Analysis of interacting partners and alternative splicing isoforms of human Splicing Factor 1

Crisci, Angela

UNIVERSITÉ DE GENÈVE
Département de biologie cellulaire

FACULTÉ DES SCIENCES
Professeur Angela Krämer

Analysis of interacting partners and alternative splicing isoforms of human Splicing Factor 1

THÈSE

présentée à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences, mention biologie

par
Angela Crisci

de
Maddaloni (CE)
(Italie)

Thèse N° 4748

Genève
Atelier de reproduction de Uni Mail
2014

UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES

Doctorat ès sciences
Mention biologie

Thèse de *Madame Angela CRISCI*

intitulée :

"Analysis of Interacting Partners and Alternative Splicing Isoforms of Human Splicing Factor 1"

La Faculté des sciences, sur le préavis de Madame A. KRÄMER, professeure ordinaire et directrice de thèse (Département de biologie cellulaire), Madame F. STUTZ, professeure associée (Département de biologie cellulaire), et Monsieur J. VALCÁRCEL, professeur (Centre de Regulació Genòmica, Universitat Pompeu Fabra, Institució Catalana de Recerca i Estudis Avançats, Barcelona, España), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 17 décembre 2014

Thèse - 4748 -

Le Doyen

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

Acknowledgments

I would like to thank:

Angela Krämer for accepting me in her laboratory and giving me the possibility to work and learn during the years I have spent here;

Juan Valcárcel and Françoise Stutz to be part of my thesis jury;

all the members of my Thesis Advisor Committee, Martine Collart, Robbie Loewith, Jean-Claude Martinou and Stéphane Thore, which helped me and supported me during my PhD;

present and past members of Angela Krämer's group for creating a nice working environment: Andrzej, Carlota, Chenna, Claudia, Flore, Ivona, Julie, Maria, Mireille, Monika, Nicolas, Paola and Patty;

the members of Bicel department for their advice and help;

all my friends in Geneva for sharing nice and unforgettable moments and for listening and being present during the difficulties.

Vorrei ringraziare i miei genitori per essermi stati vicino nonostante la distanza e per avermi sempre sostenuto nelle mie scelte. Grazie a mio fratello, per essere stato presente nei momenti più bui.

Grazie a Michele, che ha condiviso pazientemente con me gioie e dolori di questa esperienza.

List of contents

List of contents

Figures and tables

Figures and tables

List of figures

List of tables

Abbreviations

Abbreviations

°C	Celsius
aa	amino acids
Ala	alanine
AMT	aminomethyltrioxsalen
AP	alternative promoter
APA	alternative polyadenylation
AS	alternative splicing
ATP	adenosine triphosphate
BBP	branch point-binding protein
bp	base pairs(s)
BPS	branch point sequence
CC	commitment complex
cDNA	complementary DNA
ChIP	chromatin immunoprecipitation
CIP	calf intestine alkaline phosphatase
cm	centimeter
CMV	cytomegalovirus
CO_2	carbon dioxide
DMEM	Dulbecco'd modified eagle's medium
DNA	desoxyribonucleic acid
dNTP	nucleotide triphosphate
DSCAM	Down syndrome cell adhesion molecule
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
EST	expressed sequence tag
EWS	Ewing's sarcoma
FBS	fetal bovine serum
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
gSF1	genomic SF1
GST	glutathione S-transferase
HeLa	Henrietta Lacks
His_6	hexahistidine
hnRNP	heterogeneous nuclear ribonucleoprotein
HRP	horseradish peroxidase
IP	immunoprecipitation
IPTG	isopropyl-D-1-thiogalactopyranoside
ISE	intronic splicing enhancer

ISS	intronic splicing silencer
K_D	dissociation constant
kDa	kilo dalton
KH	hnRNP K homology
lncRNA	long non-coding RNA
m^7G	N7-methylated guanosine triphosphate
MBP	maltose binding protein
MEN1	multiple endocrine neoplasia
Met	methionine
mRNA	messenger RNA
MS	mass spectrometry
N	any nucleotide
NEAT1	nuclear-enriched abundant transcript 1
NE	nuclear extract
NLS	nuclear localization signal
NMD	non-sense mediated pathway
NMR	nuclear magnetic resonance
nt	nucleotide(s)
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PPT	poly-pyrimidine tract
pre-mRNA	mRNA precursor
Pro	proline
PTC	premature termination codon
QUA2	quaking homology 2
RACE	rapid amplification of cDNA ends
RNA	ribonucleic acid
RNA Pol II	RNA polymerase II
RNAi	RNA interference
RNase	ribonuclease
rpm	revolutions per minute
RRM	RNA Recognition Motif
rRNA	ribosomal RNA
RS domain	arginine/serine-rich domain
RT-PCR	reverse transcription followed by polymerase chain reaction
SAP	SAF-A/B, Acinus and PIAS
SDS	sodium dodecyl sulfate
SF1	Splicing factor 1
SFC	splicing factor compartments
SFSWAP	splicing factor suppressor of white apricot
siRNA	short interfering RNA

SMN	survival of motor neuron protein
snoRNA	small nucleolar RNA
snRNP	small nuclear ribonucleoprotein particle
SR-protein	serine/arginine-rich protein
ss	splice site(s)
SSAP	stage-specific activator protein
SWAP	suppressor of white apricot
TAP	tobacco acid pyrophosphatase
Taq	Thermus aquaticus
TLS	translocated in liposarcoma
tRNA	transfer RNA
TSS	transcription start site
U snRNA	uridine-rich small nuclear RNA
U2AF	U2 snRNP auxiliary Factor
UHM	U2AF homology motif
ULM	U2AF65 ligand motif
UTR	untranslated region
UV	ultraviolet
WW domains	domains with two highly conserved tryptophans
Y	pyrimidine
Y2H	yeast two hybrid
ZFM1	zinc finger protein in the MEN1 locus
Zn	zinc

Resumé en français

Resumé en français

Chez les organismes eucaryotes, le pré-ARNm est soumis à une réaction d'épissage, au cours de laquelle les séquences non codantes (introns) sont soustraites au pré-ARNm et les séquences codantes (exons) sont reliées entre elles. Cette réaction a lieu dans le splicéosome, formé par une série d'étapes consécutives.

Le facteur d'épissage 1 (SF1) a été initialement identifié comme une protéine nécessaire pour l'assemblage précoce du splicéosome. Les portions N-terminales et centrales de la protéine contiennent deux signaux de localisation nucléaire, un domaine d'interaction avec U2AF65, un domaine KH/QUA2 nécessaire pour la liaison de la protéine à l'ARN et un domaine «zinc knuckle ». Dans le complexe précoce du splicéosome (E), SF1 se lie à U2AF65. Ensemble ils reconnaissent deux séquences adjacentes dans le pré-ARNm: la séquence d'embranchement et une série de pyrimidines. Dans le complexe A suivant, U2 PRNpn est recruté par U2AF65 et interagit de manière stable avec la séquence d'embranchement et déplace SF1.

La fonction de SF1 dans la cellule est très complexe. Bien que montrée comme essentielle pour la viabilité des levures, des vers et des mammifères, des expériences ont révélé que la déplétion de SF1 n'affecte pas l'épissage de plusieurs pré-ARNm testés. Des études ultérieures ont montré que SF1 a un rôle cinétique dans l'assemblage du splicéosome et ne prend pas part à tous les événements d'épissage. En outre, des fonctions autres que l'épissage ont été proposées, telles que la rétention nucléaire des pré-ARNm et la régulation de la transcription.

L'objectif de la première partie de la thèse est de mieux comprendre le rôle de SF1 dans la cellule. Comme première étape vers cet objectif, les partenaires d'interaction de SF1 ont été identifiés par co-immunoprécipitation couplée à la spectrométrie de masse et par un crible double hybride. Nos résultats démontrent que SF1 fonctionne principalement dans l'épissage et interagit avec d'autres composants du splicéosome. Étonnamment, des protéines contenant des domaines SURP et associées avec U2 PRNpn ont été co-purifiées avec SF1. Les interactions protéine-protéine ont été validées *in vitro* avec des protéines recombinantes et dans des extraits nucléaires de mammifère. Nous avons identifié une région C-terminale du domaine « zinc knuckle » de SF1, qui est suffisante pour interagir avec les modules SURP. Nos résultats révèlent également que l'interaction de SF1 avec les protéines associées de U2

PRNpn est nécessaire pour la formation efficace du complexe A et confirment les études antérieures qui ont été faites sur le rôle de SF1 sur la cinétique de l'assemblage du splicéosome.

Dans la deuxième partie de la thèse, des isoformes de SF1, avec différentes extrémités N-terminales ont été analysées. De précédentes études ont montré qu'au moins six différentes isoformes de SF1 sont exprimées chez l'homme et la souris et qu'elles partagent les domaines responsables de l'assemblage du splicéosome, mais pas la partie C-terminale de la protéine. Les résultats présentés révèlent que le gène SF1 a des promoteurs alternatifs, qui donnant lieu à plusieurs ARNm SF1 qui n'ont pas les premiers exons. Nos données suggèrent qu'une isoforme SF1 N-terminale est exprimée dans des lignées cellulaires humaines et qu'il lui manque un signal de localisation nucléaire ainsi qu'un domaine d'interaction avec U2AF65.

Abstract

Abstract

In eukaryote organisms, pre-mRNA undergoes a splicing reaction, during which non-coding sequences (introns) are removed from the pre-mRNA and coding sequences (exons) are joined together. This reaction takes place in the spliceosome in a stepwise manner.

Splicing Factor 1 (SF1) was initially identified as a protein required for early spliceosome assembly. The N-terminal and central parts of the protein contain two nuclear localization signals, a U2AF65 interaction domain, a KH/QUA2 domain for RNA binding and a zinc knuckle. In the early spliceosomal complex (E), SF1 binds to U2AF65 and they both recognize adjacent sequences in the pre-mRNA, which are the branch point sequence and the polypyrimidine tract. In the following pre-splicing complex A, U2 snRNP, recruited by U2AF65, stably interacts with the branch point and displaces SF1.

The role of SF1 in the cell is very complex. Although it was shown to be essential for viability in yeast, worms and mammals, initial experiments showed that SF1 depletion did not affect splicing of several pre-mRNAs tested. Later studies demonstrated that SF1 has a kinetic role in spliceosome assembly and does not take part in all splicing events. Moreover, functions other than splicing have been suggested, such as nuclear pre-mRNA retention and transcription regulation.

The aim of the first part of the thesis was to better investigate the role of SF1 in the cell. As a first step towards this goal, SF1 interacting partners were identified by co-immunoprecipitation coupled to mass spectrometry and in yeast two-hybrid screens. Our results demonstrate that SF1 mainly functions in splicing and interacts with other spliceosomal components. Surprisingly, U2 snRNP-associated proteins containing SURP domains co-purify with SF1. Protein-protein interactions were validated *in vitro* with recombinant proteins and in mammalian nuclear extracts. We identified a region C-terminal of the zinc knuckle of SF1, which is sufficient to interact with SURP modules. Our results also show that the interaction of SF1 with U2 snRNP components is required for efficient formation of complex A, which further supports previous studies on the kinetic role of SF1 in spliceosome assembly.

An analysis of SF1 isoforms with different N termini was performed in the second part of the thesis. At least six different SF1 isoforms are expressed in humans and mice, which share the domains required for spliceosome assembly, but not the C-terminal part of the

protein. The results presented here show that the SF1 gene has alternative promoters, which give rise to several SF1 transcripts lacking the first exons. Our data suggest that an N-terminal SF1 isoform lacking the first nuclear localization signal and the U2AF65 interaction domain is expressed in human cell lines.

Introduction

1. Introduction

1.1. RNA processing

In eukaryotes, pre-mRNAs (messenger RNA precursors) are transcribed by RNA Pol II (RNA polymerase II) and undergo several maturation processes, before being translated into proteins. These maturation events are 5' capping, 3' polyadenylation and splicing. Capping of the pre-mRNA consists in the addition of an altered nucleotide, a m^7G cap (N7-methylated guanosine triphosphate), at the 5' end of the RNA molecule via an unusual 5' to 5' triphosphate linkage. The 5' cap protects the mRNA from degradation and promotes translation initiation by recruiting the ribosomes (Hocine et al, 2010). The 3' end of the pre-mRNA is formed by pre-mRNA cleavage and addition of a series of adenosine residues, which form the poly(A) tail. Similarly to the 5' cap structure, the poly(A) tail confers stability to the mRNA as well as influencing translational efficiency (Norbury, 2013).

Splicing is a process where non-coding regions (introns) are removed from the transcript and coding-regions (exons) are ligated together to form the mature mRNA. Splicing was first identified in 1977 by Philip Sharp and Richard Roberts (Berget et al, 1977; Chow et al, 1977). Sharp, Roberts and colleagues observed in the electron microscope that adenoviral mRNA-DNA hybrids gave rise to several loops of non-hybridized DNA. They concluded that genes are discontinuous and that the coding elements are separated by intervening sequences named introns. Their discovery was of fundamental importance and Sharp and Roberts were awarded with the Nobel Prize in Physiology and Medicine in 1993.

For a long time it is has been thought that pre-mRNA maturation processes are independent from each other and occur after transcription. However, in the past years it has been shown that they are coupled with transcription in space and time (Bentley, 2014). Splicing factors and enzymes involved in the 5' capping and 3' polyadenylation were shown to localize to sites of transcription and to interact with RNA Pol II. Transcripts are capped when the first 25-30 nt (nucleotides) have been transcribed by RNA Pol II and cleaved at the 3' end before transcription terminates further downstream (Hocine et al, 2010; Shi, 2012). Splicing can occur both during and after transcription (Braunschweig et al, 2013). The first evidence that splicing can occur co-transcriptionally was demonstrated in *Drosophila melanogaster* embryos, whose pre-mRNAs were shown to be spliced before transcription was completed (Beyer & Osheim, 1988). Later studies based on ChIP (chromatin immunoprecipitation)

experiments confirmed this observation by identifying splicing factors associated with chromatin through nascent pre-mRNA (Kotovic et al, 2003; Listerman et al, 2006).

Splicing is a very important cellular function that ensures the correct formation of mRNAs. If a pre-mRNA is not properly spliced, it might give rise to non-functional or toxic proteins. It is therefore not surprising that many human diseases are associated with splicing defects (Faustino & Cooper, 2003).

1.2. The splicing reaction

The splicing reaction requires the presence of short and conserved sequences in the pre-mRNA (Breathnach et al, 1978; Mount, 1982). These are the 5' (donor site) and 3' (acceptor site) ss (splice sites), the BPS (branch point sequence) and the PPT (poly-pyrimidine tract) (figure 1). These *cis*-acting elements are essential for the splicing reaction and are characterized by consensus sequences (figure 1). The 5' and 3' ss surround the exon-intron boundaries, whereas the BPS, usually located 20-40 nt upstream of the 3' ss, contains the conserved adenosine essential for the splicing reaction. The BPS is highly conserved in the budding yeast *Saccharomyces cerevisiae* (UACUAAC), but degenerate in metazoans (YUNAY, where Y is a pyrimidine and N any nucleotide) (figure 1). In metazoans, the BPS is followed by a PPT, a stretch of 10 or more pyrimidines. The PPT is also present in many yeast introns; it is composed prevalently of uracil and is less defined and shorter than the metazoan PPT (figure 1). According to the similarity of these signals with the consensus sequences, they specify "strong" or "weak" splice sites.

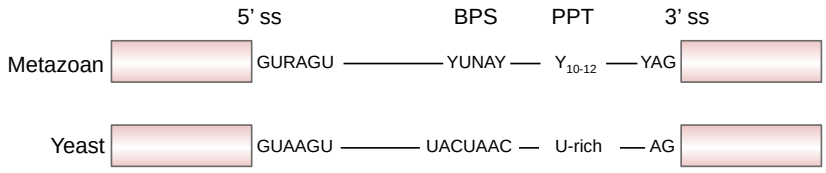


Figure 1. Conservation of splicing sequences.

Conserved sequence elements of metazoan and yeast pre-mRNAs are shown. The pink boxes are the exons and the black line the intron. The consensus sequences in metazoans and yeast at the 5' ss, BPS, and 3' ss are indicated, where N is any nucleotide, R is a purine, and Y is a pyrimidine.

From a chemical point of view, the splicing reaction occurs through two consecutive trans-esterification reactions (Padgett et al, 1984; Ruskin et al, 1984) (figure 2). In the first reaction, the 2' hydroxyl group of the conserved adenosine of the BPS attacks the phosphodiester bond at the 5' ss. This results in a free 5' exon and the exon-intron lariat intermediate product. The lariat is formed through a 2' – 5' phosphodiester bond. In the second reaction, the 3' hydroxyl of the 5' exon attacks the phosphodiester bond at the 3' ss; thus, the intron is released in the form of a lariat and the exons are joined.

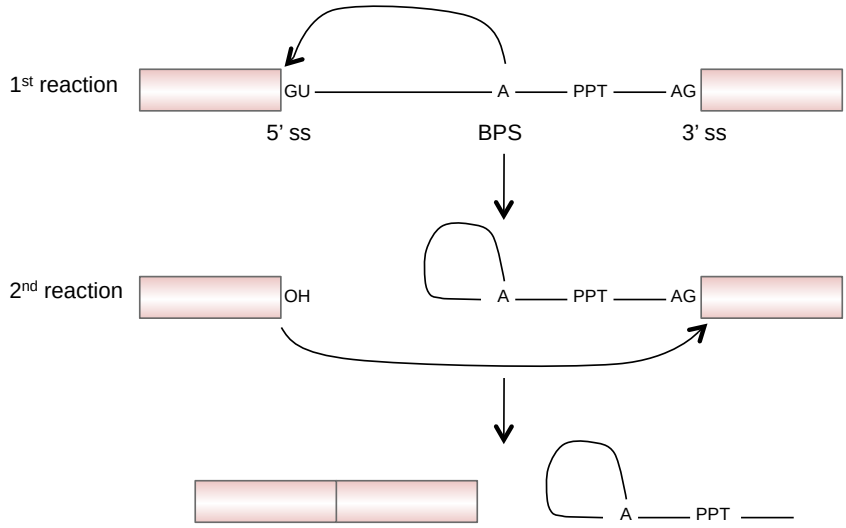


Figure 2. The chemistry of splicing.

The figure summarizes the chemistry of the splicing reaction. Exons are shown as pink boxes, the intron as a black line. The position of the 5' ss, 3' ss, BPS and PPT is shown. Arrows indicate the nucleophilic attack of the first and second trans-esterification reactions.

1.3. The spliceosome

The splicing reaction is catalyzed by a macromolecular ribonucleoprotein machinery, the spliceosome, which consists of five snRNPs (the U1, U2, U4, U5 and U6 small nuclear ribonucleoprotein particles) and hundreds of non-snRNP proteins (Wahl et al, 2009). Each snRNP contains a U snRNA (Uridine-rich small nuclear RNA) associated with seven Sm proteins (B/B', D1, D2, D3, E, F, G) and several snRNP specific proteins (figure 3).

The U1 snRNP contains three specific proteins (U1-70K, U1-A and U1-C) and the U1 snRNA, which recognizes the conserved sequence at the 5' ss of the pre-mRNA (Bringmann & Luhrmann, 1986; Zhuang & Weiner, 1986). The U2 snRNA binds to the BPS and this

snRNA-pre-mRNA interaction is stabilized by the U2 snRNP-specific proteins U2A', U2B'', SF3a and SF3b, which are described in detail in section 1.5 (Behrens et al, 1993b; Bringmann & Luhrmann, 1986). The U4 and U6 snRNPs associate with each other through base pairing to form the U4/U6 di-snRNP. U5 snRNP binds to the di-snRNP through protein-protein interactions to form the U4/U6.U5 tri-snRNP (figure 3) (Wahl et al, 2009).

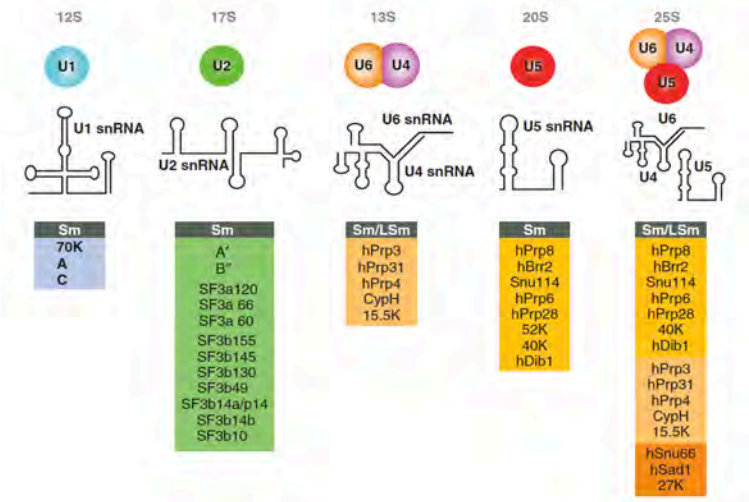


Figure 3. Protein composition and snRNA secondary structure of U snRNPs.

The figure shows a schematic representation of snRNA secondary structures. The boxes indicate the proteins associated with each snRNP. Sm and LSm protein are indicated at the top of the boxes. The figure is taken from Will & Lührmann (2011).

1.4. Spliceosome assembly

The spliceosome is highly dynamic and assembles in a stepwise manner at each cycle of splicing through the formation of intermediate complexes, termed H, E, A, B and C (figure 4) (Matera & Wang, 2014; Will & Luhrmann, 2011). Each complex is characterized by a specific protein composition, in addition to specific protein-protein, protein-RNA and RNA-RNA interactions.

The "holospliceosome" represents an alternative mode in which the spliceosome can assemble. This model hypothesizes that all five snRNPs co-exist in a unique particle, called penta-snRNP, before binding to the pre-mRNA. The penta-snRNP has been isolated from *S. cerevisiae* and human cell extracts (Malca et al, 2003; Stevens et al, 2002). The possibility that intermediate spliceosomal complexes derive from progressive dissociation of the holospliceosome due to harsh conditions used for purification has been excluded and the

possibility that the models might co-exist *in vivo* has been taken in consideration (Behzadnia et al, 2006). However, several research studies based on the application of ChIP experiments argue against the holospliceosome model (Gornemann et al, 2005; Tardiff & Rosbash, 2006). These reports have shown that snRNPs are recruited in a stepwise manner *in vivo* and accumulate at different positions along the pre-mRNA.

During spliceosome assembly, some U snRNAs undergo structural and conformational rearrangements, which are catalyzed by RNA helicases. RNA helicases are RNA-binding proteins that modify RNA secondary structure upon NTP hydrolysis. The helicases involved in the splicing reaction belong to the DEAD-box, DEAH-box or Ski2-like families (Cordin & Beggs, 2013). DEAH-box proteins hydrolyze any NTP, whereas DEAD-box and Ski2-like proteins are characterized by an ATP-specific activity.

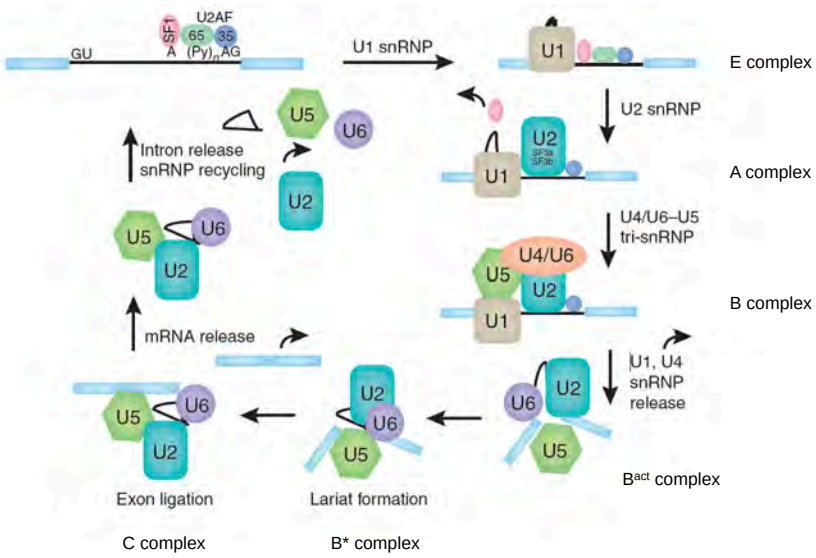


Figure 4. Spliceosome assembly.

The spliceosome assembles in a stepwise manner as described in section 1.4 and several spliceosome intermediates (E, A, B, B^{act}, B* and C) are shown. The figure is taken from Rymond (2007).

1.4.1. Complex H

Complex H is formed upon non-specific binding of hnRNPs (heterogeneous nuclear ribonucleoproteins) to the pre-mRNA. It is believed not to be specific to splicing, since its formation is ATP- and temperature-independent. Moreover complex H forms on RNAs lacking ss sequences.

1.4.2. Complex E

Spliceosome assembly starts with the recognition of the 5' ss by U1 snRNP in an ATP-independent manner (figure 4). This interaction is stabilized by U1-70K and U1-C proteins (Will & Luhrmann, 2011). The 3' portion of the intron is recognized by SF1 and U2AF (U2 snRNP Auxiliary Factor) (figure 4) (Berglund et al, 1997; Ruskin et al, 1988). The U2AF heterodimer was identified as a protein complex required for the U2 snRNP-BPS interaction and consists of a large and a small subunit, termed U2AF65 and U2AF35 according to their molecular weight (Ruskin et al, 1988). U2AF35 and U2AF65 are tightly associated with each other and interacting regions in both proteins were identified in the N-terminal domain of U2AF65 and the UHM (U2AF Homology Motif) of U2AF35 (Kielkopf et al, 2001; Zhang et al, 1992). U2AF65 binds to the PPT via two RRMs (RNA Recognition Motifs, RRM1 and RRM2), whereas U2AF35 recognizes the 3'ss. Moreover, the N-terminal RS (arginine/serine-rich) domain of U2AF65 makes contact with the BPS and stabilizes the base-pairing interaction between U2 snRNA and the pre-mRNA (Ito et al, 1999; Valcarcel et al, 1996; Zamore et al, 1992). Besides interacting with the 3' ss, U2AF35 binds via its C-terminal RS domain to two SR proteins, SC35 (SRSF2) and SF2/ASF (SRSF1) (Wu & Maniatis, 1993; Wu et al, 1999; Zuo & Maniatis, 1996). SC35 and SF2/ASF also interact with U1-70K, a protein associated with the U1 snRNP (Wu & Maniatis, 1993). Therefore, interactions of U2AF35 with SR proteins are important for juxtaposing 3' and 5' ss.

Prior to pre-mRNA binding, the UHM of U2AF65 binds a short N-terminal region of SF1 (Selenko et al, 2003). The U2AF65-SF1 interaction was shown to be cooperative, because it increases the binding affinity of SF1 and U2AF65 to the BPS and PPT, respectively (Berglund et al, 1998a).

In budding yeast, the mammalian E complex corresponds to the commitment complexes CC1 and CC2. In the CC1 complex U1 snRNP interacts with the 5' ss. CC1 is converted into

CC2 after recruitment of two non-snRNP proteins, BBP (branch point-binding protein) and Mud2 (Brow, 2002), the yeast homologues of SF1 and U2AF65, respectively (Abovich et al, 1994; Abovich & Rosbash, 1997; Rutz & Seraphin, 1999). Similar to metazoans, BBP interacts with the BPS and Mud2 (Abovich & Rosbash, 1997; Berglund et al, 1997). In budding yeast, Mud2 cross-links to pre-mRNAs containing an intact BPS sequence, most likely by interacting with BBP, and was shown to be non-essential for the splicing reaction (Abovich et al, 1994; Tang et al, 1996). The U2AF35 homolog is present in the fission yeast *Schizosaccharomyces pombe*, but not in budding yeast and binds the poorly conserved 3' ss (Wentz-Hunter & Potashkin, 1996). In *S. cerevisiae* the splice sites are brought into close proximity by interaction between BBP and Prp40, a U1 snRNP-associated protein, in the CC2 complex (Abovich & Rosbash, 1997). Similar to metazoans, all interactions forming CC1 and CC2 are ATP-independent.

In the mammalian E complex, the functional U2 snRNP is loosely associated with the spliceosome in an ATP-independent manner and this association does not require the BPS (Das et al, 2000; Hong et al, 1997; Makarov et al, 2012). In yeast, U2 snRNP has not been detected in either CC1 or CC2.

1.4.3. Complex A

Complex A is formed upon stable binding of U2 snRNP to the pre-mRNA (figure 4). An internal and highly conserved region of U2 snRNA base pairs with the BPS surrounding the conserved adenosine. This RNA-RNA interaction is very important because the conserved adenosine is bulged out from the duplex and marked for the nucleophilic attack at the 5' ss. Recognition of the BPS by U2 snRNP is ATP-dependent and requires the DEAD-box RNA helicases Prp5 and UAP56. Prp5 interacts with several U2 snRNP proteins and, in *S. cerevisiae*, it induces conformational changes in the U2 snRNA to facilitate base pairing with the BPS (O'Day et al, 1996; Will et al, 2002). UAP56 (the yeast homolog of Sub2) interacts with U2AF65 and promotes U2 snRNP recruitment to the pre-mRNA (Fleckner et al, 1997).

The helix formed between U2 snRNA and the pre-mRNA is short and destabilized by the degeneracy of the BPS and by the bulged adenosine. Therefore, U2 snRNP recruitment cannot rely exclusively on the base pairing interaction. The UHM of U2AF65 was shown to directly interact with a U2 snRNP-associated protein, SF3b155, and thus to recruit the

particle to the pre-mRNA (Gozani et al, 1998). The same domain of U2AF65 binds to SF1 in complex E (Berglund et al, 1998a; Rain et al, 1998; Selenko et al, 2003). Therefore, U2 snRNP and SF1 compete for the same binding site not only on the pre-mRNA but also on U2AF65. The presence of ATP and RNA helicases in the A complex destabilizes the SF1-U2AF65 interaction, which is replaced by SF3b155-U2AF65 with subsequent displacement of SF1 from the BPS by the U2 snRNA/pre-mRNA base pairing interaction (Thickman et al, 2006). SF1, therefore, has the function to avoid a premature association of U2 snRNP with the pre-mRNA by forming a thermodynamic barrier, which is abolished by ATP hydrolysis (Thickman et al, 2006).

1.4.4. Complexes B, B^{act} and B*

In complex B, the pre-assembled tri-snRNP U4/U6.U5 is recruited (figure 4). The U4/U6 duplex masks a region in U6 snRNA (the ACAGA motif), important for catalysis, thus preventing premature pre-mRNA cleavage. Although complex B contains all snRNPs, it is not yet catalytically active (figure 5). It undergoes several conformational and structural rearrangements to first form the activated B complex (complex B^{act}) and then, after recruitment of Prp2, complex B* (Ohrt et al, 2012). In complex B^{act}, the U4/U6 snRNA base pairing is disrupted and the U1 and U4 snRNPs are released from the pre-mRNA. The ACAGA motif of U6 snRNA is no longer hidden and thus can base pair with the 5’ ss, replacing the U1 snRNA. Sequences downstream of the ACAGA motif of U6 snRNA base pair with U2 snRNA, thus bringing the 5’ exon and the BPS in close contact. The two exons are brought in proximity by the U5 snRNP, which binds to the extremities of the 5’ and 3’ exons (figure 5) (Newman & Norman, 1992). Complex B^{act} undergoes some structural remodeling and is converted into complex B*, which can now catalyze the first splicing reaction that generates the free 5’ exon and the intron-exon lariat.

Transition from B to B^{act} requires several proteins, such as Prp28, Brr2 and Snu114. The most important rearrangements require Prp28 and Brr2. Prp28 is a DEAD-box helicase shown to destabilize the base pairing of U1 snRNA with the 5’ ss, whereas Brr2 belongs to the Ski2-like family and functions in unwinding the U4 and U6 snRNAs duplex (Kim & Rossi, 1999; Raghunathan & Guthrie, 1998; Staley & Guthrie, 1999). Brr2 and Snu114 are components of the U5 snRNP and play an important role during the catalytic activation of the spliceosome, together with the DEAH-box RNA helicase Prp2, (Cordin & Beggs, 2013;

Hacker et al, 2008). Prp2 and the cofactor Spp2 catalyze the conversion from complex B^{act} into B* upon ATP hydrolysis (Ohrt et al, 2012).

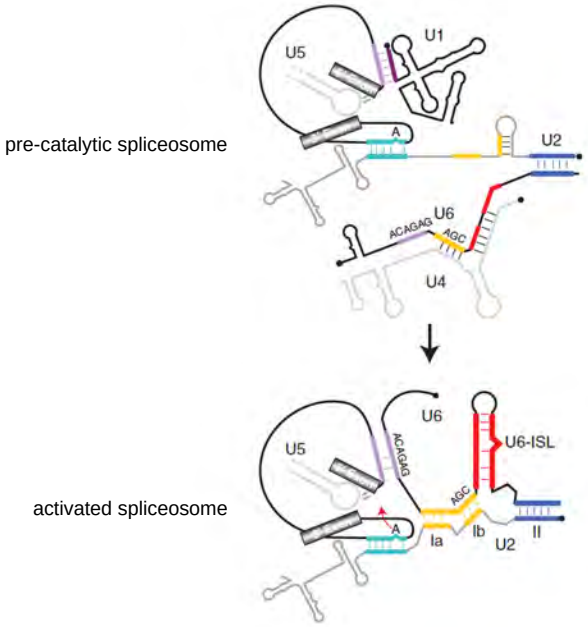


Figure 5. Rearrangements of RNA-RNA interactions in the catalytically activated spliceosome.

Exons and introns are shown as grey boxes and black lines, respectively. U snRNAs are indicated as black or grey lines and the regions involved in base pairing are highlighted in color. The capped 5' ends of the U snRNAs are represented by a dot. In the pre-catalytic spliceosome the 5' ss and the BPS base pair to U1 and U2 snRNAs, respectively. U4/U6.U5 tri-snRNP recruitment leads to several RNA rearrangements, such as unwinding of the U1 snRNA-5' ss and U4/U6 snRNAs duplexes, base pairing of U6 snRNA with the 5' ss and with U2 snRNA. In the activated spliceosome, the splice sites are brought into proximity by U2-U6 base pairing and by the interaction of U5 snRNP with the exons. The red arrow indicates the nucleophilic attack of the conserved adenosine during the first catalytic step. The figure is taken from Will & Lührmann (2011).

1.4.5. Complex C

After the first trans-esterification reaction, complex C is formed and the second splicing reaction is catalyzed (figure 4). Transition between the first and the second catalytic steps requires the DEAH-box RNA helicase Prp16. Additional proteins join the spliceosome before the second reaction, such as Prp8, Prp18 and Slu7.

After the second catalytic reaction, the DEAH-box RNA helicases Prp22 and Prp43 catalyze spliceosome disassembly and release the spliced mRNA and excised intron. Once released,

the mRNA is exported to the cytoplasm, whereas the intron-lariat is linearized and degraded or further processed to generate non-coding RNAs, such as snoRNAs and microRNAs (Ruby et al, 2007; St Laurent et al, 2012; Tycowski et al, 1993).
After disassembly, spliceosome components are recycled for further rounds of splicing.

1.5. U2 snRNP

The U2 snRNP recognizes the BPS in the A complex by a base-pairing interaction. The U2 snRNP is composed of the U2 snRNA associated with several proteins. The U2 snRNA is 189 bp (base pairs) long and contains two regions complementary to the 3' end of U6 snRNA in its 5' region, followed by a short and highly conserved sequence element (GUAGUA) that base pairs with the sequence flanking the conserved adenosine of the BPS (figure 6). Besides the Sm proteins, U2 snRNP contains several specific proteins, which are U2A' (33 kDa), U2B" (28.5 kDa) and the heteromeric complexes SF3a and SF3b (figure 6) (Behrens et al, 1993b; Bringmann & Luhrmann, 1986). U2A' and U2B'' bind stem-loop IV at the 3' end of U2 snRNA, whereas the SF3a and SF3b complexes interact with the 3' half and the 5' half of U2 snRNA, respectively (figure 6) (Krämer et al, 1999; Price et al, 1998).
Two forms of the U2 snRNP particle, of 12S and 17S, have been isolated (Behrens et al, 1993b). The 12S U2 snRNP particle consists of the U2 snRNA associated with Sm, A' and B" proteins, whereas the 17S particle also contains SF3a and SF3b. In addition, a 15S particle was identified, which lacks SF3a (Brosi et al, 1993a). The functional U2 snRNP contains all U2 snRNP-associated proteins; therefore only the 17S U2 snRNP assembles with the spliceosome (Behrens et al, 1993a).
Although the U2 snRNA base pairs with the BPS only in the A complex, the functional U2 snRNP was shown to be loosely associated in an ATP-independent manner with the E complex and this association does not require the BPS (Das et al, 2000; Hong et al, 1997; Makarov et al, 2012).

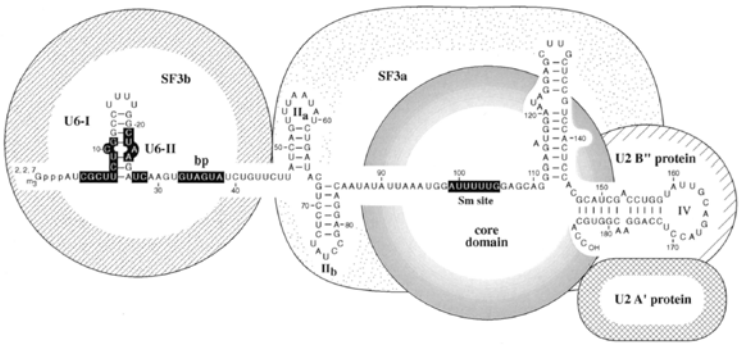


Figure 6. U2 snRNP structure.

The primary and secondary structure of U2 snRNA is shown. Black boxes represent the Sm site and regions complementary to U6 snRNA and the BPS. Binding of Sm proteins and U2 snRNP-specific proteins is indicated. The figure is taken from Nagai et al (2001).

1.5.1. SF3a

The heterotrimeric SF3a complex consists of three subunits, SF3a60, SF3a66 and SF3a120, which are present in a stoichiometry of 1:1:1 (Brosi et al, 1993b). SF3a60 and SF3a66 interact with SF3a120, but not with one another and, therefore, SF3a120 functions as a scaffold in the formation of the SF3a complex (Huang et al, 2011; Nesic & Krämer, 2001). The interactions among SF3a subunits are highly conserved across species. In yeast, the SF3a complex consists of Prp9, Prp11 and Prp21. Similar to mammals, Prp9 and Prp11 do not interact with each other, but bind Prp21 to form the heterotrimeric complex (Lin & Xu, 2012; Rain et al, 1996).

The SF3a subunits were shown to be essential for 17S U2 snRNP assembly, for the splicing reaction and for viability in mammalian cells and *S. cerevisiae* (Brosi et al, 1993a; Brosi et al, 1993b; Tanackovic & Krämer, 2005).

All SF3a subunits were shown to cross-link to the pre-mRNA in the region immediately upstream the BPS (figure 7) (Gozani et al, 1996). Thus, their main function is to recruit and anchor the U2 snRNA to the pre-mRNA by stabilizing the pre-mRNA-snRNA interaction weakened by the degeneracy of the BPS (Gozani et al, 1996).

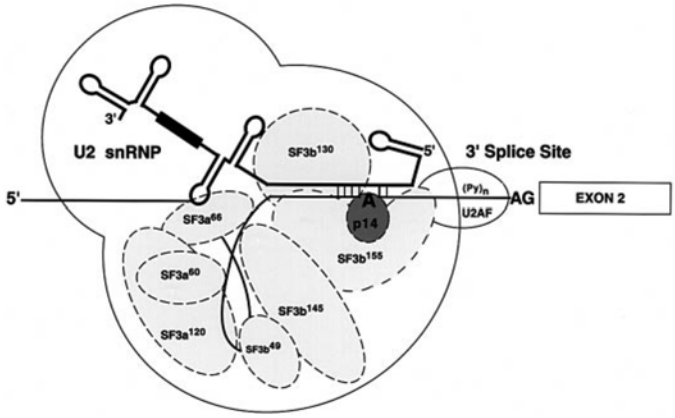


Figure 7. Model of the interaction between U2 snRNP and the BPS.

The exon and intron are shown as a box and a black line, respectively. Positions of PPT, BPS and 3' ss are indicated. The black line indicates the U2 snRNA, whereas the ellipses represent the U2-specific proteins. The figure is taken from Will et al (2001).

1.5.1.1. SF3a60

SF3a60 is part of the SF3a complex and contains several evolutionary conserved domains (figure 8) (Krämer et al, 2005).

The protein is 501 amino acids long and contains several sequences in the N-terminal part that are conserved in different species, except for *S. cerevisiae* (Nesic & Krämer, 2001). So far, no structural domains have been identified in this part of the protein. However, the first 107 amino acids interact with SF3a120 (figure 8) (Huang et al, 2011; Kuwasako et al, 2006; Nesic & Krämer, 2001). Biochemical analyses have shown that amino acids 35-107 are necessary for the interaction, whereas the first 34 amino acids stabilize the association between SF3a60 and SF3a120 (figure 8).

The central part of the protein contains a so-called SAP (SAF-A/B, Acinus and PIAS) motif (figure 8). This region is dispensable for spliceosome assembly and so far, no function in splicing has been assigned to it (Nesic & Krämer, 2001). A zinc finger domain (aa 408-431) is present in the C-terminal part of the protein and shows the highest homology between metazoan, plant and yeast SF3a60 (figure 8) (Nesic & Krämer, 2001). A region encompassing the zinc finger was shown to bind to the 15S U2 snRNP, although it is not clear whether the SF3a60 interaction with the snRNA is direct or mediated by Sm or SF3b proteins (figure 8) (Nesic & Krämer, 2001).

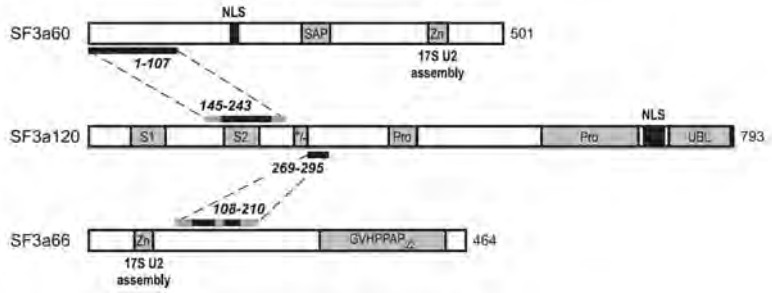


Figure 8. Schematic representation of SF3a domains and interactions.

Structural and functional domains of SF3a60, SF3a66 and SF3a120 are shown as grey and black boxes. Thick lines show the sequences involved in protein-protein interactions. Black lines indicate sequences essential for the binding, whereas the grey lines mark sequences that contribute to the interaction. NLS, nuclear localization signal; SAP, SAP domain; Zn, zinc finger; S1 and S2, SURP1 and SURP2 domain; +/-, charged residues; Pro, proline-rich sequence; UBL, ubiquitin-like domain; $GVHPPAP_{22}$, heptad repeats. The figure is taken from Huang et al (2011).

1.5.1.2. SF3a66

Together with SF3a60 and SF3a120, SF3a66 forms the heterotrimeric SF3a complex. The protein is 464 amino acids long and contains several evolutionarily conserved domains (figure 8) (Krämer et al, 2005).

The N-terminal part of the protein is highly conserved from yeast to humans and contains a zinc finger domain (aa 56-78) which, similar to SF3a60, binds the 15S U2 snRNP (figure 8) (Nesic & Krämer, 2001). However, it is not completely understood whether the zinc finger interacts directly with the U2 snRNA or through U2 proteins. The central region of SF3a66 contains an amino acid sequence (aa 108-210) that binds to SF3a120 (figure 8) (Huang et al, 2011; Nesic & Krämer, 2001). However, only amino acids 125-154 and 165-185 were shown to be essential for the binding to SF3a120; the others contribute to the interaction or optimize binding. Therefore the N-terminal and central regions of SF3a66 contain all the domains required for its function in the assembly of the mature U2 snRNP, such as the SF3a120 and 15S U2 snRNP binding sites.

The C-terminal part of SF3a66 contains 22 heptad repeats (GVHPPAP), which are dispensable for the spliceosome function of the protein (figure 8) (Nesic & Krämer, 2001). The function of this region is still unknown. The observation that it is divergent among

species and absent from the yeast orthologue suggests that it might have a specific role in different organisms (Nesic & Krämer, 2001).

1.5.1.3. SF3a120

The SF3a120 subunit has several evolutionary conserved domains (figure 8) (Krämer et al, 2005). It is 793 amino acids long and contains two SURP modules in its N-terminal part (aa 52-94 and 166-208), which are highly conserved in *S. cerevisiae* and *Caenorhabditis elegans* (figure 8) (Krämer et al, 1995). SURP domains are also called SWAP domains because they were initially identified in the *D. melanogaster* alternative splicing factor SFSWAP/SFRS8 (Suppressor-of-White-Apricot) (Denhez & Lafyatis, 1994; Spikes et al, 1994). These domains are not abundant in the human proteome and so far, only six human proteins have been shown to contain SURP domains (Kuwasako et al, 2006). Surprisingly, all SURP domain-containing proteins associated with the splicing reaction. However, to date, the exact function of SURP domains is still unknown. Biochemical analysis revealed that a region encompassing the second SURP domain of SF3a120 (aa 145-243) binds to SF3a60, suggesting that SURP modules can mediate protein-protein interactions (figure 8) (Huang et al, 2011; Nesic & Krämer, 2001). The NMR structure of SF3a60 and SF3a120 peptides (aa 71-107 and aa 134-217, respectively) has been solved and it describes that a shorter region of SF3a120, containing only amino acids 162-195 of the SURP2 module make contacts with SF3a60 (Kuwasako et al, 2006). However, *in vitro* experiments have shown that the single SURP2 sequence of SF3a120 is not sufficient for binding to SF3a60 and that further flanking sequences are necessary for the interaction. The biochemical experiments correlate with structural analyses performed with the yeast homologs Prp9 (SF3a60), Prp11 (SF3a66) and Prp21 (SF3a120), which show that contacts between Prp9 and Prp21 are mediated not only by SURP2 but also by a long helix located at its C-terminal part (Lin & Xu, 2012). So far, no function has been described for SURP1 of SF3a120 or SURP domains of other proteins.

