UNIVERSIDADE DE LISBOA

FACULDADE DE MEDICINA

New insights into alternative splicing using microarray technology

Ana Rita Fialho Grosso

Doutoramento em Ciências Biomédicas,

especialidade de Ciências Morfológicas

2009

UNIVERSIDADE DE LISBOA

FACULDADE DE MEDICINA

New insights into alternative splicing using microarray technology

Ana Rita Fialho Grosso

Doutoramento em Ciências Biomédicas,

especialidade de Ciências Morfológicas

2009

Tese orientada por
Professora Doutora Maria do Carmo Fonseca
Professor Doutor Simon Tavaré (University of Cambridge, UK)

A impressão desta dissertação foi aprovada pela Comissão Coordenadora do Conselho Científico da Faculdade de Medicina de Lisboa em reunião de 22 de Setembro de 2009.

As opiniões expressas são da exclusiva responsabilidade do seu autor.

O desenvolvimento e execução gráfica da presente dissertação foram financiados pela Fundação para a Ciência e a Tecnologia (Bolsa SFRH/ BD/22825/2005).

Resumo

Palavras-chave: *splicing* alternativo, regulação de *splicing*, bioinformática, *microarrays*

Nos eucariotas, a transcrição dos genes pela RNA Polimerase II origina moléculas precursoras de RNA mensageiro (pré-mRNA), sendo necessárias várias etapas de processamento até à formação do RNA maduro (mRNA) que é transportado para o citoplasma, onde serve de molde para a síntese proteica.

Adicionalmente, a mesma molécula de pre-mRNA pode gerar diversos tipos de mRNA funcional devido ao *splicing* alternativo. Este processo consiste na inclusão ou exclusão de regiões do pré-mRNA e é um importante mecanismo responsável pela grande variedade de proteínas nos Eucariotas (Black, 2003). Vários estudos baseados em análises de ESTs (*expressed sequence tags*) revelaram que mais de 60% dos genes humanos estão sujeitos ao *splicing* alternativo, e esse número aumentou para 80% com o aparecimento dos *microarrays* (Johnson et al., 2003; Kampa et al., 2004). Mais recentemente, as novas tecnologias de sequenciação de RNA revelaram 92 a 95% dos genes humanos sujeitos ao *splicing* alternativo (Wang et al., 2008a; Pan et al., 2008).

O *splicing* do pré-mRNA está a cargo do spliceosoma, um complexo macromolecular formado a partir de várias pequenas partículas ribonucleoproteicas nucleares (snRNPs) e outras proteínas reguladoras (Jurica and Moore, 2003; Wahl et al., 2009). Através de interacções do RNA e proteínas do complexo com o pré-mRNA, os snRNPs medeiam o reconhecimento e subsequente ligação às junções exão-intrão (sítio de *splicing*).

A selecção e identificação entre os diferentes sítios de *splicing* depende de múltiplas interacções RNA-RNA e RNA-proteína que envolvem a ligação cooperativa de várias proteínas reguladoras (reguladores em *trans*) aos sinais reguladores não codificantes na sequência do pré-mRNA (regulação em *cis*). Os sinais reguladores incluem sequências curtas e muito degeneradas localizadas nos sítios de *splicing* e outras sequências reguladoras denominadas de activadores (*enhancers*) e silenciadores (*silencers*) localizadas em exões ou intrões. Os reguladores em *trans* são designados de factores de *splicing* e usualmente classificadas como activadores ou repressores, dependendo se facilitam ou suprimem a ligação dos snRNPs aos sítios de *splicing*.

Resumo

O processo de *splicing* alternativo é regulado em resposta a vias de sinalização, e é específico a fases de desenvolvimento e tipo de tecido. De acordo com o actual modelo, a regulação do *splicing* alternativo resulta de interacções combinatórias de várias proteínas agindo positivamente e negativamente, e decisões de *splicing* específicas a um tipo de célula ou tecido resultam provavelmente de diferenças na concentração e/ou actividade destas proteínas (Matlin et al., 2005; Shin and Manley, 2004; Singh and Valcárcel, 2005). Uma previsão imediata deste modelo é que a abundância relativa das proteínas reguladoras de *splicing* deve variar de acordo com os tecidos.

Ao longo das últimas décadas começaram a ficar disponíveis dados em larga escala para abordar sistematicamente esta questão, permitindo o desenvolvimento de meta-análises computacionais. A tecnologia dos *microarrays* alterou profundamente o modo de investigação, movendo de uma abordagem gene-a-gene para estudos globais e à escala do genoma.

Neste contexto, o presente estudo teve como objectivo gerar previsões baseadas em *microarrays* para compreensão do código do *splicing* alternativo que controla e coordena o transcriptoma.

Para explorar a hipótese da expressão específica a tecidos das proteínas de *splicing*, neste trabalho analisamos padrões de expressão génica obtidos a partir de estudos de *microarray* anteriormente publicados. A análise contemplou as várias famílias de proteínas hnRNP, SR, cinases-SR, helicases de RNA, proteínas snRNP e muitas outras proteínas associadas com o *splicing* (Barbosa-Morais et al., 2006; Chen et al., 2007).

Em primeiro lugar, o estudo incidiu na alteração da expressão dos factores de *splicing* durante quatro tipos de diferenciação celular do ratinho: miogénese, adipogénese, eritropoiese e espermatogénese. Para identificar variações da expressão específicas ao tipo celular foi necessário comparar dados de *microarray* provenientes de diferentes sistemas biológicos e ensaios experimentais. Para resolver este problema, foi desenvolvida uma nova abordagem que se baseia em métodos de regressão (Grosso et al., 2008). A minha análise revelou diferenças robustas que distinguem um processo de diferenciação dos outros e estas assinaturas de expressão génica incluíam proteínas das várias famílias dos factores de *splicing*.

Em seguida, explorei as variações na expressão dos factores de *splicing* em vários tecidos humanos, de chimpanzé e de ratinho. Comparando padrões de expressão do cérebro, testículos, coração, fígado e rim, 104 genes apresentaram alterações na expressão específicas a tecidos, em pelo menos um organismo. A minha análise revelou que o maior número de factores de *splicing* diferencialmente expressos ocorreu no testículo e no cérebro, enquanto que o fígado apresentou padrões semelhantes ao rim. Curiosamente, os resultados mostram que os dois tecidos com maior quantidade de *splicing* alternativo (Yeo et al.,

2004; Pan et al., 2004; Clark et al., 2007) são os que apresentam uma maior variação na expressão de factores de *splicing*. Assim, os meus resultados mostram que assinaturas de factores de *splicing* estão correlacionadas com padrões de *splicing* específicos de tecidos.

Utilizando PCR quantitativo em tempo real confirmei 75% das previsões dos *microarrays* para a miogenese e eritropoiese e 71% para vários tecidos. De um modo geral, eu identifiquei mais de 100 genes que apresentam uma expressão diferencial associada a um determinado tecido ou processo de diferenciação. Estes resultados mostraram que todos os genes das principais famílias de factores de *splicing* apresentam expressão génica diferencial, incluindo proteínas cinases-SR e proteínas dos snRNP.

O processo de *splicing* alternativo está também associado a doenças humanas, incluindo cancro (Wang and Cooper, 2007; Cooper et al., 2009). São conhecidas várias mutações que afectam o *splicing* de oncogenes, genes supressores tumorais e outros genes relevantes para o cancro (Srebrow and Kornblihtt, 2006; Venables, 2006). Contudo, muitas das anomalias de *splicing* identificadas nas células tumorais não estão associados com mutações nos genes afectados. De facto, estudos recentes sugerem que as mudanças na expressão de factores de *splicing* podem desempenhar um papel fundamental na perturbação geral do *splicing* que ocorre em muitos cancros (Kirschbaum-Slager et al., 2004; Karni et al., 2007; Kim et al., 2008; Ritchie et al., 2008).

Com o objectivo de pesquisar eventuais associações entre a expressão de factores de *splicing* e padrões de *splicing* em cancro, realizei uma meta-análise global que integra dados em larga escala de experiências de *microarrays* em vários tipos de cancro. A evolução da tecnologia dos *microarrays* tem permitido explorar a expressão génica ao nível do exão e estudar variações dos padrões de *splicing* em grande escala (Blencowe, 2006).

Em primeiro lugar, estudei as variações de expressão génica de vários factores de *splicing* em 13 tipos de cancro: bexiga, cérebro, mama, cólon, esófago, cabeça e pescoço, rim, fígado, pulmão, neuroblastoma, próstata, tiróide e vulva. Comparando cancro e correspondentes tecidos normais foram identificadas variações para 192 genes, que codificam proteínas das famílias dos factores de *splicing* snRNPs, hnRNPs, SRS, cinases-SR, RNA-helicases e outros reguladores de *splicing*. Os meus resultados também mostraram que a maioria dos factores de *splicing* com variações na expressão eram essencialmente mais expressos em cancro (sobre-expressão), e de facto foi observado um enriquecimento de factores de *splicing* entre todos os genes sobre-expressos. Alguns genes desregulados apareciam em vários cancros, sugerindo que alguns cancros podem apresentar as mesmas variações de factores de *splicing*.

Em seguida, seleccionei um conjunto de eventos de *splicing* desregulados em cancro através da aplicação de uma metodologia de análise exaustiva a dados de *microarrays* de *splicing* de estudos anteriores em cancro do cólon e pulmão. Foram encontrados exões con-

tendo sítios de ligação para SF2/ASF obtidas a partir dados de CLIP-seq (Sanford et al., 2009) para SF2/ASF, que é sobre-expresso em ambos os cancros. Foram também identificadas pequenas sequências enriquecidas associadas com eventos de *splicing* de cancro que apresentam alguma semelhança a sítios de ligação de factores de *splicing* desregulados em cancro.

Concluíndo, o presente trabalho forneceu contribuições científicas novas e importantes para a compreensão do código do *splicing* alternativo que controla e coordena o transcriptoma. Os meus resultados reforçam o actual modelo de regulação do *splicing* alternativo, que sugere que as diferenças na abundância relativa ou actividades específicas de várias proteínas influenciam decisões no mecanismo de *splicing*. Este trabalho alargou a lista de possíveis reguladores para s*plicing* alternativo associado à diferenciação celular e ao cancro, o que desencadeia novas linhas de investigação e validação experimental.

Finalmente, o presente trabalho mostra o poder de utilizar a tecnologia dos *microarrays* e abordagens computacionais para gerar previsões para uma visão global da regulação do *splicing* alternativo.

Abstract

Keywords: alternative splicing, splicing regulation, bioinformatics, microarrays

The present study aimed to generate microarray-based predictions for understanding the alternative splicing code that controls and coordinates the transcriptome. To explore the hypothesis that splicing proteins are expressed in a tissue-specific manner, I analysed gene expression profiles from previously published microarray studies. The analysis included members of the hnRNP and SR protein families, SR protein kinases, DEAD-box RNA helicases, snRNP proteins and several splicing-related proteins (Barbosa-Morais et al., 2006; Chen et al., 2007). First, I focused on variations of splicing factors during four types of murine cell differentiation: myogenesis, adipogenesis, erythropoiesis and spermatogenesis. To identify cell-type specific variations in splicing factor expression, microarray data sets derived from different biological systems and experimental assays have to be compared. To address this issue, I developed a new approach that is based on regression modelling methods (Grosso et al., 2008). My analysis revealed robust differences that distinguish one differentiation process from the others and these gene expression signatures included members of several splicing-related protein families. Second, I explored variations in splicing factor expression across tissues from human, chimpanzee and mouse. Comparing brain, testis, heart, liver and kidney gene profiles, 104 genes showed tissue-specific expression variation in at least one organism. My analysis revealed that the highest number of highly differentially expressed splicing-related genes occurred in the testis and in the brain, whereas the liver showed higher concordance in expression of splicing-related genes relative to other tissues, namely the kidney. Interestingly, the results distinguished the two tissues previously described with highest abundance of alternatively spliced mRNA isoforms that differ by inclusion or exclusion of an exon (Yeo et al., 2004; Pan et al., 2004; Clark et al., 2007), as those with a highest variation in splicing factor expression. Thus, my findings showed that splicing factor signatures correlate with tissue-specific alternative splicing patterns. By using quantitative real time PCR, I confirmed 75% of microarray predictions for myogenesis and erythropoiesis and 71% for several tissues. Overall I identified over 100 splicing-related genes that are most highly differentially

expressed in a particular tissue or differentiation process. These results showed that genes from the main splicing factor families present differential gene expression, including SR protein kinases and snRNP proteins.

Alternative splicing is associated with several human diseases, including cancer (reviewed in Wang and Cooper, 2007; Cooper et al., 2009). Several mutations are known that affect the splicing of oncogenes, tumour suppressors and other cancer-relevant genes (Srebrow and Kornblihtt, 2006; Venables, 2006). However, many splicing abnormalities identified in cancer cells are not associated with mutations in the affected genes. Indeed, recent studies suggest that changes in splicing factor expression may play a key role in the general splicing disruption that occurs in many cancers (Kirschbaum-Slager et al., 2004; Karni et al., 2007; Kim et al., 2008; Ritchie et al., 2008). To investigate whether misregulation of splicing factor expression correlates with cancer-associated splice variants, I applied a global meta-analysis that integrates large-scale data from microarray experiments in several cancer types. Microarray technology evolution allowed to resolve exon-level gene expression and enabled large-scale profiling of mRNA splicing (reviewed in Blencowe, 2006). First, I explored gene expression variations of several splicing factors in 13 cancer types: bladder, brain, breast, colon, esophagus, head and neck, kidney, liver, lung neuroblastoma, prostate, thyroid and vulva. Comparing the cancer and corresponding normal tissue, I identified misregulation for 192 splicing-related genes encoding the major splicing protein families (snRNPs, hnRNPs, SRs, SR-kinases, RNA-helicases-like and other splicing regulators). My results also showed that the majority of differentially expressed splicing regulators were up-regulated in cancer and in fact, an enrichment of splicing factors in total overexpressed genes was detected. Some misregulations appear consistently in several cancer types, suggesting that some cancers can present common misregulated splicing factors. Afterwards, I collected a high-confidence set of misregulated splicing events in cancer by applying a comprehensive workflow analysis to splicing microarray data sets from previous studies for colon and lung cancers. I found misspliced exons containing cis-acting RNA elements obtained from CLIP-seq data (Sanford et al., 2009) for SF2/ASF, which is overexpressed in both cancers. Enriched motifs were also identified in the cancer-associated splicing events that resemble binding sites for other splicing factors found misregulated in cancer.

My work has given important and original scientific contributions for understanding the alternative splicing code that controls and coordinates the transcriptome. My results reinforce the current model for alternative splicing regulation, which postulates that differences in relative abundances or activities of multiple proteins influence specific splicing decisions. Furthermore, the large number of splicing-related genes with differential expres-

sion found in the present study raises the question of which splicing factors can indeed be responsible for alternative splicing events relevant for cell differentiation and tumorigenesis. This work extended the list of putative regulators for differential alternative splicing, which triggers new lines of research and experimental validation. Finally, this work shows the power of using microarray technology and computational approaches to generate initial predictions for a global view of alternative splicing regulation.

Acknowledgements

Though the following dissertation is an individual work, I could never have reached the end without the help, support, guidance and efforts of of several people to whom I want to express my gratitude:

Maria Carmo-Fonseca for giving me the opportunity to work in a new promising field and be a member of her research group. I acknowledge her for having provided me excellent working conditions and for the grateful discussions and encouragement throughout the work presented in this dissertation.

Simon Tavaré for accepting me as his student and for giving me the opportunity to be a member of his computational group in the University of Cambrige.

Juan Valcárcel and Benjamin Blencowe for receiving me so well at their laboratories and for the critical discussions.

I would like to acknowledge all my dear colleagues and friends at the Faculty of Medicine in Lisbon.

Special thanks to Francisco Enguita, Margarida Gama-Carvalho and Joana Cardoso for creating stimulating intellectual discussions throughout this research and for all the constructive comments and suggestions for improving my work.

Ricardo Henriques, Marisa Cabrita, Jorge Andrade, Houda Hallay, Ana Garcia-Sacristan, Marco Campinho, Sérgio de Almeida, Joana Desterro and Alexandre Teixeira for friendship, collaboration and precious help throughout this boundary road between computers and lab bench.

Ines Mollet and José Nuno Pereira for sharing the struggles of bioinformaticians.

Anita Gomes and Sandra Martins for all their help during the short period of validations with real time PCR.

Special thanks to José Braga and José Rino for many favors and companionship in Zé's office (Ricardo José and José Nuno were also part of the team).

Acknowledgements

Teresa Carvalho, Célia Carvalho, Noélia Custódio, Sérgio Marinho, Joana Borlido, Birgit Weissenboeck, Natalia Kozlova, Patricia Calado, Sandra Caldeira and João Paulo Tavanez for companionship in Carmo-Fonseca's group.

Simon Tavaré's group for receiving me so well at the University of Cambridge and for discussion on the statistical analysis. Special thanks to Natalie Thorne for all the guidance, support, constructive comments and suggestions to improve my work. Nuno Barbosa-Morais for many exciting brainstorms and encouragement throughout these last years.

Josefin Lundgren (Center for Genomic Regulation - Barcelona, Spain) and Ashraf Ibrahim (Cambridge Cancer Center - Cambridge, UK) for all the collaboration and patience teach me microarray assays.

The members of my thesis committee, Lisete Sousa, Mario Ramirez and Luis Moita for important feedback throughout my PhD.

Principal Investigators of Institute of Molecular Medicine, João Ferreira, Bruno Silva-Santos, Luis Moita and Maria Mota, for sharing their scientific interests and projects with me. Their group members, Joana Cardoso, Anita Gomes, Daniel Correia, Telma Lança, Pedro Alves, Teresa Pacheco and Celina Carret, for the very pleasant collaboration.

Special thank to my dear old colleagues but still good friends at the Faculty of Sciences in Lisbon Andreia Ferreira, Lara Carvalho, Cristiane Bastos-Silveira, Cristina Luisa, Carina Cunha, Deodália Dias, Andreia Fonseca and Sandra Botelho for all encouragement and good advices.

Finally, I am deeply grateful to my family.

My parents, Marcela and Joaquim, for ongoing support and encouragement of my scientist life, for their unconditional love and for having made this possible.

My dear brother, Nuno, for his love and kindness, for being so special and for teaching me how to become a better person.

My husband, Marco, for the active participation throughout this work (programming languages, computer problems, thesis reviewing, printing, etc.), love and patience which kept me going, and for believing that I would overtake this challenge.

Preface

In this dissertation are described the results of research work developed between 2006 and 2009 under the supervision of Prof. Dr. Maria Carmo-Fonseca from Faculty of Medicine (Lisbon University - Portugal) and Prof. Dr. Simon Tavaré from Oncology Department (Cambridge University - United Kingdom).

The main goal of this work was to explore and combine large scale data from microarray technology in order to understand the splicing code underlying alternative splicing regulation.

This dissertation is organized in five chapters.

Chapter 1 corresponds to the general introduction of this dissertation. In this chapter is exposed the pre-mRNA splicing process and the key regulators of this mechanism, focusing also on alternative splicing. Next, the microarray technology, assays and the different types of applications are described. Finally, a brief overview of the bioinformatic tools for large scale approaches is presented.

Chapter 2 presents results from a large-scale computational analysis of mRNA expression data where splicing-factor expression signatures were identified for differentiation processes and tissues derived from human, chimpanzee and mouse. The original work described in this chapter has been integrally published in: Grosso AR, Gomes AQ, Barbosa-Morais NL, Caldeira S, Thorne NP, Grech G, von Lindern M, Carmo-Fonseca M (2008) Tissue-specific splicing factor gene expression signatures, Nucleic Acids Research. 36(15):4823-32.

Chapter 3 focuses on gene expression variations of splicing factors in cancer. Results from the analysis of microarray data for several cancer types are presented, showing several misregulated splicing factors in cancer. This chapter explores also the emerging role of splicing factors in cancer, which discussion is written as a review article in: Grosso AR, Martins S, Carmo-Fonseca M. (2008) The emerging role of splicing factors in cancer, EMBO Rep, 2008 Nov;9(11):1087-93.

Chapter 4 explores splicing misregulation in cancer using splicing sensitive microarrays. Here, cancer-associated splicing profiles were identified through the use of a comprehensive

workflow analysis from splicing microarray data. In this chapter cancer-associated splice variants which appear to be regulated by splicing factors with gene expression affected in the same cancer are also described.

Chapter 5 comprises an integrative discussion of results and future perspectives, pointing out the contribution of microarray technology for understanding alternative splicing mechanism and regulation.

During my PhD I have collaborated in projects from different research groups of the Institute of Molecular Medicine in Lisbon, extending my knowledge and experience in the bioinformatics field. Although the results were not included in the present dissertation, I would like to mention some of the resultant publications or manuscripts in preparation:

- Albuquerque SS, Carret C, **Grosso AR**, Tarun AS, Peng X, Kappe SHI, Prudêncio M and Mota MM (2009). Host cell transcriptional profiling during malaria liver stage infection reveals a coordinated and sequential set of biological events, BMC Genomics, 17;10(1):270.

  Individual contribution: microarray data analysis

- Correia DV, d'Orey F, Cardoso BA, Lança T, **Grosso AR**, Debarros A, Martins LR, Barata JT, Silva-Santos B (2009). Highly active microbial phosphoantigen induces rapid yet sustained MEK/Erk- and PI-3K/Akt-mediated signal transduction in anti-tumor human gammadelta T-Cells, PLoS ONE, 4(5):e5657.

  Individual contribution: microarray data analysis

- Mollet IG, Ben-Dov C, Felício-Silva D, **Grosso AR**, Eleutério P, Alves R, Staller R, Silva TS, Carmo-Fonseca M. Unconstrained mining of mRNA and EST databases reveals increased alternative splicing complexity in the human transcriptome. *(submitted)*

  Individual contribution: improvement of mysql database performance and development of R based tools for ExonMine webserver

- Gomes AQ*, Correia DV*, **Grosso AR**, Lança T, Gomes da Silva M and Silva-Santos B. Identification of molecular markers of leukemia/ lymphoma susceptibility or resistance to gamma-delta T cell cytotoxicity. *(submitted)*

  Individual contribution: microarray data analysis

- Alves PM, Neves-Costa A*, Raquel H*, Oliveira M, **Grosso AR**, Moita C, D'Almeida B, Pacheco T, Rodrigues R, Gama-Carvalho M, Hacohen N and Moita LF. ASF/SF2 and SRp20 are negative regulators of IL-1β secretion. *(manuscript in preparation)*

Individual contribution: Exon-microarray data analysis

- Hallay H, Cardoso J, *Grosso AR*, Ferreira J, Carmo-Fonseca M. Global Analysis of splicing in a cellular model of human aging *(manuscript in preparation)*

  Individual contribution: Exon-microarray data analysis

- Cardoso J, **Grosso AR**, Carmo-Fonseca M, Ferreira J. The contribution of alternative splicing to drug-induced accelerated senescence. *(manuscript in preparation)*

  Individual contribution: Exon-microarray data analysis

- Enguita FJ, **Grosso AR**, Carmo-Fonseca M. Genome-wide screening for miRNA genes within coding exons by using support vector machine applications. *(manuscript in preparation)*

  Individual contribution: development of pipeline to integrate data, based on R language and packages

- Campinho MA, Pereira J, **Grosso AR**, Carmo-Fonseca M. ΔspliceMutation: A bioinformatical tool for prediction of mutations that disrupt Exonic Splicing Regulatory motifs (ESR) *(manuscript in preparation)*

  Individual contribution: collaboration for pipeline development based on R language and packages

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A IUPAC nucleotide code for Adenine

C IUPAC nucleotide code for Cytosine

cDNA complementary DNA

CGH Comparative genomics hybridization

DNA Deoxyribonucleic Acid

ESE Exonic Splicing Enhancer

ESS Exonic Splicing Silencer

EST Expressed Sequence Tag

G IUPAC nucleotide code for Guanine

GEO Gene Expression Omnibus

GO Gene Ontology

HGNC HUGO Gene Nomenclature Committee

hnRNPs heterogenous nuclear Ribonucleoproteins

I IUPAC nucleotide code for Inosine

ISE Intronic Splicing Enhancer

ISS Intronic Splicing Silencer

IUPAC International Union of Pure and Applied Chemistry

MIAME Minimum Information About a Microarray Experiment

mRNA messenger Ribonucleic Acid

NCBI National Center for Biotechnology Information

nt nucleotides

PABPs Poly(A)-Binding Proteins

PCA Principal Component Analysis

pre-mRNA precursor messenger RNA

PSR Probe Selection Region

R IUPAC ambiguous nucleotide code for Adenine and Guanine

RBPs RNA-binding proteins

RRM RNA Recognition Motif

S IUPAC ambiguous nucleotide code for Cytosine and Guanine

SFs Splicing Factors

SNPs Single Nucleotide Polymorphisms

snRNA small nuclear RNA

snRNPs small nuclear Ribonucleoprotein Particles

SR Serine/Arginine-rich

SREs Splicing Regulatory Elements

SRGs Splicing-Related Genes

SRPK SR Protein Kinase

SVD Singular Value Decomposition

T IUPAC nucleotide code for Thymine

U IUPAC nucleotide code for Uracil

U2AF U2 snRNP Auxiliary Factor

Y IUPAC ambiguous nucleotide code for Cytosine and Thymine/Uracil

Chapter 1

Introduction

1.1 pre-mRNA splicing

1.1.1 mRNA biogenesis

Gene expression is the process by which information from a gene is used in the synthesis of a functional product, including the main steps: transcription, several precursor messenger RNA (pre-mRNA) processing steps ($5'$ capping, splicing, $3'$ end processing and editing), export of the mature mRNA to the cytoplasm, translation into amino-acid sequence and post-translational modification (Figure 1.1). This multistep process requires several complex cellular machines responsible for each specific step in this process. Although the identification of the protein components of each of these cellular machines has been carried out independently, recent findings suggest that each one of these steps regulating gene expression is physically and functionally connected to the next, as part of a continuous process (Orphanides and Reinberg, 2002; Kornblihtt et al., 2004).

Messenger ribonucleic acid (mRNA) is the key intermediary in gene expression, which carries the genetic information transcribed from deoxyribonucleic acid (DNA) and is translated as a template for polypeptide synthesis. The mRNA is similar to the DNA molecule, except that: it is single-stranded; the base uracil (U) substitutes the base thymine (T), and the pentose sugar is a ribose.

Most genes coding for proteins in eukaryotes are transcribed by the RNA polymerase II transcription machinery, which can be generally divided in three major components: the 12-subunit polymerase, five general transcription factors (TFIIB, -D, -E, -F and -H) and the Mediator complex (reviewed in Boeger et al., 2005; Woychik and Hampsey, 2002).

Transcription takes place in three stages: initiation, elongation and termination. Initiation involves binding of transcription factors and RNA Polymerase II to the promoter, local melting (separation of the DNA strands), and forming the first phosphodiester bond.

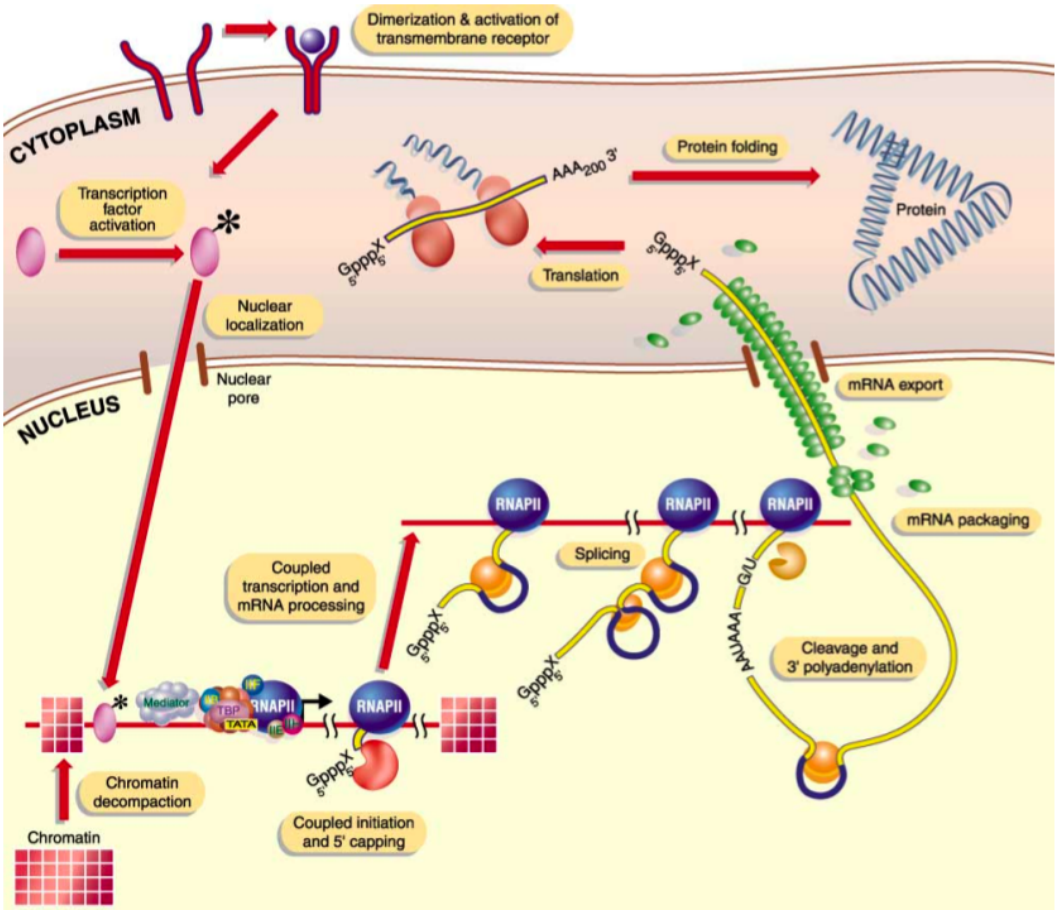


Figure 1.1: **Gene Expression.** A contemporary view representing the several steps regulating gene expression physically and functionally connected: transcription, post-transcriptional modification, translation and post-translational modification(Image adapted from Orphanides and Reinberg, 2002).

.

During elongation, the RNA polymarease II moves 5′ to 3′ along the gene sequence and extends the transcript. In termination, the transcript is released from the RNA polymerase and the polymerase from the DNA template (Orphanides and Reinberg, 2002).

The transcripts initially produced are designated as precursor messenger RNAs (pre-mRNAs) and undergo several modification steps before transport to the cytoplasm as functional mRNAs and posterior translation into proteins.

Pre-mRNA processing occurs co-transcriptionally and includes 5′-end capping, splicing, 3′-end processing and editing (Figure 1.1).

The 5′-end capping occurs soon after RNA polymerase II initiates transcription, when the transcript is about 20-30 nucleotides in length (Proudfoot et al., 2002). This capping consists of a three-step reaction: first an RNA 5′ triphosphatase hydrolyzes the triphos-

phate of the first nucleotide to a diphosphate; second a guanylyltransferase catalyzes the addition of a GMP (guanosine monophosphate) to the first nucleotide of the pre-mRNA via an unusual 5′-5′ triphosphate linkage; finally a methyltransferase methylates the N7 position of the transferred GMP. This cap serves initially to improve the stability of the mRNA (protecting the new transcript from attack by nucleases) and later serves as a binding site for proteins involved in export of the mature mRNA into the cytoplasm (Proudfoot et al., 2002).

Pre-mRNA splicing is an essential step in the regulation of gene expression due to the split nature of eukaryotic genes. Most pre-mRNAs are interrupted by long noncoding sequences named introns that must be removed in order to place the coding sequences, exons, in a protein-reading frame. The pre-mRNA splicing is described in more detail in Sections 1.1.2 and 1.1.3.

In the termination step, upon reaching the end of a gene, the newly synthesized RNA is cleaved and a polyadenosine tail of 200-250 adenosine residues is added to the 3′ end of the transcript. These 3′ poly(A) tails provide the mRNA with a binding site for a major class of regulatory factors, the poly(A)-binding proteins (PABPs). These proteins bind poly(A) using one or more RNA-recognition motifs and have several roles in mediating gene expression. PABPs are necessary for the synthesis of the poly(A) tail in the nucleus, regulating its ultimate length and stimulating maturation of the mRNA. Association with PABP is also a requirement for some mRNAs to be exported from the nucleus. Finally, PABPs promotes translation initiation and termination, recycling of ribosomes, and stability of the mRNA in cytoplasm (Mangus et al., 2003).

Besides capping, splicing and 3′-end processing, pre-mRNA undergoes another modications broadly defined as RNA editing. These modifications of the pre-mRNA can occur through insertion or deletion of nucleotides or by the substitution of bases. The most common in mammals are deamination reactions, like the conversion of C to U and of A to I (inosine, that is equivalent to G for translation process). These modications can affect both coding and non-coding (namely intronic) sequences and are suggested to regulate splicing and to have a role in processing and stability of mRNAs (reviewed in Keegan et al., 2001; Gerber and Keller, 2001).

After pre-mRNA processing, the mature mRNA is exported by factors that bind to mRNA molecules in the nucleus and direct them into the cytoplasm though interactions with proteins that line the nuclear pores (reviewed in Reed and Hurt, 2002; Cole and Scarcelli, 2006).

Translation of mRNA into protein takes place in cytoplasm on large ribonucleoprotein complexes called ribosomes (reviewed in Ramakrishnan, 2002). The process begins with the location of the start codon (AUG) by translational initiation factors in conjunction with

subunits of the ribosome and involves elongation and termination phases (recognition of stop codons UAA, UAG or UGA) (reviewed in Dever, 2002; Sonenberg and Hinnebusch, 2009). Finally, the nascent polypeptide undergoes folding and often post-translational modications to generate the final active protein (reviewed in Daggett and Fersht, 2003).

1.1.2 Spliceosome and splicing signals

Pre-mRNA splicing is carried out by the spliceosome, a macromolecular complex formed from several small nuclear ribonucleoprotein particles (snRNPs) and numerous non-snRNP splicing factors (reviewed in Jurica and Moore, 2003; Wahl et al., 2009) (Figure 1.2). Initial mass spectrometric studies of spliceosomal complexes indicated that between 150 and 300 distinct proteins copurify with spliceosomes (Zhou et al., 2002; Rappsilber et al., 2002). More recently, studies purifying spliceosomes at more defined stages of assembly and function indicated that the total number of spliceosome-associated factors is approximately 170 (reviewed in Wahl et al., 2009).

The splicing process requires several specific RNA-RNA, RNA-protein and protein-protein interactions to recognise the exon-intron junctions and remove the introns. Some of these interactions are mediated by several *cis*-acting elements, RNA sequence signals, that distinguish exons from introns, direct the spliceosome to the correct nucleotides for exon joining and intron removal, and serve as binding sites for auxiliary factors (*trans*-acting elements).

snRNPs

Each snRNP particle consists of stable small nuclear RNA (snRNA) bound by a core ring of seven different Sm or Sm-like proteins, and several particle-specific proteins (reviewed in Jurica and Moore, 2003) (Figure 1.2).

The major spliceosomal snRNPs U1, U2, U4, U5 and U6 are responsible for splicing the vast majority of pre-mRNA introns (so-called U2-type introns). A group of less abundant snRNPs, U11, U12, U4atac and U6atac, together with U5, are subunits of the so-called minor spliceosome. The minor splicing system targets a rare class of introns (U12-type or minor-class introns) (reviewed in Patel and Steitz, 2003). The snRNPs carry out a number of essential functions during splicing. Via the interactions of their RNA and protein components with the pre-mRNA, they mediate the recognition and subsequent pairing of the $5'$ and $3'$ boundaries of the intron.

The core signals that define the intron correspond to four poorly conserved sequences: the exon-intron junctions at the $5'$ and $3'$ ends of the intron (designated as $5'$ and $3'$ splice site, respectively), the polypyrimidine tract (polyimmediately preceding $3'$ splice site (a stretch of 15-25 nucleotides enriched with pyrimidine residues), and the branch point

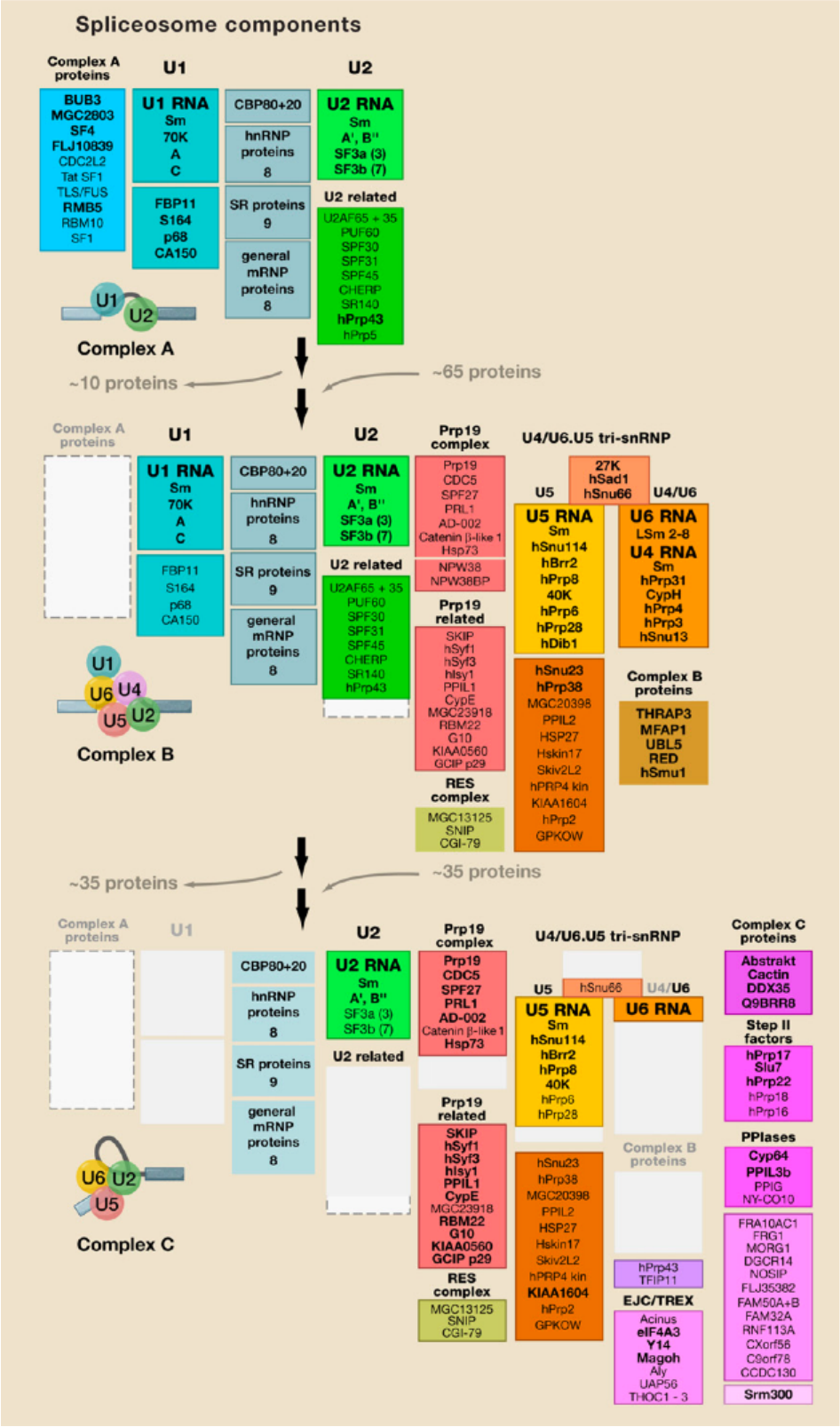


Figure 1.2: **Compositional dynamics of human spliceosomal complexes formed by snRNPs during splicing process** (Image from Wahl et al., 2009).

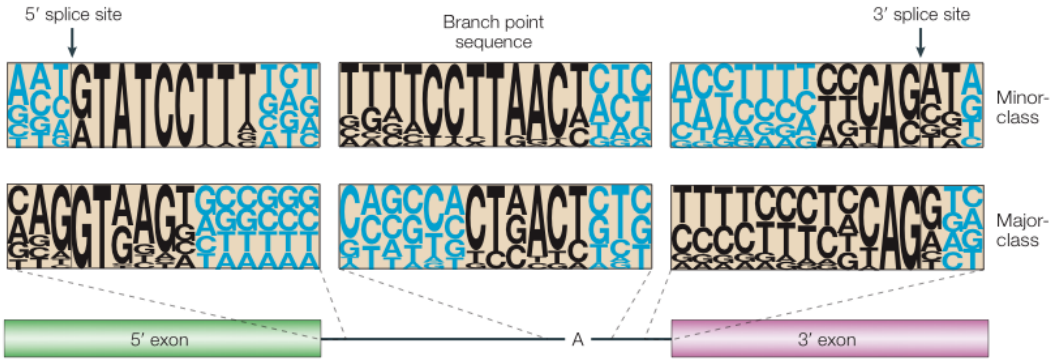


Figure 1.3: **Consensus sequences of major-class and minor-class introns.** The size of a nucleotide at a given position is proportional to the frequency of that nucleotide at that position. The positions that are thought to be involved in intronrecognition are shown in black; other positions are shown in blue. Frequencies were derived from a set of U12-type introns from various plant and animal species and from a set of mammalian U2-type introns (Image from Patel and Steitz, 2003).

sequence (which includes an adenosine residue). During the splicing reaction snRNAs interact by base paring with the splice sites and branch point regions (explained in detail below). Thus, different sequence elements are found in U2- and U12-type introns (Figure 1.3).

