B,6a' (6)	BF821614.1 (EST)	putative fragment	no start	5Ba' (40925946)

Figure II.F

U2AF26 (U2AF1L4)

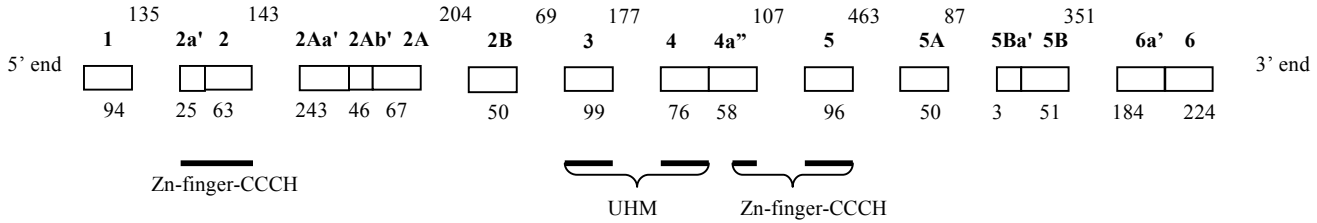
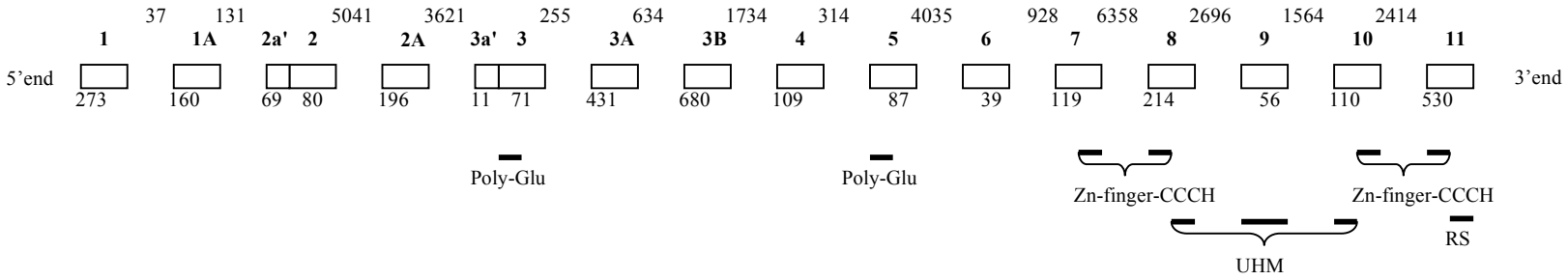


Table II.G

Protein (Gene symbol) Chr position Strand	Order of exons in isoform (orf in bold)		Accession	Status of cds (length of cds in aa), protein Accession	Start codon Exon (chr postion)#	Stop codon Exon (chr position)##
U2AF^{35}-RS2 (U2AF1L2) chrX:15568024-15601078 '+' strand	**1,2,3a',3,4,5,6,7,8,9,10,11**		NM_005089.2 (RefSeq)	complete (482aa), NP_005080.1 , Q15696	1 (15568275)	11 (15601022)
	1,2,3a',3,3B	1	BC065719.1 (mRNA)	premature stop (147aa)	1 (15568275)	3B (15579294)
	1,2,3a',3,3A	2	DA173194.1 (EST)	premature stop (81aa)	1 (15568275)	3A (15578031)
	1,2,3,3A	3	DA383795.1 (EST)	premature stop (49aa)	1 (15568275)	3 (15577691)
	1,2,3,4,5,6,7,8	(3)	CN289520.1 (EST)	premature stop (49aa)	1 (15568275)	3 (15577691)
	1,2a',2,3a',3,4	4	BE619312.1 (EST)	premature stop (24aa)	1 (15568275)	2a' (15568678)
	1A,2,3,4,5,7	5	DA261525.1 (EST)	unknown (12aa)	1A (15568364)	1A (15568415)
	2A,3a',3,4	6	CA425173.1 (EST)	unknown (16aa)	2A (15573869)	2A (15573908)

Figure II.G

U2AF^{35}-RS2 (U2AF1L2)



Appendix E – *Phylogeny of animal model organisms*

Figure taken from "Hedges, S.B. (2002). **The origin and evolution of model organisms**. *Nat Rev Genet*. 3, 838-849".

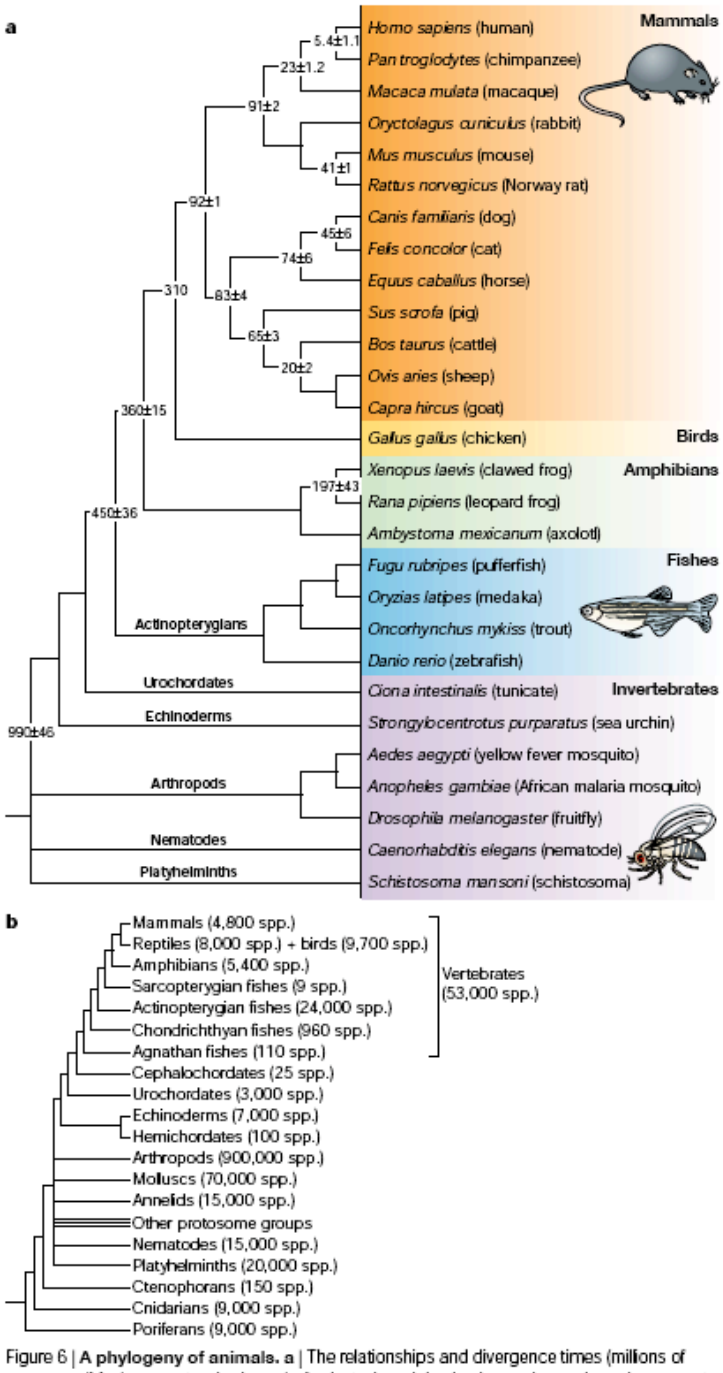


Figure 6 | **A phylogeny of animals. a** | The relationships and divergence times (millions of years ago (Mya) ± one standard error) of selected model animals are shown, based on recent multigene and multiprotein studies[51,61,84]. The fossil divergence time of birds and mammals (310 Mya) was used to calibrate the molecular clock. Branch lengths are not proportional to time. **b** | The relationships and numbers of living species, from a diversity of sources in most of the main groups.

Appendix F - *Cross species analysis of U2AF35 family*

Splicing patterns detected across four vertebrate model organisms for the U2AF small subunit family of splicing factors: Table A. U2AF[35](U2AF1), Table B. U2AF[26] (U2AF1L4) and Table C. U2AF[35]-RS2 (ZRSR2). Longest open reading frame is indicated on the splicing pattern in bold. Splicing patterns confirmed by RT-PCR by Joana Borlibo are indicated in color and underlined (see Appendix G). Introns are highlighted in yellow.

Table A - U2AF[35](U2AF1)

Organism - Gene name Chr position Strand (Assembly)	Splicing pattern (orf in bold)	mRNA Accession
Human - U2AF1 chr21: 43386135-43400785 '-' strand (hg18)	**1,2,3,4,5,6,7,8**	NM_006758.2 (RefSeq)
	1,2,2A,4,5,6,7,8	NM_001025203.1 (RefSeq)
	1,2,2A,3,4,5,6,7,8	NM_001025204.1 (RefSeq)
	1,2,2B,4,5,6,7,8	BG612658.1 (EST)
	1A,2,3,4	BE736536.1 (EST)
	1A_2,2,2A,3,4	BM827141.1 (EST)
	1A_2,2,3,4,5,6,7,8	AF370386.1 (mRNA)
	1,2,2_2A,2A,2A_2B,2B,2B_3,3,4,5,6,7,8	BC032521.2 (mRNA)
	1,2,2A,2A_2B,2B,2B_3,3	AL832665.1 (mRNA)
	4,4_5,5,5_6,6,7,8	BC013786.1 (mRNA)
	3_4,4,5,6	DA936986.1 (EST)
	3_4,4,4_5,5,5_6,6,7,8	BC013786.1 (mRNA)
	3_4,4,4_5,5,5_6,6,7,7_8,8	AK098266.1 (mRNA)
Mouse - U2af1 chr17:31784025-31795838 '-' strand (mm9)	**1,2,3,4,5, 6,7,8**	NM_024187.3 (RefSeq)
	1,2,2A,4,5,6,7,8	CA977261.1 (EST)
	1,2,2A,3,4,5,6,7,8	CX236600.1 (EST)
	1,4,5,6,7,8	AV476937.1 (EST)
	1,2,2A,2A_3	CN669619.1 (EST)
	3_4,4,5,6,7	CA978127.1 (EST)
	1,2,2_2A	CB524928.1 (EST)
Chicken - U2AF1 chr1:113163827-113180264 '-' strand (galGal3)	**1,2,3,4,5,6,7,8**	NM_204655.2 (RefSeq)
	1,2,2B,4,5,6,7,8	BU111430.1 (EST)
	1,2,2B,3,4,5,6,7,8	BU240246.1 (EST)
	1,2,2A,2B,3,4,5,6,7,8	AJ291765.1 (mRNA)
	1,2,2A,3,4,5,6,7	BM491388.1 (EST)
	1,2,2A,2B,4,5,6,7,8	BU318632.1 (EST)
Zebrafish - u2af1 chr9:4449663-4462013 '-' strand (Zv6)	**1,2,3,4,5,6,7,8,**	NM_177479.2 (RefSeq)

Table B - U2AF[26] (U2AF1L4)

Organism - Gene name Chr position Strand (Assembly)	Splicing pattern (orf in bold)	mRNA Accession
Human - U2AF1L4 chr19:40925137-40928182 '-' strand (hg18)	**1,2a',2,3,4,5,6a',6 (zf,UHM)**	NM_144987.2(RefSeq)
	1,2a',2,3,4,4a",5,6a',6 (zf,UHM,zf,RS) change of frame after 4a"	NM_001040425.1 (RefSeq)
	2A,2B,3,4,5,6a',6	BE856544.1 (EST)
	1,2a',2,2Ab',2A,3,4,5,6a',6	BM696851.1 (EST)
	1,2,2Ab',2A,3,3_4,4.5.6a',6	BC010865.1 (mRNA)
	2,2Ab',2A,3,4,5,6	AW274826.1 (EST)
	1,2,3,4,5,6	DB127360.1 (EST)
	2,2A,3,4,4a",5,6a',6	BU628789.1 (EST)
	1,2,2A,3,3_4,4,4a"	W51842.1 (EST)
	2,2A,3,4,5,6a'	AA455588.1 (EST)
	1,2,2Aa',2Ab',2A,2A_2B	BI770029.1 (EST)
	1,2a',2,2Aa',2Ab',2A,2A_2B,2B,3,4,5,6a',6	BC062474.1 (mRNA)
	1,2Ab',2A,3	BG481735.1 (EST)
	2Ab',2A,**3,4,5,5A,5B,6a',6**	BU608847.1 (EST)
	2A,**3,4,5,5A,5Ba'**,5B,6	DB338076.1 (EST)
	4,5,5A,5Ba',5B,6a'	BF821614.1 (EST)
	1,2,2Ab',2A,3,4,4a",5,5A,5A_5Ba',5Ba',5B,6a',6	AY461582.1 (mRNA)
	4,5,6,W1	AA845716.1 (EST)
	4,5,6,6a",W1a',W1	AA713510.1 (EST)
Mouse - U2af1l4 chr7:31347032-31352981 '+' strand (mm9)	1,**1a",2a',2,3,4,4a",5,6,7,8**	NM_170760.3 (RefSeq)
	1a",2a',2,3,4,4a",5,6,7,8,W1,W2,W3	BC031744.1 (mRNA)
	1,**1a",2a',2,3,4,4a",5,6,7,8**,W1,W1_W2,W2,W2a",W3	AK160974.1 (mRNA)
	1,**1a",2a',2,3,4,4a",4a"_5,5**	BY303250.1 (EST)
	1a",2a',2,3,4,4a",4a"_5,5,5a",5a"_6,6	BI152717.1 (EST)
	1a",2a',2,3,4,4a",5,5a",5a"_6,6	AA185455.1 (EST)
	1a",2a',2,3,4,4a",5,5a",6,7,8,W1,W2,W2a",W3	BC057315.1 (mRNA)
	1b"_2a',2a',2,3,4,4a",**5,6,7,8**	DV647783.1 (EST)
	1,**1a",2a',2,3,4,4a",5,6,8**	BI790606.1 (EST)
	1,**1a",2a',2,3,4,5,6,7**	DV060266.1 (EST)
	1,2,3,4,4a",5,6,7,8	CF105709.1 (EST)
	1,2,**3,4**,4a",5,5a",5a"_6,6	BE372653.1 (EST)
	1,2a',2,3,4,4a",5,5a",5a"_6,6,7	CN713072.1 (EST)
	1,2a',2,3,4,4a",5,6,8	BY709622.1 (EST)
	1a",2,3,4,4a",5,6,7	BE850892.1 (EST)
	1,**1a",2,3**,4,4a",5,6	BY276709.1 (EST)
	1,**1a",2,3**,4,4a",5,5a",5a"_6,6	CN832960.1 (EST)
	1,**1a",2,3**,4,4a",4a"_5,5,5a",5a"_6,6,7,8	DV062015.1 (EST)
	2_3,3,4,4a",5,5a",5a"_6,6	BI247514.1 (EST)
	2_3,3,4,4a",5,6	CJ167056.1 (EST)
	1a",1b",1b"_2a',2a',2,3,4,4a",4a"_5,5,5a"	BM944609.1 (EST)
	1a",1b",2a',2,3,4,4a",5,6,7,8	BE570340.1 (EST)
	1,**1a",1b"**,2a',2,3,4,4a",5,6	BC051477.1 (mRNA)
	1,**1a",1b"**,1b"_2a',2a',2,3,4,4a",4a"_5,5,6	BG863348.1 (EST)
	1a",**1b",2,3,4,4a",5,8**	AV481891.1 (EST)
	1a",1b",1b"_2a',2a',2,3,4,4a",5,8	CK020700.1 (EST)
	1a",**1b",2,3,4,4a",5,6,7**	CB840666.1 (EST)

