FUNCTIONAL ANALYSIS OF THE U2AF\textsuperscript{35} FAMILY OF SPLICING FACTORS

JORGE MIGUEL LOPES MENDES CALHEIROS ANDRADE

DOUTORAMENTO EM CIÊNCIAS BIOMÉDICAS
ESPECIALIDADE DE CIÊNCIASFUNCIONAIS

Tese orientada pela Professora Doutora Maria Carmo-Fonseca

Todas as afirmações efectuadas no presente documento são da exclusiva responsabilidade do seu autor, não cabendo qualquer responsabilidade à Faculdade de Medicina de Lisboa pelos conteúdos nele apresentados.
A impressão desta dissertação foi aprovada pelo Conselho Científico da Faculdade de Medicina da Universidade de Lisboa em reunião de 19 de Julho de 2011.

O desenvolvimento e execução gráfica da presente dissertação foram financiados pela Fundação para a Ciência e a Tecnologia (Bolsa SFRH/BD/31012/2006).
Table of Contents

PREFÁCIO ........................................................................................................................................... v
RESUMO ............................................................................................................................................... ix
SUMMARY ........................................................................................................................................... xvii
ABBREVIATIONS .............................................................................................................................. xxi

CHAPTER 1

Introduction ................................................................................................................................. 1
1.1 The Nucleus: the hallmark of eukaryotic cells ................................................................. 2
1.2 Gene expression: a multistep process ............................................................................ 2
1.3 Transcription by RNA Polymerase II ............................................................................. 4
    1.3.1 The role of chromatin during transcription ......................................................... 6
1.4 Pre-mRNA processing ....................................................................................................... 9
    1.4.1 Capping ............................................................................................................... 9
    1.4.2 Splicing ............................................................................................................. 10
    1.4.3 3’ end processing ............................................................................................ 15
1.5 Gene expression is a highly interconnected multistep process ...................................... 17
1.6 Alternative Splicing is a major regulator of gene expression ........................................ 20
    1.6.1 The splice site strength and the role of the U2AF complex in splicing regulation 22
    1.6.2 The role of cis-acting Regulatory Elements in Splice-Site Selection ............... 27
    1.6.3 Alternative Splicing coupled to NMD ................................................................. 31
    1.6.4 Alternative Splicing and Polyadenylation ......................................................... 33
    1.6.5 Splicing and Disease ......................................................................................... 36

CHAPTER 2

2.1 Diversity of human U2AF splicing factors ..................................................................... 39
2.2 Scope of this thesis ......................................................................................................... 50
    2.2.1 Objectives ....................................................................................................... 50
CHAPTER 3

3.1 The retrotransposed mouse \textit{Zrsr1} gene acquired a new function in erythroid cells

3.1.1 Summary

3.1.2 Introduction

3.2 Materials and Method

3.2.1 Microarray data sets and analysis
3.2.2 Cell culture and transfections assays
3.2.3 RT-PCR and Real-Time Quantitative PCR
3.2.4 Immunoblotting
3.2.5 Gene constructs
3.2.6 Antibodies
3.2.7 Chromatin immunoprecipitation
3.2.8 Expression of recombinant U2AF$^{35}$-family members in \textit{E.coli}
3.2.9 Expression and purification of recombinant proteins using a lentivirus System
3.2.10 Size exclusion chromatography
3.2.11 Pull-down assay
3.2.12 Immunofluorescence
3.2.13 Protein isolation and fractionation
3.2.14 Zrsr1-knockout mice
3.2.15 Hematological Analysis
3.2.16 Flow cytometry and blood smears

3.3 Results

3.3.1 U2AF-family members are differentially expressed in a Tissue-specific Manner
3.3.2 The U2AF-family member \textit{Zrsr1} is differentially up-regulated in erythropoiesis
3.3.3 Development and characterization of a anti-Zrsr1 polyclonal antibody
3.3.4 Zrsr1 expression is up-regulated in two different cell models of erythroid differentiation
3.3.5 The \textit{Zrsr1} gene is transcriptionally activated during erythroid differentiation
3.3.6 Is there a link between H3K36me histone modification and splicing?
3.3.7 Does the Zrsr1 protein interact with U2AF$^{65}$?

3.3.8 Expression and Purification of recombinant U2AF-family members in $E.\text{coli}$

3.3.9 Expression and purification of recombinant proteins in HEK293T cells

3.3.10 Zrsr1 interacts with U2AF$^{65}$ and associates with spliceosomal components

3.3.11 Subcellular localization of the Zrsr1 protein in MEL cells

3.3.12 The Zrsr1 gene is required for normal erythropoiesis

3.3.13 Erythroid-specific alternative splicing decisions are altered in Zrsr1-deficient mice

3.4 Discussion

CHAPTER 4
4.1 Concluding Remarks and Future Perspectives

4.1 Why splicing factors may regulate alternative splicing in a tissue-specific manner?

4.1.2 Is there a link between histone modifications and splicing?

4.1.3 Subcellular localization: a way to control the U2AF-related proteins function?

4.1.4 The role of Zrsr1 in erythropoiesis

REFERENCES

SUPPLEMENTARY MATERIAL
Prefácio

Nesta dissertação apresentam-se os resultados do trabalho de investigação desenvolvido entre Janeiro de 2006 e Dezembro de 2010, sob orientação da Professora Doutora Maria do Carmo Fonseca, na Unidade de Biologia Celular do Instituto de Medicina Molecular, Faculdade de Medicina da Universidade de Lisboa.

Este trabalho teve como principal objectivo aprofundar o conhecimento actual sobre a família de proteínas U2AF\textsuperscript{35}. Inicialmente foi feito um levantamento bibliográfico sobre as características estruturais desta família, cujo objectivo centrou-se na compreensão das implicações biológicas que a evolução e a diversidade destas proteínas pode exercer no controlo dos mecanismos de splicing em organismos eucariótas. Os nossos estudos focaram-se principalmente no membro Zrs1 da família de proteínas U2AF\textsuperscript{35}, sobre cuja função, evolução e distribuição celular existiam inúmeras interrogações à data do início deste projecto. Além disso, as características peculiares da família de genes U2AF\textsuperscript{35}, permitiram-nos ainda abordar a interligação entre os mecanismos de expressão gênica, através do estudo da relação entre a estrutura da cromatina e o mecanismo de splicing.

Como previsto no Artigo 41º do Regulamento de Estudos Pós-graduados da Universidade de Lisboa, a presente dissertação encontra-se redigida em língua inglesa, contendo um resumo alargado (mais de 1200 palavras) em língua portuguesa.

Esta dissertação encontra-se dividida em quatro capítulos: no primeiro capítulo - Introduction, é feita uma revisão alargada sobre os mecanismos celulares de biogênese do RNA mensageiro. Inicialmente é feita uma descrição de cada uma das etapas onde identificamos os principais intervenientes e introduzimos a grande interligação existente entre as diversas etapas da expressão gênica. No contexto onde se insere o trabalho desenvolvido, é igualmente feita uma revisão mais detalhada sobre os mecanismos de splicing bem como o papel dos principais factores envolvidos na sua regulação.
No segundo capítulo - *Diversity of human U2AF splicing factors* - apresentado sobre a forma de artigo de revisão, discutimos as características estruturais dos membros da família de proteínas U2AF, a sua evolução bem como as potenciais implicações que a diversidade desta família poderá ter na regulação do mecanismo de *splicing*. Por último, neste capítulo, introduzimos os objectivos gerais que levaram à elaboração deste trabalho de investigação.

No terceiro capítulo - *The retrotransposed mouse Zrsr1 gene acquired a new function in erythroid cells* – são apresentados resultados originais obtidos no âmbito desta tese. É dada uma especial atenção ao membro da família de genes U2AF\(^{35}\), Zrsr1. Neste contexto, é estudado o seu padrão de expressão em diversos tecidos, com especial foco no seu envolvimento no processo de diferenciação de células eritróides. Para tal, estudámos os mecanismos genéticos envolvidos no mecanismo de sobreexpressão do gene Zrsr1, identificamos interacções proteína-proteína com outros factores de *splicing* e apresentamos uma análise do fenótipo hematopoiético de um ratinho deficitário na expressão deste gene. Por último, é feita uma ligação entre a ausência de expressão do gene Zrsr1 com eventos de *splicing* alternativos específicos de células eritroïdes.

Por último, no quarto capítulo - *Concluding Remarks and Future Perspectives* - são realçadas as principais conclusões do trabalho e expostas as perspectivas que se abrem para futuros estudos.

A realização deste trabalho não teria sido possível sem a colaboração de pessoas e instituições a quem desejó expressar os meus agradecimentos.

Em primeiro lugar, desejo começar por endereçar um agradecimento especial à Professora Doutora Maria do Carmo Fonseca, pelo privilégio que me concedeu ao aceitar-me como estudante de Doutoramento na sua Unidade de investigação. Agradeço-lhe a orientação bem como as excelentes condições de trabalho que me proporcionou por forma a desenvolver os trabalhos apresentados nesta tese.

Ao Professor Doutor Francisco Enguita (Paco), que conheci ainda nos “tempos do ITQB” queria deixar uma palavra de especial de reconhecimento, não só pela sua colaboração neste trabalho, mas especialmente pela amizade e pelos inúmeros incentivos e conselhos que me ajudaram a ultrapassar a dificuldades que foram surgindo
ao longo destes anos. Ao Doutor Sérgio de Almeida (Robertão), agradeço a sua amizade bem como a sua colaboração imprescindível para a realização de uma parte do trabalho aqui apresentado. Aos dois, o meu muito obrigado. Aos meus amigos do Erasmus MC, em particular à Professora Doutora Marieke von Lindern, por me ter recebido de forma tão especial.

A todos os meus colegas de trabalho do Instituto de Medicina Molecular, quero agradecer a amizade e o apoio constante ao longo destes anos. Uma palavra especial para todos aqueles que conheci como colegas e que hoje considero também amigos: Rui Freitas (Alfredo), Marco Campinho, Marco Antunes (Jakim), Pedro Saavreda (Zé), Alexandre Teixeira, Ana Sacristan, Joana Borlido, Ana Rita Grosso, Marisa Cabrita, Sérgio Marinho, Margarida Gama-Carvalho, Sandra Martins, Joana Desterro, Célia Carvalho, Teresa Carvalho, Noélia Custódio e José Rino.

Agradeço à Fundação para a Ciência e a Tecnologia o apoio financeiro que me concedeu no âmbito de uma Bolsa de Doutoramento (Bolsa SFRH/BD/31012/2006) durante a fase inicial do trabalho aqui apresentado.

O meu último agradecimento é, obrigatoriamente, para a minha família. Meus Pais, irmão e Avós, pelo apoio incondicional e permanente, por sempre me terem proporcionado as condições ideais para garantir o meu sucesso. Sem eles seria impossível.
Resumo

**Palavras-chave:** splicing alternativo; regulação de splicing; família de genes U2AF; Zrsr1

No núcleo, durante o processo de biogênese do RNA mensageiro (mRNA), a transcrição genética pela RNA Polimerase II origina moléculas precursoras de RNA mensageiro (pré-mRNA). Estas são submetidas a várias etapas de processamento por forma a originar moléculas de mRNA, que após transporte para o citoplasma, servem de molde para a síntese proteica. Nos organismos eucariótas, a grande maioria dos genes que codificam proteínas encontram-se organizados de uma forma complexa. As sequências codificantes, os exões, que em média são constituídos por cerca de 150 pares de bases, encontram-se separados por vastas sequências não-codificantes, designadas por intrões. O mecanismo que permite o processamento do pré-mRNA no qual as sequências não codificantes são removidas é designado por splicing. Este desenvolve-se através de uma sequência altamente coordenada de interacções RNA-RNA, RNA-proteína e proteína-proteína, com a finalidade de reconhecer/delimitar as junções exão-intrão, catalisar a remoção das sequências não codificantes e promover a ligação das sequências codificantes, por forma a originar um mRNA. A reacção de splicing do pré-mRNA é efectuada por uma complexa máquina ribonucleoproteica denominada por spliceosoma (Jurica and Moore 2003). Este é constituído por cinco partículas ribonucleicas (U1, U2, U4, U5 e U6 snRNP) e por um vasto número de proteínas auxiliares, cuja função tem um papel fundamental no correcto reconhecimento dos locais de splicing bem como na catálise da reacção (Wahl, Will et al. 2009).

O processo de splicing é hoje em dia reconhecido como um mecanismo essencial na regulação da expressão genética, em particular, pela sua importância na gênese da diversidade proteómica subjacente à complexidade dos organismos metazoários (Nilsen and Graveley 2010). De facto, existem vários exemplos em que o mesmo gene ao ser submetido ao splicing alternativo, origina diversos mRNAs cuja informação codifica proteínas com funções antagónicas. Para além de contribuir para a diversidade proteica nos organismos multicelulares, a sua importância está bem patente nos diversos exemplos onde a desregulação dos mecanismos de splicing, onde se incluem mutações em sequências reguladoras e/ou alteração dos níveis de expressão de
algunas proteínas auxiliares do spliceosoma, é relacionada com o desenvolvimento de diversas patologias humanas, tais como cancro e doenças neurodegenerativas (Kim, Goren et al. 2008; Kim, Goren et al. 2008). Além disso, o processo de splicing constitui também um mecanismo importante na regulação da expressão genética, pois o processamento alternativo de um pré-mRNA pode incluir na sequência final do mRNA um codão prematuro de terminação, titulando deste modo os níveis de expressão do gene em causa (Lareau, Green et al. 2004).

O recente desenvolvimento de novas tecnologias de sequenciação de RNA permitiu demonstrar que mais de 95% dos nossos genes são submetidos ao splicing alternativo (Pan, Shai et al. 2008; Wang, Sandberg et al. 2008). O impacto que este mecanismo exerce nos organismos implica que seja um processo altamente regulado e, por isso, reveste-se de grande importância todo o conhecimento relacionado com a sua regulação (Matlin, Clark et al. 2005). Um dos principais desafios do spliceosoma prende-se com o reconhecimento exacto dos locais de splicing. Da extensa lista de proteínas auxiliares, ou factores de splicing, que estão envolvidos na regulação deste mecanismo (Jurica and Moore 2003), destaca-se o complexo heterodimérico U2AF (U2snRNP Auxiliary Factor). Este é formado por duas subunidades altamente conservadas ao longo da evolução, tendo sido identificado em diversos organismos tais como as leveduras (Potashkin, Naik et al. 1993), nemátodes (Zorio and Blumenthal 1999), moscas (Rudner, Kanaar et al. 1996) e humanos (Zamore, Patton et al. 1992). O complexo U2AF, constituído por uma subunidade de 65 kDa, U2AF₆₅, e outra de 35 kDa, U2AF₃₅, foi identificado pelo seu papel fundamental no reconhecimento dos locais de splicing bem como pela importância que desempenha nos primeiros passos de recrutamento do spliceosoma. A subunidade U2AF₃₅ reconhece o di-nucleótido AG na extremidade 3’ do intrão (Wu, Romfo et al. 1999), e a sua associação estabiliza a interacção da subunidade U2AF₆₅ com o pré-mRNA (Sickmier, Frato et al. 2006). Deste modo, a ligação do complexo U2AF, ao promover o recrutamento da partícula ribonucleica U2snRNP bem como outras proteínas auxiliares (Zamore and Green 1989; Zamore and Green 1991), é considerada como o passo essencial para o recrutamento do spliceosoma.

A conclusão dos projectos de sequenciação do genoma de vários organismos, permitiu a identificação de novos factores de splicing que poderão desempenhar um papel fundamental no processo de recrutamento do spliceosoma (Tupler, Perini et al. 2001). De particular interesse encontram-se três genes, Zrsr1 (Hatada, Sugama et al.
1993), Zrsr2 (Tronchere, Wang et al. 1997) e U2AF26 (Shepard, Reick et al. 2002), cujas proteínas apresentam características estruturais semelhantes com a subunidade U2AF35, e por isso, são consideradas como parte de um grupo de proteínas: a família de proteínas U2AF35. Dada a importância da subunidade U2AF35 no processo de recrutamento do spliceosoma, especula-se que estas novas proteínas poderão ter evoluído por forma a desempenharem um papel fundamental na regulação do processo de splicing. Apesar de ser um tema sujeito a intensa pesquisa durante as últimas décadas, os mecanismos que promovem o reconhecimento dos locais de splicing ainda não são completamente compreendidos, embora as características únicas dos vários membros da família U2AF35 sugiram que estes possam ser alvos importantes de regulação. Compreender os detalhes dos mecanismos iniciais de recrutamento do spliceosoma é importante porque estes são frequentemente alvo de regulação. Desde modo, reveste-se de grande importância todos os estudos que permitam complementar o actual conhecimento sobre o papel específico dos membros da família U2AF35.

Neste contexto, no âmbito desta tese, foi inicialmente feita uma revisão bibliográfica cujo objectivo incidiu numa análise sobre a evolução e as características estruturais das famílias de proteínas relacionadas com o factor U2AF, discutindo as implicações da sua diversidade na regulação do mecanismo de splicing em organismos complexos.

É actualmente aceite pela comunidade científica que os mecanismos de splicing alternativo são regulados em resposta a diversos estímulos externos (Matlin, Clark et al. 2005), sendo possível associar determinados padrões de splicing às diversas fases de desenvolvimento de um organismo ou a diferentes tipos de tecido (Grosso, Gomes et al. 2008). De acordo com este modelo, a regulação do processo de splicing alternativo resulta de interacções combinatorias entre várias proteínas que, após reconhecerem sequências específicas no pré-mRNA (sequências activadoras ou silenciadoras), podem favorecer ou inibir um determinado padrão de splicing. Deste modo, as decisões de splicing específicas a um tipo de célula ou tecido resultam provavelmente de diferenças na concentração e/ou actividade destas proteínas. De acordo com este modelo, a abundância relativa destas proteínas reguladoras de splicing deverá variar de acordo com o tipo de tecido e/ou processo de diferenciação celular.

Neste contexto, começamos por determinar o padrão de expressão da família de genes U2AF35 em vários tecidos bem como em diversos processos de diferenciação celular. Para tal, efectuamos uma análise comparativa de dados de microarray
provenientes de vários sistemas biológicos. Esta análise revelou que existem diferenças robustas nos padrões de expressão dos genes da família U2AF1, em particular o membro Zrsr1 que foi identificado de forma consistente com sendo diferencialmente expresso em dois sistemas biológicos de eritropoiese. Este resultado constituiu o ponto de partida para a tentativa de caracterização do gene Zrsr1 sendo que se levantou-se a hipótese de que este poderá ter evoluído funções específicas por forma a desempenhar um papel fundamental na regulação do processo de diferenciação dos eritrócitos.

A compreensão dos mecanismos de regulação da expressão genética passa não só pela caracterização qualitativa dos factores envolvidos bem como pela determinação efectiva da quantidade de mRNA transcrito num dado sistema. Deste modo, para validar os resultados provenientes da análise dos microarrays, utilizamos a técnica de PCR quantitativo em tempo real (qPCR). Dois modelos celulares de eritroblastos de ratinho (células MEL e I/11) foram expostos a diferentes agentes químicos que induzem a sua diferenciação em eritrócitos. O mRNA total foi isolado e os níveis de expressão dos membros da família de genes U2af1 ao longo do processo de diferenciação eritróide foram determinados por qPCR. Desta análise foi possível validar os resultados acima descritos pois enquanto os níveis de mRNA do gene Zrsr1 estão aumentados ao longo da diferenciação de células eritróides, verifica-se o oposto para alguns membros da família de genes U2af1. Além disso, efectuámos estudos de imunoprecipitação da cromatina utilizando um anticorpo específico para a RNA Polimerase II com o objectivo de mapear o recrutamento/associação temporal e espacial desta proteína nos genes U2af1 e Zrsr1 durante o processo de diferenciação de células MEL. Após indução da diferenciação, verificámos que existe um aumento significativo do recrutamento da RNA Polimerase II para o gene Zrsr1, verificando-se o oposto para o gene U2af1. Mais uma vez, estes resultados confirmam os dados de microarray e de qPCR, pois a indução da diferenciação eritróide promove um aumento da transcrição do gene Zrsr1, o que culmina no aumento dos níveis de mRNA deste gene. No entanto, é importante não esquecer que alterações nos níveis de expressão de um factor de splicing não se reflectem necessariamente nos níveis de expressão da proteína em questão devido aos mecanismos de regulação pós-transcricional (Boutz, Stoilov et al. 2007; Makeyev, Zhang et al. 2007). Deste modo, na continuação dos nossos estudos sobre o envolvimento da proteína Zrsr1 no processo de diferenciação de eritrócitos, desenvolvemos um conjunto de ferramentas bioquímicas, entre as quais se destaca a produção e caracterização de um anticorpo policlonal altamente específico para a
proteína Zrsr1. Após a sua caracterização, este anticorpo foi utilizado para estudar o padrão de expressão da proteína Zrsr1 durante a diferenciação de dois modelos celulares de eritropoiese. Ao analisarmos a variação da expressão desta proteína ao longo da diferenciação de células MEL e I/11, mostrámos que o aumento dos níveis de mRNA do gene Zrsr1 é acompanhado por um aumento da expressão da proteína correspondente. Assim, os nossos resultados sugerem que o membro Zrsr1 da família U2AF35 é um factor específico das células eritróides pois é diferencialmente expresso durante o processo de diferenciação deste tecido.

Estudos recentes demonstram que todos os mecanismos de expressão genética necessários para a síntese do mRNA encontram-se interligados, ou seja, ocorrem cotranscripcionalmente (Moore and Proudfoot 2009). As inúmeras interações funcionais entre as maquinarias responsáveis pela catálise de cada reacção, permitiram o desenvolvimento de diversos mecanismos de controlo de qualidade que garantem que cada passo seja executado de forma correcta e completa. Deste modo, o elevado grau de complexidade destas interações permite um maior potencial de regulação do processo de expressão genética (Maniatis and Tasic 2002). Sabe-se que os nucleossomas, as unidades básicas da cromatina, formam uma barreira ao progresso da RNA Pol II (Hodges, Bintu et al. 2009), e que determinadas modificações pós-translacionais das histonas estão associadas com os níveis de transcrição genética: o “código das histonas” (Kouzarides 2007). Recentemente, vários estudos demonstraram os mecanismos responsáveis pela interligação entre a maquinaria do splicing e a transcrição pela RNA Pol II (Li, Howe et al. 2003; Lin, Coutinho-Mansfield et al. 2008; Pandya-Jones and Black 2009). No entanto, permanece por esclarecer a relação entre a estrutura da cromatina, a transcrição pela RNA Pol II e o splicing alternativo.

O mapeamento no genoma humano da distribuição intragénica das modificações pós-translacionais da cromatina indicam que a histona H3 trimetilada na lisina 36 (H3K36me3) está presente em genes activamente transcritos (Kouzarides 2007). Por outro lado, vários estudos recentes indicam que existe um enriquecimento acentuado desta histona modificada nos exões em relação aos intrões (Kolasinska-Zwierz, Down et al. 2009; Schwartz, Meshorer et al. 2009), levantando a hipótese de que o posicionamento dos nucleossomas com determinadas modificações pós-translacionais poderá ter um papel importante no processo de reconhecimento dos exões pela maquinaria de splicing. Uma das hipóteses é que ao invés de marcar os locais de splicing acumulando-se nos exões, a estrutura da cromatina é moldada em paralelo à
transcrição como resultado do processo de *splicing*. Deste modo, colocámos a hipótese de que determinadas modificações da cromatina poderão estar associadas ao *splicing* e contribuem para a regulação da elongação pela *RNA Pol II* através da barreira imposta pela presença dos nucleosomas. Neste contexto, especula-se que o *splicing* faz parte de um mecanismo de controlo de qualidade, onde uma falha na remoção de um intrão impede a continuação da transcrição prevenindo assim a produção de mRNA defeituosos. Do mesmo modo, esta poderá ser uma das razões pela qual a maioria dos genes tem intrões e sofrem *splicing*.

Para investigar esta hipótese, utilizámos um sistema experimental onde os níveis de transcrição podem ser manipulados. Desta forma, avaliámos a deposição da histona H3K36me3, nos genes *U2af1* e *Zrsr1*, em células MEL induzidas a diferenciar. Estes dois genes constituem um excelente modelo para estudar a interligação “código das histonas”-*splicing* alternativo. Em primeiro lugar, dos nossos resultados, sabemos que existe uma variação dos níveis de transcrição destes dois genes durante a diferenciação de células MEL. Em segundo lugar, o gene *U2af1* é constituído por exões e intrões, ao passo que o gene *Zrsr1* é constituído apenas por um exão e, por isso, não é submetido ao processo de *splicing* (Hatada, Sugama et al. 1993; Hayashizaki, Shibata et al. 1994). Os nossos resultados, demonstram claramente que não existe acumulação da histona H3K36me3 no gene *Zrsr1*, ao contrário ao que se verifica no gene *U2af1*. Deste modo, este resultados sugerem que a histona H3K36me3 poderá ser um elo de ligação entre o *splicing*, a estrutura da cromatina e a transcrição pela *RNA Pol II*.

Foram ainda feitos estudos bioquímicos para determinar se a proteína Zrsr1 estabelece interacções com outros factores de *splicing*. Para tal produzimos e purificámos a proteína Zrsr1 recombinante, que foi utilizada em ensaios *in vitro*. As características estruturais altamente conservadas da proteína Zrsr1, tal como os restantes membros da família U2AF35, sugerem que esta possa interagir com a subunidade U2AF65 (Mollet, Barbosa-Morais et al. 2006). Utilizando as ferramentas bioquímicas desenvolvidas no âmbito desta tese, confirmámos que a proteína Zrsr1 estabelece interacções com a proteína U2AF65, bem como com outros factores de *splicing*, tais como, as proteínas U2AF35 e ASF/SF2. Tal como descrito para outros membros da família U2AF35 (Tronchere, Wang et al. 1997), estas interacções sugerem que a proteína Zrsr1 faça parte de um outro complexo U2AF, que poderá estar envolvido na regulação de determinados eventos de *splicing*.
Analizámos ainda a distribuição celular da proteína Zrsr1 em células MEL. Através da aplicação de métodos bioquímicos que permitem o isolamento de proteínas citoplasmáticas, nucleoplásicas e associadas com a cromatina, verificámos que a proteína Zrsr1 está presente nas três fracções descritas. Esta distribuição da proteína Zrsr1 entre o núcleo e o citoplasma, sugere a existência de funções citoplasmáticas desconhecidas, sendo que este padrão de distribuição celular foi também descrito noutros factores de splicing, nomeadamente, as proteínas U2AF65 (Gama-Carvalho, Carvalho et al. 2001) e ASF/SF2 (Sanford, Gray et al. 2004). No entanto, estudos complementares serão necessários para esclarecer a função biológica da proteína Zrsr1 no citoplasma.

Por forma a investigar a importância biológica do gene Zrsr1 no processo de diferenciação de células eritróides, analizámos ainda o sangue de ratinhos transgénicos deficitários (Knockout) na expressão da proteína Zrsr1. De todos os parâmetros hematológicos analisados, verifica-se que os ratinhos transgénicos apresentam níveis de hematócrito consideravelmente inferiores aos dos ratinhos normais. Ou seja, o sangue dos ratinhos transgénicos é constituído por eritrócitos mais pequenos, o que sugere que o gene Zrsr1 pode estar envolvido na regulação de genes que controlam o tamanho dos eritrócitos. Por último, foi ainda feito um estudo para perceber o impacto que a perda da expressão de Zrsr1 provoca em eventos de splicing alternativo específicos de células eritróides. Os nossos resultados demonstram uma desregulação do padrão de splicing alternativo do exão 8 do gene Mbnl2.

Concluindo, o presente trabalho forneceu uma contribuição científica original pois estes resultados reforçam o actual modelo de regulação da expressão genética, onde se estabelece que as diferenças na abundância relativa e/ou actividade específica de determinadas proteínas poderão influenciar decisões no mecanismo de splicing. De facto, este trabalho alargou o conhecimento actual sobre a proteína Zrsr1, um membro até ao momento não caracterizado da família U2AF35. Desta forma, o presente trabalho desencadeia novas linhas de investigação pois muito permanece por desvendar no que respeita à função das proteínas da família U2AF35, nomeadamente na regulação de eventos de splicing alternativo pela proteína Zrsr1. De facto, as evidências recolhidas ao longo deste trabalho sugerem que a proteína Zrsr1 poderá desempenhar um papel importante no processo de diferenciação de células eritróides. Nesse sentido, a determinação dos alvos de RNA da proteína Zrsr1 em eritrócitos por técnicas de
immunoprecipitação e/ou sequenciação de RNA, poderá elucidar os mecanismos de regulação que este *splicing factor* exerce no processo de diferenciação dos eritrócitos.
Summary

Removal of non-coding intron sequences from the pre-mRNA is orchestrated by a complex macromolecular machinery called the spliceosome (Jurica and Moore 2003). Assembly of the spliceosome proceeds through the formation of several intermediates and is directed by consensus sequences located at the 5’ and 3’ splice sites and at the branchpoint (Black 2003). Regulation of this assembly results in differential splice site usage and the consequential patterns of alternative splicing are not only the major source of proteome diversity in higher eukaryotes (Nilsen and Graveley 2010), but also an important mechanism that regulates protein expression by generating premature termination codons that targets the transcripts to decay (Lareau, Green et al. 2004). Correct recognition of a functional 3’splice site involves the association of the U2AF splicing factor with the pre-mRNA. U2AF is a heterodimeric protein composed by two evolutionary conserved subunits (U2AF65/U2AF35) that play a critical role in the exon definition process (Zamore and Green 1991; Zamore, Patton et al. 1992; Zhang, Zamore et al. 1992; Wu, Romfo et al. 1999; Webb and Wise 2004; Webb, Lakhe-Reddy et al. 2005). The biochemical mechanisms that control splice-site usage, and therefore alternative splicing, are complex and remain poorly understood (Matlin, Clark et al. 2005). The growing number of studies indicating that such regulation can be tissue specific (Ule, Stefani et al. 2006), driven in a developmental (Sanchez 2008) or differentiation-specific manner (Makeyev, Zhang et al. 2007), still increases the complexity of alternative splicing regulation. While U2AF2 is found to be extremely well conserved from yeast to humans, U2AF1 was shown to have alternative spliced isoforms with unknown functions and the recent discovery of a family of U2AF35 related genes in the human genome (U2AF1, U2AF1L4, Zrsr1 and Zrsr2), argues that these proteins may have evolved specific new functions important for the development of complex multicellular organisms.

To investigate the function of Zrsr1, a previously uncharacterized member of the U2AF35-family of splicing factors, we started to access the tissue distribution patterns of these genes. By analysing several microarray datasets, Zrsr1 was found to be an erythroid tissue-specific signature, arguing that this gene may have evolved specific new functions important for the differentiation of erythrocytes. To validate this results
we used two cellular models of erythroid cells (I/11 and MEL), which were induced to differentiate upon stimulation with chemical agents. We show by qPCR that the Zrsr1 gene is specifically up-regulated during erythroid differentiation while other members of the U2AF35-family (U2af1) were found to be down-regulated. We also performed ChIP experiments to map the spatial and temporal recruitment of RNA Pol II into the U2af1 and Zrsr1 genes, upon erythroid differentiation. In agreement with the microarray and qPCR data, we found an increased occupancy of RNA Pol II at the Zrsr1 promoter as well as along the gene body, in clear contrast with a lower accumulation along the U2af1 gene.

Although changes in splicing factor mRNA levels may not necessarily reflect on protein expression due to post-transcriptional regulation (Boutz, Stoilov et al. 2007; Makeyev, Zhang et al. 2007) we accessed the protein expression levels of some U2AF35 family members during erythroid differentiation. To do this we produce and characterize a rabbit polyclonal antibody specific to the Zrsr1 protein. Our results demonstrate that upon erythroid differentiation Zrsr1 protein is up-regulated, while the U2AF35 protein levels remain largely unaffected. This up-regulation of Zrsr1, raises the possibility that Zrsr1 could replace U2AF35 in the canonical U2AF-complex, allowing the formation of a distinct heterodimer which could regulate specific splicing events.