The U2-type introns in higher eukaryotes have the highly conserved signals 5′-GU... A ... Y(10-20) .... AG-3′ for 5′ splice site, branch point, polypyrimidine tract and 3′ splice site, respectively (Figure 1.3). However, these longer consensus sequences can be extended to R/GURAGU for 5′ splice site (/ denotes the exon-junction), YNYURAYY for the branck poinst and YAG for 3′ splice site. The U12-type introns lack the polypyrimidine tract upstream the 3′ splice site and contain highly conserved splicing signals: 5′-RUAUCCUUU ... UCCUAAC ... YAS-3′ for 5 splice site, branch point and 3 splice, respectively. The introns were initially named AT-AC due to the first discoveries, however recent findings revealed that most U12-type introns have GT-AG boundary sequences (reviewed in Patel and Steitz, 2003).

SR, hnRNPs and other splicing-related genes

To compensate for the short and poorly conserved nature of splice-site sequences in higher eukaryotes, recognition of higher eukaryotic introns often relies on other *cis*-acting regulatory elements located in exons and introns that act as splicing enhancers and silencers (Izquierdo and Valcárcel, 2006). These splicing regulatory elements (SREs) are binding sites for additional splicing factors that can enhance or prevent the association of the snRNPs to the adjacent splice sites. Indeed, most of the functionally important RNA-RNA

interactions formed within the spliceosome are weak and generally require the assistance of proteins to enhance their stability.

One class of factors recognizing enhancer sequences are the the Serine/Arginine-rich (SR) proteins, which contain one or more domains rich in serine and arginine residues (the RS domain) and RNA-binding domains or RNA recognition motifs (RRM). Some examples of classical SR proteins (and respective human official gene symbol (defined by HUGO Gene Nomenclature Committee - HGNC) are: ASF/SF2 (SFRS1), SC35 (SFRS2), SRp46 (SFRS2B), SRp20 (SFRS3), SRp75 (SFRS4), SRp40 (SFRS5), SRp55 (SFRS6), 9G8 (SFRS7) (reviewed in Long and Caceres, 2009).

Some additional splicing factors also contain RS domains but are not considered as classical SR proteins, referred as SR-related proteins. Among the most important SR-related proteins is the U2 snRNP auxiliary factor (U2AF) which in mammals is composed of a 35 and a 65 kDa subunit, named U2AF35 (U2AF2) and U2AF65 (U2AF1) respectively. These two subunits also contain RRM domains and bind to the polypyrimidine tract (U2AF65) and $3'$ splice site (U2AF35), promoting the stable binding of U2 snRNP to the branch site. Examples of other SR-related proteins are: U1-70K (SNRP70), SRp30c (SFRS9), hTra2β (SFRS10), hTra2α (TRA2A), p54 (SFRS11), SRrp86 or SRrp508 (SFRS12), SRm160 (SRRM1), SRm300 (SRRM2) (reviewed in Long and Caceres, 2009).

SR proteins play diverse roles in many aspects of mRNA processing including splicing, export, and translation (reviewed in Graveley, 2000). The sub-cellular localization and activity of SR proteins is modulated by extensively phosphorylation of their RS domains. Several protein kinase families have been shown to phosphorylate the RS domain of SR proteins, including the SRPK (SR protein kinase) family, the Clk/Sty family and topoisomerase (reviewed in Stamm, 2008; Long and Caceres, 2009).

Other non-snRNP splicing factors that contain also RRM domains are the heterogenous nuclear ribonucleoproteins (hnRNPs). hnRNPs can repress splicing by directly antagonizing the recognition of splice sites, or can interfere with the binding of proteins bound to enhancers, like SR proteins. There are several hnRNPs and more than 20 of them have been characterized and given alphabetical names based on size from hnRNP A1 to hnRNP U. They have been implicated in a variety of biological processes including telomere biogenesis, translation, RNA stability and splicing (reviewed in Martinez-Contreras et al., 2007). A function in splicing has been documented or proposed for more than half of the major hnRNPs: A1 (HNRNPA1), A2 (HNRPA2B1), E1 and 2 (PCBP1 and PCBP2), C (HNRNPC), F (HNRNPF), H (HNRNPH1), H' (HNRNPH2), 2H9 (HNRNPH3), G (RBMX), PTB (PTBP1), nPTB (PTBP2), K (HNRNPK), L (HNRNPL), M (HNRNPM), Q (SYNCRIP), R (HNRNPR) (Venables et al., 2008). Other hnRNP proteins like N, S and T remain poorly characterized and there is no evidence that the hnRNP proteins A0, A3,

A/B, D, DL and U play a role in splicing (reviewed in Martinez-Contreras et al., 2007).

One more class of proteins essential for splicing reaction are the proteins from RNA-helicases superfamily, namely DEAD and DEAH box families. The family names derived from the conserved motif II with the four amino-acids Aspartic acid-Glutamic acid-Alanine-Aspartic acid (DEAD in one-letter code) Aspartic acid-Glutamic acid-Alanine-Histidine (DEAH), being also known as DExH/D (where x can be any amino acid) (reviewed in Rocak and Linder, 2004). These proteins are associated with all processes involving RNA molecules, including transcription, splicing, editing, ribosome biogenesis, RNA export, translation, etc. In pre-mRNA splicing DEAD-box proteins are required for establishment of a functional spliceosome, whereas DEAH-box proteins are (indirectly) required for the *trans*-estherication reactions, the release of the mRNA, and the recycling of the spliceosome components. Examples of DExH/D proteins acting in splicing are DDX3Y, DDX23 (orthologous of yeast Prp28 protein), DDX46 (orthologous of yeast Prp5 protein), UAP56 (BAT1), DDX5, p75, Prp2, Prp16, Prp22, Prp43, DDX42 (reviewed in Linder, 2006).

1.1.3 Spliceosome assembly

Assembly begins with the binding of the U1 snRNP through base-pairing interactions of the 5′-end of the U1 snRNA to the 5′ splice site (Figure 1.4). This interaction in higher eukaryotes is stabilized by members of the SR proteins. Early assembly step also require the binding of SF1 (Splicing Factor 1) protein and the U2AF subunits to the branch point and the polypyrimidine tract, respectively. SF1 interacts with U2AF65 and the other subunit of the U2AF heterodimer, U2AF35, binds AG dinucleotide of the 3′ splice site. Together, these molecular interactions yield the spliceosomal E complex and play crucial roles in the initial recognition of the 5′ and 3′ splice sites of an intron and brings the splice sites that are to be cleaved and joined into juxtaposition.

After the formation of the spliceosomal E complex, the U2 snRNA engages in ATP-dependent manner and in a base-pairing interaction with the brach point, leading to the formation of the A complex. This base-pairing interaction is stabilized by heteromeric protein complexes of the U2 snRNP (namely SF3a and SF3b) and also by the U2AF65 protein. Association of U2 leads to the displacement of SF1 from the branch point.

After A complex formation, the U4/U6 and U5 snRNPs are recruited as a preassembled U4/U6.U5 tri-snRNP, forming the B complex (or B1). Although all snRNPs are present in the B complex, it is still catalytically inactive and requires major conformational and compositional rearrangements (catalytic activation) in order to become competent to facilitate the first out of two *trans*-esterification reactions involved in splicing. During spliceosome activation, U1 and U4 are destabilized or released, giving rise to the activated spliceosome

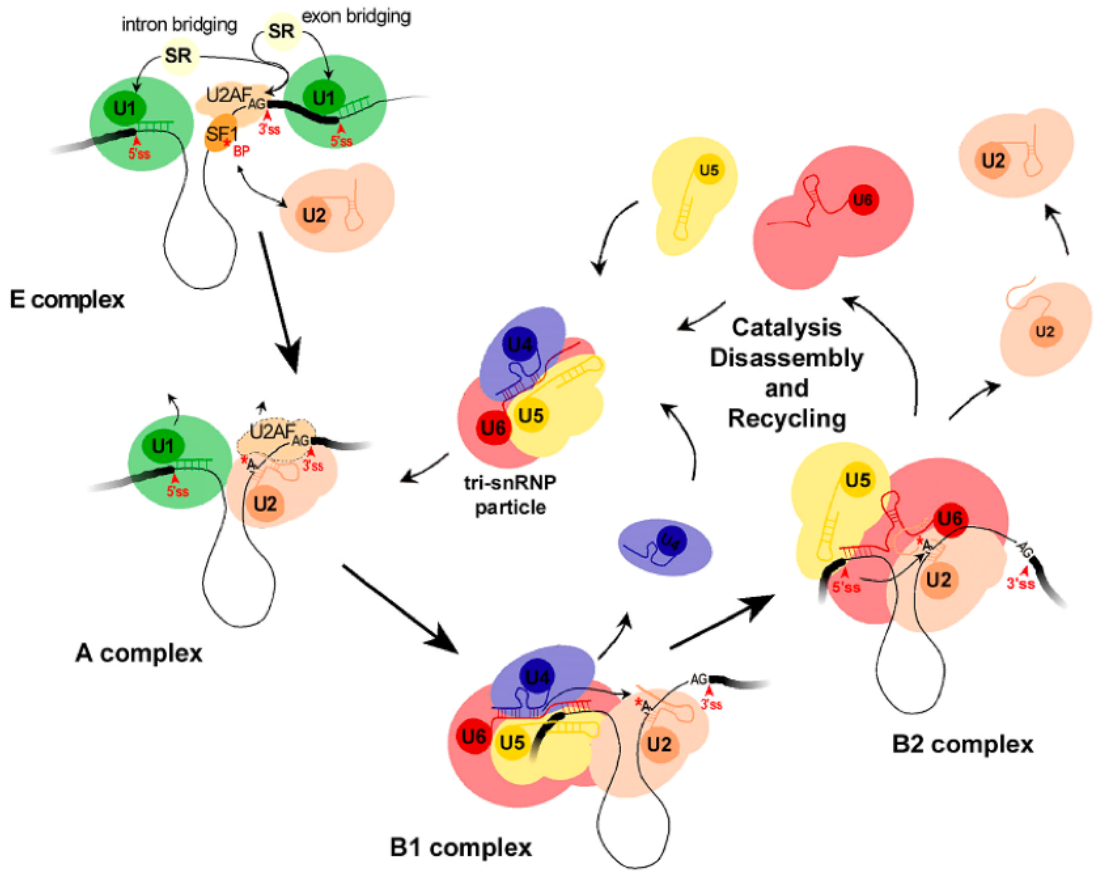


Figure 1.4: **Spliceosome Assembly.** Exons are represented by thick and introns by thin lines; protein particles (U snRNPs, U2AF and SF1) are represented by round shapes; snRNAs are depicted by the lines accompanying snRNPs; splice sites and branch point (A) are indicated (Image modified from Gama-Carvalho, 2002).

.

(the B2 or C complex). The activated spliceosome then undergoes the first catalytic step of splicing, generating the C complex. The 3′-5′-phosphodiester bond at the 5′ splice site is attacked by the 2′-hidroxyl group of the conserved intronic adenosine at the branch site. A 2′-5′ phosphodiester bond is formed, generating a intron lariat and a free 5′ exon, with a 3′-hidroxyl group. Then the second *trans*-esterification occurs, where the 3′-hidroxyl group of the 5′ exon attacks the phosphodiester bond at the 3′ splice site, releasing the intron lariat (to be degraded) and ligating the exons. After the second catalytic step, the spliceosome dissociates, releasing the mRNA and the U2, U5, U6 snRNPs to be recycled for additional rounds of splicing (reviewed in Wahl et al., 2009).

Assembly of the minor spliceosome is similar to that of the major spliceosome, with the U11, U12 and U4atac/U6atac being functionally similar to the U1, U2 and U4/U6,

respcetively. The major difference occurs at the earliest assembly step in which the U11 and U12 snRNPs form a highly stable di-snRNP that binds cooperatively to the 5 splice site and branch point, which is equivalent to the A complex of the major spliceosome (reviewed in Patel and Steitz, 2003).

1.1.4 Alternative splicing

Splicing is usually constitutive, which means that all exons are joined together in the order in which they occur in the pre-mRNA. In many genes, however, alternative splicing has also been observed, in which the exons may be combined in some other way (Figure 1.5). For example, some exon or exons may be skipped, being removed similarly to introns. However, the primary order of the exons is not altered in alternative splicing. Thus alternative splicing makes it possible for a single gene to produce more than one messenger RNA molecule, so-called isoforms.

Transcripts from a gene can undergo many different patterns of alternative splicing (Figure 1.5): transcriptional initiation at different promoters generates alternative first exons that can be joined to a common exon; alternative terminal with alternative polyadenylation sites can be joined to a common upstream exon; use of alternative $5'$ or $3'$ splice sites, exons can be extended or shortened in length; inclusion and skipping of cassette exon, inserting or deleting a portion of internal sequence; mutually exclusive splicing of cassette exons, where one exon or the other is included (but not both); intron retention where the excision of an intron is suppressed. Many genes show multiple positions of alternative splicing, creating complex combinations of exons and alternative segments and consequently different protein coding sequences. In addition, alternative splicing can mediate the repression of gene expression by stimulating the formation of transcripts subject to nonsense-mediate mRNA decay (mRNA degradation) (reviewed in Lejeune and Maquat, 2005).

Several studies based on large-scale expressed sequence tag (EST) analysis estimated that >60% of human genes undergo alternative splicing, and this number more recently increased to >80% when microarray data became available (Black, 2003; Matlin et al., 2005). More recently, high-throughput sequencing technologies are revealing that 92-94% of human genes undergo alternative splicing (Wang et al., 2008b; Sultan et al., 2008; Pan et al., 2008).

Alternative splicing relies on the same core signal sequences and regulatory sequence elements (SREs) as constitutive splicing. Indeed, alternative splicing patterns are usually determined by the presence of SREs in the regulated exon and its flanking introns. These SREs can be intronic or exonic splicing silencer elements (ISS or ESS) and intronic or exonic splicing enhancer elements (ISE or ESE). Enhancer elements promote the inclusion

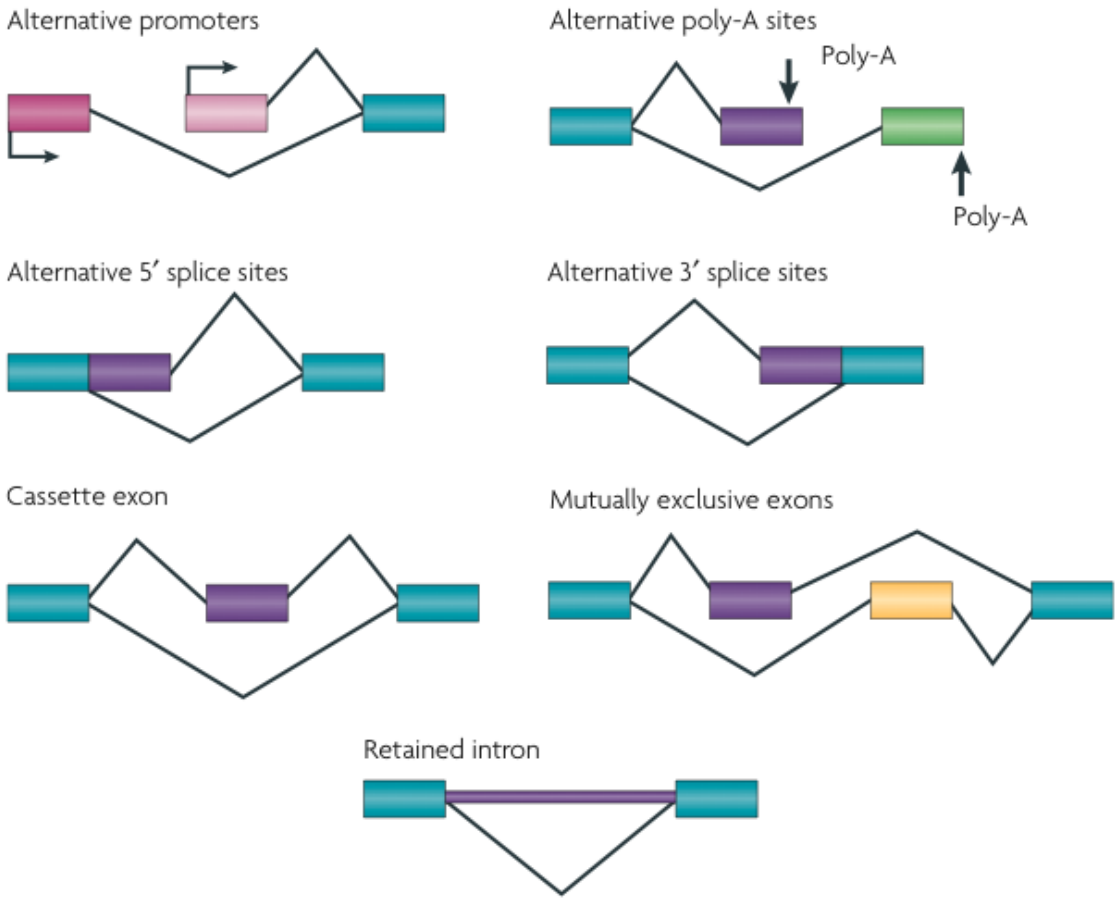


Figure 1.5: **Patterns of alternative splicing.** Constitutive exons are shown as blue boxes, whereas the remaining colors represent the different alternative patterns. The two splicing patterns for each case are represented by the lines above and bellow (Image from Li et al., 2007).

.

of an exon, and silencers promote its skipping or exclusion from the final mRNA. Many of these elements are bound by known RNA-binding proteins (RBPs), such as SR and hnRNPs. hnRNPs typically bind to ESSs and ISSs, whereas SR proteins usually bind to ESEs and ISEs (Figure 1.6)

Many more sequence elements have been identified, but their protein mediators are unknown. Most alternative exons are controlled by the balance of multiple splicing enhancer and silencer elements (reviewed in Sharma and Black, 2006).

Investigations into individual splicing events and, more recently, combined microarray and computational analyses have identified subsets of commonly regulated splicing events that share *cis*-acting elements associated only with alternative splicing. These elements are bound by RNA-binding proteins that are not generally associated with the spliceo-

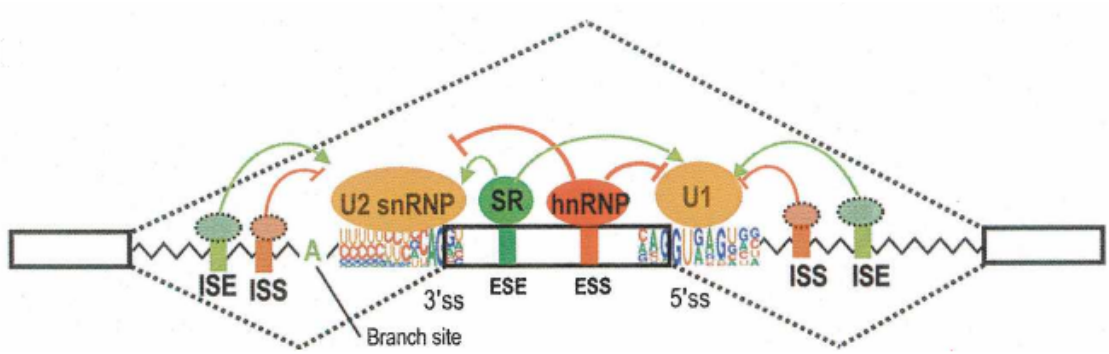


Figure 1.6: **Splicing Signals.** *Cis*-acting elements required for exon-intron recognition and splicing regulation (Image from Wang and Burge, 2008).

.

some, such as FOX 1 and 2 proteins (A2BP1, RBM9), CUG-binding proteins (CUG-BP also known as eTR-3-like or CELF proteins) (CUGBP1, CUGBP2, BRUNOL4), MBNL (MBNL), Nova proteins (NOVA1, NOVA2) NOVA and TIA proteins (TIA1, TIAR). Regulation of alternative splicing in vertebrates involves a dynamic interplay of antagonistic regulatory factors; for example, between the SR and hnRNP protein families, and between pairs of proteins including Nova-PTB, CELF-PTB, CELF-MBNL, TIA-PTB, and PTB-FOX (reviewed in Wang and Cooper, 2007).

Moreover, many elements are not strict silencers or enhancers, rather the position of an element relative to an alternative exon can determine whether it acts positively or negatively, namely Nova (Ule et al., 2006) and FOX2 proteins (Yeo et al., 2009). The authors suggested an RNA map for Nova and FOX2 proteins: binding sites located in the alternative exon (only for Nova) and upstream intron work as splicing silencers leading to exon exclusion, whereas sites in downstream intron act as splicing enchancers promoting exon inclusion.

1.2 Microarray technology

A microarray, in general, corresponds to a glass or polymer slide, onto which pre-defined sequences of DNA are attached at fixed locations. The purpose of a microarray is to detect the abundance of labelled nucleic acids in a biological sample, which will hybridize to the pre-defined sequences of DNA and measured via the label. Since thousands of different DNA molecules may be bound to a microarray it is possible to measure the abundance of many features simultaneously. There are several microarray platforms that differ in fabrication, sample preparation, experimental design and the methods for data analysis. DNA microarrays can be used to measure changes in gene expression levels, to detect

single nucleotide polymorphisms (SNPs), alternative splicing changes, etc.

1.2.1 Microarray fabrication

The use of a collection of distinct DNAs in arrays for expression profiling was first described in 1987 and these early arrays were made by spotting complementary DNA onto filter paper with a pin-spotting device (Kulesh et al., 1987). The miniaturized microarrays appeared in 1995 (Schena et al., 1995) and the commercialization of microarrays was started in 1996 by Affymetrix.

Nowadays, there are two main technologies for making microarrays: robotic spotting and *in-situ* synthesis.

The spotted DNA probes can be cDNA or DNA oligonucleotides presynthesised. The cDNA is a single-stranded DNA synthesized from mature (fully spliced) mRNA using the enzyme reverse transcriptase. The attachment chemistry can be covalent or non-covalent. In the first case, a primary aliphatic amine group (NH_2) previously added to the DNA probe, bounds to the linkers on the glass. In the non-covalent attachment, the bonding is made via electrostatic attraction between the phosphate backbone of the DNA probe and the NH_2 attached to the surface of the glass. The spotting robot itself consists of one or a series of pins arranged as a grid (pin-group) held in a cassette. The pins collect the DNA probes and spot them onto a number of different arrays. After spotting, the pins are washed and new DNA probes are collected to be printed. In the end, the surface of the array can be fixed so that no further DNA can attach to it (Stekel, 2003). This spotting technology is used most by the research laboratories that produce their own microarrays. It is the least expensive technology and only requires the spotting robot. Also, many researchers do not work with model organisms (human, mouse, chicken, yeast, etc.) and this type of microarrays allow any type of cDNA, obtained from clone libraries of non-model organisms to be printed on an array.

The other types of microarray platforms are *in-situ* synthesised oligonucleotide microarrays. For this type of technology the oligos are built up base-by-base on the surface of the array. This takes place by several rounds of synthesis where a covalent reaction occurs between the last added nucleotide and the next one. Each nucleotide added to the oligonucleotide on the glass has a protective group on its terminal position to prevent the addition of more that one base during each round of synthesis. The protective group is then converted to a hydroxyl group before the next round of synthesis. There are different methods for deprotection: photodeprotection and chemical deprotection with synthesis via inkjet technology.

The photodeprotection, also called photolitography, is the basis of the Affymetrix GeneChips. This technique uses light to unprotect the protective group to which further

nucleotides can be added. The light is directed to appropriate features using masks that allow light to pass to some areas of the array but not to others. Each round of synthesis requires a different mask. The mask set is expensive to produce, but once made, it is straightforward to produce a large number of identical arrays. Thus, the Affymetrix technology is mostly used to make large numbers of standard arrays for research in model organisms.

The other type of photodeprotection similar to that described above, directs light via micromirror arrays. This is the method used by Nimblegen and Febit technologies.

The inkjet technology uses chemical deprotection to synthesise the oligonuclotides. At each step, the appropriate nucleotide is fired onto each spot of the array. This process is computer controlled, so any oligonucleotides can be synthesised on the array simply by specifying the sequences in the computer file. Although, being highly flexible, this technology is less efficient for making large numbers of identical arrays (Stekel, 2003). This is the technology used by Rosetta, Agilent and Oxford Gene Technology.

A different approach to the conventional high-density microarrays is the bead array technology used by Illumina. The beadarrays are based on 3-micro silica beads that self assemble in microwells on either two substrates: fiber optic bundles or planar silica slides. The beads are randomly assembled on one of these two substrates and identified using decoding hybridizations. Each bead pool is covered with hundreds of thousands of copies of a specific oligonucleotide that act as the capture sequences in one of Illumina's assays. Once a bead pool is made, it is relatively straightforward to assemble and decode large numbers of arrays containing ∼50000 beads. Because individual arrays are only ∼1.4 mm in diameter, they can easily be arranged into a 96-array matrix, designed for parallel analysis of samples in standard microtiter plates (Gunderson et al., 2004).

1.2.2 Microarray assays

Microarray assays basically consist of the hybridization of nucleic acids (DNA or RNA) extracted from the tissue or cells of interest with the DNA probes on the microarray. The type of extracted sample will depend on the final goal: comparative genomic hybridization or SNPs require DNA, whereas gene expression profiling and alternative splicing studies require RNA. In the case of RNA extraction, this is afterwards converted into cDNA using the enzyme reverse transcriptase. Then, the DNA/cDNA sample is labelled with fluorescent dyes and hybridised to the microarray. After hybridization and washing, the microarray is excited by a laser and scanned at wavelengths suitable for the detection of the fluorescence intensities.

The procedures for some of these steps depend on the technology used: single or two-channel microarrays.

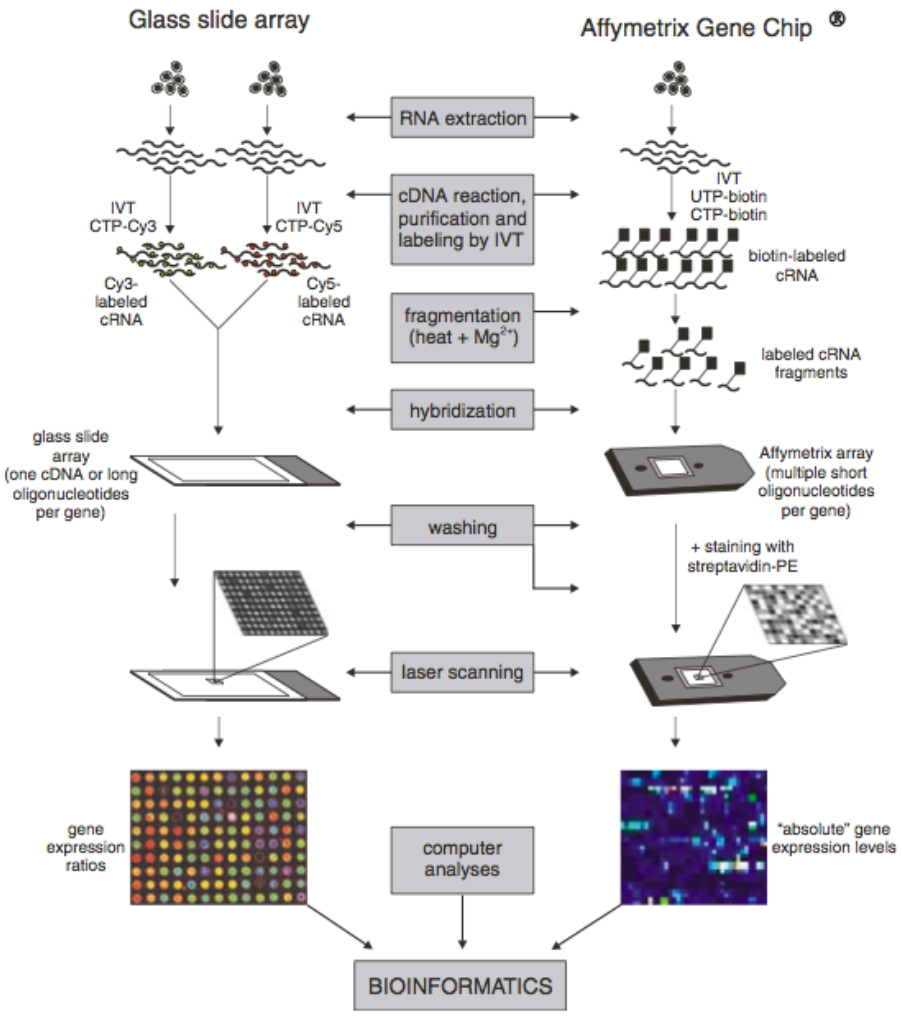


Figure 1.7: **Simplified scheme of the assay steps for two-channel (glass slide) and single-channel (Affymetrix) microarray platforms.** Figure from Staal et al. (2003).

The Affymetrix GeneChip is an example of gene expression profiling in a single-channel platform (Figure 1.7). In this assay the cDNA is converted to complementary RNA and labelled with Biotin-ddUTP. Then, the sample is hybridized into the GeneChip and the microarray is scanned. The GeneChip contains 11 to 20 probe pairs for each gene. Each probe pair has perfect-match (PM) probes to specifically hybridize with the transcripts from the intended gene, and paired mismatch (MM) probes for measuring non-specific hybridization. The idea is that the final expression value is given by the difference between the PM and MM signals and the average of the probe pairs for the same gene. However, in practice the MM probes often cross hybridise with signal from other genes and sophisticated statistics are needed to combine PM and MM data for a given gene.

The other type of microarrays are the two-channel platforms (Figure 1.7), where two samples are labelled with different dyes and competitively hybridised to the same array. The fluorescence emitted by the two dyes is detected by the microarray scanner and reflects the relative amount of the two target samples hybridised to each probe. The commonly used dyes Cy3 and Cy5 emit fluorescence in wavelengths 510-550nm and 630-660nm respectively. The scanner excites both dyes and detects the emission for each of the red (Cy5) and green (Cy3) channels.

In two-channel platforms, the final expression value is generally given by the comparison between the amount of fluorescence emitted by the two dyes.

One important requirement in the microarray assay is the replication of experiments. There are two main types of replicates: biological replicates and technical replicates. Biological replicates are arrays that use biological samples from different individual organisms, pools of organisms or flasks of cells, but yet compare the same treatments. This type of replicate allows the assessment of the natural variability of the system. Technical replicates provide information on the differences between samples from the point at which an individual sample was processed and separated independently. Using an experimental design which uses both types of replication enables us to estimate both the biological variance and the noise/error due to the technical differences in array processing (Causton et al., 2003).

The two-channel platforms typically involve a specific type of technical replication, designated as a dye-swap. These are replicate arrays with the same samples hybridised but swapped fluorescent labelling. For example, sample A is labelled with Cy3 and sample B with Cy5 in the first array, but in the second array the sample A is labelled with Cy5 and sample B with Cy3. Dye swaps are used to estimate technical dye bias in some genes, because the two fluorescent dyes may be differentially incorporated into DNA and they emit fluorescence in different wavelengths.

1.2.3 Microarray applications

Microarrays applications have evolved and transcended the initial goal of gene expression profile. Microarray can be used for assessing single nucleotide polymorphisms (SNPs), copy number variation, interactions between proteins and DNA or RNA, methylation state, alternative splicing changes, etc (Bier et al., 2008).

In a gene expression profiling experiment the abundance levels of thousands of genes may be simultaneously monitored to study the effects on gene expression of certain treatments, diseases, developmental stages, presence of pathogens or other organisms. Nowadays, the expression of approximately 29000 genes can be measured simultaneously using microarrays with probes matching only the $3'$ end of each gene or using probes spread across the full length of the gene (Robinson and Speed, 2007). The expression of several non-coding RNAs (microRNAs and small nuclear RNAs) can also be detected using specific microarrays. Non-coding RNAs are emerging as a major component of the regulatory circuitry that underlies the development and physiology of complex organisms (Li and Ruan, 2009).

Microarrays technology development have been also extremely useful for SNPs genotyping by allele-specific hybridization to oligonucleotides probes, allowing the detection of approximately 1 million SNPs in a single experiment (Syvänen, 2005).

Different microarray technologies have been used for genome-wide copy number variation detection: SNPs genotyping microarrays; comparative genomic hybridization (CGH); tiling microarrays (Carter, 2007).

Comparative Genomic Hybridization measures DNA copy number differences between a test and reference, allowing the detection of loss, gain and amplification of the copy number at the levels of chromosomes (Carter, 2007).

Tiling or high-density oligonucleotide microarrays cover the entire genome with oligonucleotide probes overlapping with some base-pairs shift. Potential uses for such whole-genome arrays include empirical annotation of the transcriptome, interactions between proteins and DNA or RNA by chromatin-immunoprecipitation-chip, analysis of alternative splicing, characterization of the methylome (the methylation state of the genome), SNPs discovery and genotyping and genome resequencing (Mockler et al., 2005).

Microarray-based methods have also been used for identification of interactions between proteins and DNA or RNA. Sequences bound to a specific protein can be isolated by chromatin immunoprecipitation (ChIP) or RNA immunoprecipitation (RIP) and these fragments can be then hybridized to tiling microarrays (ChIP-chip or RIP-chip) allowing the determination of protein binding site occupancy. For example, transcription factor binding sites throughout the genome can be identified using ChiP-chip assays. Using the same technique for other proteins allow the identification of promoter regions, enhancers,

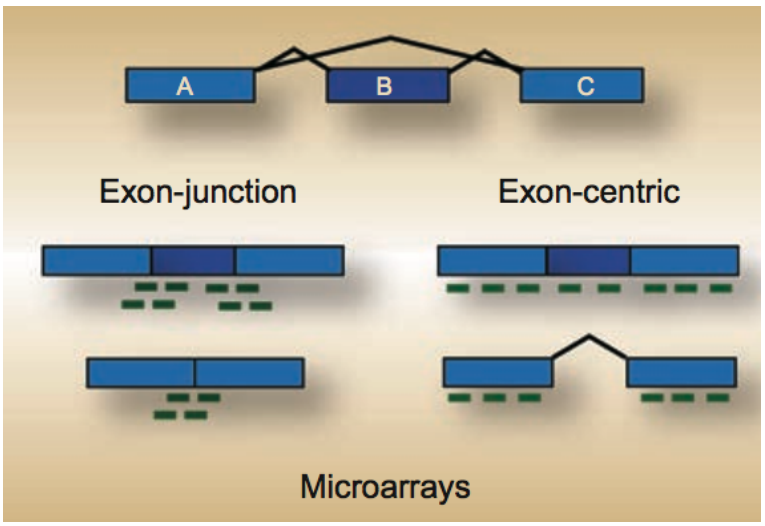


Figure 1.8: **Splicing-sensitive microarrays.** The scheme shows the probes (green) distribution across exons according to the main two types of splicing-sensitive microarrays: exon and exon-junction centric platforms (Image from McKee and Silver, 2007).

repressors and silencing elements, etc (Durand-Dubief and Ekwall, 2009). RIP-chip has been applied to identify relationships between transcripts and regulatory RNA-binding proteins, namely splicing factors (Gama-Carvalho et al., 2006).

Splicing-sensitive microarrays were also developed for large-scale identification of splicing differences between two RNA populations. These microarrays typically contain spotted oligonucleotide probes that are complementary to individual exons and/or exon-exon junctions for thousands of genes. Splicing microarray reliability can be improved by higher coverage using both exon probes and exon-junction probes (reviewed in Wang and Cooper, 2007). Several splicing microarray platforms were developed consisting essentially in two types of approaches exon and exon-junction centric platforms (Figure 1.8). The two approaches present distinct advantages in their ability to measure transcript structure due the location of the probes. The exon-centric platforms are more appropriate to identify novel splicing events since probes are designed for well annotated and predicted exons, whereas with exon-junction platforms transcript architecture directly targeting pre-determined arrangements of exons can be assessed (reviewed in McKee and Silver, 2007).

In addition, several studies have combined the use of different microarray types to answer more complex questions, namely correlation between gene expression and copy number variations (Stransky et al., 2006), gene expression profile and methylation (Martín-Subero et al., 2009), etc.

The commercially available microarray platforms present also a broad range of genomes for animal (human, mouse, rat, pig, bovine, canine, chicken, Rhesus Macaque, Xenopus

tropicalis, Xenopus laevis, Zebrafish, Drosophila, etc.), plants (Arabidopsis, Barley, Citrus, Cotton genomes Array, Maize, Medicago, rice, Soybean, Sugar Cane, tomato, grape, Wheat, etc.), fungi (Yeast - Saccharomyces cerevisiae) and bacteria (Escherichia coli, Pseudomonas aeruginosa, Staphylococcus aureus).

1.2.4 Microarray data analysis

The enormous volume of data generated by microarray technology, associated with the need to properly handle, analyse, interpret and make use of the data, has been a challenge to many researchers in statistics, computational sciences and biology. As a consequence, a great diversity of software has been developed and several statistical methods have been proposed to analyse this type of data. There are several types of programs based on different programming languages and also some web-based systems: AFM (Breitkreutz et al., 2001), Babelomics (Al-Shahrour et al., 2006), GEPAS (Montaner et al., 2006), dChip (Li and Wong, 2001), GenePattern (Reich et al., 2006), R and BioConductor systems (R Development Core Team, 2009; Gentleman et al., 2004).

Microarray assays are complex experiments with many steps requiring the use of bioinformatics tools. In a broad way these steps are: image analysis; data pre-processing and normalization; statistical analysis; database annotation.

Data Pre-Processing and Normalization

Data preprocessing and normalization programs perform raw data merging, format conversion, quality assessment and data normalization before higher level analysis.

After microarray scanning, microarray data analysis begins with image processing. Data extraction from images involves several steps: aligning and overlaying the red and green images; identification of the arrayed features; detection of saturated spots; determination of appropriate number of pixels for a spot; etc. Image processing is a critical step as this primary data collected from each experiment will be the starting point for all downstream analysis.

After the image processing we obtain the specific hybridization intensity for each probe (foreground) and the intensity of non-specific hybridisation (background). Background correction is based on the premise that background estimates represent the non-specific hybridisation of labelled target to the glass and also some natural fluorescence of the glass slide itself. Thus, each probe intensity should be corrected for the background and there are several background correction methods (Stekel, 2003).

Microarray data is usually transformed from the raw intensities into log-intensities. The are several reasons for this transformation but the main purpose is to make the data and the model error terms closer to a normal distribution (i.e. to make the distribution of

the data more symmetrical); to reduce the influence of outliers, especially when they are at one end of the distribution and to make effects that are multiplicative on the raw scale additive on a transformed scale. The ratio of the intensities between two sample types (or red and green channels in two-colour microarrays) is transformed into the difference between the logs of the intensities of both samples. It is common to use logarithm to base 2. Therefore, a log-ratio of 0 corresponds to a gene with no expression change, while +1 and -1 will correspond to a 2-fold up-regulated and down- regulated genes respectively (Stekel, 2003).

Finally, the microarray data should be normalized to adjust for effects which arise from variation in the microarray technology rather than from biological differences between the RNA samples or between the printed probes (Smyth and Speed, 2003). Microarray normalization can be split in two phases: normalization within and between-arrays. The normalization within-arrays corrects spatial effects (heterogeneity in intensities through the microarray) and dye bias in two-colour microarrays. The dye bias corresponds to a difference in the red and green probe intensities (two different fluorescent dyes are used to quantify the samples and their intensity is measured at different wavelengths) that is not the result of a biological difference in gene expression. The normalization between-arrays corrects expression intensities so that the intensities or log-ratios have similar distributions across a series of arrays and thus making it possible to compare samples hybridized in different slides (Smyth and Speed, 2003).

Class Comparison and Discovery

One of the most important goals in microarray studies is to compare pre-specified classes (or samples) and identify genes that are differentially expressed between them. Class comparison methods are supervised in the sense that they utilize the information of which specimens belong to which classes. This is in contrast to methods such as cluster analysis for class discovery which do not utilize any information about class membership (Stekel, 2003).

In class comparison, the identification of differentially expressed genes can be considered in two stages. First it is necessary to select a statistic which will rank the genes in order of evidence for differential expression. Then, a critical-value for the ranking statistic is chosen above which any value is considered to be significant. The ranking step is more important because in many microarray studies the aim is to identify a number of candidate genes for confirmation and further study. Since often only a limited number of genes can be followed up, it is sometimes more important to identify the most likely candidates than to select a large list where most genes will not be a focus. The statistics used can be the classical t-test and use the p-value for selecting the cut-off. Another approach was

described by Smyth (2004), where linear model and empirical Bayes methods are used to identify differentially expressed genes.