Table C – U2AF^{35}-RS2 (ZRSR2)

Organism - Gene name Chr position Strand (Assembly)	Splicing pattern (orf in bold)	mRNA Accession
Human - ZRSR2 chrX:15718308-15751385 '+' strand (hg18)	1,2,3a',**3,4,5,6,7,8,9,10,11**	NM_005089.2 (RefSeq)
	1,2,3a',**3,3B**	BC065719.1 (mRNA)
	1,2,3a',**3,3A**	DA173194.1 (EST)
	1,2,3,3A	DA383795.1 (EST)
	1,2,3,4,5,6,7,8	CN289520.1 (EST)
	1,2a',2,3a',3,4	BE619312.1 (EST)
	1A,2,3,4,5,7	DA261525.1 (EST)
	2A,3a',3,4	CA425173.1 (EST)
Mouse - Zrsr2 chrX:159279548-159302782 '-' strand (mm9)	1,1A*,1A',1B*,**1B',2,3a',3,4,5,6,7,8,9,10,11**	NM_009453.2 (RefSeq)
	1,1A*,1A',1B*,**1B',2,3a',3,4,5,5_6**	NM_178794.3 (RefSeq)
	1,**1B',2,3a',3**	CJ149312.1 (EST)
	1A',**1B',2,3a',3,4**	CK623156.1 (EST)
	6,7,8,9,10,10_11	BU614769.1 (EST)
	3_4,4,5,5_6	BQ745328.1 (EST)
	1B',2,3	BB578434.1 (EST)
	7_8,**8,9**,9_10	CN700280.1 (EST)
	5_6,**6,7**,7_8	CB521832.1 (EST)
	1B'_2,2,3a',3,4,5	BU530520.1 (EST)
Chicken - ZRSR2 chr1:125314753-125330984 '-' strand (galGal3)	**1,2,3,4,5,** **2,3,4,5,6,7,8,** **5,6,7,8,9,10,11**	composed from CR354065.1 (mRNA) CV890012.1 (EST) BU464289.1 (EST)
Zebrafish - ZRSR2 (duplicated gene) chr11:11601277-11611649 chr11:11390562-11401670 '-' strand (Zv6)	**1,2,3,4** **7,8,9,10,11**	EB909890.1 (EST) CT639411.2 (EST)

Appendix G - *Method PCR validation of human U2AF35 novel isoforms*

Method of RT-PCR validation performed by Joana Borlido

cDNA synthesis was carried out on Human Universal RNA (Stratagene) using Superscript II reverse transcriptase (Invitrogen). Primers were designed by using the computer software Primer Express (Applied Biosystems, Foster City, CA) and primer concentrations were optimized for each amplicon using Human Universal RNA as a template. The oligonucleotide used for each set of primers are listed in Table G1; the concentration of primers used in the reaction was of 900nM. All reactions were performed in the ABI7000 Sequence Detector (Applied Biosystems, Foster City, CA). The amplification products were then separated in a 2% agarose gel containing 0.5 μg/ml of ethidium bromide (EtBr) and quantified by ethidium bromide fluorescence in a Phosphorimager detector (Amersham Pharmacia Biotech).

Table G1 - Oligonucleotide sequence and concentration of primers used in RT-PCR

Protein (Gene Symbol)	Accession Number	Amplicon size (bp)	Primer Sequences
U2AF35 (U2AF1)	NM_006758.2 (RefSeq)	126	Forward, 5'- GCCAGACCATTGCCCTCTT -3' Reverse, 5'- CCTCCTCAAAAAACTCATCATAGTGTT -3'
	NM_001025203.1 (RefSeq)	71	Forward, 5'- GGCTGACGGCTCACACTGT -3' Reverse, 5'- CCTCCTCAAAAAACTCATCATAGTGTT -3'
	BG612658.1 (EST)	75	Forward, 5'- AGGCAGCCGGTGCCGT -3' Reverse, 5'- TTTCTGTAAAAACCTCCTCAAAAAACTC -3'
U2AF26 (U2AF1L4)	NM_001040425.1 (RefSeq)	69	Forward, 5'- GTGAGCTGTCTCCTGTCACTGACT -3' Reverse, 5'- GGGTACATTCCCCCATCTCAT -3'
	NM_144987.2 (RefSeq)	111	Forward, 5'- AGGCTGTGCACGGGAATG -3' Reverse, 5'- GACCTGCGCCTGGGTC -3'
	BU608847.1 (EST)	68	Forward, 5'- CTCTGCCTCCCAGAGATGGG -3' Reverse, 5'- TGACCTGAGGTCCGGAGTTC -3'
	DB338076.1 (EST)	76	Forward, 5'- GCAACCTCTGCCTCCCAGTA -3' Reverse, 5'- GAGCCTGAGGTCCGGAGTTC -3'
U2AF35-R2 (U2AF1L2)	NM_005089.2 (RefSeq)	98	Forward, 5'- TTTCCCGAGAAACCAAGCC -3' Reverse, 5'- CCTGAGTCTCTCAGTCGAGCAA -3'
	DA261525.1 (EST)	94	Forward, 5'- ATCGTGGGCATGGAGGATGT -3' Reverse, 5'- GCGCACCAGGAAGGAATG -3'
	CA425173.1 (EST)	74	Forward, 5'- GTTGGGATTACAGACACGCA -3' Reverse, 5'- TGCTGGAAGTGCTCTCAAGAAG -3'

6. SUPPLEMENTARY DATA

Supplementary data is found on the CD

Tables

SupplemetaryTable1.doc

SupplementaryDataFile1_2000HumanMuscleGenes.xls

SupplementaryDataFile2_Tables443MuscleGenes_U06hs48.xls

SupplementaryDataFile3_Tables420SplicingFactors_U06hs48.xls

Publications

Article_Mollet et al 2006.pdf

Article_Pacheco et al 2006.pdf

Article_Mollet et al 2008.pdf

ABBREVIATIONS

BLAT – Blast Like Alignment Tool

CPAN - Comprehensive Perl Archive Network

EST – Expressed Sequence Tag

mRNA – messenger RNA

RT-PCR – Reverse Transcriptase –Polymerase Chain Reaction

SQL – Standard Query Language

UCSC – University of California Santa Cruz

7. REFERENCES

Abovich, N., and Rosbash, M. (1997). **Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals.** *Cell* 89, 403-412.

Aguilera, A. (2005). **Cotranscriptional mRNP assembly: from the DNA to the nuclear pore.** *Curr Opin Cell Biol* 17, 242-250. Review.

Ajuh P., Kuster, B., Panov, K., Zomerdijk, J.C., Mann, M., and Lamond, A.I. (2000). **Functional analysis of the human CDC5L complex and identification of its components by mass spectrometry.** *EMBO J* 19, 6569-6581.

Allison, D.B., Cui, X., Page, G.P., and Sabripour, M. (2006). **Microarray data analysis: from disarray to consolidation and consensus.** *Nat Rev Genet* 7, 55-65. Review.

Amack, J.D., Reagan, S.R., and Mahadevan, M.S. (2002). **Mutant DMPK 3'-UTR transcripts disrupt C2C12 myogenic differentiation by compromising MyoD.** *J Cell Biol* 159, 419-429.

Ambros, V. (2001). **microRNAs: tiny regulators with great potential.** *Cell* 107, 823-826. Review.

Ambros, V. (2003). **MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing.** *Cell* 113, 673-676. Review.

Ambros V. (2004). **The functions of animal microRNAs.** *Nature* 431, 350-355. Review.

Ara, T., Lopez, F., Ritchie, W., Benech, P., and Gautheret, D. (2006). **Conservation of alternative polyadenylation patterns in mammalian genes.** *BMC Genomics* 7, 189.

Aronova, A., Bacíková, D., Crotti, L.B., Horowitz, D.S., and Schwer, B. (2007). **Functional interactions between Prp8, Prp18, Slu7, and U5 snRNA during the second step of pre-mRNA splicing.** *RNA* 13, 1437-1444.

Auboeuf, D., Dowhan, D.H., Kang, Y.K., Larkin, K., Lee, J.W., Berget, S.M., and O'Malley, B.W. (2004). **Differential recruitment of nuclear receptor coactivators may determine alternative RNA splice site choice in target genes.** *PNAS* 101, 2270-2274.

Aznarez, I., Chan, E.M., Zielenski, J., Blencowe, B.J., and Tsui, L.C. (2003). **Characterization of disease-associated mutations affecting an exonic splicing enhancer and two cryptic splice sites in exon 13 of the cystic fibrosis transmembrane conductance regulator gene.** *Hum Mol Genet* 12, 2031-2040.

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

Bartels, C., Klatt, C., Lührmann, R., and Fabrizio, P. (2002). **The ribosomal translocase homologue Snu114p is involved in unwinding U4/U6 RNA during activation of the spliceosome.** *EMBO Rep* 3, 875-880.

Bartels, C., Urlaub, H., Luhrmann, R., and Fabrizio, P. (2003). **Mutagenesis suggests several roles of Snu114p in pre-mRNA splicing.** *J Biol Chem* 278, 28324-28334.

Bartel,D.P. (2004). **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell* 116, 281-297.

Bashaw, G.J., and Baker, B.S. (1997) **The regulation of the Drosophila msl-2 gene reveals a function for Sex-lethal in translational control.** *Cell* 89, 789-798.

Batsché, E., Yaniv, M., and Muchardt, C. (2006). **The human SWI/SNF subunit Brm is a regulator of alternative splicing.** *Nat Struct Mol Biol* 13, 22-29.

Bazeley, P.S., Shepelev, V., Talebizadeh, Z., Butler, M.G., Fedorova, L., Filatov, V., and Fedorov, A. (2008). **snoTARGET shows that human orphan snoRNA targets locate close to alternative splice junctions.** *Gene* 408, 172-179.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M., and Gautheret, D. (2000). **Patterns of variant polyadenylation signal usage in human genes**. *Genome Res* 10, 1001-1010.

Beaudoing, E., and Gautheret, D. (2001). **Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data.** *Genome Res* 11**,** 1520-1526.

Behzadnia, N., Golas, M.M., Hartmuth, K., Sander, B., Kastner, B., Deckert, J., Dube, P., Will, C.L., Urlaub, H., Stark, H., and Lührmann, R. (2007). **Composition and three-dimensional EM structure of double affinity-purified, human prespliceosomal A complexes.** *EMBO J* 26, 1737-1748.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. **(**2007). **GenBank**. *Nucleic Acids Res* 35(Database issue), D21–D25.

Bentley, D. (2002). **The mRNA assembly line: transcription and processing machines in the same factory.** *Curr Opin Cell Biol* 14, 336-342. Review.

Bentley, D.L. (2005). **Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors.** *Curr Opin Cell Biol* 17, 251-256. Review.

Berget, S.M. (1995) **Exon Recognition in Vertebrate Splicing.** *J Biol Chem* 270, 2411-2414.

Berglund, J.A., Chua, K., Abovich, N., Reed, R., and Rosbash, M. (1997). **The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC.** *Cell* 89, 781-787.

Berglund, J.A., Abovich, N., and Rosbash, M. (1998a). **A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition.** *Genes Dev* 12, 858-867.

Berglund, J.A., Fleming, M.L., and Rosbash, M. (1998b). **The KH domain of the branchpoint sequence binding protein determines specificity for the pre-mRNA branchpoint sequence.** *RNA* 4, 998-1006.

Berglund, J.A., Rosbash, M., and Schultz, S.C. (2001). **Crystal structure of a model branchpoint-U2 snRNA duplex containing bulged adenosines.** *RNA* 7, 682-691.

Bird, G., Zorio, D.A., and Bentley, D.L. (2004). **RNA polymerase II carboxy-terminal domain phosphorylation is required for cotranscriptional pre-mRNA splicing and 3'-end formation.** *Mol Cell Biol* 24, 8963-8969.

Bird, G., Fong, N., Gatlin, J.C., Farabaugh, S., and Bentley, D.L. (2005). **Ribozyme cleavage reveals connections between mRNA release from the site of transcription and pre-mRNA processing.** *Mol Cell* 20, 747-758.