Our findings that upon MEL cells differentiation we are able to manipulate the transcription levels of both U2af1 and Zrsr1 opened us a new window to study the interconnection between the mechanisms of gene expression. While the impact of chromatin modifications on transcription dynamics is currently acknowledged (Hodges, Bintu et al. 2009), its crosstalk with co-transcriptional mRNA splicing remains an open question in the field. The recent finding that nucleosomes are preferentially positioned in exons, and enriched with the histone H3K36me3 modification, provides evidence for extensive functional connections between chromatin structure and pre-mRNA processing (Kolasinska-Zwierz, Down et al. 2009). To investigate this hypothesis, our model system emerge as particularly appealing for that purpose since the U2af1 gene has a classical exon-intron configuration, while the related Zrsr1 mouse gene was found to be imprinted and intronless (Hatada, Sugama et al. 1993; Hayashizaki, Shibata et al. 1994; Hatada, Kitagawa et al. 1995). Our results support the importance of the histone H3K36me3 modification in splicing since there is no accumulation of this histone mark in a intronless genes when compared to a gene with intron-exon structure. In this way, the interconnection between the gene expression mechanisms are thought to act as a
quality control surveillance mechanism where failure to complete a co-transcriptional checkpoint could stall RNA Pol II complexes, thus preventing the production of misspliced mRNAs.

Although members of the U2AF-related family of proteins like U2AF, U2AF and Zrsr2 were previously described to interact with U2AF, to date there was no experimental evidences showing that Zrsr1 is also able to establish such interaction. To investigate if Zrsr1 is able to interact with U2AF, we produce recombinant proteins to perform pull-down experiments with MEL cells extracts and also in vitro assembly studies of the U2AF complex. Our results demonstrate that Zrsr1 is able to interact with U2AF, and other splicing factors like SF1/BBP and ASF/SF2. Interestingly, we found that Zrsr1 could also pull-down U2AF, which suggests that this protein is part of a larger U2AF complex that could engage network interactions during spliceosome assembly. Although the same interaction in observed for Zrsr2 (Tronchere, Wang et al. 1997), a protein that shares 94% aminoacid homology with Zrsr1, there is still no experimental data that validates this hypothesis.

Using a biochemical approach we have also accessed the subcellular localization of the Zrsr1 protein. Although in a steady-state situation Zrsr1 is found exclusively localised in the nucleus, as determined by our immunofluorescence data, analysis of the protein distribution in cytoplasmic, nucleoplasmic and chromatin-associated fractions, revealed that like other splicing factors (Gama-Carvalho, Carvalho et al. 2001), Zrsr1 shows a nucleo-cytoplasmic subcellular localization. These results seems to suggest the involvement of Zrsr1 in new cellular functions in the cytoplasm, which opens new and exciting hypothesis regarding the function of this protein.

In this work we also accessed the role of Zrsr1 in erythropoiesis by taking advantage of an available Zrsr1-deficient mice strain. To address this question we have analysed the hematologic parameters of blood samples taken from Zrsr1 KO mice. We found that the blood from these animals is populated with smaller red blood cells, suggesting that loss of Zrsr1 affects the pathways involved in the normal differentiation of the major blood cell types. Finally, we also evaluated the effect of Zrsr1 loss in previously described erythroid specific alternative splicing events. In conclusion, in light of our findings, we suggest that Zrsr1 is a novel erythroid specific splicing factor and future work, elucidating Zrsr1 specific RNA targets, will allow us to understand how this protein controls erythroid differentiation.
ABBREVIATIONS

A - adenosine

BPS- Branch Point Sequence

C - cytidine

CBC - cap binding complex

cDNA - complementary DNA

Ceg1 - RNA guanylyltransferase

Cet1 - RNA triphosphatase

CF - cleavage factors

CFIA - cleavage and polyadenylation factor IA

CFIB - cleavage and polyadenylation factor IB

ChIP - chromatin immunoprecipitation

CPF - cleavage and polyadenylation factor

CPSF - cleavage-polyadenylation-specificity factor

CstF - cleavage stimulatory factor

CTD - carboxyl-terminal domain

DEX - Dexamethasone

DRB - 5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole

DSIF - DRB sensitivity-inducing factor

dsRNA - double-stranded RNA

EPO - Erythropoietin

ESE - exonic splicing enhancer
**Pre-mRNA** - precursor messenger RNA

**PTB** - Polypyrimidine Tract-Binding protein

**P-TEFb** - positive transcription elongation factor b

**PIC** – Pre-Initiation Complex

**Py** - polypyrimidine

**RBP’s** – RNA binding proteins

**RI** – Rnase Inhibitor

**RMM**- RNA Recognition motifs

**RNA Pol II** - RNA polymerase II

**RNA Pol II LS** - largest subunit of RNA polymerase II

**RNAi** - RNA interference

**rRNA** - ribosomal RNA

**SCF** – Stem Cell Factor

**SF** - Splicing factors

**SMN** – Survival of motor neuron protein

**snoRNA** - small nucleolar RNAs

**snRNA** - small nuclear RNA

**snRNP** - small nuclear ribonucleoprotein particle

**SR** - serine/arginine

**SS**- Splice Site

**TF** - transcription factor

**U** - uridine

**U2AF** - U2 snRNP auxiliary factor
Chapter 1

Introduction
1.1 The Nucleus: the hallmark of eukaryotic cells

The nucleus is a spherical-shaped organelle present in every eukaryotic cell. Compared to other cell organelles, the nucleus is the most prominent one, with a diameter of approximately 5 µm (Cooper and Hausman 2009). Generally, a eukaryotic cell contains only one nucleus although some specialized cells are enucleated, like the mammalian erythrocytes, while others are multinucleate (skeletal muscle fibers).

The main function of the cell nucleus is to host the genetic information and therefore it is considered to act as the control center of the cell. Essential processes like DNA replication, repair, recombination and the initial steps of gene expression (transcription and RNA processing) take place in the nucleus while the final stage of gene expression (translation) takes place in the cytoplasm (Cooper and Hausman 2009). In eukaryotes a double-layered membrane separates the contents of the nucleus from the cytoplasm and the communication between the two compartments is made through nuclear pores allowing a dynamic and selective bidirectional shuttling of regulatory factors. This separation allows the cell to prevent translation of unspliced mRNA (Gorlich and Kutay 1999). Eukaryotic mRNA’s contain non-coding sequences (introns) that must be removed before being translated into functional proteins. Without the nucleus, ribosomes would translate newly transcribed (still unprocessed) mRNA’s resulting in misfolded, non-functional and potential pathogenic proteins (Martin and Koonin 2006).

Therefore, the origin of the eukaryotic nucleus marked an important evolutionary step since the physical separation of the genome from the cytoplasm allowed the rise of distinct regulatory mechanisms that are not available in prokaryotes.

1.2 Gene expression: a multistep process

Eukaryotic gene expression is a complex multistep process by which information encoded in the DNA is used to produce functional proteins. This complex multistep process begins with transcription. During transcription, the nascent pre-mRNA is capped at the 5’ end, introns are removed by the spliceosome, and the 3’ end is cleaved and polyadenylated (Moore and Proudfoot 2009). The mature mRNA is then
released from the site of transcription and exported to the cytoplasm for translation into a functional protein.

Most of what we understand about these events has been addressed along the last decades either by classical biochemical methods complemented by recent state of the art genomics and proteomics approaches. The complexity of these events led first to a simplistic view with the different steps in the pathway from gene to protein considered as unconnected events. However, findings obtained during the last decade 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, de la Mata et al. 2004).

In this section we will introduce the main steps in gene expression as well as identify the key players on each step in eukaryotic gene expression.

**Figure 1.1- Gene Expression is a multistep process.** Representation of the contemporary view of the several steps involved in the regulation of gene expression. (Adapted from (Orphanides and Reinberg 2002)).
1.3 Transcription by RNA Polymerase II

Transcription is the first step leading to gene expression. In eukaryotes RNA polymerase II (RNA Pol II) catalyzes DNA-dependent synthesis of both mRNA precursors as well as most snRNA and microRNA’s (Orphanides and Reinberg 2002; Fuda, Ardehali et al. 2009; Moore and Proudfoot 2009). To accomplish this task RNA Pol II associates with other cofactors to assemble the so called general RNA Pol II transcriptional machinery. This huge macromolecular complex (nearly 60 subunits with ~3Mda) can be simplistic divided into three main components: a 12-subunit polymerase, able to synthesize RNA and proofreading the nascent transcript, a set of five general transcription factors (GTFs), TFIIB, -D, -E, -F and –H, responsible for promoter recognition, and a modular complex of 25 proteins called Mediator that is essential to respond to gene specific regulatory signals (Woychik and Hampsey 2002).

The transcription cycle is a multistep process that can be divided into eight distinguish major steps at which several layers of regulation are present (Figure 1.2). The transcription cycle begins with RNA Pol II assembly at the core promoter. In some cases require “promoter clearing” from nucleosomes that may block RNA Pol II and GTFs access to the DNA (step 1). A pre-initiation complex (PIC) assembles on the core promoter (step 2), the DNA is unwound by the DNA helicase XPB, a subunit of TFIIH, and the RNA Pol II initiates transcription (step 3).

Early-elongating RNA Pol II escapes/clears the core promoter and proceeds to the promoter-proximal pause region (step 4). The largest subunit of RNA Pol II carboxy-terminal domain (CTD) contains evolutionary conserved heptapeptide repeats (Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7; ranging from 26-27 in yeast and 52 in mammals) (Stiller and Hall 2002) and the transition between initiation and pausing (step 4) is marked by phosphorylation of the CTD repeats on serine 5 (Ser5) by the kinase subunit of the GTF TFIIH (CDK7 in Drosophila) (Komarnitsky, Cho et al. 2000). Following promoter escape/clearance, RNA Pol II transcribes 20-40 nucleotides and stops at the promoter-proximal pause site (Step 5). At this stage RNA Pol II is held by two transcription factors, the negative elongation factor (NELF) and the DRB sensitivity-inducing factor (DSIF) which is composed of SPT4 and 5. Efficient elongation along the body of the gene, requires additional recruitment of CDK9, a subunit of human P-TEFb that phosphorylates Ser2 of the CTD, SPT5 and NELF. This enables NELF to dissociate...
from the complex allowing RNA Pol II to escape from pause. Some GTF’s can remain associated with the promoter after RNA Pol II has escaped, forming a scaffold that allows it to initiate efficiently in successive rounds of transcription (Step 6,7,8) (Fuda, Ardehali et al. 2009; Weake and Workman 2010)).

**Figure 1.2 - The transcription cycle is a multistep process.** Step 1: chromatin opening. The repressed gene and regulatory region are entirely packaged as nucleosomes (green). An activator (orange oval) binds and recruits nucleosome remodelers to clear the promoter. Step 2: PIC formation. A second activator (yellow diamond) binds, promotes the binding of GTFs (blue rectangle) and recruits coactivators (green hexagon), facilitating Pol II (red rocket) entry to the PIC. Step 3: initiation. DNA is unwound (oval inside Pol II) at the TSS, and an open complex is formed. Step 4: promoter escape/clearance. Pol II breaks contacts with promoter-bound factors, transcribes 20–50 bases downstream of the TSS, produces an RNA (purple line) and pauses, partially mediated by SPT4SPT5 in *Drosophila* (pink pentagon) and negative elongation factor (NELF) complex (purple circle). The Ser residues at position 5 (Ser 5) of the Pol II CTD repeats are phosphorylated (red P) during this step. Step 5: escape from pausing. P-TEFb (blue triangle) is recruited directly or indirectly by the activator and phosphorylates Ser 2 of the Pol II CTD repeats, SPT5 and the NELF subunits (blue Ps). NELF dissociates from the rest of the complex. Pol II escapes from the pause, either terminating or entering productive elongation. Step 6: productive elongation. Nucleosomes are disassembled and reassembled as the Pol II elongation complex transcribes through the gene. Step 7: termination. After the Pol II complex transcribes the gene, it is removed from the DNA, and the RNA is released. Step 8: recycling. (Image from (Fuda, Ardehali et al. 2009))
1.3.1 The role of chromatin during transcription

In eukaryotic cells, chromatin is the state in which DNA is packed inside the cell nucleus (Li, Carey et al. 2007; Cairns 2009). The nucleosome, the fundamental unit of chromatin, is composed by an octamer of four core histones (H2A, H2B, H3, and H4) around which 147 base pairs of DNA are wrapped (Luger, Mader et al. 1997).

Nucleosomes compact the genome but also restrict the access of DNA-binding transcription factors and RNA Pol II, creating a balance between DNA packaging and accessibility (Li, Carey et al. 2007; Clapier and Cairns 2009). It has become increasingly apparent that modulation of chromatin structure plays an important role in the regulation of transcription in eukaryotes (Hodges, Bintu et al. 2009). In fact, it is known for some time that nucleosomes behave as a barrier to RNA Pol II, while mutant histones were shown to affect gene expression in vitro (Knezetic and Luse 1986).

Once thought of as static building blocks of chromatin structure, the nucleosomes are now clearly understood as a dynamic structure whose stability can be regulated by posttranslational modifications and enzymatic activities (Workman and Kingston 1998). Two major classes of factors have the ability to rebuild chromatin structure in a way that might impact specifically on RNA Pol II transcription: histone modifiers and chromatin remodelers.

The core histones are predominantly globular except for their N-terminal “tails”, which are unstructured (Luger, Mader et al. 1997), and both histone tails and globular domains are subjected to a vast array of posttranscriptional modifications (Kouzarides 2007). There are at least eight distinct types of modification which include methylation, acetylation, ubiquitilation, ADP-ribosylation, sumoylation and phosphorylation of histone residues. Except for acetylation (carried out by a variety of histone acetyltransferase complexes, HATs), all other modifications are usually catalyzed by a specific enzyme and occur at a specific site resulting in unique physiological roles (see Figure). In fact, the term “histone code” has been loosely used to describe the role of modifications that enable DNA functions (Kouzarides 2007).
Histones H3 and H4 are typically acetylated at active genes, and the level of histone acetylation tends to be greatest at the promoter and 5’ regions. In fact, among all histone modifications, acetylation has the most potential to unfold chromatin since it neutralizes the basic charges of the lysine (Workman and Kingston 1998). Nevertheless, a single standard code of histone acetylation remains elusive since it is also important to note that particular acetylated residues can be associated with gene inactivity (Kouzarides 2007; Li, Carey et al. 2007).

Histone methylation at specific residues as well as the extent of their modification (mono-, di- or tri-methylation) can also correlate with either transcriptional activity or repression. The role of each histone modification on transcription relies on the exact gene location to where the enzymes that catalyze such modifications are recruited (Campos and Reinberg 2009).

Figure 1.4- Distribution of several H3 histone tail methylations along active genes. Adapted from (Bell, Wirbelauer et al. 2007).
For instance, Set2 targeting to the chromatin templates occurs through binding to RNA Pol II phosphorylated on the serine 2 (Li, Howe et al. 2003). This phosphorylation pattern is a trademark of elongating RNA Pol II molecules and is present within the open reading frame of actively transcribed genes (Weake and Workman 2010). Since Set2 mediates trimethylation of histone H3K36, this modification has a positive correlation with transcription (Li, Carey et al. 2007).

The second major class of chromatin regulators are the protein complexes that use ATP hydrolysis to change histone-DNA contacts (Cairns 2009). A number of enzymes termed chromatin remodelers, reposition, reconfigure or eject nucleosomes, tailoring the way that chromatin is packaged to influence gene expression (Clapier and Cairns 2009).

![Figure 1.5 - The functions of chromatin remodelers in nucleosome dynamics](image)

**Figure 1.5 – The functions of chromatin remodelers in nucleosome dynamics.** Remodelers use ATP hydrolysis to change the histone-DNA binding properties. Remodelers from the ISWI family are involved in nucleosome assembly and organization which may mask a binding site (red) for a transcriptional activator (ACT). The SWI/SNF-family of remodelers can provide access to binding sites, mainly through nucleosome sliding or ejection from the DNA. Adapted from (Cairns 2009)

Chromatin remodeling complexes use ATP hydrolysis to change the histone-DNA binding properties (Lusser and Kadonaga 2003) and may be grouped in four families: SWI/SNF; ISWI; CHD and INO80 (Saha, Wittmeyer et al. 2006), each of them specialized for particular purposes and biological contexts. The activities of these complexes have different outputs, like unwrapping of DNA from histone octamers, moving nucleosomes to different intragenic positions and changing the accessibility of nucleosomal DNA to TFs. In addition, these remodelers can modify the nucleosome composition by promoting displacement of histones from DNA (Clapier and Cairns 2009).

While the impact of chromatin remodeling on transcription dynamics is currently acknowledged, its crosstalk with co-transcriptional mRNA processing remains an open question in the field. The recent finding that nucleosomes are preferentially positioned
in exons provided evidence for extensive functional connections between chromatin structure and pre-mRNA processing (Kolasinska-Zwierz, Down et al. 2009).

1.4 Pre-mRNA processing

1.4.1 Capping

The first RNA processing event to occur on the nascent transcript is 5’ end capping. The cap structure is found at the 5’ end of all eukaryotic mRNA’s and is formed shortly after transcription initiation, when the nascent pre-mRNA’s are about 25-30 nucleotides in length (Coppola, Field et al. 1983; Rasmussen and Lis 1993; Moteki and Price 2002). Capping involves the sequential action of three enzymatic activities: RNA 5’-triphosphatase (RTP), RNA guanylyltransferase (GT) and RNA-(guanine-N7)- methyltransferase. First the RTP removes the γ-phosphate of the first nucleotide of the pre-mRNA, followed by the transfer of GMP to the resulting diphosphate end by RNA GT. The cap structure is then finalized by the RNA (guanine-7-) methyltransferase that adds a methyl group to the N7 position of the cap guanine to form the m7G(5’)ppp(5’)N cap. In metazoans, the capping enzyme is bi-functional with both RNA 5’-triphosphatase and RNA guanylyltransferase activities, while in Saccharomyces cerevisiae, capping enzyme consists of a heterodimer of RNA triphosphatase (Cet1) and RNA guanylyltransferase (Ceg1) (Changela, Ho et al. 2001; Shuman 2001).

The 5’ m7GpppN cap plays an essential role in the life cycle of eukaryotic mRNA and is required for efficient pre-mRNA splicing, export, stability and translation. In the nucleus, the cap structure is recognized by the heterodimeric protein complex called the cap-binding complex (CBC), which is composed by two cap binding proteins (CBP20 and CBP80). After export to the cytoplasm, this association supports the pioneer round of mRNA translation after which the CBP is replaced by the eukaryotic translation initiation factor eIF-4E (Mitchell and Tollervey 2001).
1.4.2 Splicing

The low number (~20,000-25,000) and split nature of eukaryotic genes requires an important physiological mechanism capable of produce a large number of mRNA’s in order to generate the complex proteome of higher organisms (Matlin, Clark et al. 2005). Most protein coding genes from higher eukaryotes are synthesized as a precursor molecule (pre-mRNA), which must be submitted to a series of processing steps before being exported to the cytoplasm where is used as a template for protein translation (Sharp 2005). During gene expression, non-coding intervening sequences (introns) are removed from pre-mRNA, while coding sequences (exons) are joined together, to generate a mature mRNA. This process, called splicing, is orchestrated by the spliceosome, a highly conserved, dynamic and complex macromolecular machine (Wahl, Will et al. 2009), in which five small nuclear ribonucleoprotein particles (snRNP’s) and a large number of auxiliary proteins cooperate to accurately recognize exons from introns and catalyse the two steps of the splicing reaction (Jurica and Moore 2003; Matlin, Clark et al. 2005).

In metazoans, two distinct spliceosomes catalyzing pre-messenger RNA splicing have been identified. The first one, the U2-dependent or major spliceosome, is found in all eukaryotes and catalyzes the removal of U2-type introns, which are the most commonly encountered class of introns. The main building blocks of the major spliceosome are five snRNPs: U1, U2, U4; U5 and U6. Each of them contain a single uridine-rich small nuclear RNA (snRNA) that is associated to a common core of SM and other proteins characteristic of each snRNP. Active splicing complexes are also found to require additional non-snRNP proteins, also known as splicing factors, that exert auxiliary functions in the splicing reaction. In fact, mass spectrometric analysis and co-purification studies of spliceosomes identified between 150-300 non-snRNP protein components (Zhou, Licklider et al. 2002; Jurica and Moore 2003). More recently, a less abundant spliceosome, the U12-dependent “minor” spliceosome, was also found to exist in parallel with the U2-dependent “major” spliceosome in most multicellular eukaryotes (Will and Luhrmann 2005). It catalyzes the removal of a rare class of introns (U12-type) that represent less than 1% of introns in mammals (Tarn and Steitz 1996). Although less frequent, the importance of the minor spliceosome can be illustrated by the fact that U12-type introns are found in genes carrying out essential cellular functions like DNA replication and repair, transcription, RNA processing and translation (Will and
Luhrmann 2005). The U12-dependent spliceosome contains four unique snRNAs: U11, U12, U4atac, and U6atac, which are paralogs of U1, U2, U4, and U6 snRNAs of the U2-dependent spliceosome, respectively, while the U5 snRNA is shared between both spliceosomes (Patel and Steitz 2003). Proteomic analysis of the minor spliceosome tri-snRNP U4atac/U6atac.U5 revealed a remarkable similar protein composition when compared to the paralog tri-snRNP from the major spliceosome (Schneider, Will et al. 2002). This striking protein composition was also found in the U11-U12 di-snRNP, although seven proteins specific to the minor spliceosome could be found (Will, Schneider et al. 2004).

The major challenge for the spliceosome lies in locating and bringing together the sites at which the cut-and-paste reactions have to proceed with single nucleotide precision (Cartegni, Chew et al. 2002; Wang and Burge 2008). To accomplish this, the splicing machinery must recognize introns from exons within the context of the gene sequence (Moore 2000; Black 2003). Indeed, introns and intron/exon boundaries are defined within the gene sequence by a set of specific conserved elements, which are required for splicing (Stephens and Schneider, 1992). There are four short consensus sequences that define an intron: the exon–intron junction, or splice sites (ss), at the 5′ and 3′ end of introns (5′ss and 3′ss), the branch point sequence located upstream of the 3′ss and the polypyrimidine tract (Py tract) located between the 3′ss and the branch site.

![Figure 1.6– Consensus sequences of major- and minor-class introns.](image.jpg)

The letters heights at each position represent the frequency of occurrence of the corresponding nucleotides at that position. Nucleotides that are involved in intron recognition are shown in black. (Image from (Patel and Steitz 2003))

In the yeast *Saccharomyces cerevisiae*, these consensus elements are found to be extremely well conserved and the information coded by these sequences is known to be sufficient for correct recognition of the splice sites by the splicing machinery, leading to
the subsequent intron excision (Black 2003). The 5’ splice site in yeast is defined by the consensus sequence 5’-R/GUAUGU-3’ (R-purine nucleotide, either A or G and / denotes the exon-intron boundary), the branch point is invariably 5’-UACUAΔC-3’, while the 3’ splice site is defined by the 5’-CAG/N-3’ consensus sequence (Lin et al., 1985; Rymond and Rosbash, 1992). In clear contrast, in higher eukaryotes, the U2-type introns have a 5’ splice site characterized by the consensus sequence 5’-AG/GURAGU-3’ (Will and Luhrmann 2005), while the 3’ splice site follows the sequence 5’-YAG/G-3’ (Y-pyrimidine base, either C or T) and a pyrimidine-rich, 10-12 nucleotide (nt) long region upstream of the AG dinucleotide. Located 18-40 nt upstream the 3’ splice site, the branch point sequence (BPS) is characterized by a highly degenerate sequence 5’-YNYURAC-3’ (A–branch adenosine; N- any nucleotide) that contains a conserved adenosine (Reed and Maniatis 1988). On the other hand, U12-type introns, found in vertebrates, insects and plants, they lack a recognizable Py-tract and have highly conserved splicing signals: the 5’/-AUAUCCUUU-3’ for the 5’ splice site, 5’-UCCUUAAΔC-3´for the branch point sequence located 10-20 nt upstream a 3’ splice site with the degenerate sequence and 5’-YAS/-3’ (S- either C or G) (Will and Luhrmann 2005). Although essential, these short and highly degenerated elements are not sufficient for splicing (Bindereif and Green 1986), since they do not provide full specificity for splice site determination. Thus, other sequence elements, like intronic and exonic splicing enhancers or silencers, have been identified to play an important role in splice site selection and alternative splicing regulation (Blencowe 2000) (see section 1.6.2).

Mechanistically, the catalytic removal of an intron occurs through two trans-esterification steps (Figure 1.6). In the first, the 2’-hydroxyl group of the intronic branch point adenine residue attacks the phosphodiester bond of the guanosine nucleotide at the 5’ end of the intron (1st step, red arrow). At the end of this step, the 5’ end of the intron is cleaved from the upstream exon and covalently linked to the adenine, generating free 5’exon and a intron lariat-3’exon. In the second step of the splicing reaction, the free 3’-hydroxyl group from the excised exon (2nd step, red arrow) attacks the phosphodiester bond at the 3’end of the intron. This releases the intron as a free lariat and produces an RNA with the two ligated exons (Black 2003; Wahl, Will et al. 2009).

The splicing mechanism proceeds by a coordinated series of RNA–RNA, RNA–protein and protein–protein interactions which recognize exon-intron junctions leading to exon ligation and release of the intron.
INTRODUCTION

Figure 1.7- Pre-mRNA splicing reaction. A schematic pre-messenger RNA is shown on the left as a single intron (solid line) flanked by two exons. The first and second steps of splicing involve nucleophilic attacks (red arrows) on the terminal phosphodiester bonds (blue dots) by the 2′ hydroxyl of the branch-point adenosine (A) and by the 3′ hydroxyl of the upstream exon, respectively. The ligated exons and the lariat intron products are shown on the right. Adapted from (Patel and Steitz 2003)

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 trans-acting elements (auxiliary protein factors). The current model for the formation of an active spliceosome, based on several in vitro studies of different spliceosomal complexes (Konarska and Sharp 1986; Das and Reed 1999; Makarov, Makarova et al. 2002; Zhou, Licklider et al. 2002), has lead to the suggestion that it involves an ordered, stepwise assembly (see Figure 1.8) of snRNP particles on the pre-mRNA substrate (Kent and MacMillan 2002). The earliest step of the spliceosome assembly, requires the ATP-independent binding of the U1 snRNP to the 5′ ss of the intron via direct base pairing. Along with the U1-5′ ss interaction, the SF1/BBP and the U2 auxiliary factor (U2AF) are recruited to the pre-mRNA and together they recognize the BPS and the polypyrimidine tract, respectively (Nilsen 2003). The U2AF complex is an heterodimeric protein composed by two subunits (U2AF65 and U2AF35) (Zamore and Green 1991; Zamore, Patton et al. 1992). While the U2AF65 binds to the polypyrimidine tract (Sickmier, Frato et al. 2006), interacting with SF1/BBP through its C-terminal domain (Selenko, Gregorovic et al. 2003), the U2AF35 binds the AG dinucleotide at the 3′ ss (Wu, Romfo et al. 1999). Together, these interactions yield the spliceosomal E, or commitment, complex and play a crucial role in the commitment steps that triggers the pre-RNA to the splicing mechanism (Wahl, Will et al. 2009).

After the formation of the spliceosomal E complex, the U2 snRNA engages in an ATP-dependent manner a base-pairing interaction with the pre-mRNA’s BPS, leading to the formation of the A complex. At this stage the SF1/BBP protein is released
from the pre-mRNA, being replaced by the SF3b14a/p14 (Spadaccini, Reidt et al. 2006), while the U2AF heterodimer recruits the U2 snRNP to the branch point sequence (Will, Schneider et al. 2001). This interaction is stabilized by proteins complexes of the U2 snRNP, namely SF3a and SF3b (Gozani, Feld et al. 1996) and also by the arginine-serine-rich domain of U2AF65 (Valcarcel, Gaur et al. 1996). Subsequent to A complex formation, the U4/U6 and U5 snRNPs are recruited as a preassembled U4/U6.U5 tri-snRNP to form the B complex. Although all snRNP’s are present, it is still catalytic inactive, requiring major conformational changes for catalytic activation. During spliceosome activation, U1 and the U4snRNP are released, giving rise to the activated spliceosome (the B* complex) that catalyses the first splicing reaction. The intron is now in a lariat configuration, the C complex is formed, and additional conformational changes are required for the second trans-esterification reaction to occur. After this catalytic step, the spliceosome dissociates, releasing the mRNA while the U2, U5 and U6 snRNP’s are recycled for new round of splicing (Kent and MacMillan 2002; Black 2003; Wahl, Will et al. 2009).

Assembly of the U12-minor spliceosome is similar to the U2-dependent, with a major difference occurring at the earliest step. Prior to association with the pre-mRNA, the U11 and U12 snRNPs form a highly stable di-snRNP that binds cooperatively to the 5’splice site and branch point sequence. Thus, in contrast to the major spliceosome, the earliest assembly step involves formation of the A complex while the remaining appear to mirror those of the major spliceosome (Will and Luhrmann 2005).

Correct intron recognition and splicing are crucial steps in gene expression and are especially complex problems in the case of alternative splicing, where a single gene may yield multiple mRNAs and protein isoforms. Indeed, alternative splicing is a major source of metazoan proteome diversity (Maniatis and Tasic 2002; Nilsen and Graveley 2010) and is known to be an important mechanism that regulates gene expression by generating premature termination codons that targets the transcripts to non-mediated mRNA decay (Lewis, Green et al. 2003). Recent studies using high-throughput sequence technology estimates that 95-100% of human pre-mRNA’s undergo alternative splicing (Pan, Shai et al. 2008; Wang, Sandberg et al. 2008) and, not surprisingly, up to 15% of human genetic diseases arise from disruption of normal splicing patterns (Krawczak, Reiss et al. 1992). Since the spliceosome must be able to recognize and remove something like $10^5$-$10^6$ different intron sequences (Moore 2000; Black 2003), the major challenge for the splicing machinery is to ensure the precise
excision of introns. As the splicing process is central to the work presented in the following chapters, current knowledge about alternative splicing regulation and integration in the gene expression flow path will be described in more detail.

![Figure 1.8- Different steps of the major spliceosome assembly](image)

**Figure 1.8- Different steps of the major spliceosome assembly.** Cross-intron assembly and disassembly cycle of the major spliceosome. The stepwise interaction of the spliceosomal snRNPs (colored circles), in the removal of an intron from a pre-mRNA containing two exons (blue). (Adapted from (Wahl, Will et al. 2009))

### 1.4.3 3’ end processing

In eukaryotes, formation of the mature 3’ end of a mRNA involves a two step reaction where the transcript is cleaved and then polyadenylated. This universal step of gene expression (with the exception of replication-dependent histone transcripts) proceeds through the recognition of cis-acting elements in the transcript. Core polyadenylation sequence motifs recognized by the 3’ processing machinery includes the hexanucleotide AAUAAA element (or a close variant AUUAAA) found 10-30 nucleotides upstream the cleavage site (CA dinucleotide) and the U/GU-rich region located 30 nt downstream of the cleavage site (see Figure 1.9A) (Gilmartin 2005). In metazoans, the 3’ end processing machinery requires multiple protein factors, including:
the cleavage and polyadenylation specificity factor (CPSF), the cleavage stimulation factor (CstF), cleavage factor I and II (CFI and CFII) and the poly(A) polymerase (PAP). CPSF, required for both cleavage and polyadenylation, recognizes and binds to the AAUAAA hexamer while the CstF associates with the U/GU-rich region. After the cleavage reaction, polyadenylation proceeds through the recruitment of PAP to the AAUAAA-containing substrate via interaction with CPSF and this reaction is further stimulated after binding of PABPN1 protein to the nascent polyA tail until it reaches approximately 200 adenosine residues (Proudfoot and O'Sullivan 2002).