A known problem when performing statistical tests on many genes in parallel is the multiplicity of p-values. By the general definition, p is the probability of observing data as extreme as yours under the assumption that the null hypothesis is true and it also represents the probability of a false positive. So, a gene with a p-value of 0.01 will have 1% of probability of being a false positive. Although 1% is acceptable for one test, when applying several tests on the same dataset this 1% can result in a large number of false positives. Microarrays today typically have close to 20000 genes and applying the rule described above one would expect to find 200 false positives by chance. This is a statistical problem caused by the analysis of a large number of genes. There are several methods to correct this artifact: estimated false positive rate; Bonferroni correction; Benjamini and Hochberg method (Benjamini and Hochberg, 1995); etc.

The goal of class discovery methods is to create classes of specimens for which gene expression is different. These methods can be used to identify groups of specimens and then arrange the groups so that the closest groups are adjacent. There are three ways to group microarray data: analyse the genes, analyse the samples, or both. In the first case, the aim is to find the groups or sets of genes with similar expression across samples. In the case of grouping samples, each sample profile will be measured using the expression values of the genes in that specific sample. Although these two analyses are very different and used to achieve different results, from the perspective of data analysis methods they are essentially the same. The only important difference lies in the relative number of samples and genes (Stekel, 2003). One can group samples/genes using clustering methods or by dimensionality reduction.

The goal of clustering methods is to form groups such that samples within a group are more similar to one another than objects in different groups. From the clustering algorithms, the hierarchical clustering is the most widely used for gene expression data (reviewed in D'haeseleer, 2005; Kerr et al., 2008). This method produce a nested sequence of clusters which can be graphically represented with a tree, called a dendogram. The height of each horizontal line represents the distance between the two samples or clusters that it merges, with greater heights representing greater distances. The hierarchical clustering method begins with a distance matrix of the samples obtained with a similarity measure (for example, Pearson correlation or Euclidean distance). After this, the nearest samples are joined together in the tree to form a cluster. Then, a new distance matrix is obtained substituting the clustered samples by the newly formed cluster. Once more, the nearest samples or cluster of samples are clustered together and this is repeated until all samples and clusters are linked (reviewed in D'haeseleer, 2005; Kerr et al., 2008).

Principal Component Analysis (PCA)(Jollife, 1986) is a method that reduces the dimensionality of a high-dimensional data set, like microarrays with thousands of genes, while retaining as much information as possible. Basically, PCA is a multivariate procedure which rotates the existing axes to new positions in the space defined by the data, such that maximum variabilities are projected onto the axes. Essentially, the set of correlated variables are transformed into a set of uncorrelated variables, called the principal components. The principal components are linear combinations of the original variables, and they are ordered by reduced variability. Singular value decomposition (SVD) is an algorithm for computing the PCA of a non-symmetric matrix. This method is used in microarray data analysis, since the data sets have different numbers of columns (samples) and rows (genes). Methods based on principal components are not specifically intended to discover groups, like the clustering methods, but to find structure in the expression profiles. These methods have been successfully applied to microarray data (Palmer et al., 2008; Hua et al., 2009).

1.3 Bioinformatic tools for large scale approaches

During the past few years, there have been enormous advances in genomics and molecular biology. The challenge of interpreting the vast amounts of data from microarrays and other high throughput technologies has led to the development of new tools in the fields of computational biology and bioinformatics.

1.3.1 R and BioConductor

R is an open source system for statistical computation and graphics (Figure 1.9) (`http://www.r-project.org/`). It consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files. This language is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc) and graphical techniques. Moreover, R is highly extensible via *packages*. There are approximately 1900 packages available through the Comprehensive R Archive Network (CRAN) of Internet sites covering a very wide range of modern statistics and interface to other computational systems/languages. For instance, *RMySQL* is a R package interface to the MySQL databases, allowing the user to run multiple and complex queries using a local database.

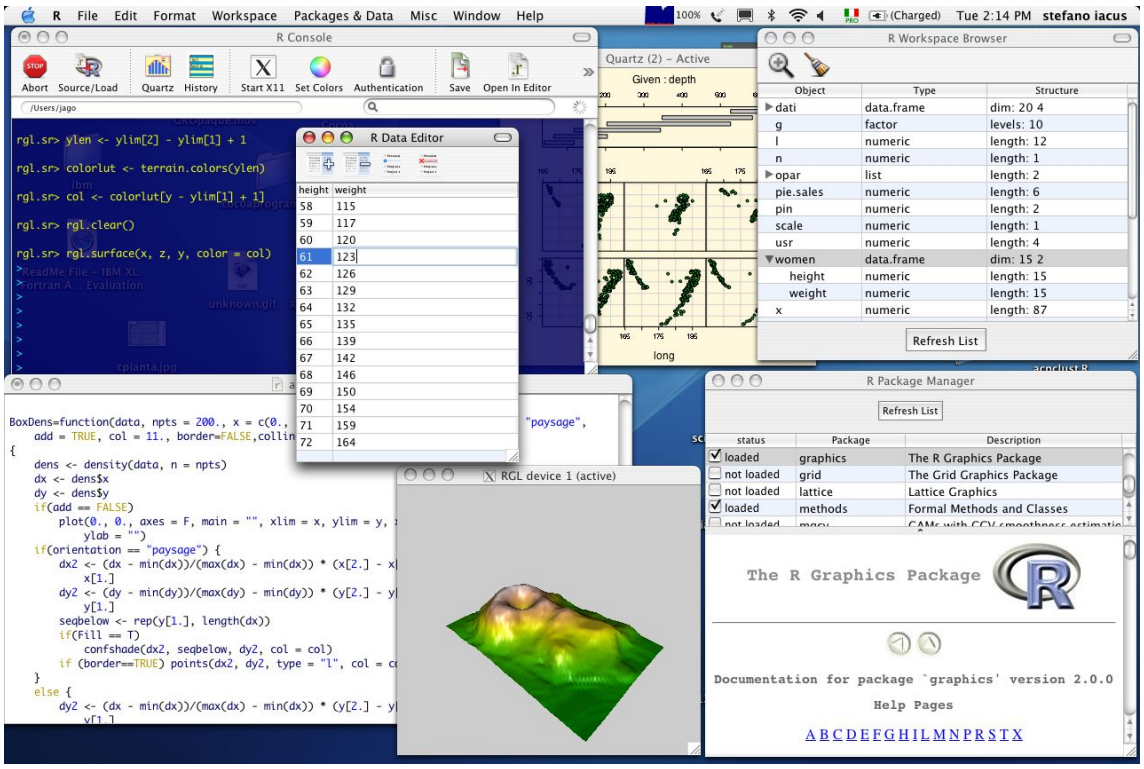


Figure 1.9: **R system for statistical computation and graphics.** Image obtained from (C) R Foundation, `http://www.r-project.org`.

Bioconductor is an open source and open development software project to provide tools for the analysis and comprehension of genomic data (`http://www.bioconductor.org`) Bioconductor is based primarily on the R programming language, but does contain contributions in other programming languages. Most Bioconductor components are distributed as R packages, which are add-on modules for R. Initially most of the Bioconductor software packages focused primarily on DNA microarray data analysis. As the project has matured, the functional scope of the software packages broadened to include the analysis of all types of genomic data, such as sequence or SNP data.

1.3.2 Molecular biology databases

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. This huge increase in genomic information has led to an absolute requirement for computerized databases.

A database is a large and structured collection of data, usually associated with computerized software designed to update, query, and retrieve components of the data stored

within the system. A simple database might be a single file containing many records, each of which includes the same type of information. Nowadays, one can find databases with information relative to genomes, nucleotide sequences, genes, SNPs, proteins, microarray data, etc. Moreover, this information is available not only for human but also for other eukaryotic species and bacteria.

Most well known Genomic and Nucleotide databases are Entrez and Nucleotide databases (developed by National Center for Biotechnology Information - NCBI, `http://www.ncbi.nlm.nih.gov/`), Ensembl (joint project between EMBL - EBI and the Wellcome Trust Sanger Institute `http://www.ensembl.org/index.html`) and UCSC Genome Browser (developed by University of California Santa Cruz, `http://genome.ucsc.edu/`). A major advantage of these well structured and updated databases is the possibility to download the entire database and run locally for multiple and complex searches.

Microarray and gene expression databases have been also growing and the more commonly used are Gene Expression Omnibus (GEO developed by NCBI, `http://www.ncbi.nlm.nih.gov/projects/geo/`) and Array Express (developed by EBI-EMBL, `http://www.ebi.ac.uk/microarray-as/ae/`). The submission to databases require microarray data to comply with MIAME guidelines (Minimum Information About a Microarray Experiment), which describes the minimum information that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment (Brazma et al., 2001). The six most critical elements contributing towards MIAME are:

1. The raw data for each hybridisation (e.g., CEL or GPR files);

2. The final processed (normalised) data for the set of hybridisations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study);

3. The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment);

4. The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridisations are technical, which are biological replicates);

5. Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number);

6. The essential laboratory and data processing protocols (e.g., what normalisation method has been used to obtain the final processed data).

The current count for available datasets on these databases is 13980 microarray experiments for GEO (corresponding to 359915 samples hybridized) and 9164 for ArrayExpress (260515 samples), from which 5620 are common to both databases.

1.4 Objectives

Differential alternative splicing has been associated with developmental stages, tissue types and cancer (Matlin et al., 2005; Shin and Manley, 2004; Singh and Valcárcel, 2005; Wang and Cooper, 2007; Cooper et al., 2009). According to the current model, regulation of alternative splicing uses combinatorial interactions of many positively and negatively acting proteins, and specific splicing decisions most likely result from differences in the concentration and/or activity of these proteins (Matlin et al., 2005; Shin and Manley, 2004; Singh and Valcárcel, 2005). One should expect differential gene expression of splicing factors through several cell types. The development of microarray technology allowed the use of large-scale studies to systematically address this question.

The main goal of the present study has been to generate microarray-based predictions for understanding the alternative splicing code that controls and coordinates the transcriptome.

First, I have aimed to systematically assess by microarray data analysis the widespread gene expression of splicing regulators during cell differentiation, in differentiated tissues and in cancer. The identification of genes with differential expression can indicate putative regulators for specific splicing decisions.

Finally, I have tried to establish the link between changes in splicing factors expression and alternative splicing profiles in cancer, combining results from gene and splicing microarrays. The analysis centered in the identification of motifs enriched in the cancer-associated splicing events that resemble binding sites for cancer misregulated splicing factors.

Chapter 2

Tissue-specific splicing factor gene expression signatures

The original work described in this chapter has been published in: Grosso AR, Gomes AQ, Barbosa-Morais NL, Caldeira S, Thorne NP, Grech G, von Lindern M, Carmo-Fonseca M (2008) Tissue-specific splicing factor gene expression signatures, Nucleic Acids Research. 36(15):4823-32.

I would like to stress that some of the results presented and discussed in this section are the product of collaborative work. Nuno Barbosa-Morais was responsible for selection of splicing-related genes and annotation of microarray probes. Marieke von Lindern and Godfrey Grech provided one of the microarray data sets for erythroid differentiation in vitro. Anita Q. Gomes and Sandra Caldeira were responsible for the cell culture for validation. Real-time quantitative PCR results were equally obtained by me and Anita Q. Gomes.

Keywords: alternative splicing; splicing regulation; splicing factor; spliceosome

Abstract: The alternative splicing code that controls and coordinates the transcriptome in complex multicellular organisms remains poorly understood. It has long been argued that regulation of alternative splicing relies on combinatorial interactions between multiple proteins, and that tissue-specific splicing decisions most likely result from differences in the concentration and/or activity of these proteins. However, large-scale data to systematically address this issue have just recently started to become available. Here we show that splicing factor gene expression signatures can be identified that reflect cell type and tissue specific patterns of alternative splicing. We used a computational approach to analyze microarray-based gene expression profiles of splicing factors from mouse, chim-

panzee and human tissues. Our results show that brain and testis, the two tissues with highest levels of alternative splicing events, have the largest number of splicing factor genes that are most highly differentially expressed. We further identified SR protein kinases and snRNP proteins among the splicing factor genes that are most highly differentially expressed in a particular tissue. These results indicate the power of generating signature-based predictions as an initial computational approach into a global view of tissue-specific alternative splicing regulation.

2.1 Introduction

Alternative splicing generates multiple mRNA products from a single gene, thereby increasing transcriptome and proteome complexity. In contrast to the prokaryotic rule of one gene-one polypeptide, alternative splicing expands the protein coding potential of eukaryotic genomes by allowing a single gene to produce proteins with different properties and distinct functions. Several studies based on large-scale expressed sequence tag (EST) analysis estimated that >60% of human genes undergo alternative splicing, and this number more recently increased to >80% when microarray data became available (Black, 2003; Matlin et al., 2005). Alternative splicing is regulated in response to signaling pathways, and is specific to a developmental stage and tissue type.

The removal of introns from precursor mRNAs requires accurate recognition of splice sites by the spliceosome, an assembly of uridine-rich small nuclear RNAs (U snRNAs) packaged as ribonucleoprotein particles (snRNPs) that function in conjunction with numerous non-snRNP proteins (Jurica and Moore, 2003; Nilsen, 2003). The selection between different splice sites on a particular pre-mRNA substrate relies on an intricate interplay involving the cooperative binding of trans-acting splicing proteins to cis-acting sequence elements in the pre-mRNA. In mammals, these cis-elements include short and highly degenerate $5'$ and $3'$ splice signals, additional regulatory sequences termed splicing enhancers and silencers located in either exons or introns, the sizes of the exons and introns and secondary structures of the pre-mRNA. The trans-acting factors are commonly classified as splicing activators or repressors depending on whether they facilitate or suppress the assembly of snRNPs onto splice sites. However, many of these factors are also essential for constitutive splicing, making it unrealistic to distinguish between proteins required for the operation and regulation of the splicing reaction (Maniatis and Tasic, 2002; Shin and Manley, 2004). Contrasting with the multitude of sequence-specific DNA-binding proteins that control transcription, there are very few known regulatory proteins that selectively control the splicing of specific genes. Although such factors exist, and a good example is the brain-specific NOVA1 protein in mammals (Jensen et al., 2000), in the vast majority

of cases splicing factors are ubiquitously expressed and modulate splicing of several genes in distinct cell types. Indeed, specificity of splicing regulation is largely achieved with nonspecific RNA-binding proteins (Singh and Valcárcel, 2005).

According to the current view, regulation of alternative splicing uses combinatorial interactions of many positively and negatively acting proteins. Tissue-specific splicing decisions could therefore result from differences in the concentration and/or activity of these proteins (Matlin et al., 2005; Shin and Manley, 2004; Singh and Valcárcel, 2005). An immediate prediction from this model is that the relative abundance of multiple splicing proteins should differ in a tissue-specific manner. To explore this idea, we performed a large-scale computational analysis of mRNA expression data obtained from DNA microarray studies of different cell types and tissues derived from human, chimpanzee and mouse. Our results show for the first time that splicing factor gene expression signatures can be identified that correlate with tissue-specific patterns of alternative splicing.

2.2 Material and Methods

2.2.1 Selection of splicing-related genes

A list of 254 human splicing-related genes and several murine orthologues was previously described (Barbosa-Morais et al., 2006). The remaining mouse genes were identified in Ensembl (Hubbard et al., 2007) (`http://www.ensembl.org`), through the Family classification and BLAST (Altschul et al., 1990) search, and by searching SwissProt (Bairoch et al., 2005) (`http://us.expasy.org/sprot/`) with appropriate keywords. Perl scripts, relying on Bioperl (Stajich et al., 2002) (`http://www.bioperl.org`) and modules from the Ensembl PERL API (Stabenau et al., 2004) were used for consistent annotation of genes and subsequent cross-linking with the Affymetrix probe set annotation. Annotation for the selected probe sets was validated with a Perl script. The first step of the pipeline consisted in BLASTing (Altschul et al., 1990) and/or BLATing (Kent, 2002) each probe against both the respective transcriptome (comprising RefSeq (Pruitt et al., 2007), GenBank (Benson et al., 2007) and transcripts from the UCSC Genome Browser database (Kuhn et al., 2007)) and genome (Mouse mm8 and Human hg18, NCBI 36)). The program subsequently parsed the outcome and extracted the associated transcriptomic and genomic annotations from the tables in the UCSC genome annotation database (Kuhn et al., 2007).

2.2.2 Microarray data pre-processing

All the microarray data analysis was done using R and several packages available from CRAN (R Development Core Team, 2009) and Bioconductor (Gentleman et al., 2004).

The raw data (CEL files) were normalized and summarized with the Robust MultiArray Average method from the *affy* package (Gautier et al., 2004). An initial quality assessment was done to remove microarrays with poor quality, using quality diagnostics with probe level models and array quality control metrics for all arrays (average background was $<$ 200, scale factors $<$ 6, percentage of present calls, RNA degradation for GAPDH and beta-actin - $3'/5'$ ratio).

2.2.3 Cell culture and real-time quantitative PCR

C2 mouse myoblasts were cultured at 30% confluence in DMEM supplemented with 20% FCS. For the differentiation experiments the cells were grown in DMEM containing 20% FCS until they reached 90% confluency. At this stage the cells were changed to low serum media (DMEM supplemented with 2% horse serum differentiation media) and allowed to differentiate for a maximum period of four days. Primary mouse erythroid progenitors were obtained from fetal livers of E12.5 mouse embryos and were subject to differentiation in stem-Pro-34 medium supplemented with Epo and iron-saturated human transferrin as described previously (Drissen et al., 2005). The C2 cell RNA samples used in the qRT-PCR experiments were collected at days 0, 1 and 2 after changing to differentiation media. Primary mouse erythroid RNA was collected at 0h, 24h, 36h, 48h and 60h after induction of differentiation. The RNA was extracted using the RNeasy extraction kit according to the manufacturer's instructions (Qiagen) and treated with RNase-free DNaseI (Roche Diagnostics) to remove any possible genomic DNA contaminant. The concentration of RNA was determined using the Nanodrop (Nucliber) and RNA quality was assessed by gel electrophoresis. Only samples yielding distinct 28S and 18S bands and A260/A280 ratios between 1.8 and 2.1 were used in this study. Production of cDNA was carried out using Superscript II reverse transcriptase following the manufacturers protocol (Invitrogen). 0.6 μg of total RNA were used in a 20 μl reaction volume. Isolated cDNA from brain, heart, kidney, liver and testes was purchased from Ambion. A total of 30 ng of cDNA was used for each SYBR Green measurement.

The primers used in the qRT-PCR assay (Annex Table A.1.1) were designed with the Primer3 programme (Rozen and Skaletsky, 2000). The cDNA was amplified in 25 μl reactions containing 50% of SYBR Green PCR master mix (Applied Biosystems). Primers were added at a final concentration of 300nM, which proved to be the best concentration for all the sets of primers tested. All reactions were performed in the ABI7000 Sequence Detector (Applied Biosystems).

The relative quantification of mRNA levels at the various C2 differentiation stages was calculated using 18S as an endogenous reference and the sample at day 0 as the calibrator. For the erythropoiesis experiments we used Rnase Inhibitor as an endogenous reference

and the sample at 0h as the calibrator. For the adult tissues experiments, RNU6A was used as the endogenous reference. The quantities obtained for each gene were extracted from a standard curve of CT versus quantity of mRNA obtained from a serial dilution of either a mix of C2 cell cDNA extracts or a mix of erythropoietic progenitors at the stages of differentiation used for the analysis. For tissue samples, the standard curve was obtained from serial dilutions of a mix of all tissues.

2.3 Results

2.3.1 Splicing factor expression during cell differentiation

To study splicing factor expression during differentiation, we first established a list of human and mouse genes associated with splicing and next we compared the corresponding expression profiles from data sets obtained from microarray studies that analyzed cell differentiation. A list containing 254 human genes associated with splicing was previously reported by Barbosa-Morais et al. (2006).

Here, we searched for the respective orthologues in the mouse genome. Both human and mouse lists contain genes that encode known splicing factors, spliceosome-associated proteins, and proteins with a domain structure similar to bona fide splicing factors (Barbosa-Morais et al., 2006). We selected transcript profiling studies performed with myotube, adipocyte and erythroid cells differentiated in vitro and whole mouse testis collected from birth to adulthood. In total we studied four distinct differentiation processes and for each process we analyzed two independent data sets covering a total of 126 arrays (Table 2.1). (Annex Table A.1.2). We identified 181 splicing-related genes (SRGs) for which 240 probe sets are present in the Affymetrix Murine Genome U74v2 platform that was used in all selected microarray studies (Annex Table A.1.3).

Table 2.1: Microarray data sets used to study mouse differentiation processes. The GEO accession number, references and number of arrays analyzed are indicated. The time range refers to the total differentiation period. The number of time points studied for each differentiation process is also indicated. For our expression analysis, T1 and T2 correspond to the indicated number of differentiation hours.

Description	**Data Set ID**	**GEO Acc. Number**	**Reference**	**Arrays Number**	**Time Range**	**Points**	**Times for Fold-Change**	
							T1	**T2**
Myogenesis	Myog1	GSE989	Tomczak et al. (2004)	23	-24h - 240h	8	24h	48h
	Myog2	GSE1984		10	0 - 48h	5	24h	48h
Adipogenesis	Adip1	GSE2192	Akerblad et al. (2005)	15	0 - 240h	4	48h	96h
	Adip2		Burton et al. (2004)	13	0 - 96h	7	48h	96h
Spermatogensis	Sperm1	GSE640	Schultz et al. (2003)	12	24h - 1440h	9	336h	720h
	Sperm2	GSE926	Shima et al. (2004)	19	0 - 1344h	11	336h	720h
Erythropoiesis	Ery1	GSE628	Welch et al. (2004)	17	0-30h	6	15h	30h
	Ery2		Von Lindern, unpubl.	17	0-60h	5	30h	60h

All expression values were obtained from Gene Expression Omnibus (Barrett et al., 2005) (`http://www.ncbi.nlm.nih.gov/projects/geo`). Data for myogenesis were obtained from published studies using the in vitro model of C2C12 myoblasts undergoing differentiation induced by serum restriction (Tomczak et al., 2004; Zhao et al., 2006). Adipocyte differentiation in vitro was induced by hormonal treatment on two distinct models: the 3T3-L1 preadipocyte cell line (Burton et al., 2004), and NIH-3T3 fibroblasts (Akerblad et al., 2005). Two distinct cell models were also used to analyze erythroid differentiation in vitro. One model consisted of G1E cells derived from GATA-1-null embryonic stem cells; these cells proliferate in culture as immature erythroblasts and undergo terminal erythroid maturation when GATA-1 function is restored (Welch et al., 2004). The other model consisted of primary erythroid progenitors from mouse fetal livers; these cells proliferate in serum-free medium under the control of erythropoietin (Epo), stem cell factor (SCF) and dexamethasone (Dex), and undergo terminal differentiation when exposed to Epo in the absence of SCF and Dex (Drissen et al., 2005). Spermatogenesis was examined in vivo (Schultz et al., 2003; Shima et al., 2004).

To test whether the two data sets corresponding to the same differentiation process were temporally synchronized, we performed a time-course analysis of the expression level of the following differentiation marker genes: the muscle specific troponin C (Tnnc1) (Hastings and Emerson, 1982) and Ca2+ channel ryanodine receptor 1 (Ryr1) (MacLennan et al., 1990); the adipogenic complement factor D adipsin (Cfd) (Djian et al., 1985) and peroxisome proliferator-activated receptor (Pparγ)(Akerblad et al., 2005); the erythroid specific markers glycophorin A (Gypa) (Lahlil et al., 2004) and Slc4a1 (Paw et al., 2003); the male germ cell lineage markers lactate dehydrogenase C (Ldhc) (Bonny et al., 1998) and phosphoglycerate kinase 2 (Pgk2) (McCarrey et al., 1996). For myogenesis, adipogenesis, and spermatogenesis the distinct data sets were approximately synchronous and were directly used as biological replicates (Figure 2.1). For erythroid differentiation, maturation of the cell type used in one study (G1E-ER4 cells) occurred significantly faster than that of primary fetal liver progenitors used in the other study. This difference was corrected considering that the last time points of both experiments were biologically equivalent (Figure 2.1).

Next, for each differentiation process, we searched for variation in expression of splicing-related genes along time. For each splicing-related gene on each data set, we estimated the Pearson correlation coefficient between expression level and differentiation time point. Only genes with absolute correlation values higher than 0.75 (p-values < 0.05, corrected for multiple hypotheses testing using the Benjamini and Hochberg method (Benjamini and Hochberg, 1995)) in both data sets were selected for further analysis. The Pearson correlation coefficients of this subset of genes were used to cluster the microarray data

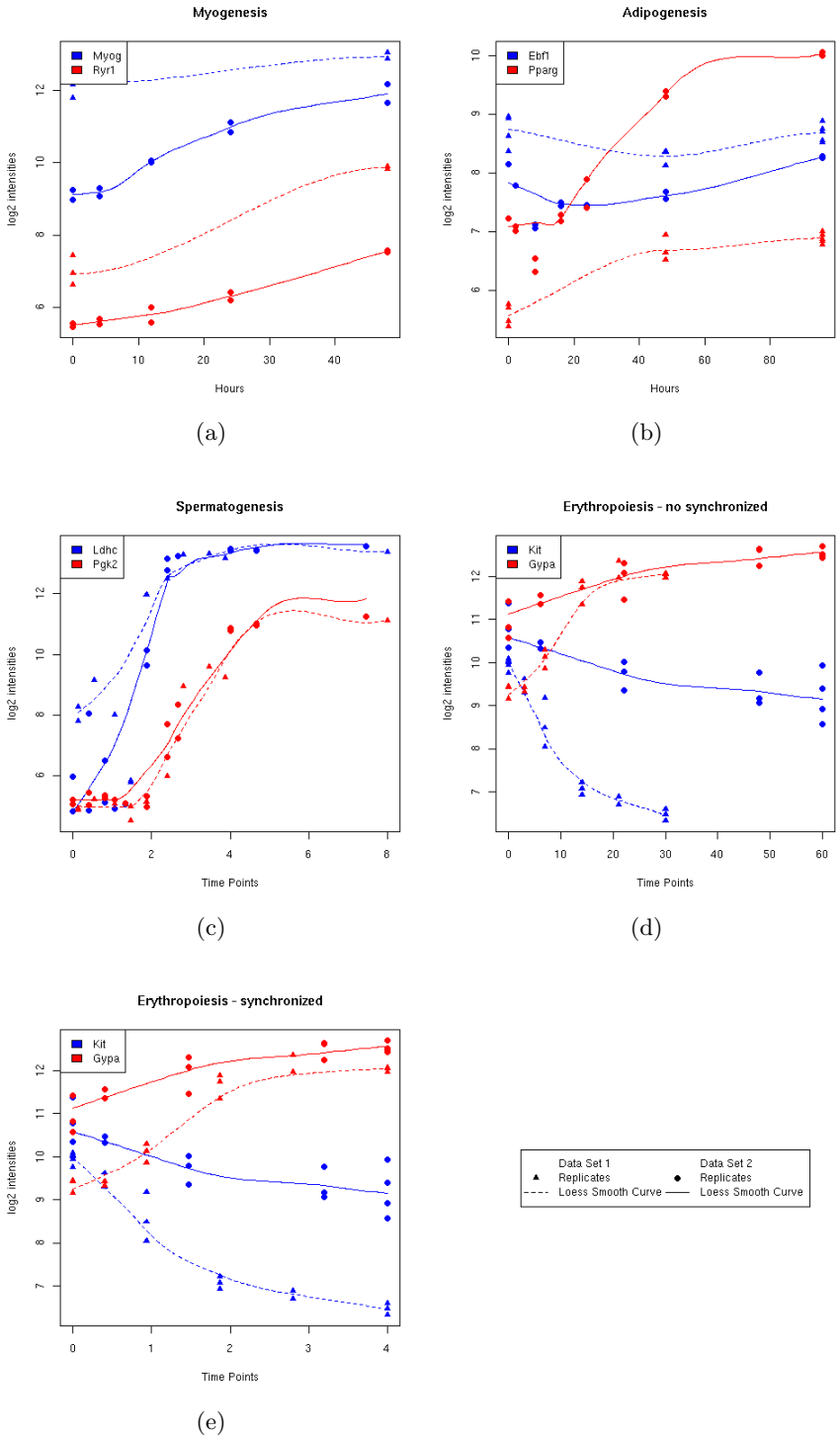


Figure 2.1: **Time-course analysis of the expression level of specific differentiation marker genes.** a) Myogenesis b) Adipogenesis c) Spermatogenesis d) Erythropoiesis non-synchronized and e) Erythropoiesis after synchronization. The gene expression changes though time for each data set are represented using the $\log_2$ intensities of the replicates (triangles and circles) and the fitted loess smooth curves (dotted and solid lines).

sets (Figure 2.2). The hierarchical clustering results revealed consistency between the two data sets for each differentiation process, indicating that similar groups of genes were found up- or down-regulated in the two independent experimental studies performed with each cell type. The only exception was found for adipogenesis data sets, where the different expression patterns for some splicing-related genes can be due to the distinct cell lines used in both experiments. As shown in Figure 1, during myotube and erythroid differentiation most splicing-related genes presented a negative correlation, meaning that the expression decreased along time. In contrast, several splicing-related genes increased expression during adipocyte and sperm cell differentiation.

2.3.2 Identification of cell-type specific variations in splicing factor expression

To identify cell-type specific variations in splicing factor expression, we had to compare microarray data sets derived from different biological systems and experimental assays. To address this issue, we developed a new approach that is based on regression modeling methods. Polynomial models were fitted to the splicing factor expression profiles along each differentiation process, and the best model was selected by the Akaike's Information Criterion in a Stepwise Algorithm (Hastie, 1992), as implemented in the *stats* package (R Development Core Team, 2009). Since the selected regression models were essentially linear or quadratic (meaning that gene expression variations were constant throughout differentiation or showed only one inflexion point), for further analysis we reduced each differentiation process to three time points, T0, T1 and T2 (Table 2.1). T0 corresponds to the time when cultured cells were switched to differentiation medium or to the first day postpartum for testis. T2 corresponds to terminally differentiated cells or adult testis, and T1 corresponds to an intermediate stage specific to each differentiation process. During myogenesis, the proliferating mononucleate myoblasts withdraw from the cell cycle and subsequently fuse to form multinucleate myotubes; we therefore considered that T1 corresponds to the time when irreversible cell cycle withdrawal occurs, approximately 24h after serum restriction (Tomczak et al., 2004). Likewise, for adipogenesis T1 corresponds to the time when cells withdraw permanently from the cell cycle at approximately 2 days after hormonal stimulation (Akerblad et al., 2005; Burton et al., 2004). In contrast, during erythropoiesis cells undergo three to four rapid cell divisions accompanied by a decrease in cell size and the accumulation of hemoglobin; in this case, we considered that T1 corresponds to the stage of proliferating capacity, which occurs at approximately 15h in GE1 cells and at 30h in fetal liver erythroid progenitors (Welch et al., 2004). Based on the observation that >99% of male germ cell-specific transcripts are first expressed during or after the occurrence of meiosis (Schultz et al., 2003), we considered the onset of sperm cell

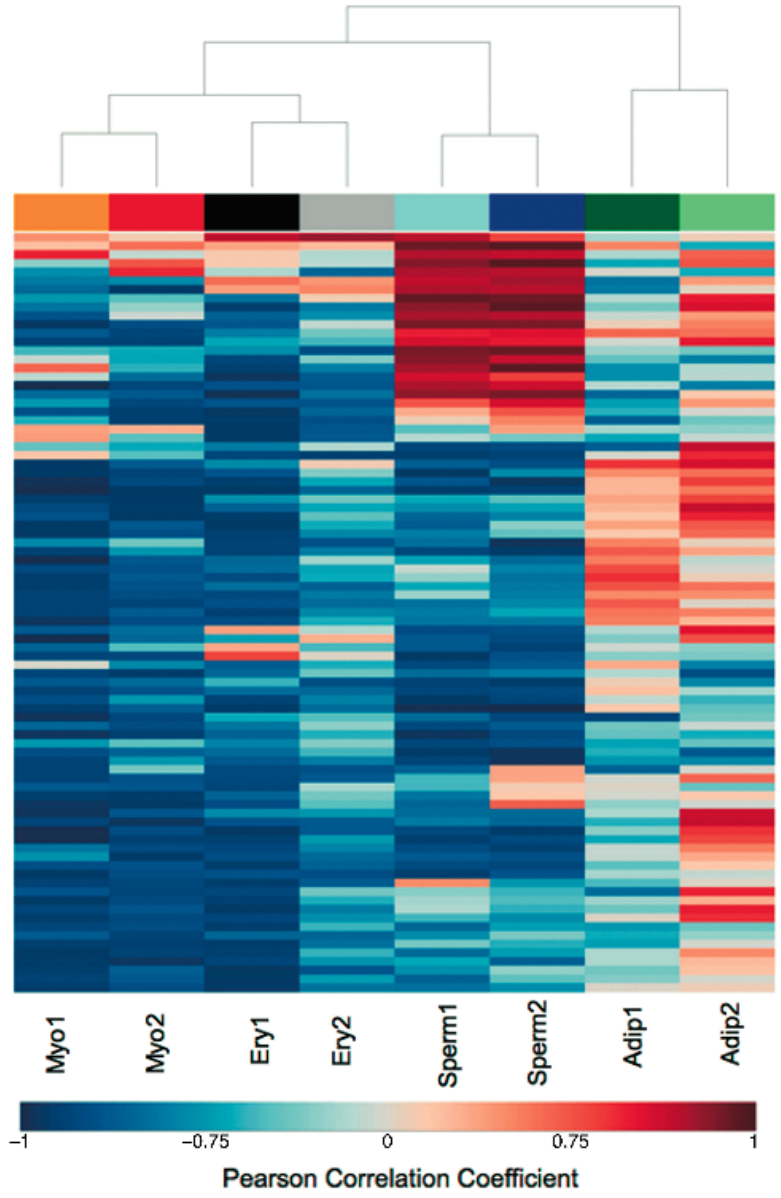


Figure 2.2: **Variation in expression of splicing-related genes during cell differentiation.** Hierarchical clustering display of Pearson correlation values between gene expression and time, for the splicing-related genes with the absolute correlation values higher than 0.75 in both data sets of at least one differentiation process. The negative and positive correlation values are represented by blue and red colours respectively.

meiosis (taking place at approximately 14 days after birth) as T1 for spermatogenesis.

For each differentiation process, the fitted models were used to predict the splicing factor expression levels at time points T0, T1 and T2. Then, to normalize the data, we estimated the fold-changes observed at T1 and T2, relative to T0. We also transformed the residual standard errors from each fitted regression model and used as weights (weights = exponential (- residual standard error)) to include confidence levels of each prediction (biological variability). Finally, the differentially expressed splicing-related genes for each T1 and T2 differentiation stage were selected using linear models and empirical Bayes methods (Smyth, 2004) as implemented in *limma* package (Smyth, 2005). The B-statistics gives the log odds of differential expression and it requires an a prior value for the estimated proportion of differentially expressed genes. To determine this value, we visually inspected the volcano plot, which compares biological significance (represented by fold-changes) with statistical significance (B-values) (Jin et al., 2001), finding the value which enabled genes to be distinguished from the majority (Conboy et al., 2007). Additionally, we verified the p-values corresponding to moderated F-statistics. Using the Benjamini and Hochberg method (Benjamini and Hochberg, 1995), all genes selected as differentially expressed had adjusted p-values lower than 0.01. To validate our approach we included in the analysis the specific differentiation marker genes used for synchronization. Up-regulation of each differentiation marker gene was specifically detected in the respective differentiation process (see Figure 2.3).

Our analysis revealed that major variations in splicing factor expression occurred at T2. The highest variation was found in spermatogenesis: 47% of total splicing-related genes were up- or down-regulated at T2 relative to T0. The genes that were statistically selected as up- or down-regulated in the different processes included members of the hnRNP and SR protein families, SR protein kinases, DEAD-box RNA helicases, snRNP proteins and several additional spliceosomal proteins (Annex Table A.1.4).

In order to validate the microarray data analysis we determined mRNA expression levels using a more sensitive method. RNA samples were obtained from C2 myoblasts and fetal liver erythroid progenitors and analyzed by quantitative real time PCR (qRT-PCR). We started by selecting 12 genes that the microarray data analysis identified as up or down-regulated during erythroid and myotube differentiation. As shown in Figure 2.4 (closed circles), expression changes were confirmed for 9 genes (75%). Thus, we obtained a validation rate of 75% among independent biological samples for genes identified as differentially expressed in our statistical analysis of microarray fold-changes. We then selected 15 other genes that were not identified as differentially expressed during myogenesis or erythropoiesis. From these, we found 4 genes down-regulated in myogenesis (Rod1, Hnrpa1, Sfrs10, Hnrpa2b1) and 10 genes down-regulated in erythropoiesis (Cugbp1, Cugbp2,

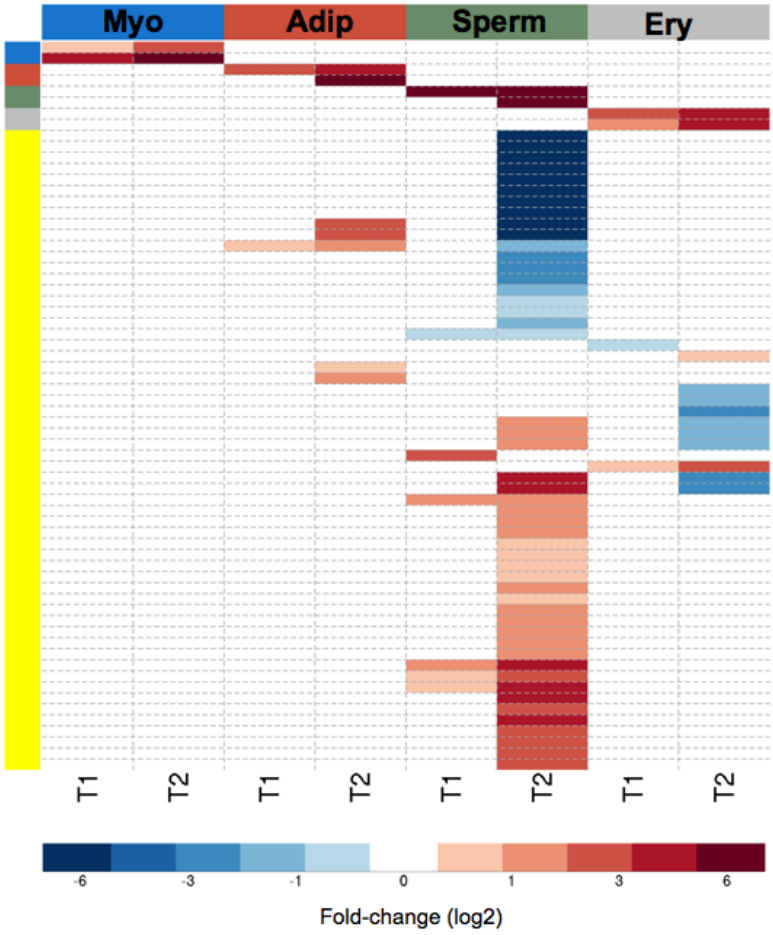


Figure 2.3: **Splicing-related gene expression signatures during cell differentiation.** Heatmap with the fold-changes ($\log_2$) observed for each gene that is most highly differentially expressed during myotube (Myo), adipocyte (Adip), sperm cell (Sperm) and erythrocyte (Ery) differentiation. The genes and respective fold-changes are presented in detail in Annex Table A.1.5. The side colours represent the splicing-related genes (yellow) and the specific differentiation marker genes for myogenesis (Ryr1, Tnnc1 in blue), adipogenesis (Pparg and Cfd in red), spermatogenesis (Ldhc, Pgk2 in green) and erythropoiesis (Gypa and Slc4a1 in grey).

Ddx17, Snrpb2, U2af1, Sfrs2, Ptbp1, Hnrpdl, Hnrpr and Wtap) (open circles in Figure 2.4). This reveals that the microarray analysis is missing several genes the expression of which is less obviously altered.

We next asked whether robust differences could be found that distinguish one differentiation process from the others. To identify genes that are most highly differentially expressed in a particular differentiation process we used linear models and empirical Bayes methods (Smyth, 2004) as described previously. Following the statistical analysis, a filter was applied to eliminate genes that were similarly differentially expressed in more than one differentiation process. A gene is considered to be part of a signature when its expression changes at least 1.5 fold ($\log_2 = 0.58$) more than in any other process. As shown in Figure 2.3 and Annex Table A.1.5, we identified gene expression signatures associated with 3 of the 4 differentiation processes. The list of genes in each signature included members of the several splicing-related protein families. The gene expression signature associated with spermatogenesis contained the highest number of genes. The signature associated with erythroid differentiation consisted of two genes (U2af1-rs1 and Prpf6), and the adipogenesis signature comprised three genes (Hnrpab, Hnrpdl, Sfrs1). No signature was associated with myotube differentiation, as the genes that were differentially expressed during myogenesis were also found differentially expressed in at least one of the other processes analyzed. This may be related to the finding that splicing factors in muscle are predominantly regulated at the post-transcriptional level (Kuyumcu-Martinez et al., 2007).