Black, D.L. (1995). **Finding splice sites within a wilderness of RNA**. *RNA* 1, 763-771.

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

Blaustein, M., Pelisch, F., Tanos, T., Muñoz, M.J., Wengier, D., Quadrana, L., Sanford, J.R., Muschietti, J.P., Kornblihtt, A.R., Cáceres, J.F., Coso, O.A., and Srebrow, A. (2005). **Concerted regulation of nuclear and cytoplasmic activities of SR proteins by AKT.** *Nat Struct Mol Biol* 12, 1037-1044.

Blencowe, B.J. (2000). **Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases.** *Trends Biochem Sci* 25,106 –110.

Blencowe, B.J. (2003). **Splicing Regulation: The Cell Cycle Connection**. *Curr Biol* 13, R149-151.

Blencowe, B.J. (2006). **Alternative Splicing: New Insights from Global Analyses.** *Cell* 126, 37-47.

Bollen, M., and Beullens, M. (2002). **Signaling by protein phosphatases in the nucleus**. *Trends Cell Biol* 12, 138-145.

Borchert, G.M., Lanier, W., and Davidson, B.L. (2006). **RNA polymerase III transcribes human microRNAs.** *Nat Struct Mol Biol* 13, 1097-1101.

Boukis, L.A., Liu, N., Furuyama, S. and Bruzik, J. P. (2004). **Ser/Arg-rich Protein-mediated Communication between U1 and U2 Small Nuclear Ribonucleoprotein Particles.** *J Biol Chem* 279, 29647-29653.

Boutz, P.L., Chawla, G., Stoilov, P., and Black, D.L. (2007b). **MicroRNAs regulate the expression of the alternative splicing factor nPTB during muscle development**. *Genes Dev* 21, 71-84.

Brow, D.A. (2002). **Allosteric Cascade Of Spliceosome Activation.** *Annu Rev Genet* 36, 333-360.

Brudno, M., Gelfand, M.S., Spengler, S., Zorn, M., Dubchak, I., and Conboy, J.G. (2001). **Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing**. *Nucleic Acids Res* 29, 2338-2348.

Buj-Bello, A., Furling, D., Tronchère, H., Laporte, J., Lerouge, T., Butler-Browne, G.S., and Mandel, J.L. (2002). **Muscle-specific alternative splicing of myotubularin-related 1 gene is impaired in DM1 muscle cells.** *Hum Mol Genet* 11, 2297-2307.

Buratti, E., and Baralle, F.E. (2004). **Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process.** *Mol Cell Biol* 24, 10505-10514. Minireview.

Burge, C.B., Tuschl, T. and Sharp, P.A. (1999). **Splicing of Precursors to mRNAs by the Spliceosomes.** In *The RNA World*, Gesteland, R.F., Atkins, J.F. & Cech, T.R. (eds). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2nd Ed., pp. 525–560.

Butler, J.E.F., and Kadonaga, J.T. (2002). **The RNA polymerase II core promoter: a key component in the regulation of gene expression**. *Genes Dev* 16, 2583-2592.

Cáceres, J.F., Screaton, G.R., and Krainer, A.R. (1998). **A specific subset of SR proteins shuttles continuously between the nucleus and the cytoplasm.** *Genes Dev* 12, 55–66.

Cáceres, J.F., and Misteli, T. (2007). **Division of labor: minor splicing in the cytoplasm.** *Cell* 131, 645-647.

Calin, G.A., and Croce, C.M. (2006). **MicroRNA signatures in human cancers.** *Nat Rev Cancer* 6, 857-866.

Carrington, J.C., and Ambros, V. (2003). **Role of microRNAs in plant and animal development.** *Science* 301, 336-338. Review.

Cartegni, L., Chew, S. L. and Krainer, A. R. (2002). **Listening to silence and understanding nonsense: exonic mutations that affect splicing.** *Nat Rev Genet* 3, 285–298.

Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q., and Krainer, A.R. (2003). **ESEfinder: A web resource to identify exonic splicing enhancers.** *Nucleic Acids Res* 31, 3568-3571.

Chan, S.P., Kao, D.I., Tsai, W.Y., and Cheng, S.C. (2003). **The Prp19p-associated complex in spliceosome activation.** *Science* 302, 279-282.

Charlet-B ,N., Savkur, R.S., Singh, G., Philips, A.V., Grice, E.A., and Cooper, T.A. (2002). **Loss of the muscle-specific chloride channel in type 1 myotonic dystrophy due to misregulated alternative splicing.** *Mol Cell* 10, 45-53.

Chasin, L.A. (2007). **Searching for splicing motifs.** *Adv Exp Med Biol* 623, 85-106. Review.

Chen, C.D., Kobayashi, R., and Helfman, D.M. (1999). **Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat b-tropomyosin gene.** *Genes & Dev* 13, 593-606.

Chen, C.H., Yu, W.C., Tsao, T.Y., Wang, L.Y., Chen, H.R., Lin, J.Y., Tsai, W.Y., and Cheng, S.C. (2002). **Functional and physical interactions between components of the Prp19p-associated complex.** *Nucleic Acids Res* 30, 1029-1037.

Chen, Y.I., Moore, R.E., Ge, H.Y., Young, M.K., Lee, T.D., and Stevens, S.W. (2007). **Proteomic analysis of in vivo-assembled pre-mRNA splicing complexes expands the catalog of participating factors.** *Nucleic Acids Res* 35, 3928-3944.

Cheng, J., Randall, A., Sweredoski, M., and Baldi, P. (2005). **SCRATCH: a Protein Structure and Structural Feature Prediction Server**. *Nucleic Acids Research* 33 (web server issue), w72-76,.

Chiara, M.D., Gozani, O., Bennett, M., Champion-Arnaud, P., Palandjian, L., and Reed, R. (1996). **Identification of proteins that interact with exon sequences, splice sites, and the branchpoint sequence during each stage of spliceosome assembly.** *Mol Cell Biol* 16, 3317-3326.

Chou, M.Y., Underwood, J.G., Nikolic, J., Luu, M.H., and Black, D.L. (2000). **Multisite RNA binding and release of polypyrimidine tract binding protein during the regulation of c-src neural-specific splicing**. *Mol Cell* 5, 949-957.

Colgan, D.F. and Manley, J.L. (1997). **Mechanism and regulation of mRNA polyadenylation.** *Genes Dev* 11, 2755-2766.

Collins, C.A., and Guthrie, C. (1999). **Allele-specific genetic interactions between Prp8 and RNA active site residues suggest a function for Prp8 at the catalytic core of the spliceosome**. *Genes Dev* 13, 1970-1982.

Collins, C.A., and Guthrie, C. (2000). **The question remains: is the spliceosome a ribozyme?** *Nat Struct Biol* 7, 850-854. Review.

Collins, C.A., and Guthrie, C. (2001). **Genetic interactions between the 5' and 3' splice site consensus sequences and U6 snRNA during the second catalytic step of pre-mRNA splicing.** *RNA* 7, 1845-1854.

Cooper, S.J., Trinklein, N.D., Anton, E.D., Nguyen, L., and Myers, R.M. (2006). **Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome**. *Genome Res* 16, 1-10.

Cooper, T.A. (2007). **Regulation of chloride ion conductance during skeletal muscle development and in disease. Focus on "Chloride channelopathy in myotonic dystrophy resulting from loss of posttranscriptional regulation for CLCN1".** *Am J*

Physiol Cell Physiol 292, C1245- C1247.

Corpet, F. (1988). **Multiple sequence alignment with hierarchical clustering**. *Nucleic Acids Res* 16, 10881–10890.

Corsini, L., Bonnal, S., Basquin, J., Hothorn, M., Scheffzek, K., Valcárcel, J., and Sattler, M. (2007). **U2AF-homology motif interactions are required for alternative splicing regulation by SPF45**. *Nat Struct Mol Biol* 14, 620-629.

Cramer, P., Pesce, C.G., Baralle, F.E., and Kornblihtt, A.R. (1997). **Functional association between promoter structure and transcript alternative splicing.** *PNAS* 94, 11456-11460.

Cramer, P., Srebrow, A., Kadener, S., Werbajh, S., de la Mata, M., Melen, G., Nogués, G., and Kornblihtt, A.R. (2001). **Coordination between transcription and pre-mRNA processing**. *FEBS Lett* 498, 179-182. Review.

Cramer, P. (2007). **Finding the right spot to start transcription.** *Nat Struct Mol Biol* 14, 686-687.

Crowder, S.M., Kanaar, R., Rio, D.C., and Alber, T. (1999). **Absence of interdomain contacts in the crystal structure of the RNA recognition motifs of Sex-lethal.** *Proc Natl Acad Sci USA* 96, 4892-4897.

Das, R., Zhou, Z., and Reed, R. (2000). **Functional association of U2 snRNP with the ATP-independent spliceosomal complex E**. *Mol Cell* 5, 779-787.

Das, R., Dufu, K., Romney, B., Feldt, M., Elenko, M., and Reed, R. (2006). **Functional coupling of RNAP II transcription to spliceosome assembly.** *Genes Dev* 20, 1100-1109.

Das, R., Yu, J., Zhang, Z., Gygi, M.P., Krainer, A.R., Gygi, S.P., and Reed, R. (2007). **SR proteins function in coupling RNAP II transcription to pre-mRNA splicing**. *Mol Cell* 26, 867-881.

Davis, B.M., McCurrach, M.E., Taneja, K.L., Singer, R.H., and Housman, D.E. (1997). **Expansion of a CUG trinucleotide repeat in the 3' untranslated region of myotonic dystrophy protein kinase transcripts results in nuclear retention of transcripts.** *Proc Natl Acad Sci U S A* 94, 7388-7393.

Deckert, J., Hartmuth, K., Boehringer, D., Behzadnia, N., Will, C.L., Kastner, B., Stark, H., Urlaub, H., and Lührmann, R. (2006). **Protein composition and electron microscopy structure of affinity-purified human spliceosomal B complexes isolated under physiological conditions.** *Mol Cell Biol* 26, 5528-5543.

Denis, M.M., Tolley, N.D., Bunting, M., Schwertz, H., Jiang, H., Lindemann, S., Yost, C.C., Rubner, F.J., Albertine, K.H., Swoboda, K.J., Fratto, C.M., Tolley, E., Kraiss, L.W., McIntyre, T.M., Zimmerman, G.A., and Weyrich, A.S. (2005). **Escaping the nuclear confines: signal-dependent pre-mRNA splicing in anucleate platelets.** *Cell* 122, 379-391.

Dhaenens, C.M., Schraen-Maschke, S., Tran, H., Vingtdeux, V., Ghanem, D., Leroy, O., Delplanque, J., Vanbrussel, E., Delacourte, A., Vermersch, P., Maurage, C.A., Gruffat, H., Sergeant, A., Mahadevan, M.S., Ishiura, S., Buée, L., Cooper, T.A., Caillet-Boudin, M.L., Charlet-Berguerand, N., Sablonnière, B., and Sergeant, N. (2008). **Overexpression of MBNL1 fetal isoforms and modified splicing of Tau in the DM1 brain: Two individual consequences of CUG trinucleotide repeats.** *Exp Neurol* 210, 467-478.

Dominguez, C. and Allain, F.H. (2006). **NMR structure of the three quasi RNA recognition motifs (qRRMs) of human hnRNP F and interaction studies with Bcl-x G-tract RNA: A novel mode of RNA recognition.** *Nucleic Acids Res* 34, 3634-3645.

Donahue, C.P., Muratore, C., Wu, J.Y., Kosik, K.S., and Wolfe, M.S. (2006). **Stabilization of the tau exon 10 stem–loop alters pre-mRNA splicing**. *J Biol Chem* 281, 23302-23306.

Dönmez, G., Hartmuth, K., Lührmann, R. (2004). **Modified nucleotides at the 5' end of human U2 snRNA are required for spliceosomal E-complex formation.** *RNA* 10, 1925-1933.

Dönmez, G., Hartmuth, K., Kastner, B., Will, C.L., and Lührmann, R. (2007). **The 5' end of U2 snRNA is in close proximity to U1 and functional sites of the pre-mRNA in early spliceosomal complexes.** *Mol Cell* 25, 399-411.

Dowhan, D.H., Hong, E.P., Auboeuf, D., Dennis, A.P., Wilson, M.M., Berget, S.M. and O'Malley, B.W. (2005). **Steroid hormone receptor coactivation and alternative RNA splicing by U2AF65-related proteins CAPERalpha and CAPERbeta**. *Mol Cell* 17, 429–439.

D'Souza, I. and Schellenberg, G.D. (2002). **Tau exon 10 expression involves a bipartite intron 10 regulatory sequence and weak 59 and 39 splice sites.** *J Biol Chem* 277, 26587-26599.

Du, H., and Rosbash, M. (2001). **Yeast U1 snRNP-pre-mRNA complex formation without U1snRNA-pre-mRNA base pairing.** *RNA* 7, 133-142.

Du, H., and Rosbash, M. (2002).**The U1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing**. *Nature* 419, 86-90.

Eddy, S.R. (1999). **Noncoding RNA genes.** *Curr Opin Genet Dev* 9, 695-699.

Eddy, S.R. (2001). **Non–coding RNA genes and the modern RNA world.** *Nat. Rev Genet* 2, 919-929.

Eddy, S.R. (2002). **Computational Genomics of Noncoding RNA Genes.** *Cell* 109, 137-140.

Edwalds-Gilbert, G., Veraldi, K.L., and Milcarek, C. (1997). **Alternative poly(A)site selection in complex transcription units: means to an end?** *Nucleic Acids Res* 25, 2547-2561.

Eliceiri, G.L. (1999). **Small nucleolar RNAs.** *Cell Mol Life Sci* 56, 22-31. Review.

Elliott, D.J., and Grellscheid, S.N. (2006). **Alternative RNA splicing regulation in the testis.** *Reproduction* 132, 811-819. Review.

