Structural Insights into Early Spliceosome Assembly by cryo-EM

DISSERTATION

for the award of the degree

"Doctor rerum naturalium" (Dr. rer. nat.)

of the Georg-August-Universität Göttingen

within the IMPRS Molecular Biology Graduate Program

of the Göttingen Graduate School for Neurosciences,

Biophysics, and Molecular Biosciences (GGNB)

submitted by

Zhenwei Zhang

from Hangzhou, China

Göttingen, 2021

Thesis Committe

Prof. Dr. Holger Stark

Structural Dynamics, Max Planck Institute for Biophysical Chemistry, Göttingen

Prof. Dr. Henning Urlaub

Bioanalytical Mass Spectrometry, Max Planck Institute for Biophysical Chemistry, Göttingen

Dr. Alexander Stein

Membrane Protein Biochemistry, Max Planck Institute for Biophysical Chemistry, Göttingen

Members of the Examination Board

1st Referee: Prof. Dr. Holger Stark

Structural Dynamics, Max Planck Institute for Biophysical Chemistry

2nd Referee: Prof. Dr. Henning Urlaub

Bioanalytical Mass Spectrometry, Max Planck Institute for Biophysical Chemistry

Prof. Dr. Reinhard Lührmann

Cellular Biochemistry, Max Planck Institute for Biophysical Chemistry

Dr. Alex Faesen

Biochemistry of Signal Dynamics, Max Planck Institute for Biophysical Chemistry

Dr. Alexander Stein

Membrane Protein Biochemistry, Max Planck Institute for Biophysical Chemistry

Prof. Dr. Kai Tittmann

Molecular Enzymology, Georg-August-University Göttingen

Date of submission of the thesis: 26th October 2021

Date of the oral examination: 8th December 2021

AFFIDAVIT

I hereby declare that this PhD thesis "Structural Insights into Early Spliceosome Assembly by cryo-EM" has been written independently with no other aids or sources than quoted. This thesis (wholly or in part) has not been submitted elsewhere for any academic award or qualification.

Zhenwei Zhang

Contents

List of Figures

Abbreviations

2D	two dimensional
3D	three dimensional
Å	Angstrom (1 Å= 1^{-10} metre)
ATP	adenosine triphosphate
BPB	β-propeller domain B
BS-A	brach site adenosine
BSL	branchpoint-interacting stem loop
BSiR	branch site interacting region
cryo-EM	cryogenic electron microscopy
cryo-ET	cryogenic electron tomography
cryo-FIB	cryogenic focused ion beam
CTF	contrast transfer function
CTR	C-terminal region
CXMS	chemical cross-linking mass spectrometry
DEAD	aspartate-glutamate-alanine-aspartate
DEAH	aspartate-glutamate-alanine-histidine
DNA	deoxyribonucleic acid
DPLD	aspartate-proline-leucine-aspartate
EM	electron microscopy
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
FRET	fluorescence resonance energy transfer
GTP	guanosine triphosphate
HEAT	Huntingtin, elongation factor 3, protein phosphatase 2A, and the yeast kinase TOR1
hn	heterogenous nuclear
ISE	intronic splicing enhancer
ISL	internal stem-loop
ISS	intronic splicing silencer
kDa	kilo daltons
KH	hnRNP K homology

LSm	like-Sm
MDS	myelodysplastic syndromes
mM	millimolar
mRNA	messenger ribonucleic acid
MS	mass spectrometry
MW	molecular weight
nm	nanometre
NMR	nuclear magnetic resonance
NTR	N-terminal region
PCA	principal component analysis
PDB	protein data bank
pH	preponderance of hyhrogen ions
PPT	polypyrimidine tract
pre-mNRA	precursor mRNA
RNA	ribonucleic acid
RNP	ribonucleoprotein
RRM	RNA recognition motif
S. cerevisiae	*Saccharomyces cerevisiae*
S. pombe	*Schizosaccharomyces pombe*
SAT	serine-alanine-threonine
SDS-PAGE	SDS-polyacrylamide gel electrophoresis
SL	stem loop
SNR	signal-to-noise ratio
snRNA	small nuclearribonuleic acid
snRNP	small nuclear ribonucleoprotein
SPA	single particle analysis
SR	serine/arginine-rich
SS	splice site
SSA	spliceostatin A
TEM	transimission electron microscopy
UHM	U2AF homology motif
ULM	UHM-ligand motif
UPR	unmodelled protein region
ZnF	Zinc finger

Acknowledgments

Since 2018, I have had the opportunity to work in the Stark department at the Max Planck Institute for Biophysical Chemistry to work on this thesis. I have been extremely lucky to have the chance to work on these amazing projects, with enthusiastic, talented, and enlightening people. This has been an amazing experience and will definitely influence me in the following years of my life. Therefore, I would like to express my sincerest gratitude to the following people who have helped, advised, and guided me during the past few years.

First of all, I would like to thank Prof. Holger Stark for giving me the chance to work in his department. Back in 2017, when I had just started working as a Master's student, he offered me the opportunity to work on a challenging project, the 17S U2 snRNP. This project has then become an essential part of this thesis and the foundation of the following projects. I am thus extremely grateful for his trust. During the past few years, he has been a great supervisor, who guided me into the cryo-EM field and widened my perspective. His great knowledge of science and electron microscopy has greatly impacted me in many ways. His support motivated me during the time when things did not work well. Thank you, Holger!

Second, I would like to thank Prof. Reinhard Lührmann for the great collaboration, his consistent advice, and support. Reinhard led me into the astonishing world of the spliceosome. I have learned so much from the endless but enlightening discussions with Reinhard. His spirit has been one of my greatest motivations, and this thesis work would not have been possible with his great knowledge of splicing and the spliceosome. I am extremely grateful to have the opportunity to work with and to learn from Reinhard.

Third, I would like to thank my collaborators in the Lührmann department, namely Dr. Cindy Will, Dr. Norbert Rigo, Dr. Vinay Kumar, Dr. Olexandr Dybkov, Dr. Romina Hofele, Dr. Dmitry E. Agafonov, and Dr. Sebastian Ludwig. They helped the spliceosome sample preparation for the EM experiments or conducted the mass

spectrometry experiments. Without their help, this thesis would not have been possible. In addition, I would like to thank Dr. Klaus Hartmuth and Dr. Berthold Kastner for the enlightening discussions. I have learned so much from you.

I would also like to thank the two other members of my thesis advisor committee (TAC), Henning Urlaub and Alexander Stein, for the discussion and help whenever necessary. During the TAC meetings, their questions and suggestions have stimulated my critical thinking. In addition, I am also grateful to the IMPRS Molecular Biology program for all the supports during the past few years.

Furthermore, I would like to thank all the current and previous members of the Stark department. Especially Dr. Karl Bertram and Dr. David Haselbach, who initially led me to the methodology of cryo-EM. And special thanks to Cole Townsend for all the technical and scientific discussions, for proofreading this thesis work, as well as for all the beers we had together outside of working hours. I am also grateful to Ka-Man Yip and Cole for creating such a pleasant environment to work in the office.

Last but not least, I am grateful to my family and friends for supporting and bringing joy to my life. Especially my parents, who have always been supportive of whatever I have chosen to pursue. I am also grateful to all the friends I met in Göttingen, with whom I did sports, played board games, and spent all the happy times together. And finally, my very special thanks to my girlfriend, Yinfeng Long, who has been encouraging, supporting, and motivating me. She has brought colour into my life, and thank you for always being on my side.

1   Summary

In eukaryotes, splicing is a process during which the non-coding introns of the pre-messenger RNA (pre-mRNA) are removed, and the protein-coding exons are ligated. The introns are defined by conserved 5' splice site (5'SS), 3'SS, and the branch site[1]. These sequences are recognized by a highly dynamic molecular machinery called the spliceosome that catalyzes the splicing reactions. The spliceosome is a large ribonucleoprotein (RNP) complex that contains five small nuclear ribonucleoproteins (snRNPs), namely U1, U2, U4, U5, and U6 snRNPs, and other non-snRNP proteins[1]. The spliceosome is assembled in a stepwise manner on the pre-mRNA, and during the early assembly phase, U1 and U2 snRNPs recognize the 5'SS and the branch site, respectively, forming the prespliceosome complex (A complex). In the A complex, the U2 snRNA base-pairs with the branch site, forming the U2-BS helix. The branch-site adenosine (BS-A), the nucleophile of the first catalytic step of splicing, is bulged out from the U2-BS helix and thereby defined. The U2-BS helix is stabilized by U2 proteins, especially the SF3b complex. As revealed by recent cryo-EM structures of fully assembled spliceosome complexes[2,3], SF3B1 (human)/Hsh155 (yeast), the major scaffolding protein of the SF3b complex, contains a ring-like HEAT domain and tightly wraps around the U2-BS helix. The closed conformation of the SF3B1/Hsh155 HEAT domain (SF3B1/Hsh155$^{\mathrm{HEAT}}$) found in the spliceosomes is distinct from the open conformation found in the crystal structure of isolated SF3b complex, but the trigger to this functionally important conformational change is unknown. The U2 snRNP also contains TAT-SF1 (Cus2 in yeast), whose function is only poorly understood. In yeast, U2 snRNA nucleotides that base pair with the branch site are initially sequestered in a branchpoint-interacting stem-loop (BSL)[4], but it is unknown whether the human U2 snRNA folds in a similar manner.

The A complex formation requires the DEAD-box ATPase Prp5 that has been shown to mediate an ATP-dependent conformational change in U2 snRNP[5,6,7,8]. However, the Prp5 mediated RNP remodelling is poorly understood due to the lack of structural information of early spliceosomal complexes, especially in humans. Furthermore, Prp5 is also implicated in proofreading the branch site during the A

complex formation. Charles Query's and Soo-Chen Cheng's lab[9][10] have shown that branch-site mutations – including a U to A mutation at position 257 (U257A) of the actin pre-mRNA, which destabilize the U2-BS helix – hinder splicing reaction and lead to an accumulation of spliceosomes at an early assembly stage, at which Prp5 is still associated. Several mutations in Prp5 have been shown to improve the splicing efficiency of the pre-mRNAs containing branch-site mutations. This has led to the proposal that Prp5 functions in assessing the fidelity of U2 base-pairing with the branch site. However, the exact mechanism of how Prp5 proofreads the branch site is unknown.

In the first part of this thesis, we report the first structure of the human 17S U2 snRNP by single-particle cryo-EM at a core resolution of 4.1 Å and its molecular architecture based on crosslinking mass spectrometry (CXMS) data[11]. Our 17S U2 snRNP structure shows that SF3B1$^{\text{HEAT}}$ adopts the open conformation, similar to the conformation identified in the crystal structure of isolated SF3b core[12][13] and that the SF3b core complex does not undergo major structural changes during U2 assembly. The open conformation of SF3B1$^{\text{HEAT}}$ is likely stabilized by the binding of U2 proteins Prp5 and TAT-SF1. Our studies further reveal that U2 snRNA forms a BSL in humans, which is sandwiched by U2 proteins, including Prp5, TAT-SF1, and SF3B1, and is inaccessible for branch site recognition. Therefore, substantial remodelling of the BSL nucleotides and displacement of BSL interacting proteins must occur for stable U2 addition into the spliceosome. By comparing the 17S U2 structure with the U2 region of later spliceosome complexes, we reveal the structural rearrangements facilitated by Prp5 that are required for stable U2-BS interaction and the A complex formation.

In the second part of this thesis, we investigate the molecular mechanism of how Prp5 contributes the branch site proofreading in yeast. We report cryo-EM structures of the yeast spliceosome intermediates assembled on actin pre-mRNAs with the BS-A deleted (Δ BS-A complex) or with the U257A mutation (U257A complex) at the branch site[14]. We show that the two complexes are structurally identical at the current resolution, and they represent a novel spliceosomal intermediate, the pre-A complex, which is formed after Prp5 mediated U2 remodelling but prior to A complex formation. Our pre-A

complexes reveal that formation of the U2-BS helix alone is not sufficient to trigger the closure of Hsh155$^{\text{HEAT}}$. Instead, insertion of the BS-A into the SF3b binding pocket is the major trigger. A comparison of our pre-A complex structures with the 17S U2 snRNP and the A complex structures reveals that the U2 snRNP undergoes a large-scale remodelling during the U2 to pre-A complex and pre-A to A complex transition. Importantly, the pre-A to A transition, which generates the U4/U6.U5 tri-snRNP binding site, is inhibited by Prp5 binding. This provides a structural explanation of why Prp5 and the tri-snRNP binding are mutually exclusive, as previously observed[10]. Our data also suggest that the displacement of Prp5 is coordinated with the docking of the U2-BS helix and the Hsh155$^{\text{HEAT}}$ closure that is trigger by the BS-A insertion. Branch-site mutations that hinder the correct insertion of the BS-A inhibit formation of the productive closed conformation of Hsh155$^{\text{HEAT}}$ and the release of Prp5, thus being stalled at the pre-A stage. Taken together, we propose that Prp5 does not proofread the branch-site sequence directly but rather the overall RNP conformation of the pre-A complex.

2 Introduction

2.1 Structural biology and electron microscopy

2.1.1 Structural biology: understanding biomolecules by visualization

Biological molecules include DNA, RNA, proteins, lipids, and carbohydrates. These molecules are the fundamentals of all known living organisms on earth and contribute to virtually every biological activity. The function of a biomolecule is underpinned by its structure. Structural biology is a study of understanding the mechanism of how biomolecules work by studying how they are built. Visualization of their three-dimensional (3D) structures allows us to interpret how the molecules are assembled, how they interact, and how they perform their specific functions. However, traditional light microscopy is not able to reach the sub-nanometer resolution where the structural details of biomolecules are visible, as it is restricted by the wavelength of visible light. Therefore, structural biology utilizes specialized methods, and the most commonly used of which are X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR), and electron microscopy (EM).

2.1.2 X-ray crystallography, NMR, and EM

A crystal is regular arrays of atoms[15]. The principle behind X-ray crystallography is to generate the diffraction pattern of a crystal by shooting an X-ray. The scattering of an X-ray is determined by the electron within the crystal, and according to diffraction theory, the electron density and the diffraction patterns are related by the fourier transform[16]. Therefore, by obtaining the diffraction pattern, the electron density within the crystal (thus the position of each atom) can be calculated. However, X-ray crystallography has certain inherent limitations. For instance, crystallizing large and flexible macromolecular complexes can be very challenging. In contrast to X-ray crystallography, NMR is a solution-based structural determination method. Therefore, it is not limited by the crystallization process. But high-resolution structural determination by NMR is limited to smaller molecules up to about 80 kDa.

Electron microscopy (EM) is another method based on wave scattering. In this scenario, due to the wave-particle duality, the fast-moving electrons behave as a wave. The electron wave is scattered after interacting with the sample. The interaction between the electrons and the sample can be best understood when consider electrons as particles. When electrons interact with the atoms in biological samples, some electrons are deflected by the nucleus of an atom due to the electrostatic field; some may collide with the atomic nuclei; and some may interact with the orbital electrons of the specimen atoms. Of the interacting electrons, those that do not lose energy are termed as elastically scattered, while those with energy lost are inelastically scattered. After interacting with the sample, the exiting electrons that carries the information of the sample can be focused and magnified by a series of lenses (i.e., the objective lens and projection lenses). Similar to glass lenses in light microscopy that focus a light beam, the electron beam can be focused by making use of electromagnets that act as lenses, generating a two-dimensional (2D) projection image of the specimen on the imaging plane. In this scenario, since the electron beam that is transmitted through the specimen contributes to the image formation, this is known as transmission electron microscopy (TEM). Although only 2D projections can be obtained TEM, methods have been developed that allow the reconstruction of 3D structures *in silico*.

Single-particle analysis (SPA) is one of the most commonly used methods for 3D structure determination of biomolecules. For SPA, the biomolecule sample is purified and imaged under an electron microscope. The resulting EM images contain the 2D projections of each biomolecule (i.e., particle images). Each 2D particle image represents a certain "view" of the purified biomolecule. Ideally, the random orientation of each biomolecule results in various "views" of the sample. By aligning these 2D particle images representing different "views" of the underlying 3D biomolecule and by working out the relative orientation of the "views" (angular determination), the 3D structure can be calculated[17–19]. SPA avoids the crystallization step and is therefore more suitable for dynamic macromolecular complexes. However, SPA is not suitable for biomolecules smaller than 50 kDa, as angular determination for small particles can be very tricky due to insufficient low-resolution signals.

2.1.3 A general workflow of single-particle cryo-EM

During the early years of EM, to increase the contrast of TEM images of thin biological samples, a thin layer of heavy metal salt was used to stain the sample (negative-staining). However, the resolution was largely limited by the grain size of the stain and by working with dried out and flattened biomolecules at room temperature in a high vacuum. The problem was later overcome by Dubochet and Alasdair McDowall[20] through the introduction of cryogenic electron microscopy (cryo-EM), in which the biological sample is embedded in vitreous ice (glass-like ice) that is formed by rapid cooling[21]. The vitrification process preserves the biological samples in a more natural state compared with samples embedded in heavy metal salt. In addition, the low temperature (i.e., the liquid nitrogen temperature of -195°C) also mitigates the radiation damage of the biological samples by the electron beam[22]. Cryo-EM has undergone dramatic development over the decades since its invention[23] and is now routinely used to determine high-resolution structures of biomolecules. A typical workflow of single-particle cryo-EM is summarized in Figure 2.1. The purified sample is firstly applied to a grid, which is a metal disc (e.g., copper or gold) cast with a fine mesh. The grid is then rapidly frozen by plunging into liquid ethane (about -188°C). This rapid freezing prevents the crystallization of water, generating vitrified ice. The cryo-grid is transferred into the electron microscope that is also cooled to liquid nitrogen temperature, and EM images are recorded by a direct electron detector with the help of automatic image acquisition software.

An EM image (i.e., a micrograph) from the detector is usually a stack of image frames. To account for the motion of biomolecules during the time of exposure, the frames are aligned by computer algorithms (motion correction). Because of the spherical aberration and defocused imaging, the recorded images do not fully represent the information in the specimen. Instead, they are modulated by the contrast transfer function (CTF)[24]. At a given acceleration voltage of the electron beam (e.g., 200 kV or 300 kV for typical biological sample data acquisition), the CTF is determined by the spherical aberration and the defocus applied. Therefore, to compensate this modulation (CTF correction), the optical parameters of EM images, including the defocus, are determined. Subsequently, the position of each particle image within the EM micrograph
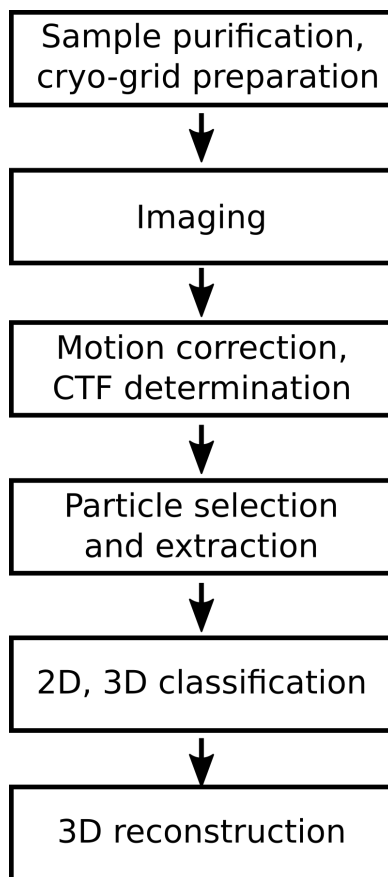
Figure 2.1: **Basic workflow of single-particle cryo-EM**

is determined by template-matching or deep-learning-based algorithms. Each particle is then computationally extracted as a smaller quadratic image, generating a large stack of single-particle images, typically consisting of several hundred thousand single-particle images. The extracted particles usually contain certain amounts of "trash" (e.g., impurities in the sample, empty boxes, ice). However, due to each particle image's low signal-to-noise ratio (SNR) and the large number of particles, it is impractical to go through each particle manually to remove the "trash" images. Instead, classification methods are used (2D classification) so that the projections representing the same relative angular orientation of the molecule are aligned and averaged. Since the noise in the image is generally randomly distributed, after classification and averaging, the signal is improved proportionally to the number of particle images used, while the noise cancels out, generating high SNR 2D class averages. After 2D classification, the classes of interest are selected, while the ones showing no reasonable structural features are discarded.

In order to reconstruct the 3D structure of a biomolecule, the relative orientations of the particle images need to be determined. The relative orientation of each 2D single-

particle projection image with respect to underlying 3D biomolecule can be parameterized by two in-plane translations and three rotations (i.e., Euler angles). To determine these unknown parameters computationally, many methods have been developed, and projection matching-based algorithms are currently the most widely used. For projection matching, an initial 3D reference is required. The 3D reference needs to resemble the overall shape of the true structure and is initially filtered to low-resolution to prevent bias. The low-resolution reference is computationally projected to every direction at certain angular intervals (angular sampling). Each particle is compared with all the reference projections and subsequently aligned and assigned to the best-matching 2D projection of the 3D reference. In this way, the in-plane translational and the Euler angle parameters of each single-particle image are determined. After all the particles are assigned and aligned, a new 3D volume can be reconstructed. In general, this new reconstruction obtained from the particle images should yield a better structure than the initial 3D reference with improved resolution. Therefore, the new reconstruction can be used as the new 3D reference, and the projection-matching algorithm can be applied iteratively. With the improvement in the resolution, the sampling interval can also be decreased (i.e., finer sampling rate) so that the assignment and the alignment of the particle images as well as the resulting 3D reconstruction are improved (3D refinement).

In addition to the unknown orientational parameters of 2D single-particle images, another challenge of high-resolution structural determination of biomolecules is the structural heterogeneity. The 3D reconstruction in SPA assumes all the 2D particle images represent views of an identical 3D structure. However, in reality, most purified biological samples contain multiple conformational and compositional states (and impurities), and thus the collected 2D particle images are mixtures of 2D projections of multiple 3D structures. Therefore, the particle images need to be "purified" *in silico* by being grouped into structurally homogeneous classes. For this purpose, 3D classification methods have been developed. With prior knowledge of the structural variability, instead of one reference, multiple 3D references representing different structural states can be used for the aforementioned projection-matching algorithm (supervised 3D classification)[25]. Thus, in addition to the orientational parameters, each particle image will also be assigned to a class, and multiple 3D structures can be reconstructed simultaneously. However, the requirement of prior structural knowledge largely limits the

general application of the supervised 3D classification. Thus, unsupervised classification methods have been developed, among which the maximum-likelihood based 3D classification[26] is now most widely applied and has been shown to be successful in both revealing the structural heterogeneity in the data set, as well as improving the resolution via grouping structurally homogeneous particles[27,28].

2.1.4   Single-particle cryo-EM of dynamic macromolecular complexes

Although certain large and dynamic complexes (e.g., the ribosome[29,30]) have been successfully crystallized and determined at high-resolution by X-ray crystallography, many dynamic macromolecular complexes are by nature structurally heterogeneous and are hard to crystallize. In the past decade, cryo-EM has become a major asset for structural biologists, as it enables not only structure determination of dynamic complexes that are hard to crystallize, but also an investigation of their conformational landscape by capturing millions of snapshots of the complex, each fixed in a specific conformational state. Recently, the development of computational algorithms such as 3D principal component analysis (PCA)-based and neural networks based methods further allowed visualization of continuous motions within biological complexes[27,28,31,32].

One of the most successful applications of single-particle cryo-EM on macromolecular complexes is the structure determination of the spliceosome (see below), which is a highly dynamic large ribonucleoprotein (RNP) complex that undergoes not only dramatic conformational but also compositional changes throughout its functional process. Although X-ray crystallography has succeeded in resolving some of the sub-complexes of the spliceosome, the inherently dynamic nature of the spliceosome has left it challenging for structural biology for decades. Only until recently have the structural details of intact spliceosomes been brought to high resolution by cryo-EM[2,3,33]. These structures of the spliceosome at different stages have substantially advanced our knowledge of its underlying molecular mechanism. An overview of splicing and current structural insights of the spliceosome are summarized in chapters 2.2 and 2.3.
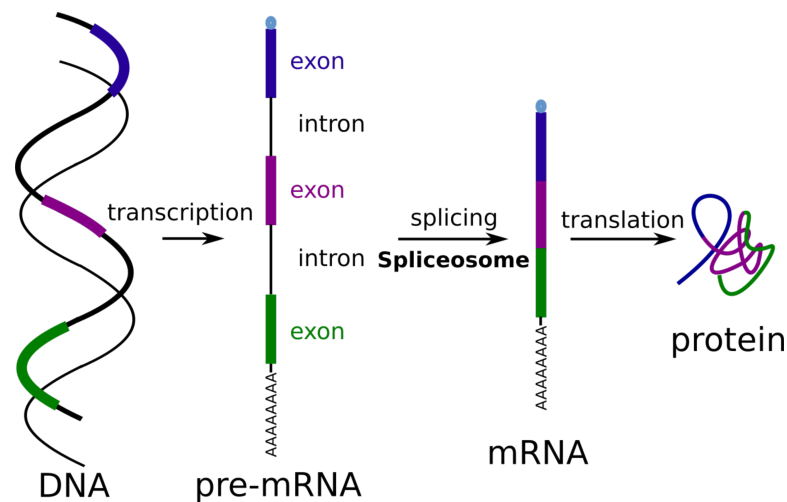
Figure 2.2: **The central dogma of molecular biology.** A schematic view of the information flow from DNA to protein. The pre-mRNA is a copy of the DNA. Splicing removes the non-coding regions and ligates the coding regions, generating translatable mRNAs that can be used by ribosomes.

2.2 Overview of splicing and spliceosome

2.2.1 Genes and the central dogma of molecular biology

The term "gene" was first introduced to describe the inheritable units in 1909[34]. However, it was not until the mid-20th century when scientists confirmed the molecule that carries genetic information was DNA[35]. Later in 1957, Francis Crick foretold the "flow" of genetic information from DNA to RNA and proteins, which is known as "the central dogma of molecular biology" – the information stored in DNA can be copied to RNA (transcription), and proteins can be synthesized using the information in RNA (translation)[36]. Later it was discovered that the protein-coding regions on DNA are not continuous but are instead interrupted by non-coding "silent" DNA, and the precursor mRNA (pre-mRNA) is initially transcribed as a copy of DNA in the nucleus[37,38]. Therefore, before translatable mRNAs can be transported to the cytoplasm, a maturation step called splicing has to convert the pre-mRNA into mature mRNA by removing non-coding regions and ligating the protein-coding parts. Hence, the pre-mRNA splicing process establishes an essential step in the "central dogma of molecular biology" (Figure 2.2). Walter Gilbert, in 1978 termed the region that will be excised during the maturation of mRNA as introns and the regions that will be expressed as exons[39].

2.2.2 Split genes and intron definition

Intron definition

Pre-mRNA introns are defined by three short and conserved *cis*-elements, namely the 5' splice site (5'SS), the 3' splice site (3'SS), and the branch site (Figure 2.3a). The 5'SS defines the border of the 5' exon and the downstream intron, with the 5'-G|GURAG-3' consensus sequence in humans and the 5'-AG|GUAUGU-3' sequence in *Saccharomyces cerevisiae* (*S.cerevisiae*, termed "yeast" from hereon, unless otherwise specified), in which | represents the exon-intron border and R represents any purine[40,41]. The 3'SS is the border of the intron and the downstream exon. Typically, the 3'SS has a sequence of YAG| in both humans and yeast, in which Y represents any pyrimidine[41]. The branch site is located ∼10-60 nucleotides upstream of the 3'SS and contains the branch-site adenosine (BS-A), which is the nucleophile for the first step of splicing reaction (see 2.2.3)[42]. While the branch site is highly conserved in yeast (UACUA**A**C, where **A** is the BS-A)[43], it is much more degenerated in humans (YNYUN**A**Y, where **A** is the BS-A and N represents any nucleotide). Therefore, additional *cis*-elements are required for humans to compensate for the degenerated branch site, such as the poly-pyrimidine tract (PPT), located downstream of the branch site[44,45].

In addition to the aforementioned classical intron (i.e., major type), in higher eukaryotes, there is a minor type of intron (~1%) that is specified by alternative cis-elements. Instead of 5'-GU and AG-3' at the 5'SS and 3'SS, the minor introns are specified by 5'-A|GU and AC|G-3'. Moreover, the minor introns lack a PPT, and their branch site is a more conserved yeast-like sequence (UCCUU**A**C, where **A** is the BS-A)[47].

Alternative splicing

In higher eukaryotes, most transcripts contain multiple splice sites that are subject to alternative selection, leading to various combinations of exons and thus different mRNA isoforms (Figure 2.3b). This process is termed alternative splicing. Alternative splicing introduces an additional layer of the regulation of gene expression and allows a more complex proteome without expanding the genome size significantly. In humans,
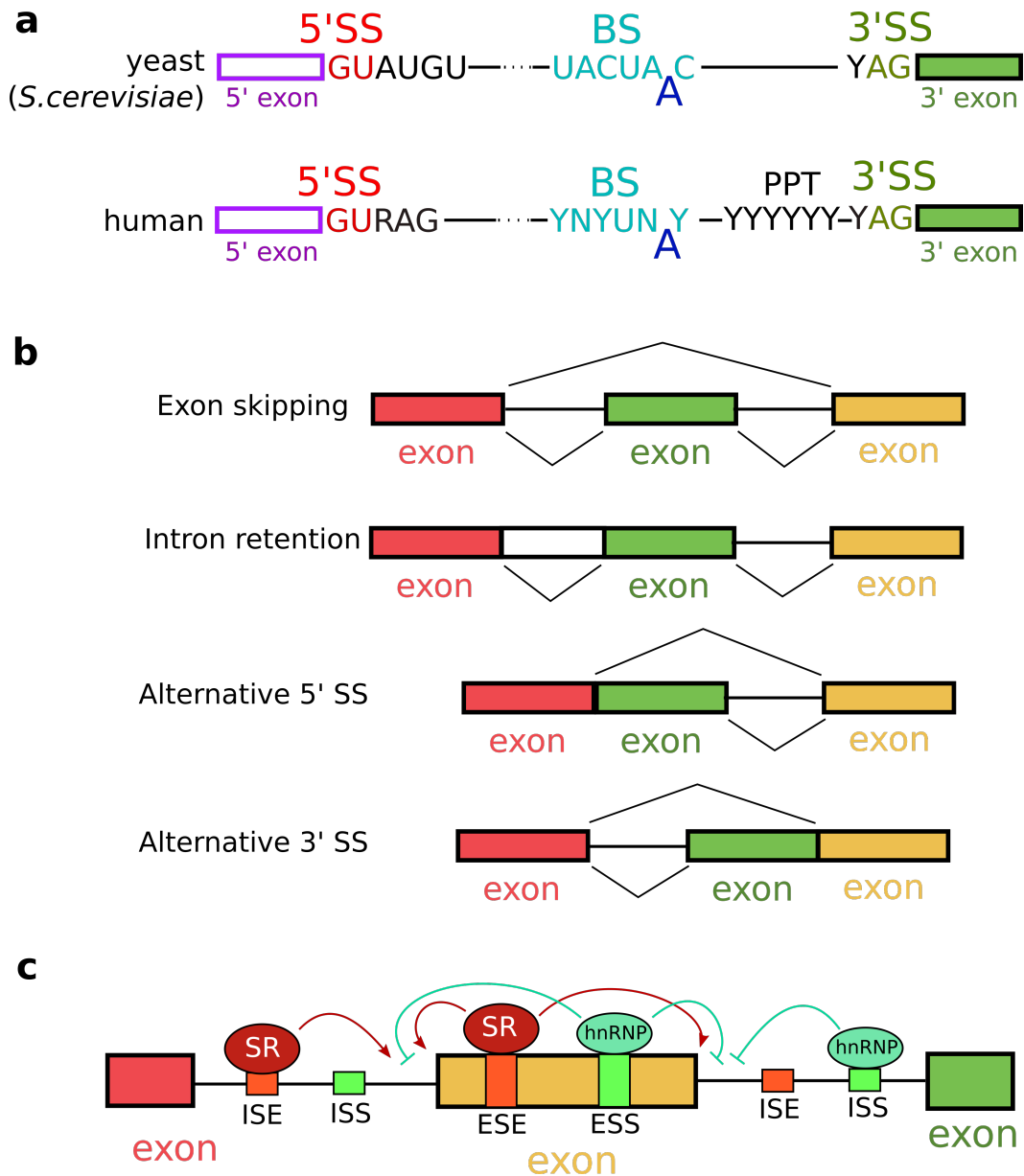
Figure 2.3: **Conserved *cis*-elements that define an intron. a.** Boxes represent the 5'
and the 3' exons. The highly conserved GU and AG at the 5'SS and the 3'SS are
colored in red and green, respectively. The branch site (BS) is colored in light
blue. The BS-A is colored in blue. "Y" represents pyrimidine; "R" represents
purines; "N" represents any nucleotide. **b.** Different modes of alternative splicing.
c. Regulation of alternative splicing. *cis*-elements include exonic splice enhancers
(ESE), silencers (ESS), as well as intronic enhancers (ISE) and silencers (ISS). The
enhancers and silencers are recognized by *trans*-acting proteins, including serine-
rich SR proteins and heterogeneous nuclear ribonucleoproteins (hnRNPs). The
regulation of alternative splicing is not completely understood. Classic regulation
model states that by binding to silencers, the hnRNPs reduce the recruitment of
the splicing machinery and thereby reduce the possibility that the nearby splice
site is used as a junction. In contrast, by binding to enhancers, SR proteins
enhance the usage of the nearby splice junction sites. However, this simple positive
and negative division of SR proteins and hnRNPs do not hold true in every case[46].
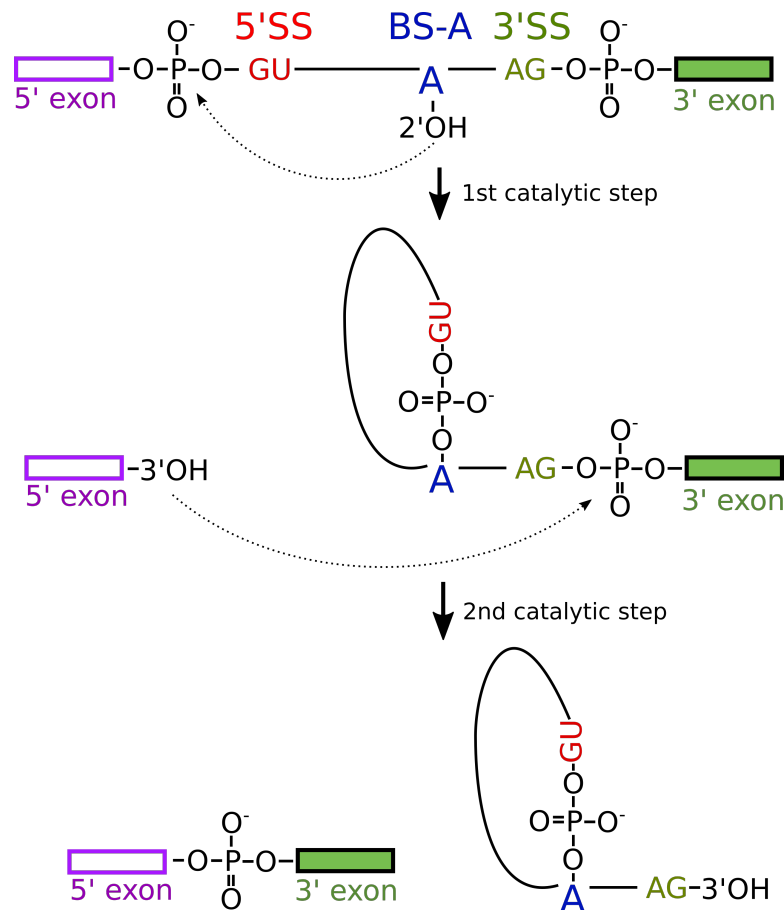
Figure 2.4: **The two-step reaction of pre-mRNA splicing.** The first catalytic step of reaction is also known as branching, during which the 2'-OH group of the BS-A attacks the 5'SS, forming the intron lariat. The second catalytic step is exon ligation, during which the 3'-OH group of the 5'SS attacks the 3'SS, ligating the two exons.

it is estimated that more than 90% of the genes are alternatively spliced[48], and specific isoforms of mRNA are usually produced in a cell- or tissue-dependent way. Alternative splicing can be achieved in multiple ways (Figure 2.3b), and is tightly regulated by *cis*-elements on the pre-mRNA and *trans*-acting proteins[49] (Figure 2.3c).