2.3.3 Tissue-specific differences in splicing factor expression

Having identified splicing factor signatures associated with cell differentiation, we next explored variations in splicing factor expression across tissues from human, chimpanzee and mouse. Available mRNA expression data was obtained from a microarray study covering five different tissues in six humans and five chimpanzees using a total of 48 hybridizations (Khaitovich et al., 2005). This study used the Affymetrix Human Genome hgu133plus2 platform containing 738 probe sets for 208 human splicing-related genes (Annex Table A.1.3). Gene expression profiles from adult mouse were obtained from a study that analyzed 24 brain regions and 10 body tissues using a total of 150 array hybridization measurements with the Affymetrix Murine mgu74av2 platform (Zapala et al., 2005) (Annex Table A.1.6).

To compare the splicing-related gene expression profiles from human, chimpanzee and mouse datasets, a linear model (Smyth, 2004) was fitted for each gene using the expression values from all microarrays and with one regression coefficient for each tissue. Thus, each regression coefficient from the model represents the expression level of the gene in a

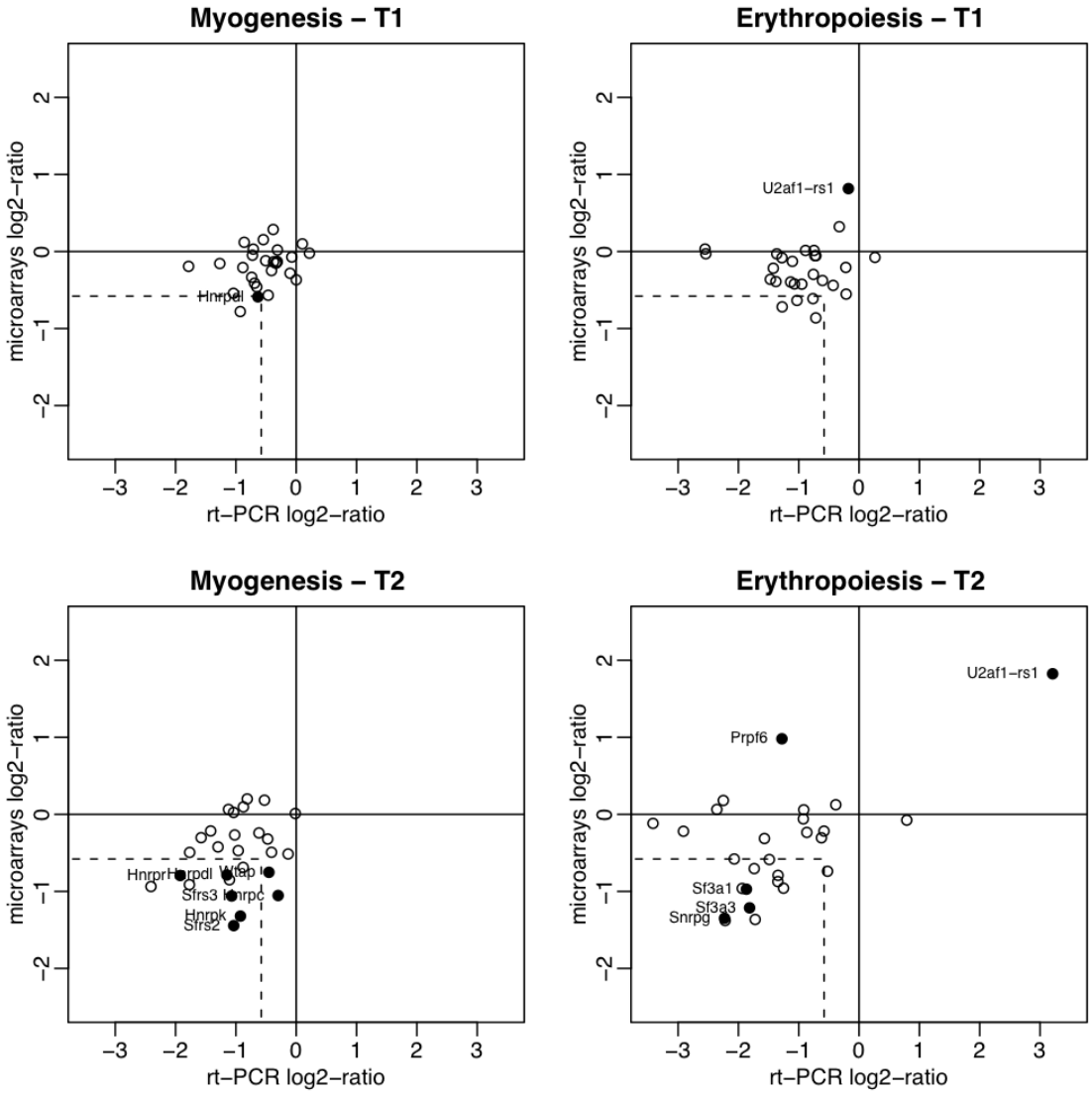


Figure 2.4: **Validation of microarray data analysis for myogenesis and erythropoiesis by quantitative real-time PCR.** The fold-changes in expression of 27 splicing factors at T1 and T2 relative to T0 are indicated. For qRT-PCR analysis, RNA samples were obtained from C2 myoblasts and fetal liver erythroid progenitors. Results are presented as means for at least three independent experiments. Results from microarray data sets are presented as the fold-changes estimated from the linear models. The dashed lines indicate the 1.5 fold-changes values (in logarithm scale) for microarray data and qRT-PCR. The differentially expressed genes selected by microarray data analysis for each differentiation stage are indicated with solid circles.

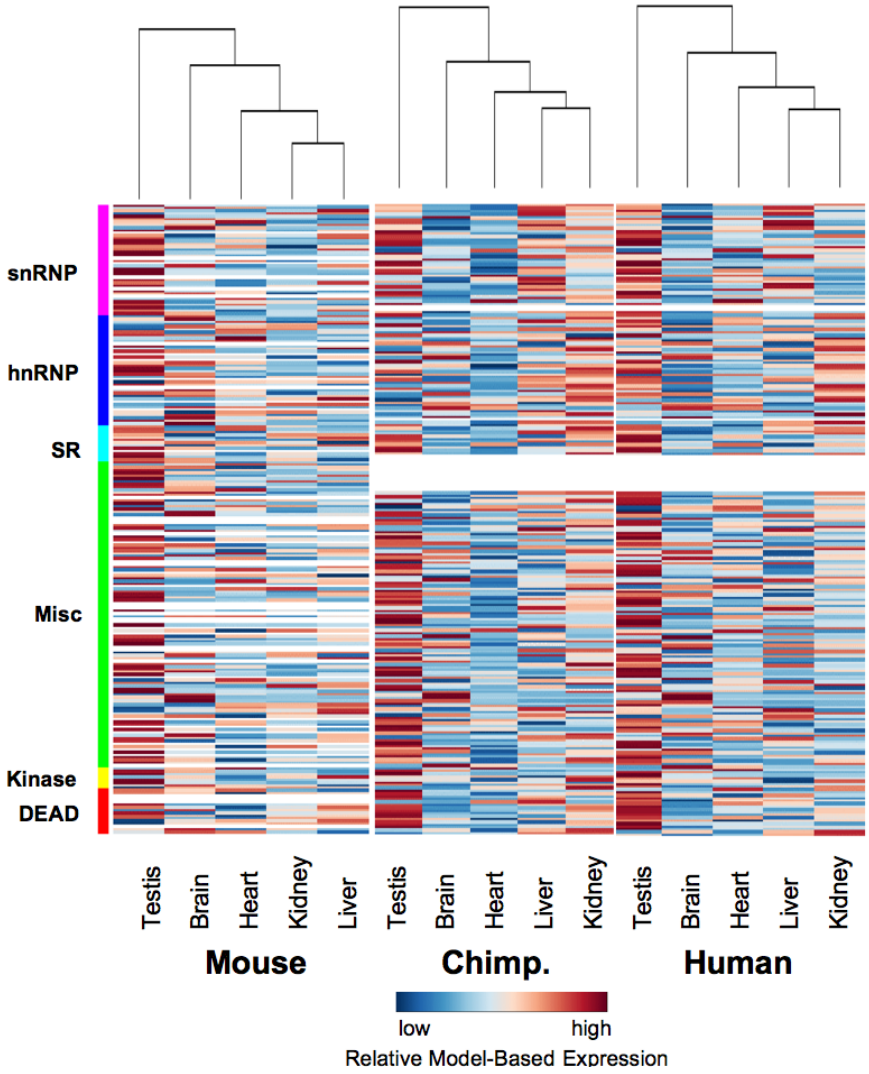


Figure 2.5: **Tissue expression profiles of splicing-related genes are similar in human, chimpanzee and mouse.** Heatmap of adult mouse, chimpanzee and human tissues using microarray-derived expression profiles of splicing-related genes. The expression value for each gene is normalized across the samples to zero mean and one standard deviation for visualization purposes. Genes with expression levels greater than the mean are colored in red and those below the mean are colored in blue. The expression values for genes that are not present in one of the microarray platforms are represented by white.

different tissue. The tissues relatedness was studied performing a hierarchical clustering analysis of the tissues expression profiles using only the splicing related genes and the non-splicing related genes. We estimated the Euclidean distance among the tissues and used hierarchical clustering with different agglomeration methods (complete, single, average, centroid and Ward) as implemented in *stats* package (R Development Core Team, 2009). The best hierarchical tree was chosen using the cophenetic correlation value. The results revealed very similar expression profiles of splicing-related genes in human and chimpanzee tissues (Figure 2.5).

For these two organisms, the testis was clearly an outlier, with low concordance in expression of splicing-related genes relative to the other tissues examined. Analysis of mouse tissues also indicated the testis as the main out-group (Figure 2.5). Most of the 24 mouse brain regions revealed high similarity in expression profiles and were mostly grouped together for both splicing-related genes and all remaining genes (Figure 2.6). Pituitary and retina appeared as an out-group of the brain cluster, and corpus plexus of the fourth ventricle (Cp4v) did not group with the remaining brain regions but rather clustered with the body tissues. Hierarchical clustering of splicing-related gene expression profiles in the 10 body tissues revealed the testis, spleen and thymus as the main out-group (Figure 2.6).

From the human, chimpanzee and mouse microarray data, we identified 154 genes that were differentially expressed between brain, testis, heart, liver and kidney (Annex Table A.1.7). From these, 7 genes were selected and 5 of them (71%) were found differentially expressed by qRT-PCR (Figure 2.7).

Similarly to the results observed during cell differentiation, the differentially expressed genes code for hnRNP and SR proteins, SR protein kinases, DEAD-box RNA helicases, snRNP proteins, and several other splicing-related proteins. From the selected 154 genes, 104 showed tissue-specific expression variation higher than 1.5-fold in at least one of the three organisms (Figure 2.8 and Annex Table A.1.8). Analysis of all mouse data sets further revealed 74 genes with highest expression variation in the 24 brain regions and 10 body tissues (Annex Table A.1.9).

As shown in Figure 2.8, testis and brain contain the highest number of splicing-related genes that are more than 1.5-fold differentially expressed. From the human and chimpanzee microarray data sets, we identified 43 genes included in the testis-specific signature and 20 in the brain signature. From the mouse studies our results reveal 49 genes in the testis signature and 6 in the brain signature. Out of the 48 genes included in the signature for spermatogenesis (Annex Table A.1.5), 27 appeared also in the adult mouse testis signature (Annex Table A.1.8 and Figure 2.9).

Concerning the brain-specific splicing factor gene expression signature, the gene list includes the previously reported brain splicing regulators PTB1, NOVA1, A2bp1/FOX1,

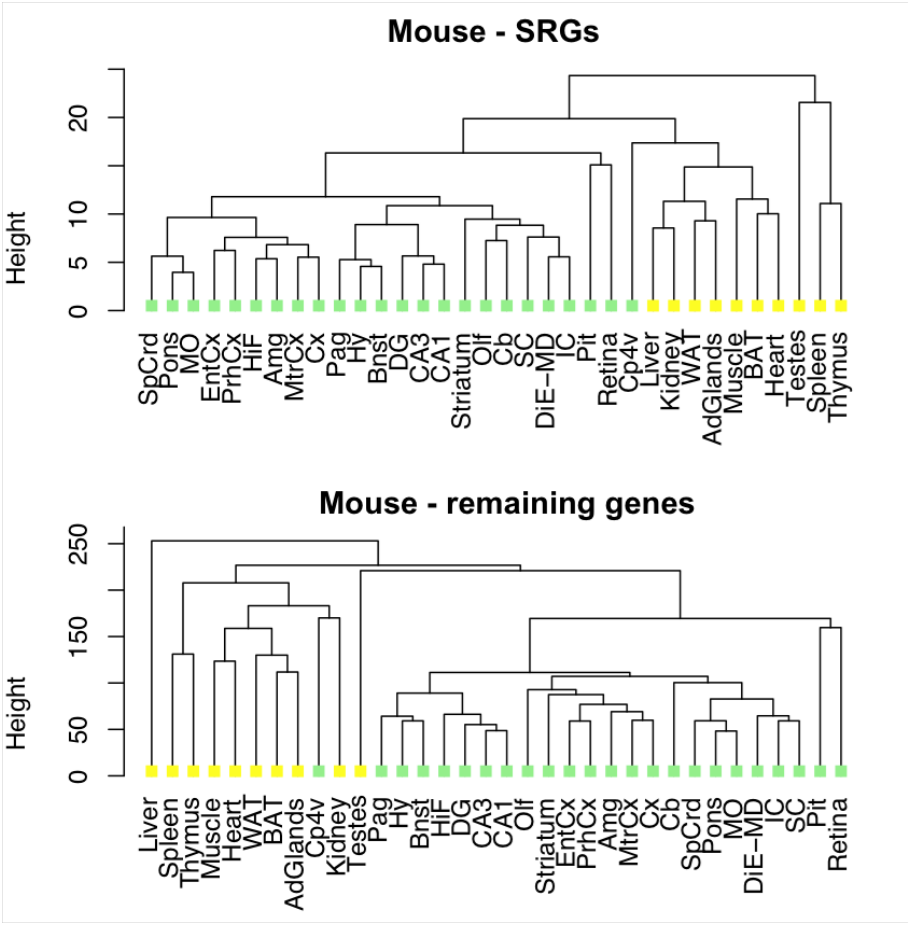


Figure 2.6: **Hierarchical clustering of 24 mouse brain regions and 10 body tissues using microarray-derived expression profiles for splicing-related genes (SRGs) and remaining genes.** The brain and body tissues are highlighted with green and yellow squares, respectively. The brain regions are identified as follows: amygdala (Amg), bed nucleus of the stria terminalis (Bnst), CA1 region of the hippocampus, CA3 region of the hippocampus, cerebellum (Cb), choroid plexus from the fourth ventricle (cp4v), cortex (Cx), dentate gyrus (DG), diencephalon and midbrain excluding hypothalamus (Hy) (DiE-MD), entorhinal cortex (EntCx), hippocampal formation (HiF), Hy, inferior colliculus (IC), medulla oblongata (MO), motor cortex (MtrCx), olfactor y bulbs (Olf), periaqueductal gray (Pag), perirhinal cortex (PrhCx), pituitary (Pit), pons, retina, spinal cord (SpCrd), striatum, and superior colliculus (SC). The height represents the value of the criterion associated with the clustering method for the particular agglomeration, i.e., the distance between each tissue.

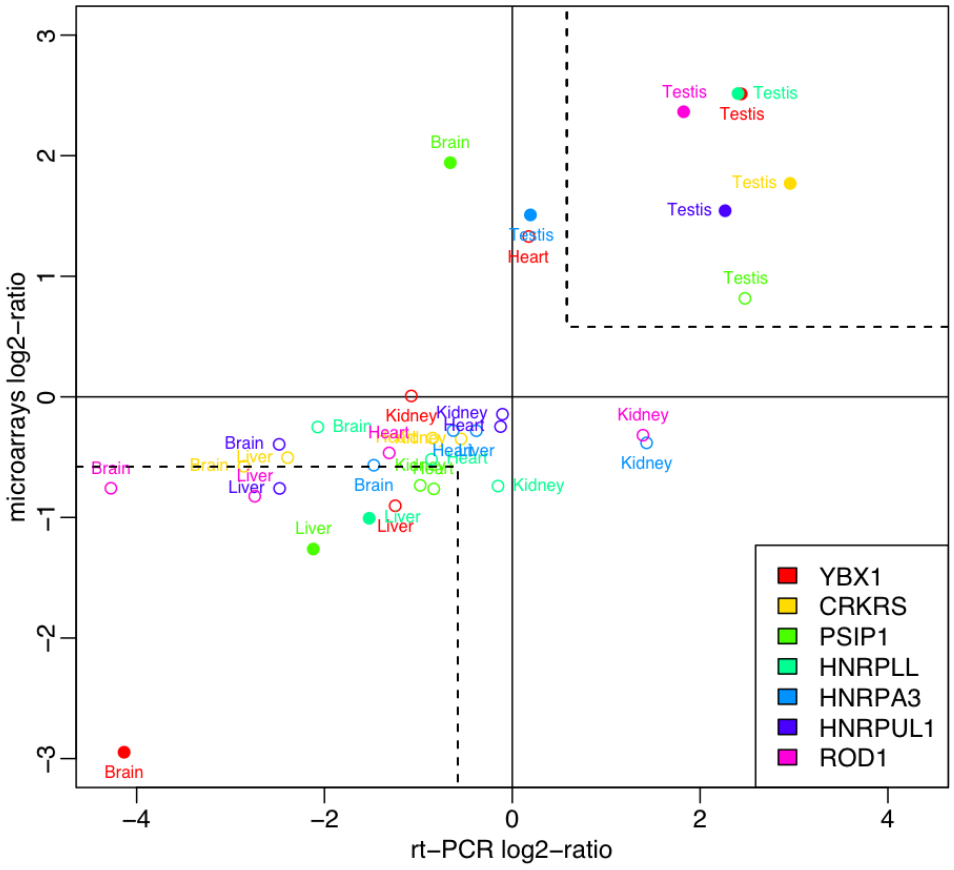


Figure 2.7: **Validation of microarray data analysis for human tissues by quantitative real-time PCR.** The fold-changes in expression of 7 splicing factors between five different tissues: brain, heart, kidney, liver and testes. qRT-PCR fold-changes were obtained from the ratios between each tissue and the average of remaining tissues. Results from microarray data sets are presented as the fold-changes estimated from the linear models. The dashed lines indicate the 1.5 fold-changes values (in logarithm scale) for microarray data and qRT-PCR. The differentially expressed genes selected by microarray data analysis for each tissue are indicated with solid circles.

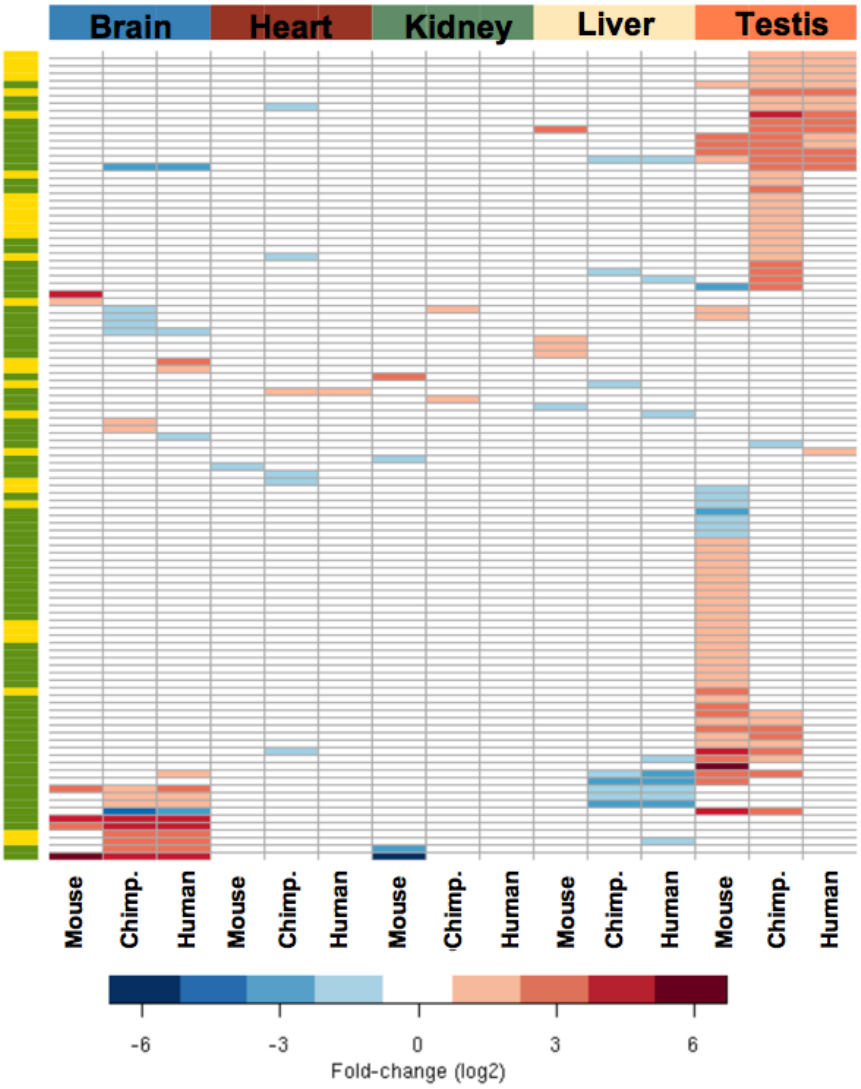


Figure 2.8: **Tissue-specific splicing-related gene expression signatures.** Heatmap indicating the fold-changes ($\log_2$) observed for each gene that is most highly differentially expressed in the five tissues examined. The left bar highlights genes that are present in both human and mouse Affymetrix platforms (green) or only in one of the two platforms (yellow). The genes and respective fold-changes are presented in detail in Annex Table A.1.8.

and members of the CELF/BRUNOL and ELAVL families. Additionally, we identified the non-SR splicing regulator Y-box protein 1 (Stickeler et al, 2001) highly down-regulated and the core snRNP protein SmN (Grimaldi et al, 1993) highly up-regulated. We detected many genes that were highly differentially expressed in chimpanzee but not in human brain, and we found two genes (TNRC4, encoding the CELF3/BRUNOL1 protein, and LSM8, encoding the U6 snRNA-associated Sm-like protein LSm8) that were, respectively, highly up- and down-regulated in human but not in chimpanzee brain.

The testis-specific signature included the splicing factor 3a subunit 2 (SF3A2) and the SR protein kinases 1 and 2 (SRPK1 and SRPK2). The genes that were common to the testis-specific signatures from all three organisms (human, chimpanzee and mouse) encode SF3A2, SRPK2, protein phosphatase 1G (PPM1G), the RNA binding protein RDBP and the heterogeneous nuclear ribonucleoprotein HNRPLL. Remarkably, 10 (37%) of the mouse genes included in the testis-signature corresponded to up-regulated snRNPs (Lsm2, Lsm4, Sf3a3, Snrpa, Snrpa1, Snrpc, Snrpd2, Snrpg, Usp39 and U5-40d).

2.4 Discussion

In this study we applied computational methods to identify tissue-specific splicing factor gene expression signatures from published microarray data sets. By using this approach we have identified over 100 splicing related genes that are most highly differentially expressed in a particular tissue or differentiation process.

Recently, several microarray-based methods have been reported for genome-wide monitoring of splicing events in mammalian tissues (Blencowe, 2006; Wang and Cooper, 2007). The increasing availability of splicing-microarray data sets will make it possible to extend our approach and systematically search for differential expression of alternatively spliced isoforms of splicing regulators. Importantly, however, changes in splicing factor mRNA levels may not necessarily reflect on protein expression due to post-transcriptional regulation (Boutz et al., 2007; Makeyev et al., 2007). Therefore, further experimental investigation on the candidate tissue-specific splicing regulators identified in this study is required to determine whether specific changes in the protein concentration and/or activity do occur.

2.4.1 Splicing factor signatures correlate with tissue-specific alternative splicing patterns

By using a method that normalizes the number of observed alternative splicing events to the EST coverage in each tissue, Yeo and colleagues found that the brain has the highest proportion (>40%) of alternatively spliced genes, followed by the liver and testis (Yeo et al., 2004). The brain and testis showed the highest levels of exon skipping, while

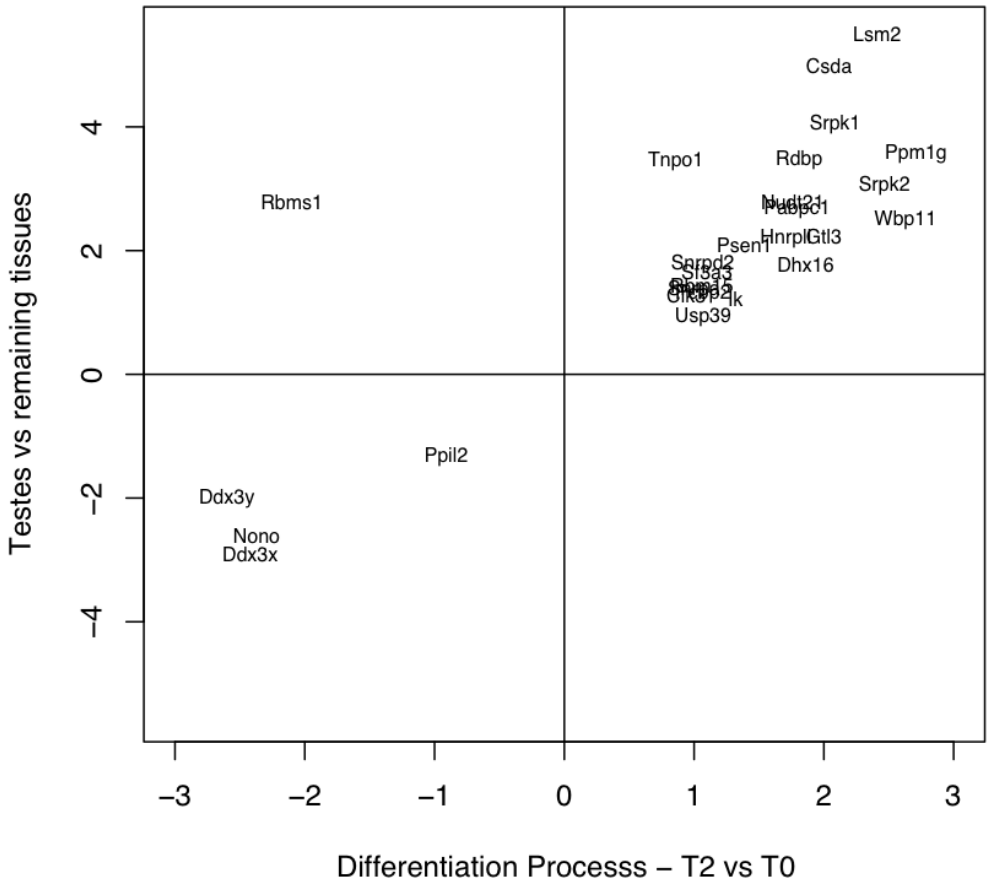


Figure 2.9: **Comparison of expression fold-changes ($\log_2$) observed for 48 mouse splicing factors during testis differentiation (T2 vs T0) and in adult tissue (testis vs remaining tissues).** The contradictory fold-changes observed for Rbms1 were obtained from two Affymetrix probes sets that hybridize in different regions of the gene (alternative exon and 3′UTR).

the liver had the highest frequency of alternative $3'$ and $5'$ splice site usage. Using a microarray platform with probes that span exon-exon junctions, Pan et al (2004) detected the largest number of tissue-dependent alternative splicing events associated with brain. A more recent analysis performed with human exon microarrays revealed that testis and brain express the largest number of probesets that are not expressed in any other tissue (Clark et al., 2007). In that study, tissue-specific probesets may be from genes that are only expressed in a single tissue, or individual exons that are included in a tissue-specific manner via alternative splicing (Clark et al., 2007).

Our analysis revealed that the highest number of highly differentially expressed splicing-related genes occurred in the testis and in the brain, whereas the liver showed higher concordance in expression of splicing-related genes relative to other tissues, namely the kidney. Thus, our results specifically distinguish the two tissues with highest abundance of alternatively spliced mRNA isoforms that differ by inclusion or exclusion of an exon, as those with a highest variation in splicing factor expression. Yeo and coauthors (2004) have also analyzed microarray expression data for 20 splicing factors of the SR, SR-related and hnRNP protein families across several human tissues and identified liver as an outlier, suggesting an involvement of this group of factors in regulation of liver-specific alternative $3'$ and $5'$ splice decisions. However, our analysis revealed that variation in expression levels of these factors is not unique to the liver.

2.4.2 SR protein kinases as tissue-specific signatures

According to a current model, small differences in concentration or activity of SR proteins may influence the choice of competing splice sites and therefore control alternative splicing (Shin and Manley, 2004). SR proteins form multiprotein complexes that bind to splicing enhancer sequences in the pre-mRNA and stabilize the assembly of the spliceosome at splice sites. One possible mechanism to affect SR protein activity is differential phosphorylation. Indeed, the phosphorylation status of Ser residues within the RS domain of SR proteins has been shown to alter protein-protein interactions and splicing activity (Prasad et al., 1999; Prasad and Manley, 2003; Xiao and Manley, 1997). Several SR-protein kinases have been identified, including SRPK and CLK/STY (Colwill et al., 1996; Gui et al., 1994). Here, we detected members of both the SRPK and CLK gene families being differently expressed in distinct cell types and tissues. In particular, the SRPK1 and SRPK2 genes were highly up-regulated during mouse spermatogenesis. Moreover, SRPK1 and SRPK2 were included in the testis-specific signature for chimpanzee and mouse (SRPK2 also found for human), whereas SRPK3 was included in the heart signature for human and chimpanzee. We therefore predict that SR protein kinases are likely to play an important role in tissue-specific alternative splicing.

2.4.3 Tissue-specific signatures include several snRNP proteins

It is generally assumed that splicing is regulated by non-snRNP proteins that modulate the association of core components of the spliceosome with the pre-mRNA. This view was for the first time questioned by an RNAi screen in Drosophila cells that unexpectedly detected changes in alternative splicing of endogenous genes after reducing the levels of core spliceosomal proteins (Park et al., 2004). These included components of the U1, U2 and U4/U6 snRNPs, and both subunits of the U2 snRNP auxiliary factor, U2AF. More recently, we used RNAi to down-regulate expression of the small subunit of U2AF in human cells and we also observed changes in alternative splicing of transcripts derived from both endogenous genes and exogenous reporter minigenes (Pacheco et al., 2006b, a). In another study, Massiello and coauthors (2006) reported that RNAi-mediated down-regulation of SAP155 (a subunit of splicing factor SF3B, which associates with the U2 snRNP) affected alternative splicing of Bcl-x transcripts. Although some of the effects on alternative splicing induced by RNAi may be indirect, it was also shown that in Saccharomyces cerevisiae substrate selectivity can be modulated by altering the kinetics of spliceosome rearrangement (Query and Konarska, 2004). Further support for the idea that fluctuations in the concentration of core spliceosomal proteins may contribute to regulate splicing is provided by the differential cell type and tissue-specific expression profiles presented in this study. Variations in expression of genes that code for Lsm, Sm and snRNP-specific proteins were detected in the course of myotube, erythroid and sperm cell differentiation. Consistent with our results, down-regulation of snRNP synthesis during myogenesis was previously demonstrated by pulse-labeling experiments (Gabanella et al., 2005). A decrease in expression of genes that encode snRNP proteins was not observed during adipogenesis, arguing that the variations detected in myogenesis are not related to the cell cycle arrest, which is common to both myotube and adipocyte differentiation. In addition to core snRNP proteins, the U2af1-rs1 gene, which encodes a protein with a high degree of homology to the small subunit of U2AF (Mollet et al., 2006), was found specifically up-regulated during erythroid differentiation. Another U2AF-related gene, U2af1-rs2, was highly up-regulated in the mouse brain. SF3A2 was further identified as part of the testis-signature for human, chimpanzee mouse, while the snRNP protein SmN appeared in the brain-signature for the three organisms. Clearly, a major task for the future will be to determine whether tissue-specific alternative splicing events are regulated by the differential expression of these snRNP and snSNP-related proteins.

Chapter 3

Cancer-specific misregulation of splicing factor gene expression

Part of the introduction and discussion of this chapter is written as a review article in: Grosso AR, Martins S, Carmo-Fonseca M. (2008) The emerging role of splicing factors in cancer, EMBO Rep, 2008 Nov;9(11):1087-93. I would like to stress that Sandra Martins and I contributed equally to this article.

The original work described in this chapter is in preparation for submission to a peer reviewed journal.

Keywords: splicing of pre-mRNA; splicing factor; spliceosome; cancer

Abstract: Recent progress in global sequence and microarray data analysis has revealed the increasing complexity of the human transcriptome. Alternative splicing generates a huge diversity of transcript variants and disruption of splicing regulatory networks is emerging as a major contributor to various diseases, including cancer. Current efforts to establish the dynamic repertoire of transcripts that are generated in health and disease are showing that many cancer-associated alternative splicing events occur in the absence of mutations in the affected genes. Rather, a growing body of evidence reveals changes in splicing factor expression that correlate with cancer development, progression and response to therapy. To explore this idea, we performed a large-scale analysis of expression profiles for several splicing factors. Cancer-specic alterations in gene expression were found for the major splicing protein families: snRNPs, hnRNPs, SRs, SR-kinases and splicing regulators. The majority of differentially expressed splicing regulators were up-regulated in cancer and some misregulations appear consistently in several cancer types.

3.1 Introduction

Removal of noncoding sequences (introns) from precursor messenger RNAs (pre-mRNAs) through splicing provides a versatile means of genetic regulation. Alternative splicing allows a single gene to generate multiple transcripts, thus expanding the transcriptome and proteome diversity in metazoans. Several studies based on large-scale expressed sequence tag (EST) analysis estimated that >60% of human genes undergo alternative splicing, and this number more recently increased to >80% when microarray data became available (Black, 2003; Matlin et al., 2005).

Intron excision is carried out by an assembly of small nuclear ribonucleoprotein particles (snRNPs) and extrinsic, non-snRNP protein splicing factors that are collectively recruited to pre-mRNAs and form the spliceosome. The initial events of spliceosome assembly require the recognition of specific sequences located at and near the $5'$ and $3'$ splice sites, which recruit the U1 and U2 snRNPs. In metazoan organisms, the splice site sequences are weakly conserved and require specific additional RNA sequence elements that function to either enhance or repress the ability of the spliceosome to recognize and select nearby splice sites (Maniatis and Tasic, 2002; Matlin et al., 2005). The multiplicity of protein-protein and protein-RNA interactions that modulate the association of the spliceosome with the pre-mRNA constitutes the basis to control alternative splicing.

A typical multiexon pre-mRNA can undergo a number of alternative splicing patterns (Black, 2003). Most exons are constitutive, meaning that they are always included in the final mRNA, but there are also regulated exons, which are sometimes included and sometimes excluded from the mRNA. Exons can also be lengthened or shortened by altering the position of one of their splice sites, or by a distinct splicing pattern that consists in failure to remove an intron, a process known as intron retention. Alternative splicing can also be coupled to differential promoter or polyadenylation site usage, giving rise to an even larger transcriptome heterogeneity.

Splicing abnormalities play an important role in human diseases such as cancer (Wang and Cooper, 2007). Several mutations are known that affect the splicing of oncogenes, tumour suppressors and other cancer-relevant genes (Srebrow and Kornblihtt, 2006; Venables, 2006), however, many splicing abnormalities identified in cancer cells are not associated with mutations in the affected genes. Rather, a growing body of evidence indicates that the splicing machinery is a major target for misregulation in cancer. According to recent bioinformatics studies, changes in splicing factor expression may play a key role in the general splicing disruption that occurs in many cancers (Kim et al., 2008; Ritchie et al., 2008).

To systematically investigate the splicing factor expression variations in cancer, we performed a large-scale analysis of expression profiles in cancer for several genes encoding

Table 3.1: **Microarray data sets used to study several cancer types.** The microarray datasets were grouped according to general cancer tissue and cancer type. GEO ("GSExxxxx") and ArrayExpress ("E-xxxxx") accession numbers and reference to published work (if available) are provided.

General Cancer Tissue	Cancer Type	Database Acession Number	References
Bladder		GSE3167	Dyrskjot et al. (2004)
Brain	Glioblastoma	E-MEXP-567	Margareto et al. (2007)
	Astrocytoma	E-MEXP-567	Margareto et al. (2007)
Breast	Invasive breast carcinoma	GSE3744	Richardson et al. (2006)
Colon		GSE4107	Hong et al. (2007)
Esophagus	Barrett's-associated adenocarcinomas	GSE1420	Kimchi et al. (2005)
Head and neck	Head and neck squamous cell carcinoma	GSE6631	Kuriakose et al. (2004)
Kidney	Renal cell carcinoma	GSE6344	Gumz et al. (2007)
Liver	Hepatocellular carcinomas	E-TABM-36	Boyault et al. (2007)
Lung	Squamous cell lung cancer	GSE3268	Wachi et al. (2005)
	Lung Adenocarcinoma	E-MEXP-231	Yap et al. (2005)
Neuroblastoma		E-MEXP-669	De Preter et al. (2006)
Prostate		GSE3325	Varambally et al. (2005)
Thyroid	Papillary thyroid cancer	GSE3678	unpublished
Vulva	Vulvar intraepithelial neoplasia	GSE5563	Santegoets et al. (2007)

proteins shown to be involved in the splicing process.

3.2 Material and Methods

3.2.1 Data selection

Microarray data sets for assessment of splicing factors expression were selected from previous studies for several cancer types (Table 3.1). We selected microarray data sets for which the hybridizations were done using Affymetrix GeneChip 3′ Expression Arrays (platforms hgu95a, hgu95av2, hgu133a, hgu133b and hgu133plus2) and the biological samples were from biopsy samples (no cell lines) for the cancer and corresponding normal tissue. A list containing genes that encode known splicing factors, spliceosome-associated proteins, and proteins with a domain structure similar to bona fide splicing factors was obtained from previous studies (Barbosa-Morais et al., 2006; Chen et al., 2007).

3.2.2 Microarray data analysis

All the microarray data analysis was done using R and several packages available from CRAN (R Development Core Team, 2009) and Bioconductor (Gentleman et al., 2004). The raw data (CEL files) for Affymetrix GeneChip 3′ Expression Arrays was normalized

and summarized with the Robust MultiArray Average method from the *affy* package (Gautier et al., 2004). An initial quality assessment was done to remove microarrays with poor quality, using quality diagnostics with probe level models and array quality control metrics for all arrays (average background was < 200, scale factors < 6, percentage of present calls, RNA degradation for GAPDH and beta-actin - $3'/5'$ ratio).

The differentially expressed splicing-related genes for each cancer type were selected using linear models and empirical Bayes methods (Smyth, 2004) as implemented in *limma* package (Smyth, 2005). The B-statistics gives the log odds of differential expression and it requires an a priori value for the estimated proportion of differentially expressed genes. To determine this value, we visually inspected the volcano plot, which compares biological significance (represented by fold-changes) with statistical significance (B-values) (Jin et al., 2001), finding the value which enabled genes to be distinguished from the majority (Conboy et al., 2007). Additionally, we verified the p-values corresponding to moderated F-statistics, adjusted using the Benjamini and Hochberg method (Benjamini and Hochberg, 1995).

3.2.3 Functional analysis

The statistically overrepresented Gene Ontology (GO) terms within up-regulated genes was assessed using GOstat (Beissbarth and Speed, 2004). The program determines which GO terms are enriched in an input gene list relative to the reference genes (all the genes present in the microarray platform) using Fisher's Exact Test (Fisher, 1935).

3.3 Results

3.3.1 Up-regulation of splicing factors in cancer

To study cancer-associated changes in splicing factor expression and other splicing-related genes we analysed 14 microarray expression data sets previously published and containing cancer samples and corresponding normal tissue (Table 3.1). The cancer data sets were classified according to general cancer tissue (bladder, brain, breast, colon, esophagus, head and neck, kidney, lung, neuroblastoma, prostate, thyroid and vulva cancers) and also the cancer type for cases previously shown to present different histological or gene expression profiles, namely glioblastoma and astrocytoma for brain cancer (Margareto et al., 2007) and squamous cell and adenocarcinoma for lung cancer (Wistuba and Gazdar, 2006).

A list of 262 splicing-related genes (SRGs) was obtained from previous studies and contained genes that encode known splicing factors, spliceosome-associated proteins, and proteins demonstrated by mass spectrometry or predicted by sequence homology to in-

teract with the pre-mRNA or the spliceosome (Barbosa-Morais et al., 2006; Chen et al., 2007) (Annex Table A.2.1).

In total 192 splicing-related genes were found to be misregulated in cancer (Figure 3.1 and Annex Table A.2.2). The genes with expression variations encoded for the major splicing protein families snRNPs, hnRNPs, SRs, SR-kinases, RNA-helicases-like and other splicing regulators.

Some cancers presented a higher number of misregulated splicing-related genes, namely bladder, brain, kidney, lung and vulva cancers (Table 3.2). Although, these cancers presented also a higher number of differentially expressed genes, the percentage of SRGs relative to all misregulated genes was higher than for the remaining cancers (2.14 - 5.01% compared to 0.30 - 1.89%).

The results revealed that the majority of misregulated splicing regulators were up-regulated in cancer (Table 3.2). Indeed, the proportion of up-regulation was higher for SRGs than when considering all genes up-regulated on each cancer type.