The central part of SF3a120 contains a short stretch of charged amino acids (aa 254-268), which are conserved in *S. cerevisiae* and whose function is still unknown (figure 8) (Lin & Xu, 2012). The interaction domain for SF3a66 was mapped to the C-terminus of the charged residues and corresponds to amino acids 269-295 (figure 8) (Huang et al, 2011). This region is well conserved between metazoans, plants and yeast. The C-terminal part of the protein is not conserved in *S. cerevisiae* and contains proline-rich sequences (aa 369-403 and 557-675)

and an ubiquitin-like domain (aa 707-790) (figure 8). So far, no function has been assigned to the C-terminal region of SF3a120.

1.5.2. The SF3b complex

The heteromeric SF3b complex is part of the U2 snRNP and, together with SF3a, it is essential for the assembly of the mature 17S U2 particle (Brosi et al, 1993a; Krämer et al, 1999).

The SF3b complex consists of the followings subunits, which are evolutionary conserved across several species: SF3b10, SF3b14a/p14, SF3b14b, SF3b49, SF3b130, SF3b145 and SF3b155 (Behrens et al, 1993b; Das et al, 1999; Will et al, 2001; Will et al, 2002).

A number of protein-protein interactions have been detected between the SF3b subunits. Biochemical and structural analyses have shown that SF3b14a/p14 stably interacts with a central region of SF3b155, containing amino acids 381-424 (Spadaccini et al, 2006; Will et al, 2001). In addition, SF3b14a/p14 has been shown to bind to SF3b130, whereas the N-terminal portion of SF3b49, containing two RRMs, interacts with SF3b145 (Champion-Arnaud & Reed, 1994; Das et al, 1999; Will et al, 2001).

All SF3b subunits, except SF3b130, cross-link to a 20-nucleotide region situated at the 5' end of the BPS (figure 7) (Staknis & Reed, 1994). SF3b14a/p14 is the only SF3b subunit shown to directly contact the BPS adenosine, whereas SF3b155 cross-links on both sides of the BPS (figure 7) (Gozani et al, 1996; Gozani et al, 1998; Will et al, 2001). Thus, together with SF3a proteins, the SF3b subunits function to stabilize the U2 snRNA interaction with the pre-mRNA (Gozani et al, 1996; Gozani et al, 1998). SF3b subunits also function in recruiting the U2 snRNP to the pre-mRNA. A N-terminal sequence of SF3b155 binds the UHM of U2AF65 (Gozani et al, 1998). Although it was initially thought that the SF3b155 binding sites for U2AF65 and SF3b14a/p14 partially overlap, further studies demonstrated that the interaction of SF3b155 with one protein does not affect the binding to the other (Cass & Berglund, 2006). The N-terminal part of SF3b155 is conserved in *S. pombe*, but not in *S. cerevisiae*. Therefore, the U2 snRNP is recruited to the pre-mRNA through direct interaction between U2AF65 and SF3b155 in humans and *S. pombe*, but not in *S. cerevisiae* (Gozani et al, 1998). In addition, an interaction between SF3b155 and U2AF35 has been detected but the interaction domains have not been mapped (Gozani et al, 1998).

1.6. Proteome diversity

Sequencing of the human genome revealed an unexpectedly low number of protein-coding genes compared to the amount of expressed proteins. This complexity is achieved because a single gene can produce several mRNAs, and therefore several proteins, by the mechanisms of alternative splicing (AS), alternative poyadenylation (APA) and alternative promoter use (AP) (figure 9) (Keren et al, 2010).

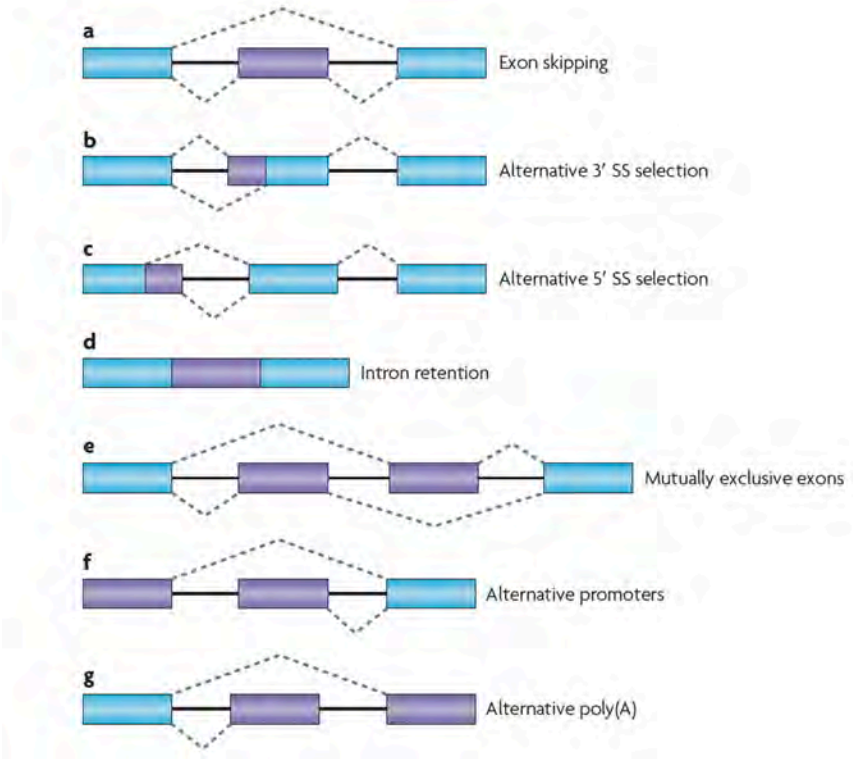


Figure 9. Different modes to increase proteome diversity.

Alternative splicing (AS) events (a, b, c, d and e), alternative promoters (f) and polyadenylation signals (g) increase proteome diversity. The schematic pre-mRNA has constitutive exons in blue boxes, alternatively spliced regions in violet boxes and introns as solid lines. Dashed lines indicate alternative decisions. The figure is taken from Keren et al (2010).

1.6.1. Alternative splicing

AS is a process in which a pre-mRNA can be spliced in several ways and produce transcript variants containing different exons and/or introns. A good example of how AS can increase

proteome complexity is represented by the DSCAM gene (Down syndrome cell adhesion molecule) of *D. melanogaster*, which produces more than 38'000 different mRNA variants. The main modes in which AS can occur are summarized in figure 9 and they are exon skipping, alternative 3' and 5' ss, intron retention and mutually exclusive alternative exons (Keren et al, 2010).

It was shown that the frequency of AS events correlates with organism complexity. In humans, high-throughput RNA sequencing technology has shown that nearly 95% of the pre-mRNAs are alternatively spliced, whereas unicellular eukaryotes, like trypanosomes, have mostly intron-less genes (Braunschweig et al, 2013; Kornblihtt et al, 2013). In *S. cerevisiae* introns are short and present in only 3% of the genes. Since most of yeast genes contain only one intron, AS is very rare in budding yeast.

As described before, splicing is catalyzed by the spliceosome, whose components recognize and bind *cis*-acting elements in the pre-mRNA (5' and 3' ss, BPS and PPT). However, these sequences are not always sufficient to initiate the splicing reaction and additional RNA elements, named ESEs and ISEs (exonic and intronic splicing enhancers), ESSs and ISSs (exonic and intronic splicing silencers), have been shown to promote and repress the splicing reaction, respectively (Braunschweig et al, 2013). Splicing enhancers and silencers are short (about 6 nt) and do not present any consensus RNA sequence. They function by binding *trans*-acting splicing regulatory factors, which activate or suppress constitutive and alternative splicing. The major *trans*-acting factors are represented by SR family and hnRNP proteins.

The SR proteins were initially identified as positive splicing regulators that bind to splicing enhancers. They function in recruiting spliceosome components to the pre-mRNA, stabilizing the interaction of snRNPs and splicing factors with consensus RNA sequences and/or antagonizing the effects of splicing silencer elements (Cartegni et al, 2002; Long & Caceres, 2009). The SR proteins promote the splicing reaction also by juxtaposing the 3' and 5' ss of the pre-mRNA through protein-protein interactions. An example is the binding of SR proteins SF2/ASF and SC35 to U1 70K at the 5' ss and with U2AF35 at the 3' ss (Long & Caceres, 2009; Wu & Maniatis, 1993; Zuo & Maniatis, 1996). Later studies have shown that SR proteins cannot only activate the splicing reaction, but also repress it and their role depends on the pre-mRNA sequence to which they bind. For example, SF2/ASF was shown to bind also to a splicing silencer, leading to an inhibition of adenovirus IIIa splicing

(Kanopka et al, 1996). On the other hand, some SR proteins were shown to function exclusively as splicing repressor, such as p54, SRp38 and SRp86 (Long & Caceres, 2009).

The function of SR proteins is antagonized by hnRNP proteins, which inhibit the splicing reaction by binding to splicing silencers and preventing the interaction of spliceosome components to an adjacent site (Martinez-Contreras et al, 2007). It has been shown that hnRNP proteins can interfere with the binding of splicing factors, including SR proteins. The antagonistic role of hnRNP A1 and SF2/ASF has been extensively described and the ratio of both classes of proteins in the nucleus was shown to be very important for splicing regulation (Long & Caceres, 2009; Martinez-Contreras et al, 2007). HnRNP proteins can also modulate the splicing reaction by formation of a loop in the pre-mRNA, by interaction of hnRNP proteins bound to different sites. This mechanism was shown to either promote exon skipping, if hnRNP proteins are binding flanking introns of the alternative exon, or facilitate exon definition, if they bind to the same intron (Martinez-Contreras et al, 2007).

Together with SR and hnRNP proteins, tissue-specific factors have been shown to control splicing, such as Nova and FOX (Kornblihtt et al, 2013). Regulation of AS is therefore tightly controlled and alterations can be associated with pathologies, such as retinitis pigmentosa, spinal muscular atrophy, myotonic dystrophy and cancer (Douglas & Wood, 2011). Indeed, almost one third of the genes involved in genetic diseases harbor mutations in *cis*- and *trans*-acting regulatory elements, which affect the formation of mRNA and lead to a dysfunctional protein (Singh & Cooper, 2012).

1.6.2. Alternative polyadenylation

Proteome complexity is also achieved by APA, which consist in the use of alternative signals of cleavage and addition of a poly(A) tail at different sites of the pre-mRNA. It was shown that around 70% of human and yeast pre-mRNAs are alternatively polyadenylated, suggesting that, unlike splicing, APA is evolutionarily conserved (Shi, 2012). APA signals can be present in the same exon or in different ones. Therefore, the usage of APA signals can generate several transcript isoforms with different coding sequences leading to distinct proteins, and/or different 3’ UTRs (untranslated regions), which influences mRNA stability and translation efficiency (Elkon et al, 2013). In addition, 3’ UTRs contain regulatory elements, such as sequences complementary to microRNAs and protein binding sites, which influence mRNA translation and stability (Kuersten & Goodwin, 2003).

1.6.3. Use of alternative promoters

Usage of APs is very common in the human genome and almost half of human and mouse genes contain APs (Baek et al, 2007). The main consequences of their usage are modifications in the coding region and/or 5' UTR of the transcripts (Davuluri et al, 2008). Therefore, APs further increase protein diversity by generating isoforms with either distinct N-terminal sequences, which can affect protein function and localization, or, more rarely, generate proteins with completely different ORFs (open reading frames). However, it has been demonstrated that most genes containing APs produce transcripts with the same coding sequences, but with different tissue/development-specific expression levels and/or 5' UTRs, which influence mRNA stability and translation efficiency (Davuluri et al, 2008). APs play an important role in regulating developmental and/or tissue-specific gene expression; several genes are characterized by a ubiquitous promoter and an AP, the latter of which is activated only in certain tissues and/or developmental stages by specific transcription factors (Landry et al, 2003).

Several reports have shown that APs can also indirectly regulate the splicing process. As discussed above, splicing and transcription are tightly coupled and it was shown that promoter structure can influence splicing decisions of the same gene, by recruiting for example splicing factors through transcription factors or by modulating RNA Pol II speed (Cramer et al, 1997). A low RNA Pol II elongation rate allows spliceosomal components to recognize and bind to weak consensus sequences of the exon and therefore to promote exon inclusion, whereas a fast elongation rate has an opposite effect (Kornblihtt, 2005). Similar to splicing, aberrant activation or silencing of APs has been associated with several pathologies, such as cancer, Parkinson's disease or schizophrenia (Davuluri et al, 2008).

1.7. SF1

SF1 was identified as a heat-resistant protein of 75 kDa and was shown to be required for the early steps of spliceosome assembly, by binding the BPS in the E complex (Arning et al, 1996; Berglund et al, 1997; Krämer, 1992). SF1 is also known as ZFM1 (Zinc Finger protein in the MEN1 locus), because it maps on chromosome 11q13, close to the locus responsible for MEN1 (Multiple Endocrine Neoplasia). SF1 is essential for embryonic development in mice and *C. elegans* and for viability in *S. cerevisiae* and mammalian cells (Abovich &

Rosbash, 1997; Mazroui et al, 1999; Rain et al, 1998; Shitashige et al, 2007b; Tanackovic & Krämer, 2005).

The SF1 gene has 14 exons, which are translated from an AUG start codon in exon 1. Several SF1 isoforms have been identified, which mainly differ in length and in the C-terminal amino acid sequences (Arning et al, 1996; Krämer et al, 1998). The SF1 isoforms share the following structural domains present in the first 356 amino acids (figure 10):

- two NLSs (nuclear localization signals; aa 15-19, 92-104);
- a ULM (U2AF65 Ligand Motif) domain (aa 15-22) that binds to the C-terminal UHM of U2AF65;
- a helix-hairpin domain (aa 45-116), which interacts with the UHM of U2AF65 and is essential for cooperative formation of the ternary SF1-U2AF65-RNA complex (Zhang et al, 2013);
- a KH (hnRNP K homology) domain and a QUA2 (Quaking Homology 2) domain (aa 135-228 and aa 236-255, respectively), essential for BPS recognition and spliceosome assembly (Rain et al, 1998);
- a zn (zinc) knuckle (aa 279-292), which increases RNA binding through non-specific interactions (Berglund et al, 1998b).

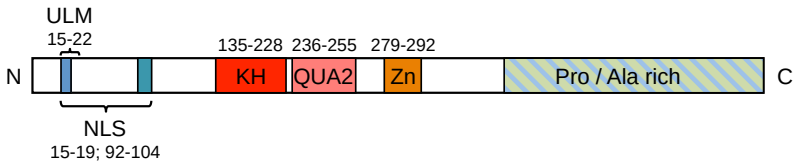


Figure 10. Schematic representation of SF1 protein domains.

All characterized SF1 domains are shown: NLS, nuclear localization signal; ULM, U2AF65 ligand motif; KH/QUA2 domain; Zn, zinc knuckle. Amino acid positions are indicated above the figure. The C-terminal part of the protein differs in length among SF1 isoforms and can have proline or alanine-rich sequences (Pro/Ala rich).

These domains are highly conserved across several species (Rain et al, 1998). However, the yeast homolog BBP has two, and not one, zn knuckles, which could explain why BBP binds to the RNA with a better affinity compared to SF1 (Berglund et al, 1998b). SF1 and BBP

have a low affinity for RNA (K_D = 30 μM and 500 nM, respectively), which is increased 20-fold when they are bound to U2AF65 (Berglund et al, 1998a; Berglund et al, 1997).

The C-terminal part of SF1 is less conserved and dispensable for spliceosome assembly (Rain et al, 1998). This region varies among different isoforms, which are generated via AS of the 3' half of the SF1 pre-mRNA (by exon skipping, intron retention, and duplication of 5' and 3' ss) (figure 11) (Krämer et al, 1998; Rain et al, 1998).

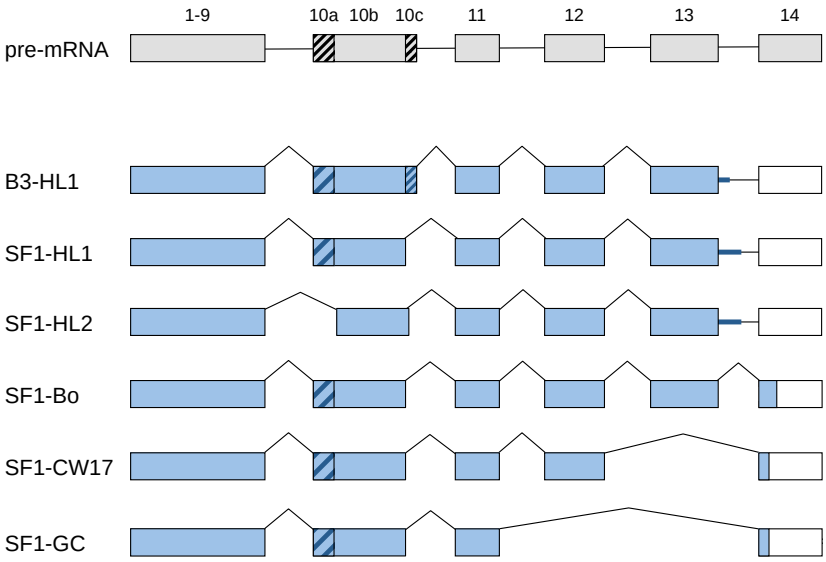


Figure 11. AS events in the 3' end of SF1 pre-mRNA.

The SF1 pre-mRNA is shown on the top. Exons are represented as grey boxes, introns as lines and alternative 3' and 5' ss as dashed boxes. AS events specific for each SF1 isoforms are shown below with the corresponding name specified on the left. Exons are indicated by blue boxes and alternative 3' and 5' ss by dashed boxes. Retained introns are shown by a thick line, where only the blue part is translated. The white boxes indicate the 3' UTR.

1.8. SF1 isoforms

Several AS events have been identified in the 3' half of the SF1 pre-mRNA, which give rise to the following SF1 isoforms: B3-HL1 (or AlaA), SF1-HL1 (ProA), SF1-HL2 (ProA-10a), SF1-Bo (ProB), SF1-CW17 (ProC) and SF1-GC (ProD) (figures 11 and 12) (Krämer et al, 1998). All SF1 isoforms share the first nine exons, which correspond to amino acids 1-356 (figure 11). The isoform transcripts differ for intron 13 retention, exon 12 and 13 skipping, 5' and 3' ss duplication of exon 10. Use of the duplicated 5' ss of exon 10 induces a frame shift in the reading frame of the protein, which leads to two main classes of SF1 isoforms

containing either Ala- or Pro-rich sequences in the C-terminal half. The Pro-rich region, present in most SF1 isoforms, was shown to be involved in protein-protein interactions (Goldstrohm et al, 2001; Lin et al, 2004).

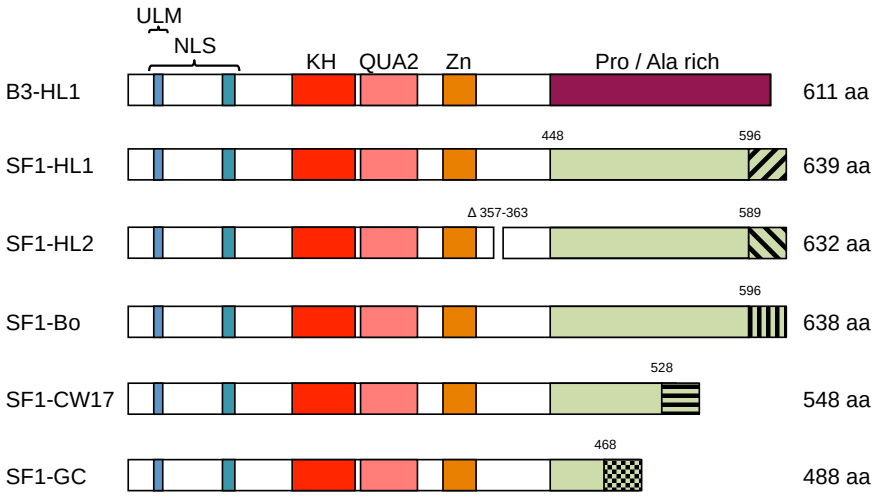


Figure 12. Schematic representation of SF1 isoforms.

The six SF1 isoforms share the same amino acid sequence until amino acid 356 and they all contain the NLS, ULM, KH-QUA2 and Zn domains. Name and length of each SF1 isoforms are indicated on the left and right, respectively. Amino acid positions are indicated above the figure. B3-HL1 contains an alanine-rich sequence (purple box). SF1-HL1, SF1-HL2, SF1-Bo, SF1-CW17 and SF1-GC share a common proline-rich sequence (green box), but differ in length and C-terminal sequence (dashed boxes).

B3-HL1 is the only Ala-rich isoform identified so far and derives from use of the distal 5' ss of exon 10 (figures 11 and 12). The mRNA contains 4 additional nt in exon 10, named exon 10c, and the B3-HL1 isoform is characterized by an Ala-rich sequence in the C-terminal part of the protein (figures 11 and 12) (Arning et al, 1996). All the remaining SF1 isoforms lack exon 10c and have a C-terminal Pro-rich sequence (figures 11 and 12).

In addition, other AS events occur in the 3' half of the SF1 pre-mRNA. B3-HL1, SF1-HL1 and SF1-HL2 retain intron 13, whereas SF1-HL2 lacks exon 10a due to use of an alternative 3' ss of exon 10 (figure 11). Exon 10a encodes only 7 amino acids and the SF1-HL2 isoform therefore has the same reading frame as SF1-HL1 (figures 11 and 12). So far, SF1-HL2 is the only isoform detected that is missing exon 10a. The SF1-Bo isoform does not retain intron 13, whereas SF1-CW17 and SF1-GC lack exon 13 and exons 12-13, respectively (figure 11).

The C-terminal SF1 isoforms shown were identified in HeLa cells. The gene structure is conserved in mouse, suggesting that the same AS events can occur in this organism yielding the same SF1 isoforms (Krämer et al, 1998). In addition, in *D. melanogaster* a novel SF1 isoform lacking exon 4 was identified. This isoform was detected only at the RNA level; Western blot analysis failed to detect any protein of the corresponding size (Mazroui et al, 1999). The transcript is likely degraded due to the presence of a PTC (premature termination codon) in exon 5. A similar alternative transcript was also identified in mammalian cells. An mRNA lacking exon 3, corresponding to exon 4 of *D. melanogaster*, but not the protein, has been detected in HeLa cells (Choleza and Krämer, unpublished results).

The SF1 isoforms share all structural domains required for spliceosome assembly, suggesting that they might all be involved in splicing. It was shown that SF1-Bo, SF1-HL1 and SF1-HL2 function in pre-spliceosome formation *in vitro* and they have the same cellular localization (Arning et al, 1996; Choleza and Krämer, unpublished results). However, SF1 isoforms show a cell-type and tissue-specific expression, suggesting that they might have different functions (Arning et al, 1996; Choleza, Tanackovic and Krämer, unpublished results). However, to date, the exact role of the SF1 isoforms in splicing and/or other cellular functions has still to be elucidated.

1.9. SF1 functions

1.9.1. The role for SF1 in splicing

The main function of SF1 is related to splicing, in particular to the formation of the E complex during the early steps of spliceosome assembly. However, later studies have shown that SF1 is not required for the splicing of all the introns. In *S. cerevisiae*, it was demonstrated that depletion of BBP does not affect pre-spliceosome formation or catalysis and that conditional mutants of BBP showed reduced splicing efficiency only of reporter pre-mRNAs containing a weak 5' ss and/or BPS (Rutz & Seraphin, 1999; Rutz & Seraphin, 2000). In *C. elegans*, it was shown that SF1, together with U2AF, regulates AS *in vivo* of a specific nematode transcript and that mutation or depletion of SF1 induces retention of intron 1 and exon 3 skipping (Ma et al, 2011). In human cells, *in vivo* analyses of SF1 RNA targets revealed that SF1 does not bind to all introns and that it affects the ratio of exon inclusion/exclusion of some pre-mRNAs tested (Corioni et al, 2011). Taking together, these

results show that SF1 is an alternative rather than a constitutive splicing factor, because it participates only in the splicing of certain introns, particularly those with weak splice sites. This would not be the first case that a constitutive splicing factor regulates AS decisions. Besides SF1, there are other proteins, such as SF3a120 and U2AF65 that were also shown to function as AS factors (Hastings et al, 2007; Park et al, 2004). Further evidence that SF1 is required for the splicing of introns with weak splice sites was provided by the analysis of spliceosome assembly in SF1-depleted HeLa cell nuclear extract. It was observed that SF1 depletion leads to decreased A complex formation only of pre-mRNAs with a weak BPS (Guth & Valcarcel, 2000). These results suggest that SF1 does not regulate the splicing of different substrates in the same way and that it promotes the conversion from E to A complex, i.e. has a kinetic role. Similar observations were also made in budding yeast. In *S. cerevisiae*, depletion of BBP to 99% does not compromise pre-spliceosome formation but leads to a reduced transition rate from CC1 to CC2 with subsequent accumulation of the CC1 complex (Rutz & Seraphin, 1999). Similar to other splicing factors, BBP is recycled at the end of the splicing reaction. When only low levels of BBP are present, the conversion of CC1 into CC2 is slowed down by the need to recycle BBP thus leading to an accumulation of CC1. Therefore, in budding yeast BBP plays an important role in the kinetics of the splicing reaction by assisting the transition of CC1 to CC2. However, the mechanism by which SF1 promotes the conversion of splicing complexes E or CC1 into A or CC2 in mammals and yeast is still unknown.

SF1 was shown to be an essential protein in yeast, humans and worms. However, since depletion of SF1 does not compromise spliceosome assembly and splicing in general, the vital function of SF1 could be indirectly related to splicing. SF1 might regulate splicing and/or AS of certain essential genes.

Analysis of in vivo targets of SF1 showed that it can function either as an activator or repressor of AS (Corioni et al, 2011). Surprisingly, SF1 RNA binding sites are not only present close to 3’ ss, but also throughout introns, in exons and 3’ UTRs, suggesting a more complex role of SF1 in splicing regulation and possibly additional functions unrelated to splicing (Corioni et al, 2011). It should be stressed that the *in vivo* experiments performed to identify RNA targets of SF1 did not discriminate between SF1 isoforms, which might have different functions, different RNA targets and different binding sites in the pre-mRNA. Therefore, the diverse RNA binding sites obtained by *in vivo* analysis could be a reflection of different functions of individual isoforms.

1.9.2. Other roles of SF1

Additional functions of SF1, unrelated to spliceosome assembly, have also been described.

1.9.2.1. SF1 and pre-mRNA retention

In *S. cerevisiae*, BBP was shown to function in nuclear pre-mRNA retention (Rutz & Seraphin, 2000), since BBP mutants show leakage of un-spliced pre-mRNA to the cytoplasm. A combination of pre-mRNA leakage with disruption of the NMD (non-sense mediated decay) pathway strongly affects yeast viability, probably due to accumulation of aberrantly spliced transcripts. However, at present, data for a similar role for SF1 in mammals are lacking.

1.9.2.2. SF1 and transcription repression

Several reports proposed a role for SF1 as a transcriptional repressor (Zhang & Childs, 1998; Zhang et al, 1998). SF1 binds SSAP (activation domain of stage-specific activator protein), a transcription factor that activates the expression of the sea urchin late H1 gene at the mid-blastula stage of embryogenesis (Zhang & Childs, 1998). SF1 negatively regulates the activity of the activation domain of SSAP, thus functioning as a transcriptional repressor. Similarly, SF1 interacts with the activation domains of EWS (Ewing's sarcoma) and TLS (translocated in liposarcoma), which are involved in Ewing's sarcoma tumor formation and function in transcription initiation (Zhang et al, 1998). As a transcriptional repressor, SF1 inhibits the transcription activation of genes targeted by EWS and TLS.

In addition, SF1 was shown to interact with the N-terminal part of the transcription factor CA150, which negatively regulates transcription elongation (Goldstrohm et al, 2001). The N-terminal half of CA150 contains WW domains, which specifically bind to the C-terminal Pro-rich region of SF1. WW domain binding to SF1 positively correlates with the repression activity of CA150 and mutations in the WW domains affect CA150 repression activity and the binding to SF1.

1.9.2.3. SF1 and non-coding RNAs

SF1 was shown to bind to lncRNAs (long non-coding RNAs), such as NEAT1/Men-ε/β (Nuclear-Enriched Abundant Transcript 1), NEAT2/MALAT-1 and Gomafu (Tsuiji et al, 2011; Choleza and Krämer, unpublished results). NEAT1 localizes to nuclear paraspeckles and functions as a scaffold for their formation (Clemson et al, 2009). NEAT1 was shown to be essential for paraspeckles assembly, since its depletion leads to the loss of paraspeckles (Clemson et al, 2009). However, NEAT1 knock-out mice do not show any particular phenotype and so far, the function of paraspeckles is unknown (Nakagawa et al, 2011). All SF1 isoforms localize in paraspeckles, however, why SF1 is present in these structures is not clear (Choleza and Krämer, unpublished results). SF1 also binds the tandem repeats of UACUAAC sequences present in the neuronal lncRNA Gomafu (Tsuiji et al, 2011). Interaction between SF1 and Gomafu is not surprising, considering that Gomafu's tandem repeats are identical to the consensus BPS. It was suggested that Gomafu might regulate splicing by competing with weak BPS sequences for the binding of SF1.

1.9.2.4. SF1 and cancer

Although SF1 is expressed ubiquitously, its expression levels are associated with tumor progression. So far, contradictory results have been shown regarding SF1 and the incidence of tumors. *sf1* +/- mice have an increased susceptibility to colon cancer, whereas in testis reduced levels of SF1 lead to a reduction of testicular tumors (Shitashige et al, 2007b; Zhu et al, 2010). SF1, therefore, has different functions according to the cell type and it acts either as a tumor suppressor or as a proto-oncogene, by regulating the AS of cell-type specific pre-mRNAs, which in turn can affect tumor growth in different ways. In addition, SF1 regulates AS of cancer-related splice variants (WISP1v and FG3R-ATII), although the exact mechanism is not clear, nor has it been shown whether this function is direct (Shitashige et al, 2007a).

1.10. Aim of the thesis

SF1 has been initially identified as a protein that functions in the earliest steps of spliceosome assembly and it was shown to be essential for viability in yeast and mammals

(Krämer, 1992). Later studies revealed a controversial role of SF1 in splicing. First, SF1 is not required for the splicing of all introns but only of introns with weak consensus splice sites, indicating that SF1 functions as an alternative rather than constitutive splicing factor (Corioni et al, 2011; Ma et al, 2011; Rutz & Seraphin, 2000). Second, in mammals and yeast SF1 was shown to have a kinetic role in spliceosome assembly (Guth & Valcarcel, 2000; Rutz & Seraphin, 1999). Third, several reports have suggested an involvement of SF1 in processes not related to splicing, such as nuclear pre-mRNA retention and transcription regulation (Goldstrohm et al, 2001; Rutz & Seraphin, 2000; Zhang & Childs, 1998; Zhang et al, 1998). Therefore the role of SF1 appears to be more complex to what was thought earlier and one of the goals of my PhD work was to clarify SF1 function by identifying novel interacting proteins.

The second part of thesis focuses on the analysis of SF1 isoforms. To date, six SF1 isoforms have been identified in mammalian cells, which arise from combinations of AS choices in the 3' half of the pre-mRNA, leading to different amino acid sequences in the C-terminal, but not N-terminal half of the protein (Arning et al, 1996; Krämer et al, 1998). Database searches of mRNAs and ESTs (expressed sequence tags) identified non-annotated alternative SF1 transcripts containing different 5' portions of the pre-mRNA, which suggested use of an AP. However, the existence of such SF1 isoforms has never been verified experimentally. The N-terminal and central parts of SF1 are common to all isoforms and contain evolutionarily conserved domains required for spliceosome assembly. Isoforms lacking any of these domains might not function in splicing and thus play a different role in the cell. The analysis of N-terminal SF1 isoforms was performed in different cell lines at the RNA and protein level. In addition, the question of whether the SF1 gene contains an AP was addressed.

Results I: Analysis of SF1-interacting proteins

2. Results I: Analysis of SF1-interacting proteins

2.1. Introduction

As described in the main introduction, SF1 was shown to function not only in splicing but also in other cellular processes, such as nuclear pre-mRNA retention and transcription regulation (Corioni et al, 2011; Goldstrohm et al, 2001; Rutz & Seraphin, 2000; Zhang & Childs, 1998; Zhang et al, 1998). This part of the thesis, presented as a manuscript, focuses on elucidating the role of SF1 in splicing and more generally in the cell by identifying its interacting partners. For this purpose, SF1 IP (immunoprecipitation) from HeLa cell lysates was performed by Flore Raleff in the Krämer lab, followed by MS (mass spectrometry) analysis in the group of Henning Urlaub (Max Planck Institute for Biophysical Chemistry, Göttingen, Germany). In addition, data from Y2H (yeast two hybrid) screens were kindly provided by Jean-Christophe Rain (Hybrigenics, Paris, France). Known and new SF1 binding partners were identified, which were classified either according to their function in splicing (results from IP/MS) or according to structural domains (Y2H data). My PhD work focused on validating the SF1 interaction with SURP domain-containing proteins by biochemical approaches. I performed the protein-protein interaction assays with the support of our technical staff, Flore Raleff and Ivona Bagdiul. SURP domains are present in proteins that function in splicing and several are associated with the U2 snRNP (Kuwasako et al, 2006; Will et al, 2002). I therefore investigated the functional relevance of these interactions by performing *in vitro* spliceosome assembly, UV-cross-linking of proteins to RNA and psoralen cross-linking.

2.2. Manuscript: Mammalian splicing factor SF1 interacts with SURP domains of U2 snRNP associated proteins

Mammalian splicing factor SF1 interacts with SURP domains of U2 snRNP-associated proteins

Angela Crisci[1], Flore Raleff[1], Ivona Bagdiul[1], Monika Raabe[2], Henning Urlaub[2,3], Jean-Christophe Rain[4] and Angela Krämer[1,*]

[1] Department of Cell Biology, Faculty of Sciences, University of Geneva, CH-1211 Geneva 4, Switzerland

[2] Bioanalytical Mass Spectrometry, Max Planck Institute for Biophysical Chemistry, D-37077 Göttingen, Germany

[3] Bioanalytics, Institute for Clinical Chemistry, University Medical Center Göttingen, D-37075 Göttingen, Germany

[4] Hybrigenics Services, F-75014 Paris, France

* To whom correspondence should be addressed. Tel: +41 22 379 6750; Fax: +41 22 379 6442; Email: angela.kraemer@unige.ch

ABSTRACT

Splicing factor 1 (SF1) functions by recognizing the branch point sequence (BPS) at the 3' splice site during the formation of early complex E. SF1 binding to the BPS pre-bulges the BPS adenosine, thought to facilitate subsequent base-pairing of the U2 snRNA with the BPS. The 65-kDa subunit of U2 snRNP auxiliary factor (U2AF65) interacts with SF1 and was shown to recruit the U2 snRNP to the spliceosome. We report here that co-immunoprecipitation experiments with anti-SF1 antibodies confirm the association of SF1 with early splicing complexes, but surprisingly almost all U2 snRNP proteins were found associated with SF1. Yeast two-hybrid (Y2H) screens identified two SURP domain-containing U2 snRNP proteins as partners of SF1. In vitro binding assays demonstrated direct interactions of a short, evolutionarily conserved region of SF1 with the SURP domains, stressing their role in protein-protein interactions. A reduction of A complex formation in SF1-depleted extracts could be rescued with recombinant SF1 containing the SURP-interaction domain, but only partial rescue was observed with SF1 lacking this sequence. Thus, SF1 can initially recruit the U2 snRNP to the spliceosome during E complex formation, whereas U2AF65 may stabilize the association of the U2 snRNP with the spliceosome at later times.

INTRODUCTION

Pre-mRNA splicing is essential for eukaryotic gene expression and one of the most versatile mechanisms to increase proteome diversity through alternative splice site choices (1). The reaction is catalyzed by the spliceosome, a multi-megadalton complex (2). Five small nuclear ribonucleoprotein particles (U1, U2, U4, U5 and U6 snRNPs) and more than 100 non-snRNP proteins assemble in a step-wise fashion on the pre-mRNA through networks of RNA-RNA, RNA-protein and protein-protein interactions.

In brief, essential sequence elements at the 5' and 3' splice sites are recognized at the onset of spliceosome assembly (2). U1 snRNP binds to the 5' splice site, whereas SF1 and the U2 snRNP auxiliary factor (U2AF) recognize distinct sequences at the 3' splice site. SF1 and the large subunit of U2AF (U2AF65) interact with one another and cooperatively bind the branch point sequence (BPS) and the adjacent polypyrimidine tract, respectively (3-7); the small U2AF subunit (U2AF35) recognizes the 3' splice site AG dinucleotide (8). Together these interactions assemble the early complex E, which is converted into pre-splicing complex A by incorporation of the U2 snRNP. This is accomplished through interaction of the U2 snRNP-associated SF3b155 with U2AF65 (9), binding of U2 snRNP proteins to and adjacent to the BPS (10-12) and base pairing of the U2 snRNA with the BPS. These events are thought to displace SF1 from the spliceosome (13). The following steps involve binding of the remaining snRNPs and additional non-snRNP proteins, juxtaposition of the splice sites and dynamic remodeling of the complexes leading to the formation of the catalytic center, followed by intron removal in two catalytic steps (2).

SF1 was initially identified as a protein required for pre-spliceosome assembly (14,15). It is evolutionarily conserved and the *Saccharomyces cerevisiae* ortholog, termed branch point-binding protein (BBP), was shown to function in the formation of the early commitment complex CC2, which together with CC1 is the equivalent of the human E complex (13). SF1 is essential in *S. cerevisiae*, *Caenorhabditis elegans* and mammalian cells (4,16-19), but surprisingly, initial RNA interference-mediated knockdown of SF1 in human cells did not affect the splicing of several pre-mRNAs tested (18). Similarly, biochemical or genetic depletion of SF1/BBP from human extracts or yeast cells only marginally affected splicing activity, possibly due to a kinetic role for SF1 in spliceosome assembly (13,20). Later experiments demonstrated that SF1 is not involved in the splicing of all introns, but influences alternative splicing decisions (19,21-23). Thus, altered or mis-splicing of essential genes in the absence of SF1 could reflect its requirement for viability. In addition to its role in splicing, SF1 has been implicated in nuclear pre-mRNA retention and transcriptional repression (24,25), functions that may contribute to or be the cause of SF1's essential phenotype.

To better understand SF1 function in splicing and potentially other processes, we set out to identify interacting proteins by co-immunoprecipitation (co-IPs) from HeLa cell nuclear extracts combined with mass spectrometry and in yeast two-hybrid (Y2H) screens with human SF1. Our results confirm that the major role for SF1 is in early stages of the splicing reaction. We demonstrate a novel function of SF1, which is essential for efficient spliceosome assembly, in the initial recruitment of the U2 snRNP through direct interactions with two U2 snRNP-associated proteins. The isolation of Y2H partners

moreover suggests interactions of SF1 with additional proteins implicated in splicing, but also in other processes and identifies domains potentially mediating these interactions.

MATERIAL AND METHODS

Co-IP of SF1-interacting proteins and mass spectrometry

HeLa cell nuclear extract (corresponding to 2 mg of total protein) was incubated in the absence or presence of 0.3 mg/ml RNase A (Sigma) for 15 min at room temperature followed by pre-clearing for 1 hr at 4°C with Protein G Sepharose (GE Healthcare) equilibrated in IP buffer (50 mM Tris-HCl, pH 7.9, 100 mM NaCl, 0.05% Nonidet P-40 [NP-40] and 0.5 mM dithiothreitol [DTT]). Unbound material was incubated for 1 hr at 4°C with Protein G Sepharose-coupled anti-SF1 (Abnova; H7536-MO1A) or control mouse IgG (Sigma, P4810), followed by centrifugation at 2,000 rpm in a microfuge for 2 min at 4°C. Unbound material was saved and the beads were washed three times with 1 ml IP buffer. Bound material was eluted with NuPAGE LDS sample buffer (Life Technologies) and separated in NuPAGE Novex 4-12% Bis-Tris protein gels (Life Technologies), followed by Coomassie blue staining. Lanes were cut into 23 slices, proteins were in-gel trypsin-digested and extracted according to Shevchenko et al. (26). Peptides were analyzed in a LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific) under standard conditions. Proteins were identified by searching fragment spectra against the NCBI non-redundant database using Mascot as search engine. The number of product ion spectra (total spectrum count) of identified peptides/proteins was compared between anti-SF1 and control-IgG samples. Proteins with less than five spectrum counts in the anti-SF1 samples and those with more than 15% of spectra in control IgG versus anti-SF1 were eliminated, in addition to proteins assumed to represent common contaminants, such as ribosomal proteins.

Yeast two-hybrid analysis

Yeast two-hybrid screens were performed with the partial coding sequence of human SF1 (amino acids 1-441; GenBank accession number gi: 295842308), which was PCR-amplified and cloned into pB27 as a C-terminal fusion to LexA (N-LexA-SCRIB-C). The construct was verified by sequencing and used as a bait to screen random-primed and dT-primed cDNA libraries constructed in pP6 (Table 2) with a mean insert size of 900 bp. Library construction has been described (27). pB27 and pP6 are derived from pBTM116 and pGADGH, respectively (28).

Baits were screened using a mating approach with YHGX13 (Y187 ade2-101::loxP-kanMX-loxP, matα) and L40ΔGal4 (mata) yeast strains as described (29). The number of clones tested and histidine-positive clones are listed in Table 2. Prey fragments of positive clones were PCR-amplified and sequenced at the 5' and 3' junctions. Preys from the dT-primed library were only sequenced at the 5' junctions. The resulting sequences were used to identify the corresponding interacting proteins in the GenBank database (http://www.ncbi.nlm.nih.gov/gene/) with a fully automated procedure. A confidence score (Predicted Biological Score, PBS) was attributed to each interaction as described (27).

Cloning procedures

DNAs were amplified with the Expand High Fidelity PCR System (Roche) according to the manufacturer's instructions and suitable primers. Correct cloning was verified by sequencing (Microsynth, Switzerland).

DNAs for bacterial expression of the GST-tagged SURP1 of SF3a120 (amino acids 38-113), SURP2 of SFSWAP (amino acids 197-272), and the single SURP domain of CHERP (amino acids 1-75) were cloned into the Gateway vector pDEST15 (Life Technologies). The plasmid encoding the GST-tagged U2AF65 UHM (amino acids 367-473) has been described (4).

Plasmids for bacterial expression of SF1 proteins were cloned as follows: C-terminal truncations carrying N-terminal His_6-tags were cloned into the BamHI and EcoRI sites of pTrcHisA (Life Technologies). N-terminal deletions were cloned into the NcoI and EcoRI sites of pETMBP-1a (a gift of Michael Sattler, Helmholtz Center and Technical University, Munich), encoding proteins with N-terminal His_6-MBP tags. Internal deletions of SF1-C370 in pETMBP-1a were generated by replacing the deleted sequences with one (Δ306-326 and ΔZn) or two (ΔKH) KpnI sites, encoding Gly and Thr. Internal sequences of SF1 (127-302, 295-335 and 304-326) were cloned into the NcoI/EcoRI sites of pETMBP-1a.

For transient expression of N-terminal GFP fusion proteins in HeLa cells, SF1 sequences were cloned into the Gateway vector pDEST53 (Life Technologies).

Templates for in vitro transcription of 3' splice site pre-mRNAs were generated by HindIII and BstEII digestion of a pBluescript plasmid encoding the region spanning exons 1 and 2 of AdML pre-mRNA (30). The 3' overhangs were blunt-ended with T4 DNA polymerase (Promega) followed by religation. The resulting pre-mRNA contains 51 nts of vector sequences, 75 nts of the 3' end of AdML intron 1 and 38 nts of exon 2. This plasmid was used to mutate the original AdML BPS (UACUUAU) to a consensus (UACUAAC) or weak (AAUUCAC) BPS with the GeneTailor Site-Directed Mutagenesis System (Life Technologies).

Bacterial expression and purification of recombinant proteins

Plasmids encoding GST-tagged SURP domains and His_6-MBP-tagged SF1 proteins were transformed into *E. coli* strain BL21 by heat shock. Plasmids encoding GST, GST-U2AF65-UHM and His_6-tagged SF1 proteins were transformed into *E. coli* strain XL1-Blue by heat shock. Proteins were expressed for 4 h at 37°C after addition of isopropyl-β-D-thiogalactopyranoside (IPTG) to a final concentration of 1 mM. Cells were harvested by centrifugation at 5,000 × g for 10 min.

Cells expressing GST-tagged proteins were lysed in PBS supplemented with complete protease inhibitors (Roche) by sonication on ice and supplemented with Triton X-100 to a final concentration of 1%. Proteins were purified with 500 µl of a 50% suspension of glutathione agarose (Sigma) equilibrated in PBS. Unbound proteins were removed by washing three times with 10 ml PBS and GST-tagged proteins were eluted with 5 mM glutathione and 50 mM Tris-HCl, pH 8.0.