2.2.3   The chemical mechanism of splicing

Although the sequences that define intron-exon borders are slightly different for the major and minor types of introns, the chemical mechanism remains identical. And for both types of introns, the three *cis*-elements are not only important for the intron definition, but also play an essential role in the chemical reactions of the splicing (Figure 2.4)[1].

The chemical nature of splicing is a two-step transesterification reaction[50,51] (Figure 2.4). The first transesterification reaction is also known as branching, during which the 2'-OH group of the BS-A acts as the nucleophile and attacks the 5'SS, generating the cleaved 5'SS and a lariat like intermediate. The second transesterification is exon ligation, during which the 3'-OH group of the 5'SS acts as the nucleophile and attacks the 3'SS, ligating the two exons and splicing out the intron lariat. The ligated exons (the mature mRNA) are exported to the cytoplasm, and the intron lariat will be subsequently degraded. In some cases, the intron lariat may be converted into circular intronic RNA (ciRNA) and plays a role in regulation[52].

The chemical mechanism of pre-mRNA splicing largely resembles that of the group II self-splicing introns found mainly in prokaryotes, which is proposed to be the ancestor of the spliceosomal introns[53,54]. It is believed that the group II introns were introduced to eukaryotic nucleic genome by endosymbiotic bacteria that gave rise to mitochondria and chloroplasts, and afterwards the structured ribozyme was degenerated and separated into small nuclear RNAs (snRNAs) that are able to reassemble into the spliceosome[54]. For the group II introns, its structured RNA regions are able to form six domains (from DI to DVI) and fold into a complex tertiary structure, so that the splice sites are precisely positioned in the RNA catalytic center, which is formed through interactions of DI, DII, DV, and DVI[55]. *In vitro*, group II introns do not need additional proteins for the self-splicing reactions[56,57]. In contrast, the pre-mRNA introns contain little conserved secondary or tertiary structural information to align the chemically reactive groups. Therefore, they require the spliceosomal proteins to act in *trans*, to fold the pre-mRNA correctly so that the reactive groups from the 5'SS, 3'SS, and the branch site are accurately positioned in the catalytic center.

2.2.4 snRNPs, the building blocks of the spliceosome

The spliceosome that catalyzes the splicing of the major type of intron is known as the major spliceosome, which has U1, U2, U4/U6, and U5 small nuclear RNP (snRNP) as its main building blocks. Each snRNP is made from one copy of its corresponding snRNA and numerous interacting proteins. The snRNAs, with the exception of the U6 snRNA, are transcribed by the RNA polymerase II, post-transcriptionally modified

and then capped with tri-methylated guanosine (m$_3$G)[58–62]. Although the sequence and the overall structure of different snRNAs vary, the U1, U2, U4, and U5 snRNA all contain a common uridine-rich site, also known as the Sm-site (Figure 2.5). This Sm-site serves as a binding site for a seven-member heteromeric protein ring, the Sm ring, which is comprised of Sm B, D1, D2, D3, E, F, and G[63]. The U6 snRNA is transcribed by the RNA polymerase III and lacks a typical Sm site[64]. Instead, it contains a uridine-rich sequence at its 3' end, on which the Sm-like (LSm) protein ring is docked. The seven-member heteromeric LSm ring is composed of LSm2, LSm3, LSm4, LSm5, LSm6, LSm7, and LSm8[65,66]. In addition to the Sm/LSm complexes, each snRNP also contains a specific set of proteins that interact with the snRNA (Figure 2.5).

The U1 snRNP contains U1-70K (Snp1 in yeast), U1-A (Mud1 in yeast), and U1-C (Yhc1 in yeast), in addition to the U1 Sm complex (Figure 2.5). U1-70K and U1-A bind stem-loop I (SLI) and SLII of the U1 snRNA, respectively, while U1-C binds to the 5' end of the U1 snRNA[67,68]. In *S. cerevisiae*, the U1 snRNA is expanded, and additional yeast-specific U1 proteins (Nam8, Prp39, Prp40, Prp42, Luc7, Snu71, and Snu56) help to stabilize the RNA expansion[69]. Except for Snu56 and Prp42, all the yeast-specific U1 proteins have human homologues and act as alternative splicing factors[69,70].

The human U2 snRNP is purified as a 17S particle under lower salt conditions (e.g., 150 mM) and contains U2-A', U2-B", the three-subunit SF3a complex (SF3A1, SF3A2, and SF3A3), and the seven-subunit SF3b complex (SF3B1, SF3B2, SF3B3, SF3B4, SF3B5, PHF5A, and SF3B6/p14), in addition to the U2 Sm complex[6]. Prp5 (also known as DDX46) and TAT-SF1 were also shown to associate with the immunopurified human 17S U2 snRNP (Figure 2.5)[6,71]. The human U2 snRNA is 187 nucleotides long and is proposed to structure into several stem-loops (see 2.3.2). Although the yeast U2 snRNA is significantly expanded (more than 1000 nucleotides), the first ~100 nucleotides are well conserved from human to yeast. Early structural studies of the human 17S U2 snRNP showed a bipartite structure, in which the SF3b complex binds to the 5' region of the U2 snRNA (i.e., upstream of the Sm site), forming the 5' domain of the particle, while the Sm complex together with U2-A' and U2-B" that bind to
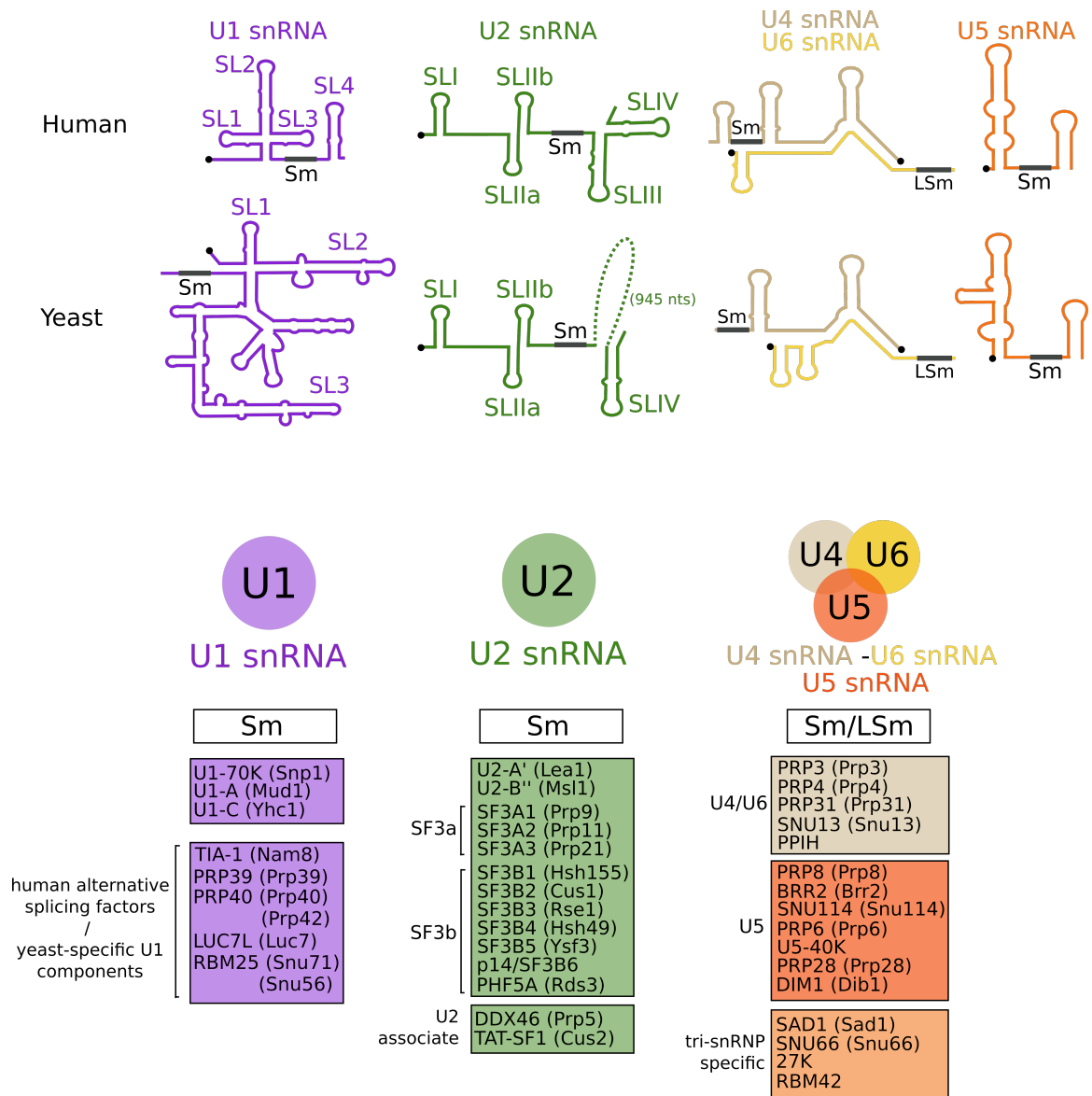
Figure 2.5: **Protein and RNA compositions of human and yeast snRNPs.** The human and yeast snRNA secondary structures are based on Will et al.,2011[1] and Plaschka et al., 2019[3], respectively. The protein composition is based on Kastner et al., 2019[2] and Plaschka et al., 2019[3]. The yeast proteins are shown in brackets. The yeast U1 snRNP contains yeast-specific U1 proteins (Nam8, Prp39, Prp40, Prp42, Luc7, Snu71, and Snu56), of which the human counterparts are not stably associated with the human U1 snRNP and are believed to function as alternative splicing factors. Snu71 and Prp42 do not have human counterparts. The black dot at the end of each snRNA represents the capped 5' end.

the 3' region forms the 3' domain. The two domains are bridged by the SF3a complex[72].

The U4, U5, and U6 snRNPs exist as a preassembled 25S U4/U6.U5 tri-snRNP, in which the U4 and U6 snRNAs are extensively base-paired (Figure 2.5). In addition to the Sm/LSm complexes, the human U4/U6 di-snRNP contains a specific set of proteins, namely Prp3, Prp31, Prp4, Snu13, and PPIH. The human U5 snRNP contains the U5 snRNA, Prp6, Prp8, Prp28, Snu114, Brr2, U5-40K, and the U5 Sm[2,71]. The tri-snRNP contains additional proteins Snu66, Sad1, 27K, and RBM42 (Figure 2.5). The yeast tri-snRNP is compositionally similar to its human counterpart, except that yeast homologs of PPIH, U5-40K, 27K, and RBM42 have not been identified, and Sad1 and Prp28 are not stably integrated into the yeast tri-snRNP[3]. In the tri-snRNP, Prp8 is the largest (2335 amino acids in humans and 2413 in yeast) and the most conserved scaffold protein and plays an essential role in catalytic core formation during the splicing process[73]. Snu114 is structurally homologous to the ribosome elongation factors EF-G/EF-2. EF-G is known to trigger a conformational change of the ribosome by binding and hydrolyzing GTP. However, a similar role of Snu114 in the spliceosome has not been identified[3]. Brr2 and Prp28 are RNA helicases that facilitate the unwinding of the U4/U6 base-pairing and U1/5'SS base-pairing, respectively (see chapter 2.2.5).

In higher eukaryotes, the minor spliceosomes recognize and process the minor type of introns. In addition to the U5 snRNP that is shared with the major spliceosomes, the minor spliceosomes contain U11, U12, U4atac, and U6atac snRNPs, which are functional analogs of U1, U2, U4, and U6 snRNP, respectively.

2.2.5   Dynamic assembly pathway of the spliceosome

Unlike other molecular machinery, like RNA polymerases or ribosomes that are preassembled with a pre-existing catalytic center, the spliceosome has to assemble *de novo* on pre-mRNAs for each splicing reaction in a stepwise manner, and the catalytic center has to form anew each time[1](Figure 2.6). During the active center formation and the catalysis, the snRNA-snRNA and pre-mRNA-snRNA networks are significantly remodelled (Figure 2.7). The RNA remodelling is facilitated by proteins.

Throughout the splicing process, more than 80 proteins are involved in yeasts and ~80 more are involved in higher eukaryotes, and the spliceosome undergoes dramatic conformational and compositional changes[71,74]. This mixture of "biogenesis" and the catalysis of the spliceosome makes the entire process highly complex, despite the rather simple chemical nature of splicing. To shed light on the complex and dynamic process mechanistically, high-resolution structural data are required. Starting from 2015, a surge of human and yeast spliceosome structures at different stages have been reported at high-resolution[2,3,33]. Combined with decades of biochemical, genetic, and structural studies, these cryo-EM structures have provided valuable insights into pre-mRNA splicing.

The splicing process can be roughly divided into four phases, namely assembly, activation, catalysis, and disassembly of the spliceosome (Figure 2.6). The assembly of the spliceosome starts with the recognition of the 5'SS by the U1 snRNP via the base-pairing between the 5'SS and the U1 snRNA 5' end, in an ATP-independent manner[75]. Meanwhile, the splicing factor SF1/mBBP and the U2AF1-U2AF2 (U2AF65-U2AF35) heterodimer bind to the branch site, the PPT, and the 3'SS of the intron in a collaborative manner[76–78], forming the spliceosomal E complex ("E" stands for "Early", also known as the commitment complex). SF1/mBBP recognizes the branch site in a sequence dependent manner via its KH-QUA2 domain[79], and U2AF binds downstream of the branch site via RNA recognition motifs (RRMs)[80]. The solution NMR structure of SF1/mBBP bound with RNA shows that the BS-A is deeply buried inside the KH-QUA2 protein and pre-bulged[79]. The pre-bulged BS-A might facilitate later U2 addition. In yeast, Msl5, the homologue of SF1/mBBP, recognizes the branch site in an almost identical way via its KH-QUA2 domain[81,82]. However, the yeast U2AF, Mud2, lacks the small subunit and forms a stable heterodimer with Msl5[83]. Similar to U2AF, Mud2 binds downstream of the branch site, but it is not essential for yeast vegetative growth[84,85].

In humans, the U2 snRNP is believed to be already loosely associated with the E complex[86,87]. In the presence of ATP, SF1/mBBP is displaced from the branch site, and the U2 snRNP is stably incorporated, forming the spliceosomal A complex[88] (also
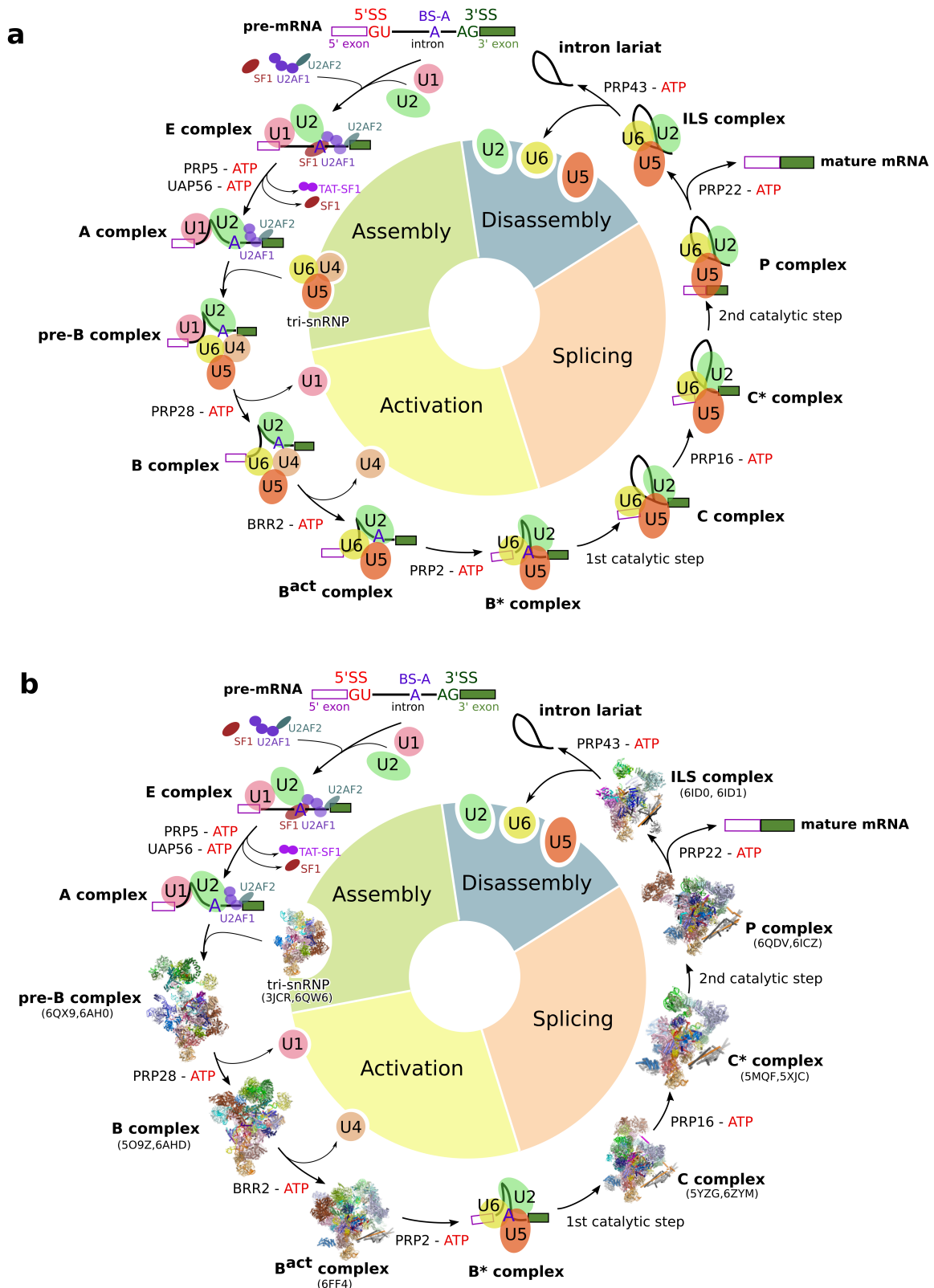
Figure 2.6: **Dynamic assembly pathway of the spliceosome. a.** A schematic overview of the splicing cycle. The U snRNPs are shown as cycles. The exons and introns are represented by colored boxes and lines, respectively. **b.** The PDB models of published cryo-EM structures of human spliceosomes at distinct stages. Prior to this study, cryo-EM structures of the human E and A complexes as well as their major subunit, the 17S U2 snRNP, have not been reported.
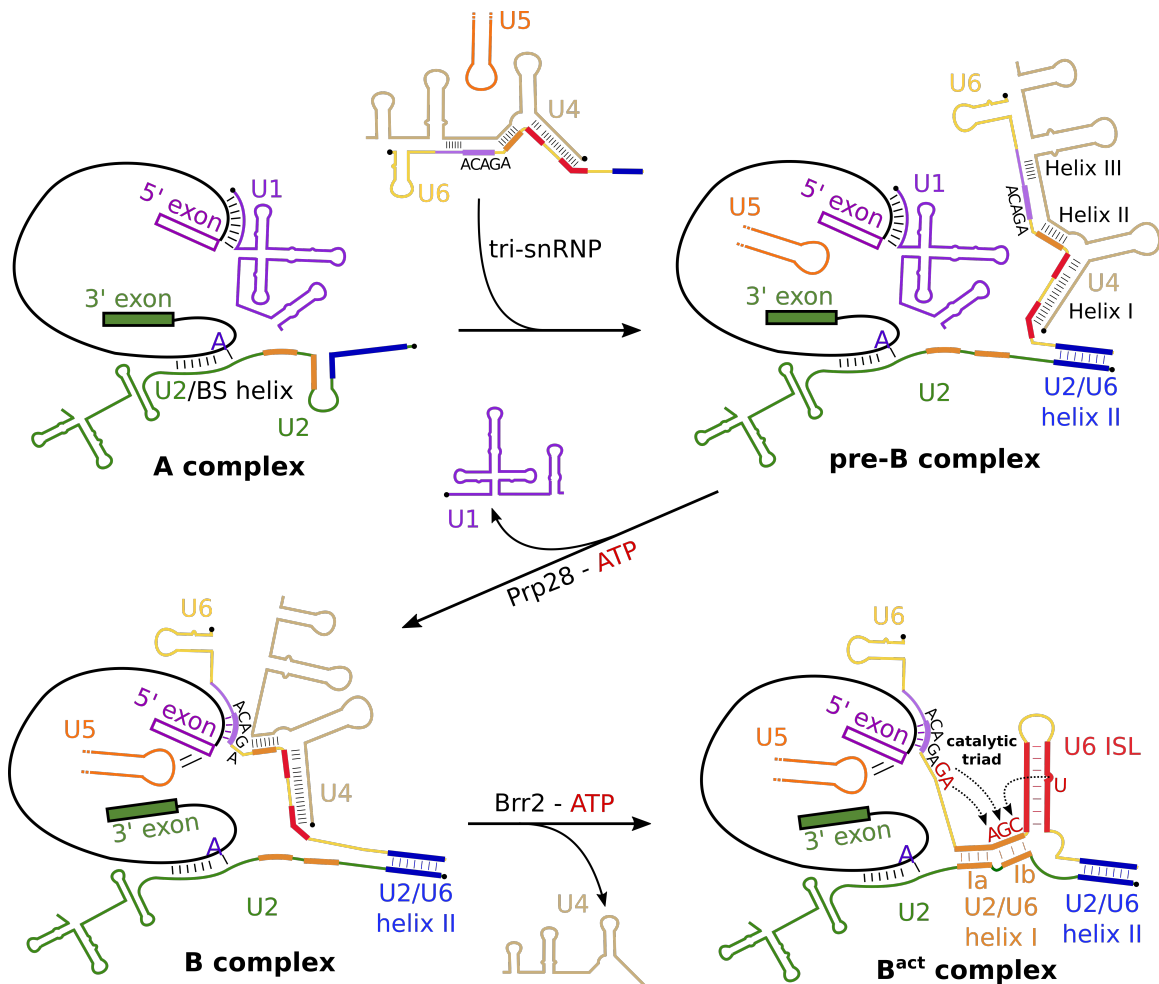
Figure 2.7: **The RNA remodelling events during spliceosome assembly and activation.** Exons are depicted as colored boxes and introns as lines. In the A complex, U1 and U2 base-pair with 5'SS and the branch site, respectively. The BS-A is highlighted in blue. After the tri-snRNP recruitment, U2 5' end base-pairs with U6 3' end, forming U2/U6 helix II. After the release of U1, U6 base-pairs with 5' SS, and U5 loop 1 interacts with 3' terminal of the 5' exon. After the release of U4, U6 refolds into the ISL, and U2 forms further base-pairing with U6, forming U2/U6 helix I (consisting of U2/U6 helix Ia and Ib). U6 ISL and U2/U6 helix I together form the catalytic triad of the spliceosome. The U6 nucleotides that contribute to the catalytic center are highlighted in red. The black dot at the end of each snRNA represents the capped 5' end.

known as the prespliceosome). In the A complex, the U2 snRNA base-pairs with the branch-site region, forming the U2-BS helix with the BS-A bulged out and thereby selected[42]. The U2-BS helix is further stabilized by the U2 SF3a and SF3b complexes, in particular, the largest subunit of the SF3b complex, SF3B1 (see 2.3.3). The formation of the A complex requires two RNA helicases UAP56 (Sub2 in yeast) and Prp5 (see 2.2.6). After the formation of the A complex, the preassembled U4/U6.U5 tri-snRNP is recruited, forming the pre-B complex, the fully assembled spliceosomal complex that contains all the five snRNPs[89].

In the pre-B complex, the U6 snRNA forms new base-pairing with the U2 snRNA (U2/U6 helix II, Figure 2.7), while the U1 snRNA remains base-paired with the 5'SS[89]. The subsequent displacement of the U1 snRNP is mediated by the RNA helicase Prp28[90], liberating the 5'SS to base-pair with the U6 ACAGA box. In addition, the U5 loop 1 base-pairs with the 3' terminal nucleotides of the 5' exon. Formation of the U6-5'SS helix and the U5-5'exon interaction demarcate the formation of the pre-catalytic B complex[91]. In the B complex, the catalytic core has not yet formed. The activation of the spliceosome converts the B complex into the activated B^{act} complex by the RNA helicase Brr2. Brr2 facilitates the unwinding of the U4/U6 helices, leading to the dissociation of U4 snRNA[92](Figure 2.7). This frees the U6 snRNA to refold into the active form, the internal stem-loop (ISL), and to engage with the U2 snRNA, forming U2/U6 helices Ia and Ib. The folding of U2 and U6 RNAs are guided by proteins, as revealed by the pre-B^{act} complexes[93]. The newly formed U6 ISL together with U2/U6 helices Ia and Ib adopts a tertiary conformation, known as the catalytic triplex. Similar to group II introns, this catalytic triplex coordinates the two catalytic magnesium ions, giving rise to the RNA catalytic center of the spliceosome[94,95]. The formation of the catalytic center features the completion of the activation phase. The catalytic center formed in the B^{act} complex is maintained and performs both steps of the splicing reaction.

Although the catalytic center has formed in the B^{act} complex, the BS-A, the nucleophile for the first step reaction, is sequestered and kept away from the catalytic center[31,96−98]. Subsequently, the RNA helicase Prp2 triggers the remodelling of the

B^{act} complex, leading to the liberation of the BS-A and B* complex formation. After the first step of splicing, the branching reaction, the C complex is formed. For the second step of the reaction to happen, the U2-BS helix has to be removed from the catalytic center to allow the 3'SS docking. This remodelling is mediated by the RNA helicase Prp16, generating the C* complex. After the 3'SS docking in the catalytic core of the C* complex, exon ligation takes place, forming the post-catalytic P complex[99]. In the disassembly phase, the P complex is disassembled by the RNA helicase Prp22, releasing the spliced mature mRNA. The resulting ILS complex is disassembled by the RNA helicase Prp43, and the U2, U5, and U6 snRNPs are recycled.

The recent cryo-EM snapshots of spliceosomal complexes at distinct stages have revealed the extremely dynamic nature of the spliceosome. In the first fully assembled spliceosomal complex, the pre-B complex, the 5'SS and the BS-A that later need to be brought together for the branching reaction are separated by more than 19 nm; and the BS-A is located more than 17 nm away from the future catalytic center (Figure 2.8). During the transition from pre-B to B complex, the 5'SS is handed over to the U6 snRNA, translocating by about 5 nm. The most dramatic RNA remodelling happens during B-to-B^{act} transition. During the process, the U6 snRNA undergoes a drastic remodelling, folding into a compact structure. The U6 nucleotides G46 and U74 that are initially separated by about 8 nm in the B complex are brought into proximity for the formation of the catalytic triad; the U2 snRNA moves closer to the tri-snRNP, which brings the BS-A closer to the catalytic center by more than 11 nm (Figure 2.8). In the B^{act} complex, the BS-A is still separated from the catalytic center by about 5 nm, and further structural remodelling is required for the first catalytic step. In addition to the drastic RNA remodelling, the spliceosomal proteins also undergo significant translocation. For instance, during pre-B to B transition, the Brr2 RNA helicase undergoes not only a translocation of more than 20 nm but also a 180° rotation; during B to B^{act} transition, the Brr2 RNA helicase further undergoes a complex rotational movement, as revealed by the recent pre-B^{act} structures[93]; concomitantly, the U2 SF3b complex undergoes about 11 nm translocation during B-to-B^{act} transition. Furthermore, the protein composition undergoes an extensive exchange. During B-to-B^{act} transition, about 25 proteins are displaced from the complex, and another 25 proteins stably dock, and many of the proteins bind in a mutually exclusive manner[1,71]. Although it
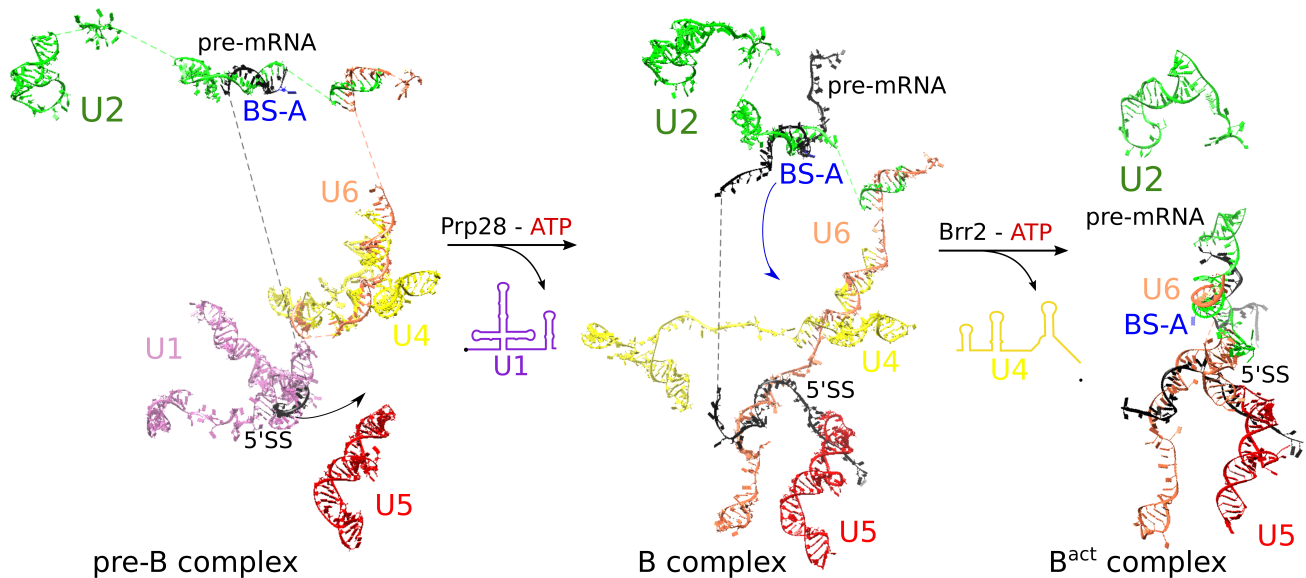
Figure 2.8: **3D view of the RNA remodelling events during spliceosome activation.**
RNA components of the pre-B complex (PDB:6QX9)[100] are shown on the left, B
complex in the middle (PDB:6AHD)[101], and B^{act} on the right (PDB:5Z58)[31,98].
The pre-mRNA, U1, U2, U4, U5, and U6 snRNAs are colored in black, purple,
green, yellow, red, and orange, respectively. The BS-A is highlighted in blue. The
translocations of the 5'SS and BS-A are depicted as arrows.

is long known that the structural remodelling requires RNA helicase action (see 2.2.6),
the mechanisms behind all the complex and dramatic structural rearrangements are
still poorly understood.

2.2.6 DExD/H-box RNA helicases

While the basic two-step splicing reaction does not require energy consumption,
the pre-mRNA-snRNA and the snRNA-snRNA network remodelling during the
splicing cycle requires ATP consumption. As briefly implicated in the last chapter,
these structural rearrangements are facilitated by the eight RNA helicases that are
able to bind and hydrolyze ATP[102]. The eight spliceosomal RNA helicase have
characteristic Asp-Glu-x-Asp/His (DExD/H) residues and all belong to superfamily
2 (SF2) helicases, which can be further divided into three categories: DEAD-box,
DEAH-box, and Ski2-like subfamilies. The three helicases important in the early
steps of spliceosome assembly, namely UAP56 (Sub2 in yeast), Prp5 (also known as
DDX46, the yeast name Prp5 will be used hereafter for simplicity), and Prp28, are
DEAD-box proteins. Brr2 belongs to the Ski2-like subfamily, and the last four helicases
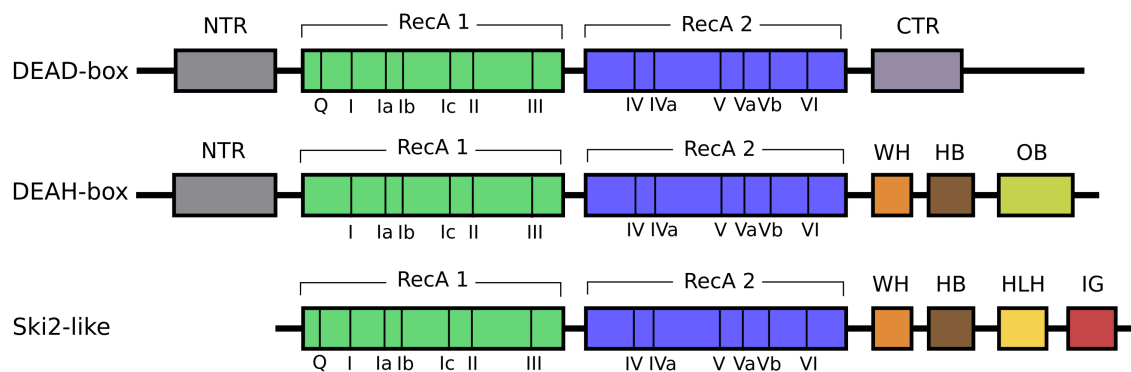
Figure 2.9: **Schematic representation of spliceosomal RNA helicases.** Domain organizations of subfamilies of spliceosomal RNA helicases. Motif Q, I, II, and IV are responsible for ATP binding and hydrolysis. Motif Ia, Ib, Ic, IV, IVa, V, and Vb are responsible for nucleic acid binding. Motif III and Va are responsible for communication between ATP-binding and nucleic-acid binding[103] (CTR, C-terminal region; HB, helix bundle domain; HLH, helix-loop-helix; IG, immunoglobulin-like; NTR, N-terminal region; OB, oligonucleotide/oligosaccharide-binding; WH, winged helix domain).

responsible for the catalysis and disassembly of spliceosomes, namely Prp2, Prp16, Prp22, and Prp43, belong to DEAH-box proteins. Although the spliceosomal RNA helicases vary in size and function, they all share the characteristic RecA domains and conserved motifs responsible for ATP binding and hydrolysis, as well as nucleic acid binding motifs (Figure 2.9). Previous biochemical studies and crystal structures of individual RNA helicases together with recent cryo-EM structures of spliceosomal complexes have provided insights into the structural consequences of the helicase activities.

Structural remodelling facilitated by RNA helicases

Prp5 is associated with the 17S U2 snRNP in humans and is believed to mediate a structural rearrangement of the U2 snRNP in an ATP-dependent way during the A complex formation[5–8]. In the absence of Cus2, the essentiality of the Prp5 ATPase activity can be bypassed[104]. Therefore Prp5 is implicated to play a role in displacing TAT-SF1 (human)/Cus2 (yeast) via its ATPase activity. During the A complex formation, UAP56 (Sub2 in yeast) is implicated in ejecting SF1 (Msl5 in yeast) from its position directly over the branch site, permitting stable U2 addition[85,105,106]. The mechanisms of Prp5 and UAP56's action are poorly understood. As DEAD-box RNA helicases, they are believed to bind ATP and promote RNA duplexes unwinding in a non-processive way[107,108]. However, the RNA target of Prp5 and UAP56 remains

to be identified. UAP56 was also shown to participate in unwinding the U4/U6 base-pairing[106]. However, the significance of this role is not established, as UAP56 was not found in purified tri-snRNP, pre-B, and B complex complexes.[91,100,109]

Prp28 plays a role in unwinding U1/5'SS helix during pre-B to B transition. The cryo-EM structures of the human pre-B complex indicate that upon ATP binding, the two RecA domains of Prp28 may clamp around the 5' end of the U1 snRNP, thus disrupting the U1/5'SS base-pairing directly[100].

Unlike DEAD-box RNA helicases that unwind RNA duplexes in a non-processive manner, Ski2-like and DEAH-box RNA helicases facilitate the unwinding of RNA duplexes or RNPs remodelling by binding to single-stranded RNA and translocating from 3' to 5'. Each step of translocation is assumed to consume one ATP[110]. In the B complex, Brr2 docks onto the single-stranded region of the U4 snRNA close to U4/U6 helix I[91]. The translocation of Brr2 on the U4 snRNA is believed to facilitate the unwinding of the U4/U6 duplex[111], displacing the U4 snRNA from the spliceosome and freeing the U6 snRNA to fold into the ISL.