One could suggest that the up-regulation of SRGs was due to an overall increase of the gene expression pathway. We therefore assessed the enrichment of Gene Ontology (GO) terms associated with gene expression (namely terms like transcription, RNA processing, translation, protein maturation and offspring terms) for all up-regulated genes. Using all the overexpressed genes (not only the splicing regulators) we select enriched GO terms (p-value < 0.01, with Benjamini correction for multiple testing) associated with gene expression and offspring terms. An overall concordance was observed between the number of up-regulated splicing regulators and the enrichment of RNA-processing related terms for each cancer type (Table 3.2). Bladder, brain, lung and vulva cancers that contained the higher number of up-regulated splicing regulators also presented an enrichment of GO terms for RNA-splicing. Some of these cancers also presented enriched GO terms related to transcription and translation, however these were not found simultaneously in the same cancer type and they were related to partial steps, namely transcription and translation initiation. Thus, these results suggest that the up-regulation is mostly confined to the RNA-processing machinery.

Although our results reflect variations at RNA level, previous studies have shown cancer-specific increased levels of proteins encoded by transcripts that we identified as up-regulated in cancers compared to the corresponding normal tissue: HNRNPA2B1 in lung (Sueoka et al., 2001), PTB (PTBP1) in glioblastoma (Jin et al., 2003a), ASF/SF2 (SFRS1) in lung (Karni et al., 2007), SRPK1 in breast (Hayes et al., 2006).

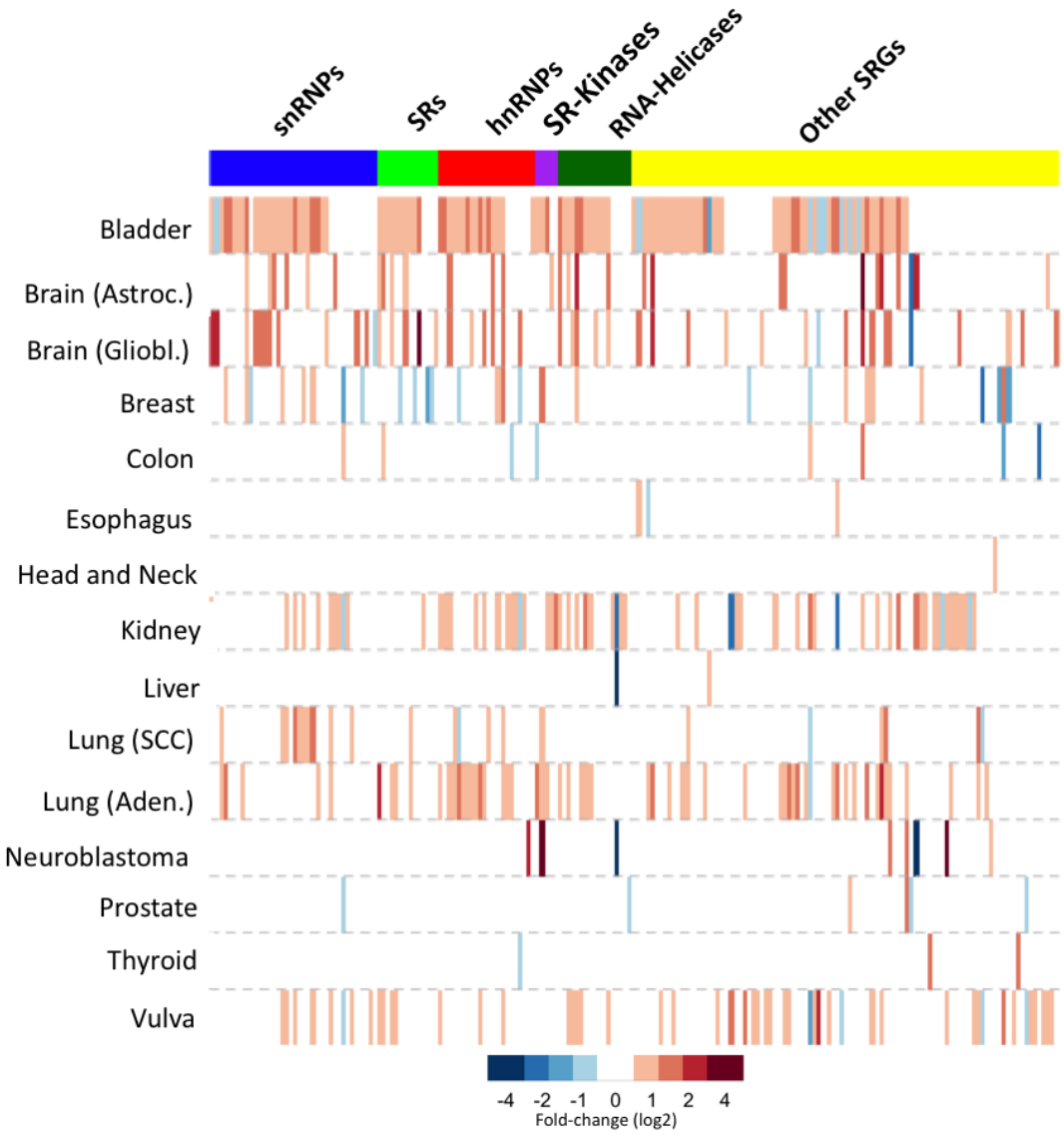


Figure 3.1: **Splicing-related genes misregulated in cancer.** The splicing factors are organized according to the major families: small nuclear ribonucleoproteins that combine with pre-mRNA and various proteins to form spliceosomes (snRNP); heterogeneous nuclear ribonucleoproteins (hnRNP); Serine/Arginine-rich proteins (SR); SR-kinases (Kinases); RNA-helicases and other splicing-related proteins (Other SRGs). Fold-changes values for up and down-regulation values in cancer are represented by red and blue colours according to legend (for more details see Annex Table A.2.2).

Table 3.2: **Genes and splicing-related genes mis-regulated in cancer.** The number of differentially expressed (DE) and up-regulated (Up-reg.) genes and splicing-related genes (SRGs) is presented. The percentage of up-regulated genes relative to all genes and only splicing related genes is also shown. GO terms enriched in up-regulated genes were obtained with GOstat (Beissbarth and Speed, 2004) and only the terms associated with gene expression (p-value < 0.01, with Benjamini correction for multiple testing) are shown (the percentage value indicates the proportion of up-regulated genes relative to all the genes present in the microarray for the enriched GO term).

Cancer Type	**Nr. of DE**		**Nr. of Up-reg.**		**Perc. of Up-reg.**		**GO terms related to gene expression for up-reg. genes**		
	Genes	**SRGS (%)**	**Genes**	**SRGs (%)**	**Genes**	**SRGs**	**Transcription**	**Splicing**	**Translation**
Bladder	2374	119 (5.01%)	1152	110 (9.55%)	48.53	92.44		RNA splicing (45.09%); spliceosome assembly (21.94%)	translational initiation (33.93%)
Brain (Gliob.)	1027	29 (2.82%)	569	28 (4.92%)	55.40%	96.55%		RNA splicing (10.92%)	
Brain (Astroc.)	995	45 (4.52%)	544	42 (7.72%)	54.67%	93.33%	transcription (8.24%); regulation of transcription (6.46%)	RNA splicing (17.82%)	
Breast	1654	28 (1.69%)	466	14 (3.00%)	28.17%	50.00%			
Colon	933	9 (0.96%)	475	5 (1.05%)	50.91%	55.56%			
Esophagus	160	3 (1.88%)	131	2 (1.53%)	81.88%	66.67%			
Head and neck	328	1 (0.30%)	165	1 (0.61%)	50.30%	100.00%			
Kidney	3281	62 (1.89%)	2066	55 (2.66%)	62.97%	88.71%			
Liver	186	2 (1.08%)	58	1 (1.72%)	31.18%	50.00%			
Lung (SC)	1026	22 (2.14%)	407	19 (4.67%)	39.67%	86.36%		RNA splicing (8.67%)	
Lung (Aden.)	1795	57 (3.18%)	1023	56 (5.47%)	56.99%	98.25%		RNA splicing (18.5%)	translational initiation (28.57%)
Neuroblastoma	554	8 (1.44%)	132	6 (4.55%)	23.83%	75.00%			
Prostate	471	6 (1.27%)	97	2 (2.06%)	20.59%	33.33%			
Thyroid	782	3 (0.38%)	271	2 (0.74%)	34.65%	66.67%			
Vulva	1932	50 (2.59%)	943	45 (4.77%)	48.81%	90.00%	transcription initiation (24.24%)	RNA splicing (16.75%)	

To evaluate the impact of the cancer-associated variations, we decided to assess the expression of the misregulated splicing-related genes in the original normal tissue. Using a microarray data set containing six of the normal human tissues here analysed (GSE1133, Su et al. (2004)) and applying a computational approach previously described (Grosso et al., 2008), we identified splicing-related genes that are most highly differentially expressed in a particular tissue (Figure 3.2). Comparing the results from both analyses we could detect that some of some up or down-regulated splicing factors in a specific tissue appear misregulated in the corresponding cancer. Namely, the up-regulated genes PTB (PTBP1) and HNRPAB in brain cancer (astrocytoma and glioblastoma) showed previously lower expression in brain comparative to other normal tissues and are up-regulated in brain cancer. In opposition, the A2BP1 (FOX1) presented higher expression in brain and it is down-regulated in brain cancer.

Similarly, HNRPA3 and SRPK1 presented lower expression in normal lung when compared to other normal tissues and they are up-regulated in lung adenocarcinoma. Thus, these splicing-related genes with tissue-specific expression profiles affected are potential candidates for alternative splicing misregulation in respective cancer tissues.

3.3.2 Cancers share common misregulated splicing factors

Our results showed splicing-related genes commonly misregulated in several cancer types (Figure 3.3). Similar up-regulation was found in at least four cancer types for genes encoding hnRNPs (HNRPA2B1, HNRNPAB, SYNCRIP - hnRNP Q), SR (SF2/ASF - SFRS1, TRA2-β1 - SFRS10, SFRS2 - SC35), SR kinases (SRPK1), snRNP associated proteins (PRPF40A, SNRPA1, SNRPB, SNRPD1, SNRPE, SNRPG, LSM5) and other regulators (YBX1, ELAVL1, NONO). Common down-regulation to three cancer types was observed for hnRNP E2 (PCBP2), CUGP2, FOX1 (A2BP1) and SNRPN.

Our results increased the number of cancer-associated misregulations for some splicing factors previously associated to a single or few cancer types: SC35 (SFRS2) up-regulation in ovary cancer (Fischer et al., 2004); TRA2-β1 (SFRS10) up-regulation in breast cancer (Watermann et al., 2006); HNRPA2B1 up-regulation in lung cancer (Sueoka et al., 2001); PTB up-regulation in ovary (He et al., 2004) and glioblastoma (Jin et al., 2003a); YBX1 up-regulation in ovary (Fischer et al., 2004); HuR (ELAVL1) up-regulation in breast (Denkert et al., 2004) and ovary cancers (Izquierdo, 2008); Sam68 (KHDRBS1) up-regulation in prostate cancer (Busà et al., 2007); hnRNP E2 (PCBP2) down-regulation in oral cancer (Roychoudhury et al., 2007); SF1 down-regulation in colon cancer (Shitashige et al., 2007b, a) and RBM5 (LUCA15) down-regulation in lung cancer (Oh et al., 2006).

Recent studies also demonstrated that some cancers could present common misregulated splicing factors. Over-expression of the splicing factor SF2/ASF (SFRS1) at protein

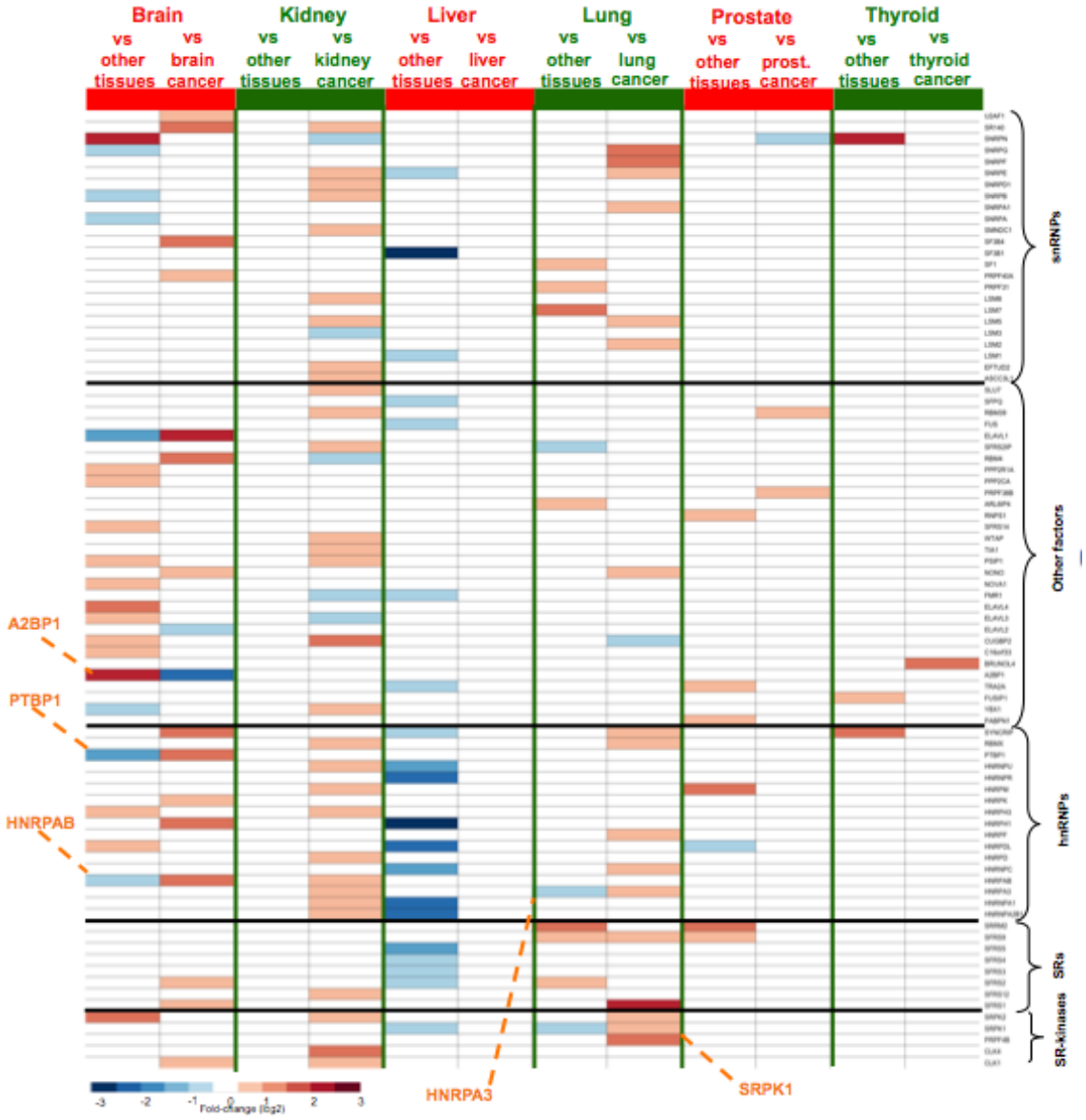


Figure 3.2: **Tissue-specific splicing factors misregulated in cancer.** For each tissue are represented the genes with expression variations (fold-changes) relative to the remaining normal tissues (first column) and misregulated in corresponding cancer (second column). Fold-change values for up and down-regulation values in cancer are represented by red and blue colours according to legend. The splicing factors are organized according to the major families: small nuclear ribonucleoproteins that combine with pre-mRNA and various proteins to form spliceosomes (snRNP); heterogeneous nuclear ribonucleoproteins (hnRNP); Serine/Arginine-rich proteins (SR); SR-kinases (Kinases); RNA-helicases and other splicing-related proteins (Other factors).

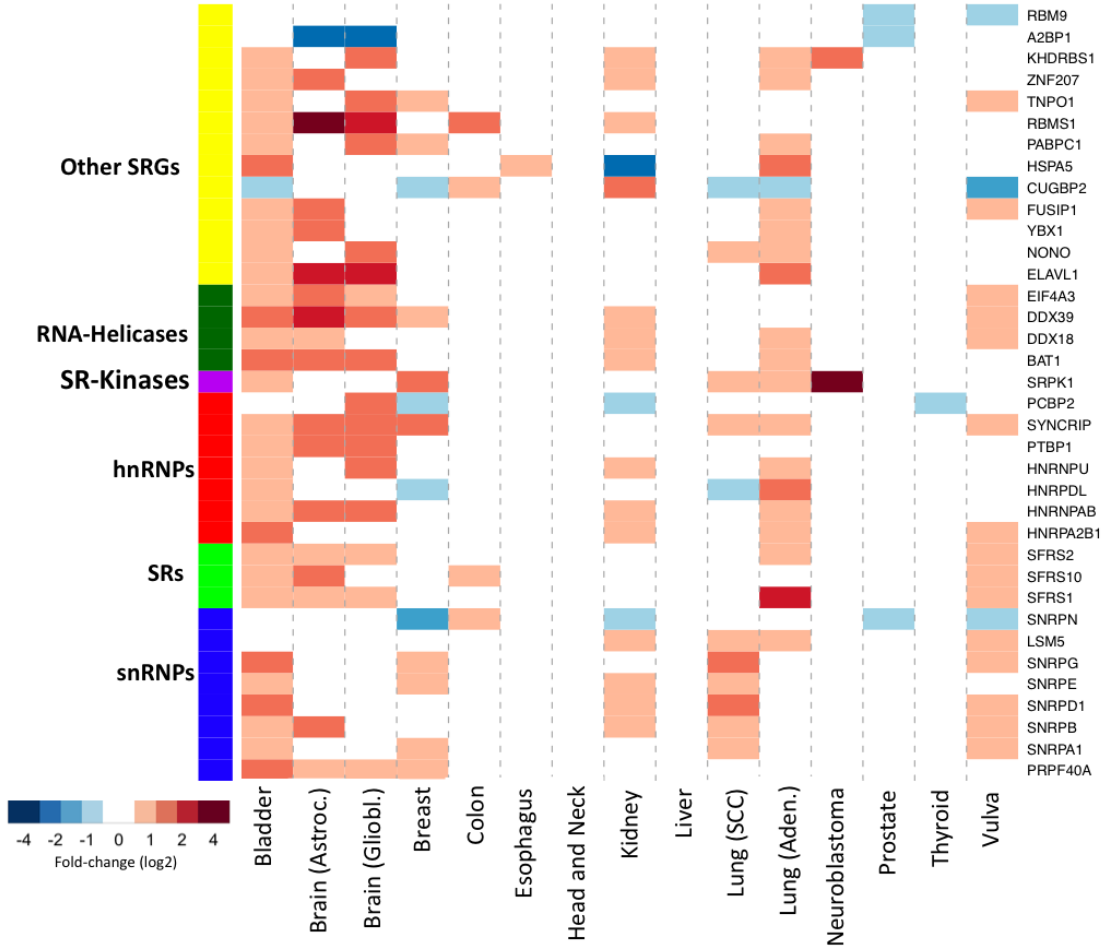


Figure 3.3: **Common misregulated splicing factors in cancer.** The splicing factors are organized according to the major families: small nuclear ribonucleoproteins that combine with pre-mRNA and various proteins to form spliceosomes (snRNP); heterogeneous nuclear ribonucleoproteins (hnRNP); Serine/Arginine-rich proteins (SR); SR-kinases (Kinases); RNA-helicases and other splicing-related proteins (Other SRGs). Fold-change values for up and down-regulation values in cancer are represented by red and blue colours according to legend (for more details see Annex Table A.2.2).

level and splicing modulation of tumour suppressor genes was recently shown for various human tumours: colon, thyroid, small intestine, kidney and lung (Karni et al., 2007). SRPK1 protein overexpression was also detected in several cancer: pancreas, breast, colon, T-cells and chronic myelogenous leukaemia (Hayes et al., 2006). Another study found FOX2 (RBM9) down-regulated in ovarian cancer and misspliced in breast cancer, leading to an overall depletion of FOX2 protein and splicing modulation (Venables et al., 2009). Interestingly, our results suggest that these variations can be extended to more cancers for SF2/ASF (bladder, brain and vulva cancers), SRPK1 (bladder, lung cancers and neuroblastoma) and FOX2 (prostate and vulva cancers).

3.4 Discussion

In this study we applied computational methods to identify splicing factor gene expression variation associated with cancer. We have identified cancer misregulation for 192 splicing-related genes encoding the major splicing protein families snRNPs, hnRNPs, SRs, SR-kinases, RNA-helicases-like and other splicing regulators. The majority of differentially expressed splicing regulators were up-regulated in cancer and some misregulations appear consistently in several cancer types.

Previously, Kirschbaum-Slager et al. (2004) using serial analysis of gene expression (SAGE) and Oncomine (microarray-based information) also observed over-expression of splicing factors as general trend in breast, colon, prostate and brain cancers. Importantly, however, changes in splicing factor mRNA levels may not necessarily reflect on protein expression due to post-transcriptional regulation, namely through mediation by miRNA regulation (Boutz et al., 2007; Makeyev et al., 2007). Several studies have shown that expression of miRNAs is altered in cancer and that there is a strong correlation between abrogated expression of miRNAs and oncogenesis (Esquela-Kerscher and Slack, 2006). Therefore, experimental investigation of the candidate cancer splicing regulators identified in this study is required to determine whether specific changes in the protein concentration and/or activity do occur.

3.4.1 Can splicing factors act as oncogenes?

Because changes in the concentration, localization and/or activity of splicing factors are known to modify the selection of splice sites (Matlin et al., 2005), it is predicted that the abnormally expressed splicing factors found in tumour cells induce the production of mRNA isoforms that were either non-existing or less abundant in normal cells. This phenomenon might contribute directly or indirectly to cancer development, progression and/or response to therapy. A recent study demonstrated for the first time that over-

Table 3.3: **Splicing factors altered in cancer and potential splicing target.**

	Name (other names)	Cancer tissue	Affected mRNA	References
Up-regulated in cancer				
SR and SR-related proteins	SF2/ASF (SFRS1)	Colon, thyroid, small intestine, kidney, lung, breast	RON, BIN1, S6K1MNK2	Ghigna et al. (2005); Karni et al. (2007)
	SC35 (SFRS2)	ovary	-	Fischer et al. (2004); Xiao et al. (2007)
	SRp20 (SFRS3)	ovary	MRP1	He et al. (2004)
	SRp40 (SFRS5)	breast	CD44	Huang et al. (2007)
	SRp55 (SFRS6)	breast	-	Karni et al. (2007)
	Tra2-1 (SFRS10)	breast	CD44	Watermann et al. (2006)
	SRm160 (SRRM1)	Thymic ephitelium, stomach and kidney	CD44	Cheng and Sharp (2006); Harn et al. (1996); Lee et al. (2003b); Wu et al. (2003)
hnRNP proteins	hnRNPA1 (HNRNPA1)	Lung, breast, ovary	-	Patry et al. (2003); Zerbe et al. (2004)
	hnRNPB1 (HNRNPA2B1)	Lung	-	Sueoka et al. (2001); Zhou et al. (1996)
	hnRNPF (HNRNPF)	colon	-	Balasubramani et al. (2006)
	hnRNPL (HNRNPL)	Esophageal cancer cell lines	-	Qi et al. (2008)
	hnRNPK (HNRNPK)	Colorectal, oral	-	Carpenter et al. (2006); Roychoudhury and Chaudhuri (2007)
	PTB (PTBP1)	Gliobastoma, ovary	FGFR1, MRP1	Jin et al. (2003a); He et al. (2004)
Other factors	YB-1 (YBX1)	Ovary	CD44	Fischer et al. (2004)
	SPF45 (RBM17)	Bladder, breast, colon, lung, ovary, pancreas and prostate	FAS	Corsini et al. (2007); Sampath et al. (2003)
	SRPK1 (SRPK1)	Pancreas, breast, colon, T-cells, Chronic myelogenous leukemia	MAP2K2	Hayes et al. (2006)
	HuR (ELAVL1)	Breast, ovary	FAS	Denkert et al. (2004); Izquierdo (2008)
	HuD (ELAVL4)	Leukaemia T-ALL	IK	Bellavia et al. (2007)
	Sam68 (KHDRBS1)	prostate	-	Busà et al. (2007)
Down-regulated in cancer				
hnRNP proteins	hnRNP E2 (PCBP2)	oral	-	Roychoudhury et al. (2007)
Other factors	U2AF35 (U2AF1)	Pancreas	CCK-B	Ding et al. (2002); Pacheco et al. (2006b)
	SF1 (SF1)	Colorectal	WISP1, FGFR3	Shitashige et al. (2007a, b)
	RBM5 (LUCA15)	lung	-	Oh et al. (2006)

expression of a splicing factor can indeed trigger malignant transformation (Karni et al., 2007). The authors showed that the splicing factor SF2/ASF is up-regulated in various human tumours and affects alternative splicing of the tumour suppressor BIN1 and the kinases MNK2 and S6K1. The resulting BIN1 isoforms lack tumor-suppressor activity, the MNK2 isoform promotes MAP kinase-independent eIF4E phosphorylation, and the S6K1 isoform has demonstrated oncogenic properties (Karni et al., 2007). This study serves as a proof-of-principle and shows that abnormally expressed splicing proteins can have oncogenic properties.

A previous study had indicated that SF2/ASF affects alternative splicing of Ron, a tyrosine kinase receptor involved in cell dissociation, motility, and matrix invasion (Ghigna et al., 2005). An alternatively spliced isoform of Ron that lacks exon 11 produces a constitutively active protein that is expressed in gastric, breast and colon cancers and induces an invasive phenotype (Collesi et al., 1996; Ghigna et al., 2005). Binding of SF2/ASF to a regulatory sequence in exon 12 stimulates skipping of exon 11, and overexpression of SF2/ASF activates cell locomotion. This effect can be reversed by specific knockdown of the alternatively spliced Ron isoform, suggesting that an up-regulation of SF2/ASF could contribute to malignant transformation by inducing alternative splicing of Ron.

Several additional splicing proteins have been detected to be up-regulated in various human tumours (Table 3.3), but in most cases the effect that these changes have on splicing regulation is unknown. In contrast, the number of splicing proteins that have been detected to be down-regulated in cancer is much lower (Table 3.3). For example, reduced expression of U2AF35 was found in pancreatic cancer cells and correlated with missplicing of the cholecystokinin-B/gastrin (CCK-B) receptor mRNA (Ding et al., 2002). Furthermore, RNAi-mediated downregulation of U2AF35 in HeLa cells has been reported to alter the ratios of alternatively spliced isoforms of transcripts encoding the oncogenic Cdc25B phosphatase, and to increase the level of Cdc25B protein (Pacheco et al., 2006b).

In conclusion, there is a growing list of splicing factors that have been found to be up- or down-regulated in cancers, as compared to the corresponding normal tissues. Nevertheless, in many cases, the available data are limited to correlations. A challenge for the future will be to determine whether these changes are directly contributing to the cancer phenotype, or they merely represent one of the multiple processes that are altered in cancer cells. A critical issue is whether cells expressing abnormal levels of certain splicing factors are positively selected for during tumour progression, as misregulated splicing factors may induce production of splice variants that encode protein isoforms with advantages such as increased proliferation, anti-apoptotic or pro-angiogenic effects, enhanced cell motility or tumor cell survival. Moreover, many RNA-binding proteins are multifunctional and their abnormal expression may have oncogenic effects that are independent from splicing. Also

unknown is what triggers the up- and down-regulation of splicing proteins. Consistent with the view that cancer-associated genetic instability is likely to play an important role in this process, over-expression of splicing factor SF2/ASF was shown to associate with amplification of the gene encoding for it (Karni et al., 2007), whereas reduced expression of RBM5 in lung cancer correlates with deletion of its gene locus at chromosomal region 3p21.3 (Oh et al., 2006). Alternatively, or additionally, splicing factor transcripts appear to be preferential targets for disrupted splicing in cancer tissues (Kim et al., 2008; Ritchie et al., 2008). Cancer-specific splicing factor isoforms could either alter the function of the protein in the cell, or reduce its level due to the introduction of a premature stop codon and nonsense-mediated decay of the mRNA.

3.4.2 Splicing factors and anticancer therapy

During the past 20 years, anticancer drug development has focused on targeted medicines that are more specifically associated with tumour cells than conventional cytotoxic drugs. Over 600 new agents are currently in the development pipeline in the hope of attaining greater anticancer activity with fewer side effects (Dancey and Chen, 2006). Still at the preclinical stage, several approaches are being explored for the correction of cancer-associated splicing abnormalities (for a comprehensive review see Pajares et al. (2007); Wang and Cooper (2007)). One strategy uses synthetically modified oligonucleotides that are able to block spliceosome assembly at specific sites, thereby preventing the generation of cancer-associated splice variants. This approach has been successfully used to shift the ratio of antiapoptotic to proapoptotic proteins produced by alternative splicing of the Bcl-x gene, thereby sensitizing refractory cancer cells to undergo apoptosis in response to chemotherapeutic drug treatment (Taylor et al., 1999). Another strategy consists of raising antibodies against epitopes that are uniquely present in the cancer-associated protein isoforms and conjugate the antibodies to tumour-cell toxins. For example, human recombinant antibodies specific to the alternatively spliced domains of tenascin-C large isoform, an abundant glycoprotein of cancer extracellular matrix that is virtually undetectable in normal adult tissues, show promising tumor-targeting properties (Brack et al., 2006).

Strategies for targeting components of the splicing machinery that are abnormally expressed in cancer are expected to be less specific because they are likely to impinge on splicing regulation in normal cells. Nevertheless, many approaches have been attempted with encouraging results. Particular attention has been devoted to the development of protein kinase inhibitors that modulate the activity of splicing factors containing RS domains, which are characterized by repeats of arginine-serine dipeptides. Phosphorylation/dephosphorylation of these serine residues are thought to act as switches that modulate the binding properties to both RNA and proteins (Singh and Valcárcel, 2005). Although

there are several known splicing factor kinases, members of the SRPK (SR protein kinase) family appear to be the most relevant in cancer (see Table 1). Down-regulation of SRPK1 expression by siRNA in cancer cell lines caused a reduction of cell proliferation and increased sensitivity to gemcitabine and cisplatin, making the approach of targeting SRPK1 a promising tool that may prove therapeutically effective for tumours that overexpress of this protein (Hayes et al., 2006, 2007). In addition to kinases, aberrant expression of splicing factors in tumour cells might be implicated in resistance to drugs commonly used in cancer therapy. For example, increased expression of the splicing factors PTB and Srp20 in ovarian cancer correlates with the production of alternatively spliced isoforms of the multidrug resistance protein 1 (mrp1) that confer increased resistance to doxorubicin (He et al., 2004). Another splicing factor highly expressed in numerous carcinomas, SPF45 (RBM17), affects the alternative splicing of the apoptosis regulator Fas (Corsini et al., 2007), and over-expression of SPF45 has been implicated in resistance to doxorubicin and vincristine (Sampath et al., 2003).

It is fully anticipated that inhibiting the function of either a splicing kinase or a splicing protein will have a pleiotropic effect by altering the splicing of numerous gene products in both cancer and normal cells. However, a well-established principle of cancer therapy is to use a combination of drugs with different mechanisms of action and resistance, at their optimal doses and according to schedules that are compatible with normal cell recovery. Thus, it may be possible to develop and optimize agents that temporarily inhibit a splicing regulator and partially correct abnormal splicing, resulting in enhanced tumour cell killing by chemotherapeutic drugs. Very recently, a proof of concept has been provided for the development of anti-tumour compounds that target the splicing machinery. Spliceostatin A (Kaida et al., 2007) and pladienolide (Kotake et al., 2007), two potent inhibitors of cycling cancer cells, target the essential splicing protein SF3b and inhibit splicing of several transcripts. Both drugs are only mildly toxic to animals and a pladienolide derivative, E7107, has already progressed to clinical trials. This moderate toxicity is probably due to partial inhibition of splicing throughout the organism, but why cancer cells are particularly vulnerable to the drugs remains unknown. Most important, these studies have defined a new mode of action in anticancer drugs and identified a ubiquitous core component of the U2 snRNP, SF3b, as a valuable new therapeutic target.

In summary, the rapid development and increasing availability of novel genome-wide tools will soon provide a catalogue of all splicing factors and all splice variants that are differentially expressed in specific cancer types and the corresponding normal tissues. Irrespectively of whether changes in splicing play a direct causative role in cancer, or act as modifiers or susceptibility factors in the oncogenic process, the identification of splicing signatures is likely to provide important markers for diagnosis, prognosis, and/or

sensitivity to treatment. A full description of all components of the splicing machinery and splicing events altered in cancer will also identify potential new targets for therapeutic approaches. However, the most challenging goal for the future will be to integrate the different layers of gene expression regulation altered in cancer and to acquire a systems biology view of the multiple molecular mechanisms that contribute to the pathophysiology of this disease.

Chapter 4

Cancer-associated splicing misregulation

The original work described in this chapter is in preparation for submission to a peer reviewed journal.

I would like to stress that some of the results presented and discussed in this section are the product of collaborative work. Inês Mollet was responsible for the annotation of exon-microarray probes according to ExonMine.

Keywords: alternative splicing; exon-microarrays; cancer; splicing regulatory sequences

Abstract: Alternative splicing generates a huge diversity of transcript variants and disruption of splicing regulatory networks is emerging as a major contributor to various diseases, including cancer. Disruption of alternative splicing in cancer cells occurs in the absence of mutations in the affected genes and current evidence indicates that the splicing machinery is a major target for misregulation in cancer. Here, we collected a high-confidence set of misregulated splicing events for colon and lung cancers by applying a comprehensive workflow analysis to splicing-sensitive microarray data. Functional analysis using distinct sets of genes split according to misregulation level revealed that some pathways are more affected by variations in transcript abundance, whereas others at alternative splicing level. We identified misspliced exons containing *cis*-acting RNA elements obtained from CLIP-seq data for SF2/ASF, which is overexpressed in both cancers. The cancer-associated splicing events also contained enriched motifs that resemble binding sites for additional cancer misregulated splicing factors.

4.1 Introduction

Alternative splicing of precursor messenger RNAs (pre-mRNAs) is a mechanism by which proteomic diversity is generated from a low number of genes. The number of human genes subject to alternative splicing increased from >60% based on ESTs-based studies to 92-95% in recent high-throughput sequencing technologies (Wang et al., 2008a; Pan et al., 2008).

Pre-mRNA splicing is carried out by the spliceosome, a macromolecular complex formed from several small nuclear ribonucleoprotein particles (snRNPs) and numerous non-snRNP splicing factors (Jurica and Moore, 2003; Wahl et al., 2009). Specific sequences located at and near the $5'$ and $3'$ splice sites are recognized by the spliceosome, triggering the splicing process. However, the splice site sequences are weakly conserved and require additional regulatory sequences termed splicing enhancers and silencers located in exons or introns (Maniatis and Tasic, 2002; Matlin et al., 2005). These additional regulatory sequences are recognized by splicing factors, which are commonly classified as splicing activators or repressors depending on whether they facilitate or suppress the assembly of snRNPs onto splice sites. Thus, regulation of alternative splicing is mediated by the cooperative binding of *trans*-acting splicing proteins to *cis*-acting sequence elements in the pre-mRNA.

Disruption of alternative splicing has been associated with several diseases, including cancer (Cooper et al., 2009). This disruption can be caused by several mutations affecting the splicing of oncogenes, tumour suppressors and other cancer-relevant genes (Srebrow and Kornblihtt, 2006; Venables, 2006). However, many splicing abnormalities identified in cancer cells are not associated with mutations in the affected genes. Indeed, recent studies suggest that changes in splicing factor expression may play a key role in the general splicing disruption that occurs in many cancers.

Karni et al. (2007) showed that splicing factor SF2/ASF (SFRS1) is up-regulated in various human tumours and affects alternative splicing of the tumour suppressor BIN1 and the kinases MNK2 and S6K1. Recent bioinformatics studies also suggest that splicing factors are not expressed at proper levels and/or their functions are impaired in cancer (Kirschbaum-Slager et al., 2004; Kim et al., 2008; Ritchie et al., 2008).

A large number of cancer-associated alternative splicing events have been identified by several studies using splicing-sensitive microarrays. however, few associations between misregulated splicing factors and splicing events in the same cancer were established (Relógio et al., 2005; Gardina et al., 2006; Thorsen et al., 2008; Zhang et al., 2006; Thorsen et al., 2008; French et al., 2007; Cheung et al., 2008; Xi et al., 2008; Soreq et al., 2008; Thorsen et al., 2008). For example, Relógio et al. (2005) showed for Hodgkin lymphoma cells that Nova2 was overexpressed and the expression was correlated with gene isoforms detected

with splicing-sensitive microarrays.

More recently, using high-throughput RT-PCR, misspliced genes were identified in ovarian and breast cancers, from which many contained binding-sites for FOX2 protein (Venables et al., 2009). These authors also showed that FOX2 is down-regulated in ovarian cancer and misspliced in breast cancer, leading to an overall depletion of FOX2 protein and splicing modulation.

Here, we used splicing-sensitive microarray data to investigate associations between changes in splicing factor expression and alternative splicing events in colon and lung cancers.

4.2 Material and Methods

4.2.1 Microarray data collection and analysis

For identification of cancer-associated alternative splicing events Affymetrix GeneChip Exon Microarray data sets were collected from previous studies for colon cancer (Gardina et al., 2006) and lung cancer (Xi et al., 2008). Microarray data were analysed using R and suitable packages available from CRAN (R Development Core Team, 2009) and Bioconductor (Gentleman et al., 2004). The raw data for the Affymetrix GeneChip Exon microarray data sets were normalized and summarized using the FIRMA method (Purdom et al., 2008) implemented in *aroma.affymetrix* package (Bengtsson et al., 2008). The statistical significance from the gene and exon expression alterations was assessed using linear models and empirical Bayes methods (Smyth, 2004) implemented in the *limma* package (Smyth, 2005). Graphical representation of Firma scores for each probeset or probe selection region (PSR) was based on annotated exons from ExonMine (Mollet et al., submitted, `http://imm.fm.ul.pt/exonmine/`).

4.2.2 Functional and pathway analysis

Functional and pathway analysis was performed using four gene sets according to misregulation level: transcript; alternative splicing; only alternative splicing and no transcript abundance variation; all misregulated genes at transcript or alternative splicing levels. The entire list of genes present in the Affymetrix GeneChip Human Exon microarray were used as control or reference set for all analyses.

The enrichment of biological functions was analysed using Ingenuity Pathway Analysis software (Ingenuity Systems, Mountain View, CA, USA). Pathways significantly affected were identified using the systems biology approach from Pathway-Express tool (Draghici et al., 2007).

4.2.3 *Ab initio* motif searches

The *ab initio* motif search was performed using the SeedSearcher algorithm (Barash et al., 2001). This algorithm identifies sequence motifs that discriminate a set of query sequences from a group of control sequences. Three groups of query sequences were defined for each cancer: exons enriched in cancer samples (inclusion in cancer), exons enriched in normal samples (exclusion in cancer), exons with splicing misregulated (inclusion or exclusion). For the control group we selected exons without splicing variations between cancer and normal tissue from genes with high overall gene expression in all samples (expressed genes).

The motif search covered the sequences for the alternative exon, 150 nucleotides of intron sequencing flanking the misspliced exon and the neighboring exons. The first 10 and last 30 nucleotides of introns were excluded from all sequences as these contain conserved signals for the constitutive splicing machinery. Only misspliced exons for which no splicing variation in the neighbor exons was detected in cancer were used. Several SeedSearcher searches were performed using different motif length and with various degrees of sequence flexibility (i.e., number of wildcards represent the possible number of degenerate nucleotides): 5 nt (zero and one wildcard), 6 nt (zero and one wildcard), 7 nt (zero, one and 2 wilcards), 8 nt (one to 3 wildcards), 9 nt (2 to 3 wildcards) and 10 nt (2 to 4 wildcards). The motifs were ranked by the statistical significance score computed and corrected for multiple hypotheses testing by SeedSearcher (Barash et al., 2001). To identify the enriched motifs with different frequencies between included and excluded exons in cancer we used Fisher's exact test (Fisher, 1935).

Statistically significant motifs were compared against a list comprising previously reported motifs associated with splicing (Cartegni et al., 2003; Martinez-Contreras et al., 2007; Gabut et al., 2008; Long and Caceres, 2009).

4.3 Results and Discussion

4.3.1 Workflow for the detection of alternative splicing

To identify misregulated splicing events in cancer, we analysed Affymetrix GeneChip Exon Microarray data sets from previous studies for colon cancer (Gardina et al., 2006) and lung cancer (Xi et al., 2008).

The Affymetrix GeneChip Human Exon 1.0 ST array can determine the expression of virtually all exons present in human genome, deriving from annotations ranging from highly curated mRNA sequences to ab-initio computational predictions (Gardina et al., 2006). The array contains approximately 5.4 million probes grouped into 1.4 million probesets or probe selection regions (PSR), for which 90% are represented by 4 probes. A

PSR usually corresponds to one exon, however some exons can contain several PSRs.

