Intron Recognition at the 3′ Splice Site by U2AF, SF1 and p14

by

Charnpal Singh Grewal

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biochemistry

University of Alberta

Abstract:

In the central dogma of biology, genetic information is stored in DNA in discrete organizational units called genes. A gene is transcribed into RNA, which is finally translated into a protein. The discontinuous split-gene structure of eukaryotes is distinct from prokaryotes and requires pre-mRNA splicing to remove silent intron sequences from the initial pre-mRNA transcript before it becomes a mature mRNA suitable for protein synthesis. This process is co-ordinated and directed by the spliceosome. The assembly, maturation, and disassembly of the spliceosome as it removes the intron progresses through a discrete and definable sequence of events, which are collectively known as the spliceosome cycle. The spliceosome cycle begins with the E (early) complex, which commits the intron to removal. In the E complex, both the 5′ and 3′ SS (splice site) of the intron are recognized by the splicing machinery. The 3′ SS is recognized by the heterotrimeric U2AF/SF1 protein complex. In eukaryotic gene expression this is a fundamentally important layer of regulation, particularly for developmentally complex eukaryotes. Despite this, no complete and integrated experimentally derived atomic model exists for the U2AF/SF1 complex in any of its potential free or RNA-bound states. Here, an experimental system was established to express milligram quantities of stable, soluble, and monodisperse U2AF dimer and U2AF/SF1 trimer variants from the fission yeast, *Schizosaccharomyces pombe* which are suitable for further biochemical and structural characterization.

Preliminary biochemical characterization of these complexes consisted of SEC-MALLS to establish the existence of the stable apo complexes from first principles, followed by spectroscopy to establish the existence and long-term stability of the RNA-bound U2AF/SF1

trimer. This was followed by EMSAs (electrophoretic mobility shift assays) with different pairing of protein complexes and 3′ SS model RNA sequences to determine the RNA binding properties of U2AF dimer and U2AF/SF1 trimer. Broadly, EMSAs reveal that SF1 has a major role in the affinity of the U2AF/SF1 complex for the 3′ SS, and that both phosphorylation of SF1 as well as the zinc knuckles of SF1 modulate sequence specificity and affinity of the U2AF/SF1 complex for the 3′ SS.

Structural characterization consisted of two sets of experiments. The first of these was SAXS to evaluate both the overall solution behaviour and scattering of these complexes in their free and RNA-bound states, as well as to generate molecular envelopes of individual complexes. The SAXS data indicate that *S. pombe* U2AF dimer and U2AF/SF1 trimer are well folded and roughly globular in both their apo and RNA-bound states. These experiments were followed up by a more advanced SAXS technique that can more accurately model conformationally flexible, multidomain protein complexes such as the U2AF dimer and U2AF/SF1 trimer than traditional *ab initio* shape reconstruction. This technique used a modified, chimeric U2AF/SF1 complex containing both human and *S. pombe* components in both its apo state as well as bound to two different 3′ SS model RNAs and combined SEC-SAXS with rigid body modelling in order to generate energy-minimized structure ensembles. These experiments indicate that all samples possess some conformational flexibility in solution, showing modest amounts of larger particles in addition to a more predominant compact form. However, a bound 3′ SS RNA clearly skews the particle distribution in solution to the more compact, globular form.

To complement the U2AF/SF1 studies in this thesis the X-ray structures of the *S. pombe* and *Candida albicans* p14/SF3B155 complexes were solved. P14 displaces SF1 during the spliceosome cycle and is tightly complexed to an essential scaffold protein in the spliceosome

called SF3B155. Therefore, structural characterization of *S. pombe* p14 is an important step towards developing a complete and integrated structural model of the spliceosome cycle which includes 3′ SS recognition by U2AF/SF1. *C. albicans* is a budding yeast whose p14 orthologue lacks the conserved aromatic residue that contacts the pre-mRNA in humans and presumably also in *S. pombe*, therefore its structure was solved since it is informative to understand alternative mechanisms for specific recognition of the conserved sequence motifs within the 3′ SS. Both the yeast structures are similar to previously published human p14/SF3B155 structures but reveal that p14 is more conformationally plastic than previously thought.

The work described in this thesis is important foundational work for understanding 3′ SS recognition since it represents the first published report of successful expression and purification of the conserved RNA-binding core of the U2AF dimer and U2AF/SF1 trimer complexes. This system is sufficiently tractable and adaptable for a variety of biochemical and structural experiments making it a promising first step towards realizing an atomic model of the U2AF dimer and U2AF/SF1 trimer complexes. This tractability makes it a promising candidate for more nuanced biochemical investigations as well. In addition to the U2AF/SF1 studies presented here, the yeast p14/SF3B155 structures also contribute to building a unified model of 3′ SS recognition within the spliceosome cycle.

Dedicated to Baba Ji.

Table of Contents:

List of Tables:

List of Figures:

Chapter 1

Chapter 2

Chapter 3

Appendix *I*

Appendix *III*

Appendix *V*

List of Abbreviations:

Δ	delta
ΔE67	Murine-specific U2AF26 mRNA and protein isoform generated by the exclusion of exons 6 and 7
ΔE7	Murine-specific U2AF26 mRNA and protein isoform generated by the exclusion of exon 7
ΔZn	*S. pombe* SF1 constructs ending at Q309, therefore lacking the two C-terminal zinc knuckles
(P)	phosphorylated
α-[^{32}P]-ATP	ATP labeled on the α phosphate group with ^{32}P
Å	angstrom
λ	wavelength
$\chi 2$	chi-square
A	adenine/adenosine/deoxyadenosine
A	branch A
A complex	pre-spliceosome complex
ACT1	actin 1
Ad2	adenovirus type 2
AdML	adenovirus major late
ALS	Advanced Light Source
AQR	aquarius
ASO	antisense oligonucleotide
AtGRP	*A. thaliana* glycine rich RNA binding protein
ATP	adenosine triphosphate
ATPase	adenosine triphosphatase
B* complex	catalytically activated complex
B^{act} complex	activated complex
Bbp	branchpoint bridging protein
B complex	pre-catalytic complex
BLAST	basic local alignment search tool
BME	β-mercaptoethanol
BPS	branchpoint sequence
branch A	branch adenosine
Brr2	bad response to refrigeration 2
C	cytosine/cytidine/deoxycytidine
C* complex	catalytic step 2 complex
CACTIN	renal carcinoma antigen NY-REN-24
CAPERα	coactivator of activating protein-1 and estrogen receptors α
CAPERβ	paralogue of U2AF65 related to CAPERα
C complex	catalytic step 1 complex
CD2	cluster of differentiation 2
CD2BP2	CD2 binding protein 2
Cdc2	cell division cycle protein 2 homologue
cDNA	complementary DNA

CELF4	CUGBP Elav-like family member 4
CFTR	cystic fibrosis transmembrane conductance regulator
CHARMM-GUI	Chemistry at Harvard Macromolecular Mechanics-Graphical User Interface
Ci	Curie
C_i complex	C complex intermediate
circRNA	exonic circular RNA
Clk	Cdc2-like kinase
CLS	Canadian Light Source
COOT	Crystallographic Object-Oriented Toolkit
COPII	coat protein complex II
cryo-EM	cryogenic electron microscopy
CTP	cytidine triphosphate
CUGBP	CUG triplet repeat, RNA binding protein
Cwc25	complexed with Cef1 25
CXMS	chemical cross-linking mass spectrometry
ddH_2O	double distilled H_2O
DEPC	diethyl pyrocarbonate
D_{max}	maximum particle radius
DNA	deoxyribonucleic acid
dRI	differential refractive index
DTT	1,4-dithiothreitol
E complex	early complex
EDC	exon definition complex
EDTA	ethylenediaminetetraacetic acid
Elav	embryonic lethal abnormal visual system
EM	electron microscopy
EMDB	Electron Microscopy Data Bank
EMSA	electrophoretic mobility shift assay
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
EST	expressed sequence tag
FAM32A	family with sequence similarity 32 member A
Filter 390 BP	390-nm band-pass filter
FMR1	fragile X mental retardation 1
FRET	fluorescence resonance energy transfer
G	guanine/guanosine/deoxyguanosine
GAJOE	Genetic Algorithm Judging Optimisation of Ensembles
GASBOR	Genetic Algorithm for SwitchBOx Routing
GPRG	Gly-Pro-Arg-Gly
GST	glutathione-S transferase
GTP	guanosine triphosphate
GYF	glycine-tyrosine-phenylalanine
HeLa	Helen Lane/Henrietta Lacks
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HIV	human immunodeficiency virus

hnRNP	heterogeneous nuclear ribonucleoprotein
HPLC	high-performance liquid chromatography
HT-SAXS	high-throughput SAXS
I(0)	forward scattering intensity measured at zero scattering angle
iCLIP	individual nucleotide resolution UV cross-linking and immunoprecipitation
IDC	intron definition complex
ILS complex	intron lariat spliceosome complex
IPTG	isopropyl β-D-thiogalactopyranoside
ISL	internal stem-loop
JCSG	Joint Center for Structural Genomics
k_BT	energy unit derived from the product of k_B (Boltzmann constant) and T (temperature in Kelvin units)
K_d	dissociation constant
kDa	kilodalton
KH	K homology
KHDRBS	KH domain containing, RNA binding, signal transduction associated
KH-QUA2	KH-quaking
KIS	kinase interacting with stathmin
lac	lactose
LB	Luria-Bertani
LIC	ligation independent cloning
LS2	large subunit 2
LSm	like Smith protein
Luc7	lethal unless cap-binding complex is produced 7
MAD	multiple-wavelength anomalous diffraction
MALLS	multi-angle laser light scattering
mAU	milli-absorbance unit
Mb	megabase
MBNL1	muscleblind-like splicing regulator 1
MBP	maltose-binding protein
MCS	multiple cloning site
mimetic	*S. pombe* SF1 constructs containing the S131E, S133E dual phosphomimetic point mutations
MPD	2-methyl-2,4-pentandiol
mRNA	messenger RNA
Msl5	Mud synthetic lethal 5
Mud2	mutant U1 die 2
MW	molecular weight
n	any nucleotide
NCBI	National Center for Biotechnology Information
Ni-NTA	nickel nitrilotriacetic acid
NKAP	nuclear factor kappa B-activating protein
NLS	nuclear localization signal
NMD	nonsense-mediated mRNA decay

NMR	nuclear magnetic resonance
NONO	non-POU domain containing octamer binding
Nova2	neuro-oncological ventral antigen 2
NTC	prp NineTeen associated complex
OD_{600}	optical density at 600 nm
ORF	open reading frame
OTAG-12	ovarian tumor associated gene-12
P(r)	pair-distance distribution function
p14	14 kDa protein
p32	32 kDa protein
P complex	post-catalytic complex
PCR	polymerase chain reaction
PDB	Protein Data Bank
PEG	polyethylene glycol
PMSF	phenylmethylsulfonyl fluoride
POU	Pit-Oct-Unc
PPT	polypyrimidine tract
pre-A complex	precursor to A complex
pre-B complex	precursor to B complex
pre-B^{act} complex	precursor to B^{act} complex
pre-mRNA	precursor mRNA
Prp	pre-mRNA processing
PSF	PTB associated splicing factor
PSI	Protein Structure Initiative
PsiPred	PSI-BLAST based secondary structure prediction
PTB	polypyrimidine tract binding protein
PTC	premature termination codon
PUF60	poly(U) binding splicing factor 60 kDa
Py tract	polypyrimidine tract
q	scattering vector
QELS	Quasi-Elastic Light Scattering
QK1	quaking homologue 1
QUA	quaking
R	purine
Rg	radius of gyration
rmsd	root-mean-square deviation
RNA	ribonucleic acid
RNP	ribonucleoprotein
RRM	RNA recognition motif
RRM1	N-terminal RRM of U2AF-L
RRM2	C-terminal RRM of U2AF-L
RS	arginine-serine dipeptide
RT-PCR	reverse transcription-PCR
RUST	regulated unproductive splicing and translation
S	Svedberg unit
S75	Superdex 75

S200	Superdex 200
SAD	single-wavelength anomalous diffraction
Sam68	Src associated in mitosis of 68 kDa
SAXS	small angle X-ray scattering
SC35	splicing component 35 kDa
SCN	suprachiasmatic nucleus
SDE2	silencing defective 2
SDS-PAGE	sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SEC	size exclusion chromatography
SEC-MALLS	size exclusion chromatography + multi-angle laser light scattering
SEC-SAXS	size exclusion chromatography + small angle X-ray scattering
SELEX	systematic evolution of ligands by exponential enrichment
SF1	splicing factor 1
SF2/ASF	splicing factor 2/ alternative splicing factor
SF3A	splicing factor 3A
SF3B	splicing factor 3B
SF3B155	SF3B 155 kDa subunit
SF3B6	SF3B subunit 6 (synonymous name for p14)
SIBYLS	Structurally Integrated BiologY for the Life Sciences
Slu7	synergistic lethal with U5 snRNA 7
Sm	Smith protein
smFRET	single-molecule fluorescence resonance energy transfer
Smy2	suppressor of myo2-66
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein
Snu71	small nuclear ribonucleoprotein associated 71
Spp2	suppressor of Prp2
SPSP	serine-proline-serine-proline
SR	serine-arginine dipeptide
Srp2	SR protein 2
SRp46	SR protein 46 kDa
SS	splice site
STAR	signal transduction and activation of RNA
T	thymine/thymidine
TBE	Tris-borate-EDTA
TCEP-HCl	tris (2-carboxyethyl) phosphine hydrochloride
TEV	tobacco etch virus
TFIIH	transcription factor IIH
THD	TIMELESS homology domain
TIA-1	T-cell-restricted intracellular antigen 1
TIAR	TIA-1-related protein
TIS11d	TPA-inducible sequence 11d
TNFR	tumor necrosis factor receptor
TPA	12-0-tetradecanoyl-phorbol-13-acetate
TRA2-Beta	transformer 2 beta homolog
Tris	tris(hydroxymethyl)aminomethane

U	uracil/uridine
U2AF	U2 auxiliary factor (heterodimer of U2AF-S and U2AF-L)
U2AF23	*S. pombe*-specific U2AF-S orthologue of apparent 23 kDa MW
U2AF26	Mammalian-specific U2AF-S paralogue of U2AF35 of apparent 26 kDa MW
U2AF35	Metazoan-specific U2AF-S orthologue of apparent 35 kDa MW
U2AF35a	Conserved metazoan-specific U2AF35 mRNA and protein isoform
U2AF35b	Conserved metazoan-specific U2AF35 mRNA and protein isoform
U2AF35c	Conserved metazoan-specific U2AF35 mRNA isoform targeted for the NMD pathway
U2AF35-RS1	Mammalian-specific U2AF35 paralogue
U2AF35-RS2	Mammalian-specific U2AF35 paralogue related to U2AF35-RS1
U2AF38	*D. melanogaster*-specific U2AF-S orthologue of apparent 38 kDa MW
U2AF50	*D. melanogaster*-specific U2AF-L orthologue of apparent 50 kDa MW
U2AF59	*S. pombe*-specific U2AF-L orthologue of apparent 59 kDa MW
U2AF65	Metazoan-specific U2AF-L orthologue of apparent 65 kDa MW
U2AF-L	large subunit of U2AF
U2AF-S	small subunit of U2AF
U2AF/SF1	heterotrimer of U2AF-S, U2AF-L, and SF1
$U2AF/SF1_{\Delta Zn,mimetic}$	phosphomimetic *S. pombe* U2AF/SF1 trimer with SF1 zinc knuckles deleted used in Chapter 2 of this thesis
$U2AF/SF1_{\Delta Zn,wt}$	wildtype *S. pombe* U2AF/SF1 trimer with SF1 zinc knuckles deleted used in Chapter 2 of this thesis
$U2AF/SF1_{mimetic}$	phosphomimetic, prototype *S. pombe* U2AF/SF1 trimer used in Chapter 2 of this thesis
$U2AF/SF1_{wt}$	wildtype, prototype *S. pombe* U2AF/SF1 trimer used in Chapter 2 of this thesis
U2OS	U2 osteosarcoma epithelial cell line
UAF-1	U2 auxiliary factor-1 (*C. elegans*-specific U2AF-L orthologue)
UBC4	ubiquitin-conjugating enzyme E2 4
UHM	U2AF homology motif
UHMK1	U2AF homology motif kinase 1
ULM	U2AF ligand motif
Urp	U2AF related protein (synonym for U2AF35-RS2)
UTP	uridine triphosphate
UTR	untranslated region
UV	ultraviolet
V	Volts
Vp	particle volume in $Å^3$
wt	wildtype
y	pyrimidine
YES	yeast extract + supplements
YPD	yeast extract peptone dextrose
ZF	zinc finger

ZF1	N-terminal ZF of U2AF-S
ZF2	C-terminal ZF of U2AF-S

Chapter 1

The split-gene structure and pre-mRNA splicing

The unifying principle and bedrock of all contemporary biology is the central dogma of molecular biology, which was proposed in 1958 by Francis Crick. Briefly, it proposes that genetic information is stored in DNA (deoxyribonucleic acid) in discrete organizational units called genes. A gene is transcribed into RNA (ribonucleic acid) by RNA polymerase, which is finally translated into a protein.

Within the context of this dogma, eukaryotic genetic organization is distinct from that of prokaryotes because it is built on a discontinuous split-gene structure in which protein-coding exon sequences within a single gene are interrupted by intron sequences. The initial intron-containing RNA transcript in eukaryotes is called pre-mRNA (precursor mRNA). The pre-mRNA transcript must pass through a cellular program called pre-mRNA splicing whereby the introns are removed and the neighbouring exons are subsequently ligated together in order to generate a mature mRNA (messenger RNA) that is suitable for translating into a functional protein; this entire process is termed pre-mRNA splicing (Fig. 1-1A) (Burge, Tuschl, & Sharp, 1999; Dunn & Rader, 2014; Kramer, 1995; Staley & Guthrie, 1998; Will & Luhrmann, 2011). After splicing is complete, the mRNA is exported from the nucleus and is finally available for translation into functional protein.

Pre-mRNA splicing was first discovered in 1977 by the laboratories of Rich Roberts and Phillip Sharp while attempting to map the location of individual mRNAs on the chromosome of adenovirus, where the presence of a single-stranded DNA loop in a hybrid of the Ad2 (adenovirus type 2) mRNA and the corresponding genomic DNA indicated that the gene structure was discontinuous on the chromosome (Berget, Moore, & Sharp, 1977; Chow, Gelinas, Broker, & Roberts, 1977).

Gene organization (both physical structure and DNA sequence) and pre-mRNA splicing are two topics of intense and universal interest, because they represent the two main layers of information control that serve as building blocks for higher-order biological organization. More pointedly, increasing our understanding of these two levels of genetic regulation will automatically increase our understanding of all biological phenomena. This is obvious in respect to splicing when considering that the split-gene structure is a ubiquitous feature of genetic organization across all eukaryotic kingdoms (over 90% of all eukaryotic genes contain introns) and regulates the faithful expression of virtually all human genes.

With respect to the general principles of biology, a fact which draws to attention the need for ongoing study of splicing is the feature of alternative splicing (see Section 1-2.). Splicing is actively manipulated by the cellular machinery through the apparatus of alternative splicing, whereby a single gene can code for several different functional proteins with variable properties depending on how intron/exon boundaries are defined, and protein variations generated from a single gene are called isoforms. This allows for spatial and temporal control of gene expression, representing a major route for expanding the proteome and profoundly shaping the evolution of developmentally complex eukaryotes.

Finally, the fundamental relationship of splicing to eukaryotic biology makes it intuitively obvious that splicing is not only of very general interest but also of great importance in human medicine because a diverse array of human diseases, including cancers, are the direct result of splicing errors (See Section 1-3.) (Anna & Monika, 2018; Chabot & Shkreta, 2016).

1-1. The chemistry of pre-mRNA splicing

There are a total of four distinct classes of introns but this thesis is exclusively based within the context of the splicing of the class called nuclear pre-mRNA introns. In this class of introns, splicing consists of two sequential transesterifications within the intron, hereafter referred to as the 1st step and 2nd step of splicing (Fig. 1-1A). In the 1st step, the 2′-hydroxyl group of a conserved, catalytic adenosine (branch adenosine or branch A for simplicity) performs a nucleophilic attack on the phosphodiester linkage preceding the guanosine at the 5′ SS (splice site) to free the 5′ exon. This nucleophilic attack generates a 2′-5′ phosphodiester linkage between the 5′ guanosine and branch A for a total of three phosphodiester linkages on the branch A (including the standard 5′ and 3′ linkages immediately upstream and downstream of the branch A) in order to create a unique circularized lariat structure within the intron which is still attached to the 3′ exon. The 2nd step entails the 3′-hydroxyl on the now freed 5′ exon performing a nucleophilic attack on the 3′ SS at the invariant AG di-nucleotide (Black, 2003; Padgett, Konarska, Grabowski, Hardy, & Sharp, 1984). This entire process ligates two exons into a continuous sequence and generates a free lariat intron which is subsequently degraded (Jacquier, 1990).

For complete biological context, all four intron classes will be touched on briefly before proceeding into a deeper study of nuclear pre-mRNA introns and the associated topics relevant to this thesis. Three of the four classes (group I, group II and nuclear pre-mRNA introns) catalyze intron removal and exon ligation via two sequential phosphotransesterification reactions. Broadly speaking, group I and group II introns are both spliced via a self-catalytic mechanism but differ in that the nucleophile attacking the 5′ SS in the 1st step is a free guanosine for group I introns whereas it is an adenosine residue forming part of the intron sequence for group II

introns (Cech, 1990; H. Nielsen & Johansen, 2009; Pyle, 2016; Ritchie, Schellenberg, & MacMillan, 2009). Nuclear pre-mRNA introns are the largest class of introns and are spliced using the same mechanism/chemistry as that of group II introns, but require the assistance of a multi-megadalton protein/RNA entity called the spliceosome to co-ordinate splicing catalysis (Rio, 1993; Valadkhan & Jaladat, 2010). The fourth and final class of introns is transfer RNA introns and these are distinct from the other three classes in that splicing is catalyzed by a splicing protein endonuclease which derives energy from ATP (adenosine triphosphate) hydrolysis to catalyze intron removal and exon-exon ligation through a mechanism comparable to the DNA ligase reaction (Calvin & Li, 2008).

Other than having the same mechanism/chemistry as each other, group II introns and nuclear pre-mRNA introns are separated by significant differences. Aside from self-catalysis, group II introns are distinguished from nuclear pre-mRNA introns in having six stem-loop structures (domains I to VI) within the intron itself and are defined by very few conserved nucleotides with catalytically important nucleotides being spread out over the entire intron (Bonen & Vogel, 2001). In contrast, nuclear pre-mRNA introns have well-defined sequence elements. The 5′ boundary of the intron is defined by a nearly invariant GU di-nucleotide, which is part of the larger and more degenerate consensus sequence, GURAGU (R = purine). With respect to this thesis, the focus is on the 3′ boundary of the intron and the sequence elements which define it. From 5′ → 3′, this boundary is defined by a degenerate BPS (branchpoint sequence) in humans, which conforms to the yUnAy (y = pyrimidine, n = any nucleotide; branch A is underlined) consensus sequence (this sequence is an invariant UACUAAC in *S. cerevisiae*), followed by a PPT (polypyrimidine tract) which is highly variable in sequence and length, followed by an invariant yAG tri-nucleotide motif which marks the intron/exon boundary (Gao,

Masuda, Matsuura, & Ohno, 2008; Reed, 1989; Senapathy, Shapiro, & Harris, 1990). Additionally, a number of non-constitutive splicing enhancer and repressor sequences within the vicinity of these well-defined sequence motifs allow the cell to modulate splicing, which is an important component of gene regulation via alternative splicing (Barash et al., 2010).

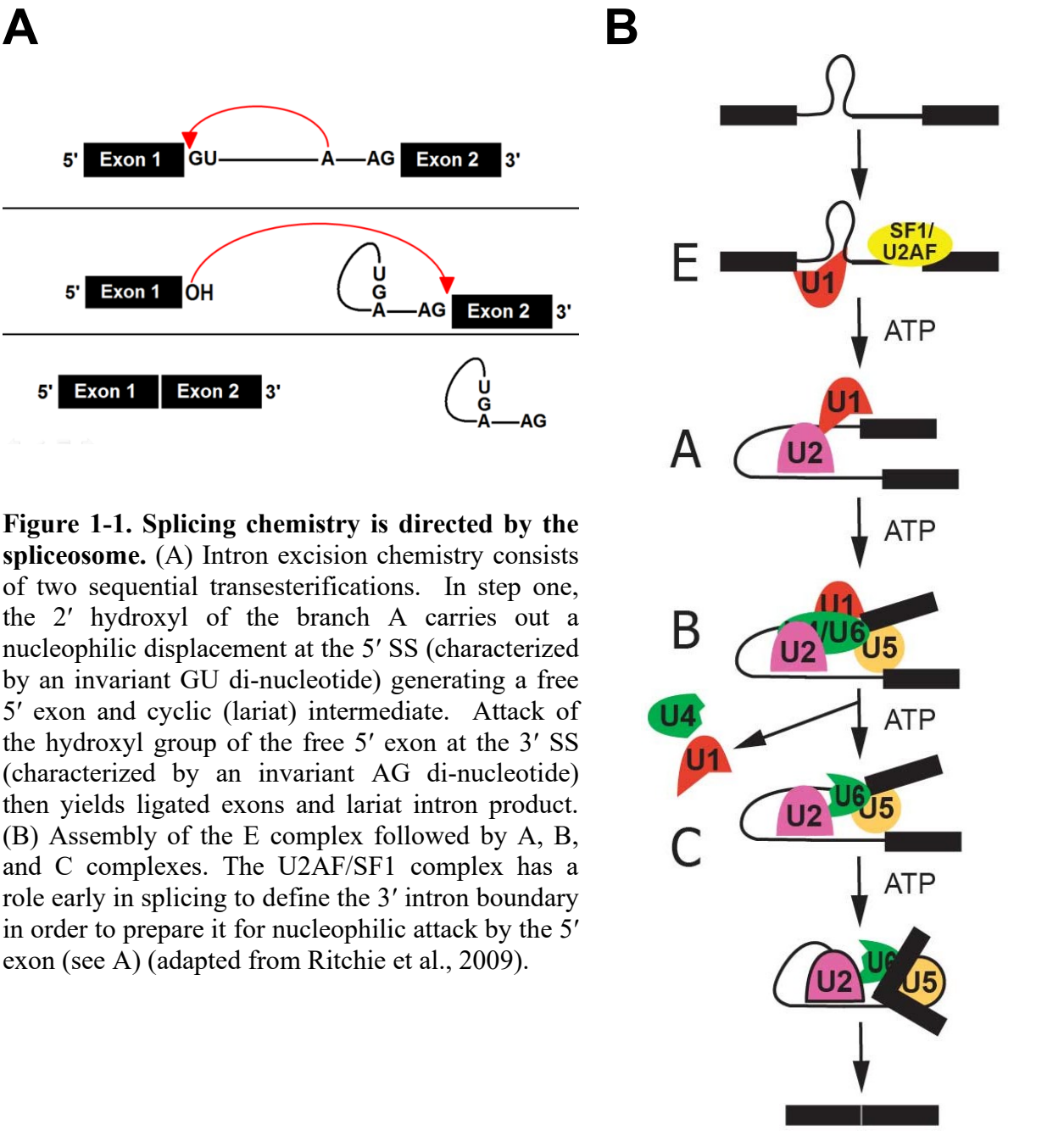


Figure 1-1. Splicing chemistry is directed by the spliceosome. (A) Intron excision chemistry consists of two sequential transesterifications. In step one, the 2′ hydroxyl of the branch A carries out a nucleophilic displacement at the 5′ SS (characterized by an invariant GU di-nucleotide) generating a free 5′ exon and cyclic (lariat) intermediate. Attack of the hydroxyl group of the free 5′ exon at the 3′ SS (characterized by an invariant AG di-nucleotide) then yields ligated exons and lariat intron product. (B) Assembly of the E complex followed by A, B, and C complexes. The U2AF/SF1 complex has a role early in splicing to define the 3′ intron boundary in order to prepare it for nucleophilic attack by the 5′ exon (see A) (adapted from Ritchie et al., 2009).

1-2. Alternative splicing in the context of higher-order biology

In the central dogma of eukaryotic biology, all genetic information must flow from DNA to pre-mRNA to mature mRNA before the cell is able to use it to execute its various functions. For this reason, it represents a comprehensive layer of information control and therefore splicing is a tightly regulated process. Constitutive splicing refers to the strict use of intron/exon boundaries for assembling splicing complexes. However, these intron/exon boundaries can be defined in a number of different ways which is a process called alternative splicing and dramatically expands proteome diversity because it allows a single gene to generate multiple messages (mRNA isoforms) each of which has the potential to code for a unique protein (protein isoforms), each with unique properties (Nilsen & Graveley, 2010).

There are several routes for alternative splicing to occur. The recognition of intron/exon boundaries by the splicing machinery may be suppressed leading to splicing out of an intron-exon-intron sequence as if it were a continuous intron (exon skipping) and mutually exclusive exons also exist in certain genes. Suppressing the recognition of intron/exon boundaries also leads to intron retention (Black, 2003). As stated before, alternative splicing is actively manipulated in order to promote or suppress various splice site selection events and there is a network of trans-acting splice site activators that recruit splicing factors to specific splice sites and splice site silencers that block the recruitment of splicing factors. The equilibrium of these activators and silencers is responsible for the selection or rejection of specific splice sites (M. Chen & Manley, 2009; C. W. Smith & Valcarcel, 2000). Because alternative splicing is regulated, mutations that interfere with it by altering the balance of protein isoforms or creating new splice sites (cryptic splice site activation) often cause disease (Ward & Cooper, 2010). The most obvious example would be the broad disease of cancer where aberrations in alternative

splicing can shift the balance towards protein isoforms that increase cell proliferation, survival and migration, thereby promoting cancer (e.g. mutations in the alternative splicing regulator SC35 are frequently observed in hematologic malignancies including 10% of myelodysplastic syndromes, 31-47% of chronic myelomonocytic leukemia, and 2% of acute myeloid leukemia) (Urbanski, Leclair, & Anczukow, 2018). This can occur either within oncogenes themselves or within splicing proteins.

Alternative splicing is especially important because the spatial and temporal control of gene expression through alternative splicing is intimately tied to the development of multicellular, developmentally complex eukaryotes. There is no strong correlation between genome size and organismal complexity, but a strong correlation exists between the prevalence of alternative splicing and organismal complexity and ~95% of human genes undergo alternative splicing (Maniatis & Tasic, 2002; Pan, Shai, Lee, Frey, & Blencowe, 2008; Taft & Mattick, 2003; E. T. Wang et al., 2008). By contrast, the model organism and microscopic nematode *Caenorhabditis elegans* consists of about a thousand cells but has about as many genes as humans, and one cell of the amoeba *Chaos chaos* has over 200x more DNA than a human cell (Ezkurdia et al., 2014; Hillier et al., 2005; Holm-Hansen, 1969).

Many protein isoforms generated by alternative splicing are known to play key regulatory roles in a variety of biological processes and can differ in a variety of subtle or dramatic ways by differing in regions responsible for functionalities such as catalysis, localization, or association with other macromolecules. Localization is a particularly important regulatory theme. Alternative inclusion/exclusion of membrane-spanning regions is a common outcome of alternative splicing, and several TNFR (tumor necrosis factor receptor) genes are regulated through alternative splicing to produce either soluble or membrane-associated signal-transducing

forms (Cascino, Fiucci, Papoff, & Ruberti, 1995; Cheng et al., 1994; Cline et al., 2004; Lainez et al., 2004; J. Michel, Langstein, Hofstadter, & Schwarz, 1998; Screaton et al., 1997; Tone, Tone, Fairchild, Wykes, & Waldmann, 2001; Xing, Xu, & Lee, 2003). Another common outcome of alternative splicing is to downregulate a gene by generating non-functional protein isoforms.

Protein isoforms are the most obvious route for alternative splicing to generate organismal complexity, but alternative splicing also regulates biological processes at the post-transcriptional level through the functional coupling of alternative splicing and the NMD (nonsense-mediated mRNA decay) pathway. In order to accomplish this, alternative splice sites generate non-functional mRNA isoforms with a PTC (premature termination codon) that targets the transcript for destruction through the NMD pathway, a mechanism called RUST (regulated unproductive splicing and translation) (Hilleren & Parker, 1999; Hillman, Green, & Brenner, 2004; Lareau, Green, Bhatnagar, & Brenner, 2004; B. P. Lewis, Green, & Brenner, 2003; Morrison, Harris, & Roth, 1997; Vilardell, Chartrand, Singer, & Warner, 2000).

In mammals, recognition of PTCs is intimately connected to splicing: a termination codon is recognized as premature if it is more than ~50 nucleotides upstream of a splicing-generated exon-exon junction (Maquat, 2004). Using this rule, one computational study inferring alternative splicing from ESTs (expressed sequence tags) revealed that 45% of alternatively spliced human genes produce at least one mRNA isoform containing a PTC that targets it to the NMD pathway, and another computational screen of annotated alternatively spliced protein isoforms within Swiss-Prot revealed that 7.9% of 1463 alternatively spliced human genes amenable to analysis generate at least one mRNA isoform predicted to enter the NMD pathway (Hillman et al., 2004; B. P. Lewis et al., 2003). Because a large percentage of alternative splicing appears to target mRNAs for degradation through the NMD pathway, RUST is likely a

widespread and generalized mechanism of post-transcriptional gene repression. Additionally, examples of RUST exist in fungi, animals, and plants, indicating that RUST either originated in their common ancestor or that it has evolved independently multiple times.

Experimentally characterized cases of RUST validate the proposed broad role of gene regulation through RUST and reveal the versatility of this mechanism in building genetic regulatory circuits. RUST is known to autoregulate genes. One example is PTB (polypyrimidine tract binding protein), which is a regulator of alternative splicing, whereby increased PTB levels cause the protein to alter the splicing of its own pre-mRNA in order to increase levels of a naturally occurring PTC-containing mRNA isoform (Wollerton, Gooding, Wagner, Garcia-Blanco, & Smith, 2004). The alternative splicing regulatory protein TRA2-Beta (transformer 2 beta homolog) also autoregulates itself through RUST (Stoilov, Daoud, Nayler, & Stamm, 2004).

RUST autoregulatory circuits can also operate indirectly, as in the case of Clk1 (Cdc2-like kinase 1). Strong evidence of intra-species and inter-species conservation of RUST is found in Clks, which phosphorylate SR proteins (see Section 1-4.1. for a summary of SR proteins) to regulate splicing, and Clk1 indirectly regulates its own activity whereby increased Clk1 activity changes the activity of one or more SR proteins, which then alters Clk1 splicing to favor a transcript with a PTC (Colwill et al., 1996; Duncan, Stojdl, Marius, & Bell, 1997; Hillman et al., 2004).

Not all RUST involving splicing regulators is autoregulatory. The *Arabidopsis thaliana* AtGRP7 (*A. thaliana* glycine rich RNA binding protein 7) gene uses RUST to autoregulate itself as well as execute heterologous regulation of another gene, AtGRP8 (Staiger, Zecca, Wieczorek Kirk, Apel, & Eckstein, 2003). Another example of heterologous RUST regulation is TIAR (TIA-1-related protein)/TIA-1 (T-cell-restricted intracellular antigen 1). TIA-1 downregulates

TIAR by inducing the splicing of a PTC-containing mRNA isoform of TIAR, which is a protein that affects regulation of both translation and alternative splicing events (Le Guiner, Gesnel, & Breathnach, 2003).

Several well-characterized examples of RUST, including those described above, regulate splicing factors (Sureau, Gattoni, Dooghe, Stevenin, & Soret, 2001). It is not yet known whether RUST disproportionately regulates splicing factors or whether this observation reflects acquisition bias. However, RUST controls many proteins other than splicing factors as well.

In addition to protein isoforms and post-transcriptional control through RUST, alternative splicing can regulate gene expression by generating multiple alternative mRNA isoforms which all code for the same protein isoform, and this class of mRNAs comprises ~20% of alternative mRNA isoforms. The potential roles of this class of mRNAs are mostly unexplored, and many large-scale studies filter out these seemingly redundant mRNA isoforms at an early preprocessing step. Alternative promoter usage may regulate the expression of alternative mRNA isoforms that differ only in their 5′ UTR (untranslated region), which are common (T. Zhang, Haws, & Wu, 2004). In contrast, 3′ UTR elements may control mRNA subcellular localization, stability, and translational efficiency (Bratu, Cha, Mhlanga, Kramer, & Tyagi, 2003; Wickens, Bernstein, Kimble, & Parker, 2002).

The co-evolutionary relationship between alternative splicing and developmentally complex, multicellular eukaryotes means that the conservation of gene regulation mechanisms through alternative splicing over long periods of time implies the existence of selective pressure to conserve function. The discovery and characterization of these conserved mechanisms across species and protein families by comparing orthologous and paralogous gene sequences provides insight into evolution and function, as well as providing a means to filter out noise from large

datasets. Examples of evolutionary conservation include several of the genes discussed above, one of which is Clks. An autoregulatory mechanism exists whereby exclusion of exon 4 introduces a frameshift and PTC, and this is conserved across the three human paralogues (Clk1-3), the three orthologues of these genes in mouse, and the single gene copy in the sea squirt, *Ciona intestinalis* (Hillman et al., 2004).

The three layers of genetic regulation discussed above which alternative splicing can control provide many opportunities for regulatory networks to emerge and co-evolve. These mechanisms of controlling gene expression through alternative splicing are not mutually exclusive, and the expression of a single gene can be controlled through multiple mechanisms and participate in multiple regulatory networks. Two examples of a single gene being controlled simultaneously by both autologous and heterologous RUST include the splice regulator proteins raver1 and CELF4 (CUGBP Elav-like family member 4) influencing the autoregulatory splicing of PTB, and SRp46 (SR protein 46 kDa) inducing a PTC variant of SC35 (splicing component 35 kDa) mRNA (Sureau et al., 2001; Wollerton et al., 2004). Genes can be regulated through alternative splicing in addition to other layers of non-splicing regulation. AtGRP7/AtGRP8 is an example of combined autologous and heterologous RUST, however AtGRP7 expression is additionally also controlled by rhythmic, oscillating transcriptional regulation via the circadian clock (Staiger & Apel, 1999).

Alternative splicing is able to generate organismal complexity by allowing genome-wide spatio-temporal control of gene expression at multiple levels. These are woven into regulatory networks that can potentially include all the other layers of non-splicing gene regulation in order to generate overlapping, redundant and very complex networks. One final consequence of alternative splicing is noted here in order to emphasize its importance and potential in enabling

and driving the evolution of organismal complexity. The use of alternative splicing to regulate splicing factors implies the existence of regulatory cascades of alternative splicing. These cascades would allow an upstream splicing or phosphorylation event in a splicing factor to alter the splicing of many downstream substrates, allowing minimal upstream input to effect large changes in the transcriptome and proteome.

1-3. Splicing and human disease

Many human diseases are caused by or associated with splicing errors and at least 15% of all genetic disorders are associated with mutated splice sites (Anna & Monika, 2018; Chabot & Shkreta, 2016; Matlin, Clark, & Smith, 2005). Changes in splicing patterns can interfere with normal human biology by generating non-functional messages or messages that generate abnormal protein, as well as altering the balance of naturally occurring mRNA/protein isoforms. These changes include frameshift mutations, deletions/insertions in a mature mRNA, and PTCs. These aberrant messages either enter the NMD pathway or generate abnormal proteins with possible deleterious effects (Lykke-Andersen & Jensen, 2015).

Aberrations in splicing are often the result of mutations in the mis-spliced gene itself. These mutations can be present either within the splice sites or the regulatory RNA elements in the pre-mRNA and they can alter alternative splicing patterns by activating cryptic splice sites or silencing canonical ones (Matlin et al., 2005). One prominent example of a disease caused by altered splicing signals is cystic fibrosis where 70% of all cases are the result of in-frame exon skipping in the CFTR (cystic fibrosis transmembrane conductance regulator) gene due to a recessive allele with a short PPT in intron 8 (Chu, Trapnell, Curristin, Cutting, & Crystal, 1993). Other examples of diseases caused through this route include Duchenne muscular dystrophy,

Becker muscular dystrophy, and spinal muscular atrophy (Bellayou et al., 2009; Habara et al., 2009; Price, Morderer, & Rossoll, 2018; L. Wan & Dreyfuss, 2017; S. Wu et al., 2018).

Finally, mutations in the proteins involved in the splicing cycle or in regulating the splicing cycle can also cause incorrect splice site selection (Faustino & Cooper, 2003). Surprisingly however, mutations in these proteins tend to produce a localized phenotype despite the fact that splicing occurs in all cell types (Lehalle et al., 2015). One example is retinitis pigmentosa, which causes vision loss through decreased photoreceptor cells. A number of cases of this disease are caused by mutations in key spliceosomal proteins, disturbing the normal alternative splicing pattern in the retina (Chakarova et al., 2002; Farkas, Grant, & Pierce, 2012; McKie et al., 2001; Vithana et al., 2001).

It is not possible to provide an exhaustive list of splicing diseases because splicing is a universal feature of all human biology and therefore splicing errors can also give rise to a wide array of diseases through a variety of mechanisms. An attempt has been made above to illustrate the main routes for how splicing errors generate disease along with giving a few examples. Diseases known to be caused by splicing errors also present a useful possibility for treatment because they can be treated by targeting the splicing errors. As an example, the potential to treat diseases caused by a disturbed balance of protein isoforms exists by using small ASOs (antisense oligonucleotides) in order to block splice site usage generating oncogenic isoforms and these drugs are currently being investigated (Havens & Hastings, 2016).

1-4. Overview of the spliceosome and stepwise assembly of splicing complexes

Most genes in higher eukaryotes contain multiple introns and an individual spliceosome is required to define and remove each individual intron. The spliceosome is necessary to

organize and direct the chemistry of splicing by proximating the catalytic sequence elements in nuclear introns and this occurs for each individual intron. The spliceosome is a large (60S), dynamic and transient assemblage of snRNPs (small nuclear ribonucleoprotein) and accessory protein factors.

A snRNP consists of an snRNA (small nuclear RNA) that directly interacts with a conserved intron sequence element and additionally also interacts with a ring of Sm (Smith) core proteins; these particles direct higher-order spliceosome assembly and maturation. In total, seven Sm proteins (SmB／SmB', SmD1, SmD2, SmD3, SmE, SmF, and SmG) are arranged onto each snRNA in order to form a snRNP. Sm proteins were initially identified as the target for antibodies present in the serum of lupus patients and individual Sm proteins were named after the patient whose serum contained antibodies specific for the Sm complex (Kattah, Kattah, & Utz, 2010). In addition to the scaffold formed by the ring of Sm proteins, snRNPs also contain unique snRNP-specific protein components (i.e. U1A, U1C and U1-70K in U1 snRNP). In total, the major spliceosome contains ~200 proteins and five unique snRNAs, named U1, U2, U4, U5 and U6; the names of the snRNPs correspond with the snRNA that they are built upon (Brow, 2002; Hoskins et al., 2011). It should be noted here for completeness that in contrast to the other snRNPs, the U6 scaffold is built from LSm (like Smith) proteins, which are homologues of the Sm proteins present in all of the other snRNPs (L. Zhou et al., 2014).

The spliceosome consists of the U1, U2 and U4/U6•U5 tri-snRNPs, each of which contains a unique RNA or in the case of the tri-snRNP three unique RNAs (Grabowski, Seiler, & Sharp, 1985; Jurica & Moore, 2003; Rappsilber, Ryder, Lamond, & Mann, 2002; Z. Zhou, Licklider, Gygi, & Reed, 2002). These snRNAs are U-rich and have a specific secondary structure (Bringmann & Luhrmann, 1986; Hinterberger, Pettersson, & Steitz, 1983).

Spliceosome assembly is directed by conserved intron sequences and proceeds in a stepwise progression (Fig. 1-1B) (Das & Reed, 1999; Konarska & Sharp, 1986; Z. Zhou et al., 2002). Assembly, activation and disassembly of the spliceosome as it progresses through the spliceosome cycle requires extensive remodelling of its structure and composition, which is driven by eight evolutionarily conserved DExD/H-box, RNA-dependent ATPases (adenosine triphosphatases)/helicases (Cordin, Hahn, & Beggs, 2012; Staley & Guthrie, 1998). This successive pathway in spliceosome assembly/dynamics both ensures fidelity of splice site selection and presents multiple points where alternative splicing events are determined.

Significant parallels exist between the RNA domains of group II introns and the RNA components of the spliceosome for nuclear pre-mRNA introns suggesting that snRNAs may have evolved from group II intron sequences, as all snRNAs (with the exception of U4) have functional counterparts in group II introns (Cech, 1986; Valadkhan, 2010).

1-4.1. Intron recognition and E complex assembly

Before the splicing cycle begins, the H complex (named after the hnRNP particles that it contains) assembles on the pre-mRNA. It contains SR (serine-arginine) proteins and hnRNP (heterogeneous nuclear ribonucleoprotein) proteins. The SR family of proteins contains 12 identified core members in humans to date, and these proteins contain a C-terminal region of repeating SR dipeptides (Z. Zhou & Fu, 2013). SR proteins bind ESE (exonic splicing enhancer) sequences within the pre-mRNA transcript, which are RNA sequence motifs that promote correct intron/exon boundary definition and splice site selection by the splicing machinery. In contrast to SR proteins, hnRNPs repress splice site usage by binding ESS (exonic splicing silencer) sequences. There are 13 families of hnRNPs expressed in humans (Busch & Hertel, 2012).

Although the H complex is not a necessary precursor to spliceosome assembly, it is important because it can compete with and regulate the canonical pathway (Bennett, Pinol-Roma, Staknis, Dreyfuss, & Reed, 1992; Blencowe & Graveley, 2010).

The intron is the basic structural scaffold upon which the spliceosome complexes are built and canonically the splicing cycle begins with the binding and base-pairing of the 5′ end of U1 snRNA with the 5′ SS (Siliciano & Guthrie, 1988). This event defines formation of the E (early) complex in an ATP-dependent fashion, upon which the intron is committed to completing the splicing cycle (Jamison, Crow, & Garcia-Blanco, 1992; Legrain, Seraphin, & Rosbash, 1988; Seraphin & Rosbash, 1989). Additionally, the E complex also contains the non-snRNP proteins SF1 (splicing factor 1), and the heterodimer U2AF (U2 auxiliary factor, consisting of a small and large subunit, hereafter referred to as U2AF-S and U2AF-L, respectively), which bind conserved sequences at the 3′ SS (Berglund, Abovich, & Rosbash, 1998; Berglund, Chua, Abovich, Reed, & Rosbash, 1997; T. Huang, Vilardell, & Query, 2002; Jamison et al., 1992; Jamison & Garcia-Blanco, 1992; Merendino, Guth, Bilbao, Martinez, & Valcarcel, 1999; Michaud & Reed, 1991; Query, Strobel, & Sharp, 1996; Seraphin & Rosbash, 1989; S. Wu, Romfo, Nilsen, & Green, 1999; Zamore, Patton, & Green, 1992; Zorio & Blumenthal, 1999a). Hereafter, the heterotrimer consisting of SF1 and the U2AF heterodimer will be referred to as U2AF/SF1.

1-4.2. Formation of the A complex

Recognition of the 3′ SS involves two steps: initial association of U2AF/SF1 within the E complex followed by recruitment of U2 snRNP to the BPS via interactions with U2AF and possibly U1 snRNP. Outside of the Sm core proteins, U2 snRNP contains two U2-specific protein sub-complexes named SF3A (splicing factor 3A) and SF3B (splicing factor 3B) that are

necessary for initial binding of U2 to the pre-mRNA, and which dissociate prior to the first step of splicing, as well as two structural proteins (U2-A′ and U2-B″) that remain associated with the spliceosome throughout the splicing cycle (Caspary & Seraphin, 1998). With respect to this thesis, SF3B is of special interest and will be described more exhaustively in Appendix *V*. Briefly, SF3B is a complex of seven unique proteins, which binds U2AF in order to recruit U2 snRNP to the BPS (Gozani, Feld, & Reed, 1996; Gozani, Potashkin, & Reed, 1998; Will et al., 2001).

U2AF/SF1 are displaced by U2 snRNP in an ATP-dependent process to form the A (pre-spliceosome) complex; a duplex between U2 snRNA and the BPS of the RNA (Fig. 1-1B) extrudes an unpaired adenosine, which is thereby selected as the nucleophile for the 1st step of splicing (Parker, Siliciano, & Guthrie, 1987; Query, Moore, & Sharp, 1994). U2 snRNP-associated protein sub-complexes SF3A and SF3B assist this process by binding flanking sequences upstream of the BPS and the SF3B component p14 (14 kDa protein), also known as SF3B6 (SF3B subunit 6), directly contacts the BPS (Query et al., 1994). By binding the BPS, SF3B keeps it sequestered until the 1st step of splicing is ready to occur (Golas, Sander, Will, Luhrmann, & Stark, 2003). Accuracy of the snRNA/BPS duplex is ensured by a proof-reading mechanism in which the U2 snRNP components SF3A, SF3B and Prp5 (a DExD/H-box RNA helicase) cooperate in order to prevent association of U2 snRNA with a sub-optimal BPS (Perriman & Ares, 2010; Y. Z. Xu & Query, 2007).

1-4.3. Formation of the B complex and B* complex

In the canonical spliceosome cycle, the A complex is followed by the B (pre-catalytic) complex. The B complex undergoes structural rearrangements in order to form the active site of

the spliceosome, converting the B complex into the B* (catalytically activated) complex, which undergoes further structural rearrangements to become the C (catalytic step 1) complex. The B, B*, and C complexes require the involvement of a unique snRNP called the U4/U6•U5 tri-snRNP which differs from U1 and U2 snRNPs in that it is built upon an assembly of three separate snRNAs instead of just one. Within the tri-snRNP, U4 and U6 duplex with each other over an extensive region, and this duplex must be unwound for splicing to proceed, a process promoted by the U5 snRNP component Brr2 (bad response to refrigeration 2), a DExD/H-box RNA helicase (Hardin, Warnasooriya, Kondo, Nagai, & Rueda, 2015; T. H. Nguyen et al., 2015; Noble & Guthrie, 1996; Staley & Guthrie, 1999; D. Xu, Nouraini, Field, Tang, & Friesen, 1996). U5 is the largest snRNP and contains seven proteins outside of the Sm assembly, including the critically important, large scaffolding protein Prp8, which forms the core of the spliceosome and interacts with Brr2 helicase (Nancollis, Ruckshanthi, Frazer, & O'Keefe, 2013).

Once the A complex is formed, U1 snRNP associates with and recruits the U4/U6•U5 tri-snRNP to the 5′ SS. Prp28 (a DExD/H-box RNA helicase) destabilizes the U1/5′ SS duplex allowing the tri-snRNP to completely displace U1 snRNA from the 5′ SS, thereby forming the B complex, which subsequently rearranges to form the C complex (Ares & Weiser, 1995; Brow, 2002; Konarska & Sharp, 1987; Madhani & Guthrie, 1994; Nilsen, 1994; P. A. Sharp, 1991; Staley & Guthrie, 1998).

In the displacement of U1 snRNA by the tri-snRNP, the U4/U6 duplex is unwound. This allows U6 to form a U2/U6 duplex, as well as a new 5′ SS/U6 duplex with the intron; the duplex with the 5′ SS is formed using the ACAGAGA box of U6 snRNA (Bindereif, Wolff, & Green, 1990; T. H. D. Nguyen et al., 2016). The formation of these new duplexes is coupled with the dissociation of U1 and U4 snRNPs from the spliceosome. Upon dissociation of U1 and U4

snRNPs, the association of U5 and U6 with the spliceosome is stabilized by the NTC (prp NineTeen associated complex) which is a complex of non-Sm spliceosomal proteins involved in the regulation of splicing. It is also required for the dissociation of the LSm proteins from U6 snRNA (Chan, Kao, Tsai, & Cheng, 2003; Hogg, McGrail, & O'Keefe, 2010). Together, these rearrangements play a role in forming the active site of the spliceosome which converts the B complex into the B* complex.

1-4.4. Splicing catalysis: formation of the C complex

Conversion of the B* complex to the C complex occurs when Prp2 (a DExD/H-box RNA helicase) displaces SF3A and SF3B from the spliceosome, thereby activating the 1st step of splicing (Lardelli, Thompson, Yates, & Stevens, 2010). Upon completion of the 1st step, Prp16 (a DExD/H-box RNA helicase) proofreads the 1st step by repressing sub-optimal splice sites and also remodels the spliceosome in order to allow the 2nd step of splicing to proceed (Koodathingal, Novak, Piccirilli, & Staley, 2010; Wahl, Will, & Luhrmann, 2009).

1-4.5. Post-spliceosomal complex & disassembly

Prp22 (a DExD/H-box RNA helicase) proofreads exon ligation by sensing aberrant substrates during the 2nd step of splicing and cooperates with Brr2 helicase to release the mature mRNA (Schwer, 2008).

In addition to the two helicases previously discussed (Prp16 & Prp22), there is a final DExD/H-box RNA helicase, Prp43, that functions to terminate defective splicing (stalled by either Prp16 or Prp22) as well as to complete a successful splicing reaction by stimulating the

release of the splicing substrate and disassembly of the spliceosome (Koodathingal et al., 2010; Pandit, Lynn, & Rymond, 2006).

1-5. Current state of the splicing field and recent structural advances using cryo-EM

With respect to the nuclear pre-mRNA splicing cycle, an extensive body of literature exists for the biochemical analyses of the spliceosome as well as structures of its various sub-components ranging from individual proteins to larger protein-protein and protein-RNA complexes. However, despite the volume and depth of this body of literature it does not provide sufficient information to piece together an integrated structural model of the spliceosome cycle. This is because the spliceosome is far too complex and dynamic to have a coherent understanding of how its components operate together to achieve splicing without a complete or near-complete, experimentally derived, atomic level or near-atomic level structural model of the spliceosome at discrete steps in the spliceosome cycle. Additionally, this integrated structural model of the spliceosome cycle is a necessary framework and foundation to resolve conflicting data in the literature.

Despite the impasse of using X-ray crystallography to achieve the spliceosome structure, this goal was realized recently using cryo-EM (cryogenic electron microscopy). This was made possible because of several improvements in both the hardware and software of cryo-EM technology each of which increase the maximum resolution attainable (Koning & Koster, 2009; Koning, Koster, & Sharp, 2018). The most important of these was the introduction of direct electron detectors, which supersede and are much more sensitive than the older method of collecting electron micrographs on photographic film. Additionally, direct electron detectors allow a fast readout reaching several hundred frames per second, allowing movies of the sample

to be recorded in place of single images which improves resolution even further by allowing the detection of individual electrons and permitting the detection and elimination of radiation-induced distortions in the sample (McMullan, Faruqi, Clare, & Henderson, 2014; Merk et al., 2016; Xuong et al., 2007). In addition to this, superior particle averaging techniques have also been developed for use in model reconstruction, again contributing to an increase in resolution (Grange, Vasishtan, & Grunewald, 2017; Pfeffer et al., 2012; T. H. Sharp, Koster, & Gros, 2016; Unverdorben et al., 2014). Together, these advances in cryo-EM technology have led to a "resolution revolution", where the number of published cryo-EM derived, near-atomic resolution structures of a variety of proteins and protein assemblies has dramatically increased, and this has included the many cryo-EM derived, near-atomic models of the spliceosome that have been published in the past several years (Fernandez-Leiro & Scheres, 2016; Kuhlbrandt, 2014).

1-5.1. Extant spliceosome structures

As noted above, until the spliceosome cycle was structurally characterized via cryo-EM, the biochemical and structural data available in the literature was insufficient to generate an integrated structural model of the spliceosome pathway. The discrete steps of the spliceosome cycle summarized in Section 1-4. are based on the information available prior to the published cryo-EM derived spliceosome structures. Because the biochemical and structural data prior to the cryo-EM derived spliceosome structures does not contain sufficient information to generate a structural model of the spliceosome cycle, this data also cannot be used to identify and define discrete states that are structurally distinct and do not reveal themselves through this data. Therefore, solving the spliceosome structure via cryo-EM has provided greater temporal

resolution to the spliceosome cycle by allowing the identification and definition of distinct spliceosome states that were previously unknown.

The first cryo-EM derived spliceosome structure to be published was the ILS (intron lariat spliceosome) complex of *S. pombe*, which is a post-catalytic state containing the intron lariat (Yan et al., 2015). One major reason why this was the first spliceosome structure to be successfully solved is that most spliceosomes purified from *S. pombe* extracts are post-catalytic, allowing for purification of sufficient quantities of a single state of the spliceosome for structure determination via cryo-EM (W. Chen et al., 2014). All of the spliceosome structures that have been solved and published so far have been from three organisms: *H. sapiens*, *S. pombe*, and *S. cerevisiae*.

In Section 1-5.1.1. below, all the U1 snRNP, U2 snRNP, and U4/U6•U5 tri-snRNP structures that have been solved via cryo-EM have been catalogued. These structures do not represent any individual step in the spliceosome cycle. However, they are very large entities that deliver major, dynamic parts to the spliceosome as it progresses through the spliceosome cycle; therefore, these structures are critical to understanding spliceosome assembly and maturation. Additionally, as with the spliceosome, these structures were not possible until recently due to limitations in cryo-EM technology.

In Section 1-5.1.2 to Section 1-5.1.7. below, all the spliceosome structures that have been solved and published up until now have been catalogued beginning with the earliest stage of the spliceosome cycle and progressing chronologically up until the spliceosome cycle is complete. The published structures of smaller spliceosomal sub-complexes have been omitted for simplicity and clarity. As stated previously, cryo-EM has allowed the resolution of the

spliceosome cycle into a larger number of discrete states, which will be described as they are introduced below.

1-5.1.1. U1 snRNP, U2 snRNP, and U4/U6•U5 tri-snRNP

One U1 snRNP structure has been solved via cryo-EM and deposited in the PDB (Protein Data Bank) (Li et al., 2017). Although several X-ray derived U1 snRNP structures and substructures have been deposited prior to the cryo-EM structure, they are lacking in completeness and/or resolution when compared to the cryo-EM structure (Kondo, Oubridge, van Roon, & Nagai, 2015; Pomeranz Krummel, Oubridge, Leung, Li, & Nagai, 2009; Weber, Trowitzsch, Kastner, Luhrmann, & Wahl, 2010).

The U2 snRNP structure has been solved via cryo-EM and deposited in the PDB under three separate PDB accession codes; two of the PDB accessions are partial structures, and the third is a combined overall structure. These three structures represent the 17S U2 snRNP (Z. Zhang et al., 2020). U2 snRNP exists in three forms (12S, 15S, and 17S); the 12S U2 snRNP associates with SF3B in order to form 15S U2 snRNP, which then associates with SF3A in order to form 17S U2 snRNP (Brosi, Groning, Behrens, Luhrmann, & Kramer, 1993; Brosi, Hauri, & Kramer, 1993). The 17S form of U2 snRNP is the active form, which binds to pre-mRNA during spliceosome assembly (Kramer, Gruter, Groning, & Kastner, 1999).

A total of seven structures have been deposited for the U4/U6•U5 tri-snRNP, which include the overall structure as well as three substructures (foot, head and body region) that form substantial portions of the complete structure (Agafonov et al., 2016; Charenton, Wilkinson, & Nagai, 2019; T. H. D. Nguyen et al., 2016; R. X. Wan et al., 2016).

All of the cryo-EM derived U1 snRNP, U2 snRNP, and U4/U6•U5 tri-snRNP structures that have been discussed are catalogued in Table 1-1 below.

Table 1-1: Cryo-EM derived U1 snRNP, U2 snRNP, and U4/U6•U5 tri-snRNP structures

Structure	PDB accession code	PDB reported resolution	Species
U1 snRNP	6N7X	3.6 Å	*S. cerevisiae*
17S U2 snRNP (5′ domain)	6Y50	4.1 Å	*H. sapiens*
17S U2 snRNP (low resolution part)	6Y53	7.1 Å	*H. sapiens*
17S U2 snRNP	6Y5Q	7.1 Å	*H. sapiens*
U4/U6•U5 tri-snRNP (foot region)	5GAM	3.7 Å	*S. cerevisiae*
U4/U6•U5 tri-snRNP (head region)	5GAO	4.2 Å	*S. cerevisiae*
U4/U6•U5 tri-snRNP (body region)	5GAP	3.6 Å	*S. cerevisiae*
U4/U6•U5 tri-snRNP (overall structure)	5GAN	3.7 Å	*S. cerevisiae*
U4/U6•U5 tri-snRNP (overall structure)	3JCM	3.8 Å	*S. cerevisiae*
U4/U6•U5 tri-snRNP (overall structure)	3JCR	7.0 Å	*H. sapiens*
U4/U6•U5 tri-snRNP (overall structure)	6QW6	2.9 Å	*H. sapiens*

1-5.1.2. H complex and E complex

No structure solution exists in the literature for the H complex. The H complex is not part of the canonical splicing pathway and has been shown to assemble on RNAs lacking functional splice sites, indicating that it is not specific to splicing substrates (Konarska & Sharp, 1986; Reed, 1990). For these reasons, the only defining characteristic of the H complex is that it consists of an RNA sequence to which one or more SR proteins and/or one or more hnRNPs is bound. Therefore, no meaningful structure solution will likely emerge for this complex because it is not a discrete complex and the indiscrete complexes classified as H complex are not expected to contain any unifying structural features.

Regarding the E complex, a study was published in 2019 in which two cryo-EM derived structure solutions were reported for the *S. cerevisiae* E complex which differ from each other in the model pre-mRNA sequence used (Li et al., 2019). One structure was assembled on the ACT1

pre-mRNA and deposited in the PDB (6N7R) with a reported resolution of 3.2 Å. ACT1 is the *S. cerevisiae* gene coding for actin and contains one intron; the transcript for this ORF (open reading frame) is widely used as a reporter for both *in vivo* and *in vitro* splicing assays (Pleiss, Whitworth, Bergkessel, & Guthrie, 2007). The ACT1 pre-mRNA sequence used consists of a 73 nucleotide 5′ exon, a 302 nucleotide intron lacking a cryptic BPS, and a 167 nucleotide 3′ exon (Li et al., 2013). The second structure was assembled on the UBC4 pre-mRNA and deposited in the PDB under the code 6N7P, with a reported resolution of 3.6 Å. UBC4 is the *S. cerevisiae* gene coding for ubiquitin-conjugating enzyme E2 4; the transcript for this ORF contains one small intron that is efficiently spliced *in vitro* as well as *in situ* during smFRET (single-molecule fluorescence resonance energy transfer) experiments (Abelson et al., 2010). The UBC4 pre-mRNA sequence used consists of a 20 nucleotide 5′ exon, a 95 nucleotide intron, and a 32 nucleotide 3′ exon (Abelson et al., 2010). These two structures are particularly relevant to this thesis, because the E complex is the only stage of the canonical spliceosome cycle where U2AF/SF1 are present and these structures will be analyzed in more detail in Section 1-5.2.3.

1-5.1.3. pre-A complex and A complex

The *S. cerevisiae* spliceosome cycle has been successfully stalled in *S. cerevisiae* cell extracts at a stage immediately prior to A complex formation; this newly identified assembly intermediate has been termed the pre-A (precursor to A) complex, and its cryo-EM structure has been solved, which provides insights into how Prp5 helicase proofreads the U2/BPS duplex (Z. W. Zhang et al., 2021). The pre-A complex consists of two major elongated domains, one comprising U1 snRNP and the other comprising U2 snRNP, the connections between the two domains consisting of two main bridges. This structure has been deposited in the PDB under

three accession codes; one represents the U1 snRNP region, one represents the U2 snRNP region, and the final accession represents the composite truncated model. In addition to the pre-A complex cryo-EM structure, one structure has been solved for the *S. cerevisiae* A complex via cryo-EM and deposited in the PDB (Plaschka, Lin, Charenton, & Nagai, 2018). These structures are catalogued below in Table 1-2.

Table 1-2: Cryo-EM derived *S. cerevisiae* pre-A complex and A complex structures

Structure	PDB accession code	PDB reported resolution
pre-A complex (U1 part)	7OQC	4.1 Å
pre-A complex (U2 part)	7OQB	9.0 Å
pre-A complex (composite model)	7OQE	5.9 Å
A complex	6G90	4.0 Å

1-5.1.4. pre-B complex and B complex

Cryo-EM has allowed the identification of a new assembly intermediate in the spliceosome cycle termed the pre-B (precursor to B) complex, which follows the A complex and precedes the B complex. In the pre-B complex, the catalytic center of the spliceosome has not yet formed, even though all the necessary components are present. In the pre-B complex, U1 snRNP and U2 snRNP associate with U4/U6•U5 tri-snRNP. The pre-mRNA has yet to be recognized by U5 snRNA or U6 snRNA. The pre-B complex undergoes remodelling to become the B complex. A total of three structures (two from human and one from *S. cerevisiae* spliceosomes) have been solved for the pre-B complex via cryo-EM and deposited in the PDB; the *S. cerevisiae* structure is represented by two separate substructure accessions in the PDB (Bai, Wan, Yan, Lei, & Shi, 2018; Charenton et al., 2019; Zhan, Yan, Zhang, Lei, & Shi, 2018b).

Prp28 helicase is essential in the events that transition the spliceosome from the pre-B complex to the B complex. Prp28 helicase transfers the 5′ SS of the pre-mRNA from U1 snRNA

to the ACAGAGA box of U6 snRNA. The new U6/5′ SS duplex triggers remodelling of a protein-RNA network in the pre-B complex to induce the relocation of Brr2 helicase and loading of U4 snRNA to the active site of Brr2 helicase. This triggers unwinding of the U4/U6 snRNA duplex allowing U6 to form the catalytic site of the spliceosome and thereby generating the B complex. Four structures have been solved for the B complex via cryo-EM and deposited in the PDB (Bai et al., 2018; Bertram, Agafonov, Dybkov, et al., 2017; Plaschka, Lin, & Nagai, 2017; Zhan et al., 2018b).

All of the cryo-EM derived pre-B complex and B complex structures discussed above are catalogued below in Table 1-3.

Table 1-3: Cryo-EM derived pre-B complex and B complex structures

Structure	PDB accession code	PDB reported resolution	Species
pre-B complex	6QX9	3.3 Å	*H. sapiens*
pre-B complex	6AH0	5.7 Å	*H. sapiens*
pre-B complex (tri-snRNP and U2 snRNP part)	5ZWM	3.4 Å	*S. cerevisiae*
pre-B complex (U1 snRNP region)	5ZWN	3.4 Å	*S. cerevisiae*
B complex	5NRL	7.2 Å	*S. cerevisiae*
B complex	5ZWO	3.9 Å	*S. cerevisiae*
B complex	5O9Z	4.5 Å	*H. sapiens*
B complex	6AHD	3.8 Å	*H. sapiens*

1-5.1.5. pre-B^{act} complex, B^{act} complex, and B* complex

With respect to events in the spliceosome cycle occurring in the transition from the B complex up to and including the B* complex, several time-resolved intermediate states have been visualized by cryo-EM. These structures have been solved from *S. cerevisiae* spliceosomes (seven PDB structure accessions) and human spliceosomes (10 PDB structure accessions). The *S. cerevisiae* structure accessions will be addressed in Section 1-5.1.5.1. and the human structure

acccessions will be addressed separately in Section 1-5.1.5.2., because the large number of time-resolved states solved for the two species do not correspond to one another.

The B^{act} (activated) complex is an important intermediate between the B complex and B* complex that was identified several years before the first cryo-EM based spliceosome structure solution; it was omitted in Fig. 1-1 and Section 1-4. for simplicity and clarity (Bessonov et al., 2010). The B^{act} complex has been solved for both *S. cerevisiae* and human spliceosomes.

The transformation of the B complex into the B^{act} complex occurs via very substantial changes in the protein composition and structure of the spliceosome, which serve to form the catalytically active U2/U6 network that characterizes the B^{act} complex. These events involve the most extensive protein and RNA rearrangements during the assembly and maturation of the splicing machinery and a dramatic exchange of spliceosomal proteins occurs: in higher eukaryotes, B^{act} formation involves the recruitment or stable integration of more than 25 proteins and the loss of about 25 proteins (Agafonov et al., 2011; Bessonov et al., 2010; Kastner, Will, Stark, & Luhrmann, 2019).

Spliceosome activation is initiated by Brr2 helicase, which unwinds the U4/U6 duplex, leading to the release of U4 (Absmeier, Santos, & Wahl, 2016). This allows U6 to form new base pairs with U2 generating U2/U6 helices Ia and Ib and also to restructure, forming an ISL (internal stem-loop). These rearrangements and other events are important to enable U6 to position two Mg^{2+} ions that participate directly in splicing catalysis (Fica et al., 2013).

Presently, little is known about the folding pathway of U2 and U6 snRNA during activation, as well as the mechanisms by which proteins assist the formation of the catalytic RNA network. The extensive remodelling of the spliceosome in this transition means that there are likely additional intermediate stages during this transition in which only a subset of proteins

are exchanged and which may contain additional, transiently interacting splicing factors that stabilize these currently unidentified intermediate states.

1-5.1.5.1. *S. cerevisiae* B^{act} complex and B* complex

Two papers were published in 2016 describing the cryo-EM derived *S. cerevisiae* B^{act} structure (Rauhut et al., 2016; Yan, Wan, Bai, Huang, & Shi, 2016). A follow up study was published in 2021 by the research group led by Yigong Shi. This study investigated the role of Prp2 helicase and its co-activator Spp2 (suppressor of Prp2) in driving the transition of the spliceosome from the B^{act} complex to the B* complex. A new cryo-EM derived structure solution of the *S. cerevisiae* B^{act} complex was described in this paper at an atomic resolution of 2.5 Å allowing the atomic identification of 12 new proteins including Prp2 helicase and Spp2 (Bai, Wan, Yan, et al., 2021).

In 2019, the research group led by Yigong Shi published a study in which four distinct cryo-EM derived structures of the *S. cerevisiae* B* complex were described (R. Wan, Bai, Yan, Lei, & Shi, 2019). These B* complexes were assembled either on ACT1 pre-mRNA, or UBC4 pre-mRNA. During the structure solution process, two discrete substrate-specific conformational states were identified for both B* complexes; B*-a1 and B*-a2 contain ACT1 pre-mRNA, whereas B*-b1 and B*-b2 contain UBC4 pre-mRNA. A major difference between these four complexes is the conformation and location of key RNA elements in the catalytic center of the spliceosome. These include the U2/BPS duplex, U5/5′ exon duplex, and 5′ SS. A dissection of these differences reveals that the precise positioning of U2/BPS duplex in the active site in the close vicinity of the 5′ SS requires stabilization by the splicing factors Cwc25 (complexed with

Cef1 25) and Yju2 in order to allow the 1st step of splicing to proceed, adding to our mechanistic understanding of the 1st step of splicing.

All of the cryo-EM derived *S. cerevisiae* B^{act} and B* complexes described above are catalogued below in Table 1-4.

Table 1-4: Cryo-EM derived *S. cerevisiae* B^{act} complex and B* complex structures

Structure	PDB accession code	PDB reported resolution
B^{act} complex	7DCO	2.5 Å
B^{act} complex	5GM6	3.5 Å
B^{act} complex	5LQW	5.8 Å
B*-a1 complex	6J6H	3.6 Å
B*-a2 complex	6J6G	3.2 Å
B*-b1 complex	6J6N	3.9 Å
B*-b2 complex	6J6Q	3.7 Å

1-5.1.5.2. Human pre-B^{act} complex and B^{act} complex

Little is known about the assembly pathway of the B^{act} complex. In order to address this gap in understanding, cryo-EM was used to solve the structure of human spliceosomes at two successive and previously uncharacterized stages that are intermediate between the B complex and B^{act} complex; both of these intermediate states are known as the pre-B^{act} (precursor to B^{act}) complex. The pre-B^{act} complex lacks a mature catalytic U2/U6 structure; the earlier intermediate complex is known as the pre-B^{act-1} complex and is a precursor of the later intermediate complex known as the pre-B^{act-2} complex. Chase experiments established that these are functional spliceosome intermediates.

A dissection and comparison of the cryo-EM structures of B, pre-B^{act-1}, pre-B^{act-2}, and B^{act} complex establishes that Brr2 helicase and U2 snRNP undergo stepwise repositioning during spliceosome activation which is necessary to proximate U2 snRNA and U6 snRNA so that the

catalytically active U2/U6 structure present in the B^{act} complex can form. SnRNP rearrangements that occur in this process are assisted by several proteins that transiently interact with the spliceosome and a number of mutually exclusive protein-protein and protein-RNA interactions that help drive the directionality of the activation process. Additionally, the scaffold protein Prp8 at the core of the spliceosome has a key role in assisting the folding of the catalytically active U2/U6 RNA network, and it accomplishes this by undergoing a conformational rearrangement (Townsend et al., 2020).

The cryo-EM structure solution of the human B^{act} complex was reported in two separate publications in 2018, and both publications report the identification and characterization of multiple time-resolved states for the B^{act} complex. In the earlier publication, eight major conformational states for the core of the human B^{act} complex were identified; three of these states were combined in order to generate a near-atomic structural model and a combination of all states was used to generate a second composite model which includes all visible densities and was deposited in the PDB as an alanine trace (Haselbach et al., 2018). The later publication reports the structural model of three distinct and successive conformations, referred to as the early, mature, and late states of the human B^{act} complex. These three states differ in a number of protein components including the orientation of the switch loop of the scaffold protein Prp8 (X. Zhang et al., 2018).

The structure solutions for the pre-B^{act-1} and pre-B^{act-2} complexes were deposited in the PDB under multiple entries as partial structures and also as overall structures and all of these accessions are summarized below in Table 1-5, as are all of the PDB accessions of the B^{act} complex discussed above.

Table 1-5: Cryo-EM derived human pre-B^{act} complex and B^{act} complex structures

Structure	PDB accession code	PDB reported resolution
pre-B^{act-1} complex (core)	7ABF	3.9 Å
pre-B^{act-1} complex	7ABG	7.8 Å
pre-B^{act-2} complex (core)	7AAV	4.2 Å
pre-B^{act-2} complex (SF3B/U2 snRNP portion)	7ABH	4.5 Å
pre-B^{act-2} complex	7ABI	8.0 Å
B^{act} complex (core)	6FF4	3.4 Å
B^{act} complex (core)	6FF7	4.5 Å
early B^{act} complex	5Z58	4.9 Å
mature B^{act} complex	5Z56	5.1 Å
late B^{act} complex	5Z57	6.5 Å

1-5.1.6. C complex, C_i complex, and C* complex

As described previously in Section 1-4., the B* complex possesses a fully formed active site and is primed for the 1st step of splicing (also known as the branching reaction) to occur, which produces an intron lariat-3′ exon intermediate and a 5′ exon. Upon completion of the 1st step of splicing, the spliceosome is referred to as the C complex; both products of the 1st step of splicing remain bound to the C complex. A total of two *S. cerevisiae* C complex structures and two human C complex structures have been solved using cryo-EM and published. These four structures are represented in the PDB by a total of six PDB accession codes, and the human C complex contains 11 more proteins than the *S. cerevisiae* C complex (Bertram et al., 2020; Galej et al., 2016; R. Wan, Yan, Bai, Huang, & Shi, 2016; Zhan, Yan, Zhang, Lei, & Shi, 2018a).

In order for the 2nd step of splicing to occur, the branched intron structure generated during the 1st step must be displaced from the catalytic center in order to allow juxtapositioning of the 2nd step reactants. This is achieved by ribonucleoprotein remodelling of the C complex in order to transform the spliceosome into the C* (catalytic step 2) complex, and this process is driven by Prp16 helicase; Prp16 helicase proofreads the 1st step and is thought to bind and pull

the single-stranded RNA sequences in a 3′ → 5′ direction (Koodathingal et al., 2010; Schwer & Guthrie, 1992; Semlow, Blanco, Walter, & Staley, 2016; Wahl et al., 2009).

It is important to note that genetic and biochemical experiments provide evidence for a two-state model of the catalytic spliceosome, in which the conformation of both the C and C* complex exist in equilibrium (Query & Konarska, 2004). A new intermediate, the C_i complex (C complex intermediate), has been identified between these two conformations and the cryo-EM structure solution of this state has been published. This structure reveals that the binding of protein factors specific to either branching or exon ligation establishes the equilibrium between these two conformations. With respect to driving the C → C* transition, exon ligation factors Slu7 (synergistic lethal with U5 snRNA 7) and Prp18 bind the C_i complex weakly prior to Prp16 helicase action thereby priming the C complex for conversion to the C* complex by Prp16 helicase. After Prp16 helicase action, pre-bound Slu7 and Prp18 bind strongly to promote exon ligation (Wilkinson, Fica, Galej, & Nagai, 2021).

With respect to the C* complex, a total of four cryo-EM derived structures have been solved and published, in which the spliceosome has been remodelled from the C conformation to the C* conformation, but the 2nd step of splicing (exon ligation) has yet to occur. Two of these structure solutions are of the *S. cerevisiae* spliceosome and two are of the human spliceosome; one of the *S. cerevisiae* structures has been deposited in the PDB under two accession codes (Bertram, Agafonov, Liu, et al., 2017; Fica et al., 2017; Yan, Wan, Bai, Huang, & Shi, 2017; X. Zhang et al., 2017).

All of the cryo-EM derived C complex, C_i complex, and C* complex structures discussed above are catalogued below in Table 1-6.

Table 1-6: Cryo-EM derived C complex, C_i complex, and C* complex structures

Structure	PDB accession code	PDB reported resolution	Species
C complex	5GMK	3.4 Å	*S. cerevisiae*
C complex (core)	5LJ3	3.8 Å	*S. cerevisiae*
C complex (overall)	5LJ5	10.0 Å	*S. cerevisiae*
C complex	5YZG	4.1 Å	*H. sapiens*
C complex (core)	6ZYM	3.4 Å	*H. sapiens*
C complex (periphery)	7A5P	5.0 Å	*H. sapiens*
C_i complex (bound to Slu7 & Prp18)	7B9V	2.8 Å	*S. cerevisiae*
C* complex (core)	5MPS	3.9 Å	*S. cerevisiae*
C* complex (core + Prp22 + U2 snRNP)	5MQ0	4.2 Å	*S. cerevisiae*
C* complex	5WSG	4.0 Å	*S. cerevisiae*
C* complex	5MQF	5.9 Å	*H. sapiens*
C* complex	5XJC	3.6 Å	*H. sapiens*

1-5.1.7. P complex and ILS complex

The 2nd step of splicing (exon ligation) occurs spontaneously in the C* complex, in which the 3′ exon is excised from the intron lariat and covalently joined to the 5′ exon. Upon completing exon ligation, the spliceosome is referred to as the P (post-catalytic) complex. In this state, the end products of splicing remain bound to the spliceosome and release of the ligated exons requires the action of Prp22 helicase in order to generate the ILS complex, which only contains the intron lariat (Company, Arenas, & Abelson, 1991; Mayas, Maita, & Staley, 2006; Schwer & Gross, 1998; Schwer & Meszaros, 2000; Semlow et al., 2016; Wagner, Jankowsky, Company, Pyle, & Abelson, 1998).

In addition to its ATPase and helicase dependent role in mRNA release, Prp22 also proofreads the 3′ SS for exon ligation, sensing aberrant substrates and allowing for alternative 3′ SS choice by an ATP-dependent mechanism (Mayas et al., 2006; Schwer, 2008; Semlow et al., 2016). It is thought that Prp22 rejects 3′ SS sequences in the C* complex by dislodging them from the active site in order to allow the selection of an alternative 3′ SS prior to exon ligation.

Prp22 may also possess a third ATP-independent role in assisting the exon ligation reaction (Ohrt et al., 2013; Schwer & Gross, 1998). Just as Prp22 is responsible for alternative 3′ SS selection as part of its proofreading role in the 2nd step of splicing, Prp16 helicase is responsible for alternative BPS selection in the B* complex as part of its proofreading role in the 1st step of splicing. However, the transient nature of the B* complex means that the mechanism for this latter activity remains to be determined (Semlow et al., 2016).

A total of five cryo-EM derived P complex structures have been published; three of these structures represent *S. cerevisiae* and two represent the human P complex (Bai, Yan, Wan, Lei, & Shi, 2017; Fica, Oubridge, Wilkinson, Newman, & Nagai, 2019; S. Liu et al., 2017; Wilkinson et al., 2017; X. F. Zhang et al., 2019). The overall organization and detailed structural features of the P complex closely resemble the C* complex in both *S. cerevisiae* and human spliceosomes. However, there are important and surprising differences between the human and *S. cerevisiae* P complex structures that have implications for the regulation of alternative 3′ SS choice during the exon ligation reaction.

In *S. cerevisiae*, Prp18 stabilizes the catalytic conformation during exon ligation but is absent in both the C* and P complexes of humans (Bertram, Agafonov, Liu, et al., 2017; Ilagan, Chalkley, Burlingame, & Jurica, 2013; Jurica, Licklider, Gygi, Grigorieff, & Moore, 2002). Despite this, depletion of Prp18 from HeLa extracts abolishes exon ligation of β-globin pre-mRNA, suggesting that Prp18 operates in humans as it does in *S. cerevisiae* in order to promote the splicing of a subset of human transcripts; similar to humans, genetic depletion of Prp18 in *S. pombe* abolishes splicing in an intron-specific manner, further supporting this possibility (Horowitz & Krainer, 1997; Vijaykrishna et al., 2016). Additionally, the human P complex structures reveal that the metazoan-specific alternative splicing factors FAM32A (family with

sequence similarity 32 member A), CACTIN (renal carcinoma antigen NY-REN-24), SDE2 (silencing defective 2), and NKAP (nuclear factor kappa B-activating protein) have been co-opted in human spliceosomes in order to fulfill roles played by Prp18 in the *S. cerevisiae* spliceosome and stabilize the catalytic conformation during exon ligation.

FAM32A promotes mRNA formation for proapoptotic genes, acting as a tumor suppressor and is also known as OTAG-12 (ovarian tumor associated gene-12); it is down-regulated in a mouse model of ovarian tumor development (X. Chen, Zhang, Aravindakshan, Gotlieb, & Sairam, 2011). NKAP is involved in T cell development, binds a genome-wide set of exon sequences, and depletion of NKAP *in vivo* reduces splicing efficiency (Burgute et al., 2014). With respect to CACTIN and SDE2, although they have been referred to as metazoan-specific factors, orthologues do exist for both proteins in *S. pombe*. In this yeast species these factors promote splicing of the same specific subset of introns. This observation, combined with their roles in the human P complex structure suggest that these proteins function cooperatively in the spliceosome (Lorenzi et al., 2015; Thakran et al., 2018). Taken together, these observations indicate the existence of mechanisms in the metazoan spliceosome to fine-tune alternative splicing at the exon ligation stage by influencing 3′ SS selection.

After the completion of both steps of splicing, Prp43 helicase releases the intron lariat, generating the ILS complex. Subsequent disassembly of the ILS complex and recycling of spliceosomal components demarcates the end of one complete splicing cycle (Arenas & Abelson, 1997; Martin, Schneider, & Schwer, 2002; Tanaka, Aronova, & Schwer, 2007; Tsai et al., 2005). A total of four cryo-EM structure solutions have been reported for the ILS complex, one from *S. cerevisiae*, one from *S. pombe*, and two from humans; the two human structures represent the ILS before (ILS1), and after (ILS2) Prp43 is loaded (R. Wan, Yan, Bai, Lei, & Shi, 2017; Yan et

al., 2015; X. F. Zhang et al., 2019). The structure and organizational features of all four structures are very similar to one another. The *S. pombe* and human ILS1 structures lack Prp43, while the *S. cerevisiae* and human ILS2 structures contain Prp43. It is likely that the *S. pombe* structure represents an early state soon after remodelling of the spliceosome by Prp22 and that the *S. cerevisiae* structure represents a very late state just prior to disassembly (as evidenced by the presence of weak EM density in the RNA-binding tunnel of Prp43). Based on the location of Prp43 in the spliceosome structures and previous biochemical studies, it is thought that Prp43 either pulls the intron lariat or the 3′ end of U6 snRNA in order to destabilize and disassemble the ILS complex (Arenas & Abelson, 1997; Bohnsack et al., 2009; Fourmann et al., 2016; Martin et al., 2002). These two mechanisms are not mutually exclusive, and it is possible that both are used.

All of the cryo-EM derived P complex and ILS complex structures discussed above are catalogued below in Table 1-7.

Table 1-7: Cryo-EM derived P complex and ILS complex structures

Structure	PDB accession code	PDB reported resolution	Species
P complex	6EXN	3.7 Å	*S. cerevisiae*
P complex	5YLZ	3.6 Å	*S. cerevisiae*
P complex	6BK8	3.3 Å	*S. cerevisiae*
P complex	6QDV	3.3 Å	*H. sapiens*
P complex	6ICZ	3.0 Å	*H. sapiens*
ILS complex	5Y88	3.5 Å	*S. cerevisiae*
ILS complex	3JB9	3.6 Å	*S. pombe*
ILS1 complex	6ID0	2.9 Å	*H. sapiens*
ILS2 complex	6ID1	2.9 Å	*H. sapiens*

1-5.2. Summary of cryo-EM derived spliceosome structures

The canonical spliceosome cycle is now almost fully structurally characterized via cryo-EM. This has allowed the integration of decades of biochemical and structural studies on the spliceosome. Additionally, cryo-EM has allowed the identification and characterization of a much larger number of time-resolved states and sub-states in the spliceosome cycle that cannot be detected through biochemical inquiry or reductionist structural studies alone. Currently, the spliceosome cycle is known to progress directionally through the following sequence of complexes: H → E → pre-A → A → pre-B → B → pre-B^{act} → B^{act} → B* → C → C_i → C* → P → ILS. Structural characterization of the spliceosome cycle has provided many new and unexpected insights. Combined with decades of previous study, this body of knowledge provides a foundation for more detailed study of the spliceosome cycle and the role of splicing in the regulation of eukaryotic gene expression.

1-5.2.1. General properties of the spliceosome

The plastic nature of the spliceosome is a functionally important feature that is dependent on several properties that are observed in the cryo-EM structures. These structures do not display uniform resolution and different regions of the spliceosome vary significantly in their reported resolution. The best-resolved parts of all structures mainly consist of the stable and well-defined catalytic U2/U6 RNP core, as well as the U5 snRNP components. Outside of the core, the next layer of splicing components are usually determined at lower resolution. This layer surrounding the core can change significantly in composition and spatial positioning as the spliceosome cycle progresses, and the lower resolution typically reported for this layer is due to its plasticity which is required to enable the many remodelling steps in the spliceosome cycle. It is important to

understand the mechanics of this plastic layer since it contributes to the functional activation and therefore likely also the regulation of the spliceosome.

Cryo-EM has revealed that the spliceosome is organized around certain kinetic properties. The thermal energy required to remodel the spliceosome and drive the spliceosome cycle forward from any one state to another is very low ($\leq 3\ k_BT$), making the spliceosome cycle stochastic. The spliceosome samples many conformations and each time a new component binds the entire conformational sampling space is modified. This mechanism transitions the spliceosome from one conformational sampling space to another on a productive path toward the catalytically active spliceosome. In this way, the extremely large conformational changes in the plastic layer of the spliceosome have direct consequences on the formation of the catalytic core (Haselbach et al., 2018).

1-5.2.2. Comparison of *S. cerevisiae* and human spliceosomes

With the exception of the initial ILS structure from *S. pombe*, all of the reported structures are derived from *S. cerevisiae* and human spliceosomes. The general architecture is largely conserved between both species. Both share a large number of evolutionarily-conserved core proteins; the main overall difference is that human spliceosomes are larger and more complex than *S. cerevisiae* spliceosomes with many additional protein components and the human orthologues of conserved proteins typically contain additional unstructured regions (Agafonov et al., 2011; Bessonov, Anokhina, Will, Urlaub, & Luhrmann, 2008; Fabrizio et al., 2009; Jurica et al., 2002; Jurica & Moore, 2003; Kastner et al., 2019; Korneta & Bujnicki, 2012; Wahl et al., 2009). Cryo-EM structures reveal how some additional human proteins are integrated into the conserved architecture of the spliceosome. However experimental

characterization of these proteins and their functions in the context of the spliceosome cycle is lacking. This also means that it has not yet been conclusively established whether these proteins are constitutive components of the human spliceosome or whether they regulate alternative splicing of a subset of pre-mRNAs in a spatio-temporally controlled fashion.

Not only are human spliceosomes larger and more complex, but the human spliceosome cycle is also more complex. In both *S. cerevisiae* and humans, protein composition changes from one state to the next as the spliceosome cycle progresses and many metazoan-specific factors associate with the human spliceosome at distinct stages in this cycle (Agafonov et al., 2011). During initial assembly, *S. cerevisiae* A complex formation depends on a minimal interaction between the U1 and U2 snRNPs, bringing the 5′ SS and BPS into one assembly. However, mammalian A complex formation is promoted and regulated by many alternative splicing factors (Pan et al., 2008; C. W. Smith & Valcarcel, 2000). The directionality of the spliceosome cycle is largely driven by a set of eight conserved DExD/H-box RNA helicases. However, the human B → B* transition additionally also requires an additional RNA helicase called AQR (aquarius), which is absent in *S. cerevisiae* indicating that conformational rearrangements necessary for catalytic activation in human spliceosomes are more complex (De et al., 2015). Finally, in the intricate cascade of RNP rearrangements that occurs during splicing catalysis, the human spliceosome cycle moves through intermediate RNP conformations not found in *S. cerevisiae*.

The simplicity and experimental tractability of *S. cerevisiae* has made it a very important model system to establish foundational principles of splicing through both biochemical and structural techniques. Accordingly, the emergence of human spliceosome structures has lagged significantly behind *S. cerevisiae*. However, it is essential to study the splicing machinery of developmentally complex, multicellular eukaryotes such as humans since the additional features

present in the splicing machinery of these organisms are bound to deliver insights into conserved mechanisms of splicing regulation and alternative splicing involved in large scale proteome expansion and spatio-temporal control of gene expression, thereby making developmentally complex, multicellular life possible.

1-5.2.3. Insights into 3′ SS recognition by U2AF, SF1 and p14

This thesis aims to investigate early events of spliceosome assembly involving U2AF, SF1 and p14 that determine the 3′ SS. With respect to the existing cryo-EM derived spliceosome structures and currently known sequence of complexes that define the directional progression of the spliceosome cycle, the E complex and pre-A complex both contain at least one of these splicing factors and require a deeper analysis. This will provide a clearly defined context for this thesis and its goals within the splicing field.

1-5.2.3.1. Limitations of *S. cerevisiae* as a model to study U2AF, SF1 and p14

Both the two E complex structures and the pre-A complex structure are derived from *S. cerevisiae* which cannot be used to accurately model the early events of splicing at the 3′ SS involving U2AF, SF1 and p14 because the mechanism of 3′ SS recognition has been simplified in this organism (reviewed below).

SF1 is well conserved between *S. cerevisiae* and humans; the yeast orthologue is called either Msl5 (Mud synthetic lethal 5) or Bbp (branchpoint bridging protein). A functional orthologue does exist in *S. cerevisiae* for U2AF-L called Mud2 (mutant U1 die 2); however, it is poorly conserved and does not possess the same domain organization as the human orthologue. In *S. cerevisiae*, intron recognition is initiated by recognition of the 5′ SS by U1 snRNP and

recognition of the BPS by Mud2-Bbp heterodimer thereby forming the E complex; the 3′ SS is not recognized until later (Abovich & Rosbash, 1997; Ruby & Abelson, 1988; Seraphin, Kretzner, & Rosbash, 1988; Siliciano & Guthrie, 1988). Bbp binds directly to the BPS and Mud2 fulfills a similar role as U2AF-L, although it is not clear how it interacts with pre-mRNA (Abovich & Rosbash, 1997; Berglund et al., 1997; Jacewicz, Chico, Smith, Schwer, & Shuman, 2015).

Despite having a functional orthologue of U2AF-L, no orthologue has been found in *S. cerevisiae* for U2AF-S (Abovich, Liao, & Rosbash, 1994; Q. Wang, Zhang, Lynn, & Rymond, 2008). In mammals, U2AF-L (which recognizes the PPT) is tightly complexed to U2AF-S (which recognizes the AG di-nucleotide at the 3′ SS); binding of U2AF-S to the AG di-nucleotide is critical for a subset of introns referred to as 'AG-dependent' where the AG is required for E complex formation and the 1st step of splicing (Merendino et al., 1999; S. Wu et al., 1999; Zorio & Blumenthal, 1999a). In *S. cerevisiae* however, all introns are AG-independent (Umen & Guthrie, 1995).

In humans, p14 is a part of the seven member SF3B complex (Section 1-4.2.), a salt-dissociable component of U2 snRNP. Although SF3B is highly conserved throughout evolution, no p14 orthologue exists in *S. cerevisiae*.

Section 1-5.2.3.2. below will address the roles of Mud2 and Bbp in the existing *S. cerevisiae* spliceosome structures. Section 1-5.2.3.3. will address the feature of 'intron definition' vs. 'exon definition' in early splicing events because it is necessary to understand this feature in order to fully appreciate the E complex structures and how they are integrated into the spliceosome cycle.

1-5.2.3.2. Mud2 and Bbp in cryo-EM derived *S. cerevisiae* spliceosome structures

The E and pre-A complex structures contain the Mud2-Bbp heterodimer, but it has not been modelled in any of them because the density is too weak to accomplish this. The E complex structures will be reviewed first, followed by a discussion of the pre-A complex structure.

As mentioned previously in Section 1-5.1.2., one of the E complexes was assembled on ACT1 pre-mRNA and the other was assembled on UBC4 pre-mRNA. The 5′ SS → BPS region of the ACT1 intron is much longer (265 nt) than the same region in UBC4 (58 nt), and the ACT1 complex structure contains a ~25 bp double helix between the 5′ SS and BPS which is absent in the UBC4 complex. This indicates that secondary structures are a key mechanism to proximate conserved intron motifs necessary for splicing catalysis so that spliceosome assembly can be initiated. Corroborating this, ACT1 is predicted to form long stem-like structures and mutations that abolish these structures inhibit splicing (Li et al., 2019).

In *S. cerevisiae*, proximation of the 5′ SS and BPS is necessary to initiate the spliceosome cycle and the U1 snRNP protein Prp40 is important in this process (Abovich & Rosbash, 1997). Prp40 forms a stable dimer with Snu71 (small nuclear ribonucleoprotein associated 71) and a trimer with Snu71-Luc7 (lethal unless cap-binding complex is produced 7) (Ester & Uetz, 2008; Gornemann et al., 2011; Li et al., 2017). The Prp40 WW domains (named so because the WW domain contains two strictly conserved tryptophans) also directly interact with the N-terminal domain of Bbp (Abovich & Rosbash, 1997; Wiesner, Stier, Sattler, & Macias, 2002). In the ACT1 complex structure, there is a large volume of weak density close to the pre-mRNA double helix, which is best interpreted as the Mud2-Bbp dimer. This density is not obvious in the UBC4 complex structure, potentially because UBC4 lacks the pre-mRNA helix that brings the BPS close to the 5′ SS. Prp40 therefore bridges both ends of the intron by interacting with the U1

snRNP components U1-70K, Snu71 and Luc7 through its FF domains (named so because the FF domain contains two strictly conserved phenylalanines) and interacting with Bbp through its WW domains. The ~60 residue linker between the WW and FF domains is predicted to be disordered, explaining why density corresponding to Mud2-Bbp is difficult to observe. The general features of the E complex structure discussed above are visualized in Fig. 1-2.

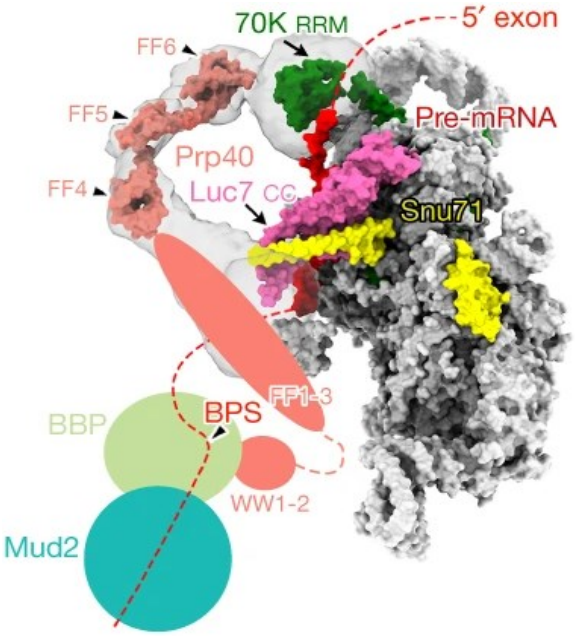


Figure 1-2. Overall structure of the *S. cerevisiae* E complex. Features from both the ACT1 and UBC4 complexes are included. Surface representations are provided for proteins that interact or possibly interact with Prp40 in different colours. Locations of proteins or protein domains that are not modelled owing to weak densities are indicated by shapes. U1-70K = green. Prp40 = salmon. Luc7 = pink. Snu71 = yellow. Pre-mRNA = red. Red dashed line represents a hypothetical path of the pre-mRNA regions that cannot be modelled (adapted from Li et al., 2019).

The *S. cerevisiae* pre-A complex structure represents a state directly before the A complex is formed and provides insight into how Prp5 helicase proofreads the U2/BPS duplex. This complex consists of two main elongated domains comprising the U1 snRNP and U2 snRNP; these two substructures are connected by two main bridges. Mud2-Bbp is present in the pre-A complex but cannot be modelled precisely based on EM density alone, presumably due to structural flexibility. CXMS (chemical cross-linking mass spectrometry) indicates that Mud2-Bbp is likely located near the U2/BPS helix and remains bound to the Prp40 WW domain.

An EM map that is low-pass filtered to 30 Å resolution reveals weak EM density below the U2/BPS helix directly downstream of the BPS that probably corresponds to Mud2-Bbp. Formation of the U2/BPS helix requires Bbp to transfer the BPS to U2 snRNA and therefore Bbp is expected to be displaced from the BPS in the pre-A complex. Retention of Mud2-Bbp close to the U2/BPS helix is consistent with the binding of Mud2 to the intron downstream of the BPS.

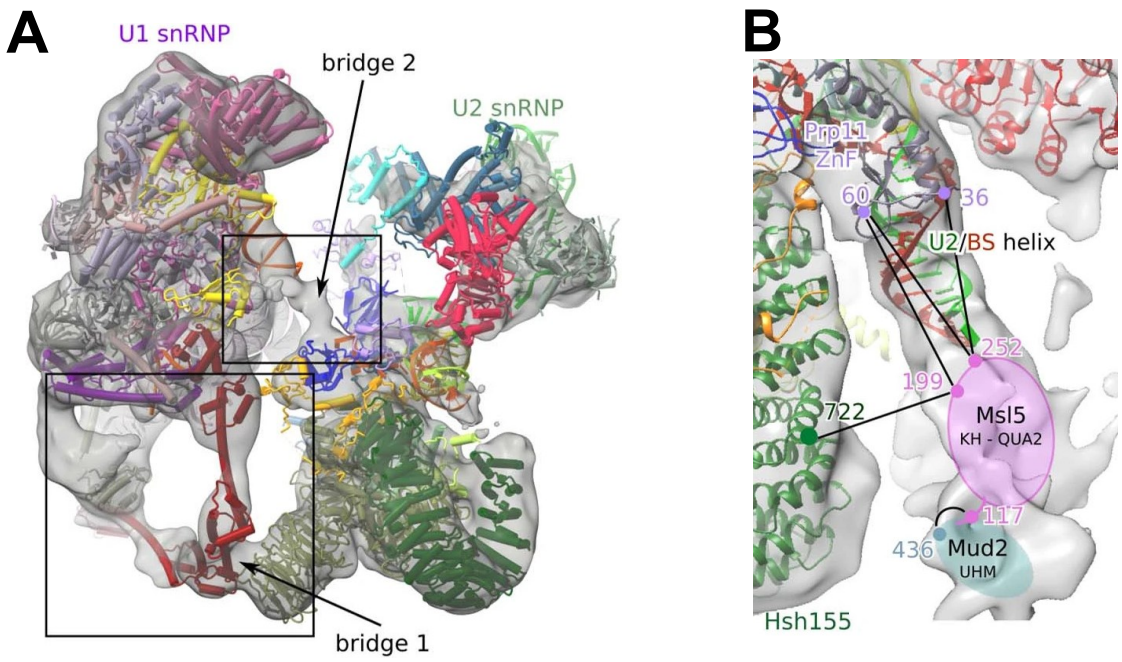


Figure 1-3. Overall structure of the *S. cerevisiae* pre-A complex. (A) Fit of the molecular model of the pre-A complex into the EM density (low-pass filtered). The two main bridges that connect U1 snRNP and U2 snRNP are indicated by arrows. (B) An EM map low-pass filtered to 30 Å resolution reveals density that probably corresponds to Mud2-Bbp. Protein crosslinks supporting the localization of Mud2-Bbp adjacent to the U2/BPS helix are shown. Numbers (colour coded to match protein colours) indicate the positions of crosslinks, which are connected by black lines. Mud2-Bbp can tentatively be positioned into weak density directly downstream of the BPS and close to the U2/BPS helix, with Mud2 being bound to the 3′ end of the intron (adapted from Zhang et al., 2021).

1-5.2.3.3. Intron definition vs. exon definition and back-splicing

A discussion of the E complex structures is not complete without discussing the phenomena of ‘intron definition’ and ‘exon definition’, which are both features of the canonical splicing pathway. In ‘intron definition’, the spliceosome is assembled across the intron to form the IDC (intron definition complex), whereas the ‘exon definition’ pathway assembles the spliceosome across an exon to form the EDC (exon definition complex).

It is assumed that the EDC must be remodelled into an IDC in order to remove introns. Existence of the EDC has been largely circumstantial, and biochemical and structural analysis of

this pathway are limited, however the EDC appears to be similar to the IDC in composition (Schneider et al., 2010; Sharma, Kohlstaedt, Damianov, Rio, & Black, 2008). It is unknown whether the two complexes differ in structural organization or how the EDC remodels into an IDC. However, the E complex architecture suggests that the same E complex can form across both introns and exons, where the E complex either connects the BPS to an upstream 5′ SS (in the IDC) or a downstream 5′ SS (in the EDC). Similarly, the A complex architecture also suggests that the same A complex can span either an intron or an exon (Plaschka et al., 2018). However, exons below a certain length will allow the EDC to form across them as either an E or A complex but will not allow progression of the spliceosome cycle into the pre-B complex because integration of the U4/U6•U5 tri-snRNP requires a degree of conformational and steric flexibility that may be hindered by a short exon (Bai et al., 2018; Plaschka et al., 2018). This may serve as a signal to remodel the spliceosome from an EDC to an IDC. In this EDC → IDC remodelling pathway for short exons, the E, pre-A, A, and pre-B complexes can form across either introns or exons. However, an EDC spanning a short exon will be unstable at the pre-B state, triggering the EDC → IDC remodelling process. This model of the canonical pathway wherein EDC → IDC remodelling occurs at the pre-B stage for short exons is consistent with previous observations in mammalian systems (Schneider et al., 2010).

The use of both intron and exon definition in *S. cerevisiae* have been confirmed to occur (Li et al., 2019). Intron definition appears to dominate in yeast, which typically contain small introns and large exons (De Conti, Baralle, & Buratti, 2013). However, exon definition dominates in vertebrates, where small exons and large introns are prevalent (Berget, 1995). The dominance of either intron or exon definition in a species is likely determined by a combination of gene architecture and many other factors.

Finally, exon definition across exons above a certain length can lead to a phenomenon called “back-splicing”, which is a non-canonical splicing pathway in which the EDC is not remodelled, and the splicing cycle is completed due to a lack of steric hindrance. This results in the 5′ SS downstream of the exon being “back-spliced” to the 3′ SS upstream of the exon thereby generating a circRNA (exonic circular RNA) as a natural by-product of splicing in diverse eukaryotic species (including *S. cerevisiae*) and prompting speculation that just like intron and exon definition in the canonical splicing pathway, non-canonical back-splicing is also an ancient and conserved feature of eukaryotic gene expression (P. L. Wang et al., 2014). It is also possible for back-splicing to occur as a result of the EDC forming across multiple exons, and in this case the resulting circRNA will consist of those multiple exons and their intervening introns.

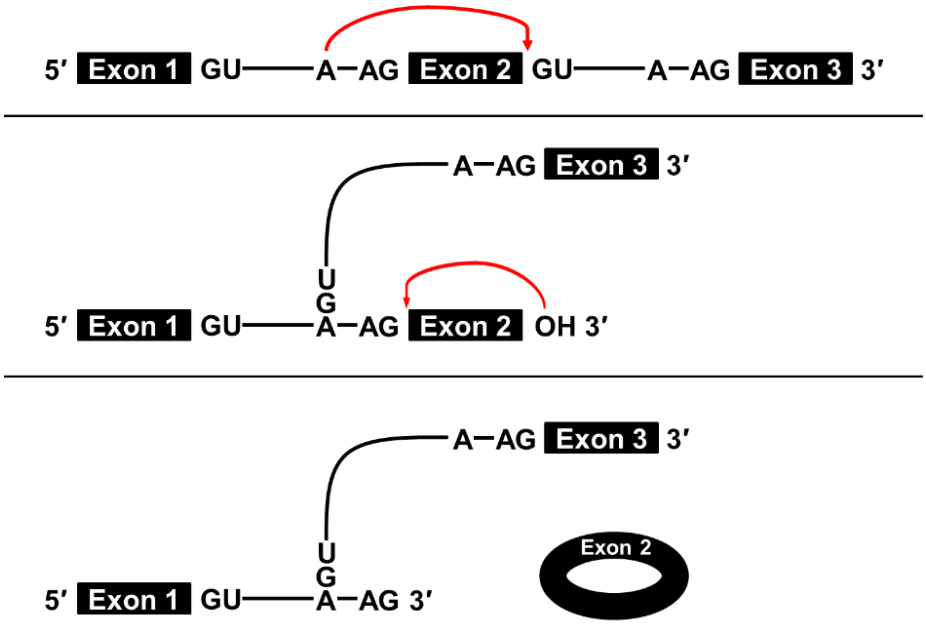


Figure 1-4. Back-splicing is a non-canonical splicing pathway that generates circRNA. Back-splicing occurs via the same two sequential transesterifications that define the canonical splicing pathway (see Fig. 1-1A for a comparison). However, in step one, the 2′ hydroxyl of the branch A carries out a nucleophilic displacement at a downstream 5′ SS instead of the upstream 5′ SS, generating a T-branch structure. In step two, the hydroxyl group at the 3′ end of the internal T-branch attacks the upstream 3′ SS, thereby yielding a pre-mRNA with an internal T-branch structure and a circRNA containing one or more exons.

CircRNAs are involved in the regulation of their host genes or microRNAs, ageing, and other disease processes (Wilusz, 2018). It is not precisely clear what signals an exon to be back-spliced into a circRNA rather than participate in canonical splicing, but the model above (which was first presented in the report of the E complex structures) in which RNA length is a key signal is supported by previous studies, and is also supported by the observation that the average exon length in circRNA is 690 nucleotides, which is much longer than the median length of 120 nucleotides for a human exon (Jeck et al., 2013; D. Liang & Wilusz, 2014; D. M. Liang et al., 2017; Mokry et al., 2010). Additionally, RNA elements (such as intronic complementary sequences flanking the exon) and RNA-binding proteins potentially promote back-splicing by

increasing the efficiency of EDC formation and proximating opposite ends of different exons, thereby enabling back-splicing across multiple exons (Wilusz, 2018).

To summarize, the architecture of the E complex structures as well as a number of experiments suggest that the same complexes are likely able to define both introns and exons in all eukaryotes and that the EDC complex either remodels to span an intron for canonical linear splicing (typically on short exons) or catalyses back-splicing to generate a circRNA (on long exons) (Li et al., 2019). Additionally, many *cis*-acting or *trans*-acting factors may act as modulators to promote or suppress a particular process. Examples include RNA, protein, nucleosomes, etc. For example, most vertebrate exons are short, which is likely the main signal for EDC remodelling; however, other factors may promote EDC → IDC remodelling on long exons and lower the efficiency of back-splicing. Although canonical splicing signals and the spliceosome are necessary for the production of circRNAs through back-splicing, the exact players and mechanisms are unknown (Starke et al., 2015).

One potential consequence of back-splicing that must be addressed is that it is an obvious route for generating splice variants of a gene in which one or more exons are excluded from the final transcript. Therefore, it is necessary to fully investigate and understand this process in order to understand its potential roles in regulated alternative splicing.

A summary has been provided here for intron definition, exon definition, and back-splicing in order to provide a more complete context for this thesis because all of these processes directly depend on definition of the 3′ SS by U2AF/SF1. The detailed biochemical and structural characterization of the roles of U2AF/SF1 in the spliceosome cycle is therefore essential to addressing gaps in our understanding of these processes.

1-6. The utility of *S. pombe* to model complex splicing phenomena in higher eukaryotes

S. pombe is an ancient yeast famously known for its role in mapping out the cell cycle and in 2002 became the 6th eukaryotic model organism to have its genome sequence and annotation published (Wood et al., 2002). Leland Hartwell and Paul Nurse were co-awarded the Nobel Prize in Physiology or Medicine (2001) for their pioneering work, which used both *S. pombe* and *S. cerevisiae* as model organisms to first identify and characterize the checkpoints of the cell cycle (Hartwell & Weinert, 1989). However, despite both being unicellular, ascomycetous yeasts with similarly sized genomes (*S. pombe* = 13.8 Mb, *S. cerevisiae* = 12.1 Mb), they are neither related nor syntenic. *S. pombe* is an ancient "basal" ascomycete (Taphrinomycetes), tracing back to the early radiative evolution of ascomycetes and perhaps close to the split between animals and fungi, making the evolutionary distance between *S. pombe* and *S. cerevisiae* of the same order as the distance between either of these yeasts and mammals (Egel, 2004; Heckman et al., 2001; Hoffman, Wood, & Fantes, 2015; Sipiczki, 2000). Since the divergence of *S. pombe* and *S. cerevisiae* ~350 million years ago, *S. pombe* has evolved more slowly than *S. cerevisiae*. Therefore, it retains more characteristics of the common ancient yeast ancestor and also shares more features with metazoan cells as well (Hoffman et al., 2015). For example, *S. cerevisiae* has lost many genes (338) that are conserved between *S. pombe* and mammals, making the proteomic content of *S. pombe* closer to that of the common yeast ancestor (Aravind, Watanabe, Lipman, & Koonin, 2000; Wood, 2006). An interesting and very useful consequence of the placement of *S. pombe* in the tree of life as a primitive yeast evolutionarily remote from *S. cerevisiae* is that when *S. pombe* shares a conserved gene with both humans and *S. cerevisiae*, the sequence conservation will generally be higher between *S. pombe* and humans

than between *S. pombe* and *S. cerevisiae*, and this also applies to splicing factors (Brennwald, Porter, & Wise, 1988; Egel, 2004).

The tractability of *S. cerevisiae* as a model organism has greatly advanced the splicing field, but it falls short as a tool to model the complex splicing phenomena and alternative splicing patterns present in higher eukaryotes that are crucial to understanding human biology. This is because *S. cerevisiae* has lost many of these features on its divergent evolutionary path. For the reasons summarized above however, *S. pombe* has retained at least a simple version of many of these phenomena. Because *S. pombe* combines the experimental tractability of *S. cerevisiae* with the complex splicing phenomena present in higher eukaryotes, it is an ideal intermediate model organism for the initial investigation, characterization, and modelling of these phenomena. At the incipient stage, these phenomena may be significantly more difficult or impossible to study using either the ultra-reductionist *S. cerevisiae* spliceosome (which is experimentally tractable but has lost these features) or the ultra-complex human spliceosome (which retains these features but presents significant technical challenges for experimentation and may be too complex to allow initial interpretation and modelling). *S. pombe* has a haploid genome of only three chromosomes and is much less likely than *S. cerevisiae* to have duplicated genes. However, it has significantly more introns (>4700) than *S. cerevisiae* with individual genes containing multiple introns (up to 15), some of which may be alternatively spliced (Forsburg, 1999; Hughes & Friedman, 2003; Okazaki & Niwa, 2000; Wood et al., 2002). Most importantly for this thesis, *S. pombe* retains the basic apparatus of higher eukaryotes responsible for initial recognition of the 3′ SS which has been lost in *S. cerevisiae*. This includes U2AF-L, U2AF-S, SF1, and the SF3B component p14. For these reasons, it has been chosen as the model system used to characterize these proteins in this thesis. Intron architecture and general splicing

features conserved between *S. pombe* and higher eukaryotes but lost in *S. cerevisiae* (with a few exceptions) are reviewed below in Section 1-6.1., and the apparatus for 3′ SS recognition conserved between *S. pombe* and higher eukaryotes but lost in *S. cerevisiae* is reviewed below in Section 1-6.2.