Esquela-Kerscher, A. and Slack, F.J. (2006). **Oncomirs - microRNAs with a role in cancer.** *Nat Rev Cancer* 6, 259-269.

Fairbrother, W.G., Yeh, R.F., Sharp, P.A. and Burge, C.B. (2002). **Predictive Identification of Exonic Splicing Enhancers in Human Genes.** *Science* 297, 1007-1013.

Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B., and Bartel, D.P. (2005). **The widespread impact of mammalian MicroRNAs on mRNA repression and evolution.** *Science* 310, 1817-1821.

Fededa JP, Petrillo E, Gelfand MS, Neverov AD, Kadener S, Nogués G, Pelisch F, Baralle FE, Muro AF, Kornblihtt AR. (2005). **A polar mechanism coordinates different regions of alternative splicing within a single gene.** *Mol Cell* 19, 393-404.

Fekete, C.A., Mitchell, S.F., Cherkasova, V.A., Applefield, D., Algire, M.A., Maag, D., Saini, A.K., Lorsch, J.R., and Hinnebusch, A.G. (2007). **N- and C-terminal residues of eIF1A have opposing effects on the fidelity of start codon selection.** *EMBO J* 26, 1602-1614.

Filipowicz, W., Bhattacharyya S.N., and Sonenberg, N. (2008). **Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?** *Nat Rev Genet* 9, 102-114.

Förch, P., Puig, O., Kedersha, N., Martínez, C., Granneman, S., Séraphin, B., Anderson, P., and Valcárcel, J. (2000). **The apoptosis-promoting factor TIA-1 is a regulator of alternative pre-mRNA splicing.** *Mol Cell* 6, 1089-1098.

Förch, P., Merendino, L., Martinez, C., and Valcarcel, J. (2001). **Modulation of msl-2 5' splice site recognition by Sex-lethal**. *RNA* 7, 1185-1191.

Fujimura, K., Kano, F., and Murata, M. (2008). **Identification of PCBP2, a facilitator of IRES-mediated translation, as a novel constituent of stress granules and processing bodies.** *RNA* 14, 425-431.

Gabut, M., Chaudhry, S., and Blencowe, B.J. (2008). **SnapShot: The splicing regulatory machinery.** *Cell* 133, 192.e1.

Gama-Carvalho, M., Barbosa-Morais, N.L., Brodsky, A.S., Silver, P.A., and Carmo-Fonseca, M. (2006). **Genome-wide identification of functionally distinct subsets of cellular mRNAs associated with two nucleocytoplasmic-shuttling mammalian splicing factors.** *Genome Biol* 7, R113.

Gao, K., Masuda, A., Matsuura, T., and Ohno, K. (2008). **Human branch point consensus sequence is yUnAy.** *Nucleic Acids Res* 36, 2257-2267.

Gardina, P.J., Clark, T.A., Shimada, B., Staples, M.K., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S., Davies, C., Williams, A., and Turpaz, Y. (2006). **Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array.** *BMC Genomics* 7, 325.

Garg, K., and Green, P. (2007). **Differing patterns of selection in alternative and constitutive splice sites.** *Genome Res* 17, 1015-1022.

Gaur, R.K., Valcárcel, J., and Green, M.R. (1995). **Sequential recognition of the pre-mRNA branch point by U2AF65 and a novel spliceosome-associated 28-kDa protein.** *RNA* 1, 407-417.

Gebauer, F., Merendino, L., Hentze, M.W., and Valcarcel, J. (1998). **The Drosophila splicing regulator sex-lethal directly inhibits translation of male-specific-lethal 2 mRNA**. *RNA* 4, 142-150.

Glover-Cutter, K., Kim, S., Espinosa, J., and Bentley, D.L. (2008). **RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes**. *Nat Struct Mol Biol* 15, 71-78.

Gooding, C., Clark, F., Wollerton, M.C., Grellscheid, S.N., Groom, H., and Smith, C.W. (2006). **A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones**. *Genome Biol* 7, R1.

Gordon, P.M., Sontheimer, E.J., and Piccirilli, J.A. (2000). **Metal ion catalysis during the exon-ligation step of nuclear pre-mRNA splicing: extending the parallels between the spliceosome and group II introns.** *RNA* 6, 199-205.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., and Ast, G. (2006).**Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers.** *Mol Cell* 22, 769-781.

Görnemann, J., Kotovic, K.M., Hujer, K., and Neugebauer, K.M. (2005), **Cotranscriptional Spliceosome Assembly Occurs in a Stepwise Fashion and Requires the Cap Binding Complex.** *Mol Cell* 19, 53-63.

Gouet, P., Courcelle, E., Stuart, D.I. and Metoz, F. (1999). **ESPript: analysis of multiple sequence alignments in PostScript**. *Bioinformatics* 15, 305–308.

Graber, J.H., Cantor, C.R., Mohr, S.C., and Smith, T.F. (1999). **Genomic detection of new yeast pre-mRNA 3'-end-processing signals.** *Nucleic Acids Res* 27, 888-894.

Grabowski, P.J. (2007). **RNA-binding proteins switch gears to drive alternative splicing in neurons**. *Nat Struct Mol Biol* 14, 577-579.

de la Grange, P., Dutertre, M., Correa, M., and Auboeuf, D. (2007). **A new advance in alternative splicing databases: from catalogue to detailed analysis of regulation of expression and function of human alternative splicing variants.** *BMC Bioinformatics* 8, 180.

Graveley, B.R., and Maniatis, T. (1998). **Arginine/serine-rich domains of SRproteins can function as activators of pre-mRNA splicing**. *Mol Cell* 1, 765-771.

Graveley, B.R., Hertel, K.J., and Maniatis, T. (1999). **SR proteins are 'locators' of the RNA splicing machinery.** *Curr Biol* 9, R6-R7.

Graveley, B.R. (2000). **Sorting out the complexity of SR protein functions**. *RNA* 6, 1197-1211.

Graveley, B.R., Hertel, K.J., and Maniatis, T. (2001). **The role of U2AF35 and U2AF65 in enhancer-dependent splicing.** *RNA* 7, 806 –818.

Graveley, B.R. (2001). **Alternative splicing: increasing diversity in the proteomic world**. *Trends Genet* 17, 100-107.

Graveley, B.R. (2005). **Coordinated control of splicing and translation.** *Nat Struct Mol Biol* 12, 1022-1023.

Grellscheid, S.N., and Smith, C.W. (2006). **An apparent pseudo-exon acts both as an alternative exon that leads to nonsense-mediated decay and as a zero-length exon.** *Mol Cell Biol* 26, 2237-2246.

Grün, D., Wang, Y-L., Langenberger, D., Gunsalus, K.C., and Rajewsky, N. (2005). **microRNA Target Predictions across Seven Drosophila Species and Comparison to Mammalian Targets.** *PLoS Comput Biol* 1, e13.

Guthrie, C., and Patterson, B. (1988). **Spliceosomal snRNAs.** *Annu Rev Genet* 22, 387-419. Review.

Hall, S.L., and Padgett, R.A. (1996). **Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns.** *Science* 271, 1716-1718.

Han, K., Yeo, G., An, P., Burge, C.B., and Grabowski, P.J. (2005). **A combinatorial code for splicing silencing: UAGG and GGGG motifs.** *PLoS Biol* 3, e158.

Handa, N., Nureki, O., Kurimoto, K., Kim, I., Sakamoto, H., Shimura, Y., Muto, Y., and Yokoyama, S. (1999). **Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein**. *Nature* 398, 579-585.

He, L. and Hannon, G.J. (2004). **MicroRNAs: small RNAs with a big role in gene regulation**. *Nat Rev Genet* 5, 522-531.

Hedges, S.B. (2002). **The origin and evolution of model organisms**. *Nat Rev Genet* 3, 838-849.

Hertel, K.J., Lynch, K.W., Hsiao, E.C., Liu, E.H., and Maniatis, T. (1996). **Structural and functional conservation of the Drosophila doublesex splicing enhancer repeat elements**. *RNA* 2, 969-981.

Hertel, K.J., Lynch, K.W., and Maniatis, T. (1997). **Common themes in the function of**

transcription and splicing enhancers. *Curr Opin Cell Biol* 9, 350 –357.

Hertel, K.J., and Graveley, B.R. (2005). **RS domains contact the pre-mRNA throughout spliceosome assembly**. *Trends Biochem Sci* 30, 115-118. Review.

Hertel, J., Hofacker, I.L., and Stadler, P.F. (2008). **SnoReport: computational identification of snoRNAs with unknown targets.** *Bioinformatics* 24, 158-64.

Hicks, M.J., Yang, C.R., Kotlajich, M.V., and Hertel, K.J. (2006). **Linking splicing to Pol II transcription stabilizes pre-mRNAs and influences splicing patterns**. *PLoS Biol* 4, e147.

Hiller, M., Zhang, Z., Backofen, R., and Stamm, S. (2007). **Pre-mRNA Secondary Structures Influence Exon Recognition.** *PLoS Genet* 3, e204.

Hirose, Y. and Manley, J.L. (2000). **RNA polymerase II and the integration of nuclear events.** *Genes Dev* 14, 1415-1429.

Hirose, T., Ideue, T., Nagai, M., Hagiwara, M., Shu, M.D., and Steitz, J.A. (2006). **A spliceosomal intron binding protein, IBP160, links position-dependent assembly of intron-encoded box C/D snoRNP to pre-mRNA splicing.** *Mol Cell* 23, 673-684.

Ho, T.H., Charlet-B, N., Poulos, M.G., Singh, G., Swanson, M.S., and Cooper, T.A. (2004). **Muscleblind proteins regulate alternative splicing.** *EMBO J* 23, 3103-3112.

Hodgkin, J. (2001). **What does a worm want with 20,000 genes?** *Genome Biol* 2(11).

Hoffman, B.E., and Grabowski, P.J. (1992). **U1 snRNP targets an essential splicing factor, U2AF65, to the 3' splice site by a network of interactions spanning the exon.** *Genes Dev* 6, 2554-2568.

Horiuchi, T., and Aigaki, T. (2006). **Alternative trans-splicing: a novel mode of pre-mRNA processing.** *Biol Cell* 98, 135-140. Review.

Horowitz, D.S., Lee, E.J., Mabon, S.A., and Misteli, T. (2002). **A cyclophilin functions in pre-mRNA splicing.** *EMBO J* 21, 470-480.

House, A.E. and Lynch, K.W. (2008). **Regulation of Alternative Splicing: More than Just the ABCs**. *J Biol Chem* 283, 1217-1221.

Howe, K.J., Kane, C.M., and Ares, M.Jr. (2003). **Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae.** *RNA* 9, 993-1006.

Hu, J., Lutz, C.S., Wilusz, J., and Tian, B. (2005). **Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation.** *RNA* 11**,** 1485-1493.

Hughes, T.A. (2006). **Regulation of gene expression by alternative untranslated regions**. *Trends Genet* 22, 119-122. Review.

Huh, G.S. and Hynes, R.O. (1994). **Regulation of alternative pre-mRNA splicing by a novel repeated hexanucleotide element**. *Genes & Dev* 8, 1561-1574.

Hui, J., Hung, L.H., Heiner, M., Schreiner, S., Neumüller, N., Reither, G., Haas, S.A., and Bindereif, A. (2005). **Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing.** *EMBO J* 24, 1988-1998.

Hung, L.H., Heiner, M., Hui, J., Schreiner, S., Benes, V., and Bindereif, A. (2007). **Diverse roles of hnRNP L in mammalian mRNA processing: A combined microarray and RNAi analysis.** *RNA* 14, 284-296.

Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Bálint, E., Tuschl, T., and Zamore, P.D. (2001). **A cellular function for the RNAinterference enzyme Dicer in the maturation of the let-7 small temporal RNA**. *Science* 293, 834–838.

Izquierdo, J.M., Majós, N., Bonnal, S., Martínez, C., Castelo, R., Guigó, R., Bilbao, D., and Valcárcel, J. (2005). **Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition.** *Mol Cell* 19, 475-484.

Jacquenet, S., Ropers, D., Bilodeau, P.S., Damier, L., Mougin, A., Stoltzfus, C.M., and Branlant, C. (2001). **Conserved stem-loop structures in the HIV-1 RNA region containing the A3 3' splice site and its cis-regulatory element: possible involvement in RNA splicing**. *Nucleic Acids Res* 29, 464-478.

Jankovic, L., Efremov, G.D., Petkov, G., Kattamis, C., George, E., Yang, K.G., Stoming, T.A., and Huisman, T.H. (1990). **Two novel polyadenylation mutations leading to beta(+)-thalassemia.** *Br J Haematol* 75, 122-126.

Jensen, K.B., Musunuru, K., Lewis, H.A., Burley, S.K., and Darnell, R.B. (2000). **The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain**. *Proc Natl Acad Sci* 97, 5740-5745.

Jin, Y., Suzuki, H., Maegawa, S., Endo, H., Sugano, S., Hashimoto, K., Yasuda, K., and Inoue, K. (2003). **A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG**. *EMBO J* 22, 905-912.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays**. *Science* 302, 2141-2144.

Jurica, M.S., and Moore, M.J. (2003). **Pre-mRNA splicing: awash in a sea of proteins.** *Mol Cell* 12, 5-14.

Juven-Gershon, T., Hsu, J.Y., Theisen, J.W., and Kadonaga, J.T. (2008). **The RNA polymerase II core promoter - the gateway to transcription.** *Curr Opin Cell Biol* Apr 22.

Kadener, S., Cramer, P., Nogués, G., Cazalla, D., de la Mata, M., Fededa, J.P., Werbajh, S.E., Srebrow, A., and Kornblihtt, A.R. (2001). **Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing.** *EMBO J* 20, 5759–5768.

Kadener, S., Fededa, J.P., Rosbash, M., and Kornblihtt, A.R. (2002). **Regulation of alternative splicing by a transcriptional enhancer through RNA pol II elongation.** *Proc Natl Acad Sci* U S A 99, 8185-8190.