Proper 3’end processing of a nascent transcript is critical for the functionality of the mature RNA. In fact, this step plays an essential role in the gene expression flow-path since it may affect the transcript’s stability, sub-cellular localization, translational efficiency and export to the cytoplasm. A well characterized example involves the IgM heavy chain mRNA, where usage of an intronic versus the normal terminal polyA site regulates the production of secretory versus membrane bound proteins (Zhao, Hyman et al. 1999). Not surprisingly, there is an increased evidence that defective 3’ processing is linked to several human diseases (Danckwardt, Hentze et al. 2008).

![Figure 1.9- Mammalian pre-mRNA 3’end processing.](A) Schematic representation of the mammalian poly(A) site with conserved sequence elements and relative distances between them. Adapted
INTRODUCTION

The cleavage and polyadenylation reactions require CPSF, CstF, two additional cleavage factors (CF I and CF II), PAP and the phosphorylated CTD (pCTD) of RNA Polymerase II. PAP together with CPSF, directs poly(A) addition. PABPN1 (PAB) binds the growing poly(A) tail, enhancing the efficiency of polyadenylation and forming 21 nm spherical particles. Adapted from (Proudfoot and O'Sullivan 2002).

1.5 Gene expression is a highly interconnected multistep process

Capping is the first RNA processing event to occur on the nascent transcript and is also the best described mechanism coupled to transcription. After the formation of the transcription initiation complex, or soon after initiation, DSIF and NELF are recruited to the transcription unit. The arrest mediated by the DSIF/NELF complex association is then overcome by the positive transcription elongating factor, P-TEFb, and the associated protein kinase, CDK9, which phosphorylates both the CTD at Ser2 and Spt5, the larger subunit of DSIF (reviewed by (Orphanides and Reinberg 2002)). In mammals, the capping enzyme is able to interact directly with the phosphorylated CTD through the guanylyltransferase domain (Ho, Sriskanda et al. 1998; Fong and Bentley 2001) and the human methyltransferase binds to the complex of human capping enzyme and phosphorylated CTD (Pillutla, Yue et al. 1998). In fact, the human capping enzyme stimulates promoter escape by countering the negative elongation factor NELF, and capping enzyme recruitment is enhanced by direct binding to the elongation factor Spt5 (Mandal, Chu et al. 2004). Therefore, capping may well be a key component of the switch that pushes RNA Pol II from abortive early elongation into fully processive elongation across the body of the gene. In this way a checkpoint to ensure timely capping of the nascent pre-mRNA before commitment to processive elongation of the transcript, ensuring that only properly capped RNA molecules are extended (Guiguen, Soutourina et al. 2007).

As mentioned, the transcription rate of RNA Pol II is known to be modulated by the phosphorylation status of the CTD. Phosphorylation on Ser5 of the CTD is associated with RNAP II stalling downstream the promoter region, whereas phosphorylation on Ser2 is associated with elongation (Lin, Marshall et al. 2002; Phatnani and Greenleaf 2006). A good example for the coupling between the gene expression events is provided by the CD44 gene. In fact, it was shown that a subunit of the human Swi–SNF complex, BRM, regulates changes in the alternative splicing of...
CD44 pre-mRNA. Upon T cell stimulation, RNAP II phosphorylated on Ser5 pauses at the variant exon region of CD44 by a mechanism requiring its association with BRM. Interestingly, this also results in increased association of BRM with components of the splicing machinery like the splicing factor Sam68, leading to inclusion of the V5 exon (Batsche, Yaniv et al. 2006). Therefore, this study shows that the mechanism for alternative splicing regulation by transcription can result from a combination of transcription elongation-related effects and differential recruitment of splicing factors (Chen and Manley 2009). In fact, some splicing factors were shown to have a critical role in RNA Pol II transcriptional elongation. SC35-depleted cells were found to have lower levels of RNA Pol II phosphorylated on Ser2, which results from an impaired recruitment of P-TEFb. Consequently, it is proposed that by affecting the dynamic recruitment p-TEFb and, therefore, the phosphorylation status of the CTD, SC35 plays a critical role in the elongating rate of RNA Pol II (Lin, Coutinho-Mansfield et al. 2008).

Another example for the cross-talk between splicing and transcription is related with the observation that in vitro, the splicing factor U2AF65 binds directly to the RNA Pol II during the transition from initiation to elongation (Ujvari and Luse 2004). Indeed, U2AF65 was found to coimmunoprecipitate with hyperphosphorylated forms of RNA Pol II (Listerman, Sapra et al. 2006).

As discussed above, variations in the posttranscriptional modifications of histone tails are implicated in the regulation of chromatin structure and function. Nucleosome positioning can affect the selection of exons through two possible scenarios. First, the nucleosome might act as a “speed bump” on the exon, which slows RNA Pol II elongation and leads to increased inclusion of that exon (Keren, Lev-Maor et al. 2010). In fact, this model is strongly supported by recent studies showing that the nucleosome behaves as a fluctuating barrier that results in RNA Pol II pausing (Hodges, Bintu et al. 2009).

Another possibility is that nucleosomes in exons have a specific set of histone modifications that enhance the interaction with the splicing machinery which enables a more efficient recognition of the exon (Tilgner, Nikolaou et al. 2009).

The increased number of genome-wide studies seems to indicate that specific histone modifications have a clear distribution profile along the genome and are conserved between species (Keren, Lev-Maor et al. 2010). For example, it was known that H3K4me3 is found near transcription start sites while H3K36me3 accumulates in the body of genes and H3K9me3 is enriched on silent genes. More recently, a novel and
striking pattern of histone modification was reported (Kolasinska-Zwierz, Down et al. 2009), where H3K36me3 is found to be preferentially associated with exons relative to introns. In this way, what could be the function of H3K36me3 in exon marking? An attractive possibility is that marked exons in chromatin provide a mechanism to facilitate efficient splicing and, therefore, “marked” exons could aid recruitment of splicing factors to chromatin. A second possibility is that the splicing machinery could regulate, directly or indirectly, H3K36 methyltransferases on the traveling RNA Pol II complex, such as Set2.

A major breakthrough is the field was provided by Misteli and co-workers were they demonstrate a direct link between histone modifications, the splicing machinery and the splice site choice (Luco, Pan et al. 2010). The human fibroblast growth factor receptor 2 (FGFR2) gene is an establish model, in which exons IIIId and IIIc undergo mutually exclusive and tissue-specific alternative splicing. In human prostate normal epithelium cells (PNT2s), exon IIIb is predominantly included, whereas in human mesenchymal stem cells, it is repressed and exon IIIc is included. The differential inclusion of these two exon is regulated by PTB, which binds to silencing elements around exon IIIb, resulting in its repression. In this model gene is was found that levels of H3K4me2, H3K9ac and H3K27ac are similar across the alternatively spliced region. In contrast H3K36me3 and H3K4me1 were found to be enriched over the FGFR2 gene in hMSC (where IIIb is repressed), whereas H3K27me3, H3K4me3 and H3K9me were reduced when compared to PNT2 cells (where the exon is included). These observations revealed for the first time a correlation between histone mark signature and PTB-dependent repression of alternatively spliced exons. In fact, whenever the levels of H3K36me3 are modulated either by overexpression or knock-down of SET2, the inclusion levels of exons IIIb are reduced or increased, respectively (Luco, Pan et al. 2010). Moreover, the physical link between H3K36me3 and PTB seems to be the protein MRG15 that is known to recognize this histone modifications and is able to interact directly with PTB. In light of these findings, it is proposed that the epigenetic information encoded by the histone modification patterns is used not only to determine the level of activity of a gene, but also transmits information that regulates alternative splicing patterns.
1.6 Alternative splicing is a major regulator of gene expression

Alternative splicing of pre-mRNAs is a powerful and versatile regulatory mechanism that is responsible not only for the expansion of the proteome functional diversity (Nilsen and Graveley 2010) but is also an important quantitative control of eukaryotic gene expression (Lopez 1998). In eukaryotic cells, splicing is often complex since the majority of mammalian pre-mRNAs are composed by multiple introns which, in some cases, have more than one 5' and/or 3' splice sites in their sequences (Black 2003; Barash, Calarco et al. 2010). Alternative splicing by the utilization of different combinations of splice sites (see Figure 1.10) can yield the production of several mRNA isoforms encoding different proteins, some of them with distinct functions and/or activities (reviewed by (Stamm, Ben-Ari et al. 2005)). A remarkable example is provided by the *Drosophila Dscam* gene, where alternative splicing is predicted to generate 38016 distinct mRNA isoforms (Schmucker, Clemens et al. 2000), that is, twice the number of predicted genes in the entire fly genome (Adams, Celniker et al. 2000). Additionally, alternative splicing of pre-mRNA’s may also generate isoforms with profound regulatory effects in the protein function. A good example is provided by the human Bcl-x gene in which alternative splicing generates two isoforms with antagonistic activities since Bcl-x(L) is an anti-apoptotic factor, whereas Bcl-x(S) can induce apoptosis (Boise, Gonzalez-Garcia et al. 1993).

![Figure 1.10- Different types of alternative splicing.](image)

*Figure 1.10- Different types of alternative splicing.* Transcripts from a pre-mRNA can undergo many different patterns of alternative splicing. Several genes show multiple positions of alternative splicing, creating complex combinations of exons and alternative segments and a large family of encoded proteins. Adapted from (Li, Lee et al. 2007)
Transcripts from a gene may be submitted to several different patterns of alternative splicing (see Figure 1.10): transcriptional initiation at different promoters may generate alternative first exons that can be joined to a common exon, while through the use of alternative 5’ or 3’ splice sites, exons can be extended or shortened in length. Inclusion and skipping of a cassette exon and mutually exclusive splicing of cassette exons are also known alternative splice patterns, while intron retention where the excision of an intron is suppressed is also found in several transcripts (Galante, Sakabe et al. 2004). Many genes show multiple positions of alternative splicing, creating complex combinations of exons and alternative segments and consequently different protein coding sequences (Black 2003; Chen and Manley 2009). Additionally, alternative splicing is also acknowledge to regulate gene expression by promoting the inclusion of premature stop codons-containing exons which triggers the transcript to nonsensemediate mRNA decay (Lejeune and Maquat 2005).

Due to the implications that alternative splicing may exert in generating the biological complexity of higher eukaryotes, and to the increased evidences that its misregulation is linked to several human diseases (Wang and Cooper 2007), deciphering the splicing code (Barash, Calarco et al. 2010) has been a subject of intense research in the last decades. A typical human gene contains relatively short exons, ranging from 50-250 base pairs, separated by much larger introns, typically with hundreds to thousands base pairs in length that on average account for more than 90% of the primary transcript (Wang and Burge 2008). Within this context, one of the major challenges for the spliceosome machinery is to accurately recognize exons from introns in the vast sequence of a pre-mRNA. Several genetic and biochemical approaches have identified cis-acting regulatory elements (pre-mRNA sequences) and trans-acting factors (auxiliary splicing factors) that are involved in the regulation of specific pre-mRNAs alternative splicing events. These studies are contributing to a better understanding of alternative splicing regulation. Therefore, in this section, we present the fundamental topics that are relevant for the regulation of alternative splicing and how this mechanism can contribute to the overall gene expression regulation in complex organisms.
1.6.1 The splice site strength and the role of the U2AF complex in splicing regulation

The major issue of both constitutive and alternative splicing is the selection of the correct splice sites within the vast sequence of a pre-mRNA. In higher eukaryotes, introns are usually much longer than exons and splice-site motifs are highly degenerated and predicted to have many matches along pre-mRNAs. Despite the vast majority of these sequences, also known as pseudo splice sites, are highly frequent they are not selected for splicing (Sun and Chasin 2000), although in some cases they can be used as a result of a mutation in the pre-mRNA (Roca, Sachidanandam et al. 2003). Given the complexity of higher eukaryotic genes and the relatively low level of splice site conservation, the precision of the splicing machinery in recognizing and pairing splice sites is impressive (Hertel 2008). Which 5´ss and 3´ss are recognized in the context of a pre-mRNA and subsequently paired by the spliceosome, clearly influences the sequence of the mRNA that is ultimately produced (Wahl, Will et al. 2009). Splice site selection in higher eukaryotes is determined by multiple factors (Reed 1996; Smith and Valcarcel 2000; Nilsen 2003). Among these, the relative strength of a given splice site has been described to play an important role in the early steps of spliceosome assembly (Wahl, Will et al. 2009).

In a typical eukaryotic pre-mRNA the 5′-splice site junction is defined by a single element of 9 nucleotides, while the 3′-splice site can be broken down to three sequence elements usually found within 40 nt upstream the 3′-exon/intron junction (Reed 1996). These elements are known as the branch-point sequence, the Py-tract, and the YAG sequence at positions -1 to -3 relative to the 3′-exon /intron junction (see Figure).

Figure 1.11- Consensus sequences that define a mammalian U2-type intron. Y, R and N, indicate pyrimidine, purine and any nucleotide, respectively. Image from (Moore 2000)
In the early steps of the spliceosome assembly, initial recognition of exon/intron junctions is based on direct interactions between the U1snRNP with the 5’splice site and the U2AF splicing factor with the Py-tract (Black 2003). Because the sequence specificity of these interactions is driven by pre-mRNA/U1snRNA interactions and U2AF binding to the Py-tract (Singh, Valcarcel et al. 1995), splice sites strengths can be classified on the basis of the complementarity between the U1snRNP with the 5´splice site and by the affinity of the U2AF complex to the Py-tract (3´-splice site).

Although the mammalian 5´splice site consensus sequence corresponds to perfect Watson–Crick base-pairing to the U1 snRNA 5´ terminus (Horowitz and Krainer 1994), individual 5´splice sites are found to exhibit considerable variation at different positions, indicating a tolerance for mismatches in U1 base pairing. Nevertheless, deviations from the consensus sequences are known to result in decreased affinity of the splicing machinery for the pre-mRNA (Smith and Valcarcel 2000).

Despite the direct base-pairing between 5´splice/U1snRNA plays an important role in the early steps of the spliceosome assembly, exceptions have been reported. U1-snRNA depleted extracts were shown to be complemented by U6 snRNA and SR proteins (Crispino, Blencowe et al. 1994; Crispino and Sharp 1995), a mechanism that may contribute to the high fidelity of splicing when U1 snRNA is present in limiting amounts.

Another relevant question in the 5´splice site selection arises when to nearby competing 5´splice sites are present in the pre-mRNA sequence (see Figure 1.12), a question that is relevant for both alternative and cryptic 5´splice site activation (Roca, Sachidanandam et al. 2005). Analysis cryptic 5´splice site in human genes, showed that in general they are weaker than the near authentic 5´splice sites (Roca, Sachidanandam et al. 2003), although some mutations are known to affect 5´splice site selection. Thalassemia-associated mutations of the 5´splice site of intron 1 of human β-globin gene, are known to activate the use of three cryptic 5´splice sites (Treisman, Orkin et al. 1983). Additionally, point mutations in β-globin exon 1 cryptic 5´splice site were found to have higher affinity to U1 snRNP and thus activate this splice site (Nelson and Green 1990). Therefore, mutations in flanking cryptic 5´ splice site may change the level of activation of constitutive 5´splice site, suggesting that the choice of a splice site is not only related to its own intrinsic strength, but might also be influenced by its flanking competitors (Xia, Bi et al. 2006). Additionally, besides the information present in the pre-mRNA, trans-acting factors may be critical to distinguish between authentic and
cryptic splice sites. In fact, the relative efficiency for the utilization of the three cryptic 5’ splice sites in human β–globin gene was shown to be modulated, either in vivo and in vitro, by ASF/SF2 and hnRNP A1 (Krainer, Conway et al. 1990; Mayeda and Krainer 1992; Caceres, Stamm et al. 1994). Therefore, selection of a 5’ splice site may involve not only the relative strength but also other sequence features (cis-acting sequence motifs) within the 5’ splice site that are binding sites for trans-acting auxiliary proteins (see section 1.6.2).

Figure 1.12- Regulation of alternative splicing through the relative 5’ss strength. Splicing patterns resulting from competition between two adjacent 5’ss. Adapted from (Roca, Sachidanandam et al. 2005).

In higher eukaryotes, typical introns have a uracil-rich stretch or Py-tract, adjacent to the 3’splice site where it serves as an important signal for both constitutive and regulated pre-mRNA splicing (Green 1991). Initial reports have identified distinct 3’ splice-site sequence arrangements that were found to promote splicing (Reed 1989). These include a short Py-tract (14 nt), followed by an essential AG-dinucleotide and a long pyrimidine stretch (26 nt) without an AG requirement (Reed 1989). Therefore, based on the relative strengths of a 3’ splice site sequence they can be classified as follows: BPS long Py-tract AG > BPS short Py-tract AG = BPS long Py-tract > BPS short Py-tract (Reed 1989).

The splicing factor responsible for the recognition of the Py-tract during the early steps of the spliceosome assembly is U2AF\textsuperscript{65}, a subunit of the U2AF complex (Zamore and Green 1989; Zamore and Green 1991). The U2AF splicing factor was first biochemically described as an uncharacterized activity essential for assisting U2 snRNP binding to the branch point of the 3’ splice site (Ruskin, Zamore et al. 1988). In fact, U2AF was considered to be an essential pre-mRNA splicing factor since depleted nuclear extracts were not able to splice β-globin pre-mRNA. U2AF was found to be an evolutionary conserved heterodimeric protein composed by two polypeptides with a relative molecular mass of 65 KDa (U2AF\textsuperscript{65}) and 35 KDa (U2AF\textsuperscript{35}) that interact in a
1:1 stoichiometry (Zamore and Green 1989). Orthologs of U2AF$^{65}$ have been observed in *Mus musculus* (Sailer et al., 1992), *C. elegans* (Zorio and Blumenthal 1999), *D. melanogaster* (Kanaar, Roche et al. 1993), and *Schizosaccharomyces pombe* (Potashkin, Naik et al. 1993), were it was shown that this gene encodes an essential protein. Indeed, deletion or mutation of the large subunit either in *S. pombe* (Potashkin, Naik et al. 1993) or *D. melanogaster* (Kanaar, Roche et al. 1993) were shown to correlate with a lethal phenotype. Functional characterization of U2AF$^{65}$ revealed that this splicing factor binds specifically to the polypurine tract/3′ splice site region of several pre-mRNA’s (Zamore and Green 1989). Although the polypurine tract is acknowledged to be a required splicing signal, there is a remarkable diversity of Py-tract sequences in mammalian pre-mRNA’s (Zamore, Patton et al. 1992). The Py-tract is known to affect splice site choice and thus the earliest steps of the spliceosome assembly (Roscigno, Weiner et al. 1993), and these effects are most likely to be a consequence of the affinity of U2AF$^{65}$ to distinct Py-tracts. In fact, U2AF$^{65}$ was found to bind diverse Py-tracts with extraordinary different affinities, showing a direct correlation between binding to RNA with both pyrimidine content and Py-tract length (Zamore, Patton et al. 1992). In fact, identification of the optimal binding site using systematic evolution of ligands by exponential enrichment (SELEX) identified the consensus binding site of U2AF$^{65}$ as UUUUuu/cCcucUUUUUucc (Singh, Valcarcel et al. 1995), providing clear evidences that U2AF$^{65}$ preferentially binds to uridine-rich degenerated sequences similar to those found in the Py-tract of most vertebrate introns. However, the RNA binding properties of U2AF$^{65}$ could not account for all the mammalian 3′ splice site arrangements (Reed 1989). Since U2AF$^{65}$ showed high affinity for degenerated pyrimidine-rich sequences alone (Singh, Valcarcel et al. 1995), and not for a Py-tract followed by a YAG, a key piece was missing to explain the sequence organization of mammalian 3′ splice sites (Moore 2000).

The human U2AF small subunit, U2AF$^{35}$, was cloned and characterized by Zhang and co-workers (Zhang, Zamore et al. 1992). The functional significance of the small subunit was illustrated by the phylogenetic conservation of this subunit between different organisms. Orthologs of the small subunit have been identified in humans (Zhang, Zamore et al. 1992), *C. elegans* (Zorio and Blumenthal 1999), *D. melanogaster* (Rudner, Kanaar et al. 1996), and *S. pombe* (Wentz-Hunter and Potashkin 1996). Genetic analysis in *S. pombe* (Wentz-Hunter and Potashkin 1996) and RNAi –mediated knockdown of the small subunit in *C. elegans* (Zorio and Blumenthal, 1999b) resulted
in a lethal phenotype, indicating that U2AF\textsuperscript{35} is an essential splicing factor required for viability. Additionally, biochemical data from the fruit fly U2AF heterodimer demonstrated that the small subunit significantly contributes to the high-affinity binding of the heterodimer (Rudner, Kanaar et al. 1996). Alone the small subunit was found to have minimal RNA binding activity but in the context of the heterodimer the small subunit increases the binding potential of the large subunit by twenty-fold. Additionally, mutations in the Py-tract of the 3’ splice site, by changing several pyrimidine nucleotides to purine, were found to increase the dependence of the large subunit on the small subunit for binding. Therefore, for the first time it was suggested that the small subunit could assist the large subunit binding to the 3’ splice site sequence through association with the pre-mRNA (Rudner, Kanaar et al. 1998).

The major breakthrough on U2AF\textsuperscript{35} came when three groups independently (Merendino, Guth et al. 1999; Wu, Romfo et al. 1999; Zorio and Blumenthal 1999) reported the basis of the U2AF\textsuperscript{35}-mediated assistance in RNA binding. These studies conclusively demonstrated that the small subunit specifically recognize the AG dinucleotide, thus elucidating the function of the small subunit in 3’ splice site recognition, explaining how it can assist U2AF\textsuperscript{65} binding to the pre-mRNA (see Figure 1.13). Remarkably, a SELEX experiment performed with the human U2AF heterodimer lead to the amplification of a sequence that is exactly the mammalian 3’ splice site (Wu, Romfo et al. 1999).

Figure 1.13- The role of the U2AF complex in the regulation of the 3´ splice site selection. (A) Consensus binding sites for UAF\textsuperscript{65} (right) and for the U2AF complex (Singh, Valcarcel et al. 1995; Wu, Romfo et al. 1999). (B) In a AG-independent intron (left) binding of U2AF\textsuperscript{65} to a strong Py-tract is
sufficient to recruit the U2snRNP to the BPS. However, in AG-dependent introns (right), when the Py-tract is short, U2AF$^{35}$ binding to the YAG sequence is required to stabilize U2AF$^{65}$ binding to the pre-mRNA and the subsequent recruitment of the U2snRNP. Adapted from (Moore 2000).

These results provided for the first time the molecular explanation for the AG dependence of some introns since strong Py-tracks efficiently bind U2AF through the large subunit and thus do not require the AG-dinucleotide or U2AF$^{35}$ for the first step of splicing (Wu, Romfo et al. 1999). In contrast, introns with weak Py-tracks have a relatively low affinity for U2AF$^{65}$, and thus require the additional contact provided by the U2AF$^{35}$/AG interaction for efficient U2AF binding and splicing (Wu, Romfo et al. 1999; Moore 2000). In fact, RNAi-mediated knock-down of each U2AF subunit was reported to inhibit weak 3’ splice site recognition of some introns, while U2AF$^{35}$ was found to regulate the selection of weak 3’ splice sites in a specific subset of cellular transcripts (Pacheco, Coelho et al. 2006). Therefore, recognition by the U2AF complex in the early steps of spliceosome assembly is considered to be the evolutionary driving force behind the mammalian 3´ splice site organization (Moore 2000).

1.6.2 The role of cis-acting Regulatory Elements in Splice-Site Selection

The small and degenerate splicing signals in higher eukaryotes is a complex problem in the context of splice site selection (Matlin, Clark et al. 2005). Since in eukaryotes, introns are outstandingly long when compared with exons (Sorek, Shamir et al. 2004), which 5´ and 3´ splice sites are correctly recognized and subsequently paired by the spliceosome clearly influences the sequence of the mRNA that is ultimately produced (Wahl, Will et al. 2009). To compensate for the short and poorly conserved nature of splice-site sequences in higher eukaryotes, recognition and selection of splice sites is in most cases influenced by flanking pre-mRNA regulatory sequences. Depending on the position and function these splicing regulatory sequences can be divided into four categories: exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs), intronic splicing enhancers (ISEs), and intronic splicing silencers (ISSs) (Izquierdo and Valcarcel, 2006). These so-called cis-acting elements mediate their effect by acting as binding sites for trans-acting regulatory elements that in turn mediate
or repress the recruitment of the spliceosome machinery to the adjacent splice site (see Figure 1.14).

**Figure 1.14- Mechanisms of protein-protein cross-talk for splice site recognition during the early steps of the spliceosome assembly.** The splicing machinery recognizes the 5' (GU) and 3' (AG) splice sites (ss) as exon flanking sequences. Binding of SR proteins to exonic splicing enhancers (ESE), recruit the U1 snRNP to the downstream 5'ss and the U2AF complex to the upstream polypyrimidine (YYYY) tract (65 kDa subunit) and 3'ss (35 kDa subunit). U2AF recruits the U2 snRNP to the branch point (A). SR proteins function in both “cross-exon” and “cross-intron” recognition complexes. (Adapted from (Maniatis and Tasic 2002).)

Several members of the multifunctional family of Serine/Arginine (SR) proteins are acknowledged to play a critical role in both constitutive and alternative splicing (Long and Caceres 2009). In fact, alternative exons are known to have weak splice sites which are inefficiently recognized by the U1 snRNP and the U2AF complex, although SR proteins bound to near cis-acting elements are described to compensate the degenerated mammalian splicing signals (Lopez 1998), and to modulate splice site selection in specific cell contexts (Ladd and Cooper 2002).

ESEs are often binding sites for SR proteins, which have roles in several steps of spliceosome assembly and acts as regulatory factors. For example, SR proteins are known to be involved in the recruitment of the U1 snRNP to the 5’ splice site and the U2AF complex and U2 snRNP to the 3’ splice site. An example for this mechanism of regulation is provided by the binding of T cell-restricted intracellular antigen 1 (TIA1) to a U-rich sequence downstream of a weak 5’ splice site which helps to recruit the U1 snRNP (Del Gatto-Konczak, Bourgeois et al. 2000). Additionally, the Src-associated in mitosis 68 Kda protein (Sam68) was shown to bind and recruit U2AF to the 3’ splice site of exon V5 of the CD44 pre-mRNA (Tisserant and Konig 2008).
Figure 1.15- Regulation of exon v5 alternative splicing in human CD44 pre-mRNA. (A) Inclusion of exon v5. In T cells that are activated by Ras signalling, SAM68 (Src-associated in mitosis 68 kD) is phosphorylated by extracellular signal-regulated kinase (ERK). Binding of SAM68 to exon v5 on CD44 pre-mRNA abolishes the repressive activity of hnRNP A1. Phosphorylated SAM68 either prevents hnRNP A1 from binding the exonic splicing silencer (ESS) by steric hindrance, or counteracts the inhibitory effect of hnRNP A1 that is bound to the ESS. These protein–protein interactions allow serine-arginine rich (SR) proteins such as alternative-splicing factor/splicing factor-2 (ASF/SF2) and the related protein transformer-2 (TRA2), to function through the exonic splicing enhancer (ESE) to enhance v5 inclusion in the final mRNA. (B) Exclusion of exon v5. In most tissues, hnRNP A1 represses exon v5 inclusion by directly interfering with essential upstream splicing factors such as U2 snRNP and U2AF35, and/or by abolishing the formation or function of the downstream ESE complex. (Image from (Shin and Manley 2004)).

In opposition to the positive effects of splicing enhancers, inhibition of splice site recognition can be achieved in many ways. Many silencers (both ISS’s and ESSs) include binding sites for hnRNPs, namely hnRNP A1, hnRNP F/H, hnRNP L and PTB/hnRNP I, or other proteins like SXL. Inhibition of a splice site recognition can be achieved by different ways. First, binding of a negative trans-acting factor can sterically block the access of a positive regulatory factor to a near enhancer, thus preventing the recruitment of snRNPs. For example, it is known that PTB can bind the Py-tract of
several genes and block the binding of the U2AF complex (Sauliere, Sureau et al. 2006; Spellman and Smith 2006). Another very well known example of simple steric inhibition of early spliceosome assembly is provided by the SXL protein. In female flies, SXL binds the Py-tract of an up-stream 3’ splice site, blocking the recruitment of the U2AF complex, which allows the selection of the immediately downstream 3’ splice site (Penalva, Lallena et al. 2001). On the other hand, inhibition of a splice site can also be achieved when upon binding of trans-acting elements the RNA adopts a secondary structure that masks splice sites or binding sites for splicing factors. In fact, it is known that upon binding of hnRNP A1 to a ISSs the alternative exon adopts a conformation (loop) that prevents further spliceosome assembly (Nasim, Hutchison et al. 2002). Another example is provided by the protein MBNL1 which binds a stem-loop within intron 4 of the cardiac troponin T pre-mRNA (Warf and Berglund 2007). In fact, MBNL1 was shown to regulate the splicing of exon 5 by competing directly with U2AF65 for binding to the 3’ end of intron 4 (Warf, Diegel et al. 2009). MBNL1 and U2AF65 seems to compete by binding to mutually exclusive RNA structures, where MBNL1 recognizes the intron as a stem-loop, whereas U2AF65 binds the same region in a single-stranded structure. Therefore mutations in the pre-mRNA structure that strengthen the stem-loop, were shown to decrease the binding of U2AF65, repressing exon 5 inclusion (Warf, Diegel et al. 2009).

Although the examples mentioned above mostly involve promotion or inhibition of early spliceosome assembly, many alternative splicing events involve a more complex interplay between positive and negative regulators. The final decision of whether an exon is included or excluded from the final mRNA sequence is tightly regulated by the relative concentration of these regulatory factors. In good agreement with this, it is known that the relative concentration of hnRNP A1 and the SR protein ASF/SF2 can influence the alternative splicing pattern of a model transcript (Mayeda and Krainer 1992; Caceres, Stamm et al. 1994), through the selection of different adjacent 5’ splice sites. The increased evidence that these splicing regulators are often expressed in a tissue specific fashion and that post-transcriptional modifications can regulate their activity, provides an additional layer of complexity and regulation. Nevertheless, these mechanisms provide a very fine tuned way of controlling splice site selection.
1.6.3 Alternative Splicing coupled to NMD

If the differential usage of different splice sites is acknowledged to be an important mechanism to generate the enormous proteomic diversity of higher eukaryotes (Nilsen and Graveley 2010), alternative splicing is also emerging as an important mechanism to regulate gene expression (Lareau, Green et al. 2004). A recently recognized mechanism that contributes to post transcriptional gene expression regulation is provided by the coupling between alternative splicing events that introduce a premature termination codon (PTC) and the consequent degradation of the mRNA by non-sense mediated decay (NMD). The process of gene expression regulation through the coupled action of alternative splicing and NMD has been termed AS-NMD (Lewis, Green et al. 2003).

During pre-mRNA splicing, exon–exon splice junctions are marked with a protein complex, termed exon junction complex (EJC), which is deposited 20 to 24 nucleotides upstream the splice junction (Le Hir, Izaurralde et al. 2000). Besides playing an important role in exporting spliced mRNA’s to the cytoplasm (Le Hir, Gatfield et al. 2001), the EJC also allows the cell to distinguish between normal termination codons, which are located in the last exon, and PTC’s found 50 to 55 nucleotides upstream a EJC (Chang, Imam et al. 2007) (see Figure 1.16).

Initial bioinformatic predictions estimated that about 25–35% of alternative exons introduce frameshifts or stop codons into the pre-mRNA (Lewis, Green et al. 2003; Stamm, Ben-Ari et al. 2005). Since approximately 75% of these exons are predicted to be subject to non-sense mediated decay, an estimated 18–25% of transcripts will be switched off by stop codons caused by alternative splicing and nonsense-mediated decay (Lewis, Green et al. 2003; Stamm, Ben-Ari et al. 2005), suggesting that AS-NMD is a widely used mechanism to control mRNA abundance.