1-6.1. Intron architecture and general splicing features conserved between *S. pombe* and higher eukaryotes but lost in *S. cerevisiae*

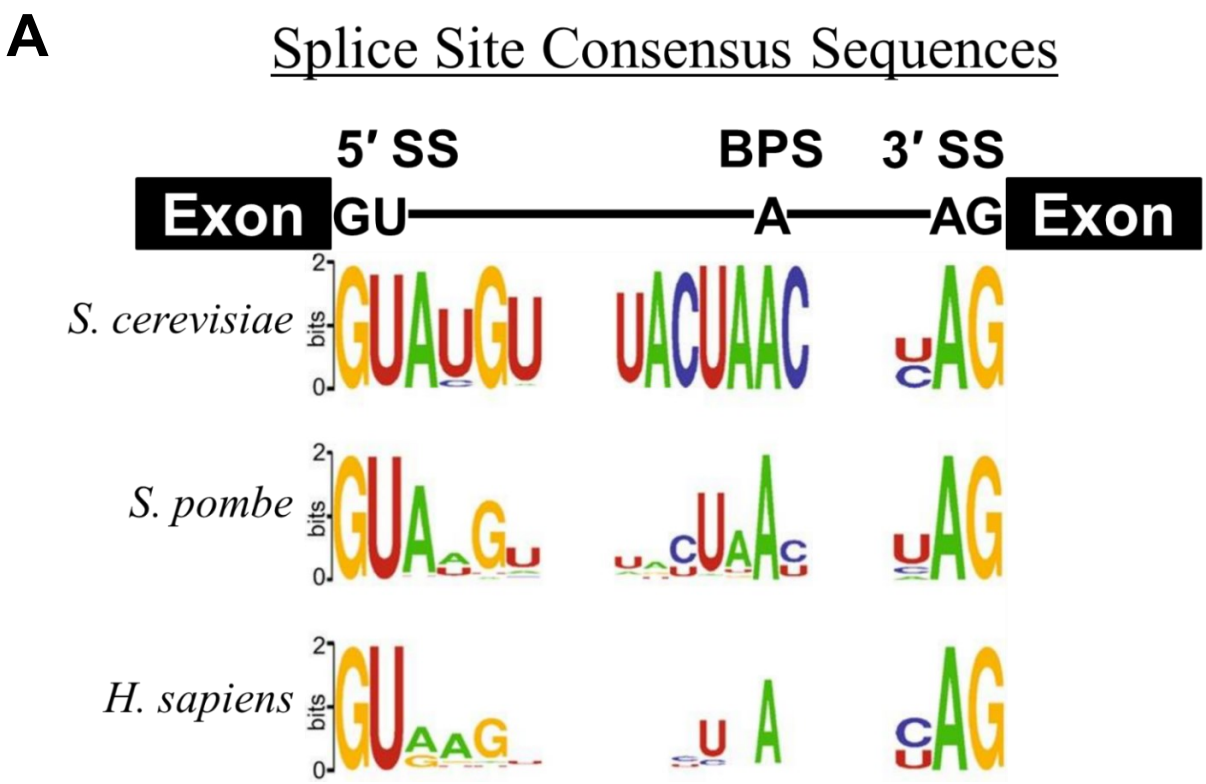
The gene architecture of *S. pombe* is closer to humans than *S. cerevisiae* is with respect to intron density. *S. pombe* genes are intron-dense (average of 0.9 introns per gene) compared to *S. cerevisiae* (average of 0.05 introns per gene), but not as intron-dense as human genes (average of 8 introns per gene). Additionally, many *S. pombe* genes contain two or more introns (a pre-requisite for exon-skipping), some of which are interrupted by extremely short microexons, similar to human genes (Scheckel & Darnell, 2015). The increased intron density of *S. pombe* genes relative to *S. cerevisiae* also reflects a higher degree of alternative splicing that more closely resembles the alternative splicing patterns of higher eukaryotes. As an example, *S. pombe* possesses orthologues of the metazoan-specific alternative splicing factors CACTIN and SDE2 (see Section 1-5.1.7.), which appear to have a role in regulating alternative splicing and are therefore functionally conserved with their human counterparts; in contrast, these factors have been lost in *S. cerevisiae* (Baldwin, Dinh, Hart, & Masson, 2013; Thakran et al., 2018).

It is estimated that ~2-3% of *S. pombe* splicing events involve exon-skipping, which is the dominant mechanism of alternative splicing predicted by exon definition (the dominant canonical splicing pathway in vertebrates) (Awan, Manfredo, & Pleiss, 2013; Bitton et al., 2015; Stepankiw, Raghavan, Fogarty, Grimson, & Pleiss, 2015). By comparison, <1% of splicing

events in *S. cerevisiae* involve alternative splicing, most of which reflect intron retention, and few reflect exon skipping (Juneau, Nislow, & Davis, 2009; Kawashima, Douglass, Gabunilas, Pellegrini, & Chanfreau, 2014; Marshall, Montealegre, Jimenez-Lopez, Lorenz, & van Hoof, 2013). The skipped exons in *S. pombe* are characterized by a relatively consensus upstream BPS, but weak downstream 5′ SS, consistent with observations in mammalian systems (Robberson, Cote, & Berget, 1990; Stepankiw et al., 2015). Some of these alternative splicing events are particularly sensitive to environmental cues which suggests that they are regulated.

Alternative splicing is more prevalent in *S. pombe* than in *S. cerevisiae* but pales in comparison to humans. This may be a result of the dominance of intron definition in *S. pombe*, just as in *S. cerevisiae* (Romfo, Alvarez, van Heeckeren, Webb, & Wise, 2000). The dominance of intron definition over exon definition creates a selection pressure for introns to remain short in order to maintain the initial pairing of splice sites across the intron. The gene architecture of *S. pombe* supports this since the natural distribution of intron lengths in both *S. pombe* and *S. cerevisiae* favors shorter introns than human genes (Fig. 1-5B) (Fair & Pleiss, 2017).

Importantly for this thesis, both the 5′ SS and 3′ SS are much more degenerate in *S. pombe* than in *S. cerevisiae*, more closely reflecting the degeneracy present in human splice sites; 5′ → 3′, the 3′ SS consists of the BPS, PPT, and AG di-nucleotide (see Section 1-1. for more detail). These sequence motifs in the 3′ SS are the binding targets for U2AF, SF1 and p14 and are responsible for directing the assembly of the spliceosome as well as controlling a number of alternative splicing events. Genome-wide patterns of intron architecture in *S. pombe*, *S. cerevisiae*, and humans with respect to splice site sequences and intron length distributions are summarized in Fig. 1-5. Interestingly, *S. cerevisiae* introns are characterized by a bimodal length distribution (Fig. 1-5B).



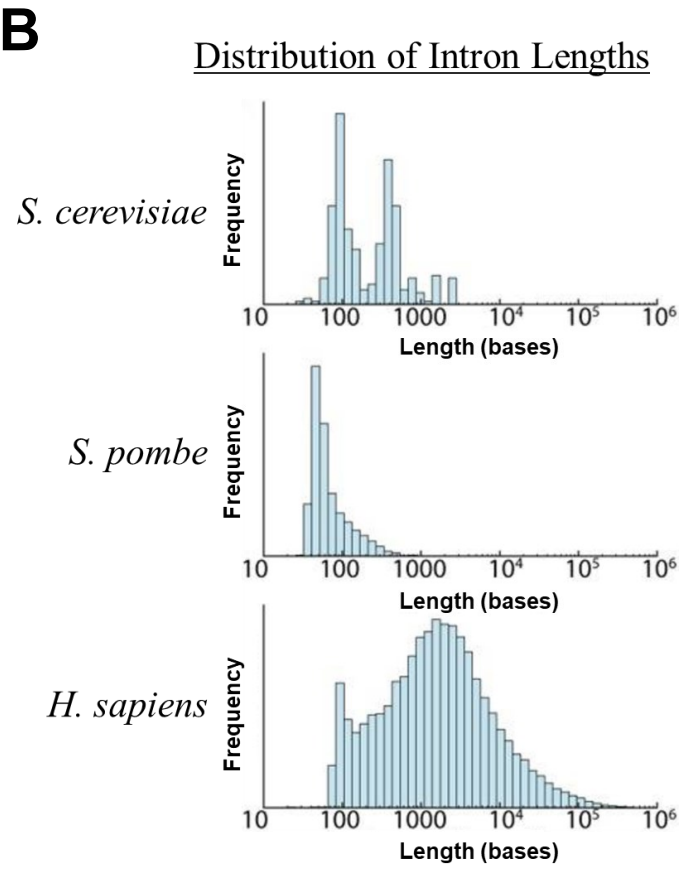


Figure 1-5. Comparison of intron architecture in *S. cerevisiae*, *S. pombe*, and humans. Intron lengths, 5′ SS sequences, BPS sequences, and 3′ SS sequences were obtained from the same source material as the original figure (Clark, Sugnet, & Ares, 2002; Gao et al., 2008; W. J. Kent et al., 2002; Stepankiw et al., 2015; Wilhelm et al., 2008). (A) Intron architecture of the consensus sequence at the 5′ SS, BPS, and 3′ SS depicted by sequence logos. The total height of each nucleotide at each position is proportional to conservation at that position (Crooks, Hon, Chandonia, & Brenner, 2004). (B) Histogram representation of length distribution of introns (adapted from Fair & Pleiss, 2017).

PPT architecture has been excluded from Fig. 1-5A for clarity due to the complexity of this sequence motif, specifically its variability in length and sequence composition. Mammalian introns typically possess a PPT between the BPS and AG di-nucleotide at the 3′ SS (Reed & Maniatis, 1985). However, the distribution of PPTs in both *S. pombe* and *S. cerevisiae* introns is quite different. A study was published characterizing the intron architecture of five diverse yeasts, including *S. pombe* and *S. cerevisiae* containing the results of a genome-wide screen of PPT sequences (using a minimal definition of six consecutive nucleotides with at least 3 uridines and no adenines) (Coolidge, Seely, & Patton, 1997; Drabenstot et al., 2003; Kupfer et al., 2004; Shelley & Baralle, 1987). The results of this screen reveal that the PPT distribution of *S. pombe* introns is unusual in that PPTs are typically located near the 5′ SS rather than within the 3′ SS. Most introns in both *S. pombe* (88.1%) and *S. cerevisiae* (93.7%) possess a PPT in the 5′ SS → BPS region, and most of these introns are located close to the 5′ SS. However, a surprisingly high percentage of *S. pombe* introns (62.1%) only have a PPT in the 5′ SS → BPS region, whereas in *S. cerevisiae*, 28.5% of introns only have a PPT in this region. In the BPS → AG di-nucleotide region where PPTs are expected based on metazoan introns, most *S. cerevisiae* introns (68.8%) contain a PPT, but only a minority of *S. pombe* introns (27.6%) contain a PPT.

The intron architecture of *S. pombe* is characterized by certain patterns that are closer to *S. cerevisiae* than humans. However, the overall intron architecture of *S. pombe* is a better representative of humans than *S. cerevisiae* is and this reflects the retention of splicing features in higher eukaryotes that have been lost in *S. cerevisiae*, including the apparatus for initial recognition of the 3′ SS.

1-6.2. Elements of the 3′ SS recognition apparatus conserved between *S. pombe* and higher eukaryotes but lost in *S. cerevisiae*

As discussed previously, *S. pombe* is a tractable model system for studying the complex splicing phenomena of higher eukaryotes that have been lost in *S. cerevisiae*. These phenomena include basic components of the 3′ SS recognition apparatus, such as SR proteins, U2AF-S, U2AF-L, and the SF3B component p14. The *S. pombe* counterparts of these proteins are discussed below.

S. pombe contains two orthologues of human SR proteins, which aid in the initial recognition and pairing of splice sites across exons by binding ESE sequences within the transcript; these proteins are thought to be master regulators of alternative splicing in plants and animals (see Section 1-4.1. for a more detailed summary) (Graveley, 2000; Kaufer & Potashkin, 2000; Kuhn & Kaufer, 2003; Long & Caceres, 2009). Interestingly, when mammalian ESE sequences are placed into *S. pombe* exons, they are able to recruit the *S. pombe* SR protein Srp2 (SR protein 2) to assist in the identification of weak upstream 3′ SSs, supporting the functional conservation of these proteins and their target sequences (Webb, Romfo, van Heeckeren, & Wise, 2005).

In addition to SR proteins, *S. pombe* also possesses highly conserved orthologues of both U2AF-L and U2AF-S, which are critical in the mammalian system but absent or non-essential in *S. cerevisiae* (Burge et al., 1999; Kaufer & Potashkin, 2000). Consistent with the presence of a U2AF-S orthologue, *S. pombe* contains AG-dependent introns, whereas *S. cerevisiae* does not (see Section 1-5.2.3.1. for more detail) (Romfo & Wise, 1997; Wentz-Hunter & Potashkin, 1996).

In order to fully understand the mechanisms of alternative splice site choice, it is necessary to study U2AF-S, U2AF-L, and SF1 because their binding activity is actively modulated by other auxiliary splicing proteins in order to selectively bind alternative 3′ SS sequence motifs (Shao et al., 2014). In humans, U2AF-S mediates protein-protein interactions involved in ESE-dependent splicing and specifically interacts with SR proteins (J. Y. Wu & Maniatis, 1993; Zuo & Maniatis, 1996). Similarly, the *S. pombe* orthologue of U2AF-S interacts with the SR protein Srp2 and mutations in the *S. pombe* U2AF/SF1 complex that cause exon skipping can be suppressed by overexpression of Srp2 consistent with a model where Srp2 contacts U2AF-S to recruit this complex to the 3′ SS (Haraguchi, Andoh, Frendewey, & Tani, 2007; Sasaki-Haraguchi et al., 2015; Webb & Wise, 2004).

In order to create a satisfactory framework to integrate findings on the U2AF/SF1 complex into the splicing cycle, it is necessary to fully characterize the SF3B complex, because it binds U2AF in order to recruit U2 snRNP to the BPS, and the SF3B component p14 directly contacts the branch A (reviewed in Section 1-4.2.). P14 is not present in *S. cerevisiae*, but *S. pombe* does contain an orthologue of p14 (PomBase systematic ID = SPBC29A3.07c), which is highly conserved with human p14 and has never been characterized prior to this thesis (Harris et al., 2022; Lock et al., 2019).

Several major advantages exist in using *S. pombe* to study the above-mentioned proteins over humans or other comparatively complex, multicellular eukaryotes. *S. pombe* only has two SR proteins whereas humans have twelve. Because they are not redundant in *S. pombe*, deep mutational scanning of SR proteins is possible in this system. Similarly, *S. pombe* only has one gene and isoform for U2AF-L, U2AF-S, and SF1, whereas humans and similarly complex,

multicellular eukaryotes have multiple paralogues and splice variants of these proteins (reviewed in Section 1-7.3.).

In the context of this thesis, the most important advantage in using *S. pombe* is that U2AF-S, U2AF-L, and SF1 are tightly complexed into a pre-assembled heterotrimer that is independent of RNA-binding or incorporation into the E complex, whereas in metazoans U2AF first binds the PPT and AG di-nucleotide, which then recruits SF1 to the BPS (T. Huang et al., 2002). This unique streamlining feature creates an opportunity to characterize the entire 3′ SS and the three proteins that recognize it as a single entity.

1-7. Overview of U2AF/SF1 in the context of splicing

The transition from the non-specific H complex to the E complex is an ATP-independent event that commits the intron to the splicing cycle which is demarcated by recognition and definition of the 5′ SS and 3′ SS by U1 snRNP and U2AF/SF1 respectively (Legrain et al., 1988; Michaud & Reed, 1991). The U2AF/SF1 complex subsequently recruits and is displaced by U2 snRNP to generate the A complex which is an ATP-dependent event (Crawford, Hoskins, Friedman, Gelles, & Moore, 2013).

Because U2AF/SF1 is directly responsible for defining the 3′ boundary of the intron and initiating the splicing cycle, it is a topic of intense interest. Anything that affects this apparatus has the potential for far reaching consequences by affecting gene expression through altered splice site choice which can subsequently affect any number of higher-order biological processes.

1-7.1. Architecture of the 3′ SS sequence

Compared to the 5′ SS which is defined by one sequence motif, the 3′ SS is more complex and degenerate and is defined by three sequence motifs which direct the early recognition and assembly events of the splicing cycle at the 3′ SS. In the 5′ → 3′ direction, these motifs are the BPS, PPT and 3′ AG di-nucleotide which demarcates the intron/exon boundary (Burge et al., 1999). These sequence motifs are essential to precisely define intron/exon boundaries in a controlled way and therefore are a critical and constitutive part of the splicing cycle. As previously discussed in Section 1-6., organism-specific differences of intron architecture exist that include the three motifs of the 3′ SS. However, the general construction of the 3′ SS is broadly conserved across eukaryotes.

The BPS consists of roughly half a dozen nucleotides containing the branch A which serves as the nucleophile in the 1st step of splicing and is required for lariat formation (Ruskin, Krainer, Maniatis, & Green, 1984). The human consensus BPS (yUnAy) is very degenerate, the *S. cerevisiae* BPS (UACUAAC) is invariant, and the *S. pombe* consensus BPS (CURAy) is intermediate between the two (Fig. 1-5A) (Gao et al., 2008; Kupfer et al., 2004; Spingola, Grate, Haussler, & Ares, 1999; M. Q. Zhang & Marr, 1994).

The PPT is downstream of the BPS and is highly variable in length and nucleotide composition though mammalian PPTs are typically 14-40 nucleotides long (Reed, 1989). It has been shown that PPTs with 11 consecutive uridines have the highest splicing efficiency, but splicing proceeds as long as the PPT is at least 9 nucleotides long and contains at least 5 consecutive uridines (Coolidge et al., 1997; Norton, 1994; Reed, 1989; Roscigno, Weiner, & Garciablanco, 1993). In *S. pombe* and humans, introns can be functionally classed as either AG-dependent or AG-independent, whereas all *S. cerevisiae* introns are AG-independent (see Section

1-5.2.3.1. and Section 1-6.2). In mammals, introns are distinguished by PPT length; AG-dependent introns have a short PPT and AG-independent introns have a long PPT (Moore, 2000; Reed, 1989).

The PPT has signaling roles in both constitutive and alternative splicing and is able to exert these effects despite being highly degenerate in both length and sequence composition (Green, 1991; McKeown, 1992; Nadalginard, Smith, Patton, & Breitbart, 1991). In metazoans, it is essential for correct intron recognition when the BPS is weak by recruiting splicing factors early during spliceosome assembly (Green, 1991). The PPT also exerts control over 3′ SS selection by promoting the use of alternative BPS sites (Mullen, Smith, Patton, & Nadalginard, 1991).

Downstream of the PTT, the 3′ SS contains an invariant AG di-nucleotide motif which demarcates the intron/exon boundary and, like the BPS, participates in splicing catalysis and is required for exon ligation (Reed, 1989). This AG di-nucleotide is nested inside the larger consensus sequence, yAG/G (slash = intron/exon boundary) (Moore, Query, & Sharp, 1993).

1-7.2. 3′ SS recognition by U2AF/SF1 in the context of the splicing cycle

The 5′ SS, BPS and 3′ AG di-nucleotide directly participate in the splicing reaction and are precisely defined in the E complex by being recognized and bound by a minimal set of specific proteins or, in the case of the 5′ SS, by being recognized and bound by U1 snRNP (Legrain et al., 1988). The sequence motifs that comprise the 3′ SS are recognized as follows: the BPS is bound by SF1, the PPT is bound by U2AF-L (U2AF65 in humans, U2AF59 in *S. pombe*), and the 3′ AG di-nucleotide is recognized by U2AF-S (U2AF35 in humans, U2AF23 in *S. pombe*) (Berglund, Abovich, et al., 1998; Berglund, Fleming, & Rosbash, 1998; T. Huang et al.,

2002; Jamison et al., 1992; Seraphin & Rosbash, 1989; Zamore et al., 1992; M. Zhang, Zamore, Carmo-Fonseca, Lamond, & Green, 1992).

Recruitment of the U2AF/SF1 complex to the 3′ SS is promoted by SR proteins, which are a significant component of the H complex, which contains members such as SC35 and SF2/ASF (splicing factor 2/ alternative splicing factor) (Reed, 1989; J. Y. Wu & Maniatis, 1993; Zamore et al., 1992). It should be noted that in addition to their regulatory role in the H complex, SR proteins are also non-snRNP associated components of the E complex (Berglund et al., 1997; Das, Zhou, & Reed, 2000; C. G. Lee, Zamore, Green, & Hurwitz, 1993; Z. Liu et al., 2001; Merendino et al., 1999; Michaud & Reed, 1991; Reed, 1989; Staknis & Reed, 1994; S. Wu et al., 1999; Zamore et al., 1992; M. Zhang et al., 1992; Zorio & Blumenthal, 1999a, 1999b). Additionally, U1 snRNP may play a role in recruiting U2AF to the PPT and AG di-nucleotide via SC35 which has been shown to bridge the 5′ SS and 3′ SS by bridging an interaction between U1 snRNP and U2AF35 (J. Y. Wu & Maniatis, 1993).

The E → A complex transition occurs when SF1 is displaced by U2 snRNP at the BPS and U2 snRNP becomes tightly associated with the BPS (Grabowski et al., 1985). This is an ATP-dependent event that is initiated by interactions between U2 snRNP and U2AF (see Section 1-4.2.).

1-7.3. U2AF/SF1 in the context of alternative splicing and higher-order biology

A complete understanding of 3′ SS recognition by U2AF/SF1 is essential to understanding the evolution and biology of developmentally complex, multicellular eukaryotes. This is because one major route for alternative splicing to occur is when the cellular machinery actively exploits splice site ambiguity at the 3′ SS in order to alter splice site choice (C. W. Smith

& Valcarcel, 2000). Because different protein isoforms generated from a single gene can have different properties, manipulation of 3′ SS choice significantly expands the proteome in a way that can be actively modulated by the cell in order to suit spatially and temporally dependent tissue and development specific needs (Ruhl et al., 2012; Staiger & Brown, 2013).

Multiple mechanisms exist to alter 3′ SS choice such as the existence of multiple paralogues and isoforms of U2AF-S, U2AF-L, and SF1 in higher eukaryotes with different spatio-temporal expression patterns and RNA-binding properties, as well as exon skipping by factors that block the binding of U2AF/SF1 to a splice site. All of these mechanisms can change the final spliced mRNA sequence and therefore the protein synthesized from a particular gene (Mollet, Barbosa-Morais, Andrade, & Carmo-Fonseca, 2006; Schellenberg, Ritchie, & MacMillan, 2008). In addition to multiple variants of U2AF-S, U2AF-L and SF1, degeneracy of the human BPS (Fig. 1-5A) is known to play a role in alternative BPS selection by SF1 (Corioni, Antih, Tanackovic, Zavolan, & Kramer, 2011; Mollet et al., 2006).

With respect to nomenclature, U2AF-S and U2AF-L orthologues and paralogues that participate in the U2AF/SF1 complex and therefore have a direct role in 3′ SS recognition are typically named according to their apparent molecular weight as assigned on an SDS-PAGE (sodium dodecyl sulfate-polyacrylamide gel electrophoresis) gel when they were first characterized.

1-7.3.1. The human U2AF-S gene family

Characterizing the U2AF-S gene family is particularly important because these proteins are directly responsible for determining the intron/exon boundary. U2AF35 is the prototype for the human U2AF-S gene family which includes the genes for U2AF35, U2AF26, U2AF35-RS1,

U2AF35-RS2/Urp (U2AF related protein) as well as additional uncharacterized ORFs (Barbosa-Morais, Carmo-Fonseca, & Aparicio, 2006; Mollet et al., 2006; Tupler, Perini, & Green, 2001).

U2AF35 was likely duplicated early in the emergence of vertebrates during the wave of whole-genome duplications 650-450 million years ago thereby creating U2AF26 (Mollet et al., 2006). On the other hand, retrotransposition of U2AF35-RS2 appears to have created U2AF35-RS1 less than 100 million years ago. Interestingly, it appears that retrotransposition of U2AF35-RS2 to create U2AF35-RS1 occurred independently in the lineages of rodents and primates after they diverged and this is supported by both the differing genomic location of U2AF35-RS1 in these two lineages, as well as the fact that the murine counterpart is imprinted whereas the human counterpart is not (Barbosa-Morais et al., 2006; Hayashizaki et al., 1994; Mollet et al., 2006; Nabetani, Hatada, Morisaki, Oshimura, & Mukai, 1997; Pearsall et al., 1996).

In vertebrates (including humans), the U2AF35 ORF has undergone an exon duplication event. The two copies of the duplicated exon are mutually exclusive and give rise to two unique functional variants: U2AF35a and U2AF35b. The high conservation of U2AF35a and U2AF35b in vertebrates from fish to humans indicates a strong selective pressure to retain these isoforms due to lineage-specific functions. U2AF35a is 9-18-fold more abundant than U2AF35b with distinct tissue-specific patterns of expression in mice varying from ~10-fold in the brain to ~20-fold in lung and skeletal muscle. U2AF35a also appears to be more efficient than U2AF35b at splicing pre-mRNA. In addition to the existence of two isoforms, multiple classes of U2AF35 mRNAs exist that vary in their polyadenylation signals as well as a conserved transcript with a PTC called U2AF35c that is targeted for the NMD pathway, suggesting that an additional layer of post-transcriptional regulation exists for modulating the expression levels and isoform ratios

of U2AF35 possibly contributing to a more finely tuned control of splicing events in different tissues (Pacheco et al., 2004).

In contrast to U2AF35, U2AF26 appears to serve a narrower and more specialized set of lineage-specific functions. U2AF26 is a vertebrate-specific paralogue found in rat, pig and cow, but there is no evidence of U2AF26 in the genomes of birds, amphibians or fish (Mollet et al., 2006). Most experimental characterization of U2AF26 until now has used murine models. U2AF26 can substitute for U2AF35 *in vitro*, however it appears to fulfill a non-redundant alternative splicing function *in vivo*, and U2AF26 is enriched in brain tissue, which is highly differentiated and experiences unusually high levels of alternative splicing (Grabowski & Black, 2001; Heyd, ten Dam, & Moroy, 2006; Shepard, Reick, Olson, & Graveley, 2002). Interestingly, U2AF35 binds with higher affinity to the AG/G intron/exon boundary sequence, whereas U2AF26 prefers either AG/C or AG/A (Heyd, Carmo-Fonseca, & Moroy, 2008; Shepard, 2004). Both the AG/C and AG/A intron/exon boundaries are enriched in alternatively spliced exons in tissues such as brain and muscle suggesting that U2AF26 is a regulator of tissue-specific alternative splicing (Stamm et al., 2000). Specifically, the ratio of U2AF26 vs. U2AF35 mRNA varies from ~3 in the brain to ~0.5 in the liver and U2AF26 pre-mRNA is alternatively spliced, generating at least three mRNA isoforms: full-length U2AF26, an isoform excluding exons 6 and 7 (ΔE67), and an isoform excluding exon 7 (ΔE7) (Heyd et al., 2008; Preussner et al., 2014; Shepard et al., 2002).

With respect to the specific alternative splicing mechanisms controlled by U2AF26, this protein is a component of a circadian and light-inducible splicing switch in the peripheral circadian clock. The circadian clock consists of two components. The central clock, which in mammals resides in the SCN (suprachiasmatic nucleus) of the brain, receives light cues from the

retina to synchronize peripheral clocks which reside in nearly every tissue and organ system tested and play an integral and unique role in each of their respective tissues by driving the circadian expression of specific genes involved in a variety of physiological functions (Albrecht, 2004; Dibner, Schibler, & Albrecht, 2010; Richards & Gumz, 2012). Specifically, the ΔE67 mRNA isoform alters the reading frame so that translation continues far into the 3′ UTR generating a protein isoform with a C terminus with homology to the *Drosophila melanogaster* clock regulator TIMELESS, hereafter referred to as a THD (TIMELESS homology domain). U2AF26-deficient mice display broad defects in circadian mRNA expression in peripheral clocks and increased phase advance adaptation following experimental jet lag indicating that light-induced U2AF26 alternative splicing stabilizes the circadian clock against abnormal changes in light/dark conditions (Preussner et al., 2014).

Total mRNA levels of U2AF26 remain constant in both the cerebellum (a brain region with well-established circadian oscillation of gene expression) and liver, however the ratio of full-length and ΔE67 mRNA oscillates in a light-inducible manner as part of the circadian clock. The ΔE67 mRNA levels vary ~5-fold in the cerebellum whereas the liver displays less pronounced cycling. However, ΔE67 is weakly expressed in the central clock and is not subject to circadian regulation suggesting that U2AF26 alternative splicing is not directly light sensitive in the SCN and that signals from the SCN control this splicing switch in peripheral clock(s). Also, the ΔE67 protein isoform has a half-life well below 3 hr whereas the full-length protein has a half-life of over 24 hr. The short half-life of the ΔE67 protein isoform enables circadian expression which requires fast protein turnover (Preussner et al., 2014).

This clock mechanism appears to be conserved. As with mice, rat U2AF26 contains a THD in the 3′ UTR in an alternative reading frame which is accessible when exons 6 and 7 are

excluded and rat U2AF26 alternative splicing is regulated in the cerebellum similarly to mouse. The human U2AF26 gene also contains a THD at the 3′ end of the ORF which is accessible through alternative splicing. However, in contrast to mouse and rat, use of this frame requires inclusion of exons 6 and 7 and the use of an alternative 3′ SS in exon 8. RT-PCR (reverse transcription-PCR) of RNA from human U2OS cells (a bone osteosarcoma epithelial cell line commonly used as a circadian rhythm model) confirms the existence of a U2AF26 mRNA isoform corresponding to this splicing pattern, suggesting that human peripheral clock(s) are regulated by a U2AF26 protein isoform containing a THD (Preussner et al., 2014).

In addition to a different RNA-binding sequence specificity than U2AF35 and its roles in activating circadian gene expression cascades by coupling alternative splicing with the circadian clock, U2AF26 also appears to regulate splicing decisions through its subcellular localization, which differs from U2AF35. The U2AF26 NLS (nuclear localization signal) differs from any known NLS and is encoded by exons 7 and 8 in the murine orthologue. This is consistent with the observation that like U2AF35, full-length U2AF26 is nuclear and undergoes nucleo-cytoplasmic shutting, whereas the ΔE7 and ΔE67 protein isoforms of U2AF26 are exclusively cytoplasmic (Heyd et al., 2008; Preussner et al., 2014). U2AF35 and U2AF26 differ primarily at the C-terminus. This region of U2AF35 consists of an RS domain (see Section 1-10.2.) and contains the NLS, whereas U2AF26 does not contain an RS domain, and translocation of full-length U2AF26 into the nucleus requires interaction between its NLS and p32 (32 kDa protein), which is involved in many functions including regulating alternative splicing by binding the SR protein SF2/ASF and changing its activity (Krainer, Mayeda, Kozak, & Binns, 1991; Petersen-Mahrt et al., 1999). Consistent with its many functions, p32 is localized to many subcellular compartments such as mitochondria, the outer cell membrane, and nucleus and it is conceivable

that it regulates the intracellular distribution of its binding partners including U2AF26 (Brokstad, Kalland, Russell, & Matthews, 2001). The exclusive cytoplasmic presence of the ΔE67 isoform of U2AF26 underlines the importance of protein localization for the molecular clockwork (Hirano et al., 2013; Yoo et al., 2013).

Finally, both U2AF35 and U2AF26 appear to regulate alternative splicing non-canonically at the translational level. The ΔE7 isoform of U2AF26 is strongly induced during activation of primary mouse T cells and appears to regulate cytoplasmic gene expression by binding the 5′ UTR of target mRNAs influencing translation as well as altering the abundance of many cytoplasmic mRNAs, which may indicate a role in controlling mRNA stability (Herdt et al., 2020; Martinez et al., 2012; Schultz, Preussner, Bunse, Karni, & Heyd, 2017). Cytoplasmic full-length U2AF35 also appears to affect translation (Herdt et al., 2020; Palangat et al., 2019).

Neither U2AF35-RS1 nor U2AF35-RS2/Urp are as thoroughly characterized as U2AF35 or U2AF26. However, murine U2AF35-RS1 shows tissue-specific expression and is predominantly expressed in the brain, especially the pyramidal neurons of the hippocampus and dental gyrus, and U2AF35-RS2/Urp is functionally distinct from U2AF35 because U2AF35 cannot complement Urp-depleted extracts (Hatada et al., 1995; Hatada, Sugama, & Mukai, 1993; Tronchere, Wang, & Fu, 1997).

Biochemical evidence indicates that U2AF35, U2AF26, and U2AF35-RS2/Urp all interact with U2AF65 (Pacheco et al., 2004; Shepard et al., 2002; Tronchere et al., 1997). In the future it will be important to study the distinct functional activities of the U2AF-S gene family in the context of the U2AF dimer and U2AF/SF1 trimer.

1-7.3.2. The human U2AF-L gene family

U2AF65 is the prototype for the human U2AF-L gene family and has two protein isoforms generated by alternative 5′ SS selection in the pre-mRNA transcript with one isoform being 3 residues shorter than the other. These isoforms are otherwise identical and biochemically indistinguishable (F. Ding, Hagan, Wang, & Grabowski, 1996). Additionally, the human U2AF-L gene family includes the paralogous genes for U2AF65, PUF60 (poly(U) binding splicing factor 60 kDa), CAPERα (coactivator of activating protein-1 and estrogen receptors α), and CAPERβ. With the exception of U2AF65, members of this gene family are splicing regulators but do not operate as a part of the U2AF/SF1 apparatus for initial 3′ SS recognition in the spliceosome cycle. Although it is unknown whether U2AF65 performs functions outside of splicing, the other human U2AF-L family members have clear roles in both splicing and transcription.

CAPERα and CAPERβ are the most recently characterized proteins related to U2AF65, and CAPERβ likely arose from a gene duplication of CAPERα during the same wave of whole-genome duplications that created U2AF26 (Dowhan et al., 2005; Mollet et al., 2006; Strausberg et al., 2002). Northern blotting has revealed multiple transcripts for both CAPERα and CAPERβ; there are at least four CAPERα mRNA transcripts and two CAPERβ transcripts. These transcripts vary in abundance across several human tissues.

CAPERα was first isolated as a novel autoantigen from a patient with liver cirrhosis who progressed to hepatocarcinoma (Imai, Chan, Kiyosawa, Fu, & Tan, 1993). Both CAPERα and CAPERβ regulate transcription and alternative splicing in a steroid hormone-dependent manner and the splicing and transcription functions are located in distinct and separable domains of the protein (Amara, Jonas, Rosenfeld, Ong, & Evans, 1982; Auboeuf et al., 2004; Dowhan et al.,

2005; Hartmuth et al., 2002; Jung, Na, Na, & Lee, 2002; Rappsilber et al., 2002; Wellmann et al., 2001). Accordingly, both CAPERα and CAPERβ expression levels are higher in the placenta and liver which both possess active steroid hormone signaling.

One possible model for the functional coupling of transcription and alternative splicing through CAPER proteins is that these proteins first interact with promoter-bound transcription factors to stimulate transcription in response to steroid hormones allowing their incorporation into the preinitiation complex which then enables direct access to the nascent RNA transcript. CAPER proteins may then interact with splicing factors required for early recognition of the 3′ SS, thereby influencing the commitment to splicing (Dowhan et al., 2005).

Human PUF60 was first isolated as a PPT-binding protein closely related to U2AF65 that was required for efficient *in vitro* RNA splicing. Around the same time, it was also identified as a modulator of TFIIH (transcription factor IIH) activity (J. Liu et al., 2000; Page-McCaw, Amonlirdviman, & Sharp, 1999). Unlike CAPERα/β, PUF60 (similar to U2AF65) is expressed in most tissues, as expected for a constitutive splicing factor (Dowhan et al., 2005). However, the *D. melanogaster* orthologue of PUF60 has roles in both constitutive and alternative splicing *in vivo* raising the possibility that human PUF60 also regulates alternative splicing (Van Buskirk & Schupbach, 2002). It is also unknown whether the transcription and splicing functions of PUF60 are coupled as they are with CAPERα/β.

1-7.3.3. The human SF1 gene family

The human SF1 gene family contains SF1 as well as four additional paralogues: QK1 (quaking homologue 1), Sam68 (Src associated in mitosis of 68 kDa)/KHDRBS1 (KH domain containing, RNA binding, signal transduction associated 1), KHDRBS2, and KHDRBS3. The

paralogues of SF1 bind pre-mRNA and have roles in splicing but, as with U2AF65-related genes, the paralogues of SF1 do not operate as a part of the U2AF/SF1 apparatus for initial 3′ SS recognition in the spliceosome cycle (Stelzer et al., 2016).

A total of twelve predicted protein isoforms have been identified and reported in the NCBI (National Center for Biotechnology Information) database for human SF1 ranging from 433-764 amino acids in total length based on mRNA transcripts detected. The RNA-binding core of these predicted protein isoforms is identical, and the differences are restricted to the region N-terminal to the ULM and the proline-rich domain C-terminal to the zinc knuckle(s) of the protein (see Section 1-8.1. for the conserved domain structure of SF1). The function of these two terminal regions is uncharacterized, and they differ significantly in length and sequence between the isoforms. Alternatively spliced protein isoforms of human SF1 are expressed in a variety of mammalian cell types and show cell type-specific expression. The C-terminal region of SF1 outside of the RNA-binding core is dispensable for spliceosome assembly *in vitro* and the corresponding C-terminal region of Bbp outside of the RNA-binding core is dispensable for viability in *S. cerevisiae* (Arning, Gruter, Bilbe, & Kramer, 1996; Caslini et al., 1997; Guth & Valcarcel, 2000; Kramer, Quentin, & Mulhauser, 1998; Rain, Rafi, Rhani, Legrain, & Kramer, 1998; Toda, Iida, Miwa, Nakamura, & Imai, 1994; Wrehlke, Schmitt-Wrede, Qiao, & Wunderlich, 1997). Together, these data suggest that the C-terminus of SF1 may be required for other functions *in vivo*. Interestingly, three of the twelve predicted protein isoforms of SF1 lack some or all of the N-terminal region needed to interface with U2AF65; isoform # 5 (GenBank accession number: NP_001171502.1) is missing the ULM that interfaces with the UHM of U2AF65, and both isoform # 7 (GenBank accession number: NP_001333338.1) and isoform # 8 (GenBank accession number: NP_001333339.1) are missing the ULM as well as the

phosphorylated SPSP domain (see Sections 1-8.1. and 1-8.2. for the conserved domain structure of SF1 and U2AF-L, respectively). Therefore, SF1 may downregulate 3′ SS recognition in certain contexts and promote alternative 3′ SS selection by generating protein isoforms that can bind the BPS but are unable to participate in the U2AF/SF1 complex thereby competing with functional protein isoforms of SF1.

The paralogues of human SF1 bind pre-mRNA and have many functions, some of which include important roles in pre-mRNA splicing, transport, and stability (Stelzer et al., 2016). Most importantly, they have roles in alternative splicing and one route by which they exert their effects on splice site selection is by binding the BPS and blocking SF1 in an isoform, target sequence, and cell type-specific manner (X. Chen et al., 2021; Farini et al., 2020; Meyer et al., 2010; Song & Richard, 2015; Tisserant & Konig, 2008; J. Z. Wang et al., 2021; Zong et al., 2014). The existence of competitive binding of the BPS by SF1 paralogues in order to promote alternative splicing parallels the potential for downregulatory activity by non-functional SF1 isoforms that cannot interact with U2AF65. Both of these phenomena further point to a functional significance of a highly degenerate BPS in higher eukaryotes such as humans.

1-7.3.4. The U2AF/SF1 gene families in the context of the evolution of higher-order biological organization

Phylogenetic analysis reveals that the origin of U2AF gene families dates back to the divergence of eukaryotes more than 1.5 billion years ago, and several proteins similar to both U2AF35 and U2AF65 have been identified in humans that have arisen through both gene duplication as well as well as retrotransposition (Barbosa-Morais et al., 2006; Mollet et al., 2006).

Very few examples of U2AF mRNA isoforms have been described in the literature. However, bioinformatics analyses reveal that, with the single exception of the U2AF35-RS1 (which is devoid of introns), all genes in the U2AF-S and U2AF-L gene families can be alternatively spliced (Mollet et al., 2006). In addition to mRNAs that are either expected to or have been demonstrated to generate functional protein, many of the predicted alternatively spliced mRNA isoforms contain PTCs and are expected to be targeted for degradation by the NMD pathway. This has already been demonstrated to occur for the U2AF35c mRNA isoform, and the conservation of this transcript indicates selective pressure due to functional roles and potential autoregulation of U2AF35 through RUST.

All aspects of genetic regulation through alternative splicing that have been discussed in Section 1-2. are represented by the U2AF/SF1 gene families. Due to its potential to effect large changes in the proteome and transcriptome however, the most important of these is arguably the existence of regulated cascades of gene expression triggered by U2AF/SF1 family members. Importantly, U2AF/SF1 family members activate these cascades through multiple means, which include non-canonical mechanisms that are unrelated to their splicing functions. One example is the potential translational functions of cytoplasmic U2AF26 and U2AF35 and their roles in mRNA stability. Several examples of splicing factors exist that localize to the cytoplasm upon alternative splicing and have functions in cytoplasmic mRNA processing and the best described are SR proteins; cytoplasmic SR proteins regulate mRNA stability and translation (Twyffels, Gueydan, & Kruys, 2011).

Alternative splicing cascades activated by U2AF/SF1 variants are not always a direct result of splicing and can emerge from the coupling of splicing with other biological functions. This opens up new and unexpected mechanisms to activate splicing cascades as well as fine-tune

their activation in response to external stimuli. One example is the THD in murine and human U2AF26 controlling peripheral clocks which is not canonically involved in splicing. Additionally, genomic information for this splicing switch is hidden in a supposedly untranslated part of the mRNA and only accessible by altering the reading frame. It has been suggested that the expression of several other splicing-regulatory proteins is regulated in a circadian manner, and proteins such as PSF (PTB associated splicing factor) and NONO (non-POU domain containing octamer binding) that regulate alternative splicing in different contexts are known to be involved in clock regulation (Duong, Robles, Knutti, & Weitz, 2011; Heyd & Lynch, 2010; Kowalska et al., 2013; McGlincy et al., 2012).

Another important example of the coupling of splicing with other biological functions is CAPERα/β in which steroid hormone-dependent transcription is coupled to alternative splicing and increasing evidence indicates that gene expression is modulated through the functional coupling of transcription and pre-mRNA processing (Dowhan et al., 2005; Han et al., 2017; Loerch, Maucuer, Manceau, Green, & Kielkopf, 2014; Mollet et al., 2006; Tari et al., 2019; Uehara et al., 2017).

The diversification of U2AF/SF1 gene families likely evolved in response to a requirement for the co-ordination of the multiple steps of gene expression in complex organisms. According to this view, as mRNA biogenesis became progressively more targeted for regulation, new gene, mRNA and protein variants emerged to diversify the functions of these proteins, and new sequence characteristics developed to couple multiple biological processes via the same protein. Currently, certain genes within the human U2AF/SF1 gene families remain uncharacterized or poorly characterized. Closing this gap will enable the search for co-regulated genes, thereby allowing the identification and characterization of discrete alternative splicing

events and larger alternative splicing cascades controlled by regulated U2AF/SF1 activity, as well as non-canonical mechanisms of regulating gene expression controlled by U2AF/SF1.

A complete understanding of the layers of genetic regulation that organize complex eukaryotic life requires a thorough characterization of the entire U2AF/SF1 family. These proteins have critical roles in alternative splicing which multiplies the genome's coding capacity and has a vast, yet largely unexplored regulatory potential (Irimia & Blencowe, 2012). In addition to this, the U2AF/SF1 family regulates gene expression through non-splicing mechanisms and the coupling of splicing with other biological functions. Finally, the examples above reveal that functionally important features of U2AF/SF1 can be hidden in numerous, often unexpected ways. This is the case with the THD in a supposedly untranslated region of U2AF26 which is only accessible through a splicing switch. For this reason, the detection of additional layers of genetic regulation controlled by these genes requires investigations into gene structure, including patterns of conservation in gene structure as U2AF/SF1 genes diverged, as this suggests selection pressure to retain features that would otherwise be overlooked and have currently unknown functional roles.

1-7.4. U2AF/SF1 in the context of human disease

Because all human protein coding RNAs undergo splicing before a protein is synthesized, it is not surprising that ~50% of human genetic diseases (including diverse cancers) are related to aberrant splicing, a number of which involve errors in 3′ SS recognition (Stenson et al., 2014). U2AF and SF1 belong to a splicing pathway where the mutation of many components is associated with disease phenotypes ranging from developmental abnormalities, degenerative eye disease, and multiple cancers (Bernier et al., 2012; W. Q. Ding, Kuntz, & Miller, 2002; McKie et

al., 2001; L. Wang et al., 2011; K. Yoshida et al., 2011). For example, it has been shown that a loss of U2AF35 activity promotes abnormal splicing in pancreatic cancer (W. Q. Ding et al., 2002). Additionally, U2AF activity is critical in the life cycle of pathogenic viruses, including HIV (human immunodeficiency virus), as they hijack both U2AF-S and U2AF-L in order to execute certain life cycle tasks in order to replicate (Domsic, Wang, Mayeda, Krainer, & Stoltzfus, 2003; Gama-Carvalho et al., 1997; Lutzelberger, Backstrom, & Akusjarvi, 2005). The translation of biochemical/structural understanding of these proteins to clinical practice is also a worthwhile and practical goal. For example, purine interruptions of PPTs at the 3′ SS cause many inherited diseases and a structure-guided mutant of U2AF65 can restore wildtype splicing to these disease-related splicing signals in tissue culture (Agrawal, McLaughlin, Jenkins, & Kielkopf, 2014).

One particularly important and well-characterized hotspot for disease-associated mutations is S34 of human U2AF35; S34F/Y mutations result in cancers and myelodysplastic syndromes by inducing aberrant splicing (Imielinski et al., 2012). Usually, the AG at the human 3′ SS is preceded by a pyrimidine and SELEX (systematic evolution of ligands by exponential enrichment) confirms that U2AF35 prefers a pyrimidine at this position (Sheth et al., 2006; S. Wu et al., 1999). However, whole-exosome sequence analysis revealed that A and C are found much more frequently at this position in S34F/Y-induced hematological malignancies; by changing the preferred 3′ SS sequence, S34F/Y mutations enhance aberrant exon inclusion thereby inducing these disorders (Ilagan et al., 2015; Kim et al., 2018; Okeyo-Owuor et al., 2015). Similarly, S34F causes aberrant alternative splicing by preferentially binding CAG in lung adenocarcinomas (Esfahani et al., 2019; Fei et al., 2016).

1-8. Overview of U2AF/SF1 architecture and organization

A large number of studies have biochemically characterized the U2AF/SF1 particle and its sub-components using various techniques and model systems. However, many ambiguities and open-ended questions remain. This is in large part due to the degeneracy of the 3′ SS because it is unclear how these proteins can accommodate such a wide variation in the length and information content of the 3′ SS. These can be resolved through high-resolution structural data, but current X-ray and solution NMR (nuclear magnetic resonance) structures are minimalist and only represent 1-3 domains ± a short oligonucleotide.

The domain organization of the U2AF/SF1 complex is discussed below in Sections 1-8.1. to 1-8.4. Section 1-9. consists of a tabular summary of all X-ray and solution NMR structures reported in the PDB to date, as well as a discussion of the most important and recent of these structures (redundant and obsolete models are not discussed for clarity and simplicity).

1-8.1. SF1 recognizes the BPS

The BPS is a conserved motif containing the branch A involved in the first transesterification of intron removal. It is defined by the sequence yUnAy in metazoans, with UACUAAC being the optimal sequence. This is recognized by SF1 (Gao et al., 2008; Pastuszak et al., 2011).

The domain structure comprising the conserved RNA-binding core of SF1 is summarized in Sections 1-8.1.1. to 1-8.1.4. below and omits the N-terminus and C-terminus of SF1 because they are poorly conserved in length and sequence composition across species and across alternatively spliced isoforms of human SF1, uncharacterized, and are not known to have well-defined roles in 3′ SS recognition. From N-terminus to C-terminus, the conserved RNA-binding

core of SF1 consists of a ULM (U2AF ligand motif), phosphorylated domain, KH-QUA2 (KH-quaking) domain, and one or two zinc knuckles (depending on the species).

The region N-terminal to the ULM of SF1 is predicted to be unstructured. In certain species such as *D. melanogaster* and *C. elegans*, this region can be classified as an RS domain (Mazroui, Puoti, & Kramer, 1999). In contrast to SR domains (see Section 1-4.1.), which are located at the C-terminus of their host protein and are characterized by repeating serine-arginine di-peptides, the RS domain is a similar domain found in a number of splicing proteins but is characterized by repeating RS (arginine-serine) di-peptides and is typically found at the N-terminus of a protein. As with the SR domain, the RS domain is presumed to be intrinsically unstructured, and the repeating di-peptide motif confers a net positive charge to this domain.

C-terminal to the zinc knuckle(s), a conserved α-helix is predicted to exist in both human and *S. pombe* SF1 according to the output generated by the PsiPred (PSI-BLAST based secondary structure prediction) server (see Section 1-10.4. and Fig. 1-20) (Buchan, Minneci, Nugent, Bryson, & Jones, 2013). C-terminal to the predicted α-helix, SF1 consists of a proline-rich domain of unknown function. However, the cell type-specific expression patterns of human SF1 isoforms (which vary only in the N- and C-termini), combined with the observation that the C-terminus of SF1 is dispensable for spliceosome assembly *in vitro*, and that the C-terminus of Bbp is dispensable for viability in *S. cerevisiae*, suggest that the proline-rich C-terminus of SF1 may be required for functions other than splicing *in vivo* (see Section 1-7.3.3.).

1-8.1.1. SF1 domain structure (N-terminus to C-terminus): ULM

The ULM of SF1 is a short unstructured coil that participates in a protein-protein dimerization interface with the UHM (U2AF homology motif) of U2AF-L and becomes structured upon interfacing with its binding partner (Selenko et al., 2003).

1-8.1.2. SF1 domain structure (N-terminus to C-terminus): Phosphorylated domain

This domain serves a regulatory role that is mediated through phosphorylation. It contains a conserved SPSP motif that is phosphorylated on the two serines; phosphorylation induces subtle conformational changes that transduce into much larger global conformational changes in the U2AF/SF1 complex by kinking and rigidifying it. These changes are responsible for modulating the interaction of the U2AF/SF1 complex with the pre-mRNA substrate. Additionally, upon phosphorylation, this domain interfaces with U2AF-L, thereby extending the interfacial region that exists between the UHM of U2AF-L and the ULM of SF1 (W. Wang et al., 2013; Y. Zhang et al., 2013).

1-8.1.3. SF1 domain structure (N-terminus to C-terminus): KH-QUA2 domain

This domain directly contacts the BPS (Z. Liu et al., 2001).

1-8.1.4. SF1 domain structure (N-terminus to C-terminus): Zinc knuckles

SF1 contains 1-2 zinc knuckles, depending on the species (1 in each of the twelve isoforms of human SF1, 2 in *S. pombe* SF1 and *S. cerevisiae* Bbp) (Garrey, Voelker, & Berglund, 2006). One of the zinc knuckles is conserved. The zinc knuckles are not well-characterized. However, in *S. cerevisiae*, the N-terminal zinc knuckle proximal to the KH-QUA2

domain enhances binding affinity to the BPS, though the role of the C-terminal zinc knuckle is unknown (Garrey et al., 2006; Rain et al., 1998). Both zinc knuckles conform to the consensus sequence cys-x2-cys-x4-his-x4-cys, a characteristic motif in retroviral nucleocapsid proteins (Darlix, Lapadattapolsky, Derocquigny, & Roques, 1995; Rain et al., 1998).

1-8.2. U2AF-L recognizes the PPT

The PPT is a stretch of sequence 10-40 nucleotides long (in metazoans) dense in pyrimidines, particularly uridines; it is recognized by U2AF-L (Reed, 1989).

The domain structure comprising the conserved RNA-binding core of U2AF-L is summarized in Sections 1-8.2.1. to 1-8.2.3. below and omits the N-terminal RS domain of U2AF-L because it is poorly conserved in length and sequence composition across species, poorly characterized, and is not known to have a well-defined role in 3′ SS recognition. From N-terminus to C-terminus, the conserved RNA-binding core of U2AF-L consists of a ULM, two RRMs (RNA recognition motifs), and UHM. Hereafter, the N-terminal RRM will be referred to as RRM1, and the C-terminal RRM will be referred to as RRM2.

The role of the RS domain has been difficult to define, however previous work published by our laboratory indicates that the RS domain is associated with the BPS (O. A. Kent, Reayi, Foong, Chilibeck, & MacMillan, 2003). Additionally, reconstitution of splicing activity in human U2AF-depleted extracts by adding U2AF65 requires the presence of the RS domain (Zamore et al., 1992). Further confounding the specific role of the RS domain is the observation that its presence is essential but redundant; studies in *D. melanogaster* demonstrate that either the RS domain of U2AF-L or the C-terminal RS domain of U2AF-S, but not both are required for

viability (Rudner, Breger, Kanaar, Adams, & Rio, 1998). In the future, it will be necessary to clarify the role of this RS domain in splicing but this goal falls outside of the scope of this thesis.

1-8.2.1. U2AF-L domain structure (N-terminus to C-terminus): ULM

This short unstructured coil participates in a protein-protein dimerization interface with the UHM of U2AF-S and becomes structured upon interfacing with its binding partner (Kielkopf, Rodionova, Green, & Burley, 2001).

1-8.2.2. U2AF-L domain structure (N-terminus to C-terminus): RRM1, RRM2

U2AF-L contains two tandem RRMs. An RRM is characterized by a β-sheet formed from four β-strands which canonically conform to the β1-α2-β2-β3-α3-β6 fold (Muto & Yokoyama, 2012). This β-sheet interacts with RNA via stacking interactions in canonical RRMs analyzed to date and is buttressed by two α-helices (Clery, Blatter, & Allain, 2008). Each of the two RRMs contain two RNP consensus motifs, which are short sequences on the 4-stranded β-sheet of the RRM, and these short consensus sequences contain aromatic residues that participate in stacking interactions with the pre-mRNA (Adam, Nakagawa, Swanson, Woodruff, & Dreyfuss, 1986; Dreyfuss, Swanson, & Pinol-Roma, 1988; Kenan, Query, & Keene, 1991; Query, Bentley, & Keene, 1989; Swanson, Nakagawa, LeVan, & Dreyfuss, 1987). The two RRMs in U2AF-L bind the PPT but the basis of this is unclear since a single RRM can only bind 3-4 bases and metazoan PPTs range from 10-40 bases (Sickmier et al., 2006). These RRMs are separated by a flexible linker which regulates their orientation relative to each other, regulates their interaction with the pre-mRNA, and the linker itself also assumes structure and interacts with the pre-mRNA (Agrawal et al., 2016; Kang et al., 2020; Mackereth et al., 2011).

1-8.2.3. U2AF-L domain structure (N-terminus to C-terminus): UHM

This domain has the same general fold as an RRM, and both folds originated from the same ancestor, however UHMs mediate protein-protein interactions in RNA-binding proteins. This UHM binds the ULM in SF1 (Selenko et al., 2003).

1-8.3. U2AF-S recognizes the AG di-nucleotide at the 3′ SS

The 3′ boundary of the intron is defined by a yAG/G motif and is recognized by U2AF-S (Merendino et al., 1999; S. Wu et al., 1999; Zorio & Blumenthal, 1999a, 1999b).

The domain structure comprising the conserved RNA-binding core of U2AF-S is summarized in Sections 1-8.3.1. to 1-8.3.3. below and omits the C-terminus of U2AF-S because it is poorly conserved in length and sequence composition across species, poorly characterized, and is not known to have a well-defined role in in 3′ SS recognition. Additionally, this C-terminal region has diversified considerably in different orthologues, paralogues, and protein isoforms of U2AF-S. For example, the C-terminus of vertebrate U2AF35 consists of an RS domain, whereas the C-terminus of one isoform of mouse, rat, and human U2AF26 consists of a THD, which is a clock regulator responsible for coupling the splicing function of U2AF26 with the regulation of peripheral clocks in order to regulate circadian mRNA expression (see Section 1-7.3.1.).

The conserved RNA-binding core of U2AF-S consists of a UHM flanked on both the N-terminus and C-terminus by a ZF (zinc finger). Hereafter, the N-terminal ZF will be referred to as ZF1, and the C-terminal ZF will be referred to as ZF2.

1-8.3.1. U2AF-S domain structure (N-terminus to C-terminus): ZF1

U2AF-S contains two CCCH-type ZFs, which create a unified binding site in order to co-operatively bind the target pre-mRNA sequence (Shepard et al., 2002; Webb & Wise, 2004; H. Yoshida et al., 2015; H. Yoshida et al., 2020).

1-8.3.2. U2AF-S domain structure (N-terminus to C-terminus): UHM

This UHM binds the ULM in U2AF-L (Birney, Kumar, & Krainer, 1993; Kielkopf et al., 2001; Shepard et al., 2002; Webb & Wise, 2004).

1-8.3.3. U2AF-S domain structure (N-terminus to C-terminus): ZF2

See description above for ZF1.

1-8.4. Arrangement of U2AF-S, U2AF-L, and SF1 domains in the U2AF/SF1 complex

The arrangement of the domains of U2AF-S, U2AF-L, and SF1 in the U2AF/SF1 complex has been summarized in Fig. 1-6 below in order to provide context for the remainder of this thesis.

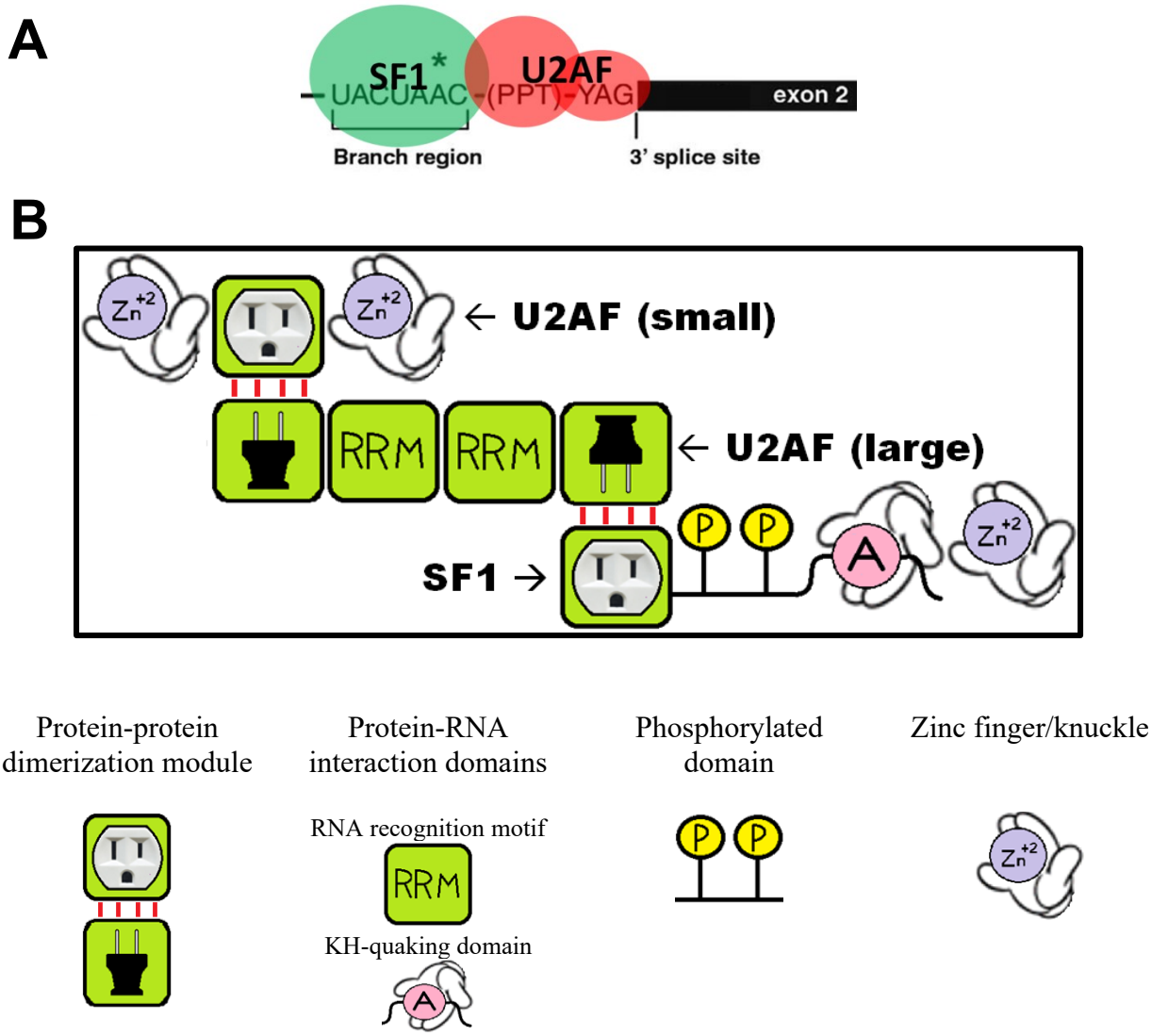


Figure 1-6. The 3′ SS is recognized by the U2AF/SF1 complex. (A) SF1 recognizes the consensus BPS yUnAy (UACUAAC represents the optimal binding site). U2AF-L recognizes the PPT. U2AF-S recognizes the yAG tri-nucleotide motif at the 3′ boundary of the intron. (B) Arrangement of the domains of U2AF-S, U2AF-L, and SF1 in the U2AF/SF1 complex. Protein domains are arranged from the N-terminus to C-terminus of each protein (left to right). Domain symbols are provided below. U2AF-S consists of a UHM flanked by two ZFs. U2AF-L consists of a N-terminal ULM, followed by two RRMs, followed by a UHM. SF1 consists of an N-terminal ULM, followed by the phosphorylated domain, followed by the KH-QUA2 domain, followed by 1-2 zinc knuckles.

1-9. Summary of experimentally derived U2AF/SF1 structure accessions in the PDB

Considerable effort has been invested over many years into experimental determination of U2AF/SF1 structures. With respect to atomic-resolution models however, success has been limited so far to the structure of 1-2 domains, with the exception of two structures consisting of 3 domains. Sections 1-9.1. to 1-9.6. below catalogue all reported structures to date. Additionally, the listed structures are analyzed in detail focusing on the most recent, comprehensive, biologically accurate, and least redundant structures.

1-9.1. U2AF dimer

A total of five X-ray structures have been deposited. The earliest was the minimal human U2AF35/U2AF65 interface (Kielkopf et al., 2001). This was followed by the *S. pombe* dimer of full-length U2AF23 and the ULM of U2AF59 (H. Yoshida et al., 2015). Finally, three structures were solved of the aforementioned *S. pombe* dimer bound to a model RNA containing the target binding sequence of U2AF-S (H. Yoshida et al., 2020). These are the only five structures containing U2AF-S and no structures exist of apo U2AF-S (summarized in Table 1-8).

Table 1-8: X-ray derived U2AF dimer structures

Structure	PDB accession code	PDB reported resolution	Species
U2AF35 (UHM) + U2AF65 (ULM)	1JMT	2.2 Å	*H. sapiens*
U2AF23 (full-length) + U2AF59 (ULM)	4YH8	1.7 Å	*S. pombe*
U2AF23 (full-length) + U2AF59 (ULM) + RNA (UAGGU)	7C06	3.0 Å	*S. pombe*
U2AF23 (full-length) + U2AF59 (ULM) + RNA (AAGGU)	7C07	3.2 Å	*S. pombe*
U2AF23 (full-length, S34Y) + U2AF59 (ULM) + RNA (UAGGU)	7C08	3.4 Å	*S. pombe*

1-9.1.1. Apo U2AF dimer

The human U2AF35/U2AF65 interface is established by W92 (U2AF65) inserting itself into a hydrophobic pocket on U2AF35 and W134 (U2AF35) inserting itself into a hydrophobic pocket on U2AF65 in a reciprocal 'lock-and-key' mechanism. The interaction of these two proteins causes the otherwise unstructured ULM of U2AF65 to assume a well-defined structure in the context of a dimer. Both the human and the more recent *S. pombe* apo structures are closely superimposable with each other (rmsd = 0.9 Å) and both have the same reciprocal Trp recognition. However, the *S. pombe* structure has a more extensive interface since it includes an interfacial region not seen in the human structure between the C-terminal α-helix of U2AF23 and the N-terminal α-helix of U2AF59 formed by hydrophobic contacts. However, the most important feature of the *S. pombe* structure is the architecture of the two ZFs because when combined with additional biochemical investigations it provides new insights into intron recognition as the ZFs are directly responsible for binding the 3′ SS.

In the *S. pombe* structure, the UHM is only a scaffold to organize the ZFs and does not directly interact with RNA. The overall structures of both ZFs resemble typical CCCH-type ZFs such as those found in TIS11d (TPA-inducible sequence 11d) or MBNL1 (muscleblind-like splicing regulator 1) (Hudson, Martinez-Yamout, Dyson, & Wright, 2004; Teplova & Patel, 2008). However, the ZFs of U2AF23 form an entirely new RNA-binding surface not observed in previously characterized ZFs. The two ZFs lie side-by-side against the β-sheet of the UHM and feature an intimate interaction with each other on the β-sheet surface. Additionally, the ZFs of U2AF23 are unique because although canonical RNA-binding CCCH-type ZFs have an almost completely conserved central aromatic amino acid residue (Phe or Tyr) between the third and fourth Zn-coordinating residues which forms stacking interactions with RNA bases, in U2AF23

only ZF2 has this residue, corresponding to F165 (Hudson et al., 2004; Teplova & Patel, 2008). Despite this, deletion of ZF1 causes a complete loss of RNA-binding in U2AF23 implying that the two ZFs cooperate to form a unified RNA-binding surface with the UHM holding them together. Additionally, with additional biochemical experiments the authors were able to demonstrate that the two ZFs create a unified positively charged area with hydrophobic knobs to form a single binding site for one consecutive RNA sequence with four pockets for base recognition and a strong preference for UAGG. These results indicate that the two ZFs recognize a 4-base RNA in a sequence-specific manner.

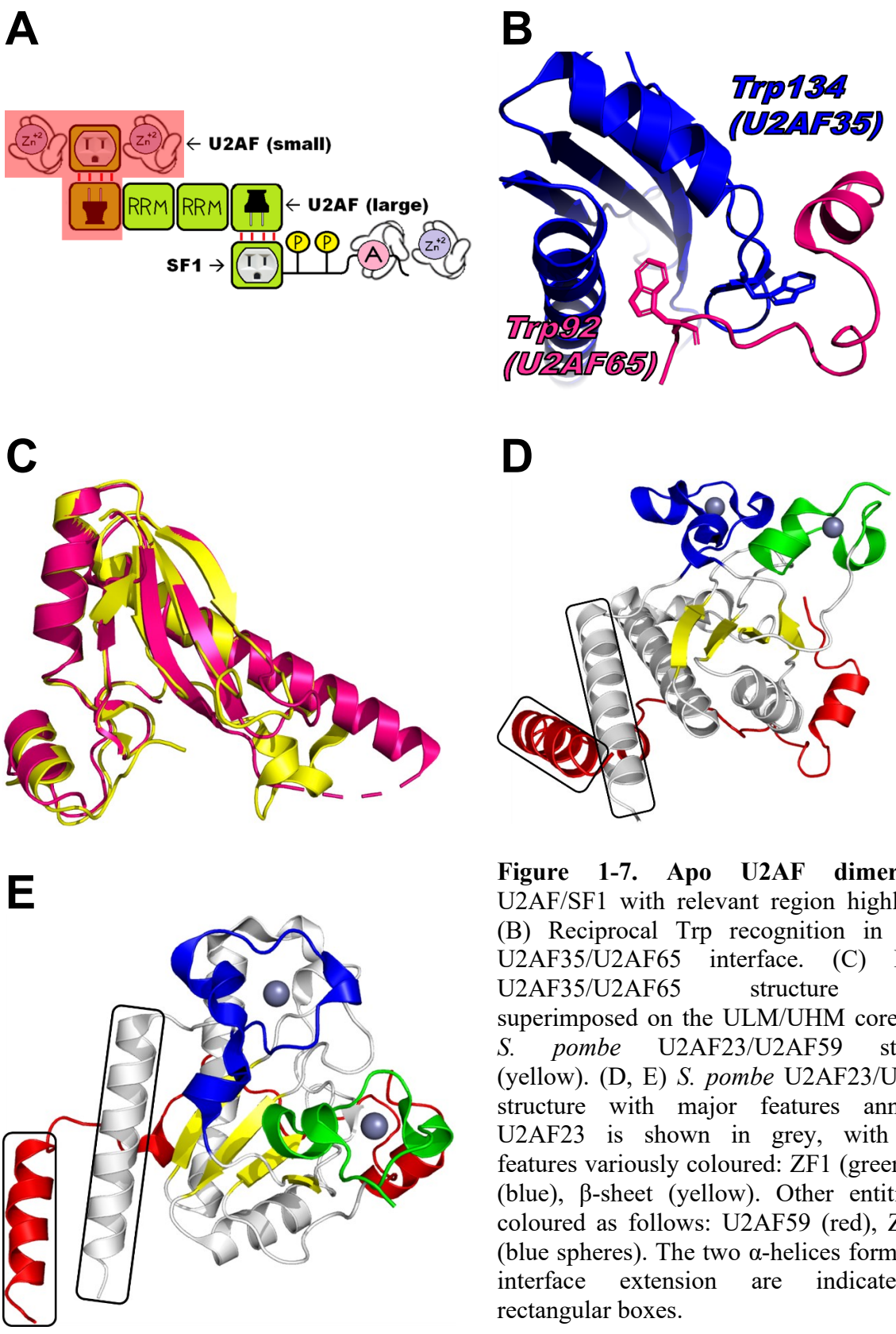


Figure 1-7. Apo U2AF dimer. (A) U2AF/SF1 with relevant region highlighted. (B) Reciprocal Trp recognition in human U2AF35/U2AF65 interface. (C) Human U2AF35/U2AF65 structure (pink) superimposed on the ULM/UHM core of the *S. pombe* U2AF23/U2AF59 structure (yellow). (D, E) *S. pombe* U2AF23/U2AF59 structure with major features annotated. U2AF23 is shown in grey, with major features variously coloured: ZF1 (green), ZF2 (blue), β-sheet (yellow). Other entities are coloured as follows: U2AF59 (red), Zn ions (blue spheres). The two α-helices forming the interface extension are indicated by rectangular boxes.

1-9.1.2. RNA-bound U2AF dimer

A follow up study to the apo *S. pombe* U2AF dimer structure was published in which the same protein complex was co-crystallized with RNA (H. Yoshida et al., 2020). A total of three structures were solved which all have the same overall structure. Unlike the apo complex which crystallized in spacegroup $C222_1$ and only contains one copy of the U2AF dimer in the asymmetric unit, all three RNA-bound structures crystallized in spacegroup $P12_11$ with nine copies of the RNA-bound U2AF dimer in the asymmetric unit.

The prototype structure is wildtype U2AF dimer bound to UAGGU RNA and establishes the basic mechanism of RNA binding. This structure reveals that all the residues involved in RNA binding, the cancer-associated S34 hotspot (see Section 1-7.4.), and a secondary disease-associated hotspot (Q157P/R mutations in human U2AF35, corresponding to Q151 in *S. pombe* U2AF23) are identical between *S. pombe* U2AF23 and human U2AF35 (Ilagan et al., 2015). Therefore, the RNA-bound *S. pombe* U2AF dimer was further used to investigate the molecular basis of altered splicing caused by S34F/Y mutations. In order to accomplish this, wildtype U2AF dimer was co-crystallized with AAGGU RNA, and a mutant complex containing U2AF23 (S34Y) was co-crystallized with UAGGU RNA. For clarity and simplicity however, the following discussion is restricted to the prototype complex.

The overall RNA-bound structure is almost the same as the apo structure in all nine copies of the asymmetric unit yielding rmsd = 0.6-1.0 Å when different copies of the RNA-bound complex in the asymmetric unit are compared with the apo complex. RNA is bound by both ZFs, and each copy in the asymmetric unit shows the same RNA contacts. The most obvious structural difference between the apo and RNA-bound structure is that M1-D14 is disordered in the apo structure whereas this region can be modeled to varying degrees in the

different copies of the asymmetric unit in the RNA-bound structure; only M1 is disordered in four copies, M1-L5 is disordered in one copy, M1-A6 is disordered in one copy, and M1-K15 is disordered in three copies. Additionally, the RNA interaction causes some residues of U2AF23 to change their side-chain configuration, especially in the ZFs.

In the discussion that follows, residues within the UAGGU sequence will be referred to by a number to indicate their position relative to the intron/exon boundary (see Fig. 1-8A). The fourteen N-terminal residues of U2AF23 are visible to varying degrees in the RNA-bound structure because L5-Y9 is stabilized in some copies of the asymmetric unit due to interactions with neighbouring molecules conferred by the crystal packing. However, L5-Y9 does not interact with RNA. Additionally, part of the N-terminus is stabilized through its interaction with RNA; specifically, the side chain of E12 interacts with the 2′-hydroxyl group of the sugar portion of -1G, and K15 interacts with the phosphate group between -3U and -2A. The RNA-bound structure is shown in Fig. 1-8B and Fig. 1-8C below with its main features annotated in order to show the basic organization of the complex.

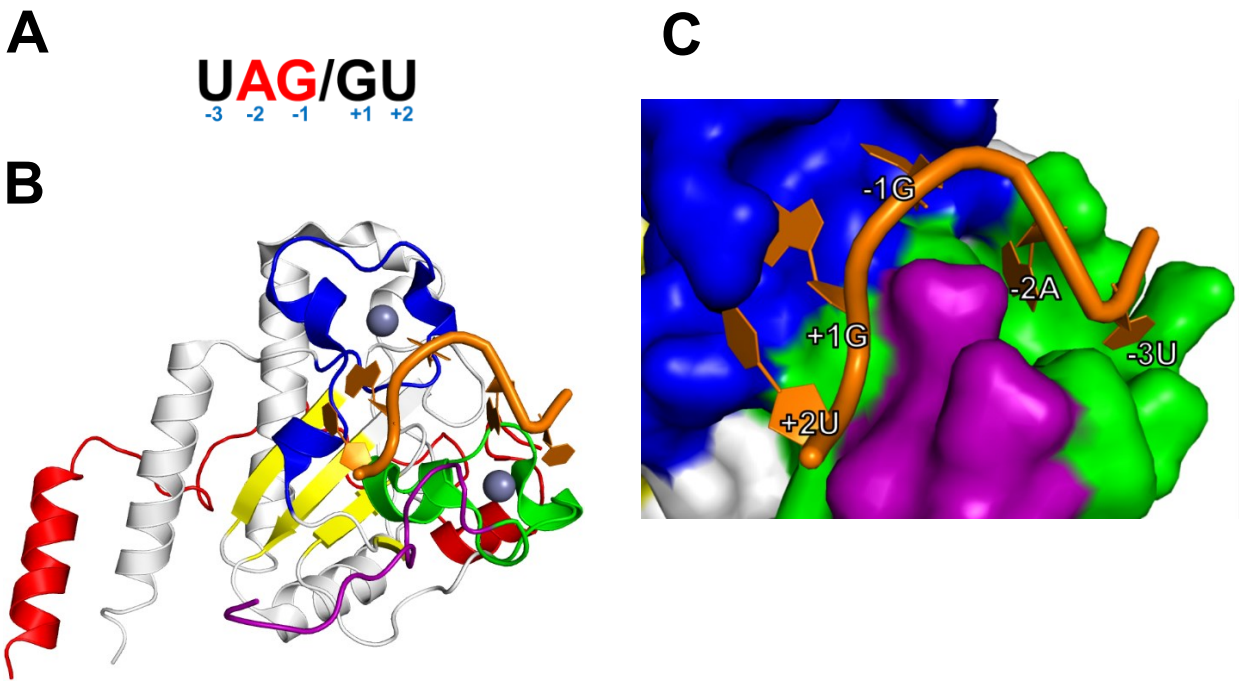


Figure 1-8. RNA-bound U2AF dimer. (A) Nucleotide numbering within the 3′ SS RNA sequence (PDB accession 7C06). (B) PDB accession 7C06 (chains A, B, and C) with major features annotated. Compare with Fig. 1-7E. U2AF23 is shown in grey, with major features variously coloured: ZF1 (green), ZF2 (blue), β-sheet (yellow), N-terminus of U2AF23 (purple; this is disordered in the apo structure). Other entities are coloured as follows: U2AF59 (red), Zn ions (blue spheres), RNA (orange). (C) Zoomed-in surface representation of (B). Nucleotides are numbered as indicated in (A).

With respect to the overall binding of the RNA, -3U, -2A, -1G, and +1G are held in place by both ZFs, and the spatial relationship between -2A and -1G is fixed by the positions of ZF1 and ZF2. Surprisingly, the ZFs contribute to the preference for bases flanking the AG di-nucleotide. However, various orientations are seen for +2U in different copies in the asymmetric unit, indicating that this nucleotide is not strictly recognized by U2AF23.

The aromatic ring of F20 in ZF1 is rotated slightly to stack with -2A. The guanidyl group of R28 in ZF1 stacks against -2A and interacts with the 2′ hydroxyl group of -3U. The aromatic ring of F165 in ZF2 is rotated slightly to stack with +1G and interacts with -1G through

perpendicular π–π stacking and interacts with +1G through parallel π–π stacking. The RNA-bound structure reveals that the RNA-binding site is formed by the co-operative interaction of both ZFs, since the aromatic ring of both F20 in ZF1 and F165 in ZF2 form the van der Waals contact surface for -1G.

At the -3U position of the RNA, the uracil is stacked with the imidazole ring of H29 and surrounded by S34, R35, and a short helix of ZF1. Stacking between the imidazole and uridine rings is crucial at this position for RNA binding. Additionally, the crystal structure indicates that U2AF23 has a strong preference for guanine at the +1 position; +1G forms a π–π interaction with F165 and a cation–π interaction with R150 involving the six-membered ring of the base. Pyrimidine (C or U) at this position would not be large enough to interact with R150 or F165. Additionally, +1G interacts directly with -1G, R145, E146, and C148 through hydrogen bonds. An adenine at this position would remove all these hydrogen bonds. Together, these features explain the strong preference for guanine at the +1 position and are consistent with earlier SELEX experiments showing that guanine is the preferred nucleotide at this position for U2AF35 (S. Wu et al., 1999).

1-9.2. Apo U2AF-L

Seven structures of apo human U2AF65 have been solved through both X-ray crystallography and solution NMR (Glasser, Agrawal, Jenkins, & Kielkopf, 2017; Ito, Muto, Green, & Yokoyama, 1999; Kang et al., 2020; Mackereth et al., 2011; Thickman, Sickmier, & Kielkopf, 2007). Additionally, three structures of apo mouse U2AF65 have been solved through both X-ray crystallography and solution NMR, but these structures remain unpublished.

Most recently, two NMR structure ensembles have been deposited for *D. melanogaster* LS2 (large subunit 2); the *D. melanogaster* orthologue of U2AF65 is called U2AF50, and LS2 is a paralogue of this gene which arose from a retrotransposition event. As with U2AF50, LS2 heterodimerizes with U2AF38 which is the U2AF35 orthologue in *D. melanogaster*. However, unlike U2AF50, LS2 is a splicing repressor and is highly enriched in testes, mediating testis-specific alternative splicing (Giot et al., 2003; Taliaferro, Alvarez, Green, Blanchette, & Rio, 2011). These structures have not been published yet.

All reported apo U2AF-L structures encompass the region C-terminal to the ULM and span regions of 1-2 domains. These structures are summarized below in Table 1-9.

Table 1-9: X-ray and NMR derived apo U2AF-L structures

Structure	PDB accession code	Solution method	PDB reported resolution/conformers submitted	Species
RRM1	5W0G	X-ray	1.1 Å	*H. sapiens*
RRM1	2HZC	X-ray	1.5 Å	*H. sapiens*
RRM1	1U2F	NMR	1 conformer	*H. sapiens*
RRM2	2U2F	NMR	1 conformer	*H. sapiens*
RRM2	3V4M	X-ray	1.8 Å	*M. musculus*
RRM2	4Z2X	X-ray	2.2 Å	*M. musculus*
RRM2	5W0H	X-ray	1.1 Å	*H. sapiens*
UHM	2M52	NMR	20 conformers	*M. musculus*
RRM1 + inter-RRM linker + RRM2	6TR0	NMR	10 conformers	*H. sapiens*
RRM1 + inter-RRM linker + RRM2	2YH0	NMR	10 conformers	*H. sapiens*
Inter-RRM linker + RRM2	7AAO	NMR	10 conformers	*D. melanogaster*
RRM2	7AAF	NMR	10 conformers	*D. melanogaster*

U2AF-L is a modular protein defined by five characterized domains (see Section 1-8.2.). The ULM has been addressed in Section 1-9.1. The UHM will be discussed in Section 1-9.4. in the context of a dimer with human SF1. Although structures derived from high-quality data are available for RRM1 and RRM2 individually, they will not be discussed here since they represent rigid bodies with a well-defined and evolutionarily conserved structure and are redundant in the context of a focused discussion when larger, multi-domain structures are available. Instead, this section will address the two NMR structure ensembles spanning the region from the N-terminus of RRM1 to the C-terminus of RRM2 (Kang et al., 2020; Mackereth et al., 2011). No X-ray structure exists representing this isolated region in its apo state because it is very flexible owing to the inter-RRM linker and it may even be impossible to obtain well-diffracting crystals of this region.

The first reported NMR structure ensemble establishes a mechanism to ensure binding specificity of the RRMs of human U2AF65 to a given PPT sequence based on the RRMs existing in either an open or closed state relative to each other, where only the open state supports RNA binding (J. R. Huang et al., 2014). This and other studies combine NMR spectroscopy, SAXS (small angle X-ray scattering), computational analysis and FRET (fluorescence resonance energy transfer) to reveal that RRM1 and RRM2 exist mainly in a dynamic arrangement of closed, inactive states, that the equilibrium between the two states is shifted to the open state upon the addition of RNA, and that the degree of the shift is commensurate with the binding affinity of the RNA to the RRMs (J. R. Huang et al., 2014; Mackereth et al., 2011; Voith von Voithenberg et al., 2016). This model of a population shift correlates PPT strength with the efficiency of splicing of a given intron. This study also establishes that the linker is intrinsically disordered and highly

flexible in solution, a feature presumably required to enable the tandem RRMs to sample a large conformational space to support the conformational population shift.

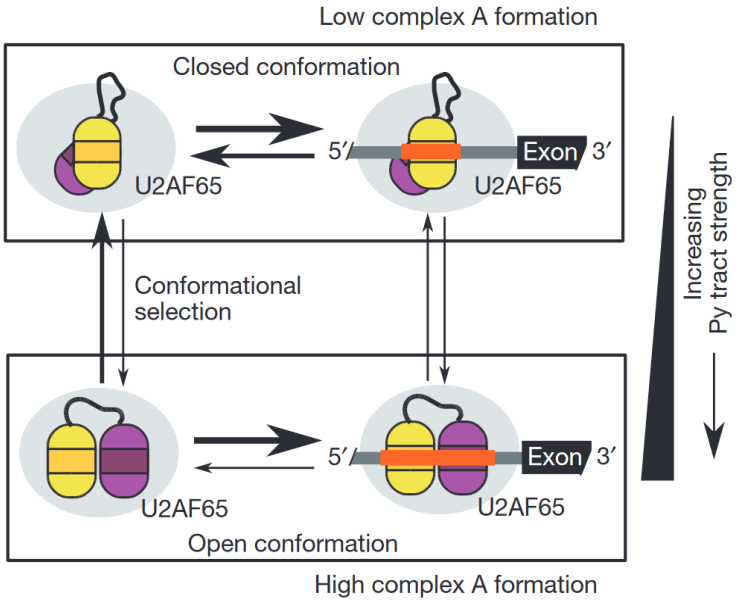


Figure 1-9. Cartoon representation of the closed/open conformation population shift model for the RRMs of human U2AF65. Image summarizes the relationship between the two conformations and PPT strength (adapted from Mackereth et al., 2011).

Despite the important insights that it provides, this NMR structure ensemble is flawed due to limitations in the data available for modelling the structures. The first oligonucleotide-bound structures of this region of human U2AF65 containing an unmodified, natural inter-RRM linker (see Section 1-9.3.) were reported several years after this apo NMR structure ensemble was published. Therefore, this ensemble does not take into account any structural information for the inter-RRM linker because none was available at the time it was modelled. After the aforementioned oligonucleotide-bound structures were reported, a newer NMR structure ensemble of apo human U2AF65 spanning the region from the N-terminus of RRM1 to the C-terminus of RRM2 was reported. This ensemble is more biologically accurate since it is modelled using new structural insights provided by the oligonucleotide-bound structures. The older ensemble is an incomplete picture because it does not provide insight into the mechanisms that reduce binding to weak and presumably non-functional sequences (Agrawal et al., 2016; Mackereth et al., 2011).

Although the inter-RRM linker is intrinsically disordered, it is evolutionarily conserved in length. Additionally, it is also conserved with respect to the presence of multiple hydrophobic aliphatic residues. These conserved features suggest that the inter-RRM linker of U2AF-L has a functional role (J. R. Huang et al., 2014; Mackereth et al., 2011). The more recent ensemble incorporates these conserved features into the model and provides a mechanism explaining how the inter-RRM linker proofreads RNA binding and significantly enhances binding specificity for strong PPTs. In order to establish this mechanism, the newer ensemble was also complemented with studies using a technique for large-scale mapping of protein-RNA interactions by *in vitro* and *in vivo* iCLIP (individual nucleotide resolution UV cross-linking and immunoprecipitation) (Sutandy et al., 2018).

Briefly, the newer ensemble confirms that RRM1 and RRM2 are well-defined but do not adopt a specific domain arrangement in solution as shown previously (J. R. Huang et al., 2014). More significantly however, the C-terminus of the linker (V250-H259) adopts a well-defined, rigid conformation in the ensemble and chemical shift differences for NMR signals of residues in RRM2 suggest that this region of the linker transiently interacts with the RNA binding surface of RRM2, thereby exerting autoinhibition by directly competing with RNA for the binding surface of RRM2. Also, iCLIP experiments establish that mutations in the linker that interfere with this function increase nonspecific binding to weak PPTs in natural pre-mRNA transcripts *in vitro* and impair splicing fidelity *in vivo* (Sutandy et al., 2018; Zarnack et al., 2013).

In addition to important new insights regarding the roles of the inter-RRM linker, the more recent ensemble also provides new insights into the sequence that flanks the region spanning RRM1 to RRM2 since it is modeled using a construct that is longer at both the N- and C-termini than the one used to generate the older ensemble. Specifically, a short α-helix flanks the N-terminus of the RRM1 core which is stabilized by interactions with D231 and Y232 at the N-terminus of the inter-RRM linker. Another short α-helix flanks the C-terminus of the RRM2 core. The more recent ensemble is summarized below in Fig. 1-10.

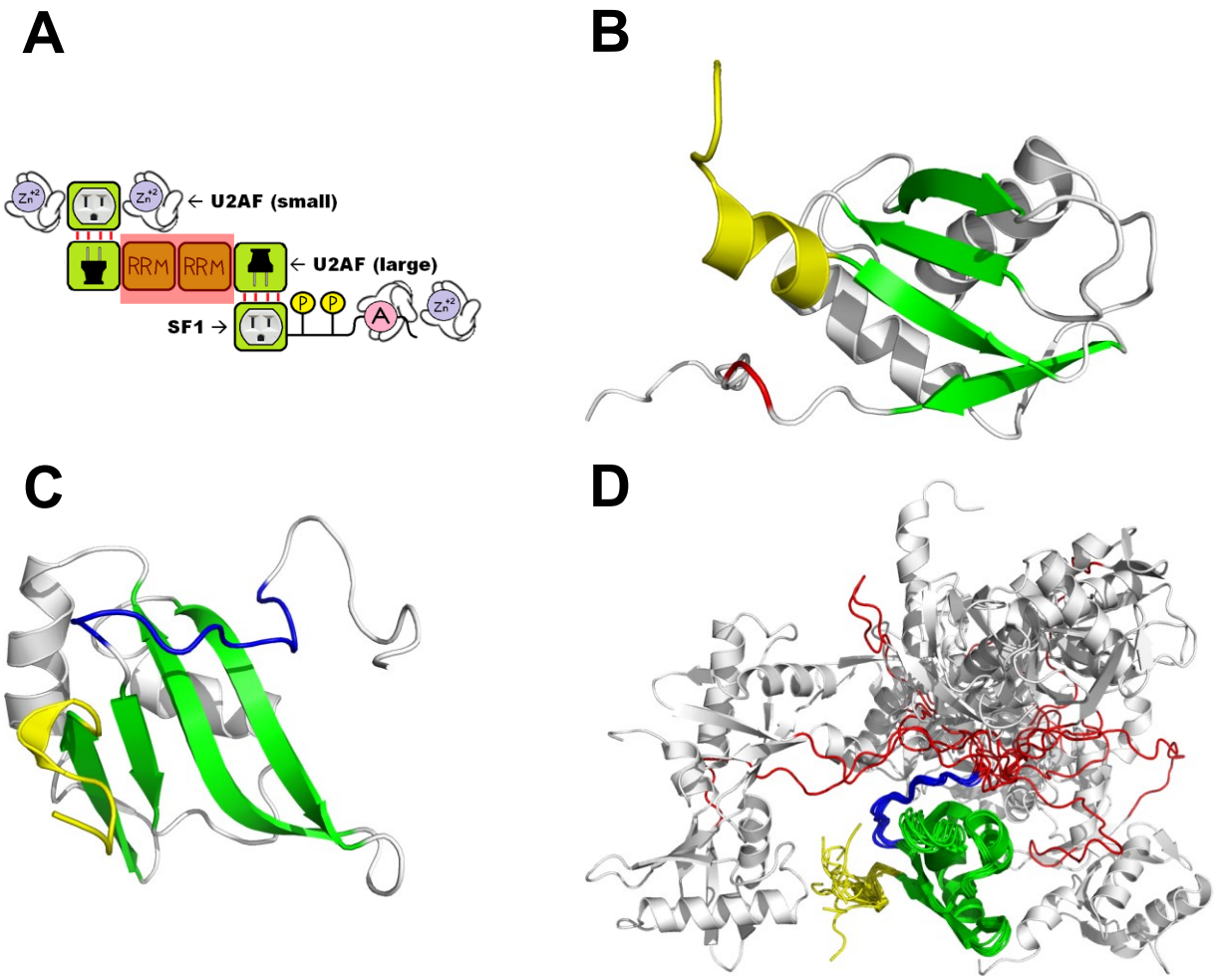


Figure 1-10. NMR structure ensemble modeling the autoinhibitory mechanism of the inter-RRM linker of human U2AF65. (A) U2AF/SF1 with relevant region highlighted. (B) RRM1 and the N-terminus of the inter-RRM linker: β-sheet (green), α-helix N-terminal to the RRM1 core (yellow), D231 and Y232 (red). (C) RRM2 and the C-terminus of the inter-RRM linker: β-sheet (green), α-helix C-terminal to the RRM2 core (yellow), C-terminus of the inter-RRM linker (blue: V250-H259). (D) Movements of functionally important regions in the inter-RRM linker during conformational shifts. The ten states of the ensemble are aligned onto RRM2: flexible N-terminus of the inter-RRM linker (red), rigid C-terminus of the inter-RRM linker which binds the RNA-binding face of RRM2 (blue: V250-H259), RRM2 (green), α-helix flanking the C-terminus of the RRM2 core (yellow).

1-9.3. Oligonucleotide-bound U2AF-L

All of these structures are human derived and consist of the two RRMs bound to a model PPT sequence. Due to several technical challenges that will be described in more detail later in this section, these structures were very difficult to obtain. The first successful structure solution was an X-ray structure of a modified U2AF65 construct with a shortened inter-RRM linker bound to a polyuridine RNA (Sickmier et al., 2006). This was used as a prototype for several later structures containing DNA (for technical reasons) to investigate the RNA-binding properties of U2AF65 (Agrawal et al., 2014; Jenkins, Agrawal, Gupta, Green, & Kielkopf, 2013). Eventually, four structures were successfully solved of the RRMs bound to a DNA/RNA hybrid oligonucleotide but with a natural inter-RRM linker (Agrawal et al., 2016). These were later used as a prototype to structurally probe two cancer-associated U2AF65 point mutations, again using a modified DNA/RNA hybrid (Maji et al., 2020). There is also one NMR structure ensemble of the RRMs with a natural inter-RRM linker bound to polyuridine RNA, however the inter-RRM linker has not been modelled correctly because this ensemble was modeled and reported before structural information on the linker was made available via X-ray structures (Mackereth et al., 2011). The structures discussed above are summarized below in Table 1-10.

Table 1-10: X-ray and NMR derived structures of oligonucleotide-bound U2AF-L

Structure	PDB accession code	Solution method	PDB reported resolution/conformers submitted
U2AF65 (Δ linker) + U12 RNA	2G4B	X-ray	2.5 Å
U2AF65 (Δ linker) + DNA	3VAF	X-ray	2.5 Å
U2AF65 (Δ linker) + DNA	3VAG	X-ray	2.2 Å
U2AF65 (Δ linker) + DNA	3VAH	X-ray	2.5 Å
U2AF65 (Δ linker) + DNA	3VAI	X-ray	2.2 Å
U2AF65 (Δ linker) + DNA	3VAJ	X-ray	1.9 Å
U2AF65 (Δ linker) + DNA	3VAK	X-ray	2.2 Å
U2AF65 (Δ linker) + DNA	3VAL	X-ray	2.5 Å
U2AF65 (Δ linker) + DNA	3VAM	X-ray	2.4 Å
U2AF65 (Δ linker, D231V) + DNA	4TU7	X-ray	2.1 Å
U2AF65 (Δ linker) + DNA	4TU8	X-ray	1.9 Å
U2AF65 (Δ linker) + DNA	4TU9	X-ray	2.0 Å
U2AF65 (natural linker) + U9 RNA	2YH1	NMR	10 conformers
U2AF65 (natural linker) + DNA/RNA hybrid	5EV1	X-ray	2.0 Å
U2AF65 (natural linker) + DNA/RNA hybrid	5EV2	X-ray	1.9 Å
U2AF65 (natural linker) + DNA/RNA hybrid	5EV3	X-ray	1.5 Å
U2AF65 (natural linker) + DNA/RNA hybrid	5EV4	X-ray	1.6 Å
U2AF65 (natural linker, G301D) + DNA/RNA hybrid	6XLX	X-ray	1.7 Å
U2AF65 (natural linker, N196K) + DNA/RNA hybrid	6XLV	X-ray	1.4 Å
U2AF65 (natural linker) + DNA/RNA hybrid	6XLW	X-ray	1.5 Å

All structures in Table 1-10 aim to model the binding of the two RRMs and inter-RRM linker to the PPT. However, despite the important insights that they can provide, the path to generating these structures is complicated and difficult due to two interrelated challenges.

First, although the isolated RRMs can be accurately described as rigid bodies and produced crystals yielding high-quality diffraction data, the inherent flexibility of the inter-RRM linker is a functional feature. Therefore, a biologically accurate structure of the region spanning from the N-terminus of RRM1 to the C-terminus of RRM2 must include a natural inter-RRM linker which is a challenge owing to the flexibility of the linker. However, because the RRMs only bind RNA in a discrete open conformation, binding this region of the protein to its target

RNA sequence is expected to convert the crystallization target from an indefinite continuum of structures into a single rigid body consisting of the protein/RNA heterodimer. However, even after addressing this issue, obtaining a sample of pure and uniform protein/RNA complex is still a major challenge because of a second technical difficulty which is that the PPT is extremely variable in length and sequence composition and the RRMs must be able to bind sequences promiscuously in order to recognize multiple variations of this sequence. Therefore, the binding of the RRMs to RNA is likely best described as an avidity effect rather than an affinity meaning that the stable binding of RNA to the RRMs is conferred by the combined strength of all bases in an RNA sequence. In addition to this, a polypyrimidine RNA sequence is expected to stably bind the RRMs in several unique registers. Therefore, it is essential to address all of the considerations above to successfully generate a suitable protein/RNA sample for crystallization.

The first structure was successfully obtained by deleting the inter-RRM linker and co-crystallizing the protein with a polyuridine RNA. Interestingly, the RNA does not form a 1:1 *cis*-complex with the protein in the crystal lattice. Instead, half of the RNA sequence binds RRM1 in one protein molecule and the other half of the RNA binds RRM2 in a second protein molecule. In this way, a *trans*-complex is formed between the protein and RNA whereby the RNA serves as a bridge between protein molecules in the crystal lattice. This structure was subsequently used as a prototype for additional structures that aimed to probe the RNA-binding properties of the RRMs more deeply by using different co-crystallization oligonucleotides as well as a D231V point mutation. Additionally, bromo-uridine and bromo-deoxyuridine were incorporated into the oligonucleotides for these newer structures because it is known that halogenated bases have unique base-stacking properties relative to their unmodified counterparts, the halogen can be successfully used for MAD (multiple-wavelength anomalous diffraction) and SAD (single-

wavelength anomalous diffraction) phasing in the structure solution process, and the bulkiness of the halogen atom can be used to fix the register of oligonucleotide binding because steric clashes can prevent effective binding in certain registers (Jiang, Sheng, Carrasco, & Huang, 2007; Shah, Wu, & Rana, 1994; Sternglanz & Bugg, 1975).

In the first reported structure of the oligonucleotide-bound RRMs as well as its derivative structures described above, the crystallization artifact and deletion of the inter-RRM linker are both non-natural features. Additionally, the inter-RRM linker is functionally important as discussed in Section 1-9.2. making these structures flawed. However, the molecular insights provided by the linker deletion structures, as well as the tools developed to obtain these structures such as the use of strategically placed halogenated bases in the oligonucleotide sequence made it possible to finally generate biologically accurate X-ray structures of the RRMs bound to oligonucleotide in a natural 1:1 *cis*-complex with a unmodified, natural inter-RRM linker. It should be noted that an NMR structure ensemble was reported representing this complex, but the inter-RRM linker has not been modelled correctly because the solution NMR-based modelling methods used to generate this ensemble are dependent on previously reported models and do not generate a structure *de novo* from raw data as crystallography does (Brunger et al., 1998; Nilges, 1995; Simon, Madl, Mackereth, Nilges, & Sattler, 2010). This ensemble was modeled and reported before the corresponding X-ray structures; therefore no structural information was available at the time to model the inter-RRM linker.

Several technical features were essential in the strategy to successfully obtain the natural structure of the oligonucleotide-bound RRMs. With respect to the protein construct itself, it was necessary to extend the N-terminus beyond the RRM1 core and the C-terminus beyond the RRM2 core because the region flanking the RRMs forms part of the interface which binds the

oligonucleotide. This was combined with the previously established insights with modified oligonucleotides for fixing the RNA binding register and sequential bootstrapping was used to optimize the co-crystallization oligonucleotides for length, the position of a Br-dU, and the identity of the terminal nucleotide (rU, dU or rC).

The first report of the 1:1 *cis*-complex of oligonucleotide-bound RRMs with a natural inter-RRM linker included a total of four structures in which the protein and oligonucleotide conformations are nearly identical to one another. These four structures each contain the same protein construct but a different co-crystallization oligonucleotide and show that RRM1, RRM2 and the inter-RRM linker all cooperate in order to recognize and bind a contiguous nine-nucleotide PPT by concomitantly recognizing the three central nucleotides of the PPT which likely coordinates the conformational arrangement of these disparate regions of the protein.

The inter-RRM linker traverses across the α-helical surface of RRM1 and the central β-strands of RRM2 and is well defined in the electron density. The linker recognizes the central nucleotide at position 5 and, as previously discussed in Section 1-9.2., the N-terminal extension outside of the RRM1 core and the C-terminal extension outside of the RRM2 core both form an α-helix. These two α-helices recognize the 3′ terminus and third nucleotide. At the opposite side of the central fifth nucleotide, the sixth nucleotide is located at the RRM1/RRM2 interface. This nucleotide twists to face away from the linker and instead inserts the uracil into a sandwich between the β2/β3 loops of RRM1 and RRM2. The discovery of nine binding sites for contiguous nucleotides was unexpected. Based on the earlier inter-RRM deletion structures, this U2AF65 construct was expected to bind the minimal seven nucleotides observed in the earlier inter-RRM deletion structures but surprisingly, the RRM2 extension and inter-RRM linker contribute newly identified central nucleotide-binding sites near the RRM1/RRM2 junction and

the RRM1 extension recognizes the 3′-terminal nucleotide (Agrawal et al., 2014; Jenkins et al., 2013; Sickmier et al., 2006). Because RRM1 and RRM2 associate with the PPT in a parallel, side-by-side arrangement, these four structures support the model previously established by the first apo NMR structure ensemble for this region of human U2AF65 in which only the open conformation of the RRMs can bind RNA (Mackereth et al., 2011).

The first four oligonucleotide-bound structures with a natural inter-RRM linker were used as a prototype for a follow up study in which three more very similar structures were reported, investigating cancer-associated mutations in human U2AF65. The structure of human U2AF65 with a natural inter-RRM linker bound in a 1:1 *cis*-complex with oligonucleotide is shown below in Fig. 1-11, with the most important structural features annotated; the structure corresponding to PDB accession code 5EV1 is used as a representative example in this figure to show the shared features of the seven very similar structures corresponding to this complex.

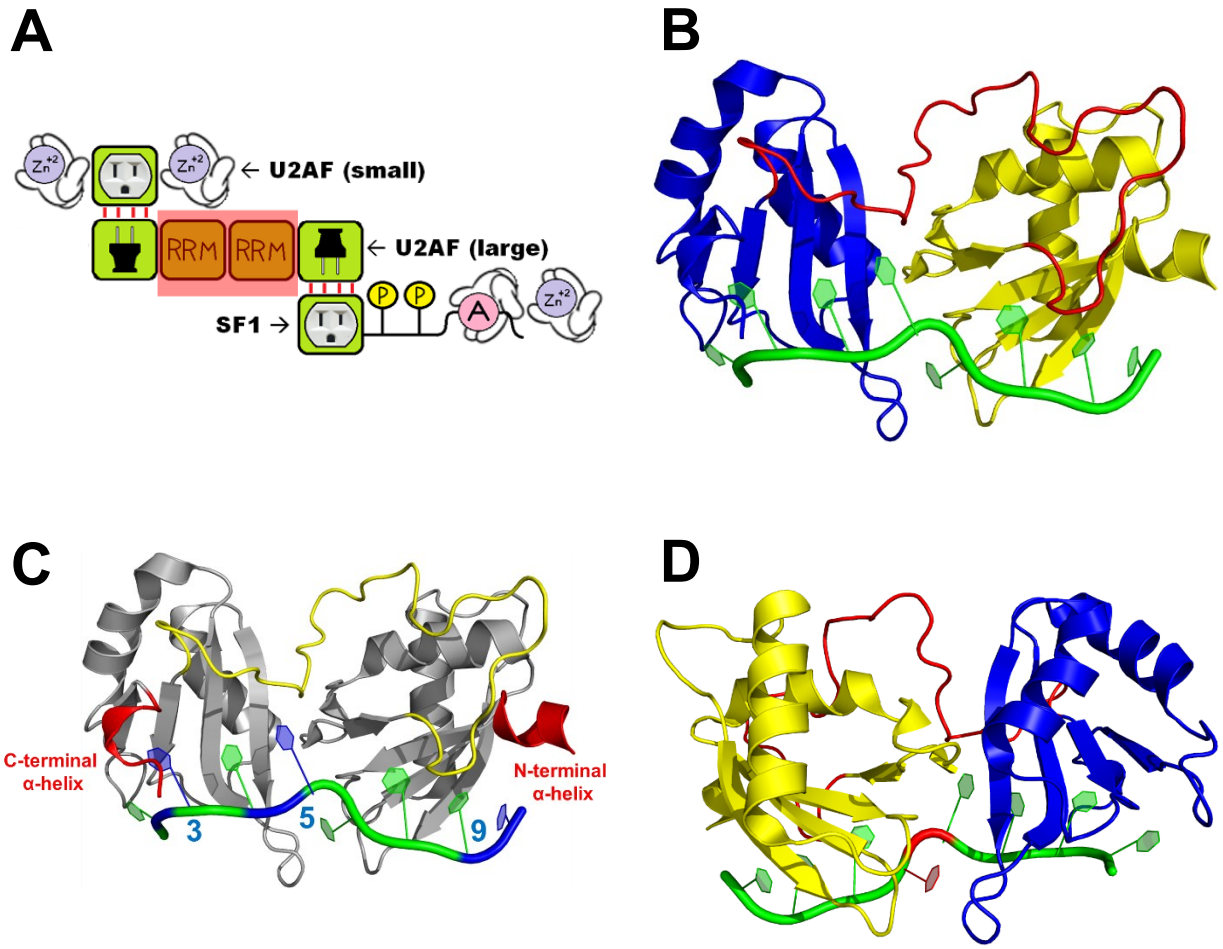


Figure 1-11. Human U2AF65 with a natural inter-RRM linker bound in a 1:1 *cis*-complex with oligonucleotide. PDB accession 5EV1 is shown. (A) U2AF/SF1 with relevant region highlighted. (B) Overall structure with general features annotated: RRM1 (yellow), inter-RRM linker (red), RRM2 (blue), oligonucleotide (green). (C) Overall structure with additional features annotated. Protein is shown in grey, with major features variously coloured: N-terminal α-helix outside of the RRM1 core and C-terminal α-helix outside of the RRM2 core (red), inter-RRM linker (yellow). Oligonucleotide is shown in green, with nucleotides at positions 3, 5 and 9 in blue. Nucleotides 3 and 9 directly contact the terminal α-helices, and nucleotide 5 directly contacts the inter-RRM linker. (D) Overall structure annotated as in (B) with the additional annotation of nucleotide 6 (red). The uracil of this nucleotide is sandwiched between the β2/β3 loops of RRM1 and RRM2.

1-9.4. Apo U2AF-L/SF1 dimer & apo SF1

A total of four structures have been deposited in the PDB of the human U2AF65/SF1 dimer. The first two structures to be deposited were NMR structure ensembles of the U2AF65/SF1 dimerization interface (Selenko et al., 2003). One decade later, two concurrent studies were published which both reported a larger structure consisting of the aforementioned interface as well as the phosphorylated domain of SF1; one study reported the unphosphorylated counterpart and the other described the phosphorylated counterpart of this larger U2AF65/SF1 dimer (W. Wang et al., 2013; Y. Zhang et al., 2013). The unphosphorylated U2AF65/SF1 counterpart was solved as an NMR structure ensemble, whereas the phosphorylated counterpart was solved through X-ray crystallography.

In addition to these two structures, both of the two concurrent studies each also reported the structure of the isolated phosphorylated domain of SF1 in its unphosphorylated state; one of these structures is an NMR structure ensemble and the other is an X-ray structure. All of the structures discussed above are summarized below in Table 1-11.

Table 1-11: X-ray and NMR derived human structures of apo U2AF65/SF1 dimer & apo SF1

Structure	PDB accession code	Solution method	PDB reported resolution/conformers submitted
U2AF65 (UHM) + SF1 (ULM)	1O0P	NMR	10 conformers
U2AF65 (UHM) + SF1 (ULM)	1OPI	NMR	10 conformers
U2AF65 (UHM) + SF1 (ULM + phosphorylated domain, unphosphorylated)	2M0G	NMR	10 conformers
U2AF65 (UHM) + SF1 (ULM + phosphorylated domain, phosphorylated)	4FXW	X-ray	2.3 Å
SF1 (phosphorylated domain, unphosphorylated)	4FXX	X-ray	2.5 Å
SF1 (phosphorylated domain, unphosphorylated)	2M09	NMR	10 conformers

U2AF-L and SF1 interact by the same general mechanism as U2AF-S with U2AF-L via a ULM/UHM dimerization in which a tryptophan on one protein fits into the hydrophobic pocket of its binding partner and vice-versa in a reciprocal 'lock-and-key' binding mechanism. This arrangement induces folding of the otherwise disordered ULM. This configuration was previous discussed in the context of the U2AF-S/U2AF-L interface (Section 1-9.1.1.).

A superimposition of the human ULM/UHM interface from both U2AF35/U2AF65 and U2AF65/SF1 (Fig. 1-12) reveals that both structures are very similar (rmsd = 1.0 Å). The main difference is the presence of an unusually long 30 residue α-helix in U2AF35 which is absent in U2AF65 in the counterpart structure.

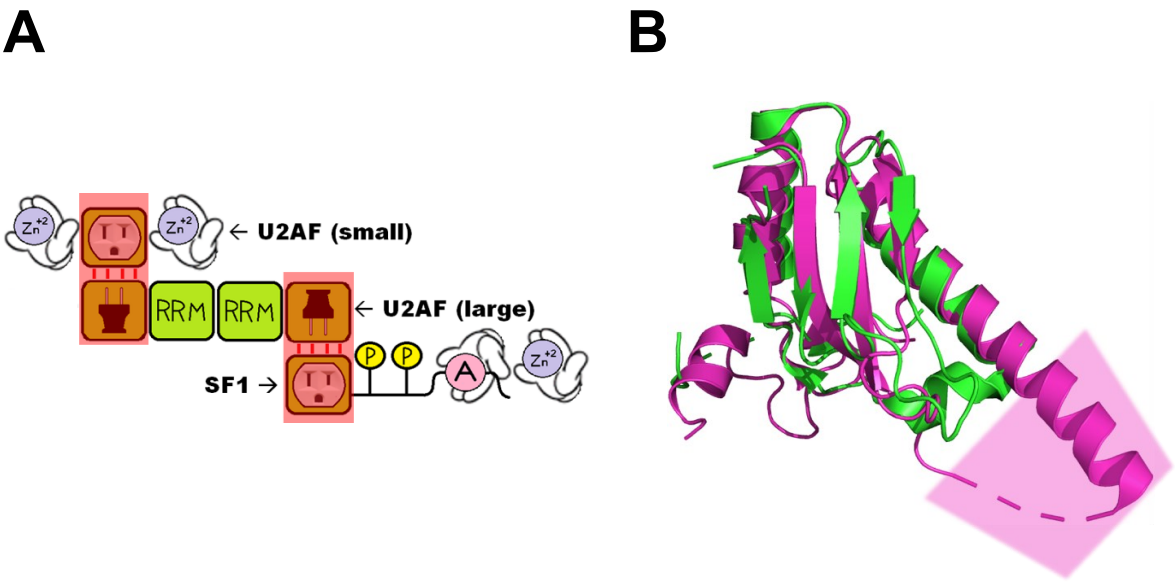


Figure 1-12. Comparison of the human ULM/UHM interface from both U2AF35/U2AF65 and U2AF65/SF1. PDB accession 1JMT (chains A, B) represents the U2AF35/U2AF65 interface and PDB accession 4FXW (chains A, B) represents the U2AF65/SF1 interface; regions in 4FXW outside of the ULM/UHM interface are omitted for clarity. (A) U2AF/SF1 with relevant region highlighted. (B) Superimposition of the human ULM/UHM interface from U2AF35/U2AF65 (magenta) and U2AF65/SF1 (green). The 30 residue α-helix in U2AF35 which is absent in U2AF65 in the counterpart structure is indicated by a magenta shadow.

C-terminal to the ULM in SF1, the phosphorylated domain also interacts with the UHM of U2AF-L and extends the interface of these two proteins via a mechanism which is regulated through phosphorylation. A multiple sequence alignment of SF1 orthologues shows that this phosphorylated domain is highly conserved (Fig. 1-20). Also, unlike most domains of U2AF-L or SF1, it does not have any known structural homologues and its function outside of phosphorylation is poorly understood. The phosphorylated domain of SF1 is a pair of α-helices that are separated by a flexible linker containing a conserved four residue 'SPSP' motif, in which the two serines are targeted for phosphorylation. It is well-established that protein phosphorylation is required for splicing and the SR protein superfamily represents the prototype for phosphorylation-dependent control of splicing (Ghosh & Adams, 2011; Stamm, 2008). Although the phosphorylation of non-SR splicing factors is less well-characterized, one important known function of non-SR phosphorylation is to regulate protein-protein interactions involving U2AF65 (Golling et al., 2002; Ruskin, Zamore, & Green, 1988). Phosphorylation of the SPSP motif of SF1 is a very important example since it is necessary *in vivo* and, additionally, the phosphorylated domain is essential for cooperative formation of the ternary U2AF65/SF1/3′ SS RNA complex (W. Wang et al., 2013). It has also been found that the SPSP motif is predominately in the phosphorylated state in proliferating human embryonic kidney cells, and subsequent phospho-proteome analyses of HeLa, lymphoma, and prostate cancer cells confirmed the prevalence of this phosphorylation suggesting that the phosphorylation state of the SPSP motif influences cancer initiation and/or progression (Beausoleil et al., 2004; Manceau et al., 2006; Myung & Sadar, 2012; Shu, Chen, Bi, Mumby, & Brekken, 2004).