Kalnina, Z., Zayakin, P., Silina, K., and Line, A. (2005). **Alterations of pre-mRNA splicing in cancer.** *Genes Chromosomes Cancer* 42, 342-357. Review.

Kambach, C., Walke, S., and Nagai, K. (1999a). **Structure and assembly of the spliceosomal small nuclear ribonucleoprotein particles.** *Curr Opin Struct Biol* 9, 222-230. Review.

Kambach, C., Walke, S., Young, R., Avis, J.M., de la Fortelle, E., Raker, V.A., Lührmann, R., Li, J., and Nagai, K. (1999b). **Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs.** *Cell* 96, 375-387.

Kan, Z., Rouchka, E.C., Gish, W.R. and States, D.J. (2001). **Gene structure prediction and alternative splicing analysis using genomically aligned ESTs**. *Genome Res* 11, 889-900.

Kanadia, R.N., Shin, J., Yuan, Y., Beattie, S.G., Wheeler, T.M., Thornton, C.A., and Swanson, M.S. (2006). **Reversal of RNA missplicing and myotonia after muscleblind overexpression in a mouse poly(CUG) model for myotonic dystrophy.** *Proc Natl Acad Sci U S A* 103, 11748-11753.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermuller, J., Hofacker, I.L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., Sementchenko, V., Tammana, H., and Gingeras, T.R. (2007). **Rna maps reveal new rna classes and a possible function for pervasive transcription**. *Science* 316,1484–1488.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D. and Kent, W.J. (2004). **The UCSC Table Browser data retrieval tool**. *Nucleic Acids Res* 32, D493-D496.

Karolchik, D., Kuhn, R.M., Baertsch, R., Barber, G.P., Clawson, H., Diekhans, M., Giardine, B., Harte, R.A., Hinrichs, A.S., Hsu, F., Kober, K.M., Miller, W., Pedersen, J.S., Pohl, A., Raney, B.J., Rhead, B., Rosenbloom, K.R., Smith, K.E., Stanke, M., Thakkapallayil, A., Trumbower, H., Wang, T., Zweig, A.S., Haussler, D., and Kent, W.J. (2008). **The UCSC Genome Browser Database: 2008 update.** *Nucleic Acids Res* 36(Database issue), D773-779.

Kashima, T., Rao, N., and Manley, J.L. (2007). **An intronic element contributes to splicing repression in spinal muscular atrophy.** *Proc Natl Acad Sci* 104, 3426-3431.

Kataoka, N., Yong, J., Kim, V.N., Velazquez, F., Perkinson, R.A., Wang, F., and Dreyfuss, G. (2000). **Pre-mRNA splicing imprints mRNA in the nucleus with a novel RNA-binding protein that persists in the cytoplasm.** *Mol Cell* 6, 673-682.

Kazazian, H.H. Jr. (2004). **Mobile elements: drivers of genome evolution.** *Science* 303, 1626-1632. Review.

Kelley, R.L., Wang, J., Bell, L., and Kuroda, M.I. (1997). **Sex lethal controls dosage compensation in Drosophila by a non-splicing mechanism**. *Nature* 387, 195-199.

Kent, W.J. (2002). **BLAT--the BLAST-like alignment tool.** *Genome Res* 12, 656-664.

Kent, O.A., Reayi, A., Foong, L., Chilibeck, K.A., MacMillan, A.M. (2003). **Structuring of the 3' splice site by U2AF65**. *J Biol Chem* 278, 50572-50577.

Kent, O.A., Ritchie, D.B. and Macmillan, A.M. (2005). **Characterization of a U2AF-Independent Commitment Complex (E') in the Mammalian Spliceosome Assembly Pathway.** *Mol Cell Biol* 25, 233-240.

Kielkopf, C.L., Rodionova, N.A., Green, M.R., and Burley, S.K. (2001). **A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer.** *Cell* 106, 595-605.

Kielkopf, C.L., Lücke, S., and Green, M.R. (2004). **U2AF homology motifs: protein recognition in the RRM world.** *Genes Dev* 18, 1513-1526. Review.

Kim, V.N. (2005). **MicroRNA biogenesis: coordinated cropping and dicing**. *Nat Rev Mol Cell Biol* 6, 376-385.

Kim, T.H., Barrera, L.O., Qu, C., Van Calcar, S., Trinklein, N.D., Cooper, S.J., Luna, R.M., Glass, C.K., Rosenfeld, M.G., Myers, R.M., and Ren, B. (2005a). **Direct isolation and identification of promoters in the human genome**. *Genome Res* 15, 830–839.

Kim, D.H., Langlois, M.A., Lee, K.B., Riggs, A.D., Puymirat, J., and Rossi, J.J. (2005b). **HnRNP H inhibits nuclear export of mRNA containing expanded CUG repeats and a distal branch point sequence.** *Nucleic Acids Res* 33, 3866-3874.

Kim, E., Magen, A., and Ast, G. (2007a). **Different levels of alternative splicing among eukaryotes.** *Nucleic Acids Res* 35, 125-131.

Kim, N., Alekseyenko, A.V., Roy, M., and Lee, C. (2007b). **The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species**. *Nucleic Acids Res* 35(Database issue), D93- D98.

Kimura, T., Nakamori, M., Lueck, J.D., Pouliquin, P., Aoike, F., Fujimura, H., Dirksen, R.T., Takahashi, M.P., Dulhunty, A.F., and Sakoda, S. (2005). **Altered mRNA splicing of the skeletal muscle ryanodine receptor and sarcoplasmic/endoplasmic reticulum Ca2+-ATPase in myotonic dystrophy type 1.** *Hum Mol Genet* 14, 2189-2200.

Kiss, T. (2001). **Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs.** *EMBO J* 20, 3617-3622. Review.

Kiss, T. (2002). **Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions.** *Cell* 109, 145-148. Review.

Kiss, T. (2004). **Biogenesis of small nuclear RNPs.** *J Cell Sci* 117, 5949-5951. Review.

Kiss, T. (2006). **SnoRNP biogenesis meets Pre-mRNA splicing.** *Mol Cell* 23, 775-776. Review.

Koch, K.S., and Leffert, H.L. (1998). **Giant hairpins formed by CUG repeats in myotonic dystrophy messenger RNAs might sterically block RNA export through nuclear pores.** *J Theor Biol* 192, 505-514.

Kochetov, A.V., Palyanov, A., Titov, I.I., Grigorovich, D., Sarai, A., and Kolchanov, N.A. (2007). **AUG_hairpin: prediction of a downstream secondary structure influencing the recognition of a translation start site.** *BMC Bioinformatics* 8, 318.

Kol, G., Lev-Maor, G., and Ast, G. (2005). **Human–mouse comparative analysis reveals that branch-site plasticity contributes to splicing regulation**. *Hum Mol Genet* 14, 1559–1568.

Konarska, M.M., and Query, C.C. (2005). **Insights into the mechanisms of splicing: more lessons from the ribosome.** *Genes Dev* 19, 2255-2260. Review.

König, H., Matter, N., Bader, R., Thiele, W., and Müller, F. (2007). **Splicing Segregation: The Minor Spliceosome Acts outside the Nucleus and Controls Cell Proliferation**. *Cell* 131, 718-729.

Kornblihtt, A.R., de la Mata, M., Fededa, J.P., Munoz, M.J., and Nogues, G. (2004). **Multiple links between transcription and splicing.** *RNA* 10, 1489-1498. Review.

Kornblihtt, A.R. (2005). **Promoter usage and alternative splicing.** *Curr Opin Cell Biol* 17, 262-8. Review.

Kozak, M. (1986) **Bifunctional messenger RNAs in eukaryotes.** *Cell* 47**,** 481-483.

Kozak, M. (1987a). **An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs.** *Nucleic Acids Res* 15, 8125-8148. Review.

Kozak, M. (1987b). **At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells**. *J Mol Biol* 196, 947-950.

Kozak, M. (1989). **The scanning model for translation: an update**. *J Cell Biol* 108, 229-241.

Kozak, M. (1990). **Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes.** *Proc Natl Acad Sci U S A* 87, 8301-8305.

Kozak, M. (1995). **Adherence to the first-AUG rule when a second AUG codon follows closely upon the first.** *Proc Natl Acad Sci U S A* 92, 7134.

Kozak, M. (1997). **Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not gener ally affected by the nucleotides in positions +5 and +6**. *EMBO J* 16, 2482-2492.

Kozak, M. (2001). **Constraints on reinitiation of translation in mammals.** *Nucleic Acids Res* 29, 5226-5232.

Kozak, M. (2002). **Pushing the limits of the scanning mechanism for initiation of translation**. *Gene* 299**,** 1-34. Review

Kozak, M. (2005). **A second look at cellular mRNA sequences said to function as internal ribosome entry sites**. *Nucleic Acids Res* 33**,** 6593-6602. Review

Kuhn, A.N., Reichl, E.M., and Brow, D.A. (2002). **Distinct domains of splicing factor Prp8 mediate different aspects of spliceosome activation.** *Proc Natl Acad Sci U S A* 99, 9145-9149.

Kuyumcu-Martinez, N.M., and Cooper, T.A. (2006). **Misregulation of alternative splicing causes pathogenesis in myotonic dystrophy.** *Prog Mol Subcell Biol* 44, 133-159. Review.

Kwon, Y.T., Kashina, A.S., and Varshavsky, A. (1999). **Alternative Splicing Results in Differential Expression, Activity, and Localization of the Two Forms of Arginyl-tRNA-Protein Transferase, a Component of the N-End Rule Pathway.** *Mol Cell Biol* 19, 182-193.

Ladd, A.N., and Cooper, T.A. (2002). **Finding signals that regulate alternative splicing in the post-genomic era.** Genome Biol 3, reviews0008.

Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., and Tuschl, T. (2003). **New microRNAs from mouse and human.** *RNA* 9, 175-179.

Lai, M.C., Lin, R.I., and Tarn, W.Y. (2003). **Differential effects of hyperphosphorylation on splicing factor SRp55.** *Biochem J* 371, 937-945.

Lallena, M.J., Chalmers, K.J., Llamazares, S., Lamond, A.I., and Valcarcel, J. (2002). **Splicing Regulation at the Second Catalytic Step by Sex-lethal Involves 3' Splice Site Recognition by SPF45.** *Cell* 109, 285-296.

Landry, J.R., Mager, D.L., and Wilhelm, B.T. (2003). **Complex controls:The role of alternative promoters in mammalian genomes**. *Trends Genet* 19, 640–648.

Lareau, L.F., Green, R.E., Bhatnagar, R.S. and Brenner S.E. (2004). **The evolving roles of alternative splicing**. *Curr Opin Struct Biol* 14, 273–282.

Laurencikiene, J., Källman, A.M., Fong, N., Bentley, D.L., and Ohman, M. (2006). **RNA editing and alternative splicing: the importance of co-transcriptional coordination.** *EMBO Rep* 7, 303-307.

Le Hir, H., Izaurralde, E., Maquat, L.E. and Moore, M.J. (2000a). **The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions.** *EMBO J* 19, 6860-6869.

Le Hir, H., Moore, M.J. and Maquat, L.E. (2000b). **Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon-exon junctions.** *Genes Dev* 14, 1098-1108.

Le Hir, H., Gatfield, D., Braun, I.C., Forler, D., and Izaurralde, E. (2001a). **The protein Mago provides a link between splicing and mRNA localization.** *EMBO Rep* 2, 1119-1124.

Le Hir, H., Gatfield, D., Izaurralde, E., and Moore, M.J. (2001b). **The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay.** *EMBO J* 20, 4987-4997.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., and Kim, V.N. (2003). **The nuclear RNase III Drosha initiates microRNA processing.** *Nature* 425, 415-419.

Lee, J.Y., Yeh, I., Park, J.Y., and Tian, B. (2007a). **PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes.** *Nucleic Acids Res* (Database issue) 35, D165-168.

Lee, Y., Lee, Y., Kim, B., Shin, Y., Nam, S., Kim, P., Kim, N., Chung, W.H., Kim, J., and Lee, S. (2007b). **ECgene: an alternative splicing database update.** *Nucleic Acids Res* 35(Database issue), D99-103

Legendre, M., and Gautheret, D. (2003). **Sequence determinants in human polyadenylation site selection.** *BMC Genomics* 4, 7.

Legendre, M., Lambert, A., and Gautheret, D. (2005). **Profile-based detection of microRNA precursors in animal genomes.** *Bioinformatics* 21, 841-845.

Leroy, O., Wang, J., Maurage, C.A., Parent, M., Cooper, T., Buée, L., Sergeant, N., Andreadis, A., and Caillet-Boudin, M.L. (2006). **Brain-specific change in alternative splicing of Tau exon 6 in myotonic dystrophy type 1.** *Biochim Biophys Acta* 1762, 460-467.

Lewis, B.P., Green, R.E., and Brenner, S.E. (2003). **Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans**. *PNAS* **100**, 189-192.

Libri, D., Piseri, A., and Fiszman, M.Y. (1991). **Tissue-Specific Splicing in Vivo of the ß-Tropomyosin Gene: Dependence on an RNA Secondary Structure**. *Science* 252, 1842-1845.

Lim, L.P., and Burge, C.B. (2001). **A computational analysis of sequence features involved in recognition of short introns.** *Proc Natl Acad Sci U S A* 98, 11193-11198.

Lim, S.R., and Hertel, K.J. (2004). **Commitment to splice site pairing coincides with A complex formation.** *Mol Cell* 15, 477-483.

Lin, X., Miller, J.W., Mankodi, A., Kanadia, R.N., Yuan, Y., Moxley, R.T., Swanson, M.S., and Thornton, C.A. (2006). **Failure of MBNL1-dependent post-natal splicing transitions in myotonic dystrophy.** *Hum Mol Genet* 15, 2087-2097.