After the B^{act} complex formation, Prp2, along with the G-patch protein Spp2, is recruited to the single-stranded intron downstream of the branch site[112,113]. Prp2 action then triggers the release of SF3a and SF3b complexes that sequester BS-A, liberating the reactants of the first step of splicing. The recent the high-resolution cryo-EM structure of the yeast B^{act} complex[110] has provided more detailed insights into the action of Prp2. With the help of the G-patch protein Spp2, Prp2 interacts with Hsh155 and Rse1 (yeast homologues of SF3B1 and SF3B3) and docks to the single-stranded intron downstream of the branch site via its RNA-binding channel. The ATP binding and hydrolysis trigger Prp2 RecA1 and RecA2 movement, leading to the translocation of Prp2 on the single-stranded RNA. The structure also shows L536 and R844 of Prp2 prevent backward sliding, thus ensuring the pre-mRNA translocates unidirectional towards the 3' end. Therefore, Prp2 triggers the remodeling of the B^{act} complex and the liberation of the BS-A from a distance, by acting from the 3' end of the pre-mRNA intron.

During the C to C* transition, Prp16 triggers the conformational changes that remove U2-BS helix from the catalytic center[101,114,115]. In the yeast C complex structure, the intron region downstream of the branch site points towards Prp16. Therefore, it was proposed that Prp16 may bind the pre-mRNA downstream of the branch site, and trigger the removal of the U2-BS helix in an ATP-dependent way[116]. The removal of the U2-BS helix may also displace two proteins, Yju2 and Cwc25, so that space is created for the 3' exon to enter the catalytic center.

After both steps of catalysis, Prp22 contributes to the release of the spliced mRNA from the spliceosome. The cryo-EM structures of the yeast C* and P complexes[117,118] show that Prp22 binds to the 3' exon via its two RecA domains and the CTD. By binding to the 3' exon, Prp22 translocates on the exons and facilitates the release of the mRNA[116].

After Prp22 action, Prp43 is required for the disassembly of the ILS complex, and the U2, U5, and U6 snRNPs are recycled. Biochemical studies in yeast[119] and the cryo-EM structures of human and yeast ILS complexes[120,121] suggest that Prp43 binds to 3' end of the U6 snRNA and triggers the disassembly of the complex from a distance. Analogous to Prp2 and Spp2, the G-patch protein Ntr1 is required for Prp43 activity[122-124]. Considering the structural conservation between Prp43 and Prp2, the translocation action of Prp43 on the single-stranded RNA is likely similar to Prp2.

fidelity control by RNA helicases

Since the spliceosome remodelling is coupled with the ATP-dependent actions of RNA helicases, many intermediate states that require RNA helicases also function as checkpoints during the splicing process. Many of the RNA helicases, including Prp5, Prp28, Prp16, Prp22, and Prp43, have been shown to contribute to proofreading the substrate[9,10,90,125,126]. A kinetic proofreading model has been proposed to explain how some of the RNA helicases play a role in fidelity control.

In this model, RNA helicases function as a "timer". For instance, Prp16 has been shown to play a role in discriminating against substrates undergoing slow 5'SS cleavage[125,126].

For a suboptimal substrate, Prp16 enables a structural rearrangement by its ATPase activity prior to 5'SS cleavage and thereby antagonizes splicing. Prp22 may contribute to proofreading the exon ligation step, possibly through a similar kinetic proofreading mechanism[127]. In the C* complex, Prp22 is recruited to the 3' exon. For a suboptimal substrate, the remodelling triggered by Prp22 would remove the 3'SS from the spliceosome before the second catalytic step, thereby preventing the exon ligation step. Nevertheless, whether the kinetic proofreading mechanism holds true for all the spliceosomal RNA helicases needs further investigation. Soo-Chen Cheng's lab[10] previously discovered that Prp5 may contribute to fidelity control of the branch site selection in an ATP-independent manner and therefore may have a different mechanism (see 2.3.5).

2.3 The U2 snRNP and the branch site recognition

As mentioned in 2.2, the U2 snRNP plays an essential role in the branch site selection during the A complex formation. In this chapter, I will summarize functional and structural insights into human and yeast U2 snRNP revealed by previous genetic, biochemical, and structural studies.

2.3.1 Biogenesis of the U2 snRNP

As briefly mentioned before in 2.2.4, spliceosomal snRNA precursors are initially transcribed by RNA polymerase II (except U6 and U6atac), with extra nucleotides at 3' end and capped with a monomethylated m7G cap[60]. snRNA precursors are then exported to the cytoplasm, where the RNP assembly takes place[60,61]. In the cytoplasm, the seven Sm proteins recognize the Sm site, forming the Sm-ring complex[128]. Association of the Sm complex is a prerequisite for 3' end trimming of the snRNA precursor and hyper-methylation of the cap, which gives rise to the 2,2,7-tri-methylated guanosine (m_3G) cap[59,129]. The m_3G cap and the Sm complex together form the nuclear localization signal, tagging the snRNA-Sm complex for re-import into the nucleus[130,131].

For the U2 biogenesis, U2-A' and U2-B" are believed to be transported into the nucleus independent of the U2 snRNA[132]. Within the Cajal body, a membrane-less organelle in the nucleus that is enriched in proteins and RNAs, U2-A' and U2-B" are recruited

to the U2 snRNA-Sm complex[133], forming a stable 12S particle. The mature U2 snRNP formation requires further recruitment of SF3a and SF3b complexes. *In vitro* reconstitution studies have shown that the SF3b complex is first recruited to the 12S particle, followed by the interaction of the SF3a complex, giving rise to the 17S U2 snRNP particle[72,134]. Immunoaffinity-purified human 17S U2 snRNPs also contain other associated factors, including Prp5 and TAT-SF1[6,71].

2.3.2 The U2 snRNA

The U2 snRNA can be roughly divided into 5' region that contains stem-loop 1 (SLI), branch-site interacting region, and SLII, and 3' region that contains the Sm site and 3' stem loops (Figure 2.10). Although the yeast U2 snRNA contains a large insertion at its 3' region, this insertion is not essential for splicing[135,136]. In fact, the shorter human snRNA can replace the yeast one *in vivo*, indicating the high conservation of the U2 snRNA among species[137].

SLI is at the 5' end of the U2 snRNA. This region undergoes dynamic remodelling during the splicing cycle (see Figure 2.7). Upon the formation of the pre-B complex, SLI must unwind for base-pairing with the U6 snRNA, forming U2/U6 helix II[89,100]. In the B^{act} complex, the U2 SLI region further base-pairs with the U6 snRNA, forming U2/U6 helix Ia and Ib[31,121] (see 2.2.5).

The branch-site interacting region (BSiR) of the U2 snRNA locates in between SLI and SLII regions. In the 17S U2 snRNP, the BSiR was initially proposed to be single-stranded[138]. More recently, studies in yeast suggest that the BSiR of U2 is sequestered in a branchpoint interaction stem-loop (BSL), resulting in a shortened SLI[4]. The predicted structure positions the 5'-GΨAGUA-3' sequence at the tip (i.e., loop region) of the BSL. Therefore, the tip of BSL may form initial base-pairing with the intron branch site. Although the highly conserved sequence would suggest a similar BSL in humans, the BSL has never been identified in humans. It was initially proposed that upon stable U2 addition into the A complex, the consensus 5'-GΨAGUA-3' sequence base-pairs with the branch site and forms a short U2-BS helix[42], in which the BS-A is bulged out. However, recent cryo-EM structures of
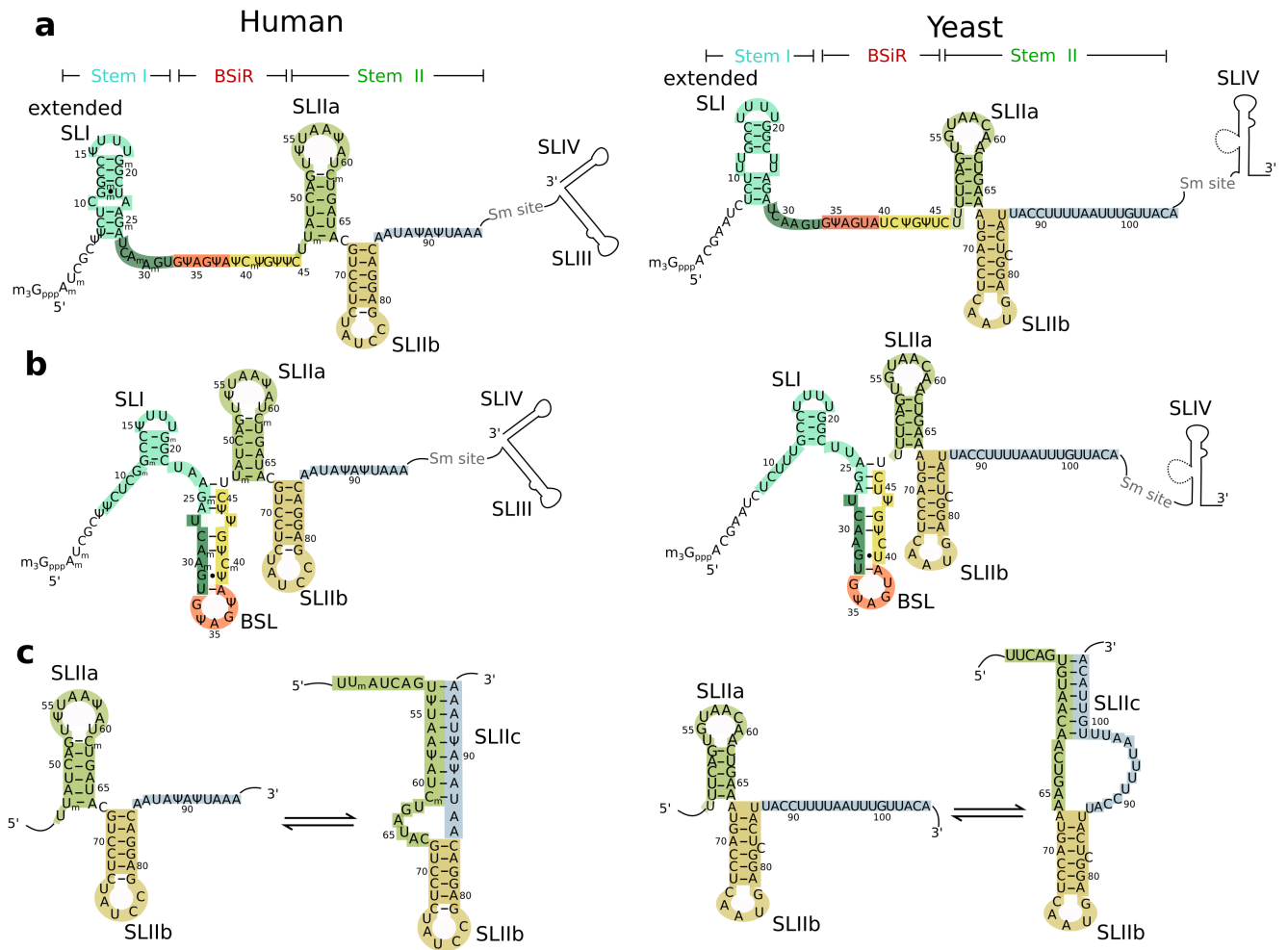
Figure 2.10: **Sequence and 2D structure of the human and yeast U2 snRNAs. a,b.** Two alternative conformations of the U2 snRNA that potentially form. The nucleotides in the extended SLI are labeled in teal; the nucleotides that later form the "*bona fide*" U2-BS helix in orange, the nucleotides that form extended part of the U2-BS helix in yellow; SLIIa in light green; SLIIb in light yellow; the nucleotides downstream of SLIIb that form part of SLIIc in light blue. Ψ is pseudouridine. **c.** Stem II region toggles between IIa/IIb and IIb/IIc conformations during splicing.

spliceosomal complexes showed an extended U2-BS helix, containing unexpected base-pairing between U2 and the intron upstream of the branch site[31,98]. But the exact function of the extended part of the U2-BS helix is currently unclear.

As with the SLI region that undergoes dynamic rearrangement during the splicing cycle, the SLII region also toggles between two mutually exclusive conformations, IIa/IIb and IIb/IIc (Figure 2.10c)[139,140]. *In vitro*, the SLIIa/SLIIc toggling is spontaneous[141]. In the spliceosome context, SLIIa is required for A complex formation and is present in the spliceosome until its catalytic activation[31,91,93,98,142]. SLIIc conformation occurs upon the B* complex formation, during which SF3a and SF3b complexes that bind to SLIIa are released by Prp2[143]. Therefore, the SLIIa/IIc remodelling is coordinated with the release of SF3a and SF3b complexes and the U2-BS helix docking into the catalytic center. Although genetic data suggest that SLIIc may transiently convert into SLIIa between first and second steps of splicing[139,140], the cryo-EM structures of B* to ISL complex suggest that after its formation, the IIb/IIc conformation is maintained until disassembly.

The human U2 snRNA 3' region contains the U2-Sm site, SLIII, and SLIV, while the yeast U2 snRNA has no SLIII and contains a large insertion at SLIV. Unlike the 5' region of the U2 snRNA that undergoes dynamic rearrangements, the U2 3' region remains rigid during the splicing cycle in both humans and yeast.

2.3.3 SF3a and SF3b, the major components of the 17S U2 snRNP

The SF3 complex was first isolated by chromatographic fractionation in 1991[144]. Further investigation of SF3 revealed two stable subcomplexes: SF3a and SF3b, which were subsequently shown to be the major components of the 17S U2 snRNP and play an essential role during A complex formation[6]. Recently, the crystal structures of the SF3a and SF3b complexes as well as the cryo-EM structures of spliceosomes have provided unprecedented structural information about the two complexes.

The SF3a complex is a heterotrimer, comprised of SF3A1, SF3A2, SF3A3 (Prp21, Prp11, and Prp9 in yeast, respectively). The crystal structure of the yeast SF3a

"core" domain provided the first structural information about the SF3a complex and showed detailed protein-protein interactions within the complex (Figure 2.11a)[145]. The overall structure is Y-shaped with the major body composed of the N-terminal domain of Prp9 and the SURP2 domain of Prp21, and a branch formed by a long α-helix of Prp21, to which the β-sandwich domain of Prp11 is bound. Despite large differences in protein sizes between corresponding yeast and human SF3a components, the structured domains and the interaction patterns are highly conserved, as shown by the yeast and human spliceosome structures. In the spliceosome context (from A to B[act] complex)[31,91,93,96–98,121,142,146], the C-terminal region of SF3A3/Prp9 binds to the SF3b complex and interacts with the 5' region of the U2 snRNA, while the N-terminal region of SF3A3/Prp9 binds to the Sm complex and SLIII (SLIV in yeast), thereby bridging the 5' and the 3' domain of the U2 snRNP. In addition, the zinc-finger (ZnF) domain of SF3A2/Prp11 stably binds to the extended part of the U2-BS helix and likely plays an important role during the U2-BS helix formation. Interestingly, in the B[act] complex, the N-terminus of SF3A2/Prp11 inserts into the catalytic center and contacts 5'SS, while the remainder of the protein – which is about 50 Å from the 5'SS – stays associated with the U2 snRNP. However, the functional significance of SF3A2/Prp11 linking the branch site and the 5'SS is yet unclear.

SF3b is comprised of SF3B1, SF3B2, SF3B3, SF3B4, SF3B5, PHF5A, and SF3B6/p14. The yeast SF3b complex (composed of Hsh155, Cus1, Rse1, Hsh49, Ysf3, Rds3) is structurally highly similar to the human counterpart but lacks the SF3B6/p14 subunit. A crystal structure of the stable core of isolated human SF3b, containing the C-terminal HEAT repeat domain of SF3B1, the WD40 protein SF3B3, SF3B5, and the ZnF protein PHF5A was recently reported (Figure 2.11a)[12]. The HEAT domain of SF3B1 (SF3B1[HEAT]) consists of 20 tandem HEAT repeats that adopt a superhelical structure and are contacted by SF3B3, SF3B5, and PHF5A (Figure 2.11a,b)[12]. SF3B3 contains 3 three WD40 (seven bladed β-propeller, WD40-A to WD40-C) domains, and binds SF3B5 and C-terminal HEAT repeats of SF3B1. In the spliceosome context, in which U2 is stably associated with the branch site, the SF3B1[HEAT] encompasses the U2-BS helix and the BS-A is accommodated in a binding pocket formed by PHF5A and residues of HEAT repeats 15-17 (H15-17), as revealed by the recent cryo-EM structures of the human and yeast B and B[act] complexes[31,96,97,121,146]. A comparison

of the structure of the SF3B1$^{\text{HEAT}}$ in the isolated complex with its structure in the human B$^{\text{act}}$ complex, shows a major difference in the conformation of the HEAT domain. While the SF3B1$^{\text{HEAT}}$ exhibits an open conformation in the isolated SF3b complex, in the spliceosomes it has a more closed conformation and tightly wraps around the U2-BS helix (Figure 2.11c)[12,31,98]. The same closed conformation of Hsh155$^{\text{HEAT}}$ (yeast homologue of SF3B1) is also found in yeast spliceosomes[96,97]. This conformational change is required to form the binding pocket for the BS-A (Figure 2.11c), and therefore it is proposed to occur upon formation of the U2-BS helix and plays a role in stabilizing the U2 base-pairing with the branch site. However, since the structure of the 17S U2 snRNP is lacking prior to this study, it is not clear if the closed conformation is indeed first generated in the A complex. Alternatively, it may even occur at an earlier stage, for example, during the 17S U2 assembly when the SF3b complex is first recruited. Furthermore, it is also not clear what triggers this functionally important conformational change of SF3B1. In addition to the HEAT domain, SF3B1 also contains N-terminal U2AF ligand motifs (ULMs) that allow it to bind to proteins containing U2AF homology motifs (UHMs), like U2AF and TAT-SF1 (see 2.3.4).

SF3B1 is of particular clinical interest, as it is one of the most common mutational targets in myelodysplastic syndromes (MDS)[147–151]. SF3B1 mutations are detected in a third of MDS patients, and most mutations are located at around H4-7, with the hotspots being E622, R625, H662, K666, K700, and G742[12,152] (Figure 2.11). These mutations are associated with usage of cryptic branch site and 3'SS located ∼20 nucleotides upstream of the canonical 3'SS, leading to aberrant splicing products[153,154]. Many hypotheses have been proposed to explain the underlying mechanism of how SF3B1 mutations lead to mis-splicing. A study in yeast suggests that these disease-related mutations are related to decreased binding affinity with Prp5, which plays an important role in proofreading the branch site[155] (see 2.3.5). Studies in humans suggest that SF3B1 hotspot mutations result in reduced SUGP1 binding, and SUGP1 mutations can mimic SF3B1 mutation effects[121,156]. However, SUGP1 is a poorly understood splicing factor that is implicated in activating a helicase that remains to be identified. Structural analysis of human and yeast spliceosomes implicates another possible mechanism, namely that the cancer-related mutations may alter the intron path
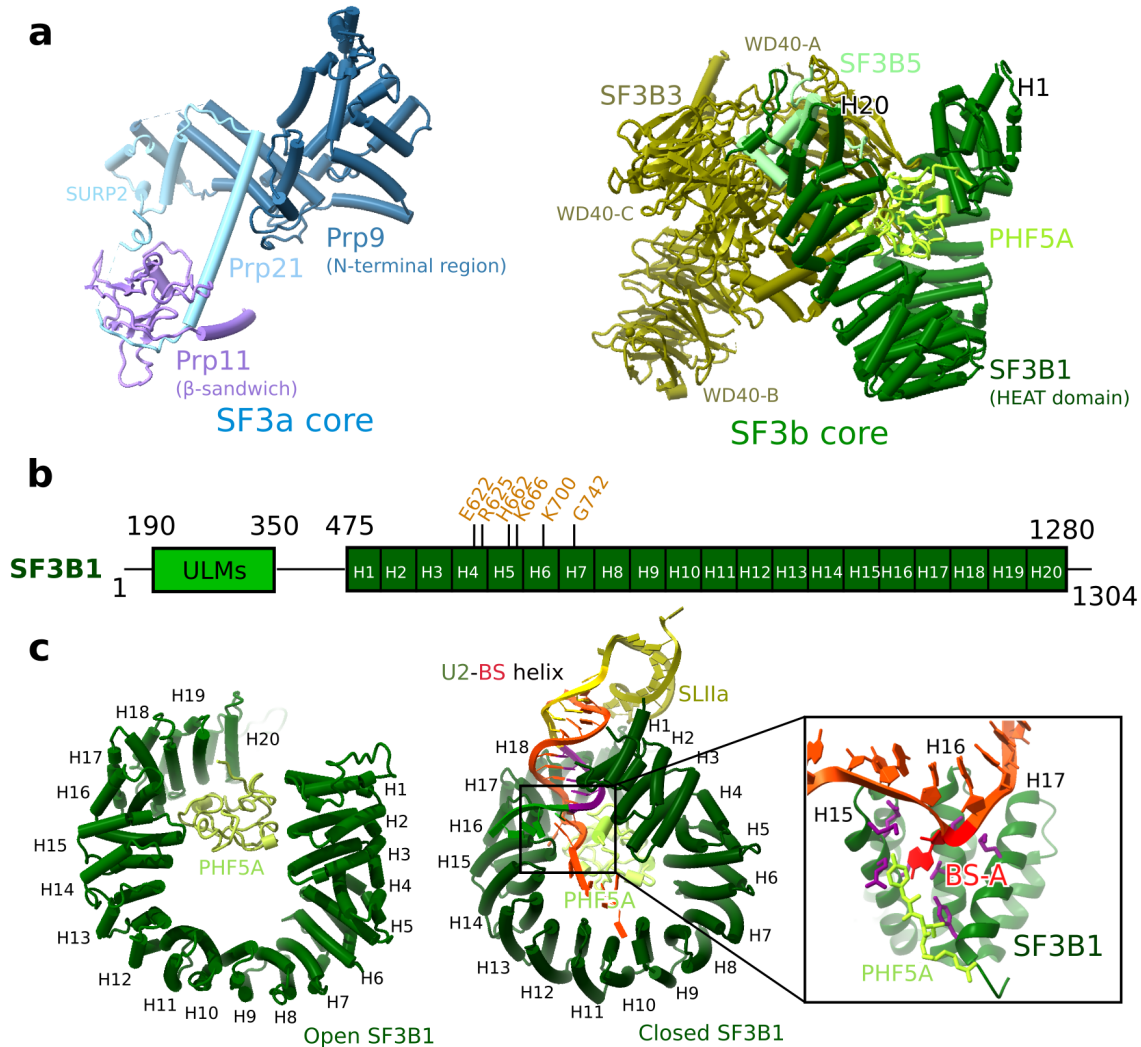
Figure 2.11: **Domain organization and structural comparison of SF3B1$^{\textbf{HEAT}}$ in isolated SF3b core and the spliceosome. a.** Crystal structures of the yeast SF3a core (4DGW, left) and the human SF3b core (5IFE, right) complexes. **b.** Domain Organization of SF3B1. Hotspot mutations are labeled in yellow. **c.** Open conformation of SF3B1$^{\text{HEAT}}$ in isolated SF3b core (left, PDB:5IFE) and closed conformation in the context of the spliceosomes, interacting with the U2-BS helix (right, PDB:6FF4). The BS-A (red) is inserted and sequestered in the protein pocket formed by SF3B1 H15-17 and PHF5A. The formation of the protein pocket requires closure of the HEAT domain.

downstream of the branch site, thereby affecting 3'SS usage[157]. But this hypothesis has not been tested experimentally. Thus, the detailed molecular mechanism of how SF3B1 mutations promote cryptic branch site selection requires further investigations.

2.3.4   TAT-SF1 and Prp5

Immunoaffinity purified human 17S U2 snRNP contains near stoichiometric amounts of TAT-SF1 and the DEAD-box ATPase Prp5[71] (Figure 2.12). TAT-SF1 (Cus2 in yeast) contains an N-terminal RRM and a C-terminal U2AF homology motif (UHM). The RRM binds U2 snRNA *in vitro*[158]. Biochemical studies in yeast show that Cus2 plays a key role in the formation of U2 snRNA SLIIa, as opposed to the competing SLIIc conformation, possibly through direct binding with U2 snRNA via its RRM[141,158]. However, although SLIIa is essential for U2 function[159], Cus2 is dispensable in yeast[104,158], and the exact molecular mechanism through which Cus2 helps the formation of SLIIa is unclear. The UHM has an RRM-like fold but has lost RNA binding ability[160]. Instead, the UHM is specialized in recognizing U2AF ligand motifs (ULMs) and thereby mediating protein-protein interactions[160]. Recent studies have also shown that the TAT-SF1/Cus2 UHM interacts with a ULM in the N-terminal domain of SF3B1/Hsh155 in both humans and yeast[161,162]. By binding to ULM, Cus2 enforces the ATPase activity of Prp5[162]. The human TAT-SF1 also contains a long C-terminal acidic domain that contains phosphorylation sites and may function in the regulation of TAT-SF1 activity[161,162].

Both human and yeast Prp5 contain a conserved N-terminal region. In *S. cerevisiae*, this N-terminal region binds the HEAT domain of Hsh155, and the conserved DPLD motif is essential for its binding[155]. The human and *S. pombe* Prp5 additionally contains an RS motif (Serine/Arginine-rich) at the very N-terminal end, which plays a role in tethering U1 and U2 snRNPs during A complex formation[163]. The C-terminal region of Prp5 contains highly conserved RecA domains. The crystal structure of the two Prp5 RecA domains adopts an unusual elongated open conformation[164]. Fluorescence resonance energy transfer (FRET) study indicates that a similar open conformation may exist in solution, but upon binding to ATP and RNA, the two RecA domains undergo a large-scale conformational change and adopt a closed conformation[165].
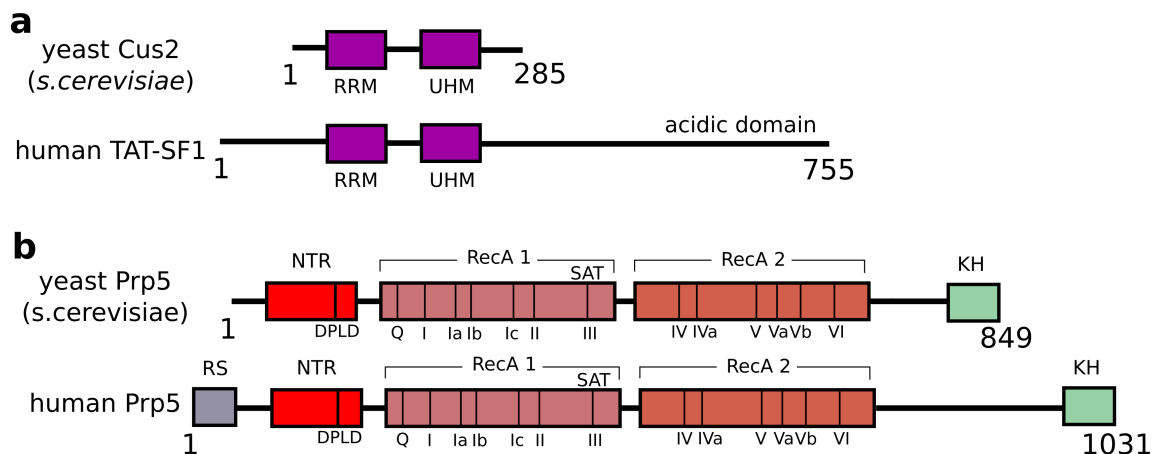
Figure 2.12: **Domain organization of Prp5 and TAT-SF1/Cus2. a.** Domain Organization of human TAT-SF1 and yeast Cus2. The human TAT-SF1 contains a long C-terminal acidic domain. **b.** Domain Organization of human and yeast Prp5. The human Prp5 contains an additional N-terminal RS domain. The conserved DPLD motif locates in the N-terminal region (NTR) and SAT motif in the RecA 1 domain. The C-terminal KH domain has an unknown function.

As the ATPase activity of Prp5 can be bypassed by Cus2 deletion or destabilization of the BSL[4,104], Prp5 is proposed to facilitate the unwinding of the BSL and the displacement of Cus2 prior/during A complex formation. However, this RNP remodelling process is poorly understood, due to limited structural information of early spliceosome complexes. In addition, Prp5 has a poorly understood ATP-independent role, as its physical presence is required even though its ATPase activity can be bypassed by Cus2 deletion[7].

2.3.5 Prp5 mediated proofreading of the branch site

In addition to RNP remodelling, Prp5 also plays an important role in fidelity control. Genetic screening of Prp5 alleles in yeast showed that some mutations of the conserved SAT motif (Figure 2.12b) in the RecA domain that lead to lower ATPase activity cause increased usage of pre-mRNA bearing a mutated branch site[9]. Therefore, it is proposed that Prp5 may contribute to proofreading the branch site by coordinating the ATP hydrolysis rate (i.e., kinetic proofreading, see 2.2.6). More recently, the Cheng lab discovered a mutually exclusive interaction of Prp5 and the tri-snRNP with the A complex in yeast[10]. Aberrant branch-site sequences lead to increased retention of Prp5 and a decreased recruitment of the tri-snRNP, suggesting that the presence

of Prp5, instead of its ATPase activity, is essential for its proofreading role. The
Cheng lab further showed that this proofreading mediated by Prp5 is irrelevant to its
ATPase activity in the absence of Cus2. However, since the cryo-EM structures of
yeast E and A complexes lack Prp5, the detailed molecular mechanisms of how Prp5
contributes to proofreading the branch site during early spliceosome assembly are
currently unclear. Therefore, a cryo-EM structure of an intermediate complex prior
to/during Prp5-mediated fidelity control would largely help to understand the proof-
reading mechanism mediated by Prp5 and by extrapolation, other DEAD-box proteins.

2.4   Rationale of the work

Distinct major spliceosomal intermediate complexes – i.e., E, A, pre-B, B, B^{act}, B*,
C, C*, P, and the ILS have been purified. Since 2015, for most of these splicing cycle
intermediates, high-resolution cryo-EM structures have been obtained. However, while
these recent structures have provided detailed insights into the catalytic activation
steps, the catalytic mechanisms, and late disassembly of the splicing process, the
structural information of early spliceosomal complexes is largely lacking, especially
in humans. Indeed, prior to this thesis, no high-resolution cryo-EM structures of the
human and the yeast isolated U2 snRNP are available. Using single-particle cryo-EM
as a tool, we aim to resolve the first high-resolution human 17S U2 snRNP structure
and shed new light on the early spliceosome assembly.

During the A complex formation, the DEAD-box protein Prp5 not only mediates RNP
rearrangements but also plays an important proofreading role during the branch site
recognition, at least in yeast. However, as the cryo-EM structures of the yeast E
and A complexes lack Prp5 density, it is currently unclear how Prp5 contributes to the
fidelity of the branch site selection mechanistically. To elucidate the poorly understood
molecular mechanism of how Prp5 contributes to branch site proofreading during the
A complex formation, we aim to resolve the structure of a yeast early spliceosome
intermediate by single-particle cryo-EM, which is after U2 addition but prior to Prp5
dissociation.

3   Results

The results section of this thesis is based on two first-author publications:

(i) "Molecular architecture of the human 17S U2 snRNP" (**Zhang et al., Nature, 2020**);

(ii) "Structural insights into how Prp5 proofreads the pre-mRNA branch site" (**Zhang et al., Nature, 2021**)

The two publications are attached below as they were originally published.

3.1 Molecular architecture of the human 17S U2 snRNP (*published manuscript*)

Zhenwei Zhang[1#], Cindy L. Will[2#], Karl Bertram[1], Olexandr Dybkov[2], Klaus Hartmuth[2], Dmitry E. Agafonov[2], Romina Hofele[3], Henning Urlaub[3,4], Berthold Kastner[2], Reinhard Lührmann[2*], and Holger Stark[1*]

[1]Department of Structural Dynamics, MPI for Biophysical Chemistry, Göttingen, Germany.

[2]Cellular Biochemistry, MPI for Biophysical Chemistry, Göttingen, Germany.

[3]Bioanalytical Mass Spectrometry, MPI for Biophysical Chemistry, Göttingen, Germany.

[4]Bioanalytics Group, Institute for Clinical Chemistry, University Medical Center Göttingen, Göttingen, Germany.

# These authors contributed equally to this work.

*Correspondence to: hstark1@gwdg.de; reinhard.luehrmann@mpibpc.mpg.de

Article

Molecular architecture of the human 17S U2 snRNP

Zhenwei Zhang[1,6], Cindy L. Will[2,6], Karl Bertram[1], Olexandr Dybkov[2], Klaus Hartmuth[2], Dmitry E. Agafonov[2], Romina Hofele[3,5], Henning Urlaub[3,4], Berthold Kastner[2], Reinhard Lührmann[2✉] & Holger Stark[1✉]

The U2 small nuclear ribonucleoprotein (snRNP) has an essential role in the selection of the precursor mRNA branch-site adenosine, the nucleophile for the first step of splicing[1]. Stable addition of U2 during early spliceosome formation requires the DEAD-box ATPase PRP5[2–7]. Yeast U2 small nuclear RNA (snRNA) nucleotides that form base pairs with the branch site are initially sequestered in a branchpoint-interacting stem–loop (BSL)[8], but whether the human U2 snRNA folds in a similar manner is unknown. The U2 SF3B1 protein, a common mutational target in haematopoietic cancers[9], contains a HEAT domain (SF3B1[HEAT]) with an open conformation in isolated SF3b[10], but a closed conformation in spliceosomes[11], which is required for stable interaction between U2 and the branch site. Here we report a 3D cryo-electron microscopy structure of the human 17S U2 snRNP at a core resolution of 4.1 Å and combine it with protein crosslinking data to determine the molecular architecture of this snRNP. Our structure reveals that SF3B1[HEAT] interacts with PRP5 and TAT-SF1, and maintains its open conformation in U2 snRNP, and that U2 snRNA forms a BSL that is sandwiched between PRP5, TAT-SF1 and SF3B1[HEAT]. Thus, substantial remodelling of the BSL and displacement of BSL-interacting proteins must occur to allow formation of the U2–branch-site helix. Our studies provide a structural explanation of why TAT-SF1 must be displaced before the stable addition of U2 to the spliceosome, and identify RNP rearrangements facilitated by PRP5 that are required for stable interaction between U2 and the branch site.

To our knowledge, at present no high-resolution, cryo-electron microscopy (cryo-EM) structure of the human 17S U2 snRNP, the major subunit of the spliceosomal E and A complexes (Extended Data Fig. 1), is available. We thus determined its 3D structure by cryo-EM (Extended Data Table 1, Extended Data Fig. 2). Consistent with previous low-resolution electron microscopy structures of isolated U2 snRNPs[12], and its overall structure in human spliceosomes[13–16], human 17S U2 exhibits a bipartite structure, with a 3′ and 5′ domain bridged by several density elements (Fig. 1a). The structure of a major part of the 5′ domain could be determined at an overall resolution of 4.1 Å (Extended Data Fig. 2). However, the 3′ domain and connecting bridges, and parts of the 5′ domain are more dynamic and thus less well-resolved (Fig. 1a, Extended Data Fig. 2). Because of the limited resolution in the more dynamic regions of 17S U2, we used an integrated structural biology approach, fitting known X-ray structures or homology models of structured regions of U2 components into the electron-microscopy density map (Extended Data Table 2), in combination with protein crosslinking coupled with mass spectrometry (Extended Data Fig. 3, Supplementary Table 1) and other biochemical data, to generate a pseudo-atomic model for the well-defined regions of the particle (Fig. 1b).