With respect to the kinase responsible for phosphorylating SF1, KIS (kinase interacting with stathmin), also known as UHMK1 (U2AF homology motif kinase 1), specifically targets the

SPSP motif in humans (Manceau, Kielkopf, Sobel, & Maucuer, 2008; Manceau et al., 2006). KIS is an unusual serine-threonine kinase not belonging to any known kinase families. It consists of an N-terminal UHM (40% identical with the UHM of human U2AF65) and a C-terminal kinase domain. It is presumed that the UHM of KIS binds the ULM of SF1 similarly to U2AF65, thereby positioning the kinase domain to phosphorylate SF1. Phosphorylation is thought to be followed by disassociation of the kinase and association of U2AF65 with SF1 (Manceau et al., 2006; Maucuer, Le Caer, Manceau, & Sobel, 2000; Maucuer et al., 1997). This kinase has not been structurally characterized.

When the SPSP motif is phosphorylated, the flexible linker between the two α-helices of the phosphorylated domain assumes a well-defined structure. This event induces a disorder-to-order transition within a U2AF65/SF1 interface that exists outside of the canonical ULM/UHM interface thereby extending the interfacial surface area between these two proteins.

Out of the two concurrent studies, one also reported a panel of four SAXS structures that provide a more complete context to understand the biological significance of all the X-ray and NMR ensemble structures (W. Wang et al., 2013). Therefore, it is useful to begin with a discussion of these SAXS structures. A total of four SAXS envelopes were reported, all containing the nearly full-length U2AF65/SF1 dimer. The envelopes for the apo and RNA-bound forms of this dimer were described in both their unphosphorylated and phosphorylated states. The RNA used was AdML (adenovirus major late) pre-mRNA, a prototype 3′ SS RNA used frequently to study splicing.

When combined with the X-ray and NMR structures containing the phosphorylated domain of SF1, the SAXS envelopes demonstrate that folding of an arginine-rich loop around the phosphorylated serines amplifies into overall conformational changes in the nearly full-length

U2AF65/SF1/3′ SS RNA complex, causing it to assume a more compact configuration. This tightly bends in a fashion that is compatible with coupling of the 5′ SS and 3′ SS early in the splicing cycle in order to support the catalytic steps of splicing and assists in the recognition and binding of sub-optimal 3′ SS RNAs. The aforementioned SAXS envelopes as well as the model of a bent U2AF65/SF1/3′ SS RNA complex which proximates the 5′ SS and 3′ SS to support splicing catalysis late in the splicing cycle are shown below in Fig. 1-13.

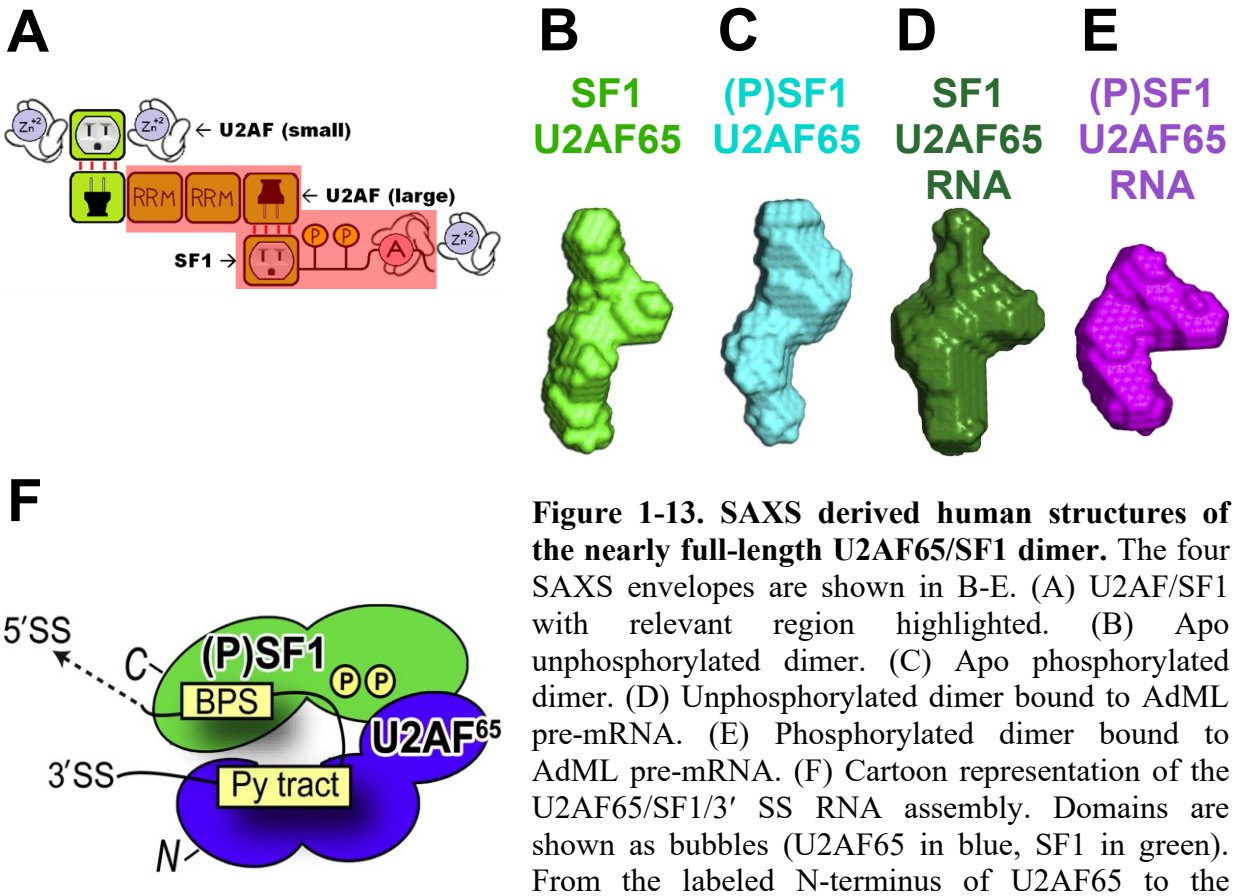


Figure 1-13. SAXS derived human structures of the nearly full-length U2AF65/SF1 dimer. The four SAXS envelopes are shown in B-E. (A) U2AF/SF1 with relevant region highlighted. (B) Apo unphosphorylated dimer. (C) Apo phosphorylated dimer. (D) Unphosphorylated dimer bound to AdML pre-mRNA. (E) Phosphorylated dimer bound to AdML pre-mRNA. (F) Cartoon representation of the U2AF65/SF1/3′ SS RNA assembly. Domains are shown as bubbles (U2AF65 in blue, SF1 in green). From the labeled N-terminus of U2AF65 to the labeled C-terminus of SF1, they are: RRM1, RRM2, UHM, ULM and phosphorylated domain, KH-QUA2 domain. Circled Ps represent phosphorylated sites on SF1 (adapted from Wang et al., 2013).

Now that an overview has been provided for the phosphorylated domain of SF1 and this domain has been placed in a more complete functional context by the SAXS envelopes, the atomic X-ray and NMR derived human structures of apo U2AF65/SF1 dimer & apo SF1 will be discussed. The structures containing the phosphorylated domain in its unphosphorylated state will be discussed in Section 1-9.4.1., and the structures containing the phosphorylated domain in its phosphorylated state will be discussed in Section 1-9.4.2.

1-9.4.1. Unphosphorylated X-ray and NMR derived human structures of apo U2AF65/SF1 dimer & apo SF1

Three structures contain the phosphorylated domain of SF1 in its unphosphorylated state. Two structures exist for the isolated phosphorylated domain in its unphosphorylated state; one is an X-ray structure and the other is an NMR structure ensemble. The third structure is an NMR structure ensemble of a dimer representing the UHM of U2AF65 bound to an SF1 construct containing both the ULM and the phosphorylated motif in its unphosphorylated state. A superimposition of SF1 from these three structures shows that SF1 consists of a helix-linker-helix configuration in all three structures, in which two rigid α-helices are held together in an anti-parallel arrangement and are connected by a flexible linker containing the SPSP motif. There is also some sequence both N-terminal and C-terminal to the helix-linker-helix which stabilizes the rigid α-helices. The three structures are very similar in these rigid areas but differ widely in the disordered inter-helix linker. Additionally, a comparison of the 10 conformers submitted for both NMR structure ensembles containing the phosphorylated domain in its unphosphorylated state also shows that this inter-helix linker is flexible. In the X-ray structure of the isolated phosphorylated domain in its unphosphorylated state there are 4 copies of SF1 in the

asymmetric unit which are close to identical and the linker is only partially modelled since residues immediately preceding the SPSP motif (P74-P81) show little electron density in the diffraction. Additionally, R93, R97, and R100, which form the arginine claw in the phosphorylated state (see Section 1-9.4.2.) do not participate in crystal packing contacts.

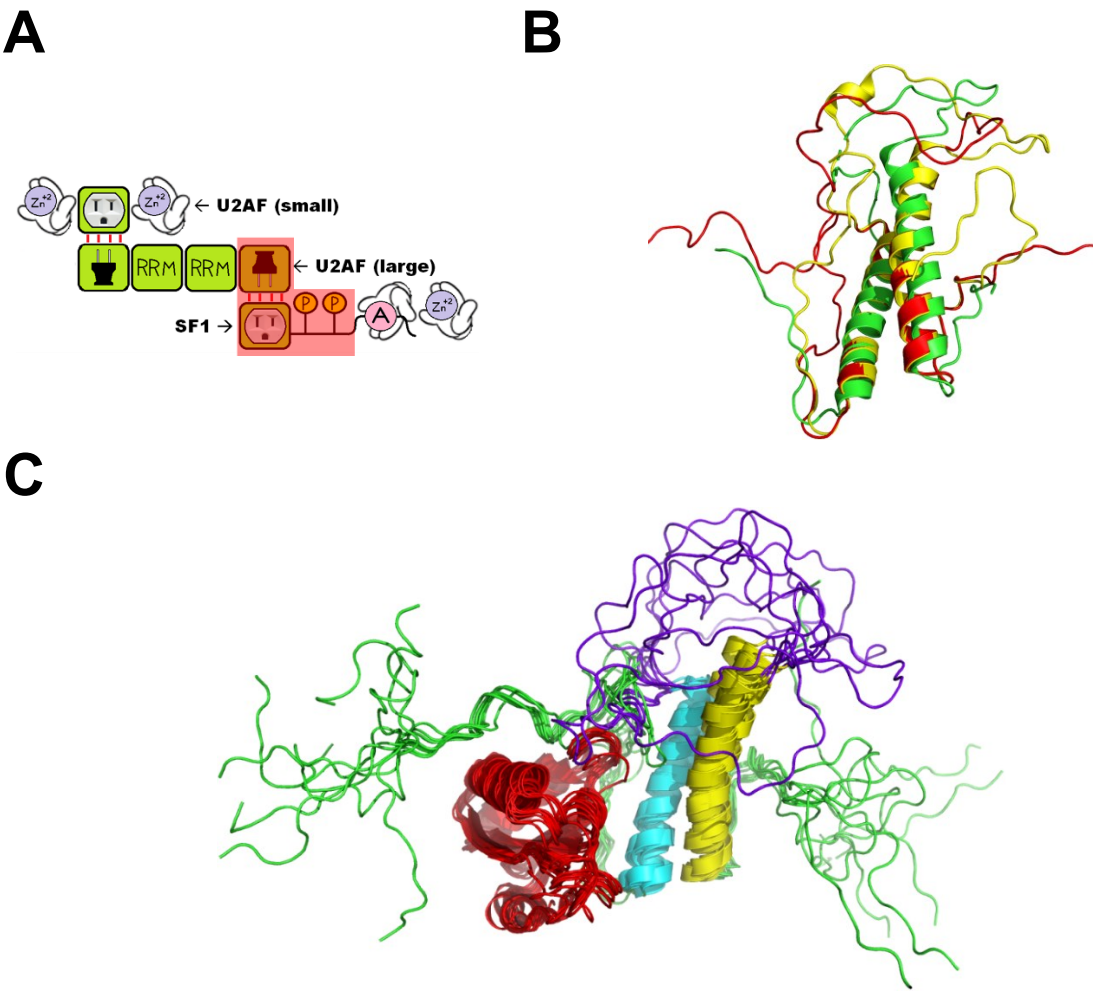


Figure 1-14. Human structures of apo U2AF65/SF1 dimer & apo SF1 in the unphosphorylated state. (A) U2AF/SF1 with relevant region highlighted. (B) The three SF1 structures containing the phosphorylated domain in its unphosphorylated state. The two NMR structures are both superimposed onto the X-ray structure. Structures are coloured as follows: PDB accession 2M09 (yellow: state 1), 2M0G (red: state 1; U2AF65 omitted for clarity), 4FXX (green: chain A). (C) Overall structure of the apo U2AF65/SF1 dimer structure ensemble (PDB accession 2M0G, 10 states submitted) with general features annotated: the UHM of U2AF65 is shown in red. SF1 is shown in green with the helix-linker-helix of the phosphorylated domain coloured as follows: N-terminal α-helix (helix α1: cyan), inter-helix linker (purple-blue), C-terminal α-helix (helix α2: yellow).

It is essential to discuss the phosphorylated domain in the context of the largest and most complete structure containing the phosphorylated domain in its unphosphorylated state (NMR; PDB code 2M0G) in order to understand the structural and functional roles of this domain within the context of the intact U2AF/SF1 assembly. Specifically, it is important to identify the network of protein interactions that stabilize the architecture of this domain and fix its positioning relative to other parts of the U2AF/SF1 complex (Fig. 1-15).

The position of helices α1 and α2 relative to each other is stabilized by hydrophobic residues every 3-4 residues in one helix contacting the other helix. Specifically, helix α1 residues A51, V54, I58, L61, and L65 contact helix α2 residues L105, L112, M116, and L119, respectively thereby stabilizing the arrangement of the two helices relative to each other. These interactions are shown in Fig. 1-15A.

In addition to the inter-helix contacts stabilizing the helix-hairpin-helix arrangement, sequence flanking the helix-linker-helix interacts with helices α1 and α2. Specifically, residues N-terminal of helix α1 (I40 and L44) form hydrophobic contacts with residues located within helix α1 (Y52, I53, and L56). These interactions are shown in Fig. 1-15B. Additionally, both helices α1 and α2 contact the sequence C-terminal to the helix-linker-helix. Residues C-terminal to helix α2 (F123, P126 and Y129) adopt an extended conformation and pack against both helix α1 (I58) and helix α2 (I113, and M116) via hydrophobic interactions and helix α2 (R109) potentially forms a salt bridge with the region C-terminal to helix α2 (D128). Contacts between helices α1/α2 to sequence C-terminal to the helix-linker-helix are shown in Fig. 1-15C.

As discussed previously, the phosphorylated domain of SF1 extends the U2AF65/SF1 interaction beyond the ULM/UHM interface. Specifically, a second hydrophobic interface is formed by the UHM of U2AF65 (M381, V458, and V460) interacting with SF1 in the region N-

terminal to helix α1 (V39, I40), and helix α1 itself (I53, L56). This interface also contains a potential salt bridge between the UHM of U2AF65 (K462) and helix α1 of SF1 (E49). The second interface has the effect of locking the orientation of the UHM of U2AF65 and SF1 relative to each other, and the contacts between the UHM of U2AF65 to SF1 in the region N-terminal to helix α1 and helix α1 itself are shown in Fig. 1-15D.

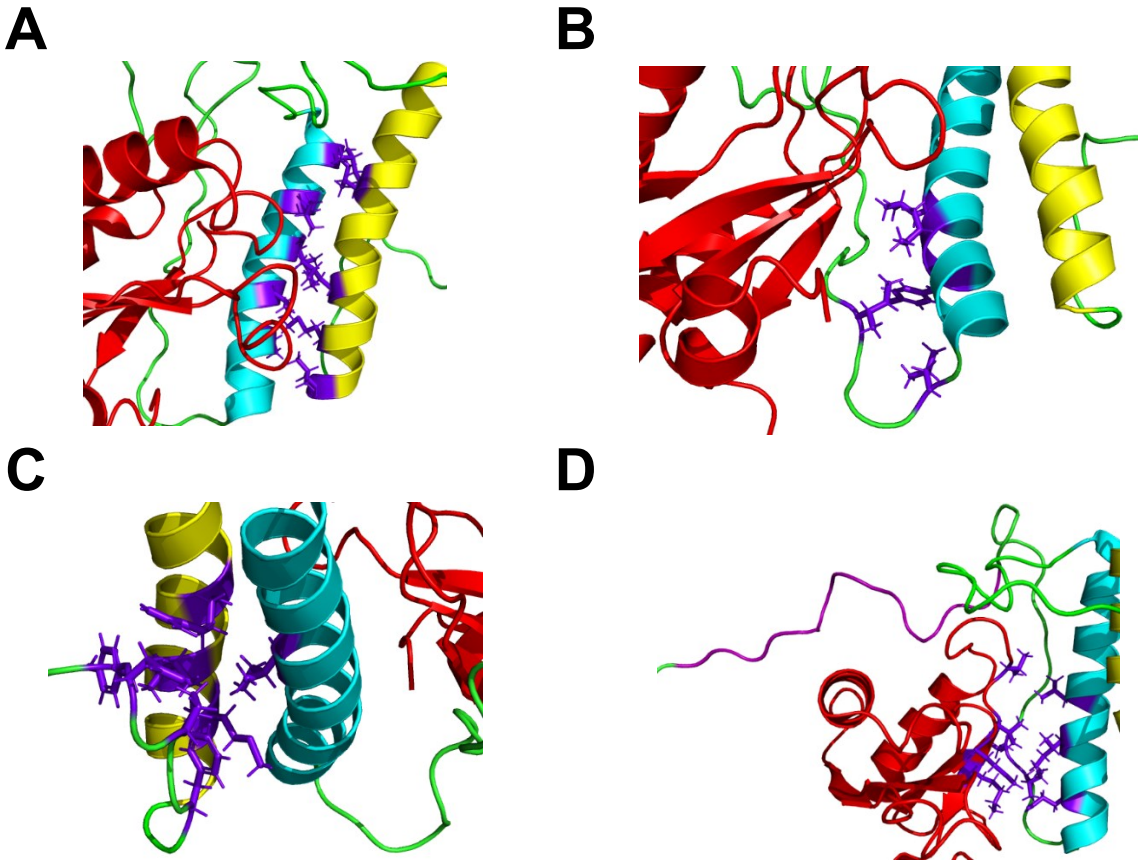


Figure 1-15. Architecture of the human phosphorylated domain of SF1 within the context of the apo U2AF65/SF1 dimer. PDB accession 2M0G is shown with general features annotated. The UHM of U2AF65 is shown in red. SF1 is shown in green with major features variously coloured: ULM (purple), helix α1 (cyan), helix α2 (yellow). Panels A-D illustrate various interfaces that stabilize the architecture and positioning of the phosphorylated domain of SF1 within the larger U2AF/SF1 complex: purple-blue sticks indicate residues forming the relevant interface. (A) Hydrophobic contacts between helix α1 and α2 are spaced every 3-4 residues, stabilizing the arrangement of the two helices. (B) Residues N-terminal of helix α1 form hydrophobic contacts with residues in helix α1. (C) Residues C-terminal of helix α2 adopt an extended conformation and pack against both helices α1 and α2 via hydrophobic interactions. Additionally, helix α2 potentially forms a salt bridge with the region C-terminal to helix α2. (D) In addition to the ULM/UHM interface, a second hydrophobic interface exists between the UHM of U2AF65 and SF1 (helix α1 and the region N-terminal to helix α1).

1-9.4.2. Phosphorylated X-ray derived human structure of apo U2AF65/SF1 dimer

There is only one X-ray structure of this complex which contains two nearly identical copies of this complex in the asymmetric unit (rmsd = 0.9 Å). A superimposition of this structure (PDB accession code 4FXW, chains A and B) and its unphosphorylated counterpart (PDB accession code 2M0G, state 1) shows close correspondence between these structures as well (rmsd = 3.2 Å, see Fig. 1-16B), indicating that although phosphorylation-induced structural changes are dramatic in the intact U2AF65/SF1 and U2AF65/SF1/3′ SS RNA complexes, they are subtle at a local level.

The focal point of the phosphorylated structure is that phosphorylation of the SPSP motif induces folding of an arginine-rich loop around the phosphorylated serines to form an arginine claw, which is a structural motif consisting of the phosphate moiety of one or more phosphoserines in the center chelated by the guanidinium groups of several arginines (Hamelberg, Shen, & McCammon, 2007). Importantly, it has been shown that the arginine-phosphate electrostatic interaction that stabilizes the arginine claw possesses unusually high stability (Woods & Ferre, 2005). The positively charged arginine claw encloses (P)S80 and (P)S82 of human SF1 and consists of three conserved arginines (R93, R97, and R100), located across from the phosphates. Additionally, positively charged side-chains of K104 and R79 are near the phosphates. The structure of the arginine claw of phosphorylated human SF1 is summarized in Fig. 1-16C.

Interestingly, there are no direct contacts between the UHM of U2AF65 and the arginine claw enclosing (P)S80 and (P)S82 suggesting that phosphorylation regulates the U2AF65/SF1/3′ SS RNA complex indirectly. In this ‘transduction’ model, formation of the arginine claw stabilizes the fold of the phosphorylated domain, and consistent with this model, phosphorylation

was found to significantly increase thermostability of the helix-linker-helix as determined by comparing the thermostability of the phosphorylated domain in both its unphosphorylated and phosphorylated states. This stabilization of the domain structure in turn itself induces a disorder-to-order transition within the U2AF65/SF1 interface thereby triggering overall conformational changes in the U2AF65/SF1/3′ SS RNA complex as discussed previously.

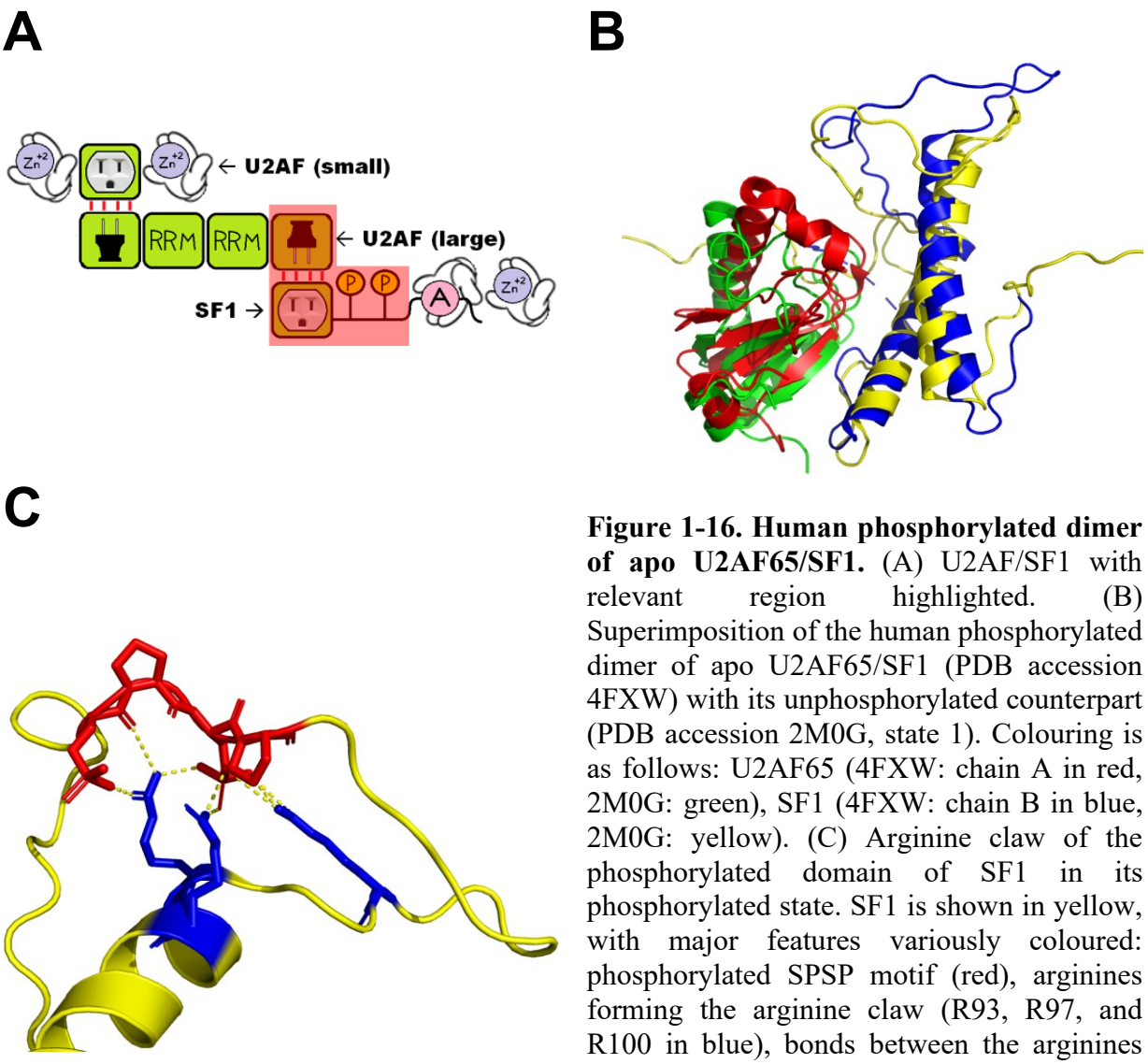


Figure 1-16. Human phosphorylated dimer of apo U2AF65/SF1. (A) U2AF/SF1 with relevant region highlighted. (B) Superimposition of the human phosphorylated dimer of apo U2AF65/SF1 (PDB accession 4FXW) with its unphosphorylated counterpart (PDB accession 2M0G, state 1). Colouring is as follows: U2AF65 (4FXW: chain A in red, 2M0G: green), SF1 (4FXW: chain B in blue, 2M0G: yellow). (C) Arginine claw of the phosphorylated domain of SF1 in its phosphorylated state. SF1 is shown in yellow, with major features variously coloured: phosphorylated SPSP motif (red), arginines forming the arginine claw (R93, R97, and R100 in blue), bonds between the arginines and SPSP motif (dashed lines).

1-9.5. RNA-bound SF1

Three structures have been reported of the KH-QUA2 domain bound in a 1:1 complex with the RNA sequence UAUACUAACAA, which contains the optimal BPS. Two are X-ray structures from *S. cerevisiae* Bbp and represent two different spacegroups of an identical protein/RNA complex; the lower resolution structure is derived from selenomethionine-substituted protein crystallized in spacegroup P422 and has one copy of the protein/RNA complex in the asymmetric unit. This structure was subsequently used for molecular replacement to derive a higher-resolution structure containing native protein, which crystallized in spacegroup $P2_1$ and has four copies of the protein/RNA complex in the asymmetric unit (Jacewicz et al., 2015). The native structure is very similar to the selenomethionine-substituted structure in all 4 copies of the asymmetric unit yielding rmsd = 0.4-0.5 Å when different copies of the native complex in the asymmetric unit are compared with the selenomethionine-substituted complex. Finally, the third structure is a human derived NMR structure ensemble, and is the oldest structure reported (Z. Liu et al., 2001). When superimposed, all 10 states submitted for the ensemble are very similar to each other, showing the most variation in the variable loop and in the UA appended 5′ terminal to the UACUAAC sequence (see Fig. 1-17B). The three structures discussed above are summarized below in Table 1-12.

Table 1-12: X-ray and NMR derived structures of RNA-bound SF1

Structure	PDB accession code	Solution method	PDB reported resolution/conformers submitted	Species
Bbp (selenomethionine-substituted) + RNA	4WAL	X-ray	2.2 Å	*S. cerevisiae*
Bbp (native) + RNA	4WAN	X-ray	1.8 Å	*S. cerevisiae*
SF1 + RNA	1K1G	NMR	10 conformers	*H. sapiens*

The structures introduced above are significant because the KH-QUA2 domain is necessary and sufficient for BPS binding and therefore these structures reveal the basic mechanism whereby SF1 recognizes the BPS. When the human structure (state 1) and selenomethionine-substituted *S. cerevisiae* structure are superimposed, they conform to the same overall configuration (rmsd = 3.3 Å) but show significant differences in the variable loop (discussed below). The discussion that follows is restricted to the human structure, since it is the most relevant model to understand the biology of humans and other multicellular, developmentally complex eukaryotes. This is because *S. cerevisiae* is a very divergent yeast species with a nearly invariant UACUAAC BPS, whereas the human BPS is much more degenerate (yUnAy), although UACUAAC is the preferred sequence (Berglund et al., 1997; Gao et al., 2008; Zhuang, Goldstein, & Weiner, 1989).

SF1 belongs to the STAR (signal transduction and activation of RNA) family of proteins; the STAR motif is composed of a KH (K homology) domain followed by a conserved C-terminal QUA2 auxiliary motif and may also contain a QUA1 motif in addition to these two components. The KH domain was first identified in hnRNP K and is a sequence-specific domain that binds to single-stranded nucleic acid. The first two structures reported of a KH domain were the NMR structure ensembles of the N-terminal KH domain of FMR1 (fragile X mental retardation 1) and the C-terminal KH domain of hnRNP K and revealed a β-α-α-β-β-α structure (Baber, Libutti, Levens, & Tjandra, 1999; Musco et al., 1997; Vernet & Artzt, 1997). The KH domain of human SF1 is similar to other KH domains; the KH fold is stabilized by conserved hydrophobic residues and a comparison with the KH domains from Nova2 (neuro-oncological ventral antigen 2) and Vigilin yields rmsd = 2.8 Å (H. A. Lewis et al., 2000; Musco et al., 1996).

The KH domain of SF1 is distinct from other KH domains because it requires the involvement of the QUA2 motif in order to specifically recognize and bind its target sequence; conserved residues in the QUA2 motif recognize the ACU trinucleotide sequence at the 5′ end of the BPS (see Fig. 1-17C). Therefore, the molecular basis of BPS recognition is a unique, enlarged KH fold defined by the combined KH-QUA2 domain which recognizes the BPS via numerous STAR-specific amino acids and this model of BPS recognition is consistent with mutations that affect RNA binding (Berglund et al., 1997; Rain et al., 1998). Finally, secondary chemical shifts indicate that the secondary structure of the KH-QUA2 domain is pre-formed and is not induced by RNA binding.

When the human structure was published, the Nova2-KH3 domain bound to a stem-loop RNA was the only other atomic-resolution structure revealing RNA recognition by a KH domain. It is important to note that in both structures, the protein/RNA interface is very hydrophobic but aromatic residues are not used for RNA recognition because of the low abundance of aromatic residues and conservation of aliphatic residues in KH domains. This is likely a general principle of RNA recognition by KH domains (H. A. Lewis et al., 2000).

The 3′ part of the BPS (UA<u>A</u>C) is specifically recognized in a hydrophobic cleft formed by the GPRG (Gly-Pro-Arg-Gly) motif and the variable loop of the KH domain which are both conserved elements among this family of proteins; the branch A is deeply buried within this cleft (see Fig. 1-17C). The residues that mediate the BPS interaction are conserved in STAR-specific residues. For example, K184 in human SF1 contacts the adenosine which is in the n-1 position to the branch A and this is conserved at the beginning of a STAR-specific extension of the variable loop. A K184A mutation abolishes BPS binding and the corresponding R185C mutation in the STAR protein How/Who induces an embryonic lethal phenotype in *D. melanogaster* (Baehrecke,

1997). The conservation and importance of these STAR-specific residues points to a conserved mode of RNA recognition by single-stranded STAR proteins.

It is striking to note that the branch A is deeply buried in the hydrophobic pocket of the KH domain; additionally, the base of the branch A is oriented towards the protein and stabilized by N6 and N1 forming two hydrogen bonds with the backbone amide and carbonyl oxygen of I177, mimicking the Watson-Crick functional groups of a uridine and shielding the branch A from interacting with other molecules. Another hydrogen bond is found between the 2′ hydroxyl of the neighbouring adenosine and the N7 of the branch A. This configuration of hydrogen bonds uniquely specifies an adenosine base at this position. The structure of Nova2-KH3 bound to a stem-loop RNA also contains this stabilization of an adenosine via Watson-Crick recognition by two hydrogen bonds with the peptide backbone mediated by equivalent positions in the β2 strand of both KH domain structures (H. A. Lewis et al., 2000). This parallel is likely a generalizable principle of adenosine recognition by varied KH domains because other KH domains also bind 3-4 nucleotide single-stranded RNA elements, most of which contain adenosine, and the STAR protein Sam68 loses its ability to bind the RNA sequence element UAAA if the adenosine in the 3rd position is mutated (H. A. Lewis et al., 1999; Q. Lin, Taylor, & Shalloway, 1997).

Overall, the BPS adopts an extended single-stranded conformation and is bound in a hydrophobic groove between QUA2, the GPRG loop, and the variable loop of the KH domain and provides clues regarding the involvement of SF1 in spliceosome complex A formation. The KH-QUA2 region and zinc knuckle of SF1 have both been suggested to bind the BPS, however NMR titrations show that the zinc knuckle does not contact the BPS and may instead interact with nucleotides located upstream of the BPS (Berglund, Abovich, et al., 1998; Berglund, Fleming, et al., 1998; Rain et al., 1998). Binding of U2AF65 to the PPT directs its RS domain to

contact the BPS and this structure is consistent with this model, suggesting that the positively charged RS domain could interact with the solvent-exposed, negatively charged phosphate backbone of the BPS (Valcarcel, Gaur, Singh, & Green, 1996).

The RNA-bound SF1 structure supports the formation of the BPS/U2 snRNA duplex in spliceosomal complex A, where the branch A is bulged out (Berglund, Rosbash, & Schultz, 2001; Query et al., 1994). The complementary U2 snRNA is able to approach the SF1/BPS complex from its accessible, open face. The burial of the branch A would help constrain and exclude it from the duplex in a 'pre-bulged' complex, and disruption of the SF1/BPS complex would trigger the transition to the BPS/U2 snRNA duplex (Guth & Valcarcel, 2000; Rutz & Seraphin, 1999).

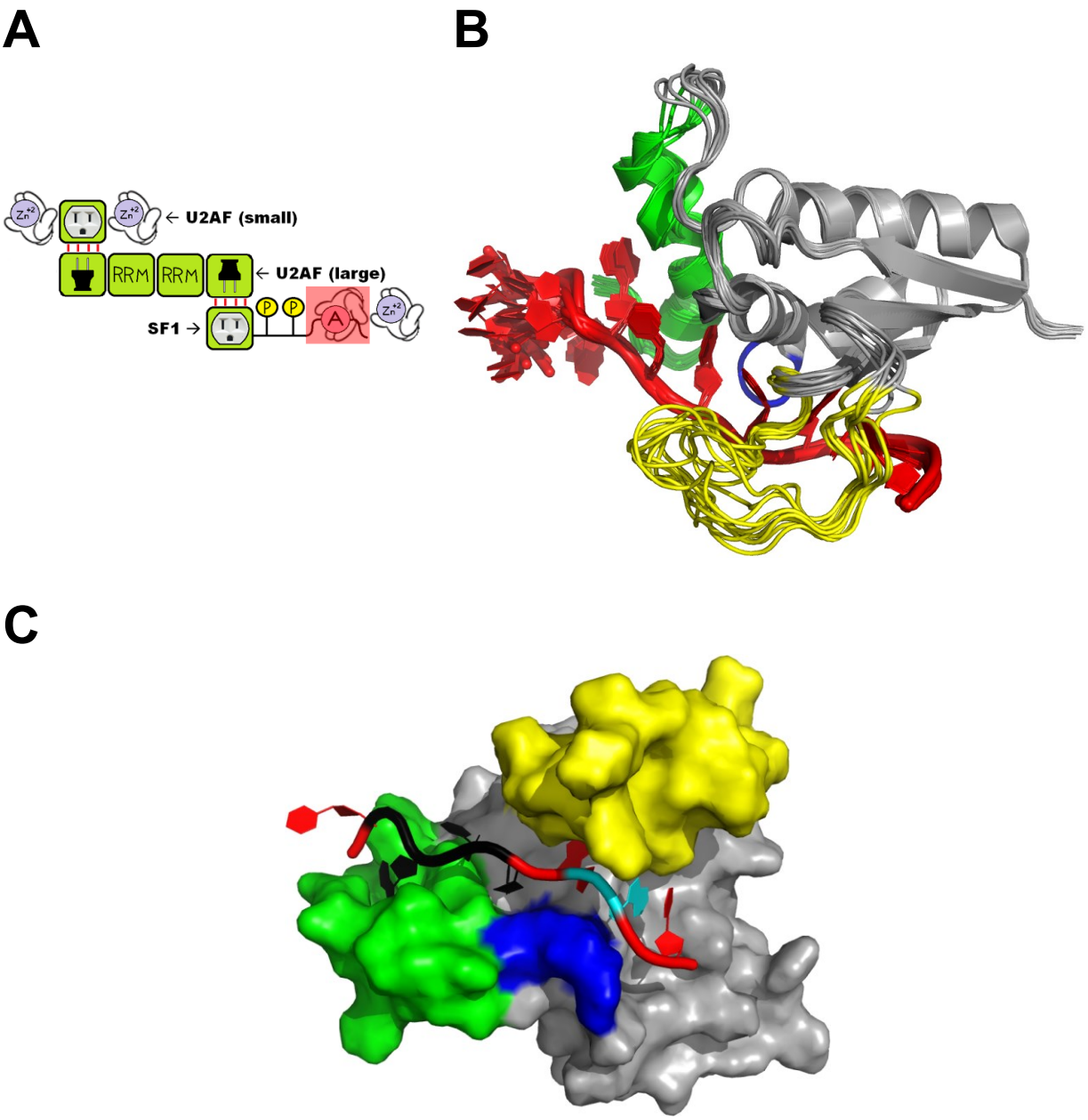


Figure 1-17. NMR structure ensemble of the KH-QUA2 domain of human SF1 bound to optimal BPS RNA. (A) U2AF/SF1 with relevant region highlighted. (B) Overall structure of the ensemble states with general features annotated. Optimal BPS RNA is shown in red. The KH domain is shown in grey, with major features variously coloured: GPRG motif (blue), variable loop (yellow). The QUA2 motif is shown in green. (C) Surface representation of (B); ensemble state 1 is shown with additional features variously coloured: ACU trinucleotide near the 5′ end (black), branch A (cyan). Terminal nucleotides outside the optimal BPS are omitted for clarity.

1-9.6. Miscellaneous structures

Three X-ray structures have been reported which contain U2AF/SF1 proteins, but which do not represent sub-assemblies of the U2AF/SF1 complex and relate to functions of these proteins outside of constitutive 3′ SS recognition. Two murine structures have been reported of the interface between the ULM of U2AF65 and UHM of CAPERα, but only one has been published (Stepanyuk et al., 2016). One *S. cerevisiae* structure has been published of an interface between Bbp and Smy2 (suppressor of myo2-66). This structure consists of a short peptide from the proline-rich domain of Bbp and the GYF (glycine-tyrosine-phenylalanine) domain of Smy2. These structures are summarized below in Table 1-13.

Table 1-13: Miscellaneous structures

Structure	PDB accession code	PDB reported resolution	Species
U2AF65 (ULM) + CAPERα (UHM)	5CXT	2.2 Å	*M. musculus*
U2AF65 (ULM) + CAPERα (UHM)	4RU2	2.2 Å	*M. musculus*
Bbp (proline-rich peptide) + Smy2 (GYF domain)	3FMA	2.5 Å	*S. cerevisiae*

1-9.6.1. Murine U2AF65/CAPERα structures

These structures are of an interface formed from the ULM of U2AF65 and the C-terminal UHM of CAPERα, so the general features are similar to the ULM/UHM structures discussed previously. Both of the two structures contain 9 copies of the dimer in the asymmetric unit and are nearly identical to each other (rmsd = 0.1 Å).

1-9.6.2. *S. cerevisiae* Bbp/Smy2 structure

The *S. cerevisiae* Bbp/Smy2 structure consists of a peptide from the proline-rich domain of Bbp bound to the GYF domain protein Smy2, which is involved in COPII (coat protein complex II) vesicle formation; COPII is a vesicle coat protein that transports proteins from the rough endoplasmic reticulum to the Golgi apparatus (Higashio, Sato, & Nakano, 2008). This is consistent with the finding that *S. cerevisiae* Bbp has roles in both pre-mRNA splicing as well as the nuclear export of mRNA (Abovich & Rosbash, 1997; Rutz & Seraphin, 2000). This structure has not been published.

GYF domain proteins are adaptor proteins named after a conserved signature motif, which is part of a larger amino acid signature that is small, versatile and responsible for interacting with proline-rich peptides. A BLAST (basic local alignment search tool) search reveals that these proteins are present in most eukaryotic species sequenced thus far, but interestingly, they have not undergone significant amplification in metazoans during evolution (Altschul, Gish, Miller, Myers, & Lipman, 1990; M. M. Kofler & Freund, 2006). Splicing and splicing-associated processes are recurring functional themes for GYF domains and GYF domain-containing proteins may mediate mRNA export from the nucleus or couple splicing with transcription/translation (M. M. Kofler & Freund, 2006).

There are two major subfamilies of GYF domains, named after the proteins in which they were first identified: human CD2BP2 (CD2 binding protein 2) and *S. cerevisiae* Smy2 protein (Freund, Dotsch, Nishizawa, Reinherz, & Wagner, 1999; Lillie & Brown, 1992; Nishizawa, Freund, Li, Wagner, & Reinherz, 1998). In eukaryotes, these two subfamilies seem to be correlated with different biological functions. CD2BP2-type GYF domains are nuclear and are involved in pre-mRNA splicing, whereas Smy2-type GYF domains are found in the cytoplasm

and its compartments and seem to be involved in translational control (Bialkowska & Kurlandzka, 2002; Giovannone et al., 2003; Huh et al., 2003; M. Kofler, Heuer, Zech, & Freund, 2004; M. Kofler, Motzny, Beyermann, & Freund, 2005; Laggerbauer et al., 2005; T. K. Nielsen, Liu, Luhrmann, & Ficner, 2007). CD2BP2 is another name for the U5 snRNP protein called U5-52K (Laggerbauer et al., 2005). Paralleling *S. cerevisiae* Bbp, evidence exists to suggest that human SF1 interacts with U5-52K (Arning et al., 1996; M. Kofler et al., 2004; M. M. Kofler & Freund, 2006; Kramer, 1992).

1-10. Project introduction and thesis overview

1-10.1. Intact *S. pombe* U2AF/SF1 complex as a model for biochemical and structural analysis

The U2AF/SF1 complex is a modular, multi-domain protein complex that binds the BPS, PPT, and yAG/G sequence motifs which, together, define the 3′ SS. Although many structures of both the apo and substrate-bound states of this complex exist, they only represent small fractions of the complete complex. Therefore, no coherent model of 3′ SS selection exists yet. The reductionist approach used up until now to study U2AF/SF1 cannot provide a satisfactory model of how the intact U2AF/SF1 assembly operates within the spliceosome cycle, because its functions are governed by higher-order principles, whereby the sum whole of the U2AF/SF1 complex and U2AF/SF1/3′ SS complexes is greater than the sum of its parts. This has already been demonstrated to be true in the current structures. For instance, the inter-RRM linker of U2AF65 directly contacts the RRMs, the PPT, and actively regulates the RNA-binding activity of U2AF65. Another example is the phosphorylated U2AF65/SF1 complex, which operates as a

molecular amplifier in order to magnify subtle phosphorylation-induced conformational changes into a global rearrangement of the complex. The cooperative nature of the domains comprising these sub-assemblies would never have been revealed by solving them as isolated structures. Because several examples already exist to demonstrate that higher-order principles govern the U2AF/SF1 and U2AF/SF1/3′ SS complexes, it is likely that additional examples exist which remain undiscovered and will only be revealed by solving the entire U2AF/SF1 and U2AF/SF1/3′ SS complexes as an intact entity. Realizing this goal is expected to resolve the unanswered questions regarding how U2AF/SF1 operates such as how U2AF-L accommodates such a wide variation in length and information content in the PPT.

Eventually, it will be necessary to establish a model system to express the human U2AF/SF1 complex in order to untangle the alternative splicing roles of these proteins and their variants. However, the *S. pombe* orthologues are an ideal starting model system for structure determination for reasons described in detail in Section 1-6. A comparison of the *S. pombe* and human orthologues of U2AF-S, U2AF-L, and SF1 via Clustal Omega alignment is provided below in Sections 1-10.2. to 1-10.4. (Madeira et al., 2019; Sievers et al., 2011). Key features in these alignments are annotated and described within the corresponding sections.

1-10.2. Comparison of U2AF-S in *S. pombe* and humans

Figure 1-18. U2AF-S alignment (human vs. *S. pombe*). GenBank accession numbers used in the alignment are M96982 (human full-length U2AF35a) and U48234 (*S. pombe* full-length U2AF23). In this figure, the central UHM is annotated in blue and is based on the human U2AF35/U2AF65 dimer structure (Kielkopf et al., 2001). The residues of the two ZFs that co-ordinate the zinc ion are annotated in red. The RS dipeptides and the glycine tract of U2AF35 are annotated in orange. The C-terminal α-helix of *S. pombe* U2AF23 is annotated in green in the alignment and is based on the apo *S. pombe* U2AF23/U2AF59 structure (H. Yoshida et al., 2015).

In the alignment in Fig. 1-18, the full-length sequences for the human and *S. pombe* orthologues have been aligned because all experiments in this thesis use either a full-length, untagged, wildtype construct of U2AF23 or a similar, slightly shorter construct ending at E194. The sequences are ~55% identical and ~81% similar over the aligned region.

U2AF23 has more potential for success with structural work because human U2AF35 consists of an RS domain at the C-terminus, which is split in two by 12 glycines, likely making this region of the protein very flexible. Additionally, this RS domain has an important role in the interaction with other proteins and/or RNA factors and in higher vertebrates, it is likely to strengthen the interaction of the two U2AF subunits with the assistance of other factors potentially complicating structural work using this orthologue as a model (Kellenberger, Stier, & Sattler, 2002; M. Zhang et al., 1992). In its place, *S. pombe* U2AF23 has an α-helix that interfaces with an α-helix corresponding to *S. pombe* U2AF59 (R107-L119), thereby forming a dimerization scaffold that extends the interface of the two proteins (H. Yoshida et al., 2015). During the incipient stages of this project, this interfacial α-helix in U2AF23 was predicted to exist from the amino acid sequence using the PsiPred secondary structure prediction server, and its presence was confirmed when the apo U2AF23/U2AF59 X-ray structure was reported (Buchan et al., 2013; H. Yoshida et al., 2015).

1-10.3. Comparison of U2AF-L in *S. pombe* and humans

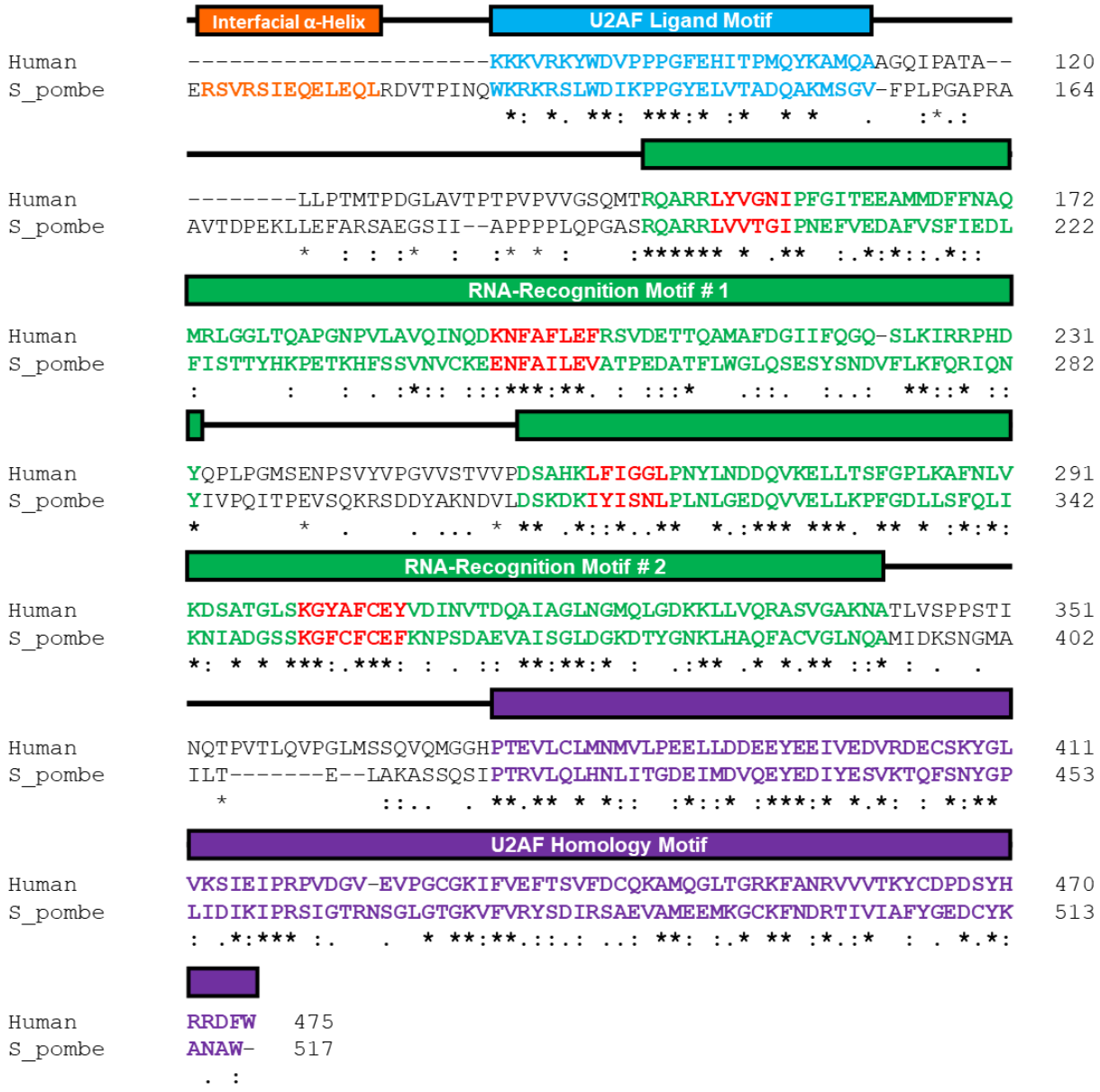


Figure 1-19. U2AF-L alignment (human vs. *S. pombe*). GenBank accession numbers used in the alignment are X64044 (human U2AF65, long isoform, K85-W475) and NP_595396 (*S. pombe* U2AF59, E106-W517). The N-terminal α-helix of *S. pombe* U2AF59 is annotated in orange and is based on the apo *S. pombe* U2AF23/U2AF59 dimer structure (H. Yoshida et al., 2015). The ULM is annotated in blue and is based on the human U2AF35/U2AF65 dimer structure (Kielkopf et al., 2001). The two RRMs are annotated in green and are based on the annotated alignment of U2AF-L orthologues provided in Selenko *et al.*, 2003. The RNP consensus motifs are annotated in red and are based on the annotated alignment of U2AF-L orthologues provided in Sickmier *et al.*, 2006. The UHM is annotated in purple and is based on the annotated alignment of U2AF-L orthologues provided in Selenko *et al.*, 2003.

The alignment in Fig. 1-19 is based on the *S. pombe* U2AF59 construct used in this thesis and omits the N-terminal RS domain of both proteins for clarity while retaining the remainder of the sequence. The RS domain has been omitted because it is poorly conserved, poorly characterized, and predicted to be unstructured. The sequences are ~31% identical and ~66% similar over the aligned region.

Protein organization is identical for both orthologues in the conserved RNA-binding core, however U2AF59 contains an α-helix between the RS domain and N-terminus of the core which interfaces with an α-helix corresponding to *S. pombe* U2AF23 (S172-A192), thereby forming a dimerization scaffold that extends the interface of the two proteins; this feature is absent in human U2AF65 (Kielkopf et al., 2001; H. Yoshida et al., 2015).

1-10.4. Comparison of SF1 in *S. pombe* and humans

Figure 1-20. SF1 alignment (human vs. *S. pombe*). GenBank accession numbers used in the alignment are NP_004621.2 (human SF1, isoform # 1, D11-P324) and AAF02214 (*S. pombe* SF1, M52-S378). The ULM is annotated in green. The phosphorylated domain is annotated in blue, and the SPSP motif within the phosphorylated domain is annotated in red. The KH-QUA2 domain is annotated in purple. The residues of the two zinc knuckles that co-ordinate the zinc ion are annotated in orange. The annotations encompassing the ULM to the conserved C-terminal zinc knuckle are based on the annotated domain structure of human SF1 provided in Wang *et al.*, 2013. The conserved α-helix predicted to flank the C-terminus of the zinc knuckle(s) by the PsiPred secondary structure prediction server is annotated in pink.

The alignment in Fig. 1-20 is based on the longest *S. pombe* SF1 construct used in this thesis and omits the region N-terminal to the ULM, as well as the proline-rich domain C-terminal to the zinc knuckle(s) in both proteins for clarity as they are poorly conserved, uncharacterized, and predicted to be unstructured. The sequences are ~44% identical and ~76% similar over the aligned region.

Protein organization is almost identical for both orthologues in the conserved RNA-binding core. However, *S. pombe* SF1 contains two zinc knuckles whereas human SF1 only contains one.

1-10.5. Expression system for downstream applications

An experimental system was successfully developed and used to generate multiple variants of both the U2AF dimer and U2AF/SF1 trimer complexes, which relies on the co-expression and co-purification of all constituent proteins from *E. coli*. This system generates milligram quantities of monodisperse protein complexes with no contamination detectable by SDS-PAGE, which are soluble and stable for at least several weeks under the appropriate storage conditions.

Development and optimization of the experimental system was not trivial. Pilot experiments which were at least partially successful were pursued and proceeded on a productive path towards eventually realizing the final reliable, efficient, optimized expression system. This system was used to obtain variants of *S. pombe* U2AF dimer and *S. pombe* U2AF/SF1 trimer.

The complexes purified using this system form the basis for all biochemical and structural investigations in this thesis. This system is described in detail in Chapter 2 and Appendix *I*.

1-10.6. Thesis overview

Chapter 1 is an introduction to this thesis and covers the current state of the splicing field in order to create context for the U2AF/SF1 sub-field. This is followed by a review of the current state of the U2AF/SF1 sub-field in order to create context for the thesis project itself. The general structure of the remainder of this thesis is provided below.

1-10.6.1. Chapter 2

Chapter 2 addresses the cloning, expression, purification, and biochemical characterization of the U2AF dimer and U2AF/SF1 trimer. All experiments described in this chapter are based on five complexes, which are catalogued in Appendix *I*, Table *I*-4. For simplicity and clarity, these complexes are referred to by abbreviated names which are also catalogued in Appendix *I*, Table *I*-4.

The biochemical characterization of these complexes first independently confirms the existence of a stable U2AF dimer and U2AF/SF1 trimer from first principles using SEC-MALLS (size exclusion chromatography + multi-angle laser light scattering), then confirms the existence of a stable RNA-bound U2AF/SF1 complex by binding the protein to a model 3′ SS RNA and evaluating the resulting complex using spectroscopy. Following this preliminary evaluation of the experimental system, EMSAs (electrophoretic mobility shift assays) were used to probe and dissect the RNA-binding properties of U2AF23, U2AF59, and SF1.

1-10.6.2. Chapter 3

Chapter 3 addresses the structural characterization of the U2AF dimer and U2AF/SF1 trimer. This chapter consists of two sets of experiments. The first of these is the use of SAXS to

generate molecular envelopes of both the *S. pombe* U2AF dimer and U2AF/SF1 trimer in their apo and RNA-bound forms, using one U2AF dimer and one U2AF/SF1 trimer complex specifically designed for this purpose.

The second set of experiments uses a more advanced SAXS technique that combines SEC-SAXS (size exclusion chromatography + small angle X-ray scattering) with rigid body modelling in order to generate ensembles of structural models that are restrained to be consistent with existing PDB structures. The SEC-SAXS experiments use a modified U2AF/SF1 trimer in which the U2AF-S and SF1 constructs are derived from *S. pombe*, and the U2AF-L construct is a *S. pombe*/human chimera in which the RRMs and inter-RRM linker of U2AF59 are replaced by those of human U2AF65 in order to permit accurate rigid body modelling using existing PDB accessions.

The cloning, expression, and purification of the three complexes used in Chapter 3 is not described in Chapter 2. However, the protocols described in Chapter 2 are transferable to these three complexes. This is because all complexes used in this thesis can be categorized into three groups based on purification behaviour, which are catalogued in Appendix *I*, Table *I*-6. Complexes belonging to the same group are purified using the same protocols and are indistinguishable from one another during purification, with the caveat that U2AF23 (M1-E194) migrates slightly faster than full length U2AF23 on an SDS-PAGE gel.

1-10.6.3. Chapter 4

This is the concluding chapter of this thesis. The insights gained in this thesis are summarized and their significance within the larger U2AF/SF1 and splicing fields are discussed. These insights are critical to developing practical ideas for future experimentation for both

biochemical and structural characterization of the eukaryotic 3′ SS recognition machinery. Therefore, several avenues are presented to further develop and experimentally investigate these themes in the future.

1-10.6.4. Appendices

Appendix *I* covers the design principles and development of the U2AF dimer and U2AF/SF1 trimer expression system in detail. Section *I*-3. of this appendix catalogues and describes the protein constructs and protein complexes used in this thesis along with a description of the experiments the various protein complexes were used for.

Appendix *II* is a catalogue of all the commercially obtained synthetic oligonucleotides used in this thesis. Section *II*-1. of this appendix catalogues and describes the U2AF dimer and U2AF/SF1 trimer cloning primers as well as the PCRs used to create U2AF-L and SF1 constructs. Although cloning of these constructs has been described in Chapter 2, the PCRs used to create the various U2AF-L and SF1 constructs are complicated and difficult to visualize. Therefore, they have been catalogued in Section *II*-1. This information has been kept separate from Chapter 2 for clarity. Section *II*-2. of this appendix catalogues and describes the oligonucleotides used to characterize U2AF dimer and U2AF/SF1 trimer complexes, which consist of the synthetic DNA oligonucleotides used as transcription templates for the 3′ SS model RNAs used in Chapter 2 as well as the RNA oligonucleotides used for both the biochemical and structural characterization of the U2AF dimer and U2AF/SF1 trimer in this thesis.

Appendix *III* describes the chimeric U2AF/SF1 trimer complex developed specifically for SEC-SAXS experiments described in Chapter 3. This information has been kept separate

from the main body of the thesis in order to maintain its continuity and logical flow due of the complicated nature of this construct and the design principles required to create it.

Appendix *IV* is a catalogue of the rigid body definitions used to build the SEC-SAXS based U2AF/SF1 model libraries described in Chapter 3.

Appendix *V* addresses the structural characterization of the SF3B component, p14. This appendix begins with an introduction covering what is currently known regarding p14 within the context of the SF3B particle, including existing human structures of the p14/SF3B155 complex in the PDB; SF3B155 (SF3B 155 kDa subunit) is the main scaffold protein of the SF3B particle and is present in all human X-ray structures because p14 in isolation is partially unfolded and insoluble and must be complexed with SF3B155 to be fully folded and well-behaved enough for characterization by X-ray crystallography. This is followed by a results section describing the crystals, X-ray data, and X-ray structures of the *S. pombe* and *C. albicans* p14/SF3B155 complexes and a discussion of the yeast X-ray structures. Appendix *V* concludes with a materials and methods section.

Chapter 2[1]

Cloning, expression, purification, and biochemical characterization of the U2AF dimer and U2AF/SF1 trimer

[1] The work presented in Chapter 2 was completed as a collaboration with Karolin Duft (Master of Science, Dept. of Biochemistry and Biology, University of Potsdam) (Duft, 2014).

2-1. Introduction

2-1.1. Cloning and co-expression of U2AF dimer and U2AF/SF1 trimer complexes

As outlined previously, the experimental system used to generate variants of both the U2AF dimer and U2AF/SF1 trimer complexes relies on the co-expression and co-purification of all constituent proteins from an *E. coli* host. Different constructs of U2AF23, U2AF-L, and SF1 were co-expressed in various combinations to generate a panel of U2AF dimer and U2AF/SF1 trimer complexes. The protocols for cloning all of the protein constructs and co-expressing all of the U2AF dimer and U2AF/SF1 trimer complexes used in this thesis are described in this chapter.

2-1.2. Purification of *S. pombe* U2AF dimer and U2AF/SF1 trimer complexes

U2AF dimer and U2AF/SF1 trimer complexes were initially purified from crude cell lysate using nickel affinity chromatography purification; the hexahistidine affinity tag was removed via TEV cleavage. Protein destined for biochemical characterization underwent a final purification using SEC (size exclusion chromatography). Protein destined for structural characterization was purified using the steps above but underwent an additional anion exchange chromatography purification step. Samples were analyzed after each purification step using SDS-PAGE and coomassie blue staining.

A total of eight different U2AF dimer and U2AF/SF1 trimer variants were used in this thesis (see Appendix *I*, Section *I*-3.4.). However, these complexes can be categorized into three groups based on purification behaviour; complexes belonging to the same group are purified using the same protocols and are indistinguishable from one another during purification, with the

caveat that U2AF23 (M1-E194) migrates slightly faster than full-length U2AF23 on an SDS-PAGE gel. The biochemical characterization of *S. pombe* U2AF dimer and U2AF/SF1 trimer described in this chapter was completed using a panel of five complexes (see Appendix *I*, Section *I*-3.4., Table *I*-4.). For clarity and simplicity and to avoid redundancy, the purification results of only these five complexes has been described.

2-1.3. Biochemical characterization of *S. pombe* U2AF dimer and U2AF/SF1 trimer complexes

Prior to biochemical characterization, purified complexes were characterized according to their physical properties. SEC-MALLS was used to establish the existence of stable and monodisperse U2AF dimer and U2AF/SF1 trimer complexes from first principles. This was followed by the spectroscopic evaluation of two RNA-bound phosphomimetic U2AF/SF1 trimer complexes after SEC purification in order to establish that the RNA-bound U2AF dimer and U2AF/SF1 trimer complexes are stable and therefore tractable for biochemical characterization. Finally, SEC purification and spectroscopy was repeated on the RNA-bound complexes after one week of storage in order to establish the long-term solubility and stability of the complexes to evaluate the tractability of the system for structural characterization.

EMSAs were used to evaluate the RNA binding properties of the *S. pombe* U2AF dimer and U2AF/SF1 trimer. The five protein complexes were paired with four different 3′ SS model RNAs. A comparison of the EMSA-derived K_d values from different protein/RNA pairings provides insight into the RNA binding properties of the U2AF dimer and U2AF/SF1 trimer complexes.

2-2. Results

2-2.1. Cloning, co-expression, and purification of U2AF dimer and U2AF/SF1 trimer

Briefly, U2AF-L constructs were cloned into the pACYC Duet-1 co-expression vector, and U2AF23 ± SF1 constructs were cloned into the pET Duet-1 co-expression vector. Subsequently, these were co-transformed into the *E. coli* expression host and co-expressed. This was followed by SEC purification. For biochemical characterization, this was sufficient purification. However, for the structural characterization experiments described in Chapter 3, an additional anion exchange purification step followed SEC.

Eluted protein was analyzed via SDS-PAGE and coomassie blue staining following nickel affinity chromatography purification. SDS-PAGE was used to confirm complete TEV cleavage of the affinity tag on U2AF-L. There is a distinct and obvious, but subtle size difference between uncleaved and cleaved U2AF-L. U2AF dimer and U2AF/SF1 trimer complexes used in this thesis can be categorized into three groups based on purification behaviour and complexes belonging to the same group are indistinguishable from one another during purification, with the caveat that U2AF23 (M1-E194) migrates slightly faster than full-length U2AF23 on an SDS-PAGE gel (see Section *I*-3.4.). The size difference between uncleaved and cleaved U2AF59 (E106-W517), as well as the size difference between U2AF23 (M1-216) and U2AF23 (M1-E194) as observed by SDS-PAGE are presented below in Fig. 2-1. Note the presence of a distinct low MW (molecular weight) band in Fig. 2-1B referred to as ‘unidentified species’. This contaminant is not present in any U2AF dimer purifications but is present in all purifications of U2AF/SF1 trimer and can only be completely eliminated by anion exchange chromatography.

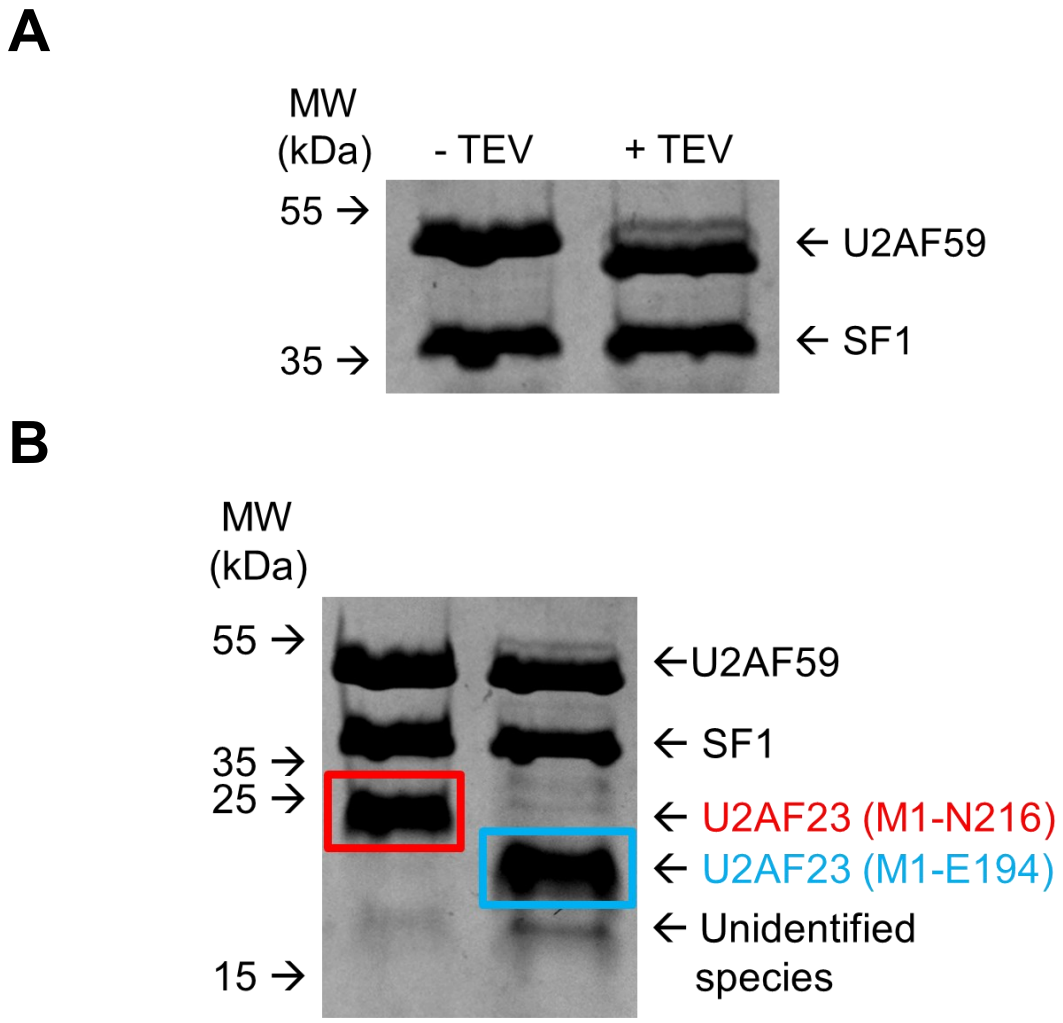


Figure 2-1. Size difference between uncleaved and cleaved U2AF59 (E106-W517), and between U2AF23 (M1-216) and U2AF23 (M1-E194). (A) Size difference between uncleaved and cleaved U2AF59 (E106-W517). The smaller protein band is SF1 (M52-E373, wildtype). (B) The complex presented in the left lane is U2AF59 (E106-W517) + SF1 (M52-E373, wildtype) + U2AF23 (M1-N216). The complex presented in the right lane is U2AF59 (E106-W517) + SF1 (M52-E373, wildtype) + U2AF23 (M1-E194). Both complexes are TEV cleaved.

2-2.1.1. SEC purification of U2AF dimer and U2AF/SF1 trimer

When the Ni-NTA purified protein is separated over the S200 column, a UV trace is generated. Fig. 2-2 superimposes a representative trace of each of the five complexes used as the basis for biochemical characterization in this chapter into one graph for a visual comparison. The traces used in Fig. 2-2 were selected because these S200 runs produced peaks of roughly equal height, providing better visual clarity when comparing the elution profile of the five complexes.

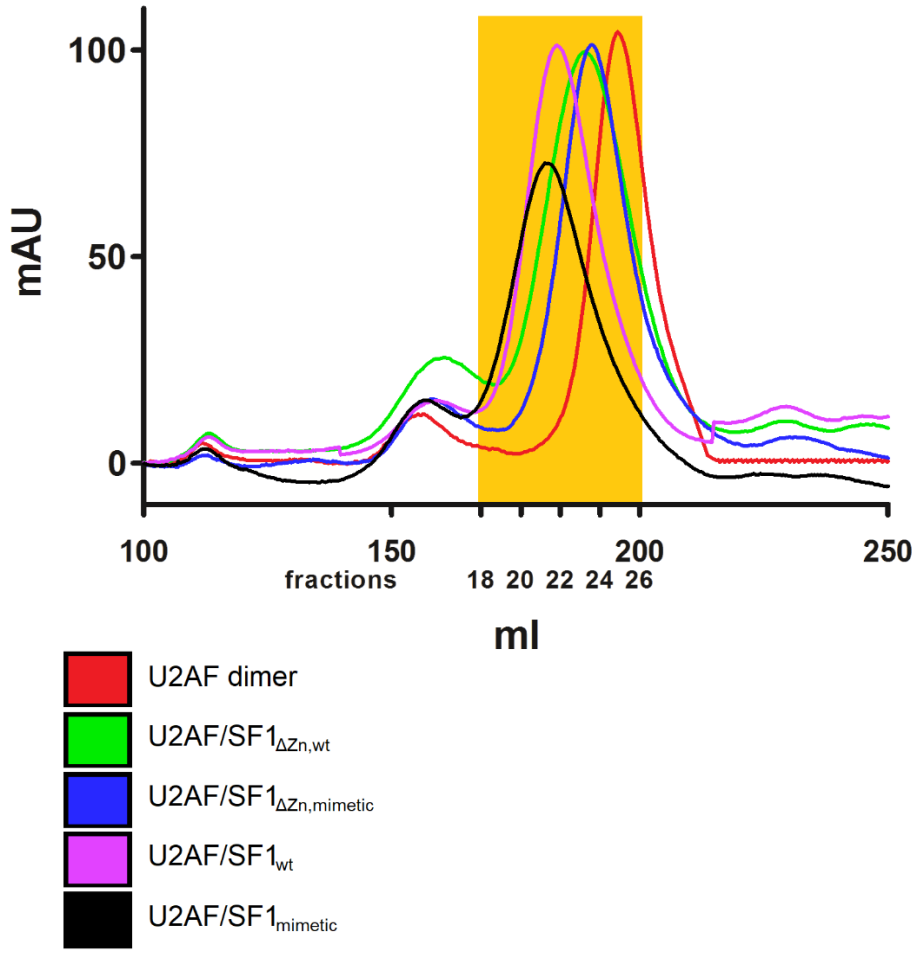


Figure 2-2. Superimposition of the S200 UV trace of each of the five complexes used in Chapter 2. Input sample for all five traces is uncleaved and was loaded onto the S200 after elution from Ni-NTA. The highlighted elution volume range covers the retention volume for all five complexes and additionally also covers the majority of the peak which contains relatively pure, monodisperse protein for all five complexes. The retention volumes for the complexes are as follows: U2AF dimer ≈195 mL, U2AF/SF1$_{\Delta Zn,wt}$ and U2AF/SF1$_{\Delta Zn,mimetic}$ ≈187 mL, U2AF/SF1$_{wt}$ and U2AF/SF1$_{mimetic}$ ≈181 mL.

The traces show a symmetrical peak for each complex, indicating that the protein components are present as either a stable 1:1 heterodimeric or a stable 1:1:1 heterotrimeric complex. This is also supported by the observation that the larger the complex, the earlier it elutes, as is expected.

Below, SEC purification results for a representative batch of protein for U2AF dimer, U2AF/SF1$_{\Delta Zn,wt}$, and U2AF/SF1$_{wt}$ are summarized in Figures 2-3, 2-4, and 2-5, respectively. In each of these figures the input sample (pooled and concentrated Ni-NTA elution), S200 UV trace, and an SDS-PAGE gel corresponding to the S200 UV trace are presented for the preparation of the corresponding complex. Note that the data for U2AF/SF1$_{\Delta Zn,wt}$ and U2AF/SF1$_{\Delta Zn,mimetic}$ are indistinguishable from each other, and the data for U2AF/SF1$_{wt}$ and U2AF/SF1$_{mimetic}$ are indistinguishable from each other. Therefore, for simplicity, SEC purification results of the two phosphomimetic U2AF/SF1 trimer complexes have been omitted.

Figure 2-3. SEC purification of U2AF dimer. (A) SDS-PAGE gel of the input sample (pooled and concentrated Ni-NTA elutions) which was loaded onto the S200. (B) S200 UV trace generated by the input sample. Fractions corresponding to the highlighted elution volume range in Fig. 2-2 are highlighted under this UV trace. (C) SDS-PAGE gel of the fractions highlighted in yellow in (B).

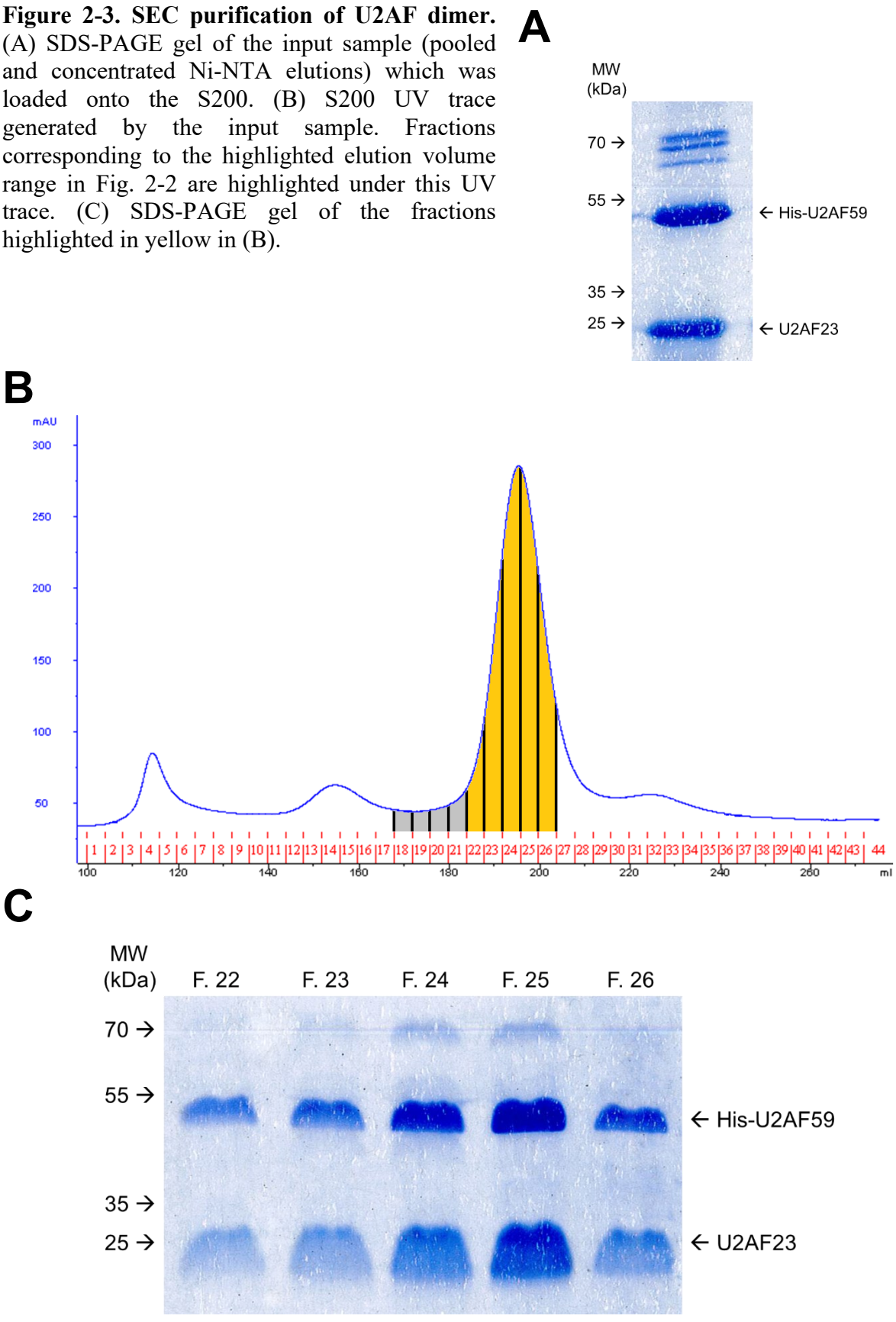


Figure 2-4. SEC purification of U2AF/SF1$_{\Delta Zn,wt}$. (A) SDS-PAGE gel of the input sample (pooled and concentrated Ni-NTA elutions) which was loaded onto the S200. (B) S200 UV trace generated by the input sample. Fractions corresponding to the highlighted elution volume range in Fig. 2-2 are highlighted under this UV trace. (C) SDS-PAGE gel of the fractions highlighted in yellow in (B).

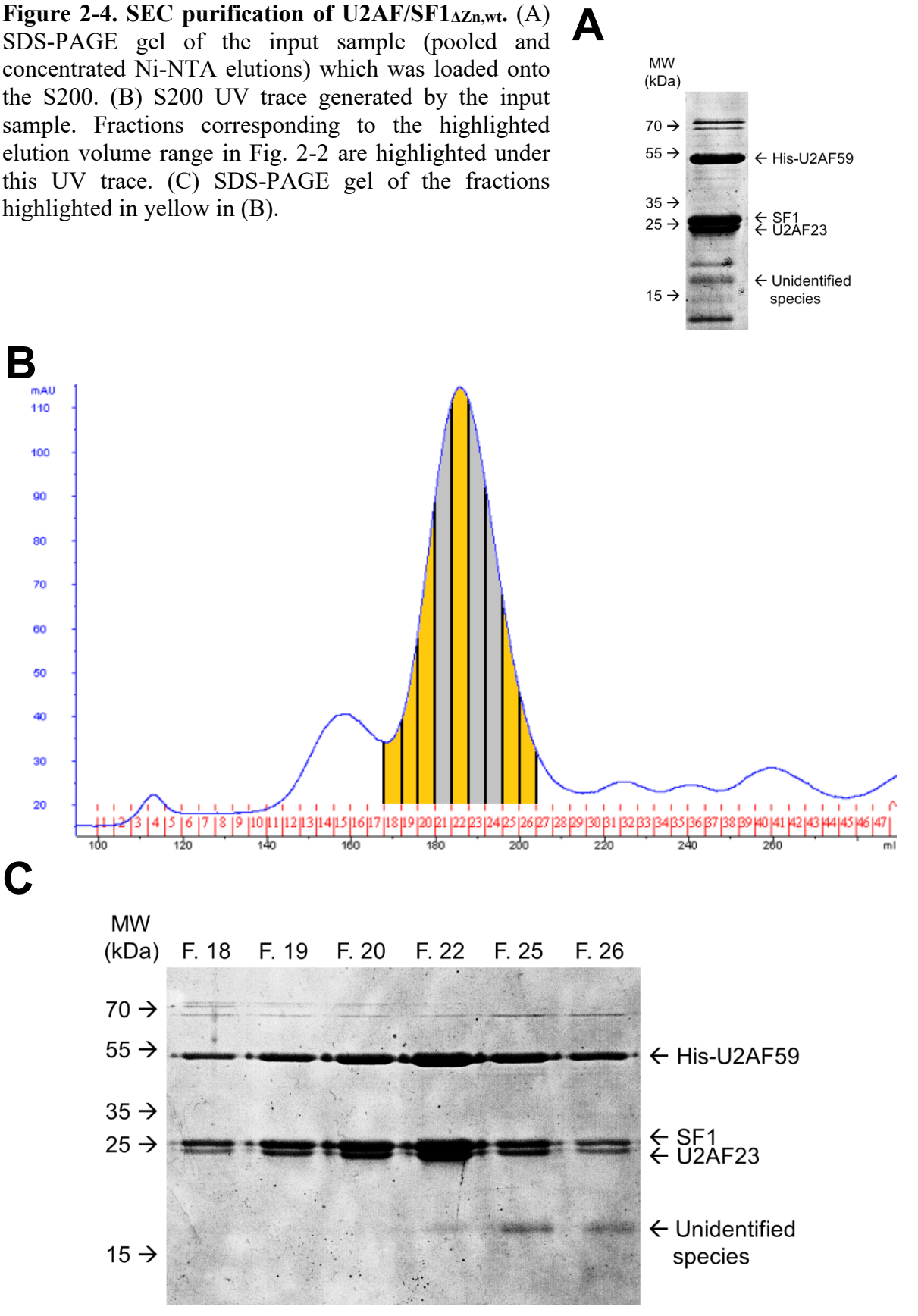
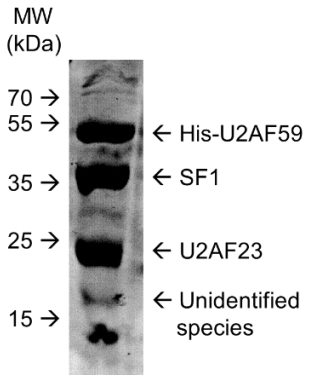
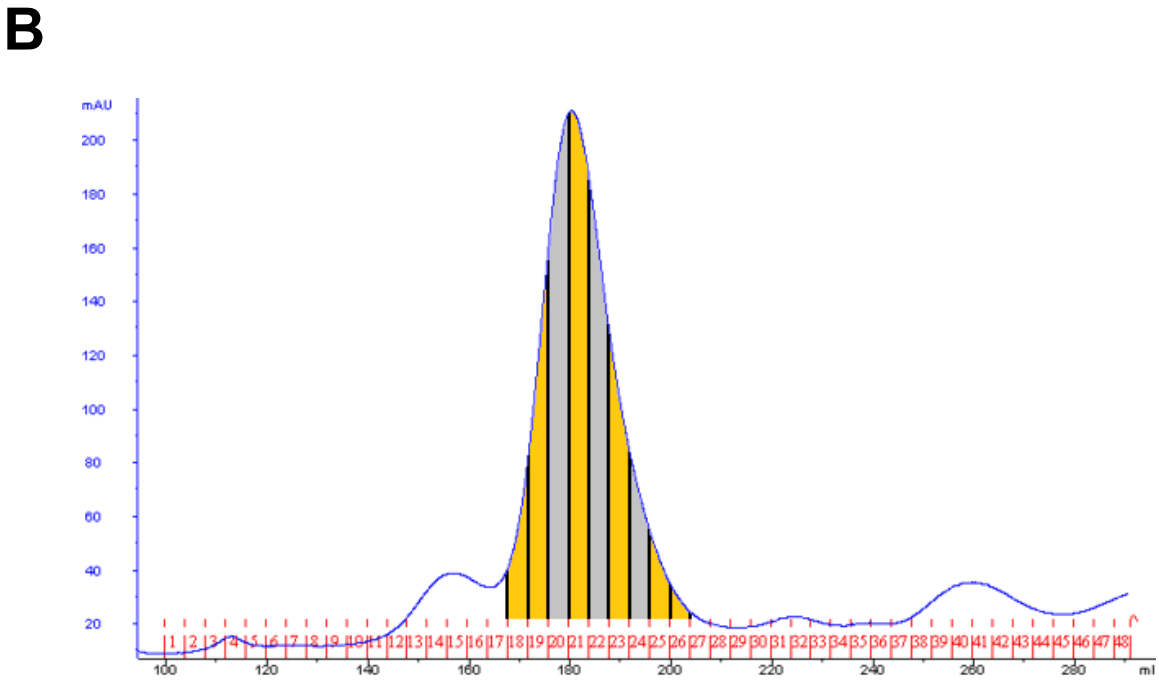
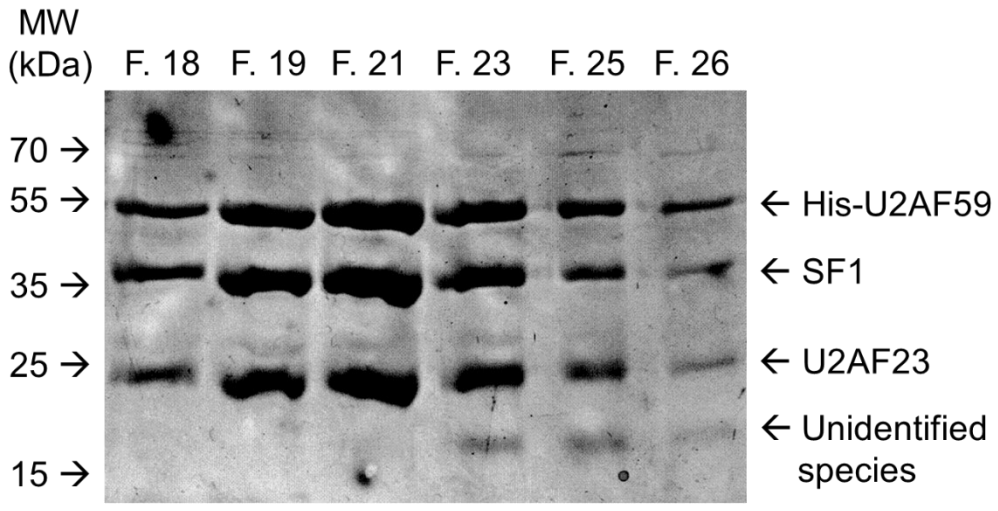


Figure 2-5. SEC purification of U2AF/SF1$_{wt}$. (A) SDS-PAGE gel of the input sample (pooled and concentrated Ni-NTA elutions) which was loaded onto the S200. (B) S200 UV trace generated by the input sample. Fractions corresponding to the highlighted elution volume range in Fig. 2-2 are highlighted under this UV trace. (C) SDS-PAGE gel of the fractions highlighted in yellow in (B).







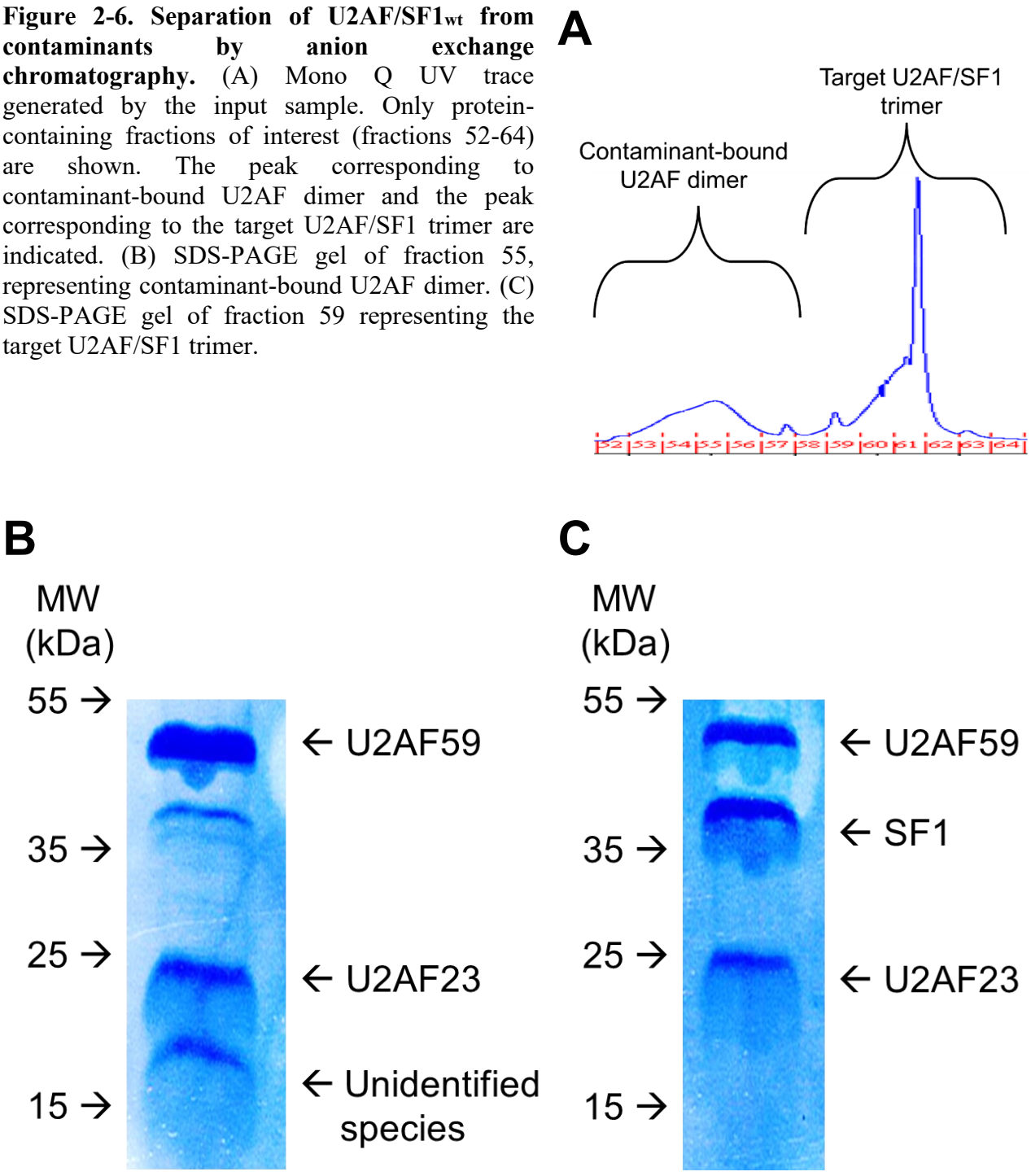
2-2.1.2. Anion exchange chromatography purification of U2AF dimer and U2AF/SF1 trimer

Following SEC, U2AF dimer and U2AF/SF1 trimer destined for structural experiments were purified using anion exchange chromatography over a Mono Q HR 5/5 column in order to ensure that the final sample was free of all contamination detectable by SDS-PAGE, SEC, or anion exchange chromatography. All U2AF dimer and U2AF/SF1 trimer complexes eluted from the anion exchange column at a salt concentration of ~100 mM NaCl.

The anion exchange protocol for the U2AF dimer and U2AF/SF1 trimer is identical, but with one critical difference: the U2AF/SF1 trimer was resolved from contaminating species using a much more gradual gradient than that used for U2AF dimer. This is required in order to eliminate the low MW contaminant previously introduced as 'unidentified species'.

When U2AF/SF1 trimer was run over the Mono Q using a sufficiently gradual gradient, the sample resolves into two distinct peaks: an earlier-eluting small and broad peak and a later-eluting tall and narrow peak. The small, broad peak corresponds to U2AF dimer, which co-elutes with the low MW contaminant, and it appears that the U2AF dimer and contaminant co-elute in stoichiometric amounts. The tall, narrow peak corresponds to the target U2AF/SF1 trimer with no detectable contamination. The Mono Q UV trace and corresponding SDS-PAGE gel for the final purification of U2AF/SF1 trimer are presented below in Figures 2-6 and 2-7. Fig. 2-6 illustrates the separation of these two complexes by anion exchange chromatography in an abbreviated form, and Fig. 2-7 shows the complete UV trace and SDS-PAGE gel of the purification shown in Fig. 2-6 for greater context.

Figure 2-6. Separation of U2AF/SF1$_{wt}$ from contaminants by anion exchange chromatography. (A) Mono Q UV trace generated by the input sample. Only protein-containing fractions of interest (fractions 52-64) are shown. The peak corresponding to contaminant-bound U2AF dimer and the peak corresponding to the target U2AF/SF1 trimer are indicated. (B) SDS-PAGE gel of fraction 55, representing contaminant-bound U2AF dimer. (C) SDS-PAGE gel of fraction 59 representing the target U2AF/SF1 trimer.



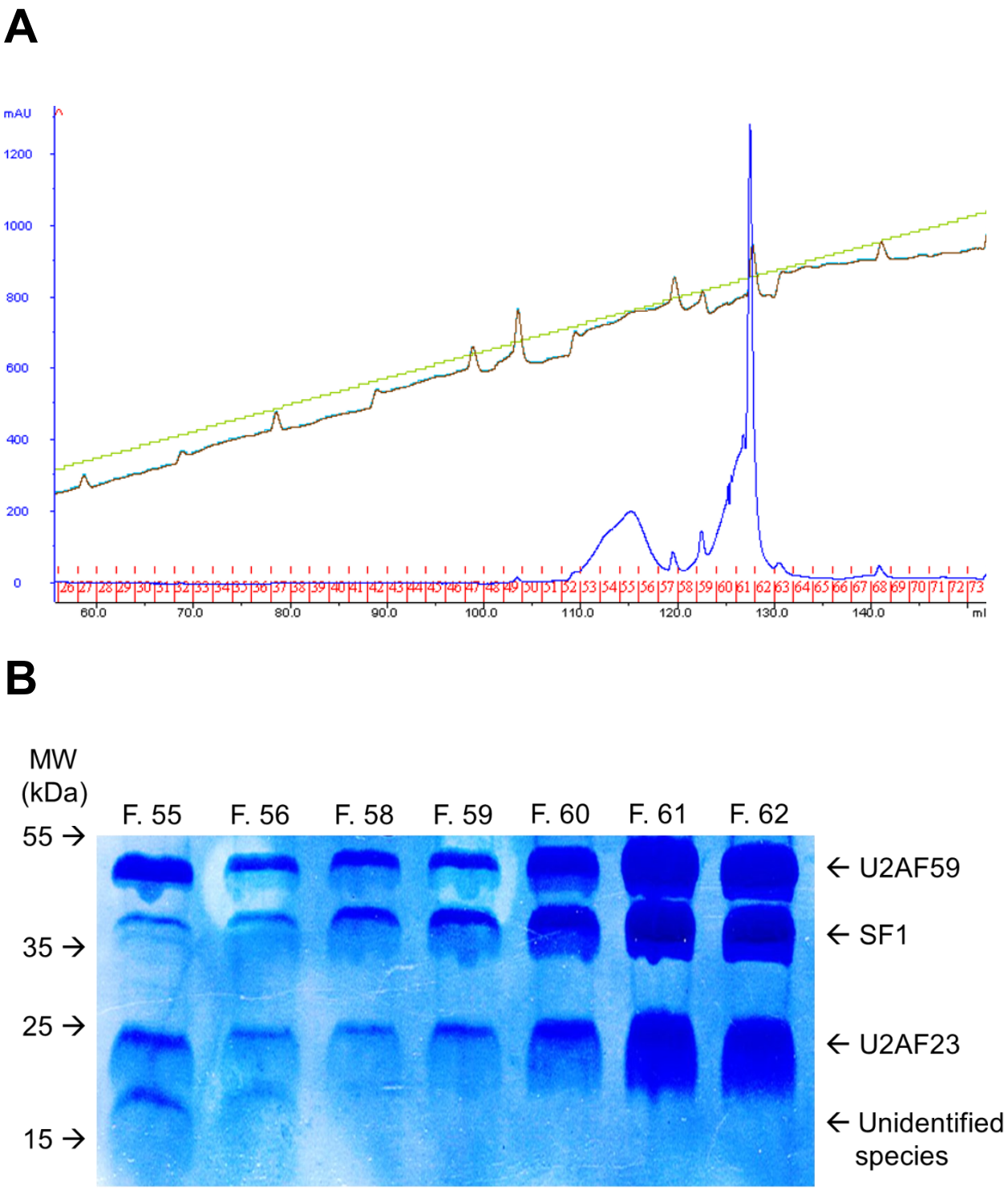


Figure 2-7. Anion exchange chromatography purification of U2AF/SF1$_{wt}$. (A) Mono Q trace generated by the input sample showing fractions 26-73 overlaid with the NaCl concentration (green line) and buffer conductivity (brown line). The peak corresponding to contaminant-bound U2AF dimer and the peak corresponding to the target U2AF/SF1 trimer described in Fig. 2-6 are clearly observable. (B) SDS-PAGE gel of protein-containing fractions.

2-2.2. SEC-MALLS characterization of U2AF dimer and U2AF/SF1 trimer

SEC-MALLS data was generated for U2AF dimer, U2AF/SF1$_{\Delta Zn,mimetic}$ trimer, and U2AF/SF1$_{mimetic}$ trimer and is presented below in Fig. 2-8, Fig. 2-9, and Table 2-1. SEC-MALLS data for U2AF/SF1$_{\Delta Zn,wt}$ trimer and U2AF/SF1$_{wt}$ trimer was not collected because when compared to their phosphomimetic counterparts, these complexes exhibit indistinguishable purification behaviour and are of almost identical size.

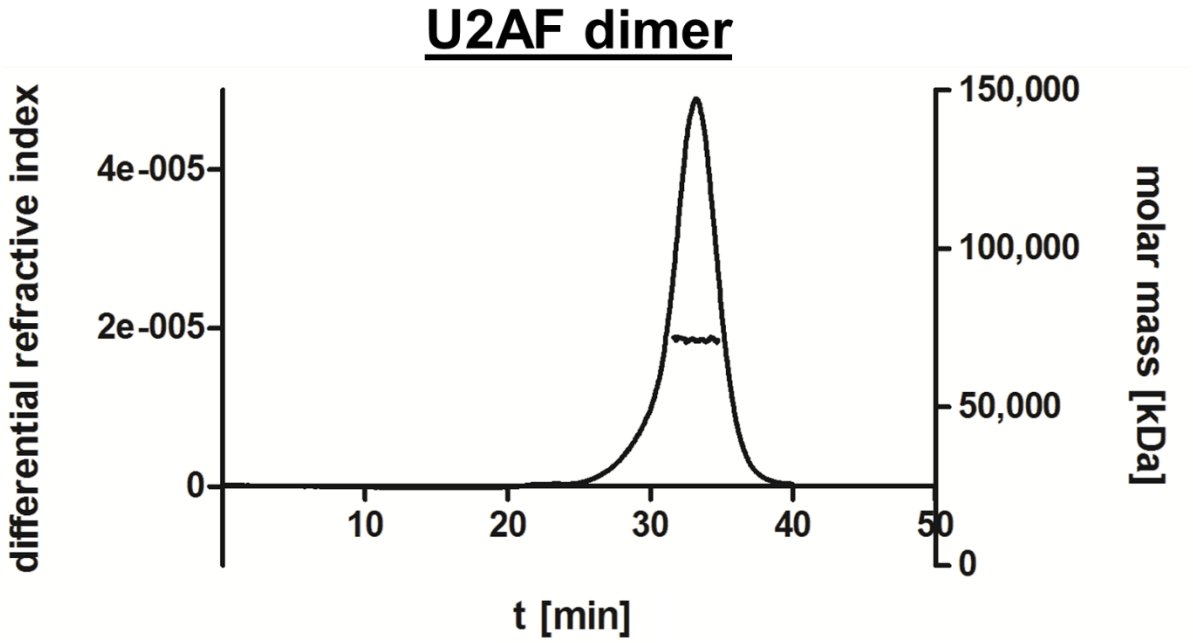


Figure 2-8. SEC-MALLS trace of U2AF dimer. The molar mass of the particle determined by MALLS (71 kDa) is overlaid with the dRI chromatogram.

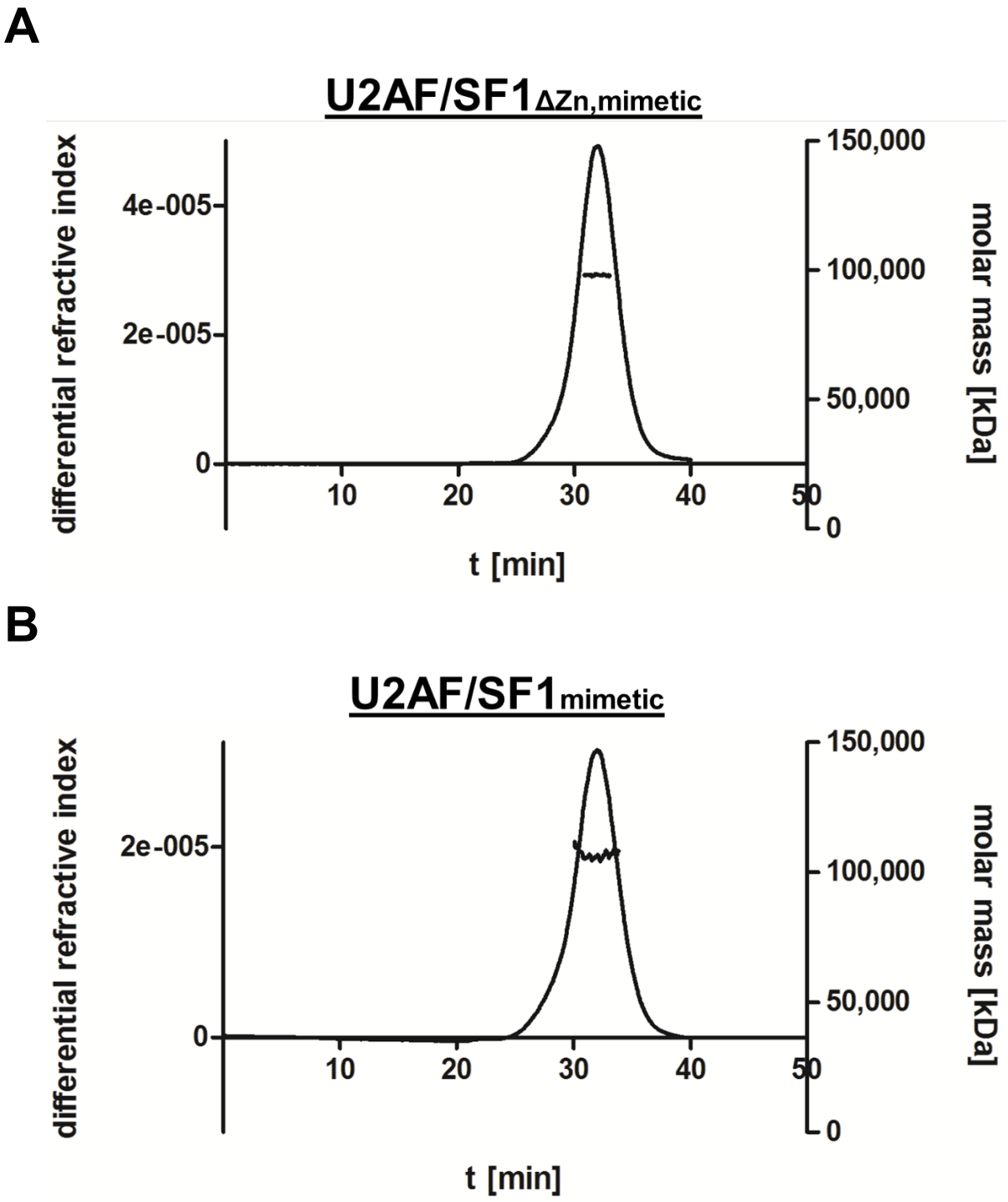


Figure 2-9. SEC-MALLS traces of U2AF/SF1 trimer. The molar masses determined by MALLS are overlaid with the dRI chromatogram. (A) SEC-MALLS trace of the $U2AF/SF1_{\Delta Zn,mimetic}$ trimer, indicating a molar mass of 98 kDa for the particle. (B) SEC-MALLS trace of the $U2AF/SF1_{mimetic}$ trimer, indicating a molar mass of 107 kDa for the particle.

Table 2-1: Comparison of sequence-based and SEC-MALLS derived MW

Complex	Sequence-based MW	SEC-MALLS derived MW
U2AF dimer	71.59 kDa	71 kDa
U2AF/SF1$_{\Delta Zn,mimetic}$	101.49 kDa	98 kDa
U2AF/SF1$_{mimetic}$	108.77 kDa	107 kDa

2-2.3. Spectroscopy of RNA-bound U2AF/SF1 trimer

The absorption spectra for free U2AF/SF1 trimer, RNA-bound U2AF/SF1 trimer, and free 3′ SS model RNA were taken and are shown below in Fig. 2-10; U2AF/SF1$_{\Delta Zn,mimetic}$ trimer and U2AF/SF1$_{mimetic}$ trimer were both assayed. The absorption spectrum for apo protein is expected to show a maximum at ~280 nm and the absorption spectrum for free RNA is expected to show a maximum at ~260 nm; if a stable protein/RNA complex exists, then its spectrum will have a maximum that is intermediate between the two, and this was observed for both U2AF/SF1 trimers (see Fig. 2-10) (Glasel, 1995; Goldfarb, Saidel, & Mosovich, 1951).

The apo state of both the U2AF/SF1$_{\Delta Zn,mimetic}$ and U2AF/SF1$_{mimetic}$ trimer displays an absorption maximum at 276 nm, and free RNA CG92 displays a maximum at ~260 nm. After incubation with RNA CG92, the maximum for U2AF/SF1$_{\Delta Zn,mimetic}$ is shifted to 270 nm, and the maximum for U2AF/SF1$_{mimetic}$ is shifted to 268 nm.

In order to assay the long-term stability of the RNA-bound U2AF/SF1 trimers, they were stored at 4°C for 7 days before being re-purified on the S200 column and concentrated; the S200 UV traces showed no aggregation and the absorption spectra of these aged, re-purified samples were similar to that of the fresh samples shown in Fig. 2-10, showing a maximum at 268 nm, indicating that the protein/RNA complex is stable and soluble over an extended period of time.

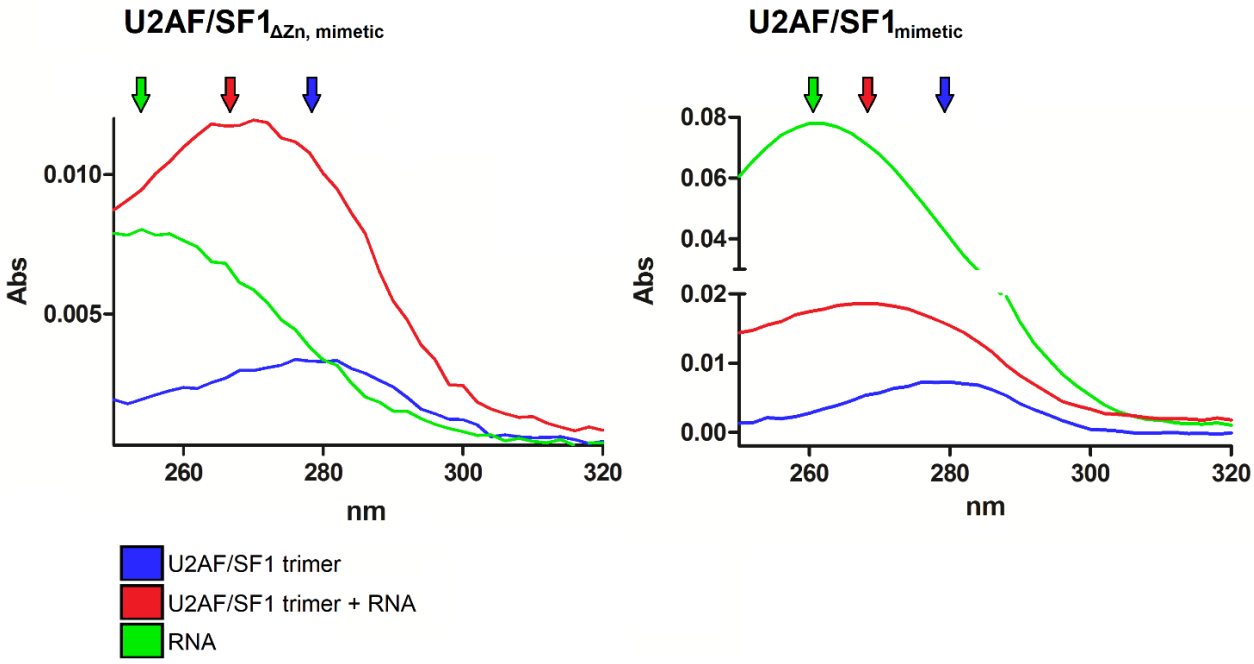


Figure 2-10. Absorption spectra of free U2AF/SF1 trimer, RNA-bound U2AF/SF1 trimer, and free 3′ SS model RNA. The absorption spectrum (250-320 nm) for U2AF/SF1$_{\Delta Zn,mimetic}$ trimer and U2AF/SF1$_{mimetic}$ trimer in their apo state, in their RNA-bound state with RNA CG92, and free RNA CG92 are shown. The absorption maxima are indicated by arrows. A colour-coded annotation is provided below the spectra.

2-2.4. EMSAs of U2AF dimer and U2AF/SF1 trimer

K_d values were determined by EMSAs (Figures 2-11 to 2-20 show a representative EMSA for each protein/RNA pairing). K_d values are summarized below in Table 2-2 along with the standard error generated by the three trials. Note that when EMSA figures are compared, in most cases the input lane is shared among several EMSA figures.

Table 2-2: K_d values for various protein/RNA pairings

Complex	Model RNA	K_d (μM)
U2AF dimer	wildtype	2.4 ± 0.2
	U12	2.0 ± 0.2
	complement	58.0 ± 2.0
U2AF/SF1$_{\Delta Zn,wt}$	wildtype	0.9 ± 0.2
	U12	0.7 ± 0.1
	scrambled BPS	16.2 ± 1.0
	complement	8.2 ± 0.1
U2AF/SF1$_{\Delta Zn,mimetic}$	wildtype	0.5 ± 0.0
	U12	1.5 ± 0.3
	scrambled BPS	9.4 ± 0.1
	complement	6.5 ± 2.1
U2AF/SF1$_{wt}$	wildtype	0.3 ± 0.0
	U12	0.5 ± 0.0
	scrambled BPS	11.6 ± 3.0
	complement	9.5 ± 0.1
U2AF/SF1$_{mimetic}$	wildtype	0.1 ± 0.0
	U12	0.5 ± 0.1
	scrambled BPS	10.3 ± 0.2
	complement	8.9 ± 1.6

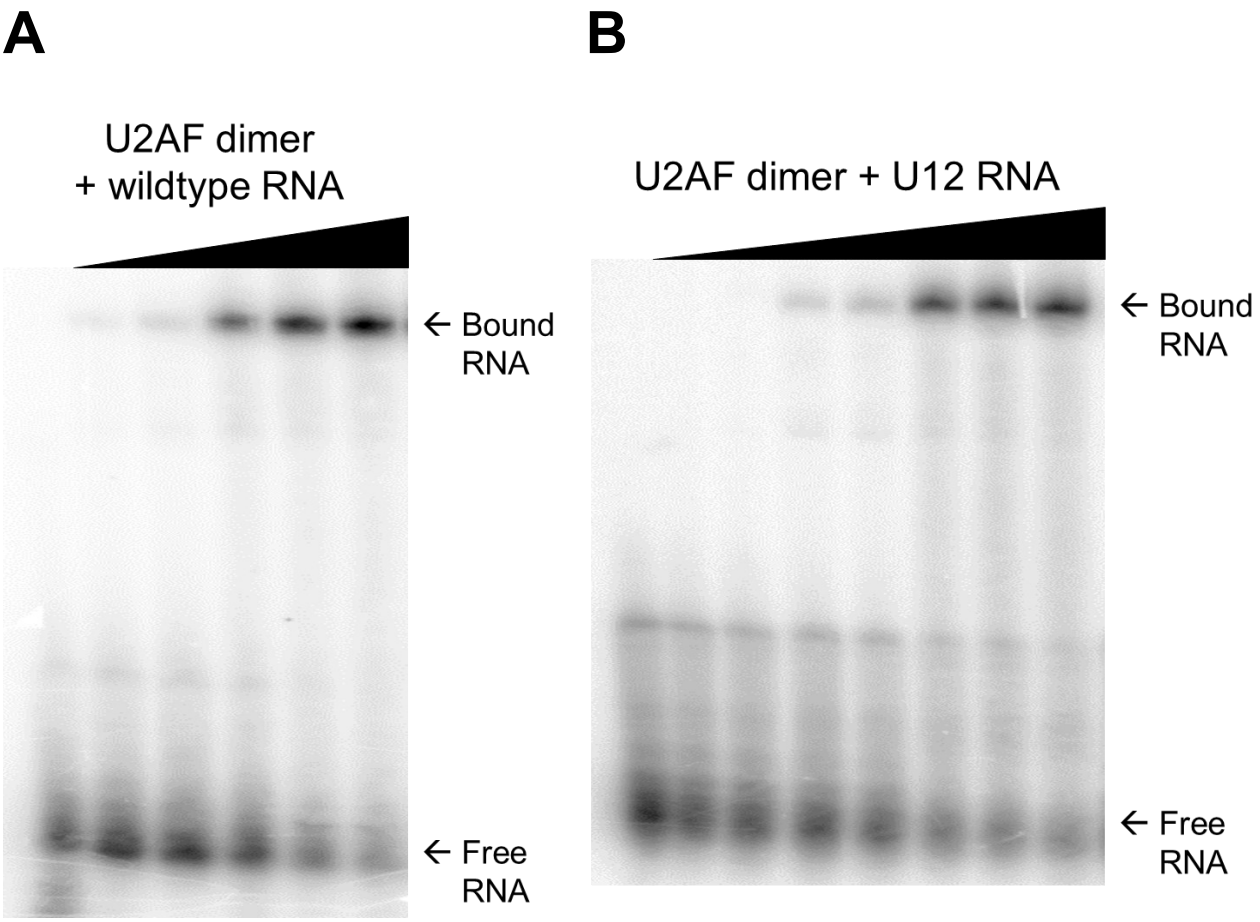


Figure 2-11. EMSAs of U2AF dimer bound to wildtype RNA and U12 RNA. (A) U2AF dimer + wildtype RNA. Titration series: 0, 0.5, 1, 5, 10, and 20 μM protein. (B) U2AF dimer + U12 RNA. Titration series: 0, 0.05, 0.1, 0.5, 1, 5, 10, and 20 μM protein.