Lipscombe, D. (2005). **Neuronal proteins custom designed by alternative splicing**. *Curr Opin Neurobiol* 15, 358-363.

Liu, H.X., Goodall, G.J., Kole, R., and Filipowicz, W. (1995). **Effects of secondary structure on pre-mRNA splicing: hairpins sequestering the 5' but not the 3' splice site inhibit intron processing in Nicotiana plumbaginifolia.** *EMBO J* 14, 377-388.

Liu, H.X., Zhang, M. and Krainer, A.R. (1998). **Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins.** *Genes Dev* 12, 1998–2012.

Liu, H.X., Chew, S.L., Cartegni, L., Zhang, M.Q., and Krainer, A.R. (2000). **Exonic splicing enhancer motif recognized by human SC35 under splicing conditions.** *Mol Cell Biol* 20, 1063-1071.

Lopez, A. J. (1998). **Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation**. *Annu Rev Genet* 32, 279-305.

Lueck, J.D., Mankodi, A., Swanson, M.S., Thornton, C.A., and Dirksen, R.T. (2007). **Muscle chloride channel dysfunction in two mouse models of myotonic dystrophy.** *J Gen Physiol* 129, 79-94.

Lund, E., Guttinger, S., Calado, A., Dahlberg, J. E. and Kutay, U. (2004). **Nuclear export of microRNA precursors**. *Science* 303, 95–98.

Lynch, K.W., and Maniatis, T. (1996). **Assembly of specific SR protein complexes on distinct regulatory elements of the Drosophila doublesex splicing enhancer**. *Genes Dev* 10, 2089-2101.

Lynch, K.W. (2004). **Consequences of regulated pre-mRNA splicing in the immune system.** *Nat Rev Immunol* 4, 931-940.

Lynch, K.W. (2006). **Cotranscriptional splicing regulation: it's not just about speed.** *Nat Struct Mol Biol* 13, 952-953.

Madhani, H.D., and Guthrie, C.(1992). **A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome.** *Cell* 71, 803-817.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2007). **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 35(Database issue), D26-D31.

Makarova, O.V., Makarov, E.M., Urlaub, H., Will, C.L., Gentzel, M., Wilm, M., and Lührmann, R. (2004). **A subset of human 35S U5 proteins, including Prp19, function prior to catalytic step 1 of splicing.** *EMBO J* 23, 2381-2391.

Maniatis, T., and Reed, R. (2002). **An extensive network of coupling among gene expression machines.** *Nature* 416, 499-506. Review.

Mankodi, A., Urbinati, C.R., Yuan, Q.P., Moxley, R.T., Sansone, V., Krym, M., Henderson, D., Schalling, M., Swanson, M.S., and Thornton, C.A. (2001). **Muscleblind localizes to nuclear foci of aberrant RNA in myotonic dystrophy types 1 and 2.** *Hum Mol Genet* 10, 2165-2170.

Mankodi, A., Takahashi, M.P., Jiang, H., Beck, C.L., Bowers, W.J., Moxley, R.T., Cannon, S.C., and Thornton, C.A. (2002). **Expanded CUG repeats trigger aberrant splicing of ClC-1 chloride channel pre-mRNA and hyperexcitability of skeletal muscle in myotonic dystrophy.** *Mol Cell* 10, 35-44.

Manley, J.L., and Tacke, R. (1996). **SR proteins and splicing control.** *Genes Dev* 10, 1569-1579. Review.

Maquat, L.E. (2004). **Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics.** *Nat Rev Mol Cell Biol* 5, 89-99. Review.

Margolis, J.M., Schoser, B.G., Moseley, M.L., Day, J.W., and Ranum, L.P. (2006). **DM2 intronic expansions: evidence for CCUG accumulation without flanking sequence or effects on ZNF9 mRNA processing or protein expression.** *Hum Mol Genet* 15, 1808-1815.

Maroney, P.A., Romfo, C.M., and Nilsen, T.W. (2000). **Functional recognition of 5' splice site by U4/U6.U5 tri-snRNP defines a novel ATP-dependent step in early spliceosome assembly.** *Mol Cell* 6, 317-328.

Maroney, P.A., Yu, Y., Fisher, J., and Nilsen, T.W. (2006). **Evidence that microRNAs are associated with translating messenger RNAs in human cells.** *Nat Struct Mol Biol* 13, 1102-1107.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). **Transcriptional regulatory elements in the human genome.** *Annu Rev Genomics Hum Genet* 7, 29-59.

de la Mata, M., Alonso, C.R., Kadener, S., Fededa, J.P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D., and Kornblihtt, A.R. (2003). **A slow RNA polymerase II affects alternative splicing in vivo.** *Mol Cell* 12, 525-532.

de la Mata, M. and Kornblihtt, A.R. (2006). **RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20.** *Nat Struct Mol Biol* 13, 973-980.

Matlin, A.J., Clark, F., and Smith, C.W. (2005). **Understanding alternative splicing: towards a cellular code.** *Nat Rev Mol Cell Biol* 6, 386-398. Review.

Matsushita, K., Tomonaga, T., Shimada, H., Shioya, A., Higashi, M., Matsubara, H., Harigaya, K., Nomura, F., Libutti, D., Levens, D. and Ochiai T. (2006). **An essential role of alternative splicing of c-myc suppressor FUSE-binding protein-interacting repressor in carcinogenesis**. *Cancer Res* 66, 1409–1417.

McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S.D., Wickens, M., and Bentley, D.L. (1997). **The C-terminal domain of RNA polymerase II couples mRNA processing to transcription.** *Nature* 385, 357-361.

McCullough, A.J., and Berget, S.M. (1997). **G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection.** *Mol Cell Biol* 17, 4562-4571.

McCullough, A.J., and Berget, S.M. (2000). **An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5' splice sites.** *Mol Cell Biol* 20, 9225-9235.

Merendino, L., Guth, S., Bilbao, D., Martinez, C., and Valcarcel, J. (1999). **Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3' splice site AG.** *Nature* 402, 838-841.

Meshorer, E., and Misteli, T. (2005). **Splicing misplaced.** *Cell* 122, 317-318. Review.

Miller, J.W., Urbinati, C.R., Teng-Umnuay, P., Stenberg, M.G., Byrne, B.J., Thornton, C.A., and Swanson, M.S. (2000). **Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy.** *EMBO J* 19, 4439-4448.

Minovitsky, S., Gee, S.L., Schokrpur, S., Dubchak, I., and Conboy, J.G. (2005). **The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons.** *Nucleic Acids Res* 33, 714-724.

Misteli, T., and Spector, D.L. (1999). **RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo.** *Mol Cell* 3, 697-705.

Misteli, T. (2000). **Cell biology of transcription and pre-mRNA splicing: nuclear architecture meets nuclear function.** *J Cell Sci* 113,1841-1849. Review.

Modafferi, E.F. and Black, D.L. (1999). **Combinatorial control of a neuron-specific exon**. *RNA* 5, 687-706.

Modrek, B., and Lee, C. (2002). **A genomic view of alternative splicing**. *Nat Genet* 30, 13-19.

Moen, P.T. Jr., Smith, K.P., and Lawrence, J.B. (1995). **Compartmentalization of specific pre-mRNA metabolism: an emerging view.** *Hum Mol Genet* 4 Spec No:1779-1789. Review.

Moore, M.J., Schwartzfarb, E.M., Silver, P.A., and Yu, M.C. (2006). **Differential recruitment of the splicing machinery during transcription predicts genome-wide patterns of mRNA splicing.** *Mol Cell* 24, 903-915.

Mollet, I., Barbosa-Morais, N.L., Andrade, J., and Carmo-Fonseca, M. (2006). **Diversity of human U2AF splicing factors.** *FEBS J* 273, 4807-4816.

Mori, D., Sasagawa, N., Kino, Y., and Ishiura, S. (2008). **Quantitative Analysis of CUG-BP1 Binding to RNA Repeats.** *J Biochem* 143, 377-383.

Moucadel, V., Lopez, F., Ara, T., Benech, P., and Gautheret, D. (2007). **Beyond the 3' end: experimental validation of extended transcript isoforms.** *Nucleic Acids Res* 35, 1947-1957.

Nakahata, S. and Kawamoto, S. (2005). **Tissue-dependent isoforms of mammalian Fox-1 homologs are associated with tissue-specific splicing activities.** *Nucleic Acids Res* 33, 2078-2089.

Napierała, M., and Krzyzosiak, W.J. (1997). **CUG repeats present in myotonin kinase RNA form metastable "slippery" hairpins.** *J Biol Chem* 272, 31079-31085.

Nasim, F.U., Hutchison, S., Cordeau, M., and Chabot, B. (2002). **High affinity hnRNP A1 binding sites and duplex-forming inverted repeats have similar effects on 59 splice site selection in support of a common looping out and repression mechanism.** *RNA* 8, 1078-1089.

Nilsen, T. W. (1998) in *RNA Structure and Function* (eds Simons, R. W. & Grunberg-Manago, M.) 279–307 (Cold Spring Harbor Laboratory Press, NY).

Nilsen, T.W. (2000). **The case for an RNA enzyme.** *Nature* 408, 782-783.

Nilsen, T.W. (2002). **The spliceosome: no assembly required?** *Mol Cell* 9, 8-9.

Nilsen, T.W. (2003). **The spliceosome: the most complex macromolecular machine in the cell?** *Bioessays* 25, 1147-1149. Review.

Nilsen, T.W. (2007). **Mechanisms of microRNA-mediated gene regulation in animal cells.** *Trends Genet* 23, 243-249.

Ng Kwang Loong, S., and Mishra, S.K. (2007). **Unique folding of precursor microRNAs: quantitative evidence and implications for de novo identification**. *RNA* 13, 170-187.

Nogués, G., Kadener, S., Cramer, P., Bentley, D., and Kornblihtt, A.R. (2002). **Transcriptional activators differ in their abilities to control alternative splicing.** *J Biol Chem* 277, 43110-43114.

Nott, A., Le Hir, H., and Moore, M.J. (2004). **Splicing enhances translation in mammalian cells: an additional function of the exon junction complex.** *Genes Dev* 18, 210-222.

Ohler, U., Yekta, S., Lim, L.P., Bartel, D.P., and Burge, C.B. (2004). **Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification.** *RNA* 10, 1309-1322.

Okoniewski, M.J., Hey, Y., Pepper, S.D., and Miller, C.J. (2007**). High correspondence between Affymetrix exon and standard expression arrays.** *BioTechniques* 24,181-185.

Pacheco, T.R., Gomes, A.Q., Barbosa-Morais, N.L., Benes, V., Ansorge, W., Wollerton, M., Smith, C.W., Valcárcel, J., and Carmo-Fonseca, M. (2004). **Diversity of vertebrate splicing factor U2AF35: identification of alternatively spliced U2AF1 mRNAS.** *J Biol Chem* 279, 27039-27049.

Pacheco, T.R., Coelho, M.B., Desterro, J.M., Mollet, I., and Carmo-Fonseca, M. (2006**). In vivo requirement of the small subunit of U2AF for recognition of a weak 3' splice site**. *Mol Cell Biol* 26, 8183-8190.

Pagani, F., Stuani, C., Zuccato, E., Kornblihtt, A.R., and Baralle, F.E. (2003). **Promoter architecture modulates CFTR exon 9 skipping.** *J Biol Chem* 278, 1511-1517.

Pascual, M., Vicente, M., Monferrer, L., and Artero, R. (2006). **The Muscleblind family of proteins: an emerging class of regulators of developmentally programmed alternative splicing.** *Differentiation* 74, 65-80. Review.

Patel, A.A., McCarthy, M., and Steitz, J.A. (2002). **The splicing of U12-type introns can be a rate-limiting step in gene expression.** *EMBO J* 21, 3804-3815.

Patel, A.A., and Steitz, J.A. (2003). **Splicing double: insights from the second spliceosome.** *Nat Rev Mol Cell Biol* 4, 960-970.

Paul, S., Dansithong, W., Kim, D., Rossi, J., Webster, N.J., Comai, L., and Reddy, S. (2006). **Interaction of muscleblind, CUG-BP1 and hnRNP H proteins in DM1-associated aberrant IR splicing.** *EMBO J* 25, 4271-4283.

Pauws, E., van Kampen, A.H., van de Graaf, S.A., de Vijlder, J.J., and Ris-Stalpers, C. (2001). **Heterogeneity in polyadenylation cleavage sites in mammalian mRNA sequences: implications for SAGE analysis.** *Nucleic Acids Res* 29, 1690-1694.

Pelisch, F., Blaustein, M., Kornblihtt, A.R., and Srebrow, A. (2005). **Cross-talk between signaling pathways regulates alternative splicing: a novel role for JNK.** *J Biol Chem* 280, 25461-25469.

Philips, A.V., Timchenko, L.T., and Cooper, T.A. (1998). **Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy.** *Science* 280, 737-741.

Pleiss, J.A., Whitworth, G.B., Bergkessel, M., and Guthrie, C. (2007). **Rapid, transcript-specific changes in splicing in response to environmental stress.** *Mol Cell* 27, 928-937.

Pollastri, G., Przybylski, D., Rost, B., and Baldi, P. (2002). **Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles.** *Proteins* 47, 228-235,.

Pozzoli, U. and Sironi, M. (2005). **Silencers regulate both constitutive and alternative splicing events in mammals**. *Cell Mol Life Sci* 62, 1579-1604.

Prasanth, K.V., and Spector, D.L. (2007). **Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum.** *Genes Dev* 21, 11-42. Review.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 35(Database issue), D61- D65.

Query, C.C., Moore, M.J., and Sharp, P.A. (1994). **Branch nucleophile selection in pre-mRNA splicing: evidence for the bulged duplex model.** *Genes Dev* 8, 587-597.

Query, C.C., Strobel, S.A., and Sharp, P.A. (1996). **Three recognition events at the branch-site adenine.** *EMBO J* 15, 1392-1402.