There is an increasing number of examples in which specific transcripts are regulated by the coupling of alternative splicing and NMD (reviewed by (Lareau, Brooks et al. 2007)), and this mechanism is thought to provide an additional layer of gene expression regulation by allowing the cell to titrate the proper level of expression for a given protein. In this way, the cell can change the levels of a productive mRNA after transcription by “shutting-off” some fraction of the already transcribed pre-mRNA into an unproductive splice form, targeting it to degradation by NMD (Lareau, Green et
al. 2004). In fact, the coupling of alternative splicing and NMD can be easily incorporated into the existing models of gene regulation since it allows the use of the alternative splicing machinery to regulate protein expression in a developmental stage- and cell-specific manner.

![Figure 1.16- Coupling between Alternative splicing and NMD](image)

Figure 1.16- Coupling between Alternative splicing and NMD. (A) The spliceosome deposits an EJC on the mRNA ~20-24 nt upstream the splice junction. On the pioneering round of translation, any in frame stop codon found more than 50 nt upstream of the splice junction triggers NMD; such a codon is called a PTC.

(B) Regulation of gene expression through alternative splicing coupled NMD. At low protein concentrations, the alternative splicing pattern generates a stable mRNA that is used for protein translation. However, when the protein concentration is too high, changes in the alternative splicing patterns introduce a PTC in the mRNA which triggers NMD and titrates the protein concentration. Adapted from (Lareau, Brooks et al. 2007; Moore and Proudfoot 2009).

A particularly interesting example of AS-NMD controlled gene expression is provided by Exportin 4 (Xpo4), a nuclear receptor which is known to play an important role in the transport of transcription factors involved in the regulation of embryonic development (Gontan, Guttler et al. 2009). This gene was found to harbor a developmentally regulated PTC-introducing exon, since the expression of Xpo4 is regulated in adult tissues by the inclusion of a PTC that triggers NDM, while alternative splicing regulated exon skipping allows Xpo4 expression in embryonic tissues (Barash, Calarco et al. 2010).

An important question concerning the gene expression regulation mechanisms is how the components of the spliceosome machinery are regulated at appropriate levels in a tissue-specific manner. This constitutes an extremely important point since the current
view of alternative splicing regulation postulates that the differential expression of splicing factors and/or its relative abundance are known to control tissue-specific alternative splicing events (Grosso, Gomes et al. 2008). Not surprisingly, there is an increasing number of splicing factors and elements of the splicing machinery (Saltzman, Kim et al. 2008) that are regulated through AS-NMD (Lareau, Brooks et al. 2007), namely SR proteins (Lareau, Inada et al. 2007) and the small subunit of the U2AF complex, which was shown to give rise to an alternative spliced transcript that contains a PTC (Pacheco, Gomes et al. 2004). A very well known example of such regulation is provided by the PTB protein. The PTB pre-mRNA is found to be alternatively spliced in order to produce two unproductive isoforms lacking exon 11, which causes a frameshift leading to a downstream PTC that triggers NMD (Wollerton, Gooding et al. 2004). Additionally to this AS-NMD mechanism, PTB was found to auto-regulate its own pre-mRNA levels since it promotes the removal of exon 11 (Wollerton, Gooding et al. 2004). Consequently, when PTB levels are high, PTB production is slowed by targeting its own transcripts for NMD while when PTB levels are low, production is accelerated by reducing the proportion of transcripts that are degraded (Rahman, Bliskovski et al. 2002; Spellman and Smith 2006). A similar auto-regulatory process has been reported for the SR protein SC35, since its overexpression promotes the alternative splicing of its own NMD-targeted isoforms in order to reduce protein production (Sureau, Gattoni et al. 2001). In fact, AS-NMD seems to be a general mechanism to regulate the expression of several proteins involved in the core spliceosome formation (Saltzman, Kim et al. 2008), while both examples above described implicates that some splicing factors are submitted to auto-regulatory loops in order to titrate their own availability in the cell. Although the prevalence of NMD-targeted splice forms has only recently become clear (Lewis, Green et al. 2003), alternative splicing and NMD coupling provides a very fine tuned mechanism to regulate the expression of a wide range of genes.

**1.6.4 Alternative Splicing and Polyadenylation**

Although it is clear the role of alternative splicing in the regulation of the , additional roles for alternative processing in the regulation of gene expression are now emerging. Consistent with EST-based bioinformatic studies (Zhang, Lee et al. 2005),
RNA-seq analysis identified tissue-specific regulation of polyadenylation sites (Wang, Sandberg et al. 2008). Alternative Polyadenylation and cleavage (APA), which can occur in both a splicing-independent mechanism (by the use of different polyadenylation sites in the terminal exon) and a splicing-dependent form (by means of mutually exclusive terminal exons, also called 3’ exon switching) (see Figure 1.17), was the potential to generate transcripts from the same pre-mRNA that are different in their 3’UTR sequences (Licatalosi and Darnell 2010). The 3’ untranslated region (3’UTR) of a mRNA has very well described functions in the stability, localization, and translation (Moore 2005) and, therefore, changes in the 3’UTR sequence provides the potential for differential regulation of mRNA expression (Licatalosi and Darnell 2010) by trans-acting regulatory proteins and/or small non-coding RNAs (Bartel and Chen 2004).

What might be the functional consequence of distinct 3’UTRs? Use of different 3’UTR sequences generated by alternative splicing and/or APA can eliminate large regulatory sequences of a given mRNA, allowing it to evade from the stronger regulatory potential of longer 3’UTRs. Besides miRNA regulation, the loss of regulatory sequences in the 3’UTR can influence mRNA nuclear export and cytoplasmic localization, as well as non-miRNA mediated changes in mRNA stability and translational efficiency (Moore 2005). In fact, alternative mRNAs that differ in their 3’UTRs can exist in different tissues or developmental stages, and several studies have shown that these mRNA isoforms can have different stability or translational active (Miyamoto, Chiorini et al. 1996; Takagaki, Seipelt et al. 1996; Lutz 2008). Bioinformatic studies predict that about half of the mammalian genes can generate multiple mRNA isoforms differing in their 3’UTRs (Beaudoing and Gautheret 2001; Zhang, Lee et al. 2005), although the extent to which differential expression of these isoforms is used to regulate mRNA and protein levels is still poorly understood.
Recent reports propose that changes in 3’UTR length by means of APA and/or alternative splicing is a coordinated mechanism for regulating the expression of many genes during T cell activation (Sandberg, Neilson et al. 2008), neuronal activation (Flavell, Kim et al. 2008) or embryonic development (Ji, Lee et al. 2009). A good example of this regulatory mechanism is provided by the Hip2 gene. Overall Hip2 mRNA expression was found to be very similar between naïve and activated T lymphocytes but, upon cellular activation, relative expression of the extended 3’UTR region decreased, leading to increased protein levels (Sandberg, Neilson et al. 2008). This regulatory mechanism seems to be correlated with the elimination of two seed sequences for mir-21 and mir-155 present in the extended 3’UTR of Hip2 mRNA in activated T cells. In fact, shortest 3’UTR mRNA isoforms were found to have greater stability, which is correlated with higher protein expression than the full-length isoforms (Mayr and Bartel 2009). Additionally, proliferating cells were found to generate alternatively spliced 3’UTRs containing fewer miRNA-binding sites (Sandberg, Neilson et al. 2008) and, strikingly, several cancer cells lines were also found to had shortest 3’UTR than non-transformed cells, suggesting that loss of 3’UTR repressive elements is an important mechanism for oncogene activation (Mayr and Bartel 2009). Another example of such regulation is provided by the T cell receptor (TCR) associated CD3 zeta (ζ) chain which is found to have decreased expression in systemic lupus erythematosus (SLE) patients. Decreased expression of this protein was
directly linked to decreased levels of the functional wild-type transcript with increased levels of an unstable 3’UTR alternatively spliced isoform (Moulton and Tsokos 2010). As the alternative spliced CD3ζ isoforms was found to lack two critical regulatory adenosine uridine-rich elements (ARE) and a translation regulatory sequence, the transcript stability and translation of this 3’UTR spliced isoforms is significantly lower when compared to the wild-type transcript (Chowdhury, Tsokos et al. 2005).

The examples above illustrate the profound effects that the regulation of the 3’UTR sequence by means of alternative splicing and/or APA may exert in the regulation of gene expression. The discovery that different 3’UTR regions can be generated in a tissue-specific fashion, adds another layer of regulation to the complex eukaryotic gene expression pathway since it allows a mRNA to escape from or be submitted to different levels of regulation. Therefore, a major challenge for the future will be to identify the exact extent that these mechanisms can exert in the regulation of gene expression in complex organisms and how much do they contribute to the development of human pathologies, which may open a new window for therapeutical targets.

1.6.5 Splicing and Disease

Splicing signals are a frequent target of mutations in genetic diseases (Wang and Cooper 2007) and cancer, since there is a growing list of mutations that affect the splicing of oncogenes, tumour suppressors and other cancer-relevant genes (Srebrow and Kornblihtt 2006; Venables 2006). Several bioinformatics studies revealed that changes in splicing factor expression/concentration may play an important role in the general splicing disruption that occurs in many cancers (Kim, Goren et al. 2008; Kim, Goren et al. 2008; Ritchie, Granjeaud et al. 2008). In fact, there is a growing body of evidence indicating that the splicing machinery is a major target for misregulation in cancer (Grosso, Martins et al. 2008). Microarray and high-throughput data analysis have detected alternative splicing signature events associated with different types of cancers (Grosso, Martins et al. 2008). A good example for this correlation between SR protein expression and cancer progression has been provided by the fact that the SR protein SF2/ASF is a proto-oncogene (Karni, de Stanchina et al. 2007). This activity of SF2/ASF was directly linked with its splicing activity, in particular, an oncogenic
isoform of ribosomal protein S6 kinase-β1 which is induced by SF2/ASF, strongly correlates with the oncogenic activity of this SR protein. In a separate study, SF2/ASF was shown to regulate alternative splicing of the tyrosine kinase receptor proto-oncogene, RON, to produce a constitutively active form (ΔRON) (Ghigna, Giordano et al. 2005). ΔRON expression is elevated in two-thirds of breast cancers and the activated receptor induces increased migration and invasiveness, properties that are characteristic of metastatic progression.

An additional example, is provided by the missplicing of the cholecystokinin-B/gastrin receptor gene which seem to be correlated with reduced expression of U2AF35 in pancreatic cancer cells (Ding, Kuntz et al. 2002). Additionally, RNAi-mediated down-regulation of U2AF35 in HeLa cells has been reported to change the ratios of alternatively spliced isoforms of transcripts encoding the oncogenic CDC25B phosphatase, and to increase the level of CDC25B protein (Pacheco, Moita et al. 2006). Although this study do not provide a direct connection between decreased levels of U2AF35 and cancer, it shows that the relative abundance of a splicing factor is correlated with the regulation of alternative splicing events of a known oncogenic transcript.

As mentioned above, splicing signals are also a frequent target of mutations in genetic diseases. A remarkable example into how a silent mutation can cause exon skipping is provided by spinal muscular atrophy (SMA) disorder, which is caused by the loss of both functional copies of the survival of motor neuron 1 (SMN1) gene (Lefebvre, Burglen et al. 1995). Although humans have two SMN genes, SMN1 and SMN2, that potentially encode indistinguishable proteins, SMN2 is only partially functional. A crucial, translationally silent single-nucleotide C→T difference between SMN1 and SMN2 at position +6 of exon 7 (C6T) results in a very inefficient inclusion of exon 7 in SMN2 mRNA (Lorson, Hahnen et al. 1999; Monani, Lorson et al. 1999). In fact, it was found that this substitution on SMN2 exon 7 causes the disruption of an ASF/SF2-dependent exonic splicing enhancer, leading to the very inefficient exon 7 inclusion in SMN2 mRNA (Cartegni and Krainer 2002).

Considering the examples provided, it is clear that misregulation of the splicing machinery is a major target for several human diseases. Therefore a major challenge for the future is to integrate the different layers of gene expression regulation altered in disease and the development of novel therapeutic approaches that could modulate the splicing machinery.
Chapter 2

Diversity of human U2AF splicing factors
Chapter 2

2.1 Diversity of human U2AF splicing factors

The heterodimeric protein U2AF was one of the first non-snRNP essential splicing factors to be identified. Correct recognition of a functional 3’splice site involves the association of the U2AF splicing factor with the pre-mRNA. U2AF is a heterodimeric protein composed by two evolutionary conserved subunits (U2AF$_{65}$/U2AF$_{35}$) that play a critical role in the exon definition process. While U2AF$_{65}$ is a very well conserved protein from yeast to humans, U2AF$_{35}$ was shown to have alternative spliced isoforms with unknown functions. Moreover, the recent discovery of a family of U2AF$_{35}$-related genes in the human genome, argues that these proteins may have evolved specific new functions important for the development of complex multicellular organisms.

In this chapter we discuss the conserved structural features that characterize the U2AF protein families, their evolutionary emergence as well as the potential implications of U2AF protein diversity in splicing regulation. This chapter has been published as a review article published in: Mollet I., Barbosa-Morais N., Andrade J., Carmo-Fonseca M., ”Diversity of human U2AF splicing factors”, FEBS Letters, 273 (2006):4807-4816.

At the end of the chapter, we introduce the main goals of this thesis.
DIVERSITY OF HUMAN U2AF SPlicing FACTORS

THE EMBO LECTURE

Diversity of human U2AF splicing factors

Based on the EMBO Lecture delivered on 7 July 2005 at the
30th FEBS Congress in Budapest

Inês Mollet, Nuno L. Barbosa-Morais, Jorge Andrade and Maria Carmo-Fonseca

Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Portugal

Keywords
CAPER, PUF60, RNA splicing: U2AF

Correspondence
M. Carmo-Fonseca, Institute of Molecular Medicine, Faculty of Medicine, Avenue Prof. Egas Moniz, 1649-028 Lisbon, Portugal
Fax: +351 21 7999412
Tel: +351 21 7999411
E-mail: carmo.fonseca@fm.ul.pt

Received 13 July 2006, revised 12 September 2006, accepted 14 September 2006

doi:10.1111/j.1742-4658.2006.05002.x

Introduction

In eukaryotes, protein-coding regions (exons) within precursor mRNAs (pre-mRNAs) are separated by intervening sequences (introns) that must be removed to produce a functional mRNA. Pre-mRNA splicing is an essential step for gene expression, and the vast majority of human genes comprise multiple exons that are alternatively spliced [1]. Alternative splicing is used to generate multiple proteins from a single gene, thus contributing to increase proteome diversity. Alternative splicing can also regulate gene expression by generating mRNAs targeted for degradation [2]. Proteins produced by alternative splicing control many physiological processes and defects in splicing have been linked to an increasing number of human diseases [1,3].

Pre-mRNA splicing occurs in a large, dynamic complex called the spliceosome. The spliceosome is composed of small nuclear ribonucleoprotein particles (the U1, U2, U4/U5/U6 snRNPs forming the major spliceosome and the U11, U12, U4atac/U6atac/U5 snRNPs forming the less abundant minor spliceosome) and more than 100 non-snRNP proteins [4]. Spliceosome assembly follows an ordered sequence of events that begins with recognition of the 5' splice site by U1 snRNP and binding of U2 AF (U2 small nuclear ribonucleoprotein auxiliary factor) to the polypyrimidine (Py)-tract and 3' splice site [5]. Human U2AF is a heterodimer composed of a 65-kDa subunit (U2AF65), which contacts the Py-tract [6–8], and a 35-kDa subunit (U2AF35), which interacts with the AG dinucleotide at the 3' splice site [9–11]. Assembly of U2AF with the pre-mRNA, which in yeast and mammals requires an interaction with the U1 snRNP [12–17], is important for subsequent recruitment of U2 snRNP to the spliceosome.

U2AF has been highly conserved during evolution. In addition, a number of U2AF-related genes are
DIVERSITY OF HUMAN U2AF SPlicing FACTORS

U2AF diversity

present in the human genome, and some are known to be alternatively spliced. Here, we review currently available information on the diversity of U2AF proteins and we discuss the resulting implications for splicing regulation.

Structural features of U2AF and U2AF-related proteins

The U2AF\textsuperscript{65} protein contains three RNA-recognition motifs or RRMs (Table 1). The two central motifs (RRM1 and RRM2) are canonical RRM domains responsible for recognition of the Py tract in the pre-mRNA, whereas the third RRM has unusual features and is specialized in protein-protein interaction. This unusual RRM-like domain, called UHM for U2AF homology motif, is present in many other splicing proteins [18]. The UHM in U2AF\textsuperscript{65} recognizes splicing factor 1 (SF1), and this cooperative protein-protein interaction strengthens the binding to the Py-tract (Fig. 1). The UHM motif was highly conserved from yeast to mammals, but, paradoxically, appears dispensable for splicing of at least certain pre-mRNAs in vitro [19]. The N-terminal amino acids 85–112 of U2AF\textsuperscript{65} interact with U2AF\textsuperscript{65}, and this association further strengthens the binding to the Py-tract [18]. Although it is not a member of the serine-arginine (SR) family of splicing factors, the U2AF\textsuperscript{65} protein further contains an arginine and serine rich (RS) domain that is required for spliceosome assembly in vitro [20,21]. Importantly, binding of U2AF\textsuperscript{65} alone is sufficient to bend the Py-tract, juxtaposing the branch region and 3' splice site [22]. Current models therefore propose an arrangement in which the C-terminus of U2AF\textsuperscript{65} is positioned proximal to the branch point, and the N-terminus is situated in the vicinity of the 3' splice site (Fig. 1).

PUF60 [poly(U)-binding factor-60 kDa] was first isolated as a protein closely related to U2AF\textsuperscript{65} that was required for efficient reconstitution of RNA splicing in vitro [23]. The homology between PUF60 and U2AF\textsuperscript{65} extends across their entire length, except for the N-terminus where PUF60 lacks a recognizable RS domain (Table 1 and Fig. 2A). CAPER\textalpha{} and CAPER\textbeta{} are the most recently characterized proteins related to U2AF\textsuperscript{65} [24]. Both have a domain organization similar to U2AF\textsuperscript{65}, except for the C-terminus of CAPER\textbeta{} which lacks the UHM domain (Table 1 and Fig. 2A).

The U2AF\textsuperscript{35} protein contains a central UHM domain (previously called \Psi-RRM) involved in the interaction with U2AF\textsuperscript{65}, flanked by two Zn\textsuperscript{2+}-binding motifs and a C-terminal RS domain (Table 2 and Fig. 1). Three-dimensional structural information revealed that, despite low primary sequence identity (23%), recognition of the respective ligands by the U2AF\textsuperscript{35}-UHM and U2AF\textsuperscript{65}-UHM domains is very similar [18]. Both the U2AF\textsuperscript{35}–U2AF\textsuperscript{65} and U2AF\textsuperscript{35}–SF1 interactions involve a critical Trp residue in the ligand sequence which inserts into a tight hydrophobic pocket created by the UHM (Fig. 3).

In the human genome there are at least three genes that encode proteins with a high degree of homology to U2AF\textsuperscript{35} (Table 2 and Fig. 2B). U2AF\textsuperscript{36} (encoded by the U2AF\textit{L}\textsuperscript{36} gene) is a 26-kDa protein bearing strong sequence similarity to U2AF\textsuperscript{35}; the N-terminal 187 amino acids are 89% identical, but the C-terminus of U2AF\textsuperscript{36} lacks the RS domain present in U2AF\textsuperscript{35} [25]. U2AF\textsuperscript{36}R1 (encoded by the U2AF\textit{L}\textsuperscript{1} gene) and

Table 1. Domain organization of U2AF\textsuperscript{65} and U2AF\textsuperscript{65}-related proteins. Domains are annotated as described in [18]; RS, ArgSer rich. The gene names approved by the HUGO Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature) have been included.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Protein</th>
<th>Domain organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>U2AF2</td>
<td>U2AF\textsuperscript{65}</td>
<td>RS ( \rightarrow ) RRM1 ( \rightarrow ) RRM2 ( \rightarrow ) UHM</td>
</tr>
<tr>
<td>SIAHBP1</td>
<td>PUF60</td>
<td>RRM1 ( \rightarrow ) RRM2 ( \rightarrow ) UHM</td>
</tr>
<tr>
<td>RNPC2</td>
<td>CAPER\textalpha{}</td>
<td>RS ( \rightarrow ) RRM1 ( \rightarrow ) RRM2 ( \rightarrow ) UHM</td>
</tr>
<tr>
<td>RBM23</td>
<td>CAPER\textbeta{}</td>
<td>RS ( \rightarrow ) RRM1 ( \rightarrow ) RRM2</td>
</tr>
</tbody>
</table>
smaller in the \textit{U2AF1L4} gene. In addition, the exon sequences of the human and mouse \textit{U2AF1L4} genes are 90\% identical at the nucleotide level, and the majority of the differences are neutral, third-position changes [25]. The evolutionary pattern for \textit{CAPER\beta} is more unusual. Among mammals, orthologs can be found for primates (chimp and rhesus) and domestic animals (dog and cow) but not for rodents. \textit{CAPER\beta} can also be found in \textit{Xenopus tropicalis}, but there is no evidence for its existence in chicken or fish. A comparison of \textit{CAPER\beta} genes from different mammals revealed that most of the exon/intron boundaries are located in the same positions as in the human \textit{CAPER\alpha} gene and the introns are found to be smaller in the \textit{CAPER\beta} gene. Given the similarities between the evolutionary histories of the \textit{U2AF\beta} and \textit{CAPER\beta} genes, it is likely that these new splicing proteins perform unique and lineage-specific functions.

Retrotransposition rather than gene duplication appears to have created the \textit{U2AF1L1} gene less than 100 million years ago. The mouse \textit{U2AF1L1} gene, which is located on chromosome 11, was formed by retrotransposition of \textit{U2AF1L2}, which is located on the X chromosome [36]. \textit{U2AF1L1} is regulated by genomic imprinting [37], and the whole gene is located in an intron of another gene, \textit{Murr1}, that is not imprinted [36]. The retrotransposition that originated the mouse \textit{U2AF1L1} gene must have occurred after mice and humans diverged, because the human ortholog of \textit{Murr1} is located on chromosome 2 and there are no \textit{U2AF1}-related genes on human chromosome 2. Indeed, the phylogenetic analysis of this family of genes indicates independent events of retrotransposition in rodents (mouse and rat) and primates (human and chimp). Similarly to the mouse gene, the human \textit{U2AF1L1} gene located on chromosome 5 is intronless whereas human \textit{U2AF1L2} is multiexonic, suggesting that it also originated by retrotransposition [28]. However, in contrast with the mouse gene, human \textit{U2AF1L1} is not imprinted [38].

**Alternative splicing and diversity of human U2AF proteins**

Our laboratory has recently reported that human transcripts encoding U2AF\textsuperscript{35} can be alternatively spliced giving rise to three different mRNA isoforms called U2AF\textsuperscript{35a}, U2AF\textsuperscript{35b}, and U2AF\textsuperscript{35c} [39]. This discovery raised the question of whether additional U2AF genes produce alternatively spliced mRNAs. Very few
DIVERSITY OF HUMAN U2AF SPlicing FACTORS

Fig. 4. Evolution of U2AF-related proteins. The possible origins of U2AF proteins are shown in relation to key metazoan evolutionary events. Solid lines represent presence of the indicated protein in all species that diverged from humans within the corresponding period of time. Dashed lines represent loss of the indicated proteins in all extant species that diverged from humans within the corresponding period of time. Dashed-dotted lines represent lineage-specific loss/preservation or appearance/absence of the indicated protein in species that diverged from humans within the corresponding period of time (ex. CAPERβ apparently disappeared from fish, birds and rodents but remained in Xenopus and some mammals; U2AF59R1 results from independent retrotranscription events affecting only primates and rodents. A star indicates that U2AF59, U2AF59, PUF60 and CAPERα genes are duplicated in teleosts, most probably as a consequence of the whole-genome duplication that occurred in ray-finned fish ~380 million years ago (Mya)).

examples of U2AF mRNA isoforms have been described in the literature. Namely, two CAPERβ mRNAs and four CAPERα mRNAs were detected in several human tissues by northern blotting [24], and a splicing variant of PUF60-FIR was identified in colorectal cancers [40]. This rarity of data prompted us to use bioinformatic search strategies to investigate alternative splicing of U2AF and U2AF-related genes. This analysis was carried out with the aid of the UCSC Genome Browser (http://genome.ucsc.edu/) [41] for the human genome assembly hg17, May 2004, NCBI Build 35. Gene regions of interest were defined by the BLAT mapping [41] of the available RefSeq transcript (RNA) sequences [42] (http://www.ncbi.nlm.nih.gov/projects/RefSeq/) for a particular gene. Using the UCSC Table Browser [43], we obtained the tables for the BLAT mappings of mRNAs and ESTs for this gene region. Making allowance only for GT_AG, GC_AG or AT_AC splice site consensus and excluding isoforms with extensive intron retentions, the non-redundant set of longest isoforms and corresponding accessions was determined. The splicing patterns obtained were cross-checked with two alternative splicing databases: the ASAP (http://bioinfo.mbi.ucla.edu/ASAP/); and the Hollywood RNA Alternative Splicing Database (http://hollywood.mit.edu).

Our analysis revealed that, with the single exception of the U2AF59L1 gene, which is devoid of introns, all genes coding for U2AF and U2AF-related proteins can be alternatively spliced (Table 3). Many alternatively spliced mRNA isoforms are predicted to contain premature stop codons and are therefore expected to be targeted for degradation by nonsense-mediated decay, as already demonstrated for U2AF35c (corresponding to RefSeq mRNA NM_001025204 in Table 3). In addition, we found evidence for several transcripts that could generate functional protein isoforms containing the conserved RRM motifs characteristic of each protein family (Table 3). Variations in activity are expected from changes in domain structure predicted for some of these isoforms, but further experimental studies are needed to address this view.

Perspectives: evolution of U2AF functions

After the discovery that U2AF59 is required to reconstitute mammalian splicing in vitro [6–8], the protein
Table 3. Predicted number of mRNA isoforms generated by alternative splicing of U2AF genes. An alternatively spliced mRNA isoform was considered confirmed if its corresponding protein sequence was annotated in RefSeq or SwissProt databases. A splicing pattern observed in an mRNA or EST was predicted to produce a premature coding sequence termination if it contained an in-frame stop codon within an internal exon. For the predicted patterns of splicing, there is redundancy in the number of accessions shown because of the fragmented nature of ESTs and some mRNAs.

<table>
<thead>
<tr>
<th>Protein (gene symbol)</th>
<th>Confirmed mRNA isoforms (accessions)</th>
<th>Predicted splicing patterns producing a premature stop codon (accessions)</th>
<th>Predicted splicing patterns of candidates for putative novel protein (accessions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U2AF&lt;sup&gt;SR&lt;/sup&gt;</td>
<td>2 (NM_007279.2, NM_00112478.1)</td>
<td>2 (CD524035.1, CR960251.1, CA488304.1)</td>
<td>2 (CD524035.1, BR034022.1)</td>
</tr>
<tr>
<td>U2AF&lt;sup&gt;2&lt;/sup&gt;</td>
<td>4 (NM_014281.3, NM_078430.1,</td>
<td>0</td>
<td>10 (BI015396.1, AL522753.3, AL514886.3, BX342032.1, AK065641.1, BO421738.1,</td>
</tr>
<tr>
<td></td>
<td>BC009734.1, BC011265.1)</td>
<td></td>
<td>BO96878.1, BG311523.8, BE303389.1, BU170641.1)</td>
</tr>
<tr>
<td>IS/AHBP1</td>
<td>5 (NM_184241.1, NM_004022.1,</td>
<td>5 (NM_184241.1, NM_184244.1, BC107866.1, BM480781.1, BE68688.1, DA15491.1,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NM_184241.1, NM_184244.1, NM_184237.1)</td>
<td></td>
<td>CA193145.1, DA37288.1, BF362717.1, D007200.1, D818523.1, BG764890.1,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DA922441.1, AV903266.1, AL183063.3)</td>
</tr>
<tr>
<td>CAPER&lt;sub&gt;a&lt;/sub&gt;</td>
<td>4 (NM_018107.3, CR596428.1, BX161440.1,</td>
<td>10 (DA582179.1, DM164389.1, BM48794.1, DA145418.1, BB3360.1, DB66416.1,</td>
<td></td>
</tr>
<tr>
<td>RNP&lt;sub&gt;C2&lt;/sub&gt;</td>
<td>AL534196.1)</td>
<td></td>
<td>AA23094.1, B1915247.1, DA929707.1, DA023922.1, CN483101.1, BX165727.1,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>BC100612.1)</td>
</tr>
<tr>
<td>CAPER&lt;sub&gt;b&lt;/sub&gt;</td>
<td>8 (NM_001025204.1, BE376536.1)</td>
<td>1 (BM69685.1, BM70076.1, AW274286.1, DB127360.1, BU52379.1, AA45888.1,</td>
<td>1 (BE65844.1, BM69685.1, BM907067.1, AW274286.1,</td>
</tr>
<tr>
<td>RRM&lt;sub&gt;23&lt;/sub&gt;</td>
<td>3 (NM_007279.2, NM_00112478.1,</td>
<td></td>
<td>DB127360.1, BU52379.1, AA45888.1, BU606847.1,</td>
</tr>
<tr>
<td></td>
<td>NM_001025204.1)</td>
<td></td>
<td>DB333076.1, BF821614.1)</td>
</tr>
<tr>
<td>U2AF&lt;sup&gt;SR&lt;/sup&gt;</td>
<td>5 (BM69685.1, BM70076.1, AW274286.1,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U2AF&lt;sup&gt;1&lt;/sup&gt;</td>
<td>2 (NM_144987.2, NM_001040425.1)</td>
<td></td>
<td>6 (BM69685.1, BM70076.1, AW274286.1, DB127360.1, BU52379.1, AA45888.1,</td>
</tr>
<tr>
<td>U2AF&lt;sup&gt;SR&lt;/sup&gt;</td>
<td>1 (BM85388.2)</td>
<td></td>
<td>BF700767.1, AW274286.1, DB127360.1, BU52379.1, AA45888.1, BU606847.1,</td>
</tr>
<tr>
<td>R2</td>
<td></td>
<td></td>
<td>DB333076.1, BF821614.1)</td>
</tr>
<tr>
<td>U2AF&lt;sup&gt;SR&lt;/sup&gt;</td>
<td></td>
<td></td>
<td>6 (BM69685.1, BM70076.1, AW274286.1, DB127360.1, BU52379.1, AA45888.1,</td>
</tr>
<tr>
<td>R2</td>
<td>4 (NM_007279.2)</td>
<td></td>
<td>BF700767.1, AW274286.1, DB127360.1, BU52379.1, AA45888.1, BU606847.1,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DB333076.1, BF821614.1)</td>
</tr>
</tbody>
</table>
DIVERSITY OF HUMAN U2AF SPlicing FACTORS

was shown to be highly conserved and its homologs are essential in *Sch. pombe* [32], *D. melanogaster* [29] and *C. elegans* [10]. Although it remains an open question whether U2AF<sup>65</sup> performs other functions in the cell in addition to its fundamental role in pre-mRNA splicing, the U2AF<sup>65</sup>-related proteins are clearly implicated in both splicing and transcription. In particular, CAPER (also known as CC1.3) was independently identified as a protein that interacts with the estrogen receptor and stimulates its transcriptional activity [44], and purified as a splicesome component capable of affecting the splicing reaction [45–47]. More recently, an additional related protein was identified, CAPERβ, and both CAPER (renamed CAPERα) and CAPERβ were shown to regulate transcription and alternative splicing in a steroid hormone-dependent manner [24]. Importantly, both CAPERα and CAPERβ are expressed at higher levels in the placenta and liver, two tissues with active steroid hormone signaling. According to one possible model, the CAPER proteins interact first with transcription factors to stimulate transcription in response to steroid hormones; by interacting with promoter-bound transcription factors, the CAPER proteins can be incorporated into the pre-initiation complex and thereby have direct access to the nascent RNA transcript; the CAPER proteins may then interact with splicing factors required for early recognition of the 3′ splice site and thereby influence the commitment to splicing [24].