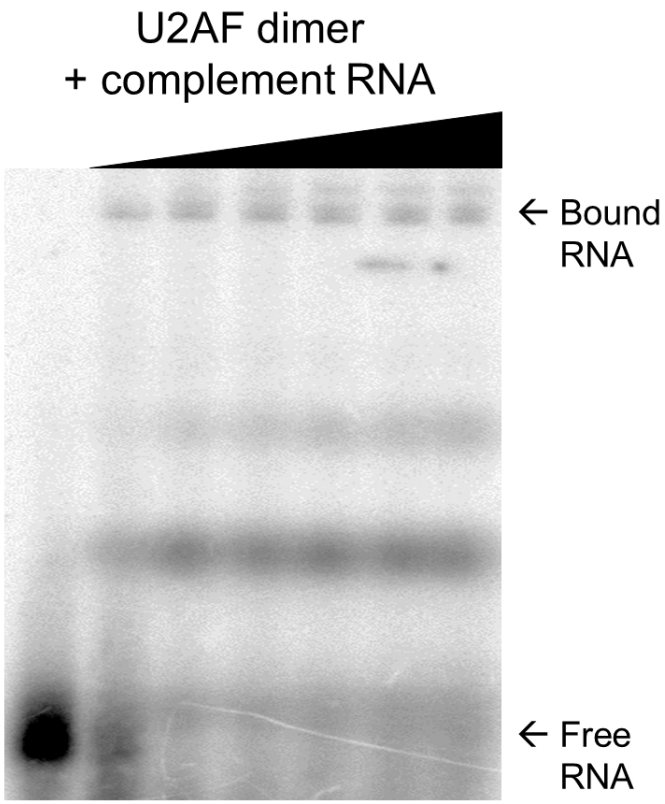


Figure 2-12. EMSA of U2AF dimer bound to complement RNA. Titration series: 0, 10, 50, 100, 150, 200, and 300 μM protein.

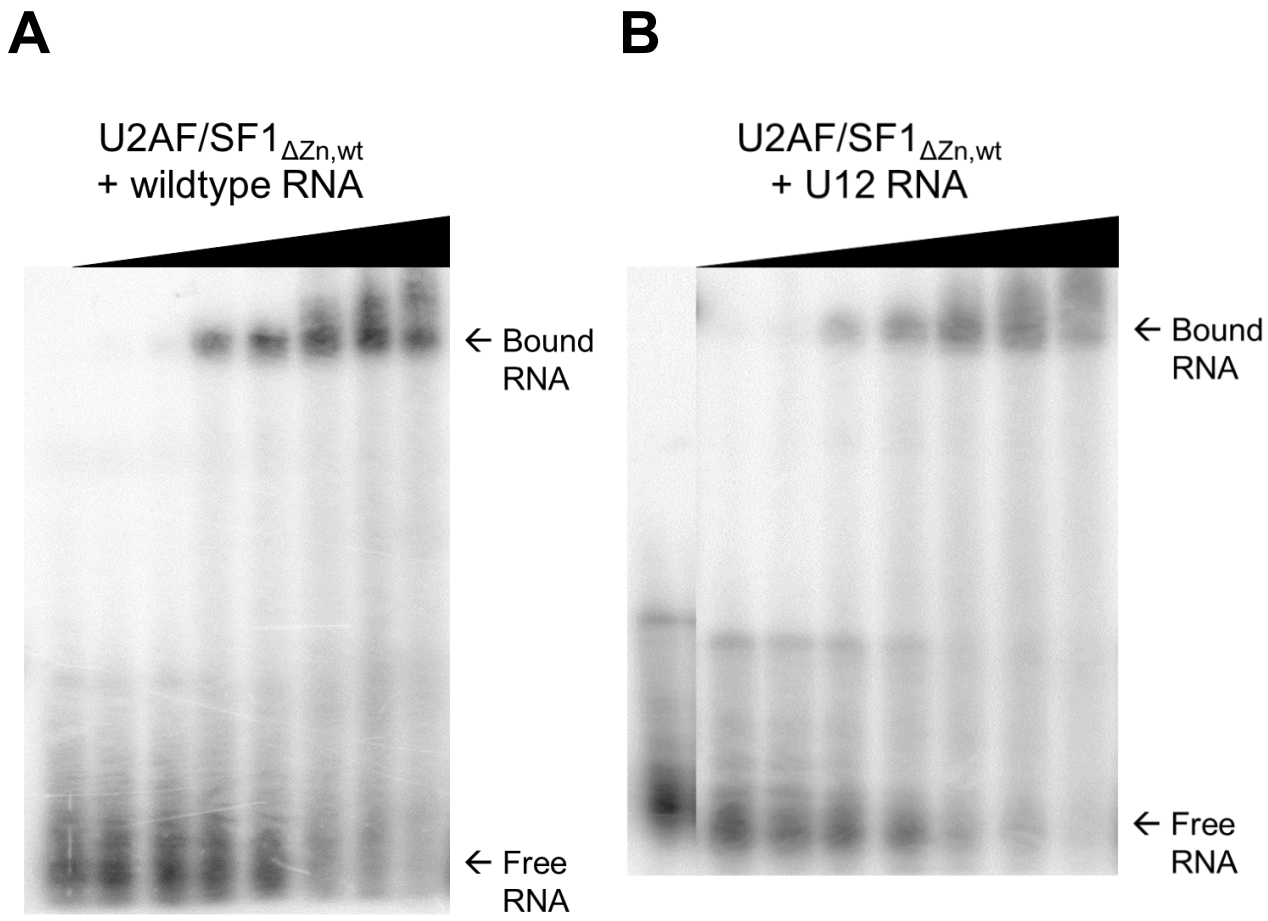


Figure 2-13. EMSAs of U2AF/SF1$_{\Delta Zn,wt}$ trimer bound to wildtype RNA and U12 RNA. (A) U2AF/SF1$_{\Delta Zn,wt}$ trimer + wildtype RNA. Titration series: 0, 0.05, 0.1, 0.5, 1, 5, 10, and 20 μM protein. (B) U2AF/SF1$_{\Delta Zn,wt}$ trimer + U12 RNA. Titration series: 0, 0.05, 0.1, 0.5, 1, 5, 10, and 20 μM protein.

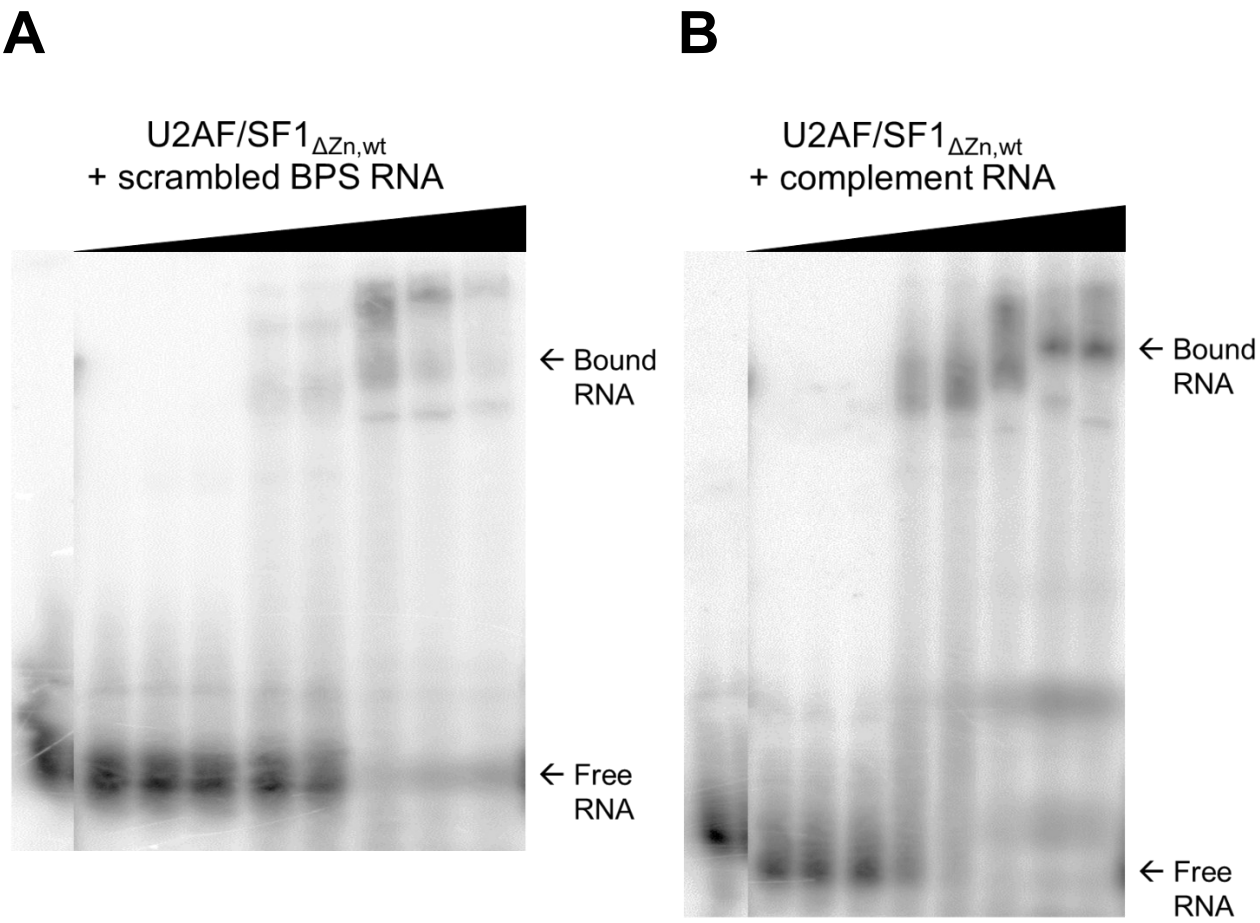


Figure 2-14. EMSAs of U2AF/SF1$_{\Delta Zn,wt}$ trimer bound to scrambled BPS RNA and complement RNA. (A) U2AF/SF1$_{\Delta Zn,wt}$ trimer + scrambled BPS RNA. Titration series: 0, 0.1, 0.5, 1, 5, 10, 50, 100, and 175 μM protein. (B) U2AF/SF1$_{\Delta Zn,wt}$ trimer + complement RNA. Titration series: 0, 0.1, 0.5, 1, 5, 10, 50, 100, and 175 μM protein.

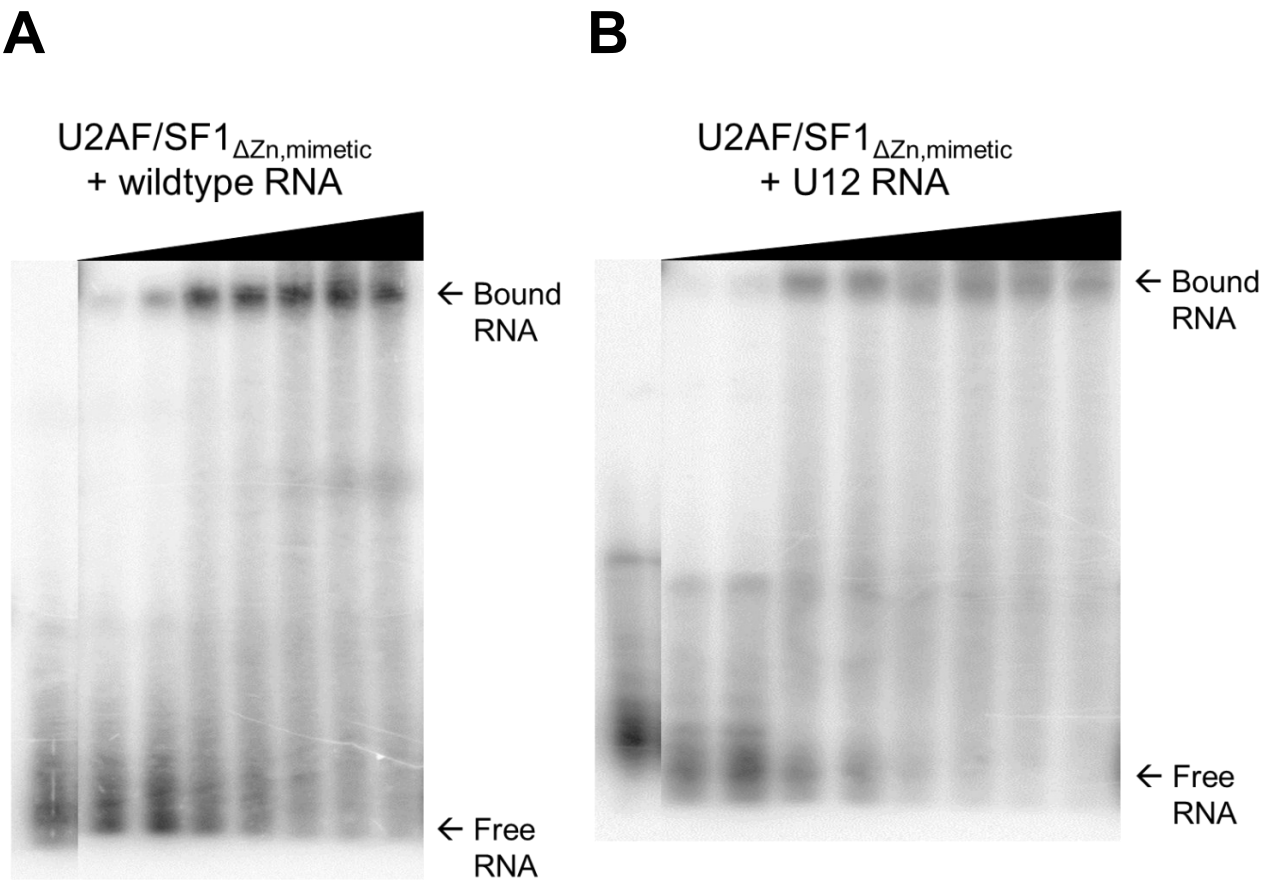


Figure 2-15. EMSAs of U2AF/SF1$_{\Delta Zn,mimetic}$ trimer bound to wildtype RNA and U12 RNA. (A) U2AF/SF1$_{\Delta Zn,mimetic}$ trimer + wildtype RNA. Titration series: 0, 0.05, 0.1, 0.5, 1, 5, 10, and 20 μM protein. (B) U2AF/SF1$_{\Delta Zn,mimetic}$ trimer + U12 RNA. Titration series: 0, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, and 20 μM protein.

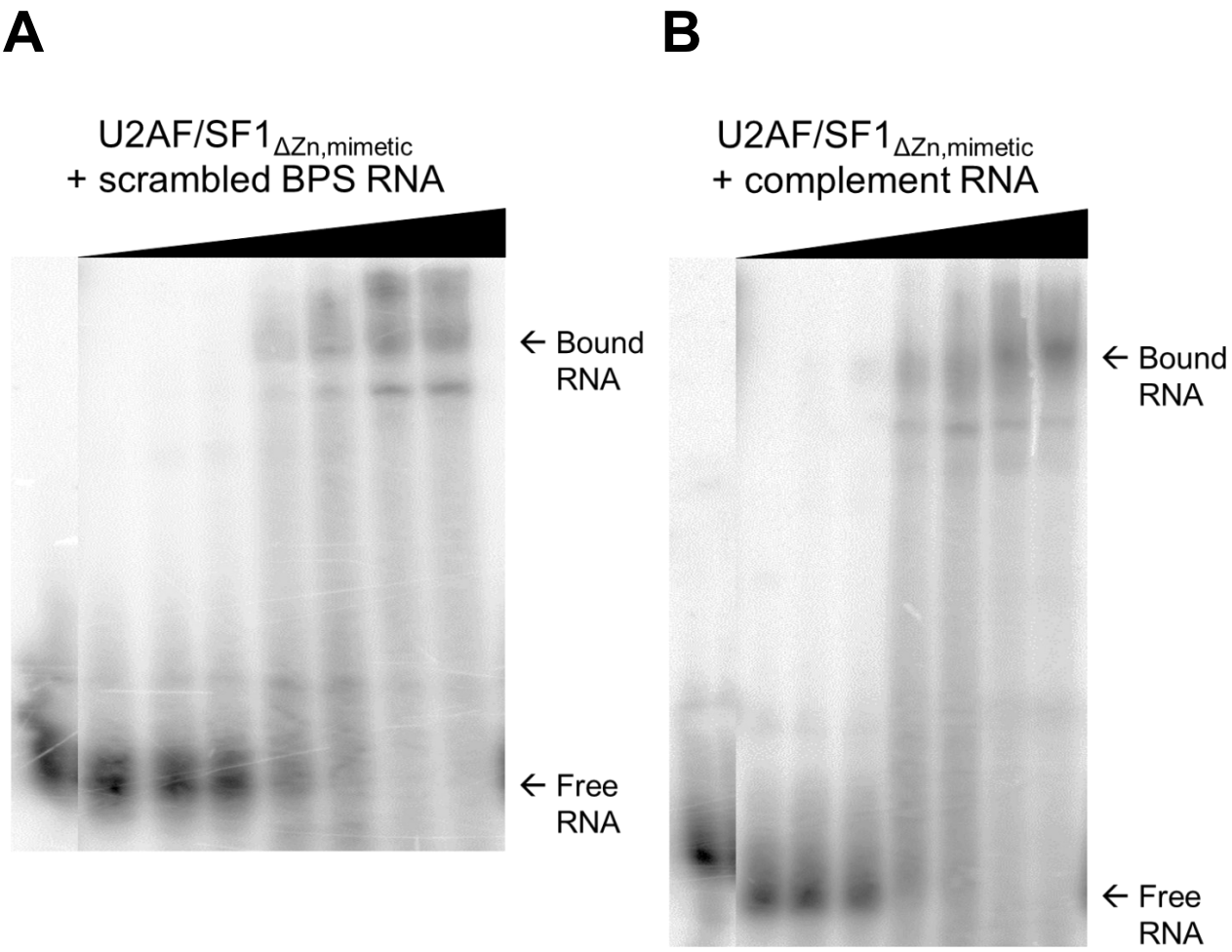


Figure 2-16. EMSAs of U2AF/SF1$_{\Delta Zn,mimetic}$ trimer bound to scrambled BPS RNA and complement RNA. (A) U2AF/SF1$_{\Delta Zn,mimetic}$ trimer + scrambled BPS RNA. Titration series: 0, 0.1, 0.5, 1, 5, 10, 50, and 100 μM protein. (B) U2AF/SF1$_{\Delta Zn,mimetic}$ trimer + complement RNA. Titration series: 0, 0.1, 0.5, 1, 5, 10, 50, and 100 μM protein.

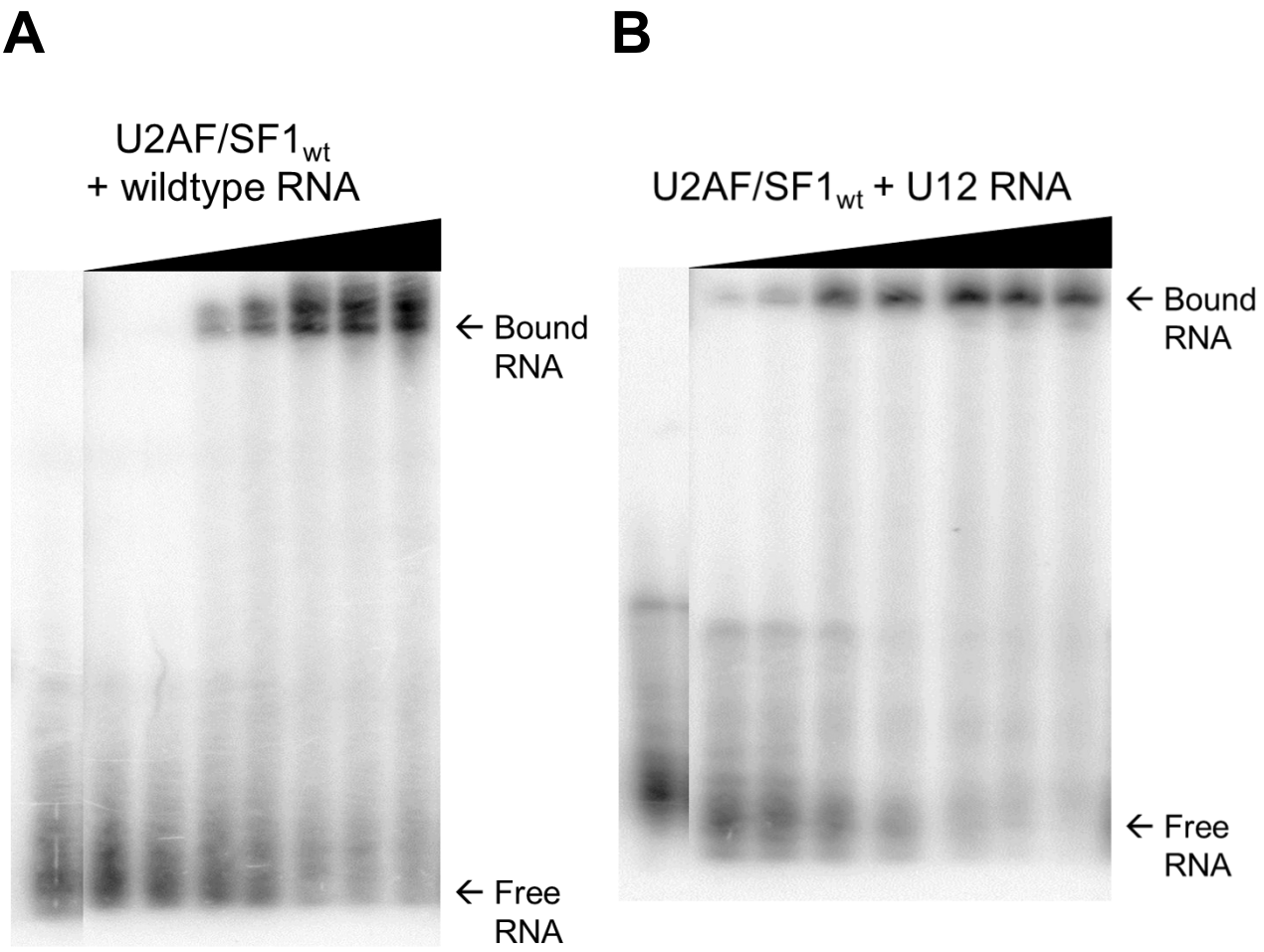


Figure 2-17. EMSAs of U2AF/SF1$_{wt}$ trimer bound to wildtype RNA and U12 RNA. (A) U2AF/SF1$_{wt}$ trimer + wildtype RNA. Titration series: 0, 0.05, 0.1, 0.5, 1, 5, 10, and 20 μM protein. (B) U2AF/SF1$_{wt}$ trimer + U12 RNA. Titration series: 0, 0.05, 0.1, 0.5, 1, 5, 10, and 20 μM protein.

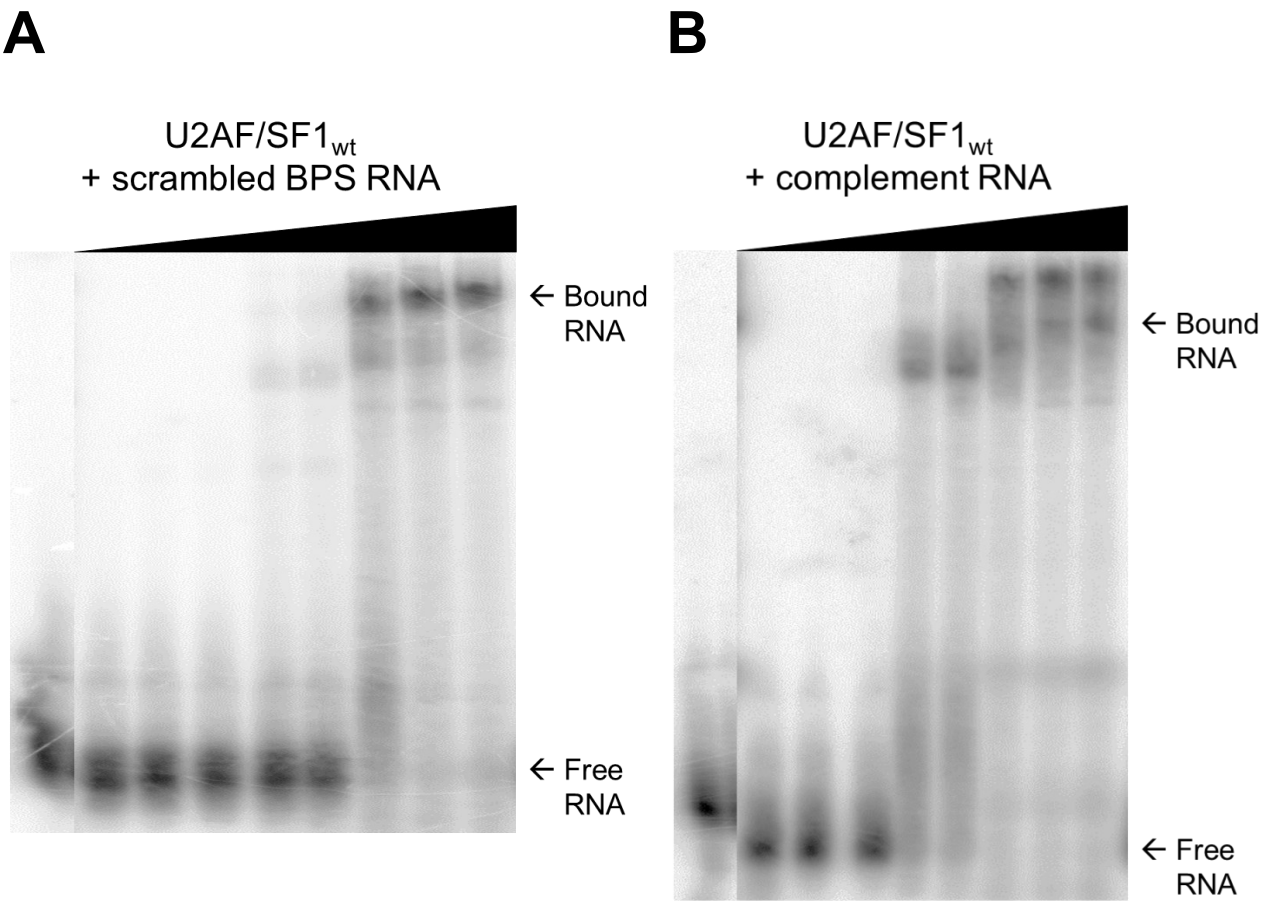


Figure 2-18. EMSAs of U2AF/SF1$_{wt}$ trimer bound to scrambled BPS RNA and complement RNA. (A) U2AF/SF1$_{wt}$ trimer + scrambled BPS RNA. Titration series: 0, 0.1, 0.5, 1, 5, 10, 50, 100, and 175 μM protein. (B) U2AF/SF1$_{wt}$ trimer + complement RNA. Titration series: 0, 0.1, 0.5, 1, 5, 10, 50, 100, and 175 μM protein.

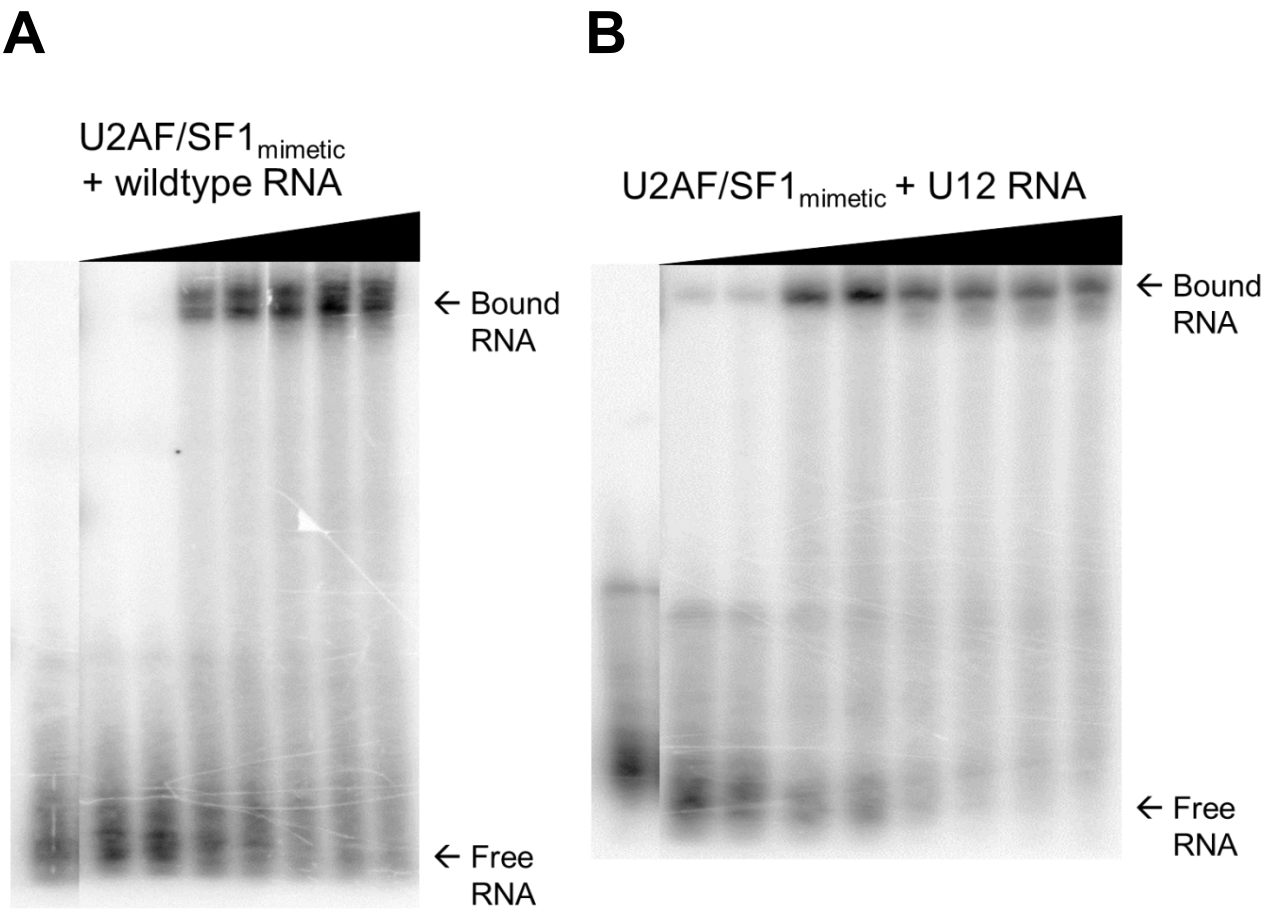


Figure 2-19. EMSAs of U2AF/SF1$_{mimetic}$ trimer bound to wildtype RNA and U12 RNA. (A) U2AF/SF1$_{mimetic}$ trimer + wildtype RNA. Titration series: 0, 0.05, 0.1, 0.5, 1, 5, 10, and 20 μM protein. (B) U2AF/SF1$_{mimetic}$ trimer + U12 RNA. Titration series: 0, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, and 20 μM protein.

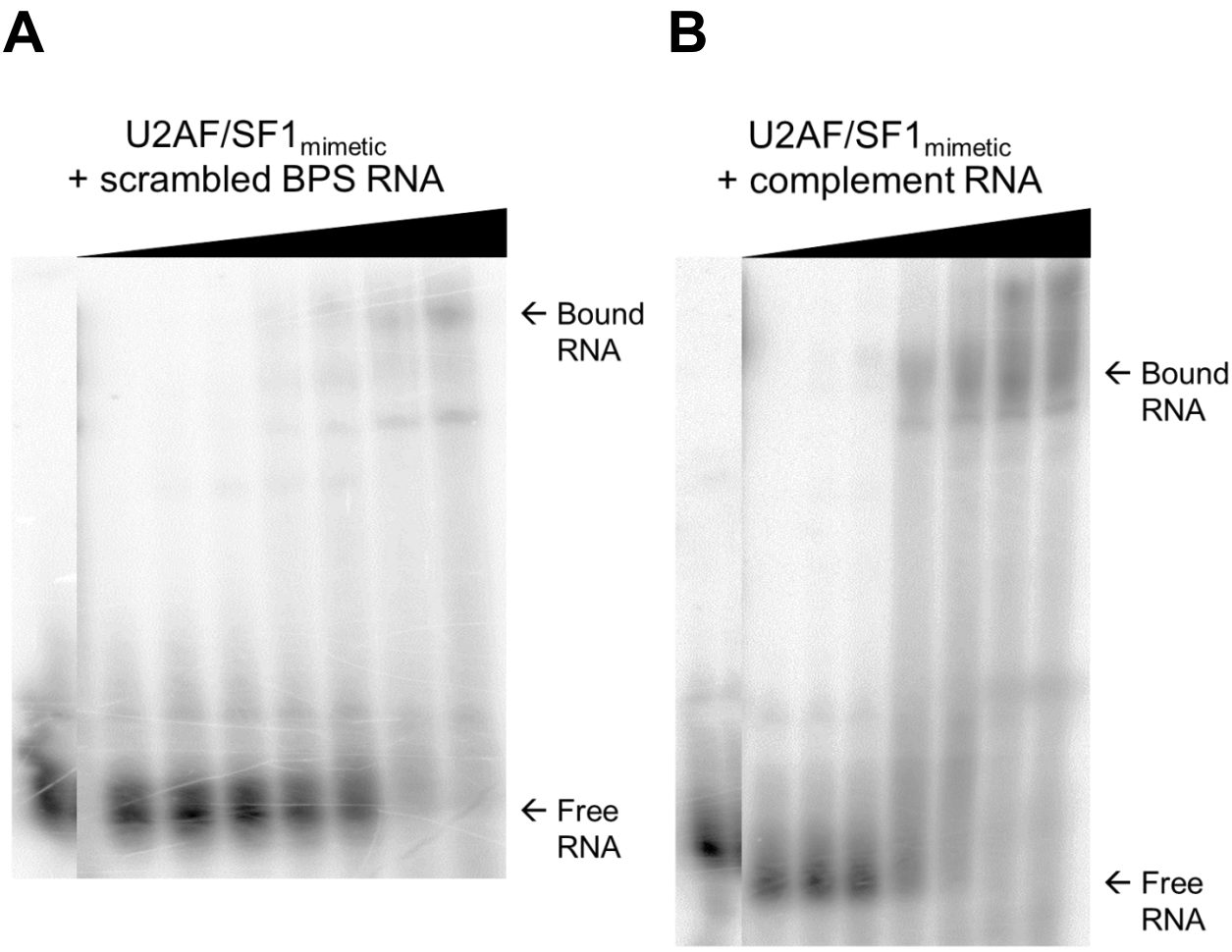


Figure 2-20. EMSAs of U2AF/SF1$_{mimetic}$ trimer bound to scrambled BPS RNA and complement RNA. (A) U2AF/SF1$_{mimetic}$ trimer + scrambled BPS RNA. Titration series: 0, 0.1, 0.5, 1, 5, 10, 50, and 100 μM protein. (B) U2AF/SF1$_{mimetic}$ trimer + complement RNA. Titration series: 0, 0.1, 0.5, 1, 5, 10, 50, and 100 μM protein.

2-3. Discussion

2-3.1. SEC purification of U2AF dimer and U2AF/SF1 trimer

The SDS-PAGE gels with the input samples show that the complex components are present in stoichiometric amounts, and that contamination is minor. Comparing the SDS-PAGE gels with the input samples to the SDS-PAGE gels for the S200 UV traces also reveals that each complex becomes significantly more pure after SEC.

As mentioned previously, each complex shows a symmetrical peak on the S200 UV trace and there is an inverse relationship between complex size and peak retention volume. The SDS-PAGE gels for the S200 UV traces of these complexes also show that the protein components of each complex are stoichiometric across the peak. All of these observations suggest the existence of a stable 1:1 U2AF heterodimer or 1:1:1 U2AF/SF1 heterotrimer.

The contaminant termed 'unidentified species' appears to be of identical size for both U2AF/SF1$_{\Delta Zn,wt}$ and U2AF/SF1$_{wt}$ and co-migrates with the U2AF/SF1 complex in a distinct pattern on the S200 UV trace. It starts to appear when the right half of the peak begins eluting, and as the peak flattens out to the baseline UV reading, the contaminant is an increasing percentage of the total eluted protein. This contaminant appears to associate with the complex, and it could only be completely removed by an anion exchange step using a very gradual gradient.

SEC purification data for these complexes was very promising for a tractable experimental system in order to pursue both biochemical and structural analysis of the U2AF dimer and U2AF/SF1 trimer. All complexes contained a generous amount of protein which is already very pure after only Ni-NTA purification as indicated by the input sample for the S200

runs. Additionally, the complexes appear to be very well-behaved, since they all produce a symmetrical peak containing stoichiometric amounts of the complex components, and the void volume has a low UV reading compared to the protein elution peak. All of the data for all of the complexes indicates a stable, soluble, monodisperse 1:1 heterodimer or 1:1:1 heterotrimer with a low percentage of contamination and aggregation. This is behaviour exhibited by an ideal model system for further study.

2-3.2. Anion exchange chromatography purification of U2AF dimer and U2AF/SF1 trimer

The low MW contaminant referred to as 'unidentified species' likely forms a 1:1:1 heterotrimeric complex with U2AF dimer, since it is stably bound and co-elutes in apparently stoichiometric amounts with the dimer (see Figures 2-6 and 2-7). The reason that it appears to form a complex with the target U2AF/SF1 trimer during SEC is because it likely associates with the U2AF dimer, which has a similar retention volume as the U2AF/SF1 trimer when purified by SEC. Additionally, since both U2AF dimer and U2AF/SF1 trimer elute at ~100 mM NaCl on the Mono Q column, the target U2AF/SF1 trimer complex and the apparent contaminant-bound U2AF dimer complex can only be separated by using a very gradual gradient in order to take advantage of the subtle difference in NaCl concentration required to elute the two complexes so that they can be completely resolved. However, the SDS-PAGE analysis (Fig. 2-7B) shows that contaminant-bound U2AF dimer and the target U2AF/SF1 trimer are not fully resolved at the boundary where the two peaks meet in this particular batch of protein. Because separation of the two complexes is subtle and occurs over a small change in ionic strength of the buffer, significant batch-to-batch variation existed in how cleanly these two complexes were resolved. However, most of the target U2AF/SF1 trimer complex was always recoverable in a completely

pure form from each batch of protein. The Mono Q UV trace and corresponding SDS-PAGE gel resulting from a Mono Q purification step of the U2AF dimer are not shown because this complex elutes as a completely pure sample without any unusual features that merit discussion and interpretation. All U2AF/SF1 trimer complexes used in this thesis show the same two peaks on the Mono Q UV trace, regardless of the protein constructs used.

The low MW contaminant is likely a degradation product of SF1 containing a region of sequence which stably associates with the U2AF dimer. This is because the contaminant is never present in U2AF dimer purifications but is always present in U2AF/SF1 trimer purifications, regardless of the protein constructs present in the complex. This possibility was not investigated however, since it is outside of the scope this thesis.

2-3.3. SEC-MALLS characterization of U2AF dimer and U2AF/SF1 trimer

Prior to using the experimental system to investigate the RNA-binding properties of U2AF23, U2AF59, and SF1, it is necessary to establish that they do in fact form a stable U2AF heterodimer and U2AF/SF1 heterotrimer. Although SEC and SDS-PAGE data may be consistent with the existence of a stable, soluble, and monodisperse U2AF heterodimer and U2AF/SF1 heterotrimer, neither SEC nor SDS-PAGE are absolute methods to establish the existence of these complexes. SDS-PAGE will denature any complex present, therefore it can only establish the existence of the protein constituents of the respective complex, even if they appear as stoichiometric bands on the gel. With respect to SEC, peak retention volume is not directly related to the MW of the macromolecular particles in solution because in addition to MW, peak retention volume is also influenced by the diffusion properties of the sample (Uliyanchenko, 2014).

A requisite for pursuing biochemical and structural experiments aimed at characterizing these complexes is to unambiguously establish their existence. Calculating the MW of the sample downstream of SEC by an absolute method and cross-referencing it with the calculated, sequence-based MW of the respective complex will accomplish this. Combining SEC with MALLS (multi-angle laser light scattering) and dRI (differential refractive index) satisfies this requirement and allows MW determination independent of peak retention volume (Mogridge, 2015; Wen, Arakawa, & Philo, 1996; Wyatt, 1993; Zimm, 1948). In this configuration, SEC is only used to resolve the species in the sample so that they enter the MALLS and dRI detector cells individually, and the peak retention time has no significance in MW determination. SEC-MALLS is considered an absolute method of MW determination, because the instruments are calibrated independently of the SEC column and do not rely on reference standards, and MW is determined directly from first principles based on the following equation:

$$M = \frac{R(0)}{Kc\left(\frac{dn}{dc}\right)^2}$$

M = Molecular weight

R(0) = The reduced Rayleigh ratio (amount of light scattered by the macromolecular particle relative to the laser intensity). This is determined by the MALLS detector and extrapolated to angle zero.

c = weight concentration. This is determined by the dRI detector.

dn/dc = Refractive index increment of the macromolecular particle (essentially the difference between the refractive index of the macromolecular particle and the buffer).

K = A system constant (Wyatt, 1993).

The SEC-MALLS data (Fig. 2-8, Fig. 2-9, and Table 2-1) unambiguously confirmed the existence of a stable and soluble U2AF dimer and U2AF/SF1 trimer. This complements and supports conclusions based on the data derived from other experiments that biochemically and structurally characterize these complexes. Additionally, it should be noted that the SEC-MALLS data was derived from samples that did not go through a final purification step using ion exchange chromatography which indicates that Ni-NTA chromatography, followed by TEV cleavage, and subsequent SEC is sufficient sample processing in order to produce a sample that is pure and monodisperse enough for biochemical characterization using EMSAs. However, the detailed nature of structural characterization and the sensitivity of the data to sample quality required a final purification step using ion exchange chromatography in order to eliminate all detectable contamination.

2-3.4. Spectroscopy of RNA-bound U2AF/SF1 trimer

The existence of a stable, soluble and monodisperse U2AF dimer and U2AF/SF1 trimer is established from first principles by SEC-MALLS, giving credibility to experiments characterizing these complexes. However, these experiments do not establish the existence of a stable and soluble protein/RNA complex, which is the final requisite for the biochemical and structural investigations that follow in this thesis.

In order to establish the existence of a stable protein/RNA complex, U2AF/SF1 trimer was mixed with a stoichiometric excess of RNA, followed by removal of unbound RNA via SEC. The phosphomimetic U2AF/SF1 trimers were selected for this experiment because they are expected to bind RNA more tightly than their wildtype counterparts (Manceau et al., 2006; Y. Zhang et al., 2013). An absorption spectrum was taken of these samples, along with free RNA

and free U2AF/SF1 trimer. An overlay of these three spectra for both phosphomimetic trimers indicates that the RNA-bound complex is stable and therefore the system is tractable for the investigation of the RNA-binding properties of U2AF dimer and U2AF/SF1 trimer.

In addition to the spectroscopy experiments described above, assessing the long-term solubility and stability of the RNA-bound U2AF dimer and U2AF/SF1 trimer complexes is important prior to structural investigation by cryo-EM, SAXS, and crystallography. Therefore, the RNA-bound phosphomimetic U2AF/SF1 trimers were evaluated via their SEC trace and absorption spectrum after 7 days of storage to determine their long-term solution behaviour. They behaved similarly to the fresh samples indicating long-term solubility and stability of the RNA-bound complexes.

In both cases, the RNA-bound complex was stable enough to be purified by SEC, and additionally was also soluble and stable over an extended period of time under the appropriate storage conditions (at least 7 days at 4°C). Confirmation of the existence of a stable RNA-bound U2AF/SF1 trimer satisfied the last requisite before commencing the EMSAs and structural investigations described in this thesis.

2-3.5. EMSAs of U2AF dimer and U2AF/SF1 trimer

The main value of the experiments described in Sections 2-3.1. to 2-3.4. is to establish the feasibility of the experimental system for investigating the RNA binding properties of the U2AF dimer and U2AF/SF1 trimer, in addition to structural characterization. EMSAs were conducted to fulfill the first aim. By allowing the calculation of a K_d (dissociation constant) value for the binding of a specific protein complex to a specific RNA sequence, these experiments provide a means to probe the RNA binding properties of the U2AF dimer and U2AF/SF1 trimer.

The prototype RNA sequence (referred to as 'wildtype') used for EMSAs is the 3′ SS of the lone intron interrupting the ORF of *S. pombe* p14 (PomBase systematic ID: SPBC29A3.07c) with the natural BPS of the intron mutated to the optimal BPS; the other model RNAs are derivatives of this wildtype RNA (Will et al., 2001). This 3′ SS was selected as a model for the EMSAs because it was successfully used in the study which first established the existence of a stable, RNA independent U2AF/SF1 heterotrimer in *S. pombe* (T. Huang et al., 2002). These RNAs were transcribed from a transcription template consisting of two synthetic reverse complementary DNA oligonucleotides duplexed to each other (see Appendix *II*-2, Table *II*-6). In addition to the BPS, PPT, and AG di-nucleotide, all of these model sequences contain several nucleotides of sequence flanking the consensus sequence motifs comprising the 3′ SS which are naturally present in the pre-p14 intron, as well as a 5′ GGG transcription artifact. All of the model RNA sequences are catalogued below in Table 2-3 with a brief description; the boundaries of the sequences representing the BPS, PPT, and AG di-nucleotide are indicated by forward slashes, and the branch A is underlined.

Table 2-3: 3′ SS model RNAs used in EMSAs

Name and sequence (5′→3′)	Description
wildtype: GGGU/UACUAAC/UUUUUUUUU/AG/UGC	Prototype 3′ SS derived from the *S. pombe* pre-p14 intron with an optimal BPS.
U12: GGGU/UACUAAC/UUUUUUUUUUUU/AG/UGC	Identical to wildtype but with a U12 PPT instead of a U9 PPT. •Used to probe the effects of PPT length on RNA binding.
scrambled BPS: GGGU/GUCGCAG/UUUUUUUUU/AG/UGC	Identical to wildtype but with a scrambled BPS not conforming to the consensus BPS sequence. •Used as a negative control for BPS binding. •The scrambled BPS was originally used in the original characterization of p14 (MacMillan et al., 1994).
complement: GGGA/AUGAUUG/AAAAAAAAAA/UC/ACG	Exact forward complement of wildtype. •Used as a comprehensive negative control for RNA binding.

Comparing K_d values of different protein/RNA pairings has the potential to give insights into the RNA binding properties of the U2AF/SF1 complex. The most logical approach for this analysis was to begin with the U2AF dimer and compare K_d values with different RNAs. Next, this analysis was repeated for the U2AF/SF1$_{\Delta Zn,wt}$ and U2AF/SF1$_{\Delta Zn,mimetic}$ trimers, which were then compared to the U2AF dimer. Finally, this analysis was repeated for U2AF/SF1$_{wt}$ and U2AF/SF1$_{mimetic}$ trimers, which were then compared to U2AF/SF1$_{\Delta Zn,wt}$ and U2AF/SF1$_{\Delta Zn,mimetic}$ trimers. Analyzing the data in this order is a systematic, logical approach progressing through a sequence of increasingly complex protein complexes, allowing the isolation of the properties of individual parts of the complex. Several phenomena are immediately obvious from the data. SF1 significantly increases affinity of U2AF/SF1 for the 3′ SS. The dual phosphomimetic point mutations in SF1 are responsible for conferring a preference to the U2AF/SF1 complex for a U9 PPT over a U12 PPT. The zinc knuckles of SF1 also confer this preference, but additionally also increase the affinity of the complex for the 3′ SS. In addition to the observations noted above, the data contains several more critical insights which are not immediately obvious and require a careful analysis of the negative control data as well (see Section 2-3.5.1.).

2-3.5.1. Extended insights into EMSA results

In order to identify patterns in the K_d values that give additional insight into the RNA binding properties of the U2AF dimer and U2AF/SF1 trimer complexes, it is useful to view the K_d values in graphical format (see Fig. 2-21 below). This is followed by a comparison of K_d values from the EMSAs conducted on progressively more complex protein complexes. This approach allows the isolation of the contribution of distinct regions of the U2AF/SF1 complex to RNA binding.

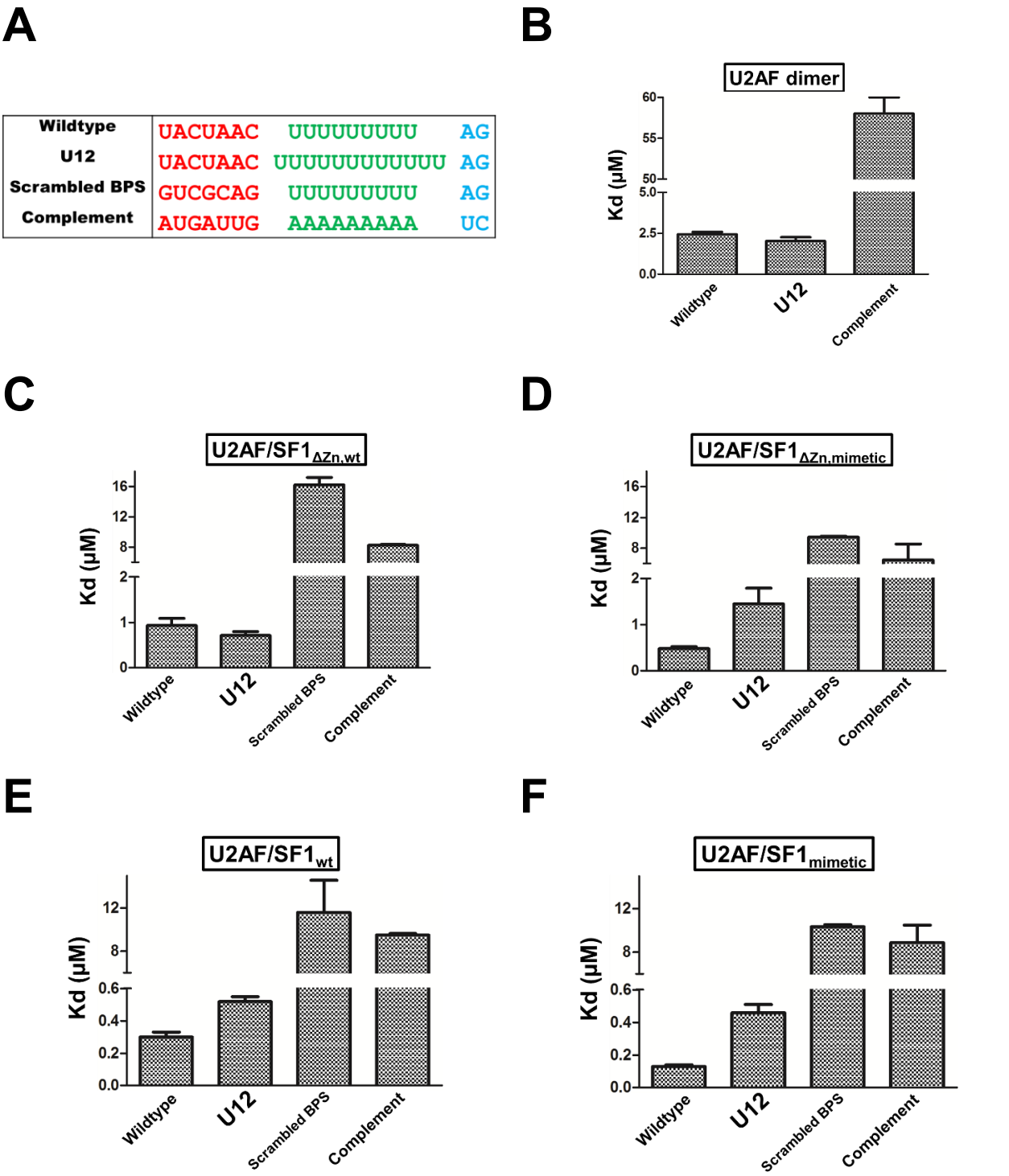


Figure 2-21. EMSA derived K_d values from Table 2-2 displayed in graphical format. (A) Sequence regions in model RNAs corresponding to 3′ SS sequence signatures. The BPS is shown in red, the PPT is shown in green, the AG di-nucleotide is shown in blue. (B) Bar graph of U2AF dimer K_d values. (C) Bar graph of U2AF/SF1$_{\Delta Zn,wt}$ trimer K_d values. (D) Bar graph of U2AF/SF1$_{\Delta Zn,mimetic}$ trimer K_d values. (E) Bar graph of U2AF/SF1$_{wt}$ trimer K_d values. (F) Bar graph of U2AF/SF1$_{mimetic}$ trimer K_d values.

2-3.5.1.1. Comparison of K_d values: U2AF dimer bound to various 3′ SS model RNAs

The K_d values for the U2AF dimer bound to both wildtype and U12 RNA are statistically identical. Additionally, the U2AF dimer binds complement RNA weakly which is not completely unexpected since U2AF59 likely possesses a degree of promiscuity in binding RNA sequences, since its target sequence varies widely in length and sequence composition. Together, these results indicate that U2AF dimer does not discriminate for PPT length and possesses promiscuous RNA binding activity.

2-3.5.1.2. Comparison of K_d values: U2AF/SF1$_{\Delta Zn,wt}$ trimer vs U2AF dimer

Compared to the U2AF dimer, U2AF/SF1$_{\Delta Zn,wt}$ has ~2.7-fold higher affinity for wildtype RNA, ~2.9-fold higher affinity for U12 RNA, and ~7-fold higher affinity for complement RNA. Together, these results indicate that cooperatively, the ULM, KH-QUA2, and phosphorylated domain (in its unphosphorylated state) of SF1 contribute significantly to the affinity of the U2AF/SF1 complex for the 3′ SS.

2-3.5.1.3. Comparison of K_d values: U2AF/SF1$_{\Delta Zn,wt}$ trimer bound to negative control RNAs

The complement RNA sequence fortuitously contains a short sequence (AUGAU) conforming to the *S. pombe* BPS consensus sequence (CURAy) at four out of five nucleotide positions. Therefore, complement RNA is in reality a negative control for U2AF rather than a general negative control for the entire 3′ SS, and as mentioned previously in Chapter 2 the scrambled BPS RNA is a negative control for SF1. Therefore, a comparison of K_d values for complement and scrambled BPS RNA is in essence a comparison of the contribution of SF1 and U2AF to 3′ SS binding in the context of a U2AF/SF1 trimer complex.

U2AF/SF1$_{\Delta Zn,wt}$ binds complement RNA with ~2-fold higher affinity than scrambled BPS RNA. This observation suggests that the binding of SF1 (M52-Q309, wildtype) to its target sequence contributes at least 2-fold more to the affinity of the U2AF/SF1 complex for the 3′ SS than the U2AF dimer binding its target sequence does. Because the fortuitous BPS in complement RNA is a sub-optimal sequence, and because SF1 (M52-Q309, wildtype) is unphosphorylated and missing the zinc knuckles, SF1 binding is expected to show an even greater contribution to 3′ SS binding when a more complete construct is used and with the presence of the optimal UACUAAC BPS sequence.

2-3.5.1.4. Comparison of K_d values: wildtype RNA vs scrambled BPS RNA

Aside from the BPS, the wildtype and scrambled BPS RNAs are identical, and both contain the PPT and AG di-nucleotide which are the target sequences for U2AF dimer. Therefore, U2AF dimer and U2AF/SF1 trimer complexes are expected to bind both wildtype and scrambled BPS RNA with similar affinity. However, the affinity of the four U2AF/SF1 trimer complexes for scrambled BPS RNA is ~3.2-fold to 6.75-fold weaker than the affinity of U2AF dimer for wildtype RNA. Conversely, the affinity of the four U2AF/SF1 trimer complexes for wildtype RNA is ~2.7-fold to 24-fold higher than the affinity of U2AF dimer for wildtype RNA. These results suggest that SF1 may proof-read the binding of U2AF dimer in the context of the U2AF/SF1 trimer through an unknown mechanism, reducing its affinity for its target sequence when the BPS is absent. This is in addition to the role of SF1 being responsible for U2AF/SF1 having significantly higher affinity to the 3′ SS than the U2AF dimer does. Together, these results indicate that SF1 modulates the binding of U2AF dimer to its target sequence and enhances the binding affinity of U2AF/SF1 for the 3′ SS when the BPS is present.

2-3.5.1.5. Comparison of K_d values: U2AF/SF1$_{\Delta Zn,wt}$ trimer vs U2AF/SF1$_{\Delta Zn,mimetic}$ trimer

Neither U2AF dimer nor U2AF/SF1$_{\Delta Zn,wt}$ trimer show a statistically significant difference between their affinity for either wildtype or U12 RNA. However, the affinity of U2AF/SF1$_{\Delta Zn,mimetic}$ for wildtype RNA is 3-fold higher than for U12 RNA. This preference is not the result of an overall increase in affinity for the 3′ SS though. Compared to U2AF/SF1$_{\Delta Zn,wt}$ trimer, U2AF/SF1$_{\Delta Zn,mimetic}$ trimer has 1.8-fold higher affinity for wildtype RNA, but ~2.1-fold lower affinity for U12 RNA. Therefore, the increase in affinity conferred by the dual point mutations in the phosphomimetic complex for one RNA is roughly commensurate with a decrease in affinity for the other. Together, these results indicate that phosphorylation of SF1 confers a preference to the U2AF dimer for shorter PPTs, at least in the context of a polyuridine PPT ranging from 9-12 nucleotides. However, phosphorylation does not confer a generalized increase in binding affinity.

2-3.5.1.6. Comparison of K_d values: U2AF/SF1$_{\Delta Zn,wt}$ and U2AF/SF1$_{\Delta Zn,mimetic}$ trimer vs U2AF/SF1$_{wt}$ trimer

Neither U2AF dimer nor U2AF/SF1$_{\Delta Zn,wt}$ trimer show a statistically significant difference between their affinity for either wildtype or U12 RNA. However, the affinity of U2AF/SF1$_{wt}$ for wildtype RNA is ~1.7-fold higher than for U12 RNA. This preference for wildtype over U12 RNA is not as pronounced as that seen for U2AF/SF1$_{\Delta Zn,mimetic}$ trimer. Together, these results indicate that the zinc knuckles of SF1 confer a preference to the U2AF dimer for shorter PPTs, at least in the context of a polyuridine PPT ranging from 9-12 nucleotides, albeit to a lesser degree than phosphorylation of SF1.

2-3.5.1.7. Comparison of K_d values: U2AF/SF1$_{\Delta Zn,wt}$ vs U2AF/SF1$_{wt}$ trimer

Compared to U2AF/SF1$_{\Delta Zn,wt}$ trimer, U2AF/SF1$_{wt}$ trimer has 3-fold higher affinity for wildtype RNA, and 1.4-fold higher affinity for U12 RNA. Together, these results indicate that the zinc knuckles of SF1 contribute significantly to the affinity of the U2AF/SF1 complex for the 3′ SS.

2-3.5.1.8. Comparison of K_d values: Binding of U2AF/SF1$_{\Delta Zn,wt}$ vs U2AF/SF1$_{\Delta Zn,mimetic}$ trimer to scrambled BPS RNA

Scrambled BPS RNA is identical to wildtype RNA, except that it has no BPS. Therefore, if the preference for a U9 PPT over a U12 PPT conferred by the phosphorylation of SF1 is dependent on the binding of SF1 to the BPS, then the affinity of both U2AF/SF1$_{\Delta Zn,wt}$ and U2AF/SF1$_{\Delta Zn,mimetic}$ trimer for scrambled BPS RNA is expected to be similar. However, compared to U2AF/SF1$_{\Delta Zn,wt}$ trimer, U2AF/SF1$_{\Delta Zn,mimetic}$ trimer has ~1.7-fold higher affinity for scrambled BPS RNA, which is similar to the increase in affinity seen for wildtype RNA when comparing these two U2AF/SF1 complexes. This result indicates that the preference of U2AF dimer for shorter PPTs conferred by phosphorylation of SF1 is independent of SF1 binding the BPS.

2-3.5.1.9. Comparison of K_d values: Binding of U2AF/SF1$_{\Delta Zn,wt}$ and U2AF/SF1$_{wt}$ trimer to scrambled BPS RNA

Scrambled BPS RNA is identical to wildtype RNA, except that it contains no BPS. Therefore, if the preference for a U9 PPT over a U12 PPT and the increase in affinity of the U2AF/SF1 complex for the 3′ SS conferred by the zinc knuckles of SF1 is dependent on the

binding of SF1 to the BPS, then the affinity of both U2AF/SF1$_{\Delta Zn,wt}$ and U2AF/SF1$_{wt}$ trimer for scrambled BPS RNA is expected to be similar. However, compared to U2AF/SF1$_{\Delta Zn,wt}$ trimer, U2AF/SF1$_{wt}$ trimer has ~1.4-fold higher affinity for scrambled BPS RNA. This result indicates that the preference of U2AF dimer for shorter PPTs and/or the increase in affinity of the U2AF/SF1 complex for the 3′ SS conferred by the zinc knuckles of SF1 is independent of SF1 binding the BPS.

2-3.5.1.10. Comparison of K_d values: U2AF/SF1$_{wt}$ and U2AF/SF1$_{\Delta Zn,mimetic}$ trimer vs U2AF/SF1$_{mimetic}$ trimer

The affinities of the U2AF/SF1$_{wt}$ and U2AF/SF1$_{mimetic}$ trimers for U12, scrambled BPS, and complement RNA are statistically identical. However, compared to U2AF/SF1$_{wt}$ trimer, U2AF/SF1$_{mimetic}$ trimer has a 3-fold higher affinity for wildtype RNA. The affinity of U2AF/SF1$_{mimetic}$ trimer for wildtype RNA is 5-fold higher than U12 RNA, which is a more pronounced preference for the U9 PPT over the U12 PPT than which is seen with either the U2AF/SF1$_{\Delta Zn,mimetic}$ or U2AF/SF1$_{wt}$ trimer. This indicates that the combined contribution of phosphorylation of SF1 and the zinc knuckles of SF1 confers a stronger preference of U2AF dimer for shorter PPTs than either feature alone.

2-3.5.1.11. Comparison of K_d values: Binding of U2AF/SF1 trimer complexes to complement RNA

Due to the fortuitous BPS in the complement RNA sequence, it is in essence a negative control for U2AF. Therefore, differences in binding affinity of the four U2AF/SF1 complexes for this RNA are best interpreted as differences in their affinity for the BPS. All four complexes

have very similar affinities for complement RNA, and this result indicates that neither phosphorylation of SF1 nor the zinc knuckles of SF1 significantly contribute to the affinity of SF1 for the BPS.

2-3.5.1.12. Comparison of K_d values: Binding of U2AF/SF1 trimer complexes to negative control RNAs

There is no significant difference between U2AF/SF1$_{\Delta Zn,mimetic}$, U2AF/SF1$_{wt}$, and U2AF/SF1$_{mimetic}$ for either of the negative control RNAs. This result indicates that the cooperative effects exerted by phosphorylation of SF1 and the zinc knuckles of SF1 only operate when both SF1 and U2AF dimer bind their target sequence.

2-3.5.1.13. Summary of extended insights into EMSA results

A careful and systematic dissection of the EMSA results yields several very interesting insights that suggest higher-order principles govern the functions of the U2AF/SF1 complex, whereby the sum whole of the U2AF/SF1 complex and U2AF/SF1/3′ SS complexes is greater than the sum of its parts. These insights require validation through additional experiments and are summarized below.

First, the RNA affinity effects are summarized. With respect to SF1, the region contained within SF1 (M52-Q309, wildtype) contributes at least 2-fold more than the U2AF dimer to the RNA binding affinity of U2AF/SF1. Phosphorylation of SF1 does not confer an increase in binding affinity, but the zinc knuckles of SF1 contribute significantly to the RNA binding affinity of U2AF/SF1.

Interestingly, SF1 appears to modulate the binding activity of U2AF. U2AF dimer does not discriminate for PPT length and possesses promiscuous RNA binding activity. Although U2AF does not discriminate for PPT length, both phosphorylation of SF1 and the zinc knuckles of SF1 confer a preference to U2AF dimer for shorter PPTs, with phosphorylation having a stronger effect. The combination of these two features confers a stronger preference than either feature alone. The aforementioned modulatory effects appear to operate independently of SF1 binding the BPS. Additionally, the increase in affinity of U2AF/SF1 for the 3′ SS due to the zinc knuckles of SF1 is also independent of SF1 binding the BPS. The modulatory effects discussed above are consistent with the observation that neither phosphorylation nor the zinc knuckles of SF1 contribute significantly to the affinity of SF1 for the BPS.

Finally, SF1 appears to proofread the binding of U2AF to its target sequence in the context of U2AF/SF1, since SF1 inhibits U2AF binding when the BPS is absent but enhances it when the BPS is present. This is consistent with the observation that the cooperative effects exerted by phosphorylation and the zinc knuckles of SF1 only operate when both U2AF and SF1 bind their target sequence.

2-4. Materials and Methods

2-4.1. Cloning of U2AF dimer and U2AF/SF1 trimer

2-4.1.1. Preparation of genomic DNA for PCR template

The ORFs for U2AF23 (PomBase systematic ID: SPAP8A3.06) and U2AF59 (PomBase systematic ID: SPBC146.07) do not contain any introns, but the ORF for *S. pombe* SF1

(PomBase systematic ID: SPCC962.06c) contains one short N-terminal intron which is outside of the conserved RNA-binding core and was therefore never taken into consideration. *S. pombe* genomic DNA was used as PCR template for cloning. In order to generate wildtype *S. pombe* genomic DNA, a patch of cells from a -80°C glycerol stock of strain JK484 (*ura4-D18 leu-32 ade6-216 his3-D1*; generous gift from Dr. Jim Karagiannis, Professor, Dept. of Biology, Western University, London, ON, Canada) was struck onto a YES (yeast extract + supplements: adenine, histidine, leucine, and uracil) + agar plate and grown overnight at 30°C (Forsburg & Rhind, 2006; Grewal, Hickmott, Rentas, & Karagiannis, 2012). After overnight incubation, the patch of cells was scraped off the plate using a pipette tip and resuspended into a microfuge tube of ddH_2O (double distilled H_2O). Subsequently genomic DNA was purified using the rapid isolation method for yeast chromosomal DNA (Hoffman, 2001).

2-4.1.2. Cloning U2AF-L into pACYC Duet-1[2]

The coding sequence for U2AF59 was PCR-amplified from wildtype *S. pombe* genomic DNA and subsequently cloned into empty pACYC Duet-1 (Novagen) using BamHI/SacI in order to generate TEV-cleavable (tobacco etch virus-cleavable), wildtype, hexahistidine-tagged U2AF59 (E106-W517).

From N-terminus to C-terminus, the *S. pombe* U2AF59/human U2AF65 chimera consists of: U2AF59 (S93-A161), QSA, U2AF65 (V137-A342), U2AF59 (M394-W517). The coding sequence for U2AF59 was PCR-amplified from previously cloned pACYC Duet-1 U2AF59 (E106-W517, wildtype), and the coding sequence for U2AF65 was PCR-amplified from a U2AF65 cDNA (complementary DNA) clone in pGEX (generous gift from Dr. Michael R.

[2] Appendix *II*-1. contains additional cloning information for U2AF-L constructs. Table *II*-2 catalogues all of the PCRs required to create these constructs, and Table *II*-3 catalogues all of the primers used to clone these constructs.

Green, Chair and Professor, Dept. of Molecular, Cell and Cancer Biology, University of Massachusetts Medical School, Worcester, MA, USA). The final expression construct for the chimera was generated in two sequential cloning steps which both used overlapping PCR in order to join the U2AF59 and U2AF65 regions into an uninterrupted ORF; these two cloning steps are described in Sections 2-4.1.2.1. and 2-4.1.2.2.

Both the U2AF59 and U2AF59/U2AF65 chimera construct are appended at the N-terminus with the following sequence, representing the TEV-cleavable hexahistidine affinity tag: MGSSHHHHHHSQDPENLYFQG.

2-4.1.2.1. First cloning step for the U2AF59/U2AF65 chimera

The first cloning step is the creation of a sub-clone consisting of a partial ORF, in which U2AF65 (V137-A342) and U2AF59 (M394-W517) are joined into a continuous sequence using overlapping PCR primers. In the first overlapping PCR step, these two sections are PCR-amplified separately using the appropriate template. In the second overlapping PCR step, the PCR template is a mixture of the two amplicons from the first step. The second overlapping PCR step joins the two sections into a continuous sequence, which is then cloned into empty pACYC Duet-1 (Novagen) using BamHI/SacI. A visual for this cloning step is provided in Appendix *III*, Section *III*-5., Fig. *III*-2.

2-4.1.2.2. Second cloning step for the U2AF59/U2AF65 chimera

The second cloning step is the creation of the final expression construct, in which U2AF59 (S93-A161) and QSA are joined into a continuous sequence with the sub-clone sequence from the first cloning step using overlapping PCR primers.

In the first overlapping PCR step, these two sections are PCR-amplified separately; U2AF59 (S93-A161) and QSA are PCR-amplified using previously cloned pACYC Duet-1 U2AF59 (E106-W517, wildtype) as template, whereas U2AF65 (V137-A342)/U2AF59 (M394-W517) are PCR-amplified using the sub-clone from first cloning step as template.

In the second overlapping PCR step, the PCR template is a mixture of the two amplicons from the first step. The second overlapping PCR step joins the two sections into a continuous sequence, which is then cloned into empty pACYC Duet-1 (Novagen) using BamHI/SacI. A visual for this cloning step is provided in Appendix *III*, Section *III*-5., Fig. *III*-3.

2-4.1.3. Cloning U2AF23 and SF1 into pET Duet-1[3]

The coding sequence for the desired U2AF23 construct was PCR-amplified from wildtype *S. pombe* genomic DNA and subsequently cloned into empty pET Duet-1 (Novagen) using either NdeI/XhoI or BglII/XhoI.

The coding sequence for SF1 constructs was PCR-amplified from wildtype *S. pombe* genomic DNA and subsequently cloned into pET Duet-1 (Novagen) into which U2AF23 had previously been cloned. SF1 was cloned by digesting the vector with NcoI/EcoRI and the SF1 amplicon with BspHI/EcoRI prior to ligating them together.[4]

Dual S131E and S133E point mutations in the phosphomimetic constructs of SF1 were introduced using overlapping PCR, however all other aspects of cloning these mutants were

[3] Appendix *II*-1. contains additional cloning information for both U2AF23 and SF1 constructs. Table *II*-1 catalogues all of the primers used to clone U2AF23 constructs. Table *II*-4 catalogues all of the PCRs required to create the SF1 constructs, and Table *II*-5 catalogues all of the primers used to clone these constructs.

[4] The isoschizomers NcoI and BspHI must be used in combination with one another in order to clone untagged SF1 because NcoI is necessary to remove the hexahistidine tag in the remaining empty MCS (multiple cloning site) of the pET Duet-1 vector; however, the cloned region of the SF1 ORF contains one NcoI site as well. The use of NcoI/BspHI appends a single methionine at the N-terminus of all SF1 constructs as a cloning artifact. Additionally, SF1 must be cloned into the vector after U2AF23, because all U2AF23 constructs are cloned using XhoI and the cloned region of the SF1 ORF contains one XhoI site.

identical to their wildtype counterparts; successful introduction of the dual point mutations was confirmed by sequencing (The Applied Genomics Core, University of Alberta).

2-4.2. Co-expression and purification of U2AF dimer and U2AF/SF1 trimer

2-4.2.1. Transformation of the *E. coli* expression strain for protein expression and purification

In order to co-express either one of the two complexes, the *E. coli* BL21-Gold expression strain was transformed with either the pACYC Duet-1 U2AF59 construct or the pACYC Duet-1 U2AF59/U2AF65 chimera construct, then plated onto LB (Luria-Bertani, VWR) + agar (Difco) plates containing 0.05 g L^{-1} kanamycin (GoldBio) + 0.034 g L^{-1} chloramphenicol (GoldBio) for plasmid selection. Subsequently, a transformant colony was selected and grown to mid-log phase in liquid culture under the same selection conditions, then used to create competent cells containing the pACYC Duet-1 expression construct.

The competent cells were either transformed with pET Duet-1 U2AF23 for expression of the U2AF dimer or pET Duet-1 (U2AF23 + SF1) for expression of the U2AF/SF1 trimer, then plated onto LB (VWR) + agar (Difco) plates containing 0.025 g L^{-1} kanamycin (GoldBio) + 0.017 g L^{-1} chloramphenicol (GoldBio) + 0.3 g L^{-1} carbenicillin (GoldBio) for plasmid selection. The resulting transformant colonies were used to express protein.

2-4.2.2. Expression of U2AF dimer and U2AF/SF1 trimer

A 25-100 mL volume of LB (VWR) broth containing 0.025 g L^{-1} kanamycin (GoldBio) + 0.017 g L^{-1} chloramphenicol (GoldBio) + 0.3 g L^{-1} carbenicillin (GoldBio) was inoculated with a

transformant colony in order to initiate a pre-culture of the required protein complex and combination of protein constructs, then grown overnight at 37°C and 210 rpm. The following day, the pre-culture was diluted into 4-6 L of LB (VWR) broth containing 0.025 g L^{-1} kanamycin (GoldBio) + 0.017 g L^{-1} chloramphenicol (GoldBio) + 0.3 g L^{-1} carbenicillin (GoldBio) that was pre-warmed to 37°C; at the time of inoculation, the incubator set temperature was fixed at 15°C in order to allow the media to gradually cool down and stabilize at 15°C by the time the cultures were ready to induce; cultures were grown at 210 rpm. Upon reaching OD_{600} (optical density at 600 nm) = 0.6-0.8, protein induction was initiated by adding IPTG (isopropyl β-D-thiogalactopyranoside) to a final media concentration of 0.5 mM and $ZnCl_2$ to a final media concentration of 100 μM, inducing for 15-18 hr at 15°C and 120 rpm.

2-4.2.3. Nickel affinity chromatography purification of U2AF dimer and U2AF/SF1 trimer

E. coli cells were harvested post-induction by centrifugation at 4°C. The pelleted cells were resuspended by stirring on ice in lysis buffer for 30 min (lysis buffer for *S. pombe* complexes = 50 mM Tris-HCl, pH 8.0, 100 mM NaCl, 5 mM BME (β-mercaptoethanol), 20 mM imidazole, 1 mM PMSF (phenylmethylsulfonyl fluoride), 0.1 mg mL^{-1} lysozyme; lysis buffer for U2AF/SF1 trimer containing U2AF59/U2AF65 chimera = 15% (v/v) glycerol, 0.5 M urea, 50 mM Tris-HCl, pH 8.0, 1 M NaCl, 5 mM BME, 25 mM imidazole, 1 mM PMSF, 0.1 mg mL^{-1} lysozyme), and lysed via sonication (Branson 450 Digital Sonifier) on ice for 30 sec at 70% output power followed by cooling the lysate on ice for 1 min; sonication followed by cooling was performed a total of 3x.

Post-sonication, the crude cell lysate was centrifuged at 10,000 rpm for 30 min at 4°C in order to clear the lysate of solid cell debris. The clarified supernatant was then incubated with 2-

3 mL of Ni-NTA (nickel nitrilotriacetic acid) resin (Thermo Scientific) for 45 min at 4°C with gentle rotation in order to keep the Ni-NTA resin in suspension. The protein-bound resin was then cleared of lysate by passing it through a gravity flow column, followed by three washes of 50 mL each with wash buffer at 4°C (20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 5 mM BME, 20 mM imidazole). After washing the resin, bound protein was eluted by gravity flow in elution buffer at 4°C (20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 5 mM BME, 250 mM imidazole). A total of three elutions of 10 mL each were collected and analyzed via SDS-PAGE (6.6% stacking gel, 16% resolving gel) followed by coomassie blue staining. Fractions containing a significant concentration of mostly pure protein were pooled and concentrated to a final volume of 0.5-2.0 mL using an Amicon Ultra Centrifugal Filter Unit with a 30 kDa pore size cutoff (MilliporeSigma) at 4°C; for a batch of protein suitable for further purification and experiments, this is either the first two fractions or all three fractions.

After concentrating the pooled elutions, the experiments that the sample was destined for determined the final purification steps. For all of the biochemical characterization experiments described in this chapter, the protein was dialyzed into imidazole-free buffer (20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 5 mM BME), followed by TEV cleavage (see Section 2-4.2.5.), followed by a final purification via SEC (see Section 2-4.2.4.). Alternatively, if the protein was destined for the structural experiments described in Chapter 3, it was first purified by SEC (see Section 2-4.2.4.), followed by TEV cleavage (see Section 2-4.2.5.), followed by a final purification step via anion exchange chromatography (see Section 2-4.2.6.).

2-4.2.4. SEC purification of U2AF dimer and U2AF/SF1 trimer

Concentrated protein (either pooled, crude affinity chromatography elution fractions or pooled, dialyzed, and TEV cleaved affinity chromatography elution fractions) was separated via SEC by running it over a HiLoad 26/60 S200 (Superdex 200) column (GE Healthcare Life Sciences) in size exclusion buffer (20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 5 mM BME) at 4°C. Fraction volumes were 4 mL each. Protein-containing fractions eluted from the column were analyzed via SDS-PAGE (6.6% stacking gel, 16% resolving gel), followed by coomassie blue staining. Fractions containing over 95% pure, monodisperse protein were pooled and concentrated to a final volume of 0.5-2.0 mL using an Amicon Ultra Centrifugal Filter Unit with a 30 kDa pore size cutoff (MilliporeSigma) at 4°C and the remaining fractions were discarded.

2-4.2.5. TEV cleavage of U2AF dimer and U2AF/SF1 trimer

Concentrated protein (either pooled and dialyzed affinity chromatography elution fractions or pooled elution fractions from SEC) was TEV cleaved by adding 25 µL of TEV at 9.6 mg mL^{-1} and mixing the sample by vortexing. The protein was incubated with TEV overnight at 4°C and complete cleavage of the hexahistidine affinity tag was confirmed by running the cleaved sample alongside its uncleaved counterpart on SDS-PAGE (6.6% stacking gel, 16% resolving gel), followed by coomassie blue staining.

2-4.2.6. Anion exchange chromatography purification of U2AF dimer and U2AF/SF1 trimer

Protein which was purified first by affinity chromatography, followed by SEC and then by TEV cleavage was finally purified by running it over a Mono Q HR 5/5 column (GE

Healthcare Life Sciences) at 4°C using an ionic strength gradient of 0-35% high salt buffer (low salt buffer = 20 mM Tris-HCl, pH 8.0, 5 mM BME; high salt buffer = 20 mM Tris-HCl, pH 8.0, 1M NaCl, 5 mM BME). U2AF dimer complexes were purified over a gradient of 70 column volumes, and U2AF/SF1 trimer complexes were purified over a much more gradual gradient of 350 column volumes. Fraction volumes were 1.5 mL each for both U2AF dimer and U2AF/SF1 trimer. Protein-containing fractions eluted from the column were analyzed via SDS-PAGE (6.6% stacking gel, 16% resolving gel), followed by coomassie blue staining. Fractions containing monodisperse protein with no detectable contamination were pooled and concentrated to a final volume of 0.25-1.0 mL using an Amicon Ultra Centrifugal Filter Unit with a 30 kDa pore size cutoff (MilliporeSigma) at 4°C and the remaining fractions were discarded.

2-4.3. SEC-MALLS characterization of U2AF dimer and U2AF/SF1 trimer

Concentrated protein was separated via SEC by running it over a Superose 12 column (GE Healthcare Life Sciences) in SEC-MALLS buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 5 mM BME) at room temperature. Upon exiting the column, particles were analyzed using a Dawn Heleos II detector (Wyatt Technology), and the ASTRA software (version 5.3.4.14, Wyatt Technology) was used to calculate the molar mass and dRI.

2-4.4. Spectroscopy of RNA-bound U2AF/SF1 trimer

TEV-cleaved U2AF/SF1$_{\Delta Zn,mimetic}$ trimer and U2AF/SF1$_{mimetic}$ trimer were both incubated with a stoichiometric excess of a synthetic, model 3′ SS RNA (CG92; see Appendix *II*, Table *II*-7**)** for 1 hr at room temperature in size exclusion buffer (20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 5 mM BME), followed by SEC over the S200 column in size exclusion buffer. The

fractions of the resulting peak, which had a retention volume similar to the free U2AF/SF1 trimer, were pooled and concentrated using an Amicon Ultra Centrifugal Filter Unit with a 10 kDa pore size cutoff (MilliporeSigma) at 4°C. The absorption spectra of the apo protein, RNA-bound protein, and free RNA for both U2AF/SF1 trimers were quantified across the wavelength range of 250-320 nm and compared.

2-4.5. EMSAs of U2AF dimer and U2AF/SF1 trimer

2-4.5.1. Transcription and purification of radio-labeled model RNAs

Radio-labeled RNA was transcribed from a duplex of two synthetic, reverse complementary, commercially obtained DNA oligonucleotides (see Appendix *II*-2, Table *II*-6). Each transcription reaction contained 0.5 μM forward template, 0.5 μM reverse template, 20 units of T7 RNA polymerase (Thermo Scientific), 80 mM Tris-HCl, pH 8.0, 12 mM NaCl, 6 mM $MgCl_2$, 2 mM spermidine, 10 mM DTT (1,4-dithiothreitol), 0.5 mM UTP, 0.5 mM CTP, 0.5 mM GTP, 15 μM ATP, and 100 μCi α-[^{32}P]-ATP (3000 Ci/mmol, Perkin Elmer). DEPC (diethyl pyrocarbonate) treated ddH_2O was used to bring the reaction to a final volume of 25 μL, and the reaction was incubated for 4 hr at 37°C.

Upon completion, the reaction was mixed with an equal volume of loading dye [0.4 mg mL^{-1} bromophenol blue, 0.4 mg mL^{-1} xylene cyanol, and 6.4 M urea, dissolved in 1x TBE (Tris-borate-EDTA) buffer], then loaded onto a 10% denaturing polyacrylamide gel (20:1 acrylamide/bisacrylamide), which was polymerized in 1x TBE buffer containing 8 M urea. The gel was run at room temperature for 2 hr at 30 watts in 1x TBE buffer. The transcribed RNA band corresponding to the expected size for the product was detected within the gel by

autoradiography, using Kodak X-OMAT film. RNA-containing gel was excised and pulverized by centrifuging it through a perforated PCR tube and collecting the gel fragments in a microfuge tube. Then the RNA was extracted by suspending and incubating the gel fragments for ~16 hr at 37°C with shaking in 440 µL of solution containing 0.3 M sodium acetate and 10% (v/v) phenol. The gel/liquid slurry was then centrifuged through a spin column in order to separate the gel and RNA-containing solution. Finally, the RNA was extracted by a phenol/chloroform/isopentyl alcohol extraction, followed by ethanol precipitation of the RNA. The purified, desiccated RNA pellet was dissolved in ddH_2O and stored at -20°C until used.

2-4.5.2. EMSAs of U2AF dimer and U2AF/SF1 trimer

In preparation for the EMSAs, U2AF dimer and U2AF trimer complexes were first dialyzed into 0.5x buffer D [10 mM HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) pH 7.9, 60 mM KCl, 10% (w/w) glycerol, 0.25 mM DTT]. The binding reactions between protein and transcribed RNA radiolabeled with α-[^{32}P]-ATP were subsequently carried out in 0.5x buffer D, which additionally also contained 2 mM $MgCl_2$, and 10 units of RNaseOUT™ (Thermo Scientific). Binding reactions were incubated for 1 hr at room temperature (O. A. Kent et al., 2003). Upon completion, binding reactions were mixed with an equal volume of loading dye (0.4 mg mL^{-1} bromophenol blue, and 0.4 mg mL^{-1} xylene cyanol, dissolved in 50 mM Tris-glycine buffer, pH 8.6), then loaded onto a 6% native polyacrylamide gel (81:1 acrylamide:bisacrylamide), which was polymerized in 50 mM Tris-glycine buffer, pH 8.6. The gel was run at 4°C for 4 hr, 20 min at 220 V in 50 mM Tris-glycine buffer, pH 8.6.

After electrophoresis, the gel was dried using a Model 583 gel dryer (Bio-Rad), then exposed overnight to a storage phosphor screen (Molecular Dynamics Inc.). The screen was

subsequently scanned using a Typhoon 9400 phosphorimager (Amersham Biosciences) at 633 nm (Filter 390 BP). The resulting image was analyzed using Image Quant (version 5.0, Molecular Dynamics) in order to generate quantifiable data that was subsequently used in order to calculate a K_d value between the protein complex and RNA.

In order to generate data for a K_d value, RNA was incubated with progressively higher protein concentrations until it was present exclusively in a protein-bound state; three trials of this titration experiment were completed for each protein/RNA combination under investigation, and this data was processed and analyzed using GraphPad Prism (version 5.01). EMSAs of wildtype and U12 RNA in combination with all five protein complexes used a titration range of 0-20 μM protein. EMSAs of complement RNA in combination with U2AF dimer used a titration range of 0-300 μM protein. EMSAs of scrambled BPS and complement RNA in combination with the U2AF/SF1$_{\Delta Zn,wt}$ and U2AF/SF1$_{wt}$ trimer complexes used a titration range of 0-175 μM protein. EMSAs of scrambled BPS and complement RNA in combination with the U2AF/SF1$_{\Delta Zn,mimetic}$ and U2AF/SF1$_{mimetic}$ trimer complexes used a titration range of 0-100 μM protein.

The relative intensity of the protein-bound RNA band and free RNA band to each another in each binding reaction were established by subtracting the background noise and these two relative intensities were then used to calculate the percentage of RNA in the binding reaction that was present in its protein-bound state. The mean and standard error were used for statistical and graphical analysis of experimental trials by GraphPad Prism and the binding affinities were fit to a non-linear regression curve using the equation ‘one-site specific binding with Hill slope’ in GraphPad Prism. Outliers were removed by the automatic outlier elimination fitting method.

Chapter 3

Structural characterization of the U2AF dimer and U2AF/SF1 trimer[5]

[5] Processing and analysis of the SEC-SAXS data presented in Chapter 3 was completed by Dr. Ross A. Edwards (Research Associate, Dept. of Biochemistry, University of Alberta, Edmonton, AB, Canada).

3-1. Introduction

Biochemical characterization of both the U2AF dimer and U2AF/SF1 trimer indicates that this is an ideal model system for atomic level structural characterization, therefore a wide variety of screening strategies were used in order to identify a crystallization hit for both the U2AF dimer and U2AF/SF1 trimer complexes (summarized in Section 3-1.1.). However, since all efforts to obtain an atomic resolution structure were unsuccessful an alternative strategy was necessary to structurally characterize the U2AF/SF1 complex. First, SAXS was used to generate a molecular envelope of both the *S. pombe* U2AF dimer and U2AF trimer complexes in their free and RNA-bound states (summarized in Section 3-1.2.). This serves as a proof of concept for atomic level characterization, as well as for more advanced SAXS-based modeling techniques. However, the traditional *ab initio* modelling technique is not appropriate to use if the target possesses conformational flexibility and the modular, multidomain structure of U2AF dimer and U2AF/SF1 trimer requires the use of a technique that will account for the conformational flexibility of these complexes. Therefore, the SAXS experiments were followed by SEC-SAXS experiments using the chimeric U2AF/SF1 complex in its free and RNA-bound states and the scattering data obtained from these experiments were combined with rigid body modeling techniques using previously solved structures described in Chapter 1 to generate ensembles of pseudo-atomic structures of the chimeric U2AF/SF1 trimer in both its free and RNA-bound states (summarized in Section 3-1.3.). SEC purification immediately prior to SAXS data collection improves data quality, which is important in advanced SAXS experiments where problems in the data potentially exert a greater downstream influence in data analysis. Firstly, SEC eliminates all potential aggregates that may form during shipping and storage. Secondly, for the RNA-bound complexes RNA was mixed in excess; SEC removes all unbound RNA, thereby

allowing data collection from 100% pure RNA-bound protein without any contaminating free protein or RNA.

3-1.1. Summary of crystallization strategies used for the U2AF dimer and U2AF/SF1 trimer complex

Several dozen different commercially available crystallization screens were attempted. Screening was attempted using both the vapor diffusion and microbatch crystallization under oil techniques. Several permutations were attempted for both techniques such as varying crystallization drop size, the ratio of protein to precipitant in the crystallization drop, attempting vapor diffusion using both sitting and hanging drops, etc. Crystallization was attempted at 4°C, room temperature, and 37°C. Protein concentration in the sample solution was varied from 0.2 mg mL^{-1} to ~50 mg mL^{-1}.

In addition to attempting to change the kinetic properties of the crystallization experiment and using a wide array of crystallization screens, various chemical additives were used. Crystallization was attempted with and without spermine, spermidine, as well as with and without $MgCl_2$. Polyamines such as spermidine positively affect the crystallization of nucleic acids by decreasing charge repulsion and stabilization of the structure, whereas Mg^{2+} positively affects the crystallization of nucleic acids by stabilization of the structure (Ferre-D'Amare & Doudna, 2001; Rould, Perona, & Steitz, 1991; Sauter et al., 1999). Because of the polyanionic nature of RNA, all of the crystallization conditions include some cations, most often Mg^{2+} and spermine. Spermine has been used for the overwhelming majority of successful RNA and RNA-protein complex crystallizations; it should initially be included in all screens for an RNA-containing crystallization target (Ferre-D'Amare & Doudna, 2001).

Finally, many different protein constructs were attempted in order to modify potential crystallization contacts. Attempts were made to crystallize both the U2AF dimer and U2AF/SF1 trimer in both their apo and RNA-bound states, using different RNAs and DNA/RNA hybrids varying widely in length and sequence composition. Various truncations at the N- and C-termini of U2AF59, U2AF23, and SF1 were attempted. This includes all of the constructs and complexes described in this thesis, as well as U2AF/SF1 complexes comprised of additional constructs not described in this thesis. Since phosphorylation of SF1 is known to rigidify the U2AF/SF1 complex and since the EMSA experiments described in Chapter 2 indicate that the phosphomimetic dual point mutation of *S. pombe* SF1 increases the affinity for the 3′ SS as expected, U2AF/SF1 complexes containing phosphomimetic SF1 were also used for crystallization trials. Additionally, since U2AF-L binds sequences varying widely in length and nucleotide composition, the chimeric complex was used in conjunction with DNA/RNA hybrids in order to fix the binding register of the RRMs.

In line with using construct modifications to change potential crystallization contacts, a well-established strategy to crystallize target proteins that refuse to do so is to attempt crystallization of the orthologues from other species, since the sequence differences are often found on the surface of the protein, potentially having a dramatic effect on crystallization (Campbell et al., 1972; Dale, Oefner, & D'Arcy, 2003). It has been observed that some proteins crystallize easily under various conditions, whereas their orthologues are very difficult to crystallize (Dale et al., 2003). The genus *Schizosaccharomyces* currently consists of 5 species: *S. pombe*, *S. octosporus*, *S. japonicus*, *S. cryophilus*, and *S. osmophilus* (Borneff, 1959; Brysch-Herzberg, Jia, Seidel, Assali, & Du, 2022; Brysch-Herzberg et al., 2019; Helston, Box, Tang, & Baumann, 2010; Lindner, 1893; Yukawa & Maki, 1931). Average amino acid identity between

all 1:1 orthologues between *S. pombe* vs *S. cryophilus* is 66%, whereas for *S. pombe* vs *S. japonicus* it is 55% (Helston et al., 2010; Rhind et al., 2011). For these reasons, the U2AF/SF1 complexes from *S. cryophilus* and *S. japonicus* were cloned, expressed, and purified using the design principles and protocols used for the *S. pombe* complex. These complexes were subsequently used for crystallization trials without success.

3-1.2. Overview of the SAXS characterization of *S. pombe* U2AF dimer and U2AF/SF1 trimer

SAXS cannot be used to generate an atomic model of the complexes but is useful and important because the scattering data used to generate the SAXS envelopes yields insights into the solution properties of the protein and protein/RNA complexes, therefore serves as a proof of concept for more detailed atomic level structural characterization through methods such as X-ray crystallography and cryo-EM. Six complexes were characterized via SAXS and are summarized below in Table 3-1, with sequence regions in model RNAs corresponding to 3′ SS sequence signatures colour-coded as follows: the BPS is shown in red, the PPT is shown in green, and the AG di-nucleotide is shown in blue.

Table 3-1: Protein and protein/RNA complexes characterized via SAXS

Complex	3′ SS model RNA sequence
Apo U2AF dimer	N/A
U2AF dimer + RNA CG120	UUUUUUUUUAGUGC
Apo U2AF/SF1 trimer	N/A
U2AF/SF1 trimer + RNA CG92	UUACUAACUUUUUUUUUAGUGC
U2AF/SF1 trimer + RNA CG109	UAUACUAACAAUUUUUUUUUUUUUUAGUGC
U2AF/SF1 trimer + RNA CG158	CUAACUUUUUUUUUAG

The synthetic 3′ SS model RNAs used in these experiments are catalogued in Appendix *II*, Table *II*-7. The prototype 3′ SS model RNA is named CG92 and was previously described in Chapter 2. The SAXS structure of RNA-bound U2AF dimer was derived using CG120, which is the CG92 sequence with the optimal BPS deleted. The SAXS structure of RNA-bound U2AF/SF1 trimer was derived using CG92 as well as two other model RNAs, CG109 and CG158. CG109 differs from CG92 in that the optimal BPS is flanked at the 5′ end by UA and at the 3′ end by AA to correspond with the optimal BPS RNA used to solve both the human RNA-bound SF1 structure, and the *S. cerevisiae* RNA-bound Bbp structure (Jacewicz et al., 2015; Z. Liu et al., 2001). Additionally, CG109 has a U14 PPT, which is longer than the U9 PPT of CG92. Finally, RNA 158 is an abbreviated version of CG92 in which the BPS is shortened to CUAAC, which conforms to the *S. pombe* BPS consensus sequence (CURAy) with no additional nucleotides. Additionally, the sequence 3′ to the AG di-nucleotide in CG158 is deleted.

3-1.3. Overview of the SEC-SAXS characterization of chimeric U2AF/SF1 trimer

SEC-SAXS scattering data was collected using the chimeric U2AF/SF1 trimer in its free state, as well as bound to all four RNAs in Table 3-1 above. However, the samples containing CG120 and CG158 behaved poorly in solution and were not carried further to the modeling stage of analysis. BilboMD was subsequently used to generate a structure ensemble of each of the three remaining complexes from the scattering data. BilboMD is a standalone web server that allows the determination of a structure ensemble of potential conformations of the U2AF/SF1 complexes. In order to accomplish this a molecular dynamics approach is used, in which known domains and domain interfaces are treated as rigid bodies, which are tethered to each other by flexible linkers. The structure ensemble is derived from a library of models that sample the

potential conformational space of the structure; these potential structures are consistent with the organization of the rigid bodies and flexible linkers with respect to one another, as well as with the scattering data (Pelikan, Hura, & Hammel, 2009).

This modelling approach cannot provide a true atomic model of the U2AF/SF1 complex. With respect to a 3D molecular model, the SAXS scattering data only contains enough data to produce a molecular envelope. Additionally, BilboMD creates model libraries based on the assumption that higher order structural organization does not exist beyond the rigid bodies used in modelling, so if higher order structural features exist in the U2AF/SF1 complex, they will not be revealed. Nonetheless, this modelling approach is a useful intermediate point between a SAXS envelope and a true atomic structure and provides useful insights into the structural features of the complex.

3-2. Results

3-2.1. SAXS characterization of *S. pombe* U2AF dimer and U2AF/SF1 trimer

3-2.1.1. Experimental SAXS curves

The experimental SAXS curves for each sample in the concentration/exposure series of apo and RNA-bound U2AF dimer are shown in Fig. 3-1, and the experimental SAXS curves for each sample in the concentration/exposure series of apo and RNA-bound U2AF/SF1 trimer are shown in Fig. 3-2.

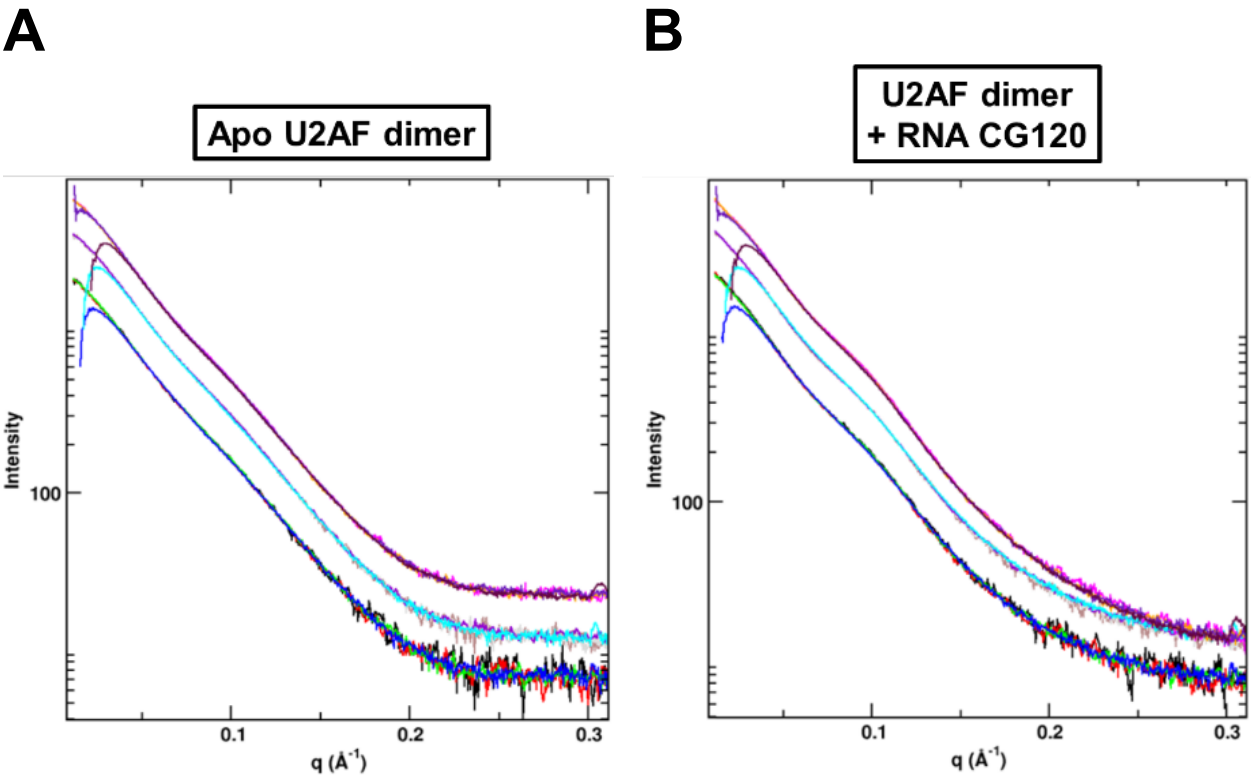


Figure 3-1. Experimental SAXS curves for apo and RNA-bound U2AF dimer. (A) Apo U2AF dimer. (B) U2AF dimer + RNA CG120.

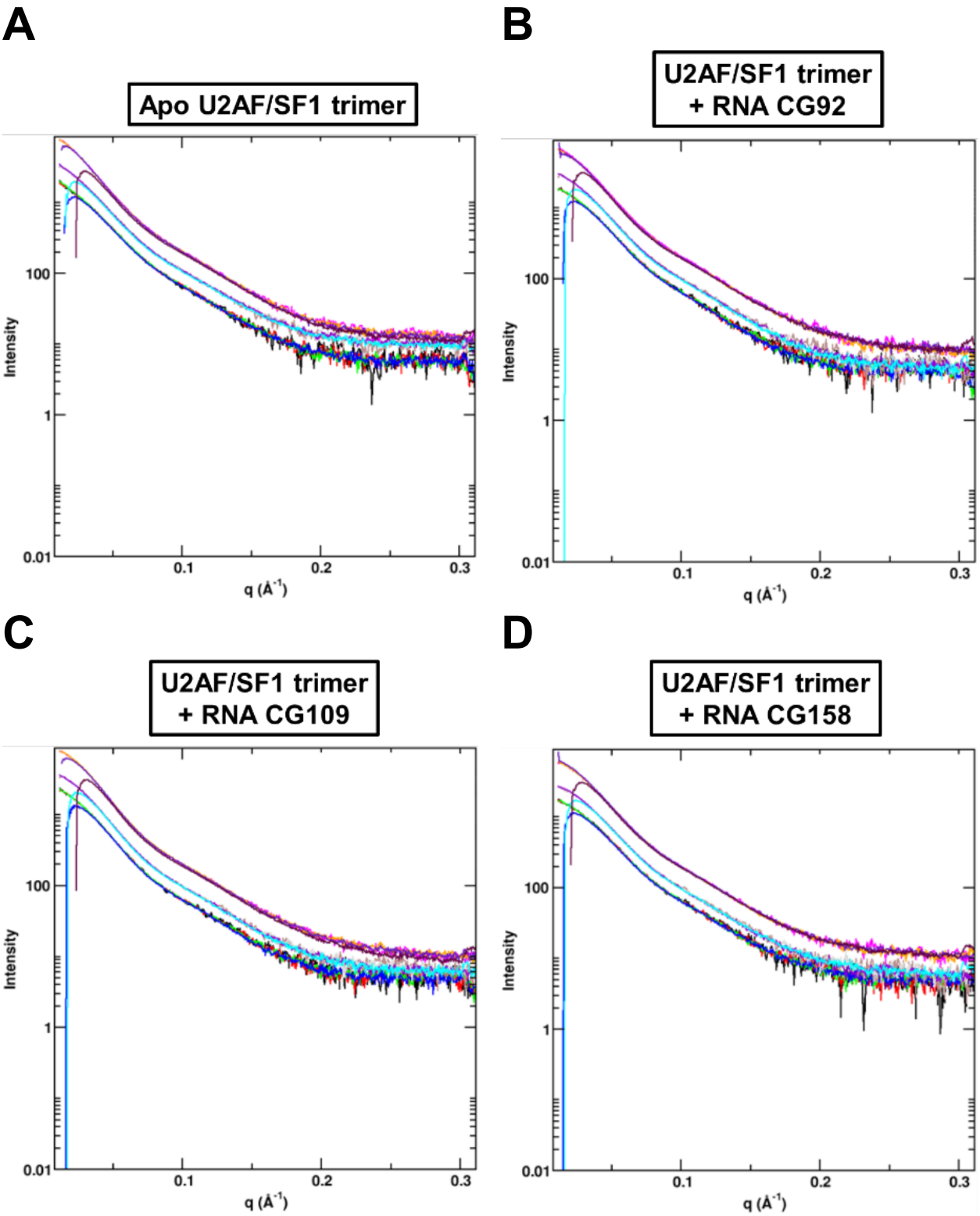


Figure 3-2. Experimental SAXS curves for apo and RNA-bound U2AF/SF1 trimer. (A) Apo U2AF/SF1 trimer. (B) U2AF/SF1 trimer + RNA CG92. (C) U2AF/SF1 trimer + RNA CG109. (D) U2AF/SF1 trimer + RNA CG158.

3-2.1.2. Guinier analysis

The statistics generated by PRIMUS for the linear Guinier plot of the selected sample for each of the six complexes are displayed below in Table 3-2. The radius of gyration calculated from the Guinier plot and its standard deviation is shown under Rg. The range of points used as the Guinier interval for the best obtained Guinier fit is shown under Points. Additionally, the sRg limits, forward scattering intensity measured at zero scattering angle I(0), and data quality estimate are shown.

Table 3-2: Guinier plot statistics

Complex	Rg (Å)	Points	sRg limits	I(0)	Quality
Apo U2AF dimer	42.79 ± 0.05 (0%)	10 to 27 (18)	0.742 to 1.184	4182.25	92%
U2AF dimer + RNA CG120	42.23 ± 0.13 (0%)	12 to 28 (17)	0.783 to 1.194	4426.48	89%
Apo U2AF/SF1 trimer	49.14 ± 0.28 (1%)	5 to 22 (18)	0.702 to 1.21	2045.28	93%
U2AF/SF1 trimer + RNA CG92	44.68 ± 0.02 (0%)	11 to 29 (19)	0.802 to 1.291	2000.7	88%
U2AF/SF1 trimer + RNA CG109	49.74 ± 0.14 (0%)	1 to 24 (24)	0.59 to 1.286	8822.37	96%
U2AF/SF1 trimer + RNA CG158	46.61 ± 0.11 (0%)	2 to 27 (26)	0.581 to 1.29	6605.81	96%

The linear Guinier plots corresponding to Table 3-2 are shown below in Fig. 3-3 to 3-5. Fig. 3-3 shows the Guinier plots for apo U2AF dimer and U2AF dimer + RNA CG120. Fig. 3-4 shows the Guinier plots for apo U2AF/SF1 trimer and U2AF/SF1 trimer + RNA CG92. Fig. 3-5 shows the Guinier plots for U2AF/SF1 trimer + RNA CG109 and U2AF/SF1 trimer + RNA CG158.

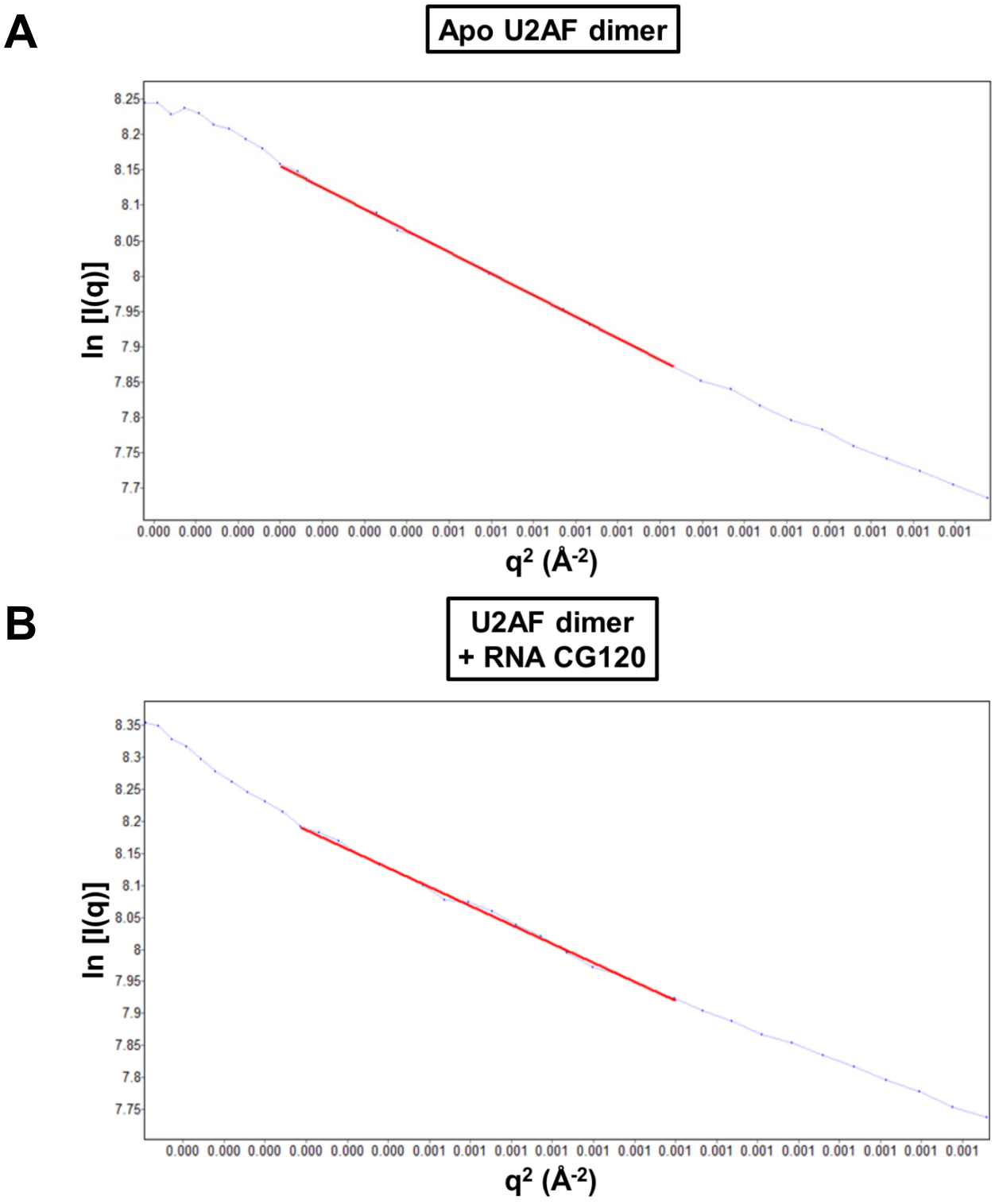


Figure 3-3. Guinier plots for apo U2AF dimer and U2AF dimer + RNA CG120. The best obtained Guinier fit is plotted as a red line. (A) Apo U2AF dimer. (B) U2AF dimer + RNA CG120.

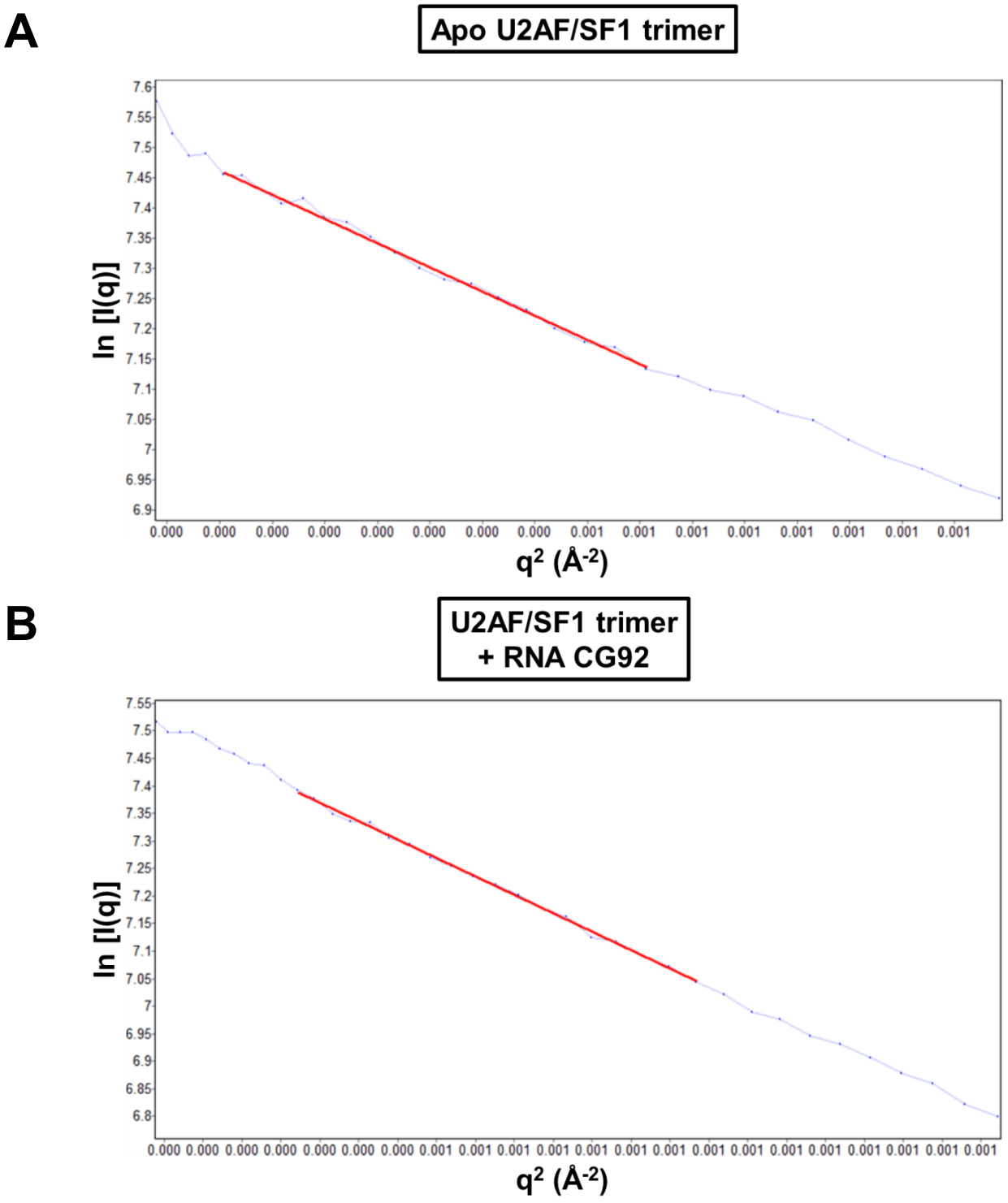


Figure 3-4. Guinier plots for apo U2AF/SF1 trimer and U2AF/SF1 trimer + RNA CG92. The best obtained Guinier fit is plotted as a red line. (A) Apo U2AF/SF1 trimer. (B) U2AF/SF1 trimer + RNA CG92.