For detection of alternative splicing and overall gene expression variation, we applied and combined methodologies previously and successfully applied for pre-processing, summarization, filtering and statistical analysis. Microarray quality, normalization and summarization were assessed using the *aroma.affymetrix* package (Bengtsson et al., 2008). Background correction and quantile normalization were applied using the entire set of main-design probes.

After pre-processing, two different types of expression indexes were computed for each gene: exon (represented by one or more probe selection region) and transcript levels. The transcript levels correspond to the amount of molecules transcribed from a single gene including various alternative splicing isoforms and it provides the baseline to compare the individual exon expression. These indexes were estimated using the FIRMA method (Purdom et al., 2008) implemented in *aroma.affymetrix* package. Briefly, the transcript level of each sample is estimated by fitting an additive model for each gene and the identification of alternative splicing events is framed as a problem of outlier detection. Thus, alternative spliced exons are identified by whether its probesets systematically deviate from the expected transcript expression level and a FIRMA score is estimated for each probeset. For the summarization we used the gene definitions from Ensembl (Hubbard et al., 2007), where 332532 probesets were previously mapped to the Ensembl annotated exons and corresponded to 23385 genes (Purdom et al., 2008).

To reduce false positive predictions of alternative splicing events, filtering steps were applied to the data prior to statistical analysis. As has been noted in previous works (Gardina et al., 2006; Purdom et al., 2008; Shah and Pallas, 2009) probesets with hybridization levels close to background (unexpressed transcripts and/or exons spliced out in all samples) can induce false positives. We filtered out probesets and transcript summary values that were found below the lower quartile of the intensity distribution for all samples (Shah and Pallas, 2009). For alternative splicing detection we removed also the probesets that hybridize at very high levels (saturated response), filtering out probesets with intensity values found above the 90th percentile for all samples. Since the FIRMA methods detects alternative splicing as deviation from the expected transcript expression, probesets close to intensity saturation will result in a decrease in sensitivity for true splicing detection (Bemmo et al., 2008).

We also re-annotate all microarray probes according to ExonMine database (Mollet et al., submitted, `http://imm.fm.ul.pt/exonmine/`) and removed all the probesets that perfectly match more that one region (co-hybridization). This database provides all splicing patterns obtained from clustering spliced ESTs and mRNAs to protein coding genes and detects a significantly higher percentage of spliced genes, isoforms and exons compared

Table 4.1: **Precision of workflow for alternative splicing events detection.** Comparison of the alternative splicing events (ASE) found in our analysis with the ASE validated in previous exon-microarray studies using RT-PCR. The number of ASE and Precision (proportion of selected ASE corresponding to True Positives) for our analysis and previous studies is also indicated.

Cancer Type	Genes with ASE validated in Previous Studies				Previous genes in our analysis	
	Reference	Confirmed	Nr.	Precision	Nr.	Precision
colon	Gardina et al. (2006)	Yes	14	0,33	12	0,71
		No	29		5	
	Thorsen et al. (2008)	Yes	6	0,27	4	0,80
		No	16		1	
lung	Xi et al. (2008)	Yes	6	0,55	3	0,75
		No	5		1	

to other recently published alternative splicing databases.

After pre-processing and filtering, we assessed the fold-changes and respective statistical significance for alterations at transcript and alternative splicing level between cancer and normal samples using linear models and empirical Bayes methods (Smyth, 2004) implemented in the *limma* package (Smyth, 2005). Transcript and alternative splicing alterations were ranked according to B-statistics and a suitable B-value cut-off was selected by visually inspecting the volcano plot, which compares biological significance (represented by fold-changes) with statistical significance (B-values) (Jin et al., 2001). Additionally, we verified the p-values corresponding to moderated t-statistics. Using the Benjamini and Hochberg method (Benjamini and Hochberg, 1995), all transcript and alternative splicing alterations selected as significant had adjusted p-values lower than 0.00005 for lung cancer and 0.05 for colon cancer. The higher p-values found for colon cancer are due to the smaller number of samples and the heterogeneous stages of cancer progression as stated in the original study (Gardina et al., 2006).

To overcome the problem of confounding multiple alternative splicing events with overall transcript variation (McKee et al., 2007), alternative splicing alterations from genes presenting changes in transcript abundance and alternative splicing for 40% of the exons were not considered for further analysis.

Finally, for visualization of results in the context of gene architecture, we graphically displayed FIRMA scores and intensity values for all probesets and samples along the gene using information from ExonMine (Figures 4.1 and 4.2). Positive FIRMA scores correspond to exon enrichment (inclusion) in the transcript, whereas negative values represent exon depletion (exclusion).

To evaluate the quality of the alternative splicing events selected we compared our results with previous exon-microarray studies where validations were perfomed by RT-PCR (Table 4.1).

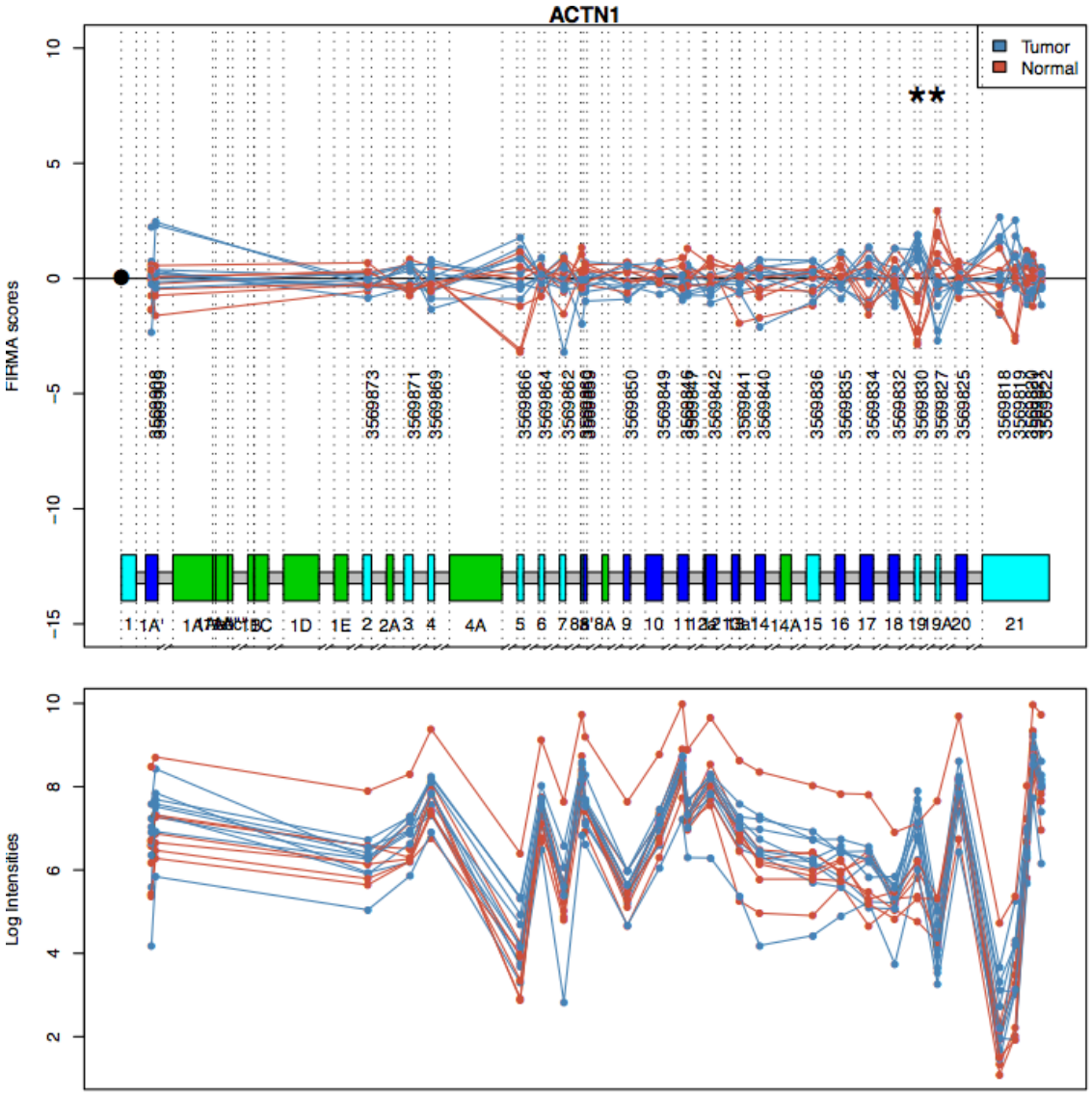


Figure 4.1: **Example of alternative splicing event found in colon cancer.** Graphical representation of the ACTN1 gene showing mutually exclusive exons: inclusion of exon 19 and exclusion of exon 19A in colon cancer (highlighted with *). The dots represent the log intensities (bottom plot) and the FIRMA scores (top plot) for each probe set and each line corresponds to a sample. The exons (constitutive exons - dark blue; alternative exons - light blue; new exons - green) and introns (gray) for ACTN1 are displayed in the top plot according to ExonMine information.

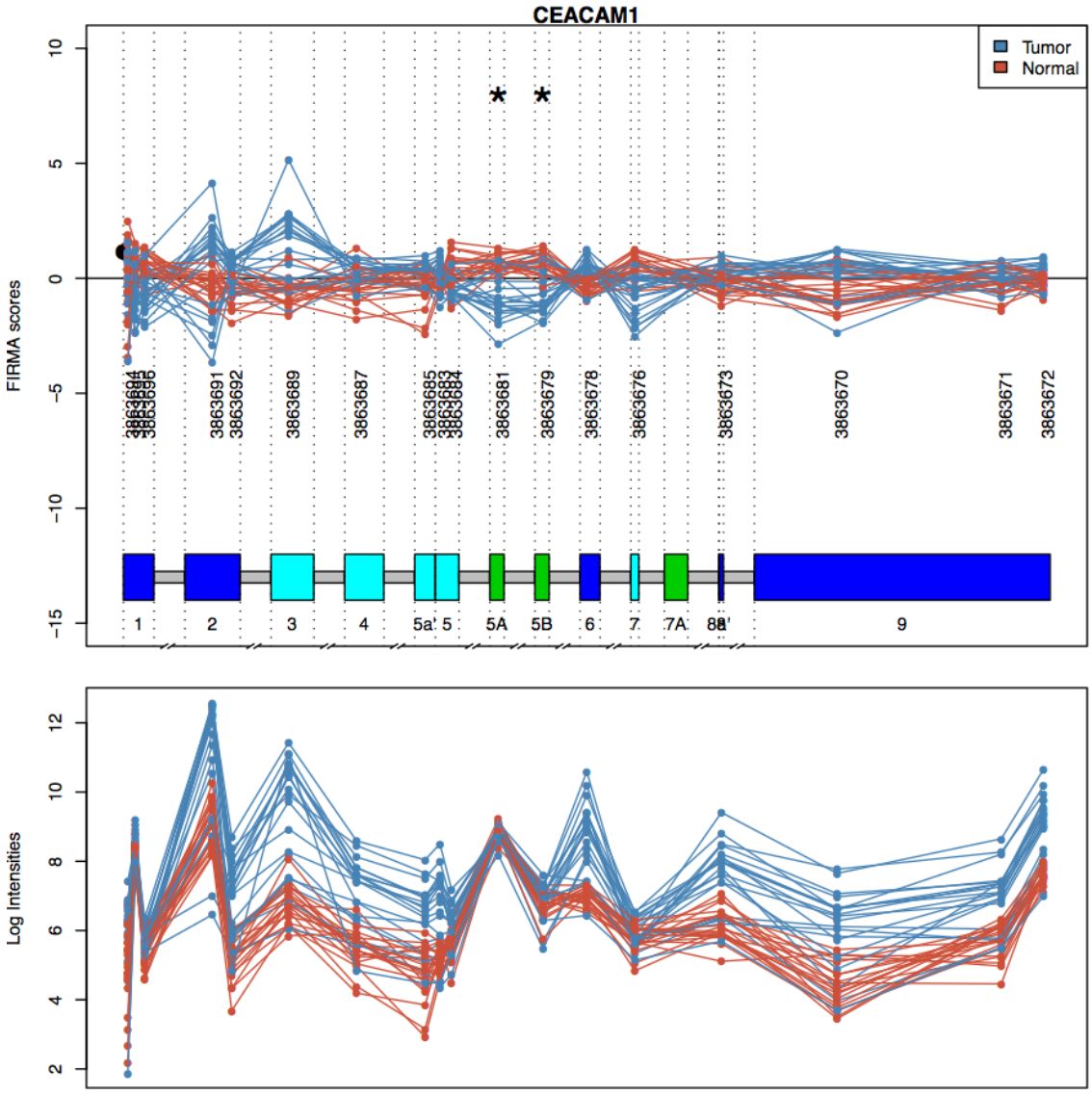


Figure 4.2: **Example of alternative splicing event found in lung cancer.** Graphical representation of the CEACAM1 gene showing exclusion of exons 5A and 5B in lung cancer (highlighted with *). The dots represent the log intensities (bottom plot) and the FIRMA scores (top plot) for each probe set and each line corresponds to a sample. The exons (constitutive exons - dark blue; alternative exons - light blue; new exons - green) and introns (gray) for CEACAM1 are displayed in the top plot according to ExonMine information.

In the original study from the colon cancer microarray data set 14 out of 43 genes were confirmed by RT-PCR (Gardina et al., 2006). 12 of the validated genes were also selected in our workflow (LGR5, ZAK, ACTN1, ATP2B4, CALD1, COL6A3, FN1, GK, MAST2, TPM1, VCL), whereas only 5 from the 29 false positives appear in our list.

In another exon-microarray study using different colon samples, 6 of 23 splicing events were confirmed (Thorsen et al., 2008) and 4 of the validated events were also selected in our analysis (ACTN1, CALD1, COL6A3 and VCL). Based on this validation our results only presented one (TPM1) of the 16 false positives. However this event was confirmed in the study referred to above (Gardina et al., 2006).

For lung cancer microarray data set 6 out of 11 genes were confirmed in the original study (Xi et al., 2008), from which 3 genes also presented splicing events in our analysis (CEACAM1, ERG and RASIP1). However, we identified the upstream exon as being alternatively spliced for ERG and RASIP1 when compared to the original study. Splicing events were identified for several exons of these two genes in our analysis suggesting more complex alternative splicing patterns.

FIRMA scores of alternative splicing events previously validated for ACTN1 in colon (Gardina et al., 2006) and CEACAM 1 in lung cancer (Xi et al., 2008) are graphically represented in Figures 4.1 and fig:CEACAM1.

Overall, these results indicate a higher precision from our workflow relative to previous studies. Although we could not detect all previously validated alternative splicing events, we should have a high confidence in our collection of alternative splicing events.

4.3.2 Cancer-associated misregulations at transcript and alternative splicing level

Using the workflow described above we identified expression variation at transcript and alternative splicing level for colon and lung cancer (Table 4.2 and Annex Tables A.3.1 - A.3.4).

A lower number of misregulated genes at both levels were identified for colon relative to lung cancer, but this can result from the modest sample size and the heterogeneous stages of cancer progression as stated in the original study (Gardina et al., 2006). However, colon cancer presented higher number of genes with changes only at alternative splicing but no variation at transcript level. Thus, the majority of misregulated genes in lung cancer undergo changes both in abundance and in transcript architecture.

Colon and lung cancers presented common misregulated genes at transcript and alternative splicing level (Table 4.2). The same misspliced exon was observed in both cancers for 29 genes and eight genes present variations only at the alternative slicing level for both cancers (COL6A3, MDK, AHNAK, JUP, PPP1R12B, SNTG2, CDCA7, CDH19).

Table 4.2: **Number of genes with variations at transcript and alternative splicing level for colon and lung cancers.** The number of genes with overall up or down-regulation is described for each cancer type. The number of genes presenting alternative splicing events (ASE) but no variation at trasncript level is also shown.

Number of genes:	Colon Cancer	Lung Cancer	Common to Both Cancers
Differentially Expressed	560	1456	154
Up-regulated	418	527	99
Down-regulated	142	929	52
With ASE (Nr. of ASE)	467 (550)	791 (1308)	29 (38)
With ASE but no variation at gene level	330	223	8

Indeed, the misspliced exon identified for COL6A3 was recently found and validated for colon, bladder and metastic prostate cancers (Thorsen et al., 2008). COL6A3 encodes a protein of the extracellular matrix and the included exon 6 in cancer most likely contains several predicted phosphorylation sites, which have potential regulatory effects on protein function (Thorsen et al., 2008). Moreover, these authors found three genes with common misregulated exons for colon, bladder and prostate cancers using exon arrays, which alternative splicing events were also detected in our results for colon cancer (CALD1,VCL and ACTN1). The previously reported results showed a clear relationship between advanced cancer stage and systematic occurrence of alternative splicing isoforms, suggesting that some of the identified splice variants could be driving forces in cancer development (Thorsen et al., 2008).

Similarly to COL6A3, our results suggest that the cancer-misregulated splicing events identified by our approach may affect protein function. According to Ensembl annotation (Hubbard et al., 2007) approximately 75% of the alternative splicing events found for colon and lung cancers were located inside the coding sequence and from which 40% most likely encode protein domains (Annex Tables A.3.3 and A.3.4).

Next, we asked whether the different sets of genes misregulated at the transcript and alternative splicing levels in cancer belong to different functional categories. The genes were split in four groups according to misregulation level: transcript; alternative splicing; only alternative splicing and no transcript abundance variation; all misregulated genes at transcript or alternative splicing levels.

Functional analysis using Ingenuity Pathway Analysis software (Ingenuity Systems, Mountain View, CA, USA) showed cancer and genetic disorder (also includes cancer) as the top first associated diseases for both cancers and all sets of genes (Benjamini-Hochberg adjusted p-value < 0.05) (Annex Table A.3.5 and A.3.6). Roughly 30% of the misregulated genes were associated with cancer disorder. Gastrointestinal disease also appeared as a top disease for all sets of genes in colon cancer (Benjamini-Hochberg adjusted p-value < 0.05).

Similar enrichment was found for both cancer types in molecular and cellular functions with top first corresponding to: cell cycle, cell death, cellular growth and proliferation (Annex Table A.3.5).

We next identified pathways that were significantly affected by applying the systems biology approach from Pathway-Express tool (Draghici et al., 2007). This method includes the classic statistical features of gene set enrichment analysis but also considerers other factors such as the magnitude of the expression changes of each gene, the position of the genes on the given pathways, the topology of the pathway that describes how these genes interact, and the type of signaling interactions between them. The method estimates an impact factor for each pathway, which corresponds to the negative log of the global probability of having both a statistically significant number of misregulated genes and a large perturbation in the given pathway. Since we could not determine the final variation for genes affected at splicing level, we performed the analysis without considering expression fold-changes.

Figure 4.3 shows the top enriched pathways according to the impact factor for each cancer type and misregulation level (False discovery rate adjusted p-value < 0.05) (Annex Tables A.3.5 and A.3.6). The majority of affected pathways were previously associated to cancer (Weinberg, 2007) and although some pathways contained misregulated genes at transcript and alternative splicing levels, a significant enrichment was only observed for some gene sets. So, most pathways appeared to be differentially affected by variations in transcript abundance or architecture. Extracelullar matrix receptor was the only pathway for which an enrichment was observed for all groups of genes, thus suggesting a high impact of changes at both levels for the component genes.

Pathways related to cell growth and death (cell cycle, p53-signaling) and replication and repair of genetic information (DNA replication, mismatch repair and nucleotide excision repair) were significantly impacted by variations of transcript abundance in colon cancer. In contrast, alternative splicing alterations appear to affect other pathways, namely: cell motility (regulation of actin cytoskeleton), immune system (Toll-like receptor signaling pathway, leukocyte transendothelial migration), cell communication (including adherens juntion, gap junction and tight junction).

Since the majority of misspliced genes in lung cancer also present changes at transcript abundance only two pathways appeared significantly affected by genes with changes only at alternative splicing level: extracellular matrix receptor and cell communication.

Alterations in transcript diversity for lung cancer affect also pathways related to signaling molecules and interaction (ECM-receptor interaction, cell adhesion molecules), signaling transduction (TGF-beta signaling pathway), cell motility (regulation of actin cytoskeleton) and cell communication (including adherens juntion, gap junction and tight

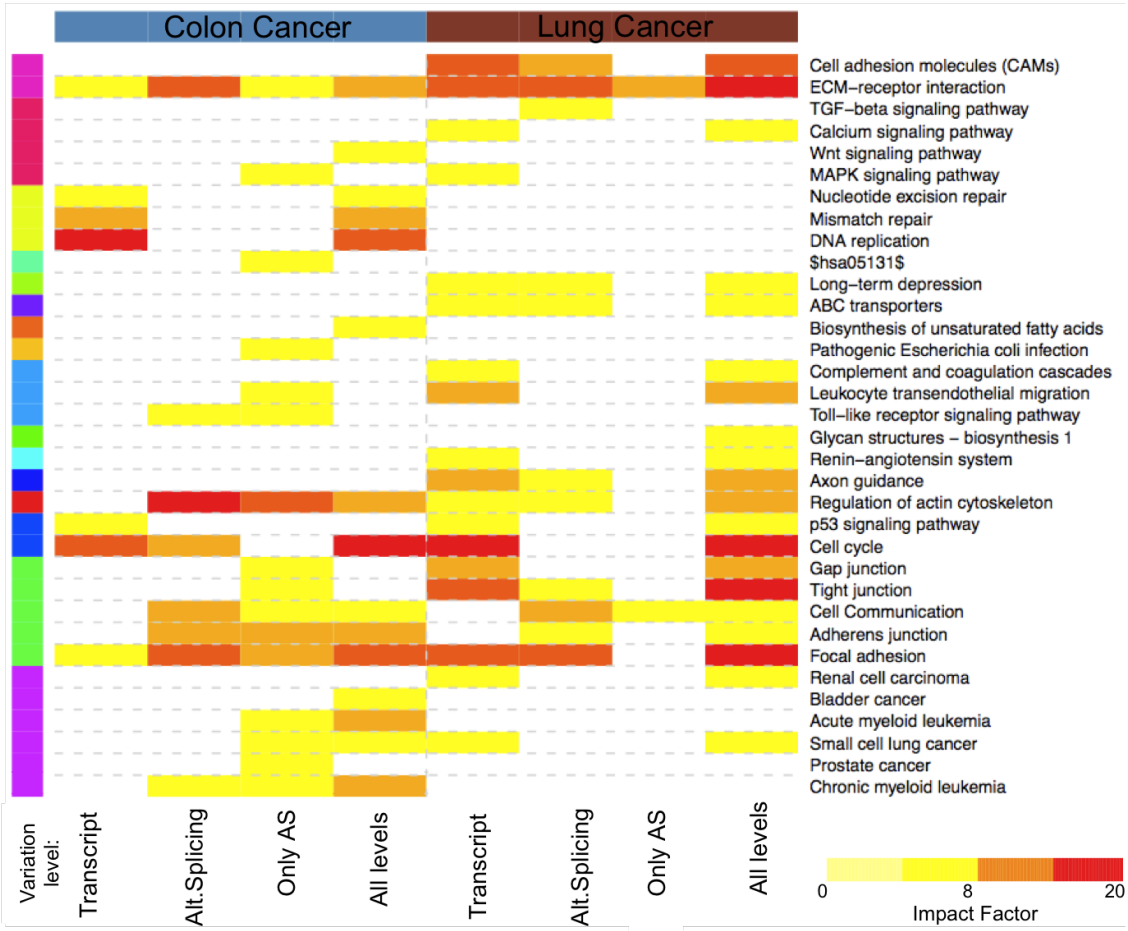


Figure 4.3: **Biological pathways associated with cancer misregulated genes.** Top pathways significantly impacted based on Pathway-Express (fdr adusted p-value < 0.05) for colon and lung cancer splitting genes according to misregulation level: transcript; alternative splicing; only alternative splicing and no transcript abundance variation; all misregulated genes at transcript or alternative splicing levels. The gradient color corresponds to the Impact Factor. The row colors identify pathways for the same biological category. More details in Annex Tables A.3.5 and A.3.6.

junction).

Both cancers appeared to be significantly affected by changes in transcript architecture from genes associated to cell communication. However only two misspliced genes were shared between the two cancers (COL6A3 and FIN1). Similarly, regulation of actin cytoskeleton appeared to be significantly affected by variations at the alternative splicing level but only four misspliced genes were common to both cancers (MYLK, FN1, PPP1R12B, IQGAP3). Association of misregulated splicing in colon, bladder and prostate cancers with regulation of actin cytoskeleton was described in previous genome-wide studies (Gardina et al., 2006; Thorsen et al., 2008). Remodeling of actin cytoskeleton is fundamental in proliferation, apoptosis, cell invasion and metastasis (reviewed in Hall, 2009).

Our results also suggest that common cancer-related pathways are misregulated at different levels. Genes associated to MAPK signaling pathway, leukocyte transendothelial migration, gap junction and tight junction are mostly affected by changes in transcript abundance for lung cancer, whereas in colon cancer alternative splicing has a higher impact. Indeed, a gene from MAPK signalling pathway, ZAK, is down-regulated in lung cancer and misspliced in colon cancer.

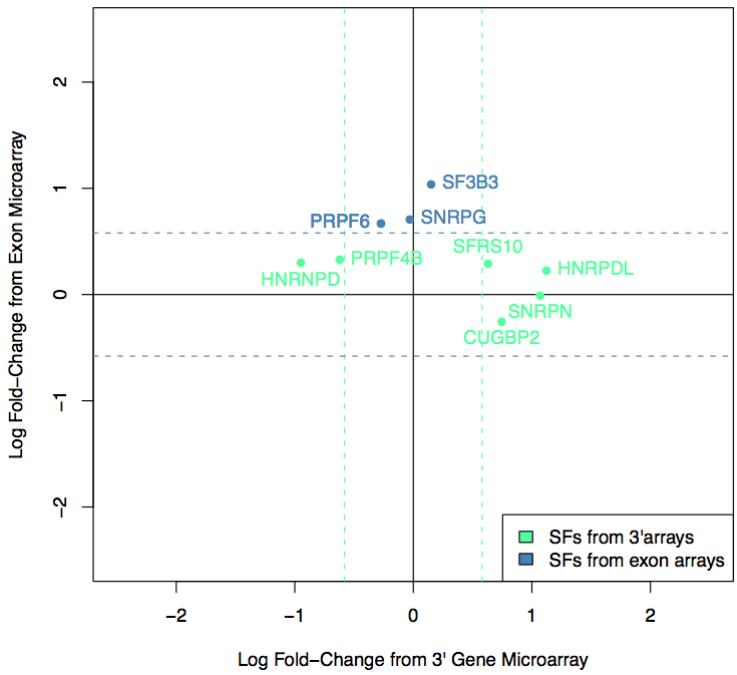
Moreover, we could also identify pathways for which genes are misregulated at transcript and alternative splicing level but the enrichment is only significant when combining the two gene sets (Wnt-signaling pathway and biosynthesis of unsaturated fatty acids for colon cancer, and Glycan structures-biosynthesis for lung cancer).

These results highlight the importance of including information of the several layers of gene expression regulation on functional and pathway analysis.

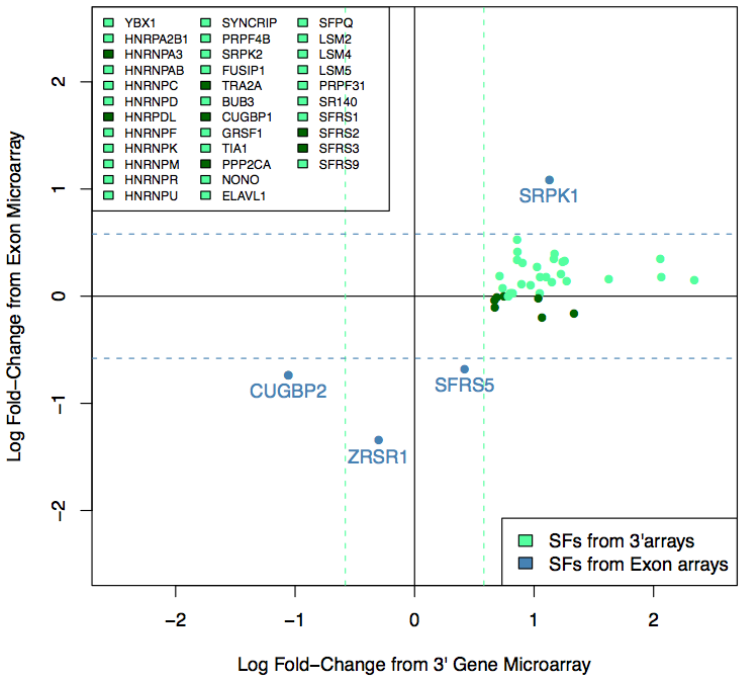
4.3.3 Misregulated splicing factors

Colon cancer presented few misregulated splicing factors, corresponding mostly to snRNP associated proteins overexpressed in cancer (PRPF6, SF3B3, SNRPG) or with alternative splicing events (PRPF8, SF3B2). Other splicing factors, like HNRNPC and NONO also presented alternative splicing events. However, no splicing factors were found in common between Exon and $3'$ microarray analysis from the previous chapter (Section 3.3.1 and Figure 4.4a).

For lung cancer few overall gene expression variations of splicing regulatory genes were also found, with only one up-regulated gene (SRPK1) and three down-regulated genes (CUGBP2, ZRSR1, SFRS5). Comparing the two microarray platforms only SRPK1 and CUGBP2 were detected in common. The majority of the splicing-related genes from $3'$ array analysis presented smaller fold-changes in exon arrays, which were not sufficient to be considered as differentially expressed (Section 3.3.1 and Figure 4.4b). We could also



(a) Colon Cancer



(b) Lung Cancer

Figure 4.4: **Common splicing-related genes between exon and 3′ microarrays.** Splicing-related genes with overall gene expression variations using different microarray platforms Affymetrix 3′ and Exon microarrays for (a) colon and (b) lung cancers.

detect alternative splicing events in lung cancer for CUGBP2 , NOVA2 and SFRS12 genes.

The few misregulated splicing factors detected by both platforms could be due to cancer samples heterogeneity (differences in tumor subtypes) or to differences in the array design (number and distribution of the probes along the gene) as previously observed (Robinson and Speed, 2007).

4.3.4 Splicing factors associated with cancer-associated alternative splicing events

SF2/ASF (SFRS1) protein overexpression was already described for several cancers, including lung (Karni et al., 2007) and colon cancers (Ghigna et al., 2005). Recently, Sanford et al. (2009) used CLIP-seq to identify *cis*-acting RNA elements recognized by SF2/ASF. Using this CLIP-seq data we found SF2/ASF target regions in 61 (5%) misspliced exons for colon and 55 (10%) for lung cancers (Figure 4.5 and Annex Tables A.3.9 and A.3.10). SF2/ASF target regions were equally distributed between inclusion and exclusion alternative splicing events.

Since alternative splicing can also involve regulatory sequences located in flanking regions (Ule et al., 2006; Martinez-Contreras et al., 2006; Fagnani et al., 2007; Wang et al., 2008a), we extended our search to the flanking introns and neighboring exons from the misspliced exon. Thus, a total number of 87 (7%) misspliced exons in colon and 97 (17%) in lung cancers contained SF2/ASF binding sites in the alternative exons and flanking regions (Figure 4.5 and Annex Tables A.3.9 and A.3.10). SF2/ASF CLIP-seq targets were observed across all neighboring regions and some alternative splicing events contained targets in multiple regions. From the 29 misspliced genes shared between both cancers only six genes seem to be targeted by SF2/ASF, corresponding all to exclusion events with different binding site distribution: alternative exon (DMN, H3F3B, SLC2A1, HDLBP), downstream adjacent intron and exon (CDCA7), both neighboring exons (TPX2). The distribution of the SF2/ASF binding sites across the several regions was similar for enriched and depleted exons. However, SF2/ASF may regulate inclusion or exclusion through the binding of adjacent constitutive exons (Ghigna et al., 2005). Indeed, Sanford et al. (2009) observed an enrichment of SF2/ASF binding sites in neighboring exons of alternative cassette exons. They proposed that SF1/ASF may play a prominent role in regulating this mode of competitive exon skipping by activating downstream splice sites. Moreover, the inclusion and exclusion of an exon results from the combination and balance of multiple interactions between splicing factors and regulatory sequences present in the alternative exon and adjacent flanking regions, including introns and neighboring exons.

Similar to Sanford et al. (2009), we observed that SF2/ASF candidate targets are enriched (Ingenuity Pathway Analysis, Benjamini-Hochberg adjusted p-value < 0.05) for

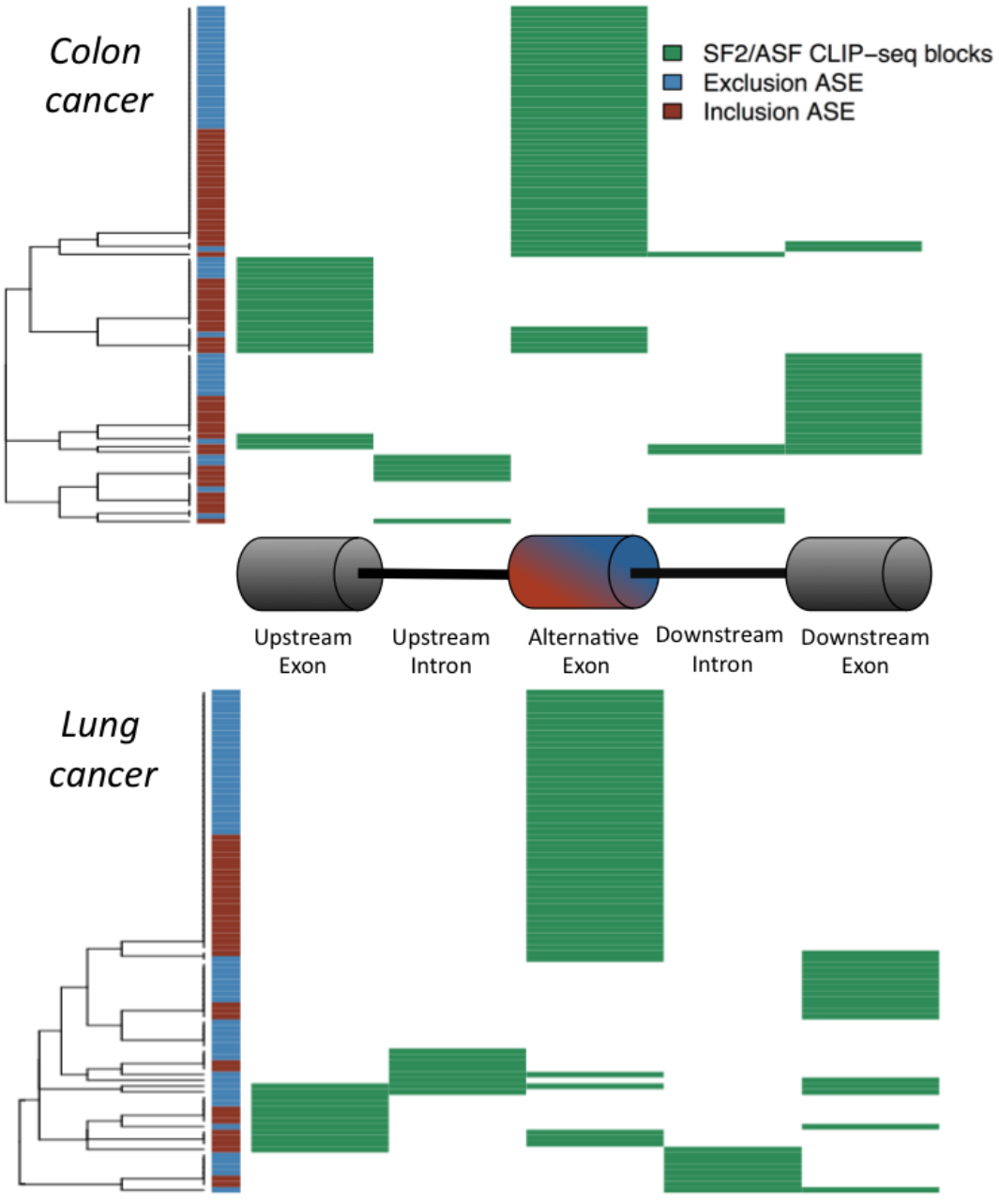


Figure 4.5: **Misspliced genes in colon and lung cancers containing CLIP-seq blocks for SF2/ASF (SFRS1).** For each alternative splicing event (ASE) is indicated the type (inclusion or exclusion) and the location of the CLIP-seq block: upstream exon, upstream intron, alternative exon, downstream intron, downstream exon. More details in Annex Tables A.3.9 and A.3.10.

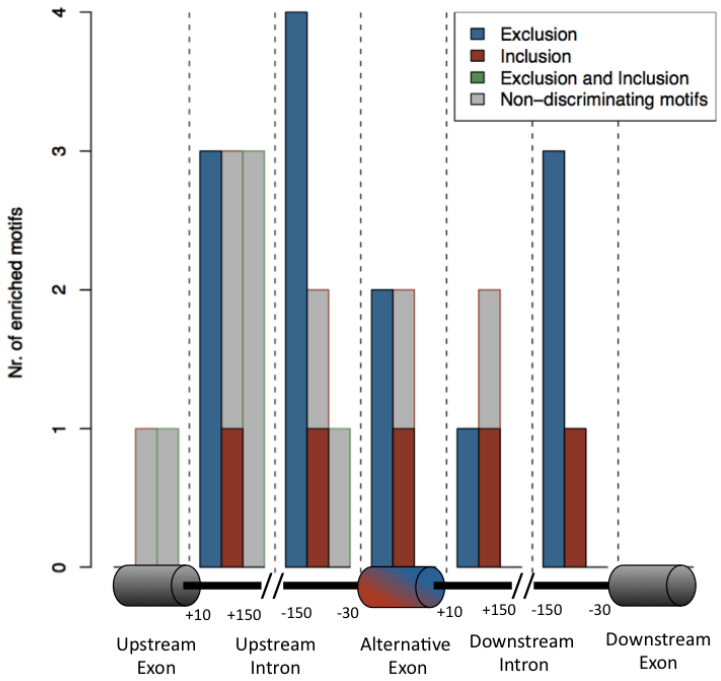
genes encoding proteins involved in gene expression (12% for lung and 15% for colon cancers) and RNA post-transcriptional modification (7% for colon cancer).

SF2/ASF CLIP-seq targets were not found for the majority of mispliced exons, but this can be explained by the fact that CLIP-seq data was obtained from embryonic kidney cells (HEK293T), not expressing a gene pool similar to colon and lung. Thus, we used the previously described consensus Position Weight Matrix (PWM) for SF2/ASF based on CLIP-seq data (Sanford et al., 2009) to find more candidate misregulated exons. Approximately 65% of the misspliced exons of each cancer contained a putative binding site based on the PWM (scores above the matching score threshold of 5.2) for SF2/ASF in the alternative exon region and the target exons increased to 99% when considering the flanking introns or neighboring exons. We did not observe a correlation between the event type and the wide distribution of putative binding sites.

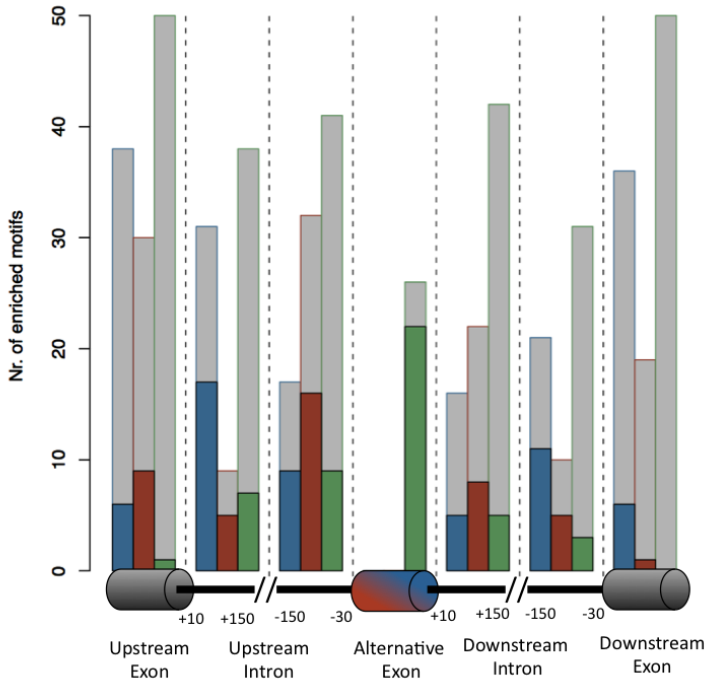
We also verified whether the cancer-associated splicing events could share more common regulatory sequences performing a motif enrichment analysis using the SeedSearcher algorithm (Barash et al., 2001). We searched for motifs that would best discriminate misspliced exons in cancer from exons with no splicing variation between cancer and normal tissue. The misspliced exons were grouped in exons enriched in cancer samples (inclusion in cancer), exons enriched in normal samples (exclusion in cancer), exons with splicing misregulated (inclusion or exclusion). Our search covered the sequences for the misspliced exon, neighboring exons and 150 nucleotides of intron sequencing flanking the exons. Only misspliced exons for which no splicing variation in the neighboring exons was detected in cancer were used, corresponding to 275 exons for colon cancer (90 excluded and 185 included) and 467 exons for lung cancer (238 excluded and 229 included). Each *ab initio* search was performed for motifs with length from five to 10 nucleotides and with various degrees of sequence flexibility as previously used by Fagnani et al. (2007).