Raker, V.A., Plessel, G., and Lührmann, R. (1996). **The snRNP core assembly pathway: identification of stable core protein heteromeric complexes and an snRNP subcore particle in vitro.** *EMBO J* 15, 2256-2269.

Ranum, L.P., and Cooper, T.A. (2006). **RNA-mediated neuromuscular disorders.** *Annu Rev Neurosci* 29, 259-277. Review.

Rappsilber, J., Ryder, U., Lamond, A.I., and Mann, M. (2002). **Large-scale proteomic analysis of the human spliceosome.** *Genome Res* 12, 1231-1245.

Reddy, R., and Busch, H. (1988). Small nuclear RNAs: RNA sequences, structure and modifications. In *Small nuclear ribonucleoprotein particles* (ed. M.L. Birnstiel), pp.1-37. Springer-Verlag, Berlin.

Reed, R. (1996). **Initial splice-site recognition and pairing during pre-mRNA splicing.** *Curr Opin Genet Dev* 6, 215-220. Review.

Reed, R., and Hurt, E. (2002). **A conserved mRNA export machinery coupled to pre-mRNA splicing.** *Cell* 108, 523-531. Review.

Reed, R. (2003). **Coupling transcription, splicing and mRNA export.** *Curr Opin Cell Biol* 15, 326-331. Review.

Rehmsmeier, M., Steffen, P., Höchsmann, M., and Giegerich, R. (2004). **Fast and effective prediction of microRNA/target duplexes.** *RNA* 10, 1507-1517.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). **The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans.*** *Nature* 403**,** 901-906.

Rigo, F. and Martinson, H.G. (2008). **Functional Coupling of Last-Intron Splicing and 3'-End Processing to Transcription In Vitro: the Poly(A) Signal Couples to Splicing before Committing to Cleavage.** *Mol Cell Biol* 28, 849-862.

Robberson, B.L., Cote, G.J., and Berget, S.M. (1990). **Exon definition may facilitate splice site selection in RNAs with multiple exons**. *Mol Cell Biol* 10, 84-94.

Roberts, G.C., Gooding, C., Mak, H.Y., Proudfoot, N.J., and Smith, C.W. (1998). **Co-transcriptional commitment to alternative splice site selection.** *Nucl Acids Res* 26, 5568-5572.

Rossi, F., Forné, T., Antoine, E., Tazi, J., Brunel, C., and Cathala, G. (1996). **Involvement of U1 small nuclear ribonucleoproteins (snRNP) in 5' splice site-U1 snRNP interaction.** *J Biol Chem* 271, 23985-23991.

Russell, A.G., Charette, J.M., Spencer, D.F., and Gray, M.W. (2006). **An early evolutionary origin for the minor spliceosome.** *Nature* 443, 863-866.

Sacco-Bubulya, P., and Spector, D.L. (2002). **Disassembly of interchromatin granule clusters alters the coordination of transcription and pre-mRNA splicing.** *J Cell Biol* 156, 425-436.

Sanford, J.R., Gray, N.K., Beckmann, K., and Cáceres, J.F. (2004). **A novel role for shuttling SR proteins in mRNA translation.** *Genes Dev* 18, 755-768.

Sanford, J.R., and Caceres, J.F. (2004). **Pre-mRNA splicing: life at the centre of the central dogma.** *J Cell Sci* 117, 6261-6263. Review.

Sanford, J.R., Ellis, J., and Cáceres, J.F. (2005). **Multiple roles of arginine/serine-rich splicing factors in RNA processing.** *Biochem Soc Trans* 33, 443-446. Review.

Sato, S., Nakamura, M., Cho, D.H., Tapscott, S.J., Ozaki, H., and Kawakami, K. (2002). **Identification of transcriptional targets for Six5: implication for the pathogenesis of myotonic dystrophy type 1.** *Hum Mol Genet* 11, 1045-1058.

Savkur, R.S., Philips, A.V., and Cooper, T.A. (2001). **Aberrant regulation of insulin receptor alternative splicing is associated with insulin resistance in myotonic dystrophy.** *Nat Genet* 29, 40-47.

Schaal, T.D., and Maniatis, T. (1999). **Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA.** *Mol Cell Biol* 19, 261-273.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 270, 467-470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., and Davis, R.W. (1996). **Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.** *Proc Natl Acad Sci U S A* 93, 10614-10619.

Schischmanoff, P.O., Yaswen, P., Parra, M.K., Lee G, Chasis JA, Mohandas N, and Conboy J.G. (1997). **Cell shape-dependent regulation of protein 4.1 alternative pre-mRNA splicing in mammary epithelial cells.** *J Biol Chem* 272, 10254-10259.

Schmittgen, T.D., Teske, S., Vessella, R.L., True, L.D., and Zakrajsek, B.A. (2003). **Expression of prostate specific membrane antigen and three alternatively spliced variants of PSMA in prostate cancer patients.** *Int J Cancer* 107, 323-329.

Schmucker, D. and Flanagan, J.G. (2004). **Generation of recognition diversity in the nervous system**. *Neuron* 44, 219-222.

Schneider, C., Will, C.L., Makarova, O.V., Makarov, E.M., and Lührmann, R. (2002). **Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions.** *Mol Cell Biol* 22, 3219-3229.

Schwer, B., and Meszaros, T. (2000). **RNA helicase dynamics in pre-mRNA splicing.** *EMBO J* 19, 6582-6591.

Schwer, B. (2001). **A new twist on RNA helicases: DExH/D box proteins as RNPases.** *Nat Struct Biol* 8, 113-116. Review.

Schwerk, C., and Schulze-Osthoff, K. (2005). **Regulation of apoptosis by alternative pre-mRNA splicing.** *Mol Cell* 19, 1-13. Review.

Sergeant, K.A., Bourgeois, C.F., Dalgliesh, C., Venables, J.P., Stevenin, J., and Elliott, D.J. (2007). **Alternative RNA splicing complexes containing the scaffold attachment factor SAFB2.** *J Cell Sci* 120(Pt 2), 309-319.

Sharma, S., Falick, A.M., and Black, D.L. (2005). **Polypyrimidine tract binding protein blocks the 5' splice site-dependent assembly of U2AF and the prespliceosomal E complex.** *Mol Cell* 19, 485-496.

Sharma, S., Kohlstaedt, L.A., Damianov, A., Rio, D.C., and Black, D.L. (2008). **Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome.** *Nat Struct Mol Biol* 15, 183-191.

Shen, H., J. L. Kan, and M. R. Green. (2004). **Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly**. *Mol Cell* 13, 367-376.

Shen, H., and Green, M.R. (2004). **A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly**. *Mol Cell* 16, 363-373.

Shen, H., and Green, M.R. (2007). **RS domain-splicing signal interactions in splicing of U12-type and U2-type introns.** *Nat Struct Mol Biol* 14, 597-603.

Shepard, J., Reick, M., Olson, S., and Graveley, B.R. (2002). **Characterization of U2AF(6), a splicing factor related to U2AF(35).** *Mol Cell Biol* 22, 221-230.

Sickmier, E.A., Frato, K.E., Shen, H., Paranawithana, S.R., Green, M.R., and Kielkopf, C.L. (2006). **Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65.** *Mol Cell* 23, 49-59.

Silva, A.L., Ribeiro, P., Inácio, A., Liebhaber, S.A., and Romão, L. (2008). **Proximity of the poly(A)-binding protein to a premature termination codon inhibits mammalian nonsense-mediated mRNA decay.** *RNA* 14, 563-576.

Singh, R., Valcárcel, J., and Green, M.R. (1995). **Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins.** *Science* 268, 1173-1176.

Singh, N.K., Singh, N.N., Androphy, E.J., and Singh, R.N. (2006). **Splicing of a critical exon of human Survival Motor Neuron is regulated by a unique silencer element located in the last intron.** *Mol Cell Biol* 26, 1333-1346.

Smale, S.T., and Kadonaga, J.T. (2003). **The RNA polymerase II core promoter.** *Annu Rev Biochem* 72, 449-479.

Smith, C.W., and Valcarcel, J. (2000). **Alternative pre-mRNA splicing: the logic of combinatorial control**. *Trends Biochem Sci* 25, 381-388.

Sontheimer, E. J., Sun, S., and Piccirilli, J. A. (1997). **Metal ion catalysis during splicing of premessenger RNA**. *Nature* 388, 801-805.

Sontheimer, E.J. (2001). **The spliceosome shows its metal.** *Nat Struct Biol* 8, 11-13.

Sorek, R. and Ast, G. (2003). **Intronic sequences flanking alternatively spliced exons are conserved between human and mouse**. *Genome Res* 13, 1631-1637.

Soret, J., and Tazi, J. (2003). **Phosphorylation-dependent control of the pre-mRNA splicing machinery.** *Prog Mol Subcell Biol* 31, 89-126. Review.

Srebrow, A., Blaustein, M., and Kornblihtt, A.R. (2002). **Regulation of fibronectin alternative splicing by a basement membrane-like extracellular matrix.** *FEBS Lett* 514, 285-289.

Staley, J.P., and Guthrie, C. (1998). **Mechanical devices of the spliceosome: motors, clocks, springs, and things.** *Cell* 92, 315–326.

Stamm, S., Riethoven, J-J.M., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Tang, Y., Barbosa-Morais, N.L., and Thanaraj, T.A. (2006). **ASD: a bioinformatics resource on alternative splicing**. *Nucleic Acids Res* 34, D46-D55.

Stark, H., Dube, P., Lührmann, R., and Kastner, B. (2001). **Arrangement of RNA and proteins in the spliceosomal U1 small nuclear ribonucleoprotein particle.** *Nature* 409, 539-542.

Sterner, D.A., Carlo, T., and Berget, S.M. (1996). **Architectural limits on split genes**. *Proc Natl Acad Sci U S A* 93, 15081-15085.

Stevens, S.W., Ryan, D.E., Ge, H.Y., Moore, R.E., Young, M.K., Lee, T.D., and Abelson, J. (2002). **Composition and functional characterization of the yeast spliceosomal penta-snRNP.** *Mol Cell* 9, 31-44.

Strausberg, R.L., and Levy, S. (2007). **Promoting transcriptome diversity.** *Genome Res* 17, 965-968.

Tacke, R., Tohyama, M., Ogawa, S. and Manley, J.L. (1998). **Human Tra2 proteins are sequence-specific activators of pre-mRNA splicing**. *Cell* 93, 139–148.

Tacke, R., and Manley, J.L. (1999). **Determinants of SR protein specificity.** *Curr Opin Cell Biol* 11, 358-362. Review.

Takuno, S., and Innan, H. (2008). **Evolution of complexity in miRNA-mediated gene regulation systems.** *Trends Genet* 24, 56-59.

Tarn, W.Y., Lee, K.R., and Cheng, S.C. (1993a). **Yeast precursor mRNA processing protein PRP19 associates with the spliceosome concomitant with or just after dissociation of U4 small nuclear RNA.** *Proc Natl Acad Sci U S A* 90, 10821-10825.

Tarn, W.Y., Lee, K.R., and Cheng, S.C. (1993b). **The yeast PRP19 protein is not tightly associated with small nuclear RNAs, but appears to associate with the spliceosome after binding of U2 to the pre-mRNA and prior to formation of the functional spliceosome.** *Mol Cell Biol* 13, 1883-1891.

Tarn, W.Y., and Steitz, J.A. (1996a). **A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro**. *Cell* 84, 801-811.

Tarn, W.Y., and Steitz, J.A. (1996b). **Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns.** *Science* 273, 1824-1832.

Tarn, W.Y., and Steitz, J.A. (1997). **Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge.** *Trends Biochem Sci* 22, 132-137. Review.

Tazi, J., Daugeron, M.C., Cathala, G., Brunel, C., and Jeanteur, P. (1992). **Adenosine phosphorothioates (ATP alpha S and ATP tau S) differentially affect the two steps of mammalian pre-mRNA splicing.** *J Biol Chem* 267, 4322-4326.

Teigelkamp, S., McGarvey, M., Plumpton, M., and Beggs, J.D. (1994). **The splicing factor PRP2, a putative RNA helicase, interacts directly with pre-mRNA.** *EMBO J* 13, 888-897.

Tian, M., and Maniatis, T. (1992). **Positive control of pre-mRNA splicing in vitro.** *Science* 256, 237-240 .

Tian, H., and Kole, R. (1995). **Selection of novel exon recognition elements from a pool of random sequences.** *Mol Cell Biol* 15, 6291-6298.

Tian, H., and Kole, R. (2001). **Strong RNA splicing enhancers identified by a modified method of cycled selection interact with SR protein.** *J Biol Chem* 276, 33833-33839.

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

Timchenko, L.T. (1999). **Myotonic dystrophy: the role of RNA CUG triplet repeats.** *Am J Hum Genet* 64, 360-364. Review.

Timchenko, N.A., Cai, Z.J., Welm, A.L., Reddy, S., Ashizawa, T., and Timchenko, L.T. (2001). **RNA CUG repeats sequester CUGBP1 and alter protein levels and activity of CUGBP1.** *J Biol Chem* 276, 7820-7826.

Timchenko, N.A., Patel, R., Iakova, P., Cai, Z.J., Quan, L., and Timchenko, L.T. (2004). **Overexpression of CUG triplet repeat-binding protein, CUGBP1, in mice inhibits myogenesis.** *J Biol Chem* 279, 13129-13139.

Trinklein, N.D., Karaöz, U., Wu, J., Halees, A., Force Aldred, S., Collins, P.J., Zheng, D., Zhang, Z.D., Gerstein, M.B., Snyder, M., Myers, R.M., and Weng, Z. (2007). **Integrated analysis of experimental data sets reveals many novel promoters in 1% of the human genome**. *Genome Res* 17, 720-731.

Tsuritani, K., Irie, T., Yamashita, R., Sakakibara, Y., Wakaguri, H., Kanai, A., Mizushima-Sugano, J., Sugano, S., Nakai, K., and Suzuki, Y. (2007). **Distinct class of putative "non-conserved" promoters in humans: Comparative studies of alternative promoters of human and mouse genes.** *Genome Res* 17, 1005-1014.