Cells expressing His_6- or His_6-MBP-tagged proteins were lysed in 10 mM Tris-HCl, pH 8.0, 100 mM NaCl, 50 mM Na phosphate, 8 M urea and complete protease inhibitors. Proteins were purified on TALON Metal Affinity Resin (BD Biosciences). Unbound material was removed by three washes with lysis buffer containing 0.4 M NaCl and 20 mM imidazole and bound proteins were eluted with 0.1 M

EDTA, pH 8.0.

All recombinant proteins were dialyzed against D buffer (31) supplemented with 3 mM $MgCl_2$ and stored at -80°C. Purified proteins were quantified by SDS-PAGE and Coomassie blue staining.

GST pull-down assays

Reactions containing 20 µl of packed glutathione-agarose beads (Sigma) and 0.056 nmole (corresponding to ~2 µg) of GST fusion proteins in a total volume of 200 µl NETN (20 mM Tris-HCl, pH 8.0, 50 mM NaCl, 0.5% NP-40, 0.5 mM EDTA) were incubated for 30 min at 4°C and washed twice with 500 µl NETN. His_6- or His_6-MBP tagged proteins were added in a two-fold molar excess. and the reaction mixture was incubated for 45 min at 4°C in a total volume of 200 µl NETN. Where indicated, purified proteins were treated with 0.3 mg/ml of RNase A for 20 min at room temperature and centrifuged at maximal speed in a microfuge for 5 min prior to GST pull-down. Unbound proteins were removed by three NETN washes. Proteins were eluted from the beads in SDS sample buffer for 5 min at 95°C and separated by SDS-PAGE followed by Western blotting (32). Input and bound proteins were detected with mouse monoclonal anti-His_6 (Sigma; H1029); the amount of GST-tagged proteins was verified by incubation with rabbit anti-GST (A.K., unpublished). Goat anti-mouse and anti-rabbit IRDye800 and 680LT (LI-COR) were used as secondary antibodies and immune complexes were visualized with the Odyssey Fc imaging system (LI-COR Biosciences).

Cell culture, transient transfection and preparation of HeLa cell lysates and nuclear extract

HeLa cells were grown at 37°C with 5% CO_2 in DMEM (Sigma) supplemented with 10% FBS (Sigma), 2 mM L-glutamine (Gibco), 100 u/ml penicillin, and 100 µg/ml streptomycin (Gibco). Cells were plated in 10-cm culture dishes 24 h prior to transfection. Cells were transfected at 60-70% confluency with polyethylenimine (Brunschwig) according to the manufacturer's instructions and collected 72 h later. Whole cell lysates were prepared by lysis of transfected cells in 50 mM Tris-HCl, pH 7.5, 100 mM NaCl, 1% NP-40, 1 mM DTT, 2 mM EDTA and complete protease inhibitors for 30 min at 4°C followed by centrifugation at 16,000 x g for 5 min at 4°C. HeLa cell nuclear extracts were prepared according to Dignam et al. (31) and dialyzed against buffer D.

Co-IPs

HeLa cell nuclear extract was incubated with 0.3 mg/ml RNase A (Sigma) for 20 min at room temperature and centrifuged in a microfuge at maximal speed for 5 min. The supernatant (corresponding to 100 µg of total protein) was incubated for 1 h at 4°C with Dynabeads Protein G (Invitrogen) coated with anti-SF1 or control mouse IgG. The unbound fraction was kept for Western blotting and beads were washed four times with 1 ml of 50 mM Tris-HCl, pH 7.5, 100 mM NaCl, 0.05% NP-40 and 0.5 mM DTT. Bound proteins were eluted with SDS sample buffer for 5 min at 95°C. Input, unbound and bound material were separated by SDS-PAGE followed by Western blotting with the following antibodies: mouse monoclonal anti-CHERP (Santa-Cruz; Sc-100650), rabbit anti-H1 (Santa-Cruz; Sc-10806), mouse monoclonal anti-SF1 (Abnova; H7536-MO1A), rabbit anti-SF3a60

(33), mouse monoclonal anti-SF3a66 (34), rabbit anti-SF3a120 (35), rabbit anti-SFSWAP (Aviva; ARP 40524), and mouse monoclonal anti-U2AF65 (Sigma; U4758). Secondary antibodies and detection of immune complexes were as above.

For precipitation of transiently expressed GFP-tagged SF1 and associated proteins, HeLa whole cell lysate (corresponding to 6 mg of total protein per reaction) was treated with RNase A as above and incubated for 1 h at 4°C with Dynabeads Protein G-coupled goat anti-GFP (a gift of Karla Neugebauer, Yale University). Washing, elution, SDS-PAGE and Western blotting were performed as above. Proteins were detected with anti-SF3a120, anti-U2AF65 and rabbit anti-GFP (Invitrogen; A-11122).

Immunodepletion

SF1 was immunodepleted from HeLa cell nuclear extract adjusted to 600 mM KCl by three passages over Dynabeads Protein G-coupled anti-SF1. The depleted extract was dialyzed against buffer D (31) and stored at -80°C. Mock depletions were performed as above in the absence of antibody.

Spliceosome assembly

AdML pre-mRNAs with a consensus or weak BPS were synthesized with T3 RNA polymerase (Promega) in the presence of [α-^{32}P]UTP and gel-purified (30,36). Spliceosome assembly was performed in triplicate in 10-µl reactions in the presence of 10% untreated, mock- or SF1-depleted HeLa nuclear extract, 10 mM Hepes-KOH, pH 7.9, 50 mM KCl, 0.05 mM EDTA, 0.25 mM DTT, 1 µg tRNA and 1.6 pmol pre-mRNA at 30°C for the times shown in the figure legends. Where indicated, reactions were complemented with 0.022, 0.22, or 2.2 pmol SF1/C370 or SF1/302. Reaction products were separated in native 4% polyacrylamide gels (37). Gels were dried and exposed to phosphorimager screens. Quantification was done with the Molecular Imager FX (BioRad) and software Quantity One V 4.2.1 (BioRad).

UV cross-linking and IP of U2AF65

Spliceosome assembly was performed as above in 20-µl reactions in the presence of 25% mock- or SF1 depleted extract and 18 pmol pre-mRNA. Samples were incubated at 30°C for 15 min, UV cross-linked and treated with RNase A as described (38). Dynabeads Protein G coupled with control IgG or anti-U2AF65 and 100 µl buffer D (31) were added and samples were incubated for 1.5 hrs at 4°C with rotation. After centrifugation, the beads were washed four times with 600 µl of 50 mM Tris-HCl, pH 7.5, 0.5 M NaCl, 1% NP-40 and 0.5 mM DTT, and once with 50 mM Tris-HCl, pH 7.5 and 1% NP-40. Bound material was eluted by boiling in SDS loading buffer for 5 min, followed by centrifugation and separation by 10% SDS-PAGE. Gels were dried, exposed to a phosphoimager screen and signals were quantified as above.

RESULTS

The U2 snRNP co-immunoprecipitates with SF1

To identify potential novel partners of SF1, we performed co-IPs of HeLa cell nuclear extracts followed by mass spectrometry (MS). Extracts were incubated in the absence or presence of RNase A, pre-cleared on Protein G Sepharose and aliquots were incubated with Protein G Sepharose-coupled control mouse IgG or a monoclonal antibody directed against the N-terminal 110 amino acids of SF1. Eluted proteins were separated by 1D PAGE (Figure S1), in-gel trypsin-digested and analyzed by LC-MSMS.

A total of 54 proteins were specifically precipitated by anti-SF1 (Table 1), four of which were previously reported to directly interact with SF1: FBP11/PRPF40A, PUF60, SPF45/RBM17 and U2AF65/U2AF2, (3,4,16,39-41). Forty-five proteins are represented in a list of 244 annotated spliceosomal proteins (see 42). Among these, 27 are associated with the U1 or U2 snRNPs or function in pre-splicing complex A assembly. Nine are part of the U4/U6, U5 or U4/U6.U5 snRNPs or function in the formation of later complexes. Components with roles in splicing catalysis were not detected. Three members of the SR family of proteins with functions early and late during spliceosome assembly were found, in addition to five proteins with roles in mRNA export or mRNA binding and one miscellaneous protein associated with the spliceosome. Thus, consistent with the function of SF1 at the initial steps of spliceosome assembly, more than half of the proteins co-precipitating with SF1 are components of early spliceosomes, compared to 52 of the annotated early spliceosomal proteins (excluding SR and hnRNP proteins), whereas only a small percentage (ten of 103 proteins) are associated with later complexes.

A closer inspection of the early spliceosomal proteins precipitated by anti-SF1 revealed the presence of nine of the 12 U2 snRNP-specific proteins, as well as seven of 11 proteins more loosely associated with the U2 snRNP (Table 1, U2 and U2-related, respectively) (43). This finding was surprising, since the U2 snRNP is thought to replace SF1 from the spliceosome during the complex E to A transition by binding of the U2 snRNP-associated SF3b155 to the U2AF65-UHM, a site occupied by SF1 in complex E (13,44-46) and base pairing of the U2 snRNA with the BPS. Moreover, SF1 has not been found among U2 snRNP-associated proteins (43). However, although the U2 snRNA base-pairs with the BPS only in complex A, the U2 snRNP has been detected already in complex E (47-49).

Nine proteins isolated by co-IP are not known to be related to splicing (Table 1). Most are nuclear proteins with functions in RNA binding and/or transcription. The presence of some of these proteins may be due to interactions with other components of the splicing machinery that are directly or indirectly associated with SF1, or they represent contaminations, since only few peptides were recovered.

SF1 interacts with U2 snRNP-associated proteins in the yeast two-hybrid (Y2H) system

Information regarding potential direct SF1 partners was obtained from Y2H screens with residues 1-441 of human SF1 as a C-terminal fusion to LexA. This portion of SF1 contains all domains required

for function in in vitro splicing assays (4), but lacks most of the Pro-rich C-terminal region, previously shown to strongly activate transcription in the Y2H system (4). cDNAs of four human and three mouse tissues or cell lines were tested for interactions (Table 2).

As summarized in Table 3, previously described SF1 partners were detected in the Y2H screens (ABL1/2, PRPF40A/FBP11, RBM17/SPF45, U2AF65 and UHMK1/KIST) (3,4,16,39,41,50), with U2AF65 showing the highest confidence score (predicted biological score, PBS; Table 3). A number of spliceosome-associated proteins and other proteins with roles in splicing were among the potential partners. However, only few proteins found in complex with SF1 by co-IP/MS were detected, suggesting that the majority of these were not direct binding partners, but co-precipitated with SF1 from HeLa cell extracts due to numerous protein-protein interactions in the spliceosome (2).

Another potential SF1 partner with a high-confidence PBS is KPNA2, a subunit of the nuclear import complex, in addition to other members of the same complex, but with a lower PBS depending on the subunit or cell type (Table 3). Moreover, KIAA0907 was found as a SF1 partner with very high confidence in all cell types tested. Roles in splicing for this protein, also termed BLOM7, have been described (51). It contains a KH domain, similar to a few other proteins found in the Y2H screens with a lower PBS. Apart from these proteins, SF1 may also directly interact with RRM- or zinc finger-containing proteins. A possible relation of these proteins to SF1 function will be described in the discussion.

Among the preys with reported roles in splicing are two U2 snRNP-associated proteins: splicing factor SF3A subunit 1 (SF3A1/SF3a120) and calcium homoeostasis endoplasmic reticulum protein (CHERP) (Table 3). SF3a120 is part of a heterotrimeric complex with SF3a66 and SF3a60 and essential for U2 snRNP function in spliceosome assembly (52). CHERP was originally identified as an endoplasmic reticulum protein involved in the regulation of intracellular Ca^{2+} homoeostasis (53), but is also loosely associated with the U2 snRNP, has been reported to localize to the nucleus and function in alternative splicing (43,54,55).

Inspection of cDNA fragments of these proteins recovered in the Y2H screens indicated that suppressor of white-apricot/Prp21 (SURP) domains (56) may mediate the interaction with SF1 (Figure 1A). SURP domains are ~40 amino acids long, often arranged in tandem and all known SURP-containing proteins are involved in splicing (57). Thus far, only the function of the second SURP domain of SF3a120 as a protein-protein interaction module has been established (57-59). The Y2H screens identified SURP1 of SF3a120 and the single SURP domain of CHERP as potential SF1 binding domains (Figure 1A). Moreover, the suppressor of white-apricot homolog (SFSWAP/SFRS8), an alternative splicing factor with two SURP domains (56,60-63) was detected in the Y2H screens (Table 3), with its second SURP domain potentially interacting with SF1 (Figure 1A). Two additional proteins containing SURP domains were found by co-IP (Table 1), the U2 snRNP-associated SURP motif-containing protein (U2SURP/SR140) and the A complex component SURP and G-patch domain-containing protein 1 (SUGP1/SF4) (43,64).

Together, these data suggest that SF1 binds U2 snRNP-associated proteins and SFSWAP via SURP domain-mediated interactions.

SF1 directly binds SURP domains of SF3a120, CHERP and SFSWAP

To validate binding of SF1 to SURP domains, we first performed co-IPs with anti-SF1 coupled to Dynabeads Protein G from RNase A-treated HeLa cell nuclear extracts and Western blotting with antibodies against the potential interacting proteins. Figure 1B shows that SF3a120, as well as SF3a60 and SF3a66 were co-precipitated with SF1; thus, the entire SF3a heterotrimer was bound. CHERP was also found in the precipitate; however, SFSWAP was not detected in the bound fraction. U2AF65, known to interact with SF1 (4,16) and used as a positive control, was bound to SF1, whereas histone H1, a negative control, was not bound. None of the proteins were bound to Dynabeads Protein G coated with non-specific IgG. These results show that SF3a120 and CHERP can be co-precipitated with SF1 from HeLa cell extracts, whereas an interaction with SFSWAP was not detected.

To determine whether SF1 binding to SF3a120, CHERP and SFSWAP was mediated by direct interactions with the SURP domains identified in the Y2H screens, we performed GST pull-down assays with GST-tagged SURP domains and recombinant His_6-tagged SF1-C370 (containing amino acids 1-370, i.e. the region common to all SF1 isoforms). His_6-SF1-C370 was bound by the SURP domains of SF3a120, CHERP and SFSWAP, as well as by the positive control, U2AF65-UHM, but not by GST alone (Figure 1C). The interactions were insensitive to RNase A treatment prior to GST pull-down, demonstrating that SURP domain binding of SF1 was not mediated by RNA. Thus, SURP1 of SF3a120, the single SURP domain of CHERP and SURP2 of SFSWAP are sufficient for direct SF1 binding, confirming the results of the Y2H screens.

An evolutionarily conserved domain in SF1 is essential for SURP domain binding

The region of SF1 responsible for the SURP domain interaction was analyzed in GST pull-down assays with His_6-tagged C- and N-terminal as well as internal deletions of SF1 (Figure 2A). Compared to SF1-C370, binding to GST-SF3a120-SURP1, CHERP-SURP and SFSWAP-SURP2 was slightly weakened with SF1-C327 and strongly reduced upon deletion to amino acid 320 (Figure 2B). Binding of SF1-C315 was barely detectable and further deletion of SF1 residues completely abolished the SURP domain interaction. Thus, the C-terminal border of the SURP interaction domain (ID) lies between SF1 residues 315-327.

SF1 mutants with N-terminal His_6 and maltose-binding protein (MBP) tags were used to further delimit the SF1 SURP ID. N-terminal deletions up to, and including the zinc knuckle (N127 to N292; Figure 2C) did not abolish binding of SF1 to the SURP domains. Thus, neither the SF1 KH/QUA2 domain nor the zinc knuckle are involved in the interaction, which was confirmed with SF1 mutants carrying internal deletions of these domains (ΔKH and ΔZn). SF1 residues 295-335 were barely sufficient for the interaction and amino acids 127-302 or 304-326 did not bind the SURP domains. Moreover, an internal deletion of residues 306-326 abolished binding. The C-terminal or internal deletions did not affect binding to GST-U2AF65-UHM, used as a positive control, whereas none of the N-terminal deletions interacted with GST-U2AF65-UHM, since the SF1 ULM was not present. Binding to GST alone was not observed. Together, these data indicate that the region between residues 293

and 327 of SF1 is required for the interaction with SURP domains.

To further test the relevance of this domain for SURP binding, GFP-tagged SF1-C370 and -C302 were transiently expressed in HeLa cells. RNase A-treated whole cell lysates were immunoprecipitated with Dynabeads Protein G-coupled anti-GFP. SDS-PAGE and Western blotting with anti-SF3a120 showed that endogenous SF3a120 interacted with GFP-SF1-C370 but not -C302, whereas U2AF65 bound both proteins (Figure 2D). Thus, the SF1 SURP-ID identified in vitro with recombinant proteins is also essential for binding SF3a120 in HeLa cell lysates.

A multiple sequence alignment of SF1 revealed the presence of a region well-conserved from *Drosophila* to humans (amino acids 293-326 of human SF1; Figure 2E) encompassing the residues that eliminated the SURP domain interaction. This region is highly conserved in mammals, *Xenopus* and zebrafish, and partial conservation of key residues (amino acids 301-321) is seen in other organisms, including *Arabidopsis*, *Schizosaccharomyces pombe* and *Dictyostelium*, with the exception of *S. cerevisiae*.

From these results we conclude that an evolutionarily conserved domain in SF1, spanning amino acids 293-327 is essential for binding SURP domains.

The SF1/SURP interaction is required for efficient early spliceosome assembly

The above results demonstrate that SF1 can bind the U2 snRNP through direct interactions with SF3a120 and CHERP. SF1 recognizes the BPS in complex E and the U2 snRNP is weakly bound to this complex before base pairing of the U2 snRNA to the BPS and pre-mRNA binding of several U2 snRNP-associated proteins in complex A (5,9,12,48,49).To test the possibility that SF1 is involved in the recruitment of the U2 snRNP to the pre-mRNA, we compared spliceosome formation in HeLa cell nuclear extracts and SF1-depleted extracts in the absence or presence of recombinant SF1 proteins with or without the SURP ID. To facilitate quantification of the results, spliceosome assembly was analyzed with an AdML pre-mRNA lacking the 5' splice site and containing 75 nts of the 3' end of intron 1 and part of exon 2. Similar splicing substrates are efficiently assembled into a 3' splice site complex related to pre-splicing complex A (termed A 3' complex hereafter) but not converted to complex B (36). In addition, the original BPS of AdML intron 1 was changed to the yeast consensus BPS, UACUAAC, which is the optimal BPS for mammalian splicing (65) and also the preferred binding site of SF1 (5,21,66).

Immunodepletion with anti-SF1 bound to Dynabeads Protein G resulted in a 91% reduction of different SF1 isoforms compared to untreated or mock-depleted nuclear extracts (Figure 3A). A 3' complex formation tested by incubation for 30 min at 30°C in triplicate experiments was reduced to an average of 44% of that seen in the mock-depleted extract (Figure 3B). The SF1-depleted extract was then complemented with equivalent concentrations of His_6-SF1-C370 or -C302 (Figure 3B), containing or lacking the SURP ID, respectively (Figures 2A and B). Addition of SF1-C-370 rescued A 3' complex formation to 160% compared to mock-depleted extract in a dose-dependent manner, whereas complex formation in the presence of SF1-C-302 was less efficient (126%; Figure 3C). The observation that the recombinant proteins rescued A 3' complex assembly to more than 100% could be due to the fact that their concentration exceeded that of endogenous SF1. Comparison of protein

levels of endogenous and recombinant proteins is difficult, due to the fact that endogenous SF1 exists in multiple isoforms (Figure 3a), and their exact contribution to splicing is not established.

A rather modest effect of SF1 depletion on spliceosome formation, despite an at least 98% reduction in SF1 protein has been reported by Guth and Valcárcel (20). The same authors showed that the effect of SF1 depletion was more pronounced with a pre-mRNA containing a weak as compared to a strong BPS. Similarly, depletion of SF1 in yeast causes more severe splicing defects with reporters containing mutated splice sites (13). We therefore repeated the experiment with the same pre-mRNA as above, but containing the weak IgM BPS (AAUUCAC) (20,67). Compared to mock-treated extract, SF1-depletion reduced complex formation to about 23% (Figure 3D). Complementation with SF1-C-370 rescued A 3' complex formation to nearly mock-depleted levels (98%), whereas addition of SF1-C-302 resulted in a markedly reduced rescue of 58%. Thus, the region of SF1 necessary for the interaction with U2 snRNP proteins is also essential for efficient A 3' complex assembly. The more marked effect seen with the pre-mRNA containing a suboptimal BPS can be explained by the fact that the consensus BPS is the preferred SF1 binding site (5,21,66).

Previous studies suggested a kinetic role for SF1 in spliceosome assembly (13,20). The above results indicate that SF1 binds two U2 snRNP proteins and the SF1 SURP-ID is required for efficient A 3' complex formation. To test whether the SF1-U2 snRNP interaction underlies the kinetic role of SF1, we performed a time course of A 3' complex formation. Pre-mRNAs with a consensus or weak BPS were incubated with mock- or SF1-depleted HeLa nuclear extract supplemented with His_6-SF1-C370 or -C302. Representative results of experiments performed in triplicate are shown in the top panels of Figures 4A and 4B. A 3' complex formation was quantified and normalized to the values for the 60-min time points of the mock-depleted extract (bottom panels). SF1 depletion resulted in a 65 or 76% reduction in A 3' complex assembly with pre-mRNAs containing a consensus or weak BPS, respectively. Addition of SF1-C370 rescued complex formation to the levels of mock-depleted extract with both pre-mRNAs. In contrast, whereas addition of SF1-C302 to reactions with the pre-mRNA containing a consensus BPS resulted in 70% A 3' complex formation, only 43% complex assembly was observed with the pre-mRNA containing a weak BPS. Thus, as shown above, the overall efficiency of A 3' complex formation is higher under conditions allowing for an interaction between SF1 and the U2 snRNP. Moreover, this interaction accelerates the kinetics of spliceosome assembly, especially when U2 snRNA/BPS base pairing is suboptimal.

SF1/C370 or C302 do not compromise U2AF65 binding to the RNA

SF1 and U2AF65 cooperatively bind the pre-mRNA and increase each others affinity for RNA (3). Thus, SF1 depletion also causes decreased binding of U2AF65 to the polypyrimidine tract with a concomitant decrease in spliceosome assembly, which likely contributes to the lower levels of A 3' complex formation seen above. To rule out the possibility that RNA binding of U2AF65 is affected in a SF1-depleted extract complemented with recombinant SF1 proteins, we performed UV cross-linking of U2AF65 in 15-min reactions containing mock- or SF1-depleted extracts and pre-mRNAs with a consensus or weak BPS. Triplicate samples were UV cross-linked, RNase A-treated and incubated with Dynabeads Protein G-coupled with control IgG or anti-U2AF65. Bound material was analyzed by

SDS-PAGE and the fraction of RNA bound to U2AF65 was quantified (Figure 5). U2AF65 was efficiently cross-linked to both pre-mRNAs in the mock-depleted samples. Only background cross-linking was detected in IPs performed with control IgG. As expected from previous results (3,38), pre-mRNA binding of U2AF65 was partially reduced upon SF1 depletion. However, addition of both, His_6-SF1-C370 or -C302, increased U2AF65-RNA binding to the levels seen in mock-depleted extract, consistent with the fact that both proteins contain the ULM and thus can bind U2AF65 (Figure 2B). We therefore conclude that the reduced levels of A 3' complex formation and slower reaction kinetics observed in SF1-depleted extracts complemented with SF1-C302 are due to the lack of interaction of SF1 with the U2 snRNP and not caused by decreased pre-mRNA binding of U2AF65.

DISCUSSION

Results from co-IP/MS experiments and Y2H screens have provided information about novel and, in part, unexpected partners of SF1. The co-IP/MS experiment demonstrated that about half of the proteins co-precipitated with SF1 from HeLa cell nuclear extracts are involved in early steps of spliceosome assembly, confirming an early role for SF1 in splicing (15,16). Unexpectedly, most of the U2 snRNP-associated proteins were co-precipitated with SF1. Y2H screens revealed potential interactions of SF1 with the U2 snRNP-associated proteins SF3a120 and CHERP, as well as the alternative splicing factor SFSWAP. Direct binding of SF1 to SURP domains of the three proteins were confirmed in GST pull-down assays and the interaction domain was mapped to an evolutionarily conserved, central region of SF1. We furthermore show that the SF1/U2 snRNP interaction is necessary for efficient spliceosome assembly, suggesting a role for SF1 in the early recruitment of the U2 snRNP to the pre-mRNA.

A novel role for SF1 in U2 snRNP recruitment

SF1 is a known component of E complexes and has been shown to leave the assembling spliceosome during the complex E to A transition (13). This evidence comes mainly from studies in *S. cerevisiae*, where two types of E complexes can easily be visualized in vitro. In the human system SF1 has only been detected in small amounts in A complexes isolated under physiological conditions and the material used for isolation of associated proteins could have contained small amounts of E complexes (68). Consistent with these data, anti-SF1 antibodies precipitated more than half of the annotated proteins present in early splicing complexes, compared to 10% of proteins associated with later complexes or functioning in splicing catalysis and post-splicing events (2,42).

The presence of almost all U2 snRNP proteins in the anti-SF1 precipitate was unexpected, since events that lead to the incorporation of the U2 snRNP into the spliceosome are thought to trigger the release of SF1 from the spliceosome at the same time. First, SF3b155, a U2 snRNP protein, interacts with the UHM of U2AF65, the same site that is bound by SF1 (44,46). Second, the U2 snRNP base pairs with the BPS and the BPS adenosine is bound by SF3b14a/p14 in complex A, most likely replacing the SF1/BPS interaction (6,10,69). However, although the U2 snRNA/BPS interaction only occurs during A complex formation, the U2 snRNP has been detected in E complexes in HeLa cell extracts (47-49). Therefore, an interaction between SF1 and the U2 snRNP via the SURP domain-containing proteins identified in the Y2H screens may occur prior to the events mentioned above and aid in the initial recruitment of the U2 snRNP to the spliceosome.

We envision the following series of events: SF1 and U2AF65 interact with one another, which involves the ULM at the N terminus of SF1 and the UHM of U2AF65 (44). This is followed by their cooperative binding to the BPS and the polypyrimidine tract at the 3' splice site (3). SF1 binds the BPS through its KH domain (6), leaving the SURP-ID available for binding SF3a120 or CHERP. The SF1/BPS interaction pre-bulges the BPS adenosine, leaving other nucleotides accessible for U2 snRNA base pairing (6). In principle, SF1 could interact with the U2 snRNP before or after binding to U2AF65, but in either case, it would position the U2 snRNP close to the BPS. The transition to complex A then occurs through replacement of the SF1/U2AF65 interaction by binding of SF3b155 to

the U2AF65 UHM, binding of SF3b14a/p14 to the BPS adenosine, base pairing of U2 snRNA to the BPS and binding of SF3a and SF3b subunits to the pre-mRNA on both sides of the BPS, tethering the U2 snRNP for later steps (9,10,12,46,69). In this scenario, SF1 initiates U2 snRNP recruitment and facilitates the events required for A complex formation by placing the U2 snRNP into the vicinity of U2AF65 and the BPS, whereas U2AF65 stabilizes the association of the U2 snRNP with the spliceosome once SF1 is released.

The interaction of SF1 with U2 snRNP-associated proteins is required for efficient spliceosome assembly, since A complex formation is strongly reduced following SF1 depletion and can be restored by addition of SF1 containing, but less so by addition of SF1 lacking the SURP-ID. This effect is not due to the disruption of the SF1/U2AF65 interaction, since both SF1 proteins contain the ULM and pre-mRNA binding of U2AF65 was not compromised in the presence of either protein. Guth and Valcárcel (20) suggested a function of SF1 in addition to increasing the affinity of U2AF65 for the polypyrimidine tract. The results presented here strongly suggest that this additional role is the binding of SF1 to the U2 snRNP, thereby directly promoting the U2 snRNP/BPS interaction.

Rescue of complex formation with SF1 lacking the SURP-ID was more efficient in the presence of a weak than a consensus BPS, as previously observed in SF1-depleted HeLa cell extracts (20). Moreover, splicing defects in *S. cerevisiae* after SF1 depletion were more pronounced with reporters containing mutated splice sites (13). These effects can be explained by a preference for the consensus BPS in mammalian splicing (65). On the one hand, SF1 has highest affinity for the consensus BPS, but can bind many sequences resembling the consensus BPS (5,21,66). On the other hand, the BPS-interacting sequence in U2 snRNA forms the most base pairs with the consensus BPS (65). Thus, the SF1/U2 snRNP interaction may be particularly important for the splicing of introns with a suboptimal BPS.

The interaction of SF1 with SURP domains

The Y2H screens identified SURP1 of SF3a120, the single SURP domain of CHERP and SURP2 of SFSWAP as SF1 interaction partners, which was confirmed with recombinant proteins. The remaining SURP domains of SF3a120 and SFSWAP could not be tested, since the bacterially expressed proteins were insoluble (data not shown). However, the following points argue that SF1 only interacts with the SURP domains tested. First, although some of the cDNAs isolated in the Y2H screens completely or partially included the other SURP domains of SF3a120 and SFSWAP, SF3a120 SURP1 and SFSWAP SURP2 were present in all of the cDNAs. Second, Kuwasako et al. (57) only obtained SURP2 of SF3a120 in soluble form when co-expressed with SF3a60. In analogy to these experiments, co-expression of SF1 and SF3a120 SURP2 did not yield soluble SURP2 (data not shown), suggesting that SF1 does not interact with SF3a120 SURP2. Third, SURP domains can be classified into two evolutionarily related subgroups (57). The SURP domains isolated in the SF1 Y2H screens belong to subgroup 1, whereas the SF3a120 SURP2 and SFSWAP SURP1 belong to subgroup 2. We therefore believe it is highly likely that SF1 only interacts with SURP domains of subgroup 1, but not those of subgroup 2.

The SF1 SURP-ID is located immediately C-terminal to the zinc knuckle. Its sequence is almost

invariant in mammals and conserved to a lesser extent in *A. thaliana* or *S. pombe*, but it is not present in *S. cerevisiae* SF1. BLAST searches did not identify a related sequence in other mammalian proteins. SF3a60 binds SURP2 of SF3a120 and single amino acid changes in SURP2 and SURP1 swap the identity of the domains, i.e. prevent SURP2 binding of SF3a60 but allow for SURP1 binding (57). It is therefore intriguing to speculate that the interaction sites in SF3a60 and SF1 show a certain degree of similarity. A sequence alignment between the SF1 and SF3a60 SURP-IDs does not reveal obvious homology (data not shown). The SF3a60 SURP-ID forms a long amphipathic α-helix (57) and an equivalent interaction surface has been described between the yeast orthologs (59). Analysis with the Phyre2 protein fold recognition server (70) predicts a helical structure for the SF1 SURP-ID (data not shown). Future structural studies should solve the question of whether the SURP-binding modes of SF1 and SF3a60 are similar.

Lack of conservation of the SF1 SURP-ID in *S. cerevisiae*

Somewhat as a surprise, the SF1 SURP-ID is not conserved in *S. cerevisiae* SF1 (BBP). In addition, SURP1 in the yeast ortholog of SF3a120 (Prp21p) lacks four amino acids close to a 3_{10} helix, which could change the spatial organization of the interaction site (57). This may suggest that BBP and Prp21p do not interact. In fact, Y2H interactions between BBP and Prp21p have not been detected (J. C. Rain, unpublished results) and are also not documented in the *Saccharomyces* Genome Database (http://www.yeastgenome.org/).

Could this interaction be dispensable in *S. cerevisiae*? The vast majority of yeast introns contain the consensus BPS, UACUAAC, (71) with a perfect complementarity to the BPS-interacting sequence of U2 snRNA (G-UAGUA; the dash indicates the missing complementarity to the bulged-out BPS adenosine). Given our observation that the SF1-U2 snRNP interaction is more important for efficient splicing of introns with a weak BPS, SF1 may thus not be required for U2 snRNP recruitment in yeast.

The SF1/SFSWAP interaction

SFSWAP was identified as an additional partner of SF1 in Y2H screens from mouse tissues. SF1 binding to its SURP2 domain was confirmed in GST pull-down assays, but we did not detect the interaction by anti-SF1 IP from HeLa cell lysates. SFSWAP is an alternative splicing regulator that also autoregulates the splicing of its own pre-mRNA (56,60-63); however, targets of its regulation remain largely unknown. Given that SF1 also acts as an alternative splicing factor (21), the two proteins may interact to regulate the splicing of perhaps only a few pre-mRNAs or only in certain tissues, which could explain the failure to detect the interaction in HeLa cells. Finding common partners with appropriate methods could solve this question.

Additional partners of SF1

Anti-SF1 antibodies co-precipitated only few proteins without a reported connection to splicing. Some of these are involved in transcription, which may be related to the published role for SF1 in transcription repression (72,73). Similarly, several transcription factors were identified as potential

partners of SF1 in the Y2H screens (e.g. zinc finger-containing proteins).

The Y2H screens were informative concerning other potential SF1 partners and structural domains mediating interactions with SF1. KPNA2 was found as a prey in all cell types tested with the highest-confidence PBS observed. Other members of this family of nuclear import factors were also detected, but with lower PBS. SF1 is a nuclear protein and therefore its interaction with the nuclear import machinery may not be surprising. However, Y2H screens with other nuclear proteins did not find KPNAs in such abundance (J.-C.R., unpublished data). The shortest selected interaction domain (ΣSID) of the KPNAs comprises armadillo repeats, which are also found in the nuclear pore complex protein Nup155, another prey of SF1. Therefore, SF1 may have a thus far unknown function related to nuclear import.

Another protein with a high PBS found in all cell types analyzed is KIAA0907/BLOM7. It was initially identified in Y2H screens with a component of the human Prp19 complex and has been shown to increase splicing in HeLa cell nuclear extracts and regulate alternative splice site decisions (51). BLOM7 contains two KH domains, the second of which is comprised in the SF1 ΣSID. Six additional proteins with KH domains covered by the ΣSID were found as SF1 preys (KHDRBS1/SAM68; KHDRBS3/SLM-2; KHSRP/FUBP2; hnRNPK; FUBP3; QKI), most of which play roles in splicing (74-78). Interestingly, SF1 itself contains a KH domain of the same type as those of KHDRBS1, KHDRBS2, QKI and BLOM7.

KIAA0562/CEP104 is a centrosomal protein of 104 kDa and another example of a prey identified in all cell types with a high PBS. The ΣSID of CEP104 does not encompass/comprise any known structural domain and a potential role for a SF1/CEP104 interaction is illusive.

In addition to these proteins, SF1 preys include RRM-containing proteins, most of which with reported functions in splicing. Interestingly, the ΣSIDs of these proteins often comprise tandem RRMs. This observation may suggest that SF1 not only binds proteins with non-canonical RRMs of the UHM class (40,41,44,79), but also other types of RRMs.

Taken together, the Y2H screens have provided a vast amount of information about possible SF1 partners and novel interaction domains, the relevance of which remains to be investigated.

To conclude, our data demonstrate a new role for SF1 in the recruitment of the U2 snRNP to the spliceosome by interaction with SURP domain-containing proteins. The Y2H screens identified other potential partners of SF1. Studying these interactions may reveal additional functions of SF1 in splicing and other cellular pathways.

ACKNOWLEDGEMENT

We thank Karla Neugebauer (Yale University) for generously providing goat anti-GFP antibodies, Michael Sattler (Helmholtz Center and Technical University, Munich) for the pETMBP-1a vector and Patricia Arboit (Proteomics Core Facility, University of Geneva) for help with MS data analysis.

REFERENCES

1. Irimia, M. and Blencowe, B.J. (2012) Alternative splicing: decoding an expansive regulatory layer. *Curr. Opin. Cell Biol.*, **24**, 323-332.
2. Wahl, M.C., Will, C.L. and Lührmann, R. (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell*, **136**, 701-718.
3. Berglund, J.A., Abovich, N. and Rosbash, M. (1998) A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. *Genes Dev.*, **12**, 858-867.
4. Rain, J.-C., Rafi, Z., Rhani, Z., Legrain, P. and Krämer, A. (1998) Conservation of functional domains involved in RNA binding and protein-protein interactions in human and *Saccharomyces cerevisiae* pre-mRNA splicing factor SF1. *RNA*, **4**, 551-565.
5. Berglund, J.A., Chua, K., Abovich, N., Reed, R. and Rosbash, M. (1997) The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell*, **89**, 781-787.
6. Liu, Z., Luyten, I., Bottomley, M., Messias, A., Houngninou-Molango, S., Sprangers, R., Zanier, K., Krämer, A. and Sattler, M. (2001) Structural basis for recognition of the intron branch site by splicing factor 1. *Science*, **294**, 1098-1102.
7. Zamore, P.D., Patton, J.G. and Green, M.R. (1992) Cloning and domain structure of the mammalian splicing factor U2AF. *Nature*, **355**, 609-614.
8. Moore, M. (2000) Intron recognition comes of AGe. *Nature Struct. Biol.*, **7**, 14-16.
9. Gozani, O., Potashkin, J. and Reed, R. (1998) A potential role for U2AF-SAP 155 interactions in recruiting U2 snRNP to the branch site. *Mol. Cell. Biol.*, **18**, 4752-4760.
10. Query, C.C., Strobel, S.A. and Sharp, P.A. (1996) Three recognition events at the branch-site adenine. *EMBO J.*, **15**, 1392-1402.
11. Will, C.L., Schneider, C., MacMillan, A.M., Katopodis, N.F., Neubauer, G., Wilm, M., Lührmann, R. and Query, C.C. (2001) A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site. *EMBO J.*, **20**, 4536-4546.
12. Gozani, O., Feld, R. and Reed, R. (1996) Evidence that sequence-independent binding of highly conserved U2 snRNP proteins upstream of the branch site is required for assembly of spliceosomal complex A. *Genes Dev.*, **10**, 233-243.
13. Rutz, B. and Séraphin, B. (1999) Transient interaction of BBP/ScSF1 and Mud2 with the splicing machinery affects the kinetics of spliceosome assembly. *RNA*, **5**, 819-831.
14. Krämer, A. and Utans, U. (1991) Three protein factors (SF1, SF3 and U2AF) function in pre-splicing complex formation in addition to snRNPs. *EMBO J.*, **10**, 1503-1509.
15. Krämer, A. (1992) Purification of splicing factor SF1, a heat-stable protein that functions in the assembly of a pre-splicing complex. *Mol. Cell. Biol.*, **12**, 4545-4552.
16. Abovich, N. and Rosbash, M. (1997) Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. *Cell*, **89**, 403-412.
17. Mazroui, R., Puoti, A. and Krämer, A. (1999) Splicing factor SF1 from *Drosophila* and *Caenorhabditis*: presence of an N-terminal RS domain and requirement for viability. *RNA*, **5**, 1615-1631.
18. Tanackovic, G. and Krämer, A. (2005) Human splicing factor SF3a, but not SF1, is essential for pre-mRNA splicing in vivo. *Mol. Biol. Cell*, **16**, 1366-1377.
19. Shitashige, M., Satow, R., Honda, K., Ono, M., Hirohashi, S. and Yamada, T. (2007) Increased susceptibility of $Sf1^{(+/-)}$ mice to azoxymethane-induced colon tumorigenesis. *Cancer Sci.*, **98**, 1862-1867.
20. Guth, S. and Valcárcel, J. (2000) Kinetic role for mammalian SF1/BBP in spliceosome assembly and function after polypyrimidine tract recognition by U2AF. *J. Biol. Chem.*, **275**, 38059-38066.

21. Corioni, M., Antih, N., Tanackovic, G., Zavolan, M. and Krämer, A. (2011) Analysis of in situ pre-mRNA targets of human splicing factor SF1 reveals a function in alternative splicing. *Nucleic Acids Res.*, **39**, 1868-1879.

22. Haraguchi, N., Andoh, T., Frendewey, D. and Tani, T. (2007) Mutations in the SF1-U2AF59-U2AF23 complex cause exon skipping in *Schizosaccharomyces pombe*. *J. Biol. Chem.*, **282**, 2221-2228.

23. Shitashige, M., Naishiro, Y., Idogawa, M., Honda, K., Ono, M., Hirohashi, S. and Yamada, T. (2007) Involvement of splicing factor-1 in beta-catenin/T-cell factor-4-mediated gene transactivation and pre-mRNA splicing. *Gastroenterology*, **132**, 1039-1054.

24. Rutz, B. and Séraphin, B. (2000) A dual role for BBP/ScSF1 in nuclear pre-mRNA retention and splicing. *EMBO J.*, **19**, 1873-1886.

25. Galy, V., Gadal, O., Fromont-Racine, M., Romano, A., Jacquier, A. and Nehrbass, U. (2004) Nuclear retention of unspliced mRNAs in yeast is mediated by perinuclear Mlp1. *Cell*, **116**, 63-73.

26. Shevchenko, A., Wilm, M., Vorm, O. and Mann, M. (1996) Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.*, **68**, 850-858.

27. Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C. *et al.* (2005) Protein interaction mapping: a Drosophila case study. *Genome Res.*, **15**, 376-384.

28. Vojtek, A.B. and Hollenberg, S.M. (1995) Ras-Raf interaction: two-hybrid analysis. *Methods Enzymol.*, **255**, 331-342.

29. Fromont-Racine, M., Rain, J.C. and Legrain, P. (1997) Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.*, **16**, 277-282.

30. Arning, S., Grüter, P., Bilbe, G. and Krämer, A. (1996) Mammalian splicing factor SF1 is encoded by variant cDNAs and binds to RNA. *RNA*, **2**, 794-810.

31. Dignam, J.D., Lebovitz, R.M. and Roeder, R.G. (1983) Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res.*, **11**, 1475-1489.

32. Kyhse-Anderson, J. (1984) Electroblotting of multiple gels. *J. Biochem. Biophys. Methods*, **10**, 203-209.

33. Krämer, A., Legrain, P., Mulhauser, F., Gröning, K., Brosi, R. and Bilbe, G. (1994) Splicing factor SF3a60 is the mammalian homologue of PRP9 of *S. cerevisiae*: the conserved zinc finger-like motif is functionally exchangeable in vivo. *Nucleic Acids Res.*, **22**, 5223-5228.

34. Brosi, R., Hauri, H.P. and Krämer, A. (1993) Separation of splicing factor SF3 into two components and purification of SF3a activity. *J. Biol. Chem.*, **268**, 17640-17646.

35. Krämer, A., Mulhauser, F., Wersig, C., Gröning, K. and Bilbe, G. (1995) Mammalian splicing factor SF3a120 represents a new member of the SURP family of proteins and is homologous to the essential splicing factor PRP21p of *S. cerevisiae*. *RNA*, **1**, 260-272.

36. Frendewey, D. and Keller, W. (1985) Stepwise assembly of a pre-mRNA splicing complex requires U-snRNPs and specific intron sequences. *Cell*, **42**, 355-367.

37. Krämer, A. (1988) Pre-splicing complex formation requires two proteins and U2 snRNP. *Genes Dev.*, **2**, 1155-1167.

38. Guth, S., Martinez, C., Gaur, R.K. and Valcárcel, J. (1999) Evidence for substrate-specific requirement of the splicing factor U2AF(35) and for its function after polypyrimidine tract recognition by U2AF(65). *Mol. Cell. Biol.*, **19**, 8263-8271.

39. Bedford, M.T., Chan, D.C. and Leder, P. (1997) FBP WW domains and the Abl SH3 domain bind to a specific class of proline-rich ligands. *EMBO J.*, **16**, 2376-2383.

40. Corsini, L., Hothorn, M., Stier, G., Rybin, V., Scheffzek, K., Gibson, T.J. and Sattler, M. (2009) Dimerization and protein binding specificity of the U2AF homology motif of the splicing factor Puf60. *J. Biol. Chem.*, **284**, 630-639.

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

42. Hegele, A., Kamburov, A., Grossmann, A., Sourlis, C., Wowro, S., Weimann, M., Will, C.L., Pena, V., Luhrmann, R. and Stelzl, U. (2012) Dynamic protein-protein interaction wiring of the human spliceosome. *Mol. Cell*, **45**, 567-580.

43. Will, C.L., Urlaub, H., Achsel, T., Gentzel, M., Wilm, M. and Lührmann, R. (2002) Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5p homologue and an SF3b DEAD-box protein. *EMBO J.*, **21**, 4978-4988.

44. Selenko, P., Gregorovic, G., Sprangers, R., Stier, G., Rhani, Z., Krämer, A. and Sattler, M. (2003) Structural basis for the molecular recognition between human splicing factors U2AF^{65} and SF1/mBBP. *Mol. Cell*, **11**, 965-976.

45. Spadaccini, R., Reidt, U., Dybkov, O., Will, C., Frank, R., Stier, G., Corsini, L., Wahl, M.C., Lührmann, R. and Sattler, M. (2006) Biochemical and NMR analyses of an SF3b155-p14-U2AF-RNA interaction network involved in branch point definition during pre-mRNA splicing. *RNA*, **12**, 410-425.

46. Thickman, K.R., Swenson, M.C., Kabogo, J.M., Gryczynski, Z. and Kielkopf, C.L. (2006) Multiple U2AF65 binding sites within SF3b155: thermodynamic and spectroscopic characterization of protein-protein interactions among pre-mRNA splicing factors. *J. Mol. Biol.*, **356**, 664-683.

47. Hong, W., Bennett, M., Xiao, Y., Feld Kramer, R., Wang, C. and Reed, R. (1997) Association of U2 snRNP with the spliceosomal complex E. *Nucleic Acids Res.*, **25**, 354-361.