We found 29 motifs enriched in misspliced exons for colon cancer and 559 for lung cancer (Figure 4.6 and Annex Tables A.3.11 and A.3.12). The difference between the two cancers could be related with the number of alternative splicing events used. Some enriched motifs were common to several regions. Using Fishers' exact test (Fisher, 1935) we selected motifs with different frequencies (p-value < 0.05) between included and excluded exons in cancer: 18 motifs for colon cancer and 145 for lung cancer (Figure 4.6 and Annex Tables A.3.11 and A.3.12). Some differentially represented motifs were only enriched when using all misspliced exons independently of event type (inclusion or exclusion). Also, some motifs found using the set of all misspliced exons appeared enriched when using only the subset of exons from the same event type.

The number of motifs was slightly higher for colon cancer in the flanking introns regions than in misspliced exon and neighboring exons regions. For lung cancer the highest number



(a) Colon Cancer



(b) Lung Cancer

Figure 4.6: **Number of motifs enriched in cancer-associated alternative slicing events.** Number of enriched motifs found for each region of misspliced exons grouped according to event type: exons enriched in cancer samples (inclusion), exons enriched in normal samples (exclusion), exons with splicing misregulated (inclusion and exclusion). Gray bars indicate the enriched motifs with non-discriminative frequencies (Fisher's test p-value > 0.05) between exlcusion and inclusion alternative splicing events.

of total motifs was found in the neighboring exons, whereas the discriminating motifs were found in more abundance in the flanking introns and misspliced exon. Interestingly, the uspstream intron regions presented higher and differential number of discriminating motifs, with more motifs in the $5'$ region for exclusive events and in the $3'$ region for inclusive events. Moreover, all discriminate motifs found in the alternative exon region revealed higher frequencies in inclusive events (Figure 4.6).

We compared the enriched motifs against a list comprising previously reported regulatory sequences associated with splicing (Cartegni et al., 2003; Martinez-Contreras et al., 2007; Gabut et al., 2008; Long and Caceres, 2009). We identified some motifs that contained or were part of described motifs and could resemble binding sites for splicing factors. However the frequency of the original binding site sequence was similar in depleted and enriched exons. Nevertheless, we could identify some cancer-associated splice variants which can be potentially regulated by splicing factors with gene expression affected in the same cancer.

Alternative exons excluded in colon cancer were enriched with the motif "GGGNCNCC" and the frequency of the motif in sequences of excluded events (23%) was significantly higher than in sequences of inclusion (1%) and control (8%). This motif contains part of the motif described for HNRNPF, HNRNPH1, HNRNPH2 and HNRNPH3 (GGGA, GGGG). Indeed, using a stricter motif "GGGRNCNCC" (R denotes A or G) the frequencies in excluded events (13%) were also higher than in included events (1%). While no misregulation was found for these genes in our microarray analysis, previous studies showed overexpression of HNRNPF at protein level in colon cancer (Balasubramani et al., 2006). These results suggest that the up-regulation of HNRNPF in colon cancer can be responsible for missplicing of some exons.

For lung cancer, the upstream intron $3'$ sequence was enriched with the motif "ACNNAGG" and the frequency of the motif in sequences of excluded events (18%) was significantly higher than in sequences of inclusion (8%) and control (10%). Two motifs for antagonistic factors match the last part of the motif for SFRS5 ("ACDGS", D denotes A, T or G; S denotes G or C) and HNRPA1 and HNRPA2B1 (TAGG). Our microarray analysis detected SFRS5 down-regulation in exon arrays and HNRPA2B1 up-regulation in $3'$ arrays. Moreover, overexpression of HNRPA1 and HNRPA2B1 at protein level was previously described (Patry et al., 2003; Zhou et al., 1996). Since the SR (ACDGS) and hnRNP (TAGG) sequences do not resemble each other, this suggest two different regulation systems for exon exclusion by absence of SFRS5 and overepxression of HNRPA1 or HNRPA2B1. However, the frequencies of these two motifs in the excluded events was slightly higher (43% for SR and 25% for hnRNP motif) than for the included sequences (38% for SR and 22% for hnRNP).

The splicing factor binding site sequences are often degenerate and lack sufficient specificity to reveal the global organization of protein-RNA interactions. CLIP-seq or splicing factor depletion approaches allow us to restrict targets containing a functional binding site from a non-functional one. Sanford et al. (2009) observed that statistically significant SF2/ASF binding sites (with scores above the matching score threshold for the consensus PWM) could not be found for some of the SF2/ASF CLIP-seq blocks. Recently, Venables et al. (2009) also observed that after FOX2 depletion only 87 out of 810 cassette exons, containing at least one nearby FOX2 binding site, presented a shift in splicing by more than 10% on average in ovarian and breast cancer cell lines. Only combining information from more specific assays (CLIP-seq, splicing factor depletion, etc) for all splicing factors will be possible to understand the alternative splicing code that controls and coordinates the transcriptome in cancer.

4.4 Conclusion

Here we collected a high-confidence set of misregulated splicing events in cancer by applying a comprehensive workflow analysis to Affymetrix Exon Microarray data sets from previous studies for colon (Gardina et al., 2006) and lung cancers (Xi et al., 2008).

Our approach revealed a different number of misspliced genes from the original studies. We increased the number of misspliced genes previously identified for colon cancer, whereas for lung cancer results our approach revealed to be more conservative, selecting less genes. Nevertheless, the comparison of our results with previous validations by RT-PCR for these cancers (Gardina et al., 2006; Thorsen et al., 2008; Xi et al., 2008) showed less false positive events, corresponding to a higher precision of our methodology.

Functional analysis revealed that 30% of the misregulated genes at transcript and alternative splicing levels and cancer types were previously associated with cancer disorder. We found different enriched pathways when using distinct sets of genes split according to misregulation level. The results may suggest that some pathways are more affected by variations in transcript abundance, whereas others at alternative splicing level. Most of the misspliced genes encoded for proteins associated with cell communication and motility, signal transduction and signaling molecules and interaction. Our results highlight the importance of including information of the several layers of gene expression regulation on functional and pathway analysis. Although differences in transcript abundance have routinely been used for gene expression profiling, it is clear that both the amount and the sequence diversity of transcripts have a high impact in cancer.

We have also identified cancer-associated splice variants which are likely to be regulated by splicing factors with gene expression affected in the same cancer. Overexpression of the

splicing factor SF2/ASF (SFRS1) was recently observed in various human tumours and shown to affect alternative splicing of both tumour suppressor genes and oncogenes (Karni et al., 2007). Using SF2/ASF CLIP-seq data previously published (Sanford et al., 2009) we found SF2/ASF binding sites for 87 (7%) misspliced exons in colon and 97 (17%) in lung cancers, located either in the alternative exon, flaking introns or neighboring exons. However, an association between alternative splicing decision and location of SF2/ASF binding sites could not be established. The cancer-associated splicing events also contained enriched motifs that match predicted binding sites for cancer misregulated splicing factors, such as HNRNPF, SFRS5, HNRPA1 and HNRPA2B1. However, identification of functional *cis*-acting RNA elements on a global scale is required to validate bioinformatic predictions.

Our finding that specific pathways are commonly affected by splicing misregulation in distinct cancers strengthens the view that splicing plays an important role in oncogenesis. In some cases, the altered splicing events correlate with abnormal levels of regulatory splicing factors that are expressed in the same tumours.

The rapid development and increasing availability of novel genome-wide tools will soon provide a catalogue of all splicing factors and all splice variants that are differentially expressed in specific cancer types and the corresponding normal tissues. Future studies will have to combine transcriptomic data with complementary approaches (namely, Chip/CLIP-seq and proteomics) to to improve our understanding of how gene expression regulatory networks are altered in cancer.

Chapter 5

Final Remarks and Future Perspectives

The present study aimed to generate microarray-based predictions for understanding the alternative splicing code that controls and coordinates the transcriptome.

We systematically assessed the widespread gene expression of splicing regulators during cell differentiation, in differentiated tissues and in cancer. By using large-scale data analysis we revealed new differential expression of several splicing-related genes, which encoded proteins may modulate cell type or tissue specific alternative splicing. Moreover, our work provided more evidences for the link between changes in splicing factors expression and alternative splicing profiles. First, our splicing factor signatures for tissues correlated with tissue-specific splicing events. Second, we identified cancer-associated splice variants which seem to be regulated by splicing factors with gene expression affected in the same cancer.

These results indicate the power of using microarray technology and computational approaches to generate initial predictions for a global view of alternative splicing regulation.

Furthermore, the large number of splicing-related genes with differential expression found in the present study raises the question of whether changes in the expression level of splicing factors regulate specific alternative splicing events that play key roles in cell differentiation and cancer. The present work has made important and original scientific contributions to solve these relevant but still open questions.

Which splicing factors may influence alternative splicing patterns in a highly specific manner?

The current hypothesis for differential alternative splicing regulation between tissues or development stages suggests that differences in relative abundances or activities of multiple proteins influence specific splicing decisions (Hanamura et al., 1998; Singh and Valcárcel, 2005). Moreover, changes in the relative expression or cellular distribution of antagonizing factors could establish a combinatorial code (Singh and Valcárcel, 2005).

Previous gene-by-gene studies addressed the differential specificity of splicing factors, focusing mainly on SR and hnRNP proteins (reviewed in Singh and Valcárcel, 2005; Long and Caceres, 2009). However, additional studies described other splicing factors inducing specific splicing (reviewed in Singh and Valcárcel, 2005), for example, ELAVL proteins in mouse brain (Lisbin et al., 2001; McKee et al., 2005), A2bp1 (also known as FOX1) in heart and brain (Shibata et al., 2000; Kiehl et al., 2001; Jin et al., 2003b), NOVA1 in mouse brain (Ule et al., 2005).

Here, we systematically assessed the widespread expression of genes encoding several splicing-related proteins. Using large-scale data we monitored their gene expression variations on different stages of myotube, adipocyte, erythroid and sperm cell differentiation and also on tissues derived from human, chimpanzee and mouse. We identified over 100 splicing-related genes that are most highly differentially expressed in a particular tissue or differentiation process. Our results showed that all genes of the main splicing factor families present differential gene expression including SR protein kinases and snRNP proteins. These results extended the list of splicing-related genes with differential gene expression, corresponding to putative regulators for cell type or tissue specific alternative splicing.

Clearly, a major task for the future will be to determine whether tissue-specific alternative splicing events are regulated by the differential expression of the genes identified in our study.

In the present work we also showed that splicing factor signatures correlate with tissue-specific alternative splicing patterns. The largest number of tissue-specific splicing factor genes was found for brain and testis, the two tissues for which highest levels of alternative splicing events were previously found based on genome-wide studies using ESTs (Yeo et al., 2004) and splicing microarrays (Pan et al., 2004; Clark et al., 2007).

Splicing microarrays have been recently developed and used for genome-wide analysis of alternative splicing. Combining splicing microarrays and computational analysis sequence motifs were found that resemble splicing factor binding sites and correlate with tissue-specific alternative splicing in mouse brain and muscle (Sugnet et al., 2006; Fagnani et al., 2007) and human muscle (Das et al., 2007). In addition, several studies combining siRNA-mediated knockdowns or conventional knockouts of splicing factors and splic-

ing microarray analyses revealed alternative exons regulated by specific splicing factors: dASF/SF2, B52/SRp55, hrp48, and PSI in fruit fly (Blanchette et al., 2005); NOVA proteins in brain mouse (Ule et al., 2005); PTB and nPTB in mouse neuronal differentiation (Boutz et al., 2007) hnRNP L in mammals (Hung et al., 2008).

Furthermore, recent studies combining the annotation of binding sites with alternative splicing patterns from splicing microarrays or high-throughput sequencing identified RNA maps for the splicing proteins NOVA (Ule et al., 2006; Licatalosi et al., 2008) and FOX (Zhang et al., 2008; Yeo et al., 2009). These RNA maps define the regulatory networks of alternative splicing and can be used to predict the outcome of alternative splicing in other genes.

Indeed, an important goal is to understand the splicing code and generate predictions for cell-type and tissue-specific splicing patterns. To achieve this, future studies will have to define cell-type and tissue-specific splicing regulatory motifs and how they function in conjunction with each other as well as with the more generally used enhancer and silencer motifs.

Which splicing factors may lead to splicing disruption in cancer?

Alternative splicing is associated with multiple human diseases including cancer (reviewed in Wang and Cooper, 2007; Cooper et al., 2009). Several mutations are known that affect the splicing of oncogenes, tumour suppressors and other cancer-relevant genes (Srebrow and Kornblihtt, 2006; Venables, 2006), however, many splicing abnormalities identified in cancer cells are not associated with mutations in the affected genes. Rather, a growing body of evidence indicates that the splicing machinery is a major target for misregulation in cancer.

Punctual cases of cancer misregulated splicing factors and connection to splicing disruption in cancer were previously described for SR, hnRNP and other regulators (see Section 3.4.1). Interestingly, it was recently shown for the first time that over-expression of a splicing factor can indeed trigger malignant transformation (Karni et al., 2007). The authors showed that the splicing factor SF2/ASF (SFRS1) is up-regulated in various human tumours and affects alternative splicing of tumour suppressor and oncogenic genes (Karni et al., 2007). This study serves as a proof-of-principle and shows that abnormally expressed splicing proteins can have oncogenic properties.

Recent bioinformatics studies also suggest that splicing factors are not expressed at proper levels and/or their functions are impaired in cancer (Kim et al., 2008; Ritchie et al., 2008).

Here, we performed a large-scale analysis of expression profiles for several splicing factors in 13 cancer types: bladder, brain, breast, colon, esophagus, head and neck, kidney,

liver, lung neuroblastoma, prostate, thyroid and vulva. We have identified 192 splicing-related genes that are differentially expressed in specific cancer types and the corresponding normal tissues. These genes encoded the major splicing protein families snRNPs, hnRNPs, SRs, SR-kinases, RNA-helicases-like and other splicing regulators. We also observed that the majority of differentially expressed splicing regulators were up-regulated in cancer and some misregulations appear consistently in several cancer types.

Consistent gene expression variations of splicing factors in several cancer was also showed for SF2/ASF (SFRS1) for various human tumours (Karni et al., 2007). Thus suggesting that some cancers can present common misregulated splicing factors.

Our results are also consistent with other analysis using on serial analysis of gene expression (SAGE) and Oncomine (microarray based information) that observed differential expression for splicing factors in four cancer types (breast, colon, prostate and brain) and over-expression was a general trend (Kirschbaum-Slager et al., 2004).

Furthermore, we showed a correlation between changes in splicing factor expression and splicing events in the same cancer. We collected a high-confidence set of misregulated splicing events in cancer by applying a comprehensive workflow analysis to splicing microarray data sets from previous studies for colon and lung cancers. We identified misspliced exons containing *cis*-acting RNA elements obtained from CLIP-seq data (Sanford et al., 2009) for SF2/ASF, which is overexpressed in both cancers. We were also able to identify motifs enriched in the cancer-associated splicing events that resemble binding sites for cancer misregulated splicing factors.

A large number of cancer-associated alternative splicing events have been reported in several studies using high-throughput technologies (Relógio et al., 2005; Gardina et al., 2006; Thorsen et al., 2008; Zhang et al., 2006; Thorsen et al., 2008; French et al., 2007; Cheung et al., 2008; Xi et al., 2008; Soreq et al., 2008; Thorsen et al., 2008). However, there is scarce information on associations between misregulated splicing factors and splicing events in the same cancer. Relógio et al. (2005) showed for Hodgkin lymphoma cells that Nova2 was overexpressed and the expression was correlated with gene isoforms detected with splicing microarrays. Recently, Venables et al. (2009) using high-throughput RT-PCR identified misspliced genes in ovarian and breast cancers, from which many contained binding-sites for FOX2 protein. They also showed that FOX2 is down-regulated in ovarian cancer and misspliced in breast cancer, leading to an overall depletion of FOX2 protein and splicing modulation.

The rapid development and increasing availability of novel genome-wide tools will soon provide a catalogue of all splicing factors and all splice variants that are differentially expressed in specific cancer types and the corresponding normal tissues. Irrespective of whether changes in splicing play a direct causative role in cancer, or act as modifiers or

susceptibility factors in the oncogenic process, the identification of splicing signatures is likely to provide important markers for diagnosis, prognosis, and/or sensitivity to treatment (see Section 3.4.2 Splicing factors and anticancer therapy). A full description of all components of the splicing machinery and splicing events altered in cancer will also identify potential new targets for therapeutic approaches.

Role and future of high-throughput technologies in alternative splicing

The emergence and development of high-throughput technologies over the last two decades has deeply modified our way of performing biological research, moving from a gene-by-gene approach to global or genome-wide studies. These technologies revealed to have a key role for the global view of alternative splicing regulation (reviewed in Lee and Roy, 2004; Blencowe, 2006; Ben-Dov et al., 2008; Hartmann and Valcárcel, 2009)

The first genome-wide studies on alternative splicing relied on alignments of expressed sequence tags (ESTs) and cDNA sequences to the genome and described alternative splice forms in a surprisingly large fraction of human genes, ranging from 40% to 60% (Mironov et al., 1999; Brett et al., 2000; Croft et al., 2000; Lander et al., 2001; Kan et al., 2001; Modrek et al., 2001). Using this approach several large databases of alternative splicing events were developed for several species: Intronerator (Kent and Zahler, 2000), ISIS (Croft et al., 2000), TAP (Kan et al., 2001), HASDB (Modrek et al., 2001), ASAP (Lee et al., 2003a), ProSplicer (Huang et al., 2003), ASD (Thanaraj et al., 2004), FASTDB (de la Grange et al., 2005), ASPIC (Bonizzoni et al., 2005), MAASE (Zheng et al., 2005), EuSplice (Bhasi et al., 2007), ExonMine (Mollet et al., submitted). However, ESTs/cDNA analysis have some limitations (reviewed in Modrek and Lee, 2002), which have been overcome by the development of microarray technology.

Microarray technology evolution has allowed us to resolve exon-level gene expression and enabled large-scale profiling of mRNA splicing (reviewed in Blencowe, 2006). Several splicing microarray platforms were developed consisting essentially in two types of approaches: exon and exon-junction centric platforms. The two approaches present distinct advantages in their ability to measure transcript structure due the location of the probes. The exon-centric platforms are more appropriate to identify novel splicing events since probes are designed for well-annotated and predicted exons, whereas with exon-junction platforms transcript architecture directly targeting pre-determined arrangements of exons can be assessed (reviewed in McKee and Silver, 2007).

These platforms have permitted the discovery of new alternative splicing events, increasing the number of affected genes to $> 80\%$ (Johnson et al., 2003; Kampa et al., 2004). Several studies using splicing microarrays identified cell type and tissue-specific alternative splicing profiles (Clark et al., 2002; Yeakley et al., 2002; Johnson et al., 2003; Wang et al.,

2003; Le et al., 2004; Pan et al., 2004; Stolc et al., 2004; Watahiki et al., 2004; Srinivasan et al., 2005; Nagao et al., 2005; Shai et al., 2006; Sugnet et al., 2006; Clark et al., 2007; Ip et al., 2007; Hartmann et al., 2009). In addition, analysis of several mouse tissues revealed that tissue-specific mechanisms of transcription and alternative splicing operate on different subsets of genes (Pan et al., 2004).

Splicing sensitive microarrays allowed also the identification of new alternative splicing events misregulated in Hodgkin lymphoma cells (Relógio et al., 2005), colon cancer (Gardina et al., 2006; Thorsen et al., 2008), prostate cancer (Li et al., 2006; Zhang et al., 2006; Thorsen et al., 2008), brain cancer (French et al., 2007; Cheung et al., 2008), lung cancer (Xi et al., 2008) MG-thymona (Soreq et al., 2008), bladder cancer (Thorsen et al., 2008).

As referred above, correlation between changes in expression of splicing factors and specific splicing events for cell type or tissue was also detected using this technology (Blanchette et al., 2005; Ule et al., 2005; Sugnet et al., 2006; Das et al., 2007; Fagnani et al., 2007; Boutz et al., 2007; Hung et al., 2008) and for cancer (Relógio et al., 2005).

Standard gene microarrays have been also used to evaluate specific expression of splicing factors in tissues (Yeo et al., 2004) and in cancer Kirschbaum-Slager et al. (2004). Indeed, in the present work we showed how microarrays can be extensively used to systematically assess the expression levels of splicing regulators during cell differentiation, in differentiated tissues and in cancer.

Although microarray technology has been revealed to be extremely useful for biological research over the last two decades, hybridization-based approaches have some limitations: detection is limited to RNA spliced patterns and genes previously identified; high background levels due to cross-hybridization; limited dynamic range of detection due to both background and saturation of signals; comparison of expression levels across different experiments is often difficult and requires complicated normalization methods (Wang et al., 2008b).

Recently, the development of novel high-throughput DNA sequencing methods has provided a new method for both mapping and quantifying transcriptomes, termed RNA-Seq (RNA sequencing) (reviewed in Wang et al., 2008b).

Although more expensive and with some data analysis issues still to be solved, RNA-seq can identify and quantify all transcript isoforms, allowing the discovery of novel splicing patterns. The first studies with RNA-seq revealed that the number of genes alternatively spliced was 92-95% (Wang et al., 2008a; Pan et al., 2008). Recently, studies using crosslinking/immunoprecipitation followed by deep RNA sequencing (HITS-CLIP) generated RNA maps for the splicing protein NOVA and FOX (Licatalosi et al., 2008; Yeo et al., 2009), which can be used to predict the outcome of alternative splicing in other genes.

Importantly, high-throughput technologies to address others layers of gene expression regulation are also emerging. Since changes in mRNA levels for transcript isoforms and splicing factors may not necessarily reflect on protein levels and function, post-transcriptional and post-translation regulation should also be considered in these genome-wide studies. Despite the difficulty of assessing the impact of alternative splicing changes in protein structure and function, differential display and sensitive mass-spectrometry studies confirmed the detection of splice variants at the protein level in large scale (Tress et al., 2008) and revealed regulatory circuits relevant to alternative splicing (Spellman et al., 2007).

In conclusion, high-throughput technologies allowing genome-wide analyses are likely to become standard tools for addressing functional, mechanistic, medical, or evolutionary questions in gene function and alternative splicing. However, the most challenging goal for the future will be to integrate the different layers of gene expression regulation to acquire a systems biology view of the multiple molecular mechanisms that might be important for cells, tissues and disease diagnosis and treatment. Therefore, future studies will require the development of new computational approaches to explore large-scale data obtained by the combinations of several technologies (e.g. cross-linking, immunoprecipitation, splicing-sensitive microarrays, tiling microarray, RNA-seq and mass-spectrometry).

Bibliography

Akerblad, P., Maansson, R., Lagergren, A., Westerlund, S., Basta, B., Lind, U., Thelin, A., Gisler, R., Liberg, D., Nelander, S., *et al.*, 2005. Gene expression analysis suggests that EBF-1 and PPARgamma2 induce adipogenesis of NIH-3T3 cells with similar efficiency and kinetics. *Physiol Genomics*, **23**(2):206–16.

Al-Shahrour, F., Minguez, P., Tárraga, J., Montaner, D., Alloza, E., Vaquerizas, J., Conde, L., Blaschke, C., Vera, J., and Dopazo, J., *et al.*, 2006. Babelomics: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res*, **34**(0):W472–6.

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D., 1990. Basic local alignment search tool. *J Mol Biol*, **215**(3):403–10.

Bairoch, A., Apweiler, R., Wu, C., Barker, W., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., *et al.*, 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res*, **33**(0):D154–9.

Balasubramani, M., Day, B., Schoen, R., and Getzenberg, R., 2006. Altered expression and localization of creatine kinase B, heterogeneous nuclear ribonucleoprotein F, and high mobility group box 1 protein in the nuclear matrix associated with colon cancer. *Cancer Res*, **66**(2):763–9.

Barash, Y., Bejerano, G., and Friedman, N., 2001. A simple hyper-geometric approach for discovering putative transcription factor binding sites. *Lecture Notes Computer Sci*, **2149**:278–293.

Barbosa-Morais, N., Carmo-Fonseca, M., and Aparício, S., 2006. Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res*, **16**(1):66–77.

Barrett, T., Suzek, T., Troup, D., Wilhite, S., Ngau, W., Ledoux, P., Rudnev, D., Lash, A., Fujibuchi, W., and Edgar, R., *et al.*, 2005. NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Res*, **33**(0):D562–6.

Beissbarth, T. and Speed, T., 2004. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**(9):1464–5.

Bellavia, D., Mecarozzi, M., Campese, A., Grazioli, P., Talora, C., Frati, L., Gulino, A., and Screpanti, I., 2007. Notch3 and the Notch3-upregulated RNA-binding protein HuD regulate ikaros alternative splicing. *EMBO J*, **26**(6):1670–80.

Bemmo, A., Benovoy, D., Kwan, T., Gaffney, J., and Majewski, J., 2008. Gene expression and isoform variation analysis using Affymetrix Exon arrays. *BMC Genomics*, **9**(1):529.

Ben-Dov, C., Hartmann, B., Lundgren, J., and Valcárcel, J., 2008. Genome-wide analysis of alternative pre-mRNA splicing. *J Biol Chem*, **283**(3):1229–33.

Bengtsson, H., Simpson, K., Bullard, J., and Hansen, K., 2008. aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. Technical Report 745, Department of Statistics, University of California.

Benjamini, Y. and Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**:289–300.

Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., and Wheeler, D., 2007. GenBank. *Nucleic Acids Res*, **35**(0):D21–5.

Bhasi, A., Pandey, R., Utharasamy, S., and Senapathy, P., 2007. Eusplice: a unified resource for the analysis of splice signals and alternative splicing in eukaryotic genes. *Bioinformatics*, **23**(14):1815–23.

Bier, F., von Nickisch-Rosenegk, M., Ehrentreich-Förster, E., Reiss, E., Henkel, J., Strehlow, R., and Andresen, D., 2008. DNA microarrays. *Adv Biochem Eng Biotechnol*, **109**(0):433–53.

Black, D., 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, **72**(0):291–336.

Blanchette, M., Green, R., Brenner, S., and Rio, D., 2005. Global analysis of positive and negative pre-mRNA splicing regulators in Drosophila. *Genes Dev*, **19**(11):1306–14.

Blencowe, B., 2006. Alternative splicing: new insights from global analyses. *Cell*, **126**(1):37–47.

Boeger, H., Bushnell, D., Davis, R., Griesenbeck, J., Lorch, Y., Strattan, J., Westover, K., and Kornberg, R., 2005. Structural basis of eukaryotic gene transcription. *FEBS Lett*, **579**(4):899–903.

Bonizzoni, P., Rizzi, R., and Pesole, G., 2005. ASPIC: a novel method to predict the exon-intron structure of a gene that is optimally compatible to a set of transcript sequences. *BMC Bioinformatics*, **6**(0):244.

Bonny, C., Cooker, L., and Goldberg, E., 1998. Deoxyribonucleic acid-protein interactions and expression of the human testis-specific lactate dehydrogenase promoter: transcription factor Sp1 plays a major role. *Biol Reprod*, **58**(3):754–9.

Boutz, P., Stoilov, P., Li, Q., Lin, C., Chawla, G., Ostrow, K., Shiue, L., Ares, B., and DL, 2007. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev*, **21**(13):1636–52.

Boyault, S., Rickman, D., de Reyniès, A., Balabaud, C., Rebouissou, S., Jeannot, E., Hérault, A., Saric, J., Belghiti, J., Franco, D., *et al.*, 2007. Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. *Hepatology*, **45**(1):42–52.

Brack, S., Silacci, M., Birchler, M., and Neri, D., 2006. Tumor-targeting properties of novel antibodies specific to the large isoform of tenascin-c. *Clin Cancer Res*, **12**(10):3200–8.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C., Causton, H., *et al.*, 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, **29**(4):365–71.

Breitkreutz, B., Jorgensen, P., Breitkreutz, A., and Tyers, M., 2001. AFM 4.0: a toolbox for DNA microarray analysis. *Genome Biol*, **2**(8):SOFTWARE0001.

Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbrück, S., Krueger, S., Reich, J., and Bork, P., 2000. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett*, **474**(1):83–6.

Burton, G., Nagarajan, R., Peterson, C., and McGehee, 2004. Microarray analysis of differentiation-specific gene expression during 3T3-L1 adipogenesis. *Gene*, **329**(0):167–85.

Busà, R., Paronetto, M., Farini, D., Pierantozzi, E., Botti, F., Angelini, D., Attisani, F., Vespasiani, G., and Sette, C., 2007. The RNA-binding protein Sam68 contributes to proliferation and survival of human prostate cancer cells. *Oncogene*, **26**(30):4372–82.

Carpenter, B., McKay, M., Dundas, S., Lawrie, L., Telfer, C., and Murray, G., 2006. Heterogeneous nuclear ribonucleoprotein K is over expressed, aberrantly localised and is associated with poor prognosis in colorectal cancer. *Br J Cancer*, **95**(7):921–7.

Cartegni, L., Wang, J., Zhu, Z., Zhang, M., and Krainer, A., 2003. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res*, **31**(13):3568–71.

Carter, N., 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet*, **39**(7):S16–21.

Causton, H., Quackenbush, J., and Brazma, A., 2003. *A Beginners guide: Microarray Gene Expression Data Analysis*. Blackwell Publishing, UK.

Chen, Y., Moore, R., Ge, H., Young, M., Lee, T., and Stevens, S., 2007. Proteomic analysis of in vivo-assembled pre-mRNA splicing complexes expands the catalog of participating factors. *Nucleic Acids Res*, **35**(12):3928–44.

Cheng, C. and Sharp, P., 2006. Regulation of CD44 alternative splicing by SRm160 and its potential role in tumor cell invasion. *Mol Cell Biol*, **26**(1):362–70.

Cheung, H., Baggerly, K., Tsavachidis, S., Bachinski, L., Neubauer, V., Nixon, T., Aldape, K., Cote, G., and Krahe, R., 2008. Global analysis of aberrant pre-mRNA splicing in glioblastoma using exon expression arrays. *BMC Genomics*, **9**(0):216.

Clark, T., Schweitzer, A., Chen, T., Staples, M., Lu, G., Wang, H., Williams, A., and Blume, J., 2007. Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol*, **8**(4):R64.

Clark, T., Sugnet, C., and Ares, 2002. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science*, **296**(5569):907–10.

Cole, C. and Scarcelli, J., 2006. Transport of messenger RNA from the nucleus to the cytoplasm. *Curr Opin Cell Biol*, **18**(3):299–306.

Collesi, C., Santoro, M., Gaudino, G., and Comoglio, P., 1996. A splicing variant of the RON transcript induces constitutive tyrosine kinase activity and an invasive phenotype. *Mol Cell Biol*, **16**(10):5518–26.

Colwill, K., Pawson, T., Andrews, B., Prasad, J., Manley, J., Bell, J., and Duncan, P., 1996. The Clk/Sty protein kinase phosphorylates SR splicing factors and regulates their intranuclear distribution. *EMBO J*, **15**(2):265–75.

Conboy, C., Spyrou, C., Thorne, N., Wade, E., Barbosa-Morais, N., Wilson, M., Bhattacharjee, A., Young, R., Tavaré, S., Lees, J., *et al.*, 2007. Cell cycle genes are the evolutionarily conserved targets of the E2F4 transcription factor. *PLoS ONE*, **2**(10):e1061.

Cooper, T., Wan, L., and Dreyfuss, G., 2009. RNA and disease. *Cell*, **136**(4):777–93.

Corsini, L., Bonnal, S., Bonna, 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**(7):620–9.

Croft, L., Schandorff, S., Clark, F., Burrage, K., Arctander, P., and Mattick, J., 2000. Isis, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat Genet*, **24**(4):340–1.

Daggett, V. and Fersht, A., 2003. The present view of the mechanism of protein folding. *Nat Rev Mol Cell Biol*, **4**(6):497–502.

Dancey, J. and Chen, H., 2006. Strategies for optimizing combinations of molecularly targeted anticancer agents. *Nat Rev Drug Discov*, **5**(8):649–59.

Das, D., Clark, T., Schweitzer, A., Yamamoto, M., Marr, H., Arribere, J., Minovitsky, S., Poliakov, A., Dubchak, I., Blume, J., *et al.*, 2007. A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing. *Nucleic Acids Res*, **35**(14):4845–57.

de la Grange, P., Dutertre, M., Martin, N., and Auboeuf, D., 2005. FAST DB: a website resource for the study of the expression regulation of human gene products. *Nucleic Acids Res*, **33**(13):4276–84.

De Preter, K., Vandesompele, J., Heimann, P., Yigit, N., Beckman, S., Schramm, A., Eggert, A., Stallings, R., Benoit, Y., Renard, M., *et al.*, 2006. Human fetal neuroblast and neuroblastoma transcriptome analysis confirms neuroblast origin and highlights neuroblastoma candidate genes. *Genome Biol*, **7**(9):R84.

Denkert, C., Weichert, W., Winzer, K., Müller, B., Noske, A., Niesporek, S., Kristiansen, G., Guski, H., Dietel, M., and Hauptmann, S., *et al.*, 2004. Expression of the ELAV-like protein HuR is associated with higher tumor grade and increased cyclooxygenase-2 expression in human breast carcinoma. *Clin Cancer Res*, **10**(16):5580–6.

Dever, T., 2002. Gene-specific regulation by general translation factors. *Cell*, **108**(4):545–56.

D'haeseleer, P., 2005. How does gene expression clustering work? *Nat Biotechnol*, **23**(12):1499–501.

Ding, W., Kuntz, S., and Miller, L., 2002. A misspliced form of the cholecystokinin-b/gastrin receptor in pancreatic carcinoma: role of reduced sellular U2AF35 and a sub-optimal 3′-splicing site leading to retention of the fourth intron. *Cancer Res*, **62**(3):947–52.

Djian, P., Phillips, M., and Green, H., 1985. The activation of specific gene transcription in the adipose conversion of 3T3 cells. *J Cell Physiol*, **124**(3):554–6.

Draghici, S., Khatri, P., Tarca, A., Amin, K., Done, A., Voichita, C., Georgescu, C., and Romero, R., 2007. A systems biology approach for pathway level analysis. *Genome Res*, **17**(10):1537–45.

Drissen, R., von Lindern, M., Kolbus, A., Driegen, S., Steinlein, P., Beug, H., Grosveld, F., and Philipsen, S., 2005. The erythroid phenotype of EKLF-null mice: defects in hemoglobin metabolism and membrane stability. *Mol Cell Biol*, **25**(12):5205–14.

Durand-Dubief, M. and Ekwall, K., 2009. Chromatin immunoprecipitation using microarrays. *Methods Mol Biol*, **529**(0):279–95.

Dyrskjot, L., Kruhoffer, M., Thykjaer, T., Marcussen, N., Jensen, J., Moller, K., and Orntoft, T., 2004. Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification. *Cancer Res*, **64**(11):4040–8.

Esquela-Kerscher, A. and Slack, F., 2006. Oncomirs - microRNAs with a role in cancer. *Nat Rev Cancer*, **6**(4):259–69.

Fagnani, M., Barash, Y., Ip, J., Misquitta, C., Pan, Q., Saltzman, A., Shai, O., Lee, L., Rozenhek, A., Mohammad, N., *et al.*, 2007. Functional coordination of alternative splicing in the mammalian central nervous system. *Genome Biol*, **8**(6):R108.

Fischer, D., Noack, K., Runnebaum, I., Watermann, D., Kieback, D., Stamm, S., and Stickeler, E., 2004. Expression of splicing factors in human ovarian cancer. *Oncol Rep*, **11**(5):1085–90.

Fisher, R. A., 1935. The logic of inductive inference. *Journal of the Royal Statistical Society Series A*, **98**:39–54.

French, P., Peeters, J., Horsman, S., Duijm, E., Siccama, I., van den Bent, M., Luider, T., Kros, J., van der Spek, P., and Sillevis Smitt, P., *et al.*, 2007. Identification of

differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays. *Cancer Res*, **67**(12):5635–42.

Gabanella, F., Carissimi, C., Usiello, A., and Pellizzoni, L., 2005. The activity of the spinal muscular atrophy protein is regulated during development and cellular differentiation. *Hum Mol Genet*, **14**(23):3629–42.

Gabut, M., Chaudhry, S., and Blencowe, B., 2008. SnapShot: The splicing regulatory machinery. *Cell*, **133**(1):192.e1.

Gama-Carvalho, M., 2002. *Nuclear Compartmentalisation of Splicing Factors: Characterisation of Molecular Signals and Role in Alternative Splicing Regulation*. PhD thesis, Universidade de Lisboa.

Gama-Carvalho, M., Barbosa-Morais, N., Brodsky, A., Silver, P., 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**(11):R113.

Gardina, P., Clark, T., Shimada, B., Staples, M., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S., *et al.*, 2006. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, **7**(0):325.

Gautier, L., Cope, L., Bolstad, B., and Irizarry, R., 2004. affy–analysis of affymetrix genechip data at the probe level. *Bioinformatics*, **20**(3):307–15.

Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.*, 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**(10):R80.

Gerber, A. and Keller, W., 2001. RNA editing by base deamination: more enzymes, more targets, new mysteries. *Trends Biochem Sci*, **26**(6):376–84.

Ghigna, C., Giordano, S., Shen, H., Benvenuto, F., Castiglioni, F., Comoglio, P., Green, M., Riva, S., and Biamonti, G., 2005. Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. *Mol Cell*, **20**(6):881–90.

Graveley, B., 2000. Sorting out the complexity of SR protein functions. *RNA*, **6**(9):1197–211.

Grosso, A., Gomes, A., Barbosa-Morais, N., Caldeira, S., Thorne, N., Grech, G., von Lindern, M., and Carmo-Fonseca, M., 2008. Tissue-specific splicing factor gene expression signatures. *Nucleic Acids Res*, **36**(15):4823–32.

Gui, J., Lane, W., and Fu, X., 1994. A serine kinase regulates intracellular localization of splicing factors in the cell cycle. *Nature*, **369**(6482):678–82.

Gumz, M., Zou, H., Kreinest, P., Childs, A., Belmonte, L., LeGrand, S., Wu, K., Luxon, B., Sinha, M., Parker, A., *et al.*, 2007. Secreted frizzled-related protein 1 loss contributes to tumor phenotype of clear cell renal cell carcinoma. *Clin Cancer Res*, **13**(16):4740–9.

Gunderson, K., Kruglyak, S., Graige, M., Garcia, F., Kermani, B., Zhao, C., Che, D., Dickinson, T., Wickham, E., Bierle, J., *et al.*, 2004. Decoding randomly ordered dna arrays. *Genome Res*, **14**(5):870–7.

Hall, A., 2009. The cytoskeleton and cancer. *Cancer Metastasis Rev*, **28**(1):5–14.

Hanamura, A., Cáceres, J., Mayeda, A., Franza, K., and AR, 1998. Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. *RNA*, **4**(4):430–44.

Harn, H., Ho, L., Shyu, R., Yuan, J., Lin, F., Young, T., Liu, C., Tang, H., and Lee, W., 1996. Soluble CD44 isoforms in serum as potential markers of metastatic gastric carcinoma. *J Clin Gastroenterol*, **22**(2):107–10.

Hartmann, B., Castelo, R., Blanchette, M., Boue, S., Rio, D., and Valcárcel, J., 2009. Global analysis of alternative splicing regulation by insulin and wingless signaling in Drosophila cells. *Genome Biol*, **10**(1):R11.

Hartmann, B. and Valcárcel, J., 2009. Decrypting the genome's alternative messages. *Curr Opin Cell Biol*, **21**(3):377–86.

Hastie, T., 1992. *Statistical Models*. Wadsworth and Brooks/Cole, California.

Hastings, K. and Emerson, 1982. cDNA clone analysis of six co-regulated mRNAs encoding skeletal muscle contractile proteins. *Proc Natl Acad Sci U S A*, **79**(5):1553–7.

Hayes, G., Carrigan, P., Beck, A., and Miller, L., 2006. Targeting the RNA splicing machinery as a novel treatment strategy for pancreatic carcinoma. *Cancer Res*, **66**(7):3819–27.

Hayes, G., Carrigan, P., and Miller, L., 2007. Serine-arginine protein kinase 1 overexpression is associated with tumorigenic imbalance in mitogen-activated protein kinase pathways in breast, colonic, and pancreatic carcinomas. *Cancer Res*, **67**(5):2072–80.

He, X., Ee, P., Coon, J., and Beck, W., 2004. Alternative splicing of the multidrug resistance protein 1/ATP binding cassette transporter subfamily gene in ovarian cancer creates functional splice variants and is associated with increased expression of the splicing factors PTB and SRp20. *Clin Cancer Res*, **10**(14):4652–60.

Hong, Y., Ho, K., Eu, K., and Cheah, P., 2007. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clin Cancer Res*, **13**(4):1107–14.