SF3B1 has an open conformation in 17S U2

The 17S U2 5′ domain consists predominantly of the SF3b complex. SF3B3, PHF5A, SF3B5, and SF3B1[HEAT], which comprise the SF3b core, are located in the most well-defined region of the particle (Fig. 1). The super-helical structure of SF3B1[HEAT], which is composed of 20 tandem HEAT repeats (HRs), exhibits an open conformation in isolated SF3b[10] (Extended Data Fig. 1d). In B and B[act] spliceosomal complexes, in which U2 stably binds the branch site (BS), SF3B1[HEAT] adopts a closed conformation, encompassing the U2–BS helix and binding the BS adenosine (BS-A) in a pocket formed by PHF5A and the HR15–HR17 of SF3B1 (Extended Data Fig. 1d). In 17S U2 and in the isolated SF3b core crystal, SF3B1[HEAT] has a very similar open conformation, except that HR16 is completely structured in 17S U2 (Extended Data Fig. 4a–c). The structural organization of SF3B5, PHF5A and SF3B3 is also highly similar (Extended Data Fig. 4d). Crosslinking suggests that SF3B6 (also known as P14 or SF3b14a) is also located in a similar region in 17S U2 and isolated SF3b (Extended Data Fig. 4e). Thus, the structure of the SF3b core does not change substantially during 17S U2 assembly and, furthermore, the crystal structure of SF3b represents a functionally relevant conformation. SF3B1 is the target of several precursor mRNA

[1]Department of Structural Dynamics, MPI for Biophysical Chemistry, Göttingen, Germany. [2]Cellular Biochemistry, MPI for Biophysical Chemistry, Göttingen, Germany. [3]Bioanalytical Mass Spectrometry, MPI for Biophysical Chemistry, Göttingen, Germany. [4]Bioanalytics Group, Institute for Clinical Chemistry, University Medical Center Göttingen, Göttingen, Germany. [5]Present address: AstraZeneca, Gaithersburg, MD, USA. [6]These authors contributed equally: Zhenwei Zhang, Cindy L. Will. ✉e-mail: reinhard.luehrmann@mpi-bpc.mpg.de; hstark1@gwdg.de
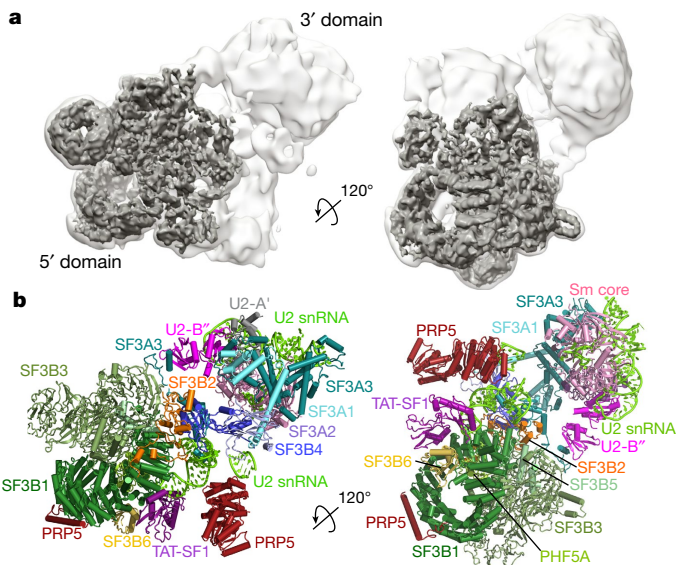
Fig. 1 | 3D cryo-EM model of human 17S U2 snRNP. a, **b**, Different views of the human 17S U2 snRNP cryo-EM density map (**a**) and molecular architecture (**b**). Dark grey denotes better resolved densities.

(pre-mRNA) splicing modulators[9]. The open conformation of SF3B1[HEAT] indicates that splicing modulators such as pladienolide B can also interact with it in the 17S U2 snRNP (Extended Data Fig. 4f), and furthermore supports the idea that the SF3B1 conformational change first occurs during stable U2 addition to the spliceosome[17]. The U2 5′ domain is connected by three main bridges to the 3′ domain (Extended Data Fig. 5), which consists of the U2 Sm core RNP, and U2 snRNA stem–loops (SL) III and IV bound by U2-A' and U2-B'' (Fig. 1, Extended Data Fig. 5). Similar to the B and B[act] complexes, SF3a proteins, which comprise part of the 3′ domain, have a key role in bridging the U2 Sm core with SF3b in 17S U2[13–16] (Extended Data Fig. 5).

A BSL forms and is bound by protein

In the spliceosome, SF3b contacts SLIIa of U2 snRNA[14,16,18]. On the basis of its length and conformation (Fig. 2, Extended Data Fig. 6a, b), we could unambiguously place U2 SLIIa, as opposed to SLIIb, into one of the two well-resolved helical density elements located close to the C-terminal HEAT repeats of SF3B1 (Extended Data Fig. 6c, d). The high conservation between yeast and humans of U2 nucleotides that form the BSL strongly suggest that the latter also forms in the human U2 snRNP[8]. The location of the helical density that flanks SLIIa indicates that it contains nucleotides 5′ of SLIIa and, as it is a direct continuation of SLIIa, that it accommodates the BSL, as opposed to the mutually

exclusive, extended SLI that is separated from SLIIa by 20 nucleotides (Extended Data Fig. 6a, b). Indeed, a modelled BSL fits well in this density (Extended Data Fig. 6e). Moreover, nucleotides 42 to 45 on one BSL strand, which form part of the extended U2–BS helix in the spliceosome, have nearly the same position relative to SLIIa and SF3B1, both in 17S U2 and the spliceosome after U2–BS helix formation (Fig. 2). The BSL is sandwiched between the SF3B1 C-terminal HEAT repeats and other proteins (Fig. 3), and its loop is also sequestered by protein and thus inaccessible for base pairing interactions. Indeed, chaperoning away reactive regions of spliceosomal RNAs and of the pre-mRNA is a common mechanism used by the spliceosome[11,19]. The base of the BSL stem is contacted by a short helix of SF3A3 (designated the separator helix), which ensures an only eight base-pair length of the BSL (Extended Data Fig. 6a). This SF3A3 helix also interacts with SLIIa and its position relative to the latter is maintained in the spliceosome until its catalytic activation[14,16,18] (Extended Data Fig. 6g, h).

PRP5 is near the BSL and encircles SF3B1

Human 17S U2 snRNPs contain TAT-SF1, the function of which in splicing remains unclear, and the DEAD-box ATPase PRP5[6,20]. Prp5 is required for stable U2 addition during formation of the spliceosomal A complex and facilitates a conformational change in U2 snRNP[2–7]. It is also implicated in the displacement of TAT-SF1 (Cus2 in *Saccharomyces cerevisiae*)[5,21], a prerequisite for U2 incorporation into the spliceosome[5,22]. However, the molecular mechanisms by which PRP5 promotes stable U2–BS interaction, and the precise nature of RNP rearrangements that it facilitates, remain poorly understood. In the 17S U2 structure, the PRP5 RecA domains can be fitted into cryo-EM density close to the BSL, U2 snRNA SLIIa and SLIIb, and adjacent to TAT-SF1 (Fig. 3, Extended Data Fig. 7a–c). Consistent with biochemical studies in yeast[23], an extended N-terminal α-helix of PRP5 interacts with SF3B1 HR9–HR12 (Fig. 3a, Extended Data Fig. 7d). Protein crosslinks suggest that PRP5 residues both N- and C-terminal of this α-helix also interact with HR12–HR15 and HR1–HR6, respectively (Fig. 3, Extended Data Fig. 7e). Thus, the N-terminal region of PRP5 seems to encompass SF3B1[HEAT], which suggests that it aids in stabilizing its open conformation. The most common (hotspot) cancer-related, point mutations in SF3B1 mainly cluster in or near HR6[10]. Notably, our crosslinking data suggest that the highly conserved PRP5 DPLD motif, which is essential for stable PRP5 interaction with U2 snRNP[24], is located in the proximity of HR6 (Extended Data Fig. 7e), consistent with studies in yeast showing that cancer-related SF3B1 mutations directly destabilize binding of PRP5[23,25].

TAT-SF1 contacts the BSL and SF3B1

RRM1 of TAT-SF1 is located adjacent to SF3B1 HR15–HR16 (Fig. 3, Extended Data Fig. 7f, g), which are thought to act as a hinge[17]. Thus, TAT-SF1 could potentially inhibit their hinge-like function and prevent the closure of SF3B1[HEAT] needed to form the BS-A binding pocket.
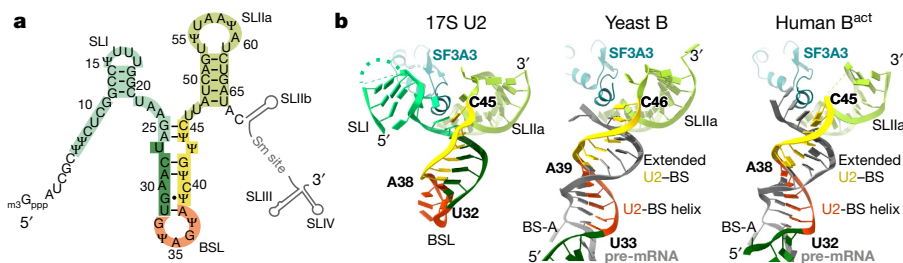
Fig. 2 | A BSL forms in human 17S U2 snRNP. a, 2D structure of human U2 snRNA. SLI nucleotides are in teal; SLIIa nucleotides are in light green. BSL nucleotides that later form the U2–BS helix are in orange, and the extended region of U2–BS helix are in yellow. Remaining BSL nucleotides are in dark green. Ψ denotes pseudouridine. Methylation of U2 bases is not shown (see Extended Data Fig. 6). **b**, 3D structure of U2 SLI, BSL and SLIIa in human 17S U2, and yeast B (PDB code 5NRL) and human B[act] (PDB code 6FF4) complexes.
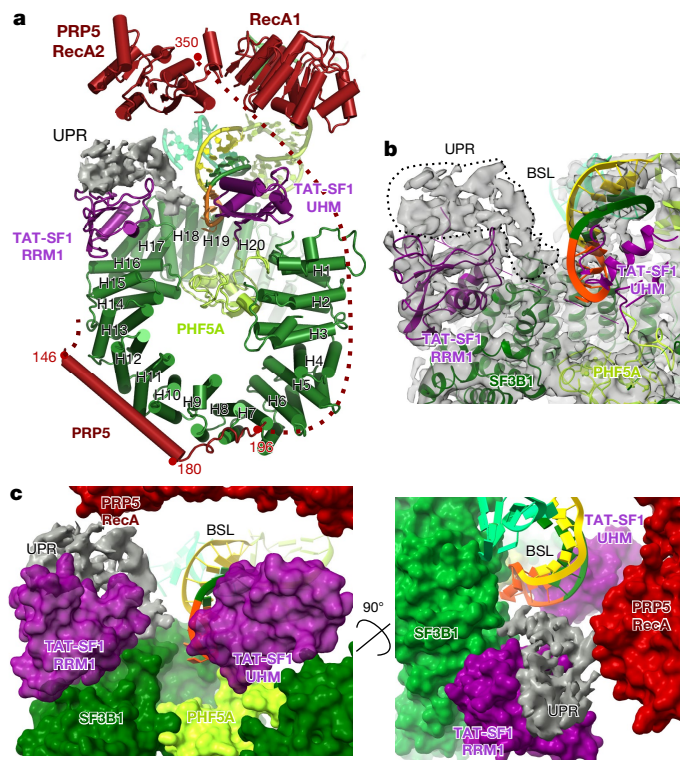
Fig. 3 | PRP5 and TAT-SF1 are located near the BSL and interact with SF3B1HEAT. a, Location of the PRP5 RecA1 and RecA2 domains, and interaction of the PRP5 α-helix with SF3B1HEAT, and TAT-SF1 RRM1 and UHM with SF3B1HEAT and the BSL. Dotted line denotes the potential path of unstructured PRP5 regions, based on crosslinking data. **b**, Fit of TAT-SF1RRM1 into cryo-EM density adjacent to SF3B1 HR15–HR16. A density element that could not be unambiguously defined, but is probably part of TAT-SF1, contacts the BSL loop. **c**, The BSL is sequestered by PRP5, TAT-SF1, SF3B1 and the unassigned protein region (UPR). Grey denotes cryo-EM density of the unassigned protein region. Coloured surfaces are derived from fitted protein models.

TAT-SF1RRM1 is closely associated with a density element that interacts directly with the BSL loop and is contacted by PRP5RecA2 (Fig. 3). This density element is probably part of TAT-SF1, as electron-microscopy density can be traced in a continuous manner from it to density comprising TAT-SF1RRM (Fig. 3). However, we cannot unambiguously identify its sequence. Regardless of its nature, this protein module helps to stabilize the position of the BSL loop close to SF3B1HEAT. TAT-SF1 contains

a U2AF homology motif (UHM) (Extended Data Fig. 7a) that has affinity for the five U2AF ligand motifs (ULMs) present in the SF3B1 N-terminal domain, but preferentially binds to ULM5, located closest to the HEAT domain[26]. Consistent with protein–protein crosslinks, TAT-SF1UHM can be fitted into a less-well resolved density element near the BSL and PRP5 RecA domains (Fig. 3, Extended Data Fig. 7c, h). Together, the location of PRP5 and TAT-SF1 suggests that they aid in stabilizing both the BSL and SF3B1 open conformation, and thus must be displaced to allow a stable U2–BS interaction.

Model of PRP5 remodelling of U2 snRNP

A comparison of our 17S U2 snRNP structure and the structure of U2 in the spliceosome indicates that a major RNP rearrangement must occur to free those U2 nucleotides that subsequently form the U2–BS helix and allow stable U2 addition during formation of the A complex (Extended Data Fig. 8a). U2 snRNP first associates with the spliceosomal E complex via protein–protein contacts, including those between PRP5 and the U1 snRNP[7], and potentially between U2 proteins and SF1 (also known as MBBP) bound to the BS, that would bring U2 into the vicinity of the BS[27] (Fig. 4). If accessible, BSL loop nucleotides could in principle form base pairs with the BS without disruption of the BSL[8]. However, in human 17S U2, the BSL is sandwiched between the PRP5 RecA domains and TAT-SF1UHM on one side, and HR16–HR19 of SF3B1 and TAT-SF1RRM1 on the other, with the unassigned protein region directly contacting the BSL loop. PRP5 is a DEAD-box ATPase and thus is expected to bind double-stranded RNA[3,28]. However, the nature of the RNA bound by PRP5 is not clear from our U2 structure and PRP5 may first bind RNA after 17S U2 interacts with the E complex. Several DEAD-box proteins disrupt protein–RNA interactions, in some cases apparently even without unwinding dsRNA[28]. Our data are consistent with a model in which, after U2 incorporation into the E complex, ATP binding would lead to engagement of PRP5 with the BSL or adjacent RNA nucleotides, with hydrolysis leading to displacement or release of BSL-interacting proteins, including TAT-SF1 and the RecA domains of PRP5 itself, without disrupting the BSL (Fig. 4a, b). This would allow initial U2–BS base pairing via the BSL loop, which would also require displacement of SF1 from the pre-mRNA BS nucleotides to free them for base pairing interactions[29]. In yeast, Prp5 ATPase activity is not required if CUS2 (the yeast TAT-SF1 homologue) is no longer present[5]. We thus propose that the BSL, which is no longer stabilized by protein, will subsequently become flexible, and thus that its unwinding and the concomitant formation of the extended U2–BS helix, may not require exogenous energy (Fig. 4b–d). However, at present we cannot exclude that PRP5 destabilizes the BSL in an ATP-dependent manner, and that this leads to the concomitant displacement of TAT-SF1 and other BSL-interacting
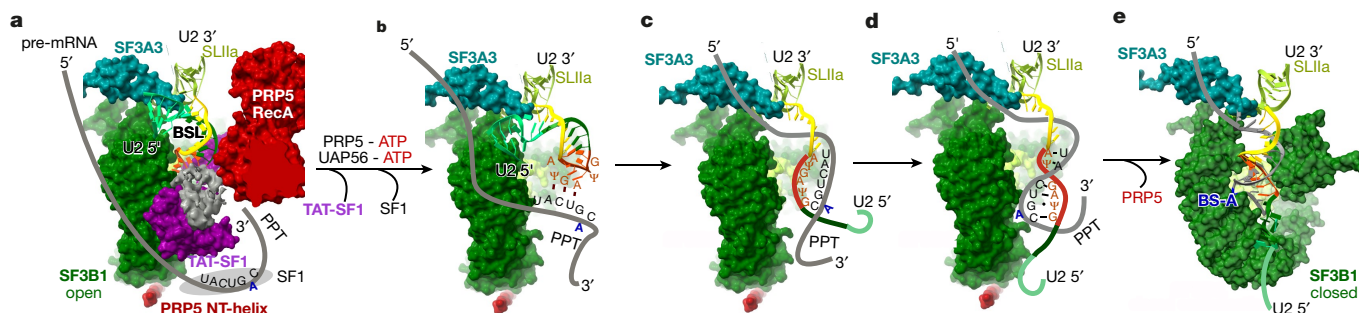
Fig. 4 | Model of PRP5-mediated remodelling events leading to stable U2–BS interaction. a–e, Model of U2 snRNP rearrangements during the transition from the spliceosomal E to A complex. U1 snRNP and U2AF are omitted for simplicity. **a, b**, ATP hydrolysis by PRP5 leads to the removal or displacement of TAT-SF1, the PRP5 RecA domains, and the unassigned protein region (grey) from the BSL, and BSL contact with SF3B1HEAT is disrupted. This frees the BSL loop for initial base pairing with the BS and leads to destabilization of its stem. **b, c**, Movement of the U2 5′ end downward aids BSL unwinding. **c, d**, Rotation of the SLI-containing U2 5′ end generates helical turns required to form the extended U2–BS helix. **d, e**, A conformational change in SF3B1 clamps the newly formed U2–BS helix within its HEAT domain and leads to PRP5 release. NT-helix, N-terminal helix; PPT, polypyrimidine tract.

proteins. For topological reasons, formation of the extended U2–BS helix would require that the 5′ strand of the BSL no longer contacts the SF3B1 C-terminal HEAT repeats. This would be a prerequisite to free the SLI-containing 5′ end of U2 snRNA for its proposed downward movement (Fig. 4b–d, Extended Data Fig. 8b, c). Rotational movement of the 5′ end of the U2 snRNA would then facilitate formation of the extended U2–BS helix (Fig. 4b–d, Extended Data Fig. 8c). The latter would then move back towards the C-terminal SF3B1 HEAT repeats such that U2 nucleotides 42–45 occupy the same position as they did before U2–BS formation, and the base of the bulged BS-A would be positioned at the hinge region between HR15 and HR17 and close to PHF5A. We propose that docking of the BS-A to the hinge region may trigger several coordinated movements of the HEAT domain leading to its closed conformation and clamping of its terminal HRs onto the extended part of the U2–BS helix (Fig. 4e).

In yeast, Prp5 also functions in an ATP-independent manner during formation of the A complex and remains bound to the latter if the BS is mutated[5,30]. In 17S U2, the N-terminal region of PRP5 makes numerous contacts with SF3B1^{HEAT} and therefore potentially stabilizes its open conformation. Thus, there may be competition between the maintenance of an open conformation, facilitated by the PRP5 N-terminal region, and the switch to a closed conformation with the BS-A tightly clamped. In this model, only if an appropriate U2–BS helix is formed and the BS-A is properly docked, will the PRP5-SF3B1^{HEAT} interaction be dissolved, leading to PRP5 release. In this way, PRP5 could potentially proofread stable accommodation of the BS-A during A complex formation. Given the weak conservation of the mammalian BS, further studies are needed to better understand how PRP5 can potentially distinguish between ideal U2–BS duplexes and those containing several mismatches.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2344-3.

1. Will, C. L. & Lührmann, R. Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* **3**, a003707 (2011).
2. Ruby, S. W., Chang, T. H. & Abelson, J. Four yeast spliceosomal proteins (PRP5, PRP9, PRP11, and PRP21) interact to promote U2 snRNP binding to pre-mRNA. *Genes Dev.* **7**, 1909–1925 (1993).
3. O'Day, C. L., Dalbadie-McFarland, G. & Abelson, J. The *Saccharomyces cerevisiae* Prp5 protein has RNA-dependent ATPase activity with specificity for U2 small nuclear RNA. *J. Biol. Chem.* **271**, 33261–33267 (1996).
4. Abu Dayyeh, B. K., Quan, T. K., Castro, M. & Ruby, S. W. Probing interactions between the U2 small nuclear ribonucleoprotein and the DEAD-box protein, Prp5. *J. Biol. Chem.* **277**, 20221–20233 (2002).
5. Perriman, R., Barta, I., Voeltz, G. K., Abelson, J. & Ares, M., Jr. ATP requirement for Prp5p function is determined by Cus2p and the structure of U2 small nuclear RNA. *Proc. Natl Acad. Sci. USA* **100**, 13857–13862 (2003).
6. Will, C. L. et al. Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5p homologue and an SF3b DEAD-box protein. *EMBO J.* **21**, 4978–4988 (2002).
7. Xu, Y. Z. et al. Prp5 bridges U1 and U2 snRNPs and enables stable U2 snRNP association with intron RNA. *EMBO J.* **23**, 376–385 (2004).
8. Perriman, R. & Ares, M., Jr. Invariant U2 snRNA nucleotides form a stem loop to recognize the intron early in splicing. *Mol. Cell* **38**, 416–427 (2010).
9. Bonnal, S., Vigevani, L. & Valcárcel, J. The spliceosome as a target of novel antitumour drugs. *Nat. Rev. Drug Discov.* **11**, 847–859 (2012).
10. Cretu, C. et al. Molecular architecture of SF3b and structural consequences of its cancer-related mutations. *Mol. Cell* **64**, 307–319 (2016).
11. Kastner, B., Will, C. L., Stark, H. & Lührmann, R. Structural insights into nuclear pre-mRNA splicing in higher eukaryotes. *Cold Spring Harb. Perspect. Biol.* **11**, a032417 (2019).
12. Krämer, A., Grüter, P., Gröning, K. & Kastner, B. Combined biochemical and electron microscopic analyses reveal the architecture of the mammalian U2 snRNP. *J. Cell Biol.* **145**, 1355–1368 (1999).
13. Bertram, K. et al. Cryo-EM structure of a pre-catalytic human spliceosome primed for activation. *Cell* **170**, 701–713.e11 (2017).
14. Haselbach, D. et al. Structure and conformational dynamics of the human spliceosomal B^{act} complex. *Cell* **172**, 454–464.e11 (2018).
15. Zhan, X., Yan, C., Zhang, X., Lei, J. & Shi, Y. Structures of the human pre-catalytic spliceosome and its precursor spliceosome. *Cell Res.* **28**, 1129–1140 (2018).
16. Zhang, X. et al. Structure of the human activated spliceosome in three conformational states. *Cell Res.* **28**, 307–322 (2018).
17. Cretu, C. et al. Structural basis of splicing modulation by antitumor macrolide compounds. *Mol. Cell* **70**, 265–273.e8 (2018).
18. Plaschka, C., Lin, P. C. & Nagai, K. Structure of a pre-catalytic spliceosome. *Nature* **546**, 617–621 (2017).
19. Papasaikas, P. & Valcárcel, J. The spliceosome: The ultimate RNA chaperone and sculptor. *Trends Biochem. Sci.* **41**, 33–45 (2016).
20. Agafonov, D. E. et al. Semiquantitative proteomic analysis of the human spliceosome via a novel two-dimensional gel electrophoresis method. *Mol. Cell. Biol.* **31**, 2667–2682 (2011).
21. Talkish, J. et al. Cus2 enforces the first ATP-dependent step of splicing by binding to yeast SF3b1 through a UHM-ULM interaction. *RNA* **25**, 1020–1037 (2019).
22. Xu, Y. Z. & Query, C. C. Competition between the ATPase Prp5 and branch region-U2 snRNA pairing modulates the fidelity of spliceosome assembly. *Mol. Cell* **28**, 838–849 (2007).
23. Tang, Q. et al. SF3B1/Hsh155 HEAT motif mutations affect interaction with the spliceosomal ATPase Prp5, resulting in altered branch site selectivity in pre-mRNA splicing. *Genes Dev.* **30**, 2710–2723 (2016).
24. Shao, W., Kim, H. S., Cao, Y., Xu, Y. Z. & Query, C. C. A. A U1-U2 snRNP interaction network during intron definition. *Mol. Cell. Biol.* **32**, 470–478 (2012).
25. Carrocci, T. J., Zoerner, D. M., Paulson, J. C. & Hoskins, A. A. SF3b1 mutations associated with myelodysplastic syndromes alter the fidelity of branchsite selection in yeast. *Nucleic Acids Res.* **45**, 4837–4852 (2017).
26. Loerch, S. et al. The pre-mRNA splicing and transcription factor Tat-SF1 is a functional partner of the spliceosome SF3b1 subunit via a U2AF homology motif interface. *J. Biol. Chem.* **294**, 2892–2902 (2019).
27. Crisci, A. et al. Mammalian splicing factor SF1 interacts with SURP domains of U2 snRNP-associated proteins. *Nucleic Acids Res.* **43**, 10456–10473 (2015).
28. Pan, C. & Russell, R. Roles of DEAD-box proteins in RNA and RNP Folding. *RNA Biol.* **7**, 667–676 (2010).
29. Liu, Z. et al. Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science* **294**, 1098–1102 (2001).
30. Liang, W. W. & Cheng, S. C. A novel mechanism for Prp5 function in prespliceosome formation and proofreading the branch site sequence. *Genes Dev.* **29**, 81–93 (2015).

Article

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Affinity purification of 17S U2 snRNPs

HeLa S3 cells were obtained from GBF, Brunswick (currently Helmholtz Center for Infection Research) and tested negative for mycoplasma. To isolate 17S U2 snRNPs, HeLa nuclear extract was prepared as previously described[31]. Extract was loaded onto a 17–30% (w/v) sucrose gradient containing GK5M buffer (20 mM HEPES-KOH, pH 7.9, 150 mM KCl, 5 mM MgCl$_2$, 0.2 mM EDTA, pH 8.0) and centrifuged in a Beckman Ti-15 zonal rotor for 40 h at 32,000 rpm. The gradient was subsequently fractionated, and those fractions comprising the 17S peak were pooled and used for anti-SF3B1 immunoaffinity purification. Under the conditions used, the 17S peak is located approximately in the middle of the gradient, whereas subsequently formed spliceosomal complexes that contain U2 sediment in the 30–45S region, which is near the bottom of the gradient, ensuring that only isolated 17S U2 particles are purified. 17S gradient fractions were diluted 1:1 with GK5M buffer and loaded onto an anti-SF3B1 affinity column. After washing with GK5M buffer containing 1.5% (v/v) sucrose, bound 17S U2 snRNPs were eluted with 0.2 mg ml^{-1} SF3B1 peptide (EQYDPFAEHRPPKIAC) in GK5M containing 1.5% sucrose. Eluates were pooled and separated on a 5–20% (w/v) sucrose gradient containing GK5M buffer and 0–0.15% glutaraldehyde, by centrifuging at 29,000 rpm for 18 h in a Surespin 630 rotor. Gradient fractions were fractionated from the bottom, and glutaraldehyde was quenched by adding 100 mM aspartate (pH 7.0) to each fraction. 17S U2 peak fractions were concentrated and the buffer exchanged with GK5M buffer, by centrifugation with Amicon 50 kDa cut-off units. SDS–PAGE and mass spectrometry indicated that the protein composition of these affinity-purified 17S U2 snRNPs is identical to that of previously described, purified human 17S U2 snRNPs[20]. Owing to the labile nature of the 17S particle and the large excess of U2 snRNP in nuclear extract, it was not possible to efficiently immunodeplete intact 17S U2 snRNPs from HeLa nuclear extract, and assay the function of our purified 17S U2 snRNPs by performing in vitro splicing complementation assays.

BS3 crosslinking of 17S U2 snRNPs and crosslink identification

After elution from the anti-SF3B1 affinity column, purified 17S U2 particles were concentrated to 0.6 μM and subsequently incubated with 60 μM BS3 for 30 min at 20 °C. After quenching by incubating with 10 mM Tris-HCl, pH 8.0, for 15 min at 20 °C, crosslinked 17S U2 particles were loaded onto a 15–30% (w/v) glycerol gradient containing G150 buffer (20 mM HEPES-KOH, pH 7.9, 150 mM KCl, 1.5 mM MgCl$_2$) and centrifuged at 4 °C for 14 h at 35,000 rpm in a TH660 rotor. Gradient fractions were fractionated from the bottom and crosslinked U2 snRNPs migrating in the 17S region of the gradient were pelleted by ultracentrifugation in an S150AT rotor, and analysed as previously described[13]. Peptides were reverse-phase extracted using Sep-Pak Vac tC18 1cc cartridges (Waters) and fractionated by gel filtration on a Superdex Peptide PC3.2/30 column (GE Healthcare). Fifty-microlitre fractions corresponding to an elution volume of 1.2–1.8 ml were analysed in a Thermo Scientific Q Exactive mass spectrometer. Protein–protein crosslinks were identified by pLink 1.23 and 2.3.5 search engines (pfind.ict.ac.cn/software/pLink) and filtered at an FDR of 1% or 5% according to the recommendations of the developer[32,33]. For simplicity, the crosslink score is represented as a negative value of the common logarithm of the original pLink score; that is, score = $-\log_{10}$('pLink Score'). An expected maximum distance between the Cα atoms of the BS3-crosslinked lysine residues is approximately 30 Å. The length of most crosslinks (93–95% at the spectral level and 85–86% at the unique crosslink level) is ≤30 Å in the presented model of the 17S U2 snRNP (see Extended Data Fig. 3). The CXMS data from Supplementary Table 1 can be downloaded and visualized by interactive 2D viewers such as xiNET http://crosslinkviewer.org/ or https://xvis.genzentrum.lmu.de/login.php, or in 3D using UCSF Chimera or PyMOL.

Electron-microscopy sample preparation and image acquisition

The purified 17S U2 sample was absorbed to a thin layer carbon film for 1 min, which was subsequently attached to R3.5/1 QUANTIFOIL grids. Four microliters of ddH$_2$O was applied to the grid and excess water was blotted away by an FEI Vitrobot loaded with pre-wet filter paper, with the setting of blotting force 8, blotting time 4 s, at 100% humidity and 4 °C, and then vitrified by plunging into liquid ethane. Cryo-EM data were acquired on a FEI Titan Krios electron microscope (Thermo Fisher Scientific), equipped with a Cs corrector, at 300 kV. The images were recorded in integration mode on a Falcon III direct electron detector at 120,700× magnification, which corresponds to a calibrated pixel size of 1.16 Å at the specimen level. Micrographs were recorded using an exposure time of 1 s and a total dose of 72 e$^-$ per Å^2. In total, 12,194 were recorded with 20 movie frames and another 28,226 micrographs were recorded with 39 frames.

Image processing

Frames were dose-weighted, aligned and summed with Motion-Cor2[34]. The defocus values of the micrographs were determined by Gctf[35]. Summed micrographs were manually evaluated in the COW-MicrographQualityChecker (http://www.cow-em.de). Micrographs with isotropic Thong rings and clear particles were processed further. In total, 27,890 out of 40,420 summed micrographs were retained for further processing. Initially, around 3.5 million particles were automatically picked using Gautomatch (http://www.mrc-lmb.cam.ac.uk/kzhang), then extracted with a box size of 360 × 360 pixels, and binned to 180 × 180 pixels (pixel size of 2.32 Å) in RELION 3.0 (http://www2.mrc-lmb.cam.ac.uk/relion/index.php/Main_Page). Several iterations of reference-free 2D classification were performed in RELION 3.0 and 'bad classes' showing fuzzy or uninterpretable features were removed, yielding around 1.43 million 'good particles'. A subset of 200,000 particles was used to generate an initial 3D map by ab initio reconstruction function in cryoSPARC[36]. The 3' domain and other peripheral parts of this map were erased in UCSF Chimera v.1.13.1[37]. The remaining 5' domain was low-pass filtered to 40 Å and used as the 3D reference for 3D classification in RELION 3.0. The approximately 1.43 million good particles from 2D classification were split into five datasets. Each dataset was 3D classified into three classes. All 612,445 particles from good 3D classes (classes that show a complete particle with fine details) were combined and subjected to the Refine3D function in RELION 3.0, with a mask around the 5' domain, resulting in a 5.7 Å resolution map. Next, using the alignment parameters from the aforementioned masked 3D refinement, the 612,445 particles were focused classified with a mask around the 5' domain, into 8 classes. The best class (with 152,078 particles) that showed clear separation of HEAT repeats was selected, re-extracted in original pixel size (1.16 Å per pixel) with the box size of 300 × 300 pixels, and subjected to another round of focused 3D classification. The best 3D class (with 120,070 particles) was then selected and refined by Refine3D in RELION 3.0 with a mask around the entire 17S U2 particle or the 5' domain. The final reconstruction of the entire particle has an average resolution of 7.1 Å and the reconstruction with a mask around 5' domain has an average resolution of 4.1 Å, based on the RELION gold-standard Fourier shell correlation. The data were acquired at two different settings (see above). Because higher averaged doses tend to result in poorer reconstruction, we checked whether removing the particles from the micrographs receiving a higher average dose (leading to 82,420 particles retained) would improve the numerical resolution and the map quality, but this was not the case (data not shown).

Model building and refinement

Templates for the U2 proteins and RNA were obtained, wherever possible from published structures (Extended Data Table 2). The SF3b core crystal structure, together with parts of SF3A3 (amino acids 392–499) and SF3B2 (amino acids 458–530, 565-598) were initially docked as rigid bodies into the 4.1 Å map of the U2 5' domain. WD40-B of SF3B3 was then manually adjusted to fit the map using Coot v. 0.8.9.2[38]. The central part of SF3b core has resolution ranging from 3.6 to 4.5 Å, which allowed manual adjustment of side chains. The other parts were locally adjusted, without manipulating secondary structures or side chains. The model of the stem of U2 SLI (nucleotides 12–14, 19–21) was built by UCSF Chimera v. 1.13.1[37] as an A-form RNA helix. The model of BSL (nucleotides 25–45) was predicted by RNAvista[39] with base-pairing restrains as proposed in a previous study for *S. cerevisiae* U2 snRNA[8]. U2 SLIIa (nucleotides 48–65) was modelled based on the SLIIa in the yeast B[act] complex, and adjusted to the human U2 snRNA sequence. These RNA models were docked into the map individually as rigid bodies and connections between the SLs were de novo built using Coot. We also built homology models for PRP5 (amino acids 146–195) and TAT-SF1[RRM1] (amino acids 127–220), using SWISS-MODEL suite[40] or SpliProt3D database[41] and guided by our crosslinking (CXMS) data, we then rigidly docked the models into the high resolution map of the 5' domain. The better resolution (approximately 5 Å) of the TAT-SF1[RRM1] region allowed local adjustment of the homology model. Models of the entire 5' domain were then combined and subjected to real space refinement in PHENIX[42], with secondary structure restraints. The models of the remaining proteins or RNA (Extended Data Table 2) were rigid body docked into the low-resolution map of the entire 17S U2 particle, based on CXMS data and the overall shape of the EM density.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The atomic coordinate files have been deposited in the Protein Data Bank (PDB) with the following accession codes: U2 5' domain (6Y50), low resolution region (6Y53) and entire 17S U2 particle (6Y5Q). The cryo-EM maps have been deposited in the Electron Microscopy Data Bank as follows: U2 5' domain (EMD-10688) and entire U2 particle (EMD-10689).