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

49. Makarov, E.M., Owen, N., Bottrill, A. and Makarova, O.V. (2012) Functional mammalian spliceosomal complex E contains SMN complex proteins in addition to U1 and U2 snRNPs. *Nucleic Acids Res*, **40**, 2639-2652.

50. Manceau, V., Swenson, M., Le Caer, J.P., Sobel, A., Kielkopf, C.L. and Maucuer, A. (2006) Major phosphorylation of SF1 on adjacent Ser-Pro motifs enhances interaction with U2AF65. *FEBS J.*, **273**, 577-587.

51. Grillari, J., Loscher, M., Denegri, M., Lee, K., Fortschegger, K., Eisenhaber, F., Ajuh, P., Lamond, A.I., Katinger, H. and Grillari-Voglauer, R. (2009) Blom7alpha is a novel heterogeneous nuclear ribonucleoprotein K homology domain protein involved in pre-mRNA splicing that interacts with SNEVPrp19-Pso4. *J. Biol. Chem.*, **284**, 29193-29204.

52. Nesic, D. and Krämer, A. (2001) Domains in human splicing factors SF3a60 and SF3a66 required for binding to SF3a120, assembly of the 17S U2 snRNP, and prespliceosome formation. *Mol. Cell. Biol.*, **21**, 6406-6417.

53. LaPlante, J.M., O'Rourke, F., Lu, X., Fein, A., Olsen, A. and Feinstein, M.B. (2000) Cloning of human Ca2+ homoeostasis endoplasmic reticulum protein (CHERP): regulated expression of antisense cDNA depletes CHERP, inhibits intracellular Ca2+ mobilization and decreases cell proliferation. *Biochem. J.*, **348 Pt 1**, 189-199.

54. Lin-Moshier, Y., Sebastian, P.J., Higgins, L., Sampson, N.D., Hewitt, J.E. and Marchant, J.S. (2013) Re-evaluation of the role of calcium homeostasis endoplasmic reticulum protein (CHERP) in cellular calcium signaling. *J. Biol. Chem.*, **288**, 355-367.

55. Sasaki-Osugi, K., Imoto, C., Takahara, T., Shibata, H. and Maki, M. (2013) Nuclear ALG-2 protein interacts with Ca2+ homeostasis endoplasmic reticulum protein (CHERP) Ca2+-dependently and participates in regulation of alternative splicing of inositol trisphosphate receptor type 1 (IP3R1) pre-mRNA. *J. Biol. Chem.*, **288**, 33361-33375.

56. Spikes, D.A., Kramer, J., Bingham, P.M. and Van Doren, K. (1994) SWAP pre-mRNA splicing regulators are a novel, ancient protein family sharing a highly conserved sequence motif with the prp21 family of constitutive splicing proteins. *Nucleic Acids Res.*, **22**, 4510-4519.

57. Kuwasako, K., He, F., Inoue, M., Tanaka, A., Sugano, S., Guntert, P., Muto, Y. and Yokoyama, S. (2006) Solution structures of the SURP domains and the subunit-assembly mechanism within the splicing factor SF3a complex in 17S U2 snRNP. *Structure*, **14**, 1677-1689.

58. Huang, C.J., Ferfoglia, F., Raleff, F. and Krämer, A. (2011) Interaction domains and nuclear targeting signals in subunits of the U2 small nuclear ribonucleoprotein particle-associated splicing factor SF3a. *J. Biol. Chem.*, **286**, 13106-13114.

59. Lin, P.C. and Xu, R.M. (2012) Structure and assembly of the SF3a splicing factor complex of U2 snRNP. *EMBO J.*, **31**, 1579-1590.

60. Denhez, F. and Lafyatis, R. (1994) Conservation of regulated alternative splicing and identification of functional domains in vertebrate homologs to the Drosophila splicing regulator, suppressor-of-white-apricot. *J. Biol. Chem.*, **269**, 16170-16179.

61. Sarkissian, M., Winne, A. and Lafyatis, R. (1996) The mammalian homolog of suppressor-of-white-apricot regulates alternative mRNA splicing of CD45 exon 4 and fibronectin IIICS. *J. Biol. Chem.*, **271**, 31106-31114.

62. Lemaire, R., Winne, A., Sarkissian, M. and Lafyatis, R. (1999) SF2 and SRp55 regulation of CD45 exon 4 skipping during T cell activation. *Eur. J. Immunol.*, **29**, 823-837.

63. Wang, J., Gao, Q.S., Wang, Y., Lafyatis, R., Stamm, S. and Andreadis, A. (2004) Tau exon 10, whose missplicing causes frontotemporal dementia, is regulated by an intricate interplay of cis elements and trans factors. *J. Neurochem.*, **88**, 1078-1090.

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

65. Zhuang, Y.A., Goldstein, A.M. and Weiner, A.M. (1989) UACUAAC is the preferred branch site for mammalian mRNA splicing. *Proc. Natl. Acad. Sci. USA*, **86**, 2752-2756.

66. Peled-Zehavi, H., Berglund, J.A., Rosbash, M. and Frankel, A.D. (2001) Recognition of RNA branch point sequences by the KH domain of splicing factor 1 (mammalian branch point binding protein) in a splicing factor complex. *Mol. Cell. Biol.*, **21**, 5232-5241.

67. Watakabe, A., Tanaka, K. and Shimura, Y. (1993) The role of exon sequences in splice site selection. *Genes Dev.*, **7**, 407-418.

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

69. MacMillan, A.M., Query, C.C., Allerson, C.R., Chen, S., Verdine, G.L. and Sharp, P.A. (1994) Dynamic association of proteins with the pre-mRNA branch region. *Genes Dev.*, **8**, 3008-3020.

70. Kelley, L.A. and Sternberg, M.J. (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.*, **4**, 363-371.

71. Schwartz, S., Silva, J., Burstein, D., Pupko, T., Eyras, E. and Ast, G. (2008) Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res.*, **18**, 88-103.

72. Zhang, D., Paley, A.J. and Childs, G. (1998) The transcriptional repressor ZFM1 interacts with and modulates the ability of EWS to activate transcription. *J. Biol. Chem.*, **273**, 18086-18091.

73. Goldstrohm, A.C., Albrecht, T.R., Sune, C., Bedford, M.T. and Garcia-Blanco, M.A. (2001) The transcription elongation factor CA150 interacts with RNA polymerase II and the pre-mRNA splicing factor SF1. *Mol. Cell. Biol.*, **21**, 7617-7628.

74. Chawla, G., Lin, C.H., Han, A., Shiue, L., Ares, M., Jr. and Black, D.L. (2009) Sam68 regulates a set of alternatively spliced exons during neurogenesis. *Mol. Cell. Biol.*, **29**, 201-213.

75. Ehrmann, I., Dalgliesh, C., Liu, Y., Danilenko, M., Crosier, M., Overman, L., Arthur, H.M., Lindsay, S., Clowry, G.J., Venables, J.P. *et al.* (2013) The tissue-specific RNA binding protein T-STAR controls regional splicing patterns of neurexin pre-mRNAs in the brain. *PLoS Genet.*, **9**, e1003474.

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

77. Venables, J.P., Koh, C.S., Froehlich, U., Lapointe, E., Couture, S., Inkel, L., Bramard, A., Paquet, E.R., Watier, V., Durand, M. *et al.* (2008) Multiple and specific mRNA processing targets for the major human hnRNP proteins. *Mol. Cell. Biol.*, **28**, 6033-6043.

78. Hall, M.P., Nagel, R.J., Fagg, W.S., Shiue, L., Cline, M.S., Perriman, R.J., Donohue, J.P. and Ares, M., Jr. (2013) Quaking and PTB control overlapping splicing regulatory networks during muscle cell differentiation. *RNA*, **19**, 627-638.

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

80. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947-2948.

FIGURE LEGENDS AND TABLES

Figure 1. Analysis of the interaction of SF1 with SURP domain-containing proteins. (A) Scheme of SURP domain-containing proteins identified in Y2H screens. The domain structure of SF3a120, CHERP and SFSWAP is shown with SURP domains indicated in dark grey. Other domains are shown in light grey: CID, RNA polymerase II-binding domain; G-patch, G-patch domain; RS domain, Arg/Ser-rich domain; UBL, ubiquitin-like domain. Numbering was taken from UniProt entries (www.uniprot.org). The smallest interaction domain (SID) deduced from cDNAs found in the Y2H screens is indicated below the proteins (numbering according to human proteins). (B) Co-IP. HeLa cell nuclear extract was incubated with Dynabeads Protein G coated with anti-SF1 or control IgG. Input (I; 10% of total), bound (B) and unbound (U) fractions were separated by 7.5% SDS-PAGE (10% for anti-H1) followed by Western blotting with antibodies against the proteins indicated on the right side of each panel. (C) GST pull-down. His_6-tagged SF1-C370 was incubated with GST alone, GST-tagged U2AF65-UHM, SF3a120-SURP1, CHERP-SURP or SFSWAP-SURP2 bound to glutathione-agarose as indicated above the figure. GST-tagged proteins were mock-treated (-) or digested with RNase A (+) as shown. The His_6-SF1-C370 input (I; 10% of total) and bound proteins were separated by 10% SDS-PAGE and Western blotted with anti-His_6 (top) and anti-GST antibodies (bottom). The migration of protein markers is indicated in kDa on the left.

Figure 2. Determination of the SURP-ID of SF1. (A) Scheme of mutant SF1 proteins. Boxes represent known protein domains: ULM, UHM ligand motif; KH/QUA2, K-homology/Quaking2 domain; Zn, zinc knuckle. Numbering above full-length SF1 refers to amino acids. The names of SF1 mutants are shown on the left; numbers on the right refer to the residues comprising the proteins. (B) and (C) GST pull-down of mutant SF1 proteins with SURP domains. GST-tagged SF3a120-SURP1, CHERP-SURP, SFSWAP-SURP2, U2AF65-UHM and GST alone, bound to glutathione agarose (as indicated on the right), were incubated with mutant His_6- (B) or His_6-MBP-tagged (C) SF1 proteins as indicated above the figures. Bound proteins were separated by 10% SDS-PAGE and Western blotted with anti-His_6 (top panels) and anti-GST (bottom panels). The input (10% of the total) is shown at the bottom. (D) Co-IP. GFP-tagged SF1/C370 and SF1/C302 were transiently expressed in HeLa cells. Total cell lysates were RNase A-treated and incubated with Dynabeads Protein G-coupled anti-GFP. Input (I; 0.5% of the total) and bound proteins (B) were separated by 7.5% SDS-PAGE and Western blotted with anti-SF3a120, anti-U2AF65 and anti-GFP as indicated on the right. The migration of protein markers (in kDa) is shown on the left. (E) Multiple sequence alignment of SF1 proteins. SF1 sequences taken from the UniProt database (www.uniprot.org) were aligned with ClustalW2 (80). The region from the zinc knuckle of human SF1 and 40 amino acids C-terminal of this domain is shown. Numbering is given for human SF1. Amino acids identical in more than 50% of the sequences are marked.

Figure 3. Deletion of the SF1 SURP ID negatively affects spliceosome assembly. (A) SF1 immunodepletion. HeLa nuclear extract (NE) was depleted by incubation with Dynabeads Protein G-

coupled anti-SF1. Serial dilutions of untreated (NE), mock-depleted (m) and SF1-depleted extracts (ΔSF1) were separated by 7.5% SDS-PAGE and Western blotted with anti-SF1. The intensities of SF1 isoforms were quantified with the Odyssey Fc imaging system (LI-COR Biosciences) and normalized to 100% untreated nuclear extract. (B) Quantification of recombinant proteins. Increasing amounts of His_6-SF1/C370 and SF1/C302 (2.2, 4.4 and 6.6 pmol) were separated by 10% SDS-PAGE followed by Coomassie blue staining. The migration of protein markers is indicated in kDa on the left. (C) and (D) Spliceosome assembly. Splicing reactions were performed in triplicate at 30°C for 30 min with AdML 3' splice site pre-mRNAs containing a consensus (C) or weak BPS (D) in the absence (-) or presence of untreated (NE), mock- (m) or SF1-depleted (ΔSF1) HeLa nuclear extracts. Reactions containing SF1-depleted extract were supplemented with increasing amounts of His_6-SF1/C370 or SF1/C302 (0.022, 0.22 or 2.2 pmol) as indicated. Reaction products were separated in native 4% polyacrylamide gels, which were dried and exposed to phosphoimager screens. A representative gel is shown on the top. Quantification of the results is shown on the bottom. "% A 3' complex" represents the ratio of A 3' complex to total lane intensities. Values were normalized to mock-treated extract.

Figure 4. Deletion of the SF1 SURP-ID affects the kinetics of spliceosome assembly. (A) and (B) Kinetics of spliceosome assembly. Splicing reactions with AdML 3' splice site pre-mRNAs containing a consensus (A) or weak BPS (B) were performed in triplicate at 30°C for the times indicated in the presence of mock- or SF1-depleted (ΔSF1) nuclear extract supplemented with 2.2 pmol His_6-SF1/C370 or SF1/C302 as shown. Samples were separated in native 4% polyacrylamide gels, which were dried and exposed to phosphoimager screens. Representative gels are shown in the top panels. The results are quantified in the bottom panels. "% A 3' complex" represents the ratio of A 3' complex to total lane intensities. Values were normalized to those of mock-treated extracts at the 60-min time points.

Figure 5. Deletion of the SF1 SURP ID does not reduce U2AF65 binding to the pre-mRNA. (A) and (B) U2AF65 UV cross-linking to AdML 3' splice site substrates. Splicing reactions containing radio-labeled RNA with a consensus (A) or weak (B) BPS, mock or SF1-depleted (ΔSF1) extracts complemented with 2.2 pmole His_6-SF1/C370 or SF1/C302 as indicated were incubated at 30°C for 15 min. Samples were UV cross-linked, RNase A-treated and immunoprecipitated with control IgG or anti-U2AF65, as indicated. RNA-protein complexes were separated by 10% SDS-PAGE. Gels were dried and exposed to phosphoimager screens. The top panels show representative results of triplicate experiments; quantifications are shown in the bottom panels. "% U2AF65-bound RNA" indicates the percentage of the intensity of the cross-linked RNAs normalized to the RNA immunoprecipitated with anti-U2AF65 from mock-treated extract.

Table 1. Proteins co-immunoprecipitated with anti-SF1.

				SF1		IgG	
Name[1]	Symbol[2]	Gene ID[3]	Group[4]	- RNase	+ RNase	- RNase	+ RNase
Sm-B/B'	SNRPB	6628	Sm	51	20	1	2
Sm-E	SNRPE	6635	Sm	48	15	1	2
Sm-G	SNRPG	6637	Sm	7	13	1	1
U1-70K	SNRNP70	6625	U1	15	10	0	0
U1-A	SNRPA	6626	U1	39	29	1	2
SRPK1	SRPK1	6732	U1	0	6	0	0
U2-A'	SNRPA1	6627	U2	27	19	0	1
U2-B"	SNRPB2	6629	U2	12	10	0	0
SF3a120*	SF3A1	10291	U2	41	47	6	5
SF3a60	SF3A3	10946	U2	23	28	1	2
SF3b155	SF3B1	23451	U2	76	66	9	5
SF3b145	SF3B2	10992	U2	47	50	0	3
SF3b130	SF3B3	23450	U2	96	93	6	7
SF3b14a	SF3B14	51639	U2	17	10	1	1
SF3b10	SF3B5	83443	U2	7	9	0	0
CHERP*	CHERP	10523	U2-related	34	15	0	0
SF3b125	DDX42	11325	U2-related	4	19	0	0
SPF31	DNAJC8	22826	U2-related	7	10	1	1
PUF60	PUF60	22827	U2-related	22	26	0	1
SPF45	RBM17	84991	U2-related	10	7	0	0
U2SURP*	SR140	23350	U2-related	41	41	1	0
U2AF65	U2AF2	11338	U2-related	27	34	0	1
FBP11	PRPF40A	55660	A	8	15	0	0
S164	RBM25	58517	A	16	14	0	0
RBM39	RBM39	9584	A	0	8	0	0
SF1	SF1	7536	A	185	193	0	0
SUGP1*	SF4	57794	A	0	6	0	0
SRp40	SFRS5	6430	SR	8	4	0	0
SRp55	SFRS6	6431	SR	8	4	0	0
9G8	SFRS7	6432	SR	22	18	3	1
CYPH	PPIH	10465	U4/U6	7	5	1	0
hPRP31	PRPF31	26121	U4/U6	6	7	0	0
hPRP28	DDX23	9416	U5	1	5	0	0
hPRP6	PRPF6	24148	U5	18	9	1	2
hSNU66	SART1	9092	U4/U6.U5	7	3	0	0
CDC5L	CDC5L	988	Prp19	5	7	0	0
hSmu1	SMU1	55234	B	1	7	0	0
SRm300	SRRM2	23524	B (act)	0	7	0	0
TOE1	TOE1	114034	C2	2	5	0	0
Acinus	ACIN1	22985	EJC/TREX	18	26	0	0
THOC2	THOC2	57187	EJC/TREX	9	4	0	0
ARS2B	ARS2	51593	mRNA	52	45	2	2
CBP80	NCBP1	4686	mRNA	24	24	1	3
ZC3H18	ZC3H18	124245	mRNA	9	11	0	0
ZNF207	ZNF207	7756	misc.	5	2	0	0
BRD2	BRD2	6046		0	18	0	0
CDC2L1	CDK11B	984		3	5	0	0
CKB	CKB	1152		5	3	0	0
EIF3A	EIF3A	8661		12	3	0	0
EIF3C	EIF3C	8663		6	1	0	0
HSPA4L	HSPA4L	22824		13	16	1	2
PCNP	PCNP	57092		8	2	0	0
SAF-B2	SAFB2	9667		7	1	0	0
ZNF598	ZNF598	90850		0	5	0	0

Proteins were separated by 1D PAGE (see Figure S1) and identified by LC-MSMS. The total spectrum count is shown. The experimental set-up is described in Materials and Methods. SF1 or IgG,

proteins precipitated by anti-SF1 antibodies or control IgG coupled to Protein G Sepharose; -RNase or +RNase, HeLa extracts were incubated without (-) or with (+) RNase A prior to IP. The raw data were filtered as explained in the Methods section. The final list of proteins was compared to 244 spliceosomal proteins annotated in table S1 of Hegele et al. (42).

[1] Protein name commonly used in the splicing field. Proteins shown in bold are known SF1 partners; underlined, proteins also found in Y2H screens (see Table 3); *, proteins with SURP domains.

[2] NCBI ENTREZ symbol.

[3] NCBI ENTREZ GeneID.

[4] Operational classification of spliceosomal proteins according to Hegele et al. (42).

Table 2. Libraries used in Y2H screens.

Library name	**Description**	**Number of clones tested**	**Histidine-positive**
CEMC7_RP[1]	Human T cell line CEMC7	53 x 10^6	70
HBMEC_RP[1]	Human bone marrow endothelial cells (cloned cells transformed by SV40)	58 x 10^6	347
HTH_RP[1]	Human thymocytes, CD4+CD8+ double positive cells from children	35 x 10^6	102
HTH_dT[2, 3]	Human thymocytes, CD4+CD8+ double positive cells from children	88 x 10^6	97
PLA_RP[1]	Human placenta	105 x 10^6	211
MANE_RP[1]	Mouse adult neurosphere cells (free-floating clusters of neural stem cells)	65 x 10^6	284
MKI_RP[1]	Total kidney, adult C57BL/J6 mice, age 8-10 weeks	102 x 10^6	144
MPC_RP[1]	Mouse pancreatic cells (beta-cell line, betaTC-tet)	71 x 10^6	143

[1] RP, random-primed

[2] dT, oligo-dT-primed

[3] Performed in the presence of 0.5 mM 3-amino-1,2,4-triazole

Table 3. SF1 partners in Y2H screens.

Gene[1] Human	Gene ID[2] Human	Gene[1] Mouse	Gene ID[2] Mouse	Predicted biological score[3] Human 1	Human 2	Human 3	Human 4	Mouse 5	Mouse 6	Mouse 7	Length[4] Human	ΣSID[5] Human	Length[4] Mouse	ΣSID[5] Mouse	Domain (Interpro)[6]
ABL1	25	Abl1	11350 (var 1)		C			D		C	1130	19-155	1142	61-230	SH3
ABL2_B	27 (var b)	Abl2	11352 (var 1)	D	A	D			D		1182	98-355	1078	62-526	SH3, SH2
ARHGAP39	80728	Arhgap39	223666					D					1109	16-306	WW
ATP6V0D1	9114	Atp6v0d1	11972						D				351	213-351	ATPase (part)
CENPI	2491				D						756	395-633			
CHERP*	10523	Cherp	27967	D				A	B	A	916	1-172	938	12-95	SURP
CIAO1	9391	Ciao1	26371						D				339	1-339	
CLTC	1213	Cltc	67300		D	D		A	D	D	1679	944-1282	1675	931-1113	TPR-like
COG2	22796	Cog2	76332						C				731	548-702	COG2_C
COL8A1	1295				D						744	565-744			C1q
DAZAP1	26528					D					378	1-257			RRM
DDB1	1642	Ddb1	13194				B		C		1140	724-893	1140	721-891	WD40 (partial)
DDX17	10521					D	D				729	172-523			Helicase_ATP-bd
DDX5	1655	Ddx5	13207		B	A		B	D		614	124-429	615	277-429	Helicase_ATP-bd (part)
DNAJA1	3301	Dnaja1	15502						D				397	255-397	DnaJ_C
EFS	10278	Efs	13644					D					560	301-560	CAS_DUF3513
FAM46C	54855						D				391	1-284			
FUBP1	8880	Fubp1	51886					B	C				642	503-520	
FUBP3	8939						D				572	98-572			KH
GIT1	28964	Git1	216963						D					2-546	Ankyrin_rpt
GOLGA2	2801	Golga2	99412						D				1026	287-1026	
GPD2	2820	Gpd2	14571					D		D			727	427-630	
HNRNPK	3190	Hnrnpk	15387					D	D				463	58-463	KH
HNRNPAB	3182			D	A	B	D				332	140-332			RRM
HNRNPD	3184	Hnrnpd	11991			C				D	355	152-355	355	141-355	RRM
HNRNPDL	9987	Hnrnpdl	50926	C	D	A	D	C			420	156-325	420	100-317	RRM
ITCH	83737	Itch	16396						D				864	320-735	HECT
KDM4B	23030				D						1096	503-1096			Znf_PHD, Tudor
KDM6A	7403	Kdm6a	22289					D					1424	63-460	TPR
KIAA0562	9731	A930027E11	230967	C	B	D	A	A	A	A	925	747-825	926	727-806	
KIAA0907	22889	2810403A07Rik	74200	A	B	A	A	A	D	B	614	211-351	612	213-328	KH?
KHDRBS1	10657	Khdrbs1	20218		D			D			443	145-319	443	99-390	KH
KHDRBS3	10656			C	D						346	20-198			KH
KHSRP	8570	Khsrp	16549					D					748	127-616	KH
KLHL32	114792	Klhl32	212390							D			587	39-226	BTB/POZ-like, BACK
KPNA1	3836	Kpna1	16646		A	B	A	A	C	A	538	252-427	538	252-443	Armadillo
KPNA2	3838	Kpna2	16647	A	A	A	A	A	A	A	529	255-492	529	245-449	Armadillo
KPNA3	3839	Kpna3	16648		B					D	521	242-446	521	80-493	Armadillo
KPNA4	3840	Kpna4	16649	B			D	D	B	B	521	111-404			Armadillo

KPNA5	3841				D						539	249-539			Armadillo
KPNA6	23633	Kpna6	16650	B	A	D	A	A	B		536	268-406	536	249-437	Armadillo
MAD2L2	10459				B						203	1-211			HORMA_DNA-bd
MBTD1	54799					D					628	18-367			Mbt
METTL2A	339175	Mettl2	52686					D	D				389	161-389	Methyltransferase
KMT2E	55904	Kmt2e	69188						D				1868	312-487	SET
MSI2	124540				D						328	83-328			RRM
NOC3L	64318	Noc3l	57753					D					807	253-529	ARM-type
NUP155	9631				D						1391	981-1338			
P4HA1	5033	P4ha1	18451							D			534	2-253	TPR
PHF19	26147				D						580	22-346			Znf_FYVE_PHD
PIK3C2B	5287					D					1634	1353-1549			Phox
PKD1	5310	Pkd1	18763					D		A			4293	1112-1227	PKD/Chitinase
PLAGL2	5326					D					496	174-425			Znf_C2H2-like
PRMT2_1	3275			D	A	B	A				433	35-155			SH3
PRPF40A	55660					C					930	96-930			WW, FF
QKI	9444	Qk	19371					C					341	10-205	KH
RALY	22913	Raly	19383							D			296	1-296	RRM
RBM5	10181				D						815	273-527			RRM (partial)
RBM17	84991			D	B	D					401	236-401			RRM
RBMS1	5937						C				406	53-234			RRM
RBMS2	5939				D						407	47-270			RRM
RBPMS	11030				D						196	1-143			RRM
RIOK2	55781	Riok2	67045					D					547	175-369	Kinase-like (part)
SCP2	6342				D						547	1-478			Thiolase-like
SF3A1*	10291	Sf3a1	67465				B	D		B	793	12-137	791	43-194	SURP
SFSWAP*	6433	Sfrs8	231769					D	C	D			945	448-502	SURP (part)
TGS1	96764	Tgs1	116940		D			D			853	63-281	853	1-330	
TIA1	7072					D					386	74-338			RRM
TIAL1	7073	Tial1	21843	D	D				C	D	375	29-300	340	1-340	RRM
TPR	7175				D						2363	1108-1323			TRP_MLP1 (part)
TTC28	23331	Ttc28	209683					D					2450	57-661	TPR
U2AF2	11338	U2af2	22185		A	A	A	A	A	C	475	360-461	475	291-393	RRM
UHMK1	127933	Kist	16589	D		B	B	B	D		419	93-419	419	95-419	RRM
VAV2	7410	Vav2	22325		D				D		839	680-839	868	549-868	SH3
TBC1D31	93594	Tbc1d31	210544					D					996	409-819	Rab-GTPase-TBC
WWP2	11060	Wwp2	66894							D			864	403-592	WW
ZPF36	7538					B					332	61-332			Znf_CCCH
ZFP36L1	677	Zfp36l1	12192				D	C	D		338	72-224	338	96-193	Znf_CCCH
ZFP36L2	678	Zfp36l2	12193				D		D		494	95-216	484	45-284	Znf_CCCH
ZMYND8	23613				D		B				1188	847-1065			Znf_MYND

Proteins interacting with SF1 in Y2H screens as detailed in Material and Methods in different human and mouse cell types are listed.

[1] NCBI ENTREZ symbols. The first column lists the gene symbols for proteins found in screens with human cells as well as the symbols for orthologs of proteins found only in screens with mouse cells. The third column lists the gene symbols only for those proteins found to interact with SF1 in mouse cells. Bold, known SF1 interactors; underlined, proteins also found by co-IP/MS; *, proteins with SURP domains; light green, spliceosome-associated proteins (42); light blue, other proteins with known roles in splicing.

[2] NCBI ENTREZ GeneID.

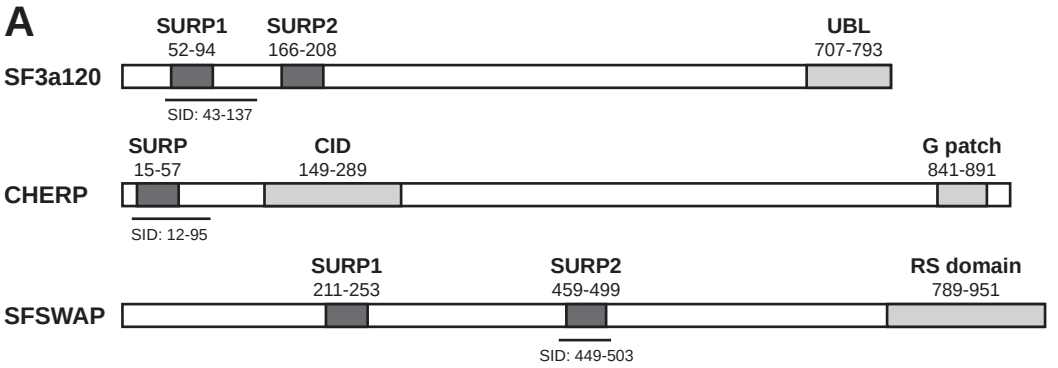
[3] Predicted Biological Score (PBS) for SF1 Y2H partners in the following cell types (see Table 2): 1, CEMC7; 2, HBMEC; 3, HTH; 4, PLA; 5, MANE; 6, MKI; 7, MPC. The qualifiers A, B, C and D indicate very high, high, good and moderate confidence in the interactions, respectively (see 27).

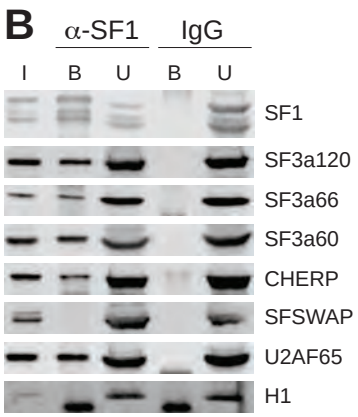
[4] Length of human and mouse proteins.

[5] Shortest selected interaction domain (ΣSID) deduced from cDNAs found in the Y2H screens.

[6] InterPro domains comprised in the ΣSID.

Figure 1


A
SF3a120
SURP1
52-94
SURP2
166-208
UBL
707-793
SID: 43-137
CHERP
SURP
15-57
CID
149-289
G patch
841-891
SID: 12-95
SFSWAP
SURP1
211-253
SURP2
459-499
RS domain
789-951
SID: 449-503


B
α-SF1
IgG
I
B
U
B
U
SF1
SF3a120
SF3a66
SF3a60
CHERP
SFSWAP
U2AF65
H1

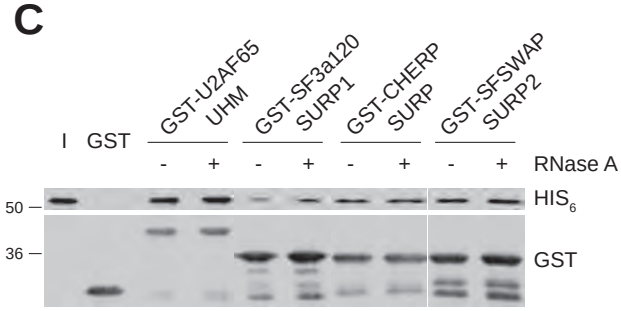
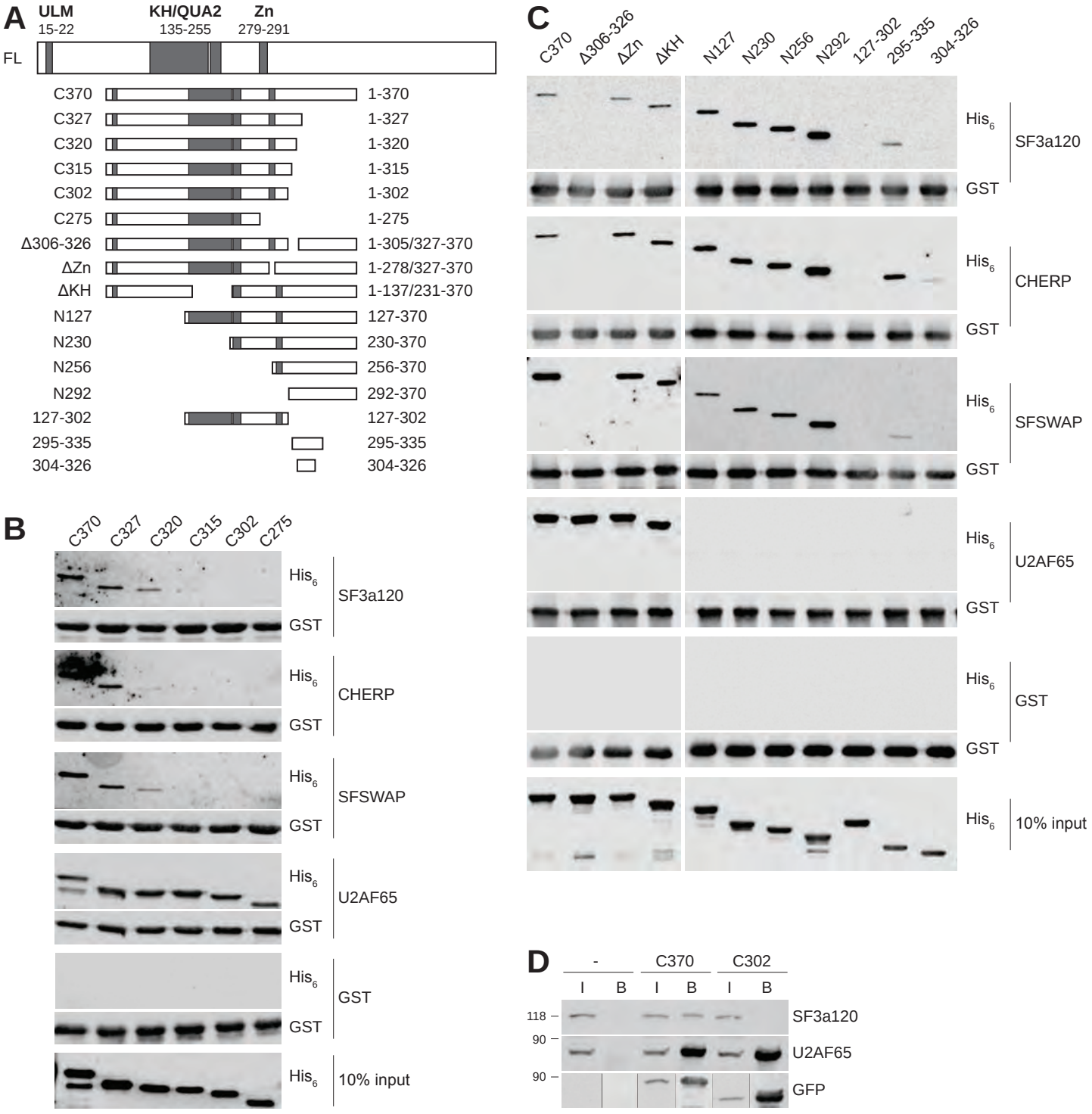

C
GST-U2AF65 UHM
GST-SF3a120 SURP1
GST-CHERP SURP
GST-SFSWAP SURP2
I
GST
- + - + - + - +
RNase A
50
36
HIS6
GST

Figure 2


A
ULM
15-22
KH/QUA2
135-255
Zn
279-291
FL
C370 1-370
C327 1-327
C320 1-320
C315 1-315
C302 1-302
C275 1-275
Δ306-326 1-305/327-370
ΔZn 1-278/327-370
ΔKH 1-137/231-370
N127 127-370
N230 230-370
N256 256-370
N292 292-370
127-302 127-302
295-335 295-335
304-326 304-326
B
C370
C327
C320
C315
C302
C275
His6
GST
SF3a120
CHERP
SFSWAP
U2AF65
GST
10% input
C
C370
Δ306-326
ΔZn
ΔKH
N127
N230
N256
N292
127-302
295-335
304-326
His6
GST
SF3a120
CHERP
SFSWAP
U2AF65
GST
10% input
D
-
C370
C302
I
B
I
B
I
B
118
90
90
SF3a120
U2AF65
GFP

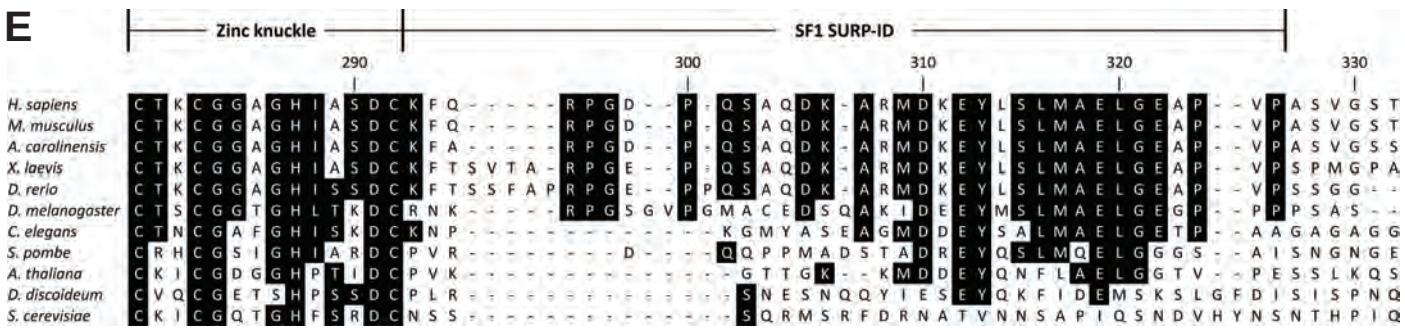

E
Zinc knuckle
SF1 SURP-ID
290
300
310
320
330
H. sapiens
M. musculus
A. carolinensis
X. laevis
D. rerio
D. melanogaster
C. elegans
S. pombe
A. thaliana
D. discoideum
S. cerevisiae

Figure 3

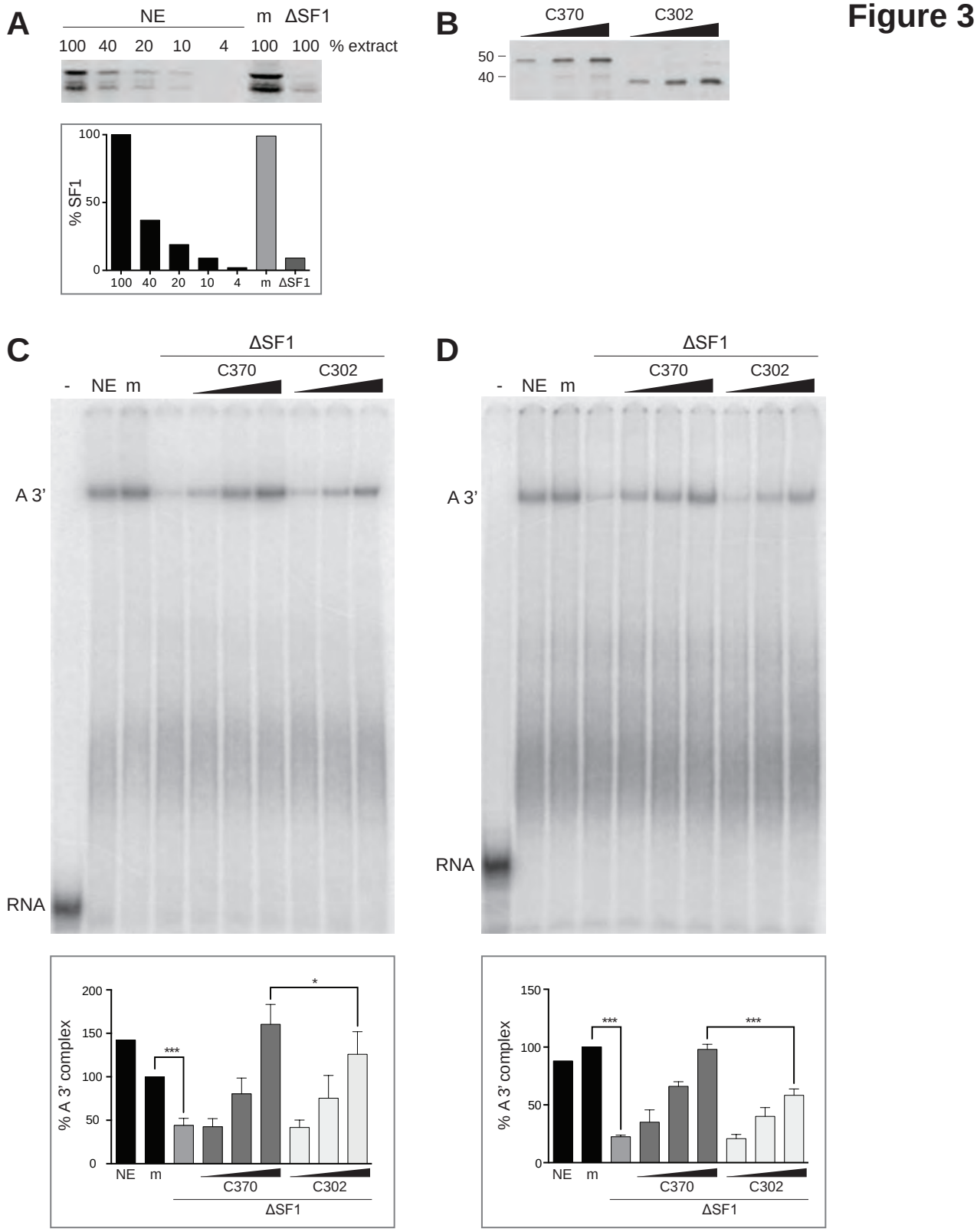
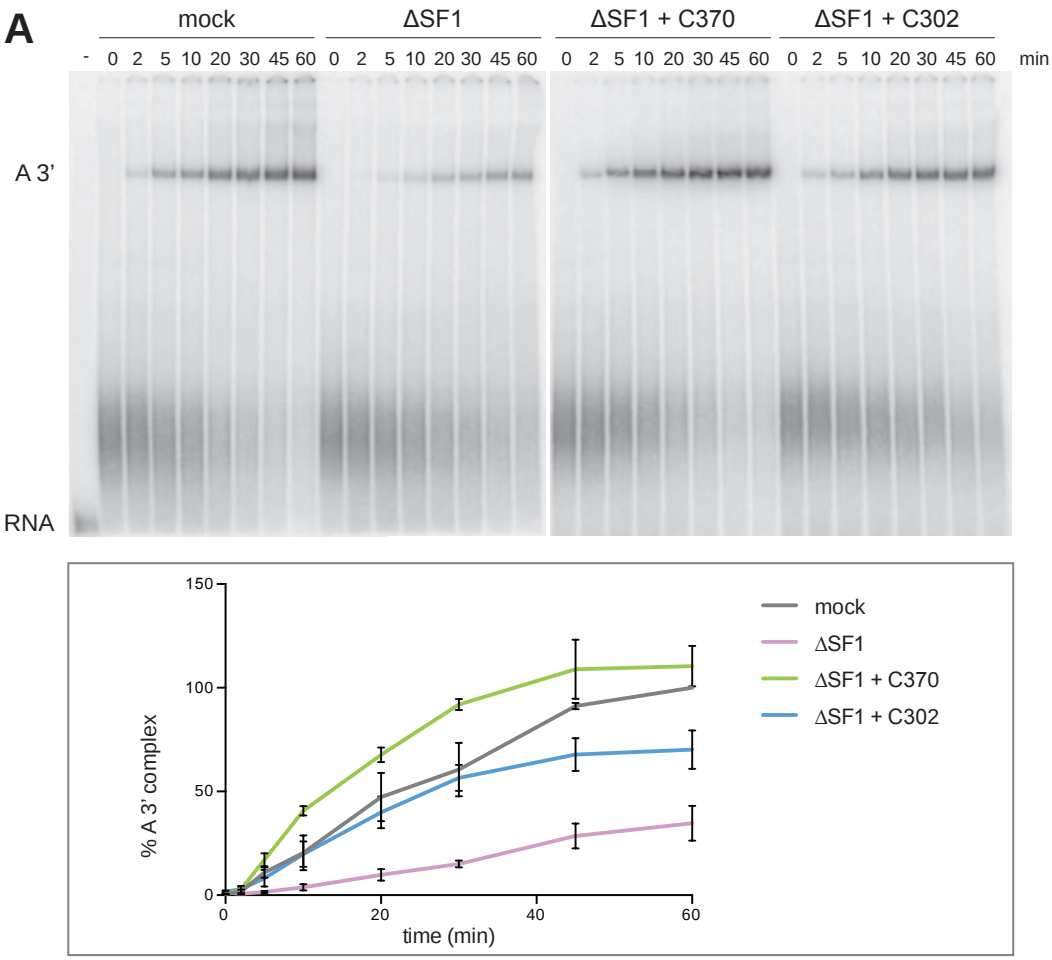

A
NE
m ΔSF1
100 40 20 10 4 100 100 % extract
% SF1
100 40 20 10 4 m ΔSF1
B
C370
C302
50
40
C
ΔSF1
C370
C302
- NE m
A 3'
RNA
% A 3' complex
NE m
C370
C302
ΔSF1
D
ΔSF1
C370
C302
- NE m
A 3'
RNA
% A 3' complex
NE m
C370
C302
ΔSF1

Figure 4


A
mock
ΔSF1
ΔSF1 + C370
ΔSF1 + C302
- 0 2 5 10 20 30 45 60 0 2 5 10 20 30 45 60 0 2 5 10 20 30 45 60 0 2 5 10 20 30 45 60 min
A 3'
RNA
% A 3' complex
time (min)
mock
ΔSF1
ΔSF1 + C370
ΔSF1 + C302