Human PUF60 was originally identified as a Py-tract-binding protein that is required, together with U2AF, for efficient reconstitution of RNA splicing *in vitro* [23]. Around the same time, the protein was also identified as a modulator of TFIIH activity and named FIR (FUSE-binding protein-interacting repressor) [48]. An interaction between PUF60/FIR and the TFIIH/p89/XBP helicase was found to repress *c-myc* transcription, and enforced expression of FIR induced apoptosis. Interestingly, a splicing variant of FIR was detected in human primary colorectal cancers, and recent data suggest that this variant may promote tumor development by disabling FIR repression of *c-myc* and opposing apoptosis [40]. Unlike the CAPER proteins, PUF60/FIR (similarly to U2AF<sup>65</sup>) is expressed in most tissues [24], as predicted for a constitutive splicing factor. Yet, the *Drosophila* ortholog of human PUF60, *Half Pint*, was found to function in both constitutive and alternative splicing in *in vivo* [49], raising the question of whether human PUF60 regulates alternative splicing. It is also unknown whether the dual function of PUF60 on transcription and splicing is coupled as in the case of the CAPER proteins or whether PUF60 affects independently the transcription and splicing of distinct genes. Although answers to these and other questions are likely to provide new clues to understanding the functional diversity of U2AF<sup>65</sup>-related proteins, we may speculate that these proteins evolved in response to a requirement for the co-ordination of the multiple steps of gene expression in complex organisms. As mRNA biogenesis became progressively more targeted for regulation, new sequence characteristics developed to allow the same molecule to engage in sequential transcriptional and splicing events, acting as coupling proteins in regulated gene expression. In agreement with this view, several other proteins related to the SR-family of splicing factors have also been associated with the coupling of transcription and splicing [50].

In contrast with U2AF<sup>65</sup>-related proteins, there is no evidence implicating the U2AF<sup>65</sup>-like proteins in any process other than splicing. Unlike U2AF<sup>65</sup>, which is essential for splicing, U2AF<sup>65</sup> is dispensable for the *in vitro* splicing of some model pre-mRNAs containing strong Py-tracts (i.e. a stretch of pyrimidines beginning at position −5 relative to the 3′ splice site and extending 10 or more nucleotides upstream into the intron) [5]. The presence of U2AF<sup>65</sup> and its interaction with U2AF<sup>65</sup> was, however, found to be essential for *in vitro* splicing of a pre-mRNA substrate with a Py-tract that deviates from the consensus [51]. Introns with nonconsensus or weak Py-tracts were previously called ‘AG-dependent’ [52]. Biochemical complementation experiments performed with extracts depleted of endogenous U2AF demonstrated that splicing of AG-dependent introns was rescued only when both U2AF subunits were added and not with U2AF<sup>65</sup> alone [11,51,53]. However, more recent work indicates that several splicing events assumed to depend critically on U2AF<sup>65</sup> did not show any defect under conditions of limited U2AF<sup>65</sup> availability *in vivo* [54,55]. Thus, the distinction between U2AF<sup>65</sup>-dependent and independent introns remains an unsolved issue.

The importance of the small subunit of U2AF *in vivo* was first shown by the finding that the *D. melanogaster* ortholog of human U2AF<sup>65</sup> (dU2AF<sup>65</sup>) is essential for viability [30]. Orthologs of U2AF<sup>65</sup> are also essential for the viability of the fission yeast *Sch. pombe* [33] and the nematode *C. elegans* [56] and for the early development of zebrafish [57]. Additional studies in both *Drosophila* and human cells further provided hints of a role for U2AF<sup>65</sup> in splicing regulation. First, loss-of-function mutations in dU2AF<sup>65</sup> affected splicing of the pre-mRNA encoding the female-specific RNA-binding protein Sex-l,ethal [58]. Second, depletion of dU2AF<sup>65</sup> by RNA interference (RNAi) affected alternative splicing of the *Drosom* gene...
transcript [59]. Third, RNAi-mediated depletion of both U2AF35a and U2AF25b isoforms in HEK293 cells altered alternative splicing of Cdc25 transcripts [55]. Sequence comparisons of U2AF35 splicing isoforms and U2AF15-related proteins revealed striking conservation of the principal signature features of UHMs (Fig. 3). Moreover, there is biochemical evidence indicating that both U2AF35a and U2AF25b splicing isoforms, U2AF35 and U2AF15-R2-Urp, can interact with U2AF65 [25,26,39]. U2AF35-R2-Urp was further shown to be functionally distinct from U2AF35 because U2AF35 cannot complement Urp-depleted extracts [26]. It was therefore proposed that the U2AF65 subunit may form diverse heterodimers with the different U2AF35-related proteins, each of them with distinct functional activities.

Many splicing regulators are thought to direct changes in the choice of splice sites by preventing the initial binding of U1 snRNP and U2AF in the early steps of spliceosome assembly [60]. Recently, the well-characterized splicing regulator poly pyrimidine tract-binding protein (PTB) was shown to repress excision of an alternatively spliced exon by preventing the 5' splice site-dependent assembly of U2AF on the 3' splice site [61]. Thus, it is possible that different U2AF variants provide a means for flexible regulation involving tissue-specific splicing choices determined by regulators such as PTB. In this regard it is noteworthy that splicing isoform U2AF35a is 9-18-fold more abundant than U2AF25b, with distinct tissue-specific patterns of expression [39], and in the mouse, the U2AF15 gene is expressed predominantly in the brain especially in the pyramidal neurons of the hippocampus and dentate gyrus [62,63]. Identifying the functional uniqueness of each U2AF35-related protein is clearly an important challenge for future research.

Concluding remarks

New biological functions are often acquired through gene duplication events, followed by the evolution of specialized gene functions, as well as by the creation and loss of different exons. Both the emergence of additional genomic copies by gene duplication and retrotransposition, and an increase in transcript diversity by alternative splicing have contributed to the generation of new U2AF-related proteins. The similarity and differences between the U2AF-related proteins imply that they have evolved distinct functions in relation to the control of gene expression in complex organisms. Clues to the biological processes in which these proteins participate may be obtained by determining their tissue expression patterns, elucidating their RNA-bind-

Acknowledgements

We thank Ben Blencowe and Margarida Gama-Carvalho for critical reading of the manuscript. This work was supported by grants from Fundação para a Ciência e Tecnologia (FCT), Portugal (POCTI/MGI/49430/2002, SFRH/BD/2914/2000), the Muscular Dystrophy Association (MDA3662), and the European Commission (EURASNET, LSHG-CT-2005-518238).

References

DIVERSITY OF HUMAN U2AF SPlicing FACTORS


DIVERSITY OF HUMAN U2AF SPlicing FACTORS


2.2 Scope of this thesis

2.2.1 Objectives

The biochemical mechanisms that control splice-site usage and, therefore, alternative splicing are complex and in large part remain poorly understood. The heterodimeric splicing factor U2AF, one of the first non-snRNP essential splicing factors to be identified, is acknowledged to play an important role in the early steps of spliceosome assembly. The recent findings that U2AF is part of an evolutionary conserved family of splicing factors which share high structural homology argues that these proteins may have evolved specific roles in the control of gene expression in multicellular organisms.

The main goal of this work was to elucidate how the diversity of the U2AF\textsuperscript{35}-family of splicing factors might contribute to the regulation of eukaryotic gene expression. My studies focused on the Zrsr1 protein, for which at the beginning of this work very little was known. In our laboratory, another PhD student (Ana Rita Grosso) used bioinformatic tools to analyze publicly available microarray datasets in order to determine if any of the U2AF\textsuperscript{35}-family members showed tissue-specific expression. This task was considered to be the starting point of this work since the current model of alternative splicing regulation has been associated with the differential expression of splicing factors in a tissue-specific fashion. In fact, some alternative splicing events are known to be dependent in the activity/concentration of some splicing factors and, according to this model, the differential expression of the U2AF\textsuperscript{35}-family could be involved in the regulation of tissue-specific RNA targets.

As described in Chapter 3, I found that the cellular levels of Zrsr1 mRNA and protein specifically increase during erythroid differentiation, and this reflects an up-regulation of transcriptional activity. I have further shown that Zrsr1 interacts directly with U2AF\textsuperscript{65} and associates with spliceosomal components suggesting that it acts as a splicing factor. Finally, I found that Zrsr1 is required for normal erythropoiesis in a knock-out mouse model, by regulating erythroid-specific splicing events.
Chapter 3

The retrotransposed mouse Zrsr1 gene acquired a new function in erythroid cells
Chapter 3

3.1 The retrotransposed mouse Zrsr1 gene acquired a new function in erythroid cells

3.1.1 Summary

In multicellular organisms alternative splicing provides a versatile mechanism to control gene expression during cell differentiation and development. Splicing is regulated by protein factors, many of which are members of multigene families. Here we characterize the properties of Zrsr1 (previously termed U2AF35-RS1), a member of the U2AF family of splicing factors that evolved after a recent retrotransposition event. We report that expression of Zrsr1 is specifically up-regulated during erythroid differentiation and that mice deficient of Zrsr1 have smaller red blood cells. We further show that Zrsr1 interacts directly with U2AF65 and controls erythroid-specific splicing decisions. Taken together our results suggest that in erythroid cells Zrsr1 competes with U2AF35 for binding to U2AF65 and thereby regulates alternative splicing events required for normal erythropoiesis.
3.1.2 Introduction

In higher eukaryotes, most protein coding genes are synthesized in the nucleus as precursor messenger RNA molecules (pre-mRNAs) that are extensively modified to produce mature mRNAs used as templates for protein production in the cytoplasm (Sharp 2005). Among other modifications, non-coding intervening sequences (introns) are removed from pre-mRNA molecules by a process termed RNA splicing. Splicing provides a versatile mechanism of genetic regulation that allows a single gene to generate multiple mRNA species through the use of alternative splice sites, and in human cells nearly all genes produce alternatively spliced mRNA products in a tissue-specific manner (Pan, Shai et al. 2008; Wang, Sandberg et al. 2008).

Splicing is carried out by the spliceosome, a highly dynamic macromolecular machine that consists of five uridine-rich small nuclear ribonucleoprotein particles (the U1, U2, U4/U6 and U5 snRNPs) and over 100 non-snRNP auxiliary proteins (Wahl, Will et al. 2009). Introns are demarcated by short consensus sequences that are recognized by base pairing with the spliceosomal snRNAs. However, these core sequences are generally poorly conserved in mammals and lack sufficient information for the spliceosome to distinguish correct from cryptic splice sites. There are additional cis-acting sequence motifs in the pre-mRNA (referred to as exonic/intronic splicing enhancers or inhibitors) that promote splice-site recognition and play a critical role in the regulation of alternative splicing (Barash, Calarco et al. 2010). Alternatively spliced mRNA products result from using one splice site over another. These decisions can be modulated by the binding of protein factors that enhance or repress the assembly of a functional spliceosome at a particular splice (Nilsen and Graveley 2010). Many splicing factors are members of multigene families (Barbosa-Morais, Carmo-Fonseca et al. 2006), and their expression differs from tissue to tissue (Grosso, Gomes et al. 2008).

The recognition of where an intron ends (i.e., the 3' splice site) is primarily carried out by an essential and highly conserved splicing factor termed U2-associated factor or U2AF (Ruskin, Zamore et al. 1988; Zamore and Green 1989). U2AF is a heterodimer composed of a large and a small subunit. The large subunit (U2AF65 in mammals) binds to the polypyrimidine tract at the 3' end of the intron (Zamore and
Green 1989), while the small subunit (U2AF$^{35}$ in mammals) binds tightly to U2AF$^{65}$ and interacts with the AG dinucleotide at the 3' splice site (Merendino, Guth et al. 1999; Wu, Romfo et al. 1999; Zorio and Blumenthal 1999). U2AF$^{65}$ additionally interacts cooperatively with the branch point adenosine-binding protein SF1/BBP (Kent, Reayi et al. 2003; Selenko, Gregorovic et al. 2003). Assembly of the U2AF complex at the 3' end of the intron allows subsequent recruitment of the U2 snRNP, thus promoting spliceosome assembly.

U2AF has been highly conserved from Schizosaccharomyces pombe to humans. Yet, the genomes of metazoans contain a family of additional genes that are highly related to those coding for the large and small U2AF subunits (Mollet, Barbosa-Morais et al. 2006). Mammalian genomes contain at least three genes that encode proteins with a high degree of homology to U2AF$^{35}$ (Mollet, Barbosa-Morais et al. 2006). Murine U2AF$^{35}$ protein is encoded by the $U2af1$ gene and U2AF$^{35}$-related proteins include U2AF$^{26}$ (encoded by the $U2af1l4$ gene), U2AF$^{35}$-RS1 and U2AF$^{35}$-RS2/Urp (encoded by the $Zrsr1$ and $Zrsr2$ genes, respectively).

The initial characterization of the human $U2af1$ gene has identified three alternative spliced transcripts encoding different U2AF$^{35}$ isoforms (U2AF$^{35}$a–c). While UAF$^{35}$c contains a PTC that triggers the resulting mRNA to decay, U2AF$^{35}$b was found to code 7 different aminoacids located at the RRM domain (Pacheco, Gomes et al. 2004). Nevertheless, U2AF$^{35}$b seems to preserve the ability to interact with the large subunit, stimulate U2AF$^{65}$ binding to a pre-mRNA and promote the U2AF splicing activity in vitro. Like the orthologs in D. Melanogaster (Rudner, Kanaar et al. 1996), S.pombe (Wentz-Hunter and Potashkin 1996) and C.elegans (Zorio and Blumenthal 1999), the human U2AF$^{35}$ was also shown to be essential for viability(Pacheco, Moita et al. 2006).

The U2AF$^{26}$ (~26 KDa) protein shares strong homology with U2AF$^{35}$, but lacks the C-terminal RS domain (Shepard, Reick et al. 2002). This splicing factor was found to replace U2AF$^{35}$ in constitutive and enhancer-dependent splicing through the direct interaction with U2AF$^{65}$, enhancing its binding to weak Py-tracts (Shepard, Reick et al. 2002). A comparison of the RNA binding specificities of U2AF$^{26}$ and U2AF$^{35}$ demonstrated that U2AF$^{26}$ binds preferentially to 3’splice sites with AG/C or AG/A nucleotides (Shepard, Reick et al. 2002) while U2AF$^{35}$ was found to bind with higher affinity to the 3’consensus splice site AG/G (Wu, Romfo et al. 1999). Interestingly,
AG/C and AG/A 3′ splice sites have been reported to be enriched in alternatively spliced exons in tissues such as brain or muscle (Stamm, Ben-Ari et al. 2005), which could point to a role of U2AF26 in regulating tissue-specific alternative splicing events. More recently, Heyd and co-workers proposed a model where U2AF26 and the transcription factor Gfi1 act antagonistically in the alternative splicing regulation of the Ptprc gene (which encodes the CD45 phosphatase). By either promoting or inhibiting the formation of the CD45RO isoform, U2AF26 was described to be involved in the regulation of antigen-dependent T cell activation (Heyd, ten Dam et al. 2006). The U2AF1L4 gene was also found to be itself submitted to alternative splicing, giving rise to a splice variant lacking exon 7 (U2AF26ΔE7) (Heyd, Carmo-Fonseca et al. 2008). In contrast to the nuclear localization of U2AF26, U2AF26ΔE7 was found present in the cytoplasm since this isoform lacks the nuclear localization signal (NLS). Such regulation may represent an independent control of the intracellular distribution and availability of U2AF26 to the splicing machinery which, in combination with the differential expression of the U2AF-family members, represents another layer of regulation in the control of alternative splicing events (Heyd, Carmo-Fonseca et al. 2008).

The biochemical data available suggests that ZRSR2 (also known as U2AF1-RS2-Urp) and U2AF35 are non-redundant proteins since both were shown to interact with U2AF65 and other SR proteins in a functionally distinct way (Tronchere, Wang et al. 1997). In fact, ZRSR2 seems to play a different role since in vitro data using a model pre-mRNA demonstrates that ZRSR2 depleted nuclear extracts are not complemented with recombinant U2AF. Unexpectedly the ZRSR2-U2AF65 complex was found to interact with U2AF35 and it is proposed that ZRSR2 may belong to a larger U2AF complex that could engage network interactions during spliceosome assembly (Tronchere, Wang et al. 1997). Although the initial reports have found ZRSR2 protein involved in splicing of a model major U2-inton (Tronchere, Wang et al. 1997), proteomic analysis identified ZRSR2 as a protein associated with the human 18S U11/U12 snRNP (Will, Schneider et al. 2004), suggesting that this protein may be involved in the in splicing of U12-type introns. Indeed, ZRSR2 was found to be recruited to a model U12-type intron 3′ splice site, where it promotes spliceosome assembly (Shen, Zheng et al. 2010). Additionally, this report also confirmed that this splicing factor also contacts the 3′ splice site of a U2-type intron, although in this case ZRSR2 was found to be specifically required for the second step of splicing. Thus, through recognition of a
common splicing element, ZRSR2 is now acknowledged to facilitate distinct steps of U2- and U12-type intron splicing (Shen, Zheng et al. 2010).

Zrsr1 was cloned by Mukai and co-workers while searching for new imprinted genes with parental-origin-specific CpG methylations (Hatada, Sugama et al. 1993). Initially named SP2, this intronless gene was named Zrsr1 (or U2af1-rs1) due to the significant sequence homology with U2AF35-like proteins. Zrsr1 was mapped into the mouse chromosome 11 (Hayashizaki, Shibata et al. 1994; Tada, Tada et al. 1994) and it was found to be expressed exclusively from the paternally inherited chromosome since the promoter of this gene was shown to be hyper-methylated in the maternal allele (Hatada, Kitagawa et al. 1995). Nevertheless, to date, the biological function of Zrsr1 is presently unknown.

Despite intense research, the mechanisms leading to splice site recognition are not fully understood, although the unique behaviour of several members of the U2AF35 family seems to indicate that some of them may be important targets for regulation. How members of this protein family acquired distinctive new functions remains unclear. Here, we characterize the properties of Zrsr1.
3.2 Materials and Methods

3.2.1 Microarray data sets and analysis

Expression profiles of U2AF$^{65}$ and U2AF$^{35}$-related genes were obtained from publicly available microarray data for differentiation processes (Grosso, Gomes et al. 2008) and humans and mouse tissues (http://www.affymetrix.com/support/technical/sample_data/). All the microarray data analysis was done using R and several packages available from CRAN (R Development Core Team, 2010) and Bioconductor (Gentleman et al 2004): affy (Gautier et al 2004), aroma.affymetrix (Bengtsson et al 2008), limma (Smyth et al 2005) and gplots (Gregory et al 2009).

3.2.2 Cell Culture and Transfection Assays

I/11 cells were grown in StemPro-34™ (Life Technologies) as described previously (Dolznig, Boulme et al. 2001; von Lindern, Deiner et al. 2001). For expansion, the medium was complemented with 0.5U/mL Epo (Ortho-Biotech, Netherlands), 100ng/mL SCF and 10$^{-6}$M dexamethasone (Sigma-Aldrich). To induce differentiation growth medium was supplemented with 5U/mL and 0.5mg/mL iron-loaded Transferrin (Intergene). MEL C88 cell line was grown in DMEM-GlutaMAX™ (Invitrogen) supplemented with 10% (v/v) fetal calf serum (FCS) (Invitrogen) and differentiation was induced by addition of 2% (v/v) DMSO (Sigma), 5% (w/v) BSA (Sigma) and 1.8x10$^{-3}$ mM Iron Dextran (Sigma), as described previously (Volloch and Housman 1982; Patel and Lodish 1987). Human embryonic kidney (HEK) 293T cells were culture in DMEM supplemented with 10% (v/v) FCS and 2mM L-glutamine. Transient transfection of HEK293T cells was performed using Lipofectin® (Invitrogen), according to the manufacturer’s instructions.

3.2.3 RT-PCR and Real-Time Quantitative PCR

RT-PCR reactions were random or OligodT primed and cDNA was produced using Superscript II Reverse Transcriptase (Invitrogen) according to the manufacturer’s instructions. PCR products were separated by agarose gel electrophoresis and detected by GelRed™ staining. All primers were designed using Primer3 software.
and gene-specific primers pairs are presented in Supplementary Material. Real-time Quantitative PCR analysis was performed in the 7500 Fast Real-Time PCR System (Applied Biosystems, Foster City, CA, USA), using SYBR Green PCR master mix (Applied Biosystems). The relative expression of each gene was calculated using a derivative of the $2^{-\Delta\Delta CT}$ method as described previously (Schmittgen, Teske et al. 2003).

### 3.2.4 Immunoblotting

Total cell protein extracts were prepared by incubating cells for 10 min at room temperature into SDS-PAGE sample buffer (62.5 mM Tris-HCl, pH 6.8, 2% SDS, 5% β-mercaptoethanol, 10% glycerol, 0.01% bromophenol blue) with 200 U/mL benzonase (Sigma-Aldrich), and then boiling for 5 min. Protein extracts were separated on 10% SDS-polyacrylamide minigels (BioRad Laboratories, Richmond, California) and transferred to nitrocellulose membranes. Western blotting was carried out following standard procedures. Membranes were washed in PBS, blocked with PBS containing 5% low fat milk for at least 1 h and incubated with specific primary antibodies diluted in PBS with 2.5% low fat milk at 4°C overnight. Membranes were then washed 3 times for 15 minutes in PBS-Tween 0.5% (v/v), incubated with appropriated secondary antibodies conjugated with horseradish peroxidase (BioRad Laboratories, Richmond, CA) and developed using the ECL chemiluminescence reaction (Amersham Buchler GmbH, Braunschweig, Germany).

### 3.2.5 Gene constructs

Mouse Zrsr1 cDNA was PCR amplified from cDNA produced from mouse Universal RNA (Stratagene) with the following primers: primer 1- 5’-atggcatcaagcagacgaggatt-3’ and primer 2- 5’-tcaggttctgtggctctggct -3’. The PCR product obtained was then cloned into pCR®2.1-TOPO® (Invitrogen) using the TOPO-TA cloning Kit® (Invitrogen). The mouse HAZrsr1-6xHis cDNA was then PCR cloned into pcDNA3 digested with *BamHI* and *XhoI*, using the following primers: primer 1- 5’-aaaaagatccatgtaccatacgatgtccagattacgctatggcatcaagcagacgaggattcct-3’ (*BamHI* site
underlined and HA-tag sequence in bold), and primer 2 5’-aaaaactcagatgtgatggtgatggtgatggtttctgtggctctggctttgtggac-3’ (XhoI site underlined and 6xHis sequence in bold). The HAU2AF35 insert was obtained from pTRE-HAU2AF35 (Pacheco, Coelho et al. 2006) and subcloned into pcDNA3 by restriction digestion with BamHI and EcoRI. The human U2AF26 cDNA was obtained from the I.M.A.G.E. clone 8992197 and used as template for PCR cloning into pGEX-4T-3 vector. PCR amplification was performed with the oligonucleotides, primer 1- 5’-aaagatcagagatgaatagcttcagat-3’ (BamHI site underlined) primers 2- 5’-ctcgag tcagaagcggccatgccagtg-3’ (XhoI site underlined), the product was digested with BamHI and XhoI and cloned into pGEX-4T-3 linearized with the same enzymes. pGEX-4T-3 U2AF35 was obtained by restriction digestion of GFP-U2AF35 with EcoRI and cloned into EcoRI linearized pGEX-4T-3 vector. Correct insert orientation was checked by DNA sequencing. PGEXU2AF65 was obtained from (Gama-Carvalho, Carvalho et al. 2001). The EGFP cDNA was PCR amplified from pEGFP-C1 vector (Clontech) with primers: primer1-5’-aaaaactcagatgtgatggtgatggtgatggtttctgtggctctggctttgtggac-3’ (BamHI site underlined and Ha-Tag sequence in bold), primer2- 5’-aaaaactcagatgtgatggtgatggtgatggtttctgtggctctggctttgtggac-3’ (XhoI site underlined and 6xHis-Tag sequence in bold) and cloned into pcDNA3 digested with BamHI and XhoI.

For lentivirus expression, mouse Zrsr1 (see chapter 3) and HAGFP-6xHis cDNA’s were subsequently subcloned from the pcDNA3 plasmid through digestion with BamHI and XhoI, and inserted into pcSC_IRES_ZsGreen vector (kindly provided by Dr. Marieke von Lindern, ErasmusMC, Netherlands) digested with BamHI and XhoI.

3.2.6 Antibodies

The following primary antibodies were used: rabbit polyclonal antibodies anti-RNA Pol II, N20 (Santa Cruz Biotechnology), anti-histone H3 (ab1791; Abcam), anti-H3K36me3 (ab9050; Abcam), anti-H3K9ac (ab10812; Abcam), anti-HA (Y11; Santa Cruz Biotechnology), anti-U2AF35 (Proteintech), anti-U2AF26 (kindly provided by Dr. Florien Heyd); anti-Zrsr1 (against C-terminal peptide N-GREEDSSPGPQSHTR-C, Pickcell Laboratories BV, The Netherlands), anti-ASF/SF2, anti-SF1 (kindly provided by Dr. Reinhardt Lührmann, Max Planck Institute for Biophysical Chemistry, Göttingen,
Germany), anti-Erk2 (ab17942; Abcam); and mouse monoclonal antibodies directed against U2AF\textsuperscript{65} (MC3, (Gama-Carvalho, Krauss et al. 1997)), green fluorescent protein (GFP; anti-GFP clones 7.1 and 13.1; Boehringer Mannheim).

3.2.7 Chromatin immunoprecipitation

ChIP was performed according to the method described previously (Listerman, Sapra et al. 2006) with some modifications. Cell extracts were sonicated with a Sanyo Soniprep 150 at an amplitude of 10 microns with six bursts of 20 seconds, resulting in 200–400 bp chromatin fragments. DNA fragments crosslinked to proteins were enriched by immunoprecipitation with anti-RNA Pol II (N20, Santa Cruz Biotechnology) and protein A sepharose beads (Sigma). Control (mock) immunoprecipitations were performed with anti-HA antibody (Santa Cruz Biotechnology). DNA from immunoprecipitated and input samples was extracted with UltraPure Phenol:Chloroform:Isoamyl Alcohol 25:24:1 (Invitrogen) and analyzed by quantitative real-time PCR, with the input consisting of a chromatin amount equivalent to that used for immunoprecipitation. The total number of cells was kept constant between experimental batches in order to yield similar amount of input chromatin and therefore avoid variability due to changes in precipitation efficiency (Dahl and Collas 2008; Dahl and Collas 2008). Primers used for quantitative real-time PCR are listed in Supplementary Table SIII. The relative occupancy of the immunoprecipitated protein at each DNA site was estimated as follows: \(2^{(Ct_{\text{mock}} - Ct_{\text{specific}})}\), where \(Ct_{\text{mock}}\) and \(Ct_{\text{specific}}\) are mean threshold cycles of qPCR done in triplicate on DNA samples from mock and specific immunoprecipitations (Nelson, Denisenko et al. 2006). For the histone mark ChIP experiments, the primers used for quantitative real-time PCR are listed in Supplementary Table SV. The fold-enrichment of the immunoprecipitated protein at each DNA site was estimated as follows: \(2^{(Ct_{\text{H3total}} - Ct_{\text{specific}})}\), where \(Ct_{\text{H3total}}\) and \(Ct_{\text{specific}}\) are mean threshold cycles of qPCR done in triplicate on DNA samples from anti-H3 total and specific immunoprecipitations (Nelson, Denisenko et al. 2006). The anti-H3 total ChIP signal was used to normalize H3K36me3 and H3K9Ac values, thus correcting for differences in nucleosome occupancy. An additional control for each specific immunoprecipitation was made with a pair of primers amplifying an intergenic
region on chromosome 16 where no annotated genes could be found (Lin, Coutinho-Mansfield et al. 2008).

3.2.8 Expression of recombinant U2AF\textsuperscript{35}-family members in \textit{E. coli}

PGEX-4T-3 constructs were transformed into chemically competent \textit{E. coli} BL21 (DE3) cells (Invitrogen) by heat-shock treatment. One mL of overnight culture, grown at 37°C in Luria-Bertani (LB) broth containing 100 μg/mL ampicillin, was used to inoculate 100 mL culture of the same mixture. Cells were grown to an OD\textsubscript{600} (optical density at λ = 600 nm) of 0.6 or 1.2, and recombinant protein expression was induced with isopropyl-β-D-thiogalactopyranoside (IPTG) to a final concentration of 0.3 mM. GST-U2AF\textsuperscript{35}, GST-U2AF\textsuperscript{65} and GST-U2AF\textsuperscript{26} protein expression was performed at 37°C and induced cells were allowed to grow for 6h. For GST-Zrsr1, \textit{E. coli} BL21 cells were induced at OD\textsubscript{600} = 1.2 and grown overnight at 18 °C. Cells were harvested by centrifugation at 4000 rpm for 10 min at 4°C, frozen in liquid nitrogen and store at -80°C until use. Thawed cells were resuspended in 20 mM Na\textsubscript{2}PO\textsubscript{4}, 150 mM NaCl, 1mM DTT plus complete protease inhibitors cocktail (Roche), pH 7.2 (lysis buffer), submitted to 5 cycles of sonication (30s each) and the cell extracts were pre-cleared by centrifugation at 14000 rpm for 30 minutes. After clarification, the GSTU2AF fusion protein was purified by glutathione affinity chromatography using a AKTA Explorer FPLC system (GE Healthcare). A glutathione–Sepharose 4B affinity column (GSTrap FF 1mL; GE HealthCare) was equilibrated with 20 mM Na\textsubscript{2}PO\textsubscript{4}, 150 mM NaCl, 1mM DTT , pH 7.2 (Buffer A) and clarified cell lysate was subsequently loaded onto the column at 0.5 mL/min. The column was then washed with at least 10 bed volumes of buffer A at 1mL/min flow rate and the bound protein was subsequently eluted with 10 mM reduced glutathione (Sigma) in 50 mM Tris buffer (pH 8.0). Fractions containing protein were pooled and concentrated by ultrafiltration and store at -80°C until use. Purity of the eluted proteins was accessed by SDS-PAGE and western blot.
3.2.9 Expression and purification of recombinant proteins using a lentivirus system

Recombinant HAZrsr16xHis and HAGFP6xHis were produced using a lentiviral overexpression system with the vector pCSC_IRES_ZsGreen. Viral production and HEK293T cells transduction were carried out following standard procedures (Moffat, Grueneberg et al. 2006). Lentiviral particles were aliquoted and stored at -80°C. Protein purification was performed using HEK293T cells. Briefly, 293T cells transduced with recombinant lentivirus were lysed by sonication in 50mM Tris pH 8.0, 100mM NaCl, 1mM PMSF, protease inhibitors cocktail, 0.1% NP 40. Samples were then centrifuged 14000 rpm for 20 minutes. Purification was performed in a AKTA Explorer system (GE-Healthcare) with a 1 ml HisTrap™ FF Column (GE-Healthcare) equilibrated with lysis buffer containing 20 mM imidazole. The column was washed with 30 mL of lysis buffer containing 20 mM imidazole, followed by 25 mL of lysis buffer containing 50 mM imidazole. Target proteins were then eluted using a linear gradient of imidazole in lysis buffer, from 50 mM to 500 mM. Protein concentration of each fraction was determined with Bradford reagent (Bio-Rad). Fractions containing higher protein concentration were pooled together and concentrated using a 10 KDa molecular weight cut-off Amicon concentrator (Milipore).

3.2.10 Size exclusion chromatography

For the in vitro assembly of the U2AF complexes, equimolar concentrations of recombinant proteins were allowed to interact overnight at 4°C. Complex formation was analysed by size exclusion chromatography using a Superdex S200 10/300 (GE Healthcare) column attached to an AKTA Explorer FPLC system (GE Healthcare). Elution fraction of 1mL were collected in PBS at 0.7 mL/min flow rate, concentrated by ultrafiltration and analysed by western-blot using specific antibodies.