31. Agafonov, D. E. et al. Molecular architecture of the human U4/U6. U5 tri-snRNP. *Science* **351**, 1416–1420 (2016).
32. Yang, B. et al. Identification of cross-linked peptides from complex samples. *Nat. Methods* **9**, 904–906 (2012).
33. Chen, Z. L. et al. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat. Commun.* **10**, 3404 (2019).
34. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
35. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
36. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
37. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
38. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
39. Antczak, M. et al. RNAvista: a webserver to assess RNA secondary structures with non-canonical base pairs. *Bioinformatics* **35**, 152–155 (2019).
40. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723 (1997).
41. Korneta, I., Magnus, M. & Bujnicki, J. M. Structural bioinformatics of the human spliceosomal proteome. *Nucleic Acids Res.* **40**, 7046–7065 (2012).
42. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
43. Behrens, S. E., Tyc, K., Kastner, B., Reichelt, J. & Lührmann, R. Small nuclear ribonucleoprotein (RNP) U2 contains numerous additional proteins and has a bipartite RNP structure under splicing conditions. *Mol. Cell. Biol.* **13**, 307–319 (1993).
44. MacMillan, A. M. et al. Dynamic association of proteins with the pre-mRNA branch region. *Genes Dev.* **8**, 3008–3020 (1994).
45. Will, C. L. et al. A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site. *EMBO J.* **20**, 4536–4546 (2001).
46. Spadaccini, R. et al. Biochemical and NMR analyses of an SF3b155-p14-U2AF-RNA interaction network involved in branch point definition during pre-mRNA splicing. *RNA* **12**, 410–425 (2006).
47. Plaschka, C., Lin, P. C., Charenton, C. & Nagai, K. Prespliceosome structure provides insights into spliceosome assembly and regulation. *Nature* **559**, 419–422 (2018).
48. Charenton, C., Wilkinson, M. E. & Nagai, K. Mechanism of 5' splice site transfer for human spliceosome activation. *Science* **364**, 362–367 (2019).
49. Darman, R. B. et al. Cancer-associated SF3B1 hotspot mutations induce cryptic 3' splice site selection through use of a different branch point. *Cell Rep.* **13**, 1033–1045 (2015).
50. Alsafadi, S. et al. Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. *Nat. Commun.* **7**, 10615 (2016).
51. Zhang, J. et al. Disease-causing mutations in SF3B1 alter splicing by disrupting interaction with SUGP1. *Mol. Cell* **76**, 82–95.e7 (2019).
52. Kielkopf, C. L., Lücke, S. & Green, M. R. U2AF homology motifs: protein recognition in the RRM world. *Genes Dev.* **18**, 1513–1526 (2004).

Extended Data Fig. 1 | Spliceosome assembly cycle and the interaction of 17S U2 with the pre-mRNA branch site. a, Early assembly and catalytic activation pathway of the spliceosome. The 17S U2 snRNP, the structure of which was determined here by cryo-EM, is indicated by an asterisk. For simplicity, the stepwise interactions of the U1, U2, U4/U6 and U5 snRNPs, and only selected non-snRNP proteins are shown. Helicases involved in conversion of the E to A complex are indicated. In the E complex, the U2AF65 UHM interacts with the ULM of SF1, and after release of SF1, it subsequently interacts with a ULM in the N-terminal region of SF3B1. This swap of UHM–ULM interactions is probably very important for positioning the BS before the conformational change in the HEAT domain clamps down on the U2–BS helix and stabilizes the U2 snRNP interaction with the intron. The U2AF65/U2AF35 dimer is released (not shown) during conversion of the A to B complex[20]. SF1 is displaced from the BS by UAP56 (either before or after PRP5 action). SF1 pre-bulges the BS-A via accommodation of the latter in its KH domain, facilitating subsequent base-pairing of U2 with the BS[29]. **b**, Base-pairing interactions between U2 snRNA and the BS and upstream intron nucleotides, that lead to bulging of the BS-A. The sequence shown is from intron 10 of the pre-mRNA for the

polypyrimidine tract-binding protein (PTB). Red shading denotes the bona fide U2–BS helix, and shows base pairs formed between human U2 snRNA and the conserved BS consensus sequence of PTB intron 10. Yellow shading denotes the extended U2–BS helix, in which the number and nature of base-pairing interactions varies depending on the pre-mRNA intron sequence. **c**, Schematic of the composition of the human 17S U2 snRNP. Only abundant U2 proteins are shown[20]. **d**, Open structure of the SF3B1 HEAT domain in the isolated SF3b complex (left) and its more closed conformation after interaction with the U2–BS helix (right). The SF3B1 HEAT domain (green) forms a super-helical structure, and in the spliceosome sequesters the U2–BS helix. The conformational change in the HEAT domain, which is required to form the BS-A binding pocket, was proposed to occur after formation of the U2–BS helix[17]. Before this study, the conformation of SF3B1 in human 17S U2 snRNP was not known. The pre-mRNA is in grey; U2 snRNA is coloured as in Fig. 2. For simplicity, the PHF5A protein that also forms part of the BS-A binding pocket is not shown. The structures of SF3B1[HEAT] in the isolated SF3b complex (PDB code 5IFE) (left) and human B[act] complex (PDB code 6FF4) (right) are aligned via HEAT repeat 20.
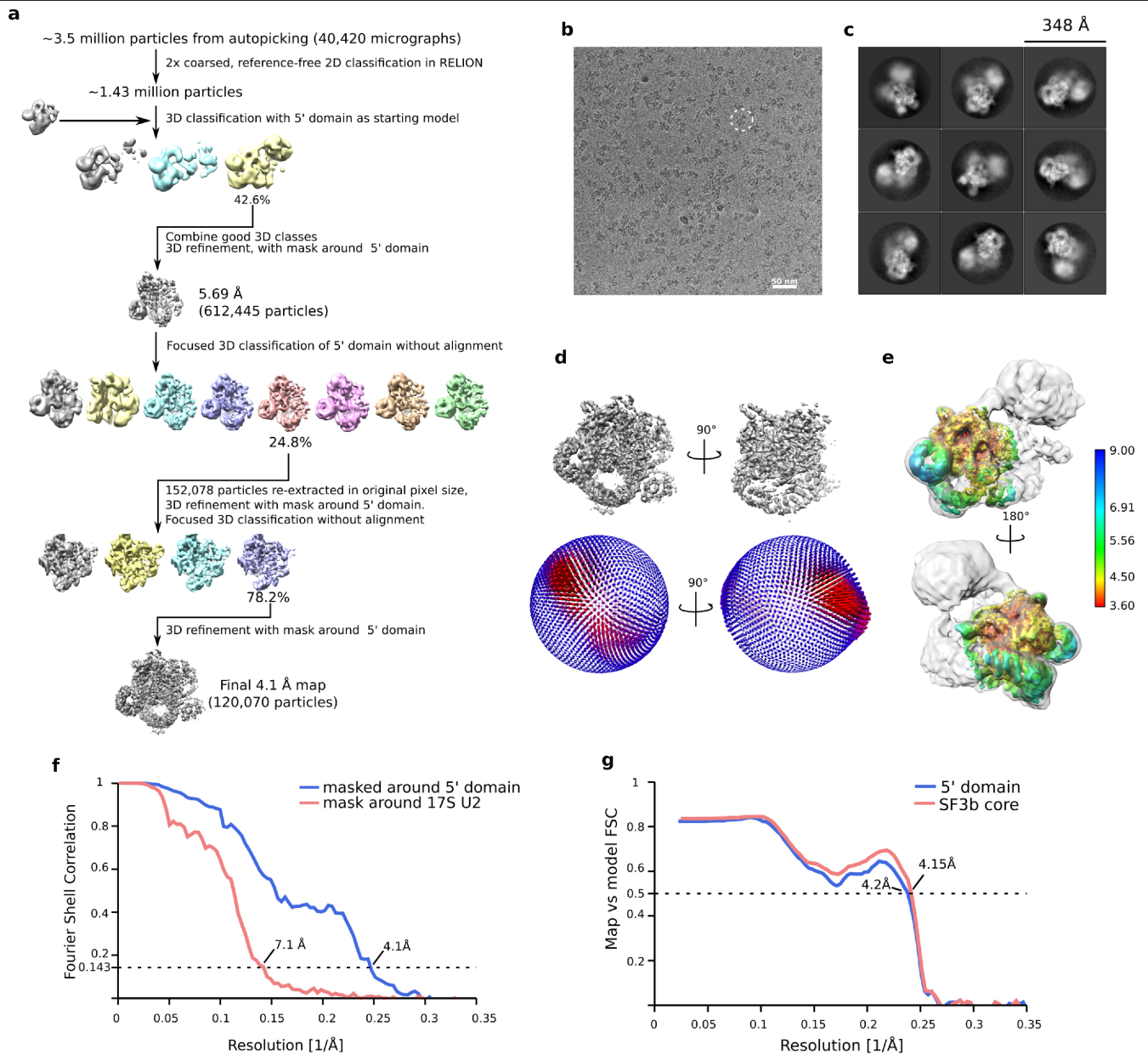
Extended Data Fig. 2 | Cryo-EM and image-processing of the human 17S U2 snRNP. a, Computation sorting scheme. All major image-processing steps are depicted. For a more detailed explanation, see 'Image processing' in the Methods. A considerable amount of conformational heterogeneity is present in all spliceosomal complexes but even more in the bipartite 17S U2 snRNP, which is structurally very labile and readily dissociates during purification[43], making its analysis by electron microscopy challenging. In addition, the bridges connecting the 5′ and 3′ domains of the U2 particle (see also Extended Data Fig. 5) have a very flexible character, leading to flexibility in the 3′ domain and the large variation in local resolution. Thus, a substantially higher number of particles was needed to generate the 17S U2 structure than is usually used for cryo-electron microscopy. **b**, Typical cryo-electron micrograph of the *Homo sapiens* 17S U2 snRNP recorded at 120,700× magnification with a Titan Krios microscope using a Falcon III direct electron detector operating in integration mode at a calibrated pixel size of 1.16 Å. **c**, Representative selection of reference-free 2D class average images depicting

17S U2 particles recorded under cryo conditions. **d**, Euler angle distribution plot of all 17S U2 particles that contributed to the final structure. Red depicts a higher relative number of particles at a certain angle. The generally uniform distribution of the particle projection angles ensures an isotropic 3D electron-microscopy density map. **e**, Local resolution estimation of the 5′ domain of 17S U2 snRNP. The 5′ domain shows a resolution between 3.6 and 9.0 Å. The map of the remaining part excluding the 5′ domain, shown as a translucent overlay, was determined at resolutions between 10 and 30 Å. **f**, Fourier shell correlation (FSC) of two independently refined half datasets, calculated using the 'PostProcessing' routine in RELION, indicates a global resolution of 7.1 Å for the entire 17S U2 snRNP, and 4.1 Å for the masked U2 5′ domain. Multibody refinement around the 3′ domain and the peripheral parts did not produce better resolved maps for those regions. **g**, Map versus model FSC curves for the 5′ domain and the SF3b core. The FSC = 0.5 criterion indicates a resolution of 4.2 Å for the U2 5′ domain, and 4.15 Å for the SF3b core.
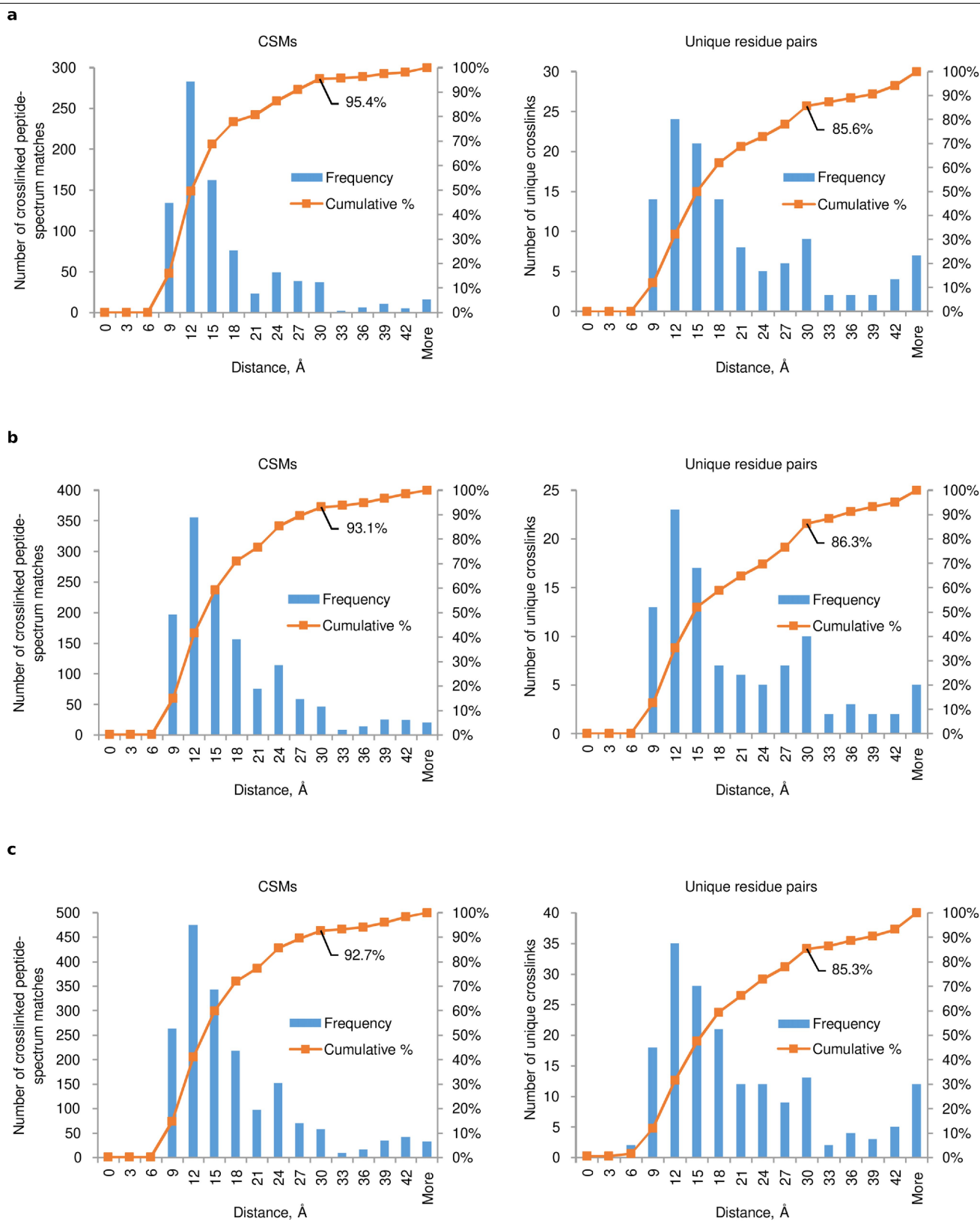
Extended Data Fig. 3 | Euclidian distances for crosslinks observed between modelled residues (Cα) of the 17S U2 snRNP. a–c, Crosslinks from a single 17S U2 crosslinking experiment (with two technical replicates) were identified by pLink2.3.5 and filtered to a false discovery rate (FDR) of 1% (**a**), pLink1.23 at an FDR of 1% (**b**), and pLink1.23 at an FDR of 5% (**c**). Calculations were performed using PyMOL2.3.4 for crosslinks with a score of at least 1. Most crosslinks (93–95% at the spectral level and 85–86% at the unique crosslink level) are consistent—that is, Cα atoms of the crosslinked amino acids are within 30 Å of each other—in the presented model of the 17S U2 snRNP. The percentage of overlength crosslinks (that is, longer than 30 Å) is slightly higher than observed for more rigid complexes, which is consistent with the known structural flexibility/dynamics of the 17S U2 snRNP.
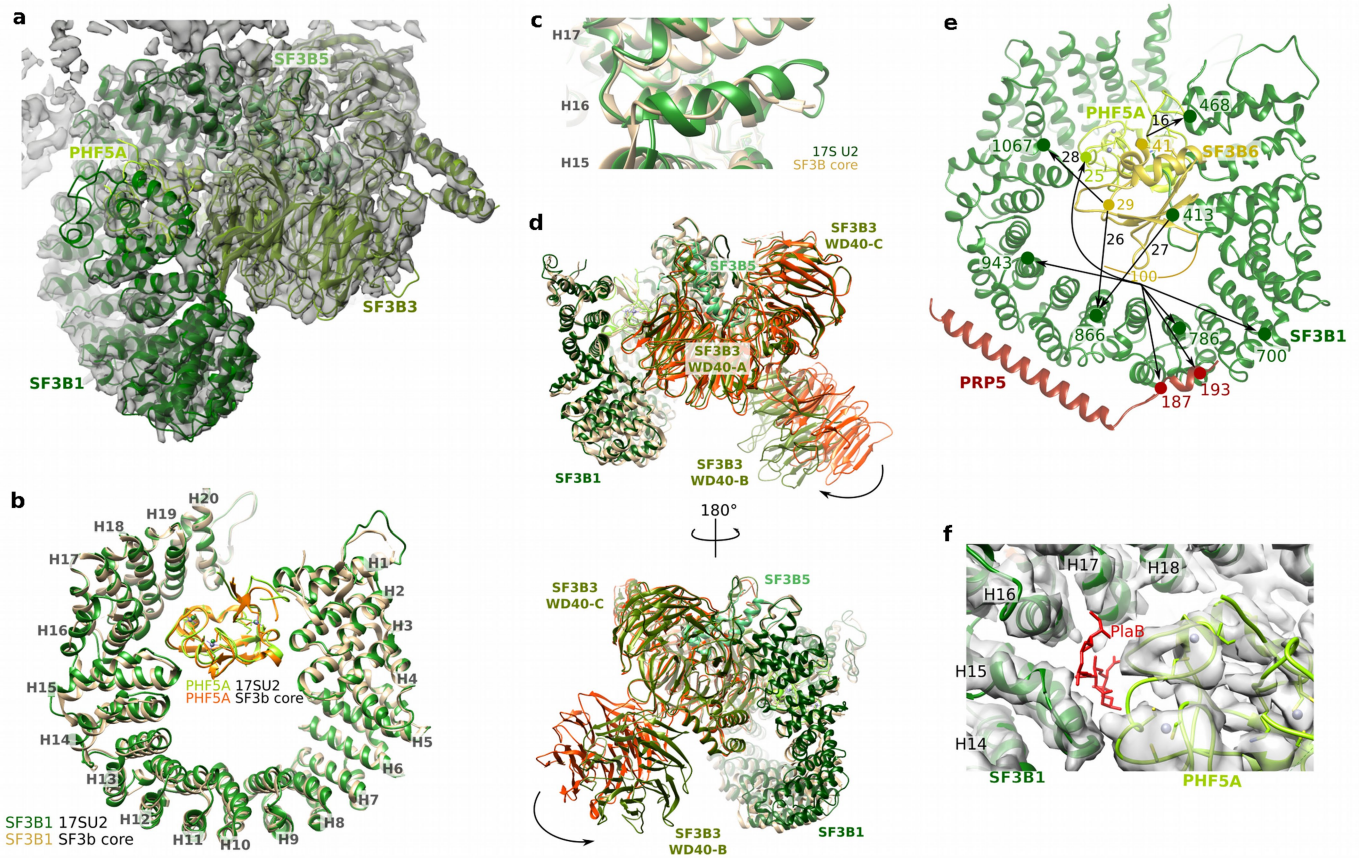
Extended Data Fig. 4 | The SF3B1 HEAT domain has an open conformation in the 17S U2 snRNP. a, Fit of the SF3b core proteins into the 17S U2 electron-microscopy density (grey). **b**, Overlay of the HEAT domain (amino acids 529–1201) of SF3B1 in 17S U2 snRNP (green) and the crystal structure of the isolated SF3b core (gold, PDB 5IFE). In 17S U2 and isolated SF3b, PHF5A contacts both N- and C-terminal regions of the HEAT domain, interacting with HEAT repeats HR2–HR3 near its N terminus, as well as HR15, HR17 and HR18 near its C terminus. PHF5A thus contacts two previously described, dynamic hinge regions (HR3–HR4 and HR15–HR16) of the HEAT domain[10], and thereby helps to stabilize the SF3B1 open conformation. **c**, Close up of SF3B1 HR16 in 17S U2 overlaid with that in isolated SF3b. HR16 is completely structured in the 17S U2 particle, but not in isolated SF3b. **d**, Overlay of the SF3b core domain in 17S U2 snRNP (green) with the crystal structure of isolated SF3b (gold; PDB code 5IFE). For clarity, the SF3B3 protein is coloured red-orange in this panel. Although the WD40-A and WD40-C domains of SF3B3 have essentially the same conformation, and clamp SF3B5 in a similar manner in both the 17S U2 snRNP and the SF3B core complex, the WD40-B domain has a slightly different position and is rotated more towards SF3B1[HEAT] in 17S U2. **e**, Multiple crosslinks were detected between SF3B6, which can be crosslinked to the BS-A in spliceosomal A complexes[44,45], and SF3B1 residues on the upper surface of the HEAT domain, as well as the PRP5 α-helix that interacts with HR9–HR12, and PHF5A (see also Supplementary Table 1). Crosslinks between SF3B1 and SF3B6

were also detected in the N-terminal region of SF3B1 located at, or near, amino acids (373–415) that are required for stable SF3B6–SF3B1 interaction[45,46]. Similar protein–protein crosslinks involving SF3B6 were observed with recombinant, intact SF3b complexes[10], which indicates that SF3B6 is located in a similar, but not firmly fixed, position both in 17S U2 and the isolated SF3b complex[10]. Numbers (colour-coded to match protein colours) indicate the positions of crosslinked lysine residues, which are connected by black arrows. The distances between the crosslinked residues in our 17S U2 model are indicated by small numbers next to the black arrows. A distance is not included if one of the crosslinked residues is present in an unstructured protein region. **f**, The site where the splicing modulator pladienolide B (PlaB) binds SF3B1 is not occupied in the 17S U2 snRNP. The crystal structure of PlaB bound to the SF3b core complex showed that the binding pocket of PlaB, which is formed by HR15–HR17 and PHF5A, is present only in the open conformation of the HEAT domain, and overlaps with the BS-A-binding pocket[17]. As the PlaB-binding pocket is present in a hinge region of the HEAT domain, it was proposed to inhibit SF3b function by preventing the conformational change in the HEAT domain needed to clamp down on the U2–BS helix[17]. There is no electron-microscopy density observed in the PlaB-binding pocket and thus it and potentially other splicing modulators can also bind SF3B1 in the 17S U2 snRNP.
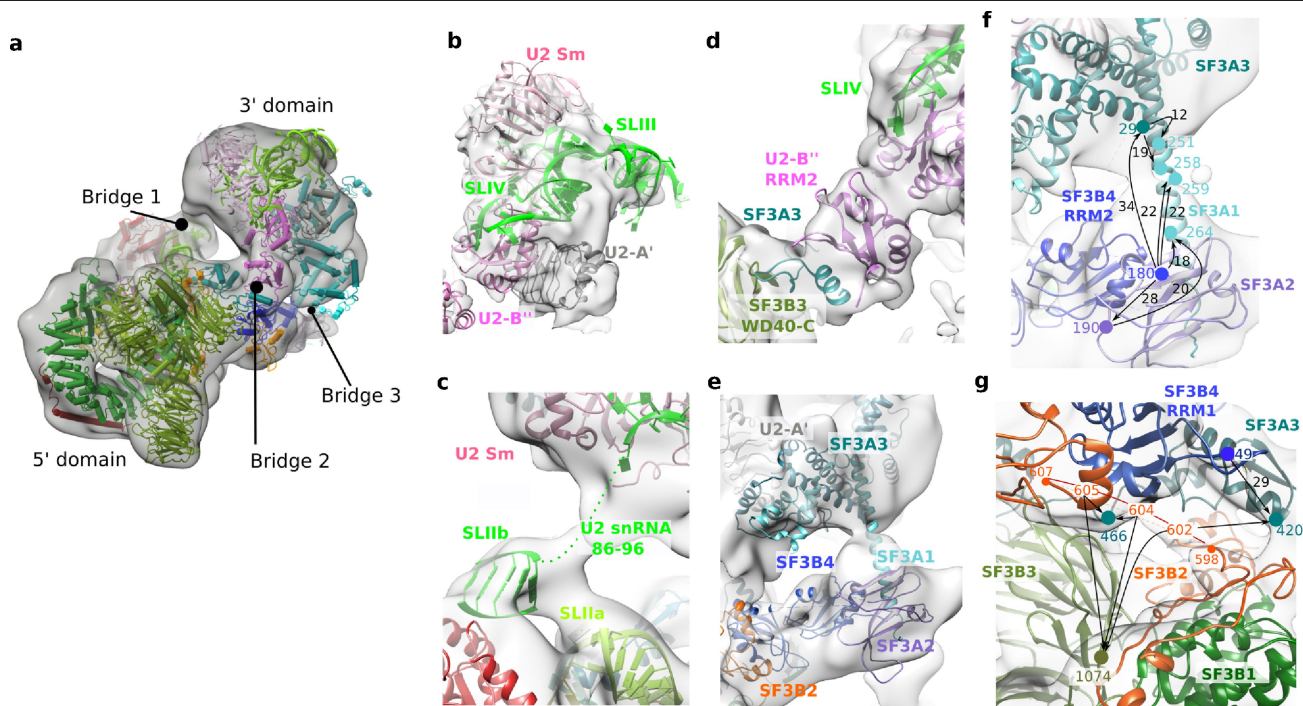
Extended Data Fig. 5 | The 3′ domain of the 17S U2 snRNP and molecular bridges connecting it to SF3b. a, The U2 3′ domain is connected to the 5′ domain by three main bridges. Fit of the entire 17S U2 molecular model into the electron-microscopy density (low-pass filtered). **b**, Fit of the U2 Sm core, U2 snRNA SLIII, and U2-A′ and U2-B″ bound to U2 snRNA SLIV. The overall structure of the U2 3′ domain does not change substantially after U2 incorporation into the spliceosome. The U2 Sm core domain is located at a similar distance from SF3B1[HEAT] as observed in the human B complex[13,15]. The resolution of the 17S U2 present in the yeast A complex[47], as well as in the human pre-B complex[15,48], is not sufficient to make meaningful, detailed comparisons of their structure with that of our 17S U2 snRNP. Furthermore, in the former complexes the molecular architecture of U2 is derived entirely from that found in yeast B or human B[act] complexes. **c**, Bridge 1 is probably composed of U2 snRNA nucleotides upstream of the Sm-binding site that connect it to SLIIb, which is also part of this bridge, as well as unassigned protein density. **d**, Bridge 2 is formed by RRM2 of U2-B″ and amino acids in the C-terminal half of SF3A3 that bridge U2-B″ and the WD40-C domain of SF3B3. **e**, The N-terminal helical domain of SF3A3 contacts the U2 Sm core. SF3A3 also interacts with SF3B2 and then extends to the U2 snRNA SLIIa and BSL (see Fig. 2). Rigid-body fitting combined with protein–protein crosslinking (see **f** and **g**) allowed us to localize near or within bridge 3, both RRM1 (amino acids 13–91) and RRM2 (amino acids 100–179) of the SF3B4 protein. SF3B4[RRM1] interacts with a short region of its binding partner SF3B2 (amino acids 607–693) and SF3B4[RRM2] extends towards the β-sandwich of SF3A2 (amino acids 118–209) and a long helical region of SF3A1 (amino acids 235–274) that also comprises part of bridge 3. SF3A1 extends from the N-terminal region of SF3A3 to the SF3A2 β-sandwich. Thus, SF3a proteins have important bridging roles in connecting the 3′ and 5′ domains of the U2 snRNP. Stable integration of SF3a into the human 17S U2 particle during its biogenesis requires the prior binding of SF3b[12] and thus the SF3a–SF3b protein contacts described above potentially have key roles in the assembly of the 17S U2 particle. **f**, **g**, Intermolecular crosslinks supporting the location in our 17S U2 model of SF3B4[RRM2] (**f**) and SF3B4[RRM1] (**g**). Numbers (colour-coded to match protein colours) indicate the positions of crosslinked lysine residues, which are connected by black arrows. The distance between the crosslinked residues in our 17S U2 model is indicated by small numbers next to the black arrows. A distance is not included if one of the crosslinked residues is present in an unstructured protein region.
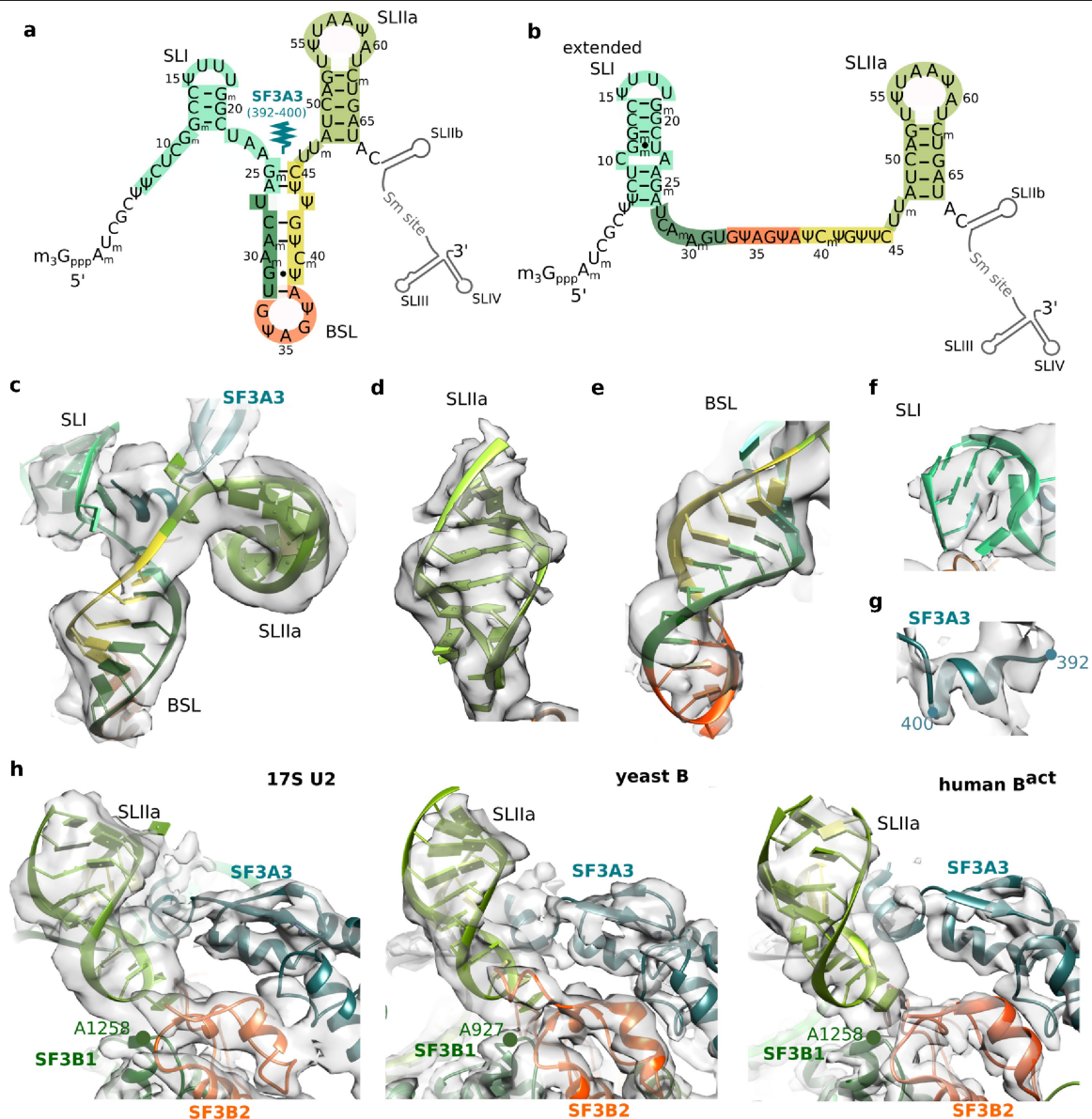
Extended Data Fig. 6 | Fit of the various stem–loops in the 5′ half of U2 snRNA. a, **b**, Alternative conformations of stem–loops potentially formed in the 5′ part of human U2 snRNA. The stem of the BSL could potentially form additional base pairs, but the presence of the SF3A3 separator helix (amino acids Y392 to H400) prevents base-pair formation beyond G25–C45. This guarantees that U46 and U47 are single-stranded and thus could potentially be involved in the proposed movement of the BSL away from SF3B1 during U2–BS helix formation (see also Extended Data Fig. 8). **c**, Fit of U2 SLIIa, BSL and SLI three-way junction into 17S U2 electron-microscopy density. **d**–**f**, Fit of the individual SLIIa (**d**), BSL (**e**) and a shortened SLI (**f**) into the electron-microscopy density. Owing to the formation of the BSL, only a shortened U2 SLI can form as nucleotides in the lower stem of an extended SLI would instead form base pairs located in the lower stem of the BSL. Thus, an extended SLI and a BSL are mutually exclusive, competing U2 snRNA conformations. The position of the remaining 10 nucleotides at the 5′ end of the U2 snRNA, which in the spliceosome form part of U2–U6 helix II, cannot be discerned. **g**, Fit of the SF3A3 separator helix into density at the base of the BSL. **h**, Similar SLIIa RNP architecture in 17S U2 snRNP, and the yeast B and human B[act] spliceosomal complexes. In human U2 snRNP, loop nucleotides of SLIIa contact amino acids of the loop connecting the two α-helices of SF3B1 HR20, as well as residues of SF3B2; two regions of the latter (amino acids 458–530 and 565–598) are located in well-resolved density close to SLIIa and SLI, respectively. These SLIIa contacts are similar to those found in yeast B and human B[act] spliceosomes. Thus, they are a major, direct anchor point for SF3b on the U2 snRNA. The poor resolution of the cryo-EM structure of the human B complex in this region does not allow for an accurate comparison.
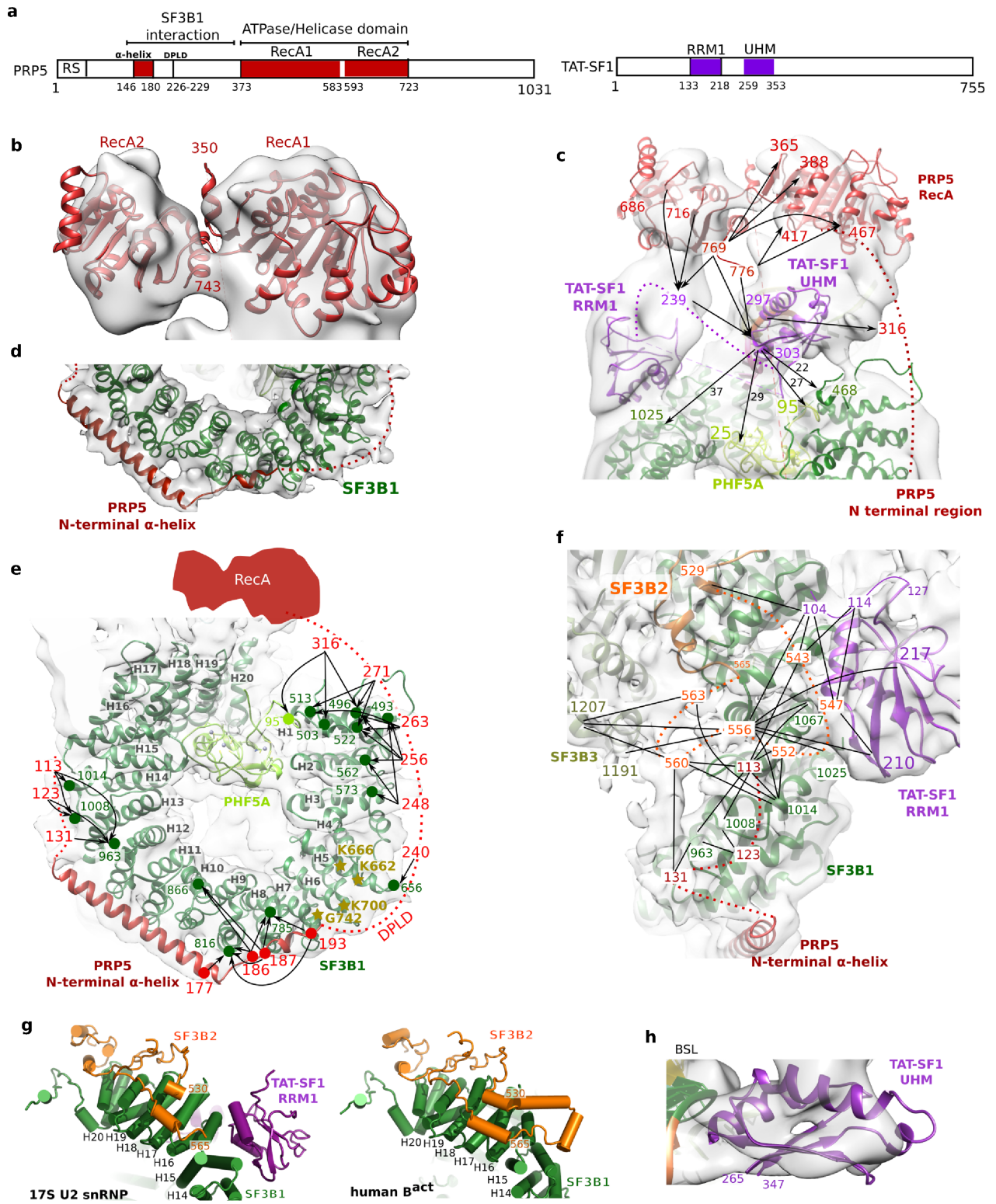
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Fit of PRP5 and TAT-SF1 and their interaction with the SF3B1 HEAT domain. a, Schematic of the domain organization of human PRP5 (left) and TAT-SF1 (right), with amino acid boundaries of each domain indicated below. **b**, Fit of the PRP5 RecA1 and RecA2 domains in an open conformation into the 17S U2 electron-microscopy density. The open (inactive) conformation of the PRP5 RecA domains seems to fit better than the closed (active) conformation, which suggests that PRP5 is inactive in 17S U2 snRNP. However, neither the resolution in this region nor intramolecular PRP5 crosslinks allow us to confidently position these two domains relative to each other and thereby distinguish between these two conformations. DEAD-box proteins typically bind double-stranded RNA via their RecA domains and facilitate local strand separation by introducing one or two sharp bends in one of the bound strands that prevent base-pairing with the complementary strand[28]. The resolution of the 17S U2 in the region where the PRP5 RecA domains are located does not allow us to discern whether PRP5 is interacting with RNA at this stage. On the basis of our structure, the RecA domains of PRP5 in 17S U2 are not close enough to the BSL to directly disrupt it, and also are not located close enough to the SLI. **c**, Protein crosslinks support the positioning of the PRP5 RecA domains and the TAT-SF1[UHM]. Crosslinks are observed between TAT-SF1[UHM] and residues in the N- and C-terminal HEAT repeats of SF3B1, in the region of PRP5 between the N-terminal α-helix and RecA domains, and in PHF5A, which supports the location of TAT-SF1[UHM] in the less-well resolved density element adjacent to the U2 BSL. **d**, Fit of the PRP5 α-helix (amino acids 146–196) into the electron-microscopy density contacting HEAT repeats 9–12. **e**, Protein crosslinks between PRP5 and SF3B1 suggest that a region spanning approximately 200 amino acids located N-terminal of the PRP5 helicase domain, wraps around most of SF3B1[HEAT]. These data are consistent with studies showing that yeast Prp5 also interacts with HR1–HR6 and HR9–HR12 of the yeast SF3B1 homologue, Hsh155[23]. Numbers (colour-coded to match protein colours) indicate the positions of crosslinked lysine residues, which are connected by black arrows. The proposed path of unstructured regions of PRP5 is indicated by a dotted line and the location of the conserved PRP5 DPLD motif is shown. The position of selected cancer-related hotspot mutations of SF3B1 (K666, K662, K700 and G742) are indicated. Point mutations in SF3B1[HEAT] are linked to various cancers and lead to the utilization of cryptic branch sites and 3' splice sites in vivo[49,50]. The exact mechanism responsible for these changes in alternative splicing is currently unknown but it has been suggested that these mutations may affect the curvature of the HEAT domain[10]. Studies in yeast showed that position-equivalent, point mutations in SF3B1 that are linked to cancer in humans, lead to loss of stable Prp5 binding to the U2 snRNP[23,25]. More recent studies indicate that in humans these cancer-related mutations destabilize the interaction of SF3B1 with the SUGP1 protein[51], which, however, is essentially absent from our 17S U2 preparations[20]. As PRP5 appears to encompass the entire HEAT domain in the human 17S U2, it is conceivable that a change in the curvature of the HEAT domain could destabilize the SF3B1–PRP5 interaction, and the absence of PRP5 may directly lead to alterations in BS selection by U2 containing SF3B1 cancer-related point mutations. Prp5 was shown to contact the SF3b complex in the U2 snRNP via a conserved DPLD motif in its N-terminal domain[24]. The most common (hotspot) cancer-related point mutations mainly cluster in or near HR6[10] and notably our crosslinking data place the DPLD motif of PRP5 (amino acids 226–229) in proximity to HR6. Thus cancer-related hotspot mutations may disrupt this essential interaction, leading to destabilization of PRP5. This, in turn, could have a detrimental effect on the function of PRP5 to facilitate the formation of a stable U2–BS interaction and/or on the proofreading activity of PRP5 (as discussed above), and thus facilitate the usage of aberrant BS and 3' splice sites. **f**, Protein crosslinks between TAT-SF1[RRM1] and neighbouring proteins. In 17S U2, TAT-SF1[RRM1] is located adjacent to SF3B1 HR15 and HR16 (Fig. 3a), and thus may also stabilize HR16, which is completely structured in 17S U2 (Extended Data Fig. 4c). Crosslinking also suggests that an unstructured loop of SF3B2 consisting of amino acids 531–564 may occupy the electron-microscopy density near SF3B1 HR12–HR15. **g**, Position of SF3B1 and SF3B2 in 17S U2 (left) and in the human B[act] complex (right). In 17S U2, TAT-SF1[RRM1] is located in the same position where two α-helices of SF3B2 are found in the B[act] complex, and thus release of TAT-SF1 would be required for the subsequent formation or repositioning of this region of SF3B2. **h**, Fit of TAT-SF1 UHM (amino acids 260–353) into the 17S U2 electron-microscopy density. Amino acids 260–353 of human TAT-SF1 were initially designated RRM2, but were later shown to comprise a UHM[52].