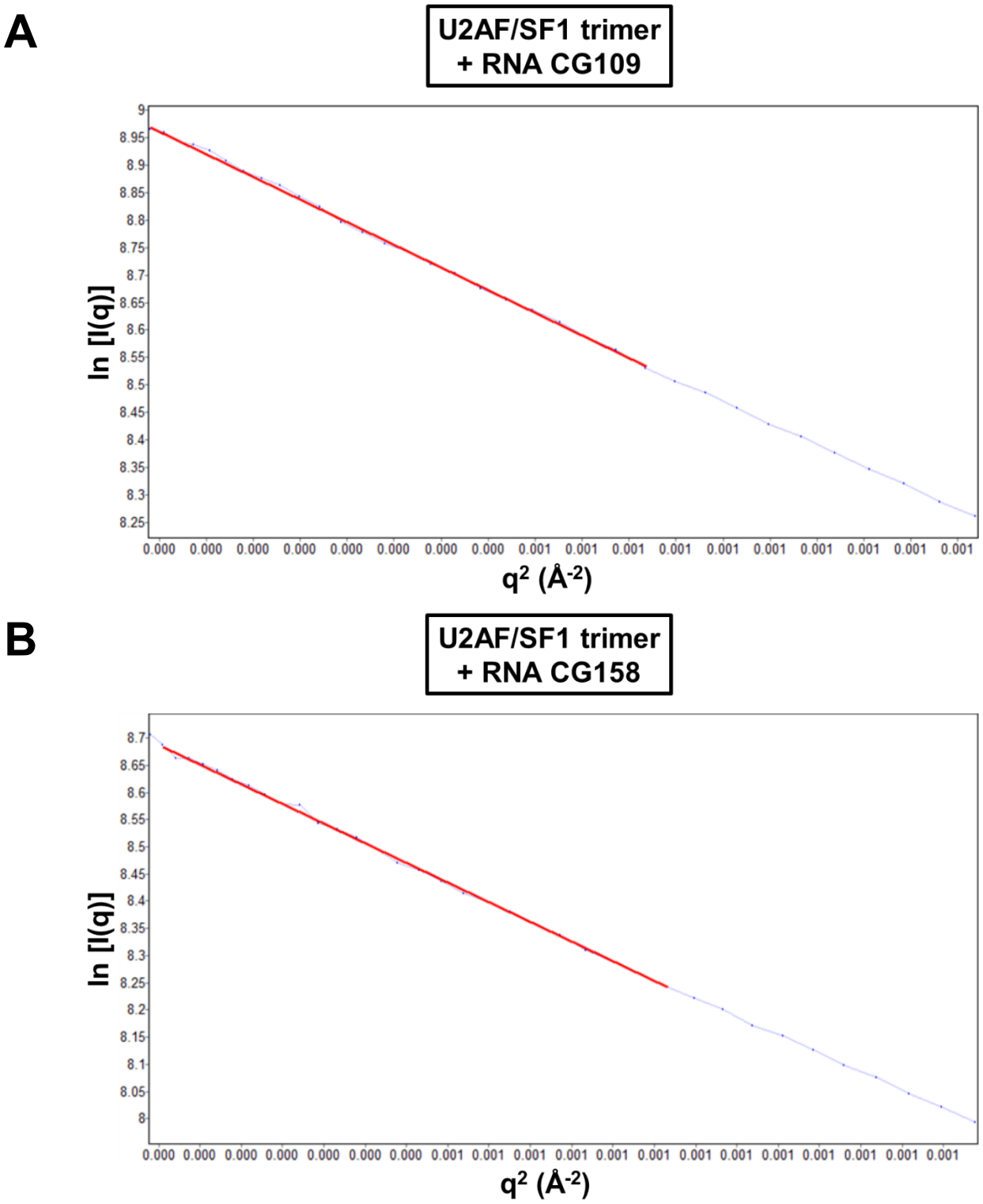


Figure 3-5. Guinier plots for U2AF/SF1 trimer + RNA CG109 and U2AF/SF1 trimer + RNA CG158. The best obtained Guinier fit is plotted as a red line. (A) U2AF/SF1 trimer + RNA CG109. (B) U2AF/SF1 trimer + RNA CG158.

3-2.1.3. Kratky analysis

The Kratky plots of apo and RNA-bound U2AF dimer are shown in Fig. 3-6, and the Kratky plots of apo and RNA-bound U2AF/SF1 trimer are shown in Fig. 3-7.

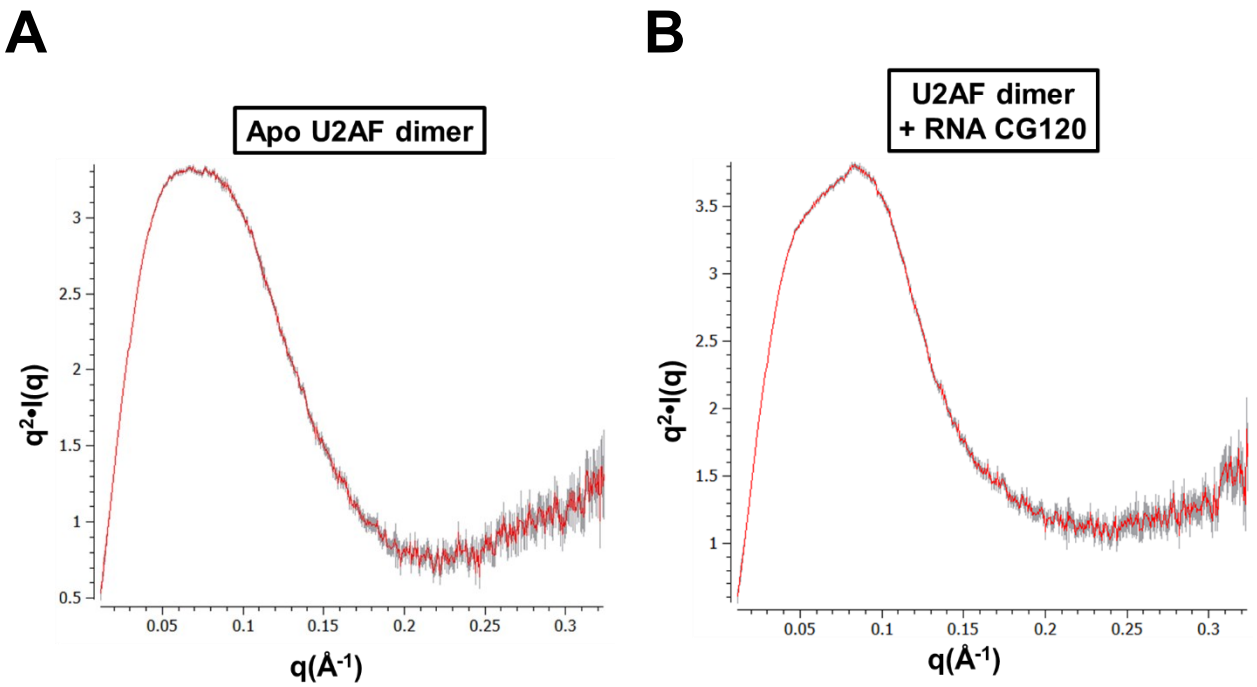


Figure 3-6. Kratky plots for apo and RNA-bound U2AF dimer. (A) Apo U2AF dimer. (B) U2AF dimer + RNA CG120.

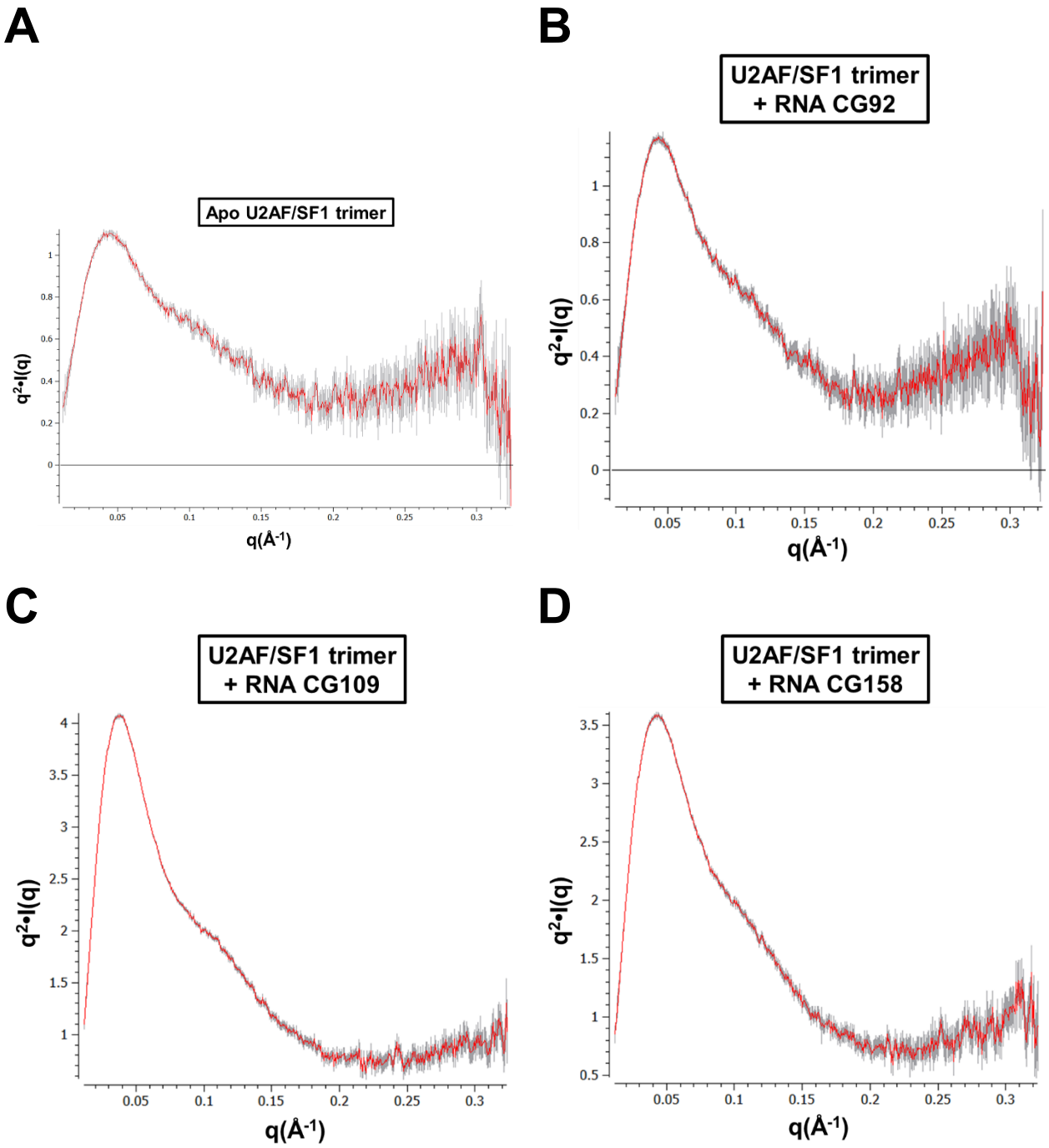


Figure 3-7. Kratky plots for apo and RNA-bound U2AF/SF1 trimer. (A) Apo U2AF/SF1 trimer. (B) U2AF/SF1 trimer + RNA CG92. (C) U2AF/SF1 trimer + RNA CG109. (D) U2AF/SF1 trimer + RNA CG158.

3-2.1.4. *Ab initio* shape reconstruction of complexes using dummy residue modelling

The D_{max} and Rg values corresponding to the P(r) function are shown below in Table 3-3.

Table 3-3: P(r) function statistics

Complex	D_{max} (Å)	Rg (Å)
Apo U2AF dimer	156.7	42.81
U2AF dimer + RNA CG120	157.5	42.23
Apo U2AF/SF1 trimer	177.5	49.14
U2AF/SF1 trimer + RNA CG92	146.9	44.68
U2AF/SF1 trimer + RNA CG109	153.8	49.74
U2AF/SF1 trimer + RNA CG158	152	46.61

All six complexes produce a bell-shaped P(r) function representative of a folded structure (not shown). This and previously shown results show that these complexes satisfy the necessary initial quality checks for the scattering data to be suitable for *ab initio* shape reconstruction. The final calculated envelopes of apo and RNA-bound U2AF dimer are shown in Fig. 3-8, and the final calculated envelopes of apo and RNA-bound U2AF/SF1 trimer are shown in Fig. 3-9.

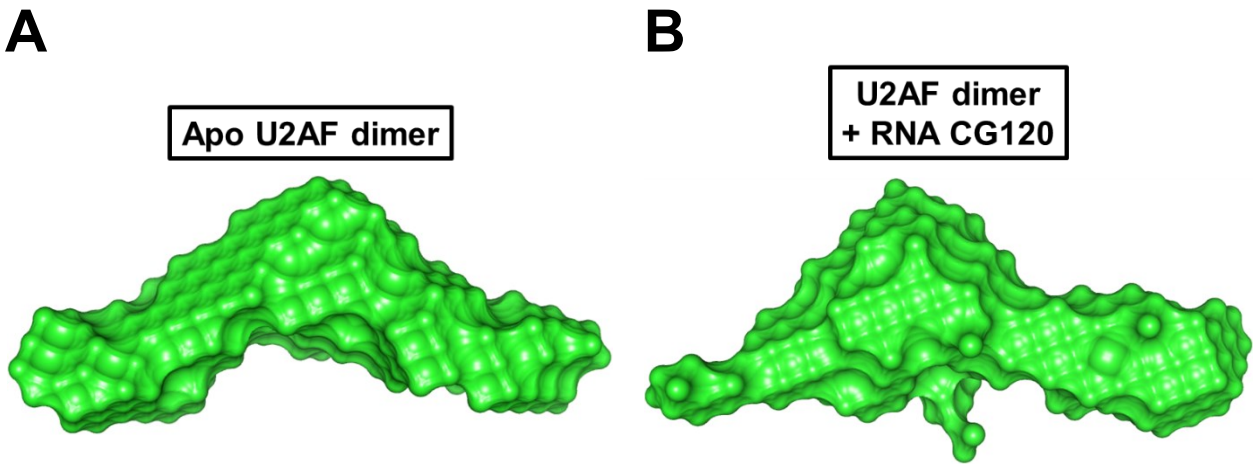


Figure 3-8. *Ab initio* shape reconstruction for apo and RNA-bound U2AF dimer. Models shown are the surface representation of the filtered average of 10 independently derived models generated from GASBOR. (A) Apo U2AF dimer. (B) U2AF dimer + RNA CG120.

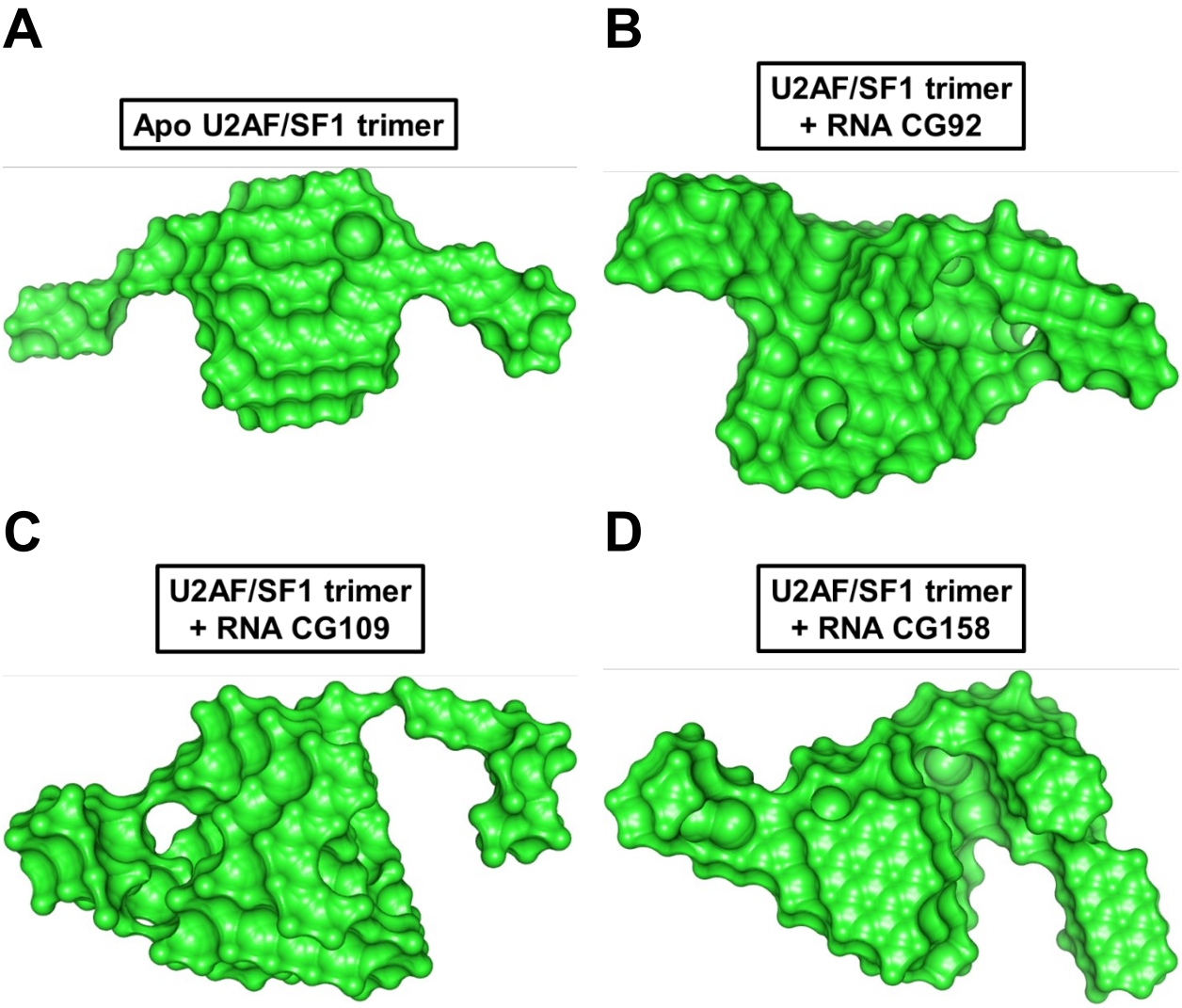


Figure 3-9. *Ab initio* shape reconstruction for apo and RNA-bound U2AF/SF1 trimer. Models shown are the surface representation of the filtered average of 10 independently derived models generated from GASBOR. (A) Apo U2AF/SF1 trimer. (B) U2AF/SF1 trimer + RNA CG92. (C) U2AF/SF1 trimer + RNA CG109. (D) U2AF/SF1 trimer + RNA CG158.

3-2.2. SEC-SAXS characterization of U2AF/SF1 chimera

3-2.2.1. SAXS analysis

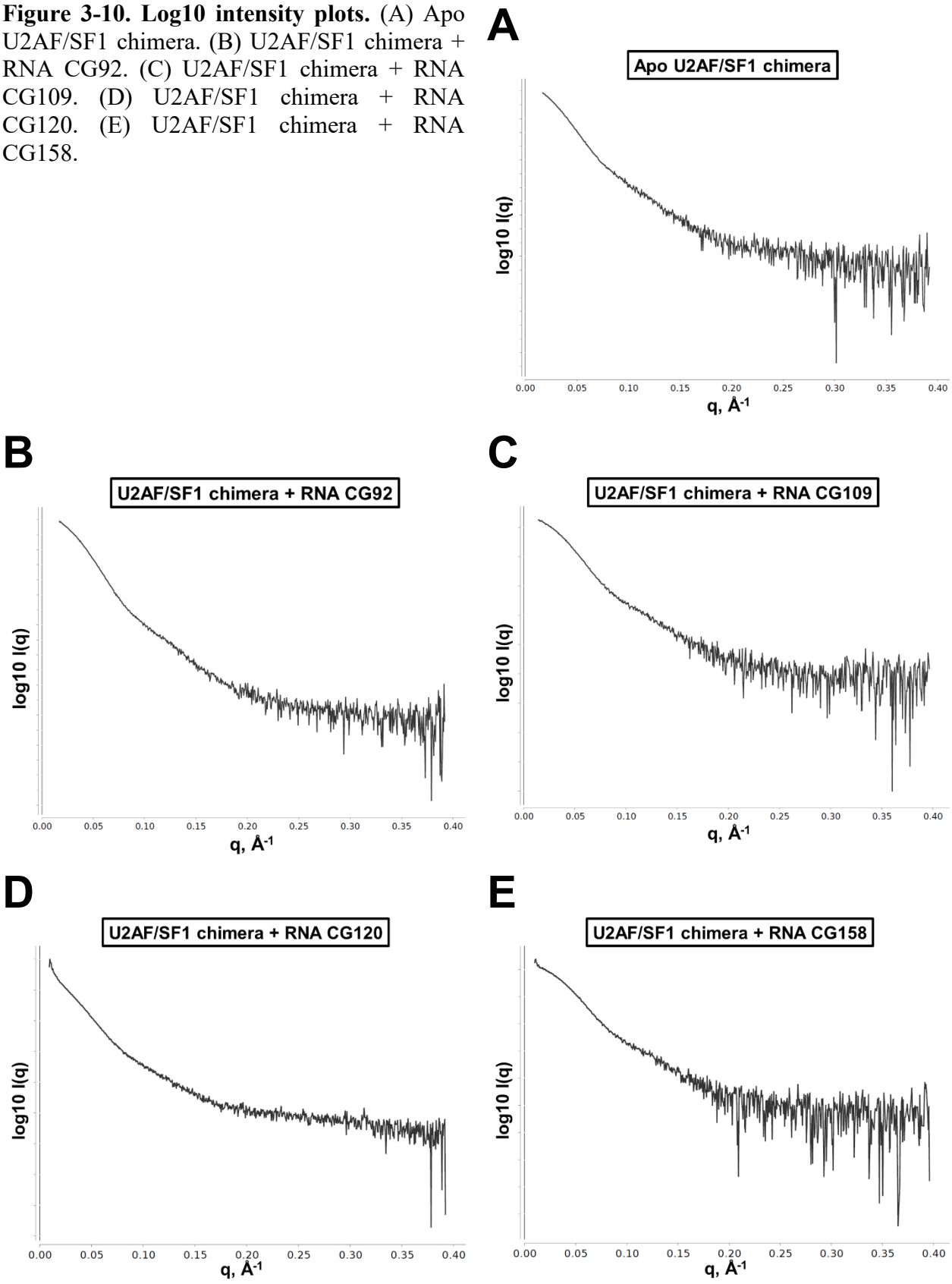
SAXS data of the chimeric U2AF/SF1 trimer complex in both its apo and RNA-bound states were used to model inter-domain interactions and conformational flexibility of the respective complexes in solution. The RNA-bound samples containing either RNA CG120 or CG158 behaved poorly in solution and were not carried forward for further analysis. The calculated SAXS parameters of the five samples are summarized below in Table 3-4.

Table 3-4: Calculated SAXS parameters

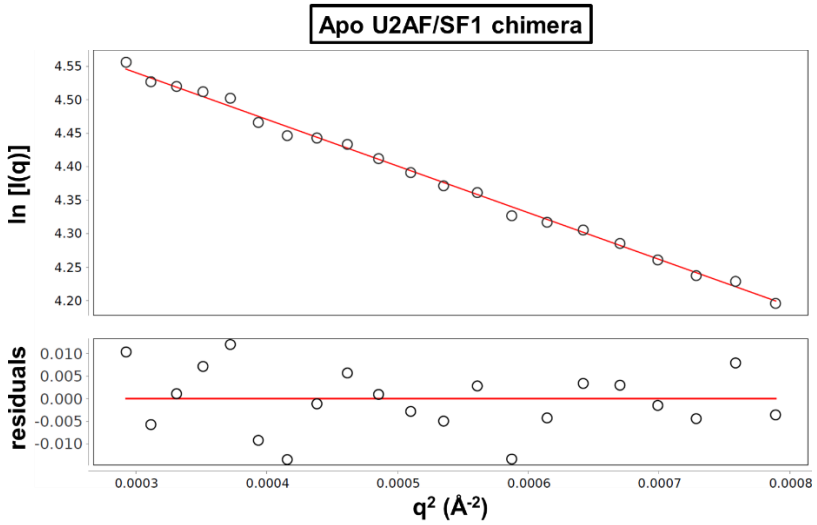
Bound RNA	apo	CG92	CG109	CG120	CG158
Starting data point	4	4	9	88	93
Ending data point	238	220	224	250	265
Start q ($Å^{-1}$)	0.017	0.017	0.014		
End q ($Å^{-1}$)	0.147	0.137	0.134		
Rg reci. (Å)	45.4	41.9	40.7	Not determined	Not determined
Rg real (Å)	44.7	40.7	40.8	Not determined	Not determined
r ave (Å)	57.3	52.6	52.7	Not determined	Not determined
D_{max} (Å)	148	139	136	Not determined	Not determined
Porod Exponent	2.6	2.9	2.5	2.9	2.6
Vp ($Å^3$)	510000	410000	536000	534000	447000

The SAXS plots are shown below (Fig. 3-10 to 3-21). Fig. 3-10 summarizes the experimental SAXS curves (Log10 intensity plots). The Guinier fitting plots (Fig. 3-11, 3-12), Kratky plots (Fig. 3-13), Porod plots (Fig. 3-14), Porod-Debye plots (Fig. 3-15), Kratky-Debye plots (Fig. 3-16), and SIBYLS plots (Fig. 3-17) are also summarized below. Fig. 3-18 and 3-19 summarize the P(r) plots; no P(r) plots were generated for RNA-bound samples containing either RNA CG120 or CG158. Fig. 3-20 and 3-21 summarize the P(r) fit plots; no P(r) fit plots were generated for RNA-bound samples containing either RNA CG120 or CG158.

Figure 3-10. Log10 intensity plots. (A) Apo U2AF/SF1 chimera. (B) U2AF/SF1 chimera + RNA CG92. (C) U2AF/SF1 chimera + RNA CG109. (D) U2AF/SF1 chimera + RNA CG120. (E) U2AF/SF1 chimera + RNA CG158.



A



B

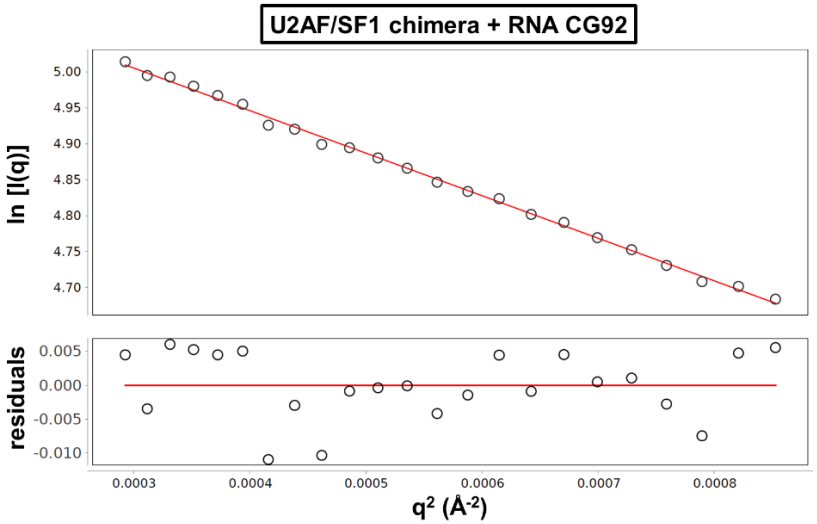


Figure 3-11. Guinier fitting plots. (A) Apo U2AF/SF1 chimera. (B) U2AF/SF1 chimera + RNA CG92.

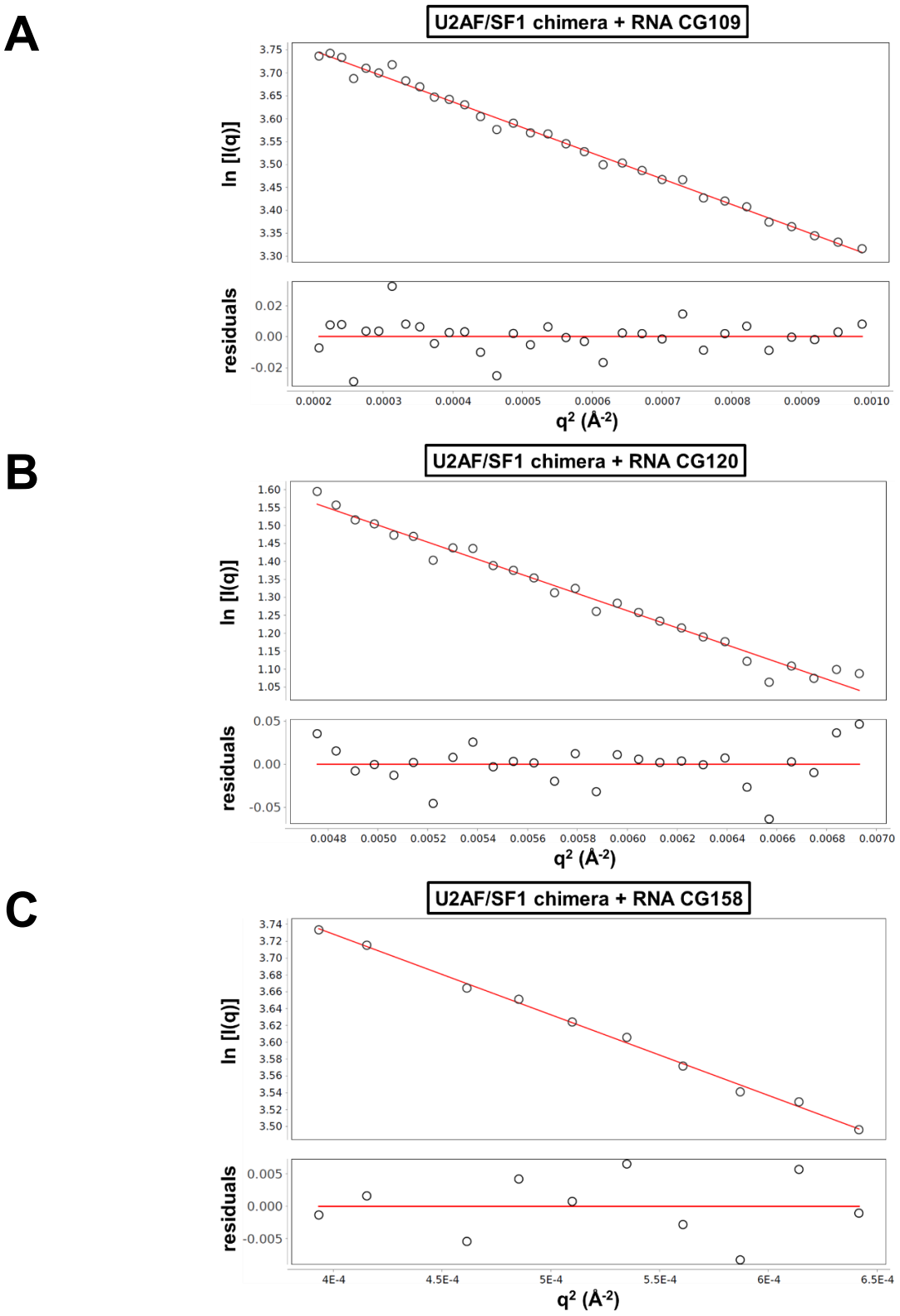


Figure 3-12. Guinier fitting plots. (A) U2AF/SF1 chimera + RNA CG109. (B) U2AF/SF1 chimera + RNA CG120. (C) U2AF/SF1 chimera + RNA CG158.

Figure 3-13. Kratky plots. (A) Apo U2AF/SF1 chimera. (B) U2AF/SF1 chimera + RNA CG92. (C) U2AF/SF1 chimera + RNA CG109. (D) U2AF/SF1 chimera + RNA CG120. (E) U2AF/SF1 chimera + RNA CG158.

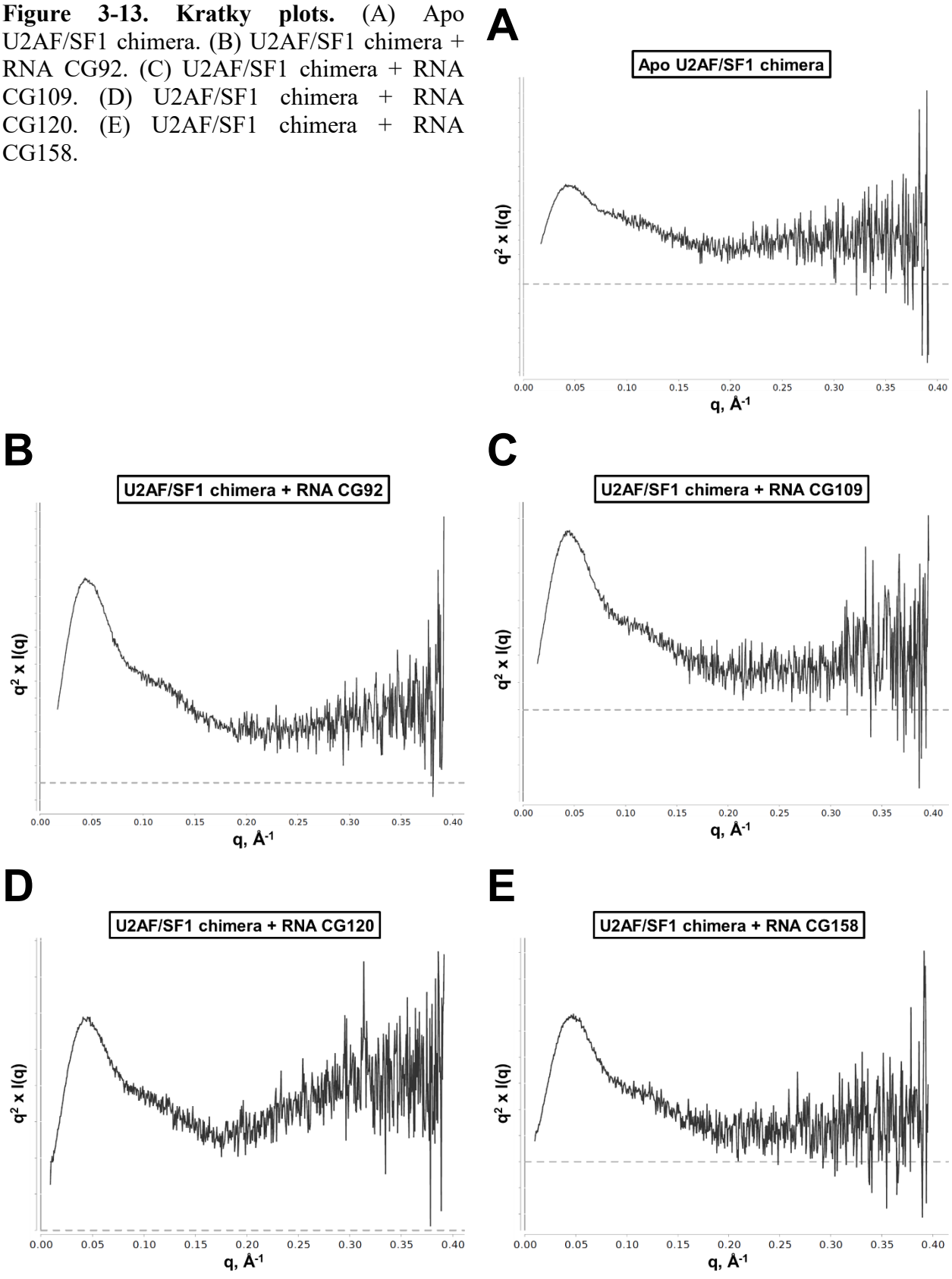


Figure 3-14. Porod plots. (A) Apo U2AF/SF1 chimera. (B) U2AF/SF1 chimera + RNA CG92. (C) U2AF/SF1 chimera + RNA CG109. (D) U2AF/SF1 chimera + RNA CG120. (E) U2AF/SF1 chimera + RNA CG158.

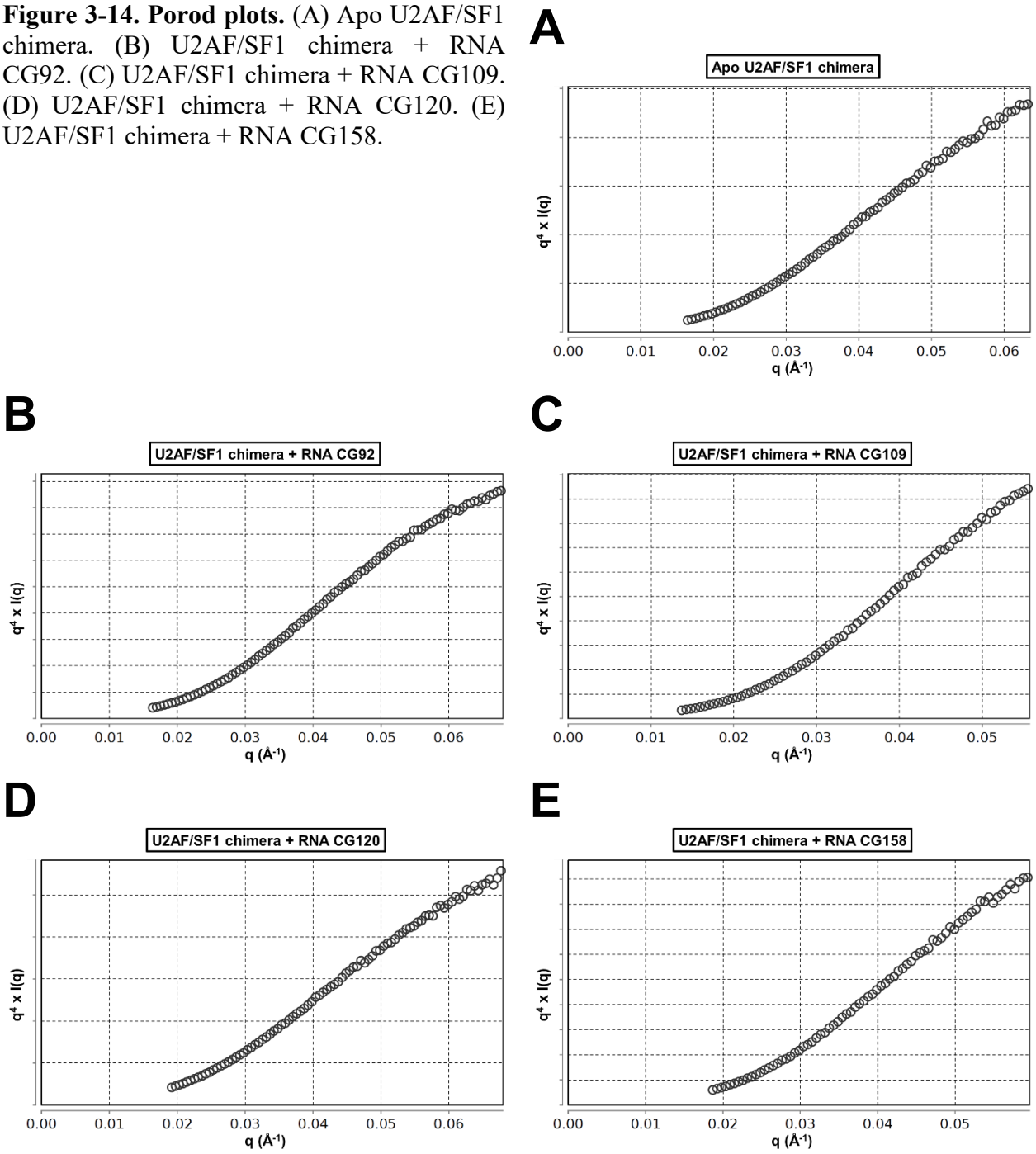


Figure 3-15. Porod-Debye plots. (A) Apo U2AF/SF1 chimera. (B) U2AF/SF1 chimera + RNA CG92. (C) U2AF/SF1 chimera + RNA CG109. (D) U2AF/SF1 chimera + RNA CG120. (E) U2AF/SF1 chimera + RNA CG158.

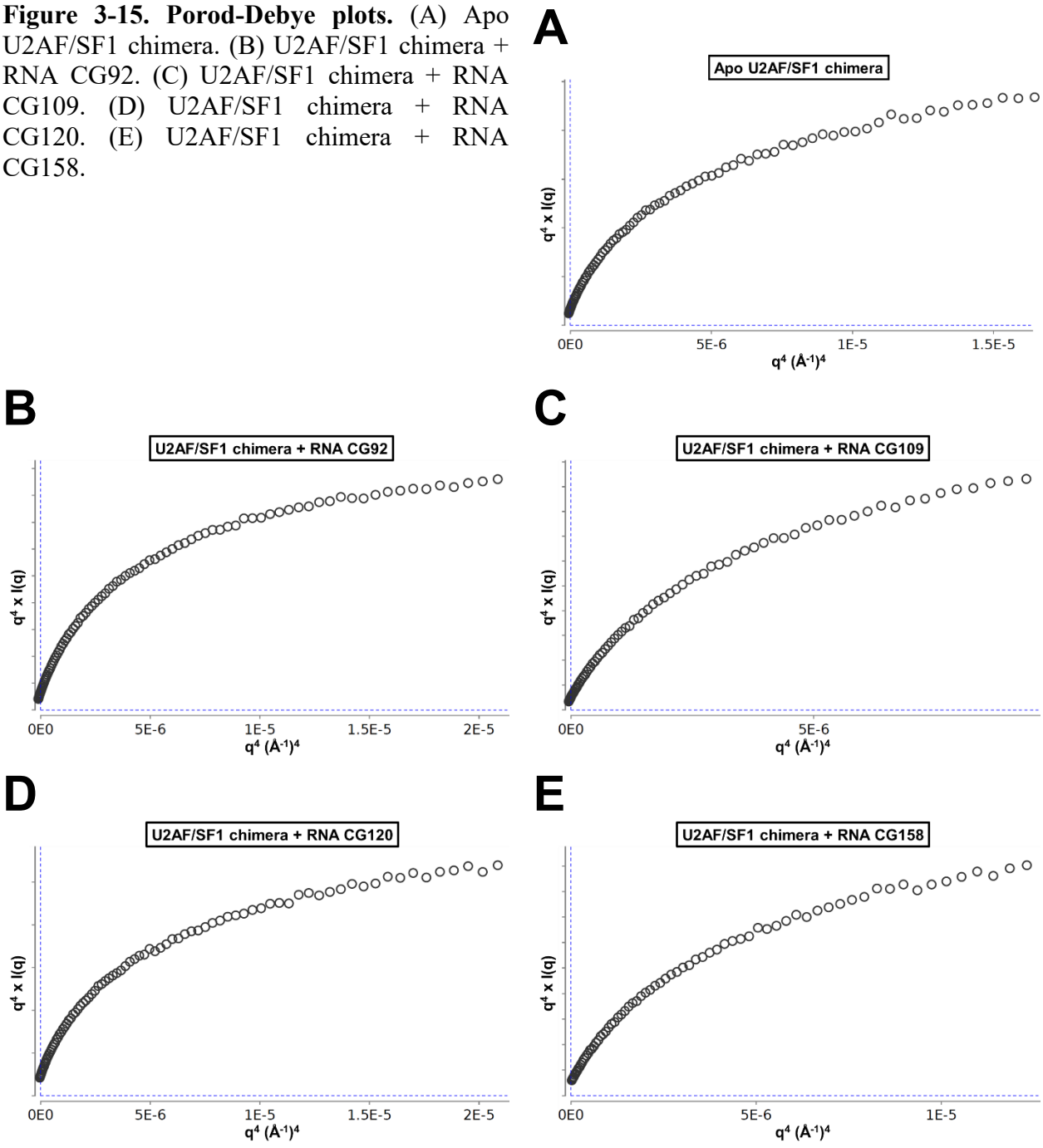


Figure 3-16. Kratky-Debye plots. (A) Apo U2AF/SF1 chimera. (B) U2AF/SF1 chimera + RNA CG92. (C) U2AF/SF1 chimera + RNA CG109. (D) U2AF/SF1 chimera + RNA CG120. (E) U2AF/SF1 chimera + RNA CG158.

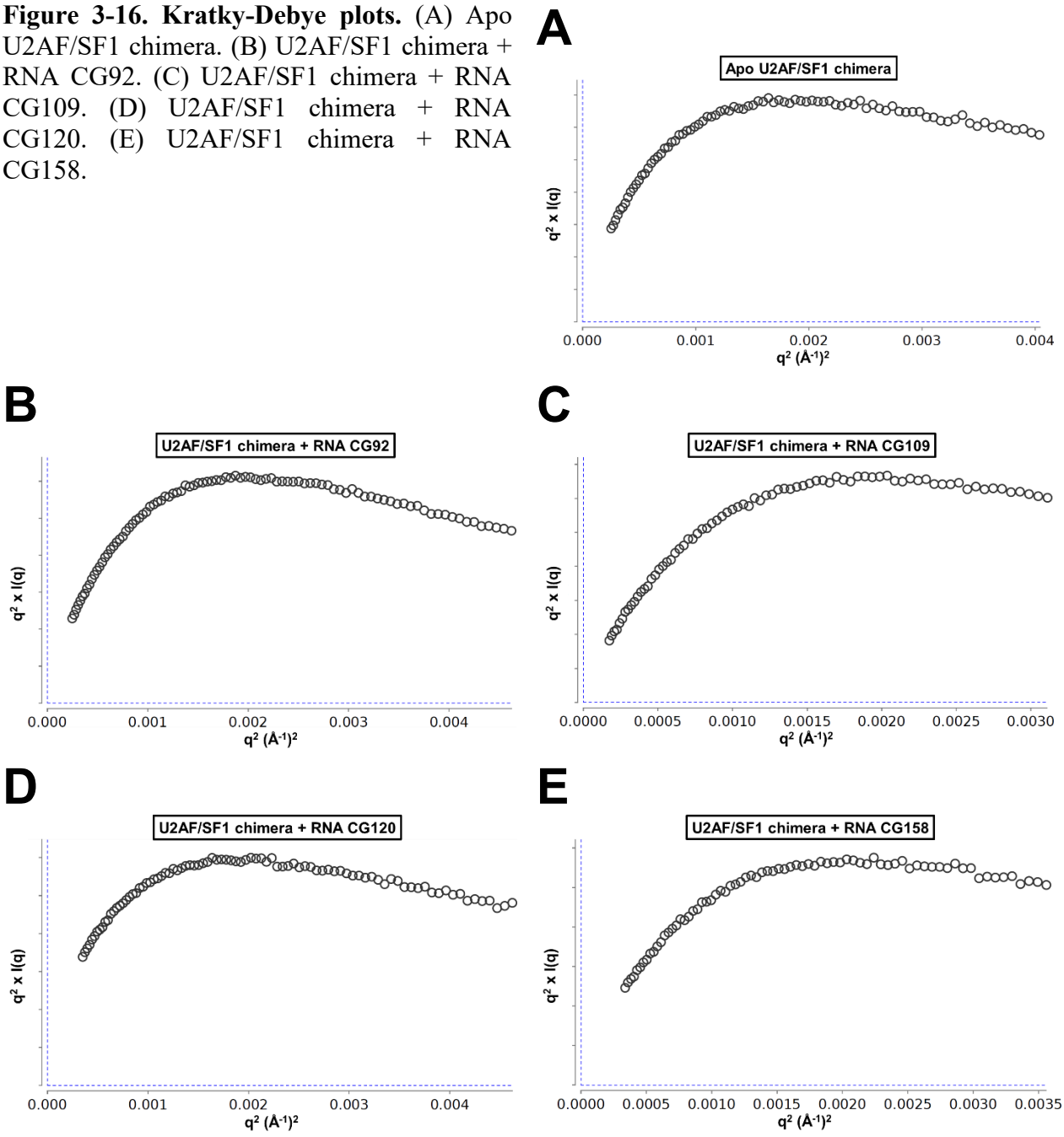
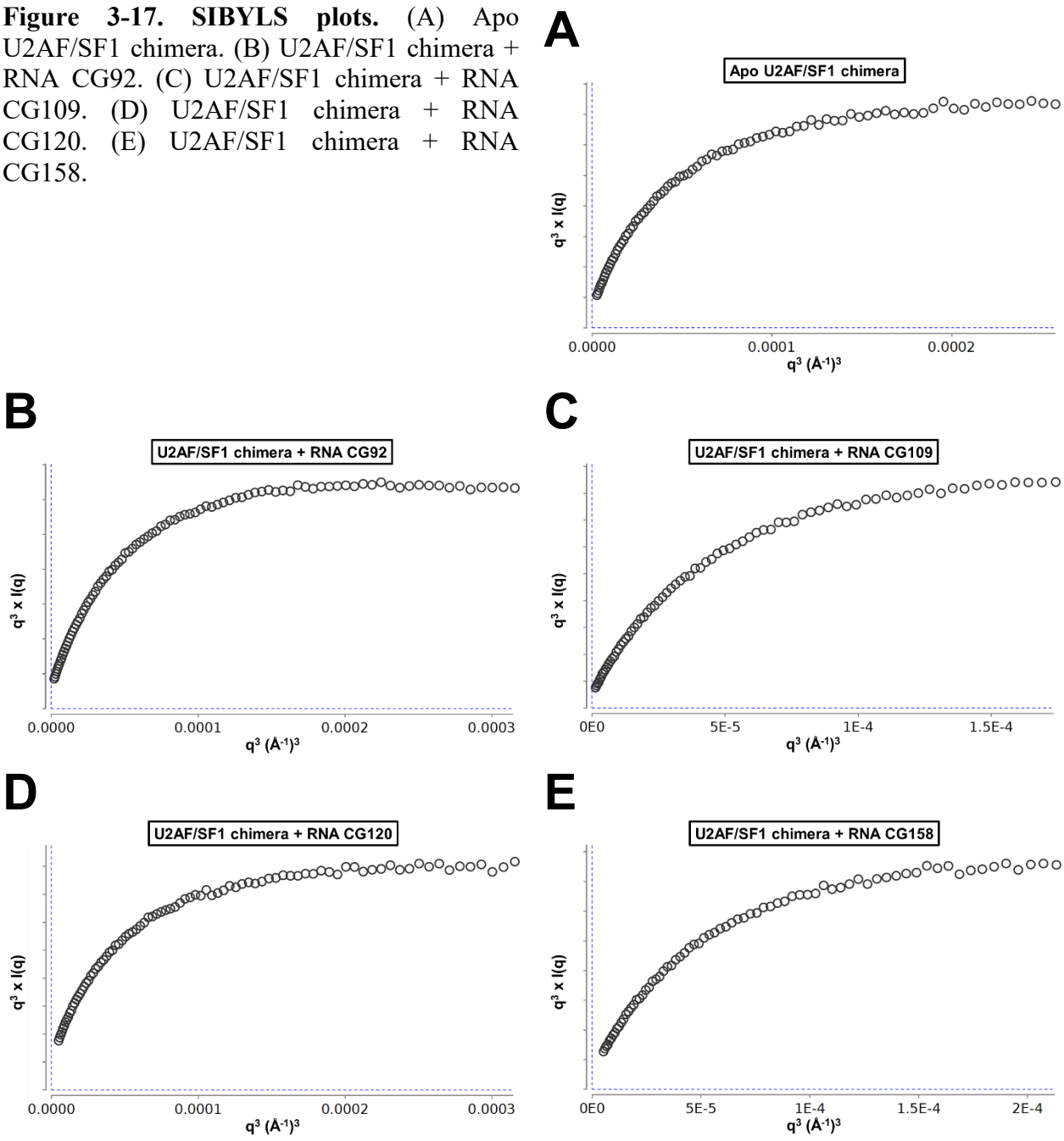


Figure 3-17. SIBYLS plots. (A) Apo U2AF/SF1 chimera. (B) U2AF/SF1 chimera + RNA CG92. (C) U2AF/SF1 chimera + RNA CG109. (D) U2AF/SF1 chimera + RNA CG120. (E) U2AF/SF1 chimera + RNA CG158.



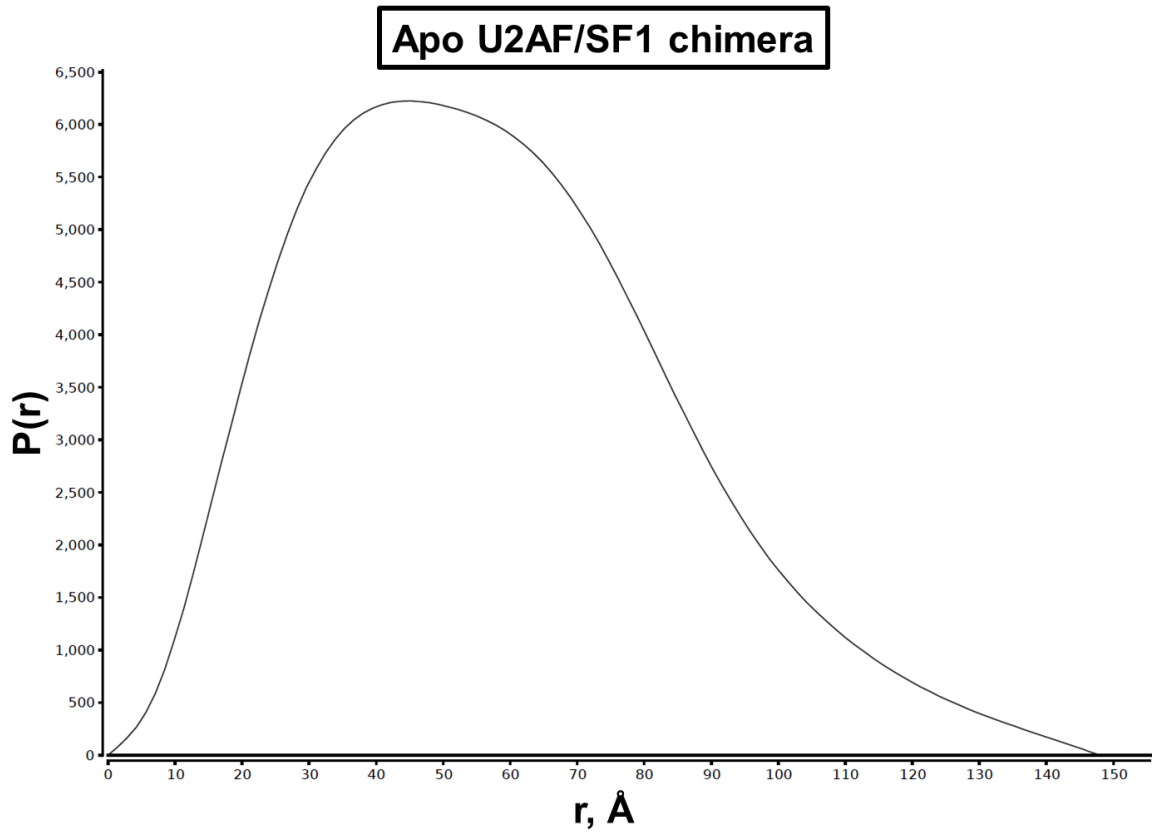


Figure 3-18. P(r) plot of apo U2AF/SF1 chimera.

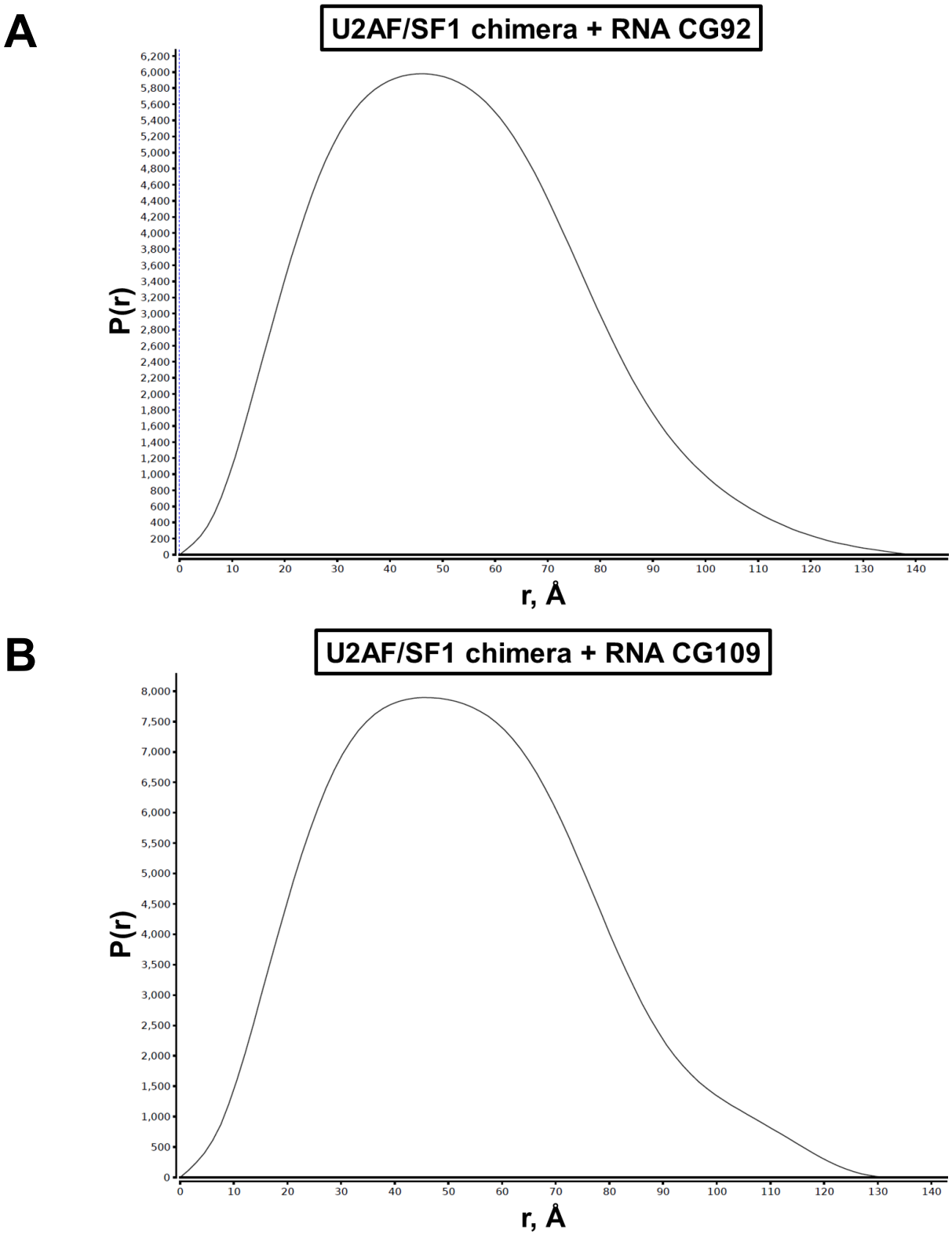


Figure 3-19. P(r) plots of RNA-bound U2AF/SF1 chimera. (A) U2AF/SF1 chimera + RNA CG92. (B) U2AF/SF1 chimera + RNA CG109.

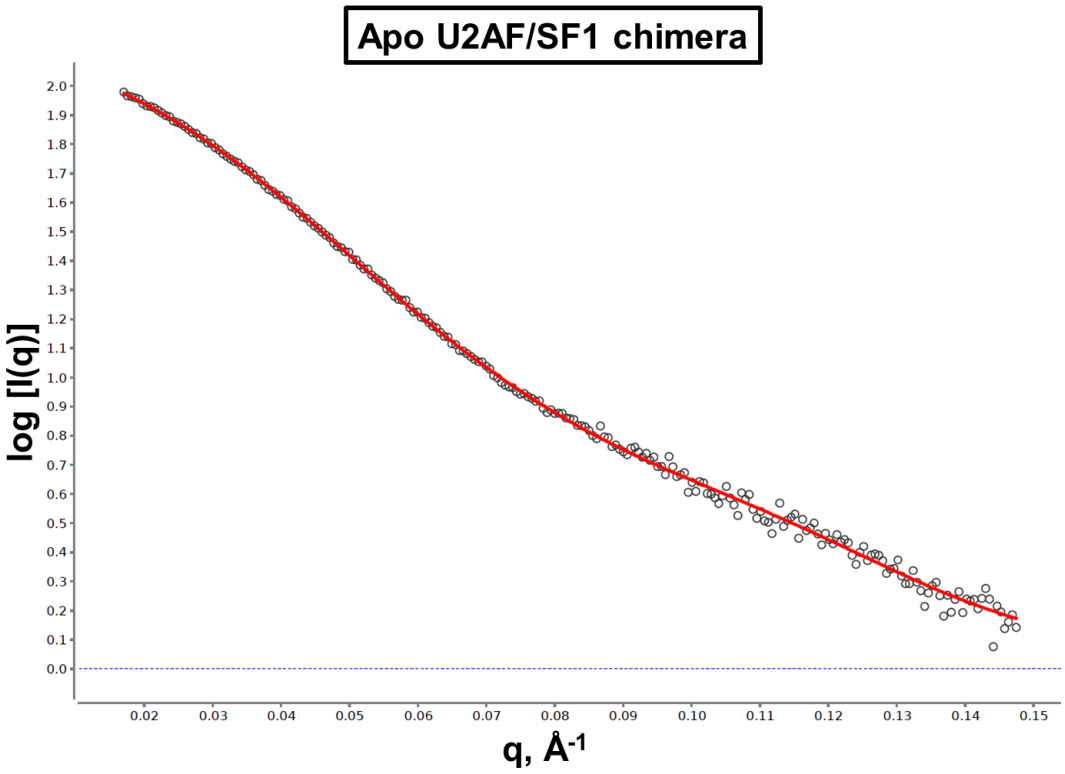


Figure 3-20. P(r) fit plot of apo U2AF/SF1 chimera.

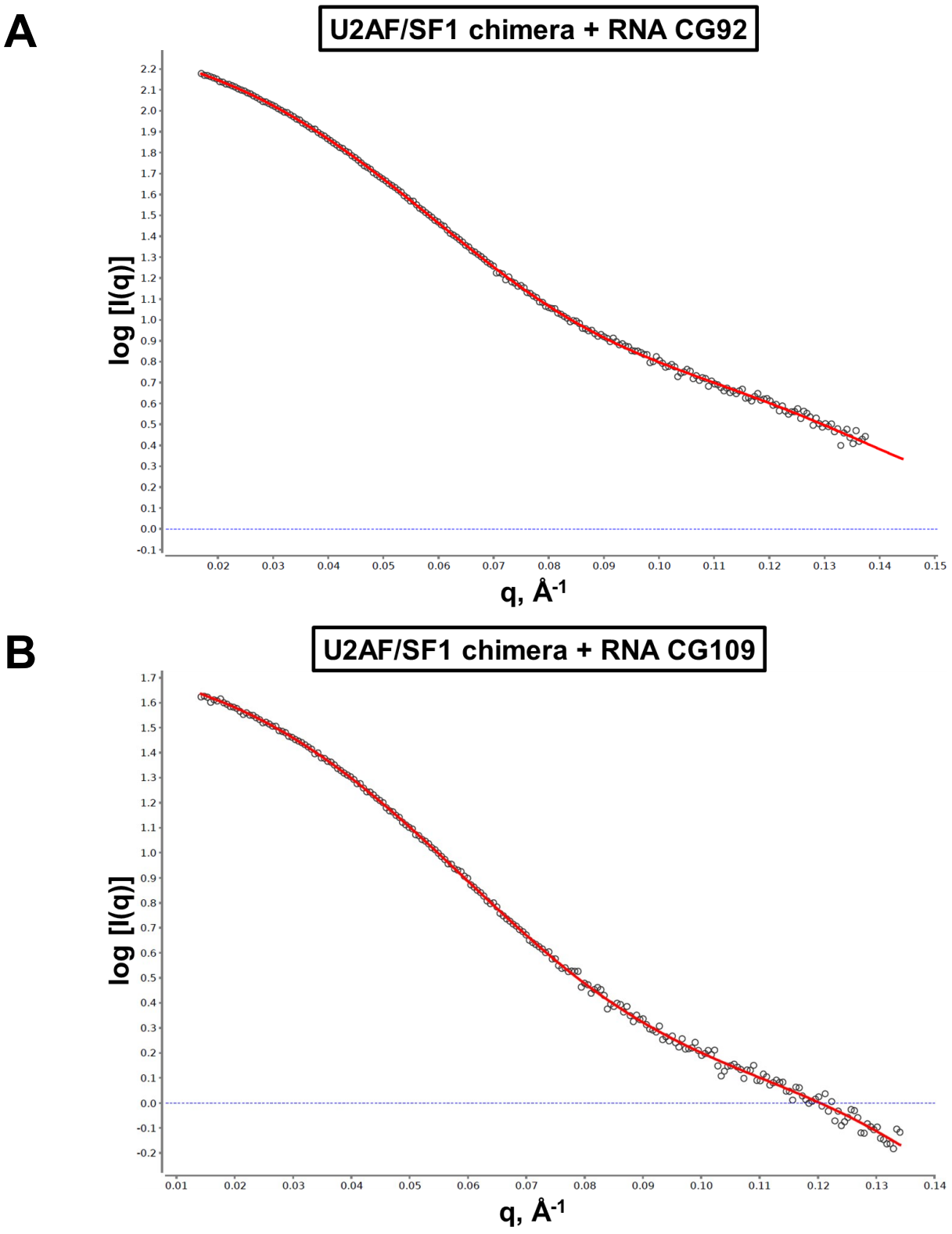


Figure 3-21. P(r) fit plots of RNA-bound U2AF/SF1 chimera. (A) U2AF/SF1 chimera + RNA CG92. (B) U2AF/SF1 chimera + RNA CG109.

3-2.2.2. Generation of model libraries

The initial model and model library for U2AF/SF1 chimera + RNA CG109 is shown below; Fig. 3-22 shows the initial model used for generating the library, and the model library itself is summarized in Fig. 3-23. For simplicity, corresponding images of the two other complexes have been omitted since the important general features are similar. Rigid body definitions used to build the model libraries are catalogued in Appendix *IV*.

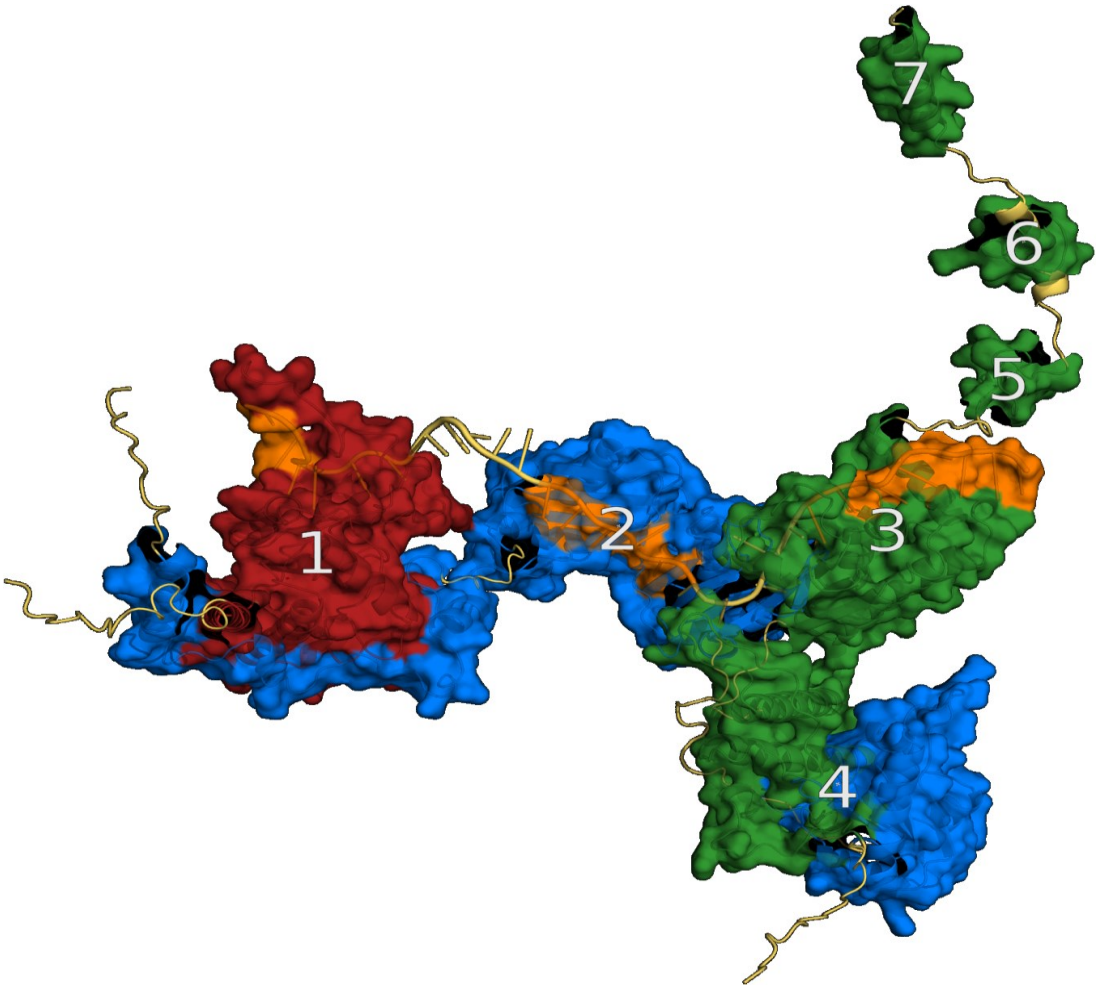
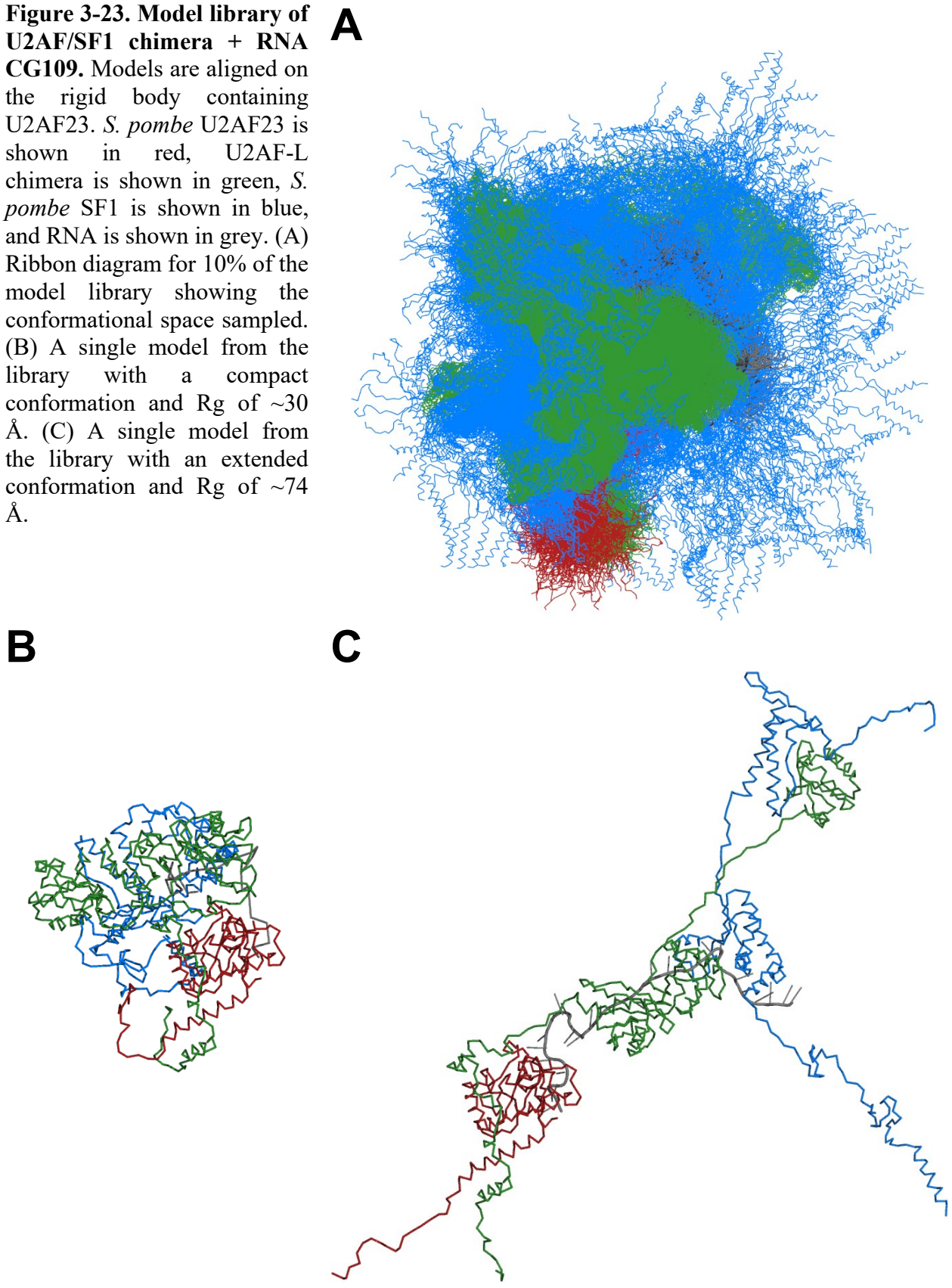


Figure 3-22. Initial model of U2AF/SF1 chimera + RNA CG109. The seven rigid bodies are numbered in order to correspond with Table *IV*-2 and are shown as a semi-transparent surface representation. Rigid bodies are coloured according to the protein chain they are part of; *S. pombe* U2AF23 is shown in red, U2AF-L chimera is shown in blue, *S. pombe* SF1 is shown in green, and RNA is shown in orange. Flexible regions and linkers outside of rigid bodies are shown in yellow.

Figure 3-23. Model library of U2AF/SF1 chimera + RNA CG109. Models are aligned on the rigid body containing U2AF23. *S. pombe* U2AF23 is shown in red, U2AF-L chimera is shown in green, *S. pombe* SF1 is shown in blue, and RNA is shown in grey. (A) Ribbon diagram for 10% of the model library showing the conformational space sampled. (B) A single model from the library with a compact conformation and Rg of ~30 Å. (C) A single model from the library with an extended conformation and Rg of ~74 Å.



3-2.2.3. Model library statistics

Statistics pertaining to building the model libraries are given below in Fig. 3-24 and 3-25, as well as in Tables 3-5 to 3-7. Fig. 3-24 shows the frequency of model inclusion in various multi-model ensembles plotted vs radius of gyration for all three complexes. Fig. 3-25 shows the frequency of model inclusion in various multi-model ensembles plotted vs maximum particle dimension for all three complexes. Table 3-5 contains the GAJOE results for apo U2AF/SF1 chimera. Table 3-6 contains the GAJOE results for U2AF/SF1 chimera + RNA CG92. Table 3-7 contains the GAJOE results for U2AF/SF1 chimera + RNA CG109.

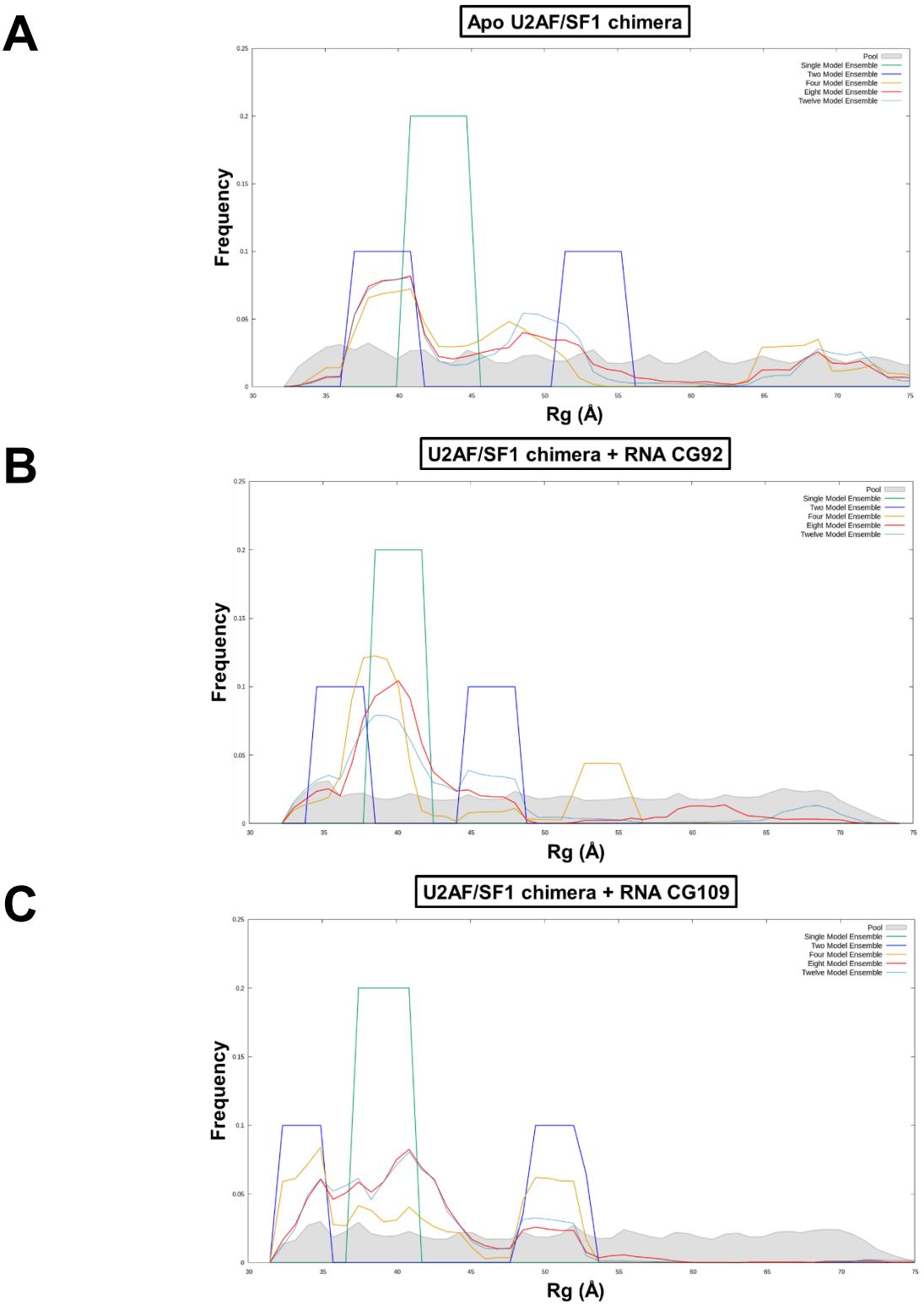


Figure 3-24. Frequency of model inclusion in various multi-model ensembles plotted vs radius of gyration. (A) Apo U2AF/SF1 chimera. (B) U2AF/SF1 chimera + RNA CG92. (C) U2AF/SF1 chimera + RNA CG109.

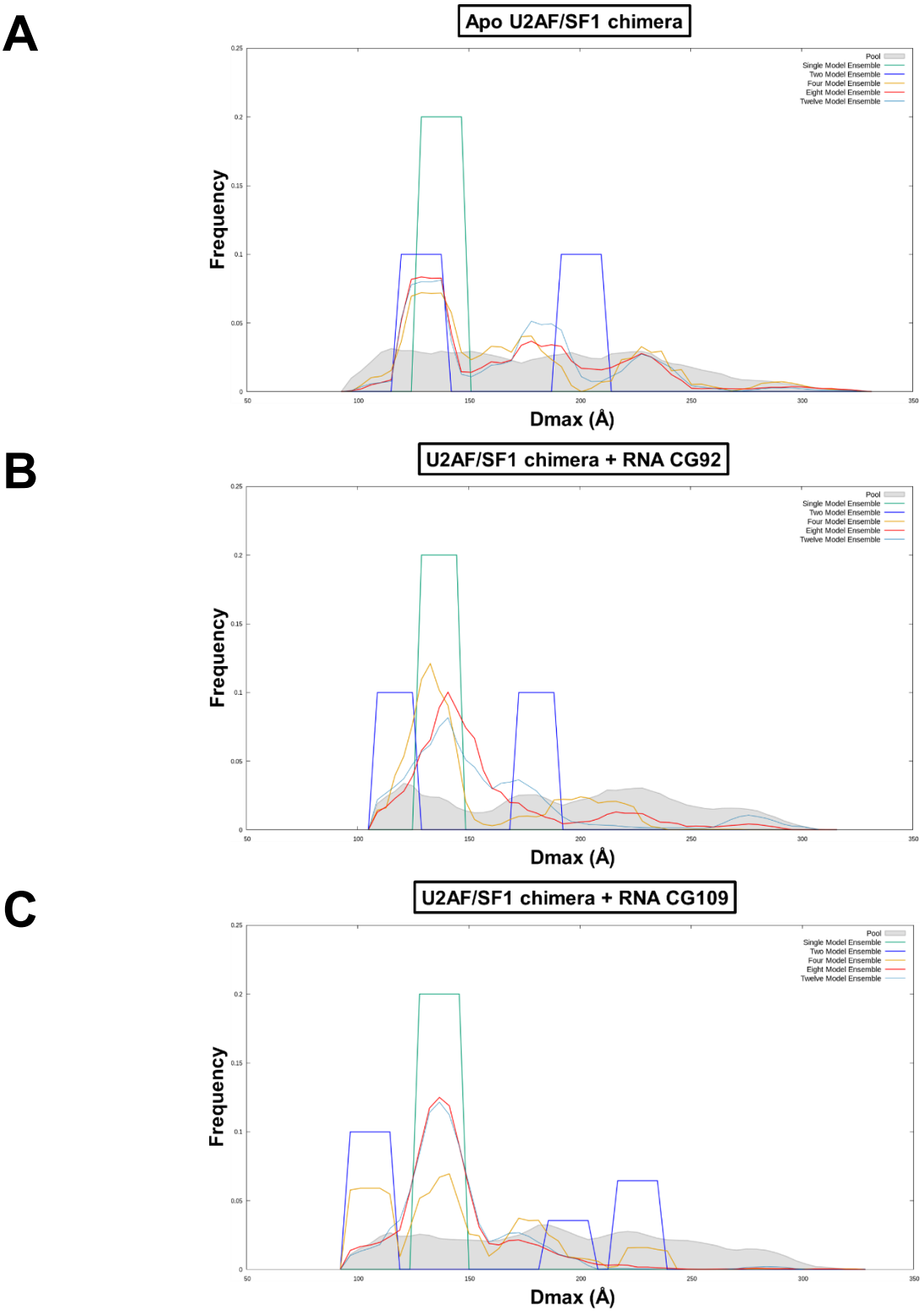


Figure 3-25. Frequency of model inclusion in various multi-model ensembles plotted vs maximum particle dimension. (A) Apo U2AF/SF1 chimera. (B) U2AF/SF1 chimera + RNA CG92. (C) U2AF/SF1 chimera + RNA CG109.

Table 3-5: GAJOE results for apo U2AF/SF1 chimera

Six independent runs with differing numbers of ensembles requested: automatically determined, 1, 2, 4, 8 and 12.

== GA001 ==			
Minimum number of curves per ensemble (min. 1) .....: 1			
-- Chi^2: 1.936			
-- Rflex (random)/Rsigma: ~85.72% (~95.47%)/0.86			
1) Model_00649	40.07	135.65	~0.50 (5/10)
2) Model_01237	48.69	177.02	~0.10 (1/10)
3) Model_04833	51.00	189.58	~0.10 (1/10)
4) Model_04835	51.83	194.20	~0.10 (1/10)
5) Model_06453	70.84	228.91	~0.10 (1/10)
6) Model_06953	73.17	274.10	~0.10 (1/10)
== GA002 ==			
Minimum number of curves per ensemble (min. 1) .....: 1			
-- Chi^2: 9.226			
-- Rflex (random)/Rsigma: ~40.35% (~95.47%)/0.00			
1) Model_00041	43.67	141.72	~1.00 (1/1)
== GA003 ==			
Minimum number of curves per ensemble (min. 1) .....: 2			
-- Chi^2: 2.050			
-- Rflex (random)/Rsigma: ~57.72% (~95.47%)/0.53			
1) Model_00047	39.66	130.37	~0.50 (1/2)
2) Model_06635	53.34	205.01	~0.50 (1/2)
== GA004 ==			
Minimum number of curves per ensemble (min. 1) .....: 4			
-- Chi^2: 1.944			
-- Rflex (random)/Rsigma: ~83.74% (~95.47%)/0.92			
1) Model_00047	39.66	130.37	~0.25 (1/4)
2) Model_00641	45.46	157.39	~0.50 (2/4)
3) Model_06616	74.95	286.79	~0.25 (1/4)
== GA005 ==			
Minimum number of curves per ensemble (min. 1) .....: 8			
-- Chi^2: 1.937			
-- Rflex (random)/Rsigma: ~85.84% (~95.47%)/0.87			
1) Model_00044	41.05	136.42	~0.12 (1/8)
2) Model_00046	39.99	132.54	~0.25 (2/8)
3) Model_00048	39.51	130.81	~0.12 (1/8)
4) Model_04838	54.16	209.72	~0.12 (1/8)
5) Model_06033	51.15	190.20	~0.12 (1/8)
6) Model_06037	54.85	210.95	~0.12 (1/8)
7) Model_06502	70.39	238.91	~0.12 (1/8)
== GA006 ==			
Minimum number of curves per ensemble (min. 1) .....: 12			
-- Chi^2: 1.932			
-- Rflex (random)/Rsigma: ~82.85% (~95.47%)/0.86			
1) Model_00639	47.12	167.55	~0.17 (2/12)
2) Model_00640	46.33	164.59	~0.08 (1/12)
3) Model_00649	40.07	135.65	~0.50 (6/12)
4) Model_06518	70.69	232.40	~0.08 (1/12)
5) Model_06954	73.51	291.73	~0.08 (1/12)
6) Model_06960	73.75	292.02	~0.08 (1/12)

Table 3-6: GAJOE results for U2AF/SF1 chimera + RNA CG92

Six independent runs with differing numbers of ensembles requested: automatically determined, 1, 2, 4, 8 and 12.

== GA001 ==
Minimum number of curves per ensemble (min. 1) .....: 1
-- Chi^2: 1.588
-- Rflex (random)/Rsigma: ~82.00% (~97.13%)/0.76

1) Model_00174	34.30	114.03	~0.08 (1/12)
2) Model_00244	34.06	115.85	~0.08 (1/12)
3) Model_01254	40.59	140.27	~0.42 (5/12)
4) Model_01456	39.58	131.66	~0.08 (1/12)
5) Model_01821	42.81	139.50	~0.08 (1/12)
6) Model_02532	46.32	170.66	~0.08 (1/12)
7) Model_03061	52.44	190.86	~0.08 (1/12)
8) Model_07105	69.69	284.22	~0.08 (1/12)

== GA002 ==
Minimum number of curves per ensemble (min. 1) .....: 1
-- Chi^2: 2.641
-- Rflex (random)/Rsigma: ~40.35% (~97.13%)/0.00

1) Model_01253	40.64	138.84	~1.00 (1/1)

== GA003 ==
Minimum number of curves per ensemble (min. 1) .....: 2
-- Chi^2: 1.731
-- Rflex (random)/Rsigma: ~57.72% (~97.13%)/0.44

1) Model_00080	36.75	119.24	~0.50 (1/2)
2) Model_02501	47.01	184.09	~0.50 (1/2)

== GA004 ==
Minimum number of curves per ensemble (min. 1) .....: 4
-- Chi^2: 1.595
-- Rflex (random)/Rsigma: ~71.22% (~97.13%)/0.58

1) Model_00050	39.72	140.04	~0.25 (1/4)
2) Model_00091	35.88	116.34	~0.25 (1/4)
3) Model_01257	40.70	136.22	~0.25 (1/4)
4) Model_03962	54.44	215.01	~0.25 (1/4)

== GA005 ==
Minimum number of curves per ensemble (min. 1) .....: 8
-- Chi^2: 1.596
-- Rflex (random)/Rsigma: ~79.35% (~97.13%)/0.68

1) Model_01064	35.88	135.92	~0.12 (1/8)
2) Model_01250	41.42	144.46	~0.12 (1/8)
3) Model_01584	39.31	152.76	~0.38 (3/8)
4) Model_01857	42.17	157.06	~0.12 (1/8)
5) Model_02870	47.04	146.34	~0.12 (1/8)
6) Model_07107	69.75	279.49	~0.12 (1/8)

== GA006 ==
Minimum number of curves per ensemble (min. 1) .....: 12
-- Chi^2: 1.588
-- Rflex (random)/Rsigma: ~82.04% (~97.13%)/0.73

1) Model_00793	35.91	120.77	~0.17 (2/12)
2) Model_01624	39.56	145.58	~0.08 (1/12)
3) Model_01626	39.31	146.72	~0.33 (4/12)
4) Model_02323	44.08	174.54	~0.08 (1/12)
5) Model_02520	46.93	187.61	~0.08 (1/12)

6) Model_02530	46.38	173.37	~0.08 (1/12)
7) Model_02933	47.34	173.87	~0.08 (1/12)
8) Model_07110	69.87	281.44	~0.08 (1/12)

Table 3-7: GAJOE results for U2AF/SF1 chimera + RNA CG109

Six independent runs with differing numbers of ensembles requested: automatically determined, 1, 2, 4, 8 and 12.

== GA001 ==
Minimum number of curves per ensemble (min. 1) .....: 1
-- Chi^2: 1.726
-- Rflex (random)/Rsigma: ~79.68% (~96.42%)/0.65

1) Model_00031	33.79	109.07	~0.12 (1/8)
2) Model_00046	41.01	141.73	~0.50 (4/8)
3) Model_00759	36.40	134.49	~0.12 (1/8)
4) Model_02413	45.71	145.19	~0.12 (1/8)
5) Model_03528	50.33	174.31	~0.12 (1/8)

== GA002 ==
Minimum number of curves per ensemble (min. 1) .....: 1
-- Chi^2: 2.615
-- Rflex (random)/Rsigma: ~40.35% (~96.42%)/0.00

1) Model_00048	39.96	140.83	~1.00 (1/1)

== GA003 ==
Minimum number of curves per ensemble (min. 1) .....: 2
-- Chi^2: 1.787
-- Rflex (random)/Rsigma: ~53.58% (~96.42%)/0.71

1) Model_00179	33.95	109.78	~0.50 (1/2)
2) Model_03308	51.32	227.48	~0.50 (1/2)

== GA004 ==
Minimum number of curves per ensemble (min. 1) .....: 4
-- Chi^2: 1.730
-- Rflex (random)/Rsigma: ~73.49% (~96.42%)/0.60

1) Model_00043	43.11	147.93	~0.25 (1/4)
2) Model_00764	36.07	136.79	~0.25 (1/4)
3) Model_00771	35.62	134.14	~0.25 (1/4)
4) Model_03452	50.63	175.45	~0.25 (1/4)

== GA005 ==
Minimum number of curves per ensemble (min. 1) .....: 8
-- Chi^2: 1.727
-- Rflex (random)/Rsigma: ~79.08% (~96.42%)/0.51

1) Model_00044	42.34	146.38	~0.50 (4/8)
2) Model_00203	33.73	108.18	~0.12 (1/8)
3) Model_00228	33.62	105.38	~0.12 (1/8)
4) Model_02280	43.35	152.89	~0.12 (1/8)
5) Model_03502	50.55	173.99	~0.12 (1/8)

== GA006 ==
Minimum number of curves per ensemble (min. 1) .....: 12
-- Chi^2: 1.726
-- Rflex (random)/Rsigma: ~78.04% (~96.42%)/0.51

1) Model_00047	40.44	140.44	~0.08 (1/12)
2) Model_00048	39.96	140.83	~0.50 (6/12)
3) Model_00766	35.90	134.84	~0.17 (2/12)
4) Model_02807	47.34	177.00	~0.08 (1/12)
5) Model_02814	47.11	167.13	~0.08 (1/12)
6) Model_03526	50.37	168.93	~0.08 (1/12)

3-3. Discussion

3-3.1. SAXS characterization of *S. pombe* U2AF dimer and U2AF/SF1 trimer

In the absence of a true atomic model, *ab initio* shape reconstruction from SAXS scattering data has the potential to provide important, albeit limited structural information with the caveat that traditional SAXS based *ab initio* modelling techniques are only appropriate if conformational flexibility does not exist. In addition to the three-dimensional model itself, the scattering data used to create this model also reveals important information about both the sample quality, as well as structural properties of the sample.

Minimal aggregation is present in the six samples, and this initial quality check indicates that they are suitable for further data processing and analysis. This is shown both by the flattening of the experimental SAXS curves at low angles, as well as the linear trend of data points towards low q on the Guinier plots.

Analysis of the scattering data suggests that all six complexes are well folded. The bell-shaped P(r) function of the six complexes is characteristic of a well folded structure. Additionally, Kratky plots can qualitatively assess the degree of flexibility and/or unfolding in the samples. The Kratky plots of all six complexes are roughly Gaussian. Globular proteins will generate a Gaussian peak, whereas highly flexible and/or unfolded proteins will have a plateau in the Kratky plot at high q.

It is expected that the U2AF/SF1 trimer complex in both its apo and RNA-bound states will have a more complex overall three-dimensional shape than the apo and RNA-bound U2AF dimer complexes, and the Kratky plots are consistent with these expectations. The four Kratky plots of apo and RNA-bound U2AF/SF1 trimer all have a shoulder on the right side of the curve,

which is characteristic of a complex, multidomain protein structure. This feature is absent in the Kratky plots of apo and RNA-bound U2AF dimer.

A careful comparison of the Rg values from the Guinier analysis has the potential to provide additional structural and functional information about the complexes which subsequently should be validated through additional experiments. One of these insights is that RNA binding does not appear to cause a compaction of the U2AF dimer, however U2AF/SF1 trimer does appear to assume a more compact structure upon RNA binding. This is because the Rg values of apo and RNA-bound U2AF dimer are almost identical, whereas there is a significant difference in Rg values between apo U2AF/SF1 trimer and the three different RNA-bound U2AF/SF1 trimer complexes. When bound to RNA CG92, the complex has an Rg that is 91% that of the apo complex, and when bound to RNA CG158, the complex has an Rg that is 95% that of the apo complex. These two RNAs both have the same 3′ SS signatures. However, when bound to RNA CG109, the Rg is nearly identical to the apo complex. CG109 is significantly longer than either CG92 or CG158. This observation suggests that the U2AF/SF1 trimer may possess structural plasticity to accommodate 3′ SS RNAs of diverse length and sequence composition.

In addition to the potential structural consequences of RNA binding, comparison of the Rg values from the Guinier analysis combined with a comparison of the *ab initio* shape reconstruction of the six complexes suggests that the U2AF dimer in both its apo and RNA-bound states is more extended/cylindrical than the U2AF/SF1 trimer in both its apo and RNA-bound states, which is more globular. These observations are consistent with previously proposed models in which the U2AF/SF1 complex is bent in order to help proximate the 5′ SS and 3′ SS. The greater mass of the U2AF/SF1 complex relative to the U2AF dimer does not translate to a commensurate increase in Rg relative to the U2AF dimer. The lowest Rg value (U2AF dimer +

RNA CG120) is 85% of the highest Rg value (U2AF/SF1 trimer + RNA CG109). However, the lowest mass (apo U2AF dimer, 68.6 kDa) is 60% of the highest mass (U2AF/SF1 trimer + RNA CG109, 115.04 kDa). Finally, as mentioned previously the *ab initio* shape reconstructions show that apo and RNA-bound U2AF dimer are more extended/cylindrical than apo and RNA-bound U2AF/SF1 trimer, which are more globular.

Together, a careful analysis of the *ab initio* models combined with the scattering data used to create them provides several useful clues into the structural and functional properties of these complexes which require validation through additional experiments. However, the main value of these experiments is that they provide a reliable proof of concept for combining SAXS based *ab initio* shape reconstruction of these complexes with rigid body modeling through BilboMD.

3-3.2. SEC-SAXS characterization of U2AF/SF1 chimera

Before building the model libraires, initial quality checks were completed for the SAXS data (see Section 3-2.2.1.). As stated previously, the RNA-bound samples containing either RNA CG120 or CG158 behaved poorly in solution and were not carried forward for further analysis. Additionally, analysis of the Porod-Debye region reveals a plateau in the SIBYLS plot (Fig. 3-17) for all samples indicating some conformational flexibility in solution.

Ensemble optimization methods use an iterative statistical approach (in this context referred to as a genetic algorithm) to fit the calculated averaged scattering of an ensemble of conformations to the experimental SAXS curve (Bernado, Mylonas, Petoukhov, Blackledge, & Svergun, 2007; Pelikan et al., 2009; Tria, Mertens, Kachala, & Svergun, 2015). The ensemble is derived from a library of structural models that sample the potential conformational space of the

structure. For simplicity, the model library is generated by treating known domains and domain interfaces as rigid bodies. In the case of multi-domain complexes such as the chimeric U2AF/SF1 trimer complex, these rigid bodies are tethered by flexible linkers.

Initial structures of the five complexes were modeled using COOT (Crystallographic Object-Oriented Toolkit) based on previously reported reductionist structures that represent discrete rigid bodies; the initial model of U2AF/SF1 chimera + RNA CG109 is provided as a visualization tool in Fig. 3-22 (Emsley, Lohkamp, Scott, & Cowtan, 2010). After all-atom energy minimization in AMBER20, a model library of 7200 conformations was generated using BilboMD for each complex over an Rg range of 30-74 Å; using U2AF/SF1 chimera + RNA CG109 as an example, the partial model library, one conformationally compact model, and one conformationally extended model is provided as a visualization tool in Figure 3-23 (Case et al., 2021; Pelikan et al., 2009).

Ensemble optimization was run using GAJOE (Genetic Algorithm Judging Optimisation of Ensembles) (Bernado et al., 2007; Tria et al., 2015). GAJOE was repeated six times for each complex, each time requesting a different maximum number of models to be included in the final ensemble. In the initial run, the algorithm chooses an optimal number of models in the ensemble. Subsequent runs were fixed at ensembles containing 1, 2, 4, 8 and 12 models. By requesting differing numbers of models per ensemble, the minimum number of models in an ensemble that best fit the data can be determined. For the apo complex, a two-model ensemble adequately describes the average structure in solution with a best fit to the experimental SAXS curve ($\chi 2$) of 2.0. Inclusion of additional models to the ensemble does not significantly improve the fit to the experimental data. For U2AF/SF1 chimera + RNA CG92 averaging a four-model ensemble ($\chi 2$ =

1.60), and for U2AF/SF1 chimera + RNA CG109 a two-model ensemble ($\chi 2 = 1.79$) best fit the data (Tables 3-5 to 3-7).

Overall trends in particle conformational flexibility based on size can be visualized by plotting the frequency of inclusion of models in ensembles during iterations of the genetic algorithm against particle dimensions (either Rg or D_{max}; see Fig. 3-24 and 3-25). While all complexes showed modest amounts of larger particles in addition to the more predominant compact form, it is clear that a bound 3′ SS RNA skews the particle distribution in solution to a more compact, globular form. The lack of an integrated atomic model of the U2AF/SF1 complex makes it difficult to interpret the larger, extended particles and more globular, compact particles in solution. However, previous studies are consistent with the existence of both compact and extended conformations of the U2AF/SF1 complex. NMR, SAXS, and X-ray structures characterizing the phosphorylated domain of SF1 show significant structural plasticity in this region, as well as the surrounding U2AF-L/SF1 interface, and this plasticity has important functional roles (W. Wang et al., 2013; Y. Zhang et al., 2013). Evidence also exists to support a model in which U2AF-L bends the pre-mRNA to proximate the catalytic nucleotides at the 3′ SS, thereby organizing them for subsequent spliceosome assembly (O. A. Kent et al., 2003). Finally, a structural model has been published in which the RRMs of U2AF-L exist in both open and closed conformations and a population shift between these two occurs upon binding RNA, which is consistent with the observation that binding of a 3′ SS RNA skews the particles in solution towards being more compact and globular (Mackereth et al., 2011).

Together, these results further support previous data indicating that the U2AF/SF1 complex is well suited for atomic level structural characterization, but additionally helps to explain why crystallization was unsuccessful. The existence of conformational flexibility is

consistent with the modular, multi-domain nature of the complex and suggests that it is likely too plastic to produce diffracting crystals. Additionally, the RNA-bound complex may be a better candidate for atomic level structural characterization through all methods since bound 3′ SS RNA skews the particle distribution in solution to a more compact, globular form, and this is expected to be a less flexible and more uniform population of particles better suited for structural characterization.

3-4. Materials and Methods

3-4.1. SAXS characterization of *S. pombe* U2AF dimer and U2AF/SF1 trimer

3-4.1.1. Sample preparation and data collection

U2AF dimer and U2AF/SF1 trimer were expressed and purified as described previously in Chapter 2. After the anion exchange purification step, protein was dialyzed into SAXS buffer (60 mM NaCl, 1mM TCEP-HCl (tris (2-carboxyethyl) phosphine hydrochloride), 20 mM Tris-HCl, pH 8.0, 2% glycerol (v/v)). For protein/RNA complexes, RNA was dissolved in SAXS buffer and mixed in a 1:1 molar ratio with the U2AF dimer or U2AF trimer complex.

In preparation for data collection, three samples of each complex were prepared at different concentrations. U2AF dimer in both its apo and RNA-bound states was diluted to 3, 6, and 9 mg mL^{-1}. U2AF/SF1 trimer in its apo state was diluted to 1.5, 3, and 4.5 mg mL^{-1}. U2AF/SF1 trimer in its RNA-bound state was diluted to 1.3, 2.7, and 4 mg mL^{-1}. In addition to the protein samples, sample buffer was also provided for data collection in order to correct the samples for background scattering.

SAXS data were collected at SIBYLS (Structurally Integrated BiologY for the Life Sciences) Beamline 12.3.1. through the HT-SAXS (high-throughput SAXS) mail-in program of the ALS (Advanced Light Source), Lawrence Berkeley National Laboratory, Berkeley, CA (Classen et al., 2013; Dyer et al., 2014; Hura et al., 2009; Putnam, Hammel, Hura, & Tainer, 2007). Data were collected for each sample at four different exposure times on a Pilatus3 2M pixel array detector (Dectris). Samples were shipped and stored at 4℃. Prior to data collection, the samples were spun at 3700 rpm for 10 minutes at 4°C. Samples were maintained at 10°C during data collection.

3-4.1.2. Data processing

The raw scattering data was received as buffer corrected DAT files, with each file corresponding to a concentration/exposure combination in the concentration/exposure series. For each of the six complexes, an overlay of all experimental SAXS curves corresponding to the complete concentration/exposure series was also provided, and an image of each protein sample and buffer blank was taken during data collection and provided with the scattering data. Alongside the experimental SAXS curves, a list of the corresponding samples was also provided for each complex noting those that displayed detector saturation during data collection; these lists were omitted from the preceding Results section for clarity and simplicity.

Each sample in the concentration/exposure series for each complex was evaluated before further processing. Buffer corrected files which were derived either from a buffer blank (used for background subtraction) or sample containing bubbles were rejected from further data processing and analysis. Similarly, buffer-subtracted samples displaying detector saturation were also rejected. Subsequent data processing and analysis was completed using software contained

within the ATSAS program suite (Manalastas-Cantos et al., 2021; Petoukhov et al., 2012).

The remaining usable samples in the concentration/exposure series for each complex were all processed using PRIMUS within the ATSAS suite in order to generate a linear Guinier plot. For each complex, the sample with the highest calculated data quality estimate from the Guinier analysis was selected for further processing and analysis (Konarev, Volkov, Sokolova, Koch, & Svergun, 2003). After assessing data quality from the Guinier analysis and selecting one sample from the concentration/exposure series for each complex, the Kratky analysis was completed using PRIMUS.

GNOM was used to calculate the pair-distance distribution function P(r), which is an indirect Fourier transform of I(0) (Svergun, 1992). The D_{max} value used to evaluate the P(r) function was adjusted to generate an Rg value very similar or identical to that derived from the Guinier analysis.

After initial quality checks, a three-dimensional molecular envelope was generated for each complex. First, the *ab initio* modeling program GASBOR (Genetic Algorithm for SwitchBOx Routing) was used in real space mode to generate models refined against the P(r) function (Lienig & Thulasiraman, 1996; Svergun, Petoukhov, & Koch, 2001). Ten rounds of model-building from different and random starting positions for dummy atoms were averaged using the program DAMAVER (Volkov & Svergun, 2003). The unfiltered model from DAMAVER was further refined by using it as a starting model for the program DAMMIN to yield the final calculated envelope (Svergun, 1999).

3-4.2. SEC-SAXS characterization of U2AF/SF1 chimera

3-4.2.1. Sample preparation and data collection

U2AF/SF1 chimera was expressed and purified as described previously in Chapter 2. After the anion exchange purification step, protein was dialyzed into SEC-SAXS buffer (100 mM NaCl, 5 mM BME, 50 μM $ZnCl_2$, 20 mM Tris-HCl, pH 8.0). For protein/RNA complexes, RNA was dissolved in SEC-SAXS buffer and mixed in a molar excess of 1:1.2 of protein to RNA. In preparation for data collection, 60 μL of the sample for each complex was prepared at 5 mg mL^{-1}. In addition to the protein samples, 10X SEC running buffer (1 M NaCl, 50 mM BME, 500 μM $ZnCl_2$, 200 mM Tris-HCl, pH 7.25) was also provided for the SEC step preceding SAXS data collection.

SEC-SAXS data were collected at SIBYLS Beamline 12.3.1. through the SEC-SAXS mail-in program of the ALS, Lawrence Berkeley National Laboratory, Berkeley, CA (Classen et al., 2013; Dyer et al., 2014; Hura et al., 2009; Putnam et al., 2007). The sample was first separated via SEC using an Agilent 1260 series HPLC (high-performance liquid chromatography) with a Shodex 803 analytical column at a flow rate of 0.5 ml/min. Eluent was then split 2 to 1 between X-ray synchrotron radiation and a series of four inline analytical instruments, which collect and analyze UV, MALLS, QELS (Quasi-Elastic Light Scattering), and differential refractive index data using Wyatt Astra 6 software.

SAXS examination of the sample as it exited the column was completed with λ=1.03 Å incident light at a sample to detector distances of 1.5 m, resulting in scattering vectors (q) ranging from 0.013-0.5 $Å^{-1}$, where the scattering vector is defined as $q = 4\pi\sin\theta/\lambda$ and 2θ is the measured scattering angle; 3 second exposures were collected for each frame over the course of

40 min.

Raw SEC-SAXS data were collected, analyzed, and merged by beamline staff. This analysis utilized those merged SAXS data files. Although in-line MALLS data were also collected and processed either automatically at the beamline or by beamline staff, that analysis was erratic and incomplete, and without access to the raw MALLS data it was not possible to reprocess so is not included in this thesis.

3-4.2.2. Data processing

SAXS data of the five complexes were analyzed using the software ScÅtter authored by Robert Rambo. Complex flexibility was assessed in the Porod-Debye region using the methodology developed and implemented in ScÅtter (Rambo & Tainer, 2011).

Atomic models of the apo and RNA-bound complexes were constructed with COOT based on existing structures from the PDB (Emsley et al., 2010). Missing regions were added based on models from Alphafold (Jumper et al., 2021; Varadi et al., 2022). CHARMM-GUI (Chemistry at Harvard Macromolecular Mechanics-Graphical User Interface) was used to prepare the molecular models for energy minimization and equilibration in AMBER20 (Brooks et al., 2009; Case et al., 2021; Jo, Kim, Iyer, & Im, 2008; J. Lee et al., 2016). Due to the potential complexity of including atomic zinc in AMBER and BilboMD, and that their overall contribution to the fitting of the SAXS data would be minimal, the four zinc atoms present in the complexes were removed from the models.

Conformational flexibility of the complexes were determined using conformational sampling and ensemble optimization. A library of potential conformations were generated using BilboMD (Pelikan et al., 2009). Known structural domains and complexes were treated as rigid

bodies and allowed to move independently relative to each other within the bounds of their tethered flexible linkers. In order to obtain a conformational library having a smooth distribution of radius of gyration (Rg) and maximum particle dimension (D_{max}), two independent runs of BilboMD were carried out for each complex, requesting conformations be generated within the Rg ranges of 30-70 Å and 34-74 Å, resulting in 7200 conformations for each complex.

Ensemble optimization was completed using GAJOE (Bernado et al., 2007; Tria et al., 2015).

Chapter 4

Conclusions and future directions

4-1. Thesis summary

In eukaryotes, genetic information is contained within chromosomes and organized into discrete units called genes which consist of protein-coding exon sequences interrupted by silent intron sequences. During gene expression, the gene sequence is transcribed into a pre-mRNA which undergoes a process called splicing in order to remove intron sequences from the final message, which is then translated by the cell in order to synthesize a protein. This process is co-ordinated and directed by a multi-megadalton protein/RNA assembly called the spliceosome which is comparable in size, complexity, and dynamic properties to the ribosome. The assembly, maturation, and disassembly of the spliceosome as it removes the intron progresses through a discrete and definable sequence of events, which are collectively known as the spliceosome cycle.

The spliceosome cycle begins with the E complex, which commits the intron to removal through the splicing process. In the E complex both the 5′ and 3′ SS of the intron are recognized by the splicing machinery through conserved motifs within the pre-mRNA sequence; the 3′ SS is defined by the BPS, PPT, and AG di-nucleotide, which are recognized by a heterotrimeric protein complex consisting of U2AF-S, U2AF-L, and SF1. U2AF-S recognizes the AG di-nucleotide, U2AF-L recognizes the PPT, and SF1 recognizes the BPS. Recognition of the 3′ SS by these splicing factors is a pivotal step because definition of this intron/exon boundary is not immutable. The definition and recognition of the 3′ SS by the U2AF/SF1 complex ultimately determines the final mRNA sequence, and therefore determines the protein sequence. This process is actively manipulated by the cell in a process called alternative splicing, which is largely responsible for the spatio-temporal regulation of gene expression that has allowed the evolution of multicellular, developmentally complex eukaryotes such as humans. Alternative

splicing patterns can be disrupted or modified in disease states which has the potential for far-reaching downstream biological consequences. Therefore, recognition of the 3′ SS by U2AF/SF1 is a fundamentally important layer of regulation in the gene expression of eukaryotes, particularly developmentally complex eukaryotes.

Despite the importance of a detailed and complete understanding of 3′ SS definition and recognition in the context of splicing, all existing atomic structures of this apparatus only represent small portions of the U2AF/SF1 complex and therefore provide a very fragmented understanding of how this entity operates. There is no complete experimentally derived atomic model of the conserved RNA-binding core of U2AF/SF1 in any of its potential free or RNA-bound states. Until and unless this goal is realized, our understanding of pre-mRNA splicing will remain incomplete and cannot be accurately integrated into the context of higher-order biological processes dependent on the spatio-temporal regulation of gene expression patterns such as organismal development, disease, evolution, etc. In order to close this gap, the conserved RNA-binding core of both the U2AF dimer and U2AF/SF1 heterotrimer were cloned, expressed, and used as an experimental model in this study, using the orthologues of these proteins present in the fission yeast *S. pombe*.

A protocol was successfully established in order to prepare milligram quantities of pure, soluble, stable and monodisperse U2AF dimer and U2AF/SF1 trimer suitable for further biochemical and structural characterization. A requisite for credible characterization of these complexes is to unambiguously establish their existence by an absolute method. To this end, SEC-MALLS was used to determine the MW of these complexes directly from first principles, thereby confidently establishing the existence of an experimental system suitable as a foundation for further characterization.

Biochemical characterization of the U2AF dimer and U2AF/SF1 trimer consisted of spectroscopic analysis of RNA-bound U2AF/SF1 trimer as well as a dissection of the RNA-binding properties of U2AF dimer and U2AF/SF1 trimer via EMSA analysis. Spectroscopy was used to determine the stability of the protein/RNA complex formed from U2AF/SF1 trimer and a model 3′ SS RNA, because stability of the protein/RNA complex is necessary for EMSA-based analyses, as well as for structural studies of RNA binding. It was determined that the protein/RNA complex is stable enough to survive SEC purification as well as storage at 4°C for at least 7 days. Subsequently, EMSAs were completed between various pairings of protein and RNA: U2AF dimer and four variations of the U2AF/SF1 trimer were titrated against two 3′ SS model RNAs and two negative control model RNAs. The resulting K_d values were compared in order to dissect the RNA-binding properties of the U2AF/SF1 complex, revealing that SF1 significantly increases the affinity of the U2AF/SF1 complex for the 3′ SS, that the dual phosphomimetic point mutations in SF1 are responsible for conferring a preference of the U2AF/SF1 complex for a U9 PPT over a U12 PPT, and that the zinc knuckles of SF1 also confer this preference, but additionally also increase the affinity of the complex for the 3′ SS. In addition to the observations noted above, the data contained several more critical insights which were not immediately obvious and required a careful analysis of the negative control data as well (see Section 2-3.5.1.).