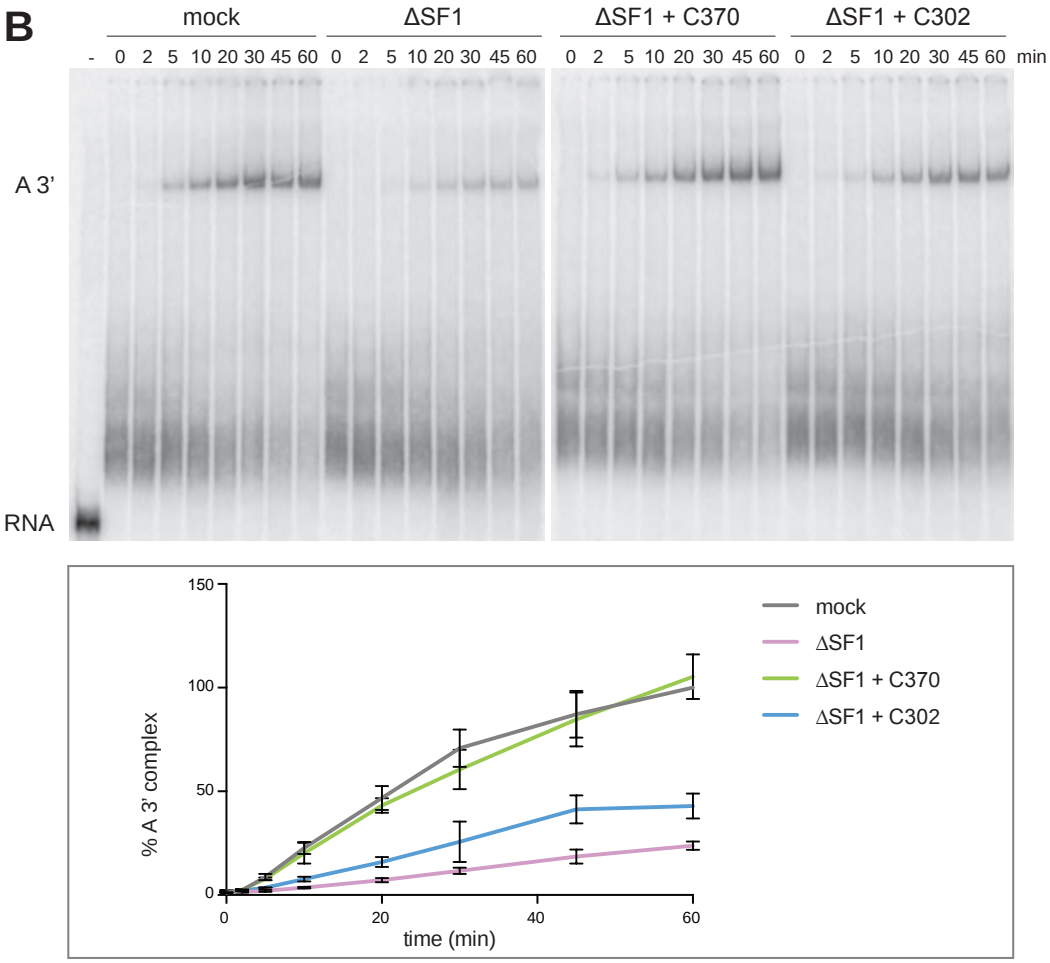
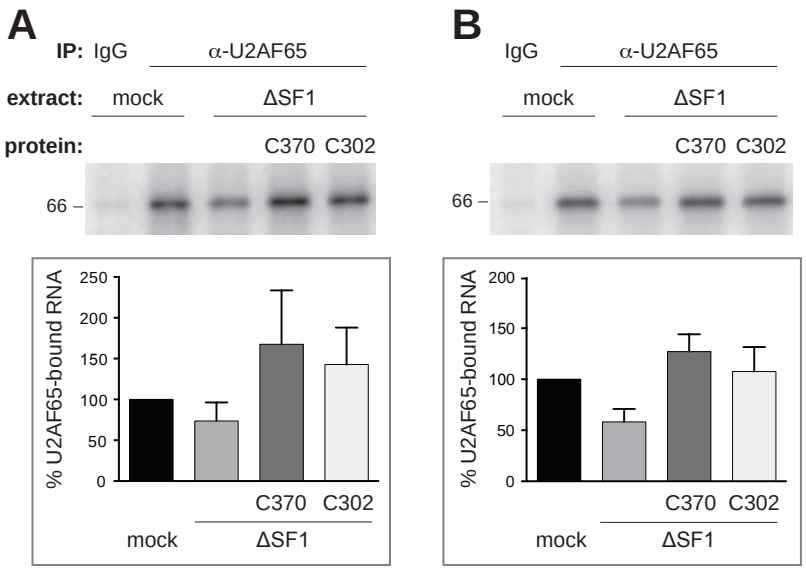

B
mock
ΔSF1
ΔSF1 + C370
ΔSF1 + C302
- 0 2 5 10 20 30 45 60 0 2 5 10 20 30 45 60 0 2 5 10 20 30 45 60 0 2 5 10 20 30 45 60 min
A 3'
RNA
% A 3' complex
time (min)
mock
ΔSF1
ΔSF1 + C370
ΔSF1 + C302

Figure 5


A
IP: IgG
α-U2AF65
extract: mock
ΔSF1
protein:
C370 C302
66 –
% U2AF65-bound RNA
250
200
150
100
50
0
C370 C302
mock
ΔSF1
B
IgG
α-U2AF65
mock
ΔSF1
C370 C302
66 –
% U2AF65-bound RNA
200
150
100
50
0
C370 C302
mock
ΔSF1

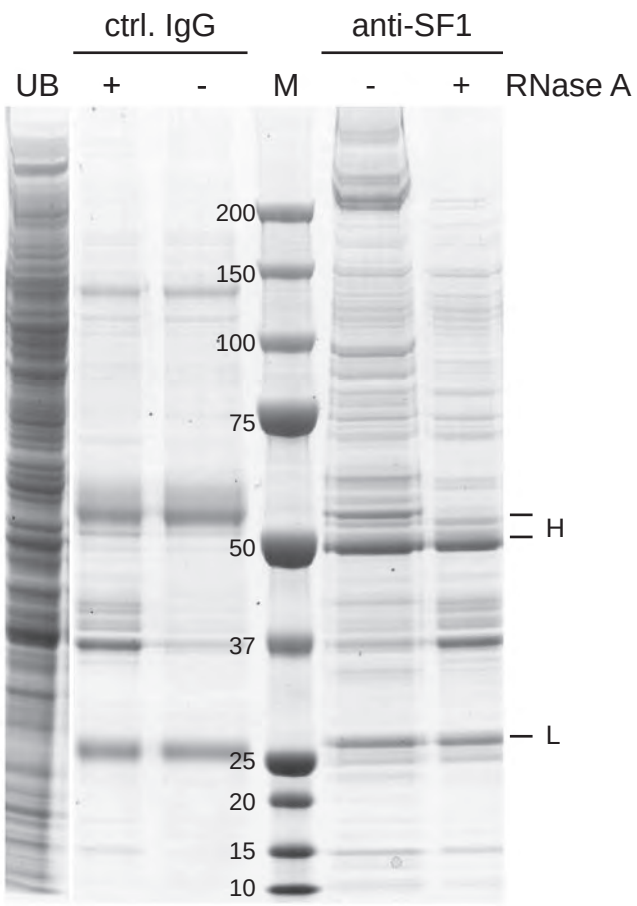


Figure S1. Co-immunoprecipitation of HeLa cell proteins with anti-SF1 antibodies.
A HeLa cell lysate was incubated in the absence (-) or presence (+) of RNase A followed by precipitation of proteins with control (ctrl.) IgG or anti-SF1 antibodies coupled to Protein G Sepharose, as indicated on top of the figure. Bound proteins were separated by 1D PAGE and stained with Coomassie blue. Unbound (UB) proteins are shown in the first lane. The migration of marker proteins (M; in kDa) is indicated. Proteins in the + RNase samples were subjected to mass spectrometry.

2.3. Additional experiments

2.3.1. Material and methods

2.3.1.1. Cell culture

HeLa cells were grown in a humid chamber at 37°C with 5% CO_2 in DMEM (Sigma) supplemented with 10% FBS (fetal bovine serum, Sigma), 1% penicillin-streptomycin (Gibco) and 1% L-glutamine (Gibco).

2.3.1.2. Cloning procedures of DNA constructs used for the co-expression

DNA sequences of SURP1 and SURP2 of SF3a120 (aa 38-113 and aa 134-217, respectively), SF3a60 (aa 1-107) and SF1 (aa 1-320) were amplified from 293T cell cDNA with the Expand High Fidelity Kit (Roche) according to the manufacturer's instructions. The following cycling conditions were used: 94°C for 2 min, 35 cycles at 94°C for 15 sec, 59°C for 30 sec, 72°C for 2 min and finally 72°C for 10 min. For the SURP sequences, the following cycling condition was applied: 94°C for 2 min, 35 cycles at 94°C for 15 sec, 59°C for 30 sec, 72°C for 30 sec and finally 72°C for 3 min. The primer sequences with restriction sites are summarized in table 1. The PCR products were analyzed in agarose gels containing 0.5 µg/ml ethidium bromide (Sigma) and gel-purified. SF1 and SF3a60 PCR products were cloned into the BamHI/EcoRI or HindIII sites of a modified pACYCDuet-1 vector, containing His_6 and the ORF of MBP (Maltose-binding protein) at the N terminus. SF3a120 SURP1 and 2 were cloned into the EcoRI/HindIII sites of a modified pET vector, containing a GST and His_6 tag at the N terminus. Vectors were kindly provided by Stéphane Thore (INSERM, Bordeaux, France). Constructs were sequenced with T7 and T7 terminator sequencing primers (Microsynth, Switzerland).

Primer name	Sequence 5' ⇒ 3'	Orientation	Restriction site
SF1_C4_F	CGCGGATCC ATGGCGACCGGAGCGAACGC	sense	BamHI
SF1_C4_R	CCGGAATTC CTAGAGCAGTCCTTGTCTTT	antisense	EcoRI
3a60_F	CGCGGATCC ATGGAGACAATACTGGAGCA	sense	BamHI
3a60_R	CCCAAGCTT TTA TGACATTGGCACACAGAT	antisense	HindIII
SF3_S1_F	CCGGAATTC CCAGTTGTGGGGATTATTTA	sense	EcoRI
SF3_S1_R	CCCAAGCTT TTA GATGGCGGCGGACGGCTCCT	antisense	HindIII
SF3_S2_F	CCGGAATTC GCCCAAGTAATCCAAGAGAC	sense	EcoRI
SF3_S2_R	CCCAAGCTT TTA CTTGGTGTACTGTTCCACTA	antisense	HindIII

Table 1. Primers for the cloning of DNA constructs used for co-expression experiments.

Sequences listed in table 1 include the restriction site, which is underlined, the stop codon, when present, and the complementary region.

2.3.1.3. Expression and co-expression of recombinant proteins

Plasmids encoding GST-tagged SURP domains, His_6-MBP-tagged SF1 protein and SF3a60 were transformed into the *E. coli* strain BL21 by heat shock. Co-expression of SF1 and SF3a60 with SURP domains was performed as previously described (Kuwasako et al, 2006). Proteins were expressed for 4 h at 37°C after addition of IPTG to a final concentration of 1 mM. Cells were harvested by centrifugation at 3,800 × g for 10 min.

Cells expressing recombinant proteins were lysed in PBS supplemented with complete protease inhibitors (Roche) by sonication on ice and supplemented with Triton X-100 to a final concentration of 1%. Lysates were centrifuged at 9,500 rpm for 10 min at 4°C in a microfuge to separate soluble (supernatant) from non-soluble proteins (pellet). Soluble proteins were purified with a 50% suspension of glutathione agarose (Sigma) equilibrated in PBS. Unbound proteins were removed by washing three times with PBS supplemented with 0.05% NP-40. Bound proteins were eluted with SDS sample buffer and assayed by SDS-PAGE and Coomassie staining.

2.3.1.4. Psoralen cross-linking

Psoralen cross-linking experiments were performed in 20-µl reactions in the presence of 26.6 pmol [α-^{32}P]-UTP-labeled RNA, 25% HeLa nuclear extract and 20 ng/µl of the psoralen derivative AMT (aminomethyltrioxsalen, Sigma). The reactions were incubated for 15 min at 30°C and irradiated on ice for 10 min at 365 nm. A 2-mm glass plate was placed between the UV source and the samples to block shortwave irradiation. Cross-linked RNAs were analyzed in a 4% denaturing polyacrylamide gel. The identification of cross-linked products was performed as described (Sauliere et al, 2006).

2.3.2. Results

2.3.2.1. SF1 does not interact with SF3a120-SURP2

In the manuscript (section 2.2) we have shown that SF1 binds to the SF3a120 subunit of SF3a *in vitro* and in a mammalian nuclear extract. As described in the main introduction (section 1.5.1.3), SF3a120 has two SURP modules in the N-terminal part and the results presented in the manuscript have demonstrated that the first SURP domain (SURP1) is sufficient to mediate the interaction with SF1. On the other hand, a region encompassing the second SURP domain (SURP2) has been shown previously to bind to SF3a60 (Huang et al, 2011; Krämer et al, 1995). Although the two SURP motifs of SF3a120 are highly conserved and characterized by a similar tertiary structure, they have different properties (Kuwasako et al, 2006). Recombinant SURP1 is soluble, whereas SURP2 is soluble only upon co-expression with SF3a60 as ligand, but not when it is expressed alone. Since binding of SF1 to SF3a120-SURP2 could not be tested, because the recombinant protein was insoluble, we decided to indirectly investigate whether the two proteins interact with each other by analyzing SURP2 solubility upon co-expression with SF1. We speculated that the recombinant SURP2 domain would become soluble if it interacts with SF1, as it was shown previously for SF3a60 (Kuwasako et al, 2006).

SF1 and SF3a60 coding sequences were cloned into a pACYAC vector, encoding an N-terminal His_6-MBP tag, whereas SURP domains were cloned into a pET vector, encoding an N-terminal GST tag. The vectors used for the co-expression have a compatible origin of replication and different antibiotic resistance genes. First, GST-tagged SURP1 (38 kDa) and SURP2 (39 kDa), His_6-MBP-tagged SF1 (80 kDa) and SF3a60 (57 kDa) were individually expressed in *E. coli* and soluble (supernatant) and non-soluble (pellet) proteins were obtained after centrifugation of the total lysates. SDS-PAGE analysis and Coomassie staining showed that all recombinant proteins are present not only in the pellet but also in the supernatant, except for GST-tagged SURP2 which is only found in the non-soluble pellet fraction, consistent with previous reports (figure 13A) (Kuwasako et al, 2006). Next, GST-tagged SURP1 and SURP2 were co-expressed with either His_6-MBP-tagged SF1 or SF3a60 and soluble and non-soluble proteins were analyzed as described above (figure 13B). The results show that co-expression of GST-tagged SURP1 with SF1 and SF3a60 does not affect the solubility of the protein (figure 13B). In contrast, GST-tagged SURP2 is partially soluble

upon co-expression with His_6-MBP-tagged SF3a60, but not with His_6-MBP-tagged SF1 (cf. figures 13A and B).

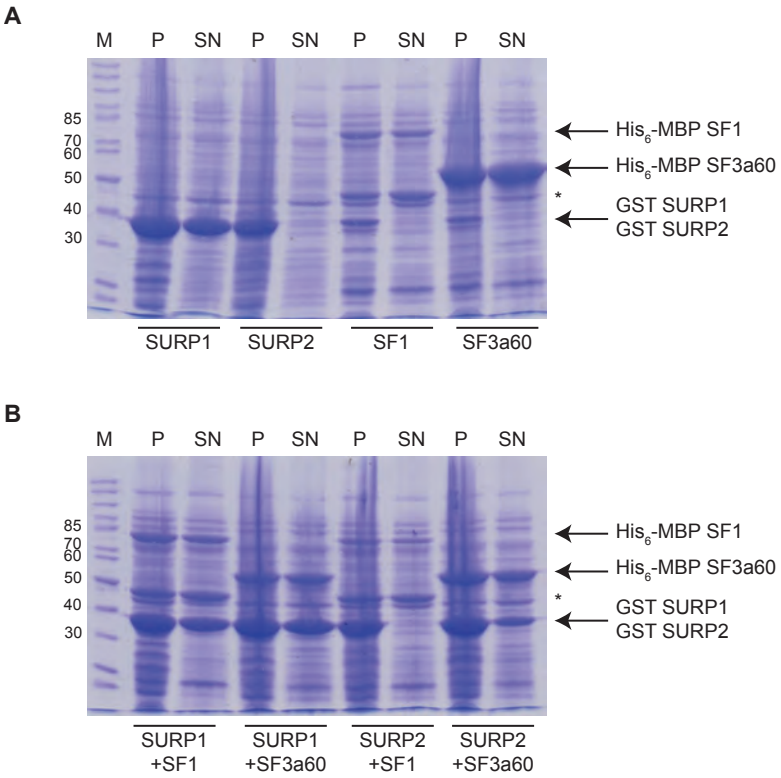


Figure 13. Bacterial co-expression of GST-tagged SF3a120 SURP domains with His_6-MBP-tagged SF1 and SF3a60.

(A) GST-tagged SURP1 and 2, His_6-MBP-tagged SF1 and SF3a60 were expressed in *E. coli*. Soluble (SN) and non-soluble (P) proteins were fractionated and equal amounts were analyzed by 10% SDS-PAGE, followed by Coomassie staining. Recombinant over-expressed proteins are indicated on the right and below the gel. The asterisk (*) indicates the His_6-MBP tag detected in the soluble and non-soluble fractions upon over-expression of SF1. The migration of protein markers (M) is indicated in kDa on the left. (B) GST-tagged SURP domains were co-expressed with either His_6-MBP-tagged SF3a60 or SF1. Solubility of recombinant proteins was assayed by SDS-PAGE and Coomassie staining as in (A). Recombinant over-expressed proteins are indicated on the right and combinations of co-expressed proteins are indicated below the gel. The asterisk (*) indicates the His_6-MBP tag detected in the soluble and non-soluble fractions upon over-expression of SF1 in combination with other proteins. The migration of protein markers (M) is indicated in kDa on the left.

The soluble fractions analyzed by SDS-PAGE in figures 13A and B were further processed by GST agarose purification and the purified proteins were analyzed by SDS-PAGE and

Coomassie staining. Expression of individual recombinant proteins in *E. coli* only yields purified GST-tagged SURP1, since GST-tagged SURP2 alone is not soluble and SF1 and SF3a60 do not have GST tags (figure 14). In the co-expressed samples SF1 co-purifies with SF3a120-SURP1, which further confirms the results presented in the manuscript (figure 14). Surprisingly, SURP1 weakly binds also to SF3a60 (figure 14), although the SURP1 DNA sequences cloned into the pET vector and corresponding to amino acid 38-113, were shown to not be involved in the interaction (Huang et al, 2011). Possible explanations about the binding of SF3a120 SURP1 and SF3a60 binding will be discussed in section 4. Our data also show that GST-tagged SURP2 solubility was affected by co-expression with SF3a60, but not with SF1 (figure 13B). SF3a60, indeed, co-purifies with SF3a120-SURP2 in a 1:1 stoichiometry, as expected, whereas no protein was purified when SURP2 and SF1 were co-expressed (figure 14) (Huang et al, 2011; Kuwasako et al, 2006; Nesic & Krämer, 2001).

In conclusion, our experiments have confirmed that recombinant SF3a120-SURP2 is not soluble alone and that its solubility increases upon co-expression with its ligand SF3a60. In contrast, co-expression of SF3a120-SURP2 with SF1 does not influence SURP2 solubility, suggesting that the recombinant SURP2 module does not interact with SF1.

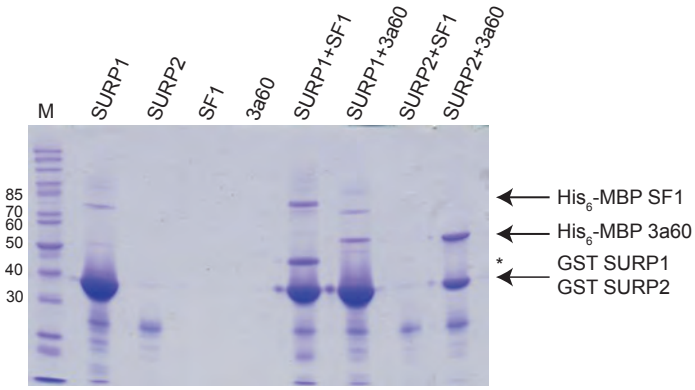


Figure 14. GST-pull down assay of SF3a120 SURP1 and SURP2.

GST-purification was performed on the soluble fractions analyzed in figures 13A and 13B. Bound proteins were analyzed by 10% SDS-PAGE and Coomassie staining. Purified proteins are indicated on the right of the gel. The asterisk (*) indicates the His_6-MBP tag detected in the soluble and non-soluble fractions upon over-expression of SF1 in combination with other proteins. The migration of protein markers (M) is indicated in kDa on the left.

2.3.2.2. SF1 deletion affects U2 snRNA base pairing to the RNA

We have shown that SF1 interacts with U2 snRNP-associated proteins via SURP domains and that depletion of SF1 leads to reduced A complex formation during spliceosome assembly. Since the A complex forms upon binding of the U2 snRNP to the BPS, we decided to examine whether the U2 snRNA interaction with the pre-mRNA is compromised by SF1 depletion with a psoralen cross-linking experiment. Psoralen (AMT) is a chemical compound that crosslinks double-stranded nucleic acid after exposure to UV light. AMT was added to splicing reactions containing untreated HeLa cell nuclear extract and *in vitro* transcribed AdML 3' RNA to cross-link U snRNAs to the pre-mRNA (Hannon et al, 1992; Hausner et al, 1990). After incubation of the splicing reactions for 15 min at 30°C and exposure to UV light, RNAs were extracted and resolved in denaturing polyacrylamide gels (figure 15). The results show the presence of two main bands, one sharp and the other one more diffused, characterized by a slower migration compared to free RNA (figure 15, lane 5). Their appearance depends on the presence of nuclear extract (NE), AMT and UV light (figure 15, lanes 1, 2, 3 and 4), suggesting that the bands are indeed cross-linked products. To identify the snRNA(s) cross-linked to the pre-mRNA, splicing reactions were incubated with DNA oligonucleotides complementary to different regions of U2 snRNA and RNase H, which digests RNA-DNA duplexes. Treatment with an oligonucleotide complementary to the U2 snRNA region that base pairs with the BPS (nt 28-42 of U2 snRNA) leads to the disappearance of the two bands (figure 15, lane 7), suggesting that they both contain U2 snRNA cross-linked to the 3' RNA. Incubation with an oligonucleotide complementary to a U2 snRNA region that does not base pair with the intron (5' end of U2 snRNA, nt 1-15), reduces but does not completely affect the formation of the lower cross-linked product (figure 15, lane 8), demonstrating that the band corresponds to U2 snRNA cross-linked to the 3' RNA. In addition, the higher band is almost completely abolished, which suggests a cross-linking event in the 5' portion of U2 snRNA (figure 15, lane 8). As discussed in the introduction (section 1.4.4.), the 5' end of U2 snRNA base pairs to U6 snRNA in the catalytic B^{act} complex and digestion of this region compromises the interaction between the two snRNPs (Hausner et al, 1990). Therefore, our results suggest that the higher cross-linked band corresponds to U2 snRNA bound to the intron and U6 snRNA. Addition of only RNase H to the splicing reactions did not compromise the formation of cross-linked products (figure 15, lane 6).

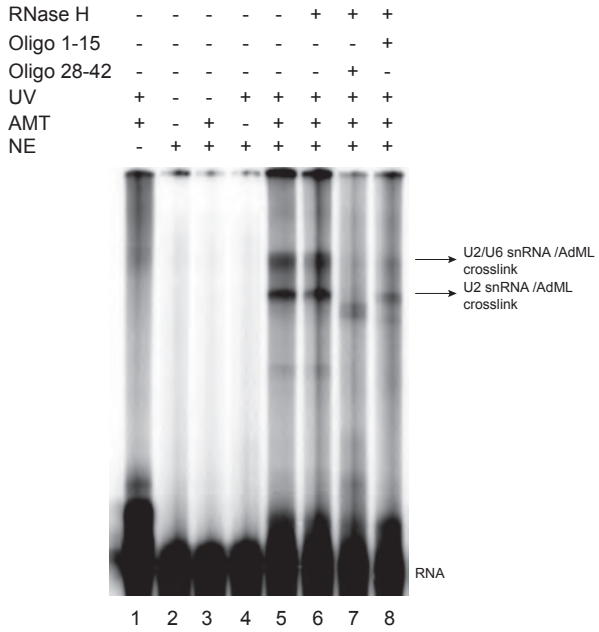


Figure 15. Psoralen cross-linking of U2 snRNA to the AdML 3' RNA.

Psoralen cross-linking was carried out with (lanes 2 to 8) or without (lane 1) HeLa cell nuclear extract (NE) in the presence of ^{32}P-labeled AdML 3' RNA. AMT was added to the reaction and the mixture was exposed to UV light (as indicated above the figure). Identification of cross-linked products was performed by incubating purified RNAs with RNAse H and DNA oligonucleotides complementary to different regions in the U2 snRNA (lanes 7 and 8). The mock control was incubated only with RNase H (lane 6). RNAs were purified and loaded onto denaturing polyacrylamide gels. Free RNA and cross-linked products are indicated on the right.

We next analyzed the interaction of U2 snRNA with 3' RNA in an SF1-depleted nuclear extract. Splicing reactions were performed by incubating AMT with untreated (NE), mock (m) and SF1-depleted (ΔSF1 NE) nuclear extract supplemented with recombinant SF1 C370 or C302. After exposure to UV light, RNAs were purified and resolved in denaturing polyacrylamide gels. U2 snRNA cross-links to 3' RNA in the wild type and mock-depleted nuclear extract (figures 16A and B, lanes 2 and 3), whereas SF1 depletion almost completely abolishes U2 snRNA base pairing with the 3' RNA (figures 16A and B, lane 4). The U2 snRNA interaction with the BPS is restored after addition to the depleted extract of SF1 C370 and C302 (figures 16A and B, lanes 5 and 6). Similar results were obtained when the reactions were carried out with RNA containing either a consensus or weak BPS (figures 16A and B). Although we did not perform any quantification, we observed a slight decrease

in the interaction of U2 snRNA with 3' RNA when the SF1-depleted extract was supplemented with recombinant SF1 C302, lacking the SURP-ID. Similar to the spliceosome assembly experiments presented in the manuscript, this effect is more evident with the 3' RNA containing a suboptimal BPS (figure 16B, lane 6).

In conclusion, our results show that depletion of SF1 strongly affects A complex formation by compromising U2 snRNA base pairing with the BPS. This effect is specific for SF1 and the interaction of U2 snRNA with the RNA is restored to control levels after addition of recombinant SF1 proteins that are able to interact with SURP domain-containing proteins.

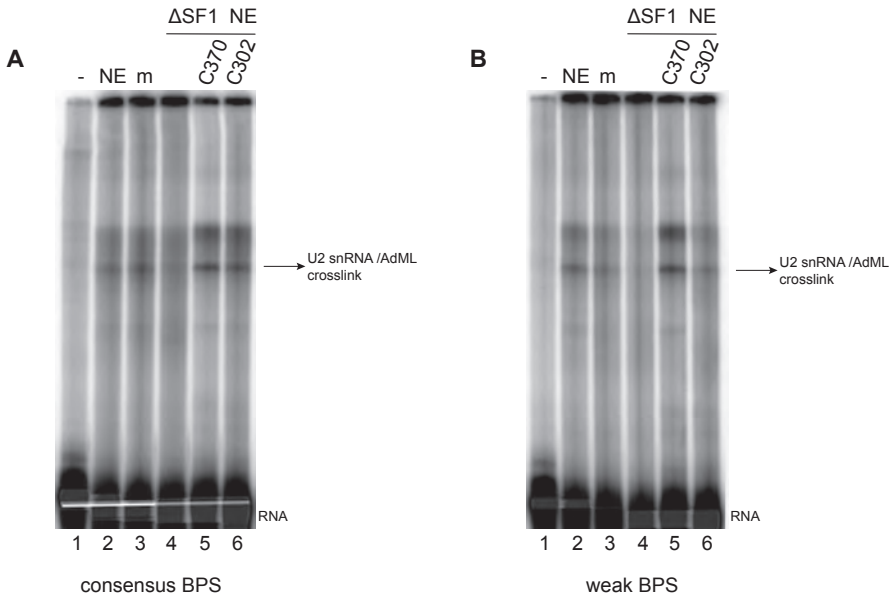


Figure 16. SF1 depletion affects U2 snRNA cross-linking to the AdML 3' RNA.

Psoralen cross-linking was performed with untreated (NE, lane 2), mock-treated (m, lane 3) and SF1-depleted nuclear extract (ΔSF1 NE, lane 4 to 6) in the presence of ^{32}P-labeled AdML 3' RNA. SF1 C370 (lane 5) and C302 (lane 6) were added to reactions containing the depleted extract. RNAs were purified and loaded onto denaturing polyacrylamide gels. Free RNA and cross-linked products are indicated on the right. Psoralen cross-linking was performed on pre-mRNA with a consensus (A) or weak (B) BPS, as indicated below the figure.

2.3.3. Conclusion

The results presented in this chapter support the findings reported in the manuscript regarding the interaction of SF1 with SF3a120. Our data indirectly show that SF1 interacts with SF3a120 only via SURP1 and not via the SURP2 domain. In addition, we have demonstrated that, during the spliceosome assembly, depletion of SF1 leads to a reduced U2 snRNP interaction with the pre-mRNA, with consequently low levels of pre-spliceosomal A complex. Our results, therefore, confirm the hypothesis that SF1 binds to SF3a120 in the earliest step of the spliceosome assembly, presumably in the E complex, and facilitates the interaction of U2 snRNA with the BPS.

Results II: Analysis of SF1 isoforms

3. Results II: Analysis of SF1 isoforms

3.1. Introduction

As described above, the SF1 pre-mRNA is extensively alternatively spliced, due to the use of exon skipping, intron inclusion, as well as duplicated 5' and 3' ss. The AS events of SF1 are mainly found in the 3' half of the pre-mRNA and give rise to numerous SF1 isoforms. In mammalian cells, several SF1 isoforms have been identified, which vary in the amino acid sequences and length of the C-terminal halves (Arning et al, 1996; Krämer et al, 1998). So far, a single AS event has been described in the 5' half of SF1 pre-mRNA, which gives rise to a transcript that lacks exon 3 and is very likely degraded by the NMD pathway due a PTC in exon 5 (Choleza and Krämer, unpublished results).

Alternative SF1 transcripts lacking the first exons have been reported in the ExonMine database (Mollet et al, 2010). ExonMine derives from an algorithm that aligns ESTs and mRNAs to genomic sequences to detect un-annotated splice variants. The ESTs are characterized by a short length (maximum 800 bp) and they result from the sequencing of cloned cDNA libraries prepared from various organisms, tissues and cell lines. Since they derive from cDNAs, ESTs represent portions of expressed genes. Three SF1 transcripts containing either exon 3A or exon 3B have been reported in the ExonMine database (figure 17B). Exons 3A and 3B are alternative exons located in intron 3 at the genomic level and are therefore usually spliced out (figure 17A). Transcript i contains exons 3A, 4, 5, 6 and 7, whereas transcript ii contains a retained intron between exons 3A and 3B (named intron 3A_3B), exons 3B, 4, 5 and 6 (figure 17B). Due to the short length of ESTs, the complete sequences of the corresponding cDNAs have not been annotated and additional exons might be present on each side of the transcripts. However, it was shown that, in humans, 45% of un-annotated exons reported in ExonMine, such as exon 3A, are first exons (Mollet et al, 2010). Therefore, it is likely that transcript i is transcribed by an AP. In addition, a third transcript, named transcript iii, has been reported in the ExonMine database and contains the constitutive exons 1, 2, 3 and 4 and the alternative exon 3B (figure 17B).

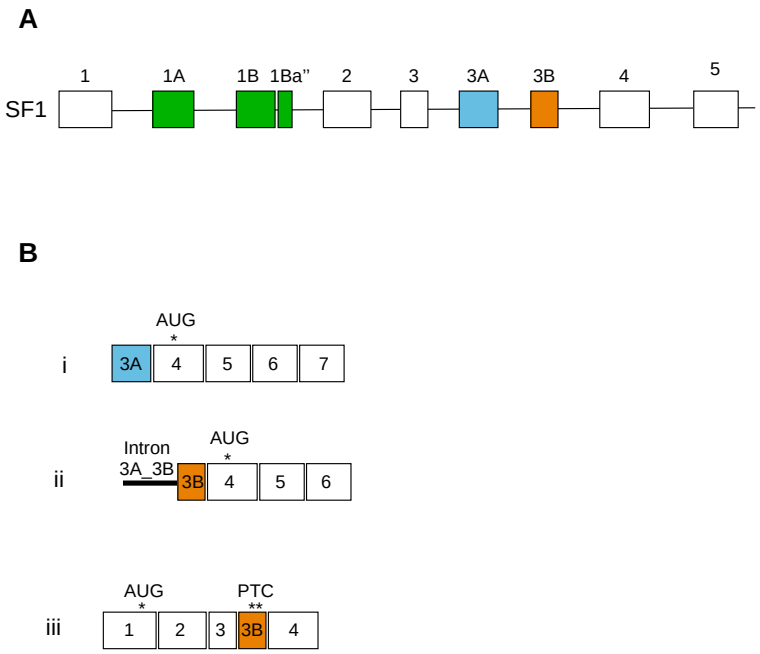


Figure 17. 5' end alternative SF1 transcripts associated on ExonMine.

(A) Schematic representation of the 5' portion of the SF1 gene. Constitutive and alternative exons are represented as empty and colored boxes, respectively. Introns are shown as lines. (B) Three different SF1 transcripts annotated in the ExonMine database are shown (i, ii and iii). Constitutive and alternative exons are indicated as empty and colored boxes, respectively. The intron is represented with a black line. The position of AUGs and a PTC is indicated.

Although transcripts i and ii lack exon 1 and the first AUG, they can be translated from a second AUG present in exon 4, which is in frame with the first one. Translation of short SF1 transcripts not containing the first three exons would lead to SF1 isoforms lacking the N-terminal 115 amino acids, which will be named Δ115 SF1. Such isoforms would have the same amino acid sequence as the other SF1 isoforms, but would lack the ULM, NLS1 and 2 (figure 18). Since they retain the KH-QUA2 domain, they can still bind RNA, but their nuclear localization and interaction with U2AF65 would be affected. Together with the KH/QUA2 domain, the ULM and NLS are required for the splicing function of SF1 and their absence in Δ115 SF1 suggests that the isoforms might not be involved in spliceosome assembly and thus have a different function in the cell. In contrast, inclusion of exon 3B in transcripts iii leads to a PTC in the same exon and thus to a non-functional protein.

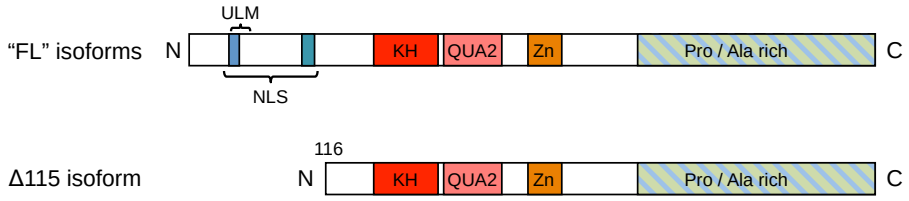


Figure 18. Comparison of protein organization between "FL" and N-terminal isoforms.

Structural and evolutionarily conserved SF1 domains are present in all C-terminal isoforms (ULM, U2AF65 ligand motif; NLS, nuclear localization signal; KH/QUA2 domain; Zn, zinc knuckle). N-terminal isoforms, corresponding to transcripts i and ii, start at amino acid 116, as indicated above the figure, and they lack the NLSs and the ULM.

Since the existence of SF1 transcripts containing exon 3A and/or 3B has never been tested experimentally, we decided to analyze their expression levels in different human cell lines by RT-PCR with specific primers. To investigate whether the newly identified alternative exons are first exons, the SF1 gene was screened for the presence of APs with bioinformatic tools and several portions of genomic sequences were tested for their transcription activity with a luciferase assay. In addition, 5' RACE was performed to identify the 5' sequence of alternative SF1 transcripts. Finally, expression of SF1 isoforms was analyzed at the protein level, taking advantage of SF1 antibodies that recognize different epitopes of the protein.

3.2. Materials and Methods

3.2.1. Cell culture

HeLa and 293T cells were grown in a humid chamber at 37°C with 5% CO_2 in DMEM supplemented with 10% FBS, 1% penicillin-streptomycin and 1% L-glutamine.

3.2.2. Cytoplasmic RNA extraction

HeLa and 293T cells were washed twice with cold PBS. Cells were lysed in a buffer containing 10 mM Tris-HCl pH 7.5, 140 mM NaCl, 1.5 mM $MgCl_2$, 0.5% NP-40 and 20 mM DTT in the presence of RNase inhibitors (Roche). The lysate was centrifuged at 6,000 rpm for 3 min at 4°C in a microfuge to remove nuclei. Additional centrifugation at maximum speed was performed for 10 min to remove mitochondria and membrane particles. 1% SDS, 10 mM EDTA pH 8.0 and proteinase K were added to the supernatant and incubated at 37°C for 30 min. Phenol-chloroform extraction was performed.

Isolated RNAs were treated with RQ1-DNase (Promega) for 30 min at 37°C and extracted with phenol-chloroform. RNAs were quantified with the Nanodrop (NanoDrop Technologies, Wilmington, DE, USA).

3.2.3. Reverse transcription (RT)

Equal amounts of RNA were reverse transcribed. cDNA was prepared with 100 ng of random primers (Promega), 10 mM each of dNTPs (Life Technologies) and 2 µg RNA for 5 min at 65°C, followed by addition of RNase inhibitors, 10 mM DTT, 1 x First-Strand buffer, 200 units of SuperScript II RT (Life Technologies) and incubated for 1 h at 42°C. The enzyme was inactivated at 70°C for 15 min.

3.2.4. Semi-quantitative PCR (Polymerase Chain Reaction)

Analysis of SF1 isoforms was performed by semi-quantitative PCR with HeLa or 293T cell cDNAs with Taq Polymerase (Life Technologies) according to the manufacturer's

instructions. The following cycling conditions were used: 94°C for 5 min, 35 cycles at 94°C for 45 sec, 56°C for 45 sec, 72°C for 3 min and finally 72°C for 10 min. Oligonucleotide sequences are summarized in table 2. PCR products were analyzed in agarose gels containing 0.5 μg/ml of ethidium bromide, purified from the gels, cloned into pGEM-T Easy (Promega) and sequenced with T7 sequencing primer (Microsynth, Switzerland).

Primer name	Sequence 5' ⇒ 3'	Orientation
SF1 exon 3A	TGATCTAGCCGATGCCTGTT	sense
SF1 exon 6	ACTGCCTTTTTGACGTTCTC	antisense
SF1 exon 2	CGCTGGAACCAAGACACAAT	sense
SF1 exon 4	GTGCGGAACTCTCGGGT	antisense
SF1 exon 3B	GGGAGCAGCCGAGTTCAGT	sense
SF1 intron 3A_3B	TCATTTAGCCGTCATTAGCC	sense
SF1 exon 5	TCATGACTTTATCACTCACACGTGTTG	antisense
GAPDH-for	CCATCACCATCTTCCAGGAG	sense
GAPDH-rev	CCTGCTTCACCACCTTCTTG	antisense

Table 2. Primers used for SF1 isoforms analysis.

3.2.5. Prediction of promoters

A partial genomic SF1 (gSF1) sequence was scanned for the presence of putative promoters. The sequence includes 2,075 bp upstream and 8,138 bp downstream the start codon, including the exon 4. Three user-friendly bioinformatic tools were chosen: PROMOTER SCAN (Prestridge, 1995), CpG Island Searcher (Takai & Jones, 2002) and Cister (Frith et al, 2001).

3.2.6. Cloning procedures

3.2.6.1. DNA constructs used for the analysis of SF1 isoforms

SF1 sequences were amplified from pcDNA3.1/SF1-HL1 and pcDNA3.1/SF1-B3-HL1 with the Expand High Fidelity Kit according to the manufacturer's instructions. The following conditions were used: 94°C for 2 min, 35 cycles at 94°C for 15 sec, 56°C for 30 sec, 72°C for 3 min and finally 72°C for 10 min. Oligonucleotide sequences with restriction sites are summarized in table 3. The PCR products were analyzed in agarose gels containing 0.5 µg/ml of ethidium bromide, purified from the gels, cloned into pcDNA3.1/HisC (Life Technologies) and sequenced with T7 and BGH-rev sequencing primers (Microsynth, Switzerland).

Primer name	Sequence 5' ⇒ 3'	Orientation	Restriction site
SF1_ex4	CGGGGTACC GTTGCACTCAATCCGGATTT	sense	KpnI
B3HL1	CTAGTCTAGA GAAGGGTCCATGGGAGGCGG	antisense	XbaI
HL1	CTAGTCTAGA GTTCTGTGGTGGAGGCGGTG	antisense	XbaI

Table 3. Primers used for analysis of SF1 isoforms.

Sequences listed in table 3 include the restriction site, which is underlined, and the complementary region.

3.2.6.2. DNA constructs used for luciferase assays

gSF1 sequences were amplified from HeLa DNA with the Expand High Fidelity Kit according to the manufacturer's instructions. The following cycling conditions were used: 94°C for 2 min, 35 cycles at 94°C for 15 sec, 60°C for 30 sec, 72°C for 3 min and finally 72°C for 7 min. Constructs 2 and 4 have a high GC content and were amplified with Takara LA Taq with GC buffer (Takara) according to the manufacturer's instructions. The following PCR conditions were used: 94°C for 1 min, 35 cycles at 94°C for 30 sec, 60°C for 30 sec, 72°C for 3 min and finally 72°C for 7 min. Oligonucleotide sequences with restriction sites are summarized in table 4. The PCR products were analyzed in agarose gels containing 0.5 µg/ml of ethidium bromide, purified from agarose gel, cloned into pGL3 (a gift from Didier

Picard, University of Geneva, Switzerland) and sequenced with plasmid sequence specific primers (Microsynth, Switzerland). Construct 3 was made by NcoI digestion of construct 4 and self-ligation.

Primer name	Sequence 5' ⇒ 3'	Orientation	Restriction site
Prom_0_F	CGGGGTACC TTGAGGTTCCTGTCTTTGC	sense	KpnI
Prom_0_R	CCCAAGCTT GCACTTCATTTGTTACTGCG	antisense	HindIII
Prom_01_R	GGAAGATCT CCTCTGCGGCGGCTTCTCCT	antisense	BglII
Prom_1_F	CGGGGTACC TCGCTCCGTCATAGAGTTCG	sense	KpnI
Prom_1_R	GGAAGATCT AAACCCTGCCTTGCTGTA	antisense	BglII
Prom_2_F	CGGGGTACC GAACCAGAATGGGAACCGTA	sense	KpnI
Prom_2_R	CATGCCATGG GTTTAGGAGGTCTTTGCCCG	antisense	NcoI
Prom_3_F	CGGGGTACC GCAAAGACCTCCTAAACCTACG	sense	KpnI
Prom_3_R	CATGCCATGG GGCGTGTAGAAGAATGGAATGT	antisense	NcoI
Prom_4_F	CCCAAGCTT GCACAGCCTCATTGGTCAA	sense	HindIII
Prom_4_R	CATGCCATGG GGCACTAAACAGTTGTGGTCTA	antisense	NcoI
Prom_5_F	CGGGGTACC CTTCCTGGATAACTGGGATTAG	sense	KpnI
Prom_5_R	CATGCCATGG CCTCGCTATTGTAGATGGGC	antisense	NcoI

Table 4. Primers used for SF1 luciferase constructs.

Sequences listed in table 4 include the restriction site, which is underlined, and the complementary region.

3.2.7. Transfection procedures

3.2.7.1. Over-expression of DNA constructs for luciferase assays

293T cells were seeded into 24-well plates. Cells were transfected with 2 ng of cytomegalovirus (CMV)-Renilla reporter vector (gift from Didier Picard, University of Geneva, Switzerland) and 10 ng of gSF1 cloned into pGL3 by calcium phosphate

transfection at about 70% confluency (Jordan et al, 1996). The culture medium was changed 24 h later. Cells were lysed 48 h after transfection and processed immediately. gSF1 constructs used for the luciferase assay are summarized in table 5.

Name	Insert
1	gSF1 (bp 130 – 1,714)
2	gSF1 (bp 130 – 2,039)
3	gSF1 (bp 1,632 – 2,074)
4	gSF1 (bp 1,632 – 2,955)
5	gSF1 (bp 2,895 – 4,853)
6	gSF1 (bp 4,837 – 6,212)
7	gSF1 (bp 6,139 – 8,185)
8	gSF1 (bp 8,271 – 10,094)

Table 5. DNA constructs used for luciferase assays.

3.2.7.2. Over-expression of DNA constructs for analysis of SF1 isoforms

293T cells were seeded into 10-cm plates and transfected with 10 µg of SF1 DNA constructs by calcium phosphate transfection as mentioned above (section 3.2.7.1) (table 6).

Name	Insert
short SF1-B3HL1	hsSF1 B3-HL1 (starting from nt 640)
short SF1-HL1	hsSF1 HL1 (starting from nt 640)

Table 6. DNA constructs used for the analysis of SF1 isoforms.

3.2.8. Luciferase assays

Fresh cell lysates were assayed for both Firefly and Renilla luciferase activities with the Dual-Luciferase Reporter Assay (Promega) according to the manufacturer's instructions. Firefly and Renilla luciferase activity were measured with the Plate Chameleon (Hidex, Turku, Finland). Firefly luciferase activities were normalized to the internal control, the Renilla luciferase activities.

3.2.9. 5' Rapid amplification of cDNA ends (RACE)

5' RACE was performed on HeLa cytoplasmic RNA with the FirstChoice RLM-RACE Kit (Applied Biosystems) according to the manufacturer's instructions. Briefly, cytoplasmic RNA was treated with CIP (calf intestine alkaline phosphatase) to remove free 5' phosphates from rRNA, tRNA, fragmented mRNA and genomic DNA. Then the RNA was treated with TAP (tobacco acid pyrophosphatase) to remove the cap structure from full length mRNA, leaving a 5'-monophosphate. An RNA adapter was ligated to the 5' monophosphate of the RNA with T4 RNA ligase and reverse transcription was performed with random primers (figure 19).