Article
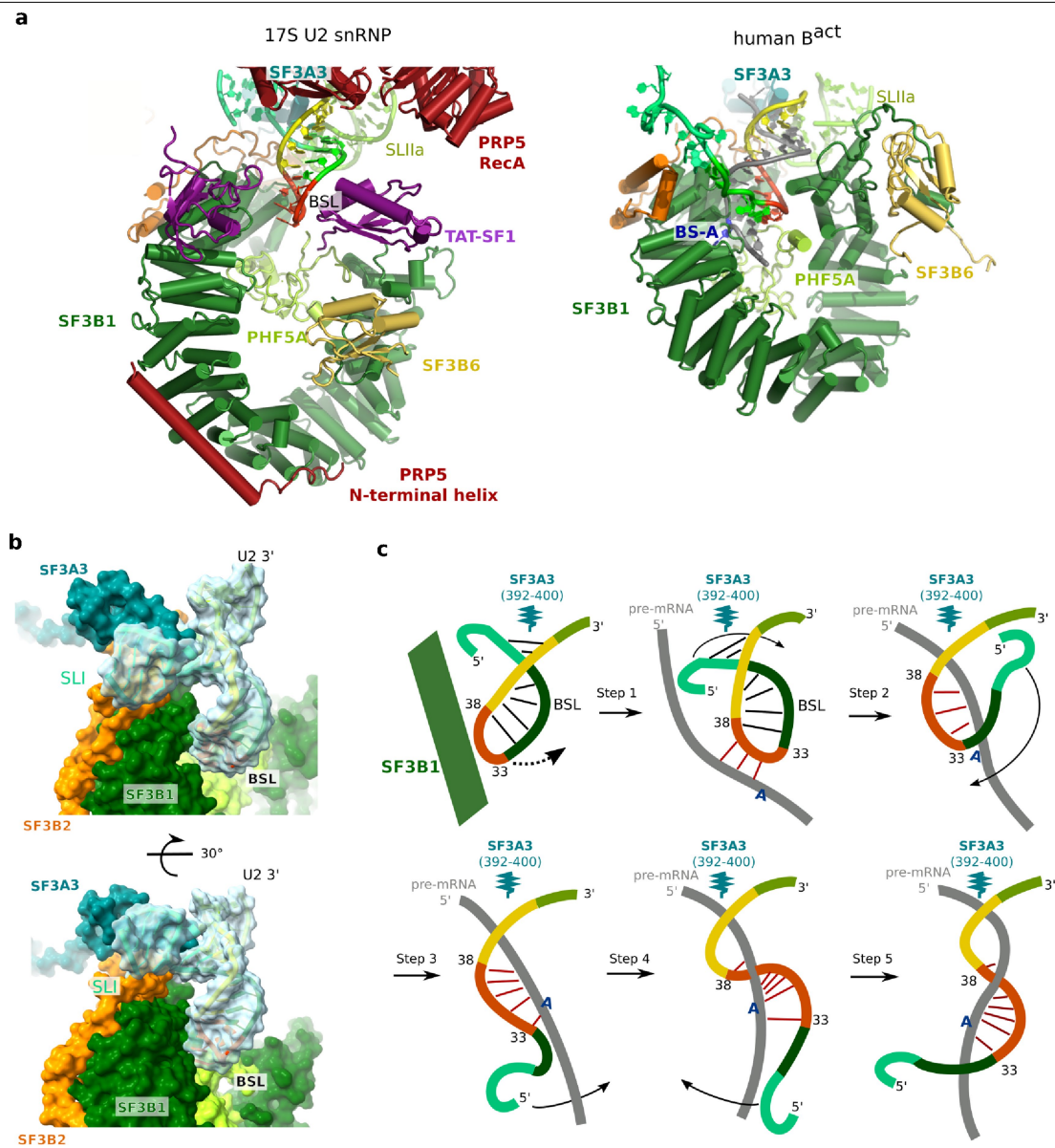
Extended Data Fig. 8 | RNP rearrangements and movements of the U2 snRNA required for formation of the extended U2–BS helix. a, Spatial orientation of the BSL and neighbouring proteins in the 17S U2 (left), and the subsequently rearranged U2 snRNA and U2 proteins after formation of the U2/BS helix and stable U2 incorporation into the spliceosome (right). Although the human B^{act} structure (PDB code 6FF4) is shown, the spatial organization of the shown U2 components is similar in B and presumably also human A complexes. **b**, Movement of U2 SLI behind the BSL requires repositioning of the latter. Coloured surfaces are derived from the fitted protein models, with the RNA depicted as a combination of a transparent surface model and RNA helix. The movement of the 5′ end of U2 required for formation of the extended U2–BS helix would require release of the BSL from SF3B1. In the absence of the latter, the 5′ end of U2 snRNA, including the short SLI, which is topologically located above the BSL stem, would have to be threaded through a very narrow opening between the BSL and the SF3B1 HEAT domain in order to unwind the BSL, a scenario that is unlikely. Thus, the remodelling of the U2 BSL into an extended U2–BS helix will probably occur in a conformational state of the U2 snRNP in which the BSL is moved away from the C-terminal HEAT repeats. This movement could involve a rotation around U2 nucleotides U46 and/or U47, which link the BSL to the stem of SLIIa and appear to be maintained in a single-stranded conformation by SF3A3 (Extended Data Fig. 6). In this respect, it is intriguing that the short SF3A3 separator helix is situated at the same place in 17S U2 and in human B^{act} complexes. Moreover, it probably has very similar roles in 17S U2 and the spliceosome. That is, in B^{act}, the SF3A3 separator helix also determines the length of the extended part of the U2–BS helix and facilitates the movement of U2 snRNA and the intron away from each other. The SF3A3 separator helix probably maintains its contact with the U2 SLIIa, ensuring the stable interaction of the SF3b complex with the U2 snRNA. This has the advantage that the newly formed extended U2–BS complex can swing back towards the HEAT domain with SF3A3 still bound to the SLIIa. Docking of the extended U2–BS helix to the SF3B1 C-terminal HEAT repeats is potentially facilitated by initial interactions with the backbone of the first 5–6 nucleotides of the intron downstream of the BS. **c**, Movements of the 5′ end of U2 that enable formation of the extended U2–BS helix. Step 1: ATP hydrolysis by PRP5 disrupts protein contacts with the BSL and presumably also with the 5′ end of the U2 snRNA including SLI. This allows base pairing between the BSL loop nucleotides and the BS of the pre-mRNA intron. Step 2: the destabilized BSL unwinds and the 5′ end of U2 moves behind the BSL and downward, with a twisting motion that repositions the pre-mRNA behind the unwound BSL. Movement of the 5′ end downward allows the formation of additional base pairs with the pre-mRNA. As both ends of the pre-mRNA intron appear to be fixed by protein and snRNP interactions, formation of the helical conformation of the U2–BS and extended U2–BS probably involves first movement of the 5′ end of U2 behind the pre-mRNA (step 4), followed by a movement across the pre-mRNA (that is, a twisting rotation of the 5′ end of U2 around the pre-mRNA) (step 5). Numbers indicate the position of selected U2 nucleotides.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	17S U2 5' domain (EMD-10688) (PDB 6Y50)	17S U2 low resolution part (EMD-10689) (PDB 6Y53)	17S U2 particle (EMD-10689) (PDB 6Y5Q)
Data collection and processing			
Magnification	120,700	120,700	120,700
Voltage (kV)	300	300	300
Electron exposure (e⁻/Å²)	72	72	72
Defocus range (μm)	1-4	1-4	1-4
Pixel size (Å)	1.16	1.16	1.16
Symmetry imposed	C1	C1	C1
Initial particle images (no.)	3.5 million	3.5 million	3.5 million
Final particle images (no.)	120,070	120,070	120,070
Map resolution (Å)	4.1	7.1	7.1
FSC threshold	0.143	0.143	0.143
Map resolution range (Å)			
5'domain	3.6-9	---	4-9
TAT-SF1 RRM1	5-6	---	6-9
TAT-SF1 UHM	---	15	15
PRP5 RecA domains	---	20	20
3'domain	---	20	20
SF3A	---	30	30
Refinement			
Initial model used (PDB code)	n/a	n/a	n/a
Model resolution (Å)	4.2	---	---
FSC threshold	0.5		
Model resolution range (Å)	4.2	---	---
Map sharpening B factor (Å²)	-241	-450	-450
Model composition			
Non-hydrogen atoms	14917	11306	22867
Protein residues	2537	2226	4763
Ligands	3	0	0
B factors (Å²)			
Protein	54.74	---	---
Ligand	196.62	---	---
R.m.s. deviations			
Bond lengths (Å)	0.004	---	---
Bond angles (°)	1.18	---	---
Validation			
MolProbity score	2.05	---	---
Clashscore	13.78	---	---
Poor rotamers (%)	0	---	---
Ramachandran plot			
Favored (%)	93.94	---	---
Allowed (%)	5.90	---	---
Disallowed (%)	0.16	---	---

Article

Extended Data Table 2 | Summary of modelled proteins and RNA in the human 17S U2 structure

Protein/RNA	Chain ID	UniProt ID	Total residues	Modeled Residue	Template	modeling approach
SF3A1	6	Q15459	793	160-286	4DGW	Docked
SF3A2	7	Q15428	464	104-209	4DGW	Docked
SF3A3	9	Q12874	501	1-362	4DGW	Docked
				392-499	6FF4	Docked and adjusted
SF3B1	u	O75533	1304	399-419	2FHO	Docked
				463-1304	6EN4	Docked and adjusted
SF3B2	8	Q13435	895	458-598	6FF4	Docked and adjusted
				607-658, 681-693	5LSB	Docked
SF3B3	v	Q15393	1217	1-1204	6EN4	Docked and adjusted
SF3B4	o	Q15427	424	11-89, 101-181	5LSB	Docked
SF3B5	x	Q9BWJ5	86	15-80	6EN4	Docked and adjusted
PHF5A	y	Q7RTV0	110	6-98	6EN4	Docked and adjusted
SF3B6/p14	z	Q9Y3B4	125	17-93	2FHO	Docked
A'	a	P09661	255	2-163	5MQF	Docked
B"	b	P08579	225	6-98	5MQF	Docked
				149-225	6FF4	Docked
Sm D2	h	P62316	118	19-116	5MQF	Docked
Sm F	i	P62306	86	2-75	5MQF	Docked
Sm E	j	P62304	92	14-92	5MQF	Docked
Sm G	k	P62308	76	3-76	5MQF	Docked
Sm D3	l	P62318	126	2-84	5MQF	Docked
Sm B	m	P14678	240	4-49,63-87	5MQF	Docked
Sm D1	n	P62314	119	1-82	5MQF	Docked
PRP5	p	Q7L014	1031	146-195		Predicted model, Docked
				350-743	4LJY	Docked and adjusted
TAT-SF1	q	O43719	755	127-220		Predicted model, docked and adjusted
				260-347	2DIT	Docked
U2snRNA	2	NR_002716	188	12-14,19-21		de novo, Docked
				22-24		de novo
				25-45		Predicted model, Docked
				46-47		de novo
				48-65	5GM6	Docked
				69-73, 81-85		de novo, Docked
				97-184	5MQF	Docked

RNA and protein regions were modelled and fit into the electron-microscopy density as indicated.

3.2 Structural insights into how Prp5 proofreads the pre-mRNA branch site (*published manuscript*)

Zhenwei Zhang[1], Norbert Rigo[2], Olexandr Dybkov[2], Jean-Baptiste Fourmann[2], Cindy L. Will[2], Vinay Kumar[2], Henning Urlaub[3,4], Holger Stark[1]*,and Reinhard Lührmann[2]*

[1]Department of Structural Dynamics, MPI for Biophysical Chemistry, Göttingen, Germany.

[2]Cellular Biochemistry, MPI for Biophysical Chemistry, Göttingen, Germany.

[3]Bioanalytical Mass Spectrometry, MPI for Biophysical Chemistry, Göttingen, Germany.

[4]Bioanalytics Group, Institute for Clinical Chemistry, University Medical Center Göttingen, Göttingen, Germany.

*Correspondence to: hstark1@gwdg.de; reinhard.luehrmann@mpibpc.mpg.de

Article

Structural insights into how Prp5 proofreads the pre-mRNA branch site

Zhenwei Zhang[1], Norbert Rigo[2], Olexandr Dybkov[2], Jean-Baptiste Fourmann[2], Cindy L. Will[2], Vinay Kumar[2], Henning Urlaub[3,4], Holger Stark[1✉] & Reinhard Lührmann[2✉]

During the splicing of introns from precursor messenger RNAs (pre-mRNAs), the U2 small nuclear ribonucleoprotein (snRNP) must undergo stable integration into the spliceosomal A complex—a poorly understood, multistep process that is facilitated by the DEAD-box helicase Prp5 (refs. [1–4]). During this process, the U2 small nuclear RNA (snRNA) forms an RNA duplex with the pre-mRNA branch site (the U2–BS helix), which is proofread by Prp5 at this stage through an unclear mechanism[5]. Here, by deleting the branch-site adenosine (BS-A) or mutating the branch-site sequence of an actin pre-mRNA, we stall the assembly of spliceosomes in extracts from the yeast *Saccharomyces cerevisiae* directly before the A complex is formed. We then determine the three-dimensional structure of this newly identified assembly intermediate by cryo-electron microscopy. Our structure indicates that the U2–BS helix has formed in this pre-A complex, but is not yet clamped by the HEAT domain of the Hsh155 protein (Hsh155[HEAT]), which exhibits an open conformation. The structure further reveals a large-scale remodelling/repositioning of the U1 and U2 snRNPs during the formation of the A complex that is required to allow subsequent binding of the U4/U6.U5 tri-snRNP, but that this repositioning is blocked in the pre-A complex by the presence of Prp5. Our data suggest that binding of Hsh155[HEAT] to the bulged BS-A of the U2–BS helix triggers closure of Hsh155[HEAT], which in turn destabilizes Prp5 binding. Thus, Prp5 proofreads the branch site indirectly, hindering spliceosome assembly if branch-site mutations prevent the remodelling of Hsh155[HEAT]. Our data provide structural insights into how a spliceosomal helicase enhances the fidelity of pre-mRNA splicing.

To isolate a spliceosome assembly intermediate formed directly before the A complex that still contains Prp5 (Extended Data Fig. 1a), we carried out splicing in *S. cerevisiae* cell extracts with an actin (*Act*) pre-mRNA in which the BS-A is deleted (Extended Data Fig. 1b). With this ΔBS-A substrate, splicing is blocked before catalytic step 1 (Extended Data Fig. 1c), consistent with previous results[6]. Spliceosomal complexes formed on ΔBS-A *Act* pre-mRNA lack the U4/U6.U5 tri-snRNP, but contain stoichiometric amounts of U1 and U2 snRNPs (Extended Data Fig. 1d). The proteins Prp5, Msl5 and Mud2 are also abundant, whereas Cus2 is absent (Extended Data Fig. 1e and Supplementary Table 1), indicating that these complexes stall after Prp5 hydrolyses ATP, but before the tri-snRNP has docked. We next carried out single-particle cryo-electron microscopy (cryo-EM) and determined the structure of the ΔBS-A complex at an average resolution of 5.9 Å (ranging from roughly 4.5 Å for U1 to approximately 15 Å for U2) (Extended Data Table 1 and Extended Data Fig. 2). Further classification and multibody refinement improved the resolution of the stable U1 snRNP region and adjacent U2 5′ region to 4.1 Å and 8.3 Å, respectively (Extended Data Fig. 2). By fitting known X-ray structures of spliceosome components into the EM density map

(Extended Data Table 2), together with protein crosslinking coupled with mass spectrometry (CXMS) (Supplementary Table 2), we generated a three-dimensional (3D) model of the ΔBS-A complex (Fig. 1). This complex consists of two major elongated domains—comprising the U1 and bipartite U2 snRNPs—that are connected by two main bridges (Fig. 1).

The U2–BS helix is not clamped by Hsh155

Hsh155[HEAT] adopts a closed conformation after U2 has integrated stably into the spliceosome, clamping the U2–BS helix and binding the bulged BS-A in a pocket formed by Rds3 (PHF5A in humans; see Extended Data Fig. 1f for a summary of yeast and human protein names) and HEAT repeats 15–17 of Hsh155 (refs. [7–10]) (Extended Data Fig. 1g). It is unclear at present what triggers this functionally important structural change. In the ΔBS-A complex, Hsh155[HEAT] exhibits an open conformation, as in the isolated human 17S U2 (ref. [11]) but in notable contrast to its conformation in yeast A complexes (Figs. 1, 2a and Extended Data Fig. 3a) and pre-B, B and B[act] complexes[7,8,12,13]. Stem–loop (SL) IIa of U2 snRNA is

[1]Department of Structural Dynamics, MPI for Biophysical Chemistry, Göttingen, Germany. [2]Cellular Biochemistry, MPI for Biophysical Chemistry, Göttingen, Germany. [3]Bioanalytical Mass Spectrometry, MPI for Biophysical Chemistry, Göttingen, Germany. [4]Bioanalytics Group, Institute for Clinical Chemistry, University Medical Center Göttingen, Göttingen, Germany. ✉e-mail: hstark1@gwdg.de; reinhard.luehrmann@mpibpc.mpg.de
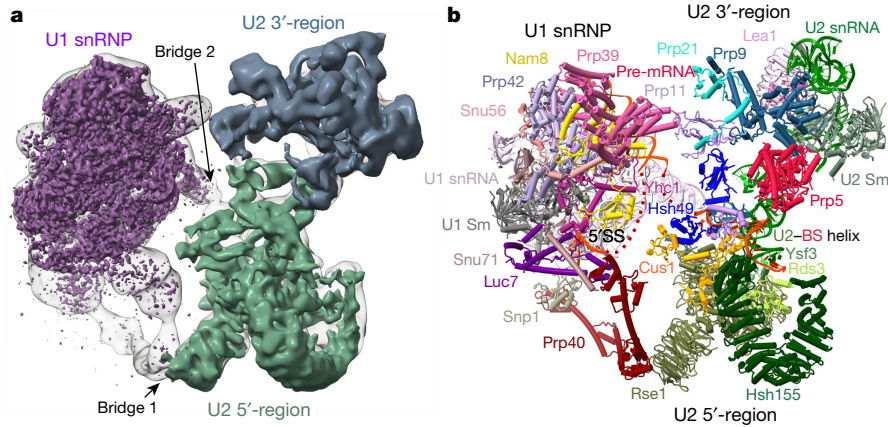
Fig. 1 | Three dimensional cryo-EM model of the yeast pre-A complex.
a, **b**, EM density map (**a**) and molecular architecture (**b**) of the *S. cerevisiae* spliceosomal pre-A complex. **a**, Purple, better-resolved U1 density; grey blue and green, better-resolved densities of the 3′- and 5′-regions of U2 snRNP; translucent grey, cryo-EM map of the pre-A complex.

bound by Hsh155, Cus1 and Prp9 in the ΔBS-A complex, in a similar manner to that seen in the human 17S U2 snRNP and subsequently formed spliceosomal complexes (Extended Data Fig. 3b). An extended helical density element is located directly upstream of SLIIa. This element is longer than the U2 branchpoint-interacting stem-loop (BSL), which is found in the isolated U2 snRNP and sequesters U2 nucleotides that base pair with the branch site[11,14] (Extended Data Fig. 3c, d). A modelled, extended U2–BS helix—lacking a bulged BS-A and comprising 13 base pairs—fits well into this density element (Extended Data Fig. 3c), indicating that an extended U2–BS helix has formed. On the basis of CXMS data, the Prp11 zinc finger (Prp11^{ZnF}) could be positioned at the top of the

U2–BS helix (Extended Data Fig. 3e, f), akin to its position in A, pre-B, B and B^{act} complexes[7,8,12,13,15]. In the ΔBS-A complex, the U2–BS helix is located further away from the carboxy (C)-terminal HEAT repeats of Hsh155^{HEAT} compared with its position in the A to B^{act} complexes, and it is not sequestered by Hsh155^{HEAT} (Fig. 2a and Extended Data Fig. 3g). Thus, formation of the U2–BS helix alone does not appear to trigger closure of Hsh155^{HEAT}. As the U2–BS helix has formed, but Hsh155^{HEAT} still exhibits an open conformation and Prp5 is stably bound (see below), we conclude that ΔBS-A complexes are stalled at a pre-A-complex stage, after Prp5-mediated formation of the U2–BS helix, but during/before it carries out its proofreading function.
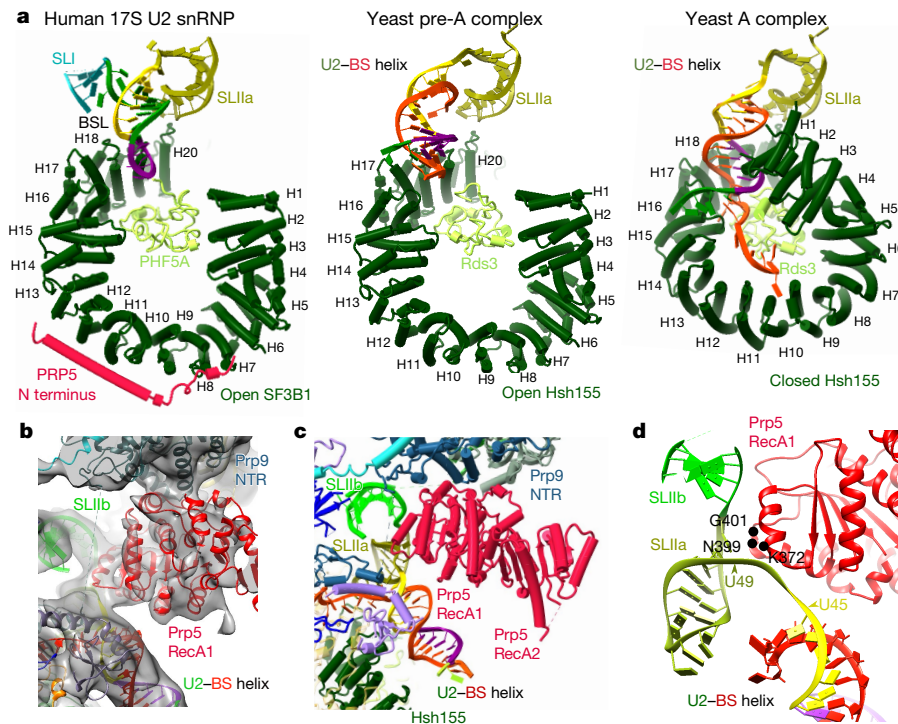
Fig. 2 | The Hsh155 HEAT domain has an open conformation in the pre-A complex. a, Conformation of the SF3B1 and Hsh155 HEAT domains and position of the U2–BS helix and U2 snRNA SLI and SLIIa in human 17S U2 snRNP (Protein DataBank (PDB) (https://www.rcsb.org) accession number 6Y5Q) and in the *S. cerevisiae* pre-A and A complexes (PDB 6G90). These domains were aligned via Hsh155 heat repeats 19–20, Rse1^{BPA} and U2 SLIIa. Olive green, SLIIa nucleotides; reddish orange, pre-mRNA branch-site nucleotides; purple, BSL nucleotides that later form the U2–BS helix; yellow, BSL nucleotides forming the extended part of the U2–BS helix; dark green, remaining BSL nucleotides; blue, SLI. **b**, Fit of Prp5^{RecA1} into the pre-A EM density. **c**, Location of the Prp5 RecA1 and RecA2 domains in the pre-A complex. **d**, Prp5^{RecA1} contacts U2 snRNA nucleotides that connect the U2–BS helix to U2 SLIIa. The positions of Prp5 amino acids (located outside of the SAT motif) that when mutated suppress branch-site mutations are indicated in black.

Article

Location of Prp5 in the pre-A complex

Prp5 was initially proposed to 'proofread' formation of the U2–BS helix by coordinating the rate of U2–BS base pairing with its ATPase activity[5]. However, more recent studies uncovered a correlation between increased Prp5 retention in early spliceosomal complexes and decreased tri-snRNP recruitment, suggesting that the physical presence of Prp5, rather than its ATPase activity, has a key role in its proofreading function[16]. In the human 17S U2 snRNP, the Prp5 RecA (Prp5[RecA]) domains sequester the BSL together with the TAT–SF1 protein and the C-terminal HEAT repeats of SF3B1 (ref. [11]), thereby preventing formation of the U2–BS helix. We find here that, in the yeast pre-A complex, Cus2 (human TAT–SF1) has dissociated and the Prp5[RecA] domains are located between the U2 3′-region and the U2–BS helix (Fig. 2b, c and Extended Data Fig. 4a–d). Compared with their position in 17S U2 snRNP, the Prp5[RecA] domains in the pre-A complex are located further away from Hsh155[HEAT], with RecA1 fitting well into the niche formed by the amino (N)-terminal region of Prp9 (Prp9[NTR]), with which it interacts, and the U2–BS helix (Fig. 2b, c and Extended Data Fig. 4c–e). Moreover, Prp5[RecA1] now contacts the U2 snRNA strand that connects SLIIa and the U2–BS helix (Fig. 2d). This is consistent with the crosslinks of Prp5 to this region (nucleotides 45–49) of yeast U2 snRNA observed previously with a Prp5-associated *S. cerevisiae* spliceosomal complex (designated the Prp5-associated intermediate complex, or FIC) formed on a pre-mRNA with a mutated branch site[16]. Finally, CXMS indicates that the NTR of Prp5 interacts extensively with Hsh155 HEAT repeats 1–7 in the pre-A complex (Extended Data Fig. 4a, b), consistent with previous biochemical studies[17].

Pre-A formation involves U2 remodelling

Comparison of the structures of the yeast pre-A complex and human 17S U2 snRNP suggests that, in addition to Prp5[RecA], there are other changes in the organization of U2 components during pre-A formation. Relative to its position in 17S U2, the U2 3′-region has moved (Extended Data Fig. 4d, e) and its new location in the pre-A complex is stabilized by newly formed contacts between the U2 3′- and 5′-regions (Extended Data Fig. 4d, e). This movement is a prerequisite for formation of the pre-A complex, as it generates the binding pocket for the repositioned Prp5[RecA1] (Fig. 2b, c and Extended Data Fig. 4). In the pre-A complex, Prp5[RecA1] establishes new contacts with the shifted 3′-region by interacting with Prp9[NTR] (Fig. 2b, c and Extended Data Fig. 4), preventing the further movement of the U2 3′-region that is ultimately required to allow the tri-snRNP to dock to the A complex (see below).

Prp40 bridges U1 and U2 snRNP in pre-A complex

The U1 snRNP structure in the pre-A complex (Extended Data Fig. 5a) is highly similar to that observed in the yeast E, A and pre-B complexes[13,15,18], indicating that U1 does not undergo substantial remodelling during early spliceosome assembly. As in the aforementioned complexes, base pairing between the 5′-splice site and the U1 snRNA is also stabilized by Yhc1 and Luc7 in the pre-A complex (Extended Data Fig. 5b). Although only three FF domains of Prp40 (domains 4–6) could be located in the yeast E complex[18], CXMS allowed us to map all six of the FF domains of Prp40 in the pre-A complex (Fig. 1 and Extended Data Fig. 5c, d). FF1 and FF6 bind Luc7 and Snp1, respectively, tethering Prp40 to U1, while FF2–FF5 form an extended binding platform that interacts with numerous proteins, including Snu71 and U2 Rse1. The interaction of FF4 with the WD40 β-propeller domain B (BPB) of Rse1 forms a bridge between the U1 and U2 snRNPs (denoted bridge 1) (Extended Data Figs. 5d, 6a, b) that is not observed in yeast A complexes[15]. Bridge 1 also contains the C-terminal region of Snu71, as numerous crosslinks between it and the FF2 and FF3 domains of Prp40, as well as with Rse1[BPB], are detected (Extended Data Fig. 5d). In the pre-A

complex, U1 and U2 are connected by a second bridge comprising intron nucleotides upstream of the branch site (Extended Data Fig. 6a, c). Although the Prp40 WW domain and Msl5–Mud2 could not be localized based solely on the EM density, CXMS indicates that Msl5–Mud2 is likely to be located near the U2–BS in the pre-A complex, and furthermore remains bound to the Prp40 WW domain (Extended Data Fig. 6d and Supplementary Table 2).

Pre-A complex with a U257A branch-site mutation

Mutations in the conserved yeast branch-site sequence upstream of the BS-A that weaken the U2–BS interaction—including a U-to-A mutation at position 257 of *Act* pre-mRNA (two nucleotides upstream of the BS-A) (Extended Data Fig. 1b)—do not completely block splicing but do lead to the accumulation of spliceosomes in which Prp5 is retained but the tri-snRNP has not yet joined[5,16]. To determine whether this mutation stalls spliceosome assembly at the pre-A stage, we purified the complexes that form on *Act* U257A pre-mRNA and determined their cryo-EM structure (Extended Data Fig. 7). The RNA and protein compositions of the purified U257A and ΔBS-A complexes were identical (Extended Data Fig. 1d, e and Supplementary Table 1), and an overlay of the U257A and ΔBS-A complexes revealed a highly similar, if not identical, structure at the present level of resolution (Extended Data Fig. 7). Indeed, the structural model of the ΔBS-A pre-A complex fits well without further adjustment into the EM density of the U257A complex (Extended Data Fig. 7). Thus, with the U257A mutant, an extended U2–BS helix has also formed, Hsh155[HEAT] is in an open conformation, and Prp5[RecA1] is docked to Prp9[NTR] and situated close to the U2 snRNA, indicating that U257A complexes are also stalled at the same pre-A stage. The highly similar structure of both pre-A complexes indicates that they represent a physiologically relevant intermediate that—at least in the case of U257A—can also progress along the wild-type spliceosome assembly pathway.

Dynamics of the pre-A to A transition

Comparison of our pre-A complex with the previously published yeast A complex[15] reveals that the transition from the pre-A to the A complex involves large-scale remodelling that requires displacement of Prp5 (Fig. 3, Extended Data Fig. 8 and Supplementary Video 1). First, the U2 3′-region rotates by roughly 55° relative to the U2 5′-region during A-complex formation. In the pre-A complex, this rotational movement is prohibited by the Prp5[RecA] domains, which bind in a mutually exclusive manner with the new position of Prp9[NTR], the long α-helix of Prp21 and the Prp11 β-sandwich, in the subsequently formed A complex. Second, U1 snRNP rotates by roughly 45° during the transition from the pre-A to the A complex, such that Prp39 now interacts with Lea1. A prerequisite for U1 movement is the dissociation of Prp40 from Rse1, and thus the apparent dissolution of U1–U2 bridge 1 (Fig. 3). The repositioning of U1 and the U2 3′-region is essential to generate the binding platform needed to dock the U4/U6.U5 tri-snRNP during formation of the pre-B complex (Fig. 3). Our studies thus provide a structural explanation for why the docking of tri-snRNP is inhibited when Prp5 is retained in yeast prespliceosomes[16].