Structural characterization consisted of two sets of experiments. The first of these was SAXS to evaluate both the overall solution behaviour and scattering of *S. pombe* U2AF dimer and U2AF/SF1 trimer complexes in their free and RNA-bound states, as well as generate molecular envelopes of individual complexes. The SAXS data indicate that all of the complexes analyzed are well folded and roughly globular. Additionally, a comparison of the Rg values from

the Guinier analysis combined with analysis of the *ab initio* shape reconstruction of the complexes suggests that the U2AF dimer in both its apo and RNA-bound states is more extended/cylindrical than the U2AF/SF1 trimer in both its apo and RNA-bound states, which is more globular. These observations are consistent with previously proposed models in which the U2AF/SF1 complex is bent in order to help proximate the 5′ SS and 3′ SS (O. A. Kent et al., 2003; W. Wang et al., 2013).

These experiments were followed up by a more advanced SAXS technique that can more accurately model conformationally flexible, multidomain protein complexes such as the U2AF dimer and U2AF/SF1 trimer than traditional *ab initio* shape reconstruction. This technique involved analysis of a modified, chimeric U2AF/SF1 complex containing both human and *S. pombe* components in both its apo state as well as bound to two different 3′ SS model RNAs and combined SEC-SAXS with rigid body modelling in order to generate energy-minimized structure ensembles. These experiments indicate that all samples possess some conformational flexibility in solution, showing modest amounts of larger particles in addition to a more predominant compact form. However, a bound 3′ SS RNA clearly skews the particle distribution in solution to the more compact, globular form. Together, these results further support previous data indicating that the U2AF/SF1 complex is well suited for atomic level structural characterization, but additionally helps to explain why crystallization was unsuccessful. The existence of conformational flexibility is consistent with the modular, multi-domain nature of the complex and suggests that it is likely too plastic to produce diffracting crystals. Additionally, the RNA-bound complex may be a better candidate for atomic level structural characterization through all methods since bound 3′ SS RNA skews the particle distribution in solution to a more

compact, globular form, and this is expected to be a less flexible and more uniform population of particles.

The results summarized above represent significant progress in closing the gap towards understanding 3′ SS recognition by U2AF/SF1 by solving major technical obstacles. However, 3′ SS recognition by U2AF/SF1 is only one of two parts in understanding the roles of U2AF/SF1 in the spliceosome cycle. In order to fit our understanding of U2AF/SF1 into the spliceosome cycle, it is also necessary to understand the events connected to displacement of U2AF/SF1 following initial 3′ SS definition from the splicing machinery. With respect to this aim, this study also includes the X-ray structure of the *S. pombe* orthologue of the conserved splicing protein p14, which directly contacts the branch A within the BPS via stacking interactions mediated by a conserved aromatic residue after U2AF/SF1 has been completely displaced. Additionally, the X-ray structure has also been solved for the p14 orthologue present in the yeast species *C. albicans* because it lacks the conserved aromatic residue, and it is important to understand alternative mechanisms that exist for specific recognition of the branch A. These two new yeast structures and previously published human p14 structures are all similar to one other. However, there are important differences between the yeast structures and the human structures outside of the RRM, which forms most of the protein, revealing that p14 is more plastic than previously thought, although the functional consequences of this observation remain to be explored.

Additionally, the ability to easily clone, express and purify many truncations and mutants of this complex permits the investigation of specific properties of this entity deeply and purposefully through many types of experiments. Although no atomic structure of U2AF/SF1 was generated in this study, this experimental system is the most promising foundational work published thus far to allow this goal to be achieved in the future through a suitable method such

as cryo-EM. Ultimately, development and use of this system is expected to finally provide a unified atomic model of U2AF/SF1 and p14 in the context of the splicing cycle. In addition to the U2AF/SF1 work in this this study, the X-ray structure of *S. pombe* p14 is also important, because it validates *S. pombe* as a useful model organism to study early events of the spliceosome cycle at the 3′ SS by providing evidence to support the assertion that multiple aspects of this apparatus are conserved between *S. pombe* and higher eukaryotes, such as humans. This is important, since *S. pombe* is a simple organism that is well developed as a model system due to its historical use in foundational studies of the eukaryotic cell cycle, with many convenient and versatile cell biology and molecular genetic tools being available. Additionally, many aspects of splicing related to the 3′ SS apparatus are highly simplified in *S. pombe*; for example, unlike higher eukaryotes such as humans, *S. pombe* has no paralogues or alternatively spliced isoforms for U2AF, SF1 or p14, and *S. pombe* also undergoes far fewer splicing events. Together, these features of *S. pombe* combined with the investigations presented here reinforce this thesis as important foundational work in understanding early events of pre-mRNA splicing at the 3′ boundary of introns controlled by the conserved splicing factors U2AF, SF1 and p14.

4-2. Future directions: biochemical characterization of the U2AF dimer and U2AF/SF1 trimer

4-2.1. Extended EMSA based analyses of the U2AF dimer and U2AF/SF1 trimer

The EMSA results indicate that currently unknown higher-order principles govern the RNA binding behaviour of the U2AF/SF1 complex that cannot be explained by existing structural models. The range of U2AF dimer and U2AF/SF1 trimer variants used, and the range

of model 3′ SS RNAs described here was limited and could be extended. Insights yielded by the EMSAs require validation through further experimentation, however they are also are able to inform more focused investigations.

Improvements are possible in the precision and accuracy of the K_d measurements by implementing two changes to the protocols. First, although the protein used for EMSAs was very pure the U2AF/SF1 trimer samples will contain varying amounts of contaminant-bound U2AF dimer. This contamination can be completely eliminated by including a final anion exchange chromatography purification step. Second, precision and accuracy can be improved by increasing the number of trials.

Scope exists for additional EMSA experiments that should be informative. With respect to the sequence motifs comprising the 3′ SS, the model RNA sequences tested only two BPS sequences (including the fortuitous BPS in the complement RNA), two PPT sequences, and one yAG sequence at the intron/exon boundary. EMSAs could be performed with model RNAs encompassing the full variation of nucleotide composition for these three sequence motifs. Additionally, the PPT is particularly variable because natural PPTs not only vary widely in their nucleotide composition but also encompass a very broad range of sequence lengths.

In addition to testing additional model RNA sequences, EMSAs could be completed on the phosphorylated counterparts of the U2AF/SF1 trimer complexes used. Phosphomimetic SF1 constructs were used as a proxy to test the effects of phosphorylation on SF1, however the EMSAs could be repeated using phosphorylated SF1 because the RNA binding properties of phosphorylated SF1 may differ from those of phosphomimetic SF1. These are important experiments to complete because the EMSA results suggest that introducing the S131E and S133E dual point mutations to SF1 confer a preference to the U2AF/SF1 complex for a U9 PPT

over a U12 PPT. Because of its potential to alter splice site choice, phosphorylation of SF1 may play roles in alternative splicing.

The EMSAs completed thus far provide the rationale for performing several EMSAs that have the potential to yield additional insights. Comparison of the K_d values of different protein/RNA pairings indicated that protein and RNA components of the U2AF/SF1 complex influence the binding properties of each other in unexpected ways. For example, it appears that SF1 affects the RNA binding properties of U2AF dimer and that this is dependent on SF1 binding its target sequence. Therefore, it is important to complete EMSAs using both the U2AF59 monomer and the U2AF59/SF1 dimer in order to isolate the effects of U2AF23 and U2AF59. Additionally, negative controls for both the PPT and BPS were tested, but not for the AG di-nucleotide at the intron/exon boundary and it would be useful to design and test a negative control for this sequence motif as well.

Finally, analysis of the EMSA results indicate that EMSAs completed using complement RNA should be repeated with two new RNA sequences. The complement RNA was originally intended to serve as a generalized negative control for all three sequence motifs comprising the 3′ SS. However due to the fortuitous BPS, it is more accurately a negative control for U2AF dimer. Therefore, the EMSAs should be repeated with a true negative control RNA in which the fortuitous BPS has been eliminated, as well as a redesigned negative control RNA for U2AF binding in which the fortuitous BPS has been replaced with the optimal BPS. Comparing the K_d values for U2AF/SF1 trimer binding to either scrambled BPS RNA or a true negative control RNA will help determine if SF1 can bind RNA promiscuously.

4-2.2. Extended biochemical characterization of the U2AF dimer and U2AF/SF1 trimer

The tractability of the experimental system makes it very useful for biochemical characterization through a variety of experiments. The ability to quickly and easily clone and express different constructs of U2AF23, U2AF59, and SF1 into this system in order to create various complexes creates the potential to tailor experiments for an isolated investigation of individual protein features within the context of the U2AF dimer and U2AF/SF1 trimer, ranging from point mutations implicated in human disease to multiple domains.

Aside from the investigations described in this thesis, many other experiments can be used to characterize these complexes. Only one gene and isoform exist for U2AF23, U2AF59, and SF1 in *S. pombe*, therefore these proteins are not redundant. Additionally, many molecular biological and genetics tools have been developed for use in *S. pombe*, and it is relatively simple to create strains of *S. pombe* expressing an affinity tagged construct or mutant of any gene of interest expressed from the native promoter and gene locus. Together, these features open the possibility to perform interesting and informative experiments that are not possible in more complex organisms, such as deep mutational scanning of U2AF23, U2AF59, and SF1 combined with analysis of the effects on splicing.

Finally, since the experimental system is easily adapted to produce different variants of *S. pombe* U2AF dimer and U2AF/SF1 trimer complexes, as well as a chimeric complex containing both human and *S. pombe* components, it is likely that the system can be successfully adapted for the human counterpart complexes. This has significant potential to help us understand the alternative splicing functions of different paralogues and isoforms of human U2AF-S, U2AF-L, and SF1.

Ultimately however, it would be useful to obtain an atomic model of the U2AF dimer and U2AF/SF1 trimer in their various apo and RNA-bound states. This would include structures containing both unphosphorylated and phosphorylated SF1, and the RNA-bound structures would need to be solved with a variety of 3′ SS model RNAs in order to understand how U2AF dimer and U2AF/SF1 complexes faithfully recognize the 3′ SS while accommodating a wide variation of sequence length and information content. Ultimately, this is the only way to interpret the findings from the EMSAs described in Chapter 2 and can also be used to make predictions about the complex and inform additional biochemical and structural experiments.

4-3. Future directions: structural characterization of the U2AF dimer and U2AF/SF1 trimer

Biochemical characterization of both the U2AF dimer and U2AF/SF1 trimer indicates that this is an ideal model system for atomic level structural characterization. The three techniques currently useful and available for atomic level structural characterization are NMR, X-ray crystallography, and cryo-EM.

NMR is not a viable option for solving the U2AF dimer or U2AF/SF1 trimer structure, because it is generally restricted to proteins smaller than 30 kDa (Gauto et al., 2019). Solution NMR based structure determination is particularly challenging for proteins larger than 50 kDa (only three NMR-based PDB accessions in this range) (Bertelsen, Chang, Gestwicki, & Zuiderweg, 2009; Gauto et al., 2019; Schwieters et al., 2010; Tugarinov, Choy, Orekhov, & Kay, 2005).

Of the three techniques, X-ray crystallography is the preferred method for atomic resolution structure solution and ~90% of all entries in the PDB are X-ray structures (Gauto et

al., 2019). However, this technique is contingent on being able to obtain diffracting crystals of the target macromolecule or macromolecular assembly and many proteins simply do not crystallize. Extensive screening was performed over the course of several years in order to identify a crystallization hit for both the U2AF dimer and the U2AF/SF1 trimer complexes, which consisted of several thousand crystallization trays. However, a successful crystallization hit condition was never identified.

Despite the recent improvements in cryo-EM, this technology does not yet routinely generate atomic-level structural data, and *de novo* atomic level structure determination from cryo-EM data is not yet the rule. The year 2016 was particularly productive for single-particle cryo-EM, with 961 entries in the EMDB (Electron Microscopy Data Bank). However, only 20% of these deposited maps had a resolution of 3.5 Å or better, making them well suited for *de novo* atomic level structure determination (Gauto et al., 2019).

Because it was not possible to obtain an atomic resolution structure, it was necessary to resort to the SAXS and SEC-SAXS based modeling techniques described in Chapter 3 to structurally characterize the U2AF dimer and U2AF/SF1 trimer.

U2AF and SF1 only participate in the earliest assembly stages of the spliceosome cycle and the cryo-EM structures of these states are from *S. cerevisiae* which does not contain U2AF. Specifically, two E complex structures and one pre-A complex structure have been reported, and only one of the E complex structures contains weak density corresponding to Mud2-Bbp which did not allow modelling of these proteins (Li et al., 2019; Z. W. Zhang et al., 2021).

The E complex structure is the assembly state which must be structurally characterized in order to understand initial recognition of the 3′ SS. It is important to note that initial attempts to use E complex purified from *S. cerevisiae* cell extracts were unsuccessful as they were too

heterogeneous for structural determination. A suitable sample was obtained by assembling the E complex *in vitro* from purified pre-mRNA, U1 snRNP, and co-expressed/co-purified Mud2-Bbp heterodimer (Li et al., 2019). Therefore, in order to successfully obtain the E complex structure from a suitable eukaryotic species containing U2AF/SF1, it may be necessary to express and purify the U2AF/SF1 complex by itself, followed by reconstitution of the E complex from its constituent parts, which is a major technical obstacle to overcome.

Human spliceosomes are far more complex than *S. cerevisiae*, and humans have multiple variants for all three components of the U2AF/SF1 complex, therefore it is very unlikely that a human E complex structure will be reported in the near future due to technical challenges. However, an important intermediate target towards achieving this ultimate goal is a biologically accurate and integrated model of 3′ SS recognition by solving the free and RNA bound structures of the intact U2AF/SF1 complex from a suitable model system, such as the *S. pombe* system described in this thesis. This will achieve several important objectives. First, it is expected to resolve at least some of the outstanding questions that extant, partial structures fail to answer. Second, a tractable and convenient model system to study U2AF/SF1 has been established in this thesis which has been demonstrated to be easily adapted for many variants of this complex; these include subcomponents (i.e. U2AF dimer), mutational variants (i.e. phosphomimetic U2AF/SF1 complex), and complexes containing various truncations of the constituent proteins. Perhaps most importantly, it has been demonstrated that this system is easily adapted to generate the counterpart U2AF/SF1 complex from the fission yeast species *S. cryophilus* and *S. japonicus*, and a chimeric complex with components from both *S. pombe* and human orthologues of these proteins. Additionally, these complexes were successfully obtained using the first cloning strategy and purification attempt with no troubleshooting required. Therefore, it is highly likely

that this system can be adapted to study various human U2AF/SF1 complexes composed of different paralogues and isoforms of the human counterpart proteins. This holds great promise to understand the alternative splicing roles of U2AF/SF1 in developmentally complex eukaryotes. Third, this system and the insights it provides will be able to inform strategies to finally achieve an E complex structure of a eukaryotic species containing the U2AF/SF1 complex. Fourth, the results from Chapter 3 indicate structural plasticity of the U2AF/SF1 complex in both its free and RNA bound states, however the range and functional role of this plasticity remains to be seen. Solving the structures of free and RNA bound U2AF/SF1 and solving the same structures in the context of the E complex may reveal biologically important conformational differences that provide insight into the transformations that occur in the transition from the non-specific H complex to the E complex.

References

Abelson, J., Blanco, M., Ditzler, M. A., Fuller, F., Aravamudhan, P., Wood, M., . . . Walter, N. G. (2010). Conformational dynamics of single pre-mRNA molecules during in vitro splicing. *Nat Struct Mol Biol, 17*(4), 504-512. doi:10.1038/nsmb.1767

Abovich, N., Liao, X. C., & Rosbash, M. (1994). The yeast MUD2 protein: an interaction with PRP11 defines a bridge between commitment complexes and U2 snRNP addition. *Genes Dev, 8*(7), 843-854. doi:10.1101/gad.8.7.843

Abovich, N., & Rosbash, M. (1997). Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. *Cell, 89*(3), 403-412. doi:10.1016/s0092-8674(00)80221-4

Absmeier, E., Santos, K. F., & Wahl, M. C. (2016). Functions and regulation of the Brr2 RNA helicase during splicing. *Cell Cycle, 15*(24), 3362-3377. doi:10.1080/15384101.2016.1249549

Adam, S. A., Nakagawa, T., Swanson, M. S., Woodruff, T. K., & Dreyfuss, G. (1986). mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence. *Mol Cell Biol, 6*(8), 2932-2943. doi:10.1128/mcb.6.8.2932

Agafonov, D. E., Deckert, J., Wolf, E., Odenwalder, P., Bessonov, S., Will, C. L., . . . Luhrmann, R. (2011). Semiquantitative proteomic analysis of the human spliceosome via a novel two-dimensional gel electrophoresis method. *Mol Cell Biol, 31*(13), 2667-2682. doi:10.1128/MCB.05266-11

Agafonov, D. E., Kastner, B., Dybkov, O., Hofele, R. V., Liu, W. T., Urlaub, H., . . . Stark, H. (2016). Molecular architecture of the human U4/U6.U5 tri-snRNP. *Science, 351*(6280), 1416-1420. doi:10.1126/science.aad2085

Agrawal, A. A., McLaughlin, K. J., Jenkins, J. L., & Kielkopf, C. L. (2014). Structure-guided U2AF65 variant improves recognition and splicing of a defective pre-mRNA. *Proc Natl Acad Sci U S A, 111*(49), 17420-17425. doi:10.1073/pnas.1412743111

Agrawal, A. A., Salsi, E., Chatrikhi, R., Henderson, S., Jenkins, J. L., Green, M. R., . . . Kielkopf, C. L. (2016). An extended U2AF(65)-RNA-binding domain recognizes the 3' splice site signal. *Nat Commun, 7*, 10950. doi:10.1038/ncomms10950

Albrecht, U. (2004). The mammalian circadian clock: a network of gene expression. *Front Biosci, 9*, 48-55. doi:10.2741/1196

Allain, F. H., Bouvet, P., Dieckmann, T., & Feigon, J. (2000). Molecular basis of sequence-specific recognition of pre-ribosomal RNA by nucleolin. *EMBO J, 19*(24), 6870-6881. doi:10.1093/emboj/19.24.6870

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/S0022-2836(05)80360-2

Amara, S. G., Jonas, V., Rosenfeld, M. G., Ong, E. S., & Evans, R. M. (1982). Alternative RNA processing in calcitonin gene expression generates mRNAs encoding different polypeptide products. *Nature, 298*(5871), 240-244. doi:10.1038/298240a0

Anna, A., & Monika, G. (2018). Splicing mutations in human genetic disorders: examples, detection, and confirmation. *Journal of Applied Genetics, 59*(3), 253-268. doi:10.1007/s13353-018-0444-7

Aravind, L., Watanabe, H., Lipman, D. J., & Koonin, E. V. (2000). Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc Natl Acad Sci U S A, 97*(21), 11319-11324. doi:10.1073/pnas.200346997

Arenas, J. E., & Abelson, J. N. (1997). Prp43: An RNA helicase-like factor involved in spliceosome disassembly. *Proc Natl Acad Sci U S A, 94*(22), 11798-11802. doi:10.1073/pnas.94.22.11798

Ares, M., Jr., & Weiser, B. (1995). Rearrangement of snRNA structure during assembly and function of the spliceosome. *Prog Nucleic Acid Res Mol Biol, 50*, 131-159. doi:10.1016/s0079-6603(08)60813-2

Arning, S., Gruter, P., Bilbe, G., & Kramer, A. (1996). Mammalian splicing factor SF1 is encoded by variant cDNAs and binds to RNA. *RNA, 2*(8), 794-810. Retrieved from <Go to ISI>://WOS:A1996VD46700006

Aslanidis, C., & de Jong, P. J. (1990). Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Res, 18*(20), 6069-6074. doi:10.1093/nar/18.20.6069

Auboeuf, D., Dowhan, D. H., Kang, Y. K., Larkin, K., Lee, J. W., Berget, S. M., & O'Malley, B. W. (2004). Differential recruitment of nuclear receptor coactivators may determine alternative RNA splice site choice in target genes. *Proceedings of the National Academy of Sciences of the United States of America, 101*(8), 2270-2274. doi:10.1073/pnas.0308133100

Awan, A. R., Manfredo, A., & Pleiss, J. A. (2013). Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. *Proceedings of the National Academy of Sciences of the United States of America, 110*(31), 12762-12767. doi:10.1073/pnas.1218353110

Baber, J. L., Libutti, D., Levens, D., & Tjandra, N. (1999). High precision solution structure of the C-terminal KH domain of heterogeneous nuclear ribonucleoprotein K, a c-myc transcription factor. *J Mol Biol, 289*(4), 949-962. doi:10.1006/jmbi.1999.2818

Baehrecke, E. H. (1997). who encodes a KH RNA binding protein that functions in muscle development. *Development, 124*(7), 1323-1332. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9118803

Bai, R., Wan, R., Wang, L., Xu, K., Zhang, Q., Lei, J., & Shi, Y. (2021). Structure of the activated human minor spliceosome. *Science, 371*(6535). doi:10.1126/science.abg0879

Bai, R., Wan, R., Yan, C., Jia, Q., Lei, J., & Shi, Y. (2021). Mechanism of spliceosome remodeling by the ATPase/helicase Prp2 and its coactivator Spp2. *Science, 371*(6525). doi:10.1126/science.abe8863

Bai, R., Wan, R., Yan, C., Lei, J., & Shi, Y. (2018). Structures of the fully assembled Saccharomyces cerevisiae spliceosome before activation. *Science, 360*(6396), 1423-1429. doi:10.1126/science.aau0325

Bai, R., Yan, C., Wan, R., Lei, J., & Shi, Y. (2017). Structure of the Post-catalytic Spliceosome from Saccharomyces cerevisiae. *Cell, 171*(7), 1589-1598 e1588. doi:10.1016/j.cell.2017.10.038

Baldwin, K. L., Dinh, E. M., Hart, B. M., & Masson, P. H. (2013). CACTIN is an essential nuclear protein in Arabidopsis and may be associated with the eukaryotic spliceosome. *Febs Letters, 587*(7), 873-879. doi:10.1016/j.febslet.2013.02.041

Banerjee, A., & Verdine, G. L. (2006). A nucleobase lesion remodels the interaction of its normal neighbor in a DNA glycosylase complex. *Proc Natl Acad Sci U S A, 103*(41), 15020-15025. doi:10.1073/pnas.0603644103

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., . . . Frey, B. J. (2010). Deciphering the splicing code. *Nature, 465*(7294), 53-59. doi:10.1038/nature09000

Barbosa-Morais, N. L., Carmo-Fonseca, M., & Aparicio, S. (2006). Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res, 16*(1), 66-77. doi:10.1101/gr.3936206

Beausoleil, S. A., Jedrychowski, M., Schwartz, D., Elias, J. E., Villen, J., Li, J., . . . Gygi, S. P. (2004). Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A, 101*(33), 12130-12135. doi:10.1073/pnas.0404720101

Bellayou, H., Hamzi, K., Rafai, M. A., Karkouri, M., Slassi, I., Azeddoug, H., & Nadifi, S. (2009). Duchenne and Becker muscular dystrophy: contribution of a molecular and immunohistochemical analysis in diagnosis in Morocco. *J Biomed Biotechnol, 2009*, 325210. doi:10.1155/2009/325210

Bennett, M., Pinol-Roma, S., Staknis, D., Dreyfuss, G., & Reed, R. (1992). Differential binding of heterogeneous nuclear ribonucleoproteins to mRNA precursors prior to spliceosome assembly in vitro. *Mol Cell Biol, 12*(7), 3165-3175. doi:10.1128/mcb.12.7.3165

Berget, S. M. (1995). Exon recognition in vertebrate splicing. *J Biol Chem, 270*(6), 2411-2414. doi:10.1074/jbc.270.6.2411

Berget, S. M., Moore, C., & Sharp, P. A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A, 74*(8), 3171-3175. doi:10.1073/pnas.74.8.3171

Berglund, J. A., Abovich, N., & Rosbash, M. (1998). A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. *Genes Dev, 12*(6), 858-867. doi:10.1101/gad.12.6.858

Berglund, J. A., Chua, K., Abovich, N., Reed, R., & Rosbash, M. (1997). The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell, 89*(5), 781-787. doi:10.1016/s0092-8674(00)80261-5

Berglund, J. A., Fleming, M. L., & Rosbash, M. (1998). The KH domain of the branchpoint sequence binding protein determines specificity for the pre-mRNA branchpoint sequence. *RNA, 4*(8), 998-1006. doi:10.1017/s1355838298980499

Berglund, J. A., Rosbash, M., & Schultz, S. C. (2001). Crystal structure of a model branchpoint-U2 snRNA duplex containing bulged adenosines. *RNA, 7*(5), 682-691. doi:10.1017/s1355838201002187

Bernado, P., Mylonas, E., Petoukhov, M. V., Blackledge, M., & Svergun, D. I. (2007). Structural characterization of flexible proteins using small-angle X-ray scattering. *Journal of the American Chemical Society, 129*(17), 5656-5664. doi:10.1021/ja069124n

Bernier, F. P., Caluseriu, O., Ng, S., Schwartzentruber, J., Buckingham, K. J., Innes, A. M., . . . Parboosingh, J. S. (2012). Haploinsufficiency of SF3B4, a component of the pre-mRNA spliceosomal complex, causes Nager syndrome. *Am J Hum Genet, 90*(5), 925-933. doi:10.1016/j.ajhg.2012.04.004

Bertelsen, E. B., Chang, L., Gestwicki, J. E., & Zuiderweg, E. R. (2009). Solution conformation of wild-type E. coli Hsp70 (DnaK) chaperone complexed with ADP and substrate. *Proc Natl Acad Sci U S A, 106*(21), 8471-8476. doi:10.1073/pnas.0903503106

Bertram, K., Agafonov, D. E., Dybkov, O., Haselbach, D., Leelaram, M. N., Will, C. L., . . . Stark, H. (2017). Cryo-EM Structure of a Pre-catalytic Human Spliceosome Primed for Activation. *Cell, 170*(4), 701-713 e711. doi:10.1016/j.cell.2017.07.011

Bertram, K., Agafonov, D. E., Liu, W. T., Dybkov, O., Will, C. L., Hartmuth, K., . . . Luhrmann, R. (2017). Cryo-EM structure of a human spliceosome activated for step 2 of splicing. *Nature, 542*(7641), 318-323. doi:10.1038/nature21079

Bertram, K., El Ayoubi, L., Dybkov, O., Agafonov, D. E., Will, C. L., Hartmuth, K., . . . Luhrmann, R. (2020). Structural Insights into the Roles of Metazoan-Specific Splicing Factors in the Human Step 1 Spliceosome. *Molecular Cell, 80*(1), 127-+. doi:10.1016/j.molcel.2020.09.012

Bessonov, S., Anokhina, M., Krasauskas, A., Golas, M. M., Sander, B., Will, C. L., . . . Luhrmann, R. (2010). Characterization of purified human Bact spliceosomal complexes reveals compositional and morphological changes during spliceosome activation and first step catalysis. *RNA, 16*(12), 2384-2403. doi:10.1261/rna.2456210

Bessonov, S., Anokhina, M., Will, C. L., Urlaub, H., & Luhrmann, R. (2008). Isolation of an active step I spliceosome and composition of its RNP core. *Nature, 452*(7189), 846-850. doi:10.1038/nature06842

Bialkowska, A., & Kurlandzka, A. (2002). Proteins interacting with Lin1p, a putative link between chromosome segregation, mRNA splicing and DNA replication in Saccharomyces cerevisiae. *Yeast, 19*(15), 1323-1333. doi:10.1002/yea.919

Bindereif, A., Wolff, T., & Green, M. R. (1990). Discrete domains of human U6 snRNA required for the assembly of U4/U6 snRNP and splicing complexes. *EMBO J, 9*(1), 251-255. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/2136831

Birney, E., Kumar, S., & Krainer, A. R. (1993). Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. *Nucleic Acids Res, 21*(25), 5803-5816. doi:10.1093/nar/21.25.5803

Bitton, D. A., Atkinson, S. R., Rallis, C., Smith, G. C., Ellis, D. A., Chen, Y. Y., . . . Bahler, J. (2015). Widespread exon skipping triggers degradation by nuclear RNA surveillance in fission yeast. *Genome Res, 25*(6), 884-896. doi:10.1101/gr.185371.114

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem, 72*, 291-336. doi:10.1146/annurev.biochem.72.121801.161720

Blencowe, B. J., & Graveley, B. R. (2010). *Alternative Splicing in the Postgenomic Era*: Springer New York.

Bohnsack, M. T., Martin, R., Granneman, S., Ruprecht, M., Schleiff, E., & Tollervey, D. (2009). Prp43 Bound at Different Sites on the Pre-rRNA Performs Distinct Functions in Ribosome Synthesis. *Molecular Cell, 36*(4), 583-592. doi:10.1016/j.molcel.2009.09.039

Bolivar, F., Rodriguez, R. L., Greene, P. J., Betlach, M. C., Heyneker, H. L., Boyer, H. W., . . . Falkow, S. (1977). Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. *Gene, 2*(2), 95-113. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/344137

Bonen, L., & Vogel, J. (2001). The ins and outs of group II introns. *Trends Genet, 17*(6), 322-331. doi:10.1016/s0168-9525(01)02324-1

Borneff, J. (1959). Zentralblatt Bakteriologie, Parasitenkunde, Infektions-krankheiten und Hygiene. *Gustav Fischer Verlag, Stuttgart, I. orig, 176*, 193-194.

Bratu, D. P., Cha, B. J., Mhlanga, M. M., Kramer, F. R., & Tyagi, S. (2003). Visualizing the distribution and transport of mRNAs in living cells. *Proc Natl Acad Sci U S A, 100*(23), 13308-13313. doi:10.1073/pnas.2233244100

Brennwald, P., Porter, G., & Wise, J. A. (1988). U2 Small Nuclear-Rna Is Remarkably Conserved between Schizosaccharomyces-Pombe and Mammals. *Molecular and Cellular Biology, 8*(12), 5575-5580. doi:Doi 10.1128/Mcb.8.12.5575

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*(13), 3509-3516. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/2951249

Brokstad, K. A., Kalland, K. H., Russell, W. C., & Matthews, D. A. (2001). Mitochondrial protein p32 can accumulate in the nucleus. *Biochem Biophys Res Commun, 281*(5), 1161-1169. doi:10.1006/bbrc.2001.4473

Brooks, B. R., Brooks, C. L., 3rd, Mackerell, A. D., Jr., Nilsson, L., Petrella, R. J., Roux, B., . . . Karplus, M. (2009). CHARMM: the biomolecular simulation program. *J Comput Chem, 30*(10), 1545-1614. doi:10.1002/jcc.21287

Brosi, R., Groning, K., Behrens, S. E., Luhrmann, R., & Kramer, A. (1993). Interaction of mammalian splicing factor SF3a with U2 snRNP and relation of its 60-kD subunit to yeast PRP9. *Science, 262*(5130), 102-105. doi:10.1126/science.8211112

Brosi, R., Hauri, H. P., & Kramer, A. (1993). Separation of splicing factor SF3 into two components and purification of SF3a activity. *J Biol Chem, 268*(23), 17640-17646. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8349644

Brow, D. A. (2002). Allosteric cascade of spliceosome activation. *Annu Rev Genet, 36*, 333-360. doi:10.1146/annurev.genet.36.043002.091635

Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., . . . Warren, G. L. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr, 54*(Pt 5), 905-921. doi:10.1107/s0907444998003254

Brysch-Herzberg, M., Jia, G. S., Seidel, M., Assali, I., & Du, L. L. (2022). Insights into the ecology of Schizosaccharomyces species in natural and artificial habitats. *Antonie Van Leeuwenhoek*

International Journal of General and Molecular Microbiology, 115(5), 661-695. doi:10.1007/s10482-022-01720-0

Brysch-Herzberg, M., Tobias, A., Seidel, M., Wittmann, R., Wohlmann, E., Fischer, R., . . . Peter, G. (2019). Schizosaccharomyces osmophilus sp. nov., an osmophilic fission yeast occurring in bee bread of different solitary bee species. *FEMS Yeast Res, 19*(4). doi:10.1093/femsyr/foz038

Buchan, D. W., Minneci, F., Nugent, T. C., Bryson, K., & Jones, D. T. (2013). Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res, 41*(Web Server issue), W349-357. doi:10.1093/nar/gkt381

Burge, C. B., Tuschl, T., & Sharp, P. A. (1999). Splicing of precursors to mRNAs by the Spliceosome. In R. F. Gesteland, T. R. Cech, & J. F. Atkins (Eds.), *The RNA World, 2nd Ed.* (pp. 525-560). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Burgute, B. D., Peche, V. S., Steckelberg, A. L., Glockner, G., Gassen, B., Gehring, N. H., & Noegel, A. A. (2014). NKAP is a novel RS-related protein that interacts with RNA and RNA binding proteins. *Nucleic Acids Research, 42*(5), 3177-3193. doi:10.1093/nar/gkt1311

Busch, A., & Hertel, K. J. (2012). Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdiscip Rev RNA, 3*(1), 1-12. doi:10.1002/wrna.100

Calvin, K., & Li, H. (2008). RNA-splicing endonuclease structure and function. *Cell Mol Life Sci, 65*(7-8), 1176-1185. doi:10.1007/s00018-008-7393-y

Campbell, J. W., Duee, E., Hodgson, G., Mercer, W. D., Stammers, D. K., Wendell, P. L., . . . Watson, H. C. (1972). X-ray diffraction studies on enzymes in the glycolytic pathway. *Cold Spring Harb Symp Quant Biol, 36*, 165-170. doi:10.1101/sqb.1972.036.01.023

Cancilla, M. T., He, M. M., Viswanathan, N., Simmons, R. L., Taylor, M., Fung, A. D., . . . Erlanson, D. A. (2008). Discovery of an Aurora kinase inhibitor through site-specific dynamic combinatorial chemistry. *Bioorg Med Chem Lett, 18*(14), 3978-3981. doi:10.1016/j.bmcl.2008.06.011

Cascino, I., Fiucci, G., Papoff, G., & Ruberti, G. (1995). Three functional soluble forms of the human apoptosis-inducing Fas molecule are produced by alternative splicing. *J Immunol, 154*(6), 2706-2713. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7533181

Case, D. A., Aktulga, H. M., Belfon, K., Ben-Shalom, I., Brozell, S. R., Cerutti, D. S., . . . Duke, R. E. (2021). *Amber 2021*: University of California, San Francisco.

Caslini, C., Spinelli, O., Cazzaniga, G., Golay, J., De Gioia, L., Pedretti, A., . . . Rambaldi, A. (1997). Identification of two novel isoforms of the ZNF162 gene: a growing family of signal transduction and activator of RNA proteins. *Genomics, 42*(2), 268-277. doi:10.1006/geno.1997.4705

Caspary, F., & Seraphin, B. (1998). The yeast U2A'/U2B complex is required for pre-spliceosome formation. *EMBO J, 17*(21), 6348-6358. doi:10.1093/emboj/17.21.6348

Cech, T. R. (1986). The generality of self-splicing RNA: relationship to nuclear mRNA splicing. *Cell, 44*(2), 207-210. doi:10.1016/0092-8674(86)90751-8

Cech, T. R. (1990). Self-splicing of group I introns. *Annu Rev Biochem, 59*, 543-568. doi:10.1146/annurev.bi.59.070190.002551

Chabot, B., & Shkreta, L. (2016). Defective control of pre-messenger RNA splicing in human disease. *J Cell Biol, 212*(1), 13-27. doi:10.1083/jcb.201510032

Chakarova, C. F., Hims, M. M., Bolz, H., Abu-Safieh, L., Patel, R. J., Papaioannou, M. G., . . . Bhattacharya, S. S. (2002). Mutations in HPRP3, a third member of pre-mRNA splicing factor genes, implicated in autosomal dominant retinitis pigmentosa. *Hum Mol Genet, 11*(1), 87-92. doi:10.1093/hmg/11.1.87

Chan, S. P., Kao, D. I., Tsai, W. Y., & Cheng, S. C. (2003). The Prp19p-associated complex in spliceosome activation. *Science, 302*(5643), 279-282. doi:10.1126/science.1086602

Chang, A. C., & Cohen, S. N. (1978). Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. *J Bacteriol, 134*(3), 1141-1156. doi:10.1128/jb.134.3.1141-1156.1978

Charenton, C., Wilkinson, M. E., & Nagai, K. (2019). Mechanism of 5' splice site transfer for human spliceosome activation. *Science, 364*(6438), 362-367. doi:10.1126/science.aax3289

Chen, M., & Manley, J. L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol, 10*(11), 741-754. doi:10.1038/nrm2777

Chen, W., Shulha, H. P., Ashar-Patel, A., Yan, J., Green, K. M., Query, C. C., . . . Moore, M. J. (2014). Endogenous U2.U5.U6 snRNA complexes in S. pombe are intron lariat spliceosomes. *RNA, 20*(3), 308-320. doi:10.1261/rna.040980.113

Chen, X., Yin, J., Cao, D., Xiao, D., Zhou, Z., Liu, Y., & Shou, W. (2021). The Emerging Roles of the RNA Binding Protein QKI in Cardiovascular Development and Function. *Front Cell Dev Biol, 9*, 668659. doi:10.3389/fcell.2021.668659

Chen, X., Zhang, H., Aravindakshan, J. P., Gotlieb, W. H., & Sairam, M. R. (2011). Anti-proliferative and pro-apoptotic actions of a novel human and mouse ovarian tumor-associated gene OTAG-12: downregulation, alternative splicing and drug sensitization. *Oncogene, 30*(25), 2874-2887. doi:10.1038/onc.2011.11

Cheng, J., Zhou, T., Liu, C., Shapiro, J. P., Brauer, M. J., Kiefer, M. C., . . . Mountz, J. D. (1994). Protection from Fas-mediated apoptosis by a soluble form of the Fas molecule. *Science, 263*(5154), 1759-1762. doi:10.1126/science.7510905

Chow, L. T., Gelinas, R. E., Broker, T. R., & Roberts, R. J. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell, 12*(1), 1-8. doi:10.1016/0092-8674(77)90180-5

Chu, C. S., Trapnell, B. C., Curristin, S., Cutting, G. R., & Crystal, R. G. (1993). Genetic basis of variable exon 9 skipping in cystic fibrosis transmembrane conductance regulator mRNA. *Nat Genet, 3*(2), 151-156. doi:10.1038/ng0293-151

Clark, T. A., Sugnet, C. W., & Ares, M. (2002). Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science, 296*(5569), 907-910. doi:DOI 10.1126/science.1069415

Classen, S., Hura, G. L., Holton, J. M., Rambo, R. P., Rodic, I., McGuire, P. J., . . . Tainer, J. A. (2013). Implementation and performance of SIBYLS: a dual endstation small-angle X-ray scattering and macromolecular crystallography beamline at the Advanced Light Source. *Journal of Applied Crystallography, 46*(Pt 1), 1-13. doi:10.1107/S0021889812048698

Clery, A., Blatter, M., & Allain, F. H. (2008). RNA recognition motifs: boring? Not quite. *Curr Opin Struct Biol, 18*(3), 290-298. doi:10.1016/j.sbi.2008.04.002

Cline, M. S., Shigeta, R., Wheeler, R. L., Siani-Rose, M. A., Kulp, D., & Loraine, A. E. (2004). The effects of alternative splicing on transmembrane proteins in the mouse genome. *Pac Symp Biocomput*, 17-28. doi:10.1142/9789812704856_0003

Colwill, K., Pawson, T., Andrews, B., Prasad, J., Manley, J. L., Bell, J. C., & Duncan, P. I. (1996). The Clk/Sty protein kinase phosphorylates SR splicing factors and regulates their intranuclear distribution. *Embo Journal, 15*(2), 265-275. doi:DOI 10.1002/j.1460-2075.1996.tb00357.x

Company, M., Arenas, J., & Abelson, J. (1991). Requirement of the RNA helicase-like protein PRP22 for release of messenger RNA from spliceosomes. *Nature, 349*(6309), 487-493. doi:10.1038/349487a0

Coolidge, C. J., Seely, R. J., & Patton, J. G. (1997). Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Research, 25*(4), 888-895. doi:DOI 10.1093/nar/25.4.888

Cordin, O., Hahn, D., & Beggs, J. D. (2012). Structure, function and regulation of spliceosomal RNA helicases. *Curr Opin Cell Biol, 24*(3), 431-438. doi:10.1016/j.ceb.2012.03.004

Corioni, M., Antih, N., Tanackovic, G., Zavolan, M., & Kramer, A. (2011). Analysis of in situ pre-mRNA targets of human splicing factor SF1 reveals a function in alternative splicing. *Nucleic Acids Res, 39*(5), 1868-1879. doi:10.1093/nar/gkq1042

Corn, J. E., & Berger, J. M. (2007). FASTDXL: a generalized screen to trap disulfide-stabilized complexes for use in structural studies. *Structure, 15*(7), 773-780. doi:10.1016/j.str.2007.05.006

Crawford, D. J., Hoskins, A. A., Friedman, L. J., Gelles, J., & Moore, M. J. (2013). Single-molecule colocalization FRET evidence that spliceosome activation precedes stable approach of 5' splice site and branch site. *Proc Natl Acad Sci U S A, 110*(17), 6783-6788. doi:10.1073/pnas.1219305110

Cretu, C., Schmitzova, J., Ponce-Salvatierra, A., Dybkov, O., De Laurentiis, E. I., Sharma, K., . . . Pena, V. (2016). Molecular Architecture of SF3b and Structural Consequences of Its Cancer-Related Mutations. *Mol Cell, 64*(2), 307-319. doi:10.1016/j.molcel.2016.08.036

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res, 14*(6), 1188-1190. doi:10.1101/gr.849004

Cudney, R., Patel, S., Weisgraber, K., Newhouse, Y., & McPherson, A. (1994). Screening and optimization strategies for macromolecular crystal growth. *Acta Crystallogr D Biol Crystallogr, 50*(Pt 4), 414-423. doi:10.1107/S0907444994002660

Dale, G. E., Oefner, C., & D'Arcy, A. (2003). The protein as a variable in protein crystallization. *Journal of Structural Biology, 142*(1), 88-97. doi:10.1016/S1047-8477(03)00041-8

Darlix, J. L., Lapadattapolsky, M., Derocquigny, H., & Roques, B. P. (1995). First Glimpses at Structure-Function-Relationships of the Nucleocapsid Protein of Retroviruses. *Journal of Molecular Biology, 254*(4), 523-537. doi:DOI 10.1006/jmbi.1995.0635

Das, R., & Reed, R. (1999). Resolution of the mammalian E complex and the ATP-dependent spliceosomal complexes on native agarose mini-gels. *RNA, 5*(11), 1504-1508. doi:10.1017/s1355838299991501

Das, R., Zhou, Z., & Reed, R. (2000). Functional association of U2 snRNP with the ATP-independent spliceosomal complex E. *Mol Cell, 5*(5), 779-787. doi:10.1016/s1097-2765(00)80318-4

De Conti, L., Baralle, M., & Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA, 4*(1), 49-60. doi:10.1002/wrna.1140

De, I., Bessonov, S., Hofele, R., dos Santos, K., Will, C. L., Urlaub, H., . . . Pena, V. (2015). The RNA helicase Aquarius exhibits structural adaptations mediating its recruitment to spliceosomes. *Nat Struct Mol Biol, 22*(2), 138-144. doi:10.1038/nsmb.2951

Deo, R. C., Bonanno, J. B., Sonenberg, N., & Burley, S. K. (1999). Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell, 98*(6), 835-845. doi:10.1016/s0092-8674(00)81517-2

Dibner, C., Schibler, U., & Albrecht, U. (2010). The mammalian circadian timing system: organization and coordination of central and peripheral clocks. *Annu Rev Physiol, 72*, 517-549. doi:10.1146/annurev-physiol-021909-135821

Ding, F., Hagan, J. P., Wang, Z., & Grabowski, P. J. (1996). Biochemical properties of a novel U2AF65 protein isoform generated by alternative RNA splicing. *Biochem Biophys Res Commun, 224*(3), 675-683. doi:10.1006/bbrc.1996.1083

Ding, W. Q., Kuntz, S. M., & Miller, L. J. (2002). A misspliced form of the cholecystokinin-B/gastrin receptor in pancreatic carcinoma: role of reduced sellular U2AF35 and a suboptimal 3'-splicing site leading to retention of the fourth intron. *Cancer Res, 62*(3), 947-952. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11830556

Domsic, J. K., Wang, Y., Mayeda, A., Krainer, A. R., & Stoltzfus, C. M. (2003). Human immunodeficiency virus type 1 hnRNP A/B-dependent exonic splicing silencer ESSV antagonizes binding of U2AF65 to viral polypyrimidine tracts. *Mol Cell Biol, 23*(23), 8762-8772. doi:10.1128/mcb.23.23.8762-8772.2003

Donnelly, M. I., Zhou, M., Millard, C. S., Clancy, S., Stols, L., Eschenfeldt, W. H., . . . Joachimiak, A. (2006). An expression vector tailored for large-scale, high-throughput purification of recombinant proteins. *Protein Expr Purif, 47*(2), 446-454. doi:10.1016/j.pep.2005.12.011

Dowhan, D. H., Hong, E. P., Auboeuf, D., Dennis, A. P., Wilson, M. M., Berget, S. M., & O'Malley, B. W. (2005). Steroid hormone receptor coactivation and alternative RNA splicing by U2AF65-related proteins CAPERalpha and CAPERbeta. *Mol Cell, 17*(3), 429-439. doi:10.1016/j.molcel.2004.12.025

Drabenstot, S. D., Kupfer, D. M., White, J. D., Dyer, D. W., Roe, B. A., Buchanan, K. L., & Murphy, J. W. (2003). FELINES: a utility for extracting and examining EST-defined introns and exons. *Nucleic Acids Res, 31*(22), e141. doi:10.1093/nar/gng141

Dreyfuss, G., Swanson, M. S., & Pinol-Roma, S. (1988). Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. *Trends Biochem Sci, 13*(3), 86-91. doi:10.1016/0968-0004(88)90046-1

Ducruix, A., & Giegé, R. (1992). Crystallization of Nucleic Acids and Proteins: A Practical Approach. In. Oxford [England] ;: IRL Press at Oxford University Press.

Duft, K. (2014). *M.Sc. Thesis: The Role of U2AF in pre-mRNA Splicing.* (M.Sc.). University of Potsdam, Potsdam, Germany.

Duncan, P. I., Stojdl, D. F., Marius, R. M., & Bell, J. C. (1997). In vivo regulation of alternative pre-mRNA splicing by the Clk1 protein kinase. *Mol Cell Biol, 17*(10), 5996-6001. doi:10.1128/MCB.17.10.5996

Dunn, E. A., & Rader, S. D. (2014). Pre-mRNA splicing and the spliceosome: assembly, catalysis, and fidelity. In A. Sesma & T. von der Haar (Eds.), *Fungal RNA Biology* (pp. 27-57). Cham: Springer.

Duong, H. A., Robles, M. S., Knutti, D., & Weitz, C. J. (2011). A molecular mechanism for circadian clock negative feedback. *Science, 332*(6036), 1436-1439. doi:10.1126/science.1196766

Dyer, K. N., Hammel, M., Rambo, R. P., Tsutakawa, S. E., Rodic, I., Classen, S., . . . Hura, G. L. (2014). High-throughput SAXS for the characterization of biomolecules in solution: a practical approach. *Methods Mol Biol, 1091*, 245-258. doi:10.1007/978-1-62703-691-7_18

Dziembowski, A., Ventura, A. P., Rutz, B., Caspary, F., Faux, C., Halgand, F., . . . Seraphin, B. (2004). Proteomic analysis identifies a new complex required for nuclear pre-mRNA retention and splicing. *Embo Journal, 23*(24), 4847-4856. doi:10.1038/sj.emboj.7600482

Egel, R. (2004). *The Molecular Biology of Schizosaccharomyces pombe: Genetics, Genomics and Beyond*. Heidelberg, Germany: Springer-Verlag.

Emsley, P., Lohkamp, B., Scott, W. G., & Cowtan, K. (2010). Features and development of Coot. *Acta Crystallographica Section D-Biological Crystallography, 66*, 486-501. doi:10.1107/S0907444910007493

Erlanson, D. A., Braisted, A. C., Raphael, D. R., Randal, M., Stroud, R. M., Gordon, E. M., & Wells, J. A. (2000). Site-directed ligand discovery. *Proc Natl Acad Sci U S A, 97*(17), 9367-9372. doi:10.1073/pnas.97.17.9367

Erlanson, D. A., Lam, J. W., Wiesmann, C., Luong, T. N., Simmons, R. L., DeLano, W. L., . . . O'Brien, T. (2003). In situ assembly of enzyme inhibitors using extended tethering. *Nature Biotechnology, 21*(3), 308-314. doi:10.1038/nbt786

Erlanson, D. A., McDowell, R. S., He, M. M., Randal, M., Simmons, R. L., Kung, J., . . . Hansen, S. K. (2003). Discovery of a new phosphotyrosine mimetic for PTP1B using breakaway tethering. *Journal of the American Chemical Society, 125*(19), 5602-5603. doi:10.1021/ja034440c

Eschenfeldt, W. H., Lucy, S., Millard, C. S., Joachimiak, A., & Mark, I. D. (2009). A family of LIC vectors for high-throughput cloning and purification of proteins. *Methods Mol Biol, 498*, 105-115. doi:10.1007/978-1-59745-196-3_7

Esfahani, M. S., Lee, L. J., Jeon, Y. J., Flynn, R. A., Stehr, H., Hui, A. B., . . . Diehn, M. (2019). Functional significance of U2AF1 S34F mutations in lung adenocarcinomas. *Nature Communications, 10*. doi:ARTN 5712

10.1038/s41467-019-13392-y

Ester, C., & Uetz, P. (2008). The FF domains of yeast U1 snRNP protein Prp40 mediate interactions with Luc7 and Snu71. *BMC Biochem, 9*, 29. doi:10.1186/1471-2091-9-29

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., . . . Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics, 23*(22), 5866-5878. doi:10.1093/hmg/ddu309

Fabrizio, P., Dannenberg, J., Dube, P., Kastner, B., Stark, H., Urlaub, H., & Luhrmann, R. (2009). The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. *Mol Cell, 36*(4), 593-608. doi:10.1016/j.molcel.2009.09.040

Fair, B. J., & Pleiss, J. A. (2017). The power of fission: yeast as a tool for understanding complex splicing. *Current Genetics, 63*(3), 375-380. doi:10.1007/s00294-016-0647-6

Farini, D., Cesari, E., Weatheritt, R. J., La Sala, G., Naro, C., Pagliarini, V., . . . Sette, C. (2020). A Dynamic Splicing Program Ensures Proper Synaptic Connections in the Developing Cerebellum. *Cell Rep, 31*(9), 107703. doi:10.1016/j.celrep.2020.107703

Farkas, M. H., Grant, G. R., & Pierce, E. A. (2012). Transcriptome Analyses to Investigate the Pathogenesis of RNA Splicing Factor Retinitis Pigmentosa. *Retinal Degenerative Diseases, 723*, 519-525. doi:10.1007/978-1-4614-0631-0_65

Faustino, N. A., & Cooper, T. A. (2003). Pre-mRNA splicing and human disease. *Genes & Development, 17*(4), 419-437. doi:10.1101/gad.1048803

Fei, D. L., Motowski, H., Chatrikhi, R., Prasad, S., Yu, J., Gao, S. J., . . . Varmus, H. (2016). Wild-Type U2AF1 Antagonizes the Splicing Program Characteristic of U2AF1-Mutant Tumors and Is Required for Cell Survival. *Plos Genetics, 12*(10). doi:ARTN e1006384

10.1371/journal.pgen.1006384

Fernandez-Leiro, R., & Scheres, S. H. (2016). Unravelling biological macromolecules with cryo-electron microscopy. *Nature, 537*(7620), 339-346. doi:10.1038/nature19948

Ferre-D'Amare, A. R., & Doudna, J. A. (2001). Methods to crystallize RNA. *Curr Protoc Nucleic Acid Chem, Chapter 7*, Unit 7 6. doi:10.1002/0471142700.nc0706s00

Fica, S. M., Oubridge, C., Galej, W. P., Wilkinson, M. E., Bai, X. C., Newman, A. J., & Nagai, K. (2017). Structure of a spliceosome remodelled for exon ligation. *Nature, 542*(7641), 377-+. doi:10.1038/nature21078

Fica, S. M., Oubridge, C., Wilkinson, M. E., Newman, A. J., & Nagai, K. (2019). A human postcatalytic spliceosome structure reveals essential roles of metazoan factors for exon ligation. *Science, 363*(6428), 710-714. doi:10.1126/science.aaw5569

Fica, S. M., Tuttle, N., Novak, T., Li, N. S., Lu, J., Koodathingal, P., . . . Piccirilli, J. A. (2013). RNA catalyses nuclear pre-mRNA splicing. *Nature, 503*(7475), 229-+. doi:10.1038/nature12734

Forsburg, S. L. (1999). The best yeast? *Trends Genet, 15*(9), 340-344. doi:10.1016/s0168-9525(99)01798-9

Forsburg, S. L., & Rhind, N. (2006). Basic methods for fission yeast. *Yeast, 23*(3), 173-183. doi:10.1002/yea.1347

Fourmann, J. B., Dybkov, O., Agafonov, D. E., Tauchert, M. J., Urlaub, H., Ficner, R., . . . Luhrmann, R. (2016). The target of the DEAH-box NTP triphosphatase Prp43 in Saccharomyces cerevisiae spliceosomes is the U2 snRNP-intron interaction. *Elife, 5*. doi:ARTN e15564

10.7554/eLife.15564

Freund, C., Dotsch, V., Nishizawa, K., Reinherz, E. L., & Wagner, G. (1999). The GYF domain is a novel structural fold that is involved in lymphoid signaling through proline-rich sequences. *Nature Structural Biology, 6*(7), 656-660. doi:Doi 10.1038/10712

Froehler, B. C. (1986). Deoxynucleoside H-Phosphonate Diester Intermediates in the Synthesis of Internucleotide Phosphate Analogs. *Tetrahedron Letters, 27*(46), 5575-5578. doi:Doi 10.1016/S0040-4039(00)85269-7

Fromme, J. C., Banerjee, A., Huang, S. J., & Verdine, G. L. (2004). Structural basis for removal of adenine mispaired with 8-oxoguanine by MutY adenine DNA glycosylase. *Nature, 427*(6975), 652-656. doi:10.1038/nature02306

Galej, W. P., Wilkinson, M. E., Fica, S. M., Oubridge, C., Newman, A. J., & Nagai, K. (2016). Cryo-EM structure of the spliceosome immediately after branching. *Nature, 537*(7619), 197-201. doi:10.1038/nature19316

Gama-Carvalho, M., Krauss, R. D., Chiang, L., Valcarcel, J., Green, M. R., & Carmo-Fonseca, M. (1997). Targeting of U2AF65 to sites of active splicing in the nucleus. *J Cell Biol, 137*(5), 975-987. doi:10.1083/jcb.137.5.975

Gao, K., Masuda, A., Matsuura, T., & Ohno, K. (2008). Human branch point consensus sequence is yUnAy. *Nucleic Acids Res, 36*(7), 2257-2267. doi:10.1093/nar/gkn073

Garrey, S. M., Voelker, R., & Berglund, J. A. (2006). An extended RNA binding site for the yeast branch point-binding protein and the role of its zinc knuckle domains in RNA binding. *J Biol Chem, 281*(37), 27443-27453. doi:10.1074/jbc.M603137200

Gauto, D. F., Estrozi, L. F., Schwieters, C. D., Effantin, G., Macek, P., Sounier, R., . . . Boisbouvier, J. (2019). Integrated NMR and cryo-EM atomic-resolution structure determination of a half-megadalton enzyme complex. *Nat Commun, 10*(1), 2697. doi:10.1038/s41467-019-10490-9

Ghosh, G., & Adams, J. A. (2011). Phosphorylation mechanism and structure of serine-arginine protein kinases. *Febs Journal, 278*(4), 587-597. doi:10.1111/j.1742-4658.2010.07992.x

Gilbert, H. F. (1995). Thiol/disulfide exchange equilibria and disulfide bond stability. *Methods Enzymol, 251*, 8-28. doi:10.1016/0076-6879(95)51107-5

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., . . . Rothberg, J. M. (2003). A protein interaction map of Drosophila melanogaster. *Science, 302*(5651), 1727-1736. doi:10.1126/science.1090289

Giovannone, B., Lee, E., Laviola, L., Giorgino, F., Cleveland, K. A., & Smith, R. J. (2003). Two novel proteins that are linked to insulin-like growth factor (IGF-I) receptors by the Grb10 adapter and modulate IGF-I signaling. *Journal of Biological Chemistry, 278*(34), 31564-31573. doi:10.1074/jbc.M211572200

Glasel, J. A. (1995). Validity of nucleic acid purities monitored by 260nm/280nm absorbance ratios. *Biotechniques, 18*(1), 62-63. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7702855

Glasser, E., Agrawal, A. A., Jenkins, J. L., & Kielkopf, C. L. (2017). Cancer-Associated Mutations Mapped on High-Resolution Structures of the U2AF2 RNA Recognition Motifs. *Biochemistry, 56*(36), 4757-4761. doi:10.1021/acs.biochem.7b00551

Golas, M. M., Sander, B., Will, C. L., Luhrmann, R., & Stark, H. (2003). Molecular architecture of the multiprotein splicing factor SF3b. *Science, 300*(5621), 980-984. doi:10.1126/science.1084155

Goldfarb, A. R., Saidel, L. J., & Mosovich, E. (1951). The ultraviolet absorption spectra of proteins. *J Biol Chem, 193*(1), 397-404. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/14907727

Golling, G., Amsterdam, A., Sun, Z. X., Antonelli, M., Maldonado, E., Chen, W. B., . . . Hopkins, N. (2002). Insertional mutagenesis in zebrafish rapidly identifies genes essential for early vertebrate development. *Nature Genetics, 31*(2), 135-140. doi:10.1038/ng896

Gornemann, J., Barrandon, C., Hujer, K., Rutz, B., Rigaut, G., Kotovic, K. M., . . . Seraphin, B. (2011). Cotranscriptional spliceosome assembly and splicing are independent of the Prp40p WW domain. *RNA, 17*(12), 2119-2129. doi:10.1261/rna.02646811

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*(2), 233-243. doi:10.1101/gad.10.2.233

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*(8), 4752-4760. doi:10.1128/mcb.18.8.4752

Grabowski, P. J., & Black, D. L. (2001). Alternative RNA splicing in the nervous system. *Prog Neurobiol, 65*(3), 289-308. doi:10.1016/s0301-0082(01)00007-7

Grabowski, P. J., Seiler, S. R., & Sharp, P. A. (1985). A multicomponent complex is involved in the splicing of messenger RNA precursors. *Cell, 42*(1), 345-353. doi:10.1016/s0092-8674(85)80130-6

Grange, M., Vasishtan, D., & Grunewald, K. (2017). Cellular electron cryo tomography and in situ sub-volume averaging reveal the context of microtubule-based processes. *J Struct Biol, 197*(2), 181-190. doi:10.1016/j.jsb.2016.06.024

Graveley, B. R. (2000). Sorting out the complexity of SR protein functions. *RNA, 6*(9), 1197-1211. doi:10.1017/s1355838200000960

Green, M. R. (1991). Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. *Annu Rev Cell Biol, 7*, 559-599. doi:10.1146/annurev.cb.07.110191.003015

Grewal, C., Hickmott, J., Rentas, S., & Karagiannis, J. (2012). A conserved histone deacetylase with a role in the regulation of cytokinesis in Schizosaccharomyces pombe. *Cell Div, 7*(1), 13. doi:10.1186/1747-1028-7-13

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*(48), 38059-38066. doi:10.1074/jbc.M001483200

Habara, Y., Takeshima, Y., Awano, H., Okizuka, Y., Zhang, Z., Saiki, K., . . . Matsuo, M. (2009). In vitro splicing analysis showed that availability of a cryptic splice site is not a determinant for alternative splicing patterns caused by +1G-->A mutations in introns of the dystrophin gene. *J Med Genet, 46*(8), 542-547. doi:10.1136/jmg.2008.061259

Hamelberg, D., Shen, T., & McCammon, J. A. (2007). A proposed signaling motif for nuclear import in mRNA processing via the formation of arginine claw. *Proc Natl Acad Sci U S A, 104*(38), 14947-14951. doi:10.1073/pnas.0703151104

Han, T., Goralski, M., Gaskill, N., Capota, E., Kim, J., Ting, T. C., . . . Nijhawan, D. (2017). Anticancer sulfonamides target splicing by inducing RBM39 degradation via recruitment to DCAF15. *Science, 356*(6336). doi:10.1126/science.aal3755

Handa, N., Nureki, O., Kurimoto, K., Kim, I., Sakamoto, H., Shimura, Y., . . . Yokoyama, S. (1999). Structural basis for recognition of the tra mRNA precursor by the sex-lethal protein. *Nature, 398*(6728), 579-585. Retrieved from <Go to ISI>://WOS:000079754700046

Haraguchi, N., Andoh, T., Frendewey, D., & Tani, T. (2007). Mutations in the SF1-U2AF59-U2AF23 complex cause exon skipping in Schizosaccharomyces pombe. *J Biol Chem, 282*(4), 2221-2228. doi:10.1074/jbc.M609430200

Hardin, J. W., Warnasooriya, C., Kondo, Y., Nagai, K., & Rueda, D. (2015). Assembly and dynamics of the U4/U6 di-snRNP by single-molecule FRET. *Nucleic Acids Res, 43*(22), 10963-10974. doi:10.1093/nar/gkv1011

Harris, M. A., Rutherford, K. M., Hayles, J., Lock, A., Bahler, J., Oliver, S. G., . . . Wood, V. (2022). Fission stories: using PomBase to understand Schizosaccharomyces pombe biology. *Genetics, 220*(4). doi:10.1093/genetics/iyab222

Hartmann, B., & Valcarcel, J. (2009). Decrypting the genome's alternative messages. *Curr Opin Cell Biol, 21*(3), 377-386. doi:10.1016/j.ceb.2009.02.006

Hartmuth, K., Urlaub, H., Vornlocher, H. P., Will, C. L., Gentzel, M., Wilm, M., & Luhrmann, R. (2002). Protein composition of human prespliceosomes isolated by a tobramycin affinity-selection method. *Proc Natl Acad Sci U S A, 99*(26), 16719-16724. doi:10.1073/pnas.262483899

Hartwell, L. H., & Weinert, T. A. (1989). Checkpoints - Controls That Ensure the Order of Cell-Cycle Events. *Science, 246*(4930), 629-634. doi:DOI 10.1126/science.2683079

Haselbach, D., Komarov, I., Agafonov, D. E., Hartmuth, K., Graf, B., Dybkov, O., . . . Stark, H. (2018). Structure and Conformational Dynamics of the Human Spliceosomal B(act) Complex. *Cell, 172*(3), 454-464 e411. doi:10.1016/j.cell.2018.01.010

Hatada, I., Kitagawa, K., Yamaoka, T., Wang, X., Arai, Y., Hashido, K., . . . Mukai, T. (1995). Allele-specific methylation and expression of an imprinted U2af1-rs1 (SP2) gene. *Nucleic Acids Res, 23*(1), 36-41. doi:10.1093/nar/23.1.36

Hatada, I., Sugama, T., & Mukai, T. (1993). A new imprinted gene cloned by a methylation-sensitive genome scanning method. *Nucleic Acids Res, 21*(24), 5577-5582. doi:10.1093/nar/21.24.5577

Havens, M. A., & Hastings, M. L. (2016). Splice-switching antisense oligonucleotides as therapeutic drugs. *Nucleic Acids Res, 44*(14), 6549-6563. doi:10.1093/nar/gkw533

Hayashizaki, Y., Shibata, H., Hirotsune, S., Sugino, H., Okazaki, Y., Sasaki, N., . . . Chapman, V. M. (1994). Identification of an Imprinted U2af Binding-Protein Related Sequence on Mouse Chromosome-11 Using the Rlgs Method. *Nature Genetics, 6*(1), 33-40. doi:DOI 10.1038/ng0194-33

He, C. A., & Verdine, G. L. (2002). Trapping distinct structural states of a protein/DNA interaction through disulfide crosslinking. *Chemistry & Biology, 9*(12), 1297-1303. doi:Pii S1074-5521(02)00283-1

Doi 10.1016/S1074-5521(02)00283-1

Heckman, D. S., Geiser, D. M., Eidell, B. R., Stauffer, R. L., Kardos, N. L., & Hedges, S. B. (2001). Molecular evidence for the early colonization of land by fungi and plants. *Science, 293*(5532), 1129-1133. doi:10.1126/science.1061457

Helston, R. M., Box, J. A., Tang, W., & Baumann, P. (2010). Schizosaccharomyces cryophilus sp. nov., a new species of fission yeast. *FEMS Yeast Res, 10*(6), 779-786. doi:10.1111/j.1567-1364.2010.00657.x

Herdt, O., Reich, S., Medenbach, J., Timmermann, B., Olofsson, D., Preussner, M., & Heyd, F. (2020). The zinc finger domains in U2AF26 and U2AF35 have diverse functionalities including a role in controlling translation. *Rna Biology, 17*(6), 843-856. doi:10.1080/15476286.2020.1732701

Heyd, F., Carmo-Fonseca, M., & Moroy, T. (2008). Differential isoform expression and interaction with the P32 regulatory protein controls the subcellular localization of the splicing factor U2AF26. *J Biol Chem, 283*(28), 19636-19645. doi:10.1074/jbc.M801014200

Heyd, F., & Lynch, K. W. (2010). Phosphorylation-dependent regulation of PSF by GSK3 controls CD45 alternative splicing. *Mol Cell, 40*(1), 126-137. doi:10.1016/j.molcel.2010.09.013

Heyd, F., ten Dam, G., & Moroy, T. (2006). Auxiliary splice factor U2AF26 and transcription factor Gfi1 cooperate directly in regulating CD45 alternative splicing. *Nat Immunol, 7*(8), 859-867. doi:10.1038/ni1361

Higashio, H., Sato, K., & Nakano, A. (2008). Smy2p participates in COPII vesicle formation through the interaction with Sec23p/Sec24p subcomplex. *Traffic, 9*(1), 79-93. doi:10.1111/j.1600-0854.2007.00668.x

Hilleren, P., & Parker, R. (1999). Mechanisms of mRNA surveillance in eukaryotes. *Annu Rev Genet, 33*, 229-260. doi:10.1146/annurev.genet.33.1.229

Hillier, L. W., Coulson, A., Murray, J. I., Bao, Z. R., Sulston, J. E., & Waterston, R. H. (2005). Genomics in C-elegans: So many genes, such a little worm. *Genome Research, 15*(12), 1651-1660. doi:10.1101/gr.3729105

Hillman, R. T., Green, R. E., & Brenner, S. E. (2004). An unappreciated role for RNA surveillance. *Genome Biology, 5*(2), R8. doi:10.1186/gb-2004-5-2-r8

Hinterberger, M., Pettersson, I., & Steitz, J. A. (1983). Isolation of small nuclear ribonucleoproteins containing U1, U2, U4, U5, and U6 RNAs. *J Biol Chem, 258*(4), 2604-2613. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/6185498

Hirano, A., Yumimoto, K., Tsunematsu, R., Matsumoto, M., Oyama, M., Kozuka-Hata, H., . . . Fukada, Y. (2013). FBXL21 Regulates Oscillation of the Circadian Clock through Ubiquitination and Stabilization of Cryptochromes. *Cell, 152*(5), 1106-1118. doi:10.1016/j.cell.2013.01.054

Hoffman, C. S. (2001). Preparation of yeast DNA. *Curr Protoc Mol Biol, Chapter 13*, Unit13 11. doi:10.1002/0471142727.mb1311s39

Hoffman, C. S., Wood, V., & Fantes, P. A. (2015). An Ancient Yeast for Young Geneticists: A Primer on the Schizosaccharomyces pombe Model System. *Genetics, 201*(2), 403-423. doi:10.1534/genetics.115.181503

Hogg, R., McGrail, J. C., & O'Keefe, R. T. (2010). The function of the NineTeen Complex (NTC) in regulating spliceosome conformations and fidelity during pre-mRNA splicing. *Biochem Soc Trans, 38*(4), 1110-1115. doi:10.1042/BST0381110

Holm-Hansen, O. (1969). Algae: amounts of DNA and organic carbon in single cells. *Science, 163*(3862), 87-88. doi:10.1126/science.163.3862.87

Horowitz, D. S., & Krainer, A. R. (1997). A human protein required for the second step of pre-mRNA splicing is functionally related to a yeast splicing factor. *Genes & Development, 11*(1), 139-151. doi:DOI 10.1101/gad.11.1.139

Hoskins, A. A., Friedman, L. J., Gallagher, S. S., Crawford, D. J., Anderson, E. G., Wombacher, R., . . . Moore, M. J. (2011). Ordered and dynamic assembly of single spliceosomes. *Science, 331*(6022), 1289-1295. doi:10.1126/science.1198830

Huang, H., Chopra, R., Verdine, G. L., & Harrison, S. C. (1998). Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science, 282*(5394), 1669-1675. doi:10.1126/science.282.5394.1669

Huang, H. F., Harrison, S. C., & Verdine, G. L. (2000). Trapping of a catalytic HIV reverse transcriptase template:primer complex through a disulfide bond. *Chemistry & Biology, 7*(5), 355-364. doi:Doi 10.1016/S1074-5521(00)00113-7

Huang, J. R., Warner, L. R., Sanchez, C., Gabel, F., Madl, T., Mackereth, C. D., . . . Blackledge, M. (2014). Transient electrostatic interactions dominate the conformational equilibrium sampled by multidomain splicing factor U2AF65: a combined NMR and SAXS study. *Journal of the American Chemical Society, 136*(19), 7068-7076. doi:10.1021/ja502030n

Huang, T., Vilardell, J., & Query, C. C. (2002). Pre-spliceosome formation in S.pombe requires a stable complex of SF1-U2AF(59)-U2AF(23). *EMBO J, 21*(20), 5516-5526. doi:10.1093/emboj/cdf555

Hudson, B. P., Martinez-Yamout, M. A., Dyson, H. J., & Wright, P. E. (2004). Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nat Struct Mol Biol, 11*(3), 257-264. doi:10.1038/nsmb738

Hughes, A. L., & Friedman, R. (2003). Parallel evolution by gene duplication in the genomes of two unicellular fungi. *Genome Res, 13*(6A), 1259-1264. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12901373

Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., & O'Shea, E. K. (2003). Global analysis of protein localization in budding yeast. *Nature, 425*(6959), 686-691. doi:10.1038/nature02026

Hura, G. L., Menon, A. L., Hammel, M., Rambo, R. P., Poole, F. L., 2nd, Tsutakawa, S. E., . . . Tainer, J. A. (2009). Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat Methods, 6*(8), 606-612. doi:10.1038/nmeth.1353

Ilagan, J. O., Chalkley, R. J., Burlingame, A. L., & Jurica, M. S. (2013). Rearrangements within human spliceosomes captured after exon ligation. *RNA, 19*(3), 400-412. doi:10.1261/rna.034223.112

Ilagan, J. O., Ramakrishnan, A., Hayes, B., Murphy, M. E., Zebari, A. S., Bradley, P., & Bradley, R. K. (2015). U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Res, 25*(1), 14-26. doi:10.1101/gr.181016.114

Imai, H., Chan, E. K., Kiyosawa, K., Fu, X. D., & Tan, E. M. (1993). Novel nuclear autoantigen with splicing factor motifs identified with antibody from hepatocellular carcinoma. *J Clin Invest, 92*(5), 2419-2426. doi:10.1172/JCI116848

Imielinski, M., Berger, A. H., Hammerman, P. S., Hernandez, B., Pugh, T. J., Hodis, E., . . . Meyerson, M. (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell, 150*(6), 1107-1120. doi:10.1016/j.cell.2012.08.029

Irimia, M., & Blencowe, B. J. (2012). Alternative splicing: decoding an expansive regulatory layer. *Curr Opin Cell Biol, 24*(3), 323-332. doi:10.1016/j.ceb.2012.03.005

Ito, T., Muto, Y., Green, M. R., & 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*(16), 4523-4534. doi:10.1093/emboj/18.16.4523

Jacewicz, A., Chico, L., Smith, P., Schwer, B., & Shuman, S. (2015). Structural basis for recognition of intron branchpoint RNA by yeast Msl5 and selective effects of interfacial mutations on splicing of yeast pre-mRNAs. *RNA, 21*(3), 401-414. doi:10.1261/rna.048942.114

Jacquier, A. (1990). Self-splicing group II and nuclear pre-mRNA introns: how similar are they? *Trends Biochem Sci, 15*(9), 351-354. doi:10.1016/0968-0004(90)90075-m

Jamison, S. F., Crow, A., & Garcia-Blanco, M. A. (1992). The spliceosome assembly pathway in mammalian extracts. *Mol Cell Biol, 12*(10), 4279-4287. doi:10.1128/mcb.12.10.4279

Jamison, S. F., & Garcia-Blanco, M. A. (1992). An ATP-independent U2 small nuclear ribonucleoprotein particle/precursor mRNA complex requires both splice sites and the polypyrimidine tract. *Proc Natl Acad Sci U S A, 89*(12), 5482-5486. doi:10.1073/pnas.89.12.5482

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., . . . Sharpless, N. E. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA, 19*(2), 141-157. doi:10.1261/rna.035667.112

Jenkins, J. L., Agrawal, A. A., Gupta, A., Green, M. R., & Kielkopf, C. L. (2013). U2AF65 adapts to diverse pre-mRNA splice sites through conformational selection of specific and promiscuous RNA recognition motifs. *Nucleic Acids Res, 41*(6), 3859-3873. doi:10.1093/nar/gkt046

Jiang, J., Sheng, J., Carrasco, N., & Huang, Z. (2007). Selenium derivatization of nucleic acids for crystallography. *Nucleic Acids Res, 35*(2), 477-485. doi:10.1093/nar/gkl1070

Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). CHARMM-GUI: a web-based graphical user interface for CHARMM. *J Comput Chem, 29*(11), 1859-1865. doi:10.1002/jcc.20945

Johnson, A. A., Santos, W., Pais, G. C., Marchand, C., Amin, R., Burke, T. R., Jr., . . . Pommier, Y. (2006). Integration requires a specific interaction of the donor DNA terminal 5'-cytosine with glutamine 148 of the HIV-1 integrase flexible loop. *J Biol Chem, 281*(1), 461-467. doi:10.1074/jbc.M511348200

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583-589. doi:10.1038/s41586-021-03819-2

Juneau, K., Nislow, C., & Davis, R. W. (2009). Alternative splicing of PTC7 in Saccharomyces cerevisiae determines protein localization. *Genetics, 183*(1), 185-194. doi:10.1534/genetics.109.105155

Jung, D. J., Na, S. Y., Na, D. S., & Lee, J. W. (2002). Molecular cloning and characterization of CAPER, a novel coactivator of activating protein-1 and estrogen receptors. *J Biol Chem, 277*(2), 1229-1234. doi:10.1074/jbc.M110417200

Jurica, M. S., Licklider, L. J., Gygi, S. R., Grigorieff, N., & Moore, M. J. (2002). Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis. *RNA, 8*(4), 426-439. doi:10.1017/s1355838202021088

Jurica, M. S., & Moore, M. J. (2003). Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell, 12*(1), 5-14. doi:10.1016/s1097-2765(03)00270-3

Kang, H. S., Sanchez-Rico, C., Ebersberger, S., Sutandy, F. X. R., Busch, A., Welte, T., . . . Sattler, M. (2020). An autoinhibitory intramolecular interaction proof-reads RNA recognition by the essential splicing factor U2AF2. *Proc Natl Acad Sci U S A, 117*(13), 7140-7149. doi:10.1073/pnas.1913483117

Kapust, R. B., & Waugh, D. S. (1999). Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Science, 8*(8), 1668-1674. doi:DOI 10.1110/ps.8.8.1668

Kastner, B., Will, C. L., Stark, H., & Luhrmann, R. (2019). Structural Insights into Nuclear pre-mRNA Splicing in Higher Eukaryotes. *Cold Spring Harbor Perspectives in Biology, 11*(11). doi:ARTN a032417

10.1101/cshperspect.a032417

Kattah, N. H., Kattah, M. G., & Utz, P. J. (2010). The U1-snRNP complex: structural properties relating to autoimmune pathogenesis in rheumatic diseases. *Immunol Rev, 233*(1), 126-145. doi:10.1111/j.0105-2896.2009.00863.x

Kaufer, N. F., & Potashkin, J. (2000). Analysis of the splicing machinery in fission yeast: a comparison with budding yeast and mammals. *Nucleic Acids Res, 28*(16), 3003-3010. doi:10.1093/nar/28.16.3003

Kawashima, T., Douglass, S., Gabunilas, J., Pellegrini, M., & Chanfreau, G. F. (2014). Widespread use of non-productive alternative splice sites in Saccharomyces cerevisiae. *PLoS Genet, 10*(4), e1004249. doi:10.1371/journal.pgen.1004249

Kellenberger, E., Stier, G., & Sattler, M. (2002). Induced folding of the U2AF35 RRM upon binding to U2AF65. *FEBS Lett, 528*(1-3), 171-176. doi:10.1016/s0014-5793(02)03294-5

Kenan, D. J., Query, C. C., & Keene, J. D. (1991). RNA recognition: towards identifying determinants of specificity. *Trends Biochem Sci, 16*(6), 214-220. doi:10.1016/0968-0004(91)90088-d

Kent, O. A., Reayi, A., Foong, L., Chilibeck, K. A., & MacMillan, A. M. (2003). Structuring of the 3' splice site by U2AF65. *J Biol Chem, 278*(50), 50572-50577. doi:10.1074/jbc.M307976200

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research, 12*(6), 996-1006. doi:10.1101/gr.229102

Kielkopf, C. L., Rodionova, N. A., Green, M. R., & Burley, S. K. (2001). A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. *Cell, 106*(5), 595-605. doi:10.1016/s0092-8674(01)00480-9

Kim, S., Park, C., Jun, Y., Lee, S., Jung, Y., & Kim, J. (2018). Integrative Profiling of Alternative Splicing Induced by U2AF1 S34F Mutation in Lung Adenocarcinoma Reveals a Mechanistic Link to Mitotic Stress. *Molecules and Cells, 41*(8), 733-741. doi:10.14348/molcells.2018.0176

Kofler, M., Heuer, K., Zech, T., & Freund, C. (2004). Recognition sequences for the GYF domain reveal a possible spliceosomal function of CD2BP2. *J Biol Chem, 279*(27), 28292-28297. doi:10.1074/jbc.M402008200

Kofler, M., Motzny, K., Beyermann, M., & Freund, C. (2005). Novel interaction partners of the CD2BP2-GYF domain. *J Biol Chem, 280*(39), 33397-33402. doi:10.1074/jbc.M503989200

Kofler, M. M., & Freund, C. (2006). The GYF domain. *FEBS J, 273*(2), 245-256. doi:10.1111/j.1742-4658.2005.05078.x

Komazin-Meredith, G., Santos, W. L., Filman, D. J., Hogle, J. M., Verdine, G. L., & Coen, D. M. (2008). The positively charged surface of herpes simplex virus UL42 mediates DNA binding. *J Biol Chem, 283*(10), 6154-6161. doi:10.1074/jbc.M708691200

Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J., & Svergun, D. I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *Journal of Applied Crystallography, 36*, 1277-1282. doi:10.1107/S0021889803012779

Konarska, M. M., & Sharp, P. A. (1986). Electrophoretic separation of complexes involved in the splicing of precursors to mRNAs. *Cell, 46*(6), 845-855. doi:10.1016/0092-8674(86)90066-8

Konarska, M. M., & Sharp, P. A. (1987). Interactions between small nuclear ribonucleoprotein particles in formation of spliceosomes. *Cell, 49*(6), 763-774. doi:10.1016/0092-8674(87)90614-3

Kondo, Y., Oubridge, C., van Roon, A. M., & Nagai, K. (2015). Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *Elife, 4*. doi:10.7554/eLife.04986

Koning, R. I., & Koster, A. J. (2009). Cryo-electron tomography in biology and medicine. *Ann Anat, 191*(5), 427-445. doi:10.1016/j.aanat.2009.04.003

Koning, R. I., Koster, A. J., & Sharp, T. H. (2018). Advances in cryo-electron tomography for biology and medicine. *Ann Anat, 217*, 82-96. doi:10.1016/j.aanat.2018.02.004

Koodathingal, P., Novak, T., Piccirilli, J. A., & Staley, J. P. (2010). The DEAH box ATPases Prp16 and Prp43 cooperate to proofread 5' splice site cleavage during pre-mRNA splicing. *Mol Cell, 39*(3), 385-395. doi:10.1016/j.molcel.2010.07.014

Korneta, I., & Bujnicki, J. M. (2012). Intrinsic disorder in the human spliceosomal proteome. *PLoS Comput Biol, 8*(8), e1002641. doi:10.1371/journal.pcbi.1002641

Kowalska, E., Ripperger, J. A., Hoegger, D. C., Bruegger, P., Buch, T., Birchler, T., . . . Brown, S. A. (2013). NONO couples the circadian clock to the cell cycle. *Proc Natl Acad Sci U S A, 110*(5), 1592-1599. doi:10.1073/pnas.1213317110

Krainer, A. R., Mayeda, A., Kozak, D., & Binns, G. (1991). Functional expression of cloned human splicing factor SF2: homology to RNA-binding proteins, U1 70K, and Drosophila splicing regulators. *Cell, 66*(2), 383-394. doi:10.1016/0092-8674(91)90627-b

Kramer, A. (1992). Purification of Splicing Factor Sf1, a Heat-Stable Protein That Functions in the Assembly of a Presplicing Complex. *Molecular and Cellular Biology, 12*(10), 4545-4552. Retrieved from <Go to ISI>://WOS:A1992JP79800034

Kramer, A. (1995). The Biochemistry of Pre-mRNA Splicing. In A. I. Lamond (Ed.), *Pre-mRNA Processing* (pp. 35-64). Austin, TX: Landes Bioscience.

Kramer, 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*(7), 1355-1368. doi:10.1083/jcb.145.7.1355

Kramer, 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*(1), 29-37. doi:10.1016/s0378-1119(98)00058-4

Kuhlbrandt, W. (2014). Biochemistry. The resolution revolution. *Science, 343*(6178), 1443-1444. doi:10.1126/science.1251652

Kuhn, A. N., & Kaufer, N. F. (2003). Pre-mRNA splicing in Schizosaccharomyces pombe: regulatory role of a kinase conserved from fission yeast to mammals. *Curr Genet, 42*(5), 241-251. doi:10.1007/s00294-002-0355-2

Kupfer, D. M., Drabenstot, S. D., Buchanan, K. L., Lai, H., Zhu, H., Dyer, D. W., . . . Murphy, J. W. (2004). Introns and splicing elements of five diverse fungi. *Eukaryot Cell, 3*(5), 1088-1100. doi:10.1128/EC.3.5.1088-1100.2004

Kuwasako, K., Dohmae, N., Inoue, M., Shirouzu, M., Taguchi, S., Guntert, P., . . . Yokoyama, S. (2008). Complex assembly mechanism and an RNA-binding mode of the human p14-SF3b155 spliceosomal protein complex identified by NMR solution structure and functional analyses. *Proteins, 71*(4), 1617-1636. doi:10.1002/prot.21839

Laggerbauer, B., Liu, S., Makarov, E., Vornlocher, H. P., Makarova, O., Ingelfinger, D., . . . Luhrmann, R. (2005). The human U5 snRNP 52K protein (CD2BP2) interacts with U5-102K (hPrp6), a U4/U6.U5 tri-snRNP bridging protein, but dissociates upon tri-snRNP formation. *RNA, 11*(5), 598-608. doi:10.1261/rna.2300805

Lainez, B., Fernandez-Real, J. M., Romero, X., Esplugues, E., Canete, J. D., Ricart, W., & Engel, P. (2004). Identification and characterization of a novel spliced variant that encodes human soluble tumor necrosis factor receptor 2. *International Immunology, 16*(1), 169-177. doi:10.1093/intimm/dxh014

Lardelli, R. M., Thompson, J. X., Yates, J. R., 3rd, & Stevens, S. W. (2010). Release of SF3 from the intron branchpoint activates the first step of pre-mRNA splicing. *RNA, 16*(3), 516-528. doi:10.1261/rna.2030510

Lareau, L. F., Green, R. E., Bhatnagar, R. S., & Brenner, S. E. (2004). The evolving roles of alternative splicing. *Curr Opin Struct Biol, 14*(3), 273-282. doi:10.1016/j.sbi.2004.05.002

Le Guiner, C., Gesnel, M. C., & Breathnach, R. (2003). TIA-1 or TUR is required for DT40 cell viability. *Journal of Biological Chemistry, 278*(12), 10465-10476. doi:10.1074/jbc.M212378200

Lee, C. G., Zamore, P. D., Green, M. R., & Hurwitz, J. (1993). RNA annealing activity is intrinsically associated with U2AF. *J Biol Chem, 268*(18), 13472-13478. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7685763

Lee, J., Cheng, X., Swails, J. M., Yeom, M. S., Eastman, P. K., Lemkul, J. A., . . . Im, W. (2016). CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J Chem Theory Comput, 12*(1), 405-413. doi:10.1021/acs.jctc.5b00935

Lee, S., Radom, C. T., & Verdine, G. L. (2008). Trapping and structural elucidation of a very advanced intermediate in the lesion-extrusion pathway of hOGG1. *Journal of the American Chemical Society, 130*(25), 7784-+. doi:10.1021/ja800821t

Legrain, P., Seraphin, B., & Rosbash, M. (1988). Early commitment of yeast pre-mRNA to the spliceosome pathway. *Mol Cell Biol, 8*(9), 3755-3760. doi:10.1128/mcb.8.9.3755

Lehalle, D., Wieczorek, D., Zechi-Ceide, R. M., Passos-Bueno, M. R., Lyonnet, S., Amiel, J., & Gordon, C. T. (2015). A review of craniofacial disorders caused by spliceosomal defects. *Clin Genet, 88*(5), 405-415. doi:10.1111/cge.12596

Lesley, S. A., & Wilson, I. A. (2005). Protein production and crystallization at the joint center for structural genomics. *J Struct Funct Genomics, 6*(2-3), 71-79. doi:10.1007/s10969-005-2897-2

Lewis, B. P., Green, R. E., & Brenner, S. E. (2003). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A, 100*(1), 189-192. doi:10.1073/pnas.0136770100

Lewis, H. A., Chen, H., Edo, C., Buckanovich, R. J., Yang, Y. Y., Musunuru, K., . . . Burley, S. K. (1999). Crystal structures of Nova-1 and Nova-2 K-homology RNA-binding domains. *Structure, 7*(2), 191-203. doi:10.1016/S0969-2126(99)80025-2

Lewis, H. A., Musunuru, K., Jensen, K. B., Edo, C., Chen, H., Darnell, R. B., & Burley, S. K. (2000). Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. *Cell, 100*(3), 323-332. doi:10.1016/s0092-8674(00)80668-6

Li, X., Liu, S., Jiang, J., Zhang, L., Espinosa, S., Hill, R. C., . . . Zhao, R. (2017). CryoEM structure of Saccharomyces cerevisiae U1 snRNP offers insight into alternative splicing. *Nat Commun, 8*(1), 1035. doi:10.1038/s41467-017-01241-9

Li, X., Liu, S., Zhang, L., Issaian, A., Hill, R. C., Espinosa, S., . . . Zhao, R. (2019). A unified mechanism for intron and exon definition and back-splicing. *Nature, 573*(7774), 375-380. doi:10.1038/s41586-019-1523-6

Li, X., Zhang, W., Xu, T., Ramsey, J., Zhang, L., Hill, R., . . . Zhao, R. (2013). Comprehensive in vivo RNA-binding site analyses reveal a role of Prp8 in spliceosomal assembly. *Nucleic Acids Res, 41*(6), 3805-3818. doi:10.1093/nar/gkt062

Liang, D., & Wilusz, J. E. (2014). Short intronic repeat sequences facilitate circular RNA production. *Genes Dev, 28*(20), 2233-2247. doi:10.1101/gad.251926.114

Liang, D. M., Tatomer, D. C., Luo, Z., Wu, H., Yang, L., Chen, L. L., . . . Wilusz, J. E. (2017). The Output of Protein-Coding Genes Shifts to Circular RNAs When the Pre-mRNA Processing Machinery Is Limiting. *Molecular Cell, 68*(5), 940-+. doi:10.1016/j.molcel.2017.10.034

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkoczi, G., Chen, V. B., Croll, T. I., . . . Adams, P. D. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallographica Section D-Structural Biology, 75*, 861-877. doi:10.1107/S2059798319011471

Lienig, J., & Thulasiraman, K. (1996). GASBOR: A genetic algorithm approach for solving the switchbox routing problem. *Journal of Circuits Systems and Computers, 6*(4), 359-373. doi:Doi 10.1142/S0218126696000248

Lillie, S. H., & Brown, S. S. (1992). Suppression of a myosin defect by a kinesin-related gene. *Nature, 356*(6367), 358-361. doi:10.1038/356358a0

Lin, Q., Taylor, S. J., & Shalloway, D. (1997). Specificity and determinants of Sam68 RNA binding. Implications for the biological function of K homology domains. *J Biol Chem, 272*(43), 27274-27280. doi:10.1074/jbc.272.43.27274

Lin, Y., & Kielkopf, C. L. (2008). X-ray structures of U2 snRNA-branchpoint duplexes containing conserved pseudouridines. *Biochemistry, 47*(20), 5503-5514. doi:10.1021/bi7022392

Lindner, P. (1893). Schizosaccharomyces pombe n. sp., ein neuer Gährungserreger. *Wochenschrift für Brauerei, 10*, 1298-1300.

Liu, J., He, L., Collins, I., Ge, H., Libutti, D., Li, J., . . . Levens, D. (2000). The FBP interacting repressor targets TFIIH to inhibit activated transcription. *Mol Cell, 5*(2), 331-341. doi:10.1016/s1097-2765(00)80428-1

Liu, S., Li, X., Zhang, L., Jiang, J., Hill, R. C., Cui, Y., . . . Zhao, R. (2017). Structure of the yeast spliceosomal postcatalytic P complex. *Science, 358*(6368), 1278-1283. doi:10.1126/science.aar3462

Liu, Z., Luyten, I., Bottomley, M. J., Messias, A. C., Houngninou-Molango, S., Sprangers, R., . . . Sattler, M. (2001). Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science, 294*(5544), 1098-1102. doi:10.1126/science.1064719

Lock, A., Rutherford, K., Harris, M. A., Hayles, J., Oliver, S. G., Bahler, J., & Wood, V. (2019). PomBase 2018: user-driven reimplementation of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. *Nucleic Acids Res, 47*(D1), D821-D827. doi:10.1093/nar/gky961

Loerch, S., Maucuer, A., Manceau, V., Green, M. R., & Kielkopf, C. L. (2014). Cancer-relevant splicing factor CAPERalpha engages the essential splicing factor SF3b155 in a specific ternary complex. *J Biol Chem, 289*(25), 17325-17337. doi:10.1074/jbc.M114.558825

Long, J. C., & Caceres, J. F. (2009). The SR protein family of splicing factors: master regulators of gene expression. *Biochemical Journal, 417*, 15-27. doi:10.1042/Bj20081501

Lorenzi, L. E., Bah, A., Wischnewski, H., Shchepachev, V., Soneson, C., Santagostino, M., & Azzalin, C. M. (2015). Fission yeast Cactin restricts telomere transcription and elongation by controlling Rap1 levels. *Embo Journal, 34*(1), 115-129. doi:10.15252/embj.201489559

Lutzelberger, M., Backstrom, E., & Akusjarvi, G. (2005). Substrate-dependent differences in U2AF requirement for splicing in adenovirus-infected cell extracts. *J Biol Chem, 280*(27), 25478-25484. doi:10.1074/jbc.M413737200

Lykke-Andersen, S., & Jensen, T. H. (2015). Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Biol, 16*(11), 665-677. doi:10.1038/nrm4063

Mackereth, C. D., Madl, T., Bonnal, S., Simon, B., Zanier, K., Gasch, A., . . . Sattler, M. (2011). Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. *Nature, 475*(7356), 408-411. doi:10.1038/nature10171

MacMillan, A. M., Query, C. C., Allerson, C. R., Chen, S., Verdine, G. L., & Sharp, P. A. (1994). Dynamic association of proteins with the pre-mRNA branch region. *Genes Dev, 8*(24), 3008-3020. doi:10.1101/gad.8.24.3008

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., . . . Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research, 47*(W1), W636-W641. doi:10.1093/nar/gkz268

Madhani, H. D., & Guthrie, C. (1994). Dynamic RNA-RNA interactions in the spliceosome. *Annu Rev Genet, 28*, 1-26. doi:10.1146/annurev.ge.28.120194.000245

Maji, D., Glasser, E., Henderson, S., Galardi, J., Pulvino, M. J., Jenkins, J. L., & Kielkopf, C. L. (2020). Representative cancer-associated U2AF2 mutations alter RNA interactions and splicing. *J Biol Chem, 295*(50), 17148-17157. doi:10.1074/jbc.RA120.015339

Manalastas-Cantos, K., Konarev, P. V., Hajizadeh, N. R., Kikhney, A. G., Petoukhov, M. V., Molodenskiy, D. S., . . . Franke, D. (2021). ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *Journal of Applied Crystallography, 54*(Pt 1), 343-355. doi:10.1107/S1600576720013412

Manceau, V., Kielkopf, C. L., Sobel, A., & Maucuer, A. (2008). Different requirements of the kinase and UHM domains of KIS for its nuclear localization and binding to splicing factors. *J Mol Biol, 381*(3), 748-762. doi:10.1016/j.jmb.2008.06.026

Manceau, V., Swenson, M., Le Caer, J. P., Sobel, A., Kielkopf, C. L., & Maucuer, A. (2006). Major phosphorylation of SF1 on adjacent Ser-Pro motifs enhances interaction with U2AF65. *FEBS J, 273*(3), 577-587. doi:10.1111/j.1742-4658.2005.05091.x

Maniatis, T., & Tasic, B. (2002). Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature, 418*(6894), 236-243. doi:10.1038/418236a

Maquat, L. E. (2004). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol, 5*(2), 89-99. doi:10.1038/nrm1310

Marshall, A. N., Montealegre, M. C., Jimenez-Lopez, C., Lorenz, M. C., & van Hoof, A. (2013). Alternative splicing and subfunctionalization generates functional diversity in fungal proteomes. *PLoS Genet, 9*(3), e1003376. doi:10.1371/journal.pgen.1003376

Martin, A., Schneider, S., & Schwer, B. (2002). Prp43 is an essential RNA-dependent ATPase required for release of lariat-intron from the spliceosome. *J Biol Chem, 277*(20), 17743-17750. doi:10.1074/jbc.M200762200

Martinez, N. M., Pan, Q., Cole, B. S., Yarosh, C. A., Babcock, G. A., Heyd, F., . . . Lynch, K. W. (2012). Alternative splicing networks regulated by signaling in human T cells. *RNA, 18*(5), 1029-1040. doi:10.1261/rna.032243.112

Matlin, A. J., Clark, F., & Smith, C. W. (2005). Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol, 6*(5), 386-398. doi:10.1038/nrm1645

Maucuer, A., Le Caer, J. P., Manceau, V., & Sobel, A. (2000). Specific Ser-Pro phosphorylation by the RNA-recognition motif containing kinase KIS. *European Journal of Biochemistry, 267*(14), 4456-4464. doi:DOI 10.1046/j.1432-1327.2000.01493.x

Maucuer, A., Ozon, S., Manceau, V., Gavet, O., Lawler, S., Curmi, P., & Sobel, A. (1997). KIS is a protein kinase with an RNA recognition motif. *Journal of Biological Chemistry, 272*(37), 23151-23156. doi:DOI 10.1074/jbc.272.37.23151

Mayas, R. M., Maita, H., & Staley, J. P. (2006). Exon ligation is proofread by the DExD/H-box ATPase Prp22p. *Nat Struct Mol Biol, 13*(6), 482-490. doi:10.1038/nsmb1093

Mazroui, R., Puoti, A., & Kramer, A. (1999). Splicing factor SF1 from Drosophila and Caenorhabditis: presence of an N-terminal RS domain and requirement for viability. *RNA, 5*(12), 1615-1631. doi:10.1017/s1355838299991872

Mccoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., & Read, R. J. (2007). Phaser crystallographic software. *Journal of Applied Crystallography, 40*, 658-674. doi:10.1107/S0021889807021206

McGlincy, N. J., Valomon, A., Chesham, J. E., Maywood, E. S., Hastings, M. H., & Ule, J. (2012). Regulation of alternative splicing by the circadian clock and food related cues. *Genome Biology, 13*(6), R54. doi:10.1186/gb-2012-13-6-r54

McIntosh, B. K., Renfro, D. P., Knapp, G. S., Lairikyengbam, C. R., Liles, N. M., Niu, L., . . . Hu, J. C. (2012). EcoliWiki: a wiki-based community resource for Escherichia coli. *Nucleic Acids Res, 40*(Database issue), D1270-1277. doi:10.1093/nar/gkr880

McKeown, M. (1992). Alternative mRNA splicing. *Annu Rev Cell Biol, 8*, 133-155. doi:10.1146/annurev.cb.08.110192.001025

McKie, A. B., McHale, J. C., Keen, T. J., Tarttelin, E. E., Goliath, R., van Lith-Verhoeven, J. J., . . . Inglehearn, C. F. (2001). Mutations in the pre-mRNA splicing factor gene PRPC8 in autosomal dominant retinitis pigmentosa (RP13). *Hum Mol Genet, 10*(15), 1555-1562. doi:10.1093/hmg/10.15.1555

McMullan, G., Faruqi, A. R., Clare, D., & Henderson, R. (2014). Comparison of optimal performance at 300 keV of three direct electron detectors for use in low dose electron microscopy. *Ultramicroscopy, 147*, 156-163. doi:10.1016/j.ultramic.2014.08.002

Merendino, L., Guth, S., Bilbao, D., Martinez, C., & Valcarcel, J. (1999). Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3' splice site AG. *Nature, 402*(6763), 838-841. doi:10.1038/45602

Merk, A., Bartesaghi, A., Banerjee, S., Falconieri, V., Rao, P., Davis, M. I., . . . Subramaniam, S. (2016). Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. *Cell, 165*(7), 1698-1707. doi:10.1016/j.cell.2016.05.040

Meyer, N. H., Tripsianes, K., Vincendeau, M., Madl, T., Kateb, F., Brack-Werner, R., & Sattler, M. (2010). Structural basis for homodimerization of the Src-associated during mitosis, 68-kDa protein (Sam68) Qua1 domain. *J Biol Chem, 285*(37), 28893-28901. doi:10.1074/jbc.M110.126185

Michaud, S., & Reed, R. (1991). An ATP-independent complex commits pre-mRNA to the mammalian spliceosome assembly pathway. *Genes Dev, 5*(12B), 2534-2546. doi:10.1101/gad.5.12b.2534

Michel, H. (1991). *Crystallization of Membrane Proteins* (1st ed.). Boca Raton, FL, USA: CRC Press.

Michel, J., Langstein, J., Hofstadter, F., & Schwarz, H. (1998). A soluble form of CD137 (ILA/4-1BB), a member of the TNF receptor family, is released by activated lymphocytes and is detectable in sera of patients with rheumatoid arthritis. *European Journal of Immunology, 28*(1), 290-295. doi:Doi 10.1002/(Sici)1521-4141(199801)28:01<290::Aid-Immu290>3.3.Co;2-J

Mogridge, J. (2015). Using light scattering to determine the stoichiometry of protein complexes. *Methods Mol Biol, 1278*, 233-238. doi:10.1007/978-1-4939-2425-7_14

Mokry, M., Feitsma, H., Nijman, I. J., de Bruijn, E., van der Zaag, P. J., Guryev, V., & Cuppen, E. (2010). Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. *Nucleic Acids Res, 38*(10), e116. doi:10.1093/nar/gkq072

Mollet, I., Barbosa-Morais, N. L., Andrade, J., & Carmo-Fonseca, M. (2006). Diversity of human U2AF splicing factors. *FEBS J, 273*(21), 4807-4816. doi:10.1111/j.1742-4658.2006.05502.x

Moore, M. J. (2000). Intron recognition comes of AGe. *Nature Structural Biology, 7*(1), 14-16. doi:10.1038/71207

Moore, M. J., Query, C. C., & Sharp, P. A. (1993). Splicing of precursors to messenger RNAs by the spliceosome. In R. F. Gesteland & J. F. Atkins (Eds.), *The RNA World, 1st Ed.* (pp. 303-357). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Morrison, M., Harris, K. S., & Roth, M. B. (1997). smg mutants affect the expression of alternatively spliced SR protein mRNAs in Caenorhabditis elegans. *Proc Natl Acad Sci U S A, 94*(18), 9782-9785. doi:10.1073/pnas.94.18.9782

Mullen, M. P., Smith, C. W. J., Patton, J. G., & Nadalginard, B. (1991). Alpha-Tropomyosin Mutually Exclusive Exon Selection - Competition between Branchpoint Polypyrimidine Tracts Determines Default Exon Choice. *Genes & Development, 5*(4), 642-655. doi:DOI 10.1101/gad.5.4.642

Musco, G., Kharrat, A., Stier, G., Fraternali, F., Gibson, T. J., Nilges, M., & Pastore, A. (1997). The solution structure of the first KH domain of FMR1, the protein responsible for the fragile X syndrome. *Nature Structural Biology, 4*(9), 712-716. doi:10.1038/nsb0997-712

Musco, G., Stier, G., Joseph, C., Castiglione Morelli, M. A., Nilges, M., Gibson, T. J., & Pastore, A. (1996). Three-dimensional structure and stability of the KH domain: molecular insights into the fragile X syndrome. *Cell, 85*(2), 237-245. doi:10.1016/s0092-8674(00)81100-9

Muto, Y., & Yokoyama, S. (2012). Structural insight into RNA recognition motifs: versatile molecular Lego building blocks for biological systems. *Wiley Interdiscip Rev RNA, 3*(2), 229-246. doi:10.1002/wrna.1107

Myung, J. K., & Sadar, M. D. (2012). Large scale phosphoproteome analysis of LNCaP human prostate cancer cells. *Mol Biosyst, 8*(8), 2174-2182. doi:10.1039/c2mb25151e

Nabetani, A., Hatada, I., Morisaki, H., Oshimura, M., & Mukai, T. (1997). Mouse U2af1-rs1 is a neomorphic imprinted gene. *Molecular and Cellular Biology, 17*(2), 789-798. doi:Doi 10.1128/Mcb.17.2.789

Nadalginard, B., Smith, C. W. J., Patton, J. G., & Breitbart, R. E. (1991). Alternative Splicing Is an Efficient Mechanism for the Generation of Protein Diversity - Contractile Protein Genes as a Model System. *Advances in Enzyme Regulation, 31*, 261-286. doi:Doi 10.1016/0065-2571(91)90017-G

Nancollis, V., Ruckshanthi, J. P., Frazer, L. N., & O'Keefe, R. T. (2013). The U5 snRNA internal loop 1 is a platform for Brr2, Snu114 and Prp8 protein binding during U5 snRNP assembly. *J Cell Biochem, 114*(12), 2770-2784. doi:10.1002/jcb.24625

Newby, M. I., & Greenbaum, N. L. (2002). Sculpting of the spliceosomal branch site recognition motif by a conserved pseudouridine. *Nature Structural Biology, 9*(12), 958-965. doi:10.1038/nsb873

Nguyen, T. H., Galej, W. P., Bai, X. C., Savva, C. G., Newman, A. J., Scheres, S. H., & Nagai, K. (2015). The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature, 523*(7558), 47-52. doi:10.1038/nature14548

Nguyen, T. H. D., Galej, W. P., Bai, X. C., Oubridge, C., Newman, A. J., Scheres, S. H. W., & Nagai, K. (2016). Cryo-EM structure of the yeast U4/U6.U5 tri-snRNP at 3.7 A resolution. *Nature, 530*(7590), 298-302. doi:10.1038/nature16940

Nielsen, H., & Johansen, S. D. (2009). Group I introns: Moving in new directions. *RNA Biol, 6*(4), 375-383. doi:10.4161/rna.6.4.9334

Nielsen, T. K., Liu, S., Luhrmann, R., & Ficner, R. (2007). Structural basis for the bifunctionality of the U5 snRNP 52K protein (CD2BP2). *J Mol Biol, 369*(4), 902-908. doi:10.1016/j.jmb.2007.03.077

Nilges, M. (1995). Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J Mol Biol, 245*(5), 645-660. doi:10.1006/jmbi.1994.0053

Nilsen, T. W. (1994). RNA-RNA interactions in the spliceosome: unraveling the ties that bind. *Cell, 78*(1), 1-4. doi:10.1016/0092-8674(94)90563-0

Nilsen, T. W., & Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature, 463*(7280), 457-463. doi:10.1038/nature08909

Nishizawa, K., Freund, C., Li, J., Wagner, G., & Reinherz, E. L. (1998). Identification of a proline-binding motif regulating CD2-triggered T lymphocyte activation. *Proc Natl Acad Sci U S A, 95*(25), 14897-14902. doi:10.1073/pnas.95.25.14897

Noble, S. M., & Guthrie, C. (1996). Identification of novel genes required for yeast pre-mRNA splicing by means of cold-sensitive mutations. *Genetics, 143*(1), 67-80. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8722763

Norton, P. A. (1994). Polypyrimidine Tract Sequences Direct Selection of Alternative Branch Sites and Influence Protein-Binding. *Nucleic Acids Research, 22*(19), 3854-3860. doi:DOI 10.1093/nar/22.19.3854

O'Shea, E. K., Rutkowski, R., Stafford, W. F., 3rd, & Kim, P. S. (1989). Preferential heterodimer formation by isolated leucine zippers from fos and jun. *Science, 245*(4918), 646-648. doi:10.1126/science.2503872

Ohrt, T., Odenwalder, P., Dannenberg, J., Prior, M., Warkocki, Z., Schmitzova, J., . . . Luhrmann, R. (2013). Molecular dissection of step 2 catalysis of yeast pre-mRNA splicing investigated in a purified system. *RNA, 19*(7), 902-915. doi:10.1261/rna.039024.113

Okazaki, K., & Niwa, O. (2000). mRNAs encoding zinc finger protein isoforms are expressed by alternative splicing of an in-frame intron in fission yeast. *DNA Res, 7*(1), 27-30. doi:10.1093/dnares/7.1.27

Okeyo-Owuor, T., White, B. S., Chatrikhi, R., Mohan, D. R., Kim, S., Griffith, M., . . . Graubert, T. A. (2015). U2AF1 mutations alter sequence specificity of pre-mRNA binding and splicing. *Leukemia, 29*(4), 909-917. doi:10.1038/leu.2014.303

Otwinowski, Z., & Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol, 276*, 307-326. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/27754618

Oubridge, C., Ito, N., Evans, P. R., Teo, C. H., & Nagai, K. (1994). Crystal structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature, 372*(6505), 432-438. doi:10.1038/372432a0

Pacheco, T. R., Gomes, A. Q., Barbosa-Morais, N. L., Benes, V., Ansorge, W., Wollerton, M., . . . Carmo-Fonseca, M. (2004). Diversity of vertebrate splicing factor U2AF35: identification of alternatively spliced U2AF1 mRNAS. *J Biol Chem, 279*(26), 27039-27049. doi:10.1074/jbc.M402136200

Padgett, R. A., Konarska, M. M., Grabowski, P. J., Hardy, S. F., & Sharp, P. A. (1984). Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. *Science, 225*(4665), 898-903. doi:10.1126/science.6206566

Page-McCaw, P. S., Amonlirdviman, K., & Sharp, P. A. (1999). PUF60: a novel U2AF65-related splicing activity. *RNA, 5*(12), 1548-1560. doi:10.1017/s1355838299991938

Palangat, M., Anastasakis, D. G., Fei, D. L., Lindblad, K. E., Bradley, R., Hourigan, C. S., . . . Larson, D. R. (2019). The splicing factor U2AF1 contributes to cancer progression through a noncanonical role in translation regulation. *Genes & Development, 33*(9-10), 482-497. doi:10.1101/gad.319590.118

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet, 40*(12), 1413-1415. doi:10.1038/ng.259

Pandit, S., Lynn, B., & Rymond, B. C. (2006). Inhibition of a spliceosome turnover pathway suppresses splicing defects. *Proc Natl Acad Sci U S A, 103*(37), 13700-13705. doi:10.1073/pnas.0603188103

Parker, R., Siliciano, P. G., & Guthrie, C. (1987). Recognition of the TACTAAC box during mRNA splicing in yeast involves base pairing to the U2-like snRNA. *Cell, 49*(2), 229-239. doi:10.1016/0092-8674(87)90564-2

Parks, T. D., Leuther, K. K., Howard, E. D., Johnston, S. A., & Dougherty, W. G. (1994). Release of Proteins and Peptides from Fusion Proteins Using a Recombinant Plant-Virus Proteinase. *Analytical Biochemistry, 216*(2), 413-417. doi:DOI 10.1006/abio.1994.1060

Pastuszak, A. W., Joachimiak, M. P., Blanchette, M., Rio, D. C., Brenner, S. E., & Frankel, A. D. (2011). An SF1 affinity model to identify branch point sequences in human introns. *Nucleic Acids Res, 39*(6), 2344-2356. doi:10.1093/nar/gkq1046

Pearsall, R. S., Shibata, H., Brozowska, A., Yoshino, K., Okuda, K., deJong, P. J., . . . Held, W. A. (1996). Absence of imprinting in U2AFBPL, a human homologue of the imprinted mouse gene U2afbp-rs. *Biochem Biophys Res Commun, 222*(1), 171-177. doi:10.1006/bbrc.1996.0716

Pelikan, M., Hura, G. L., & Hammel, M. (2009). Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen Physiol Biophys, 28*(2), 174-189. doi:10.4149/gpb_2009_02_174

Perriman, R., & Ares, M., Jr. (2010). Invariant U2 snRNA nucleotides form a stem loop to recognize the intron early in splicing. *Mol Cell, 38*(3), 416-427. doi:10.1016/j.molcel.2010.02.036

Petersen-Mahrt, S. K., Estmer, C., Ohrmalm, C., Matthews, D. A., Russell, W. C., & Akusjarvi, G. (1999). The splicing factor-associated protein, p32, regulates RNA splicing by inhibiting ASF/SF2 RNA binding and phosphorylation. *EMBO J, 18*(4), 1014-1024. doi:10.1093/emboj/18.4.1014

Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., . . . Svergun, D. I. (2012). New developments in the ATSAS program package for small-angle scattering data analysis. *Journal of Applied Crystallography, 45*(Pt 2), 342-350. doi:10.1107/S0021889812007662

Pfeffer, S., Brandt, F., Hrabe, T., Lang, S., Eibauer, M., Zimmermann, R., & Forster, F. (2012). Structure and 3D arrangement of endoplasmic reticulum membrane-associated ribosomes. *Structure, 20*(9), 1508-1518. doi:10.1016/j.str.2012.06.010

Plaschka, C., Lin, P. C., Charenton, C., & Nagai, K. (2018). Prespliceosome structure provides insights into spliceosome assembly and regulation. *Nature, 559*(7714), 419-422. doi:10.1038/s41586-018-0323-8

Plaschka, C., Lin, P. C., & Nagai, K. (2017). Structure of a pre-catalytic spliceosome. *Nature, 546*(7660), 617-621. doi:10.1038/nature22799

Pleiss, J. A., Whitworth, G. B., Bergkessel, M., & Guthrie, C. (2007). Transcript specificity in yeast pre-mRNA splicing revealed by mutations in core spliceosomal components. *PLoS Biol, 5*(4), e90. doi:10.1371/journal.pbio.0050090

Pomeranz Krummel, D. A., Oubridge, C., Leung, A. K., Li, J., & Nagai, K. (2009). Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution. *Nature, 458*(7237), 475-480. doi:10.1038/nature07851

Portmann, S., Grimm, S., Workman, C., Usman, N., & Egli, M. (1996). Crystal structures of an A-form duplex with single-adenosine bulges and a conformational basis for site-specific RNA self-cleavage. *Chem Biol, 3*(3), 173-184. doi:10.1016/s1074-5521(96)90260-4

Preussner, M., Wilhelmi, I., Schultz, A. S., Finkernagel, F., Michel, M., Moroy, T., & Heyd, F. (2014). Rhythmic U2af26 alternative splicing controls PERIOD1 stability and the circadian clock in mice. *Mol Cell, 54*(4), 651-662. doi:10.1016/j.molcel.2014.04.015

Price, P. L., Morderer, D., & Rossoll, W. (2018). RNP Assembly Defects in Spinal Muscular Atrophy. *Adv Neurobiol, 20*, 143-171. doi:10.1007/978-3-319-89689-2_6

Putnam, C. D., Hammel, M., Hura, G. L., & Tainer, J. A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q Rev Biophys, 40*(3), 191-285. doi:10.1017/S0033583507004635

Pyle, A. M. (2016). Group II Intron Self-Splicing. *Annu Rev Biophys, 45*, 183-205. doi:10.1146/annurev-biophys-062215-011149

Query, C. C., Bentley, R. C., & Keene, J. D. (1989). A common RNA recognition motif identified within a defined U1 RNA binding domain of the 70K U1 snRNP protein. *Cell, 57*(1), 89-101. doi:10.1016/0092-8674(89)90175-x

Query, C. C., & Konarska, M. M. (2004). Suppression of multiple substrate mutations by spliceosomal prp8 alleles suggests functional correlations with ribosomal ambiguity mutants. *Mol Cell, 14*(3), 343-354. doi:10.1016/s1097-2765(04)00217-5

Query, C. C., Moore, M. J., & Sharp, P. A. (1994). Branch nucleophile selection in pre-mRNA splicing: evidence for the bulged duplex model. *Genes Dev, 8*(5), 587-597. doi:10.1101/gad.8.5.587

Query, C. C., Strobel, S. A., & Sharp, P. A. (1996). Three recognition events at the branch-site adenine. *EMBO J, 15*(6), 1392-1402. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8635472

Rain, J. C., Rafi, Z., Rhani, Z., Legrain, P., & Kramer, 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-a Publication of the Rna Society, 4*(5), 551-565. doi:Doi 10.1017/S1355838298980335

Rambo, R. P., & Tainer, J. A. (2011). Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers, 95*(8), 559-571. doi:10.1002/bip.21638

Rappsilber, J., Ryder, U., Lamond, A. I., & Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. *Genome Res, 12*(8), 1231-1245. doi:10.1101/gr.473902

Rauhut, R., Fabrizio, P., Dybkov, O., Hartmuth, K., Pena, V., Chari, A., . . . Luhrmann, R. (2016). Molecular architecture of the Saccharomyces cerevisiae activated spliceosome. *Science, 353*(6306), 1399-1405. doi:10.1126/science.aag1906

Reed, R. (1989). The organization of 3' splice-site sequences in mammalian introns. *Genes Dev, 3*(12B), 2113-2123. doi:10.1101/gad.3.12b.2113

Reed, R. (1990). Protein composition of mammalian spliceosomes assembled in vitro. *Proc Natl Acad Sci U S A, 87*(20), 8031-8035. doi:10.1073/pnas.87.20.8031

Reed, R., & Maniatis, T. (1985). Intron sequences involved in lariat formation during pre-mRNA splicing. *Cell, 41*(1), 95-105. doi:10.1016/0092-8674(85)90064-9

Rhind, N., Chen, Z., Yassour, M., Thompson, D. A., Haas, B. J., Habib, N., . . . Nusbaum, C. (2011). Comparative functional genomics of the fission yeasts. *Science, 332*(6032), 930-936. doi:10.1126/science.1203357

Richards, J., & Gumz, M. L. (2012). Advances in understanding the peripheral circadian clocks. *FASEB J, 26*(9), 3602-3613. doi:10.1096/fj.12-203554

Rio, D. C. (1993). Splicing of pre-mRNA: mechanism, regulation and role in development. *Current Opinion in Genetics & Development, 3*(4), 574-584. doi:10.1016/0959-437x(93)90093-5

Ritchie, D. B., Schellenberg, M. J., & MacMillan, A. M. (2009). Spliceosome structure: Piece by piece. *Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms, 1789*(9-10), 624-633. doi:10.1016/j.bbagrm.2009.08.010

Robberson, B. L., Cote, G. J., & Berget, S. M. (1990). Exon Definition May Facilitate Splice Site Selection in Rnas with Multiple Exons. *Molecular and Cellular Biology, 10*(1), 84-94. doi:Doi 10.1128/Mcb.10.1.84

Romfo, C. M., Alvarez, C. J., van Heeckeren, W. J., Webb, C. J., & Wise, J. A. (2000). Evidence for splice site pairing via intron definition in Schizosaccharomyces pombe. *Mol Cell Biol, 20*(21), 7955-7970. doi:10.1128/MCB.20.21.7955-7970.2000

Romfo, C. M., & Wise, J. A. (1997). Both the polypyrimidine tract and the 3' splice site function prior to the first step of splicing in fission yeast. *Nucleic Acids Research, 25*(22), 4658-4665. doi:DOI 10.1093/nar/25.22.4658

Roscigno, R. F., Weiner, M., & Garciablanco, M. A. (1993). A Mutational Analysis of the Polypyrimidine Tract of Introns - Effects of Sequence Differences in Pyrimidine Tracts on Splicing. *Journal of Biological Chemistry, 268*(15), 11222-11229. Retrieved from <Go to ISI>://WOS:A1993LD46600075

Rould, M. A., Perona, J. J., & Steitz, T. A. (1991). Structural basis of anticodon loop recognition by glutaminyl-tRNA synthetase. *Nature, 352*(6332), 213-218. doi:10.1038/352213a0

Ruby, S. W., & Abelson, J. (1988). An early hierarchic role of U1 small nuclear ribonucleoprotein in spliceosome assembly. *Science, 242*(4881), 1028-1035. doi:10.1126/science.2973660

Rudner, D. Z., Breger, K. S., Kanaar, R., Adams, M. D., & Rio, D. C. (1998). RNA binding activity of heterodimeric splicing factor U2AF: at least one RS domain is required for high-affinity binding. *Molecular and Cellular Biology, 18*(7), 4004-4011. doi:Doi 10.1128/Mcb.18.7.4004

Ruhl, C., Stauffer, E., Kahles, A., Wagner, G., Drechsel, G., Ratsch, G., & Wachter, A. (2012). Polypyrimidine tract binding protein homologs from Arabidopsis are key regulators of alternative splicing with implications in fundamental developmental processes. *Plant Cell, 24*(11), 4360-4375. doi:10.1105/tpc.112.103622

Ruskin, B., Krainer, A. R., Maniatis, T., & Green, M. R. (1984). Excision of an Intact Intron as a Novel Lariat Structure during Pre-Messenger Rna Splicing Invitro. *Cell, 38*(1), 317-331. doi:Doi 10.1016/0092-8674(84)90553-1

Ruskin, B., Zamore, P. D., & Green, M. R. (1988). A Factor, U2af, Is Required for U2 Snrnp Binding and Splicing Complex Assembly. *Cell, 52*(2), 207-219. doi:Doi 10.1016/0092-8674(88)90509-0

Rutz, B., & Seraphin, B. (1999). Transient interaction of BBP/ScSF1 and Mud2 with the splicing machinery affects the kinetics of spliceosome assembly. *RNA, 5*(6), 819-831. doi:10.1017/s1355838299982286

Rutz, B., & Seraphin, B. (2000). A dual role for BBP/ScSF1 in nuclear pre-mRNA retention and splicing. *EMBO J, 19*(8), 1873-1886. doi:10.1093/emboj/19.8.1873

Sasaki-Haraguchi, N., Ikuyama, T., Yoshii, S., Takeuchi-Andoh, T., Frendewey, D., & Tani, T. (2015). Cwf16p Associating with the Nineteen Complex Ensures Ordered Exon Joining in Constitutive Pre-mRNA Splicing in Fission Yeast. *PLoS One, 10*(8), e0136336. doi:10.1371/journal.pone.0136336

Sauter, C., Ng, J. D., Lorber, B., Keith, G., Brion, P., Hosseini, M. W., . . . Giege, R. (1999). Additives for the crystallization of proteins and nucleic acids. *Journal of Crystal Growth, 196*(2-4), 365-376. doi:Doi 10.1016/S0022-0248(98)00852-5

Scheckel, C., & Darnell, R. B. (2015). Microexons--tiny but mighty. *EMBO J, 34*(3), 273-274. doi:10.15252/embj.201490651

Schellenberg, M. J., Dul, E. L., & MacMillan, A. M. (2011). Structural model of the p14/SF3b155 . branch duplex complex. *RNA, 17*(1), 155-165. doi:10.1261/rna.2224411

Schellenberg, M. J., Edwards, R. A., Ritchie, D. B., Kent, O. A., Golas, M. M., Stark, H., . . . MacMillan, A. M. (2006). Crystal structure of a core spliceosomal protein interface. *Proc Natl Acad Sci U S A, 103*(5), 1266-1271. doi:10.1073/pnas.0508048103

Schellenberg, M. J., Ritchie, D. B., & MacMillan, A. M. (2008). Pre-mRNA splicing: a complex picture in higher definition. *Trends Biochem Sci, 33*(6), 243-246. doi:10.1016/j.tibs.2008.04.004

Schneider, M., Will, C. L., Anokhina, M., Tazi, J., Urlaub, H., & Luhrmann, R. (2010). Exon Definition Complexes Contain the Tri-snRNP and Can Be Directly Converted into B-like Precatalytic Splicing Complexes. *Molecular Cell, 38*(2), 223-235. doi:10.1016/j.molcel.2010.02.027

Schultz, A. S., Preussner, M., Bunse, M., Karni, R., & Heyd, F. (2017). Activation-Dependent TRAF3 Exon 8 Alternative Splicing Is Controlled by CELF2 and hnRNP C Binding to an Upstream Intronic Element. *Molecular and Cellular Biology, 37*(7). doi:ARTN e00488

10.1128/MCB.00488-16

Schwer, B. (2008). A conformational rearrangement in the spliceosome sets the stage for Prp22-dependent mRNA release. *Mol Cell, 30*(6), 743-754. doi:10.1016/j.molcel.2008.05.003

Schwer, B., & Gross, C. H. (1998). Prp22, a DExH-box RNA helicase, plays two distinct roles in yeast pre-mRNA splicing. *EMBO J, 17*(7), 2086-2094. doi:10.1093/emboj/17.7.2086

Schwer, B., & Guthrie, C. (1992). A conformational rearrangement in the spliceosome is dependent on PRP16 and ATP hydrolysis. *EMBO J, 11*(13), 5033-5039. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/1464325

Schwer, B., & Meszaros, T. (2000). RNA helicase dynamics in pre-mRNA splicing. *EMBO J, 19*(23), 6582-6591. doi:10.1093/emboj/19.23.6582

Schwieters, C. D., Suh, J. Y., Grishaev, A., Ghirlando, R., Takayama, Y., & Clore, G. M. (2010). Solution structure of the 128 kDa enzyme I dimer from Escherichia coli and its 146 kDa complex with HPr using residual dipolar couplings and small- and wide-angle X-ray scattering. *Journal of the American Chemical Society, 132*(37), 13026-13045. doi:10.1021/ja105485b

Screaton, G. R., Xu, X. N., Olsen, A. L., Cowper, A. E., Tan, R. S., McMichael, A. J., & Bell, J. I. (1997). LARD: A new lymphoid-specific death domain containing receptor regulated by alternative pre-mRNA splicing. *Proceedings of the National Academy of Sciences of the United States of America, 94*(9), 4615-4619. doi:DOI 10.1073/pnas.94.9.4615

Selenko, P., Gregorovic, G., Sprangers, R., Stier, G., Rhani, Z., Kramer, A., & Sattler, M. (2003). Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP. *Mol Cell, 11*(4), 965-976. doi:10.1016/s1097-2765(03)00115-1

Semlow, D. R., Blanco, M. R., Walter, N. G., & Staley, J. P. (2016). Spliceosomal DEAH-Box ATPases Remodel Pre-mRNA to Activate Alternative Splice Sites. *Cell, 164*(5), 985-998. doi:10.1016/j.cell.2016.01.025

Senapathy, P., Shapiro, M. B., & Harris, N. L. (1990). Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. *Methods Enzymol, 183*, 252-278. doi:10.1016/0076-6879(90)83018-5

Seraphin, B., Kretzner, L., & Rosbash, M. (1988). A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5' cleavage site. *EMBO J, 7*(8), 2533-2538. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/3056718

Seraphin, B., & Rosbash, M. (1989). Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. *Cell, 59*(2), 349-358. doi:10.1016/0092-8674(89)90296-1

Shah, K., Wu, H. Y., & Rana, T. M. (1994). Synthesis of Uridine Phosphoramidite Analogs - Reagents for Site-Specific Incorporation of Photoreactive Sites into Rna Sequences. *Bioconjugate Chemistry, 5*(6), 508-512. doi:DOI 10.1021/bc00030a005

Shao, C., Yang, B., Wu, T., Huang, J., Tang, P., Zhou, Y., . . . Fu, X. D. (2014). Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome. *Nat Struct Mol Biol, 21*(11), 997-1005. doi:10.1038/nsmb.2906

Sharma, S., Kohlstaedt, L. A., Damianov, A., Rio, D. C., & Black, D. L. (2008). Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nat Struct Mol Biol, 15*(2), 183-191. doi:10.1038/nsmb.1375

Sharp, P. A. (1991). "Five easy pieces". *Science, 254*(5032), 663. doi:10.1126/science.1948046

Sharp, T. H., Koster, A. J., & Gros, P. (2016). Heterogeneous MAC Initiator and Pore Structures in a Lipid Bilayer by Phase-Plate Cryo-electron Tomography. *Cell Rep, 15*(1), 1-8. doi:10.1016/j.celrep.2016.03.002

Shelley, C. S., & Baralle, F. E. (1987). Deletion Analysis of a Unique 3' Splice Site Indicates That Alternating Guanine and Thymine Residues Represent an Efficient Splicing Signal. *Nucleic Acids Research, 15*(9), 3787-3799. doi:DOI 10.1093/nar/15.9.3787

Shepard, J. (2004). *Electronic Dissertation: Characterization of U2AF26, a Paralog of the Splicing Factor U2AF35.* (Ph.D.). University of Texas Southwestern Medical Center, Dallas, Texas, USA.