Nested PCR was performed on cDNAs with Taq polymerase (Life Technologies) according to the manufacturer's instructions with the forward primer annealing to the adapter (provided by the kit) and reverse sequence-specific primers, summarized in table 7. The following cycling conditions were used: 94°C for 5 min, 35 cycles at 94°C for 30 sec, 60°C for 45 sec, 72°C for 3 min and finally 72°C for 7 min. PCR products were analyzed on 2% agarose gels containing 0.5 μg/ml of ethidium bromide, were purified from the gels, cloned into pGEM-T Easy and sequenced with T7 sequencing primers.

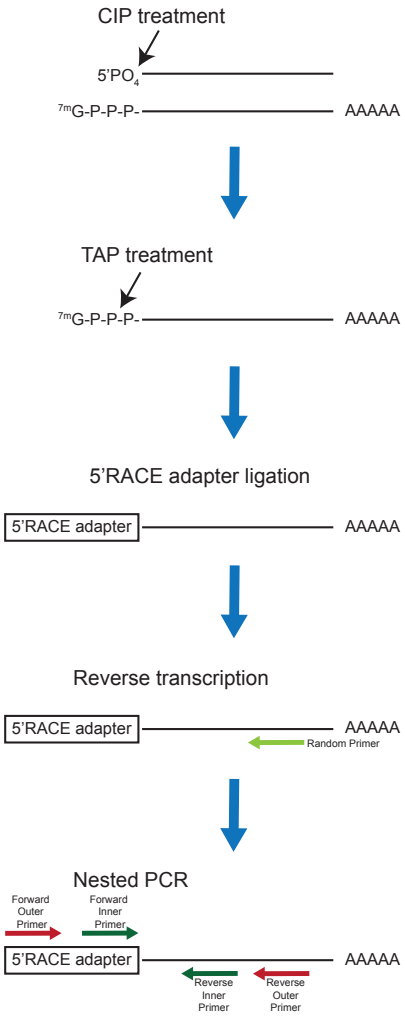


Figure 19. Scheme of the FirstChoice RLM-RACE procedure.

Primer name	Sequence 5' ⇒ 3'	Orientation
SF1 exon 6	ACTGCCTTTTTGACGTTCTC	antisense
SF1 exon 5	TCATGACTTTATCACTCACACGTGTTG	antisense
SF1 exon 3B	AGTTGAGAGAACCTTGCCTAGTTG	antisense
SF1 exon 4	GTGCGGAACTCTCGGGT	antisense
SF1 exon 3A	CCTCACAACAGTTTAACCTC	antisense

Table 7. Primers used in the nested PCR.

3.2.10. Polysome profile analysis

Before harvesting, HeLa cells were incubated with 100 µg/ml of cycloheximide (Sigma) for 10 min at 37°C. HeLa cells were washed with cold PBS supplemented with cycloheximide, collected and stored at -80°C. Cell pellets were processed for cytoplasmic extraction. Samples were resuspended in hypotonic buffer A (10 mM Hepes-KOH pH 7.9, 10 mM KCl, 1.5 mM $MgCl_2$, 1 mM DTT) supplemented with RNase inhibitors and 100 µg/ml cycloheximide and incubated on ice for 20 min. After addition of buffer B (buffer A containing 2.5% NP-40), samples were incubated on ice for 10 min and centrifuged at 3,400 rpm for 4 min in a microfuge. The supernatant, which represents the cytoplasmic fraction, was centrifuged at maximal speed for 20 min. The supernatant was diluted 1:1 with buffer A* (10 mM Hepes-KOH, 300 mM NaCl, 10 mM DTT, 1.5 mM $MgCl_2$) supplemented with RNase inhibitors and 100 µg/ml cycloheximide. Protein concentrations were measured with the Bradford reagent (Bio-Rad).

1.3 mg of protein was layered onto 11 ml of 15-50% sucrose gradients containing 50 mM Tris-HCl pH 7.5, 50 mM NH_4Cl, 5 mM $MgCl_2$, 1 mM DTT and centrifuged at 35,000 rpm in a Beckman SW41 rotor at 4°C for 3.3 h. As a control, to dissociate polysomes, 20 mM EDTA was added to the samples before loading onto the gradients. Gradient fractions of 1 ml were collected and gradient analysis was monitored at 254 nm. Total RNA was precipitated from mono- (fractions 3, 4 and 5) and polysome (fractions 7, 8, 9 and 10) fractions by the addition of 1 volume of cold isopropanol and incubated at -80°C overnight. The RNA was pelleted and purified.

3.2.11. Protein extraction

HeLa and 293T cells were washed twice with cold PBS and collected. Total protein extract was prepared by lysing the cell pellet with RIPA buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA pH 8.0, 0.5% NP-40) supplemented with complete protease inhibitors. Samples were incubated on ice for 10 min and sonicated 3 times for 20 sec with a break of 2 min. Samples were centrifuged at maximum speed in a microfuge and the protein concentrations of the supernatant were measured with the Bradford reagent (Bio-Rad).

3.2.12. Immunoprecipitation (IP)

HeLa and 293T cell protein extracts (600 µg) were incubated three times for 1 h at 4°C with Dynabeads Protein G (Life Technologies) coated with an anti-SF1 antibody (Abnova). The beads of the three incubations were washed three times with washing buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 1 mM EDTA pH 8.0, 0.01% NP-40) and bound proteins were eluted with SDS sample buffer for 5 min at 95°C. Input (30 µg), the third unbound fraction (30 µg) and 10% of the bound material were analyzed by Western blotting.

3.13. SDS-PAGE and Western blotting

Proteins were denatured by incubation with SDS loading buffer for 5 min at 95°C, separated by SDS-PAGE and transferred to nitrocellulose membranes by semi-dry blotting (Whatman, Protran). Membranes were stained for 1 min with 0.1% Ponceau S in 5% acetic acid, washed in TN-Tween buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl and 0.05% Tween 20) and blocked with 5% milk in TN-Tween. Membranes were probed with primary and secondary antibodies (table 8) diluted in blocking buffer. Membranes were washed three times with TN-Tween after incubation with antibodies. Antibody signals were detected with the SuperSignal kit (Pierce).

Primary Antibody	Host	Dilution	Company
SF1	mouse	1:50'000	Abnova
SF1-24D1	mouse	1:200	Corioni et al, 2011
His_6	mouse	1:2'000	Sigma
Secondary Antibody			
HRP-conjugated α-mouse	rabbit	1:2'000	Dako

Table 8. Antibodies.

3.2.14. Primers

Primers used in this study were purchased at Microsynth, Switzerland.

3.3. Results

3.3.1. Identification of exon 3A and/or 3B containing SF1 transcripts

Alternative SF1 transcripts containing exons 3A and/or 3B have been reported in the ExonMine database (Mollet et al, 2010). To validate their existence and analyze their expression levels in human cell lines, cytoplasmic HeLa and 293T RNAs were extracted and converted into cDNAs. RT-PCR was performed with forward primers complementary to exon 3A, exon 3B or intron 3A_3B and a reverse primer complementary to exon 6 (figure 20A). PCR products were separated on agarose gels, purified, cloned and sequenced.

As shown in figures 20B and C, RT-PCR for exons 3A and exon 6 produces, with both HeLa and 293T cDNAs, three main bands that correspond to different SF1 transcripts. The upper band of 2,179 bp corresponds to a SF1 mRNA containing exons 3A, 3B, 4, 5, 6 and introns 3A_3B, 3B_4, the band of 1,124 bp lacks intron 3B_4, whereas the band of 557 bp contains only exons 3A, 4, 5 and 6. The band of 2,000 bp (observed with HeLa cDNA) represents a PCR artifact.

RT-PCR performed with primers complementary to exons 3B and exon 6 resulted in three PCR products (figure 20D). The upper band of 1,537 bp corresponds to SF1 transcripts containing exons 3B, 4, 5, 6 and intron 3B_4. The intensity of the band is higher in 293T cDNA. Although very faint, the same band is present also in HeLa cDNA. The PCR product of 482 bp contains only exons 3B, 4, 5 and 6, whereas the band of about 1,250 (with 293T cDNA) is a PCR artifact.

Amplification of the region between intron 3A_3B and exon 6 shows the presence of two main SF1 transcripts, one of 1,598 bp and the other one of 543 bp (figure 20E). The former contains exons 3B, 4, 5 and 6 and introns 3A_3B, 3B_4, whereas the latter comprises exons 3B, 4, 5 and 6 and intron 3A_3B. PCR artifacts of about 1250 and 400 bp were detected with HeLa cDNA.

Negative controls were performed without adding any cDNA to the PCR reactions (figure 20B, C, D, E, lane -).

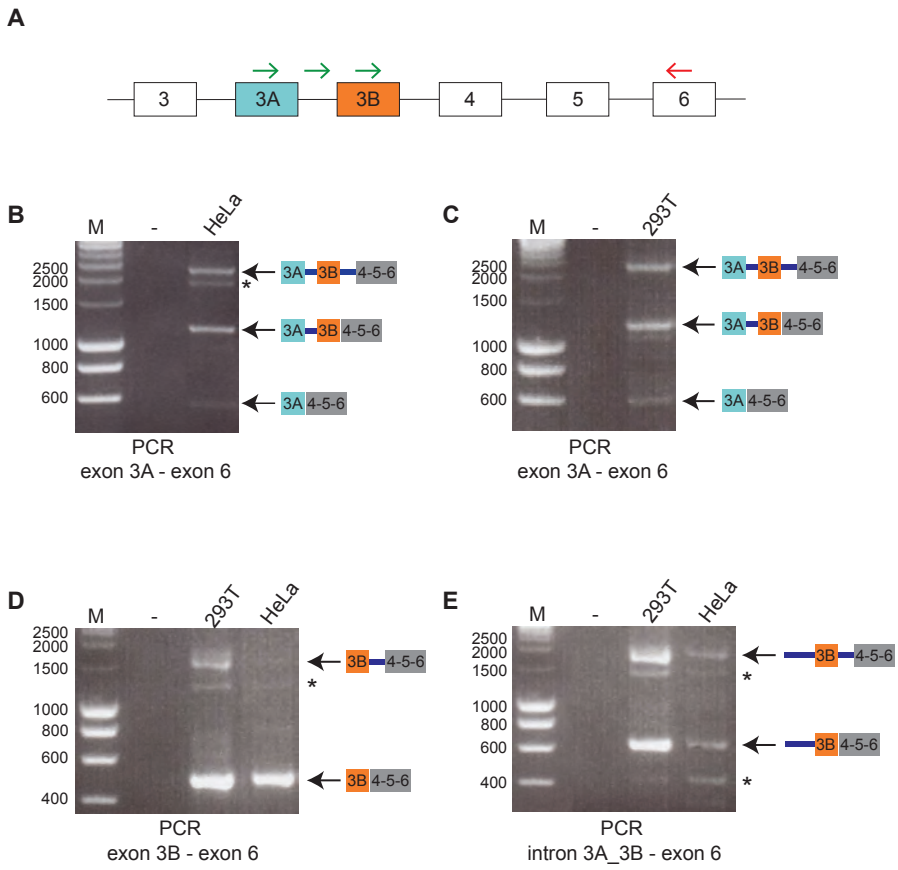


Figure 20. Identification of exon 3A and/or 3B-containing SF1 transcripts.

(A) Part of gSF1 is shown. Empty boxes represent constitutive exons, colored boxes alternative exons and lines introns. Green and red arrows indicate the position of forward and reverse primers, respectively. (B, C, D, E) RT-PCR was performed with HeLa and 293T cDNAs. The sets of primers used for the amplification are indicated under each figure. A negative control was performed without addition of cDNA to the PCR reaction (-). PCR products were separated on 1.5% agarose gels. Asterisks (*) show PCR artifacts. Arrows indicate novel SF1 transcripts, which were gel-purified, cloned and sequenced. DNA size markers (M) are indicated on the left in bp.

In summary, these data indicate the expression of several SF1 transcripts containing the annotated alternative exons 3A, 3B and retained introns in HeLa and 293T cells. In addition, we identified novel SF1 transcripts not described in ExonMine. In total, parts of three potentially new SF1 mRNAs were isolated (summarized in figure 21). PCR product A contains exons 3A, 3B, 4, 5, 6 and introns 3A_3B and 3B_4, whereas PCR product B lacks intron 3B_4. PCR product C lacks retained introns and contains exon 3A in addition to the common exons 4, 5 and 6 (figure 21). Since introns between exons 4, 5 and 6 were not

amplified, the possibility of genomic DNA contamination can be excluded. PCR product D could represent a portion of PCR product B or transcript iii (figures 17 and 21).

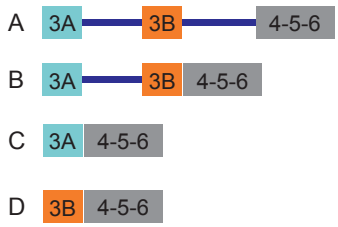


Figure 21. Summary of newly identified SF1 PCR products containing alternative exons 3A and 3B.

Four different PCR products have been identified in HeLa and 293T cells, named A, B, C and D. Constitutive and alternative exons are shown as grey or colored boxes, whereas introns are represented by a blue line.

Exons 3A and 3B were identified as novel alternative exons. Since in humans the majority of un-annotated exons identified with the ExonMine algorithm were shown to be first exons (Mollet et al, 2010), we investigated whether additional SF1 exons were present upstream of exons 3A and 3B. For this purpose, RT-PCR with primers complementary to exons 2 and 4 was performed on the same cDNAs used for the previous analysis (figure 22A). Only one PCR product of 241 bp was detected, which corresponds to SF1 exons 2, 3 and 4 (figure 22B). No bands of larger size containing alternative exons 3A and/or 3B and/or retained introns, were detected. Transcript iii annotated in ExonMine and containing exons 1, 2, 3, 3B and 4 was not amplified.

In conclusion, these results indicate that SF1 transcripts containing exon 3A do not have an upstream exon 2, suggesting the probability of an AP in the SF1 gene, whereas transcript iii might be expressed at a very low level and hence amplified only with sequence-specific primers.

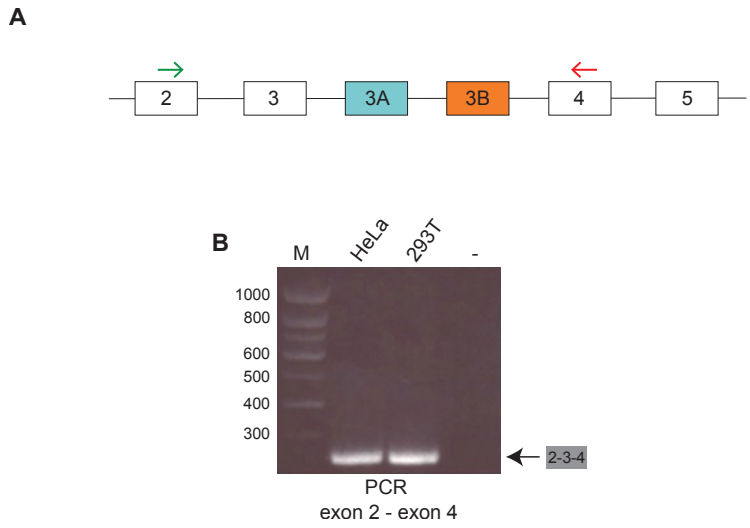


Figure 22. Analysis of exon 3A and/or 3B-containing SF1 transcripts.

(A) Part of gSF1 is shown. Empty boxes indicate constitutive exons, whereas colored boxes are alternative exons and lines introns. Green and red arrows indicate the position of forward and reverse primers, respectively. (B) RT-PCR was performed with HeLa and 293T cDNAs with primers complementary to exons 2 and 4, as indicated under the figure. A negative control was performed without addition of cDNA to the PCR reaction (-). PCR products were separated on a 1.5% agarose gel. The arrow indicates the only SF1 transcript amplified. DNA size markers (M) are indicated on the left in bp.

3.3.2. Identification of actively translated SF1 transcripts

The above results show that alternative SF1 transcripts containing exons 3A and/or 3B are expressed in HeLa and 293T cells. To investigate whether they are translated, we analyzed their presence in polysomes of a HeLa cytoplasmic extract, after fractionation of actively translated mRNAs from non-translated ones.

Translation initiates with the association of the small (40S) and the large (60S) ribosomal subunits to the mRNA to form the 80S complex (monosome). During translation elongation, further ribosomes are recruited to the same mRNA forming polysomes, which contain from two to over twenty ribosomes associated with a mRNA.

The population of HeLa cell polysomes was size-fractionated by sucrose density gradient centrifugation. On the top of the gradient we can distinguish peaks of 40S, 60S and 80S, which were initially collected in separated fractions (3, 4 and 5) and then combined to represent monosomes. Polysomes are found at the bottom of the gradient, from fraction 6 to 10, which were combined as the polysomes fraction (figure 23).

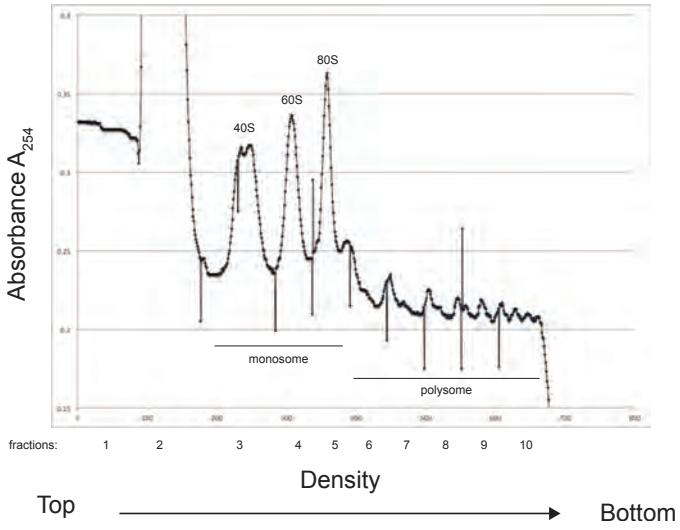


Figure 23. Polysome profile.

HeLa cell polysomes were size-fractionated by sucrose gradient centrifugation. 40S, 60S, 80S, monosome and polysome fractions are indicated on the graph.

RNAs extracted from the monosome and polysome fractions were reverse-transcribed. PCR amplification of GAPDH transcripts was carried out to test the efficiency of cDNA preparation. Expected PCR products of 591 bp were visualized on agarose gels in the monosome and polysome fractions, indicating that GAPDH mRNAs are translated (figure 24A). No band was detected when RT-PCR was performed without cDNA or with cDNA made without reverse transcriptase, confirming the absence of DNA contamination (figure 24A, lane RT- and -). We next performed RT-PCR with primers specific for exons 1 and 6 of SF1, which revealed a band of 653 bp only in the polysome fraction (figure 24B). The PCR internal control without cDNA did not show any band (figure 24B, lane -).

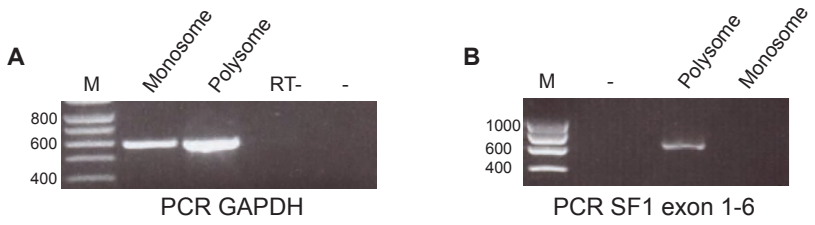


Figure 24. Analysis of actively translated transcripts.

(A and B) PCR was performed on monosome and polysome cDNAs. The sets of primers used for the amplification are indicated under the figures. A negative control was performed without adding cDNA to the PCR reaction (-). A PCR carried out on cDNA produced

without RT polymerase was used as an additional negative control (RT-). PCR products were separated on 1.5% agarose gels. DNA size markers (M) are indicated in bp.

Next, RT-PCR with forward primers specific for exon 3A, exon 3B or intron 3A_3B and a reverse primer complementary to exon 5 was performed. Surprisingly, amplification of the region between exon 3A and exon 5 yields only one band of about 350 bp, which corresponds to a SF1 mRNA containing exons 3A, 4 and 5 (figure 25B, lane polysome). No additional band of SF1 transcripts retaining introns 3A_3B and 3B_4 was detected (cf. figures 20B/C and 25B). Similarly, RT-PCR with primers specific for exons 3B and 5 shows the presence of only one product of about 200 bp in the polysome fraction, which corresponds to the sequence of exons 3B, 4 and 5 (figure 25B, lane polysome). No band of about 1,500 bp that contains intron 3B_4 was detected (cf. figures 25B and 20D). In addition, no PCR product was amplified with primer complementary to intron 3A_3B sequence (figure 25B).

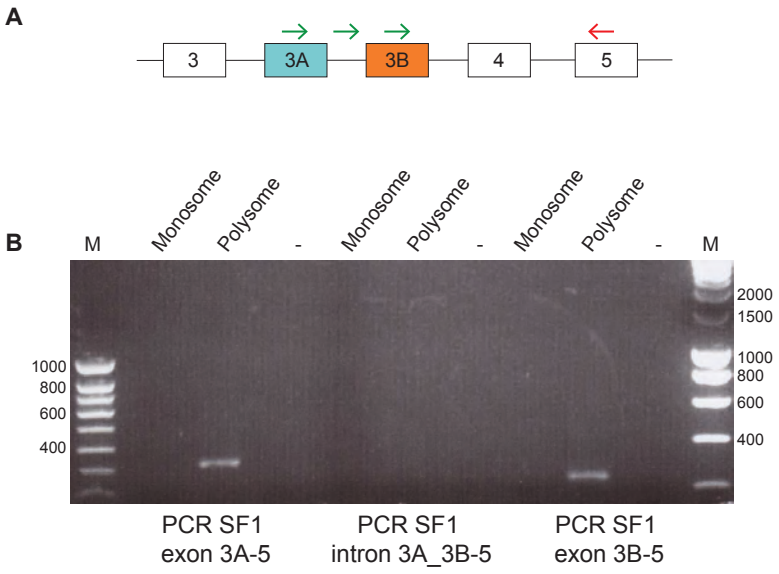


Figure 25. Identification of translated exon 3A and/or 3B-containing SF1 transcripts.

(A). Part of gSF1 is shown. Empty boxes indicate constitutive exons, whereas colored boxes are alternative exons and lines introns. Green and red arrows indicate the position of forward and reverse primers, respectively. (B) PCR was performed with monosome and polysome cDNAs. The sets of primers used for the amplification are indicated under the figure. A negative control was performed without adding cDNA to the PCR reaction (-). PCR products were separated on a 1.5% agarose gel. DNA size markers (M) are indicated in bp.

In conclusion, these results show that not all newly identified SF1 transcripts are translated into proteins. Only two SF1 transcripts, containing either exon 3A or 3B (figure 21, PCR products C and D) followed by exons 4 and 5, but not the transcripts retaining introns (figure 21, PCR products A and B), are present in the polysome fractions and therefore translated into proteins. In addition, we have observed that the exon 3B-containing transcript is amplified only with a primer specific for exon 3B and not for exon 3A or intron 3A_3B, suggesting that it contains only sequences complementary to the exon 3B oligonucleotide but not to the others. This PCR product might correspond to transcript iii, which contains a PTC in exon 5. In mammals, non-sense transcripts undergo NMD during the first round of translation and to date it is not known whether in this phase the mRNA is associated with one or more ribosomes (Maquat, 2004). Therefore, although transcript iii is not efficiently translated, it might still be associated with polysomes.

3.3.3. Prediction of APs in SF1 gene

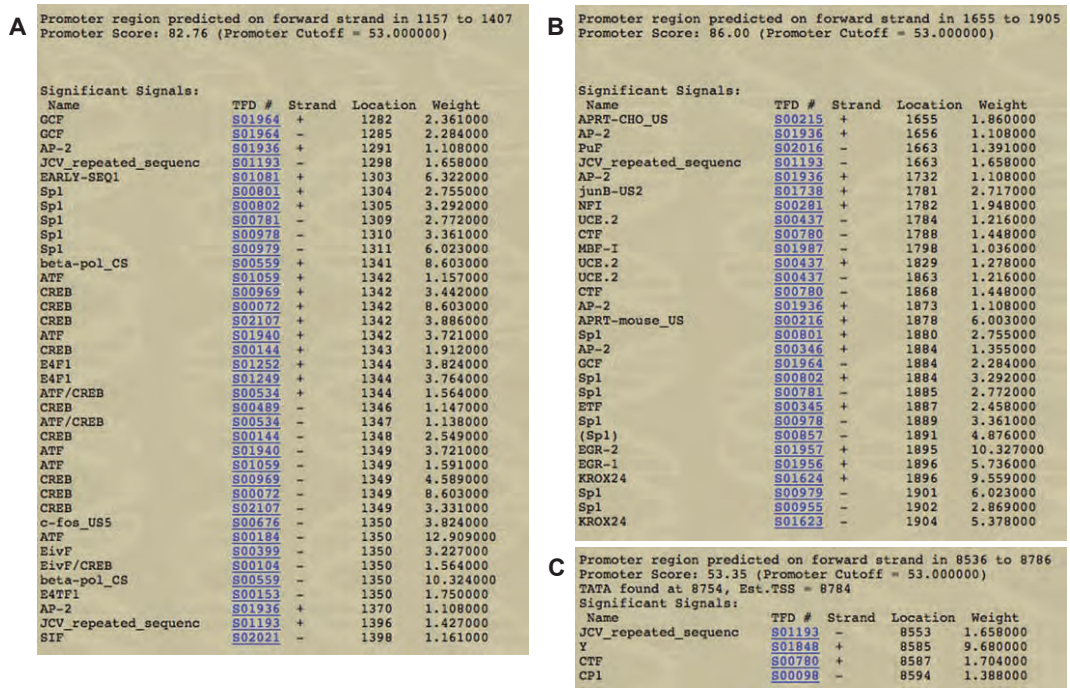
Since no SF1 transcript containing alternative exon 3A was amplified by RT-PCR with primers complementary to exons 2 and 4, we hypothesized that an AP could be present in the SF1 gene. To investigate this possibility, the genomic sequence of SF1 including 2,075 bp upstream and 8,138 bp downstream of the start codon up to exon 4 was scanned with three AP prediction tools: PROMOTER SCAN (Prestridge, 1995), CpG Island Searcher (Takai & Jones, 2002) and Cister (Frith et al, 2001).

3.3.3.1. PROMOTER SCAN

PROMOTER SCAN finds putative promoter sequences, by screening for transcription factor binding sites associated with eukaryotic RNA Pol II promoter sequences. It indicates the position of a predicted promoter and a TATA box when present. A list of transcription factors is also provided, with their weight signal and the strand and position to which they bind. The signal weight indicates the probability that a particular transcription factor is found bound to the promoter compared to non-promoter sequences. A high weight value indicates high probability that the transcription factor is found only in promoter sequence and, therefore, it discriminates between promoter and non-promoter sequences. The promoter

cutoff was set to 53 so that PROMOTER SCAN recognizes 70% of promoter sequences with a false positive rate of about 1/5,600 bp.

PROMOTER SCAN identified three promoters in the SF1 sequence analyzed (figure 26). The first promoter has a score of 82.76 and is located between 1,157 and 1,407 bp, i.e. in the region immediately upstream of exon 1 (figure 26A). The second one has a score of 86 and is located in exon 1, between 1,655 and 1,905 bp (figure 26B). PROMOTER SCAN also identified a third putative promoter with a score of 53.35 located in intron 3A_3B, between 8,536 and 8,786 bp (figure 26C). Considering that the promoter cutoff is 53, the third predicted promoter would have a very low transcription activity.



A

Promoter region predicted on forward strand in 1157 to 1407
Promoter Score: 82.76 (Promoter Cutoff = 53.000000)

Significant Signals:

Name	TFD #	Strand	Location	Weight
GCF	S01964	+	1282	2.361000
GCF	S01964	-	1285	2.284000
AP-2	S01936	+	1291	1.108000
JCV_repeated_sequenc	S01193	-	1298	1.658000
EARLY-SEQ1	S01081	+	1303	6.322000
Sp1	S00801	+	1304	2.755000
Sp1	S00802	+	1305	3.292000
Sp1	S00781	-	1309	2.772000
Sp1	S00978	-	1310	3.361000
Sp1	S00979	-	1311	6.023000
beta-pol_CS	S00559	+	1341	8.603000
ATF	S01059	+	1342	1.157000
CREB	S00969	+	1342	3.442000
CREB	S00072	+	1342	8.603000
CREB	S02107	+	1342	3.886000
ATF	S01940	+	1342	3.721000
CREB	S00144	+	1343	1.912000
E4F1	S01252	+	1344	3.824000
E4F1	S01249	+	1344	3.764000
ATF/CREB	S00534	+	1344	1.564000
CREB	S00489	-	1346	1.147000
ATF/CREB	S00534	-	1347	1.138000
CREB	S00144	-	1348	2.549000
ATF	S01940	-	1349	3.721000
ATF	S01059	-	1349	1.591000
CREB	S00969	-	1349	4.589000
CREB	S00072	-	1349	8.603000
CREB	S02107	-	1349	3.331000
c-fos_US5	S00676	-	1350	3.824000
ATF	S00184	-	1350	12.909000
EivF	S00399	-	1350	3.227000
EivF/CREB	S00104	-	1350	1.564000
beta-pol_CS	S00559	-	1350	10.324000
E4TF1	S00153	-	1350	1.750000
AP-2	S01936	+	1370	1.108000
JCV_repeated_sequenc	S01193	+	1396	1.427000
SIF	S02021	-	1398	1.161000

B

Promoter region predicted on forward strand in 1655 to 1905
Promoter Score: 86.00 (Promoter Cutoff = 53.000000)

Significant Signals:

Name	TFD #	Strand	Location	Weight
APRT-CHO_US	S00215	+	1655	1.860000
AP-2	S01936	+	1656	1.108000
PuF	S02016	-	1663	1.391000
JCV_repeated_sequenc	S01193	-	1663	1.658000
AP-2	S01936	+	1732	1.108000
junB-US2	S01738	+	1781	2.717000
NFI	S00281	+	1782	1.948000
UCE.2	S00437	-	1784	1.216000
CTF	S00780	-	1788	1.448000
MBF-I	S01987	-	1798	1.036000
UCE.2	S00437	+	1829	1.278000
UCE.2	S00437	-	1863	1.216000
CTF	S00780	-	1868	1.448000
AP-2	S01936	+	1873	1.108000
APRT-mouse_US	S00216	+	1878	6.003000
Sp1	S00801	+	1880	2.755000
AP-2	S00346	+	1884	1.355000
GCF	S01964	-	1884	2.284000
Sp1	S00802	+	1884	3.292000
Sp1	S00781	-	1885	2.772000
ETF	S00345	+	1887	2.458000
Sp1	S00978	-	1889	3.361000
(Sp1)	S00857	-	1891	4.876000
EGR-2	S01957	+	1895	10.327000
EGR-1	S01956	+	1896	5.736000
KROX24	S01624	+	1896	9.559000
Sp1	S00979	-	1901	6.023000
Sp1	S00955	-	1902	2.869000
KROX24	S01623	-	1904	5.378000

C

Promoter region predicted on forward strand in 8536 to 8786
Promoter Score: 53.35 (Promoter Cutoff = 53.000000)
TATA found at 8754, Est.TSS = 8784
Significant Signals:

Name	TFD #	Strand	Location	Weight
JCV_repeated_sequenc	S01193	-	8553	1.658000
Y	S01848	+	8585	9.680000
CTF	S00780	+	8587	1.704000
CP1	S00098	-	8594	1.388000

Figure 26. Promoters identified by PROMOTER SCAN.

Each table represents the promoter predicted by PROMOTER SCAN. Two main promoters are identified in the region immediately upstream of exon 1 (A) and in exon 1 (B), whereas a third weak promoter is predicted in intron 3A_3B (C). For each promoter the position on the forward strand and the promoter score are reported. A list of transcription factors, locations of their binding sites and weight value is also provided.

3.3.3.2. CpG Island Searcher

CpG Island Searcher screens for dinucleotide clusters of CpGs (called CpG islands) in DNA sequences and the results are shown as a graphical map of CpG dinucleotide distribution. CpG islands are present in the promoters of approximately 40% of mammalian genes (Larsen et al, 1992). CpG islands are defined as a minimum of a 500-bp region of DNA with a GC content higher than 50% and a ratio of observed CpG/expected CpG (ObsCpG/ExpCpG) greater or equal to 0.6 (Takai & Jones, 2002). The CpG Island Searcher identified a cluster of CpGs in the SF1 DNA sequence in the region corresponding to sequences upstream of exon1 to intron 1_2 (figure 27).

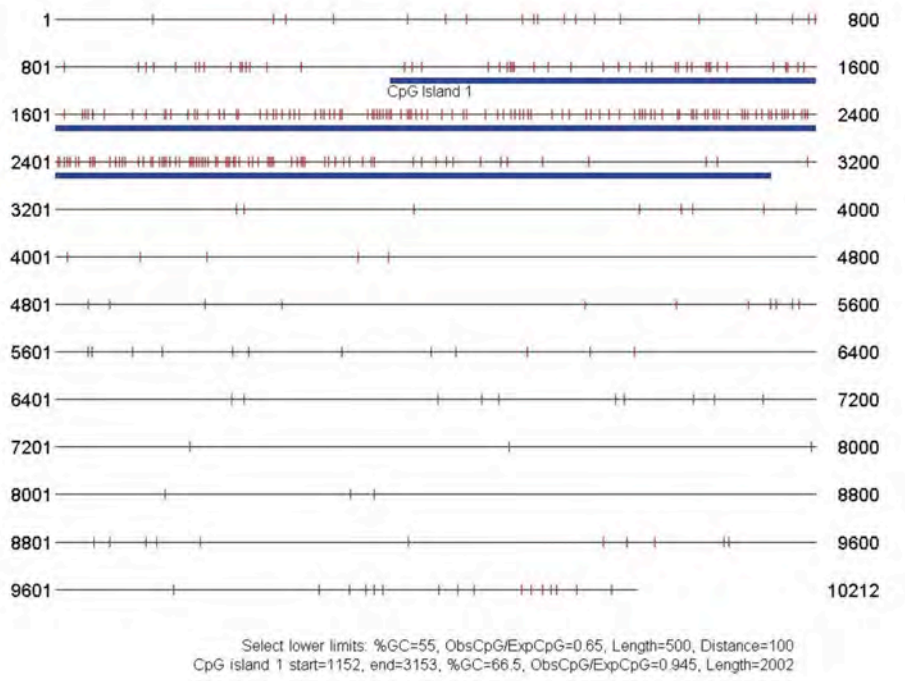


Figure 27. CpG islands in the SF1 gene.

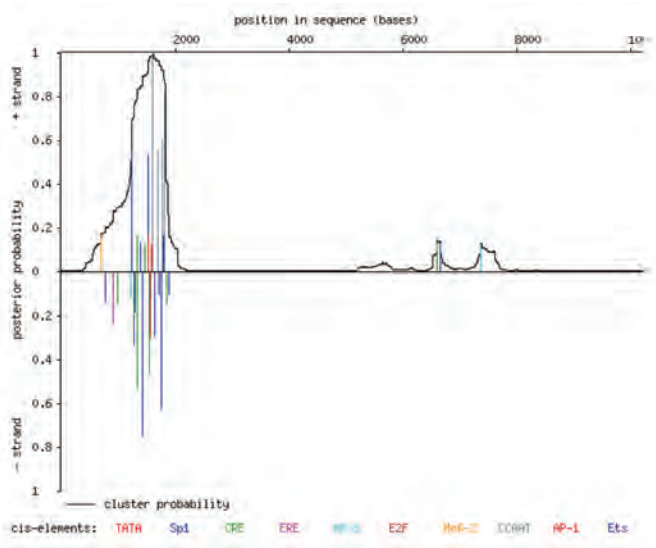
The gSF1 sequence analyzed with CpG Island Searcher is shown as a line. The underlined sequence has CpG islands. Parameters used for the search are indicated below. Position, % GC, ObsCpG/ExpCpG and length of the CpG island identified are also shown.

3.3.3.3. Cister

Cister (Cis-element Cluster Finder) predicts putative promoters by searching for clusters of *cis*-elements, which are DNA regions that regulate the expression of genes located on the same DNA strand. The results are displayed in a plot, where the colored lines indicate

probabilities that regulatory factors bind to *cis*-elements at these positions. Each color corresponds to a different binding site, as described in the figure legend (figure 28). The black curve indicates the overall probability of being within a cluster of *cis*-elements bound by their factors. Lines in the upper half of the plot indicate *cis*-elements on the direct strand, and lines in the lower half refer to the complementary strand.

Cister identified two clusters in the SF1 sequence: the first one includes exon 1 and the region immediately upstream, whereas the second one, of lower probability, encompasses exon 3 and sequences upstream and downstream (figure 28).



Possible functional cis-elements

type	position	strand	sequence
CCAAT	1609 to 1624	+	cgcagccaatgagaga
CCAAT	1812 to 1827	+	ctccgccaaatcggagg
Ets	1429 to 1439	-	cactcactgct
Ets	1752 to 1762	-	catttcctcgc
CCAAT	1775 to 1790	+	ttcggccaatcccagc
CCAAT	1698 to 1713	+	agtaacaaatgaagtg
Ets	1231 to 1241	+	aacaggaagtc
CRE	1337 to 1348	-	gttgacgtcact
Sp1	1522 to 1534	+	tgggggagggag
LSF	1541 to 1555	-	aaggctggaaccagc
Sp1	1286 to 1298	-	agccccccccctg
E2F	1563 to 1574	-	aatccccgcgaca
Sp1	1651 to 1663	-	cgccccaccccat
ERE	912 to 925	-	ggtcaggctggtct
LSF	1274 to 1288	-	ccggcgcgcgccagc
Sp1	1298 to 1310	-	gctccccgccccaa
CRE	1337 to 1348	+	gttgacgtcact
Ets	1795 to 1805	+	atcaagaagga
TATA	1534 to 1548	+	gtagaaaaaggctgg
Mef-2	707 to 718	+	agttatttttat
LSF	980 to 994	-	caggcgtgagccacc
LSF	1855 to 1869	-	ccggccaaatccagg
Sp1	787 to 799	-	acatccgcctccc
Sp1	1398 to 1410	+	gagaggcggtttt
Ets	6647 to 6657	+	ggaaggaagag
CRE	6593 to 6604	+	ggtgacgtaaca
LSF	1461 to 1475	+	tctcgttccggctgg
AP-1	1591 to 1601	+	attgattgagt
NF-1	7366 to 7383	+	ctttggccaaaagccccg
NF-1	1212 to 1229	-	ttttttttttttccaaaa
Sp1	1888 to 1900	-	cggcccgcccccg
Sp1	1722 to 1734	-	gacacctcccagc

Figure 28. Cister analysis.

Cister results are shown in a plot that represents the probability of a cluster of regulatory elements. The colored lines indicate probabilities that regulatory factors bind DNA at these positions. Each color corresponds to a different type of regulatory elements, as shown in the legend. Lines in the upper half of the plot indicate *cis*-elements on the + strand, and lines in the lower half refer to the complementary (-) strand. The table lists the positions and sequences of regulatory elements.