Mechanism of proofreading by Prp5

The cryo-EM structures presented here provide structural insights into the mechanism by which Prp5 proofreads the U2–BS helix (Fig. 4). The pre-A and 17S U2 structures are consistent with a model in which, after U2 interacts with the E complex, ATP hydrolysis by Prp5 leads to release of Cus2 and unwinding of the BSL. This allows formation of the U2–BS helix and repositioning of the U2 3′-region and Prp5[RecA], generating the pre-A complex (Fig. 4). The new proofreading (or rather, 'fidelity checkpoint') position of Prp5[RecA] in the pre-A complex transiently prevents the further movement of the U2 3′-domain needed to form an A
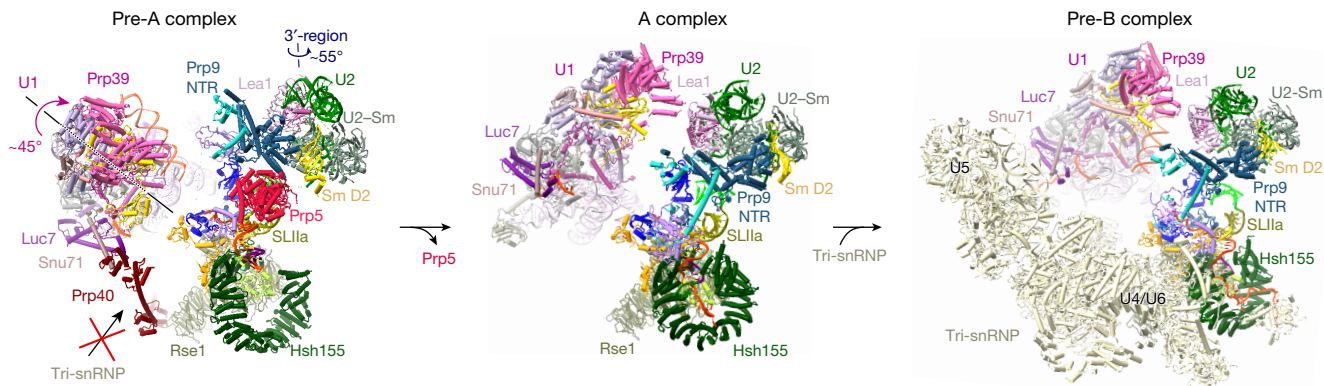
Fig. 3 | Prp5 blocks the repositioning of U1 and U2 snRNPs that is required to form the tri-snRNP-binding site. Molecular organization of U1 and U2 snRNPs in *S. cerevisiae* pre-A, A (PDB 6G90) and pre-B complexes (PDB 5ZWM and PDB 5ZWN). Movements of U1 and U2 snRNPs during the pre-A to A transition are indicated by curved arrows. All structures were aligned as in Fig. 2. For simplicity, the U1 snRNA stem-loops in the poorly resolved region of the U1 snRNP are not shown in the pre-A, A and pre-B complexes.

complex that can subsequently bind the tri-snRNP. As deletion of the BS-A hinders the closure of Hsh155[HEAT], but does not affect the stability of the U2–BS helix per se, correct binding of the bulged BS-A by Hsh-155[HEAT] and Rds3 is likely to be a major trigger for the conformational change in Hsh155[HEAT]. Furthermore, in the pre-A complex, the U2–BS helix is probably flexible, enabling it to intermittently move closer to Hsh155[HEAT], which 'probes' for its presence. Thus, we propose that when a stable U2–BS with a bulged BS-A is formed, movement of the U2–BS into the open Hsh155[HEAT] leads to insertion of the BS-A into its binding pocket and closure of the HEAT domain (Fig. 4). Previous mutational analyses of Hsh155 indicated that alignment of the U2–BS duplex with conserved, positively charged amino acids in the C-terminal half of Hsh155 is crucial for closure[19], and this alignment could thus help to properly position the bulged BS-A in its binding pocket.

Closure of Hsh155[HEAT] would destabilize not only the Prp5[NTR] that binds to it, but also the Prp5[RecA] domains (Fig. 4). Although the latter do not interact with Hsh155[HEAT], the coordinated movement of Hsh155[HEAT] and Prp5[NTR] and the U2 snRNA nucleotides contacted by Prp5[RecA1] could also lead to the displacement of Prp5[RecA] and subsequent release of the entire Prp5 protein. Release of Prp5 would then allow rotation of the U2 3′-region and formation of an A complex (Fig. 4). If the U2–BS helix lacks a bulged BS-A, binding of the latter by Hsh155[HEAT] and Rds3 would be blocked (Fig. 4). This would prevent the closure of Hsh155[HEAT] and release of Prp5, blocking the progression of spliceosome assembly and potentially targeting the stalled complex for discard (Fig. 4). As branch-site mutations that destabilize the U2–BS duplex also hinder spliceosome assembly and lead to Prp5 retention[5,16], the stability of the U2–BS duplex per se may also affect the conformational state of Hsh155[HEAT] in pre-A complexes. A destabilized U2–BS helix could also potentially prevent proper bulging of the BS-A[20], and in this way hinder closure of Hsh155[HEAT].

Various mutations in Prp5 can suppress branch-site mutations, including those in the Prp5 DPLD motif (which is important for the interaction of Prp5 with the U2 snRNP[21]) and also mutations in its SAT motif and those in adjacent regions of RecA1 that have no effect on ATPase activity, such as K372E, N399D and G401E[5,21]. The Prp5[N399D] mutant, the SAT mutant TAG and the DPLD mutant AAAA have reduced affinity for yeast prespliceosomes and enhance the binding of tri-snRNPs to spliceosomes[16]. Our pre-A structures provide insight into how some of these mutations may destabilize Prp5 binding. In pre-A complexes, the Prp5[NTR] containing the DPLD motif still interacts with Hsh155[HEAT] (Extended Data Fig. 4b), and thus mutation of this motif is likely to destabilize the Prp5[NTR]–Hsh155[HEAT] interaction. In pre-A complexes, amino acids K372, N399 and G401 are located in the Prp5 region that interacts with the
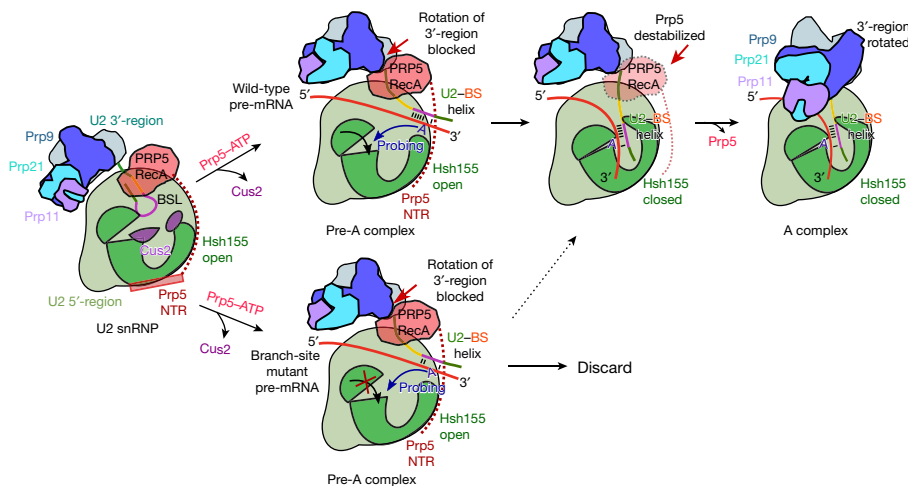
Fig. 4 | Model for Prp5-mediated proofreading of the U2–BS helix. For simplicity, in this depiction of our proposed mechanism of proofreading by Prp5, Rds3 of the BS-A-binding pocket is not shown, and the U1 snRNP has been omitted from the spliceosomal pre-A and A complexes. The closure of Hsh155[HEAT] and release of Prp5 may be highly coordinated events that occur simultaneously rather than sequentially. The dashed arrow indicates that some mutations in the branch-site sequence, including U257A, do not completely abolish conversion of the pre-A complex into the A complex.

Article

single-stranded U2 snRNA between SLIIa and the U2–BS helix (Fig. 2d). Consistent with less-stable Prp5 binding, these mutations introduce a negative charge that would destabilize the RecA1–U2 snRNA interaction. Although the SAT motif is not located at a Prp5–protein interface, some SAT mutations might alter the conformation of RecA1, thereby indirectly destabilizing its interaction with Prp9[NTR]. Indeed, several SAT mutations alter the equilibrium between the open and closed conformations of the Prp5[RecA] domains[22]. Together, our results indicate that Prp5 does not proofread the U2–BS helix directly, but instead proofreads the RNP conformation of pre-A complexes, and hinders progression of spliceosome assembly if mutations in the branch site alter the formation of a productive, closed conformation of Hsh155[HEAT].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03789-5.

1. Ruby, S. W., Chang, T. H. & Abelson, J. Four yeast spliceosomal proteins (PRP5, PRP9, PRP11, and PRP21) interact to promote U2 snRNP binding to pre-mRNA. *Genes Dev.* **7**, 1909–1925 (1993).
2. Abu Dayyeh, B. K., Quan, T. K., Castro, M. & Ruby, S. W. Probing interactions between the U2 small nuclear ribonucleoprotein and the DEAD-box protein, Prp5. *J. Biol. Chem.* **277**, 20221–20233 (2002).
3. Perriman, R., Barta, I., Voeltz, G. K., Abelson, J. & Ares, M. Jr. ATP requirement for Prp5p function is determined by Cus2p and the structure of U2 small nuclear RNA. *Proc. Natl Acad. Sci. USA* **100**, 13857–13862 (2003).
4. Xu, Y. Z. et al. Prp5 bridges U1 and U2 snRNPs and enables stable U2 snRNP association with intron RNA. *EMBO J.* **23**, 376–385 (2004).
5. Xu, Y. Z. & Query, C. C. Competition between the ATPase Prp5 and branch region-U2 snRNA pairing modulates the fidelity of spliceosome assembly. *Mol. Cell* **28**, 838–849 (2007).
6. Smith, D. J., Konarska, M. M. & Query, C. C. Insights into branch nucleophile positioning and activation from an orthogonal pre-mRNA splicing system in yeast. *Mol. Cell* **34**, 333–343 (2009).
7. Rauhut, R. et al. Molecular architecture of the *Saccharomyces cerevisiae* activated spliceosome. *Science* **353**, 1399–1405 (2016).
8. Yan, C., Wan, R., Bai, R., Huang, G. & Shi, Y. Structure of a yeast activated spliceosome at 3.5 Å resolution. *Science* **353**, 904–911 (2016).
9. Haselbach, D. et al. Structure and conformational dynamics of the human spliceosomal B[act] complex. *Cell* **172**, 454–464.e11 (2018).
10. Zhang, X. et al. Structure of the human activated spliceosome in three conformational states. *Cell Res.* **28**, 307–322 (2018).
11. Zhang, Z. et al. Molecular architecture of the human 17S U2 snRNP. *Nature* **583**, 310–313 (2020).
12. Plaschka, C., Lin, P. C. & Nagai, K. Structure of a pre-catalytic spliceosome. *Nature* **546**, 617–621 (2017).
13. Bai, R., Wan, R., Yan, C., Lei, J. & Shi, Y. Structures of the fully assembled *Saccharomyces cerevisiae* spliceosome before activation. *Science* **360**, 1423–1429 (2018).
14. Perriman, R. & Ares, M. Jr. Invariant U2 snRNA nucleotides form a stem loop to recognize the intron early in splicing. *Mol. Cell* **38**, 416–427 (2010).
15. Plaschka, C., Lin, P. C., Charenton, C. & Nagai, K. Prespliceosome structure provides insights into spliceosome assembly and regulation. *Nature* **559**, 419–422 (2018).
16. Liang, W. W. & Cheng, S. C. A novel mechanism for Prp5 function in prespliceosome formation and proofreading the branch site sequence. *Genes Dev.* **29**, 81–93 (2015).
17. Tang, Q. et al. SF3B1/Hsh155 HEAT motif mutations affect interaction with the spliceosomal ATPase Prp5, resulting in altered branch site selectivity in pre-mRNA splicing. *Genes Dev.* **30**, 2710–2723 (2016).
18. Li, X. et al. A unified mechanism for intron and exon definition and back-splicing. *Nature* **573**, 375–380 (2019).
19. Carrocci, T. J., Paulson, J. C. & Hoskins, A. A. Functional analysis of Hsh155/SF3b1 interactions with the U2 snRNA/branch site duplex. *RNA* **24**, 1028–1040 (2018).
20. Kennedy, S. D., Bauer, W. J., Wang, W. & Kielkopf, C. L. Dynamic stacking of an expected branch point adenosine in duplexes containing pseudouridine-modified or unmodified U2 snRNA sites. *Biochem. Biophys. Res. Commun.* **511**, 416–421 (2019).
21. Shao, W., Kim, H. S., Cao, Y., Xu, Y. Z. & Query, C. C. A U1-U2 snRNP interaction network during intron definition. *Mol. Cell. Biol.* **32**, 470–478 (2012).
22. Beier, D. H. et al. Dynamics of the DEAD-box ATPase Prp5 RecA-like domains provide a conformational switch during spliceosome assembly. *Nucleic Acids Res.* **47**, 10842–10851 (2019).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Preparation of yeast whole-cell extracts

Yeast whole-cell extracts were prepared from the *Saccharomyces cerevisiae* 3.2.AID/CRL2101 strain (*MAT*alpha, *prp2-1, ade2, his3, lys2-801, ura3*) (a gift from R.-J. Lin)[23]. Yeast were grown in a 100-litre fermenter and extracts were prepared as previously described[7].

Affinity purification of pre-A complexes

Uncapped actin pre-mRNA lacking the branch-site adenosine (ΔBS-A *Act* pre-mRNA) was tagged at its 5′-end with three MS2 aptamers and transcribed in vitro using T7 RNA polymerase from a template prepared with the QuikChange II site-directed mutagenesis kit (Agilent). Yeast ΔBS-A pre-A spliceosomal complexes were assembled for 45 min at 23 °C in a 175 ml splicing reaction containing 40% yeast whole-cell extract and 1.8 nM ΔBS-A *Act* pre-mRNA with prebound MBP–MS2 fusion protein. The splicing reaction was subsequently chilled on ice and cleared by centrifugation for 10 min at 9,000 rpm at 4 °C in a Fibrelite F14-14 × 50 cy rotor (Thermo Fisher Scientific). It was then loaded onto two columns, each packed with 600 µl amylose resin (New England BioLabs) that were pre-equilibrated with GK75 buffer (20 mM HEPES-KOH pH 7.9, 1.5 mM MgCl$_2$, 75 mM KCl, 5% glycerol, 0.01% NP40, 0.5 mM dithiothreitol (DTT) and 0.5 mM phenylmethylsulfonyl fluoride (PMSF)). The matrix with bound complexes was washed with 3 ml GK75 buffer and spliceosomes were eluted with 15 mM maltose in GK75 buffer. For electron microscopy, peak elution fractions containing approximately 40 pmol of spliceosomal complexes were pooled (1 ml total volume), crosslinked with 0.2 mM BS3 (Thermo Fisher) for 1 h on ice, and loaded onto a 17 ml linear 10–30% (*v/v*) glycerol/0–0.1% glutaraldehyde gradient containing GK75 buffer. Samples were centrifuged for 17 h at 24,400 r.p.m. in a SureSpin 630 rotor (Sorvall) and collected manually from the top in 28 fractions of 555 µl each. Crosslinking was stopped by adding 50 mM glycine, pH 7.7, and incubating for 2 h on ice. Fractions were analysed by Cherenkov counting in a Tri-Carb 2100TR scintillation counter (Packard). Two peak fractions containing 2.4 pmol of spliceosomal complexes were buffer-exchanged to GK75 with no glycerol, and concentrated to 250 µl in an Amicon Ultra-0.5 centrifugal filter unit Ultracel-50 (Merck), and then used for preparation of cryo-EM grids. Pre-A complexes assembled on the U257A *Act* pre-mRNA were purified as described above, with the following modifications. The template for in vitro transcription of the U257A *Act* pre-mRNA was purchased from Genscript. Pre-A complexes were assembled in a 252 ml splicing reaction containing 1.7 nM pre-mRNA; before preparation of cryo-EM grids, three peak gradient fractions were buffer exchanged to GK75 buffer containing 0.3% glycerol and samples were concentrated to 100 µl.

RNA and protein composition of pre-A complexes

To determine the RNA and protein composition of the pre-A complexes, we purified the complexes essentially as described above, except that we used a 12-ml splicing reaction, we washed spliceosomes bound to an amylose matrix with 10 ml GK150 and 10 ml GK75 buffer, and we did not incubate the eluted complexes with BS3. Furthermore, the complexes were fractionated on a 10–30% glycerol gradient lacking glutaraldehyde by centrifugation in a TH660 rotor (Thermo Fisher Scientific) for 16 h at 21,500 r.p.m. RNA and proteins were separated on NuPAGE 4–12% Bis-Tris gels (Invitrogen) and visualized by staining with SYBR Gold (Invitrogen) and Coomassie, respectively. The entire lanes were cut into 23 slices (680 fmol ΔBS-A pre-A) or 16 slices (170 fmol U257A pre-A) and proteins were in-gel digested with trypsin overnight. Resulting peptides were separated on a C18 column using an UltiMate3000 (Dionex)

ultrahigh performance liquid chromatography system, and analysed by electrospray ionization mass spectrometry in a Thermo Scientific Q Exactive HF (ΔBS-A pre-A) or Orbitrap Exploris 480 mass spectrometer (U257A pre-A). Data were acquired using Thermo Exactive Series 2.8 SP1 and Orbitrap Exploris 480 3.0 software. The U257A pre-A complex was measured in duplicate, and a sum of both measurements is shown in Supplementary Table 1. Proteins were identified by searching fragment spectra against the *S. cerevisiae* Genomic Database (SGD; https://www.yeastgenome.org) using Mascot v.2.3.02 as a search engine. For immunoblotting, proteins were separated on denaturing 4–12% NuPAGE gels, transferred to Amersham Protran 0.2-µm nitrocellulose membranes (Cytiva), immunostained with an Amersham ECL Western Blotting Detection Kit (Cytiva), and visualized with an Amersham Imager 680 (Cytiva). Antibodies against the yeast Prp5 and Lea1 proteins were provided by S.-C. Cheng.

Protein–protein crosslinking and identification

For CXMS experiments, spliceosomes were assembled in a 400-ml (experiment 1) or 300-ml (experiment 2) splicing reaction containing 40% yeast whole-cell extract from the 3.2.AID/CRL2101 strain. Following MS2 affinity selection, purified spliceosomal complexes were crosslinked with 250 µM BS3 for 1 h at 8 °C in a total volume of 3 ml. The reaction was split in half and loaded onto two 17-ml 10–30% (*v/v*) glycerol gradients and centrifuged in a Surespin 630 rotor (Thermo Fisher Scientific) for 16 h at 24,400 r.p.m. The gradients were fractionated by hand from the top into 28 fractions. Three peak fractions from each gradient, containing approximately 15 pmol of pre-A complexes, were pooled and the crosslinked complexes were pelleted by ultracentrifugation in a S100-AT4 rotor (Thermo Fisher Scientific) and analysed as previously described[24]. After tryptic digestion, peptides were reverse-phase extracted using Sep-Pak Vac tC18 1cc cartridges (Waters) and fractionated by gel filtration on a Superdex Peptide PC3.2/30 column (GE Healthcare). Next, 50 µl fractions corresponding to an elution volume of 1.2–1.8 ml were analysed in triplicate on a Thermo Scientific Q Exactive HF-X (Experiment 1) or Orbitrap Exploris 480 mass spectrometer (Experiment 2) using Thermo Exactive Series 2.9 and Orbitrap Exploris 480 1.1 software, respectively. Protein–protein crosslinks were identified using the pLink 2.3.9 search engine (pfind.ict.ac.cn/software/pLink) and filtered at a false discovery rate (FDR) of 1% or 5% according to the developer's recommendations[25,26].

EM sample preparation and image acquisition

Purified ΔBS-A or U257A pre-A complexes were absorbed for 15 min or 25 min, respectively, to a thin-layer carbon film that was subsequently attached to R3.5/1 Quantifoil grids. Next, 3.8 µl of sample buffer was applied to the grid and excess buffer was removed using an FEI Vitrobot loaded with pre-wetted filter paper, with a blotting force of 7, blotting time of 6.5 s, at 100% humidity and 4 °C. The sample was subsequently vitrified by plunging into liquid ethane. All cryo-EM data of the ΔBS-A pre-A complex and dataset 1 of the U257A pre-A complex were acquired at 300 kV on a FEI Titan Krios electron microscope (Thermo Fisher Scientific), equipped with a Cs corrector. Images were recorded in integration mode at ×120,700 magnification, corresponding to a calibrated pixel size of 1.16 Å at the specimen level, using a Falcon III direct electron detector. Micrographs were recorded via a Thermo Fischer EPU 2.1, using an exposure time of 1.02 s with 40 movie frames and a total dose of 44 e$^-$ per Å^2. In total, 87,604 and 9,170 micrographs were recorded for the ΔBS-A pre-A complex and dataset 1 of the U257A pre-A complex, respectively. Dataset 2 of the U257A pre-A complex (18,332 micrographs) was acquired at 300 kV on an FEI Titan Krios electron microscope (Thermo Fisher Scientific), in integration mode at a calibrated pixel size of 1.06 Å at the specimen level, using a Falcon III direct electron detector. Micrographs were recorded via a Thermo Fischer EPU 2.1, using an exposure time of 1.02 s with 40 movie frames and a total dose of 58 e$^-$ per Å^2.

Article

Image processing

Frames were dose-weighted, aligned and summed with MotionCor 2.0 (ref. [27]). The defocus values and equiphase averaging (EPA) of the micrographs were determined using Gctf[28]. Micrographs with a defocus range of 1 μm to 4 μm and a resolution of better than 4.3 Å based on EPA estimation were retained for further processing. For the ΔBS-A pre-A complex, 74,230 out of 87,604 summed micrographs were further processed. Initially, approximately 5.4 million particles were automatically picked using Gautomatch (https://www2.mrc-lmb.cam.ac.uk/research/locally-developed-software/zhang-software/). They were then extracted with a box size of 440 × 440 pixels, and binned to 110 × 110 pixels (pixel size of 4.64 Å) in RELION 3.0 (http://www2.mrc-lmb.cam.ac.uk/relion/index.php/Main_Page). Several iterations of reference-free two-dimensional (2D) classification were performed in RELION 3.0, and 'bad classes' showing fuzzy or uninterpretable features were removed, yielding 3,143,491 'good particles'. A subset of 308,419 particles was used to generate an initial 3D map using the ab initio reconstruction function in cryoSPARC[29]. Using the ab initio model, this subset of particles was further 3D classified into five classes in RELION 3.0. The best class showing clear features of U1 and U2 snRNP was selected, and the more flexible U2 snRNP region of this map was erased using UCSF Chimera v.1.13.1 (ref. [30]). The remaining U1 snRNP was low-pass filtered to 35 Å and used as 3D reference for further 3D classification for the entire dataset. The retained 3,143,491 particles after 2D classification of the entire dataset were split into 3 subsets and each subset was 3D classified into 5 classes. The best classes of each subset were combined, yielding 986,393 particles, which were then centred and re-extracted to 200 × 200 pixels (pixel size of 2.32 Å) and further classified into four classes. After 20 iterations of consensus classification (7.5° sampling interval without local search), a mask was placed around the U1 snRNP and the local angular search range was limited to 20° with a finer sampling interval of 3.7°. The best class (with 504,547 particles) was selected, centred and re-extracted with an original pixel size of 1.16 Å with a box size of 400 × 400 pixels and refined with a mask around the U1 snRNP, resulting in a map at 4.3 Å resolution. Next, using the alignment parameters from the aforementioned masked 3D refinement, the 504,547 particles were focus classified with a mask around the high-resolution U1 core, into 4 classes. The best class (containing 226,656 particles) was selected and refined into a map of the entire pre-A complex with an average resolution of 5.9 Å (ranging from roughly 4.5 Å at the U1 region to roughly 15 Å at the U2 region). The U1 and U2 regions were further improved by multibody refinement to 4.1 Å and 10 Å respectively. To further improve the U2 region, we re-extracted the 504,547 particles into a smaller box size of 140 × 140 pixels (pixel size of 2.32 Å) with the U2 snRNP centred, and 3D classified into 5 classes with a mask around the U2 snRNP. The best class, with 160,894 particles, was selected and multibody refined with masks around the U2 5′- and 3′-regions, resulting in a map at 8.3 Å resolution for the 5′-region and one at 9.5 Å resolution for the 3′-region. All the aforementioned resolutions were estimated on the basis of the RELION gold-standard Fourier shell correlation (0.143 criterion).

For the U257A pre-A complex, initially 460,854 and 869,306 particles were extracted from dataset 1 and dataset 2, respectively, and rescaled to 110 × 110 pixels, to the same pixel size of 4.64 Å, in RELION 3.0 (http://www2.mrc-lmb.cam.ac.uk/relion/index.php/Main_Page). After several iterations of reference-free 2D classification, 697,892 'good particles' from the two datasets were combined and classified into four classes by 3D classification with only the U1 part as the starting model, to avoid model bias. No class resembling the structure of the mature A complex was observed. Three classes had no discernible structural features of U1 or U2 snRNPs. One class (of 240,145 particles) clearly exhibited the structure of a pre-A complex, and was selected, centred and re-extracted with a pixel size of 2.32 Å with a box size of 220 × 220 pixels. Re-extracted particles were further 3D classified into four classes with a mask around the U1 part, yielding one class that showed clear secondary structures. This class (80,853 particles) was selected and refined into a map of the entire pre-A complex with an average resolution of 10.4 Å. Multibody refinement improved the U1 part to 7.5 Å. To improve the U2 part, we further classified the 80,853 particles with a mask around the U2 part into four classes, and two classes showing clear U2 density were combined, 3D refined and multibody refined, yielding the U2 part of the structure with 13 Å resolution. All of the aforementioned resolutions were estimated on the basis of the RELION gold-standard Fourier shell correlation (0.143 criterion).

ΔBS-A pre-A model building and refinement

Templates for the U1 and U2 proteins and RNA were obtained wherever possible from published structures (Extended Data Table 2). The U1 snRNP components, except Prp40, were initially docked as rigid bodies into the 4.1 Å EM map of the U1 region. In the central part of the U1 snRNP (resolution ranging from 3.7 Å to 4.3 Å), side chains were manually adjusted into the map using Coot v.0.8.9.2 (ref. [31]). The entire model of the U1 snRNP, excluding Prp40, was combined and subjected to real-space refinement in PHENIX v.1.13-2998 (ref. [32]), with secondary-structure restraints. The solution structure of the Prp40 FF1 domain and the homology model of domains FF2–FF6 predicted by the SWISS-MODEL suite[33] were truncated to polyalanine, docked into the pre-A map as rigid bodies, and were not refined owing to the limited resolution. The model of Hsh155 (H1–H15) was based on human SF3B1 (H1–H15) but with the sequence changed to that of yeast Hsh155. The model of the ΔBS-A U2/BS helix (U2 nucleotides 32–46; Act1 pre-mRNA nucleotides 254–268) was generated by deleting the BS-A from the model of the wild-type U2–BS helix using Coot. All U2 snRNP components were docked into the U2 map as rigid bodies without further adjustments, except that Prp9 (amino acids 328–362) and Prp21 (amino acids 173–192) were slightly adjusted using Coot to better fit the EM density, and the linker between the U2–BS helix and SLIIa (U2 nucleotides 47–49) was de novo modelled using Coot. All modelled components in the U2 region were modelled as polyalanine and were not refined owing to the limited resolution. The structural model for the ΔBS-A pre-A complex was fit into the EM density obtained for complexes formed on the U257A mutant Act1 pre-mRNA. The video showing the structural dynamics seen during the transition from the pre-A to the A complex was generated using ChimeraX v1.1.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The coordinate files have been deposited in the Protein Data Bank (https://www.rcsb.org) as follows: U1 snRNP region, PDB accession number 7OQC; U2 snRNP region, PDB 7OQB; and composite truncated model of the pre-A complex, PDB 7OQE. The cryo-EM maps have been deposited in the Electron Microscopy Data Bank (https://www.ebi.ac.uk/pdbe/emdb/) as follows: U1 snRNP region of the ΔBS-A pre-A complex, EMD accession number 13029, and of the U257A pre-A complex, EMD 13031; U2 snRNP region of the ΔBS-A pre-A complex, EMD 13028, and of the U257A pre-A complex, EMD 13032; and overall reconstruction of the ΔBS-A pre-A complex, EMD 13033, and of the U257A pre-A complex, EMD 13030. We used the S. cerevisiae Genome Database (SGD; https://www.yeastgenome.org) in this study.

23. Yean, S. L. & Lin, R. J. U4 small nuclear RNA dissociates from a yeast spliceosome and does not participate in the subsequent splicing reaction. *Mol. Cell. Biol.* **11**, 5571–5577 (1991).
24. Bertram, K. et al. Cryo-EM structure of a human spliceosome activated for step 2 of splicing. *Nature* **542**, 318–323 (2017).
25. Yang, B. et al. Identification of cross-linked peptides from complex samples. *Nat. Methods* **9**, 904–906 (2012).

26. Chen, Z. L. et al. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat. Commun.* **10**, 3404 (2019).
27. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
28. Zhang, K. Gctf: real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
29. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
30. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
31. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
32. Adams, P. D. et al. *PHENIX*: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
33. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723 (1997).
34. Kastner, B., Will, C. L., Stark, H. & Lührmann, R. Structural insights into nuclear pre-mRNA splicing in higher eukaryotes. *Cold Spring Harb. Perspect. Biol.* **11**, a032417 (2019).
35. Wilkinson, M. E., Charenton, C. & Nagai, K. RNA splicing by the spliceosome. *Annu. Rev. Biochem.* **89**, 359–388 (2020).
36. Semlow, D. R. & Staley, J. P. Staying on message: ensuring fidelity in pre-mRNA splicing. *Trends Biochem. Sci.* **37**, 263–273 (2012).
37. Cordin, O. & Beggs, J. D. RNA helicases in splicing. *RNA Biol.* **10**, 83–95 (2013).
38. Kistler, A. L. & Guthrie, C. Deletion of MUD2, the yeast homolog of U2AF65, can bypass the requirement for sub2, an essential spliceosomal ATPase. *Genes Dev.* **15**, 42–49 (2001).
39. Zhang, M. & Green, M. R. Identification and characterization of yUAP/Sub2p, a yeast homolog of the essential human pre-mRNA splicing factor hUAP56. *Genes Dev.* **15**, 30–35 (2001).
40. O'Day, C. L., Dalbadie-McFarland, G. & Abelson, J. The *Saccharomyces cerevisiae* Prp5 protein has RNA-dependent ATPase activity with specificity for U2 small nuclear RNA. *J. Biol. Chem.* **271**, 33261–33267 (1996).
41. Wang, Q., Zhang, L., Lynn, B. & Rymond, B. C. A. A BBP-Mud2p heterodimer mediates branchpoint recognition and influences splicing substrate abundance in budding yeast. *Nucleic Acids Res.* **36**, 2787–2798 (2008).
42. Jacewicz, A., Chico, L., Smith, P., Schwer, B. & Shuman, S. 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**, 401–414 (2015).
43. Cretu, C. et al. Structural basis of splicing modulation by antitumor macrolide compounds. *Mol. Cell* **70**, 265–273.e8 (2018).
44. Lin, P. C. & Xu, R. M. Structure and assembly of the SF3a splicing factor complex of U2 snRNP. *EMBO J.* **31**, 1579–1590 (2012).
45. Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D* **74**, 531–544 (2018).
46. Ester, C. & Uetz, P. The FF domains of yeast U1 snRNP protein Prp40 mediate interactions with Luc7 and Snu71. *BMC Biochem.* **9**, 29 (2008).
47. Caspary, F. & Séraphin, B. The yeast U2A'/U2B complex is required for pre-spliceosome formation. *EMBO J.* **17**, 6348–6358 (1998).

Competing interests The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03789-5.
Correspondence and requests for materials should be addressed to H.S. or R.L.
Peer review information *Nature* thanks Aaron Hoskins, Navtej Toor and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.