Shepard, J., Reick, M., Olson, S., & Graveley, B. R. (2002). Characterization of U2AF(6), a splicing factor related to U2AF(35). *Mol Cell Biol, 22*(1), 221-230. doi:10.1128/mcb.22.1.221-230.2002

Sheth, N., Roca, X., Hastings, M. L., Roeder, T., Krainer, A. R., & Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Research, 34*(14), 3955-3967. doi:10.1093/nar/gkl556

Shu, H., Chen, S., Bi, Q., Mumby, M., & Brekken, D. L. (2004). Identification of phosphoproteins and their phosphorylation sites in the WEHI-231 B lymphoma cell line. *Mol Cell Proteomics, 3*(3), 279-286. doi:10.1074/mcp.D300003-MCP200

Sickmier, E. A., Frato, K. E., Shen, H., Paranawithana, S. R., Green, M. R., & Kielkopf, C. L. (2006). Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. *Mol Cell, 23*(1), 49-59. doi:10.1016/j.molcel.2006.05.025

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W. Z., . . . Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology, 7*. doi:ARTN 539

10.1038/msb.2011.75

Siliciano, P. G., & Guthrie, C. (1988). 5' splice site selection in yeast: genetic alterations in base-pairing with U1 reveal additional requirements. *Genes Dev, 2*(10), 1258-1267. doi:10.1101/gad.2.10.1258

Simon, B., Madl, T., Mackereth, C. D., Nilges, M., & Sattler, M. (2010). An efficient protocol for NMR-spectroscopy-based structure determination of protein complexes in solution. *Angew Chem Int Ed Engl, 49*(11), 1967-1970. doi:10.1002/anie.200906147

Sipiczki, M. (2000). Where does fission yeast sit on the tree of life? *Genome Biology, 1*(2), REVIEWS1011. doi:10.1186/gb-2000-1-2-reviews1011

Skrzypek, M. S., Binkley, J., Binkley, G., Miyasato, S. R., Simison, M., & Sherlock, G. (2017). The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Res, 45*(D1), D592-D596. doi:10.1093/nar/gkw924

Smith, C. W., & Valcarcel, J. (2000). Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem Sci, 25*(8), 381-388. doi:10.1016/s0968-0004(00)01604-2

Smith, D. J., Query, C. C., & Konarska, M. M. (2007). trans-splicing to spliceosomal U2 snRNA suggests disruption of branch site-U2 pairing during pre-mRNA splicing. *Mol Cell, 26*(6), 883-890. doi:10.1016/j.molcel.2007.05.020

Song, J., & Richard, S. (2015). Sam68 Regulates S6K1 Alternative Splicing during Adipogenesis. *Mol Cell Biol, 35*(11), 1926-1939. doi:10.1128/MCB.01488-14
Sousa, R. (1995). Use of glycerol, polyols and other protein structure stabilizing agents in protein crystallization. *Acta Crystallogr D Biol Crystallogr, 51*(Pt 3), 271-277. doi:10.1107/S0907444994014009
Spadaccini, R., Reidt, U., Dybkov, O., Will, C., Frank, R., Stier, G., . . . 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*(3), 410-425. doi:10.1261/rna.2271406
Spingola, M., Grate, L., Haussler, D., & Ares, M., Jr. (1999). Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. *RNA, 5*(2), 221-234. doi:10.1017/s1355838299981682
Staiger, D., & Apel, K. (1999). Circadian clock-regulated expression of an RNA-binding protein in Arabidopsis: characterisation of a minimal promoter element. *Mol Gen Genet, 261*(4-5), 811-819. doi:10.1007/s004380050025
Staiger, D., & Brown, J. W. (2013). Alternative splicing at the intersection of biological timing, development, and stress responses. *Plant Cell, 25*(10), 3640-3656. doi:10.1105/tpc.113.113803
Staiger, D., Zecca, L., Wieczorek Kirk, D. A., Apel, K., & Eckstein, L. (2003). The circadian clock regulated RNA-binding protein AtGRP7 autoregulates its expression by influencing alternative splicing of its own pre-mRNA. *Plant J, 33*(2), 361-371. doi:10.1046/j.1365-313x.2003.01629.x
Staknis, D., & Reed, R. (1994). SR proteins promote the first specific recognition of Pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex. *Mol Cell Biol, 14*(11), 7670-7682. doi:10.1128/mcb.14.11.7670
Staley, J. P., & Guthrie, C. (1998). Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell, 92*(3), 315-326. doi:10.1016/s0092-8674(00)80925-3
Staley, J. P., & Guthrie, C. (1999). An RNA switch at the 5' splice site requires ATP and the DEAD box protein Prp28p. *Mol Cell, 3*(1), 55-64. doi:10.1016/s1097-2765(00)80174-4
Stamm, S. (2008). Regulation of alternative splicing by reversible protein phosphorylation. *J Biol Chem, 283*(3), 1223-1227. doi:10.1074/jbc.R700034200
Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O., & Zhang, M. Q. (2000). An alternative-exon database and its statistical analysis. *DNA Cell Biol, 19*(12), 739-756. doi:10.1089/104454900750058107
Stanojevic, D., & Verdine, G. L. (1995). Deconstruction of GCN4/GCRE into a monomeric peptide-DNA complex. *Nature Structural Biology, 2*(6), 450-457. doi:10.1038/nsb0695-450
Starke, S., Jost, I., Rossbach, O., Schneider, T., Schreiner, S., Hung, L. H., & Bindereif, A. (2015). Exon circularization requires canonical splice signals. *Cell Rep, 10*(1), 103-111. doi:10.1016/j.celrep.2014.12.002
Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., . . . Lancet, D. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics, 54*, 1 30 31-31 30 33. doi:10.1002/cpbi.5
Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A., & Cooper, D. N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet, 133*(1), 1-9. doi:10.1007/s00439-013-1358-4
Stepankiw, N., Raghavan, M., Fogarty, E. A., Grimson, A., & Pleiss, J. A. (2015). Widespread alternative and aberrant splicing revealed by lariat sequencing. *Nucleic Acids Res, 43*(17), 8488-8501. doi:10.1093/nar/gkv763
Stepanyuk, G. A., Serrano, P., Peralta, E., Farr, C. L., Axelrod, H. L., Geralt, M., . . . Williamson, J. R. (2016). UHM-ULM interactions in the RBM39-U2AF65 splicing-factor complex. *Acta Crystallogr D Struct Biol, 72*(Pt 4), 497-511. doi:10.1107/S2059798316001248

Sternglanz, H., & Bugg, C. E. (1975). Relationship between Mutagenic and Base-Stacking Properties of Halogenated Uracil Derivatives - Crystal-Structures of 5-Chlorouracil and 5-Bromouracil. *Biochimica Et Biophysica Acta, 378*(1), 1-11. doi:Doi 10.1016/0005-2787(75)90130-6

Stoilov, P., Daoud, R., Nayler, O., & Stamm, S. (2004). Human tra2-beta1 autoregulates its protein concentration by influencing alternative splicing of its pre-mRNA. *Hum Mol Genet, 13*(5), 509-524. doi:10.1093/hmg/ddh051

Stols, L., Gu, M., Dieckman, L., Raffen, R., Collart, F. R., & Donnelly, M. I. (2002). A new vector for high-throughput, ligation-independent cloning encoding a tobacco etch virus protease cleavage site. *Protein Expr Purif, 25*(1), 8-15. doi:10.1006/prep.2001.1603

Strausberg, R. L., Feingold, E. A., Grouse, L. H., Derge, J. G., Klausner, R. D., Collins, F. S., . . . Pro, M. G. C. M. (2002). Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proceedings of the National Academy of Sciences of the United States of America, 99*(26), 16899-16903. doi:10.1073/pnas.242603899

Studier, F. W., Rosenberg, A. H., Dunn, J. J., & Dubendorff, J. W. (1990). Use of T7 Rna-Polymerase to Direct Expression of Cloned Genes. *Methods in Enzymology, 185*, 60-89. Retrieved from <Go to ISI>://WOS:A1990MC41900006

Sureau, A., Gattoni, R., Dooghe, Y., Stevenin, J., & Soret, J. (2001). SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs. *Embo Journal, 20*(7), 1785-1796. doi:DOI 10.1093/emboj/20.7.1785

Sutandy, F. X. R., Ebersberger, S., Huang, L., Busch, A., Bach, M., Kang, H. S., . . . Konig, J. (2018). In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. *Genome Res, 28*(5), 699-713. doi:10.1101/gr.229757.117

Svergun, D. I. (1992). Determination of the Regularization Parameter in Indirect-Transform Methods Using Perceptual Criteria. *Journal of Applied Crystallography, 25*, 495-503. doi:Doi 10.1107/S0021889892001663

Svergun, D. I. (1999). Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophysical Journal, 76*(6), 2879-2886. doi:Doi 10.1016/S0006-3495(99)77443-6

Svergun, D. I., Petoukhov, M. V., & Koch, M. H. (2001). Determination of domain structure of proteins from X-ray solution scattering. *Biophys J, 80*(6), 2946-2953. doi:10.1016/S0006-3495(01)76260-1

Swanson, M. S., Nakagawa, T. Y., LeVan, K., & Dreyfuss, G. (1987). Primary structure of human nuclear ribonucleoprotein particle C proteins: conservation of sequence and domain structures in heterogeneous nuclear RNA, mRNA, and pre-rRNA-binding proteins. *Mol Cell Biol, 7*(5), 1731-1739. doi:10.1128/mcb.7.5.1731

Taft, R. J., & Mattick, J. S. (2003). Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. *Genome Biology, 5*(1), P1. doi:10.1186/gb-2003-5-1-p1

Taliaferro, J. M., Alvarez, N., Green, R. E., Blanchette, M., & Rio, D. C. (2011). Evolution of a tissue-specific splicing network. *Genes Dev, 25*(6), 608-620. doi:10.1101/gad.2009011

Tanaka, N., Aronova, A., & Schwer, B. (2007). Ntr1 activates the Prp43 helicase to trigger release of lariat-intron from the spliceosome. *Genes Dev, 21*(18), 2312-2325. doi:10.1101/gad.1580507

Tari, M., Manceau, V., de Matha Salone, J., Kobayashi, A., Pastre, D., & Maucuer, A. (2019). U2AF(65) assemblies drive sequence-specific splice site recognition. *EMBO Rep, 20*(8), e47604. doi:10.15252/embr.201847604

Teplova, M., & Patel, D. J. (2008). Structural insights into RNA recognition by the alternative-splicing regulator muscleblind-like MBNL1. *Nat Struct Mol Biol, 15*(12), 1343-1351. doi:10.1038/nsmb.1519

Thakran, P., Pandit, P. A., Datta, S., Kolathur, K. K., Pleiss, J. A., & Mishra, S. K. (2018). Sde2 is an intron-specific pre-mRNA splicing regulator activated by ubiquitin-like processing. *Embo Journal, 37*(1), 89-101. doi:10.15252/embj.201796751

Thickman, K. R., Sickmier, E. A., & Kielkopf, C. L. (2007). Alternative conformations at the RNA-binding surface of the N-terminal U2AF(65) RNA recognition motif. *J Mol Biol, 366*(3), 703-710. doi:10.1016/j.jmb.2006.11.077

Thickman, K. R., Swenson, M. C., Kabogo, J. M., Gryczynski, Z., & 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*(3), 664-683. doi:10.1016/j.jmb.2005.11.067

Tholen, J., Razew, M., Weis, F., & Galej, W. P. (2022). Structural basis of branch site recognition by the human spliceosome. *Science, 375*(6576), 50-57. doi:10.1126/science.abm4245

Tisserant, A., & Konig, H. (2008). Signal-regulated Pre-mRNA occupancy by the general splicing factor U2AF. *PLoS One, 3*(1), e1418. doi:10.1371/journal.pone.0001418

Toda, T., Iida, A., Miwa, T., Nakamura, Y., & Imai, T. (1994). Isolation and characterization of a novel gene encoding nuclear protein at a locus (D11S636) tightly linked to multiple endocrine neoplasia type 1 (MEN1). *Hum Mol Genet, 3*(3), 465-470. doi:10.1093/hmg/3.3.465

Tone, M., Tone, Y., Fairchild, P. J., Wykes, M., & Waldmann, H. (2001). Regulation of CD40 function by its isoforms generated through alternative splicing. *Proceedings of the National Academy of Sciences of the United States of America, 98*(4), 1751-1756. doi:DOI 10.1073/pnas.98.4.1751

Townsend, C., Leelaram, M. N., Agafonov, D. E., Dybkov, O., Will, C. L., Bertram, K., . . . Luhrmann, R. (2020). Mechanism of protein-guided folding of the active site U2/U6 RNA during spliceosome activation. *Science, 370*(6523). doi:10.1126/science.abc3753

Trakhanov, S., & Quiocho, F. A. (1995). Influence of divalent cations in protein crystallization. *Protein Sci, 4*(9), 1914-1919. doi:10.1002/pro.5560040925

Tria, G., Mertens, H. D., Kachala, M., & Svergun, D. I. (2015). Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ, 2*(Pt 2), 207-217. doi:10.1107/S205225251500202X

Tronchere, H., Wang, J., & Fu, X. D. (1997). A protein related to splicing factor U2AF35 that interacts with U2AF65 and SR proteins in splicing of pre-mRNA. *Nature, 388*(6640), 397-400. doi:10.1038/41137

Tsai, R. T., Fu, R. H., Yeh, F. L., Tseng, C. K., Lin, Y. C., Huang, Y. H., & Cheng, S. C. (2005). Spliceosome disassembly catalyzed by Prp43 and its associated components Ntr1 and Ntr2. *Genes Dev, 19*(24), 2991-3003. doi:10.1101/gad.1377405

Tugarinov, V., Choy, W. Y., Orekhov, V. Y., & Kay, L. E. (2005). Solution NMR-derived global fold of a monomeric 82-kDa enzyme. *Proc Natl Acad Sci U S A, 102*(3), 622-627. doi:10.1073/pnas.0407792102

Tupler, R., Perini, G., & Green, M. R. (2001). Expressing the human genome. *Nature, 409*(6822), 832-833. doi:10.1038/35057011

Twyffels, L., Gueydan, C., & Kruys, V. (2011). Shuttling SR proteins: more than splicing factors. *Febs Journal, 278*(18), 3246-3255. doi:10.1111/j.1742-4658.2011.08274.x

Uehara, T., Minoshima, Y., Sagane, K., Sugi, N. H., Mitsuhashi, K. O., Yamamoto, N., . . . Owa, T. (2017). Selective degradation of splicing factor CAPERalpha by anticancer sulfonamides. *Nat Chem Biol, 13*(6), 675-680. doi:10.1038/nchembio.2363

Uliyanchenko, E. (2014). Size-exclusion chromatography-from high-performance to ultra-performance. *Anal Bioanal Chem, 406*(25), 6087-6094. doi:10.1007/s00216-014-8041-z

Umen, J. G., & Guthrie, C. (1995). The second catalytic step of pre-mRNA splicing. *RNA, 1*(9), 869-885. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8548652

Unverdorben, P., Beck, F., Sledz, P., Schweitzer, A., Pfeifer, G., Plitzko, J. M., . . . Forster, F. (2014). Deep classification of a large cryo-EM dataset defines the conformational landscape of the 26S proteasome. *Proc Natl Acad Sci U S A, 111*(15), 5544-5549. doi:10.1073/pnas.1403409111

Urbanski, L. M., Leclair, N., & Anczukow, O. (2018). Alternative-splicing defects in cancer: Splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics. *Wiley Interdiscip Rev RNA, 9*(4), e1476. doi:10.1002/wrna.1476

Valadkhan, S. (2010). Role of the snRNAs in spliceosomal active site. *RNA Biol, 7*(3), 345-353. doi:10.4161/rna.7.3.12089

Valadkhan, S., & Jaladat, Y. (2010). The spliceosomal proteome: at the heart of the largest cellular ribonucleoprotein machine. *Proteomics, 10*(22), 4128-4141. doi:10.1002/pmic.201000354

Valcarcel, J., Gaur, R. K., Singh, R., & Green, M. R. (1996). Interaction of U2AF65 RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA [corrected]. *Science, 273*(5282), 1706-1709. doi:10.1126/science.273.5282.1706

Van Buskirk, C., & Schupbach, T. (2002). Half pint regulates alternative splice site selection in Drosophila. *Dev Cell, 2*(3), 343-353. doi:10.1016/s1534-5807(02)00128-4

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., . . . Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research, 50*(D1), D439-D444. doi:10.1093/nar/gkab1061

Vernet, C., & Artzt, K. (1997). STAR, a gene family involved in signal transduction and activation of RNA. *Trends Genet, 13*(12), 479-484. doi:10.1016/s0168-9525(97)01269-9

Vijaykrishna, N., Melangath, G., Kumar, R., Khandelia, P., Bawa, P., Varadarajan, R., & Vijayraghavan, U. (2016). The Fission Yeast Pre-mRNA-processing Factor 18 (prp18+) Has Intron-specific Splicing Functions with Links to G1-S Cell Cycle Progression. *J Biol Chem, 291*(53), 27387-27402. doi:10.1074/jbc.M116.751289

Vilardell, J., Chartrand, P., Singer, R. H., & Warner, J. R. (2000). The odyssey of a regulated transcript. *RNA, 6*(12), 1773-1780. doi:10.1017/s135583820000145x

Vithana, E. N., Abu-Safieh, L., Allen, M. J., Carey, A., Papaioannou, M., Chakarova, C., . . . Bhattacharya, S. S. (2001). A human homolog of yeast pre-mRNA splicing gene, PRP31, underlies autosomal dominant retinitis pigmentosa on chromosome 19q13.4 (RP11). *Mol Cell, 8*(2), 375-381. doi:10.1016/s1097-2765(01)00305-7

Voith von Voithenberg, L., Sanchez-Rico, C., Kang, H. S., Madl, T., Zanier, K., Barth, A., . . . Lamb, D. C. (2016). Recognition of the 3' splice site RNA by the U2AF heterodimer involves a dynamic population shift. *Proc Natl Acad Sci U S A, 113*(46), E7169-E7175. doi:10.1073/pnas.1605873113

Volkov, V. V., & Svergun, D. I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. *Journal of Applied Crystallography, 36*, 860-864. doi:10.1107/S0021889803000268

Wagner, J. D., Jankowsky, E., Company, M., Pyle, A. M., & Abelson, J. N. (1998). The DEAH-box protein PRP22 is an ATPase that mediates ATP-dependent mRNA release from the spliceosome and unwinds RNA duplexes. *EMBO J, 17*(10), 2926-2937. doi:10.1093/emboj/17.10.2926

Wahl, M. C., Will, C. L., & Luhrmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. *Cell, 136*(4), 701-718. doi:10.1016/j.cell.2009.02.009

Wan, L., & Dreyfuss, G. (2017). Splicing-Correcting Therapy for SMA. *Cell, 170*(1), 5. doi:10.1016/j.cell.2017.06.028

Wan, R., Bai, R., Yan, C., Lei, J., & Shi, Y. (2019). Structures of the Catalytically Activated Yeast Spliceosome Reveal the Mechanism of Branching. *Cell, 177*(2), 339-351 e313. doi:10.1016/j.cell.2019.02.006

Wan, R., Yan, C., Bai, R., Huang, G., & Shi, Y. (2016). Structure of a yeast catalytic step I spliceosome at 3.4 A resolution. *Science, 353*(6302), 895-904. doi:10.1126/science.aag2235

Wan, R., Yan, C., Bai, R., Lei, J., & Shi, Y. (2017). Structure of an Intron Lariat Spliceosome from Saccharomyces cerevisiae. *Cell, 171*(1), 120-132 e112. doi:10.1016/j.cell.2017.08.029

Wan, R. X., Yan, C. Y., Bai, R., Wang, L., Huang, M., Wong, C. C. L., & Shi, Y. G. (2016). The 3.8 angstrom structure of the U4/U6.U5 tri-snRNP: Insights into spliceosome assembly and catalysis. *Science, 351*(6272), 466-475. doi:10.1126/science.aad6466

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., . . . Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature, 456*(7221), 470-476. doi:10.1038/nature07509

Wang, J. Z., Fu, X., Fang, Z., Liu, H., Zong, F. Y., Zhu, H., . . . Hui, J. (2021). QKI-5 regulates the alternative splicing of cytoskeletal gene ADD3 in lung cancer. *J Mol Cell Biol, 13*(5), 347-360. doi:10.1093/jmcb/mjaa063

Wang, L., Lawrence, M. S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K., . . . Wu, C. J. (2011). SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med, 365*(26), 2497-2506. doi:10.1056/NEJMoa1109016

Wang, P. L., Bao, Y., Yee, M. C., Barrett, S. P., Hogan, G. J., Olsen, M. N., . . . Salzman, J. (2014). Circular RNA Is Expressed across the Eukaryotic Tree of Life. *PLoS One, 9*(3). doi:ARTN e90859

10.1371/journal.pone.0090859

Wang, Q., Zhang, L., Lynn, B., & Rymond, B. C. (2008). A BBP-Mud2p heterodimer mediates branchpoint recognition and influences splicing substrate abundance in budding yeast. *Nucleic Acids Res, 36*(8), 2787-2798. doi:10.1093/nar/gkn144

Wang, W., Maucuer, A., Gupta, A., Manceau, V., Thickman, K. R., Bauer, W. J., . . . Kielkopf, C. L. (2013). Structure of phosphorylated SF1 bound to U2AF(6)(5) in an essential splicing factor complex. *Structure, 21*(2), 197-208. doi:10.1016/j.str.2012.10.020

Wang, X., & Tanaka Hall, T. M. (2001). Structural basis for recognition of AU-rich element RNA by the HuD protein. *Nature Structural Biology, 8*(2), 141-145. doi:10.1038/84131

Ward, A. J., & Cooper, T. A. (2010). The pathobiology of splicing. *J Pathol, 220*(2), 152-163. doi:10.1002/path.2649

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., . . . Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research, 46*(W1), W296-W303. doi:10.1093/nar/gky427

Webb, C. J., Romfo, C. M., van Heeckeren, W. J., & Wise, J. A. (2005). Exonic splicing enhancers in fission yeast: functional conservation demonstrates an early evolutionary origin. *Genes Dev, 19*(2), 242-254. doi:10.1101/gad.1265905

Webb, C. J., & Wise, J. A. (2004). The splicing factor U2AF small subunit is functionally conserved between fission yeast and humans. *Mol Cell Biol, 24*(10), 4229-4240. doi:10.1128/mcb.24.10.4229-4240.2004

Weber, G., Trowitzsch, S., Kastner, B., Luhrmann, R., & Wahl, M. C. (2010). Functional organization of the Sm core in the crystal structure of human U1 snRNP. *EMBO J, 29*(24), 4172-4184. doi:10.1038/emboj.2010.295

Wellmann, S., Taube, T., Paal, K., Graf, V. E. H., Geilen, W., Seifert, G., . . . Seeger, K. (2001). Specific reverse transcription-PCR quantification of vascular endothelial growth factor (VEGF) splice variants by LightCycler technology. *Clin Chem, 47*(4), 654-660. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11274014

Wen, J., Arakawa, T., & Philo, J. S. (1996). Size-exclusion chromatography with on-line light-scattering, absorbance, and refractive index detectors for studying proteins and their interactions. *Anal Biochem, 240*(2), 155-166. doi:10.1006/abio.1996.0345

Wentz-Hunter, K., & Potashkin, J. (1996). The small subunit of the splicing factor U2AF is conserved in fission yeast. *Nucleic Acids Res, 24*(10), 1849-1854. doi:10.1093/nar/24.10.1849

Wickens, M., Bernstein, D. S., Kimble, J., & Parker, R. (2002). A PUF family portrait: 3 ' UTR regulation as a way of life. *Trends in Genetics, 18*(3), 150-157. doi:Pii S0168-9525(01)02616-6

Doi 10.1016/S0168-9525(01)02616-6

Wiesner, S., Stier, G., Sattler, M., & Macias, M. J. (2002). Solution structure and ligand recognition of the WW domain pair of the yeast splicing factor Prp40. *J Mol Biol, 324*(4), 807-822. doi:10.1016/s0022-2836(02)01145-2

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., . . . Bahler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature, 453*(7199), 1239-U1239. doi:10.1038/nature07002

Wilkinson, M. E., Fica, S. M., Galej, W. P., & Nagai, K. (2021). Structural basis for conformational equilibrium of the catalytic spliceosome. *Mol Cell, 81*(7), 1439-1452 e1439. doi:10.1016/j.molcel.2021.02.021

Wilkinson, M. E., Fica, S. M., Galej, W. P., Norman, C. M., Newman, A. J., & Nagai, K. (2017). Postcatalytic spliceosome structure reveals mechanism of 3'-splice site selection. *Science, 358*(6368), 1283-1288. doi:10.1126/science.aar3729

Will, C. L., & Luhrmann, R. (2011). Spliceosome structure and function. *Cold Spring Harb Perspect Biol, 3*(7). doi:10.1101/cshperspect.a003707

Will, C. L., Schneider, C., MacMillan, A. M., Katopodis, N. F., Neubauer, G., Wilm, M., . . . Query, C. C. (2001). A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site. *EMBO J, 20*(16), 4536-4546. doi:10.1093/emboj/20.16.4536

Wilusz, J. E. (2018). A 360 degrees view of circular RNAs: From biogenesis to functions. *Wiley Interdiscip Rev RNA, 9*(4), e1478. doi:10.1002/wrna.1478

Wollerton, M. C., Gooding, C., Wagner, E. J., Garcia-Blanco, M. A., & Smith, C. W. (2004). Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Mol Cell, 13*(1), 91-100. doi:10.1016/s1097-2765(03)00502-1

Wood, V. (2006). How to get the most from fission yeast genome data: a report from the 2006 European Fission Yeast Meeting Computing Workshop. *Yeast, 23*(13), 905-912. doi:10.1002/yea.1419

Wood, V., Gwilliam, R., Rajandream, M. A., Lyne, M., Lyne, R., Stewart, A., . . . Nurse, P. (2002). The genome sequence of Schizosaccharomyces pombe. *Nature, 415*(6874), 871-880. doi:10.1038/nature724

Woods, A. S., & Ferre, S. (2005). Amazing stability of the arginine-phosphate electrostatic interaction. *J Proteome Res, 4*(4), 1397-1402. doi:10.1021/pr050077s

Wrehlke, C., Schmitt-Wrede, H. P., Qiao, Z., & Wunderlich, F. (1997). Enhanced expression in spleen macrophages of the mouse homolog to the human putative tumor suppressor gene ZFM1. *DNA Cell Biol, 16*(6), 761-767. doi:10.1089/dna.1997.16.761

Wu, J. Y., & Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell, 75*(6), 1061-1070. doi:10.1016/0092-8674(93)90316-i

Wu, S., Li, Y. L., Cheng, N. Y., Wang, C., Dong, E. L., Lu, Y. Q., . . . Chen, W. J. (2018). c.835-5T>G Variant in SMN1 Gene Causes Transcript Exclusion of Exon 7 and Spinal Muscular Atrophy. *J Mol Neurosci, 65*(2), 196-202. doi:10.1007/s12031-018-1079-1

Wu, S., Romfo, C. M., Nilsen, T. W., & Green, M. R. (1999). Functional recognition of the 3' splice site AG by the splicing factor U2AF35. *Nature, 402*(6763), 832-835. doi:10.1038/45590

Wurtele, H., Tsao, S., Lepine, G., Mullick, A., Tremblay, J., Drogaris, P., . . . Raymond, M. (2010). Modulation of histone H3 lysine 56 acetylation as an antifungal therapeutic strategy. *Nature Medicine, 16*(7), 774-780. doi:10.1038/nm.2175

Wyatt, P. J. (1993). Light-Scattering and the Absolute Characterization of Macromolecules. *Analytica Chimica Acta, 272*(1), 1-40. doi:Doi 10.1016/0003-2670(93)80373-S

Xing, Y., Xu, Q., & Lee, C. (2003). Widespread production of novel soluble protein isoforms by alternative splicing removal of transmembrane anchoring domains. *FEBS Lett, 555*(3), 572-578. doi:10.1016/s0014-5793(03)01354-1

Xu, D., Nouraini, S., Field, D., Tang, S. J., & Friesen, J. D. (1996). An RNA-dependent ATPase associated with U2/U6 snRNAs in pre-mRNA splicing. *Nature, 381*(6584), 709-713. doi:10.1038/381709a0

Xu, Y. Z., & Query, C. C. (2007). Competition between the ATPase Prp5 and branch region-U2 snRNA pairing modulates the fidelity of spliceosome assembly. *Mol Cell, 28*(5), 838-849. doi:10.1016/j.molcel.2007.09.022

Xuong, N. H., Jin, L., Kleinfelder, S., Li, S. D., Leblanc, P., Duttweiler, F., . . . Ellisman, M. (2007). Future directions for camera systems in electron microscopy. *Cellular Electron Microscopy, 79*, 721-739. doi:10.1016/S0091-679x(06)79028-8

Yan, C., Hang, J., Wan, R., Huang, M., Wong, C. C., & Shi, Y. (2015). Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science, 349*(6253), 1182-1191. doi:10.1126/science.aac7629

Yan, C., Wan, R., Bai, R., Huang, G., & Shi, Y. (2016). Structure of a yeast activated spliceosome at 3.5 A resolution. *Science, 353*(6302), 904-911. doi:10.1126/science.aag0291

Yan, C., Wan, R., Bai, R., Huang, G., & Shi, Y. (2017). Structure of a yeast step II catalytically activated spliceosome. *Science, 355*(6321), 149-155. doi:10.1126/science.aak9979

Yates, A. D., Allen, J., Amode, R. M., Azov, A. G., Barba, M., Becerra, A., . . . Flicek, P. (2022). Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. *Nucleic Acids Res, 50*(D1), D996-D1003. doi:10.1093/nar/gkab1007

Yoo, S. H., Mohawk, J. A., Siepka, S. M., Shan, Y., Huh, S. K., Hong, H. K., . . . Takahashi, J. S. (2013). Competing E3 ubiquitin ligases govern circadian periodicity by degradation of CRY in nucleus and cytoplasm. *Cell, 152*(5), 1091-1105. doi:10.1016/j.cell.2013.01.055

Yoshida, H., Park, S. Y., Oda, T., Akiyoshi, T., Sato, M., Shirouzu, M., . . . Obayashi, E. (2015). A novel 3' splice site recognition by the two zinc fingers in the U2AF small subunit. *Genes Dev, 29*(15), 1649-1660. doi:10.1101/gad.267104.115

Yoshida, H., Park, S. Y., Sakashita, G., Nariai, Y., Kuwasako, K., Muto, Y., . . . Obayashi, E. (2020). Elucidation of the aberrant 3' splice site selection by cancer-associated mutations on the U2AF1. *Nat Commun, 11*(1), 4744. doi:10.1038/s41467-020-18559-6

Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., . . . Ogawa, S. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature, 478*(7367), 64-69. doi:10.1038/nature10496

Yukawa, M., & Maki, T. (1931). Schizosaccharomyces japonicus nov. spec La Bul Sci Falkultato Terkultura Kjusu Imp Univ. *Fukuoka, Japan, 4*, 218-226.

Zamore, P. D., Patton, J. G., & Green, M. R. (1992). Cloning and domain structure of the mammalian splicing factor U2AF. *Nature, 355*(6361), 609-614. doi:10.1038/355609a0

Zarnack, K., Konig, J., Tajnik, M., Martincorena, I., Eustermann, S., Stevant, I., . . . Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell, 152*(3), 453-466. doi:10.1016/j.cell.2012.12.023

Zhan, X., Yan, C., Zhang, X., Lei, J., & Shi, Y. (2018a). Structure of a human catalytic step I spliceosome. *Science, 359*(6375), 537-545. doi:10.1126/science.aar6401

Zhan, X., Yan, C., Zhang, X., Lei, J., & Shi, Y. (2018b). Structures of the human pre-catalytic spliceosome and its precursor spliceosome. *Cell Research, 28*(12), 1129-1140. doi:10.1038/s41422-018-0094-7

Zhang, M., Zamore, P. D., Carmo-Fonseca, M., Lamond, A. I., & Green, M. R. (1992). Cloning and intracellular localization of the U2 small nuclear ribonucleoprotein auxiliary factor small subunit. *Proc Natl Acad Sci U S A, 89*(18), 8769-8773. doi:10.1073/pnas.89.18.8769

Zhang, M. Q., & Marr, T. G. (1994). Fission yeast gene structure and recognition. *Nucleic Acids Res, 22*(9), 1750-1759. doi:10.1093/nar/22.9.1750

Zhang, T., Haws, P., & Wu, Q. (2004). Multiple variable first exons: a mechanism for cell- and tissue-specific gene regulation. *Genome Res, 14*(1), 79-89. doi:10.1101/gr.1225204

Zhang, X., Yan, C., Hang, J., Finci, L. I., Lei, J., & Shi, Y. (2017). An Atomic Structure of the Human Spliceosome. *Cell, 169*(5), 918-929 e914. doi:10.1016/j.cell.2017.04.033

Zhang, X., Yan, C., Zhan, X., Li, L., Lei, J., & Shi, Y. (2018). Structure of the human activated spliceosome in three conformational states. *Cell Research, 28*(3), 307-322. doi:10.1038/cr.2018.14

Zhang, X. F., Zhan, X. C., Yan, C. Y., Zhang, W. Y., Liu, D. L., Lei, J. L., & Shi, Y. G. (2019). Structures of the human spliceosomes before and after release of the ligated exon. *Cell Research, 29*(4), 274-285. doi:10.1038/s41422-019-0143-x

Zhang, Y., Madl, T., Bagdiul, I., Kern, T., Kang, H. S., Zou, P., . . . 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*(2), 1343-1354. doi:10.1093/nar/gks1097

Zhang, Z., Will, C. L., Bertram, K., Dybkov, O., Hartmuth, K., Agafonov, D. E., . . . Stark, H. (2020). Molecular architecture of the human 17S U2 snRNP. *Nature, 583*(7815), 310-313. doi:10.1038/s41586-020-2344-3

Zhang, Z. W., Rigo, N., Dybkov, O., Fourmann, J. B., Will, C. L., Kumar, V., . . . Luhrmann, R. (2021). Structural insights into how Prp5 proofreads the pre-mRNA branch site. *Nature, 596*(7871), 296-+. doi:10.1038/s41586-021-03789-5

Zhao, Z., McKee, C. J., Kessl, J. J., Santos, W. L., Daigle, J. E., Engelman, A., . . . Kvaratskhelia, M. (2008). Subunit-specific protein footprinting reveals significant structural rearrangements and a role for N-terminal Lys-14 of HIV-1 Integrase during viral DNA binding. *J Biol Chem, 283*(9), 5632-5641. doi:10.1074/jbc.M705241200

Zhou, L., Hang, J., Zhou, Y., Wan, R., Lu, G., Yin, P., . . . Shi, Y. (2014). Crystal structures of the Lsm complex bound to the 3' end sequence of U6 small nuclear RNA. *Nature, 506*(7486), 116-120. doi:10.1038/nature12803

Zhou, Z., & Fu, X. D. (2013). Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma, 122*(3), 191-207. doi:10.1007/s00412-013-0407-z

Zhou, Z., Licklider, L. J., Gygi, S. P., & Reed, R. (2002). Comprehensive proteomic analysis of the human spliceosome. *Nature, 419*(6903), 182-185. doi:10.1038/nature01031

Zhuang, Y. A., Goldstein, A. M., & Weiner, A. M. (1989). UACUAAC is the preferred branch site for mammalian mRNA splicing. *Proc Natl Acad Sci U S A, 86*(8), 2752-2756. doi:10.1073/pnas.86.8.2752

Zimm, B. H. (1948). The Scattering of Light and the Radial Distribution Function of High Polymer Solutions. *Journal of Chemical Physics, 16*(12), 1093-1099. doi:Doi 10.1063/1.1746738

Zong, F. Y., Fu, X., Wei, W. J., Luo, Y. G., Heiner, M., Cao, L. J., . . . Hui, J. (2014). The RNA-binding protein QKI suppresses cancer-associated aberrant splicing. *PLoS Genet, 10*(4), e1004289. doi:10.1371/journal.pgen.1004289

Zorio, D. A., & Blumenthal, T. (1999a). Both subunits of U2AF recognize the 3' splice site in Caenorhabditis elegans. *Nature, 402*(6763), 835-838. doi:10.1038/45597

Zorio, D. A., & Blumenthal, T. (1999b). U2AF35 is encoded by an essential gene clustered in an operon with RRM/cyclophilin in Caenorhabditis elegans. *RNA, 5*(4), 487-494. doi:10.1017/s1355838299982225

Zuo, P., & Maniatis, T. (1996). The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. *Genes Dev, 10*(11), 1356-1368. doi:10.1101/gad.10.11.1356

Appendix *I*

Design principles and development of the U2AF dimer and U2AF/SF1 trimer expression system

The development of the U2AF dimer and U2AF/SF1 trimer expression system starting from the very beginning of this project will be briefly reviewed below in Section *I*-1., and the final optimized system will be summarized in Section *I*-2. This system was used to produce multiple variants of the U2AF dimer and U2AF/SF1 trimer complex which were used in this thesis through the use of different protein constructs. These complexes and their constituent protein constructs are catalogued and described in detail in Section *I*-3.

I-1. Development stage of the expression system

A large number of techniques and strategies were attempted while developing the expression system, and the majority of these will not be discussed below. For clarity and focus, this discussion is limited to strategies that proceed as successive steps on a productive path towards the final system used as the basis for experiments in Chapters 2 and 3. All work to develop the system was done using *E. coli* as a host. In all vectors used, expression is driven by T7 polymerase under the control of the lac (lactose) promoter, and all N-terminal affinity tag(s) can be removed through cleavage by TEV protease (Parks, Leuther, Howard, Johnston, & Dougherty, 1994; Studier, Rosenberg, Dunn, & Dubendorff, 1990). All attempts to express U2AF23 and SF1 included $ZnCl_2$ in the induction step due to the presence of zinc-binding domains in these proteins, and all productive steps to develop the expression system involved purifying protein under non-denaturing conditions.

The first pilot experiments consisted of expressing each protein in isolation from the pMCSG9 LIC (ligation independent cloning) vector which appends a hexahistidine tag, followed by an MBP (maltose-binding protein) tag, followed by a TEV cleavage site at the N-terminus of the protein construct; pMCSG9 is a derivative of pMCSG7, which appends a TEV-cleavable

hexahistidine tag to the N-terminus of the protein construct (Aslanidis & de Jong, 1990; Donnelly et al., 2006; Eschenfeldt, Lucy, Millard, Joachimiak, & Mark, 2009; Stols et al., 2002). Although a compatible affinity purification tag for a particular protein construct can only be determined through trial and error, MBP was chosen as an affinity tag for initial pilot experiments to establish the behaviour of U2AF23, U2AF59 and SF1 in isolation because it is unusually effective at increasing the solubility of poorly-behaved target proteins, significantly more so than GST (glutathione-S transferase), another routinely used affinity tag, and in certain cases promotes the correct folding of the target protein into its biologically active form (Kapust & Waugh, 1999). Therefore, even if these proteins are poorly behaved (displaying significant degradation and/or insolubility), MBP provides advantages that increase the likelihood of purifying enough of these proteins to assess their properties and behaviour.

Full-length pMCSG9-U2AF23 produced very large quantities of intact, soluble protein with no precipitation and little to no contamination detectable by SDS-PAGE. However, SEC purification indicated that the protein exists as a mixture of oligomers. All attempts to purify this protein as a monomer were unsuccessful. Attempts to express and purify full-length pMCSG9-U2AF59 were unsuccessful as this construct generates very little intact protein which is present in the aggregate fraction when SEC-purified. Most of the purified protein consists of multiple degradation products and free his-MBP.

Before attempting to express *S. pombe* SF1, both the human and *S. pombe* SF1 sequences were analyzed using the PsiPred secondary structure prediction server in order to inform construct boundaries since the N-terminus and C-terminus are poorly conserved, uncharacterized, predicted to be unstructured, and are not known to have any role in splicing. The N-terminus of both human and *S. pombe* SF1 up to and including the ULM is predicted to be

intrinsically disordered, so the N-terminal boundary of the pilot construct began at M52 in order to truncate this region while retaining the ULM. C-terminal to the zinc knuckle(s), both human and *S. pombe* SF1 contain a region of conserved residues predicted to form an α-helix (see Fig. 1-20). C-terminal to this predicted structural element, the PsiPred output predicts the remainder of the C-terminus in both orthologues to be intrinsically disordered. Therefore, the pilot construct ended at E373 in order to delete the C-terminus while retaining this predicted structural element. Large quantities of intact pMCSG9-SF1 (M52-E373) were purified with little contamination (as determined by SDS-PAGE). However, the protein contained insoluble aggregates that remained in suspension even after prolonged centrifugation and the soluble component eluted exclusively as aggregate during SEC purification.

Since the RS domain of U2AF-L is poorly conserved, predicted to be unstructured, and has poorly defined roles in splicing, a new construct was designed, pMCSG9-U2AF59 (K129-W517) in which the RS domain was deleted which generated pure, soluble, monomeric protein. To further build upon this promising result, another construct was designed, pMCSG7-U2AF59 (E106-W17); because sequence analysis by the PsiPred server predicts that U2AF59 (S111-E117) forms an α-helix, the N-terminus of this construct was extended to retain this predicted structural element. This α-helix was subsequently confirmed to exist when the apo *S. pombe* U2AF23/U2AF59 X-ray structure was reported. This construct also produced pure, soluble, monomeric protein, which indicated that the ideal solution behaviour of pMCSG9-U2AF59 (K129-W517) is not dependent on the solubilizing and stabilizing properties of MBP.

Because both U2AF23 and SF1 interact with U2AF59, the solubility problems of these proteins when expressed in isolation may be caused by a partially unfolded structure and these proteins may require interaction with U2AF59 to assume a fully folded structure and full

solubility under non-denaturing conditions. Therefore, pMCSG7-U2AF59 (E106-W517) was used as the foundation to build a larger complex through co-expression. The first successful purification of a higher-order complex was realized by cloning full-length, untagged U2AF23 into MCS2 (multiple cloning site # 2) of the co-expression vector, pACYC Duet-1 (Novagen), co-transforming and co-expressing this construct with pMCSG7-U2AF59 (E106-W517), and co-purifying both U2AF59 and U2AF23 as a pre-formed dimer using the hexahistidine tag present on U2AF59. Protein purified from this system consisted of a mixture of U2AF59 monomer and U2AF dimer with little to no contamination as determined by SDS-PAGE; SEC purification revealed that the U2AF dimer was exclusively present as a monodisperse particle.

To further develop the expression system and purify the U2AF/SF1 trimer, untagged SF1 (M52-Q309), and SF1 (M52-E373) were both cloned into the empty MCS1 of the U2AF23-containing vector from the previous step; the M52-Q309 construct was created as a tool to help investigate the role of the zinc knuckles of SF1. When co-transformed and co-expressed with the U2AF59-containing vector, both SF1 constructs produced stable, soluble, and monodisperse U2AF/SF1 trimer as determined by SEC purification. However, the SEC trace revealed that the protein sample consisted of a mixture of U2AF59 monomer, U2AF dimer, U2AF59/SF1 dimer, and U2AF/SF1 trimer. Attempts to isolate U2AF/SF1 trimer from this mixture were impractical and inefficient.

In order to bypass the problem of multiple protein species from the affinity purification step, adjusting the relative expression levels of U2AF23, U2AF59, and SF1 was evaluated as a potential solution. Conveniently, U2AF59 interacts with both U2AF23 and SF1, and is the affinity-tagged component of the complex. Therefore, reducing the expression of U2AF59 and/or increasing the expression of both U2AF23 and SF1 in the *E. coli* host was expected to resolve

the problem of multiple co-purified species by saturating both the U2AF23 and SF1 interaction sites on U2AF59.

The amount of target protein retrieved from an *E. coli*-based expression system is influenced by multiple factors, one of which is the copy number of the expression plasmid; pMCSG7 contains a pBR322-based origin of replication, which maintains the plasmid copy number at 15-20 copies per cell, whereas pACYC Duet-1 contains a p15A origin of replication, which maintains the plasmid copy number at 10-12 copies per cell (Bolivar et al., 1977; Chang & Cohen, 1978; McIntosh et al., 2012). All other factors being equal, this fact suggests that switching the origin of replication for the U2AF59 host vector and U2AF23/SF1 host vector may resolve this problem, at least partially. U2AF59 was re-cloned into pACYC Duet-1, and U2AF23/SF1 were re-cloned into the dual expression vector pET Duet-1 (Novagen), which contains the pBR322 origin of replication. These new expression vectors were co-transformed and co-expressed as before, and the problem was completely resolved. This step represents the final optimized system.

I-2. Optimized U2AF dimer and U2AF/SF1 trimer expression system

The foundation for all experiments in this thesis is two protein complexes: the U2AF heterodimer and the U2AF/SF1 heterotrimer. Different protein constructs were used in various combinations to generate variants of these two foundational complexes for a more detailed experimental dissection of the properties of these complexes.

Although the *S. pombe* U2AF/SF1 complex exists *in vivo* as tightly associated, pre-formed complex in the absence of RNA, experiments were also completed using the U2AF dimer as a model for the metazoan system because in metazoans the U2AF dimer initially

recognizes the 3′ SS in the absence of SF1. The basic organization of the protein complexes produced from the expression system is summarized below in Fig. *I*-1; Fig. *I*-1B represents the U2AF/SF1 complex containing the longest, prototype SF1 construct, although shorter SF1 constructs omitting the zinc knuckles were also used.

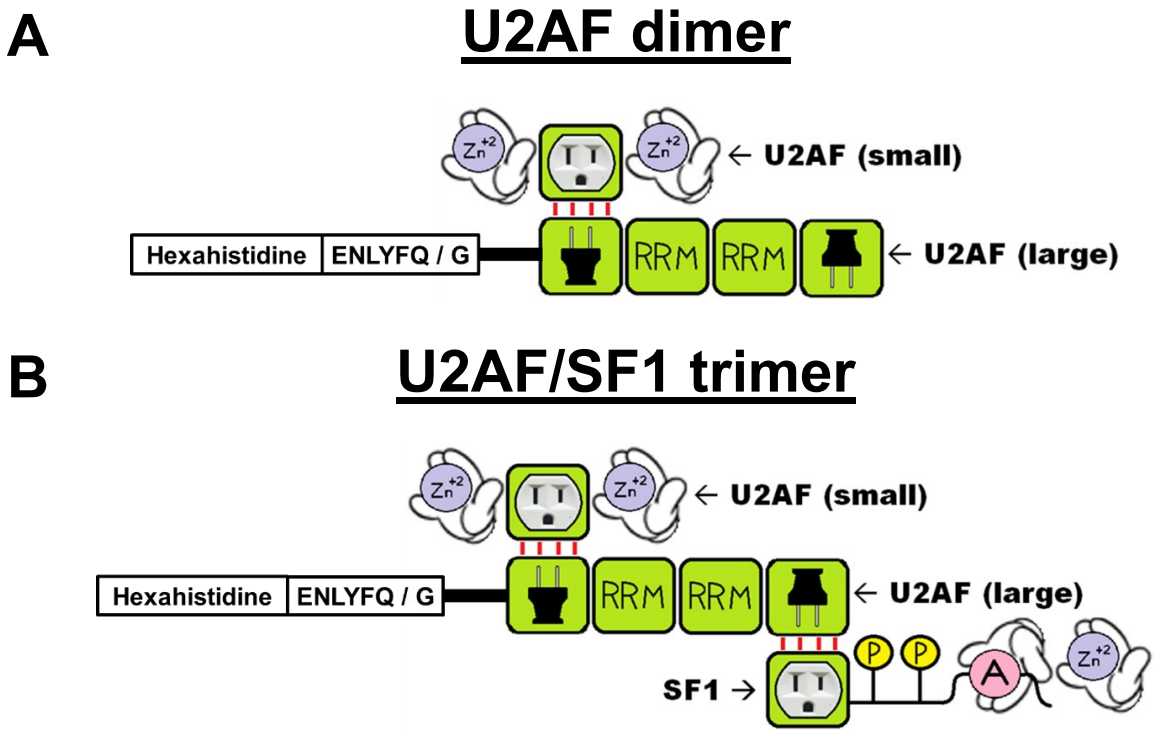


Figure *I*-1. Organizational diagram of the optimized U2AF dimer and U2AF/SF1 trimer expression system. The basic organization of protein domains is summarized for the U2AF dimer in (A) and the U2AF/SF1 trimer in (B). Included in this diagram is the location of the hexahistidine affinity tag and TEV cleavage site. See Fig. 1-6. for a legend of domain symbols.

I-3. Catalogue of protein constructs and protein complexes used

Protein constructs used in this thesis are catalogued and described in Sections *I*-3.1. to *I*-3.3. and the complexes themselves which were assembled from these constructs are catalogued and described in Section *I*-3.4 along with the experiments they were used for.

I-3.1. Catalogue of U2AF23 constructs used

Three constructs were used all of which were wildtype, untagged constructs and two of which were full-length constructs. The first construct designed was a full-length construct cloned with BglII/XhoI. This was a component in all five complexes used for biochemical characterization the U2AF dimer and U2AF/SF1 trimer in Chapter 2. Subsequently, this construct was re-cloned using NdeI/XhoI to eliminate the N-terminal Met-Ala-Asp-Leu cloning artifact introduced by BglII. This updated construct was a component of the U2AF-L chimera complex that formed the basis of all SEC-SAXS experiments described in Chapter 3.

The third construct used was cloned with NdeI/XhoI and is missing the disordered region that follows the C-terminal α-helix of U2AF23. This construct is a component of the two complexes that form the basis of all SAXS experiments described in Chapter 3.

Structural experiments are based on the two constructs that do not contain an N-terminal cloning artifact as they are expected to generate cleaner data. The disordered C-terminus was deleted in complexes used for SAXS for the same reason. Ideally, this construct would also have been used for SEC-SAXS experiments, however, a U2AF/SF1 complex containing the U2AF-L chimera will only produce stable and soluble protein if the full-length U2AF23 construct is used as a component and will not if the truncated U2AF23 construct is used.

The constructs described above are catalogued below in Table *I*-1 with their molecular weights.

Table *I*-1: Catalogue of U2AF23 constructs used

Construct	MW
M1-N216 (cloned using NdeI/XhoI)	25.03 kDa
M1-N216 (cloned using BglII/XhoI)	25.46 kDa
M1-E194	22.47 kDa

I-3.2. Catalogue of U2AF-L constructs used

Two wildtype constructs were used. One is TEV-cleavable, hexahistidine-tagged U2AF59 (E106-W517) which was a component in all complexes, except for one complex designed exclusively for the SEC-SAXS experiments. The complex forming the basis of all SEC-SAXS experiments contained the TEV-cleavable, hexahistidine-tagged U2AF-L chimera construct (see Appendix *III* for details).

The constructs described above are catalogued below in Table *I*-2 with their molecular weights.

Table *I*-2: Catalogue of U2AF-L constructs used

Construct	MW
U2AF59 (E106-W517), uncleaved	48.54 kDa
U2AF59 (E106-W517), cleaved	46.13 kDa
U2AF-L chimera, uncleaved	46.87 kDa
U2AF-L chimera, cleaved	44.46 kDa

I-3.3. Catalogue of SF1 constructs used

All constructs used were untagged. Two wildtype constructs and their phosphomimetic counterparts were used for all experiments, with the exception of the SEC-SAXS experiments. The prototype construct is wildtype SF1 (M52-E373). In Chapter 2, comparing the wildtype constructs with their phosphomimetic counterparts was used to investigate the roles of phosphorylation of SF1, and comparing the constructs ending at Q309 with those ending at E373 was used to investigate the role of the zinc knuckles of SF1.

One construct was specifically designed for the SEC-SAXS experiments described in Chapter 3. This construct is five residues longer at the C-terminus as a precaution since U2AF/SF1 trimer containing the U2AF-L chimera will only generate stable, soluble protein if U2AF23 is not truncated and if part of the disordered N-terminus of U2AF59 is retained. Additionally, this SF1 construct is a phosphomimetic. Phosphorylation of SF1 rigidifies the overall structure of the U2AF/SF1 complex and increases the interface between U2AF-L/SF1. The EMSAs described in Chapter 2 also reveal that the phosphomimetic SF1 constructs have increased affinity for the target pre-mRNA, which is consistent with the previously published effects of phosphorylation of SF1. Together, these observations indicate that phosphomimetic mutants of SF1 are a reliable system to model phosphorylated SF1, and that a phosphomimetic SF1 construct is useful as a tool to assist rigid body modelling using SEC-SAXS derived data by contributing additional restraints to the relative positioning of rigid bodies.

The constructs described above are catalogued below in Table *I*-3 with their molecular weights.

Table *I*-3: Catalogue of SF1 constructs used

Construct	MW
M52-Q309, wildtype	29.82 kDa
M52-Q309 (S131E, S133E)	29.90 kDa
M52-E373, wildtype	37.10 kDa
M52-E373 (S131E, S133E)	37.18 kDa
M52-S378 (S131E, S133E)	37.56 kDa

I-3.4. Catalogue of U2AF dimer and U2AF/SF1 trimer variants used

Biochemical investigations into the U2AF dimer and U2AF/SF1 trimer are described in Chapter 2. All experiments described in Chapter 2 are based on one U2AF dimer complex and four U2AF/SF1 trimer complexes. All five complexes contain untagged, full-length U2AF23 cloned with BglII/XhoI and TEV-cleavable, hexahistidine-tagged U2AF59 (E106-W517). The four U2AF/SF1 trimer complexes differ only with respect to the SF1 construct used. The five complexes forming the basis of Chapter 2 are catalogued below in Table *I*-4, which additionally also includes the abbreviated names that are used when appropriate for simplicity and clarity.

Table *I*-4: U2AF dimer and U2AF/SF1 trimer variants used in Chapter 2

Description	SF1 construct	Abbreviation
U2AF dimer	n/a	n/a
wildtype U2AF/SF1 trimer with SF1 zinc knuckles deleted	M52-Q309, wildtype	U2AF/SF1$_{\Delta Zn,wt}$
phosphomimetic U2AF/SF1 trimer with SF1 zinc knuckles deleted	M52-Q309 (S131E, S133E)	U2AF/SF1$_{\Delta Zn,mimetic}$
wildtype, prototype U2AF/SF1 trimer	M52-E373, wildtype	U2AF/SF1$_{wt}$
phosphomimetic, prototype U2AF/SF1 trimer	M52-E373 (S131E, S133E)	U2AF/SF1$_{mimetic}$

Structural investigations into the U2AF dimer and U2AF/SF1 trimer are described in Chapter 3 and consist of SAXS and SEC-SAXS experiments. The SAXS experiments are based on one wildtype U2AF dimer complex and one wildtype U2AF/SF1 trimer complex, whereas the

SEC-SAXS experiments are exclusively based on one U2AF/SF1 complex specifically designed for this purpose containing the U2AF-L chimera construct and a phosphomimetic SF1 construct. The three complexes forming the basis of Chapter 3 are catalogued below in Table *I*-5.

Table *I*-5: U2AF dimer and U2AF/SF1 trimer variants used in Chapter 3

Description	U2AF23 construct	U2AF-L construct	SF1 construct
wildtype U2AF dimer containing truncated U2AF23	M1-E194	U2AF59 (E106-W517)	n/a
wildtype, prototype U2AF/SF1 trimer containing truncated U2AF23	M1-E194	U2AF59 (E106-W517)	M52-E373, wildtype
prototype U2AF/SF1 trimer containing U2AF-L chimera and phosphomimetic SF1	M1-N216	U2AF-L chimera	M52-S378 (S131E, S133E)

The cloning, expression, and purification of U2AF dimer and U2AF/SF1 trimer complexes is described in Chapter 2 and all complexes described above in Table *I*-4 and Table *I*-5 can be categorized into three groups based on purification behaviour; complexes belonging to the same group are purified using the same protocols and are indistinguishable from one another during purification, with the caveat that U2AF23 (M1-E194) migrates slightly faster than full-length U2AF23 on an SDS-PAGE gel. The groupings are provided in Table *I*-6 below.

Table *I*-6: Catalogue of U2AF dimer and U2AF/SF1 trimer variants grouped by expression and purification behaviour

Grouping	U2AF23 construct	U2AF-L construct	SF1 construct
U2AF dimer	M1-N216	U2AF59 (E106-W517)	n/a
	M1-E194	U2AF59 (E106-W517)	n/a
U2AF/SF1 trimer with SF1 zinc knuckles deleted	M1-N216	U2AF59 (E106-W517)	M52-Q309, wildtype
	M1-N216	U2AF59 (E106-W517)	M52-Q309 (S131E, 133E)
prototype U2AF/SF1 trimer	M1-N216	U2AF59 (E106-W517)	M52-E373, wildtype
	M1-E194	U2AF59 (E106-W517)	M52-E373, wildtype
	M1-N216	U2AF59 (E106-W517)	M52-E373 (S131E, S133E)
	M1-N216	U2AF-L chimera	M52-S378 (S131E, S133E)

With respect to the purification results in Chapter 2, they are presented from the complexes described in Table *I*-4, as these were the first complexes successfully designed and characterized.

Appendix *II*

Commercially obtained synthetic oligonucleotides used in this thesis

II-1. U2AF dimer and U2AF/SF1 trimer cloning primers

Table *II*-1 is a catalogue of all the primers used to clone the U2AF23 constructs listed in Table *I*-1. Table *II*-2 is a catalogue of all the PCRs required to create the two U2AF-L constructs listed in Table *I*-2, and Table *II*-3 is a catalogue of all the primers used to clone these U2AF-L constructs. Table *II*-4 is a catalogue of all the PCRs required to create the SF1 constructs listed in Table *I*-3, and Table *II*-5 is a catalogue of all the primers used to clone these SF1 constructs.

Table *II*-1: U2AF23 cloning primers

Oligonucleotide name and sequence (5′→3′)	Oligonucleotide description
CG29: GCGCGCAGATCTCATGGCAAGTCATTTGGCA	Forward primer, U2AF23 beginning at M1. Digest with BglII.
CG30: ATATCTCGAGTTATCATTAATTTTTGCGTTCAGCAGTTACACTGAC	Reverse primer, U2AF23 ending at N216. Digest with XhoI.
CG240: GCGCGCCATATGGCAAGTCATTTGGCA	Forward primer, U2AF23 beginning at M1. Digest with NdeI.
CG271: GCGCGCCTCGAGTTATCATTCTTCTGCTGCATTTAA	Reverse primer, U2AF23 ending at E194. Digest with XhoI.

Table *II*-2: PCRs used to create U2AF-L constructs

Construct	PCR	Amplicon	Primers
U2AF59 (E106-W517, wildtype)	Standard PCR	U2AF59 (E106-W517, wildtype) ORF	CG57 fwd CG32 rev
U2AF-L chimera	First cloning step: First overlapping PCR step	U2AF65 (V137-A342)	CG251 fwd CG252 rev
		U2AF59 (M394-W517)	CG253 fwd CG32 rev
	First cloning step: Second overlapping PCR step	U2AF65 (V137-A342) + U2AF59 (M394-W517)	CG251 fwd CG32 rev
	Second cloning step: First overlapping PCR step	U2AF59 (S93-A161) + QSA	CG229 fwd CG263 rev
		U2AF65 (V137-A342) + U2AF59 (M394-W517)	CG262 fwd CG32 rev
	Second cloning step: Second overlapping PCR step	U2AF59 (S93-A161) + QSA + U2AF65 (V137-A342) + U2AF59 (M394-W517)	CG229 fwd CG32 rev

Table *II*-3: U2AF-L cloning primers

Oligonucleotide name and sequence (5′→3′)	Oligonucleotide description
CG32: ATATGAGCTCTTATCATCACCATGCATTAGCTTTATAGCAATCCT	Reverse primer, digest with SacI.
CG57: ATATGGATCCGGAAAATTTGTATTTTCAAGGTGGTGAAAGAAGCGTAAGGTCTATCG	Forward primer, digest with BamHI.
CG229: GCGCGCGGATCCGGAAAATTTGTATTTTCAAGGTTCATCTGTCGGAAGAAGT	Forward primer, digest with BamHI.
CG251: GCGCGCGGATCCCAGATGACCAGACAAGCCCG	Forward primer, digest with BamHI.
CG252: AGACTTATCTATCATGGCATTCTTGGCTCC	Reverse overlapping PCR primer.
CG253: GGAGCCAAGAATGCCATGATAGATAAGTCT	Forward overlapping PCR primer.
CG262: CCTTTGCCAGGCGCTCAGTCTGCAGTGCCCGTTGTTGGGAGCCAGATGACCAGACAA	Forward overlapping PCR primer.
CG263: TTGTCTGGTCATCTGGCTCCCAACAACGGGCACTGCAGACTGAGCGCCTGGCAAAGG	Reverse overlapping PCR primer.

Table *II*-4: PCRs used to create SF1 constructs

Construct	PCR	Primers
M52-Q309, wildtype	Standard PCR	CG53 fwd CG55 rev
M52-Q309 (S131E, S133E)	First overlapping PCR step (N-terminal piece)	CG53 fwd CG82 rev
	First overlapping PCR step (C-terminal piece)	CG83 fwd CG55 rev
	Second overlapping PCR step	CG53 fwd CG55 rev
M52-E373, wildtype	Standard PCR	CG53 fwd CG56 rev
M52-E373 (S131E, S133E)	First overlapping PCR step (N-terminal piece)	CG53 fwd CG82 rev
	First overlapping PCR step (C-terminal piece)	CG83 fwd CG56 rev
	Second overlapping PCR step	CG53 fwd CG56 rev
M52-S378 (S131E, S133E)	First overlapping PCR step (N-terminal piece)	CG53 fwd CG82 rev
	First overlapping PCR step (C-terminal piece)	CG83 fwd CG241 rev
	Second overlapping PCR step	CG53 fwd CG241 rev

Table *II*-5: SF1 cloning primers

Oligonucleotide name and sequence (5′→3′)	Oligonucleotide description
CG53: GCGCGCTCATGATGGATCATAGACCGGATG	Forward primer, SF1 beginning at M52. Digest with BspHI.
CG55: GCGCGCGAATTCTTATCACTGATTTTCATCG TCTCGTAAAG	Reverse primer, SF1 ending at Q309. Digest with EcoRI.
CG56: GCGCGCGAATTCTTATCACTCTTGCATAAG ACTTTGATATTCC	Reverse primer, SF1 ending at E373. Digest with EcoRI.
CG82: AGGAGGTTCAGGTTCCCGTTCACGGTGATG	Reverse overlapping PCR primer to introduce S131E, S133E mutation.
CG83: ACGGGAACCTGAACCTCCTCCGCAATACGA	Forward overlapping PCR primer to introduce S131E, S133E mutation.
CG241: GCGCGCGAATTCTTATCACGATCCCCCACC AAGCTCTTGCATAAGACTTTGATATTCC	Reverse primer, SF1 ending at S378. Digest with EcoRI.

II-2. Oligonucleotides used to characterize U2AF dimer and U2AF/SF1 trimer complexes

Table *II*-6 below catalogues all of the commercially obtained synthetic DNA oligonucleotides used as template to transcribe 3′ SS model RNAs for the biochemical characterization of U2AF dimer and U2AF/SF1 trimer complexes in this thesis and provides a brief description of their use. The T7 RNA polymerase promoter sequence is underlined.

Table *II*-7 below catalogues all of the commercially obtained synthetic RNA oligonucleotides used for both the biochemical and structural characterization of the U2AF dimer and U2AF/SF1 trimer in this thesis and provides a brief description of their 3′ SS sequence signatures.

Table *II*-6: Synthetic DNA oligonucleotides used as transcription template for 3′ SS model RNAs

Oligonucleotide name and sequence (5′→3′)	Oligonucleotide description
CG74: TAATACGACTCACTATAGGGTTACTAACTTTTTTTTTAGTGC	Forward transcription template for the wildtype *S. pombe* pre-p14 model intron with an optimal BPS.
CG75: GCACTAAAAAAAAAGTTAGTAACCCTATAGTGAGTCGTATTA	Reverse complement of CG74.
CG76: TAATACGACTCACTATAGGGTGTCGCAGTTTTTTTTTAGTGC	Forward transcription template for the *S. pombe* pre-p14 model intron with a scrambled BPS.
CG77: GCACTAAAAAAAAACTGCGACACCCTATAGTGAGTCGTATTA	Reverse complement of CG76.
CG78: TAATACGACTCACTATAGGGTTACTAACTTTTTTTTTTTTAGTGC	Forward transcription template for the *S. pombe* pre-p14 model intron with an optimal BPS and U12 intron.
CG79: GCACTAAAAAAAAAAAAAGTTAGTAACCCTATAGTGAGTCGTATTA	Reverse complement of CG78.
CG80: TAATACGACTCACTATAGGGAATGATTGAAAAAAAAATCACG	Forward transcription template for the model intron corresponding to a forward complement of the intron transcribed using CG74 and CG75.
CG81: CGTGATTTTTTTTTCAATCATTCCCTATAGTGAGTCGTATTA	Reverse complement of CG80.

Table *II*-7: Synthetic 3′ SS model RNAs

Oligonucleotide name and sequence (5′→3′)	Oligonucleotide description
CG92: UUACUAACUUUUUUUUUAGUGC	•Wildtype *S. pombe* pre-p14 model intron with an optimized BPS. •Identical to the RNA transcribed using CG74 and CG75 as transcription template but lacking the GGG transcription artifact at the 5′ end of the sequence.
CG109: UAUACUAACAAUUUUUUUUUUUUUUUAGUGC	•Optimized BPS corresponds to the sequence used to solve both the human RNA-bound SF1 structure, and the *S. cerevisiae* RNA-bound Bbp structure. •PPT consists of a U14 sequence. •AG di-nucleotide and 3′ tri-nucleotide artifact are identical to CG92.
CG120: UUUUUUUUUAGUGC	•Identical to CG92, but with the BPS deleted.
CG158: CUAACUUUUUUUUUAG	•Identical to CG92 but with abbreviated 5′ and 3′ termini. •BPS is shortened to conform to the *S. pombe* BPS consensus sequence (CURAy) with no additional nucleotides. •Sequence 3′ to the AG di-nucleotide is deleted.

Appendix *III*

Design principles of the U2AF-L chimera construct

The SEC-SAXS experiments described in Chapter 3 are combined with rigid body modelling in order to approximate the spatial positioning of the domains that comprise the U2AF/SF1 complex. Because this modelling technique requires rigid bodies, it is dependent on the existence of experimentally derived atomic structures.

Although the RNA-binding cores of U2AF-S, U2AF-L and SF1 are conserved between humans and *S. pombe*, accurate modelling of the U2AF-L/RNA interaction within the context of the U2AF/SF1 complex is confounded by the large degree of variation within both the length and sequence composition of the PPT. Though several protein/oligonucleotide structures containing both RRMs and a natural inter-RRM linker have been solved for human U2AF65, the determination of these structures was only possible by fixing the binding register with unique DNA/RNA hybrids specifically designed for this purpose. Because six of these structures contain 8 nucleotides and one structure contains 9 nucleotides, it is still unknown how the RRMs accommodate length variation in the PPT, and how the binding register is determined for PPTs longer than this.

Although U2AF-L is conserved between humans and *S. pombe*, the sequence differences between U2AF65 and U2AF59 in the RRMs and inter-RRM linker, as well as significant differences in intron architecture between humans and *S. pombe* in the PPT (see Section 1-6.1.) indicate that the human U2AF65/oligonucleotide structures may not accurately represent the corresponding *S. pombe* U2AF59/PPT interaction. Since U2AF65 and U2AF59 may possess different PPT binding preferences, the DNA/RNA hybrids co-crystallized with U2AF65 may not bind U2AF59 in the same register.

In order to address the limitations in using U2AF65 derived rigid bodies to model the *S. pombe* U2AF/SF1 complex, a U2AF-L chimera was created. Fig. *III*-1 below is an annotated

alignment of metazoan U2AF-L orthologues from *C. elegans*, *H. sapiens*, and *D. melanogaster*. This alignment is followed by and is a supplement to Sections *III*-1. to *III*-4. below, which describe the sequence structure of this construct from N-terminus to C-terminus, as well as the basis for including each distinct component of this construct. Finally, a visual representation of the cloning strategy has been provided in Section *III*-5.

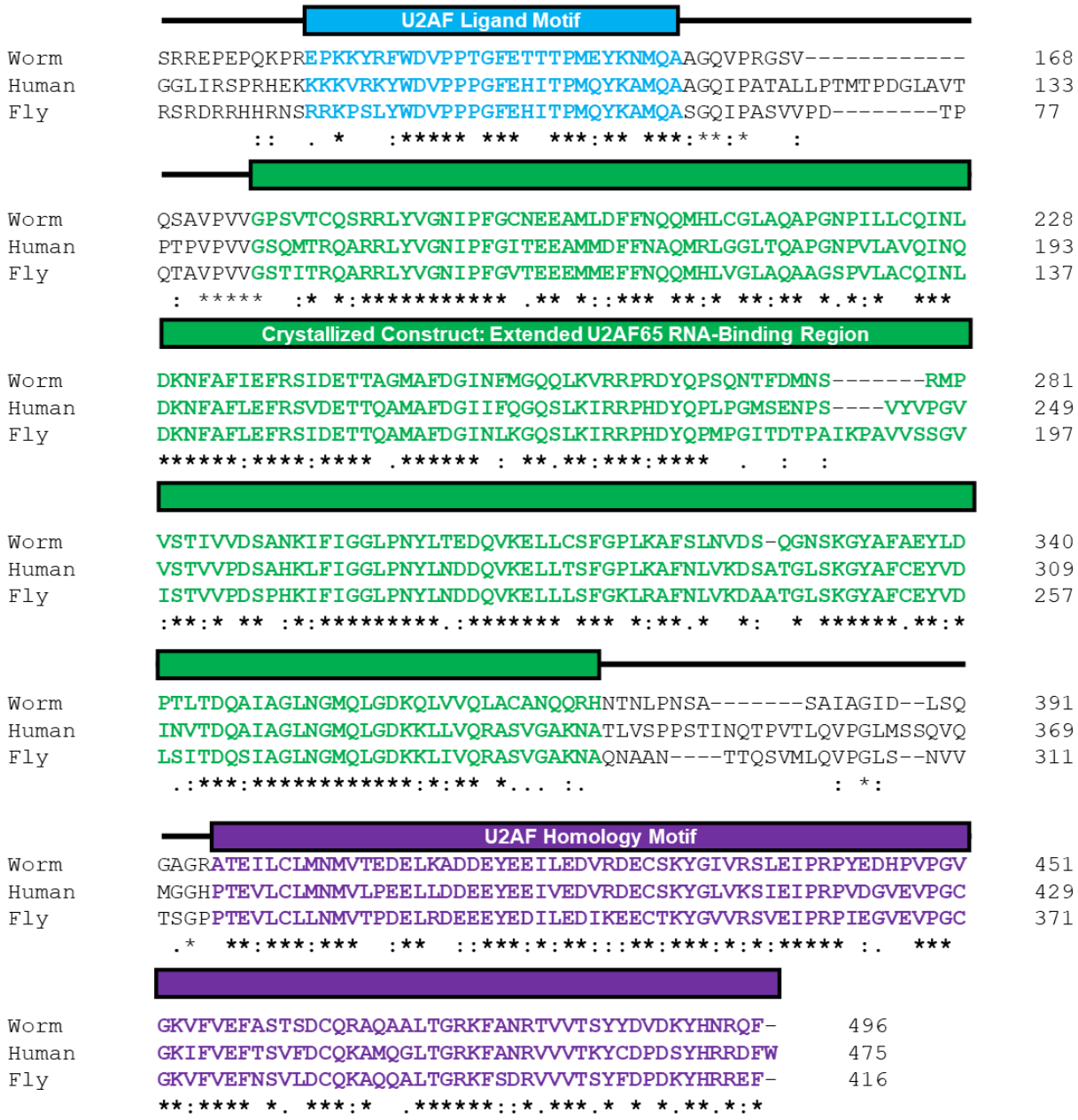


Figure *III*-1. Alignment of metazoan U2AF-L orthologues from *C. elegans*, *H. sapiens*, and *D. melanogaster*. GenBank accession numbers used in the alignment are NP_001022967.1 (*C. elegans* UAF-1, S121-F496), X64044 (*H. sapiens* U2AF65, long isoform, G74-W475), and NP_476891.1 (*D. melanogaster* U2AF50, isoform A, R26-F416). The ULM is annotated in blue and is based on the human U2AF35/U2AF65 dimer structure (Kielkopf et al., 2001). The crystallized construct for the extended U2AF65 RNA-binding region is annotated in green and is based on the four X-ray structures published in Agrawal *et al.*, 2016. The UHM is annotated in purple and is based on the annotated alignment of U2AF-L orthologues provided in Selenko *et al.*, 2003.

This alignment omits the N-terminal RS domain of all three proteins for clarity while retaining the remainder of the sequence. The RS domain has been omitted because it is poorly conserved, poorly characterized, and predicted to be unstructured. The three sequences are ~56% identical and ~77% similar with one another with respect to the total length of the aligned region.

III-1. U2AF-L chimera sequence structure (N-terminus to C-terminus): U2AF59 (S93-A161)

The chimera ULM is U2AF59 derived and is part of a rigid body consisting of full-length U2AF23 and the ULM of U2AF59 (see Table 1-8). This sequence represents the construct used to crystallize the *S. pombe* U2AF23/U2AF59 dimer and contains both the ULM and α-helix which interface with U2AF23. S93-S104 are disordered in these structures. However, truncating the chimera to E106 to mimic the U2AF59 construct in this thesis is unsuccessful, since no soluble protein is produced. These extra residues are necessary in the chimera in order to generate stable and soluble U2AF/SF1 complex.

III-2. U2AF-L chimera sequence structure (N-terminus to C-terminus): QSA

Aligning three metazoan orthologues of U2AF-L reveals high conservation of the ULM and RRM1, but considerable length and sequence divergence in the intervening sequence (Fig. *III*-1). *C. elegans* UAF-1 (U2 auxiliary factor-1) contains the shortest linker, consisting of the sequence QSAVPVV, and this linker was used in the chimera to connect the rigid bodies containing the ULM and RRM1. QSA is unique to *C. elegans*, whereas VPVV is invariant between all three species in the alignment and is addressed below.

The rigid body containing U2AF59 ends at A161, and the rigid body containing U2AF65 begins at G141. Based on the alignment of U2AF65 and U2AF59 (see Fig. 1-19), these rigid bodies are separated by a linker consisting of 21 residues in U2AF65 and 29 residues in U2AF59. This linker is neither conserved nor is it characterized, therefore it is unknown whether it contains any structure or has any role in the relative positioning of these two rigid bodies with respect to distance or orientation. Since no constraints exist to fix the relative distance or orientation of these rigid bodies in the SEC-SAXS envelopes, this linker was deleted to restrain their relative movement. However, some of the linker may be required for protein stability, solubility, and/or functionality. With respect to informing how much of the linker can be deleted and where the sequence boundaries for deletion should be, the alignment of several metazoan U2AF-L orthologues (Fig. *III*-1) was more informative than the human/*S. pombe* alignment and reveals that the *C. elegans* orthologue is missing most of the linker. Therefore, the *C. elegans* orthologue was used as the basis for designing this sequence juncture in the chimera.

III-3. U2AF-L chimera sequence structure (N-terminus to C-terminus): U2AF65 (V137-A342)

The chimera RRMs and inter-RRM linker are U2AF65 derived, and taken together as a single entity, comprise a second rigid body (see Table 1-10). V137-V140 is part of the linker connecting two rigid bodies, and G141-A342 is the protein component of the rigid body consisting of the RRMs and natural inter-RRM linker bound to oligonucleotide.

III-4. U2AF-L chimera sequence structure (N-terminus to C-terminus): U2AF59 (M394-W517)

As previously discussed, RRM2 is part of a second rigid body. C-terminal to this, the UHM of U2AF-L forms a third rigid body in a complex with the ULM of SF1 (see Table 1-11). Similar to the ULM/RRM1 juncture, a linker separates RRM2 and the UHM. It is neither conserved nor is it characterized, therefore no constraints exist to position these two rigid bodies relative to each other with respect to distance or orientation. Neither the human/*S. pombe* nor the metazoan alignment of U2AF-L provides any indication of whether this linker can be safely deleted, therefore it has been retained in the chimera.

The alignment of human U2AF65 and *S. pombe* U2AF59 shows that U2AF65 T343 corresponds to U2AF59 M394. Therefore, the final C-terminal section of the chimera consists of U2AF59 (M394-W517), which contains the linker connecting RRM2 and the UHM as well as the UHM itself; RRM2 is part of a larger rigid body as described above, and the UHM is part of the final rigid body, which also contains the ULM of SF1. The linker from U2AF59 was used to connect these rigid bodies, since it is 9 residues shorter than the U2AF65 linker (see Fig. 1-19), thereby limiting their relative movement in the modelling process since no other constraints are available.

III-5. Cloning steps to create the U2AF-L chimera expression construct

Fig. *III*-2 below summarizes the first cloning step used to generate the expression construct. Fig. *III*-3 below summarizes the second and final cloning step used to generate the expression construct.

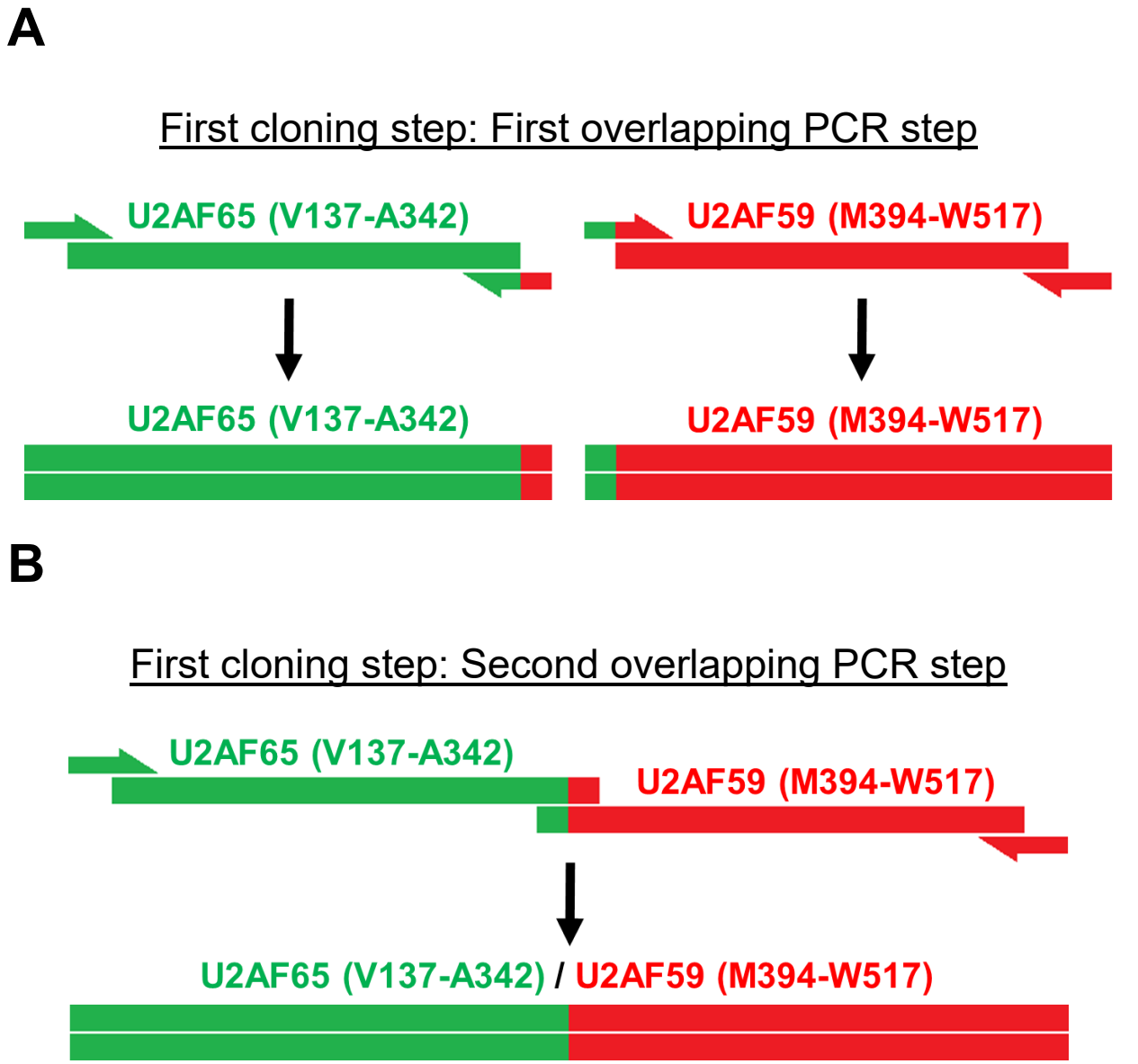


Figure *III*-2. First cloning step for the U2AF59/U2AF65 chimera. The first cloning step is to create a subclone consisting of a partial ORF in which U2AF65 (V137-A342) and U2AF59 (M394-W517) are joined into a continuous sequence using overlapping PCR primers. Sequence regions in the primers, template, and amplicon are colour-coded, with green corresponding to U2AF65 and red corresponding to U2AF59. (A) In the first overlapping PCR step, these two sections are PCR-amplified separately using the appropriate template. (B) In the second overlapping PCR step, the PCR template is a mixture of the two amplicons from the first step. The second overlapping PCR step joins the two sections into a continuous sequence.

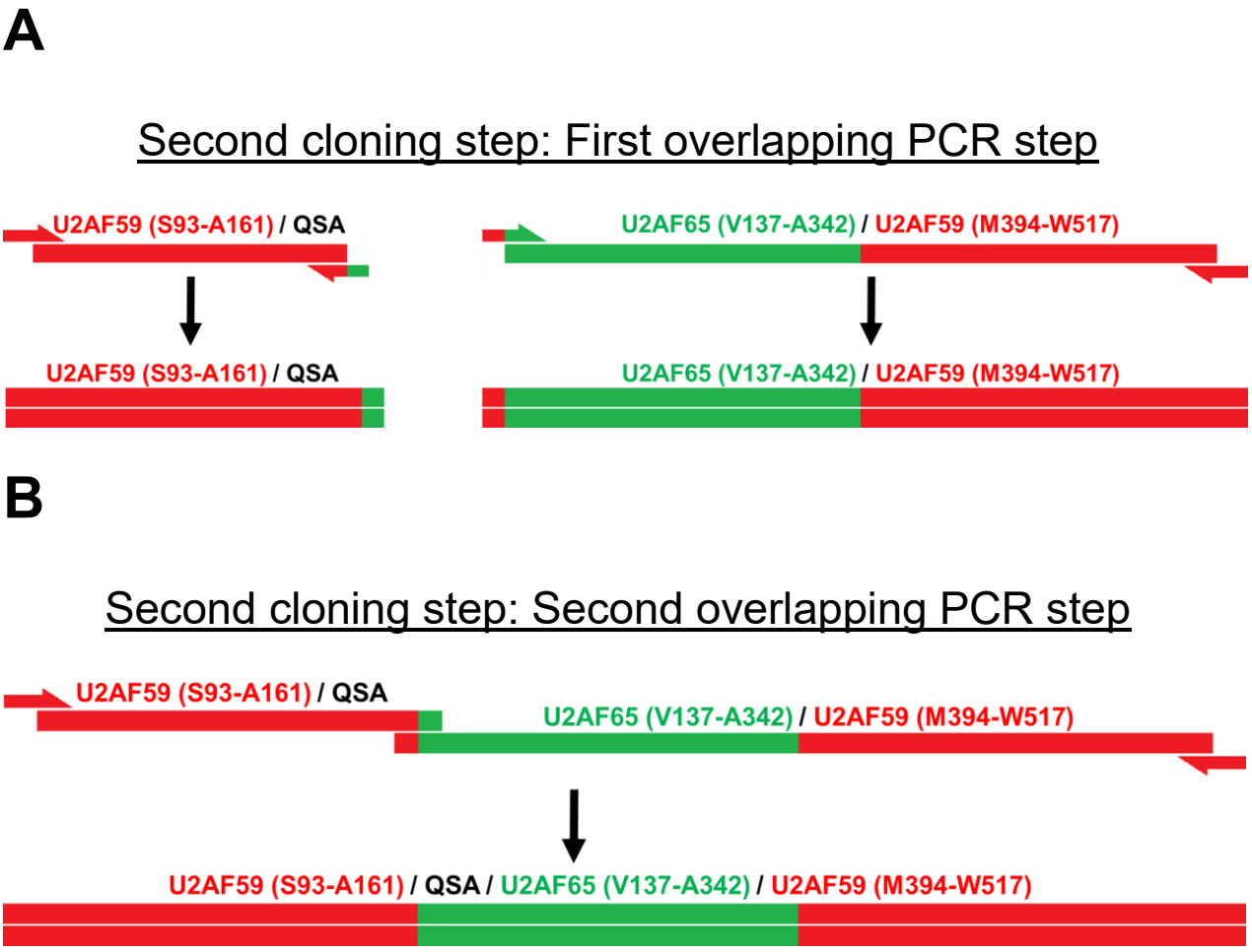


Figure *III*-3. Second cloning step for the U2AF59/U2AF65 chimera. The second cloning step is to create the final expression construct in which U2AF59 (S93-A161)/QSA and the sub-clone sequence from the first cloning step are joined into a continuous sequence using overlapping PCR primers. Sequence regions in the primers, template, and amplicon are colour-coded, with green corresponding to U2AF65 and red corresponding to U2AF59. (A) In the first overlapping PCR step, these two sections are PCR-amplified separately using the appropriate template. (B) In the second overlapping PCR step, the PCR template is a mixture of the two amplicons from the first step. The second overlapping PCR step joins the two sections into a continuous sequence.

Appendix *IV*

Rigid body definitions used to build SEC-SAXS based U2AF/SF1 model libraries

The original rigid bodies used for building apo U2AF/SF1 chimera models are summarized below in Table *IV*-1, and the original rigid bodies used for building models of both RNA-bound complexes are summarized below in Table *IV*-2. Rigid bodies 5 and 6 are the zinc knuckles of SF1, and rigid body 7 is the conserved α-helix predicted to flank the C-terminus of the zinc knuckles.

Table *IV*-1: Original rigid bodies used for models of apo U2AF/SF1 chimera

Rigid body	Structure	PDB accession code	Species
1	U2AF23 (full-length) + U2AF59 (ULM)	4YH8	*S. pombe*
2	U2AF65 (RRM1 + natural inter-RRM linker + RRM2)	6TR0	*H. sapiens*
3	SF1 (KH-QUA2 domain) + optimal BPS RNA	1K1G	*H. sapiens*
4	U2AF65 (UHM) + SF1 (ULM + phosphorylated domain, phosphorylated)	4FXW	*H. sapiens*
5	Alphafold-generated	n/a	n/a
6	Alphafold-generated	n/a	n/a
7	Alphafold-generated	n/a	n/a

Table *IV*-2: Original rigid bodies used for models of RNA-bound U2AF/SF1 chimera

Rigid body	Structure	PDB accession code	Species
1	U2AF23 (full-length) + U2AF59 (ULM) + RNA (UAGGU)	7C06	*S. pombe*
2	U2AF65 (RRM1 + natural inter-RRM linker + RRM2) + DNA/RNA hybrid	5EV1	*H. sapiens*
3	SF1 (KH-QUA2 domain) + optimal BPS RNA	1K1G	*H. sapiens*
4	U2AF65 (UHM) + SF1 (ULM + phosphorylated domain, phosphorylated)	4FXW	*H. sapiens*
5	Alphafold-generated	n/a	n/a
6	Alphafold-generated	n/a	n/a
7	Alphafold-generated	n/a	n/a

Tables *IV*-3 to *IV*-5 contain the rigid body definitions used to build models of the three complexes. In these tables, macromolecule chain names correspond to the constituent chains of the complexes as follows: PROA corresponds to *S. pombe* U2AF23, PROB corresponds to the U2AF-L chimera, PROC corresponds to *S. pombe* SF1, and RNAA corresponds to the bound 3′ SS RNA.

Table *IV*-3: Rigid body definitions for apo U2AF/SF1 chimera

```
define fixed1 sele ( resid 1:194 .and. segid PROA) end
define fixed2 sele ( resid 15:70 .and. segid PROB) end
cons fix sele fixed1 .or. fixed2 end

define rigid1 sele ( resid 80:276 .and. segid PROB) end
shape desc dock2 rigid sele rigid1 end

define rigid1 sele ( resid 136:251 .and. segid PROC) end
shape desc dock3 rigid sele rigid1 end

deine rigid1 sele ( resid 300:403 .and. segid PROB) end
define rigid2 sele ( resid 20:122 .and. segid PROC) end
shape desc dock4 rigid sele rigid1 .or. rigid2 end

define rigid1 sele ( resid 259:275 .and. segid PROC) end
shape desc dock5 rigid sele rigid1 end

define rigid1 sele ( resid 284:300 .and. segid PROC) end
shape desc dock6 rigid sele rigid1 end

define rigid1 sele ( resid 310:326 .and. segid PROC) end
shape desc dock7 rigid sele rigid1 end
```

Table *IV*-4: Rigid body definitions for U2AF/SF1 chimera + RNA CG92

```
define fixed1 sele ( resid 1:194 .and. segid PROA) end
define fixed2 sele ( resid 15:70 .and. segid PROB) end
define fixed3 sele ( resid 18:22 .and. segid RNAA) end
cons fix sele fixed1 .or. fixed2 .or. fixed3 end

define rigid1 sele ( resid 80:276 .and. segid PROB) end
define rigid2 sele ( resid 13:17 .and. segid RNAA) end
shape desc dock2 rigid sele rigid1 .or. rigid2 end

define rigid1 sele ( resid 136:251 .and. segid PROC) end
define rigid2 sele ( resid 1:7 .and. segid RNAA) end
shape desc dock3 rigid sele rigid1 .or. rigid2 end

define rigid1 sele ( resid 300:403 .and. segid PROB) end
define rigid2 sele ( resid 20:122 .and. segid PROC) end
shape desc dock4 rigid sele rigid1 .or. rigid2 end

define rigid1 sele ( resid 259:275 .and. segid PROC) end
shape desc dock5 rigid sele rigid1 end

define rigid1 sele ( resid 284:300 .and. segid PROC) end
shape desc dock6 rigid sele rigid1 end

define rigid1 sele ( resid 310:326 .and. segid PROC) end
shape desc dock7 rigid sele rigid1 end
```

Table *IV*-5: Rigid body definitions for U2AF/SF1 chimera + RNA CG109

```
define fixed1 sele ( resid 1:194 .and. segid PROA) end
define fixed2 sele ( resid 15:70 .and. segid PROB) end
define fixed3 sele ( resid 25:30 .and. segid RNAA) end
cons fix sele fixed1 .or. fixed2 .or. fixed3 end

define rigid1 sele ( resid 80:276 .and. segid PROB) end
define rigid2 sele ( resid 14:18 .and. segid RNAA) end
shape desc dock2 rigid sele rigid1 .or. rigid2 end

define rigid1 sele ( resid 136:251 .and. segid PROC) end
define rigid2 sele ( resid 1:8 .and. segid RNAA) end
shape desc dock3 rigid sele rigid1 .or. rigid2 end

define rigid1 sele ( resid 300:403 .and. segid PROB) end
define rigid2 sele ( resid 20:122 .and. segid PROC) end
shape desc dock4 rigid sele rigid1 .or. rigid2 end

define rigid1 sele ( resid 259:275 .and. segid PROC) end
shape desc dock5 rigid sele rigid1 end

define rigid1 sele ( resid 284:300 .and. segid PROC) end
shape desc dock6 rigid sele rigid1 end

define rigid1 sele ( resid 310:326 .and. segid PROC) end
shape desc dock7 rigid sele rigid1 end
```

Appendix *V*

Structural characterization of the *S. pombe* and *C. albicans* p14/SF3B155 complexes[6]

[6] Crystallographic data used to solve the X-ray structure of the *C. albicans* p14/SF3B155 complex presented in Appendix *V* was collected by Dr. Muhammad Bashir Khan (Research Associate, Dept. of Biochemistry, University of Alberta, Edmonton, AB, Canada). The X-ray structures of both the *S. pombe* and *C. albicans* p14/SF3B155 complexes presented in Appendix *V* were solved as a collaboration with Dr. Ross A. Edwards (Research Associate, Dept. of Biochemistry, University of Alberta, Edmonton, AB, Canada).

V-1. Introduction

During the E → A complex transition of the spliceosome, SF1 is displaced by the protein called SF3B155, which is the largest subunit and central scaffold protein of the SF3B particle; SF3B is itself a subcomponent of U2 snRNP (Gozani et al., 1998; Will et al., 2001). The N-terminus of SF3B155 contains several ULMs and potentially operates as a platform for A complex assembly by binding multiple UHM-containing proteins (Thickman, Swenson, Kabogo, Gryczynski, & Kielkopf, 2006). As discussed previously in this thesis, the UHM of U2AF-L interfaces with the ULM of SF1; since SF3B155 also directly interacts with U2AF-L in the A complex in both humans and *S. pombe* via its ULM, the U2AF-L/SF1 and U2AF-L/SF3B155 interactions are mutually exclusive. The stable association of U2 snRNP with the pre-mRNA generates a duplex between U2 snRNA and the pre-mRNA branch region, thereby specifying the branch A as the nucleophile for the first step of splicing by bulging it out (Query et al., 1994).

Although the BPS is critical to splicing it is highly degenerate in metazoans, therefore specific mechanisms must exist to target U2 snRNA to the BPS to ensure fidelity in forming the bulged duplex to specify the branch A as the nucleophile for the first step of splicing. Since U2AF binding precedes SF3B155 (and U2 snRNP) binding in the splicing cycle, U2AF recruits U2 snRNP to the BPS through direct interactions with SF3B155 (Gozani et al., 1998). It has been proposed that in metazoans, direct interaction between the RS domain of U2AF65 and the BPS may achieve annealing of U2 snRNA to the BPS, and other SF3A and SF3B subunits anchor U2 snRNP tightly to the pre-mRNA (Gozani et al., 1996; Valcarcel et al., 1996).

P14 is a protein that strongly interacts with SF3B155 and is the only protein that directly interacts with the branch A in the fully assembled spliceosome; p14 is a constituent of U2 snRNP and U12 snRNP, a component of the minor spliceosome responsible for splicing a relatively rare

subset of introns (Query et al., 1996; Will et al., 2001). Although the specific role of p14 in the splicing cycle is still unknown, it may function to sequester the branch A until the active site of the spliceosome has fully formed.

Experiments using a synthetic pre-mRNA substrate in which the photo-activatable crosslinker benzophenone is tethered to the branch A revealed a strong crosslink to p14 that appeared in complex A and persisted within the fully assembled spliceosome; subsequent experiments showed that p14 crosslinked directly to the branch A in complexes A through C indicating an intimate association between p14 and the pre-mRNA at the catalytic core of the mammalian spliceosome (MacMillan et al., 1994; Query et al., 1996). Because p14 displaces SF1 from the BPS early in the spliceosome cycle, a unified understanding of how 3′ SS recognition by U2AF/SF1 is integrated into the spliceosome cycle requires that the functions of p14 are defined. Additionally, since the intimate association between p14 and the branch A persists into the fully assembled spliceosome, p14 may contribute to the architecture of the active site.

P14 is a highly conserved protein. It is partially unfolded and insoluble in isolation, therefore solution of the X-ray structure of human p14 required forming a complex with a peptide of SF3B155; this complex is very stable and in contrast to isolated p14 it is soluble and tractable for X-ray crystallography and other biochemical investigations. The X-ray structure reveals that p14 consists of an RRM appended at the C-terminus by two successive α-helices. In this structure, the bound SF3B155 peptide consists of a long N-terminal α-helix and an interfacial region that interacts with p14 (Schellenberg et al., 2006). Subsequently the NMR structure of a human p14/SF3B155 complex was also reported with the same model for the folded core of the complex (Kuwasako et al., 2008).

A follow up study to the initial apo X-ray structure reported a very similar X-ray structure of the human p14/SF3B155 complex with adenine bound in a pocket on the RRM of p14, which stacks onto a highly conserved aromatic residue in RNP2 (Schellenberg, Dul, & MacMillan, 2011). Together, the X-ray structures reported in these two studies in combination with supporting experiments reveal specific recognition of the branch A by p14 and establish the orientation of the bulged duplex on the surface of p14, revealing that the p14•duplex interaction must be disrupted prior to the first step of splicing.

As discussed previously in this thesis, *S. cerevisiae* lacks orthologues for both U2AF-S and U2AF-L, but it additionally also lacks any identifiable p14 orthologue (Dziembowski et al., 2004). In contrast *S. pombe* has a p14 orthologue. Interestingly, the commensal fungus *Candida albicans* also has clear orthologues for U2AF-S, U2AF-L, and p14, but in contrast to human and *S. pombe* p14 the aromatic residue that stacks with the branch A is not conserved and corresponds to a leucine. Therefore, p14/SF3B155 complexes from *S. pombe* and *C. albicans* were cloned and co-expressed, purified, crystallized, and their X-ray structures were solved. This was completed in order to gain additional insight into p14 by comparing the counterpart structures from human, *S. pombe*, and *C. albicans*. Additionally, it is important to establish *S. pombe* and *C albicans* as viable alternative model systems to study this apparatus so that they can be used in the future to gain mechanistic insights into p14 since *S. pombe* is a much more simple and tractable system than humans and understanding branch A recognition by the more divergent orthologue present in *C. albicans* may reveal alternate mechanisms of branch A recognition in different organisms.

V-2. Results

V-2.1. Identification of p14 and SF3B155 orthologues in *S. pombe* and *C. albicans*

V-2.1.1. Sequence alignment-based orthologue comparison and construct design of p14

A comparison of the human, *S. pombe* and *C. albicans* orthologues of p14 via Clustal Omega alignment is provided below (Madeira et al., 2019; Sievers et al., 2011). Key features in this alignment are annotated and described following the alignment.