In summary, the bioinformatic tools predicted a strong promoter in the region encompassing the 5' UTR, exon 1 and intron 1_2, i.e. the location of the promoter of full-length SF1 transcripts (figure 29). However, the programs also identified sequences located further downstream in the SF1 gene that could function as APs, such as sequences within exon 3, intron 3_3A and intron 3A_3B (figure 29). Therefore the prediction results support the hypothesis of an AP for the transcription of shorter SF1 transcripts containing exons 3A and/or 3B.

```
aagatgtatttgtatattacaggtattggctctttatcatgtatggcaaatatttttcctcatctatca    70
           10        20        30        40        50        60
ttgtctcttgacttttatcagtgttaattttcggggtaatcaagtatatctttttacagtttgaggttcc   140
           80        90       100       110       120       130
tgtctttgcagtctctcaacccctagaataccttatagtctcctaaattatattctaaatgttttgttg   210
          150       160       170       180       190       200
tttaaatttaagattctacgccatttatcaacgaaatcatcttttcctgattgaaataaaatactacatt   280
          220       230       240       250       260       270
tgtggtatttccggattctgttgcactaatgtgtctttcctaccaacaatatcatattgttttattattg   350
          290       300       310       320       330       340
tggtttttgtttttgagacggagtctcatgctgttgcccaggctggagtacagtggtgcgatctcagct   420
          360       370       380       390       400       410
cactacacctccgcctcctgggttcaagcaattctcctgcctcagcctcctgagaagctgggattatagg   490
          430       440       450       460       470       480
cgtctgccaccacgctcggctaatttttgtactttagtagagacgggggtttcaccgtgttggtcaggt   560
          500       510       520       530       540       550
tggtctcgaactcctgacctcaggtgatccccccgccttggcctcccagtgccttggcactggcctattg   630
          570       580       590       600       610       620
tgttttatatactatgtttgtcatctggaaaggcaagctcctgctccgtcttgtaaagtttcccagtac   700
          640       650       660       670       680       690
tgtagtagttatttttatttttatttttgacagtctcgctctgtcacccaggctgaagtgcagtggcccc   770
          710       720       730       740       750       760
atctcggctcactgcaacctccgcctcccgggttcaagcgattctcttgcctcatcctcctgggaagctg   840
          780       790       800       810       820       830
ggattgcaggagtgcaccactacacccaactaatttttgtatttttcgtagagacggggtttcgccatgt   910
          850       860       870       880       890       900
tggtcaggctggtctcgaccttctgaccttgtgattcgcccgcctcggcctgtcaaagtgctgggattac   980
          920       930       940       950       960       970
aggcgtgagccaccgcgcccggccgtagttattgttttatacgattttaagatcatttaattaaatcctc  1050
          990      1000      1010      1020      1030      1040
tgaactccgtggagattccaatagaagtgtgtatttgtaatttttgaactgacaaccttctgttacctttt 1120
         1060      1070      1080      1090      1100      1110
catctaatatgttgttgcatcagttcaaatactgttttaaatccttcgataagtcgtctctattcgtagc  1190
         1130      1140      1150      1160      1170      1180
ttttgggtcacttttttttttttttttttttccaaaaaaacaggaagtcttttacctgcagcgtcaa  1260
         1200      1210      1220      1230      1240      1250
ctgagccggtttccggcgcgcgccagcccctcccctgctcccgccccaaggttctgcgttaccatggga  1330
         1270      1280      1290      1300      1310      1320
acaaaggttgacgtcacttagggcttccttgctggacccaagggtcgatggtgaagggagacgggtggag  1400
         1340      1350      1360      1370      1380      1390
aggcggtttgctccatctacgcatgcgcacttcctgctccaggactccccgccgtaaatctcgttccg  1470
         1410      1420      1430      1440      1450      1460
gctgggccttttccgccgcgactctcgcttaatcccggagaaactgccccctgggggaggggagtagaaa  1540
         1480      1490      1500      1510      1520      1530
aaggctggaaccagcgacagccaatcccgcgacactacaacgcaggcgagattgattgagtccaccaccg  1610
         1550      1560      1570      1580      1590      1600
cagccaatgagaGAGCTCGCCGTCGCTCCGTCATAGAGTTCGCCCCACCCCATCCCCTCCTTTCTGGACT  1680
         1620      1630      1640      1650      1660      1670
CGGAGCTCAGTTCACGCAGTAACAAATGAAGTGCGCGCTGCGACACCTCCCAGCCCACCGAACTCCGCCG  1750
         1690      1700      1710      1720      1730      1740
```

```
CCATTTCCTCGCTTGGCCTAACGGTTCGGCCAATCCCAGCGCGCATCAAGAAGGACTGAGGCTCCGCCAA  1820
          1760      1770      1780      1790      1800      1810
TCGGAGGCCGCCGATTTCGACCCTTCGCCTCGGCCCGGCCCAATCCAGGCCCCGGCCCCGCCGCCCCCGG  1890
          1830      1840      1850      1860      1870      1880
CCCGCCCCCCGCGGTGCCCTCTCTCCTCCCTCTTTGTGCGTCTCGCGCCGCCGCCGCCCGCCGCGTGAGAG  1960
          1900      1910      1920      1930      1940      1950
GACGGGCTCCGCGCGCTCCGGCAGCGCATTCGGGTCCCCTCCCCCCGGGAGGCTTGCGAAGGAGAAGCCG  2030
          1970      1980      1990      2000      2010      2020
CCGCAGAGGAAAAGCAGGTGCCGGTGCCTGTCCCCGGGGGCGCCATGGCGACCGGAGCGAACGCCACGCC  2100
          2040      2050      2060      2070      2080      2090
GTTGGgtaagctgggccccccgggctgcccccgctccccgggcccaaaggaggcttcgagccgggccct  2170
          2110      2120      2130      2140      2150      2160
ccgcctgcacccgggcctggggcggggctgctggctcccgggccgccgccgcccccgcccgtgggctcga  2240
          2180      2190      2200      2210      2220      2230
tgggggccctgaccgcggccttcccttcgcggcgctgcctccgccgcctccgccgccgccccctccgcc  2310
          2250      2260      2270      2280      2290      2300
tccccagcagccgCCGCCGCCTCCACCGCCACCGTCCCCCGGCGCCTCGTACCCGCCGCCGCAGCCGCCC  2380
          2320      2330      2340      2350      2360      2370
CCTCCGCCGCCGCTCTACCAGCGCGTGTCGCCGCCGCAGCCGCCGCCACCCCAGCCGCCGCGTAAGGACC  2450
          2390      2400      2410      2420      2430      2440
AGCAGCCGGGCCCGGCCGGCGGCGGAGGAGgtgagtcgggccgagagagcgcgagaggcgtcgcgcggac  2520
          2460      2470      2480      2490      2500      2510
gggcccggccgcctgaggggcgcgcggcggcggcgacggcggctgcacgcggcaggggcgcgcgtgcgcg  2590
          2530      2540      2550      2560      2570      2580
cacggggaggcccgggcggaggcgcccctgggcgcgcgcggcctggttctagggcctcgagtgcgcacgc  2660
          2600      2610      2620      2630      2640      2650
gcgctgtgggggggagggctgggcgggcgggaggcgctctgtgcgcagacgcagtgaggtggccgggctgt  2730
          2670      2680      2690      2700      2710      2720
gcggcgcaactgtatagactgggggcccactattgctgtccagtccgctggtgtcgggaacaggggtggc  2800
          2740      2750      2760      2770      2780      2790
gggtgttcagcgctggacggggtgctagctgtgaatgccagcagggcgtgaaagcctcacctgatgacgt  2870
          2810      2820      2830      2840      2850      2860
gcagcgactcccttctggtcttgaggaaccagaatgggaaccgtaccaaggatagctcctctgcttctac  2940
          2880      2890      2900      2910      2920      2930
agcaaggcagggttttcatccgggagaggttggcatgcttcagatgatggttgggtctgatctttcttgg  3010
          2950      2960      2970      2980      2990      3000
tctcttttggggagagaatttagtttgttttttttttaaatctgttcagtatgtttagatctcagtct  3080
          3020      3030      3040      3050      3060      3070
gatcggctgccctgtcgctttgttttaaaaagcttttccctctgatcctgtacccctttgggaaatagaag  3150
          3090      3100      3110      3120      3130      3140
gactcttttgtggcaaatgtagggagtttctgcctatgtacggggaaaggagaacttggctgctattgta  3220
          3160      3170      3180      3190      3200      3210
taggttgttgaatttttgatagatccttttaagagttcagaccctcagtgtccaactggccaggaaaact  3290
          3230      3240      3250      3260      3270      3280
gaatgctgatcttggctaaacttatgaagacaaaggtgatcttgttgaagaattgggcttttcccacttg  3360
          3300      3310      3320      3330      3340      3350
gtagggtgtaactaggtttagttgtggtccggcctctcgtacaaaagtgactggtgaggggagtggaggt  3430
          3370      3380      3390      3400      3410      3420
gggtggagcccccagtcccatcaggactgctttacaaatctatgattcttcagtgtatttttgtctcccc  3500
          3440      3450      3460      3470      3480      3490
tggggatttcagagcactcatggttgatgtagtatctatctaacactctctgcagccatatgcattaaga  3570
          3510      3520      3530      3540      3550      3560
ttaccacgtttatatatctgacagtttttctagacaaaagaaggtgaaatgttaacttgcctgaccaagg  3640
          3580      3590      3600      3610      3620      3630
accacataccttgtagtggccataagctttaaatctctgatctctgaggtttgttttctgttttttaa  3710
          3650      3660      3670      3680      3690      3700
tgtttaagttttaggttaagctctcagcttcacagtctgctgagtttctgtgtagtaccttaaatacct  3780
          3720      3730      3740      3750      3760      3770
tagctgtctgttgataactacttgtatttacatcggatggattttgtctctgccttttagacttcccaag  3850
          3790      3800      3810      3820      3830      3840
TAAGAAGCGGAAGAGGAGCCGCTGGAACCAAGACACAATGGAACAGAAGACAGTGATTCCAGGAATGCCT  3920
          3860      3870      3880      3890      3900      3910
ACAGTTATTCCCCCTGGACTTACTCGAGAACAAGAAAGAGCTTATATAGgtaaatatacgttctgcatcc  3990
          3930      3940      3950      3960      3970      3980
aaatgcactttgagattgttacgatttattctgagcaagatctgcatttttgaaagtttgagattgtag  4060
          4000      4010      4020      4030      4040      4050
ttatctttttgagatgagatactttcacgactttggtatcatctgtcagttttgcccagtgagttctg  4130
          4070      4080      4090      4100      4110      4120
```

```
catttgcagctccttgttttggtctgttcgtaactgcagtgttctccatgagtatcaggaaggtagggtt  4200
           4140      4150      4160      4170      4180      4190
tcacttaggagtaagaacagtccccagtccagcagccacccttttcagctgctgttcattgccagttgat  4270
           4210      4220      4230      4240      4250      4260
gaggtgagtgtcatctgcctctctggacaggccccagtggagaacaccgcaggtactgtaaaccaagtac  4340
           4280      4290      4300      4310      4320      4330
ttttcacagcgtggcctttcaccataattttatatttctccccctaaaatgtggtcttttacttttttata  4410
           4350      4360      4370      4380      4390      4400
ggtactttctgaaagaaaaaaaaatgacaaaaagtggttctgaaagtgtgtttacaatttacccctttt  4480
           4420      4430      4440      4450      4460      4470
attttcttgaatttaagagtttaaccttcccttcaacatcttttctggtttggtctatggagagctgtct  4550
           4490      4500      4510      4520      4530      4540
taaaatgaaggctcttgtatttacttttaaagaactctgctgggaaagggatattgcatggatatggaat  4620
           4560      4570      4580      4590      4600      4610
gaagtgctaggataaaaaggttttatttgctgggaggttttgagaattaatgactatccacagtggaatt  4690
           4630      4640      4650      4660      4670      4680
gaaattaacctcaaaattgctttttttggaaaaaaaagttttttaatgcaagttttatctttatatacctgt  4760
           4700      4710      4720      4730      4740      4750
gcagctgtctcaaaaaagttcactttataaacatttcctactaaccaaataaacatccccttttggaggaa  4830
           4770      4780      4790      4800      4810      4820
attcgggcaaagacctcctaaacctacgaggctgtggccttcacaaccaaaaagcttcaaggccagcatg  4900
           4840      4850      4860      4870      4880      4890
ttactacttttgagggtcttaatttcctgtttttagactgctgagaatttgaggcgttggaattgtct  4970
           4910      4920      4930      4940      4950      4960
gtctcctttcacagtactactaatacttgttctttcccctactgttcctgtctactcccacccattccgt  5040
           4980      4990      5000      5010      5020      5030
tgtgtctcatattagactagtctgggtctttacatccttgtgaaatgcccctagtggagtgatatcccaa  5110
           5050      5060      5070      5080      5090      5100
gaacacaagcatacttagaagatgggaagagcacaaattctttatatctccatttctgtttttgtggagg  5180
           5120      5130      5140      5150      5160      5170
gctagttagttcactgggccctaaggactatttatagctataatcaaaatagaatcaaaggagaggtgta  5250
           5190      5200      5210      5220      5230      5240
gttctccacagaagagaatgacccttttttatttagatcaggccccacatctccatagtcagctaggaaga  5320
           5260      5270      5280      5290      5300      5310
ctcaccttctcccctcagtgaaggtgactttcacctcgagttgttgagcttaatactcaactgcaaagtc  5390
           5330      5340      5350      5360      5370      5380
agaacttttaggcataaaggagcttggggaaccatgtgtcttactacttttctgtctcccacggtacat  5460
           5400      5410      5420      5430      5440      5450
actgctttgtccaggatgccctcaatatctttttttttttttttttgagagactgagtcttgctgtgtcg  5530
           5470      5480      5490      5500      5510      5520
ccaggctggagtgtagtggtgcgatctcggctcactgcaacctgcgtctcccgggttcaagtgattctcc  5600
           5540      5550      5560      5570      5580      5590
cacctcagcctcccaagtagctgggactacaggcgcacgccaccatgcccagctaatttttttttgtagtt  5670
           5610      5620      5630      5640      5650      5660
ttagtagagacggggtttcaccatgttggccaggatggtctcgatctcttgtcctcatgatctgcctgcc  5740
           5680      5690      5700      5710      5720      5730
ttggcctcccaaagtactgggattacaggtgtgagccactgtgcccggccagtatttctttcgataata  5810
           5750      5760      5770      5780      5790      5800
atgggttttcatttcttcagggttttagtgtgtgggtgtggtatagccctgtggatgtatttaaagttag  5880
           5820      5830      5840      5850      5860      5870
aaggaatctttggttagacacgcattaatcctatcttggagccagctcagttttaacccaacctagtaac  5950
           5890      5900      5910      5920      5930      5940
atttgcccttattggagatgcctgctgaaaattctcagatgctgcgaaggtaaggaacaggtggttattc  6020
           5960      5970      5980      5990      6000      6010
cgcatgttgaagaaagctgctctggctctgcagttcccattatcttggaaaatcagctgtgtaacatttc  6090
           6030      6040      6050      6060      6070      6080
cctggcgtcactggcaacttttgtgaaaccctcatctctcttttcatgcacagcctcattggtcaaggc  6160
           6100      6110      6120      6130      6140      6150
ccgagggccaagtacattgtccttgaacacacattccattcttctacacgcctgctccccatatgaacag  6230
           6170      6180      6190      6200      6210      6220
tgagcctgaaaacaaattcttagcactttatggttactggtttattaagatatgtggtttgtcatttaat  6300
           6240      6250      6260      6270      6280      6290
atttgaataaatatttcatataattgaataagagctgcaggtattgggagacccttactagtacaggtct  6370
           6310      6320      6330      6340      6350      6360
ggttgtacagaaagagtatgtacaaatttctataaggtgttgaattttgttattaaatgtcttatttaag  6440
           6380      6390      6400      6410      6420      6430
gaattttaaatgtattaatgaaaatgatctttgctcaattggtatttttccttgtcccatgttaattggg  6510
           6450      6460      6470      6480      6490      6500
```

tctcattttgggctatgtttaacatttaatgaagaaaaaactcaattcttgccctcaagtaatttgatgg 6580
6520 6530 6540 6550 6560 6570
gagtcgggtattggtgacgtaacatggtttatggttctgggtaccaatggtaagcaagccttggtgggga 6650
6590 6600 6610 6620 6630 6640
ggaagaggtgaaacacctcctaataattgcttggatgttagccttagtgaacagtctttattgctaaccc 6720
6660 6670 6680 6690 6700 6710
ctcttccaacaaactgtttgcttccttaagtgataaccattttgaagtgaaaaaaacctatagatgtaaa 6790
6730 6740 6750 6760 6770 6780
aggactcctgccgataggtggttgggggtttttgcttcaatgtgttcactgatcagtcgtccctttcaaa 6860
6800 6810 6820 6830 6840 6850
gtctccgagttcttcatgcttgttgttgagatttgagatgtttattggtaacattgtctttttttccatcc 6930
6870 6880 6890 6900 6910 6920
ttttttttctttcctaacctggtttcttccagTGCAACTGCAGATAGAAGACCTGACTCGTAAACTGCGC 7000
6940 6950 6960 6970 6980 6990
ACAGGAGACCTGGGCATCCCCCCTAACCCTGAGGACAGgttgggaaacagttttaaccagctgacagtaa 7070
7010 7020 7030 7040 7050 7060
cgcttttacctttttctcaatacgttttactggtggcaacatattgtgtcagttcaataaaatcatagta 7140
7080 7090 7100 7110 7120 7130
tgacgactgtaccctgaccactttggggattaaggtttaattagaaatatgtatggtcagtcaacagtta 7210
7150 7160 7170 7180 7190 7200
gaaaattggaataaaacttaaaaatgttgggagaataaaaagttacaggaaaatggacaactggtttcca 7280
7220 7230 7240 7250 7260 7270
attggtaagcatgtagatcttaacctgatttacccacccttaggggacagtgaggtggcccgttaacctg 7350
7290 7300 7310 7320 7330 7340
ccttttcagtacctgctttggccaaaagcccagggcatagagtttttctcacaggtacacagtgcaggga 7420
7360 7370 7380 7390 7400 7410
ctaaaaatatttttgagcatcactgaacttgagcaagccaggagctctggtcagctcattgctgcagtct 7490
7430 7440 7450 7460 7470 7480
tcctgctcatggttttgagatggtttcaaaatcagatgagatttctaaccaacttctttcactgccagtt 7560
7500 7510 7520 7530 7540 7550
aactttcagccttggccattgccttgggggtgttgggaggggccagaatattctcacctattttgggaag 7630
7570 7580 7590 7600 7610 7620
gcatgcaaccaaccctaaattactttttttttttccttaaagggagcgtgggcctcaaacagctgtggtc 7700
7640 7650 7660 7670 7680 7690
atgtcatttcagtttgcagtttctctcacctcagaaaaataaattgtgtgggggagatgaacttttttaa 7770
7710 7720 7730 7740 7750 7760
actggaattttattagctgattgaacagctaaaattggtcatttaattagaagggcagtaagtctctccc 7840
7780 7790 7800 7810 7820 7830
tctcactgaattttctcttagtacatacagaatagttgaaaagactttgctttaaaaagtagtgtttttg 7910
7850 7860 7870 7880 7890 7900
gtttttgttacccagttaatagcttattaaatcaggaagatgatgtaatggctttagaagtcactcactc 7980
7920 7930 7940 7950 7960 7970
atgcatctttcttccgcccctttgccaggaaacattaacttccttgtttctccatattgactctttaattt 8050
7990 8000 8010 8020 8030 8040
gacttactacatttgctatattttacctatagatagattttggttgtagggttgtgttgtgtttcgtttc 8120
8060 8070 8080 8090 8100 8110
ttttgaagatttttttttttttaaatttacattgtaggtcatagaccacaactgtttagtgccttttt 8190
8130 8140 8150 8160 8170 8180
ggccttttccctttaataatatgtctacaggtcaccagtcaattcctccaaaggttgaattttaacaa 8260
8200 8210 8220 8230 8240 8250
accttctttgcttcctggataactgggattagaatttAGTTGATGTAGCCGATGCCTGTTGGAGGTGAAG 8330
8270 8280 8290 8300 8310 8320
AAGTCGGGCCAGTCTAACCTCTAGAATACTAGCTTTTGCCTTTCACTTCTGCAATTTGGGAGGAGGTTAA 8400
8340 8350 8360 8370 8380 8390
ACTGTTGTGAGGCCTTCACATGGCCAATTGGGTGCTGgtaagacttgtttttgtaactagctaagaatgc 8470
8410 8420 8430 8440 8450 8460
ttaattctagacagctgtggggtgggaattagaaatgtcaagcatgttagaaataaaaccagagcagggt 8540
8480 8490 8500 8510 8520 8530
ttttcctcccttctattcagttttgaaagaggccctaatgcctgattggttttgcttttgagcccaggtt 8610
8550 8560 8570 8580 8590 8600
ggagggaaatgcatgggttttcttaattaagataatttcttaaaaaggtgttatatcacttaagccactt 8680
8620 8630 8640 8650 8660 8670
tattgttattttgatgctgcttaacttgcacagtaaaaacatgtttaaaacttttcttgtagacttaaat 8750
8690 8700 8710 8720 8730 8740
gtataagatggtagtctaatgagctggggtatgtttgctttgttcactcttaatggattgtgctaacag 8820
8760 8770 8780 8790 8800 8810
tttatacaaatgctagaaacggaggtgtcatttagccgtcattagccatgaacaaacctgacctttttatt 8890
8830 8840 8850 8860 8870 8880

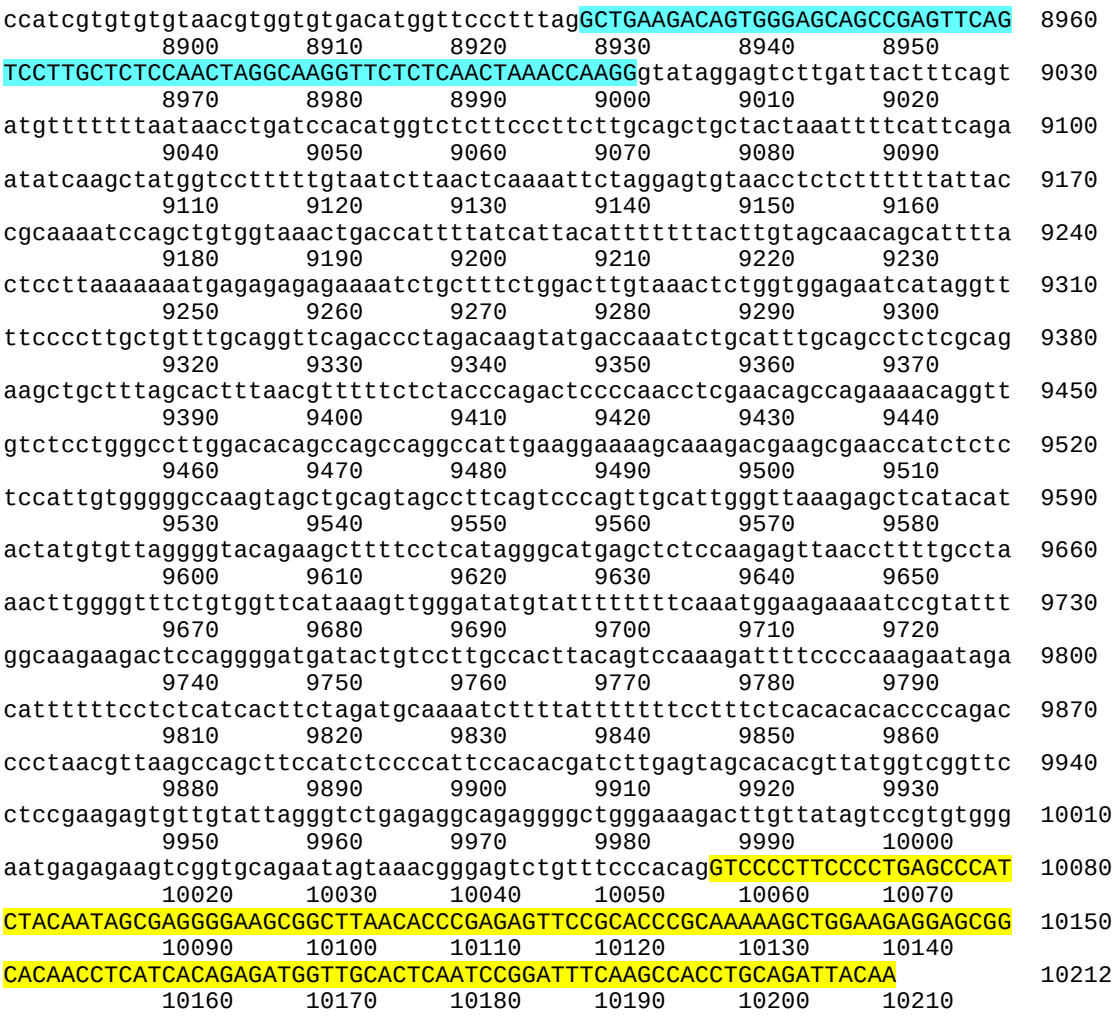


Figure 29. Summary of AP prediction in SF1 gene.

The SF1 DNA sequence (2,075 bp upstream and 8,138 bp downstream of the start codon, for a total length of 10,212 bp) used for the screening of APs is shown. Introns are reported in lower case, whereas exons are in upper case. The start codon in exon 1 is boxed. Constitutive exons are highlighted in yellow, known alternative exons in green and new alternative exons in light blue. Underlined sequences are the putative promoters identified by PROMOTER SCAN; sequences with putative promoter activity predicted by CpG Island Searcher and Cister are shown in bold and in italic, respectively.

3.3.4. Analysis of the transcription activity of SF1 genomic sequences

The prediction analyses shown above identified two main regions in the SF1 gene with a high probability to function as promoters. The first region encompasses the 5' UTR, exon 1 and intron 1_2 sequences and corresponds to the promoter of full-length transcripts, whereas the second one encompasses intron 3_3A, exon 3 and intron 3A_3B sequences.

To confirm the hypothesis of the presence of APs in SF1 gene, we tested the transcription activity of different portions of gSF1 sequences in human cell lines with a luciferase assay. Eight different regions of the SF1 gene were amplified according to the prediction studies and cloned into the pGL3 vector N-terminal of the luciferase gene (figure 30A). A list of DNA constructs used for the assay is shown below and the indicated bp numbers refer to the gSF1 sequence shown above.

- plasmid 1: 2 kb upstream of exon 1 and the first 92 bp of exon 1 (bp 130 – 1,714);
- plasmid 2: 2 kb upstream of exon 1 and the first 417 bp of exon 1, upstream of the ATG (bp 130 – 2,039);
- plasmid 3: the first 417 bp of exon 1, upstream of the ATG (bp 1,632 – 2,074);
- plasmid 4: exon 1 and the first 850 bp of intron 1_2 (bp 1,632 – 2,955);
- plasmid 5: exon 2 and 945 and 894 bp of flanking upstream and downstream introns (bp 2,895 – 4,853);
- plasmid 6: 1,376 bp of intron 2_3 sequence (bp 4,837 – 6,212);
- plasmid 7: exon 3 and 824 and 1,147 bp of upstream and downstream flanking introns (bp 6,139 – 8,185);
- plasmid 8: 27 bp of intron 3_3A, the complete exon 3A, intron 3A_3B, exon 3B, intron 3B_4 and the first 35 bp of exon 4 (bp 8,271 – 10,094).

Empty vector and plasmids containing SF1 sequences were transiently transfected into 293T cells together with a Renilla construct, which was used as a control of transfection efficiency. Cells were harvested after 48 h and luciferase activity was measured immediately.

Each transfection experiment was performed in triplicate and all experiments were repeated three times. The data were normalized to Renilla luciferase activity and expressed as a fold change compared to empty vector (figure 30B).

The strongest luciferase activity was detected when plasmid 2, which consists of the 5’ portion of the exon 1 sequence and 2 kb upstream of exon 1, was transfected into 293T cells (figure 30B). This sequence contains the promoter of the known SF1 transcripts and, according to the bioinformatic tools also has the highest probability to function as a promoter (figures 26A, 27, 28). Deletion of the 3’ part of exon 1 strongly reduces luciferase expression (plasmid 1), suggesting that exon 1 carries important information for promoter activity (figure 30B). This observation is confirmed by the levels of promoter activity of exon 1 alone (plasmid 3), which reach values comparable to those of plasmid 1 (figure 30B).

Luciferase activity was further increased when exon 1 and part of intron 1_2 were fused to the luciferase gene (plasmid 4) and correlates with the prediction studies that estimated a promoter activity in exon 1 and the downstream intron (figures 26A, 27, 28, 29). Intron 1_2 has a high content of GC-rich sequences, a characteristic usually associated with mammalian promoters (Larsen et al, 1992).

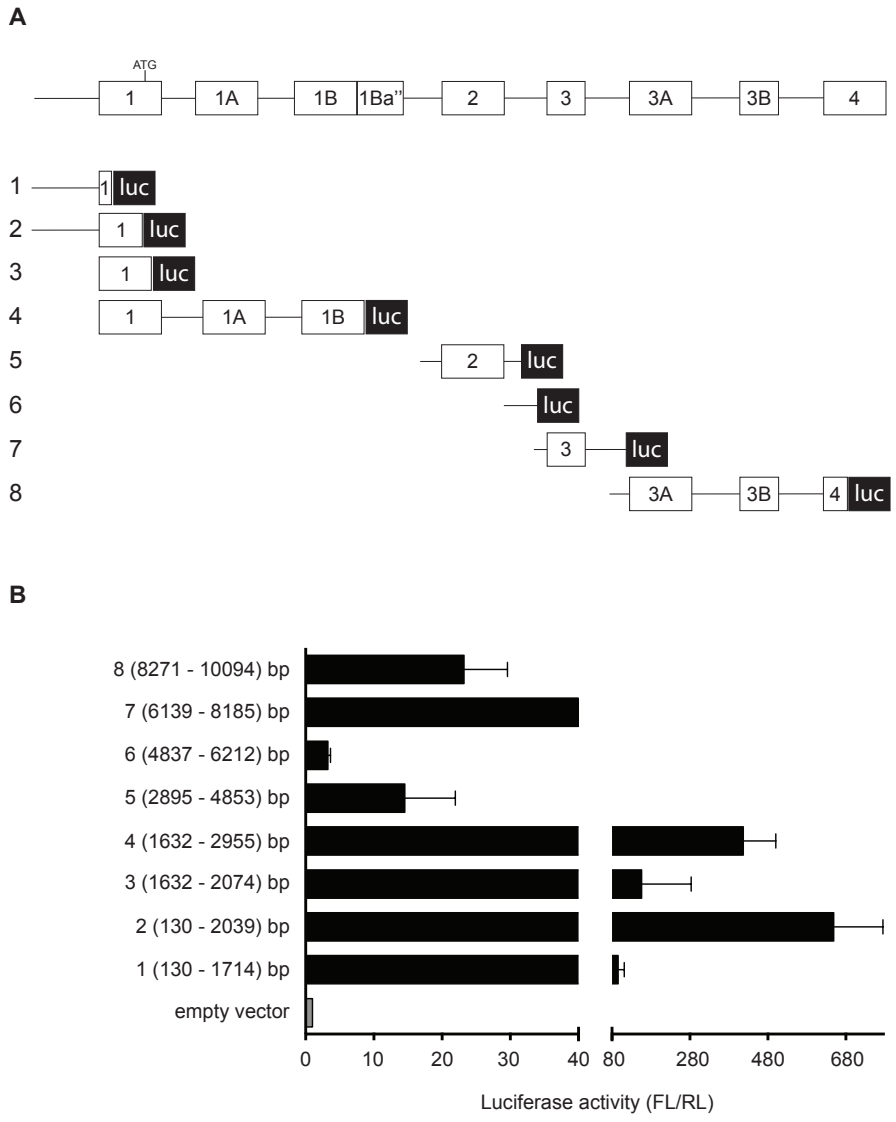


Figure 30. Transcription activity of SF1 DNA sequences.

(A) Part of the gSF1 sequence is shown. Exons and alternative exons are represented as boxes and introns as lines. SF1 DNA sequences cloned upstream of the luciferase gene are shown. (B) Luciferase assays were performed in triplicate and the firefly (FL) luciferase values were normalized to Renilla (RL) luciferase values and expressed as the change fold compared to empty vector. Nucleotide numbers of each clone are shown on the left.

Expression of plasmids 5 and 6, containing exon 2 sequences and flanking introns, resulted in low levels of luciferase activity in agreement with the bioinformatic predictions (figure 30B). The observed luciferase activity of plasmid 5 may be due to residual GC-rich sequences present at the 5' of exon 2. Indeed, the absence of these sequences resulted in a decrease of luciferase expression to background levels (plasmid 6).

High luciferase activity was also observed with plasmids 7 and 8 (figure 30B). Plasmid 7 includes exon 3 and part of the flanking introns, which, according to Cister, might contain regulatory elements, whereas the genomic sequences of plasmid 8 (exons 3A, 3B and introns 3A_3B, 3B_4) were predicted by PROMOTER SCAN to function as a promoter, although with a low score (figure 26C).

In summary, the results obtained with the luciferase assay are consistent with the prediction analysis: the strongest luciferase activity was associated with plasmids containing the region upstream of exon 1, exon 1 itself and part of the first intron. Significant, but lower luciferase activity was also observed with clones containing exons 3, 3A and 3B and introns 3_3A, 3A_3B.

3.3.5. Analysis of the SF1 transcription start site (TSS)

Analyses performed with bioinformatic tools and luciferase assays support the hypothesis that the SF1 gene might have APs. Therefore, we tried to identify the exact TSS of SF1 transcripts containing alternative exons 3A and 3B by performing 5' RACE experiments, based on the oligo-capping method. Briefly, cytoplasmic HeLa RNA was extracted, decapped and ligated to a RNA adapter at the 5' end (figure 19). After conversion into cDNA, nested PCR was performed with forward primers complementary to the 5' RNA adapter and reverse primers specific for the SF1 sequence (figure 31A).

First, to determine the TSSs of full-length SF1 transcripts, 5' RACE was performed and followed by nested PCR with reverse primers complementary to exon 6 and exon 4, for the outer and inner PCR, respectively. As shown in figure 31B, inner RT-PCR amplification gives rise to a band of about 400 bp. As a negative control, a cDNA sample not treated with TAP was used (figure 31B, C and D, lane -). RNAs not treated with TAP have an intact cap structure, which prevents ligation with the RNA adapter and thus RT-PCR amplification. The 400-bp band was gel-purified, cloned into pGEM-T Easy and twelve different clones were analyzed by sequencing. All of them are SF1-specific and contain exons 1, 2, 3 and 4.

However, inspection of the sequences did not identify a single TSS, but ten different TSSs in exon 1 (figure 32, red nucleotides). Some of these were classified as major TSSs, because they were present in several colonies. Bands of about 300 and 1,000 bp were not sequenced (figure 31B, bands marked with white asterisks).

The presence of multiple TSSs in the SF1 gene is consistent with genome-wide analyses performed on mammalian promoters, which showed that only 10-15% of genes have TATA-boxes and their transcription initiates at a defined site (Carninci et al, 2006). The majority of mammalian promoters, such as that of the SF1 gene, are characterized by CpG islands and dispersed TSSs in a region of about 50 to 100 bp (Carninci et al, 2006; Juven-Gershon & Kadonaga, 2010). However, to exclude the possibility that multiple TSSs are due to a technical issue in the PCR amplification of a region with a high GC content (such as SF1 exon 1), we performed nested RT-PCR with LA Takara polymerase supplemented with a buffer specific for GC-rich sequences. PCR amplification using this buffer yields several bands, which were analyzed by sequencing and do not correspond to SF1 (figure 31C, lane +).

Next, 5' RACE for identification of 5' ends of shorter SF1 transcripts was performed. Outer PCR was carried out with a primer specific for exon 5 and inner PCR with primers specific for exons 3A or 3B (figure 31A). Amplification with the primer complementary to exon 3A shows several bands between 400 and 1,000 bp, but none of these are SF1-specific (figure 31D, lane +). Amplification with primer specific for exon 3B shows the presence of a main band of 400 bp (figure 31D). The PCR product was gel-purified, cloned and sequenced. Two clones were analyzed and two different TSSs were identified in exon 1 (figure 32, blue nucleotides). All clones share exons 1, 2, 3 and 3B sequences, which correspond to transcript iii annotated in the ExonMine database (figure 17).

In conclusion, we identified multiple TSSs for full-length SF1 mRNAs and transcript iii. According to the clones analyzed, these SF1 transcripts do not share the same TSSs, although they concentrate within the region of 100 nucleotides upstream of the ATG. No TSS has been identified for the transcripts containing exon 3A.

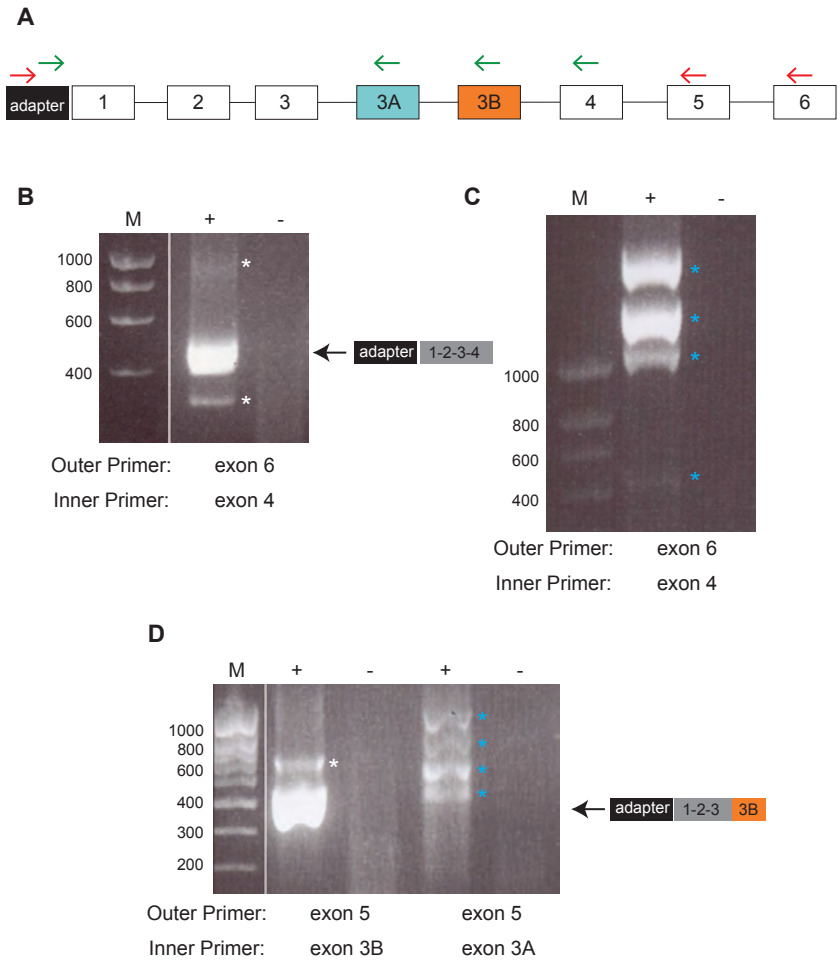


Figure 31. TSS identification of SF1 transcripts.

(A) Part of gSF1 is shown. Empty boxes represent constitutive exons, colored boxes alternative exons and lines introns. The black box corresponds to the 5' adapter ligated during 5' RACE. Red and green arrows indicate the position of outer and inner primers, respectively. (B, C and D) 5' RACE was performed with HeLa RNA and nested PCR was carried out with outer and inner primers indicated under each figure. PCR products were resolved on 2% agarose gels. Lane + indicates TAP-treated RNAs, whereas RNAs not treated with TAP were used as negative controls (-). Arrows indicate SF1-specific transcripts, which were gel-purified, cloned and sequenced. Bands labeled with blue asterisks (*) were shown to be not SF1-specific after sequencing. White asterisks (*) indicate bands that were not sequenced. DNA size markers (M) are indicated on the left in bp.

```
gagctcgccgtcgctccgtcatagagttcgccccatcccctcctttctggactcggagctcagttcacgc  70
           10        20        30        40        50        60
agtaacaatgaagtgcgcgctgcgacacctcccagcccaccgaactccgccgccatttcctcgcttggcc  140
           80        90        100       110       120       130
taacggttcggccaatcccagcgcgcatcaagaaggactgaggctccgccaatcggaggccgccgatttc  210
           150       160       170       180       190       200
gacccttcgcctcggcccggcccaatccaggccccggccccgccgcccccggcccgcccccgcggtgccc  280
           220       230       240       250       260       270
tctctcctccctctttgtgcgtctcgcgccgccgccgcccgccgcgtgagaggacgggctccgcgcgctc  350
           290       300       310       320       330       340
cggcagcgcattcgggtcccctcccccgggaggcttgcgaaggagaagccgccgcaggaggaaagcagg  420
           360       370       380       390       400       410
tgccggtgcctgtccccgggggcgccATGGCGACCGGAGCGAACGCCACGCCGTTGG               477
           430       440       450       460       470
```

Figure 32. Summary of TSSs identified in SF1 transcripts.

The sequence of SF1 exon 1 with the ATG codon in bold is shown. The 5' UTR is shown in lower case, the coding sequence is in upper case. Nucleotides labeled in red are TSSs identified by 5' RACE for the full-length SF1 transcripts (figure 31B). Red and underlined nucleotides represent the major TSS, found in several clones. Blue nucleotides are TSSs of SF1 transcripts containing the alternative exon 3B (figure 31D).

3.3.6. Analysis of N-terminal SF1 isoforms at the protein level

So far we analyzed alternative SF1 variants at the RNA level. We have shown that HeLa and 293T cells express several SF1 transcripts containing exons 3A and/or 3B and that two of these are associated with polysomes. Although we failed to identify their 5' sequences with 5' RACE, these transcripts were not amplified with a forward primer specific for exon 2, suggesting that they do not contain the first exons. Their translation could occur from an AUG codon present in exon 4, leading to a Δ115 SF1 isoform lacking the NLS and ULM domains.

We therefore proceeded with the analysis of a potential Δ115 SF1 isoform at the protein level with a strategy that takes advantage of differences in the N-terminal part of the proteins. The Δ115 SF1 isoform contains the epitope (in the KH domain) necessary to bind the monoclonal 24D1 anti-SF1 antibody (Corioni et al, 2011), but not the commercial anti-SF1 Abnova antibody raised against the N-terminal 110 amino acids of SF1 (figure 33).

Figure 33. SF1 epitopes recognized by Abnova and 24D1 antibody.

Schematic representation of full length (top) and N-terminal (bottom) SF1 isoforms. The epitopes of anti-SF1 Abnova (yellow) and 24D1 (violet) antibodies are located in the first 110 amino acids of SF1 and the KH domain, respectively. IP performed with Dynabeads coupled with anti-SF1 Abnova antibody depletes full-length SF1 isoforms from the extract and leads to an enrichment of shorter SF1 isoforms in the unbound fraction.

Therefore, we decided to deplete SF1 isoforms containing the first 110 amino acids by incubating HeLa and 293T total extracts with Dynabeads coupled to the SF1 Abnova antibody and analyze the unbound fractions with the 24D1 antibody to detect potential residual SF1 isoforms. For efficient depletion total extracts were incubated three times with antibody-coated Dynabeads. The beads were washed and bound material was eluted and analyzed by SDS-PAGE together with input and unbound fractions. Western blotting with the Abnova antibody was performed to confirm SF1 depletion. Three major bands of about 70-90 kDa were detected in the input lane (I), which correspond to the main SF1 isoforms (figure 34). Unfortunately, due to the complexity of SF1 isoforms, it is impossible to determine with certainty which isoform is represented by a given band. Most of the SF1 isoforms bind to the Dynabeads in the first incubation step (B1), whereas only very little is present in the following two incubations (B2 and B3) (figure 34, top panel). No SF1 isoforms are detected with the Abnova antibody in the unbound fraction (U3) collected after the third incubation, indicating that the IP was efficient (figure 34, top panel). In parallel, Western blot analysis with the 24D1 antibody was performed. It shows three main bands in the input and only two in the first bound fraction, suggesting that SF1 isoforms of the middle band did not bind efficiently to the Abnova antibody. A significantly lower amount of SF1 isoforms is detected in B2 and B3. In contrast to the membrane incubated with the Abnova

antibody, a band of about 80 kDa is detected in U3 with the 24D1 antibody in both HeLa and 293T cells (figure 34). This band might correspond to the Δ115 SF1 isoform.

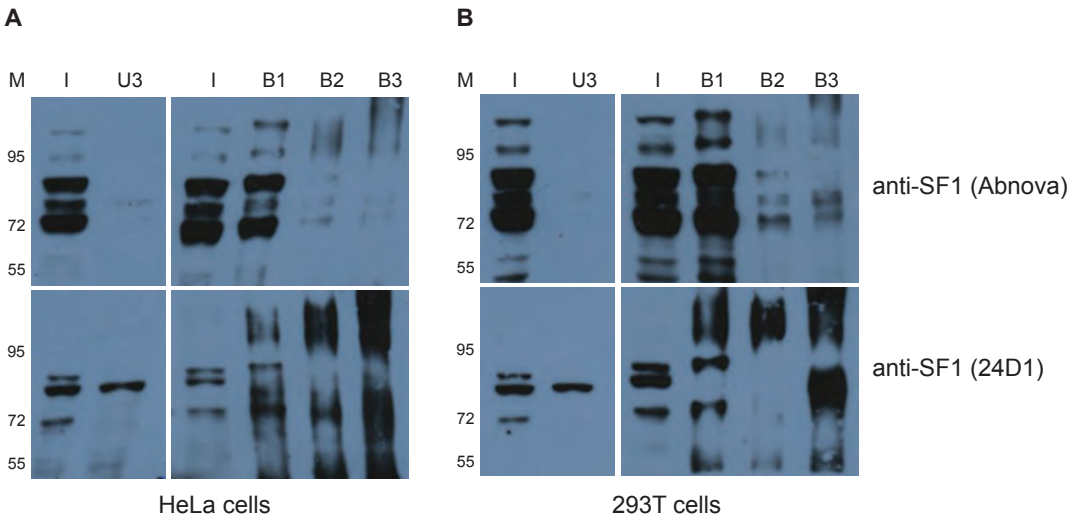


Figure 34. Immunoprecipitation of SF1 isoforms.

Immunoprecipitation was performed with Dynabeads Protein G coupled with anti-SF1 Abnova antibody with HeLa and 293T total extracts. Input (I; 5% of the total), bound (B, 10%) and unbound (U, 5%) material was separated by 7.5% SDS-PAGE and Western blotted with anti-SF1 Abnova and 24D1 antibodies. The migration of protein markers is indicated in kDa on the left.

Although the C-terminal amino acid sequence of a potential Δ115 SF1 isoform is not known, the predicted molecular weight (40-55 kDa) is smaller than the single band detected in the unbound fraction of the IP experiment. However, the Δ115 SF1 isoform might undergo post-translational modification, which can affect the migration of proteins in SDS gels. To test this possibility, we cloned the coding sequence of the Δ115 SF1 isoform into pcDNA3.1/HisC, containing an N-terminal His_6-tag, which does not alter much the size of the protein. Since the C-terminal part of this isoform is not known at present, we cloned the coding sequence of the two longest Ala- and Pro-rich SF1 isoforms, SF1-B3 and SF1-HL. According to their amino acid sequence and their N-terminal tag, the molecular weight of the over-expressed proteins should be 56.16 and 58.55 kDa, respectively. The plasmids were transfected into 293T cells and total extracts were analyzed by SDS-PAGE. Western blot analysis with anti-His_6 shows two bands of about 60-75 kDa in the SF1-HL and B3 samples,

corresponding to the transiently expressed SF1 isoforms (figure 35, left panel). No His_6-tagged proteins were expressed when cells were mock-treated or transfected with empty vector. Western blot analysis with anti-SF1 24D1 detects both endogenous and transiently expressed SF1 isoforms (figure 35, right panel). Although the over-expressed proteins migrate higher compared to their theoretical molecular weight, none of them corresponds to the ~80-kDa SF1 isoform identified by IP in the unbound fraction (cf. figures 34 and 35).

Overall our results have shown the existence of an SF1 isoform of about 80 kDa recognized by anti-SF1 24D1 (i.e. containing the KH domain) and not by anti-SF1 Abnova (i.e. lacking the first 110 amino acids). Therefore, this protein could correspond to the Δ115 SF1 isoform. However, over-expression of the coding sequence of the Δ115 SF1 isoform and analysis by SDS-PAGE revealed that the two proteins do not have the same molecular weight and therefore we exclude this possibility. Since no other 24D1-reactive bands have been detected in the unbound fraction, our data could not confirm the expression of the Δ115 SF1 isoform at the protein level.

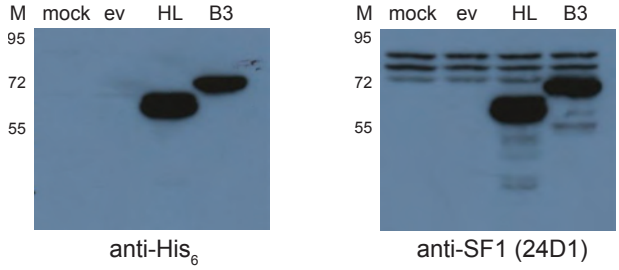


Figure 35. Analysis of the molecular weight of N-terminal SF1 isoforms.

Empty vector (ev) and two His-tagged SF1 isoforms (HL and B3) starting from the AUG present in exon 4 were transiently transfected into 293T cells. The mock sample was incubated only with the transfection reagent. 30 µg of total extract were separated by 10% SDS-PAGE and incubated with anti-His_6 and anti-SF1 24D1 antibodies. The migration of protein markers is indicated in kDa on the left.

3.4. Conclusion

In conclusion, we have verified the existence of several SF1 transcripts containing alternative exons 3A, 3B and retained introns in two human cell lines and demonstrated that only two of these transcripts are associated with polysomes. In this thesis, we have also shown that these alternative SF1 transcripts are very likely transcribed by an AP, although we have not been able to identify their 5' sequences. Moreover, our data indicate that the SF1 gene is characterized, similarly to most mammalian genes, by dispersed TSSs. We have also analyzed the SF1 gene for the presence of APs and have identified two main regions with promoter activity: one of them encompasses the 5' UTR, exon 1 and intron 1_2, whereas the other one contains exon 3, intron 3_3A and intron 3A_3B sequences.

Discussion

4. DISCUSSION

4.1. Analysis of SF1 interacting partners

The included manuscript already contains a discussion and therefore here I will discuss the supplementary results added in the thesis, trying to explain them all together.

The commitment of an intron to undergo the splicing reaction starts with the interaction of spliceosome components with specific pre-mRNA sequences, such as BPS, 3' and 5' ss. The BPS contains the conserved adenosine required for the first trans-esterification reaction and during spliceosome assembly it is recognized by several spliceosome components, such as SF1 and U2 snRNP (Will & Luhrmann, 2011). First, in the E complex, the KH domain of SF1 makes contacts with the pre-mRNA and the BPS adenosine is buried in a hydrophobic pocket, which prevents any interaction of the adenosine with other molecules (Liu et al, 2001). The SF1-BPS interaction leads to a "pre-bulged" conformation of the pre-mRNA that facilitates the base pairing interaction of U2 snRNA with nucleotides adjacent the BPS adenosine in the A complex (Liu et al, 2001). In addition, SF1 was shown to promote the kinetics of the splicing reaction by facilitating U2 snRNP binding to the pre-mRNA and it was hypothesized that it might exert this function by interacting with other splicing components, such as the U2 snRNP (Guth & Valcarcel, 2000; Rutz & Seraphin, 1999). However, the exact molecular mechanism by which this process can occur has been never described.