3.2.11 Pull-down assay

Recombinant HAZrsr1-6xHis and HAGFP6xHis purified from HEK293T cells were immobilized in 50 uL of Ni²⁺-sepharose beads (GE Healthcare). The beads were washed three times with 20 mM Tris pH 8.0, 200 mM NaCl, 0.1% NP-40, 0.1% Triton X-100, and incubated for 4 h at 4°C in the same buffer supplemented with 500 µL of undifferentiated or induced to differentiate for 48h MEL C88 cells extract. The beads
were then washed three times with the same buffer and bound proteins were eluted with 20mM Tris pH 8.0, 200 mM NaCl, 250mM imidazole, concentrated by ultrafiltration and analyzed by Western blotting.

### 3.2.12 Immunofluorescence

MEL cells were washed in serum-free medium, applied to poly-l-lysine (Sigma) coated coverslips. (Electron Microscopic Science) and allowed to sit for 5 min at room temperature. Cells were fixed in 3.7% paraformaldehyde in PBS for 10 min and the coverslips were rinsed three times in PBS containing 0.1 M glycine (rinsing buffer) to remove unbound cells. Permeabilization was performed in rinsing buffer containing 0.05% Triton X-100 for 10 min, followed by three washes in the same buffer without detergent. Cells were then incubated in rinsing buffer for an additional 10 min followed by incubation in blocking buffer (PBS containing 0.5 mM glycine, 0.2% fish skin gelatin) for 1 h (Ji, Jayapal et al. 2008). These cells were then incubated with the indicated antibody for immunofluorescence microscopy in blocking buffer for 1 h. Cells were washed three times in PBS 0.05% Tween20 (washing buffer) followed by DAPI and fluorochrome-conjugated secondary antibody staining for 30 min. After washing three times, images were acquired on a Zeiss LSM 510 META confocal microscope using the PlanApochromat 63x/1.4 objective.

### 3.2.13 Protein isolation and fractionation

Nuclear and cytoplasmic protein fractions were isolated as described (Wang, Zhu et al. 2006). Nuclear proteins were further fractionated into chromatin-associated and nucleoplasmic fractions as described (Wuarin and Schibler 1994; West, Proudfoot et al. 2008; Pandya-Jones and Black 2009). Briefly, MEL cells were washed 2 times with ice cold PBS 1x, the pellet was resuspended very gently in 1mL of RSB buffer (10 mM Tris pH 7.4, 10mM NaCl, 3mM MgCl₂) and incubated 3 min. on ice. Cells were centrifuged for 3 min, 4 ºC at 4000 rpm, the supernatant was discarded and the pellet was resuspended in 150 uL of RSBG40 buffer (10 mM Tris-HCl pH 7.4, 10mM NaCl, 3mM MgCl₂, 10% (v/v) glycerol, 0.5% NP40 (v/v) supplemented just before use with 100U/mL of RNAseOUT™ (Invitrogen) and 0.5mM DTT). After centrifugation
for 3 min, 4°C at 7000 rpm the supernatant was collected to a new Eppendorf tube and resuspended with 1mL of Trizol™ (Cytoplasmic fraction, used for protein isolation according to manufacturer instructions). The nuclei pellet was resuspended in 50 uL of prechilled glycerol buffer (20 mM Tris-HCl pH 7.9, 75 mM NaCl, 0.5 mM EDTA, 0.85 mM DTT, 0.125 mM PMSF, 50% (v/v) glycerol) by gentle flicking of the tube. Lysis of the nuclei was performed by adding 450 uL of cold nuclei lysis buffer (20 mM HEPES pH 7.6, 0.3M NaCl, 0.2 mM EDTA, 1 mM DTT, 7.5 mM MgCl2, 1M Urea, 1% NP40 (v/v)), followed by 5s vortex, a 10 min. incubation on ice and a 5 min., 4°C centrifugation at 14000rpm. The nucleoplasmic fraction was collected to a new eppendorf and resuspended in 900 uL of Trizol™ for protein extraction. Sedimented chromatin pellet was resuspended in 200 uL of 10 mM Tris pH 7.4, 0.5M NaCl, 10 mM MgCl2 supplemented with 20 U of DNaseI (Roche) and incubated 30 min at 37°C (Chromatin fraction). Chromatin-associated proteins were extracted by adding 1 mL of Trizol™.

3.2.14 Zrsr1-knockout mice

Zrsr1 KO mice were kindly provided by Dr. M. Katsuki (University of Tokyo). Routine genotyping was carried out by multiplex PCR with genomic DNA isolated from tails and with primers described in Supplementary material.

3.2.15 Hematological Analysis

Hematological analysis was performed by the CDVET-Lab, Lisbon. Peripheral blood was collected into heparinized tubes by cardiac puncture.

3.2.16 Flow cytometry

Blood from adult mice was collected and stained with FC-Block, PE-conjugated CD71 and FITC-conjugated TER119 antibodies (BD Pharmingen, San Diego, CA) for 20 minutes at room temperature. After washing with PBS/0.5% BSA, FACS analysis was carried out using a FACSCalibur (BD Biosciences) and data analysis performed with FlowJo (TreeStar, Eugene, OR).
3.3 Results

3.3.1 U2AF-family members are differentially expressed in a tissue-specific manner

To determine if the relative expression of the U2AF related genes is tissue specific, we analyzed data from microarray experiments obtained from publicly available data sets. Expression of U2AF2, U2AF1, Zrsr1, Zrsr2 and U2AF1L4 mRNA’s (which encode U2AF65, U2AF35, Zrsr1, Zrsr2 and U2AF26 proteins, respectively) was accessed in a variety of human (Figure 3.1A) and mouse (Figure 3.1B) tissues which include breast, cerebellum, heart, kidney, liver, muscle, pancreas, prostate, spleen, testes and thyroid. As already discussed in previous chapters of this thesis, the U2AF family members are extremely well conserved proteins and not surprisingly the expression patterns across all tissues are very similar between human and mouse. U2AF2 and U2AF1 family members were found to be highly expressed in all tissues analyzed, which seems to correlate with the essential nature of these two genes in higher eukaryotes. In good agreement with previous studies (Pacheco, Gomes et al. 2004) here the expression ratio between U2AF1/U2AF2 was found to be similar in skeletal muscle and heart, we also find this correlation in our data analysis. Therefore, it is clear that U2AF1 is the major U2AF35-family member expressed in all the tissues examined when compared to Zrsr1, Zrsr2 and U2AF1L4. Our microarray data analysis also showed that U2AF1L4 is highly expressed in cerebellum, heart, spleen and testis, with low expression in the liver. These results seems to correlate with previous studies (Shepard, Reick et al. 2002) were this U2AF35-family member was also found to be highly expressed in cerebellum, heart and testis. Moreover, a recent report (Heyd, ten Dam et al. 2006) showed that U2AF1L4 is involved in the splicing regulation of the CD45 gene, which is found on all nucleated hematopoietic cells (Hermiston, Xu et al. 2003) and therefore, likely to be highly expressed in the spleen. Nevertheless, U2AF1L4 is also found to be differentially expressed in several human and mouse tissues. Regarding Zrsr1, our data indicate that is highly expressed in the cerebellum when compared to the other tissues specially heart, liver and testis. Again, our data seems to be consistent with previous reports which indicate that Zrsr1 is predominantly expressed in the brain (Hatada, Sugama et al. 1993), specifically in the hyppocampus and dental gyrus (Hatada,
Kitagawa et al. 1995). Like Zrsr1, Zrsr2 is also found to be highly expressed in the cerebellum and very low expressed in the other tissues. In agreement with this, other studies showed that Zrsr2 is abundantly expressed in the brain (Grosso, Gomes et al. 2008).

Taken all together, these observations suggest that the expression of the U2AF family members could be regulated in a tissue specific manner.

3.3.2 The U2AF-family member Zrsr1 is differentially up-regulated in erythropoiesis

The tissue specific expression profile of the U2AF-family members prompted us to investigate if there is any correlation between the expression patterns of these genes with different cellular differentiation programs. We had previously shown that Zrsr1 expression increases during erythroid differentiation but remains low in myotubes and adipocytes (Grosso, Gomes et al. 2008). Here, we analyzed Zrsr1 transcription profiles in hematopoietic progenitor cells undergoing multilineage differentiation (Bruno, Hoffmann et al. 2004). Having determined that Zrsr1 expression increases during
erythroid differentiation but remains low in myotubes and adipocytes (Grosso, Gomes et al. 2008), we analyzed Zrsr1 transcription profiles in hematopoietic progenitor cells undergoing multilineage differentiation. Multipotent progenitor cells derived from mouse bone marrow (Spooncer, Heyworth et al. 1986) were cultured under conditions that produced >95% pure populations of erythroid and neutrophilic cells (Bruno, Hoffmann et al. 2004). Cultures were sampled at 0, 4, 8, 16, 24, 48, 72, and 168 h of differentiation and analyzed with Affymetrix GeneChip MG-U74Av2 arrays (Bruno, Hoffmann et al. 2004). An analysis of the expression behaviour of the genes that encode for U2AF65, U2AF35 and U2AF35-related proteins is presented in Figure 3.2A. Expression of Zrsr1 increases during erythroid differentiation but remains unchanged in the neutrophil differentiation series. In contrast, U2af1 expression is reduced during erythroid differentiation and no major changes are detected for the remaining genes (Fig. 3.2A).

Figure 3.2– Gene Expression signatures of U2AF-related genes during cell differentiation. (A) Expression levels of genes belonging to the family of U2AF-related genes determined by microarray analysis in two independent datasets during differentiation of erythroid cells and neutrophils as already described by (Grosso, Gomes et al. 2008). (B) Gene expression profile of the U2AF-related genes in the whole mice embryo, at stage E14.5. In situ hybridization with probes against the U2af2, Zrsr1, U2af1 and
U2af1l4 genes which encode the U2AF65, Zrsr1, U2AF35 and U2AF26 proteins respectively. Zrsr1 shows high expression in the fetal liver (zoomed windows) of 14.5 dpc mice embryos. Images obtained from the GenePaint database (Visel, Thaller et al. 2004).

From our results, the striking difference for the U2AF-related gene expression signature across all the differentiation processes analyzed is provided by the Zrsr1 member that was found specifically up-regulated during erythroid differentiation. Since a gene is considered to be a specific tissue signature when the expression changes at least 1.5-fold more than in any other process (Grosso, Gomes et al. 2008), as outlined in Figure 3.2A, Zrsr1 is considered to be an erythroid tissue signature.

To determine whether expression of the Zrsr1 gene associates with erythropoiesis in vivo, we searched the GenePaint atlas of gene expression patterns in the mouse embryo (Visel, Thaller et al. 2004). When comparing the spatiotemporal expression pattern of the U2AF-related genes in the mice embryo at the development stage E14.5, (Figure 3.2B), Zrsr1 is clearly detected in the fetal liver. Taking in account that the fetal liver is a major source of erythroid progenitors during the murine embryonic development (Dolznig, Boulme et al. 2001), this observation further supports our finding that Zrsr1 is an erythroid tissue specific gene and, therefore, could be an important factor involved in the regulation of erythropoiesis.

Although microarrays provide a widely used tool for whole genome profiling, the quality of the results obtained are known to be affected by several factors (Morey, Ryan et al. 2006). Therefore, this gene signature associated with erythroid differentiation (up-regulation of the Zrsr1 mRNA), should be validated by using cellular models that recapitulates the erythroid differentiation program and quantitative real-time PCR (qPCR), which is a commonly used validation tool for confirming gene expression results obtained from microarrays (Morey, Ryan et al. 2006). Additionally, accessing the U2AF35-family members protein expression profile during erythroid differentiation should also provide an alternative method to investigate if Zrsr1 is an erythroid tissue specific signature.
3.3.3 Development and characterization of an anti-Zrsr1 polyclonal antibody

Due to the lack of an available anti-Zrsr1 antibody, we decided to produce a specific antibody against this protein. The U2AF35-family members show a high sequence homology and, to avoid cross-reaction, we started to perform sequence based alignments in order to find unique peptide regions of the mouse Zrsr1 protein. The C-terminal sequence N-GREEDSSPGQSQSHRT-C (residues 412-428 highlighted in yellow in Figure 3.3) was found to be unique to mouse Zrsr1 protein.

Mouse Zrsr1 is a 428-amino acid protein with a predicted molecular weight of 51 KDa. Based on the primary sequence as well as on the motifs shared with other...
proteins (Kielkopf, Lucke et al. 2004; Mollet, Barbosa-Morais et al. 2006), Zrsr1 is considered to be a modular protein composed by three domains: a central RRM domain, flanked by two ZnF-C3H1 binding motifs and a C-terminal low complexity region (Figure 3.4A). The peptide N-GREEDSSPGPQSQSHRT-C (residues 412-428 from the protein sequence) was found to be unique to mouse Zrsr1 (Figure 3.4) and was used to immunize two rabbits (PickCell Laboratories BV, The Netherlands). Affinity purification of Zrsr1 specific antibodies (pC5506) from immune serum was made against the peptide epitope.

Figure 3.4 - Polyclonal antibody against Zrsr1. (A) Schematic representation of the domain organization of Zrsr1 and U2af35 proteins, showing the C-terminal peptide used to immunize rabbits for the polyclonal antibody production. LCR, low complexity region; CCD, coiled-coil domain; ZnF-C3H1, zinc-finger domain; RRM, RNA-recognition domain. (B) Analysis by western blot with antibodies against Zrsr1 and U2af35 of total protein extracts from transiently transfected HEK293T cells expressing Zrsr1 or U2af35 proteins. Mock lane correspond to protein extracts from cells transfected with the empty vector. In the lower panel the additional detection band corresponds to the endogenous U2AF35 protein. (C) Immunodetection by western blot of Zrsr1 protein in C2, 3T3 and I/11 mouse cell lines. Protein molecular weight markers are indicated on the left.
Purified polyclonal anti-Zrsr1 antibody (pC5506) specificity was screened by western blotting analysis in several human and mouse cell lines (Figure 3.4B and C). The synthetic peptide used for immunization is only found in mouse Zrsr1 protein and, therefore, is not expected that pC5506 cross-reacts with the human protein. The first screen was made in human HEK293T cells transiently transfected with either HAU2AF\(^{35}\) or mHAZrsr1 (where \(m\) denotes \textit{Mus musculus}), and the negative control was done with the empty vector (Figure 3.4B). When probing the nitrocellulose membrane with anti-HA tag antibody two bands with approximately 37 and 50 KDa are obtained and represent specific band products for exogenous transfected HAU2AF\(^{35}\) and mHAZrsr1, respectively. When probing the membrane with pC5506 antibody a single band around 50 KDa is obtained in mHAZrsr1 transfected cells, which is not detectable in either mock or HAU2AF\(^{35}\) transfected cells (Figure 3.4B), indicating that the antibody is specific for mZrsr1. To test if the pC5506 is able to recognize endogenous Zrsr1 protein we used three different mouse cell lines (Figure 3.4C). Using HEK293T cells transfected with HAZrsr1 as a positive control we were able to detect endogenous Zrsr1 in mouse erythroblast (I/11 cells) but not in myoblasts (C2) or fibroblasts (3T3). Interestingly, these results are not totally unexpected since Zrsr1 was found to be an erythroid specific tissue signature (Grosso, Gomes et al. 2008) and the detection of this protein in erythroblast strongly correlates it that observation. Taken all together, these results indicate that pC5506 strongly reacts and specifically recognizes the mouse Zrsr1 protein.

3.3.4 Zrsr1 expression is up-regulated in two different cell models of erythroid differentiation

In order to validate the microarray data analysis, we took advantage of two very well established murine cellular models of erythroid differentiation, I/11 and the murine erythroleukaemia (MEL) cells. I/11 are primary cells isolated from fetal livers of p53/-/ mice (Dolznig, Boulme et al. 2001) that are known to proliferate under the presence of erythropoietin (Epo), stem cell factor (SCF) and dexamethasone (Dex), and undergo
terminal erythroid differentiation when exposed to high concentrations of Epo and iron-loaded transferrin (von Lindern, Deiner et al. 2001).

To investigate if the Zrsr1 mRNA levels are up-regulated during I/11 cells erythroid differentiation we induced them to differentiate by complementing the culture medium with of Epo and Iron-loaded Transferrin. Total RNA was extracted from cell pellets, reverse transcribed using random primers and qRT-PCR analysis was performed at several differentiation time points (see Table SI for primers information). The results (Figure 3.5A) are presented as fold-changes at the indicated time points relative to undifferentiated cells (TO) and normalized against the mRNA levels of the RNase inhibitor (RI) gene (Blazquez-Domingo, Grech et al. 2005). Our results clearly demonstrate that during erythroid differentiation of I/11 cells the mRNA levels of the Zrsr1 gene are up-regulated (1.5-fold increase), while other U2AF35-family members (U2af1) were found to be down-regulated. Therefore, the microarray data was validated by qRT-PCR since the same expression pattern of the U2AF-family members is observed.

Figure 3.5- Zrsr1 expression is up-regulated upon erythroid differentiation of I/11 cells. I/11 cells were induced to differentiate, and samples were harvested for total mRNA isolation and total protein at the indicated time points. (A) mRNA transcript levels determined by qRT-PCR of U2af1 and Zrsr1 genes during in vitro differentiation of I/11 cells. The results are given as a fold-change ratio (Log2) comparative to time zero. The threshold cycle values of the gene RNase inhibitor (RI) were used for normalization. (B) Expression of the U2AF35-family members was examined by western blot in the induced differentiation of I/11 cells. Erk-2 protein was used as loading control.
We next asked if the same results could be obtained in other cell model of erythroid differentiation. MEL cells are murine virus-transformed spleen hematopoietic cells, arrested at the pro-erythroblast stage of differentiation, that can be induced with various chemical agents to undergo terminal erythroid differentiation (Tsiftsoglou, Pappas et al. 2003; Tsiftsoglou, Pappas et al. 2003). In order to investigate if the Zrsr1 mRNA levels are also up-regulated during MEL cells differentiation we induced them to differentiate by supplementing with DMSO, Iron-Dextran and BSA (Volloch and Housman 1982 to the culture medium. Erythroid differentiation was accessed by visual inspection of the cells pellets since upon induction of differentiation, very high levels of haemoglobin accumulates in the cytoplasm, giving a pink-red color to the differentiated cell pellets (Murray, 1991 #613; Patel and Lodish 1987).

Figure 3.6- Zrsr1 expression is up-regulated upon erythroid differentiation of MEL C88 cells. MEL C88 cells were induced to differentiate, and samples were harvested for total mRNA isolation and total protein at the indicated time points. (A) mRNA transcript levels determined by qRT-PCR of U2af1 and Zrsr1 genes during in vitro differentiation of MEL cells. The results are given as a fold-change ratio (Log2) comparative to time zero. The threshold cycle values of the gene RNase inhibitor (RI) were used for normalization. (B) Expression of the U2AF35-family members was examined by western blot in the induced differentiation of MEL cells. β-Actin protein was used as loading control and ASF/SF2 was used as a differentiation control according to (Yang, Huang et al. 2005).
Total RNA was extracted, reversed transcribed with random primers and the qRT-PCR was performed at the indicated differentiation time points. The results, are represented as fold-change comparative to undifferentiated cells (T0), and normalized against the mRNA levels of the RI gene (Blazquez-Domingo, Grech et al. 2005). The results show that like for I/11 cells, the Zrsr1 mRNA transcript levels are up-regulated upon erythroid differentiation of MEL cells (Figure 3.6A).

Next, we asked whether differentiated erythroid cells contain more Zrsr1 protein than undifferentiated cells. This represents a relevant question since previous studies have demonstrated that changes in splicing factors mRNA levels may not necessarily reflect on the protein expression due to post-transcriptional regulation (Boutz, Stoilov et al. 2007; Makeyev, Zhang et al. 2007). Therefore, we next accessed the protein expression profile of the U2AF35-family members upon erythroid differentiation in both I/11 and MEL cellular models. Upon erythroid differentiation of I/11 cells the Zrsr1 protein levels are clearly up-regulated when compared to non-differentiated cells (Figure 3.5B). In fact, these results are in good agreement with the microarray and qRT-PCR data since increased mRNA expression results in increased protein expression. The same is true for the other erythroid system used since in MEL cells the Zrsr1 protein levels are also found up-regulated upon induction of differentiation. Notably, protein levels of U2AF65 and U2AF35 remain largely unaffected, whereas splicing factor ASF/SF2 accumulates in differentiated MEL cells, as previously described (Yang, Huang et al. 2005). We therefore conclude that differentiation of erythroid precursor cells leads to upregulation of Zrsr1 mRNA and protein.

3.3.5 The Zrsr1 gene is transcriptionally activated during erythroid differentiation

To determine whether upregulation of Zrsr1 mRNA results from transcriptional activation, we compared levels of Zrsr1 transcripts being synthesized by RNA Polymerase II (Pol II) in erythroid precursors before and after differentiation. We used a fractionation technique that was initially described by Wuarin and Schibler (Wuarin and Schibler 1994) and subsequently modified in the Proudfoot and Black laboratories (Dye, Gromak et al. 2006; West, Proudfoot et al. 2008; Pandya-Jones and Black 2009). The protocol takes advantage of the fact that once Pol II initiates transcription it forms a
tight complex with the DNA template that resists treatment with urea and mild detergent. The extraction procedure does not dissociate histones from DNA and therefore the chromatin remains highly compacted and can be sedimented (with associated nascent RNA) by low-speed centrifugation. Molecules of mRNA that have been released from chromatin are found in the nucleoplasmic supernatant fraction, and mRNAs that were exported from the nucleus are detected in the cytoplasmic fraction. Equal amounts of RNA were taken from cytoplasmic, nucleoplasmic, and chromatin-associated fractions, reverse transcribed with random primers and PCR amplified using primer pairs indicated in Fig. 3.7. As a control for the fractionation procedure, we used primers that distinguish unspliced and spliced forms of the \( P120 \) gene.

As expected, unspliced \( P120 \) RNA is exclusively detected in the chromatin fraction whereas the spliced form is present in the chromatin and nucleoplasmic fractions and is most abundant in the cytoplasm (Fig. 3.7).

**Figure 3.7- Identification of Zrsr1 poly(A) site.** (A) Schematic representation of the mouse Zrsr1 gene. Open reading frame (ORF) is represented by the dark grey box were start (ATG) and stop (TGA) codons are shown. Relative position of the three putative polyA sites is indicated by the three hexanucleotide elements that could be recognized by the 3'processing machinery. (B) RNA oligo(dT) transcribed from cytoplasmic (Cyto), nucleoplasmic (Nucleo) and chromatin-associated (Chroma) fractions was submitted to semi-quantitative PCR with the indicated primer pairs. GAPDH is used for loading control, mice genomic DNA for PCR positive control and P120 gene for fractionation control.
This is consistent with the view that most mammalian pre-mRNAs are spliced co-transcriptionally, i.e., while still associated with chromatin (Pandya-Jones and Black 2009). Amplification of Zrsr1 mRNA using a pair of primers targeted to the coding sequence of the gene shows enrichment of transcripts in the chromatin fraction of MEL cells that have been induced to differentiate for 48h (Fig. 3.7). Higher levels of Zrsr1 mRNA are also detected in the cytoplasm of differentiated cells, as expected taking into account that more RNA is being synthesized. mRNA amplified using a pair of primers targeted to the sequence downstream of the poly(A) site is exclusively detected in the chromatin fraction (Fig. 3.7), as expected for transcripts that have not yet been cleaved and therefore remain tethered to the DNA template (Dye, Gromak et al. 2006; West, Proudfoot et al. 2008). Higher levels of uncleaved Zrsr1 RNA are detected in differentiated cells, consistent with the view that transcription was activated upon induction of erythroid differentiation.

Genes that are actively transcribed tend to have a higher density of polymerases associated with the DNA template. We therefore asked whether the distribution of Pol II along the Zrsr1 gene changes during erythroid differentiation. Chromatin immunoprecipitation (ChIP) analysis was performed using a polyclonal antibody (N-20) that recognizes the N-terminal region of the large subunit of Pol II in a phosphorylation-independent manner. For comparison, we also analyzed the distribution of Pol II along the U2af1 gene. Cellular DNA was sonicated to yield DNA fragments from 200 to 400 bp and primer pairs were used to amplify immunoprecipitated fragments across the promoter region, three regions along the gene body and a region after the poly(A) site (Fig. 3.8). For the U2af1 gene, primer pairs were used to amplify the promoter region, exon 2, exon 3, exon 6 and a region past the poly(A) site of the (Fig. 3.8). Results are depicted as relative occupancy of Pol II at the indicated genomic DNA region normalized to a control immunoprecipitation with non-immune IgG. Differentiated MEL cells show a significant increase in Pol II occupancy at the promoter region of the Zrsr1 gene (Fig. 3.8), further suggesting that transcription of this gene is activated during erythroid differentiation. In contrast, Pol II occupancy at the U2af1 promoter is lower in differentiated cells (Fig. 3.8), consistent with the lower levels of mRNA detected by microarray (Fig. 3.2A) and qRT-PCR analysis (Fig. 3.6A).
3.3.6 Is there a link between H3K36me histone modification and splicing?

While the majority of the U2AF\textsuperscript{35}-related genes have a classical exon-intron configuration, with known alternative spliced products (Pacheco, Gomes et al. 2004; Heyd, Carmo-Fonseca et al. 2008), the related Zrsr1 mouse gene was shown to be imprinted and intronless (Hatada, Sugama et al. 1993; Hayashizaki, Shibata et al. 1994; Hatada, Kitagawa et al. 1995). Our findings that upon erythroid differentiation the transcription levels of \textit{Zrsr1} (intronless) and \textit{U2af1} (intron-exon structure) genes can be modulated, provided us an excellent system to study two very important and open questions in field: how does the chromatin architecture influence co-transcriptional splicing? Is there a correlation between the histone mark H3K36me3, known to be enriched in exons (Kolasinska-Zwierz, Down et al. 2009), with alternative splicing?
To investigate this hypothesis, we induced MEL cells to differentiate with DMSO and accessed by ChIP analysis the distribution of the H3K36me3 histone modification on different regions of \textit{U2af1} and \textit{Zrsr1} after 48 hours of differentiation.

To map the temporal and spatial deposition of histone H3K36me3 into \textit{Zrsr1} and \textit{U2af1} genes we used polyclonal antibodies that specifically recognize this histone modification. Cellular DNA was sonicated to yield DNA fragments from to 200 to 400 bp and high spatial resolution of the immunoprecipitated fragments was achieved by designing specific primer sets for quantitative real-time PCR to several regions along the \textit{Zrsr1} and \textit{U2af1} genes. Primer pairs were used to amplify product fragments across the promoter associated region, exon 3 and exon 6 of the \textit{U2af1} gene and for the promoter associated region and body of the \textit{Zrsr1} gene (Figure 3.9A).

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure3.9.png}
\caption{Distribution of the H3K36me3 and H3K9Ac histone modifications on \textit{U2af1} and \textit{Zrsr1} genes upon erythroid differentiation} (A) Schematic diagram of the gene regions amplified by primer sets for \textit{U2af1} and \textit{Zrsr1} genes, respectively. (B) and (C) MEL C88 cells were induced to differentiate for 48h and aliquots of sheared chromatin from the same number of cells were used for ChIP assay with anti-H3K36me3 (B) or H3K9Ac (C) antibodies. Results are expressed as ‘relative occupancy’, calculated as enrichment from specific immunoprecipitations normalized to enrichment from anti-H3total immunoprecipitations. All histograms depict mean and standard deviation from at least three independent experiments.
\end{figure}
Our results show that for the *U2af1* gene, the histone H3K36me3 levels decreased after differentiation, which most likely reflects the lower transcriptional activity of this gene upon MEL cells differentiation. In clear contrast, the same histone mark located along the *Zrsr1* gene did not show any variation in response to DMSO stimulation (Figure 3.9B), even though the transcription of this gene is found to be increased. Noteworthy, the intronless gene did not display any increase of H3K36me3 along its length in comparison to the significant enrichment observed in the internal exons of its intron-containing family member (Figure 3.9B). In addition to the trimethylation of histone H3K36, we have also evaluated the distribution of acetylated histone H3K9 (H3K9ac) in our experimental system. As expected from its positive correlation with the gene expression levels (Kouzarides 2007; Li, Carey et al. 2007), H3K9ac decreased in the promoter region of *U2af1* during differentiation while, in *Zrsr1* this histone modification was found to be increased (Figure 3.9C). Taken all together, these results suggests that the influence of the exon-intron structure in the response to transcription modulation and splicing is a specific feature of the H3K36me3 histone modification.
3.3.7 Does the Zrsr1 protein interact with U2AF$^{65}$?

The interaction between U2AF$^{65}$ and U2AF$^{35}$ involves a critical tryptophan residue (Kielkopf, Lucke et al. 2004) that is conserved in all U2AF$^{35}$-related proteins. It is therefore expected that all members of the U2AF$^{35}$-family interact with U2AF$^{65}$ (Mollet, Barbosa-Morais et al. 2006). Experimental evidence indicating that U2AF$^{65}$ binds to Zrsr2 and U2AF$^{26}$ was previously reported (Tronchere, Wang et al. 1997) (Heyd, ten Dam et al. 2006). On the other hand, to date there was no data available showing if Zrsr1 is able or not to interact with the U2AF$^{65}$ subunit. To address this question we first produce and purge recombinant U2AF-related proteins in order to test \textit{in vitro} if we are able to isolate the Zrsr1-U2AF$^{65}$ complex.

3.3.8 Expression and Purification of recombinant U2AF-family members in \textit{E.coli}

Fusion proteins were expressed in \textit{E.coli} BL-21 (DE3) strain at 37$^\circ$C or 18$^\circ$C (see text and figures for details) induced with 0.3mM of IPTG and purified with the AKAT Explorer FPLC system (GE-Healthcare) using a GSTrap FF column. The results presented bellow concern the U2AF-related recombinant fusion proteins GSTU2AF$^{65}$, GSTU2AF$^{35}$, GSTU2AF$^{26}$ and GSTZrsr1.

The overexpression of GSTU2AF$^{65}$ was performed in the \textit{E.coli} strain BL21 using the pGEX-2X expression vector. For this construct the strategy used for expression and purification is outlined on Figure 5.2A.

The GSTU2AF$^{65}$ fusion proteins were purified from clarified lysates using a GSTrap FF 1 mL column attached to AKTA Explorer system, pre-equilibrated with lysis buffer. Loaded GSTrap column was washed with 1×PBS until the absorbance baseline stabilized, after which the bound GSTU2AF$^{65}$ was eluted from the column until the Abs280nm baseline stabilized (see Figure 3.11B). Collected elution fractions were analyzed by SDS-PAGE and western-blot, by using a monoclonal antibody against U2AF$^{65}$ (MC3). Our results show that by using the GST fusion protein system in the \textit{E.coli} host, we are able to overexpress and purify recombinant GSTU2AF$^{65}$ although the elution occurs along with two major contaminants bands (with a Mw $\sim$ 28KDa), which, most likely, represent truncated forms of the GST construct.
Figure 3.11–Overexpression and affinity purification of recombinant U2AF^{65}. (A) Strategy used to produce and purify U2AF^{65} as a recombinant protein. *E.coli* BL-21 cells were induced with 0.3mM of IPTG during 4h at 37°C. Cells were harvested and sonicated for protein purification. (B) FPLC chromatogram recording at Abs280nm the protein levels (solid line), and the Glutathione gradient (0-10mM; dashed line) used to recover the recombinant protein bound to the GSTrap column. (C) 10% SDS-PAGE of the eluted fractions. (D) Selected elution fractions were blotted into a nitrocellulose membrane and probed with an anti-U2AF^{65} antibody (MC3). Protein molecular weights markers are indicated on the left. mAU, arbitrary mili-absorbance Units.