Hua, Y., Tang, Z., Tu, K., Zhu, L., Li, Y., Xie, L., and Xiao, H., 2009. Identification and target prediction of miRNAs specifically expressed in rat neural tissue. *BMC Genomics*, **10**(0):214.

Huang, C., Shen, C., Wang, H., Wu, P., and Cheng, C., 2007. Increased expression of SRp40 affecting CD44 splicing is associated with the clinical outcome of lymph node metastasis in human breast cancer. *Clin Chim Acta*, **384**(1):69–74.

Huang, H., Horng, J., Lee, C., and Liu, B., 2003. Prosplicer: a database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data. *Genome Biol*, **4**(4):R29.

Hubbard, T., Aken, B., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., *et al.*, 2007. Ensembl 2007. *Nucleic Acids Res*, **35**(0):D610–7.

Hung, L., Heiner, M., Hui, J., Schreiner, S., Benes, V., and Bindereif, A., 2008. Diverse roles of hnRNP l in mammalian mRNA processing: a combined microarray and RNAi analysis. *RNA*, **14**(2):284–96.

Ip, J., Tong, A., Pan, Q., Topp, J., Blencowe, B., and Lynch, K., 2007. Global analysis of alternative splicing during T-cell activation. *RNA*, **13**(4):563–72.

Izquierdo, J., 2008. Hu antigen R (HuR) functions as an alternative pre-mRNA splicing regulator of Fas apoptosis-promoting receptor on exon definition. *J Biol Chem*, **283**(27):19077–84.

Izquierdo, J. and Valcárcel, J., 2006. A simple principle to explain the evolution of pre-mRNA splicing. *Genes Dev*, **20**(13):1679–84.

Jensen, K., Dredge, B., Stefani, G., Zhong, R., Buckanovich, R., Okano, H., Yang, Y., and Darnell, R., 2000. Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron*, **25**(2):359–71.

Jin, W., Bruno, I., Xie, T., Sanger, L., and Cote, G., 2003a. Polypyrimidine tract-binding protein down-regulates fibroblast growth factor receptor 1 alpha-exon inclusion. *Cancer Res*, **63**(19):6154–7.

Jin, W., Riley, R., Wolfinger, R., White, K., Passador-Gurgel, G., and Gibson, G., 2001. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nat Genet*, **29**(4):389–95.

Jin, Y., Suzuki, H., Maegawa, S., Endo, H., Sugano, S., Hashimoto, K., Yasuda, K., and Inoue, K., 2003b. A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *EMBO J*, **22**(4):905–12.

Johnson, J., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P., Armour, C., Santos, R., Schadt, E., Stoughton, R., and Shoemaker, D., *et al.*, 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**(5653):2141–4.

Jollife, I., 1986. *Principal Component Analysis*. Springer-Verlag, New York.

Jurica, M. and Moore, M., 2003. Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell*, **12**(1):5–14.

Kaida, D., Motoyoshi, H., Tashiro, E., Nojima, T., Hagiwara, M., Ishigami, K., Watanabe, H., Kitahara, T., Yoshida, T., Nakajima, H., *et al.*, 2007. Spliceostatin A targets SF3b and inhibits both splicing and nuclear retention of pre-mRNA. *Nat Chem Biol*, **3**(9):576–83.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., *et al.*, 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res*, **14**(3):331–42.

Kan, Z., Rouchka, E., Gish, W., and States, D., 2001. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res*, **11**(5):889–900.

Karni, R., de Stanchina, E., Lowe, S., Sinha, R., Mu, D., and Krainer, A., 2007. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nat Struct Mol Biol*, **14**(3):185–93.

Keegan, L., Gallo, A., and O'Connell, M., 2001. The many roles of an RNA editor. *Nat Rev Genet*, **2**(11):869–78.

Kent, W., 2002. BLAT–the BLAST-like alignment tool. *Genome Res*, **12**(4):656–64.

Kent, W. and Zahler, A., 2000. The intronerator: exploring introns and alternative splicing in Caenorhabditis elegans. *Nucleic Acids Res*, **28**(1):91–3.

Kerr, G., Ruskin, H., Crane, M., and Doolan, P., 2008. Techniques for clustering gene expression data. *Comput Biol Med*, **38**(3):283–93.

Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachmann, M., and Pääbo, S., 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science*, **309**(5742):1850–4.

Kiehl, T., Shibata, H., Vo, T., Huynh, D., and Pulst, S., 2001. Identification and expression of a mouse ortholog of A2BP1. *Mamm Genome*, **12**(8):595–601.

Kim, E., Goren, A., and Ast, G., 2008. Insights into the connection between cancer and alternative splicing. *Trends Genet*, **24**(1):7–10.

Kimchi, E., Posner, M., Park, J., Darga, T., Kocherginsky, M., Karrison, T., Hart, J., Smith, K., Mezhir, J., Weichselbaum, R., *et al.*, 2005. Progression of Barrett's metaplasia to adenocarcinoma is associated with the suppression of the transcriptional programs of epidermal differentiation. *Cancer Res*, **65**(8):3146–54.

Kirschbaum-Slager, N., Lopes, G., Galante, P., Riggins, G., and de Souza, S., 2004. Splicing factors are differentially expressed in tumors. *Genet Mol Res*, **3**(4):512–20.

Kornblihtt, A., de la Mata, M., Fededa, J., Munoz, M., and Nogues, G., 2004. Multiple links between transcription and splicing. *RNA*, **10**(10):1489–98.

Kotake, Y., Sagane, K., Owa, T., Mimori-Kiyosue, Y., Shimizu, H., Uesugi, M., Ishihama, Y., Iwata, M., and Mizui, Y., 2007. Splicing factor SF3b as a target of the antitumor natural product pladienolide. *Nat Chem Biol*, **3**(9):570–5.

Kuhn, R., Karolchik, D., Zweig, A., Trumbower, H., Thomas, D., Thakkapallayil, A., Sugnet, C., Stanke, M., Smith, K., Siepel, A., *et al.*, 2007. The ucsc genome browser database: update 2007. *Nucleic Acids Res*, **35**(0):D668–73.

Kulesh, D., Clive, D., Zarlenga, D., and Greene, J., 1987. Identification of interferon-modulated proliferation-related cDNA sequences. *Proc Natl Acad Sci U S A*, **84**(23):8453–7.

Kuriakose, M., Chen, W., He, Z., Sikora, A., Zhang, P., Zhang, Z., Qiu, W., Hsu, D., McMunn-Coffran, C., Brown, S., *et al.*, 2004. Selection and validation of differentially expressed genes in head and neck cancer. *Cell Mol Life Sci*, **61**(11):1372–83.

Kuyumcu-Martinez, N., Wang, G., and Cooper, T., 2007. Increased steady-state levels of CUGBP1 in myotonic dystrophy 1 are due to PKC-mediated hyperphosphorylation. *Mol Cell*, **28**(1):68–78.

Lahlil, R., Lécuyer, E., Herblot, S., and Hoang, T., 2004. SCL assembles a multifactorial complex that determines glycophorin A expression. *Mol Cell Biol*, **24**(4):1439–52.

Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.*, 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**(6822):860–921.

Le, K., Mitsouras, K., Roy, M., Wang, Q., Xu, Q., Nelson, S., and Lee, C., 2004. Detecting tissue-specific regulation of alternative splicing as a qualitative change in microarray data. *Nucleic Acids Res*, **32**(22):e180.

Lee, C., Atanelov, L., Modrek, B., and Xing, Y., 2003a. ASAP: the Alternative Splicing Annotation Project. *Nucleic Acids Res*, **31**(1):101–5.

Lee, C. and Roy, M., 2004. Analysis of alternative splicing with microarrays: successes and challenges. *Genome Biol*, **5**(7):231.

Lee, S., Harn, H., Lin, T., Yeh, K., Liu, Y., Tsai, C., and Cheng, Y., 2003b. Prognostic significance of CD44v5 expression in human thymic epithelial neoplasms. *Ann Thorac Surg*, **76**(1):213–8; discussion 218.

Lejeune, F. and Maquat, L., 2005. Mechanistic links between nonsense-mediated mRNA decay and pre-mrna splicing in mammalian cells. *Curr Opin Cell Biol*, **17**(3):309–15.

Li, C. and Wong, W., 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, **98**(1):31–6.

Li, H., Wang-Rodriguez, J., Nair, T., Yeakley, J., Kwon, Y., Bibikova, M., Zheng, C., Zhou, L., Zhang, K., Downs, T., *et al.*, 2006. Two-dimensional transcriptome profiling: identification of messenger RNA isoform signatures in prostate cancer from archived paraffin-embedded cancer specimens. *Cancer Res*, **66**(8):4079–88.

Li, Q., Lee, J., and Black, D., 2007. Neuronal regulation of alternative pre-mRNA splicing. *Nat Rev Neurosci*, **8**(11):819–31.

Li, W. and Ruan, K., 2009. MicroRNA detection by microarray. *Anal Bioanal Chem*, **394**(4):1117–24.

Licatalosi, D., Mele, A., Fak, J., Ule, J., Kayikci, M., Chi, S., Clark, T., Schweitzer, A., Blume, J., Wang, X., *et al.*, 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**(7221):464–9.

Linder, P., 2006. Dead-box proteins: a family affair–active and passive players in RNP-remodeling. *Nucleic Acids Res*, **34**(15):4168–80.

Lisbin, M., Qiu, J., and White, K., 2001. The neuron-specific RNA-binding protein elav regulates neuroglian alternative splicing in neurons and binds directly to its pre-mRNA. *Genes Dev*, **15**(19):2546–61.

Long, J. and Caceres, J., 2009. The SR protein family of splicing factors: master regulators of gene expression. *Biochem J*, **417**(1):15–27.

MacLennan, D., Duff, C., Zorzato, F., Fujii, J., Phillips, M., Korneluk, R., Frodis, W., Britt, B., and Worton, R., 1990. Ryanodine receptor gene is a candidate for predisposition to malignant hyperthermia. *Nature*, **343**(6258):559–61.

Makeyev, E., Zhang, J., Carrasco, M., and Maniatis, T., 2007. The microRNA mir-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol Cell*, **27**(3):435–48.

Mangus, D., Evans, M., and Jacobson, A., 2003. Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome Biol*, **4**(7):223.

Maniatis, T. and Tasic, B., 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**(6894):236–43.

Margareto, J., Leis, O., Larrarte, E., Idoate, M., Carrasco, A., and Lafuente, J., 2007. Gene expression profiling of human gliomas reveals differences between GBM and LGA related to energy metabolism and notch signaling pathways. *J Mol Neurosci*, **32**(1):53–63.

Martín-Subero, J., Kreuz, M., Bibikova, M., Bentink, S., Ammerpohl, O., Wickham-Garcia, E., Rosolowski, M., Richter, J., Lopez-Serra, L., Ballestar, E., *et al.*, 2009. New insights into the biology and origin of mature aggressive b-cell lymphomas by combined epigenomic, genomic, and transcriptional profiling. *Blood*, **113**(11):2488–97.

Martinez-Contreras, R., Cloutier, P., Shkreta, L., Fisette, J., Revil, T., and Chabot, B., 2007. hnRNP proteins and splicing control. *Adv Exp Med Biol*, **623**(0):123–47.

Martinez-Contreras, R., Fisette, J., Nasim, F., Madden, R., Cordeau, M., and Chabot, B., 2006. Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. *PLoS Biol*, **4**(2):e21.

Massiello, A., Roesser, J., and Chalfant, C., 2006. SAP155 binds to ceramide-responsive RNA cis-element 1 and regulates the alternative 5′ splice site selection of Bcl-x pre-mRNA. *FASEB J*, **20**(10):1680–2.

Matlin, A., Clark, F., and Smith, C., 2005. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*, **6**(5):386–98.

McCarrey, J., Kumari, M., Aivaliotis, M., Wang, Z., Zhang, P., Marshall, F., and Vandeberg, J., 1996. Analysis of the cDNA and encoded protein of the human testis-specific PGK-2 gene. *Dev Genet*, **19**(4):321–32.

McKee, A., Minet, E., Stern, C., Riahi, S., Stiles, C., and Silver, P., 2005. A genome-wide in situ hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain. *BMC Dev Biol*, **5**(0):14.

McKee, A., Neretti, N., Carvalho, L., Meyer, C., Fox, E., Brodsky, A., and Silver, P., 2007. Exon expression profiling reveals stimulus-mediated exon use in neural cells. *Genome Biol*, **8**(8):R159.

McKee, A. and Silver, P., 2007. Systems perspectives on mRNA processing. *Cell Res*, **17**(7):581–90.

Mironov, A., Fickett, J., and Gelfand, M., 1999. Frequent alternative splicing of human genes. *Genome Res*, **9**(12):1288–93.

Mockler, T., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S., and Ecker, J., 2005. Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, **85**(1):1–15.

Modrek, B. and Lee, C., 2002. A genomic view of alternative splicing. *Nat Genet*, **30**(1):13–9.

Modrek, B., Resch, A., Grasso, C., and Lee, C., 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res*, **29**(13):2850–9.

Mollet, I., Barbosa-Morais, N., Andrade, J., and Carmo-Fonseca, M., 2006. Diversity of human U2AF splicing factors. *FEBS J*, **273**(21):4807–16.

Mollet, I., Ben-Dov, C., Felicio-Silva, D., Grosso A.R., Eleutrio, P., Alves, R., Staller, R., Silva, T., and Carmo-Fonseca, M., 2009. Unconstrained mining of mRNA and EST databases reveals increased alternative splicing complexity in the human transcriptome. *submitted*, .

Montaner, D., Tárraga, J., Huerta-Cepas, J., Burguet, J., Vaquerizas, J., Conde, L., Minguez, P., Vera, J., Mukherjee, S., Valls, J., *et al.*, 2006. Next station in microarray data analysis: GEPAS. *Nucleic Acids Res*, **34**(0):W486–91.

Nagao, K., Togawa, N., Fujii, K., Uchikawa, H., Kohno, Y., Yamada, M., and Miyashita, T., 2005. Detecting tissue-specific alternative splicing and disease-associated aberrant splicing of the PTCH gene with exon junction microarrays. *Hum Mol Genet*, **14**(22):3379–88.

Nilsen, T., 2003. The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, **25**(12):1147–9.

Oh, J., Razfar, A., Delgado, I., Reed, R., Malkina, A., Boctor, B., and Slamon, D., 2006. 3p21.3 tumor suppressor gene H37/Luca15/RBM5 inhibits growth of human lung cancer cells through cell cycle arrest and apoptosis. *Cancer Res*, **66**(7):3419–27.

Orphanides, G. and Reinberg, D., 2002. A unified theory of gene expression. *Cell*, **108**(4):439–51.

Pacheco, T., Coelho, M., Desterro, J., Mollet, I., and Carmo-Fonseca, M., 2006a. In vivo requirement of the small subunit of U2AF for recognition of a weak 3′ splice site. *Mol Cell Biol*, **26**(21):8183–90.

Pacheco, T., Moita, L., Gomes, A., Hacohen, N., and Carmo-Fonseca, M., 2006b. RNA interference knockdown of hU2AF35 impairs cell cycle progression and modulates alternative splicing of Cdc25 transcripts. *Mol Biol Cell*, **17**(10):4187–99.

Pajares, M., Ezponda, T., Catena, R., Calvo, A., Pio, R., and Montuenga, L., 2007. Alternative splicing: an emerging topic in molecular and clinical oncology. *Lancet Oncol*, **8**(4):349–57.

Palmer, R., Barbosa-Morais, N., Gooding, E., Muralidhar, B., Thornton, C., Pett, M., Roberts, I., Schneider, D., Thorne, N., Tavaré, S., *et al.*, 2008. Pediatric malignant germ cell tumors show characteristic transcriptome profiles. *Cancer Res*, **68**(11):4239–47.

Pan, Q., Shai, O., Lee, L., Frey, B., and Blencowe, B., 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, **40**(12):1423–5.

Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A., Mohammad, N., Babak, T., Siu, H., Hughes, T., Morris, Q., *et al.*, 2004. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol Cell*, **16**(6):929–41.

Park, J., Parisky, K., Celotto, A., Reenan, R., and Graveley, B., 2004. Identification of alternative splicing regulators by RNA interference in Drosophila. *Proc Natl Acad Sci U S A*, **101**(45):15974–9.

Patel, A. and Steitz, J., 2003. Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol*, **4**(12):960–70.

Patry, C., Bouchard, L., Labrecque, P., Gendron, D., Lemieux, B., Toutant, J., Lapointe, E., Wellinger, R., and Chabot, B., 2003. Small interfering RNA-mediated reduction in heterogeneous nuclear ribonucleoparticule A1/A2 proteins induces apoptosis in human cancer cells but not in normal mortal cell lines. *Cancer Res*, **63**(22):7679–88.

Paw, B., Davidson, A., Zhou, Y., Li, R., Pratt, S., Lee, C., Trede, N., Brownlie, A., Donovan, A., Liao, E., *et al.*, 2003. Cell-specific mitotic defect and dyserythropoiesis associated with erythroid band 3 deficiency. *Nat Genet*, **34**(1):59–64.

Prasad, J., Colwill, K., Pawson, T., and Manley, J., 1999. The protein kinase Clk/Sty directly modulates SR protein activity: both hyper- and hypophosphorylation inhibit splicing. *Mol Cell Biol*, **19**(10):6991–7000.

Prasad, J. and Manley, J., 2003. Regulation and substrate specificity of the SR protein kinase Clk/Sty. *Mol Cell Biol*, **23**(12):4139–49.

Proudfoot, N., Furger, A., and Dye, M., 2002. Integrating mRNA processing with transcription. *Cell*, **108**(4):501–12.

Pruitt, K., Tatusova, T., and Maglott, D., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **35**(0):D61–5.

Purdom, E., Simpson, K., Robinson, M., Conboy, J., Lapuk, A., and Speed, T., 2008. FIRMA: a method for detection of alternative splicing from exon array data. *Bioinformatics*, **25**(15):1707–14.

Qi, Y., He, Q., Ma, Y., Du, Y., Liu, G., Li, Y., Tsao, G., Ngai, S., and Chiu, J., 2008. Proteomic identification of malignant transformation-related proteins in esophageal squamous cell carcinoma. *J Cell Biochem*, **104**(5):1625–35.

Query, C. and Konarska, M., 2004. Suppression of multiple substrate mutations by spliceosomal prp8 alleles suggests functional correlations with ribosomal ambiguity mutants. *Mol Cell*, **14**(3):343–54.

R Development Core Team, 2009. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramakrishnan, V., 2002. Ribosome structure and the mechanism of translation. *Cell*, **108**(4):557–72.

Rappsilber, J., Ryder, U., Lamond, A., and Mann, M., 2002. Large-scale proteomic analysis of the human spliceosome. *Genome Res*, **12**(8):1231–45.

Reed, R. and Hurt, E., 2002. A conserved mRNA export machinery coupled to pre-mRNA splicing. *Cell*, **108**(4):523–31.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J., 2006. GenePattern 2.0. *Nat Genet*, **38**(5):500–1.

Relógio, A., Ben-Dov, C., Baum, M., Ruggiu, M., Gemund, C., Benes, V., Darnell, R., and Valcárcel, J., 2005. Alternative splicing microarrays reveal functional expression of neuron-specific regulators in Hodgkin lymphoma cells. *J Biol Chem*, **280**(6):4779–84.

Richardson, A., Wang, Z., De Nicolo, A., Lu, X., Brown, M., Miron, A., Liao, X., Iglehart, J., Livingston, D., and Ganesan, S., *et al.*, 2006. X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell*, **9**(2):121–32.

Ritchie, W., Granjeaud, S., Puthier, D., and Gautheret, D., 2008. Entropy measures quantify global splicing disorders in cancer. *PLoS Comput Biol*, **4**(3):e1000011.

Robinson, M. and Speed, T., 2007. A comparison of affymetrix gene expression arrays. *BMC Bioinformatics*, **8**(0):449.

Rocak, S. and Linder, P., 2004. Dead-box proteins: the driving forces behind RNA metabolism. *Nat Rev Mol Cell Biol*, **5**(3):232–41.

Roychoudhury, P. and Chaudhuri, K., 2007. Evidence for heterogeneous nuclear ribonucleoprotein K overexpression in oral squamous cell carcinoma. *Br J Cancer*, **97**(4):574–5; author reply 576.

Roychoudhury, P., Paul, R., Chowdhury, R., and Chaudhuri, K., 2007. HnRNP E2 is downregulated in human oral cancer cells and the overexpression of hnRNP E2 induces apoptosis. *Mol Carcinog*, **46**(3):198–207.

Rozen, S. and Skaletsky, H., 2000. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, chapter Primer3 on the WWW for general users and for biologist programmers, pages 365–38. Humana Pres, 1st edition.

Sampath, J., Long, P., Shepard, R., Xia, X., Devanarayan, V., Sandusky, G., Perry, D., AH, W., M, R., M, M., *et al.*, 2003. Human SPF45, a splicing factor, has limited expression in normal tissues, is overexpressed in many tumors, and can confer a multidrug-resistant phenotype to cells. *Am J Pathol*, **163**(5):1781–90.

Sanford, J., Wang, X., Mort, M., Vanduyn, N., Cooper, D., Mooney, S., Edenberg, H., and Liu, Y., 2009. Splicing factor SFRS1 recognizes a functionally diverse landscape of rna transcripts. *Genome Res*, **19**(3):381–94.

Santegoets, L., Seters, M., Helmerhorst, T., Heijmans-Antonissen, C., Hanifi-Moghaddam, P., Ewing, P., van Ijcken, W., van der Spek, P., van der Meijden, W., and Blok, L., *et al.*, 2007. HPV related VIN: highly proliferative and diminished responsiveness to extracellular signals. *Int J Cancer*, **121**(4):759–66.

Schena, M., Shalon, D., Davis, R., and Brown, P., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**(5235):467–70.

Schultz, N., Hamra, F., and Garbers, D., 2003. A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. *Proc Natl Acad Sci U S A*, **100**(21):12201–6.

Shah, S. and Pallas, J., 2009. Identifying differential exon splicing using linear models and correlation coefficients. *BMC Bioinformatics*, **10**(0):26.

Shai, O., Morris, Q., Blencowe, B., and Frey, B., 2006. Inferring global levels of alternative splicing isoforms using a generative model of microarray data. *Bioinformatics*, **22**(5):606–13.

Sharma, S. and Black, D., 2006. Maps, codes, and sequence elements: can we predict the protein output from an alternatively spliced locus? *Neuron*, **52**(4):574–6.

Shibata, H., Huynh, D., and Pulst, S., 2000. A novel protein with RNA-binding motifs interacts with ataxin-2. *Hum Mol Genet*, **9**(9):1303–13.

Shima, J., McLean, D., McCarrey, J., and Griswold, M., 2004. The murine testicular transcriptome: characterizing gene expression in the testis during the progression of spermatogenesis. *Biol Reprod*, **71**(1):319–30.

Shin, C. and Manley, J., 2004. Cell signalling and the control of pre-mRNA splicing. *Nat Rev Mol Cell Biol*, **5**(9):727–38.

Shitashige, M., Naishiro, Y., Idogawa, M., Honda, K., Ono, M., Hirohashi, S., and Yamada, T., 2007a. Involvement of splicing factor-1 in beta-catenin/T-cell factor-4-mediated gene transactivation and pre-mRNA splicing. *Gastroenterology*, **132**(3):1039–54.

Shitashige, M., Satow, R., Honda, K., Ono, M., Hirohashi, S., and Yamada, T., 2007b. Increased susceptibility of Sf1(+/-) mice to azoxymethane-induced colon tumorigenesis. *Cancer Sci*, **98**(12):1862–7.

Singh, R. and Valcárcel, J., 2005. Building specificity with nonspecific RNA-binding proteins. *Nat Struct Mol Biol*, **12**(8):645–53.

Smyth, G., 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, **3**(0):Article3.

Smyth, G. and Speed, T., 2003. Normalization of cDNA microarray data. *Methods*, **31**(4):265–73.

Smyth, G. K., 2005. Limma: linear models for microarray data. In Gentleman, R., Carey, V., Dudoit, S., and R. Irizarry, W. H., editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York.

Sonenberg, N. and Hinnebusch, A., 2009. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, **136**(4):731–45.

Soreq, L., Gilboa-Geffen, A., Berrih-Aknin, S., Lacoste, P., Darvasi, A., Soreq, E., Bergman, H., and Soreq, H., 2008. Identifying alternative hyper-splicing signatures in MG-thymoma by exon arrays. *PLoS ONE*, **3**(6):e2392.

Spellman, R., Llorian, M., and Smith, C., 2007. Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. *Mol Cell*, **27**(3):420–34.

Srebrow, A. and Kornblihtt, A., 2006. The connection between splicing and cancer. *J Cell Sci*, **119**(0):2635–41.

Srinivasan, K., Shiue, L., Hayes, J., Centers, R., Fitzwater, S., Loewen, R., Edmondson, L., Bryant, J., Smith, M., Rommelfanger, C., *et al.*, 2005. Detection and measurement of alternative splicing using splicing-sensitive microarrays. *Methods*, **37**(4):345–59.

Staal, F., van der Burg, M., Wessels, L., Barendregt, B., Baert, M., van den Burg, C., van Huffel, C., Langerak, A., van der Velden, V., Reinders, M., *et al.*, 2003. DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers. *Leukemia*, **17**(7):1324–32.

Stabenau, A., McVicker, G., Melsopp, C., Proctor, G., Clamp, M., and Birney, E., 2004. The Ensembl core software libraries. *Genome Res*, **14**(5):929–33.

Stajich, J., Block, D., Boulez, K., Brenner, S., Chervitz, S., Dagdigian, C., Fuellen, G., Gilbert, J., Korf, I., Lapp, H., *et al.*, 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, **12**(10):1611–8.

Stamm, S., 2008. Regulation of alternative splicing by reversible protein phosphorylation. *J Biol Chem*, **283**(3):1223–7.

Stekel, D., 2003. *Microarray Bioinformatics*. Cambridge University Press, UK.

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M., Rifkin, S., Hua, S., Herreman, T., Tongprasit, W., Barbano, P., *et al.*, 2004. A gene expression map for the euchromatic genome of Drosophila melanogaster. *Science*, **306**(5696):655–60.

Stransky, N., Vallot, C., Reyal, F., Bernard-Pierrot, I., de Medina, S., Segraves, R., de Rycke, Y., Elvin, P., Cassidy, A., Spraggon, C., *et al.*, 2006. Regional copy number-independent deregulation of transcription in cancer. *Nat Genet*, **38**(12):1386–96.

Su, A., Wiltshire, T., Batalov, S., Lapp, H., Ching, K., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., *et al.*, 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, **101**(16):6062–7.

Sueoka, E., Sueoka, N., Goto, Y., Matsuyama, S., Nishimura, H., Sato, M., Fujimura, S., Chiba, H., and Fujiki, H., 2001. Heterogeneous nuclear ribonucleoprotein B1 as early cancer biomarker for occult cancer of human lungs and bronchial dysplasia. *Cancer Res*, **61**(5):1896–902.

Sugnet, C., Srinivasan, K., Clark, T., O'Brien, G., Cline, M., Wang, H., Williams, A., Kulp, D., Blume, J., Haussler, D., *et al.*, 2006. Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLoS Comput Biol*, **2**(1):e4.

Sultan, M., Schulz, M., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., *et al.*, 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**(5891):956–60.

Syvänen, A., 2005. Toward genome-wide SNP genotyping. *Nat Genet*, **37 Suppl**(0):S5–10.

Taylor, J., Zhang, Q., Wyatt, J., and Dean, N., 1999. Induction of endogenous Bcl-xS through the control of Bcl-x pre-mRNA splicing by antisense oligonucleotides. *Nat Biotechnol*, **17**(11):1097–100.

Thanaraj, T., Stamm, S., Clark, F., Riethoven, J., Le Texier, V., and Muilu, J., 2004. ASD: the Alternative Splicing Database. *Nucleic Acids Res*, **32**(0):D64–9.

Thorsen, K., Sørensen, K., Brems-Eskildsen, A., Modin, C., Gaustadnes, M., Hein, A., Kruhøffer, M., Laurberg, S., Borre, M., Wang, K., *et al.*, 2008. Alternative splicing

in colon, bladder, and prostate cancer identified by exon-array analysis. *Mol Cell Proteomics*, **7**(7):1214–24.

Tomczak, K., Marinescu, V., Ramoni, M., Sanoudou, D., Montanaro, F., Han, M., Kunkel, L., Kohane, I., and Beggs, A., 2004. Expression profiling and identification of novel genes involved in myogenic differentiation. *FASEB J*, **18**(2):403–5.

Tress, M., Bodenmiller, B., Aebersold, R., and Valencia, A., 2008. Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biol*, **9**(11):R162.

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B., and Darnell, R., 2006. An RNA map predicting Nova-dependent splicing regulation. *Nature*, **444**(7119):580–6.

Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., *et al.*, 2005. Nova regulates brain-specific splicing to shape the synapse. *Nat Genet*, **37**(8):844–52.

Varambally, S., Yu, J., Laxman, B., Rhodes, D., Mehra, R., Tomlins, S., Shah, R., Chandran, U., Monzon, F., Becich, M., *et al.*, 2005. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell*, **8**(5):393–406.

Venables, J., 2006. Unbalanced alternative splicing and its significance in cancer. *Bioessays*, **28**(4):378–86.

Venables, J., Klinck, R., Koh, C., Gervais-Bird, J., Bramard, A., Inkel, L., Durand, M., Couture, S., Froehlich, U., Lapointe, E., *et al.*, 2009. Cancer-associated regulation of alternative splicing. *Nat Struct Mol Biol*, **16**(6):670–77.

Venables, J., Koh, C., Froehlich, U., Lapointe, E., Couture, S., Inkel, L., Bramard, A., Paquet, E., Watier, V., Durand, M., *et al.*, 2008. Multiple and specific mRNA processing targets for the major human hnRNP proteins. *Mol Cell Biol*, **28**(19):6033–43.

Wachi, S., Yoneda, K., and Wu, R., 2005. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, **21**(23):4205–8.

Wahl, M., Will, C., and Lührmann, R., 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell*, **136**(4):701–18.

Wang, E., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S., Schroth, G., and Burge, C., 2008a. Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**(7221):470–6.

Wang, G. and Cooper, T., 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet*, **8**(10):749–61.

Wang, H., Hubbell, E., Hu, J., Mei, G., Cline, M., Lu, G., Clark, T., Siani-Rose, M., Ares, M., Kulp, D., *et al.*, 2003. Gene structure-based splice variant deconvolution using a microarray platform. *Bioinformatics*, **19 Suppl 1**(0):i315–22.

Wang, Z. and Burge, C., 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, **14**(5):802–13.

Wang, Z., Gerstein, M., and Snyder, M., 2008b. RNA-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**(1):57–63.

Watahiki, A., Waki, K., Hayatsu, N., Shiraki, T., Kondo, S., Nakamura, M., Sasaki, D., Arakawa, T., Kawai, J., Harbers, M., *et al.*, 2004. Libraries enriched for alternatively spliced exons reveal splicing patterns in melanocytes and melanomas. *Nat Methods*, **1**(3):233–9.

Watermann, D., Tang, Y., Zur Hausen, A., Jäger, M., Stamm, S., and Stickeler, E., 2006. Splicing factor Tra2-beta1 is specifically induced in breast cancer and regulates alternative splicing of the CD44 gene. *Cancer Res*, **66**(9):4774–80.

Weinberg, R., 2007. *The biology of cancer*. Garland Science, Abingdon, UK.

Welch, J., Watts, J., Vakoc, C., Yao, Y., Wang, H., Hardison, R., Blobel, G., Chodosh, L., and Weiss, M., 2004. Global regulation of erythroid gene expression by transcription factor GATA-1. *Blood*, **104**(10):3136–47.

Wistuba, I. and Gazdar, A., 2006. Lung cancer preneoplasia. *Annu Rev Pathol*, **1**(0):331–48.

Woychik, N. and Hampsey, M., 2002. The RNA polymerase II machinery: structure illuminates function. *Cell*, **108**(4):453–63.

Wu, S., Sun, G., Hsieh, D., Chen, A., Chen, H., Chang, S., and Yu, D., 2003. Correlation of CD44v5 expression with invasiveness and prognosis in renal cell carcinoma. *J Formos Med Assoc*, **102**(4):229–33.

Xi, L., Feber, A., Gupta, V., Wu, M., Bergemann, A., Landreneau, R., Litle, V., Pennathur, A., Luketich, J., and Godfrey, T., *et al.*, 2008. Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer. *Nucleic Acids Res*, **36**(20):6535–47.

Xiao, R., Sun, Y., Ding, J., Lin, S., Rose, D., Rosenfeld, M., Fu, X., and Li, X., 2007. Splicing regulator SC35 is essential for genomic stability and cell proliferation during mammalian organogenesis. *Mol Cell Biol*, **27**(15):5393–402.

Xiao, S. and Manley, J., 1997. Phosphorylation of the ASF/SF2 RS domain affects both protein-protein and protein-RNA interactions and is necessary for splicing. *Genes Dev*, **11**(3):334–44.

Yap, Y., Lam, D., Luc, G., Zhang, X., Hernandez, D., Gras, R., Wang, E., Chiu, S., Chung, L., Lam, W., *et al.*, 2005. Conserved transcription factor binding sites of cancer markers derived from primary lung adenocarcinoma microarrays. *Nucleic Acids Res*, **33**(1):409–21.

Yeakley, J., Fan, J., Doucet, D., Luo, L., Wickham, E., Ye, Z., Chee, M., and Fu, X., 2002. Profiling alternative splicing on fiber-optic arrays. *Nat Biotechnol*, **20**(4):353–8.

Yeo, G., Coufal, N., Liang, T., Peng, G., Fu, X., and Gage, F., 2009. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol*, **16**(2):130–7.

Yeo, G., Holste, D., Kreiman, G., and Burge, C., 2004. Variation in alternative splicing across human tissues. *Genome Biol*, **5**(10):R74.

Zapala, M., Hovatta, I., Ellison, J., Wodicka, L., Del Rio, J., Tennant, R., Tynan, W., Broide, R., Helton, R., Stoveken, B., *et al.*, 2005. Adult mouse brain gene expression patterns bear an embryologic imprint. *Proc Natl Acad Sci U S A*, **102**(29):10357–62.

Zerbe, L., Pino, I., Pio, R., Cosper, P., Dwyer-Nield, L., Meyer, A., Port, J., Montuenga, L., and Malkinson, A., 2004. Relative amounts of antagonistic splicing factors, hnRNP A1 and ASF/SF2, change during neoplastic lung growth: implications for pre-mRNA processing. *Mol Carcinog*, **41**(4):187–96.

Zhang, C., Li, H., Fan, J., Wang-Rodriguez, J., Downs, T., Fu, X., and Zhang, M., 2006. Profiling alternatively spliced mRNA isoforms for prostate cancer classification. *BMC Bioinformatics*, **7**(0):202.

Zhang, C., Zhang, Z., Castle, J., Sun, S., Johnson, J., Krainer, A., and Zhang, M., 2008. Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. *Genes Dev*, **22**(18):2550–63.

Zhao, P., Caretti, G., Mitchell, S., McKeehan, W., Boskey, A., Pachman, L., Sartorelli, V., and Hoffman, E., 2006. Fgfr4 is required for effective muscle regeneration in vivo. Delineation of a MyoD-Tead2-Fgfr4 transcriptional pathway. *J Biol Chem*, **281**(1):429–38.

Zheng, C., Kwon, Y., Li, H., Zhang, K., Coutinho-Mansfield, G., Yang, C., Nair, T., Gribskov, M., and Fu, X., 2005. MAASE: an alternative splicing database designed for supporting splicing microarray applications. *RNA*, **11**(12):1767–76.

Zhou, J., Mulshine, J., Unsworth, E., Scott, F., Avis, I., Vos, M., and Treston, A., 1996. Purification and characterization of a protein that permits early detection of lung cancer. identification of heterogeneous nuclear ribonucleoprotein-A2/B1 as the antigen for monoclonal antibody 703D4. *J Biol Chem*, **271**(18):10760–6.

Zhou, Z., Licklider, L., Gygi, S., and Reed, R., 2002. Comprehensive proteomic analysis of the human spliceosome. *Nature*, **419**(6903):182–5.

Appendix

Annex Tables

Annex tables are presented only on digital support (attached cd) and they are organized according to chapters. The content of each each supplementary table is described below.

Tissue-specific splicing factor gene expression signatures

Annex Table A.1.1 Sequences of the primers used for qRT-PCR.

Annex Table A.1.2 Identification and description of the hybridizations analysed for the differentiation processes.

Annex Table A.1.3 List of human splicing-related genes and respective murine orthologues. The corresponding Affymetrix probes, Ensembl identifiers, protein family and description are also indicated.

Annex Table A.1.4 Differentially expressed splicing-related genes at T1 and T2 differentiation stages. Gene Expression fold-changes ($\log_2$) at T1 and T2 relative to T0 and respective B-statistics from Empirical Bayes are indicated. Statistical significant fold-changes are highlighted in bold.

Annex Table A.1.5 Splicing-related gene signatures and respective $\log_2$ fold-changes (differentiation process vs remaining processes).

Annex Table A.1.6 Identification and description of the hybridizations analysed for adult tissues.

Annex Table A.1.7 List of differentially expressed genes found for each tissue. The genes are identified with the name for Human and Mouse organisms and the fold-changes are in $\log_2$ (tissue vs remaining tissues). References for previous studies reporting association with splicing or tissue specificity are indicated.

Annex Table A.1.8 Tissue-specific expression signatures for splicing-related genes (human, chimpanzee and mouse microarray data). The genes are identified with the

name for Human and Mouse organisms and the fold-changes are in $\log_2$ (tissue vs remaining tissues). References for previous studies reporting association with splicing or tissue specificity are indicated.

Annex Table A.1.9 Tissue-specific expression signatures for splicing-related genes (mouse microarray data). Microarray probes, gene name and respective fold-changes in $\log_2$ (tissue vs remaining tissues) are indicated.

Cancer-specific misregulation of splicing factor gene expression

Annex Table A.2.1 List of human splicing-related genes and respective murine orthologues. The corresponding Ensembl identifiers, protein family and description are also indicated.

Annex Table A.2.2 Differentially expressed splicing-related genes comparing cancer and corresponding normal tissue. Gene Expression fold-changes ($\log_2$) are indicated.

Cancer-associated splicing misregulation

Annex Table A.3.1 Genes with expression variation at gene level for colon cancer. The fold-changes ($\log_2$) and B-values are indicated.

Annex Table A.3.2 Genes with expression variation at gene level for lung cancer. The fold-changes ($\log_2$) and B-values are indicated.

Annex Table A.3.3 Genes with expression variation at exon level for colon cancer. The fold-changes ($\log_2$) and B-values are indicated. The misspliced exon is identified according to ExonMine and Ensembl databases. For each exon information is also provided relative to presence in Ensembl transcripts, part of coding sequence (cds) and of protein domain.

Annex Table A.3.4 Genes with expression variation at exon level for lung cancer. The fold-changes ($\log_2$) and B-values are indicated. The misspliced exon is identified according to ExonMine and Ensembl databases. For each exon information is also provided relative to presence in Ensembl transcripts, part of coding sequence (cds) and of protein domain.

Annex Table A.3.5 Biological functions associated with misregulated genes in colon cancer. Top functions enriched for colon cancer using different gene sets split according to misregulation level.

Annex Table A.3.6 Biological functions associated with misregulated genes in lung cancer. Top functions enriched for lung cancer using different gene sets split according to misregulation level.

Annex Table A.3.7 Biological pathways associated with misregulated genes in colon cancer. Top pathways significantly impacted based on Pathway-Express (FDR adusted p-value < 0.05) for colon cancer using different gene sets split according to misregulation level.

Annex Table A.3.8 Biological pathways associated with misregulated genes in lung cancer. Top pathways significantly impacted based on Pathway-Express (FDR adusted p-value < 0.05) for lung cancer using different gene sets split according to misregulation level.

Annex Table A.3.9 Misspliced genes in colon cancer containing CLIP-seq blocks for SF2/ASF (SFRS1).

Annex Table A.3.10 Misspliced genes in lung cancer containing CLIP-seq blocks for SF2/ASF (SFRS1).

Annex Table A.3.11 Motifs enriched in alternative slicing events for colon cancer. Enriched motifs on each region of misspliced exons grouped with: exons enriched in cancer samples (inclusion in cancer), exons enriched in normal samples (exclusion in cancer), exons with splicing misregulated (inclusion or exclusion). P-value from Fisher test applied in the comparison of frequencies for inclusion and exclusion events is also indicated. Motifs that contained or were part of described motifs and could resemble binding sites for splicing factors are identified.

Annex Table A.3.12 Motifs enriched in alternative slicing events for lung cancer. Enriched motifs on each region of misspliced exons grouped with: exons enriched in cancer samples (inclusion in cancer), exons enriched in normal samples (exclusion in cancer), exons with splicing misregulated (inclusion or exclusion). P-value from Fisher's test applied in the comparison of frequencies for inclusion and exclusion events is also indicated. Motifs that contained or were part of described motifs and could resemble binding sites for splicing factors are identified.