4.1.1. SF1 interaction with SURP modules

Results presented in this thesis show that SF1 interacts *in vitro* and in mammalian cells with two SURP domain-containing proteins, CHERP and SF3a120. Both proteins associate with U2 snRNP and contain one or two SURP domains, respectively (Krämer et al, 1995; Kuwasako et al, 2006; Will et al, 2002). Our data confirm the analysis obtained from Y2H screens and show that SURP modules are necessary to mediate the interaction of CHERP and SF3a120 with SF1. However, the Y2H data demonstrate that SF1 does not bind to all SURP modules, but only to a subset of them. A good example is represented by SF3a120,

which has two SURP domains in the N-terminal part of the protein. Although they share a similar protein structure, they have different properties and binding specificity (Kuwasako et al, 2006). Our data demonstrate that SF1 interacts with SF3a120 SURP1 and not with SURP2, indicating that the SF1 binding is specific for the first SURP domain of SF3a120.

SURP domains contain around 40 amino acids; among them, twelve are highly conserved and play an important role in the formation of the tertiary structure (Kuwasako et al, 2006). Phylogenic analyses have shown that SURP domains can be divided into two subgroups (Kuwasako et al, 2006). The first subgroup contains SF3a120 SURP1, SWAP SURP2 and the single SURP domains of CHERP and SR140, whereas SF3a120 SURP2 and SWAP SURP1 belong to the second subgroup. Interestingly, SF1 binds to SURP modules present in subgroup 1, suggesting that they might share some important features required for the interaction. For SF3a60-SF3a120 binding it was shown that a key amino acid of SURP2 (Leu169) is essential for the protein-protein interaction (Kuwasako et al, 2006). Mutation of this single amino acid strongly alters the SURP2 properties and the binding to SF3a60. It is possible that a similar mechanism occurs also for the interaction between SF1 and SURP1 and further investigations are required for the identification of amino acids crucial for the binding.

4.1.2. Role of SF1 in recruiting U2 snRNP to the pre-mRNA

Our data have revealed a novel protein-protein interaction between two spliceosome components, SF1 and the U2 snRNP. This finding leads to some re-evaluations regarding the molecular dynamics occurring during the early steps of the spliceosome assembly and suggests that SF1 might promote the association of the U2 snRNP with the pre-mRNA.

Previously, it has been shown that the UHM of U2AF65 recruits the U2 snRNP to the pre-mRNA by direct interaction with SF3b155 (Cass & Berglund, 2006; Gozani et al, 1998). However, these data are in contrast to past studies that have detected the association of U2 snRNP in the E complex in a BPS-independent manner (Das et al, 2000; Hong et al, 1997; Makarov et al, 2012). As described in section 1.4.3, SF1 and U2 snRNP recognize the same pre-mRNA sequence (i.e. the BPS) and compete with one another for the same binding site in U2AF65 (Berglund et al, 1998a; Cass & Berglund, 2006; Rain et al, 1998; Selenko et al, 2003). Since the UHM of U2AF65 interacts with SF1 in the E complex, it is very unlikely that U2AF65 would be responsible for the early recruitment of U2 snRNP in the E complex.

These data, therefore, suggest that U2 snRNP interacts with other spliceosomal components, which can facilitate its association with the pre-mRNA. It has been proposed that SF1 might be one of them and this hypothesis was further supported by studies on the kinetic role of SF1 during the spliceosome assembly (Guth & Valcarcel, 2000). It was shown that low amount of SF1 leads to reduced A complex levels and that SF1 promotes the conversion from E to A complex with an unknown molecular mechanism (Guth & Valcarcel, 2000; Rain et al, 1998). Data presented in this thesis confirm that SF1 depletion strongly reduces A complex formation and show a significant decrease of U2 snRNA base pairing to the BPS in the absence of SF1. Our findings are consistent with previous studies and provide further evidence on the important role of SF1 in promoting the interaction of U2 snRNP with the BPS.

Our data have also shown that the SURP-ID is required for efficient A complex formation and that its deletion partially compromises the U2 snRNP interaction with the pre-mRNA. We have shown that the SURP-ID of SF1 is an evolutionarily conserved domain and binds to two SURP domain-containing proteins associated with U2 snRNP. The SURP-ID and hence the SF1-U2 snRNP interaction are required for efficient association of U2 snRNP with the BPS. Therefore, our data show that SF1 recruits the particle to the pre-mRNA by interacting with U2 snRNP components, i.e. SF3a120 and CHERP. Our data are in agreement with previous work performed by Rain and colleagues (Rain et al, 1998). They have observed a complete reconstitution of A complex in an SF1-depleted nuclear extract by addition of recombinant SF1 proteins containing the complete or the partial SURP-ID (SF1 C5, aa 2-465; SF1 C4, aa 2-320). Addition of a recombinant protein lacking the SURP-ID and the zinc knuckle (SF1 C3, aa 2-278) lead to slightly reduced levels of SF1 activity. However, internal deletion of the zinc knuckle showed A complex levels comparable to SF1 C5 and C4. Taking in consideration the results presented in this thesis, we conclude that the reduced splicing activity observed with SF1 C3 is not related to the absence of the zinc knuckle, but to that of the SURP-ID, which at that time had not been identified.

Although some reports suggested that SF1 might facilitate U2 snRNP recruitment, experimental evidence to confirm this hypothesis was lacking, which is now provided with this study. We propose a mechanism through which SF1 can function in the early recruiting of U2 snRNP to the pre-mRNA by protein-protein interactions and therefore exerting a kinetic role during spliceosome assembly. The U2 snRNP recruitment model proposed in this thesis might not apply to every pre-mRNA, but most likely depends on the nature of the

BPS. Consistent with previous reports (Guth & Valcarcel, 2000), we have shown that SF1 plays an important role in facilitating U2 snRNP base pairing with pre-mRNAs containing a weak rather than consensus BPS. For these substrates, which are quite abundant in the human genome, the U2 snRNA/BPS duplex is further destabilized by several mismatches and therefore several proteins might be required to promote A complex formation.

Results presented in this thesis provide novel insight into the early steps of the spliceosome assembly and suggest a novel mechanism through which SF1 can regulate AS decisions. Modulation of splicing kinetics is crucial for AS decisions and is regulated by several splicing factors (Wang et al, 2010; Witten & Ule, 2011). By recruiting U2 snRNP to the BPS and promoting the formation of the A complex, SF1 might prevent the binding of other splicing regulators to the pre-mRNA and therefore reduce their effects on alternative splice site choice. A similar mechanism has been proposed for TIA proteins, which promote A complex formation by enhancing the U1 snRNP interaction with the 5' ss (Wang et al, 2010; Witten & Ule, 2011).

Previous studies have shown that SF1 can both enhance and silence alternative splicing decisions and SF1 has been shown to bind not only to the BPS, but also to 3' UTRs, exons and all over introns (Corioni et al, 2011). Similar to other splicing regulators, such as TIA and Nova proteins, SF1 might regulate splicing decisions according to the position where it binds on the pre-mRNA (Ule et al, 2006; Wang et al, 2010; Witten & Ule, 2011). The data presented here support a model in which SF1 recognizes the BPS and enhances exon inclusion by promoting the U2 snRNP interaction with the pre-mRNA. On the other hand, SF1 can have an opposite function in AS decisions and therefore promote exon skipping when binding elsewhere on the pre-mRNA.

4.1.3. Analysis of the interaction between SF3a120 and SF3a60

Our data show that SF3a120-SURP1 (aa 38-113) weakly binds to SF3a60. The interaction between SF3a120 and SF3a60 has been extensively studied at the structural and biochemical levels. A NMR study performed with SF3a120 and SF3a60 peptides (aa 134-217 and 71-107, respectively) has shown that the SURP2 module (aa 166-208) makes contact with SF3a60 and is sufficient for the interaction (Kuwasako et al, 2006). Biochemical studies have shown that additional sequences, besides the one of the SURP domain, are involved in the binding and that amino acids 145-224 of SF3a120 are sufficient for the interaction with

SF3a60 (Huang et al, 2011). Surprisingly, our data show a very weak binding between amino acids 38-113 of SF3a120 and residues 1-107 of SF3a60, which is in contrast to previous studies. It has been shown by GST-pull down experiments that SF3a60 does not interact with SURP1 (aa 50-103) (Huang et al, 2011). In addition, Kuwasako and colleagues did not detect any binding between SURP1 (aa 48-110) and SF3a60 (71-107) upon protein co-expression (Huang et al, 2011). It might be possible that the binding observed in our experiments is unspecific, since only a low amount of SF3a60 co-purified with the highly over-expressed SURP1 domain, whereas SF3a60 co-purified in a 1:1 ratio with SURP2. However, the possibility that additional sequences situated between the two SURP modules (aa 104-163) could contribute in a minor role to the interaction should also be taken into consideration. Previous biochemical studies have not tested whether amino acids 104-163 contribute to and/or stabilize the binding between the proteins. In addition, the SURP1 domain tested in our co-expression experiment contains three more amino acids compared to the one tested in Kuwasako's work, which might explain the differences observed in the co-expression experiments.

4.2. Analysis of SF1 isoforms

Alternative splicing increases the complexity of the human genome by generating several proteins from a single gene. Initially, analysis of ESTs and microarray experiments estimated that 60-75% of human genes are alternatively spliced (Johnson et al, 2003; Modrek & Lee, 2002). However, these approaches had some limitations (such as protocol differences, low coverage and sensitivity of conventional sequencing) that were overcome with more recent high-throughput sequencing technologies (Pan et al, 2008). To date, it is estimated that 95% of human transcripts are alternative spliced and bioinformatic tools have been developed to help researchers to deal with the high number of alternative isoforms (Mollet et al, 2010; Richard et al, 2010). One of them is the ExonMine database, which was developed in 2010 with a less-constrained method compared to other AS databases and hence allows the identification of many novel isoforms that have not been detected before (Mollet et al, 2010). ExonMine is based on the alignment of ESTs and mRNA transcripts to the genomic sequence with the aim to summarize all the alternatively spliced transcripts expressed for every single gene (Mollet et al, 2010).

The SF1 pre-mRNA undergoes AS and to date at least six isoforms have been isolated from human cell lines and tissues (Arning et al, 1996; Krämer et al, 1998). SF1 isoforms share all structural domains required for the splicing reaction, which are conserved across many species and located in the N-terminal and central parts of the protein. So far a single isoform that changes the N-terminal portion has been identified in *D. melanogaster* lacking exon 4, which corresponds to human SF1 exon 3 (Mazroui et al, 1999). Skipping of exon 4 leads to a PTC and degradation of the protein by the NMD pathway. A similar transcript has also been described in humans, but no corresponding protein was detected by Western blotting, suggesting that also the mammalian transcript undergoes degradation (Choleza and Krämer, unpublished data).

Thirty different SF1 transcripts have been reported in the ExonMine database. This is not surprising considering that the combination of AS events at the 3' end of the SF1 pre-mRNA can give rise to an enormous number of isoforms. Among the SF1 isoforms, we became interested in transcripts containing alternative exons 3A and/or 3B and lacking exons 1, 2 and 3. It was shown that the majority of un-annotated exons, such as exon 3A, function as first exons and therefore transcription of these alternative SF1 transcripts might be regulated by an AP situated downstream the canonical one (Mollet et al, 2010). Although the transcripts lack the first exon and the start codon, they could still be translated using a second AUG in exon 4, which would give rise to SF1 isoforms lacking the first 115 amino acids that cover the two NLSs and the ULM (Δ115 SF1). Both Met (methionine) codons are in the same frame and hence the isoforms would have the same amino acid sequence. These isoforms would be the first N-terminal isoforms described for SF1 that could be translated without a PTC and NMD degradation and we wished to investigate whether they play a role in splicing and whether they have a different function in the cell.

4.2.1. Identification of alternative SF1 transcripts and their 5' end sequences

Results presented in this study have shown that novel SF1 transcripts containing alternative exons 3A, 3B and introns 3A_3B and 3B_4 are expressed in two different human cell lines derived from cervix (HeLa cells) and kidney (293T cells).

In total, we have identified four different SF1 transcripts, which are summarized in figure 21 (PCR products A, B, C and D; PCR product D corresponds to transcript iii of figure 17). Our results have confirmed the expression of two transcripts annotated in ExonMine (figure 17B,

transcripts i and ii in section 3.1) and have identified additional transcripts that were not reported in the database (cf. figures 17B and 21). The newly identified SF1 transcripts show different expression levels in the cell lines tested, confirming previous findings on tissue-type specific expression of SF1 isoforms (Arning et al, 1996). Our results have also confirmed that a SF1 transcript containing exons 1, 2, 3, 3B, 4, 5 (transcript iii, figure 17) is expressed in human cell lines. Due to the inclusion of alternative exon 3B, this transcript carries a PTC in exon 5 and therefore is very likely degraded by NMD. Although the alternative SF1 transcripts were identified in cytoplasmic RNA, only those that do not retain introns (transcript iii and PCR product A) associate with polysomes (figure 24B).

It has been shown that the majority of un-annotated exons identified by ExonMine, such as the alternative exon 3A, are first exon, which suggests that the SF1 gene may have several promoters (Mollet et al, 2010). The identification of alternative SF1 transcripts described here is based on RT-PCR analysis with SF1-specific primers, which does not allow to conclude whether additional exons are present at the 5' and 3' ends of the amplified product and whether the SF1 gene has APs. However, RT-PCR performed with primers specific for exons 2 and 4 (figure 22) does not amplify any exon 3A-containing transcripts, suggesting that none of the newly identified transcripts contain upstream exons and further supports the hypothesis of the existence of APs in the SF1 gene. We have analyzed the promoter activity of the SF1 gene and shown that genomic sequences corresponding to exon 3 and intron 3A_3B are characterized by weak promoter activity, which could regulate the expression of SF1 transcripts containing exon 3A. Unfortunately, we did not succeed in the identification of their 5' end sequences (figure 31). The question why we were not able to identify their TSS remains to be answered, considering that the transcripts are expressed in human cell lines and associated with polysomes. The approach used in this thesis for the identification of the 5' sequences is based on the oligo-capped method, which consists in replacing the m^7G cap with an RNA adapter. RT-PCR amplification is then carried out with a set of primers complementary to the adapter and to SF1 sequences. Therefore only capped mRNAs are amplified. It is very unlikely that transcripts containing exon 3A do not have the m^7G cap, since all mRNAs are transcribed by RNA Pol II and acquire the cap co-transcriptionally (Cho et al, 1997; McCracken et al, 1997). However, we cannot exclude the possibility of technical problems that could prevent their identification. Hence, the experiments could be repeated with other commercial kits, which are based on different approaches. Two alternative strategies to identify the 5' sequence of pre-mRNAs are offered by Takara and

Life Technologies. The 5'-Full RACE core set (Takara) consists in cDNA production followed by circularization of the cDNA and nested PCR amplification with two sets of sequence-specific primers. Similar to the strategy used in our study, the 5´ RACE System for Rapid Amplification of cDNA Ends (Life Technologies) is based on the addition of a homopolymeric tail to the 3' end of the cDNA, which corresponds to the 5' end of the mRNA. Then, nested PCR amplification is performed with primers complementary to the homopolymeric tail and to the cDNA. In both cases, the mRNA is immediately converted into cDNA and the cloning of the 5' end sequences does not depend on the m^7G cap structure. However, a disadvantage of these techniques is the possibility that the full 5' sequences are not cloned because the reverse transcriptase might not reach the end of the transcript.

In contrast, the FirstChoice RLM-RACE Kit (Applied Biosystem) was used with success for the cloning of 5' end sequence of transcript iii. Several TSSs were identified in exon 1, which shows that the SF1 gene belongs to the class of mammalian genes with promoters characterized by CpG islands and dispersed initiation (Carninci et al, 2006; Juven-Gershon & Kadonaga, 2010). However, our results are not consistent with other data presented in this study. No PCR product containing exon 3B was detected when RT-PCR was performed with primers complementary to exons 2 and 4 and transcript iii could be amplified only when a primer specific for exon 3B was used. We speculate that transcript iii might have an unfavorable secondary structure and/or a low expression level, which may affect its amplification with SF1 primers other than those specific for exon 3B.

4.2.2. Identification of SF1 N-terminal isoforms

We have also investigated the expression of the Δ115 SF1 isoform at the protein level in human cell lines. Identification of individual SF1 isoforms by Western blotting is complex, because isoforms share most of the amino acid sequence and any anti-SF1 antibody recognizes more than one isoforms. To distinguish between SF1 isoforms containing or lacking the N terminus, a strategy based on the combination of two different anti-SF1 antibodies (Abnova and 24D1) was used. The commercial anti-SF1 Abnova antibody was raised against the first 110 amino acids of SF1, whereas the epitope of anti-SF1 24D1 is located in the KH domain in the central part of the protein (Corioni et al, 2011). Since the Δ115 SF1 isoform lacks the first 115 amino acids, it also misses the epitope for the

recognition by the anti-SF1 Abnova antibody. Our results indicate that HeLa and 293T cells express a protein of ~80 kDa, which retains the epitope for the interaction with anti-SF1 24D1 but not the one for binding to anti-SF1 Abnova. This protein could correspond to the Δ115 SF1 isoform, but this possibility was excluded due to differences in the molecular weight.

In conclusion, the results presented in this study do not confirm the expression of the Δ115 SF1 isoform, despite the fact that corresponding mRNAs are transcribed and associated with polysomes. It is possible that the expression levels of the protein are below the detection limit and therefore Western blotting analysis would not reveal it.

4.2.3. Alternative promoters in SF1 gene

APs, together with AS and APA, contribute to increase proteome complexity. AP usage is very common in the human genome and it has been estimated that almost 40-50% of human and mouse genes have APs (Baek et al, 2007; Davuluri et al, 2008; Sun et al, 2011). Analysis of APs in the SF1 gene has been performed with bioinformatic tools and a luciferase assay. Our results show the presence of two main regions with promoter activity. A weak promoter is predicted in the region spanning exon 3, intron 3_3A and intron 3A_3B, whereas a strong one has been identified in the region encompassing the 5' UTR, exon 1 and intron 1_2. The first one might be responsible for the transcription of the alternative SF1 variants described above, while the region spanning the 5' UTR, exon 1 and intron 1_2 contains the canonical SF1 promoter of full-length transcripts. However, the luciferase assay experiments have demonstrated high promoter activity also in the intron 1_2 sequence, suggesting the possibility of two adjacent promoters in this part of the SF1 gene (figure 30, cf. clones 1 and 4, as well as 3 and 4). This hypothesis is further supported by a high content of GC-rich sequences in the first intron, usually a characteristic of mammalian promoters, and by the observation that a SF1 variant lacking only exon 1 has been annotated in the Ensembl and ExonMine databases (Larsen et al, 1992). This SF1 transcript contains two alternative exons 1B and 1Ba'', which are not separated by any intron and located between alternative exon 1A and exon 2 (figure 17A). The analysis performed in this study on the 5' end of SF1 transcripts identified several TSSs in exon 1, but none in exons 1B and 1Ba''. However, nested PCR performed with inner primers complementary to exon 4 shows additional bands, apart form the one that was sequenced, which were not processed (figure 31B, white

asterisks). It would be interesting to investigate whether these bands are SF1-specific and whether they contain TSSs in the alternative exons 1B and 1Ba''.

The transcript containing alternative exons 1B and 1Ba'' lacks the AUG present in the exon 1 and gives rise to a putative SF1 isoform, which initiates at Met 27 in exon 2. Use of this Met residue would generate a SF1 isoform of 613 amino acids, which lacks the first NLS and the ULM. This putative isoform of 65.4 kDa is termed Δ26 SF1 below, since it lacks the first 26 amino acids. However, its expression has never been confirmed experimentally. It is known that SF1 proteins undergo post-translational modifications with subsequent increase of the molecular weight observed by SDS-PAGE and, therefore, the Δ26 SF1 isoform might correspond to the ~80 kDa protein identified in this study (Manceau et al, 2006; Zhang et al, 2013). However, further analyses are required to determine whether anti-SF1 Abnova, which recognizes an epitope in the first 110 amino acids, binds to the Δ26 SF1 isoform. The Δ26 SF1 isoform might not function in splicing since its localization and interaction with U2AF65 are compromised. Previous data from our lab have demonstrated that depletion of only the first NLS is sufficient to affect the nuclear localization of SF1 protein (Bagdiul and Krämer, unpublished results).

In conclusion, results presented in this thesis support the hypothesis that SF1 variants are transcribed from APs and that a N-terminal SF1 isoform lacking the first NLS and ULM (i.e. Δ26 SF1) is expressed in human cell lines. However, further experiments are required to confirm our observations. Identification of the total number of SF1 isoforms is important to clarify the role of SF1 in the cell. So far, several functions have been proposed for SF1 and some of them are not related to splicing, such as nuclear pre-mRNA retention and transcription regulation (Goldstrohm et al, 2001; Rutz & Seraphin, 2000; Zhang & Childs, 1998). SF1 isoforms could have completely different roles in the cell and some of them, like Δ26 SF1, might not be involved at all in splicing. Therefore elucidating the expression pattern of SF1 isoforms and their role in the cell might represent an important step in understanding the different SF1 functions.

Collectively the results presented in this thesis confirm that SF1 mainly functions as a splicing factor and provide novel insight into the mechanisms of early spliceosome assembly. Biochemical analyses provided experimental evidence about the binding of SF1 to SURP domain-containing proteins associated with the U2 snRNP. The interaction between

SF1 and U2 snRNP-associated proteins represents an important discovery that clarifies the kinetic role of SF1 in spliceosome assembly. In addition, data obtained during my thesis have also opened new possibilities about the presence of APs in the SF1 gene, which give rise to novel N-terminal isoforms.

References

5. References

Abovich N, Liao XC, Rosbash M (1994) The yeast MUD2 protein: an interaction with PRP11 defines a bridge between commitment complexes and U2 snRNP addition. *Genes Dev* 8: 843-854

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

Arning S, Gruter P, Bilbe G, Krämer A (1996) Mammalian splicing factor SF1 is encoded by variant cDNAs and binds to RNA. *RNA* 2: 794-810

Baek D, Davis C, Ewing B, Gordon D, Green P (2007) Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. *Genome Res* 17: 145-155

Behrens SE, Galisson F, Legrain P, Luhrmann R (1993a) Evidence that the 60-kDa protein of 17S U2 small nuclear ribonucleoprotein is immunologically and functionally related to the yeast PRP9 splicing factor and is required for the efficient formation of prespliceosomes. *Proc Natl Acad Sci U S A* 90: 8229-8233

Behrens SE, Tyc K, Kastner B, Reichelt J, Luhrmann R (1993b) Small nuclear ribonucleoprotein (RNP) U2 contains numerous additional proteins and has a bipartite RNP structure under splicing conditions. *Mol Cell Biol* 13: 307-319

Behzadnia N, Hartmuth K, Will CL, Luhrmann R (2006) Functional spliceosomal A complexes can be assembled in vitro in the absence of a penta-snRNP. *RNA* 12: 1738-1746

Bentley DL (2014) Coupling mRNA processing with transcription in time and space. *Nat Rev Genet* 15: 163-175

Berget SM, Moore C, Sharp PA (1977) Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A* 74: 3171-3175

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

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

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

Beyer AL, Osheim YN (1988) Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev* 2: 754-765

Braunschweig U, Gueroussov S, Plocik AM, Graveley BR, Blencowe BJ (2013) Dynamic integration of splicing within gene regulatory pathways. *Cell* 152: 1252-1269

Breathnach R, Benoist C, O'Hare K, Gannon F, Chambon P (1978) Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proc Natl Acad Sci U S A* 75: 4853-4857

Bringmann P, Luhrmann R (1986) Purification of the individual snRNPs U1, U2, U5 and U4/U6 from HeLa cells and characterization of their protein constituents. *EMBO J* 5: 3509-3516

Brosi R, Groning K, Behrens SE, Luhrmann R, Krämer A (1993a) Interaction of mammalian splicing factor SF3a with U2 snRNP and relation of its 60-kD subunit to yeast PRP9. *Science* 262: 102-105

Brosi R, Hauri HP, Krämer A (1993b) Separation of splicing factor SF3 into two components and purification of SF3a activity. *J Biol Chem* 268: 17640-17646

Brow DA (2002) Allosteric cascade of spliceosome activation. *Annu Rev Genet* 36: 333-360

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 38: 626-635

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

Cass DM, Berglund JA (2006) The SF3b155 N-terminal domain is a scaffold important for splicing. *Biochemistry* 45: 10092-10101

Champion-Arnaud P, Reed R (1994) The prespliceosome components SAP 49 and SAP 145 interact in a complex implicated in tethering U2 snRNP to the branch site. *Genes Dev* 8: 1974-1983

Cho EJ, Takagi T, Moore CR, Buratowski S (1997) mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. *Genes Dev* 11: 3319-3326

Chow LT, Gelinas RE, Broker TR, Roberts RJ (1977) An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* 12: 1-8

Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, Lawrence JB (2009) An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol Cell* 33: 717-726

Cordin O, Beggs JD (2013) RNA helicases in splicing. *RNA Biol* 10: 83-95

Corioni M, Antih N, Tanackovic G, Zavolan M, Krämer A (2011) Analysis of in situ pre-mRNA targets of human splicing factor SF1 reveals a function in alternative splicing. *Nucleic Acids Res* 39: 1868-1879

Cramer P, Pesce CG, Baralle FE, Kornblihtt AR (1997) Functional association between promoter structure and transcript alternative splicing. *Proc Natl Acad Sci U S A* 94: 11456-11460

Das BK, Xia L, Palandjian L, Gozani O, Chyung Y, Reed R (1999) Characterization of a protein complex containing spliceosomal proteins SAPs 49, 130, 145, and 155. *Mol Cell Biol* 19: 6796-6802

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

Davuluri RV, Suzuki Y, Sugano S, Plass C, Huang TH (2008) The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet* 24: 167-177

Denhez F, Lafyatis R (1994) Conservation of regulated alternative splicing and identification of functional domains in vertebrate homologs to the Drosophila splicing regulator, suppressor-of-white-apricot. *J Biol Chem* 269: 16170-16179

Douglas AG, Wood MJ (2011) RNA splicing: disease and therapy. *Brief Funct Genomics* 10: 151-164

Elkon R, Ugalde AP, Agami R (2013) Alternative cleavage and polyadenylation: extent, regulation and function. *Nat Rev Genet* 14: 496-506

Faustino NA, Cooper TA (2003) Pre-mRNA splicing and human disease. *Genes Dev* 17: 419-437

Fleckner J, Zhang M, Valcarcel J, Green MR (1997) U2AF65 recruits a novel human DEAD box protein required for the U2 snRNP-branchpoint interaction. *Genes Dev* 11: 1864-1872

Frith MC, Hansen U, Weng Z (2001) Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics* 17: 878-889

Goldstrohm AC, Albrecht TR, Sune C, Bedford MT, Garcia-Blanco MA (2001) The transcription elongation factor CA150 interacts with RNA polymerase II and the pre-mRNA splicing factor SF1. *Mol Cell Biol* 21: 7617-7628

Gornemann J, Kotovic KM, Hujer K, Neugebauer KM (2005) Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. *Mol Cell* 19: 53-63

Gozani O, Feld R, Reed R (1996) Evidence that sequence-independent binding of highly conserved U2 snRNP proteins upstream of the branch site is required for assembly of spliceosomal complex A. *Genes Dev* 10: 233-243

Gozani O, Potashkin J, Reed R (1998) A potential role for U2AF-SAP 155 interactions in recruiting U2 snRNP to the branch site. *Mol Cell Biol* 18: 4752-4760

Guth S, Valcarcel J (2000) Kinetic role for mammalian SF1/BBP in spliceosome assembly and function after polypyrimidine tract recognition by U2AF. *J Biol Chem* 275: 38059-38066

Hacker I, Sander B, Golas MM, Wolf E, Karagoz E, Kastner B, Stark H, Fabrizio P, Luhrmann R (2008) Localization of Prp8, Brr2, Snu114 and U4/U6 proteins in the yeast tri-snRNP by electron microscopy. *Nat Struct Mol Biol* 15: 1206-1212

Hannon GJ, Maroney PA, Yu YT, Hannon GE, Nilsen TW (1992) Interaction of U6 snRNA with a sequence required for function of the nematode SL RNA in trans-splicing. *Science* 258: 1775-1780

Hastings ML, Allemand E, Duelli DM, Myers MP, Krainer AR (2007) Control of pre-mRNA splicing by the general splicing factors PUF60 and U2AF(65). *PLoS One* 2: e538

Hausner TP, Giglio LM, Weiner AM (1990) Evidence for base-pairing between mammalian U2 and U6 small nuclear ribonucleoprotein particles. *Genes Dev* 4: 2146-2156

Hocine S, Singer RH, Grunwald D (2010) RNA processing and export. *Cold Spring Harb Perspect Biol* 2: a000752

Hong W, Bennett M, Xiao Y, Feld Kramer R, Wang C, Reed R (1997) Association of U2 snRNP with the spliceosomal complex E. *Nucleic Acids Res* 25: 354-361

Huang CJ, Ferfoglia F, Raleff F, Krämer A (2011) Interaction domains and nuclear targeting signals in subunits of the U2 small nuclear ribonucleoprotein particle-associated splicing factor SF3a. *J Biol Chem* 286: 13106-13114

Ito T, Muto Y, Green MR, Yokoyama S (1999) Solution structures of the first and second RNA-binding domains of human U2 small nuclear ribonucleoprotein particle auxiliary factor (U2AF(65)). *EMBO J* 18: 4523-4534

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302: 2141-2144

Jordan M, Schallhorn A, Wurm FM (1996) Transfecting mammalian cells: optimization of critical parameters affecting calcium-phosphate precipitate formation. *Nucleic Acids Res* 24: 596-601

Juven-Gershon T, Kadonaga JT (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol* 339: 225-229

Kanopka A, Muhlemann O, Akusjarvi G (1996) Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. *Nature* 381: 535-538

Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* 11: 345-355

Kielkopf CL, Rodionova NA, Green MR, Burley SK (2001) A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. *Cell* 106: 595-605

Kim DH, Rossi JJ (1999) The first ATPase domain of the yeast 246-kDa protein is required for in vivo unwinding of the U4/U6 duplex. *RNA* 5: 959-971

Kornblihtt AR (2005) Promoter usage and alternative splicing. *Curr Opin Cell Biol* 17: 262-268

Kornblihtt AR, Schor IE, Allo M, Dujardin G, Petrillo E, Munoz MJ (2013) Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol* 14: 153-165

Kotovic KM, Lockshon D, Boric L, Neugebauer KM (2003) Cotranscriptional recruitment of the U1 snRNP to intron-containing genes in yeast. *Mol Cell Biol* 23: 5768-5779

Krämer A (1992) Purification of splicing factor SF1, a heat-stable protein that functions in the assembly of a presplicing complex. *Mol Cell Biol* 12: 4545-4552

Krämer A, Ferfoglia F, Huang CJ, Mulhaupt F, Nesic D, Tanackovic G (2005) Structure-function analysis of the U2 snRNP-associated splicing factor SF3a. *Biochem Soc Trans* 33: 439-442

Krämer A, Gruter P, Groning K, Kastner B (1999) Combined biochemical and electron microscopic analyses reveal the architecture of the mammalian U2 snRNP. *J Cell Biol* 145: 1355-1368

Krämer A, Mulhauser F, Wersig C, Groning K, Bilbe G (1995) Mammalian splicing factor SF3a120 represents a new member of the SURP family of proteins and is homologous to the essential splicing factor PRP21p of Saccharomyces cerevisiae. *RNA* 1: 260-272

Krämer A, Quentin M, Mulhauser F (1998) Diverse modes of alternative splicing of human splicing factor SF1 deduced from the exon-intron structure of the gene. *Gene* 211: 29-37

Kuersten S, Goodwin EB (2003) The power of the 3' UTR: translational control and development. *Nat Rev Genet* 4: 626-637

Kuwasako K, He F, Inoue M, Tanaka A, Sugano S, Guntert P, Muto Y, Yokoyama S (2006) Solution structures of the SURP domains and the subunit-assembly mechanism within the splicing factor SF3a complex in 17S U2 snRNP. *Structure* 14: 1677-1689

Landry JR, Mager DL, Wilhelm BT (2003) Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet* 19: 640-648

Larsen F, Gundersen G, Lopez R, Prydz H (1992) CpG islands as gene markers in the human genome. *Genomics* 13: 1095-1107

Lin KT, Lu RM, Tarn WY (2004) The WW domain-containing proteins interact with the early spliceosome and participate in pre-mRNA splicing in vivo. *Mol Cell Biol* 24: 9176-9185

Lin PC, Xu RM (2012) Structure and assembly of the SF3a splicing factor complex of U2 snRNP. *EMBO J* 31: 1579-1590

Listerman I, Sapra AK, Neugebauer KM (2006) Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. *Nat Struct Mol Biol* 13: 815-822

Liu Z, Luyten I, Bottomley MJ, Messias AC, Houngninou-Molango S, Sprangers R, Zanier K, Krämer A, Sattler M (2001) Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science* 294: 1098-1102

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

Ma L, Tan Z, Teng Y, Hoersch S, Horvitz HR (2011) In vivo effects on intron retention and exon skipping by the U2AF large subunit and SF1/BBP in the nematode Caenorhabditis elegans. *RNA* 17: 2201-2211

Makarov EM, Owen N, Bottrill A, Makarova OV (2012) Functional mammalian spliceosomal complex E contains SMN complex proteins in addition to U1 and U2 snRNPs. *Nucleic Acids Res* 40: 2639-2652

Malca H, Shomron N, Ast G (2003) The U1 snRNP base pairs with the 5' splice site within a penta-snRNP complex. *Mol Cell Biol* 23: 3442-3455

Manceau V, Swenson M, Le Caer JP, Sobel A, Kielkopf CL, Maucuer A (2006) Major phosphorylation of SF1 on adjacent Ser-Pro motifs enhances interaction with U2AF65. *FEBS J* 273: 577-587

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

Martinez-Contreras R, Cloutier P, Shkreta L, Fisette JF, Revil T, Chabot B (2007) hnRNP proteins and splicing control. *Adv Exp Med Biol* 623: 123-147

Matera AG, Wang Z (2014) A day in the life of the spliceosome. *Nat Rev Mol Cell Biol* 15: 108-121

Mazroui R, Puoti A, Krämer A (1999) Splicing factor SF1 from Drosophila and Caenorhabditis: presence of an N-terminal RS domain and requirement for viability. *RNA* 5: 1615-1631

McCracken S, Fong N, Rosonina E, Yankulov K, Brothers G, Siderovski D, Hessel A, Foster S, Shuman S, Bentley DL (1997) 5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Genes Dev* 11: 3306-3318

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

Mollet IG, Ben-Dov C, Felicio-Silva D, Grosso AR, Eleuterio P, Alves R, Staller R, Silva TS, Carmo-Fonseca M (2010) Unconstrained mining of transcript data reveals increased alternative splicing complexity in the human transcriptome. *Nucleic Acids Res* 38: 4740-4754

Mount SM (1982) A catalogue of splice junction sequences. *Nucleic Acids Res* 10: 459-472

Nagai K, Muto Y, Pomeranz Krummel DA, Kambach C, Ignjatovic T, Walke S, Kuglstatter A (2001) Structure and assembly of the spliceosomal snRNPs. Novartis Medal Lecture. *Biochem Soc Trans* 29: 15-26

Nakagawa S, Naganuma T, Shioi G, Hirose T (2011) Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice. *J Cell Biol* 193: 31-39

Nesic D, Krämer A (2001) Domains in human splicing factors SF3a60 and SF3a66 required for binding to SF3a120, assembly of the 17S U2 snRNP, and prespliceosome formation. *Mol Cell Biol* 21: 6406-6417

Newman AJ, Norman C (1992) U5 snRNA interacts with exon sequences at 5' and 3' splice sites. *Cell* 68: 743-754

Norbury CJ (2013) Cytoplasmic RNA: a case of the tail wagging the dog. *Nat Rev Mol Cell Biol* 14: 643-653

O'Day CL, Dalbadie-McFarland G, Abelson J (1996) The Saccharomyces cerevisiae Prp5 protein has RNA-dependent ATPase activity with specificity for U2 small nuclear RNA. *J Biol Chem* 271: 33261-33267

Ohrt T, Prior M, Dannenberg J, Odenwalder P, Dybkov O, Rasche N, Schmitzova J, Gregor I, Fabrizio P, Enderlein J, Luhrmann R (2012) Prp2-mediated protein rearrangements at the catalytic core of the spliceosome as revealed by dcFCCS. *RNA* 18: 1244-1256

Padgett RA, Konarska MM, Grabowski PJ, Hardy SF, Sharp PA (1984) Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. *Science* 225: 898-903

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40: 1413-1415

Park JW, Parisky K, Celotto AM, Reenan RA, Graveley BR (2004) Identification of alternative splicing regulators by RNA interference in Drosophila. *Proc Natl Acad Sci U S A* 101: 15974-15979

Prestridge DS (1995) Predicting Pol II promoter sequences using transcription factor binding sites. *J Mol Biol* 249: 923-932

Price SR, Evans PR, Nagai K (1998) Crystal structure of the spliceosomal U2B"-U2A' protein complex bound to a fragment of U2 small nuclear RNA. *Nature* 394: 645-650

Raghunathan PL, Guthrie C (1998) RNA unwinding in U4/U6 snRNPs requires ATP hydrolysis and the DEIH-box splicing factor Brr2. *Curr Biol* 8: 847-855

Rain JC, Rafi Z, Rhani Z, Legrain P, Krämer A (1998) Conservation of functional domains involved in RNA binding and protein-protein interactions in human and Saccharomyces cerevisiae pre-mRNA splicing factor SF1. *RNA* 4: 551-565

Rain JC, Tartakoff AM, Krämer A, Legrain P (1996) Essential domains of the PRP21 splicing factor are implicated in the binding to PRP9 and PRP11 proteins and are conserved through evolution. *RNA* 2: 535-550

Richard H, Schulz MH, Sultan M, Nurnberger A, Schrinner S, Balzereit D, Dagand E, Rasche A, Lehrach H, Vingron M, Haas SA, Yaspo ML (2010) Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Res* 38: e112

Ruby JG, Jan CH, Bartel DP (2007) Intronic microRNA precursors that bypass Drosha processing. *Nature* 448: 83-86

Ruskin B, Krainer AR, Maniatis T, Green MR (1984) Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. *Cell* 38: 317-331

Ruskin B, Zamore PD, Green MR (1988) A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. *Cell* 52: 207-219

Rutz B, Seraphin B (1999) Transient interaction of BBP/ScSF1 and Mud2 with the splicing machinery affects the kinetics of spliceosome assembly. *RNA* 5: 819-831

Rutz B, Seraphin B (2000) A dual role for BBP/ScSF1 in nuclear pre-mRNA retention and splicing. *EMBO J* 19: 1873-1886

Rymond B (2007) Targeting the spliceosome. *Nat Chem Biol* 3: 533-535

Sauliere J, Sureau A, Expert-Bezancon A, Marie J (2006) The polypyrimidine tract binding protein (PTB) represses splicing of exon 6B from the beta-tropomyosin pre-mRNA by directly interfering with the binding of the U2AF65 subunit. *Mol Cell Biol* 26: 8755-8769

Selenko P, Gregorovic G, Sprangers R, Stier G, Rhani Z, Krämer A, Sattler M (2003) Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP. *Mol Cell* 11: 965-976

Shi Y (2012) Alternative polyadenylation: new insights from global analyses. *RNA* 18: 2105-2117

Shitashige M, Naishiro Y, Idogawa M, Honda K, Ono M, Hirohashi S, Yamada T (2007a) Involvement of splicing factor-1 in beta-catenin/T-cell factor-4-mediated gene transactivation and pre-mRNA splicing. *Gastroenterology* 132: 1039-1054

Shitashige M, Satow R, Honda K, Ono M, Hirohashi S, Yamada T (2007b) Increased susceptibility of Sf1(+/-) mice to azoxymethane-induced colon tumorigenesis. *Cancer Sci* 98: 1862-1867

Singh RK, Cooper TA (2012) Pre-mRNA splicing in disease and therapeutics. *Trends Mol Med* 18: 472-482

Spadaccini R, Reidt U, Dybkov O, Will C, Frank R, Stier G, Corsini L, Wahl MC, Luhrmann R, Sattler M (2006) Biochemical and NMR analyses of an SF3b155-p14-U2AF-RNA interaction network involved in branch point definition during pre-mRNA splicing. *RNA* 12: 410-425

Spikes DA, Kramer J, Bingham PM, Van Doren K (1994) SWAP pre-mRNA splicing regulators are a novel, ancient protein family sharing a highly conserved sequence motif with the prp21 family of constitutive splicing proteins. *Nucleic Acids Res* 22: 4510-4519

St Laurent G, Shtokalo D, Tackett MR, Yang Z, Eremina T, Wahlestedt C, Urcuqui-Inchima S, Seilheimer B, McCaffrey TA, Kapranov P (2012) Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. *BMC Genomics* 13: 504

Staknis D, Reed R (1994) Direct interactions between pre-mRNA and six U2 small nuclear ribonucleoproteins during spliceosome assembly. *Mol Cell Biol* 14: 2994-3005

Staley JP, Guthrie C (1999) An RNA switch at the 5' splice site requires ATP and the DEAD box protein Prp28p. *Mol Cell* 3: 55-64

Stevens SW, Ryan DE, Ge HY, Moore RE, Young MK, Lee TD, Abelson J (2002) Composition and functional characterization of the yeast spliceosomal penta-snRNP. *Mol Cell* 9: 31-44

Sun H, Wu J, Wickramasinghe P, Pal S, Gupta R, Bhattacharyya A, Agosto-Perez FJ, Showe LC, Huang TH, Davuluri RV (2011) Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. *Nucleic Acids Res* 39: 190-201

Takai D, Jones PA (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* 99: 3740-3745

Tanackovic G, Krämer A (2005) Human splicing factor SF3a, but not SF1, is essential for pre-mRNA splicing in vivo. *Mol Biol Cell* 16: 1366-1377

Tang J, Abovich N, Rosbash M (1996) Identification and characterization of a yeast gene encoding the U2 small nuclear ribonucleoprotein particle B" protein. *Mol Cell Biol* 16: 2787-2795

Tardiff DF, Rosbash M (2006) Arrested yeast splicing complexes indicate stepwise snRNP recruitment during in vivo spliceosome assembly. *RNA* 12: 968-979

Thickman KR, Swenson MC, Kabogo JM, Gryczynski Z, Kielkopf CL (2006) Multiple U2AF65 binding sites within SF3b155: thermodynamic and spectroscopic characterization of protein-protein interactions among pre-mRNA splicing factors. *J Mol Biol* 356: 664-683

Tsuiji H, Yoshimoto R, Hasegawa Y, Furuno M, Yoshida M, Nakagawa S (2011) Competition between a noncoding exon and introns: Gomafu contains tandem UACUAAC repeats and associates with splicing factor-1. *Genes Cells* 16: 479-490

Tycowski KT, Shu MD, Steitz JA (1993) A small nucleolar RNA is processed from an intron of the human gene encoding ribosomal protein S3. *Genes Dev* 7: 1176-1190

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

Valcarcel J, Gaur RK, Singh R, Green MR (1996) Interaction of U2AF65 RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA [corrected]. *Science* 273: 1706-1709

Wahl MC, Will CL, Luhrmann R (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell* 136: 701-718

Wang Z, Kayikci M, Briese M, Zarnack K, Luscombe NM, Rot G, Zupan B, Curk T, Ule J (2010) iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol* 8: e1000530

Wentz-Hunter K, Potashkin J (1996) The small subunit of the splicing factor U2AF is conserved in fission yeast. *Nucleic Acids Res* 24: 1849-1854

Will CL, Luhrmann R (2011) Spliceosome structure and function. *Cold Spring Harb Perspect Biol* 3

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

Will CL, Urlaub H, Achsel T, Gentzel M, Wilm M, Luhrmann R (2002) Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5p homologue and an SF3b DEAD-box protein. *EMBO J* 21: 4978-4988

Witten JT, Ule J (2011) Understanding splicing regulation through RNA splicing maps. *Trends Genet* 27: 89-97

Wu JY, Maniatis T (1993) Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* 75: 1061-1070

Wu S, Romfo CM, Nilsen TW, Green MR (1999) Functional recognition of the 3' splice site AG by the splicing factor U2AF35. *Nature* 402: 832-835

Zamore PD, Patton JG, Green MR (1992) Cloning and domain structure of the mammalian splicing factor U2AF. *Nature* 355: 609-614

Zhang D, Childs G (1998) Human ZFM1 protein is a transcriptional repressor that interacts with the transcription activation domain of stage-specific activator protein. *J Biol Chem* 273: 6868-6877

Zhang D, Paley AJ, Childs G (1998) The transcriptional repressor ZFM1 interacts with and modulates the ability of EWS to activate transcription. *J Biol Chem* 273: 18086-18091

Zhang M, Zamore PD, Carmo-Fonseca M, Lamond AI, Green MR (1992) Cloning and intracellular localization of the U2 small nuclear ribonucleoprotein auxiliary factor small subunit. *Proc Natl Acad Sci U S A* 89: 8769-8773

Zhang Y, Madl T, Bagdiul I, Kern T, Kang HS, Zou P, Mausbacher N, Sieber SA, Krämer A, Sattler M (2013) Structure, phosphorylation and U2AF65 binding of the N-terminal domain of splicing factor 1 during 3'-splice site recognition. *Nucleic Acids Res* 41: 1343-1354

Zhu R, Heaney J, Nadeau JH, Ali S, Matin A (2010) Deficiency of splicing factor 1 suppresses the occurrence of testicular germ cell tumors. *Cancer Res* 70: 7264-7272

Zhuang Y, Weiner AM (1986) A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell* 46: 827-835

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