It is very well described that for some constructions multiple protein bands may result from partial degradation of the GST fusion protein by proteases or co-purification of host proteins with the GST fusion protein due to oversonication. Nevertheless, these two possibilities most likely do not represent what happens in the case of our GSTU2AF^{65} construct since: i) the BL21 *E.coli* is a protease deficient strain and, ii) the most common host protein that co-purifies with GST fusion proteins due to oversonication is Dna K (Rial and Ceccarelli 2002), a protein with a Mw of 69 KDa. On the other hand, the presence of contaminant bands with a molecular weight similar to free GST can also be caused by translational pausing at the junction between GST and the fusion partner (\(\_\_\_\_\_\_\_\_\_\_\)). Therefore, this may be the reason for the presence the full-length protein and the additional contaminant bands, with an Mw of ~28KDa in the elution
fractions. Nevertheless, we were able to overexpress and purify recombinant GSTU2AF$^{35}$.

The overexpression of GSTU2AF$^{35}$ and GSTU2AF$^{26}$ was also performed in the *E.coli* strain BL21 using the pGEX-4T-3 expression vector. The strategy used for expression and purification of these construct is outlined on Figure 3.12A and 3.13A.

![Figure 3.12 - Overexpression and affinity purification of recombinant U2AF$^{35}$.](image)

**Figure 3.12–Overexpression and affinity purification of recombinant U2AF$^{35}$.** (A) Strategy used to produce and purify U2AF$^{35}$ as a recombinant protein. *E.coli* BL-21 cells were induced with 0.3mM of IPTG during 4h at 37ºC. Cells were harvested and sonicated for protein purification. (B) FPLC chromatogram recording at Abs280nm the protein levels (solid line), and the Glutathione gradient (0-10mM; dashed line) used to recover the recombinant protein bound to the GSTrap column. mAU, arbitrary mili-absorvance Units. (C) 10% SDS-PAGE of the eluted fractions. (D) Selected elution fractions were blotted into a nitrocellulose membrane and probed with a anti-U2AF$^{35}$ antibody. Protein molecular weights markers are indicated on the left.

Both recombinant proteins were purified from clarified lysates with a GSTrap FF 1 mL column attached to AKTA Explorer System, using the same conditions as described for the GSTU2AF$^{85}$ construct. Collected elution fractions were analyzed by SDS-PAGE and western-blot, by using a rabbit polyclonal antibody against U2AF$^{35}$ (Figure 3.12D) or U2AF$^{26}$ (Figure 3.13D), respectively. Although the yields for these
two construct are much lower when compared with the purification of GSTU2AF\textsuperscript{65}, our results show that we are also able to overexpress and purify these two recombinant protein in \textit{E.coli}. In fact, GSTU2AF\textsuperscript{35} recombinant protein is eluted from the GSTTrap column as a single band, as observed in the SDS-PAGE analysis of the collected fractions (Figure 3.12C). On the other hand, for the GSTU2AF\textsuperscript{26} fusion protein, several contaminant bands co-purifies with the full-length protein (Figure 3.13C), most likely due to the reason already discussed above. Nevertheless, our strategy to produce recombinant U2AF\textsuperscript{26} as fusion protein with GST proved to be useful since attempts to purify recombinant U2AF\textsuperscript{26} alone from \textit{E.coli} or Sf9 cells yielded only insoluble protein (Shepard, Reick et al. 2002).

Figure 3.13–Overexpression and affinity purification of recombinant U2AF\textsuperscript{26}. (A) Strategy used to produce and purify U2AF\textsuperscript{26} as a recombinant protein. \textit{E.coli} BL-21 cells were induced with 0.3mM of IPTG during 4h at 37°C. Cells were harvested and sonicated for protein purification. (B) FPLC chromatogram recording at Abs280nm the protein levels (solid line), and the Glutathione gradient (0-10mM; dashed line) used to recover the recombinant protein bound to the GSTrap column. mAU, arbitrary milli-absorvance Units. (C) 10% SDS-PAGE of the eluted fractions. (D) Selected elution fractions were blotted into a nitrocellulose membrane and probed with a anti-U2AF\textsuperscript{26} antibody. Protein molecular weights markers are indicated on the left.
Regarding the GSTZrsr1 recombinant protein, our initial attempts of overexpression and purification were performed with the same conditions, as described for the other U2AF-related proteins. The Zrsr1 gene was cloned into the pGEX-4T-3 expression vector, which was chemically transformed into the *E. coli* BL21 strain. The recombinant bacteria were grown at 37°C to an OD600 (optical density at λ = 600 nm) of 0.6 and the expression was induced with IPTG to a final concentration of 0.3 mM. A total culture volume of 1000 mL was used, and the cells were allowed to grow during four hours at 37°C. After purification with a GSTrap FF 1 mL column attached to a AKTA explorer system, the collected elution fractions were analyzed by SDS-PAGE (Figure 3.14C). Our results for the GSTZrsr1 protein induced at 37°C clearly shows that this are not the ideal conditions for the overexpression of this construct since the majority of the recombinant protein is purified as GST truncation (Figure 3.14C).

![Figure 3.14–Overexpression and affinity purification of recombinant Zrsr1. (A) Strategy used to produce and purify Zrsr1 as a recombinant protein. *E. coli* BL-21 cells were induced with 0.3mM of IPTG during 4h at 37°C. Cells were harvested and sonicated for protein purification. (B) FPLC chromatogram recording at Abs280nm the protein levels (solid line), and the Glutathione gradient (0-10mM; dashed line) used to recover the recombinant protein bound to the GSTrap column. mAU, arbitrary mili-absorvance Units. (C) 10% SDS-PAGE of the eluted fractions. Protein molecular weights markers are indicated on the left.](image-url)
In such cases, a variety of growth parameters should be investigated, either solely or in combination, in order to improve the yield of non-degraded fusion protein in the soluble fraction. Usually, the steps to investigate include: i) lowering the growth temperature between 18°C and 30°C, and ii) changing the induction conditions (either the OD and the time of induction). Therefore, next we screened different expression conditions for the GSTZrsr1 construct. *E. Coli* BL-21 were grown at 37°C to an OD$_{600}$ of 1.2 and the culture was cooled down to 18°C before induction with IPTG to a final concentration of 0.3 mM. The culture was allowed to grow overnight at 18°C after which the cells were harvested. The recombinant protein was purified from sonicated and clarified lysates with a GSTTrap FF 1 mL column attached to a AKTA Explorer System, and the collected elution fractions were analyzed by SDS-PAGE and western-blot, by using our rabbit polyclonal antibody against Zrsr1 (Figure 3.15D).

![Figure 3.15–Overexpression and affinity purification of recombinant Zrsr1. (A) Strategy used to produce and purify Zrsr1 as a recombinant protein. *E.coli* BL-21 cells were induced with 0.3mM of IPTG overnight at 18°C. Cells were harvested and sonicated for protein purification. (B) FPLC chromatogram recording at Abs280nm the protein levels (solid line), and the Glutathione gradient (0-10mM; dashed line) used to recover the recombinant protein bound to the GSTrap column. mAU, arbitrary mili-absorvance Units. (C) 10% SDS-PAGE of the eluted fractions. Protein molecular weights markers are indicated on the left.](image-url)
Using different expression conditions, our results show that we are able to overexpress and purify full-length recombinant GSTZrsr1 protein, although the elution occurs along with a major contaminant band (with a Mw ~ 28KDa), which, most likely, represent a truncated form of the GST construct since in a 10% SDS-PAGE this protein runs at a higher molecular weight when compared to GST alone (Figure 3.15C). Taken all together, our results demonstrate that we are able to express and purify recombinant U2AF-related proteins using the GST-fusion system.

Although the production of recombinant proteins in *E.coli* has several advantages when compared to other expression systems (which include rapid cell growth, inexpensive culture media, easy to scale-up cultures and high expression levels), there are some important disadvantages that should be taken into account. Even if the high expression levels obtained with *E.coli* expression systems are consider to be an advantage, in some cases this system yields insoluble proteins in form of inclusion bodies and, therefore, difficult to recover as functional proteins. Although these problems can be overcome with the use of fusion proteins that helps to prevent aggregation of the recombinant protein, the main disadvantage of bacterial systems is that they lack all post-translational modifications that might be important for the function of the protein. Therefore, with the aim of develop biochemical tools to study the U2AF-related proteins, we also tried to set up a mammalian expression system to produce recombinant proteins.

### 3.3.9 Expression and purification of recombinant proteins in HEK293T cells

For the production of active recombinant proteins there are several alternative protocols to the *E.coli* system that have been extensively developed over the years. As an example, mammalian cells have been successively used as an expression system for the production of recombinant splicing factors (Cazalla, Sanford et al. 2005). The advantage of this system relies on the fact that recombinant proteins produced in
mammalian cells will have the relevant post-translational modifications which can profoundly affect the protein properties (Walsh and Jefferis 2006). Here, we describe a simple and cost efficient method for the expression and purification of recombinant proteins in mammalian cells. Our method enables the expression of double HA- and 6xHIS-tagged proteins in lentiviral transduced HEK 293T cells grown in monolayer. The advantages of this method arise from the fact that the use of expensive transfection reagents is minimized and the mammalian cells are stable integrated with the gene of interest (instead of normal transient transfection).

As a proof of principle, we have cloned, transduced and purified recombinant HAGFPHis and mouse HAZrsr1His protein in HEK293T cells (Figure 3.16).

**Figure 3.16- Lentiviral stable transduction of HEK293T cells.** (A) Schematic representation of the lentiviral construct used. Mouse Zrsr1 and GFP double tagged with HA- and hexa-histidine were expressed under the control of the CMV promoter. An internal ribosome entry site (IRES) allows double expression of both both constructs and GFP, used as transduction efficiency control. (B) Transduced HEK293T cells stably expressing GFP. (C) Western blot analysis of transduced HEK293T cells with the indicated antibodies. Protein molecular weights markers are indicated on the left.

HEK293T cells expressing either HAGFPHis or HAZrsr1His were lysed, extracts were clarified by centrifugation and applied into a HisTrap column attached to a AKTA
Explorer System. To reduce non-specific contaminants binding, the cellular extract was loaded into the nickel column with 20mM imidazole and further washed with 50mM imidazole. After extensive washing, recombinant proteins were eluted with a linear imidazole gradient (50-250mM), and the collected fractions were concentrated and analyzed by SDS-PAGE and western-blot.

Our results demonstrate that using this strategy we are able to express and purify HAGFPHis (Figure 3.17) and mouse HAZrsr1His (Figure 3.18) recombinant proteins. Taken all together, we have shown that we have developed new biochemical tools that will allow us to access if the Zrsr1 protein is able to interact with the U2AF65 subunit and to identified protein interaction partners of Zrsr1.

**Figure 3.17- Affinity Purification of recombinant GFP in mammalian cells.** HEK293T cells were transduced with lentivirus encoding HAGFPHis and whole cellular extracts were prepared for protein purification using a HisTrap column. (A) Affinity column purification chromatogram in which the black trace represents the absorbance at 280nm. The dashed line represents the imidazole percentage throughout the purification procedure (B) SDS-PAGE (top panel) of the eluted fractions and western blot analysis using the indicated antibody. Protein molecular weights markers are indicated on the left.
Figure 3.18- Affinity Purification of recombinant Zrsr1 in mammalian cells. HEK293T cells were transduced with lentivirus encoding HAGFPHis and whole cellular extracts were prepared for protein purification using a HisTrap column. (A) Affinity column purification chromatogram in which the black trace represents the absorbance at 280nm. The dashed line represents the imidazole percentage throughout the purification procedure. mAU, arbitrary mili-absorvance Units. (B) SDS-PAGE (top panel) of the eluted fractions and western blot analysis using the indicated antibody. Protein molecular weights markers are indicated on the left. mAU, arbitrary mili-absorvance Units.

3.3.10 Zrsr1 interacts with U2AF\textsuperscript{65} and associates with spliceosomal components

To determine whether Zrsr1 can bind directly to U2AF\textsuperscript{65}, we used purified recombinant proteins and size exclusion chromatography (Figure 3.19). When individual members of U2AF\textsuperscript{35}-family were applied to a Sephadex S-200 column, the elution profile revealed a single peak with a narrow shape (dashed lines, Figure 3.19). Unexpectedly, all proteins eluted rapidly with a retention time close to the exclusion limit of the column. This observation suggests that members of the U2AF\textsuperscript{35}-family have a tendency to self-aggregate \textit{in vitro} forming high molecular weight (>200 kDa) complexes. In agreement with these results, previous reports suggested that both U2AF\textsuperscript{65} and U2AF\textsuperscript{35} self-interact \textit{in vivo} and \textit{in vitro} (Chusainow, Ajuh et al. 2005; Rino, Desterro et al. 2008). For binding assays, equimolar amounts of recombinant U2AF\textsuperscript{65} and a member of the U2AF\textsuperscript{35}-family were incubated overnight and then applied to a Sephadex S-200 column. Analysis of eluted fractions by western-blot shows first, a shift in the elution profile relative to fraction 1 where the single protein elutes; and second, that U2AF\textsuperscript{65} co-elutes with U2AF\textsuperscript{35}, U2AF\textsuperscript{26} and Zrsr1 (Figure 3.19A, B, C).
We therefore conclude that recombinant U2AF\textsuperscript{65} and Zrsr1 interact with each other \textit{in vitro}. Next, we asked whether Zrsr1 associates with spliceosomal proteins \textit{in vivo}. We performed pull-down experiments from MEL cell extracts with immobilized Zrsr1 recombinant protein (Figure 3.20A). Zrsr1 tagged with HA and hexa-histidine was expressed in HEK293T cells, purified by affinity chromatography and immobilized on Ni\textsuperscript{2+}-sepharose beads. The beads were then incubated with cellular extracts from undifferentiated and differentiated MEL cells. Cell extracts were pre-treated with RNase A to avoid detection of protein interactions mediated by RNA. Pulled-down complexes were eluted, concentrated and analyzed by western-blotting (Figure 3.20B). As a control, we performed the pull-down experiment using immobilized green fluorescent protein (GFP). The results show that Zrsr1 associates with U2AF\textsuperscript{65}, SF1 and ASF/SF2, but not with hnRNAPA1. Similar results were observed with extracts from undifferentiated (T0) and differentiated (T48) MEL cells. An exception was ASF/SF2, which appears less enriched in pull-downs from differentiated cells. U2AF\textsuperscript{35} was also pulled down, but in much lower amount than U2AF\textsuperscript{65}, suggesting that Zrsr1 is replacing U2AF\textsuperscript{35} when it binds to U2AF\textsuperscript{65}. 
Figure 3.19 Zrsr1 binds to U2AF^{65}. Recombinant U2AF^{65}, U2AF^{35}, U2AF^{26} and Zrsr1 proteins were expressed in *Escherichia coli*, purified and applied to a Sephadex S-200 column. The size exclusion chromatography profile of the indicated individual proteins is represented by dashed lines on left panels, and Western blot analysis of the indicated fractions is shown on right panels. Recombinant purified U2AF^{65} was mixed with equimolecular amount of U2AF^{35} (A), U2AF^{26} (B), and Zrsr1 (C) and allowed to interact overnight at 4°C prior to size fractionation in a Sephadex S-200 column. The chromatography profile of each protein mix is represented by solid lines on left panels, and Western blot analysis of the indicated fractions is shown on right panels. The antibodies used on Western blots are indicated on the right.
3.3.11 Subcellular localization of the Zrsr1 protein in MEL cells

At steady state, SR proteins like the U2AF complex are known to be localized in the nucleus and colocalize in nuclear speckles (Gama-Carvalho, Krauss et al. 1997). Our immunofluorescence data demonstrate that in a steady state situation, Zrsr1 is predominantly localized in the nucleus of MEL cells (Figure 3.20A). Moreover, consistent with the results already described, upon erythroid differentiation Zrsr1 protein levels are clearly up-regulated as seen by the increase of the nuclear immunofluorescence signal (Figure 3.21A).
Figure 3.21 - Subcellular localization of Zrsr1 upon erythroid differentiation of MEL cells. (A) MEL cells were induced to differentiate with DMSO for 48 h. Undifferentiated and differentiated cells were fixed with formaldehyde, permeabilized with Triton X-100 and hybridized with anti-Zrsr1 antibody (c). DNA was stained with DAPI (a) and F-actin with Phalloidin-Alexa488 (b). Bar 10 µm. (B) Subcellular distribution of Zrsr1 upon erythroid differentiation. MEL cells either induced to differentiate with DMSO or undifferentiated were fractionated to isolate the cytoplasmic (Cyto), nucleoplasmic (Nucleo) and chromatin (Chroma) associated proteins. Each fraction was blotted into a nitrocellulose membrane and probed with the indicated antibodies. Protein molecular weight markers are shown on the left.

Despite the nuclear localization of Zrsr1 in a steady state situation, it is known that some SR proteins shuttle continuously between the nucleus and the cytoplasm (Gama-Carvalho, Carvalho et al. 2001; Sanford, Gray et al. 2004). To investigate the subcellular localization of the Zrsr1 protein we used a biochemical approach to isolate cytoplasmic, nucleoplasmic and chromatin associated proteins (Pandya-Jones and Black 2009) in undifferentiated and differentiated MEL cells (Figure 3.21B). Consistent with previous reports, western-blot analysis revealed that histone H3 was exclusively detected in the chromatin fraction (Pandya-Jones and Black 2009). Our results demonstrates that Zrsr1 shows a nucleocytoplasmic localization, with high levels of protein present in the cytoplasm, when compared to the U2AF$^{65}$ and U2AF$^{35}$. In agreement with previous studies both subunits of the U2AF complex, U2AF$^{65}$ and
U2AF\textsuperscript{35}, were found mainly in the nuclear fractions (nucleoplasm and chromatin associated fractions), although cytoplasmic U2AF was also detected. These results are consistent with previous reports were U2AF was shown to have a nucleocytoplasmic distribution (Gama-Carvalho, Krauss et al. 1997), where the U2AF\textsuperscript{65} subunit is detected in the cytoplasmic fraction associated with spliced RNAs (Gama-Carvalho, Barbosa-Morais et al. 2006), suggesting that U2AF-related proteins, like Zrsr1, may be involved in novel cellular functions besides splicing.

3.3.12 The Zrsr1 gene is required for normal erythropoiesis

To investigate the role of Zrsr1 in vivo we analysed a mouse strain in which the entire coding region of the Zrsr1 gene has been deleted and replaced with a PGK-gpt-neo cassette via homologous recombination. (Sunahara, Nakamura et al. 2000) (Figure 3.22).

\textbf{Figure 3.22- Targeted disruption of the Zrsr1 gene.} Schematic of Zrsr1 deletion strategy and partial restriction map of the mouse Zrsr1 locus. Shaded dark grey rectangles represent two exons of the mouse Commd1 gene. Light grey rectangle, represents the intronless Zrsr1 gene. The orientation of the transcription is indicated by arrowheads within exons of the genes. The targeting vector contains the flanking genomic sequences, a neomycin gene (neo), and a thymidine kinase gene (tk). Homologous recombination is indicated by crossed lines and should result in the predicted disrupted allele. Primers used for genotyping and RT-PCR are indicated by small arrows in the Zrsr1 exon and in the targeting vector.
Western blot analysis of total protein extracts isolated from spleen confirmed that knockout mice do not express Zrsr1 protein (Figure 3.23). Although viable and fertile, the knockout mice do not show any visible abnormal phenotype.

Figure 3.23- Analysis of the Zrsr1 KO mice. (A) Genotyping the Zrsr1 KO mice. Genomic DNA isolated from tail was analyzed by PCR (see text for details). DNA weight markers are indicated on the left. (B) Immunoblot analysis of Zrsr1 in total protein extracts isolated from WT and Zrsr1KO spleens. β-Actin was used for loading control. Protein molecular weights are shown on the left.

Haematological analysis shows that the number of red blood cells (Figure 3.24A) and reticulocytes (Figure 3.24B) is not significantly affected. However, compared to normal littermates, the Zrsr1 (-/-) mice have lower haematocrit (Figure 3.24C), reduced mean corpuscle volume (Figure 3.24D) and average erythrocyte area (Figure 3.24F). Despite the smaller erthyrocytes, the knockout mice have normal haemoglobin levels (Sup. TableI). The knockout mice show additionally a reduced number of lymphocytes, normal number of monocytes and increased spleen size (Sup. TableI), which is a characteristic feature of haematological stress (Kam, Ou et al. 1999). In conclusion, we observe that red blood cells circulating in the blood of knockout mice are smaller than normal, indicating a role for Zrsr1 in erythrocyte maturation.
3.3.13 Erythroid-specific alternative splicing decisions are altered in Zrsr1-deficient mice

Alternative splicing has been reported to modulate gene expression in late erythroblasts (Welch, Watts et al. 2004; Keller, Addya et al. 2006; Yamamoto, Clark et al. 2009). Based on these recent findings several novel erythroid stage-specific alternative splicing events were reported. Among the exons found to be specifically included is a novel alternative exon (exon 8; see Figure 3.25A) which is predicted to introduce changes in the reading frame of the MBNL2 (muscleblind-like 2) gene (Yamamoto, Clark et al. 2009). This splicing switch promotes the inclusion of a novel penultimate coding exon which inserts a new peptide and also changes the reading frame.
frame of \textit{Mbnl2}. Similar specific alternative splicing events were also found for \textit{HNRPLL} (heterogeneous nuclear ribonucleoprotein L-like) and \textit{SNRP70} (U1 small nuclear ribonucleoprotein 70K), although it still remains unclear the exact biological role of these novel truncated proteins in the control of erythropoiesis. To determine the effect of \textit{Zrsr1} loss in the control of erythroid specific alternative splicing events, we have isolate total RNA from spleen and bone marrow of both wild-type and \textit{Zrsr1} KO mice and accessed by RT-PCR the relative inclusion/exclusion amounts of the novel \textit{Mbnl2} exon. Total RNA was Oligo-(dT) reverse transcribed and alternative splicing diagnostic primers were used to address if \textit{Zrsr1} is involved in the regulation of the erythroid specific exon 8 in the \textit{Mbnl2} gene.

Figure 3.25- Analysis of the splicing pattern of \textit{Mbnl2} gene in wild type and \textit{Zrsr1} KO mice. (A) Transcript map of the two \textit{Mbnl2} gene isoforms. Primers used to analyze the different exon combinations are indicated in the figure. (B) Splicing analysis by RT-PCR of the \textit{Mbnl2} gene in wild-type and \textit{Zrsr1} knock-out mice in bone marrow and spleen. The analysed mRNA isoforms are indicated in the right hand side of the graph. GAPDH gene was used as a normalization control. These results were reproducible in at least three independent animals. (C,D) Quantification by image densitometry of the results in (B). Intensities for each \textit{Mbnl2} RNA species detected in bone marrow (C) and spleen (D) were normalized to GAPDH.
As shown in Figure 3.25B, we found a clear shift in the pattern of *Mbnl2* exon 8 inclusion in mice deficient for Zrsr1, particularly in the bone marrow (Figure 3.25B and C). Therefore, this results demonstrate that Zrsr1 is involved in the control of previously described specific erythroid splicing events.

### 3.4 Discussion

Alternative splicing is known to be regulated at different steps of spliceosome assembly by different splicing factors that rely on *cis*-acting elements (Jurica and Moore 2003; Mollet, Barbosa-Morais et al. 2006; Chen and Manley 2009). Correct spliceosome assembly is a tightly regulated sequence of events were the U2AF splicing factor and their related proteins play an important role (Ruskin, Zamore et al. 1988; Zamore and Green 1989; Zamore and Green 1991; Zamore, Patton et al. 1992; Zhang, Zamore et al. 1992; Wu, Romfo et al. 1999; Kent, Reayi et al. 2003; Heyd, ten Dam et al. 2006; Hastings, Allemand et al. 2007). There is a growing list of evidences suggesting that alternative splicing uses combinatorial interactions of many positively and negatively acting proteins to regulate tissue specific splicing events (Matlin, Clark et al. 2005; Licatalosi and Darnell 2010). In this context, *Zrsr1* emerge as a particular appealing candidate since in the beginning of our studies little was known about the biological functions of this gene.

The different tissue expression profile of the U2AF35-related genes, as shown by our microarray data analysis in several human and mouse tissues, seems to support the idea that these proteins may regulate important functions in a tissue-specific fashion. We now provide evidence that the previously uncharacterized member of the U2AF35-family, *Zrsr1*, is a novel erythroid tissue specific gene signature. As obtained by microarray data analysis from two independent datasets, *Zrsr1* was found consistently up-regulated upon erythroid differentiation. In fact, since the expression of *Zrsr1* changes at least 1.5-fold more than in any other differentiation process, it is considered to be an erythroid tissue signature (Grosso, Gomes et al. 2008). In good agreement with these results, is the spatiotemporal expression pattern of *Zrsr1* in the mice embryo. In fact, the *in situ* hybridization data further reinforces our results since *Zrsr1* is found to be highly expressed in the fetal liver, which is known to be a major place for embryonic
erythropoiesis (McGrath and Palis 2008). In order to validate the microarray data analysis, we used two very well established murine cellular models of erythroid differentiation, I/11 (von Lindern, Deiner et al. 2001) and MEL (Volloch and Housman 1982; Patel and Lodish 1987) cells. Using qRT-PCR for both cell models, we were able to demonstrate that Zrsr1 mRNA levels are up-regulated during terminal erythroid differentiation, while other U2AF-related genes were found to be down-regulated. Interestingly, previous reports have described that U2AF1 regulates transcripts encoding key proteins essential for cell cycle regulation while RNAi-mediated knockdown of U2AF1 was shown to impair cell proliferation with massive delay in mitotic progression (Pacheco, Moita et al. 2006). In this way, decreased transcript levels of U2AF1 could be related with terminal erythroid differentiation since this process is characterized by a terminal cell-cycle arrest (Kiyokawa, Richon et al. 1993; Dolznig, Boulme et al. 2001).

Additionally, our results demonstrating a transcriptional activation of the Zrsr1 upon erythroid induction were also confirmed by increased levels of H3K9ac at the promoter region. In fact, acetylation of histone H3 lysine 9 is a post-transcriptional modification well known to occur in nucleosomes associated with promoters of actively transcribed genes (Liang, Lin et al. 2004). Contrasting, decreased levels of H3K9ac at the U2af1 promoter are correlated with decreased transcriptional activity (lower RNA Pol II occupancy) upon erythroid differentiation, and thus, with down regulation of the mRNA levels. Therefore, we show that regulation of the U2AF35-family members upon erythroid differentiation might be controlled at the chromatin level and, most likely, other histone modification are involved in the expression regulation of these genes.

To study a possible interconnection between histone modifications and co-transcriptional pre-mRNA events, we also performed ChIP experiments to investigate the distribution of the histone H3K36me3 modification along a intron-containing gene (U2af1) and a introless gene (Zrsr1), in our model cell system were we show that the transcriptional levels of both genes can be manipulated. Our results seems to support the connection between the histone H3K36me3 modification and splicing since there is no accumulation of this histone mark in a intonless genes when compared to a gene with intron-exon structure. Moreover, increased levels of transcription in the intonless gene does not change the levels of H3K36me3 arguing that the splicing machinery may directly regulate the deposition of this histone mark. In this way, accessing the distribution of H3K36me in a model system where the spliceosome could be easily
inhibited, could shed some light about these important questions. Additionally, genome-wide studies unrevealing other histone modifications that may be involved in exclusion or inclusion of an exon in a given mRNA will give us more details about the relationship between the “histone-code” and alternative splicing.

Although our evidences showing that Zrsr1 is an erythroid tissue-specific signature are inferred from microarray and qPCR data, it is known that changes in splicing factor mRNA levels may not necessarily reflect on protein expression due to post-transcriptional regulation (Boutz, Stoilov et al. 2007; Makeyev, Zhang et al. 2007). In this way, to access the protein expression profile of the U2AF-family members we produce and characterize a polyclonal anti-Zrsr1 antibody which was found to be highly specific against this protein. Western-blot analysis of total protein extracts from either I/11 and MEL cells, shows that upon erythroid differentiation Zrsr1 protein is found consistently up-regulated. Interestingly this up-regulation of Zrsr1 combined with unaltered levels of U2AF35, raises the possibility that Zrsr1 could replace U2AF35 in the canonical U2AF-complex, allowing the formation of a distinct heterodimers which could regulate specific splicing events.

The conserved structural features of the U2AF35-family members (Mollet, Barbosa-Morais et al. 2006) most likely makes them able to interact with U2AF65. In fact, some U2AF35-related proteins like U2AF26 (Heyd, ten Dam et al. 2006) and Zrsr2 (Tronchere, Wang et al. 1997) where shown to interact with U2AF65. To access if Zrsr1 is also able to associate with U2AF65, we used two different strategies to demonstrate the interaction U2AF65/Zrsr1. Additionally, from our gel filtration based assay to assemble the U2AF-complexes, we were able to isolate complexes with distinct protein stoichiometry, which argues in favour of the existence of U2AF complexes with different U2AF65 and U2AF35-related composition has already been reported in previous studies (Rino, Desterro et al. 2008). Nevertheless, biological significance of such complexes remain to elucidate but they may be involved in the recognition of 3’ splice sites with distinct relative strengths (Reed and Maniatis 1988; Reed 1989). Our pull-down experiments with recombinant protein, further supports our results since we show that Zrsr1 is able to interact with U2AF65. Moreover, we show the interaction of Zrsr1 with other splicing factors, namely SF1/BBP, a splicing factor known to be involved in the 3’ splice site recognition (Guth and Valcarcel 2000) and ASF/SF2, an SR protein known to be involved in erythroid specific splicing events (Yang, Huang et al. 2005).
Interestingly, Zrsr1 could also pull-down U2AF^{35}, which could suggest that this protein is part of a larger U2AF complex that could engage network interactions during spliceosome assembly. Although the same interaction in observed for Zrsr2 (Tronchere, Wang et al. 1997), a protein that shares 94% aminoacid homology with Zrsr1, the biological significance of such interaction remains to elucidate.

The subcellular distribution of splicing factors has been shown to be complex and dynamic, and there is an increasing number of evidences showing that this may play an important role in the functional regulation of several splicing factors (Gama-Carvalho, Carvalho et al. 2001; Sanford, Gray et al. 2004; Gama-Carvalho, Barbosa-Morais et al. 2006). Our immunofluorescence data demonstrate that in a steady state situation, Zrsr1 is predominantly localized in the nucleus of MEL cells, which is in agreement with previous studies were U2AF was shown to have a nuclear localization (Gama-Carvalho, Krauss et al. 1997). However, when analyzing the subcellular distribution of Zrsr1 using a biochemical approach that allows to fractionate cytoplasmic, nucleoplasmic and chromatin-associated proteins, Zrsr1 was found present in all fractions, indicating a nucleo-cytoplasmic subcellular localization. This localization of Zrsr1 is not totally surprising since it is known that other SR proteins like U2AF^{65} (Gama-Carvalho, Carvalho et al. 2001) and ASF/SF2 (Sanford, Gray et al. 2004) have the same behavior. Unexpectedly, Zrsr1 was found highly present in the cytoplasm when compared to U2AF^{65} and U2AF^{35} raising the possibility that Zrsr1 may play an important function in the cytoplasm. Therefore, the subcellular localization of Zrsr1 argues that its functions are not limited to the nucleus, and like other shuttling SR proteins (Gama-Carvalho, Carvalho et al. 2001; Sanford, Gray et al. 2004; Gama-Carvalho, Barbosa-Morais et al. 2006), Zrsr1 may have additional roles in mRNA transport and/or in cytoplasmic events such as mRNA localization, stability, or regulation of translation.

The differentiation program of erythroblasts into mature erythrocytes requires an orchestrated gene expression program that insures accurate production of the appropriate stage-specific proteome as the cells become progressively more specialized (Chen, Liu et al. 2009). During terminal differentiation to enucleated and haemoglobinised erythrocytes, red blood cells progenitors undergo drastic changes in the proteome (Tsiftsoglou, Pappas et al. 2003; Tsiftsoglou, Pappas et al. 2003). To accomplish such changes, the intrinsic gene expression program must be modulated by
the expression of erythroid specific transcription factors, alternative mRNA splicing specific events and by tight translational regulation (Testa 2004).