Tycowski, K.T., Kolev, N.K., Conrad, N.K., Fok,V., and Steitz, J.A. (2006). **The Ever-Growing World of Small Nuclear Ribonucleoproteins**. In *The RNA World,* Gesteland, R.F., Cech, T.R., and Atkins, J.F. (eds). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 3rd Ed., pp. 327-368.

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

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B.J., and Darnell, R.B. (2006). **An RNA map predicting Nova-dependent splicing regulation.** *Nature* 444, 580-586.

Underwood, J.G., Boutz, P.L., Dougherty, J.D., Stoilov, P., and Black, D.L. (2005). **Homologues of the Caenorhabditis elegans Fox- 1 protein are neuronal splicing regulators in mammals.** *Mol Cell Biol* 25, 10005-10016.

Urlaub, H., Raker, V.A., Kostka, S., and Lührmann, R. (2001). **Sm protein-Sm site RNA interactions within the inner ring of the spliceosomal snRNP core structure.** *EMBO J* 20, 187-196.

Valadkhan, S. (2005). **snRNAs as the catalysts of pre-mRNA splicing.** *Curr Opin Chem Biol* 9, 603-608. Review.

Valcárcel, J., Gaur, R.K., Singh, R., and Green, M.R. (1996). **Interaction of U2AF65 RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA.** *Science* 273, 1706-1709.

Valcárcel, J., and Green, M.R. (1996). **The SR protein family: pleiotropic functions in pre-mRNA splicing.** *Trends Biochem Sci* 21, 296-301. Review.

van Solinge, W.W., Lind, B., van Wijk, R., Hart, H.C., and Kraaijenhagen, R.J. (1996). **Clinical expression of a rare beta-globin gene mutation co-inherited with haemoglobin E-disease.** *Eur J Clin Chem Clin Biochem* 34, 949-954.

Vilardell, J., and Valcárcel, J. (2007). **Powering a two-stroke RNA engine**. *Nat Struct Mol Biol* 14, 574-576.

Villa, T., Pleiss, J.A., and Guthrie, C. (2002). **Spliceosomal snRNAs: Mg2+-Dependent Chemistry at the Catalytic Core?** *Cell* 109, 149-152.

Wagner, E.J., Baraniak, A.P., Sessions, O.M., Mauger, D., Moskowitz, E., and Garcia-Blanco, M.A. (2005). **Characterization of the intronic splicing silencers flanking FGFR2 exon IIIb.** *J Biol Chem* 280, 14017-14027.

Walke, S., Bragado-Nilsson, E., Séraphin, B., and Nagai, K. (2001). **Stoichiometry of the Sm proteins in yeast spliceosomal snRNPs supports the heptamer ring model of the core domain.** *J Mol Biol* 308, 49-58.

Wall, L., Christiansen, T., and Orwant, J. (2000). **Programming Perl.** (3rd Ed). O'Reilly.

Wang, J., and Manley, J.L. (1997). **Regulation of pre-mRNA splicing in metazoa**. *Curr Opin Genet Dev* 7, 205-211.

Wang, C., Chua, K., Seghezzi, W., Lees, E., Gozani, O., and Reed, R. (1998). **Phosphorylation of spliceosomal protein SAP 155 coupled with splicing catalysis.** *Genes Dev* 12, 1409-1414.

Wang, X., Bruderer, S., Rafi, Z., Xue, J., Milburn, P.J., Krämer, A., and Robinson, P.J. (1999). **Phosphorylation of splicing factor SF1 on Ser20 by cGMP-dependent protein kinase regulates spliceosome assembly.** *EMBO J* 18, 4549-4559.

Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M. and Burge, C. B. (2004). **Systematic Identification and Analysis of Exonic Splicing Silencers.** *Cell* 119, 831–845.

Wang, J., Smith, P.J., Krainer, A.R., and Zhang, M.Q. (2005). **Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes.** *Nucleic Acids Res* 33, 5053-5062.

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

Wang, Z., and Burge, C.B. (2008). **Splicing regulation: from a parts list of regulatory elements to an integrated splicing code.** *RNA* 14, 802-813. Review.

Warf, M.B., and Berglund, J.A. (2007). **MBNL binds similar RNA structures in the CUG repeats of myotonic dystrophy and its pre-mRNA substrate cardiac troponin T.** *RNA* 13, 2238-2251.

Wassarman KM, Zhang A, Storz G. (1999). **Small RNAs in Escherichia coli.** *Trends Microbiol* 7, 37-45. Review.

Weiner, A.M.(1993). **mRNA splicing and autocatalytic introns: distant cousins or the products of chemical determinism?** *Cell* 72, 161-164. Review.

Will, C.L., and Lührmann, R. (1997). **Protein functions in pre-mRNA splicing.** *Curr Opin Cell Biol* 9, 320-328. Review.

Will, C.L., Schneider, C., Reed, R., and Lührmann, R. (1999). **Identification of both shared and distinct proteins in the major and minor spliceosomes.** *Science* 284, 2003-2005.

Will, C.L., Schneider, C., MacMillan, A.M., Katopodis, N.F., Neubauer, G., Wilm, M., Lührmann, R., and Query, C.C. (2001). **A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site.** *EMBO J* 20, 4536-4546.

Will, C.L., and Lührmann, R. (2001). **Spliceosomal UsnRNP biogenesis, structure and function.** *Curr Opin Cell Biol* 13, 290-301. Review.

Will, C.L., and Luhrmann, R. (2005). **Splicing of a rare class of introns by the U12-dependent spliceosome.** *Biol Chem* 386, 713-724.

Will, C.L., and Lührmann, R. (2006). **Spliceosome structure and function.** In *The RNA World,* Gesteland, R.F., Cech, T.R., and Atkins, J.F. (eds). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 3rd Ed., pp. 369–400.

Wood, A.J., Schulz, R., Woodfine, K., Koltowska, K., Beechey, C.V., Peters, J., Bourc'his, D,. and Oakey, R.J. (2008). **Regulation of alternative polyadenylation by genomic imprinting**. *Genes Dev* 22, 1141-1146.

Woychik, N.A., and Hampsey, M. (2002). **The RNA polymerase II machinery: structure illuminates function.** *Cell* 108, 453-463. Review.

Wu, S., Romfo, C.M., Nilsen, T.W., and Green, M.R. (1999). **Functional recognition of the 3' splice site AG by the splicing factor U2AF35.** *Nature* 402, 832-835.

Wu, L., and Belasco, J.G. (2008). **Let Me Count the Ways: Mechanisms of Gene Regulation by miRNAs and siRNAs.** *Mol Cell* 29, 1-7. Review.

Xing, Y., and Lawrence, J.B. (1993). **Nuclear RNA tracks: structural basis for transcription and splicing?** *Trends Cell Biol* 3, 346-353.

Xu, Y.Z., Newnham, C.M., Kameoka, S., Huang, T., Konarska, M.M., and Query, C.C. (2004). **Prp5 bridges U1 and U2 snRNPs and enables stable U2 snRNP association with intron RNA.** *EMBO J* 23, 376-385.

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

Yeakley, J.M., Morfin, J.P., Rosenfeld, M.G. and Fu, X.D. (1996). **A complex of nuclear proteins mediates SR protein binding to a purine-rich splicing enhancer**. *Proc Natl Acad Sci USA* 93, 7582–7587.

Yean, S.-L., Wuenschell, G., Termini, J. and Lin, R.-J. (2000). **Metal-ion coordination by U6 small nuclear RNA contributes to catalysis in the spliceosome**. *Nature* 408, 881-884.

Yeo, G., Hoon, S., Venkatesh, B., and Burge, C.B. (2004). **Variation in sequence and organization of splicing regulatory elements in vertebrate genes.** *Proc Natl Acad Sci U S A* 101, 15700-15705.

Yong, J., Wan, L., and Dreyfuss, G. (2004). **Why do cells need an assembly machine for RNA-protein complexes?** *Trends Cell Biol* 14, 226-232.

Yu, Y. T., Scharl, E. C., Smith, C. M. and Steitz,J. A. (1999). **The growing world of small nuclear ribonucleoproteins**. In *The RNA World* (ed.Gesteland, R. F., Cech, T. R. and Atkins, J. F.), pp. 487-524. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

Kan, Z., States, D., and Gish, W. (2002) **Selecting for functional alternative splices in ests.** *Genome Res* 12, 1837-1845.

Zeng, C., and Berget, S.M. (2000). **Participation of the C-Terminal Domain of RNA Polymerase II in Exon Definition during Pre-mRNA Splicing.** *Mol Cell Biol* 20, 8290-8301.

Zhang, X.H. and Chasin, L.A. (2004). **Computational definition of sequence motifs governing constitutive exon splicing**. *Genes Dev* 18, 1241-1250.

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

Zhang, L., Ding, L., Cheung, T.H., Dong, M.Q., Chen, J., Sewell, A.K., Liu, X., Yates, JR. 3rd, and Han, M. (2007). **Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2.** *Mol Cell* 28, 598-613.

Zhao, R., Shen, J., Green, M.R., MacMorris, M., and Blumenthal, T. (2004). **Crystal structure of UAP56, a DExD/H-box protein involved in pre-mRNA splicing and mRNA export.** *Structure* 12, 1373-1381.

Zhou, Z., Licklider, L.J., Gygi, S.P., and Reed, R. (2002). **Comprehensive proteomic**

analysis of the human spliceosome. *Nature* 419, 182–185.

Zhu, J., Mayeda, A., and Krainer, A.R. (2001). **Exon identity established through differential antagonism between exonic splicing silencerbound hnRNP A1 and enhancer-bound SR proteins.** *Mol Cell* 8, 1351-1361.

Zuo, P., and Maniatis, T. (1996). **The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing**. *Genes Dev* 10, 1356-1368.

Websites

W1. **CPAN DBI module** [http://search.cpan.org/~timb/DBI-1.604/DBI.pm]
W2. **CPAN Graph module** [http://search.cpan.org/~jhi/Graph-0.84/lib/Graph.pod]
W3. **Ensembl Genome Browser** [http://www.ensembl.org/index.html]
W4. **Gene Expression Omnibus** [http://www.ncbi.nlm.nih.gov/geo]
W5. **R Development Core Team: R - A Language and Environment for Statistical Computing** [http://www.R-project.org]. Vienna, Austria: *R Foundation for Statistical Computing* 2007
W6. **The UCSC Genome Browser** [http://genome.ucsc.edu/]
W7. **Scratch Protein Predictor** [http://www.ics.uci.edu/~baldig/scratch]
W8. **Sfold** [http://sfold.wadsworth.org/srna.pl]
W9. **Mpsrch** [http://www.ebi.ac.uk/MPsrch/]
W10. **Affymetrix: Exon Array Design Datasheet.** [http://www.affymetrix.com/support/technical/datasheets/exon_arraydesign_datasheet.pdf]

Genomic sequences

The chromosome sequences for the genome assembly of each organism used were obtained from the UCSC Genome Browser [1]: human, *Homo sapiens* [2]; chimp, *Pan troglodytes* [3]; rhesus, *Macaca mulatta* [4]; mouse, *Mus musculus* [5]; rat, *Rattus norvegicus* [6,7]; cow, *Bos taurus* [8]; dog, *Canis lupus familiaris* [9]; chicken, *Gallus gallus* [10]; frog, *Xenopus tropicalis* [11]; zebrafish, *Danio rerio* [12]; fruit fly, *Drosophila melanogaster* [13,14]; nematode worm, *Caenorhabditis elegans* [15]; and sea squirt, *Ciona intestinalis* [16]:

1. **The UCSC Genome Browser** [http://genome.ucsc.edu/]

2. International Human Genome Sequencing Consortium. (2001). **Initial sequencing and analysis of the human genome**. *Nature* 409**,** 860-921

3. The Chimpanzee Sequencing and Analysis Consortium. (2005). **Initial sequence of the chimpanzee genome and comparison with the human genome**. *Nature* 437**,** 69-87

4. Rhesus Macaque Genome Sequencing and Analysis Consortium. (2007). **Evolutionary and Biomedical Insights from the Rhesus Macaque Genome.** *Science* 316, 222-234

5. Mouse Genome Sequencing Consortium. (2002). **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 420, 520-562

6. Havlak, P., Chen, R., Durbin, K.J., Egan, A., Ren, Y., Song, X.Z., Weinstock, G.M., and Gibbs, R.A. (2004). **The Atlas genome assembly system**. *Genome Res* 14, 721-732.

7. Rat Genome Sequencing Project Consortium. (2004). **Genome sequence of the Brown Norway rat yields insights into mammalian evolution**. *Nature* 428, 493-521.

8. International Bovine BAC Mapping Consortium. (2007). **A physical map of the bovine genome.** *Genome Biol* 8, R165

9. Lindblad-Toh, K., et al. (2005). **Genome sequence, comparative analysis and haplotype structure of the domestic dog**. *Nature* 438, 803-819

10. International Chicken Genome Sequencing Consortium. (2004). **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution**. *Nature* 432, 695-716

11. **The Xenopus tropicalis Genome Project** [http://genome.jgi-psf.org/Xentr4/Xentr4.home.html]

12. **The Danio rerio Sequencing Project** [http://www.sanger.ac.uk/Projects/D_rerio/]

13. **Berkeley Drosophila Genome Project** [http://www.fruitfly.org/]

14. Celniker, S.E., and Rubin, G.M. (2003). **The Drosophila melanogaster genome.** *Annu Rev Genomics Hum Genet* 4, 89-117. Review

15. **Genome Sequencing Centre - Caenorhabditis elegans** [http://genome.wustl.edu/genome.cgi?GENOME=Caenorhabditis%20elegans]

16. Dehal, P., et al. (2002). **The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins**. *Science* 298, 2157-2167