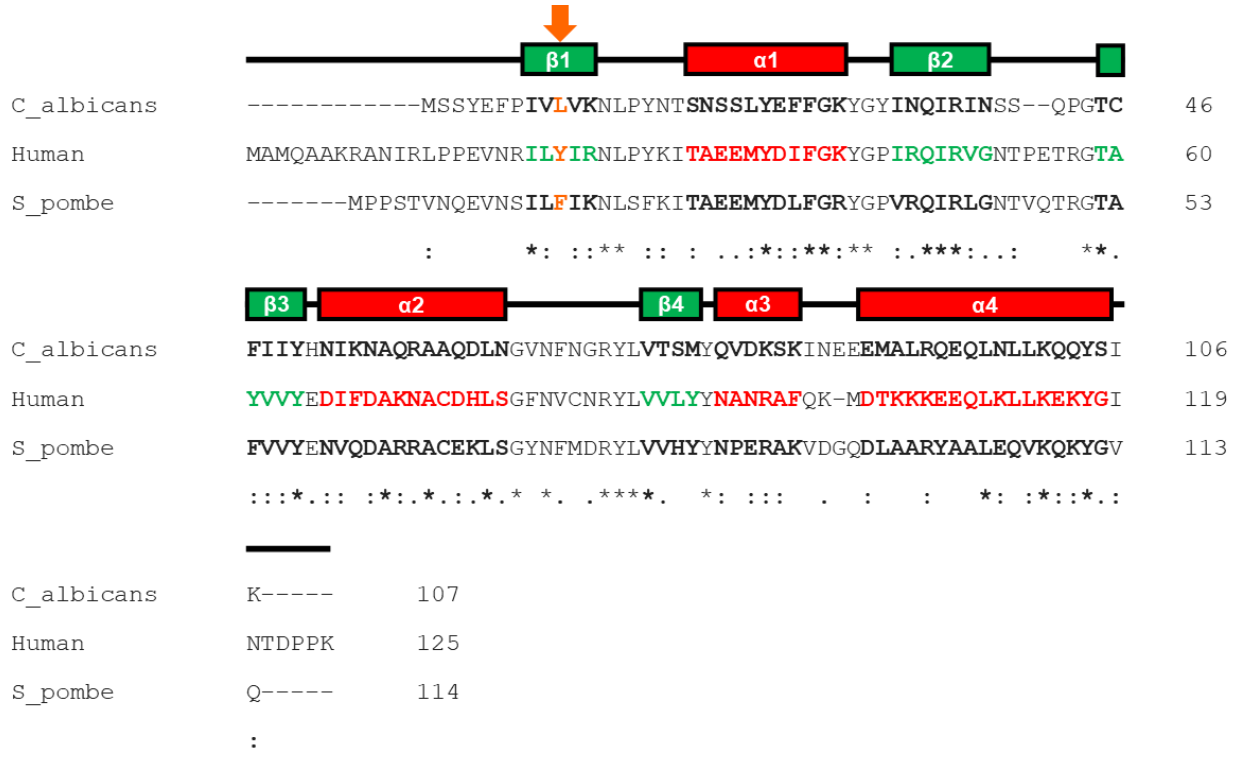


Figure *V*-1. P14 alignment (human vs. *S. pombe* vs. *C. albicans*). In this figure, the secondary structure of human p14 is annotated based on the apo p14/SF3B155 X-ray structure (PDB accession 2F9D, Chain A); α-helices are annotated in red, and β-strands are annotated in green (Schellenberg et al., 2006). The conserved aromatic residue that stacks with the branch A is annotated in orange and indicated with an arrow (human = Y22, *S. pombe* = F15, *C. albicans* = L10).

In the alignment in Fig. *V*-1, the full-length sequences for human, *S. pombe*, and *C. albicans* p14 orthologues have been aligned, because all the X-ray structures were derived from either a full-length construct (human and *S. pombe*), or a nearly full-length construct (*C. albicans*: S2-K107). *S. pombe* p14 was N-terminally tagged with a TEV-cleavable hexahistidine tag and *C. albicans* p14 was N-terminally tagged with a non-cleavable hexahistidine tag.

For the original construct design *C. albicans* p14 was N-terminally tagged with a TEV-cleavable hexahistidine tag, however when the *C. albicans* p14/SF3B155 complex was purified under the same conditions as the *S. pombe* p14/SF3B155 complex it was prone to precipitation. In order to rectify this problem, the NaCl concentration of all buffers was increased, urea was added to all buffers (except the final sample buffer used in crystallization), and the TEV cleavage and ion exchange chromatography purification steps were eliminated. Since the TEV cleavage step was eliminated, the cleavage site was deleted in the final construct used for crystallization in order to shorten the disordered N-terminus of the *C. albicans* p14 construct, which may interfere with crystallization.

The human and *S. pombe* sequences are ~55% identical and ~79% similar over the aligned region, the human and *C. albicans* sequences are ~35% identical and ~80% similar over the aligned region, and the *S. pombe* and *C. albicans* sequences are ~34% identical and ~80% similar over the aligned region. Together these results reveal that the human and *S. pombe* orthologues of p14 share roughly the same degree of conservation with *C. albicans* p14, and that human and *S. pombe* p14 are much more highly conserved with each other than either one is to *C. albicans* p14.

V-2.1.2. Sequence alignment-based orthologue comparison and construct design of SF3B155

SF3B155 is a much larger and more complex protein than p14. An organizational diagram has been provided below for context before describing the individual features of SF3B155 in more detail. This diagram is based on a Clustal Omega alignment of the full-length orthologues of SF3B155 from human, *S. pombe*, and *C. albicans* (Madeira et al., 2019; Sievers et al., 2011). Because of the length and complexity of SF3B155, the complete alignment of the full-length sequences is omitted from this section for clarity and instead two alignments of the region corresponding to the interface with p14 are shown (Fig. *V*-3 and *V*-4).

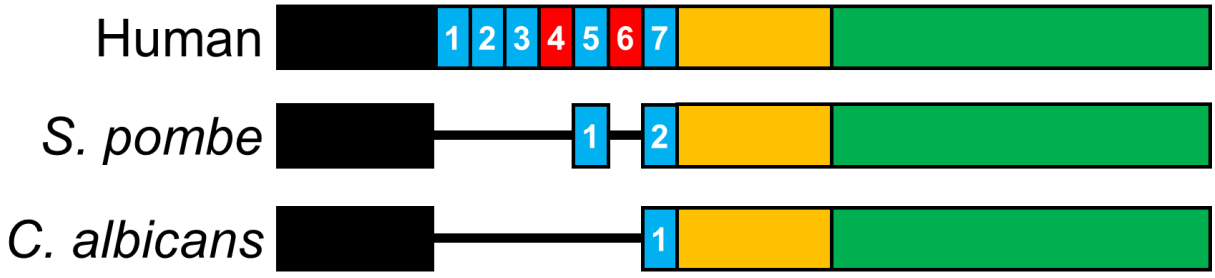


Figure *V*-2. Domain organization of SF3B155 (human vs. *S. pombe* vs. *C. albicans*). Protein domains are arranged from the N-terminus to C-terminus of each orthologue (left to right). The N-terminus of SF3B155 begins with a conserved region of unknown function (black), followed by 1-7 tandem ULMs that are either functional (blue) or non-functional (red). The tandem ULMs are followed by the region that interfaces with p14 (yellow), followed by a solenoid (green).

A Clustal Omega alignment of the human, *S. pombe*, and *C. albicans* orthologues of SF3B155 reveals that the N-terminus is partially conserved however, this region is uncharacterized. This is followed by a variable number of tandem ULMs, depending on the species. Human SF3B155 contains 7 tandem ULMs; the 4th and 6th ULMs of human SF3B155 do not contain the 'lock-and-key' tryptophan and are non-functional (Thickman et al., 2006). *S.*

pombe SF3B155 contains two functional ULMs; in a Clustal Omega alignment *S. pombe* ULM 1 corresponds to human ULM 5, and *S. pombe* ULM 2 corresponds to human ULM 7. *C. albicans* SF3B155 only contains one functional ULM corresponding to *S. pombe* ULM 2 and human ULM 7 in a Clustal Omega alignment.

The tandem ULMs are followed by the region that interfaces with p14. This is followed by a large and highly conserved solenoid that serves as the structural scaffold and binding platform for the other proteins forming the SF3B particle (Cretu et al., 2016). In order to determine suitable co-expression constructs for *S. pombe* and *C. albicans* SF3B155 the full-length orthologues of human, *S. pombe*, and *C. albicans* SF3B155 were aligned, and region from this alignment corresponding to the human interfacial region with p14 is shown annotated below (Fig. *V*-3).

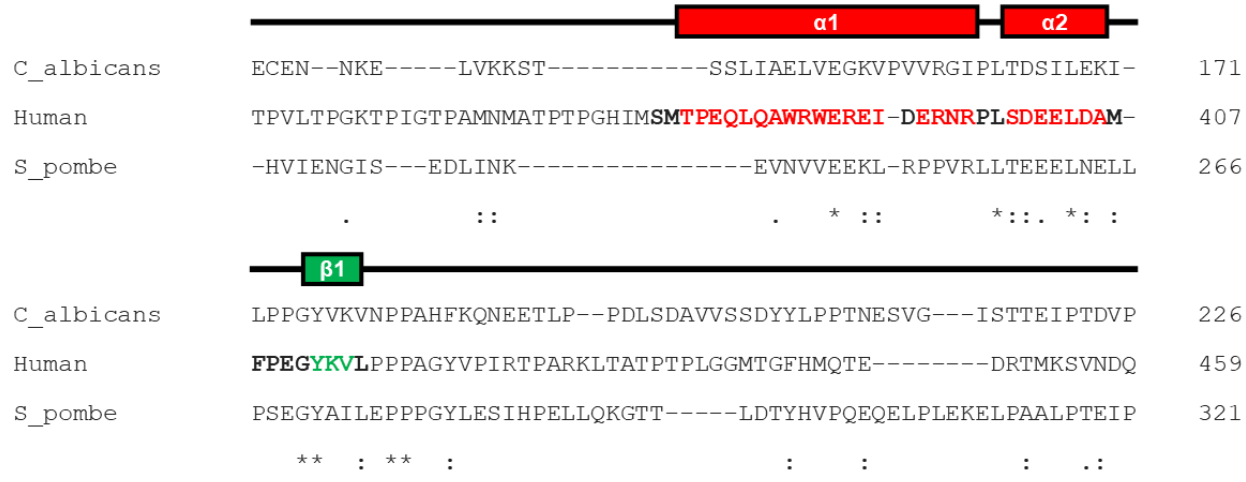


Figure *V*-3. SF3B155 alignment (human vs. *S. pombe* vs. *C. albicans*). The secondary structure of human SF3B155 is annotated based on the apo p14/SF3B155 X-ray structure (PDB accession 2F9D, Chain P); modeled residues are bolded, α-helices are annotated in red, and β-strands are annotated in green (Schellenberg et al., 2006).

When the full-length orthologues of SF3B155 are aligned, the p14 interaction region of human SF3B155 shows poor homology to the *S. pombe* and *C. albicans* orthologues. Additionally there are significant gaps in the alignment for the region surrounding the p14 interaction region of human SF3B155. This makes it difficult to logically determine ideal construct boundaries for the *S. pombe* and *C. albicans* SF3B155 co-expression constructs. Therefore, a more limited and focused Clustal Omega alignment was completed. This alignment was based on the human p14 interaction region and included a roughly equal amount of sequence from each species surrounding this region in order to force the sequences to align without large gaps. Additionally, all three SF3B155 sequences used for the alignment were run through the PsiPred secondary structure prediction server. Combined with the limited alignment, this provided much more clarity to construct design (see Fig. *V*-4 below).

Figure *V*-4. Limited alignment and PsiPred-based secondary structure prediction of SF3B155 (human vs. *S. pombe* vs. *C. albicans*). The secondary structure of human SF3B155 is labelled above the alignment and annotated in cyan based on the apo p14/SF3B155 X-ray structure (PDB accession 2F9D, Chain P) (Schellenberg et al., 2006). The PsiPred-based secondary structure prediction of human, *S. pombe*, and *C. albicans* SF3B155 is annotated in red (α-helices) and green (β-strands). A stretch of 8 residues that is highly conserved between human and *S. pombe* SF3B155, but which was excluded from the construct used to derive the human X-ray structure is annotated in grey.

Based on the limited alignment and PsiPred-based secondary structure prediction, *S. pombe* SF3B155 (N244-I284) was chosen as the co-expression construct. The N-terminal boundary of this construct includes part of the predicted α-helix outside of the p14/SF3B155

interface corresponding to α1 of the human X-ray structure but does not include the full predicted α-helix. This is because α1 in human SF3B155 does not contribute to the human p14/SF3B155 interface, nor does it play any known role in the RNA binding functions of p14. Furthermore, α1 is a long and possibly flexible protrusion from the otherwise mostly globular p14/SF3B155 complex which must be incorporated into the crystal lattice. Therefore, it is expected that shortening this predicted α-helix in *S. pombe* SF3B155 will potentially produce a complex that crystallizes more readily. The C-terminal boundary of this construct includes more sequence than the corresponding construct used to crystallize the human p14/SF3B155 complex in order to include a stretch of 8 residues that is highly conserved between human and *S. pombe* p14. Additionally, the presence of three tandem prolines in this region for both human and *S. pombe* SF3B155 suggests the presence of a conserved α-helix.

Based on the limited alignment and PsiPred-based secondary structure prediction, *C. albicans* SF3B155 (S144-A183) was chosen as the co-expression construct. The N-terminal boundary of this construct begins at the beginning of the predicted α-helix outside of the p14/SF3B155 interface and no attempt was made to shorten this potential α-helix because the secondary structure prediction for *C. albicans* SF3B155 in the region corresponding to human α1 does not indicate an obvious long α-helix as the *S. pombe* secondary structure prediction does. The C-terminal boundary of this construct includes the predicted β-strand corresponding to β1 in the human X-ray structure.

V-2.2. Cloning, co-expression, and purification of the *S. pombe* and *C. albicans* p14/SF3B155 complexes

Briefly, for both the *S. pombe* and *C. albicans* complexes, p14 and SF3B155 were both cloned into the same pET Duet-1 co-expression vector, which was then subsequently transformed into the *E. coli* expression host and co-expressed. For the *S. pombe* complex co-expression was followed by nickel affinity chromatography purification, followed by SEC purification, followed by TEV cleavage of the N-terminal hexahistidine affinity tag on *S. pombe* p14, followed by anion exchange chromatography purification, and completion of the purification protocol with a final SEC purification. For the *C. albicans* complex co-expression was followed by nickel affinity chromatography purification, followed by completion of the purification protocol with SEC purification.

V-2.2.1. Purification of the *S. pombe* p14/SF3B155 complex

Fractions eluted from the Ni-NTA column were analyzed via SDS-PAGE and coomassie blue staining (Fig. *V*-5). When Ni-NTA purified protein is separated over the S75 column, a UV trace is generated, showing a main peak with a shoulder on the right side, although these are indistinguishable from each other when analyzed via SDS-PAGE and coomassie blue staining (Fig. *V*-6).

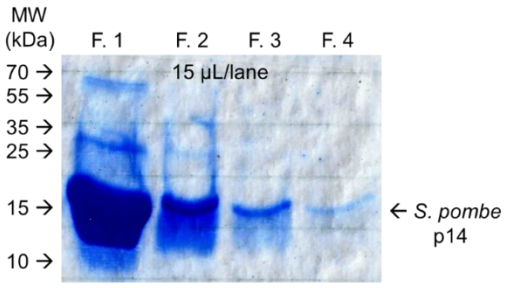


Figure *V*-5. Ni-NTA purification of the *S. pombe* p14/SF3B155 complex. SDS-PAGE gel shows the four fractions eluted.

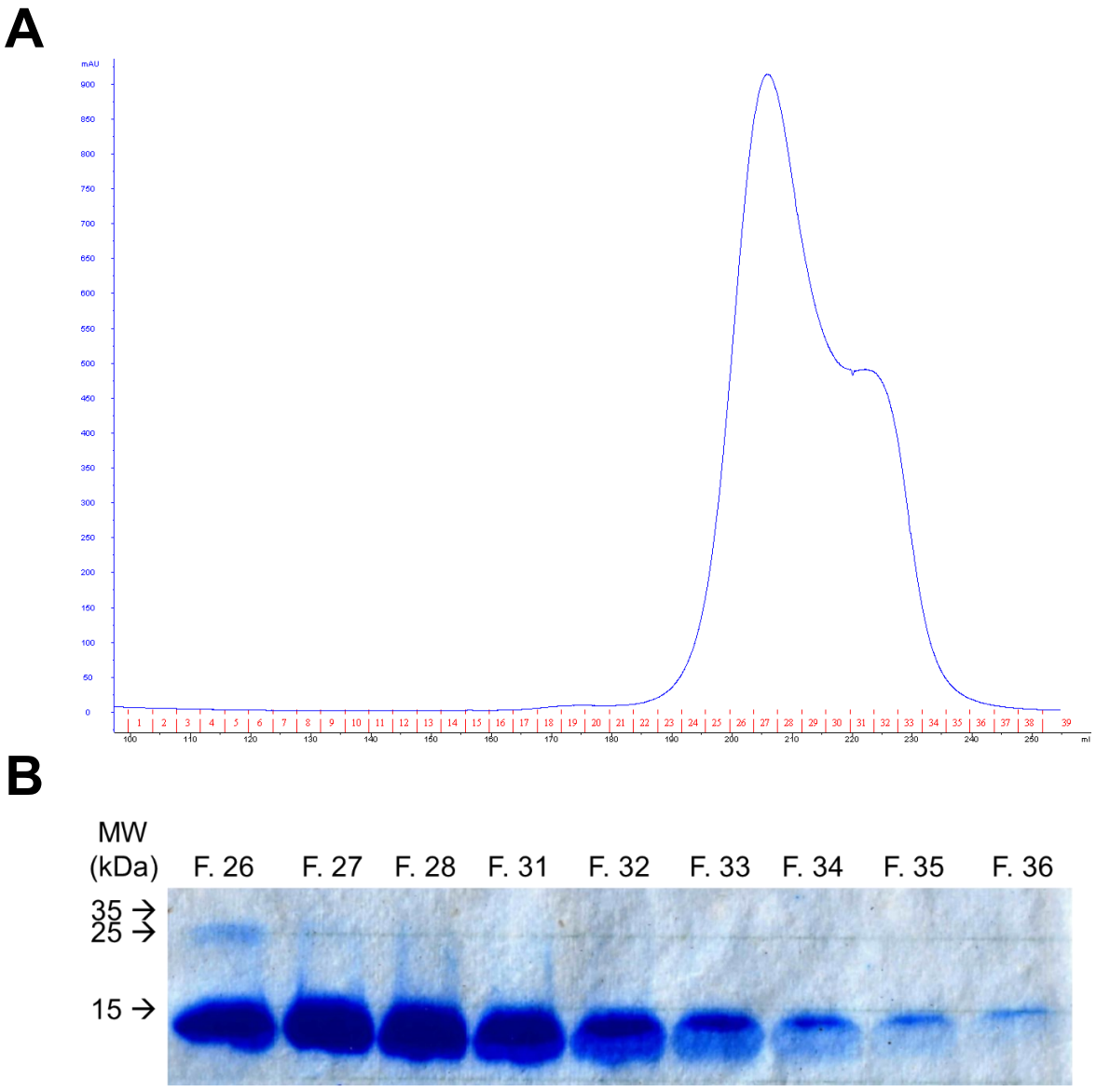


Figure *V*-6. SEC purification of the *S. pombe* p14/SF3B155 complex. (A) S75 UV trace generated by the input sample (pooled and concentrated Ni-NTA elutions). (B) SDS-PAGE gel of S75 fractions.

S. pombe SF3B155 is not observable via SDS-PAGE, though the sample behaviour is consistent with the presence of SF3B155, and it is present in the *S. pombe* X-ray structure. Following SEC purification *S. pombe* p14 was TEV cleaved. SDS-PAGE was used to confirm complete cleavage of the affinity tag (Fig. *V*-7). There is a distinct and obvious, but subtle size difference between uncleaved and cleaved p14. Following TEV cleavage, the *S. pombe* complex was purified using anion exchange chromatography over a Mono Q HR 5/5 column in order to ensure that the final sample was free of all contamination detectable by SDS-PAGE, SEC, or anion exchange chromatography. The resulting UV trace showed a sharp peak containing the complex that was subsequently analyzed via SDS-PAGE and coomassie blue staining (Fig. *V*-8).

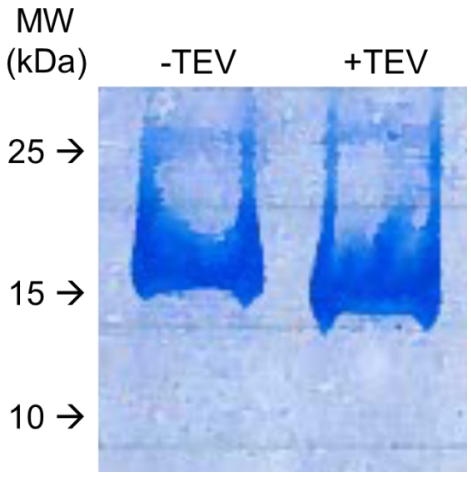


Figure *V*-7. Size difference between uncleaved and cleaved *S. pombe* p14.

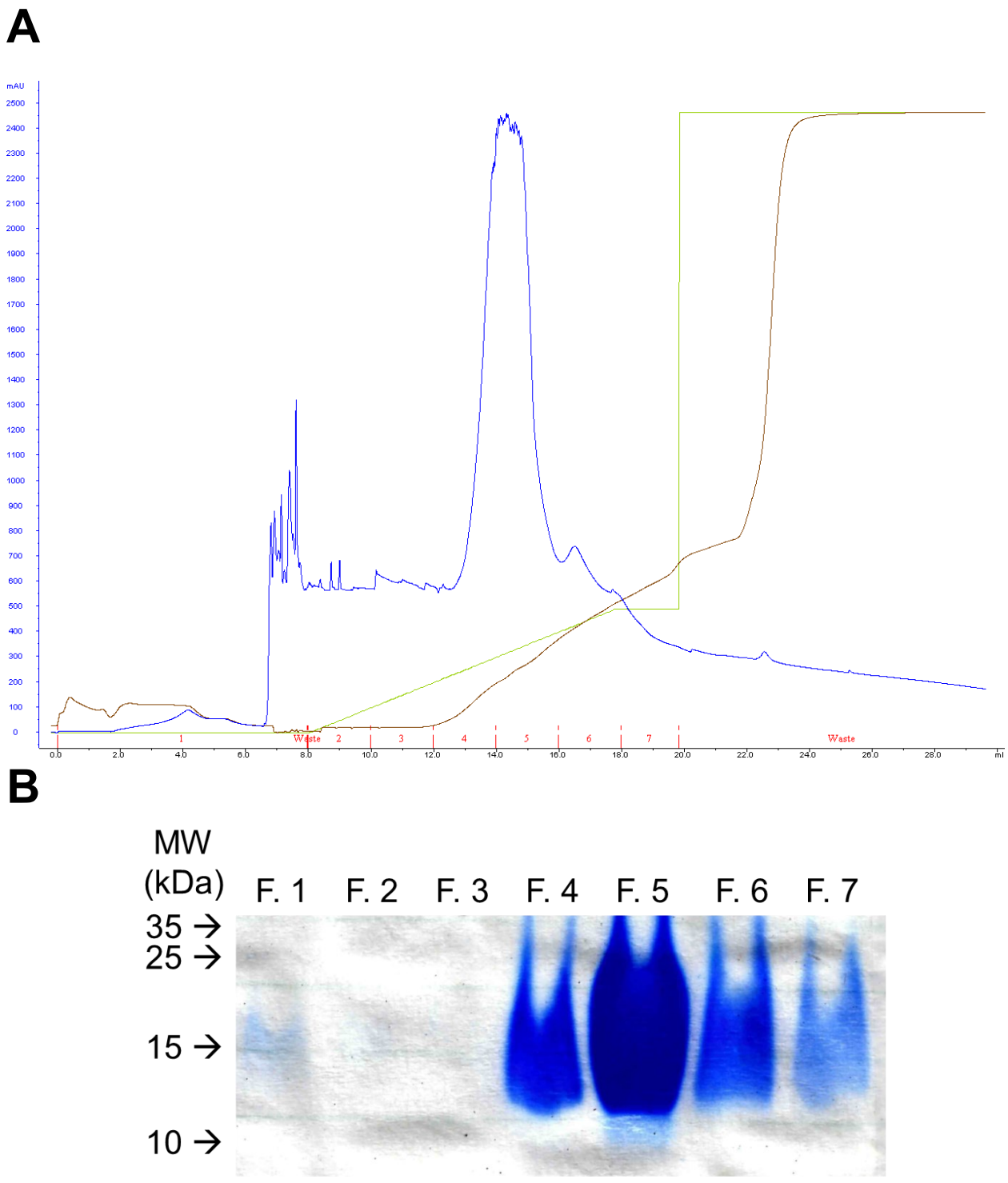


Figure *V*-8. Anion exchange chromatography purification of the *S. pombe* p14/SF3B155 complex. (A) Mono Q UV trace generated by the input sample showing fractions overlaid with the NaCl concentration (green line) and buffer conductivity (brown line). (B) SDS-PAGE gel of fractions.

The final SEC purification following anion exchange chromatography has been omitted for simplicity; both the UV trace and corresponding SDS-PAGE gel of the eluted S75 fractions is very similar to the first SEC purification (Fig. *V*-6).

V-2.2.2. Purification of the *C. albicans* p14/SF3B155 complex

Fractions eluted from the Ni-NTA column were analyzed via SDS-PAGE and coomassie blue staining (Fig. *V*-9). When Ni-NTA purified protein is separated over the S75 column, a UV trace is generated, showing a symmetrical peak for the *C. albicans* p14/SF3B155 complex, which was subsequently analyzed via SDS-PAGE and coomassie blue staining (Fig. *V*-10).

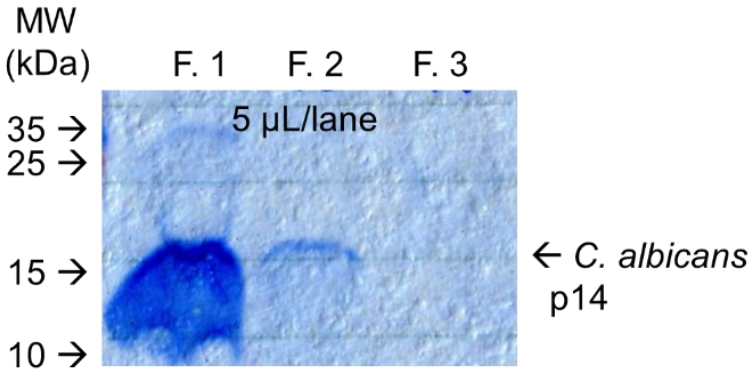


Figure *V*-9. Ni-NTA purification of the *C. albicans* p14/SF3B155 complex. SDS-PAGE gel shows the three fractions eluted.

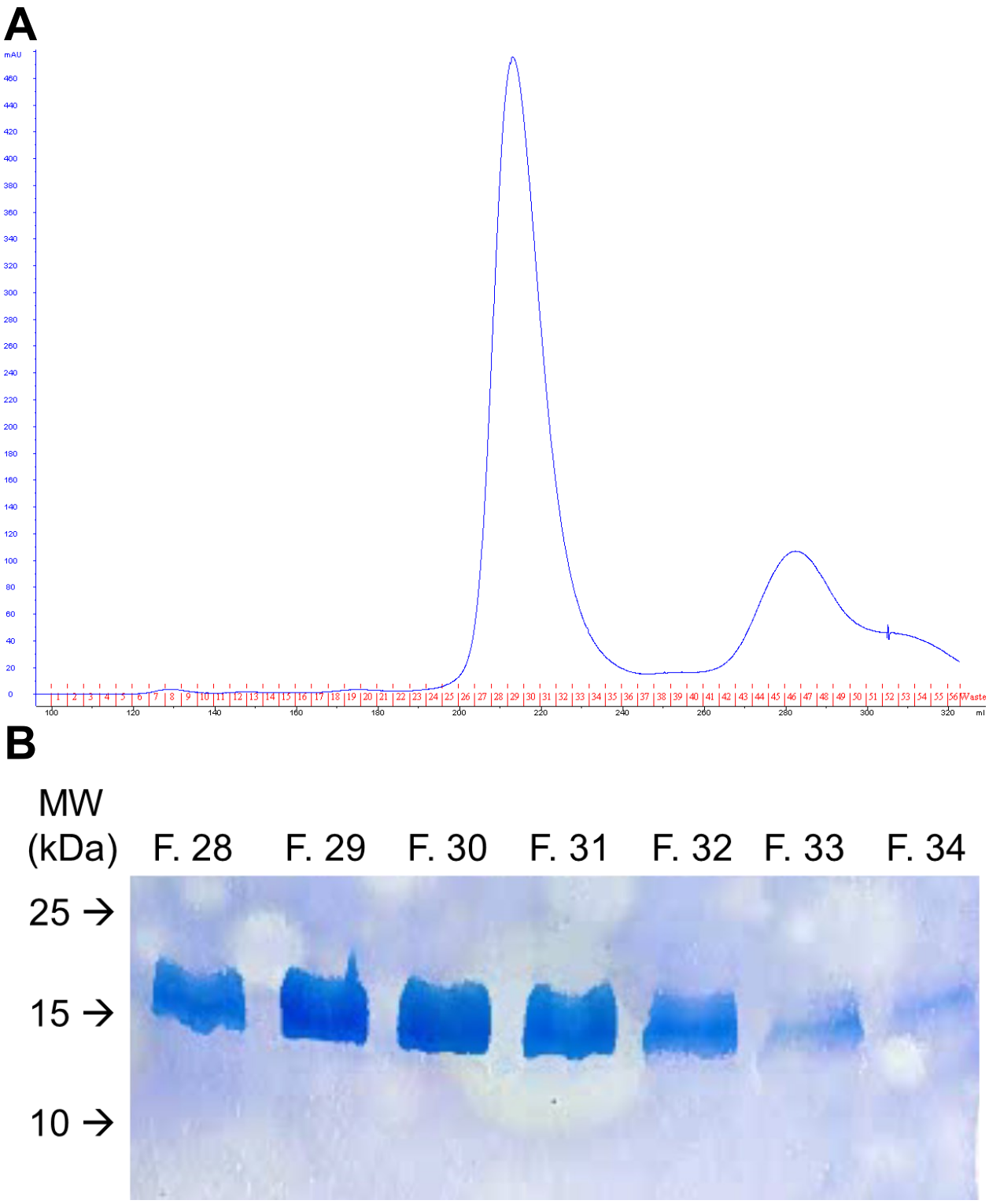


Figure *V*-10. SEC purification of the *C. albicans* p14/SF3B155 complex. (A) S75 UV trace generated by the input sample (pooled and concentrated Ni-NTA elutions). (B) SDS-PAGE gel of S75 fractions.

As with the *S. pombe* p14/SF3B155 complex, *C. albicans* SF3B155 is not observable via SDS-PAGE, though the sample behaviour is consistent with the presence of SF3B155, and it is present in the *C. albicans* X-ray structure.

V-2.3. Crystallization and structure solution of the *S. pombe* and *C. albicans* p14/SF3B155 complexes

V-2.3.1. Screening and optimization of crystals of the *S. pombe* p14/SF3B155 complex

From the crystallization screens one hit was identified producing large three-dimensional crystals after one week of incubation (JCSG Core IV condition 41: 1.5 M $LiSO_4$, 100 mM HEPES (free acid), pH 7.5; Fig. *V*-11).

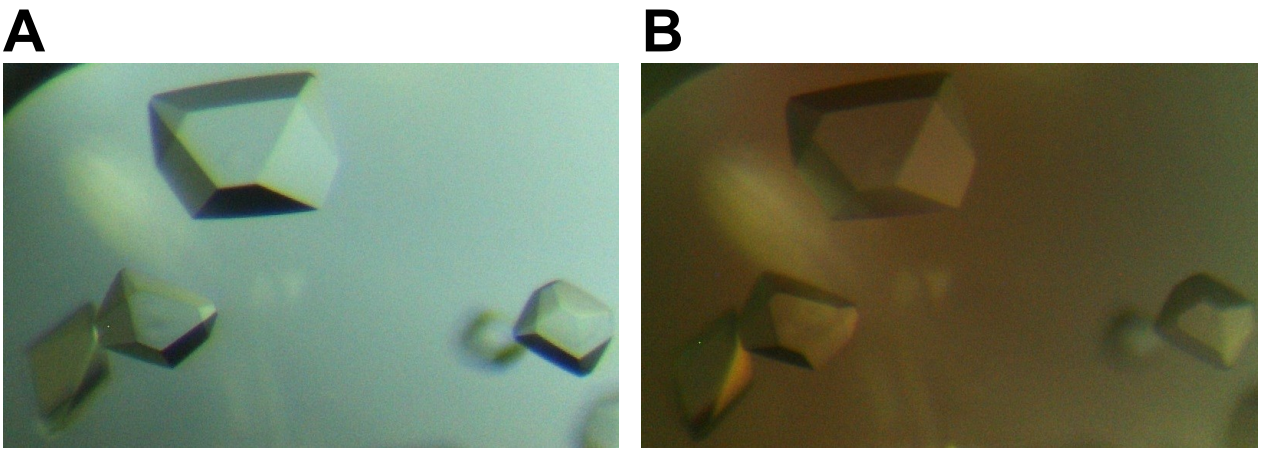


Figure *V*-11. *S. pombe* p14/SF3B155 crystallization hit (JCSG Core IV condition 41). (A) Crystals from the screening tray. (B) Same as (A) but viewed under a polarizer.

These crystals were suitable for data collection without further optimization however, they were destroyed while developing of the cryoprotection protocol and attempts to reproduce the crystallization hit using conventional methods were unsuccessful. A commercially available additive screen is available containing a library 96 unique reagents and excipients that can affect the solubility and crystallization of both soluble and membrane proteins by perturbing and manipulating sample-sample and sample-solvent interactions, as well as by perturbing water structure (Additive Screen HT, Hampton Research, Catalogue number: HR2-138). This can alter and improve both solubility and crystallization of a sample and numerous reports exist for the use of additives to improve the quality and size of macromolecular crystals (Cudney, Patel, Weisgraber, Newhouse, & McPherson, 1994; Ducruix & Giegé, 1992; H. Michel, 1991; Sousa, 1995; Trakhanov & Quiocho, 1995). When a crystallization tray was prepared based on the hit condition using this additive screen several additives successfully reproduced the crystals. Out of these, NaI was chosen for reproducing the crystals for data collection (Fig. *V*-12).

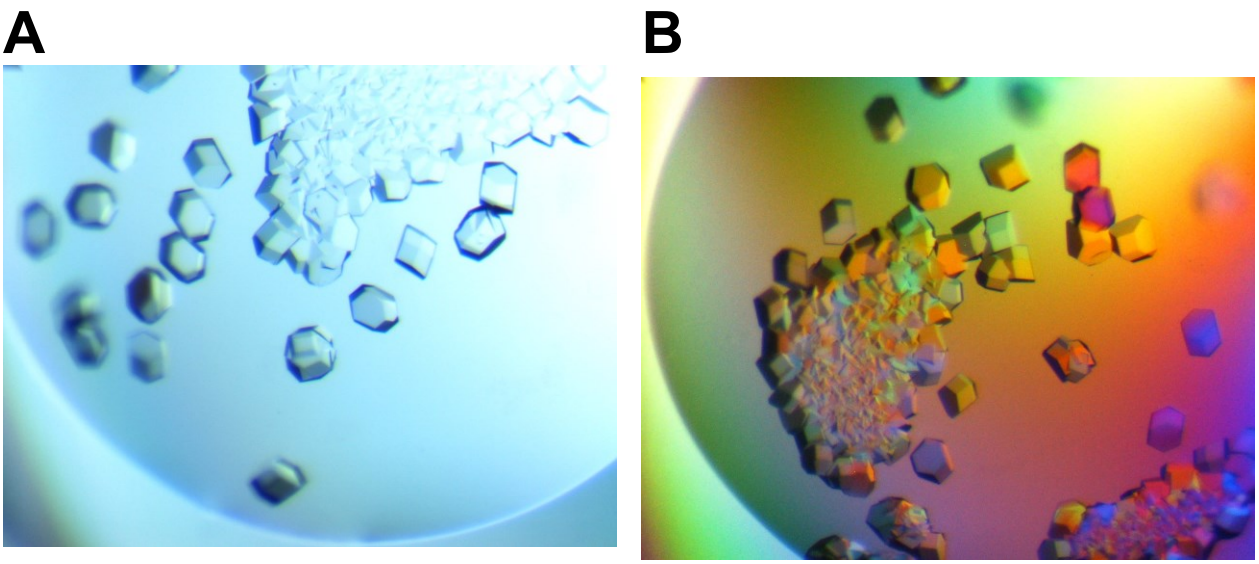


Figure *V*-12. *S. pombe* p14/SF3B155 crystals grown for data collection using NaI as an additive. (A) Representative crystals. (B) Representative crystals viewed under a polarizer.

V-2.3.2. Screening and optimization of crystals of the *C. albicans* p14/SF3B155 complex

From the crystallization screens multiple potential hits were identified producing crystals that grew in three distinct habits, each of which potentially represents a unique crystal form of the *C. albicans* p14/SF3B155 complex.

Twenty-five hits were identified producing needles growing as both individual crystals as well as needle clusters after several days of incubation. These crystals require significant optimization before they can be used for data collection. The condition producing the largest and most uniform crystals from these hits corresponded to JCSG Core II condition 52 (1.0 M LiCl, 100 mM MES (free acid), pH 6.0, 10% (w/v) PEG 6000); Fig. *V*-13).

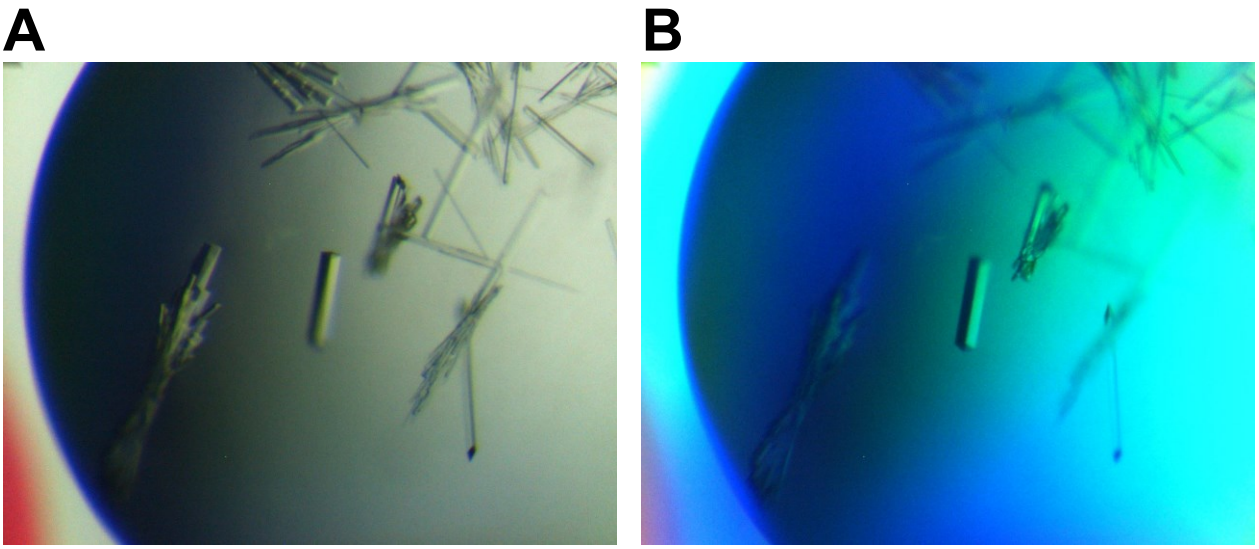


Figure *V*-13. Potential needles of *C. albicans* p14/SF3B155. Crystals formed in JCSG Core II condition 52 and were photographed several days after preparing the screening tray. (A) Crystals from the screening tray. (B) Same as (A) but viewed under a polarizer.

One hit was identified producing individual rods after four weeks of incubation. These crystals are likely suitable for data collection without further optimization. The hit condition corresponded to JCSG Core III condition 80 (100 mM Citric acid, pH 4.0, 30% (w/v) PEG 6000; Fig *V*-14).

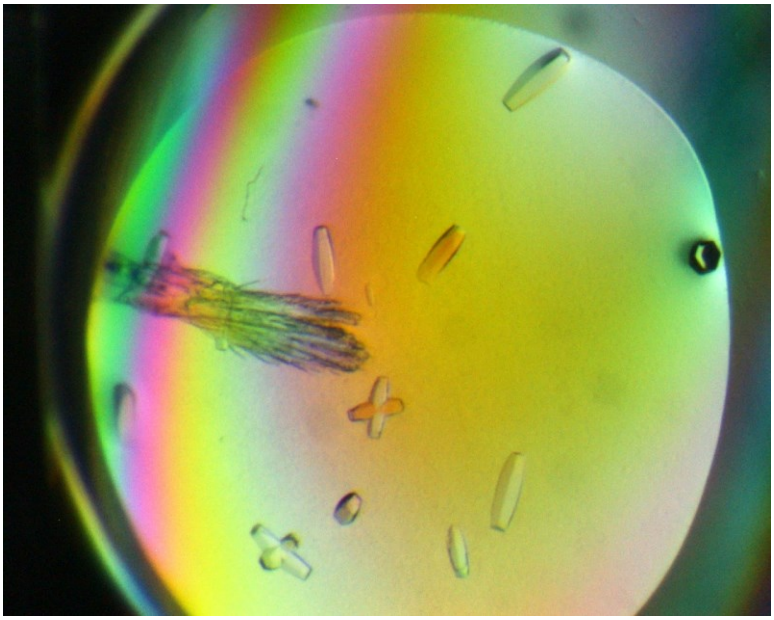


Figure *V*-14. Potential rods of *C. albicans* p14/SF3B155. Crystals formed in JCSG Core III condition 80 and were photographed 4 weeks after preparing the screening tray. Crystals were photographed under a polarizer.

One hit was identified producing a crystal shower consisting of hundreds of diamonds after one week of incubation (JCSG Core IV condition 42: 100 mM HEPES (free acid), pH 7.5, 4.3 M NaCl). These crystals were far too small to use for data collection without further optimization. This hit condition was chosen for data collection following optimization since the crystals both grew relatively quickly and had a three-dimensional habit, which is ideal for data collection. As with the *S. pombe* p14/SF3B155 crystals, these crystals were optimized using the commercially available additive screen (Additive Screen HT, Hampton Research, Catalogue number: HR2-138). Several additives successfully produced large diamonds. Out of these, NaI was chosen for producing crystals for data collection (Fig. *V*-15).

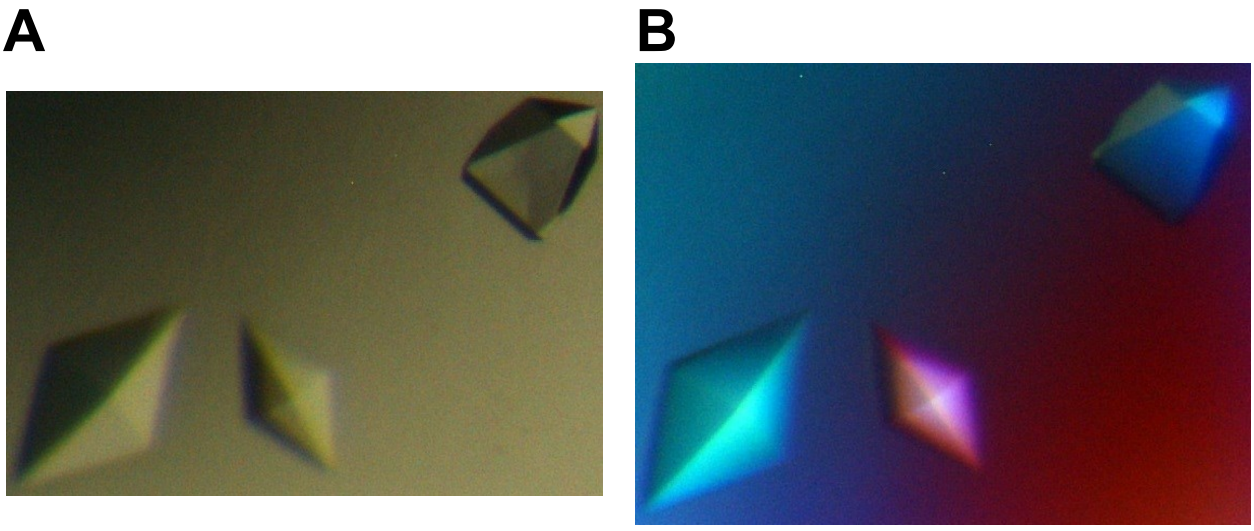


Figure *V*-15. *C. albicans* p14/SF3B155 crystals grown for data collection using NaI as an additive. (A) Representative crystals. (B) Representative crystals viewed under a polarizer.

The optimized *C. albicans* p14/SF3B155 crystals were fragile and showed visible damage upon contact with all cryoprotectant solutions that were attempted while developing the cryoprotection protocol. Since the mother liquor itself was potentially a suitable cryoprotectant due to the high concentration of NaCl, an attempt was made to harvest crystals directly from the crystallization drop and this successfully produced diffraction of sufficiently high quality to solve the structure. However, NaCl crystallized within the protein crystals, contaminating the diffraction. Therefore, in the optimized cryoprotection protocol the crystallization drop was covered with paraffin oil immediately after removing the sealing tape and prior to retrieving the crystal in order to prevent the evaporation of water and NaCl crystallization, as well as remove solvent on the surface of the protein crystal.

V-2.3.3. Data collection and structure solution of the *S. pombe* and *C. albicans* p14/SF3B155 complexes

The crystallographic data collection and refinement statistics for both complexes are summarized below (Table *V*-1). This is followed by an overview of the general features of both X-ray structures as they relate to the crystals themselves, as well as biological implications of the protein structures (Sections *V*-2.3.3.1. and *V*-2.3.3.2.).

Table *V*-1: Crystallographic data collection and refinement statistics for the *S. pombe* and *C. albicans* p14/SF3B155 complexes[7]

	S. pombe complex[8]	*C. albicans* complex[9]
Data collection statistics		
Wavelength (Å)	1.542	1.0332
Resolution range (Å)	43.77-1.9 (1.968-1.9)	45.29-3.111 (3.223-3.111)
Space group	$P2_12_12_1$	$P4_12_12$
Unit cell dimensions		
a, b, c (Å)	57.721, 67.154, 82.092	74.92, 74.92, 174.618
α, β, γ (°)	90, 90, 90	90, 90, 90
Total reflections	284812 (11987)	88533 (8729)
Unique reflections	25778 (2540)	9316 (292)
Multiplicity	11.0 (4.7)	9.5 (9.7)
Completeness (%)	99.59 (98.75)	99.55 (32.05)
Mean I/sigma(I)	16.50 (2.99)	14.24 (1.01)
Wilson B-factor	23.75	72.09
R-merge	0.0748 (0.5541)	0.3116 (4.139)
R-meas	0.07791 (0.6238)	0.3276 (4.347)
R-pim	0.02102 (0.28)	0.0996 (1.31)
$CC_{1/2}$	0.999 (0.845)	0.988 (0.215)
CC*	1 (0.957)	0.997 (0.594)
Structure refinement statistics		
Reflections used in refinement	25716 (2536)	8088 (292)
Reflections used for R-free	1271 (131)	411 (18)
R-work	0.1878 (0.2562)	0.2419 (0.3437)
R-free	0.2229 (0.3095)	0.2912 (0.3281)

[7] Data was collected from a single crystal for both complexes. Statistics for the highest-resolution shell are shown in parentheses.
[8] Molecular replacement was completed using PDB accession 2F9D as the search model.
[9] Molecular replacement was completed using PDB accession 2F9D as the search model.

CC(work)	0.947 (0.883)	0.899 (0.650)
CC(free)	0.940 (0.870)	0.836 (0.550)
Number of non-hydrogen atoms	2076	1521
macromolecules	1892	1520
ligands	27	1
solvent	157	0
Protein residues	230	189
RMS(bonds)	0.005	0.007
RMS(angles)	0.95	1.07
Ramachandran favored (%)	98.65	100.00
Ramachandran allowed (%)	0.90	0.00
Ramachandran outliers (%)	0.45	0.00
Rotamer outliers (%)	0.48	2.34
Clashscore	2.11	2.96
Average B-factor	35.13	71.34
macromolecules	34.49	71.23
ligands	81.85	241.41
solvent	34.84	n/a
Number of TLS groups	16	17

V-2.3.3.1. X-ray structure summary of the *S. pombe* p14/SF3B155 complex

As with all the previously reported human p14/SF3B155 X-ray structures, the *S. pombe* X-ray structure contains two copies of the p14/SF3B155 complex in the asymmetric unit (Schellenberg et al., 2011; Schellenberg et al., 2006). Superimposing these two copies of the complex with each other yields rmsd = 0.7 Å.

P14 is similar between the two copies in the structure. Two successive α-helices append the C-terminus of the RRM in the human X-ray structure. However, in the *S. pombe* structure neither of the two copies of p14 show electron density for these two α-helices, which exist within a solvent channel.

SF3B155 is dissimilar between the two copies of the *S. pombe* structure in the N-terminal region outside of the interfacial core that interacts with p14 but the two copies are otherwise similar. One copy (chain Q) is almost fully modelled (V245-I284) and the N-terminus outside of the interfacial core contains a short α-helix. The other copy (chain P) is less

completely modelled (P253-I284). The C-terminal region included in the *S. pombe* SF3B155 construct based on the PsiPred secondary structure prediction (E275-I284) is modelled in both copies and as predicted, contains a short α-helix; this region is similar in both copies.

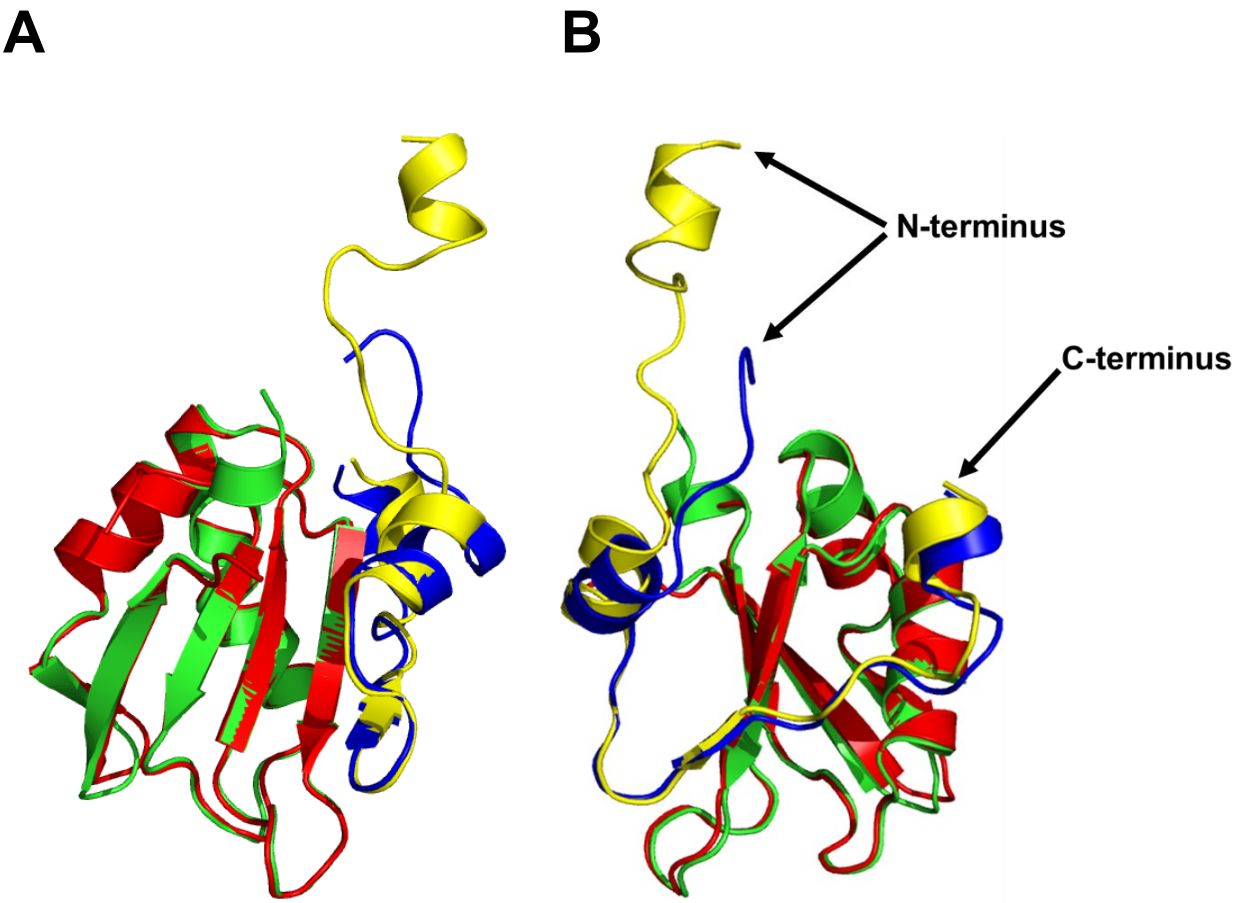


Figure *V*-16. X-ray structure summary of the *S. pombe* p14/SF3B155 complex. The two copies of the complex in the asymmetric unit are superimposed. The first copy is coloured as follows: chain A (p14) in green, chain Q (SF3B155) in yellow. The second copy is coloured as follows: chain B (p14) in red, chain P (SF3B155) in blue. (A) View of the complex showing the β-sheet of the RRM of p14. (B) View of the complex showing the labeled N- and C-termini of SF3B155.

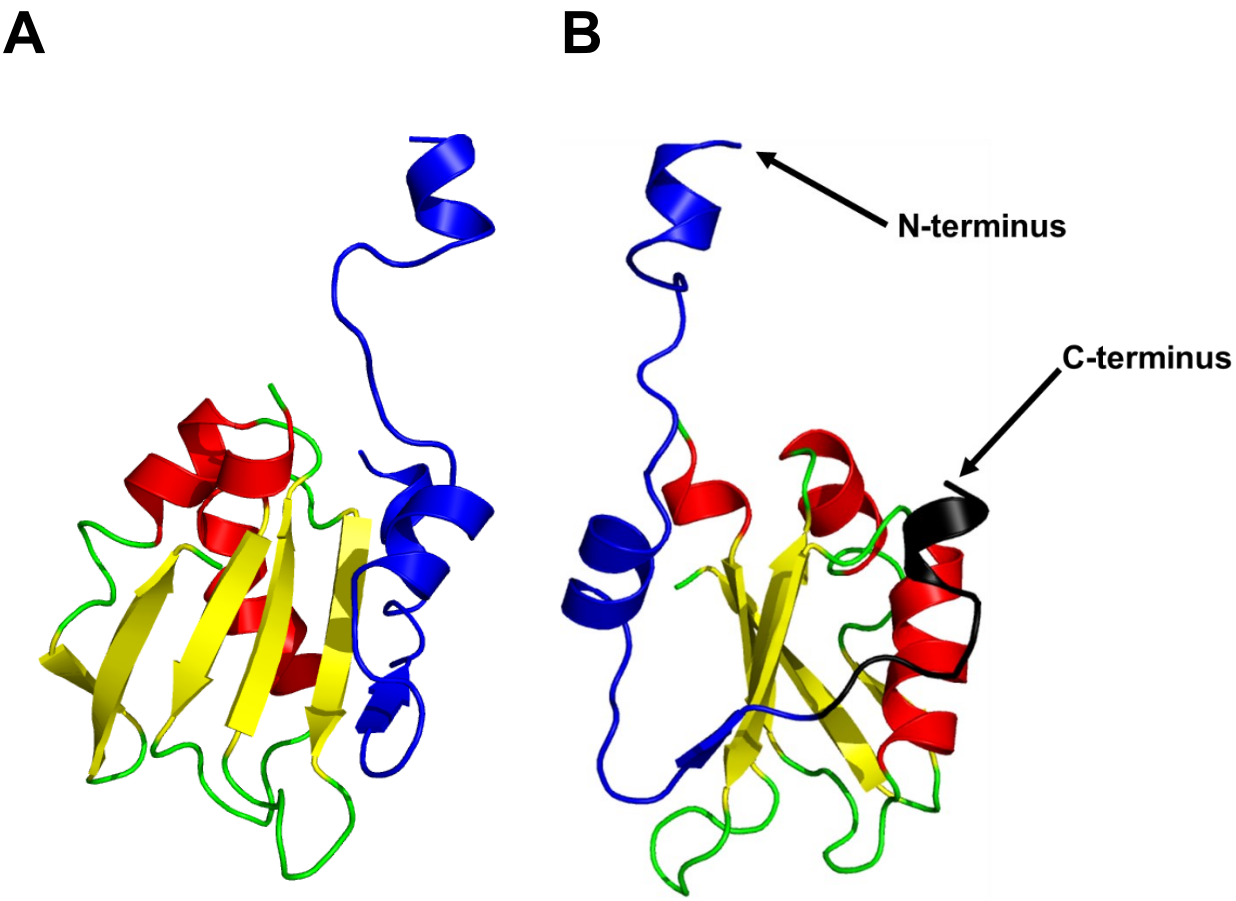


Figure *V*-17. X-ray structural features of the *S. pombe* p14/SF3B155 complex. The copy of the complex in the asymmetric unit corresponding to chain A (p14) and chain Q (SF3B155) is shown. P14 is coloured based on secondary structure (α-helices in red, β-strands in yellow, loops in green). SF3B155 is coloured in blue. (A) View of the complex showing the β-sheet of the RRM of p14. (B) View of the complex showing the labeled N- and C-termini of SF3B155. In this view E275-I284 of SF3B155 is coloured in black.

V-2.3.3.2. X-ray structure summary of the *C. albicans* p14/SF3B155 complex

As with all the previously reported human p14/SF3B155 X-ray structures as well as the *S. pombe* X-ray structure, the *C. albicans* X-ray structure contains two copies of the p14/SF3B155 complex in the asymmetric unit (Schellenberg et al., 2011; Schellenberg et al., 2006). Superimposing these two copies of the complex with each other yields rmsd = 0.6 Å.

Both copies of the *C. albicans* complex in the asymmetric unit are similar to one another across the entire model. As with the *S. pombe* structure, neither copy of the *C. albicans* complex

shows electron density for the two α-helices appending the C-terminus of the RRM of p14. Both copies of SF3B155 only show electron density for the region that interfaces with p14.

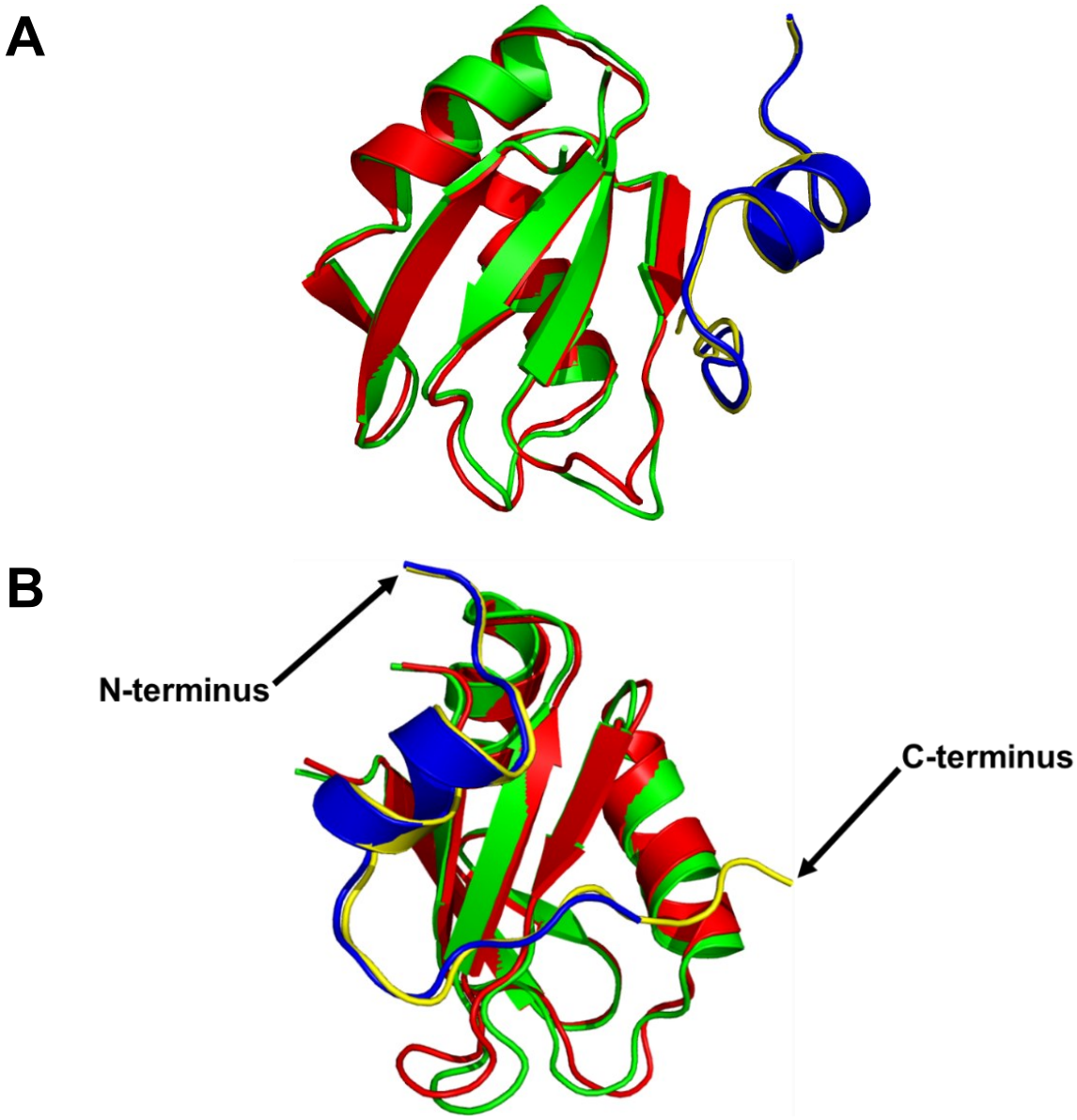


Figure *V*-18. X-ray structure summary of the *C. albicans* p14/SF3B155 complex. The two copies of the complex in the asymmetric unit are superimposed. The first copy is coloured as follows: chain A (p14) in green, chain P (SF3B155) in yellow. The second copy is coloured as follows: chain B (p14) in red, chain Q (SF3B155) in blue. (A) View of the complex showing the β-sheet of the RRM of p14. (B) View of the complex showing the labeled N- and C-termini of SF3B155.

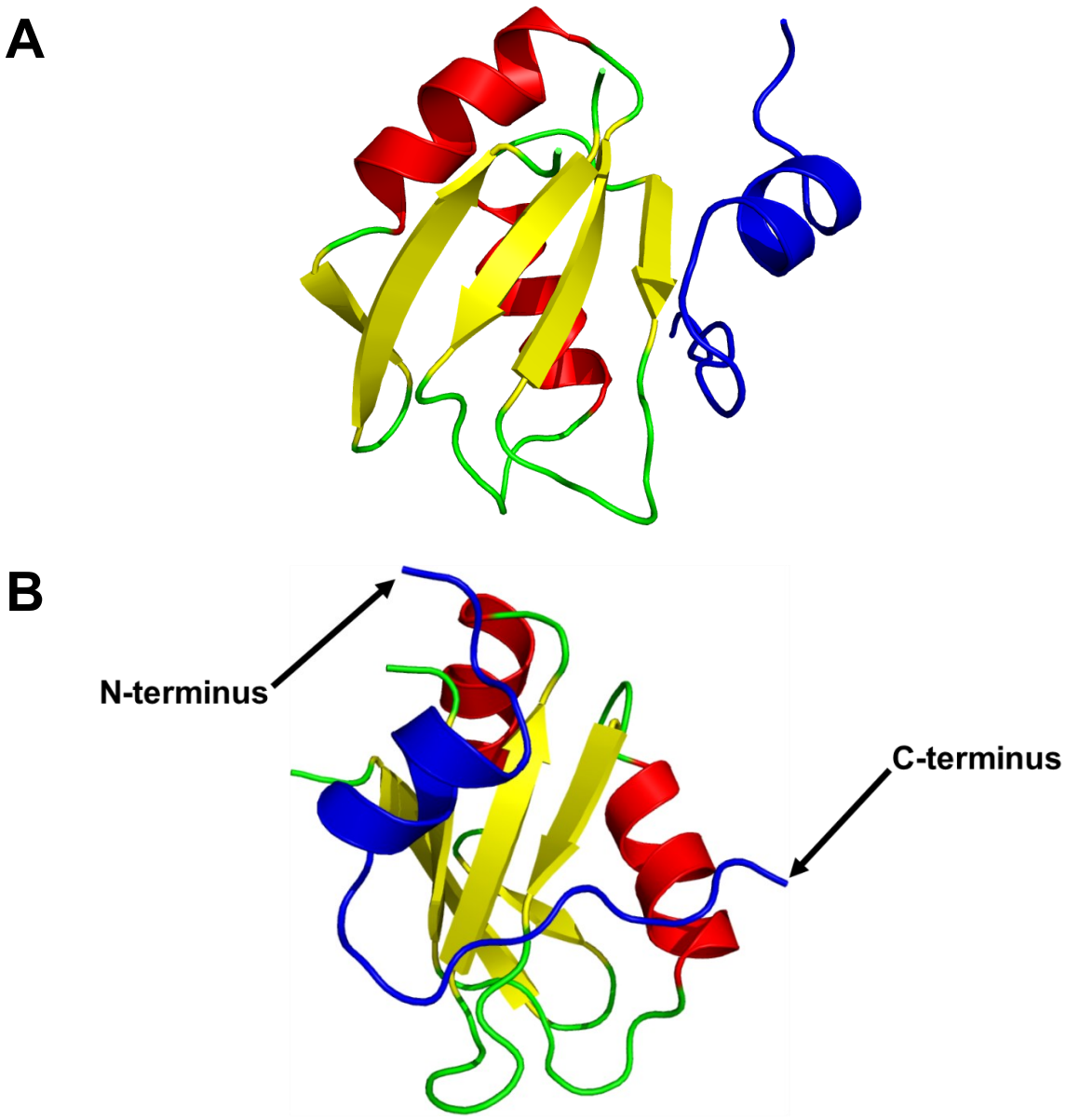


Figure *V*-19. X-ray structural features of the *C. albicans* p14/SF3B155 complex. The copy of the complex in the asymmetric unit corresponding to chain A (p14) and chain P (SF3B155) is shown. P14 is coloured based on secondary structure (α-helices in red, β-strands in yellow, loops in green). SF3B155 is coloured in blue. (A) View of the complex showing the β-sheet of the RRM of p14. (B) View of the complex showing the labeled N- and C-termini of SF3B155.

V-2.3.3.3. Comparison of the human, *S. pombe*, and *C. albicans* p14/SF3B155 X-ray structures

All the previously reported human p14/SF3B155 X-ray structures are of the same crystal form, containing two copies of the complex in the asymmetric unit. Both copies are essentially identical and the most significant difference is in α1 of SF3B155 (see Fig. *V*-20 below); superimposing the two copies from the original X-ray structure (PDB accession 2F9D) with each other yields rmsd = 0.3 Å (Schellenberg et al., 2006). As mentioned previously α3 and α4 of p14 which are outside of the RRM fold are modelled in the human structure and show electron density, unlike the counterpart complexes from *S. pombe* and *C. albicans* (see Fig. *V*-21 below).

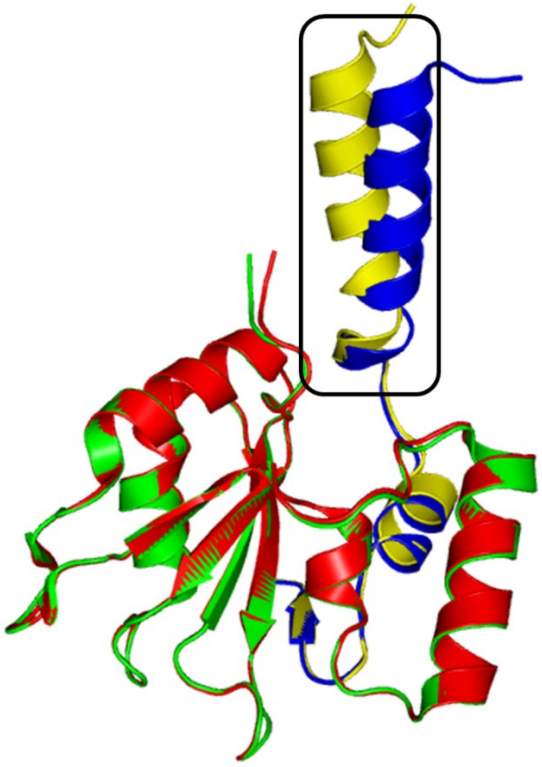


Figure *V*-20. X-ray structure summary of the human p14/SF3B155 complex. PDB accession 2F9D is shown. The two copies of the complex in the asymmetric unit are superimposed. The first copy is coloured as follows: chain A (p14) in green, chain P (SF3B155) in yellow. The second copy is coloured as follows: chain B (p14) in red, chain Q (SF3B155) in blue. A rectangular box is used to indicate α1 of SF3B155 in the two copies of the complex.

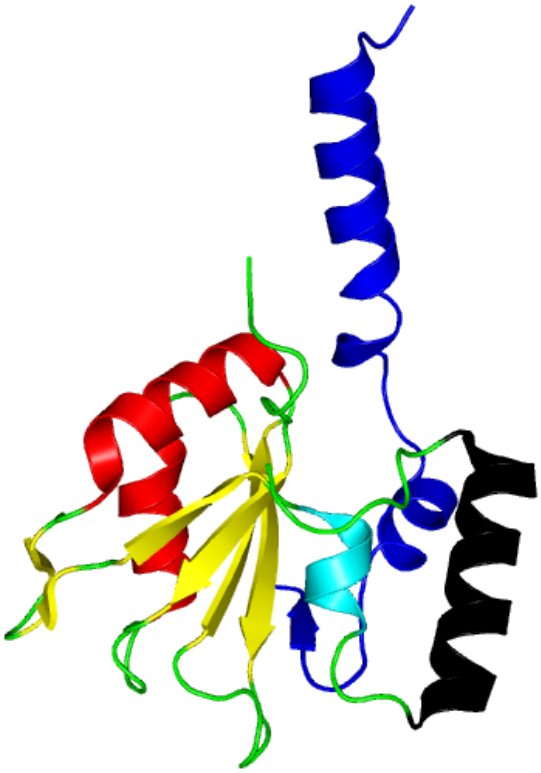


Figure *V*-21. X-ray structural features of the human p14/SF3B155 complex. The copy of the complex in the asymmetric unit corresponding to chain A (p14) and chain P (SF3B155) of PDB accession 2F9D is shown. P14 is coloured based on secondary structure (α-helices within the RRM in red, α3 in cyan, α4 in black, β-strands in yellow, loops in green). SF3B155 is coloured in blue.

When comparing the structures of the three species, superimposing the human structure (PDB accession 2F9D, chains A and P) and *S. pombe* structure (chains A and Q) yields rmsd = 0.8 Å, superimposing the human structure (PDB accession 2F9D, chains A and P) and *C. albicans* structure (chains A and P) yields rmsd = 1.1 Å, and superimposing the *S. pombe* structure (chains A and Q) and *C. albicans* structure (chains A and P) yields rmsd = 1.4 Å (see Figures *V*-22 and *V*-23 below). This shows that the human and *S. pombe* structures are more similar each other than either one is to the *C. albicans* structure, which reflects the degree of conservation of p14 across the three species.

Figure *V*-22. X-ray structure comparison of human, *S. pombe*, and *C. albicans* p14. The p14/SF3B155 complex of all three species is shown. Both the *S. pombe* and *C. albicans* complexes are superimposed onto the human complex. The human complex (PDB accession 2F9D) is coloured as follows: chain A (p14) in red, chain P (SF3B155) in black. The *S. pombe* complex is coloured as follows: chain A (p14) in green, chain Q (SF3B155) in black. The *C. albicans* complex is coloured as follows: chain A (p14) in yellow, chain P (SF3B155) in black.

Figure *V*-23. X-ray structure comparison of human, *S. pombe*, and *C. albicans* SF3B155. The p14/SF3B155 complex of all three species is shown. Both the *S. pombe* and *C. albicans* complexes are superimposed onto the human complex. The N-terminus of human SF3B155 and the C-terminus of *S. pombe* SF3B155 are indicated. The human complex (PDB accession 2F9D) is coloured as follows: chain A (p14) in black, chain P (SF3B155) in red. The *S. pombe* complex is coloured as follows: chain A (p14) in black, chain Q (SF3B155) in green. The *C. albicans* complex is coloured as follows: chain A (p14) in black, chain P (SF3B155) in yellow.

The structure superimposition of the two copies of the complex in the asymmetric unit for the human, *S. pombe*, or *C. albicans* complexes, as well as the structure superimposition of the human, *S. pombe*, and *C. albicans* complexes with one another reveals that together, the RRM of p14 and the interfacial region of SF3B155 which interacts with p14 form a single rigid body. Regions of p14 and SF3B155 outside of this rigid core of the complex either show significant differences both between the two copies in the asymmetric unit as well as between species or cannot be modelled due to a lack of electron density caused by flexibility.

Because α3 and α4 of human p14 do not show electron density in either the *S. pombe* or *C. albicans* counterpart complexes, this region of p14 is likely more plastic than previously thought, and the disposition of these α-helices in the human structure may be an artifact caused by crystal packing forces. This indicates that in a biologically relevant context, partial occlusion of the RNA-binding face of the RRM by α3 and α4 either does not occur, or only occurs in certain contexts. Similarly, the regions of SF3B155 outside of the interfacial region show significant differences across the six X-ray structures of the complex.

With respect to the conserved aromatic residue that stacks with the branch A in human p14, the human and *S. pombe* structures are very similar, although *C. albicans* p14 lacks an aromatic residue at this position and it is not clear how this organism specifically recognizes the branch A (see Fig. *V*-24 below).

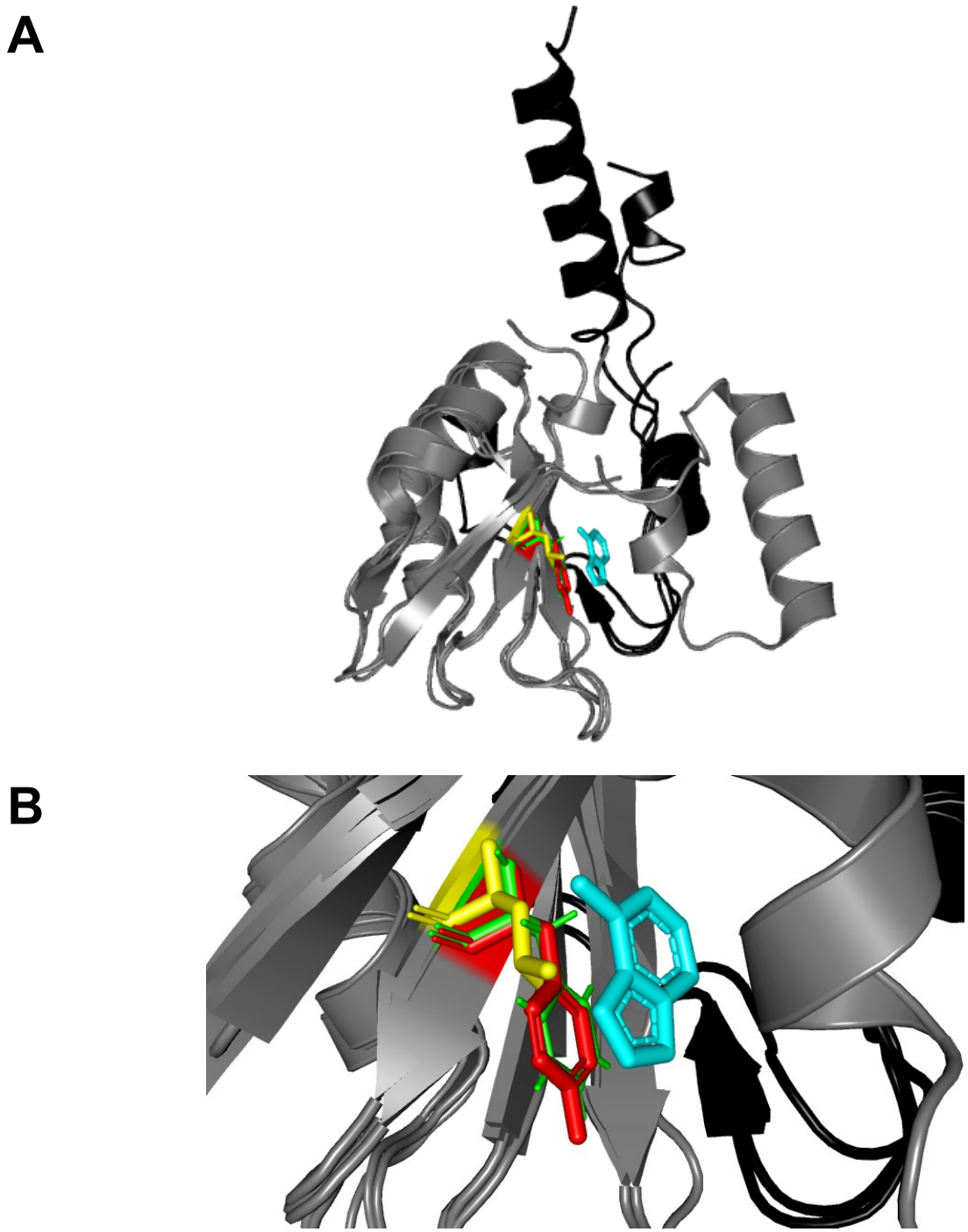


Figure *V*-24. Comparison of the conserved aromatic residue in human, *S. pombe*, and *C. albicans* p14. The p14/SF3B155 complex of all three species is shown. Both the *S. pombe* complex (chains A and Q) and the *C. albicans* complex (chains A and P) were superimposed onto the human adenine-bound complex (PDB accession 3LQV: chains A and P). The conserved aromatic residue is shown as sticks. P14 is shown in grey, with the conserved aromatic residue variously coloured: Y22 (human) in red, F15 (*S. pombe*) in green, L10 (*C. albicans*) in yellow. Other entities are coloured as follows: SF3B155 (black), bound adenine in the human structure (cyan).

V-3. Discussion

V-3.1. Summary of p14/SF3B155 structures

The previously reported X-ray structures of the human p14/SF3B155 complex combined with the *S. pombe* and *C. albicans* X-ray structures reported here indicate that aside from the rigid body core of the complex, it is much more plastic than previously thought. In addition to these X-ray structures, several cryo-EM structures of both the human U2 snRNP and spliceosome have been reported which include p14. These U2 structures include a substrate-bound A-like U2 snRNP (PDB accession 7Q4O), and 17S U2 snRNP (PDB accessions 6Y53 and 6Y5Q) (Tholen, Razew, Weis, & Galej, 2022; Z. Zhang et al., 2020). The spliceosome structures include the pre-B complex (PDB accession 6AH0), B complex (PDB accession 6AHD), pre-B^{act-1} and pre-B^{act-2} complexes (PDB accessions 7ABG, 7ABH, and 7ABI), B^{act} complex (PDB accessions 5Z56, 5Z57, 5Z58, 6FF4, and 6FF7) and minor B^{act} complex (PDB accession 7DVQ) (Bai, Wan, Wang, et al., 2021; Haselbach et al., 2018; Townsend et al., 2020; Zhan et al., 2018b; X. Zhang et al., 2018). P14 and the region of SF3B155 corresponding to the human X-ray structure (T379-L415) have been aligned onto the human p14/SF3B155 X-ray structure in order to discern the flexibility and conformational changes that occur in p14 and the interfacing region of SF3B155 (see Figures *V*-25 and *V*-26 below).

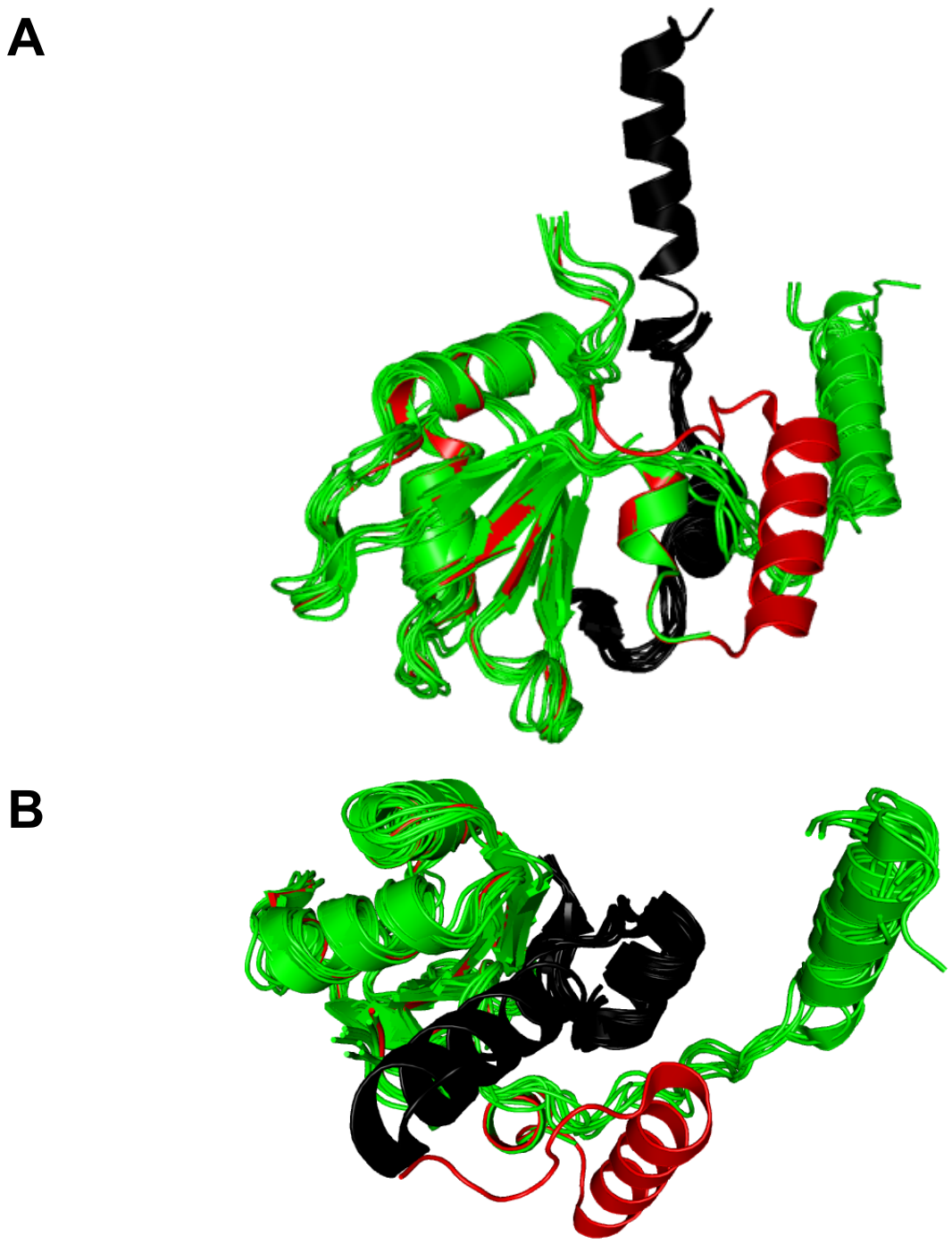


Figure *V*-25. Comparison of X-ray and cryo-EM derived models of human p14. All cryo-EM structures (PDB accessions 7Q4O, 6Y53, 6Y5Q, 6AH0, 6AHD, 7ABG, 7ABH, 7ABI, 5Z56, 5Z57, 5Z58, 6FF4, 6FF7, and 7DVQ) are trimmed to only include p14 and the modelled region of SF3B155 spanning T379-L415. All cryo-EM structures are superimposed onto the X-ray structure (PDB accession 2F9D, chains A and P). The cryo-EM structures are coloured as follows: p14 in green, SF3B155 in black. The X-ray structure is coloured as follows: chain A (p14) in red, chain P (SF3B155) in black. (A) Front view. (B) Top-down view.

Figure *V*-26. Comparison of X-ray and cryo-EM derived models of human SF3B155. All cryo-EM structures (PDB accessions 7Q4O, 6Y53, 6Y5Q, 6AH0, 6AHD, 7ABG, 7ABH, 7ABI, 5Z56, 5Z57, 5Z58, 6FF4, 6FF7, and 7DVQ) are trimmed to only include p14 and the modelled region of SF3B155 spanning T379-L415. All cryo-EM structures are superimposed onto the X-ray structure (PDB accession 2F9D, chains A and P). The cryo-EM structures are coloured as follows: p14 in black, SF3B155 in green. The X-ray structure is coloured as follows: chain A (p14) in black, chain P (SF3B155) in red.

The cryo-EM structures corroborate conclusions from the *S. pombe* and *C. albicans* X-ray structures. Specifically, p14 is very plastic outside of the RRM core, and α1 of SF3B155 is also plastic. Interestingly, in the cryo-EM structures α4 of p14 is swung out relative to the X-ray structure such that α3 assumes a loop. Previous attempts to obtain crystals containing a branch duplex RNA were not successful with the human p14/SF3B155 complex. However, the minimal SF3B155 peptide necessary to stabilize p14 has been defined via the X-ray structures solved from the original crystal form. Since α1 of S3B155 is a flexible, extended structural feature which participates in crystal packing, shortening or eliminating this region is expected to produce

a more globular particle that crystallizes more easily, and which will undoubtedly crystallize in a different crystal form or forms since α1 forms crystal contacts. Additionally, there is a lot of scope to use the *S. pombe* complex and *C. albicans* complex to study RNA-bound p14. The *C. albicans* complex produced three separate crystal forms, and only one of them was solved. Finally, the crystals used to solve the *C. albicans* structure possess a very high solvent content and α3 and α4 of p14 exist within a very large solvent cavity. This means that it may be possible to develop a protocol to soak a short RNA or other oligonucleotide into the crystal to solve the RNA-bound structure. This will also provide clarity on whether the disposition of α3 and α4 of human p14 observed in the X-ray structures is purely a crystallization artifact, or whether RNA-binding induces this region to collapse into a rigid structure.

V-3.2. Biochemical characterization of the *S. pombe* and *C. albicans* p14/SF3B155 complexes

In canonical RRMs the β-sheet represents the RNA-binding surface of the domain and in all RRM/RNA complexes characterized prior to p14, single stranded RNA binds across the β-sheet of the RRM (Allain, Bouvet, Dieckmann, & Feigon, 2000; Deo, Bonanno, Sonenberg, & Burley, 1999; Handa et al., 1999; Kenan et al., 1991; Oubridge, Ito, Evans, Teo, & Nagai, 1994; X. Wang & Tanaka Hall, 2001). However, in the human p14/SF3B155 X-ray structure a significant portion of the β-sheet in p14 is occluded (Schellenberg et al., 2006).

Crosslinking experiments between a minimal RNA and the p14/SF3B155 complex revealed that the branch A crosslinks to a region of human p14 containing β-1 of the RRM and the RNP2 consensus region but not the SF3B155 peptide (Schellenberg et al., 2006). Importantly a portion of RNP2 is exposed within a pocket on the otherwise occluded surface and a highly

conserved aromatic residue within RNP2 is exposed at the base of the pocket (Y22 in human p14). Further mutational, crosslinking and X-ray studies revealed that Y22 in human p14 directly interacts with the branch A (Schellenberg et al., 2006). Because much of the p14 β-sheet is occluded, either a rearrangement occurs upon RNA binding, or the branch duplex interacts with p14/SF3B155 noncanonically.

A follow up study to the aforementioned publication reported a structural model of the p14•bulged duplex interaction developed from a combination of X-ray crystallography, biochemical comparison of a panel of disulfide crosslinked p14/RNA complexes, and SAXS. In this study, a third p14/SF3B155 structure was reported which is essentially the same as the first two except that p14 contains C83S and C74V point mutations, the R96 side chain is swung out from the surface of the complex, and adenine was soaked into the crystals. In this structure, adenine is observed to stack onto Y22 in the previously characterized binding pocket of p14 (Schellenberg et al., 2011). Combined with the other experiments in this study, this structure reveals specific recognition of the branch A by p14 and establishes the orientation of the bulged duplex bound on the protein surface. Additionally, the branch A is buried within p14 as expected, and the p14•duplex interaction must be disrupted prior to the first step of splicing.

However, the *S. pombe* and *C. albicans* X-ray structures presented here suggest that partial occlusion of the RNA binding face of the RRM of p14 is an artifact of crystal packing forces in the human X-ray structure and that p14 is significantly more plastic than previously thought. This does not exclude the possibility that occlusion of the RRM binding face occurs in specific contexts though.

A disulfide tethering approach was used to biochemically characterize the human p14/pre-mRNA interaction and help develop the p14•bulged duplex structural model

(Schellenberg et al., 2011). This was necessary to address problems of affinity/specificity and operates by stabilizing the complex between thiol derivatized RNA and p14 with a single exposed cysteine residue. Specifically, positioning of the branch A in the adenine-bound structure combined with previously reported X-ray and NMR structures of bulged duplex RNAs was used to crudely model the interaction between the human p14/SF3B155 complex and a bulged duplex RNA (Berglund et al., 2001; Y. Lin & Kielkopf, 2008; Newby & Greenbaum, 2002; Schellenberg et al., 2011). Minimizing steric clashes with the protein surface generated two possible orientations of the RNA related by a 180° rotation.

Briefly, single cysteine mutants of the human p14/SF3B155 complex were designed based on the X-ray structure and crude model to sample the phosphodiester backbone with respect to both possible RNA orientations on the proposed RNA binding surface. In combination with this, model bulged RNA hairpins were chemically synthesized containing a single, commercially available 2′-deoxy-H-phosphonate precursor, thereby mimicking the pre-mRNA/U2 snRNA duplex (Froehler, 1986). These were derivatized with a thiol by oxidation of the H-phosphonate with cystamine disulfide during synthesis, which was then reduced and protected with Ellman's reagent. Finally, cysteine mutant protein complexes were equilibrated with the mixed Ellman's•RNA disulfides under mild reducing conditions, producing the desired RNA-protein disulfide in variable yields as determined by SDS-PAGE and confirmed by RNase treatment. From the panel of cysteine mutants, the mutant which generated the highest yield of RNA-protein disulfide and was most kinetically stable to reduction (N25C) best represented the preferred binding orientation of the bulged duplex on the protein surface. Relative disulfide stability to reduction has been demonstrated to reflect the specificity of the protein•ligand

interaction in tethered complexes and is a measure of relative binding affinity (Gilbert, 1995; O'Shea, Rutkowski, Stafford, & Kim, 1989; Stanojevic & Verdine, 1995).

The length of the RNA and position of the modification were based on the initial crude model of the protein/RNA interaction, in which the bulged nucleotide within the p14 pocket sets the register for duplex association with the protein surface. A highly conserved pseudouridine in U2 snRNA base-pairs with the nucleotide 5′ to the branch A and has been proposed to stabilize the extrahelical disposition of the branch A (Newby & Greenbaum, 2002). Therefore, two RNAs were tested, one containing a pseudouridine to match this position and another with cytidine at the same position; disulfide formation between both of these RNAs and the p14 cysteine mutants were indistinguishable suggesting that replacing pseudouridine with cytidine is a valid model of the bulged duplex.

This technique was originally developed by Verdine and coworkers and has been very successfully used to characterize protein•DNA complexes and adopted by others to accelerate the screening of drug-target interactions (Banerjee & Verdine, 2006; Cancilla et al., 2008; Corn & Berger, 2007; Erlanson et al., 2000; Erlanson, Lam, et al., 2003; Erlanson, McDowell, et al., 2003; Fromme, Banerjee, Huang, & Verdine, 2004; He & Verdine, 2002; H. Huang, Chopra, Verdine, & Harrison, 1998; H. F. Huang, Harrison, & Verdine, 2000; Johnson et al., 2006; Komazin-Meredith et al., 2008; S. Lee, Radom, & Verdine, 2008; Zhao et al., 2008).

In order to biochemically characterize the counterpart p14/SF3B155 complexes from *S. pombe* and *C. albicans*, it is important to use the disulfide tethering approach on these complexes as well. The affinity of the human p14/SF3B155 complex for short model RNA duplexes is weak (>100 mM) and there is a lack of specificity for a bulged duplex over double or single stranded RNA (Spadaccini et al., 2006). This likely reflects a cooperative mechanism of BPS recognition

(SF3B155 directly interacts with nucleotides proximal to the branch A), and the fact that p14 likely interacts with multiple bulged duplex structures due to the highly degenerate nature of the human BPS consensus sequence (Gao et al., 2008; Gozani et al., 1998).

Not only does *S. cerevisiae* lack U2AF-S and U2AF-L orthologues, it also lacks any identifiable p14 orthologue (Dziembowski et al., 2004). It is interesting to note that the *S. cerevisiae* BPS is invariant, whereas the human BPS is highly degenerate (Burge et al., 1999; Gao et al., 2008). BPS plasticity is important for splicing regulation in humans and mutations are associated with various disease states caused by splicing errors (Hartmann & Valcarcel, 2009).

It has been suggested that a Prp2-dependent rearrangement displaces both SF3A and SF3B from the branch region prior to the first step of splicing in *S. cerevisiae*, and that SF3 association with the pre-mRNA prevents premature nucleophilic attack by the branch A until the correct orientation of the proper substrate has been achieved (Lardelli et al., 2010). The structural model reported for the human p14/SF3B155 complex bound to bulged duplex RNA and resulting requirement for disruption of the protein/RNA interaction is consistent with these observations.

It is interesting to note that in *S. cerevisiae* (which lacks a p14 orthologue), the branch nucleophile can attack U2 snRNA in the presence of a weak 5′ SS where presumably the positioning of the proper substrate has been decoupled from spliceosome activation (D. J. Smith, Query, & Konarska, 2007). Additionally, bulged RNA duplexes have been shown to be susceptible to hydrolysis 3′ to the bulged residue due to backbone geometry (Portmann, Grimm, Workman, Usman, & Egli, 1996). Therefore, the p14•branch duplex association may be a regulatory step preventing aberrant chemistry during assembly of the spliceosome active site.

A complete understanding of p14 function in BPS recognition requires an atomic resolution structure of p14 bound to a bulged RNA duplex. The disulfide crosslinking approach

described previously has scope for this purpose as it can be used to trap tenuous protein/RNA complexes and has been used previously to generate protein/DNA X-ray structures (Fromme et al., 2004; H. Huang et al., 1998).

V-4. Materials and Methods

V-4.1. Identification of p14 and SF3B155 orthologues in *S. pombe* and *C. albicans*

In order to identify the *S. pombe* and *C. albicans* orthologues for p14, a BlastP search was completed within the EnsemblFungi Genome Browser using the amino acid sequence of human p14 (GenBank accession number: NP_057131.1) (Yates et al., 2022). The *S. pombe* p14 orthologue was identified as the ORF corresponding to Pombase systematic ID SPBC29A3.07c; this ORF is interrupted by a single intron (Harris et al., 2022; Lock et al., 2019). The *C. albicans* p14 orthologue was identified as the ORF corresponding to the *Candida* Genome Database systematic name C3_05540C_A (reference strain: *C. albicans* SC5314); this ORF consists of a single uninterrupted exon (Skrzypek et al., 2017).

In order to identify the *S. pombe* and *C. albicans* orthologues for SF3B155, a BlastP search was completed within the EnsemblFungi Genome Browser using the amino acid sequence of human SF3B155 (GenBank accession number: NP_036565.2) (Yates et al., 2022). The *S. pombe* SF3B155 orthologue was identified as the ORF corresponding to Pombase systematic ID SPAC27F1.09c; this ORF is interrupted by two introns in the N-terminal region (Harris et al., 2022; Lock et al., 2019). The *C. albicans* SF3B155 orthologue was identified as the ORF corresponding to the *Candida* Genome Database systematic name C4_03150W_A (reference

strain: *C. albicans* SC5314); this ORF consists of a single uninterrupted exon (Skrzypek et al., 2017).

V-4.2. Cloning of the *S. pombe* and *C. albicans* p14/SF3B155 complexes

V-4.2.1. Preparation of genomic DNA for PCR template

Genomic DNA was used as PCR template for cloning. In order to generate wildtype *S. pombe* genomic DNA, a patch of cells from a -80°C glycerol stock of strain JK484 (*ura4-D18 leu-32 ade6-216 his3-D1*; generous gift from Dr. Jim Karagiannis, Professor, Dept. of Biology, Western University, London, ON, Canada) was struck onto a YES + agar plate and grown overnight at 30°C (Forsburg & Rhind, 2006; Grewal et al., 2012). In order to generate wildtype *C. albicans* genomic DNA, a patch of cells from a -80°C glycerol stock of strain SN152 (generous gift from Dr. Michael Schultz, Professor, Dept. of Biochemistry, University of Alberta, Edmonton, AB, Canada) was struck onto a YPD (yeast extract peptone dextrose) + agar plate and grown overnight at 30°C (Wurtele et al., 2010). After overnight incubation of either *S. pombe* or *C. albicans*, the patch of cells was scraped off the plate using a pipette tip and resuspended into a microfuge tube of ddH_2O. Subsequently genomic DNA was purified using the rapid isolation method for yeast chromosomal DNA (Hoffman, 2001).

V-4.2.2. Cloning the *S. pombe* p14/SF3B155 complex into pET Duet-1

The *S. pombe* p14 ORF was PCR-amplified from wildtype *S. pombe* genomic DNA and subcloned into empty pET Duet-1 (Novagen) using overlapping PCR in order to remove the

single intervening intron. This subclone was then used as PCR template to clone the full-length protein into empty pET Duet-1 (Novagen) using BamHI/SacI.

Subsequently the *S. pombe* SF3B155 construct was PCR-amplified from wildtype *S. pombe* genomic DNA and cloned into pET Duet-1 (Novagen) into which the final *S. pombe* p14 construct had previously been cloned; the two introns interrupting the ORF are N-terminal to the coding region corresponding to the construct and were therefore never taken into consideration for the cloning strategy. SF3B155 was cloned using NdeI/BglII.

The completed co-expression vector expresses full-length *S. pombe* p14 with a TEV-cleavable, N-terminal hexahistidine affinity tag (cleavage site is ENLYFQG), as well as *S. pombe* SF3B155 (N244-I284) appended at the N-terminus by a methionine.

V-4.2.3. Cloning the *C. albicans* p14/SF3B155 complex into pET Duet-1

The *C. albicans* p14 ORF was PCR-amplified from wildtype *C. albicans* genomic DNA and cloned into empty pET Duet-1 (Novagen) using NcoI/SacI; the forward primer contains sequence corresponding to a non-cleavable hexahistidine affinity tag. Subsequently the *C. albicans* SF3B155 construct was PCR-amplified from wildtype *C. albicans* genomic DNA and cloned into pET Duet-1 (Novagen) into which the *C. albicans* p14 construct had previously been cloned. SF3B155 was cloned using NdeI/XhoI.

The completed co-expression vector expresses nearly full-length *C. albicans* p14 (M1 was deleted) appended at the N-terminus by a non-cleavable hexahistidine affinity tag (appended sequence is MGHHHHHH), as well as *C. albicans* SF3B155 (S144-A183) appended at the N-terminus by a methionine.

V-4.3. Co-expression and purification of the *S. pombe* and *C. albicans* p14/SF3B155 complexes

V-4.3.1. Co-expression of the *S. pombe* and *C. albicans* p14/SF3B155 complexes

In order to co-express either one of the two complexes, *E. coli* BL21-Gold expression strain was transformed with the appropriate pET Duet-1 vector, then plated onto LB (VWR) + agar (Difco) plates containing 0.05 g L^{-1} kanamycin (GoldBio) + 0.3 g L^{-1} carbenicillin (GoldBio) for plasmid selection; the resulting transformant colonies were used as inoculum for liquid culture. A 25-100 mL volume of LB (VWR) broth containing 0.05 g L^{-1} kanamycin (GoldBio) + 0.3 g L^{-1} carbenicillin (GoldBio) was inoculated with a transformant colony in order to initiate a pre-culture of the required protein complex, then grown overnight at 37°C and 210 rpm. The following day, the pre-culture was diluted into 4-6 L of LB (VWR) broth containing 0.05 g L^{-1} kanamycin (GoldBio) + 0.3 g L^{-1} carbenicillin (GoldBio) that was pre-warmed to 37°C; at the time of inoculation, the incubator set temperature was fixed at 15°C in order to allow the media to gradually cool down and stabilize at 15°C by the time the cultures were ready to induce; cultures were grown at 210 rpm. Upon reaching OD_{600} = 0.4-0.6, protein induction was initiated by adding IPTG (GoldBio) to a final media concentration of 0.5 mM, inducing for 15-18 hr at 15°C and 120 rpm.

V-4.3.2. Purification of the *S. pombe* p14/SF3B155 complex

V-4.3.2.1. Nickel affinity chromatography purification of the *S. pombe* p14/SF3B155 complex

E. coli cells were harvested post-induction by centrifugation at 4°C. The pelleted cells were resuspended by stirring on ice in lysis buffer for 30 min (10% (v/v) glycerol, 20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 5 mM BME, 25 mM imidazole, 1 mM PMSF, 0.1 mg mL^{-1} lysozyme), and lysed via sonication (Branson 450 Digital Sonifier) on ice for 30 sec at 70% output power followed by cooling the lysate on ice for 1 min; sonication followed by cooling was performed a total of 3x. Post-sonication, the crude cell lysate was centrifuged at 10,000 rpm for 30 min at 4°C in order to clear the lysate of solid cell debris. The clarified supernatant was then incubated with 2-3 mL of Ni-NTA resin (Thermo Scientific) for 45 min at 4°C with gentle rotation in order to keep the Ni-NTA resin in suspension. The protein-bound resin was then cleared of lysate by passing it through a gravity flow column, followed by three washes of 50 mL each with wash buffer at 4°C (10% (v/v) glycerol, 20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 5 mM BME, 25 mM imidazole). After washing the resin, bound protein was eluted by gravity flow in elution buffer at 4°C (10% (v/v) glycerol, 20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 5 mM BME, 250 mM imidazole). A total of four elutions of 10 mL each were collected and analyzed via SDS-PAGE (6.6% stacking gel, 20% resolving gel) followed by coomassie blue staining.

V-4.3.2.2. First SEC purification of the *S. pombe* p14/SF3B155 complex

Fractions from the nickel affinity chromatography purification containing a significant concentration of mostly pure protein were pooled and concentrated to a final volume of 0.5-2.0

mL using an Amicon Ultra Centrifugal Filter Unit with a 3 kDa pore size cutoff (MilliporeSigma) at 4°C. Concentrated protein was separated via SEC by running it over a HiLoad 26/60 S75 (Superdex 75) column (GE Healthcare Life Sciences) in size exclusion buffer (20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 5 mM BME) at 4°C. Fraction volumes were 4 mL each. Protein-containing fractions eluted from the column were analyzed via SDS-PAGE (6.6% stacking gel, 20% resolving gel), followed by coomassie blue staining.

V-4.3.2.3. TEV cleavage of the *S. pombe* p14/SF3B155 complex

Protein containing fractions from the first SEC purification were pooled and concentrated to a final volume of ~0.5 mL using an Amicon Ultra Centrifugal Filter Unit with a 3 kDa pore size cutoff (MilliporeSigma) at 4°C and the remaining fractions were discarded. Concentrated protein was TEV cleaved by adding 50 µL of TEV at 9.6 mg mL^{-1} and mixing the sample by vortexing. The protein was incubated with TEV overnight at 4°C and complete cleavage of the hexahistidine affinity tag was confirmed by running the cleaved sample alongside its uncleaved counterpart on SDS-PAGE (6.6% stacking gel, 20% resolving gel), followed by coomassie blue staining.

V-4.3.2.4. Anion exchange chromatography purification of the *S. pombe* p14/SF3B155 complex

TEV cleaved protein was purified by running it over a Mono Q HR 5/5 column (GE Healthcare Life Sciences) at 4°C using an ionic strength gradient of 0-20% high salt buffer (low salt buffer = 20 mM Tris-HCl, pH 8.0, 5 mM BME; high salt buffer = 20 mM Tris-HCl, pH 8.0, 1M NaCl, 5 mM BME) over a gradient of 10 column volumes. Fraction volumes were 1.5 mL

each. Protein-containing fractions eluted from the column were analyzed via SDS-PAGE (6.6% stacking gel, 20% resolving gel), followed by coomassie blue staining.

V-4.3.2.5. Second SEC purification of the *S. pombe* p14/SF3B155 complex

Fractions from the Mono Q column containing protein with no detectable contamination were pooled and concentrated to a final volume of 0.5-2 mL using an Amicon Ultra Centrifugal Filter Unit with a 3 kDa pore size cutoff (MilliporeSigma) at 4°C and the remaining fractions were discarded. Concentrated protein was separated via SEC by running it over a HiLoad 26/60 S75 column (GE Healthcare Life Sciences) in size exclusion buffer (5% (v/v) glycerol, 10 mM Tris-HCl, pH 8.0, 100 mM NaCl, 2 mM TCEP-HCl) at 4°C. Fraction volumes were 4 mL each. Protein-containing fractions eluted from the column were analyzed via SDS-PAGE (6.6% stacking gel, 20% resolving gel), followed by coomassie blue staining. Protein containing fractions were pooled and concentrated to a final volume of 0.75 mL using an Amicon Ultra Centrifugal Filter Unit with a 3 kDa pore size cutoff (MilliporeSigma) at 4°C and the remaining fractions were discarded. Concentrated protein was quantified using the absorption co-efficient, diluted with sample buffer to 40 mg mL^{-1}, then aliquoted into thin-walled PCR tubes which were flash-frozen in liquid N_2 and stored at -80°C for later use in crystallographic studies.

V-4.3.3. Purification of the *C. albicans* p14/SF3B155 complex

V-4.3.3.1. Nickel affinity chromatography purification of the *C. albicans* p14/SF3B155 complex

E. coli cells were harvested post-induction by centrifugation at 4°C. The pelleted cells were resuspended by stirring on ice in lysis buffer for 30 min (15% (v/v) glycerol, 2 M urea, 50 mM Tris-HCl, pH 8.0, 1 M NaCl, 5 mM BME, 20 mM imidazole, 1 mM PMSF, 0.1 mg mL^{-1} lysozyme) and lysed via sonication (Branson 450 Digital Sonifier) on ice for 30 sec at 70% output power followed by cooling the lysate on ice for 1 min; sonication followed by cooling was performed a total of 3x. Post-sonication, the crude cell lysate was centrifuged at 10,000 rpm for 30 min at 4°C in order to clear the lysate of solid cell debris. The clarified supernatant was then incubated with 2-3 mL of Ni-NTA resin (Thermo Scientific) for 45 min at 4°C with gentle rotation in order to keep the Ni-NTA resin in suspension. The protein-bound resin was then cleared of lysate by passing it through a gravity flow column, followed by three washes of 50 mL each with wash buffer at 4°C (15% (v/v) glycerol, 2 M urea, 50 mM Tris-HCl, pH 8.0, 1 M NaCl, 5 mM BME, 20 mM imidazole). After washing the resin, bound protein was eluted by gravity flow in elution buffer at room temperature (15% (v/v) glycerol, 2 M urea, 50 mM Tris-HCl, pH 8.0, 1 M NaCl, 5 mM BME, 250 mM imidazole). A total of three elutions of 10 mL each were collected and analyzed via SDS-PAGE (6.6% stacking gel, 20% resolving gel) followed by coomassie blue staining.

V-4.3.3.2. SEC purification of the *C. albicans* p14/SF3B155 complex

Fractions from the nickel affinity chromatography purification containing a significant concentration of mostly pure protein were pooled and concentrated to a final volume of 0.5-2.0 mL using an Amicon Ultra Centrifugal Filter Unit with a 3 kDa pore size cutoff (MilliporeSigma) at room temperature. Concentrated protein was separated via SEC by running it over a HiLoad 26/60 S75 column (GE Healthcare Life Sciences) in size exclusion buffer (5% (v/v) glycerol, 10 mM Tris-HCl, pH 8.0, 500 mM NaCl, 2 mM TCEP-HCl) at 4°C. Fraction volumes were 4 mL each. Protein-containing fractions eluted from the column were analyzed via SDS-PAGE (6.6% stacking gel, 20% resolving gel), followed by coomassie blue staining. Protein containing fractions were pooled and concentrated to a final volume of 0.5 mL using an Amicon Ultra Centrifugal Filter Unit with a 3 kDa pore size cutoff (MilliporeSigma) at room temperature and the remaining fractions were discarded. Concentrated protein was quantified using the absorption co-efficient, diluted with sample buffer to 25 mg mL^{-1}, then aliquoted into thin-walled PCR tubes which were flash-frozen in liquid N2 and stored at -80°C for later use in crystallographic studies.

V-4.4. Crystallization and structure solution of the *S. pombe* and *C. albicans* p14/SF3B155 complexes

V-4.4.1. Crystallization and structure solution of the *S. pombe* p14/SF3B155 complex

V-4.4.1.1. Screening and optimization of crystals of the *S. pombe* p14/SF3B155 complex

The *S. pombe* complex was screened with sitting drop vapour diffusion at room temperature using eight commercially available screening kits (Qiagen): The JCSG (Joint Center for Structural Genomics) Core I, II, III, and IV Suites, PEGs (polyethylene glycols) and PEGs II Suites, MPD (2-methyl-2,4-pentandiol) Suite, and Nucleix Suite (Lesley & Wilson, 2005). The reservoir contained 70 µL of precipitant solution and the crystallization drop was prepared by overlaying 0.5 µL protein solution with 0.5 µL precipitant solution. Three-dimensional diamonds suitable for data collection grew to their maximum dimensions after one week of incubation in JCSG Core IV condition 41.

In order to successfully reproduce the hit, Additive Screen HT (Hampton Research, Catalogue number: HR2-138) was used. A sitting drop vapour diffusion crystallization tray was prepared and incubated at room temperature. The reservoir contained 100 µL of precipitant solution and the crystallization drop was prepared by overlaying 0.5 µL protein solution with 0.5 µL precipitant solution; 100 mM NaI was selected as the additive producing ideal crystals.

Crystals used for data collection were grown using sitting drop vapour diffusion at room temperature. The reservoir contained 100 µL of precipitant solution (1.6 M $LiSO_4$, 100 mM NaI, 100 mM HEPES (free acid, Qiagen), pH 7.5) and the crystallization drop was prepared by

overlaying 0.5 µL protein solution with 0.5 µL precipitant solution. The crystallization tray was incubated for one week before harvesting crystals.

V-4.4.1.2. Data collection, model building, and refinement for the *S. pombe* p14/SF3B155 complex

The crystal used for data collection was soaked in cryoprotectant solution for 5 min (5-20 % (v/v) glycerol, 10 mM Tris-HCl, pH 8.0, 100 mM NaCl, 2 mM TCEP-HCl, 1.6 M $LiSO_4$, 100 mM NaI, 100 mM HEPES (free acid, Qiagen), pH 7.5). This was repeated a total of 4x with four different solutions, with each soak increasing the glycerol concentration in a stepwise series from 5→10→15→20% (v/v) glycerol. After the final soak, the crystal was flash frozen and stored in liquid N_2 until data collection.

Data were collected from a single crystal at 110 K using a Rigaku Micromax-007HF CuKα rotating anode with a Dectris Pilatus3R 200k-A detector. A total of 657 0.25° oscillations were indexed and scaled to 1.9 Å using HKL2000 (Otwinowski & Minor, 1997). The model was built and refined using COOT and phenix.refine in the PHENIX suite to a free R-factor of 0.211 (Emsley et al., 2010; Liebschner et al., 2019).

V-4.4.2. Crystallization and structure solution of the *C. albicans* p14/SF3B155 complex

V-4.4.2.1. Screening and optimization of crystals of the *C. albicans* p14/SF3B155 complex

The *C. albicans* complex was screened with sitting drop vapour diffusion at room temperature using four commercially available screening kits (Qiagen): The JCSG Core I, II, III, and IV suites (Lesley & Wilson, 2005). The reservoir contained 100 µL of precipitant solution

and the crystallization drop was prepared by overlaying 0.5 µL protein solution with 0.5 µL precipitant solution. Multiple hit conditions were identified; the hit corresponding to JCSG Core IV condition 42 was chosen for optimization and produced a crystal shower of hundreds of three-dimensional diamonds after one week of incubation.

In order to optimize the hit, Additive Screen HT (Hampton Research, Catalogue number: HR2-138) was used. A sitting drop vapour diffusion crystallization tray was prepared and incubated at room temperature. The reservoir contained 100 µL of precipitant solution and the crystallization drop was prepared by overlaying 0.5 µL protein solution with 0.5 µL precipitant solution; 100 mM NaI was selected as the additive producing ideal crystals.

Crystals used for data collection were grown using sitting drop vapour diffusion at room temperature. The reservoir contained 100 µL of precipitant solution (4 M NaCl, 100 mM NaI, 100 mM HEPES (free acid, Qiagen), pH 7.5) and the crystallization drop was prepared by overlaying 1.5 µL protein solution with 1 µL precipitant solution. The crystallization tray was incubated for one week before harvesting crystals.

V-4.4.2.2. Data collection, model building, and refinement for the *C. albicans* p14/SF3B155 complex

The crystal used for data collection was cryoprotected by covering the crystallization drop with paraffin oil immediately after removing the sealing tape followed by retrieving the crystal, which was then was flash frozen and stored in liquid N_2 until data collection.

Data from a single crystal were collected at the CLS (Canadian Light Source, Saskatoon, SK, Canada). Data were indexed and scaled to 3.1 Å in space group $P4_12_12$. An initial model for molecular replacement was constructed using SWISSMODEL based on the structure of human

p14 (PDB accession 2F9D) with the C-terminal domain deleted (Waterhouse et al., 2018). Molecular replacement was carried out using PHASER with an initial solution placing two molecules in the asymmetric unit having a Z-score of 10.1 (Mccoy et al., 2007). After initial refinement of p14, SF3B155 was added to the model. The human peptide from 2F9D was mutated to the *C. albicans* sequence followed by morphing and rebuilding using phenix.autobuild. The structure was further refined and built to a final free R-factor of 0.286.