The impact of alternative splicing in the control of erythropoiesis is suggested by the findings that 35% of erythroid-specific genes have alternative 5´ exons that increase the complexity within the N-terminal regions of the transcripts studied (Tan, Mohandas et al. 2006). More recently, new erythroid stage-specific alternative splicing switches were identified (Yamamoto, Clark et al. 2009). Among the found regulated transcripts are three splicing factors (SNRP70, HNRPLL, MBNL2), which may suggest that the regulation of specific alternative splicing events is critical for late erythroid differentiation (Yamamoto, Clark et al. 2009). The in vitro data collected in our work, strongly supports our hypothesis that the Zrsr1 gene is not only a tissue specific signature but also could be involved in the regulation of erythroid cells differentiation.

To investigate in vivo the role of Zrsr1 in erythropoiesis we analyzed the blood composition of a Zrsr1-deficient mice. Although these mice do not show any clear and visible phenotype, the blood cells counts of Zrsr1-deficient mice shows that they produced fewer and smaller red blood cells. Zrsr1 was mapped into the mouse chromosome 11 (Hayashizaki, Shibata et al. 1994; Tada, Tada et al. 1994) and it was found to be expressed exclusively from the paternally inherited chromosome due to promoter hiper-methylation in the maternal allele (Hatada, Kitagawa et al. 1995). Interestingly, mice carrying maternal duplication or paternal deficiency for proximal chromosome 11 are smaller than normal mice (Cattanach and Kirk 1985). Therefore, it is intriguing to argue that Zrsr1 could a be candidate gene that may function in the regulation of erythrocyte size, perhaps as a splicing factor that is involved in the regulation of specific genes involved in the control of cell size. In fact, our results seems to support this hypothesis since red blood cells from Zrsr1-deficient mice are smaller than in wild-type animals. Additionally, we have found that Zrsr1-deficient mice show abnormal splicing patterns of a previously characterized erythroid specific splicing event (Yamamoto, Clark et al. 2009). In fact, we show that Zrsr1 is involved in the regulation of the erythroid specific exon 8 in the Mbnl2. Mammalian Mbnl proteins are described to contain two pairs of highly conserved zinc fingers, which bind to pre-mRNA in order to regulate alternative splicing (Pascual, Vicente et al. 2006). Since alternative splicing is regulated through the use of combinatorial interactions between splicing factors, Zrsr1 may have evolve in order to regulate specific alternative splicing
events of mRNAs that code for splicing factors important for erythroid cells differentiation. Nevertheless, although the exact mechanisms by which Zrsr1 promotes its biological functions are still elusive our results provide evidences to support the idea that this gene is not only an erythroid tissue-specific signature, but also may have evolve new functions in the control of erythrocytes size by regulating erythroid specific alternative splicing events. Therefore, identification of more Zrsr1 RNA targets will be a major challenge in the future since it may provide more clues about the exact biological role of Zrsr1 in the control of key erythroid cells genes.
THE RETROTRANSPOSED MOUSE ZRSR1 GENE ACQUIRED A NEW FUNCTION IN ERYTHROID CELLS
Chapter 4

Concluding Remarks and Future Perspectives
Chapter 4

4.1 Concluding Remarks and Future Perspectives

In the past few years several exciting new findings have significantly changed our view about the complex network of events that coordinates gene expression. The availability of human (Lander, Linton et al. 2001) and mouse (Waterston, Lindblad-Toh et al. 2002) genome sequences opened a new era in the RNA world.

The low number (~20,000-25,000) and split nature of eukaryotic genes require an important physiological mechanism capable of producing a large number of mRNA’s in order to generate the complex proteome of higher organisms (Nilsen and Graveley 2010). In eukaryotes, during gene expression, non-coding sequences (introns) are removed from pre-mRNA, while coding sequences (exons) are joined together, to generate a mature mRNA. This process, called alternative splicing, is now accepted as the major mechanism to generate the proteome diversity (Nilsen and Graveley 2010). In fact, recent studies using high-throughput sequencing data indicate that as much as 95-100% of human pre-mRNA’s are processed to generate multiple mRNA’s (Pan, Shai et al. 2008; Wang, Sandberg et al. 2008). Noteworthy, is the fact that the extent of alternative splicing is now correlated with organism complexity. *C. elegans*, *D. melanogaster* and mammals have about 20000 (1998), 14000 (Adams, Celniker et al. 2000) and 20000 ( Clamp, Fry et al. 2007) genes, respectively, but mammals are clearly much more complex than worms or flies. Nevertheless, it is known that mammals have more alternative splicing events when compared to worms (Hillier, Reinke et al. 2009) or flies ( Stolc, Gauhar et al. 2004). In this way, it is very tempting to speculate that “non-conserved” changes in splicing patterns or splicing factors are likely to be responsible for the different complexity between species.

The biochemical mechanisms that control splice-site usage, and therefore alternative splicing, are complex and in large part remain poorly understood (Matlin, Clark et al. 2005). It seems clear that there cannot be specific and distinct factors dedicated to each of the more than 100,000 alternative splicing decisions that occur in human cells (Nilsen and Graveley 2010). Therefore, it is expected that only a small
number of factors are specifically dedicated to one or a few alternative splicing events, as the remarkable mechanism of sex-determination in *D. melanogaster* (Sanchez 2008).

Despite intense research, the mechanisms leading to splice site recognition are not fully understood, although the unique behaviour of several members of the U2AF\textsuperscript{35} family seems to indicate that some of them may be important targets for regulation.

Understanding the details of early events in spliceosome assembly is important because splicing is frequently regulated during these steps. Therefore, within this context our research aimed to complement the current knowledge about the biological role of the U2AF\textsuperscript{35}-related protein family members. In particular, we focused in the Zrsr1 splicing factor, which at the beginning of this work it’s functions and clues to the biological roles where this protein may be involved were still unknown. In this chapter, we present the main conclusions that can be drawn from our results, and discuss then in an integrated perspective.

### 4.1.1 Why splicing factors may regulate alternative splicing in a tissue-specific manner?

The development of high-throughput technologies in the past recent years is allowing us to have a more complete spatial and temporal gene expression chart of splicing regulators during development, cell differentiation (Grosso, Gomes et al. 2008) or in human pathologies like cancer (Grosso, Martins et al. 2008). It is now accepted that differential expression of splicing factors is an important mechanism that is involved in the control of cell type or tissue specific alternative splicing events. Indeed, 50% or more of alternative splicing isoforms are known to be differently expressed among tissues (Wang, Sandberg et al. 2008), which is in agreement with the view that alternative splicing is submitted to tissue-specific regulation. Moreover, it is currently acknowledge that differences in relative abundances or activities of multiple proteins influence specific splicing decisions. A good example is provided by the findings that the relative concentration of hnRNP A1 and the SR protein ASF/SF2 can influence the alternative splicing pattern of a model transcript (Mayeda and Krainer 1992; Caceres, Stamm et al. 1994). The increased evidence that these splicing regulators are often expressed in a tissue specific fashion and that post-transcriptional modifications can
CONCLUDING REMARKS AND FUTURE PERSPECTIVES

regulate their activity, provides an additional layer of complexity and regulation. Indeed, it is believed that these mechanisms provide a very fine tuned way of controlling splice site selection.

Although this model of regulation is very well accepted in the field, very few examples of tissue-specific splicing factors have been identified. Notable exceptions are Nova (Buckanovich, Posner et al. 1993), nPTB (Polydorides, Okano et al. 2000), Fox1 and Fox 2 (Jin, Suzuki et al. 2003), ESRP1 and ESRP2 (Warzecha, Sato et al. 2009) and nSR100 (Calarco, Superina et al. 2009), which are all expressed in a tissue-specific manner.

In our work we were very interested in the U2AF35-related family of proteins. The similarity and differences between the U2AF-related proteins, as discussed in our review (Mollet, Barbosa-Morais et al. 2006) imply that they might have evolved distinct new functions in the control of gene expression in complex organisms. Clues to the biological processes in which these proteins participate may be obtained by determining their tissue expression pattern. In this way, in our work we started by systematically assess the tissue distribution of the U2AF35-related genes in human and mouse tissues as well as in several differentiation processes. By analyzing microarray data sets our results show that Zrsr1, a previously uncharacterized member of the U2AF35-family, is up-regulated during erythropoiesis and, therefore, is consider to be an erythroid tissue-specific signature. These results were further validated in two cell models of erythroid differentiation, supporting the hypothesis that the differentially expression of splicing factors might regulate tissue specific splicing events. In this way, it is clear that a major task for the future will be to determine more tissue-specific splicing factors and identify which events are regulated by the differential expression of the genes identified in our study.

Future studies about alternative splicing regulation should also focus on how splicing factors are able to switch key splicing events during development, differentiation or in response to an external stimuli, and how misregulation of alternative splicing leads to disease. Complete characterization of tissue-specific patterns of expression is a fundamental issue to define mechanisms of alternative splicing regulation in different cell types. Moreover, identification of the regulatory motifs that are recognized by each splicing factor will be also an important approach to complete a more clear picture into the regulatory mechanisms of alternative splicing.
Although the information about the exact regulatory motifs which are recognized by a given splicing factor is still very limited, there is a growing number of alternative splicing regulators with known binding sites. In fact, the development of CLIP-seq methods allowed the identification of RNA maps for the splicing proteins NOVA (Ule et al., 2006; Licatalosi et al., 2008), FOX (Zhang et al., 2008; Yeo et al., 2009), and ASF/SF2 (Sanford, Coutinho et al. 2008). These RNA maps define the regulatory networks of alternative splicing and can be used to predict the outcome of alternative splicing in other genes. In a near future, these type of information should be used to create databases which should include comprehensive information on the expression patterns of splicing factors and the definition of potential targets with positions of binding motifs on these transcripts.

4.1.2 Is there a link between histone modifications and splicing?

The observation that pre-mRNA splicing can occur cotranscriptionally, while the RNA is still attached to the DNA by RNA Pol II, indicates that splicing and transcription are at least temporally and mechanistically coupled. In fact, there are studies showing that exons in the nascent transcript become tethered to the elongating transcription complex as they emerge from RNA Pol II (Dye, Gromak et al. 2006). It seems possible that this tethering mechanism could be required for a given exonic sequence to be included in the final mRNA. Indeed, some authors speculate that when RNA Pol II quickly transcribes an exon, they escape the tethering mechanism(s) and are skipped. However, when exons are transcribed slowly, they are efficiently tethered and the sequence that is included in the mRNA (Nilsen and Graveley 2010). There are now growing evidences that the chromatin structure influences transcription (Knezetic and Luse 1986; Hodges, Bintu et al. 2009), but its role in subsequent RNA processing mechanisms is still poorly understood. How eukaryotic genomes are manipulated within a chromatin environment is a current and fundamental issue in biology.

Remarkable advances have been made in the recent years in unrevealing the role of specific chromatin modifications in the outcome of alternative splicing. Several genome-wide studies started to map specific histone modifications that are enriched in exons. From these reports, H3K36me3 (Kolasinska-Zwierz, Down et al. 2009), H3K79me1, H4K20me1 and H2BK5me1 (Schwartz, Meshorer et al. 2009) were found
to be present in nucleosomes positioned in exons. Despite these observations raised the possibility of a combinatorial cross-talk between the “histone-code” and the splicing machinery, it remained to establish a causal relationship.

The CHD1 chromatin remodelling ATPase, which is known to interact with H3K4me3, was shown to bind spliceosomal components, while knockdown of CHD1 and reduction of H3K4me3 levels are correlated with changes in the efficiency of splicing (Sims, Millhouse et al. 2007). In addition, changes in two histone modifications, H3K9me2 and H3K27me3, were correlated with splicing patterns (Allo, Buggiano et al. 2009).

Our findings that upon MEL cells differentiation we are able to manipulate the transcription levels of both U2AF35 and Zrsr1 opened us a window to study the interconnection between the mechanisms of gene expression. In fact, to investigate this hypothesis, our model system emerge as particularly appealing for that purpose since the U2AF35 gene has a classical exon-intron configuration, while the related Zrsr1 mouse gene was shown to be imprinted and intronless (Hatada, Sugama et al. 1993; Hayashizaki, Shibata et al. 1994; Hatada, Kitagawa et al. 1995). Our results support the importance of the histone H3K36me3 modification in splicing since there is no accumulation of this histone mark in a intonless genes when compared to a gene with intron-exon structure. In good agreement with our results, a recent study made considerable advances in establishing the causal role of histone modifications and splicing (Luco, Pan et al. 2010). In fact, the histone H3K36me3 modification was shown to have an important role in a splicing pattern since this histone mark promotes the recruitment of PTB to the pre-mRNA, which leads to the inclusion of the exon in the final mRNA (Luco, Pan et al. 2010).

The interconnection between the gene expression mechanisms seems to argue that they evolved as a quality control surveillance mechanism. Failure to complete a co-transcriptional checkpoint could for instance result in the production of missspliced mRNAs which could be lead to the development of a pathology. The histone methyltransferase SET2, which is responsible for the H3K36me3 histone modification, was found to be inactivated in sporadic clear renal cell carcinomas (Duns, van den Berg et al. 2010), and to have lower levels of expression in human breast cancers (Al Sarakbi, Sasi et al. 2009; Newbold and Mokbel 2010). Could the loss of SET2, hence lower levels of the histone mark H3K36me3, lead to a misregulation of splicing patterns
which could be implicated in the process of carcinogenesis? These are still open questions, but these observations seem to highlight the significance of the histone H3K36me3 modification as an important link between the gene expression steps in which misregulation can lead to human pathologies.

Although all the studies performed so far seem to argue for a direct link between the histone code and the splicing machinery, it remains elusive how histone marks affect splice site choice. Is there an epigenetic memory contained in the histone-code that not only determine the activity level of a gene but also regulates the alternative splicing patterns during physiological process such as development and differentiation? Is there a combinatorial cross-talk between different histone modifications in the outcome of splicing? Or the readout of each histone modification is independently exerted?

By using genome-wide studies, cataloguing and mapping the histone-code should in a near future give us more clues about the close relationship of the histone code and the control of alternative splicing events.

4.1.3 Subcellular localization: a way to control the U2AF-related proteins function?

In this work we have shown for the first time that like other U2AF-related proteins, Zrsr1 has a nucleocytoplasmic localization. This behaviour of Zrsr1 is in agreement with previous reports where it was found that other SR proteins like U2AF65 (Gama-Carvalho, Carvalho et al. 2001) and ASF/SF2 (Sanford, Gray et al. 2004) have also nucleocytoplasmic localization. Moreover, splicing factors like SRp20 and 9G8 are known to promote nucleo-cytoplasmic export of mRNA (Huang and Steitz 2001). In this context, one hypothesis is that this subcellular localization behaviour could be used as a mechanism to regulate the nuclear availability of a given splicing factor, similarly to what has been proposed for hnRNPA1 (van der Houven van Oordt, Diaz-Meco et al. 2000). On the other hand, the cytoplasmatic localization of an alternatively spliced isoform of U2AF26 and the known nucleo-cytoplasmic shuttling of U2AF26 and U2AF35 raises the question whether the U2AF35-related proteins may have other functions than their known role in alternative splicing. It was recently found that U2AF65 is associated with a specific subset of fully spliced mRNAs in the cytoplasm (Gama-Carvalho,
Barbosa-Morais et al. 2006), and it appears possible that this is also the case for Zrsr1, U2AF$^{26}$ and U2AF$^{35}$. However, whether the binding of cytoplasmic RNAs by U2AF$^{35}$-like subunits has any functional implication in mRNA transport and/or in cytoplasmic events such as mRNA localization, stability, or regulation of translation remains to be shown.

A growing number of examples make it increasingly clear that distinct steps of gene expression are performed and coordinated by multifunctional proteins (Lunde, Moore et al. 2007). Many proteins that have been identified as splicing factors could be as easily been described as export or transcription factors, and vice-versa. Our results and the previously known data provide evidence that the U2AF-related proteins may also be part of this group. Therefore, further research on this topic is clearly necessary to understand the true meaning of these observations.

4.1.4 The role of Zrsr1 in erythropoiesis

Since the definitive identification of RNA Binding Proteins (RBPs) and the discovery of their consensus motifs, the list of RBPs and the multitude of functions in which they participate has expanded enormously (Lunde, Moore et al. 2007). In recent years, biochemical approaches combined with genetic experiments and bioinformatic analysis of several sequenced genomes revealed a vast array of RBPs about which little is known. From what we have learned so far, it is clear that RBPs are critical components of the gene expression pathway in eukaryotes. Their capacity to regulate every aspect of the biogenesis and function of RNAs is remarkable. In fact, RBPs function in every aspect of RNA biology, from transcription, pre-mRNA splicing and polyadenylation to RNA modification, transport, localization, translation and turnover (Glisovic, Bachorik et al. 2008). However, it is also clear, that a great deal of information is still lacking about the structure of RBPs, their mode of interaction with RNAs and the specific arrangements of these proteins in the RNP complex assemblies that they form on pre-mRNAs and mRNAs. Despite the remarkable progress that has already been made, there is an enormous number of RBPs that remain to be characterized and the development of new tools to study them, will provide in a near future more insights about the specific role of each RBP.
The present work has made original scientific contributions to understand the functional uniqueness of the U2AF$^{35}$-family of splicing factors. From these, Zrsr1, a protein of unknown functions, emerge as particularly appealing since it was found differentially expressed during the differentiation of erythrocytes. In this way, we can speculate that this protein may have evolved specific functions in the regulation of erythroid specific splicing events.

One of the best studied examples of erythroid tissue-specific regulated pre-mRNA splicing is the stage-specific alternative splicing switch of exon 16 in the 4.1R gene (Baklouti, Huang et al. 1996). Alternative splicing of exon 16 is a tightly regulated process (Gee, Aoyagi et al. 2000) since it is excluded in the erythroid progenitor cells but efficiently included in late erythroblasts (Baklouti, Huang et al. 1996). This alternative splicing event regulation is known to be functionally relevant since exon 16 inclusion yields 4.1R protein isoforms with high affinity for spectrin and actin, resulting in increased mechanical stability the erythroid membrane (Horne, Huang et al. 1993; Discher, Winardi et al. 1995). Mechanistically, this alternative splicing event was found to be modulated by changes in expression of antagonistic splicing factors, in particular, a decrease in expression of the splicing inhibitory factor hnRNP A1 (Hou, Lersch et al. 2002) relative to that of stimulatory factors Fox-2 (Ponthier, Schluepen et al. 2006) and SF2/ASF (Yang, Huang et al. 2005). Although few examples are known, it seems reasonable to assume that changes in splicing factor activity and/concentration could regulate not only the examples described but also a subset of other alternative splicing events that may be important for the erythroid differentiation alternative splicing program. In fact, our findings that Zrsr1 interacts with other splicing factors and regulates specific erythroid splicing events seems to indicate that this protein could be involved in the regulation of erythroid differentiation. However, as discussed, RBP are multifunctional proteins that are likely in to co-ordinate different steps of gene expression. Therefore, proteins that have been identified as essential splicing factors could easily be described as transcription factors, and vice-versa (Ladomery 1997). Although answers to these and other questions are likely to provide new clues about the functional diversity of U2AF$^{35}$-related proteins, like Zrsr1, we may argue that these proteins evolved in response to the needs of coordinating the multiple steps of gene expression in complex organisms.
The questions and hypothesis raised by this work should, in future studies, be addressed by correlating the functional specificity of the Zrsr1 protein with the RNA targets in erythroid cells. In this context, HITS-CLIP (Licatalosi, Mele et al. 2008) or RNA-Seq, using the Zrsr1 KO mice as a model, could allow us to shed some light about the mechanism(s) by which Zrsr1 is involved in the regulation of erythropoiesis.
References
References


REFERENCES


References


REFERENCES


Supplementary Material
### qPCR Primers

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mU2af1 isoform a Fw</td>
<td>TAGCCAGACCATGCCCTCTT</td>
</tr>
<tr>
<td>mU2af1 isoform a Rv</td>
<td>CGAAACCGTCAGCAGACT</td>
</tr>
<tr>
<td>mU2af1 isoform b Fw</td>
<td>TCCCCAAAAACAGTGACAGA</td>
</tr>
<tr>
<td>mU2af1 isoform b Rv</td>
<td>TCATCATAGTGCTCCCTGACATCTC</td>
</tr>
<tr>
<td>mZrsr1 Fw</td>
<td>TGCCAGACGCCCTCTC</td>
</tr>
<tr>
<td>mZrsr1 Rv</td>
<td>CCAAAATTGCAAACCTC</td>
</tr>
<tr>
<td>mZrsr2 Fw</td>
<td>AGGCTGAAAATGAGCTGGAA</td>
</tr>
<tr>
<td>mZrsr2 Rv</td>
<td>GCCTCGATCTCTTCAGTGACATCTC</td>
</tr>
<tr>
<td>U2af1L4 Fw</td>
<td>ACCGACTTTACGCCAGACCC</td>
</tr>
<tr>
<td>U2af1L4 Rv</td>
<td>CATCTGCAAGTCCGGCTGTG</td>
</tr>
<tr>
<td>mRI Fw</td>
<td>TCCATGTCAGCAGTGCTGAG</td>
</tr>
<tr>
<td>mRI Rv</td>
<td>TCAGGACACTGCAAAGACCA</td>
</tr>
<tr>
<td>mGAPDH Fw</td>
<td>CATTGTGGAAGGCTCATGAC</td>
</tr>
<tr>
<td>mGAPDH Rv</td>
<td>GCCCCACGGCCATC</td>
</tr>
</tbody>
</table>

**Table SI** - Sequence of primers used for qPCR.

### RT-PCR Primers

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRs1 Promoter Fw</td>
<td>TACGTACTGCGATAACCGA</td>
</tr>
<tr>
<td>mRs1 Promoter Rv</td>
<td>AAGCTTGAACGCTGAGTC</td>
</tr>
<tr>
<td>mRS1 Fw</td>
<td>TGCCAGACGCCCTCCTC</td>
</tr>
<tr>
<td>mRS1 Rv</td>
<td>CCACAAATTGCAAACCTC</td>
</tr>
<tr>
<td>mRs1 PolyA1 Fw</td>
<td>AACTAGCGGGCCCTGGATT</td>
</tr>
<tr>
<td>mRs1 PolyA1 Rv</td>
<td>AGGAGTTACGACATTGTC</td>
</tr>
<tr>
<td>mRs1 PolyA2 Fw</td>
<td>ATCGCATGCGAACATTAGT</td>
</tr>
<tr>
<td>mRs1 PolyA2 Rv</td>
<td>CACCAGTCACAGCAGACTGTC</td>
</tr>
<tr>
<td>mRs1 PolyA3 Fw</td>
<td>TGTCACAAAACAGGAAAGTG</td>
</tr>
<tr>
<td>mRs1 PolyA3 Rv</td>
<td>ATGGCAACTACGTGGCTAT</td>
</tr>
<tr>
<td>P120 intronF Fw</td>
<td>GATGATGCTGTGTCACCTGAGTC</td>
</tr>
<tr>
<td>P120 intronF Rv</td>
<td>CAGGTCCTCTGGAAGCCGCATTAG</td>
</tr>
<tr>
<td>mGAPDH Fw</td>
<td>CATTGTGGAAGGCTCATGAC</td>
</tr>
<tr>
<td>mGAPDH Rv</td>
<td>GCCCCACGGCCATC</td>
</tr>
</tbody>
</table>
Table SII- Sequence of primers used for RT-PCR.

<table>
<thead>
<tr>
<th>qPCR ChIP Primers</th>
<th>Sequence (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mU2AF35 Promoter Fw</td>
<td>TTCCCAATTGGACAGAGTTCG</td>
</tr>
<tr>
<td>mU2AF35 Promoter Rv</td>
<td>CAAGTATGCTCCATTTTTCG</td>
</tr>
<tr>
<td>mU2AF35 Exon2 Fw</td>
<td>TCGTTGCAACAATAAACCCAA</td>
</tr>
<tr>
<td>mU2AF35 Exon2 Rv</td>
<td>GGTTAGGTCAAATCCCTCT</td>
</tr>
<tr>
<td>mU2AF35 Exon3 Fw</td>
<td>GAGGGAAAATACCCGCTCTC</td>
</tr>
<tr>
<td>mU2AF35 Exon3 Rv</td>
<td>AATGTTCAGAGGGAATGG</td>
</tr>
<tr>
<td>mU2AF35 Exon6 Fw</td>
<td>CAACAGGCTGGTTGCTC</td>
</tr>
<tr>
<td>mU2AF35 Exon6 Rv</td>
<td>CACTCCACTGCTGTGCTC</td>
</tr>
<tr>
<td>mZrsr1 Promoter Fw</td>
<td>TACGTACTGGGATAACCCGA</td>
</tr>
<tr>
<td>mZrsr1 Promoter Rv</td>
<td>AAGCTTGAACCAGCTGATC</td>
</tr>
<tr>
<td>mZrsr1 Primer2 Fw</td>
<td>GAGGAGACGAAGGACAGAG</td>
</tr>
<tr>
<td>mZrsr1 Primer2 Rv</td>
<td>CACGCTCCCCGTTTTATTGTA</td>
</tr>
<tr>
<td>mZrsr1 Primer3 Fw</td>
<td>AGGGAGTCCGAGAGGAAGAG</td>
</tr>
<tr>
<td>mZrsr1 Primer3 Rv</td>
<td>TAGCTGGGCTCAGGTTGCT</td>
</tr>
<tr>
<td>mZrsr1 Primer4 Fw</td>
<td>AACTAGGCTGCGCCTGTGGAT</td>
</tr>
<tr>
<td>mZrsr1 Primer4 Rv</td>
<td>AGGGGTTCAGACATTTGTC</td>
</tr>
<tr>
<td>mZrsr1 Primer5 Fw</td>
<td>ATCGCATGCCAGAAATAAGT</td>
</tr>
<tr>
<td>mZrsr1 Primer5 Rv</td>
<td>CACCAGTCACAGCACACTGTC</td>
</tr>
</tbody>
</table>

Table SIII- Sequence of primers used for ChIP-qPCR.

<table>
<thead>
<tr>
<th>RT-PCR Primers</th>
<th>Sequence (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr137</td>
<td>GTGGTAAATCCAGTTAGATAG</td>
</tr>
<tr>
<td>E1r</td>
<td>GGGTTATAGAATGGATGTTA</td>
</tr>
<tr>
<td>pMC1neoF1</td>
<td>CCACACCGCTACCTAATA</td>
</tr>
<tr>
<td>RS1F1</td>
<td>AGCAGTCCAGTCCACAAAG</td>
</tr>
<tr>
<td>RS1R1</td>
<td>GATGGAGTCACACCTTCTT</td>
</tr>
<tr>
<td>mMbnl2E7Fw</td>
<td>ACTACAGCGGCTCGTACC</td>
</tr>
<tr>
<td>mMbnl2E9Rv</td>
<td>CATGCAATTGGTGCAATT</td>
</tr>
<tr>
<td>mMbnl2E8Fw</td>
<td>CTACACGCGCCAACCTCT</td>
</tr>
<tr>
<td>mMbnl2E8Rv</td>
<td>CTGATTTGGCTGGCTGT</td>
</tr>
</tbody>
</table>

Table SIV- Sequence of primers used for Zrsr1-deficient mice genotyping.
### qPCR ChIP Primers

<table>
<thead>
<tr>
<th>Primer</th>
<th>Sequence (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mU2AF&lt;sup&gt;35&lt;/sup&gt; Promoter Fw</td>
<td>TTCCCAATTGGACAGAGTCG</td>
</tr>
<tr>
<td>mU2AF&lt;sup&gt;35&lt;/sup&gt; Promoter Rv</td>
<td>CAAGTATTTCCGCCATTTTCG</td>
</tr>
<tr>
<td>mU2AF&lt;sup&gt;35&lt;/sup&gt; Exon3 Fw</td>
<td>GAGGGAAATAACCCGCTCTC</td>
</tr>
<tr>
<td>mU2AF&lt;sup&gt;35&lt;/sup&gt; Exon3 Rv</td>
<td>AATGTTCAGAGGGCAATGG</td>
</tr>
<tr>
<td>mU2AF&lt;sup&gt;35&lt;/sup&gt; Exon6 Fw</td>
<td>CAACAGCTGTGTGTTGCTC</td>
</tr>
<tr>
<td>mU2AF&lt;sup&gt;35&lt;/sup&gt; Exon6 Rv</td>
<td>CACTCCACTGCTGTTGCT</td>
</tr>
<tr>
<td>mZrsr1 Promoter Fw</td>
<td>TACGTACTGCGATAACCGA</td>
</tr>
<tr>
<td>mZrsr1 Promoter Rv</td>
<td>AAGCTTGAACCAGCTGAGTC</td>
</tr>
<tr>
<td>mZrsr1 Fw</td>
<td>TGCCAAGCAAGCCCTCCTCT</td>
</tr>
<tr>
<td>mZrsr1 Rv</td>
<td>CCACAAATTGCAACCTTCCA</td>
</tr>
<tr>
<td>PTB1 InG&lt;sub&gt;F&lt;/sub&gt;</td>
<td>GCAGTTCTGATTCTGAGGAGG</td>
</tr>
<tr>
<td>PTB1 InG&lt;sub&gt;R&lt;/sub&gt;</td>
<td>CTGGAGCAGACATGTAAGAGG</td>
</tr>
</tbody>
</table>

**Table SV**- Sequence of primers used for ChIP-qPCR.
**Table I: Blood cell counts for wild-type and Zrsr1-knockout mice**; WBC, white blood cell count (x10³/mL); RBC, red blood cell count (x10⁶/µL); HGB, hemoglobin concentration (g/dL); HCT, hematocrit (%); MCV, mean corpuscle volume (fL); MCH, mean corpuscle hemoglobin (pg); MCHC, mean corpuscle hemoglobin concentration (g/dL); PLTS, number of platelets (x10³/µL); MPV, mean platelet volume (fL); LYMP, lymphocyte number (x10³/µL); MONO, monocyte number (x10³/µL), SPLSZ, Spleen Size (Spleen weight/Body weigh; (mg/g)). SD, standard deviation. All data are the means ±SD.

<table>
<thead>
<tr>
<th></th>
<th>Zrsr1 (+/+)</th>
<th>SD</th>
<th>Zrsr1 (-/-)</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBC</td>
<td>7.6</td>
<td>0.96</td>
<td>3.91</td>
<td>0.37</td>
</tr>
<tr>
<td>RBC</td>
<td>10.23</td>
<td>0.59</td>
<td>9.45</td>
<td>0.60</td>
</tr>
<tr>
<td>HGB</td>
<td>16.06</td>
<td>0.77</td>
<td>15.2</td>
<td>0.46</td>
</tr>
<tr>
<td>HCT</td>
<td>45.33</td>
<td>0.95</td>
<td>39.47</td>
<td>1.75</td>
</tr>
<tr>
<td>MCV</td>
<td>44.3</td>
<td>0.87</td>
<td>39.67</td>
<td>0.97</td>
</tr>
<tr>
<td>MCH</td>
<td>15.7</td>
<td>0.173</td>
<td>15.17</td>
<td>1.35</td>
</tr>
<tr>
<td>MCHC</td>
<td>35.53</td>
<td>1.46</td>
<td>37.5</td>
<td>3.02</td>
</tr>
<tr>
<td>PLTS</td>
<td>543</td>
<td>73</td>
<td>668</td>
<td>180</td>
</tr>
<tr>
<td>MPV</td>
<td>7.85</td>
<td>2.05</td>
<td>7.66</td>
<td>0.34</td>
</tr>
<tr>
<td>LYMP</td>
<td>4.53</td>
<td>0.73</td>
<td>2.98</td>
<td>0.52</td>
</tr>
<tr>
<td>MONO</td>
<td>0.203</td>
<td>0.17</td>
<td>0.22</td>
<td>0.09</td>
</tr>
<tr>
<td>SPLSZ</td>
<td>2.5925</td>
<td>0.043</td>
<td>3.63</td>
<td>0.1412</td>
</tr>
</tbody>
</table>