Article
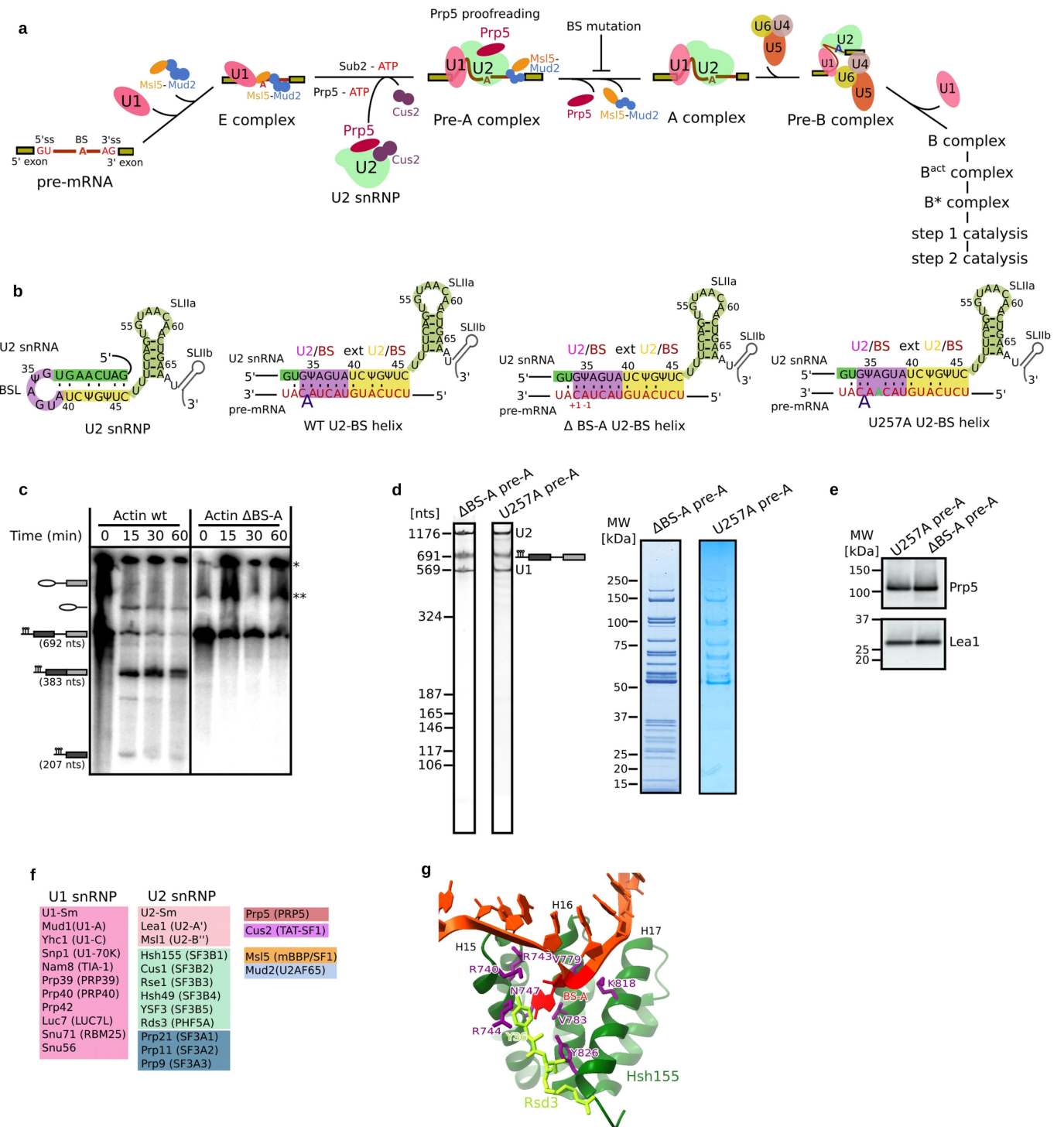
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Biochemical characterization of *S. cerevisiae* pre-A spliceosomal complexes. a, Early assembly of the *S. cerevisiae* spliceosome. Whereas Prp5 and Tat-SF1 (Cus2 in yeast) are stable components of the human 17S U2 snRNP, they appear to be less-stably associated with the yeast U2 snRNP. The spliceosome undergoes numerous structural and compositional rearrangements during its assembly and catalysis of pre-mRNA splicing[34,35]. Conserved DEXH/D-box RNA helicases are important driving forces for these rearrangements, and also ensure the proper recognition of the branch site (BS) and the 5′- and 3′-splice sites (ss) via proofreading mechanisms[36,37]. Initially an E complex is formed in an ATP-independent manner. In the yeast E complex (also denoted the commitment complex), the 5′-ss is bound by U1 snRNP, and the BS and 3′-end of the intron are bound by a heterodimer of Msl5 and Mud2. RNP rearrangements that lead to the stable association of U2 snRNP and enable the formation of a U2–BS helix—in which an adenosine is bulged, specifying it as the nucleophile for catalytic step 1 of splicing—require the ATP-dependent action of the DEAD-box RNA helicases Sub2 (refs. [38,39]; UAP56 in humans) and Prp5 (refs. [1–4,40]). U2 nucleotides that base pair with the BS are initially sequestered in a stem-loop structure denoted the BSL[11,14]. Sub2 may free the BS region by displacing Msl5 (refs. [41,42]), while Prp5 has been proposed to displace U2 snRNP proteins, including Cus2 (TAT–SF1 in humans), from the BSL[11,14]. This frees U2 nucleotides to base pair with the BS, and leads to the formation of the A complex with stably bound U2 snRNP. **b**, Structure of the BSL and U2–BS helices formed on an *Act* pre-mRNA wild-type (WT) BS (UACUA(**A**)C, where the BS-A is in bold), ΔBS-A (UACUA**C**) or U257A (UACAA(**A**)C) branch site. Note that the exact conformation of the U257A U2–BS helix is not clear. The U2–BS helix is highlighted in purple, and the extended U2–BS helix, in which the number and nature of base-pairing interactions varies depending on the pre-mRNA intron sequence, is highlighted in yellow. **c**, Deletion of the BS adenosine from the *Act* pre-mRNA stalls splicing before the first catalytic step. Splicing was performed in two independent experiments in yeast extract for 30 min at 23 °C with wild-type (lane 1) or ΔBS-A (lane 2) *Act* pre-mRNA containing MS2 aptamers for affinity purification. *Position of the loading well. **Band artefact not related to pre-mRNA splicing. For gel source data, see Supplementary Fig. 1. **d**, RNA (left gels) and protein (right gels) composition of purified yeast pre-A complexes formed on ΔBS-A and U257A *Act* pre-mRNA. RNA and protein were analysed on NuPAGE gels and visualized by staining with SYBR Gold or Coomassie, respectively, in two independent experiments. Note that fewer picomoles of the U257A pre-A complex were loaded onto the gel and, as a consequence, proteins of lower molecular weight are poorly or not at all visible. **e**, Prp5 is present in both U257A and ΔBS-A pre-A complexes. Proteins from affinity-purified U257A or ΔBS-A pre-A complexes (as indicated above each lane) were analysed by western blotting in two independent experiments with antibodies against *S. cerevisiae* Prp5 and Lea1 (used to ensure equal loading). **f**, Proteins localized in the *S. cerevisiae* pre-A complex and their human homologues (shown in parentheses). Only U1–70K, U1–A and U1–C have been identified as stable components of human U1 snRNPs. Human homologues of Snu56 and Prp42 have not been identified. **g**, Residues forming the BS-A-binding pocket. The bulged BS-A is bound in a pocket composed of residues R744, N747, V783 and Y826 of Hsh155 and residue Y36 of Rds3 (refs. [7,8,43]). The BS-A ribose and 5′-phosphate are also located near Hsh155 residues K740 and K818, respectively. Most of these residues are evolutionarily highly conserved. A Hsh155 K818A mutation is lethal, as are mutations in residues of Hsh155 that contact the backbone of nucleotides directly adjacent to the bulged BS-A[19]. However, many of the Hsh155 residues forming the BS-A binding pocket are nonessential. That is, single alanine substitutions at K740, R744, N747 and V783 do not affect yeast viability, but they do affect recognition of the branch site[19]. Substitutions with bulkier amino acids decrease the use of nonconsensus branch sites, whereas substitutions with smaller amino acids increase usage[19].

Article
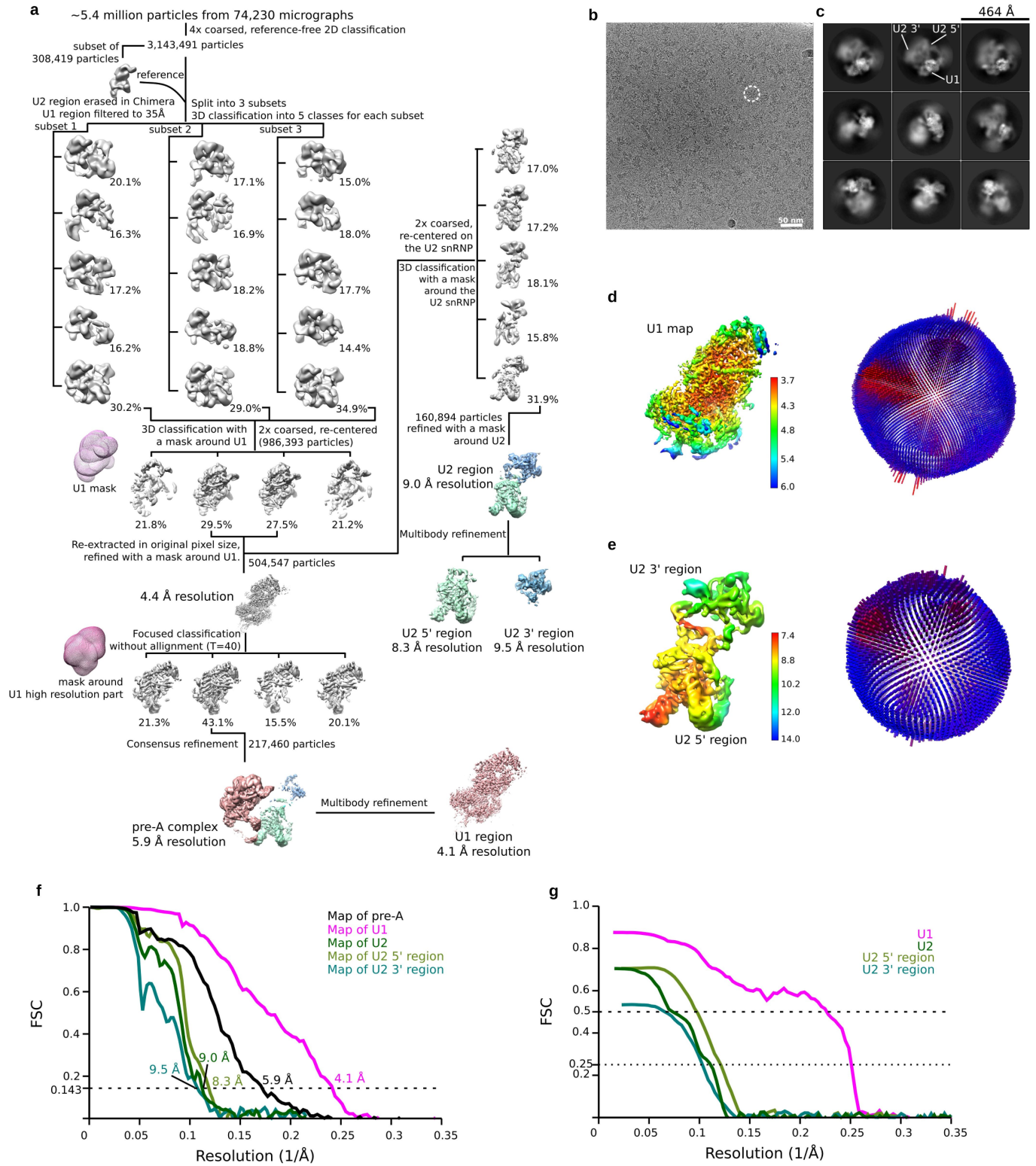
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Cryo-EM and image processing for the ΔBS-A pre-A complex. a, Computation sorting scheme. All major image-processing steps are depicted. For a more detailed explanation, see the Methods section on 'Image processing'. **b**, Typical cryo-EM micrograph (out of a total of 74,230) of the *S. cerevisiae* pre-A complex recorded at ×120,700 magnification with a Titan Krios microscope using a Falcon III direct electron detector operating in integration mode at a calibrated pixel size of 1.16 Å. **c**, Representative cryo-EM 2D class averages of the yeast pre-A complex reveal considerable flexibility between the U1 snRNP and the U2 snRNP. **d**, Left, local-resolution estimations of the cryo-EM reconstruction of the U1 snRNP. Right, plot showing the distribution of orientations for the particles contributing to the U1 reconstruction. **e**, Left, local-resolution estimations of the cryo-EM reconstructions of the U2 5′- and 3′-regions. Right, plot showing the distribution of orientations for the particles contributing to the reconstructions of the U2 5′- and 3′-regions. **f**, A Fourier shell correlation (FSC), calculated using the 'post-processing' routine in RELION, indicates a global resolution of 5.9 Å for the entire yeast pre-A complex and of 4.1 Å, 8.3 Å and 9.5 Å for the multibody-refined U1, U2 5′- and U2 3′-regions, respectively. The resolution was limited to 5.9 Å on average, owing to the movement of the U1 and U2 snRNPs relative to each other. Signal subtraction combined with local refinement improved the more stable U1 snRNP part to 4.1 Å resolution, whereas the bipartite U2 snRNP exhibited considerable internal flexibility and was refined only to 9 Å resolution. As the U2 5′-region—which is composed of the SF3b proteins, U2 SLIIa/b and the U2–BS helix—is attached to the more stable U1 snRNP and is therefore less flexible, further local classification and refinement improved the resolution of the U2 5′-region to 8.3 Å. The U2 3′-region—which comprises the U2 Sm core, U2 SLII, and Lea1 and U2-B″ bound to U2 SLIV, plus the adjacent SF3a core (that is, regions of Prp9, Prp11 and Prp21 whose crystal structure has been determined previously[44])—does not contact the U1 snRNP and therefore exhibits the greatest flexibility relative to the rest of the complex. **g**, Map versus model FSC curves for the U1, U2, U2 5′- and U2 3′-regions, using PHENIX mtriage[45].
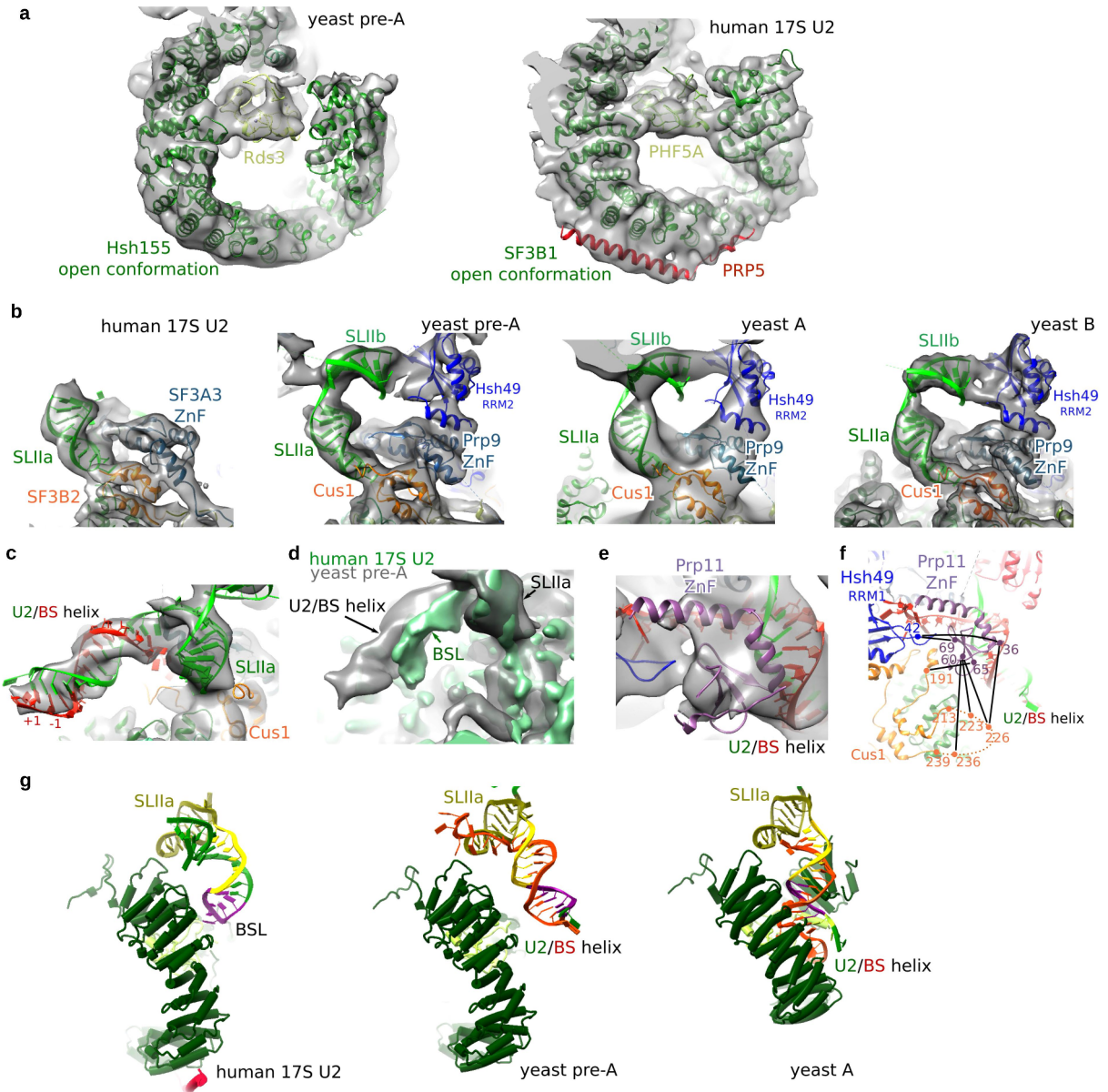
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Open conformation of Hsh155[HEAT] and spatial organization of the U2–BS helix and U2 SLIIa/b in the pre-A complex. a, Left, fit of Hsh155[HEAT] and Rds3 to the pre-A EM density; and right, fit of SF3B1, PHF5A and the N-terminal helix of Prp5 to the human 17S U2 density (PDB accession code 6Y5Q). Previous biochemical studies showed that the N-terminal region (NTR) of yeast Prp5 binds to HEAT repeats (HRs) 1–6 and HR 9–12 of Hsh155 (ref. [17]), and in the human 17S U2 snRNP cryo-EM structure, binding to SF3B1 HR 9–12 involves a long α-helix of the human Prp5[NTR] (ref. [11]). EM density that would accommodate an analogous α-helix of the yeast Prp5[NTR] is not apparent in the pre-A complex. However, CXMS indicates that the Prp5[NTR] still interacts extensively with Hsh155, albeit solely with HR 1–7 (see Extended Data Fig. 4b). Base pairing of U2 snRNA with the pre-mRNA branch site (BS) is stabilized by the major scaffolding protein SF3B1 (Hsh155 in *S. cerevisiae*) of the SF3b heteromeric complex, whose HEAT domain undergoes a conformational change during early spliceosome assembly. Hsh155[HEAT] and SF3B1[HEAT] exhibit an open conformation in the pre-A complex and the 17S U2 snRNP, respectively. The closed conformations of the HEAT domains of human SF3B1 and of yeast Hsh155 observed in the A to B[act] spliceosomal complexes are very similar. Likewise, the open conformation, which we observe here for the first time for Hsh155, also appears to be highly similar in human U2 snRNP and the *S. cerevisiae* pre-A complex. **b**, Similar spatial organization of U2 snRNA SLIIa, Prp9[ZnF] (SF3A3 in humans) and Cus1 (SF3B2 in humans) in the human 17S U2 snRNP (PDB 6Y5Q) and *S. cerevisiae* pre-A, A (PDB 6G90) and B (PDB 5NRL) complexes. Aligned via U2 SLIIa and HR 19–20 of SF3B1/Hsh155. In the pre-A complex, SLIIb can be localized downstream of SLIIa and is bound by RRM2 of Hsh49. **c**, Fit of a modelled 13-base-pair extended U2–BS helix, lacking a bulged A, to EM density adjacent to SLIIa in the pre-A complex. **d**, Overlay of EM density accommodating the U2–BS helix in the yeast pre-A complex (grey) and EM density accommodating the BSL in human 17S U2 (green) (PDB 6Y5Q). Aligned via U2 SLIIa and HR 19–20 of SF3B1/Hsh155. The sequences of the *S. cerevisiae* and human U2 BSLs are highly conserved[14], allowing a meaningful comparison with the fit of the yeast U2–BS helix. Although the BSL in yeast and human is predicted to form a 9-base-pair stem[14], in the human 17S U2 snRNP, the base of the BSL stem is contacted by a short helix of SF3A3 (designated the separator helix), which ensures that the stem is only 8 base pairs in length[11]. **e**, Fit in the pre-A EM density of the Prp11[ZnF] at the top of the U2–BS helix. **f**, Protein crosslinks supporting the positioning of the Prp11[ZnF] in the pre-A complex. Numbers (colour coded to match protein colours) indicate the positions of crosslinked lysine residues, which are connected by black lines. The Prp11[ZnF] appears to act independently of the other SF3a proteins and to dock to the end of the extended U2–BS helix concomitantly with, or soon after, its formation. As the Prp11[ZnF] has thus far been observed at this position solely after formation of the U2–BS helix, its location in the pre-A complex is consistent with the conclusion that a U2–BS helix has formed. It is likely that the Prp11[ZnF] and the Prp9 separator helix may cooperate in keeping additional intron nucleotides from interacting with the U2 snRNA, and at the same time stabilizing the end of the U2–BS helix. During clamping of the U2–BS helix by Hsh155[HEAT], Prp11[ZnF] moves together with the U2–BS helix and remains associated with the end of the helix in the A, pre-B, B and B[act] complexes. **g**, Side view showing that the U2–BS helix is located further away from the C-terminal HRs of Hsh155[HEAT] in the pre-A complex compared with its position in the *S. cerevisiae* A complex and the position of the BSL in human 17S U2 snRNP. Aligned via Hsh155 HR 19–20, Rse1[BPA] and U2 SLIIa. Olive green, SLIIa nucleotides; red orange, pre-mRNA BS nucleotides; purple, BSL nucleotides that later form the U2–BS helix; yellow, BSL nucleotides forming the extended part of the U2–BS helix; dark green, the remaining BSL nucleotides; blue, SLI. Movement away from Hsh155[HEAT] would be needed to free the SLI-containing 5′-end of U2 snRNA to undergo the rotational movements necessary to generate an extended U2–BS helix. During the transition from the pre-A to the A complex, the U2–BS helix moves back towards the Hsh155 C-terminal HRs such that the corresponding region of the U2 snRNA that contacts the C-terminal HRs in 17S U2 is located in a similar position in the A complex.
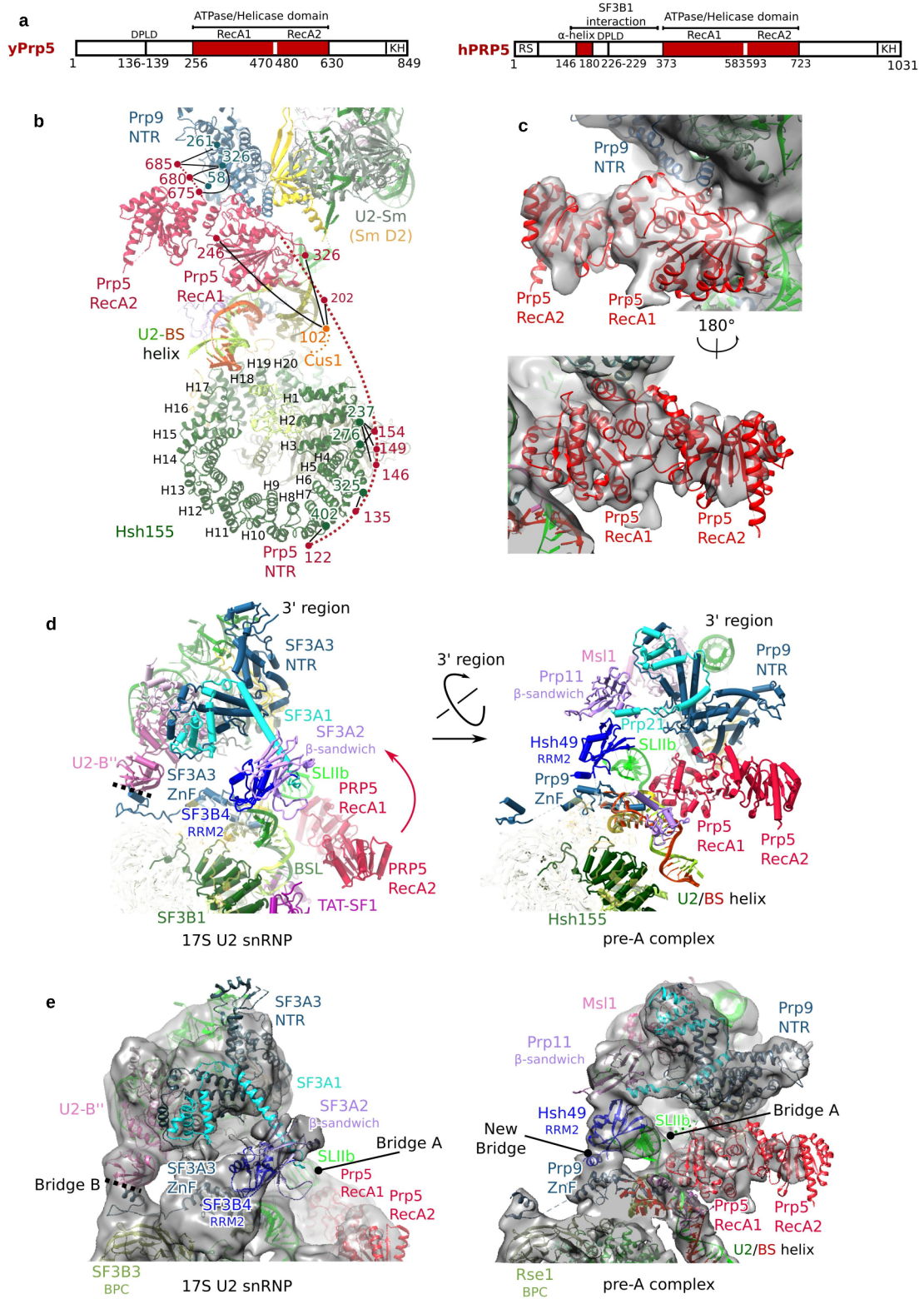
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Repositioning of the Prp5 RecA domains and the U2 3′-region during formation of the pre-A complex. a, Domain organization of the *S. cerevisiae* (y) and human (h) DEAD-box helicase Prp5, with the amino-acid boundaries of each domain indicated below. **b**, Protein crosslinks support the positions of the Prp5 NTR and RecA domains in the pre-A complex. Numbers (colour coded to match protein colours) indicate the positions of crosslinked lysine residues, which are connected by black lines. The proposed path of Prp5 amino acids located more N-terminally of the RecA domains is indicated by a dashed line. That the Prp5 NTR and RecA1 domains, but not RecA2 (and presumably also its C-terminal region), interact with other pre-A components is consistent with previous studies showing that, after destabilization of the U2 BSL, Prp5^{NTR} and Prp5^{RecA1} are sufficient for the subsequent ATP-independent function of Prp5 during A-complex formation[14]. **c**, Two different views of the fit of the Prp5 RecA1 and RecA2 domains in an open conformation into the pre-A EM density (low-pass filtered to 10 Å). A closed conformation of the Prp5^{RecA} domain does not fit well to the EM density (not shown). The open conformation of Prp5 found in the pre-A complex indicates that, after ATP hydrolysis, the RecA domains are able to transit spontaneously from the closed conformation back to the open conformation while probably remaining bound to U2. **d**, **e**, The positions of the Prp5^{RecA} domains and the U2 3′-region plus SF3a proteins, relative to SF3b, are different in the human 17S U2 snRNP and the yeast pre-A complex. Aligned via U2 SLIIa and HR 19–20 of SF3B1/Hsh155. A cryo-EM structure of an isolated *S. cerevisiae* U2 snRNP is currently lacking. However,

the high conservation of the sequence of yeast U2 proteins and their human homologues, and the similar structures of their conserved domains, suggests that the molecular architecture of the isolated U2 snRNP is similar in *S. cerevisiae* and humans. Thus, a comparison of the structures of the human 17S U2 snRNP and yeast pre-A complex reveals structural remodelling that the U2 snRNP most likely undergoes during formation of the pre-A complex. An alignment of the U2 5′-region in both complexes suggests that the U2 3′-region is repositioned after U2 stably interacts during formation of the pre-A complex. Specifically, the U2 3′-domain (that is, the 3′-region minus the SF3a core) and the Prp9^{NTR} rotate towards the Prp5^{RecA} domains, whereas the Prp11 β-sandwich and Hsh49^{RRM2} move towards Prp9^{ZnF}. The shifted position of the U2 3′-region is stabilized by different molecular bridges formed between the U2 3′- and 5′-regions. In the pre-A complex, the bridge formed by U2-B″ RRM2, Prp9 (human SF3A3) and Rse1^{BPC} (human SF3B3) in the 17S U2 snRNP (denoted bridge B) is disrupted, which allows the 3′-region to move further away from the Rse1^{BPC}. This then allows Hsh49^{RRM2} to dock on top of the Prp9^{ZnF}, and by binding to Prp9 on one side and the Prp11 β-sandwich domain on the other, a new bridge involving Hsh49^{RRM2} is formed. Moreover, in the isolated human 17S U2 snRNP, U2 SLIIb forms a second bridge (denoted bridge A) between the U2 3′- and 5′-regions that is not stabilized and is only poorly resolved. By contrast, in the pre-A complex, Hsh49^{RRM2} now binds to the loop of SLIIb and thereby stabilizes the position of SLIIb.
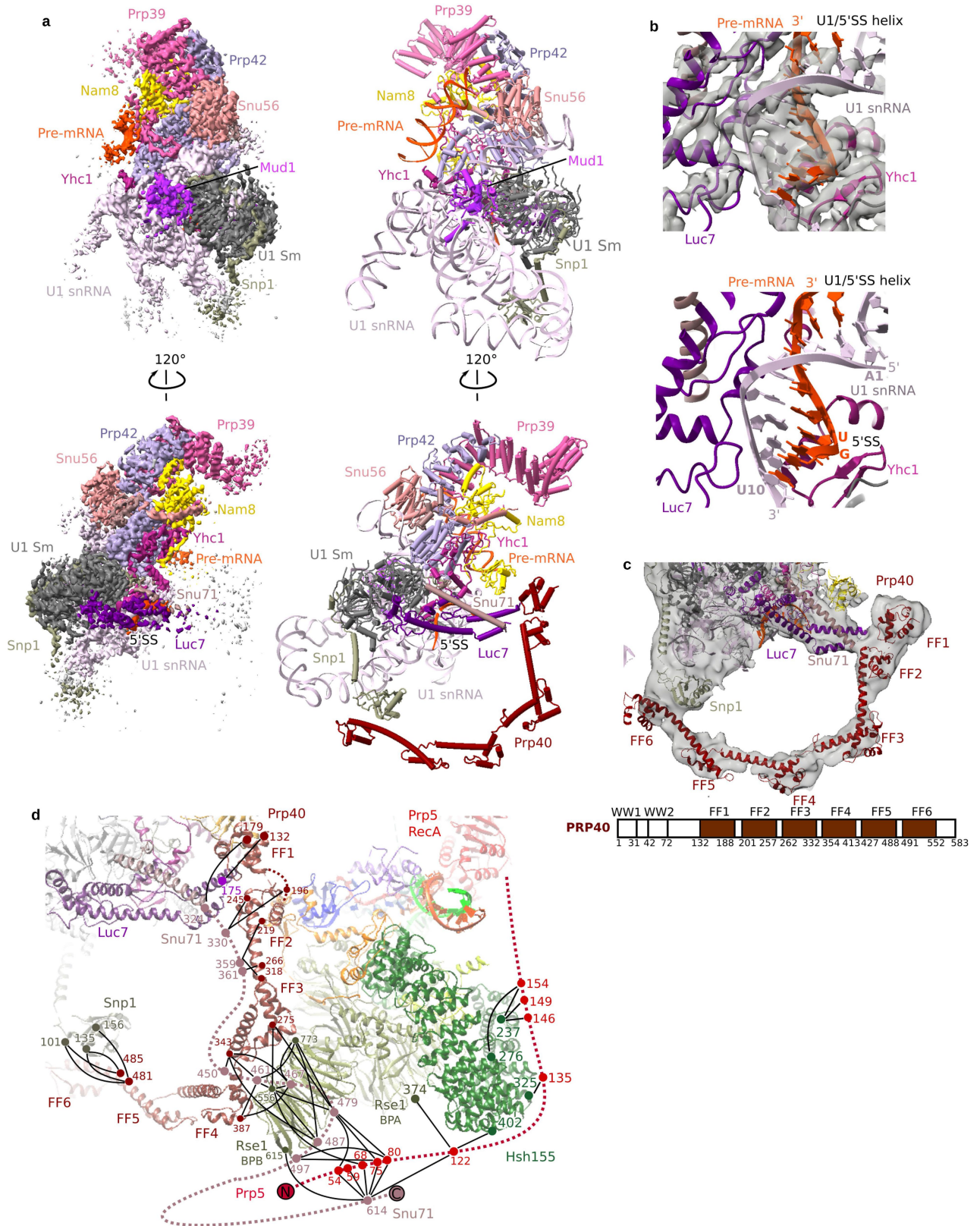
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Molecular architecture of the U1 snRNP in the pre-A complex. a, Two different views of the spatial organization of the yeast U1 snRNP, with the density shown on the left and the molecular model on the right. **b**, The U1–5′ss helix is stabilized in the pre-A complex by Luc-7 and Yhc1, in the same manner as in the yeast E and A complexes. Top, fit of the U1–5′ss helix plus Luc7 and Yhc1 to the pre-A EM density. Bottom, the U1–5′ss helix and adjacent proteins. **c**, Fit of the Prp40 FF1–6 domains in the pre-A EM density. Top, domain organization of the *S. cerevisiae* Prp40 protein; below, amino-acid boundaries of each domain. WW, domain containing two conserved tryptophans that are spaced 20–22 amino acids apart; FF, domain containing two conserved phenylalanines at its N and C termini. **d**, Protein crosslinks between Prp40, Prp5[NTR] and other pre-A-complex proteins. Numbers (colour coded to match protein colours) indicate the positions of crosslinked lysine residues, which are connected by black lines. Prp40, Snu71 and Luc7 form a stable trimer[46] that in the cryo-EM structure of the yeast E complex[18] bridges the U1 snRNP to the branch site, and we show here that they also help to bridge U1 snRNP to U2 during the early stages of prespliceosome formation. In humans and in *Schizosaccharomyces pombe*, Prp5 facilitates formation of the A complex by bridging the U1 and U2 snRNPs[4], with the Prp5 N-terminal RS domain interacting with proteins of the SF3b complex[21]. Although *S. cerevisiae* Prp5 lacks an N-terminal RS domain, CXMS data indicate that its N terminus also interacts with Snu71 and Rse1[BPB]. Therefore, the bridge formed by Rse1 and the Prp40–Luc7–Snu71 trimer in the *S. cerevisiae* pre-A complex probably serves as an anchoring point for Prp5's N terminus.
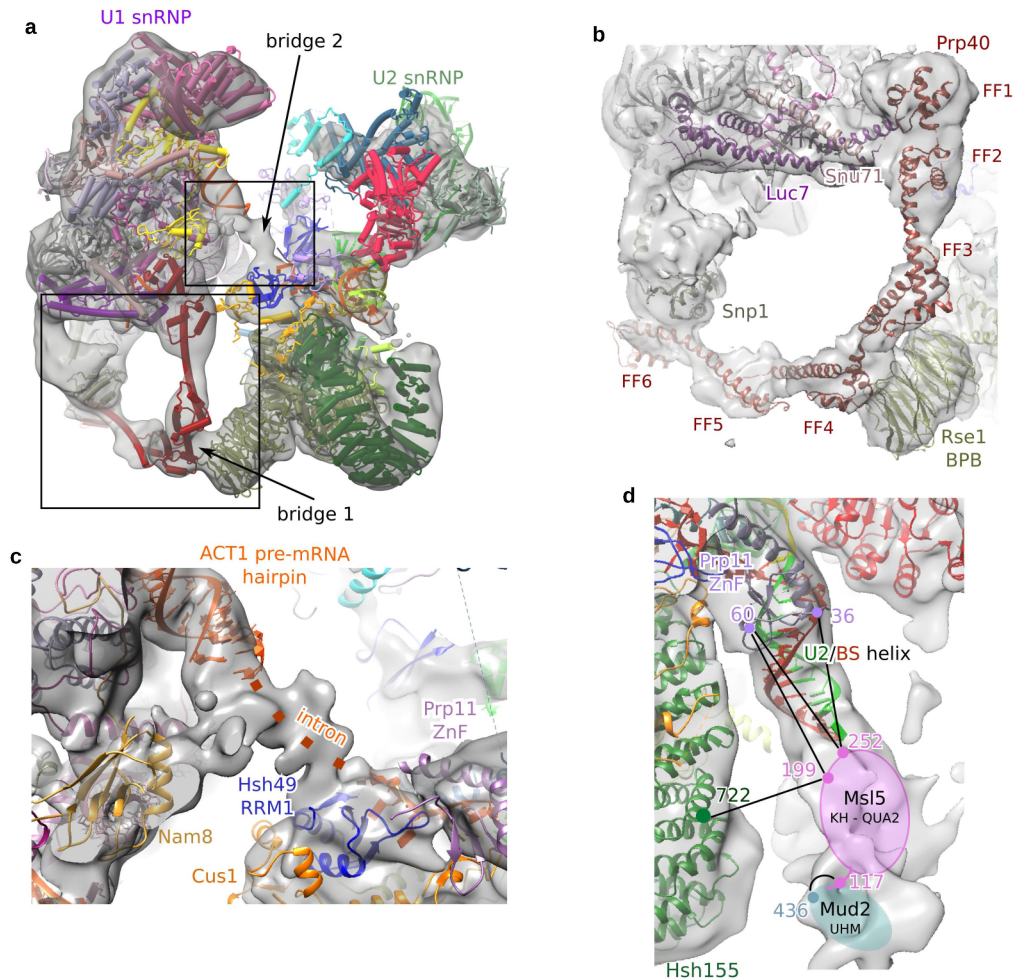
Extended Data Fig. 6 | Molecular bridges connecting the U1 and U2 snRNPs in the pre-A complex. a, Fit of the molecular model of the entire pre-A complex into the EM density (low-pass filtered). The two main bridges that connect the U1 and U2 snRNPs are indicated by arrows. The boxes indicate the regions expanded in **b, c. b**, Close-up of bridge 1 formed mainly by the interaction of Prp40 with Rse1. Bridge 1 is disrupted during the transition from the pre-A to the A complex (see Fig. 3 and Supplementary Video 1). Deletion analyses of Prp40 showed that although FF domains 3–6 are dispensable for yeast viability, they convey a considerable growth disadvantage when absent[46]. There is also evolutionary conservation of the presence of four or more of the FF domains in Prp40 from various organisms[46], suggesting that FF3 and FF4 have important roles during spliceosome assembly and/or splicing. Our crosslinking data indicate that Snu71 also extensively contacts Rse1[BPB] and FF2 of Prp40 (see Extended Data Fig. 5d). It is conceivable that, in the absence of Prp40 FF3–FF4, Snu71 still interacts with Rse1[BPB], the latter being a protein–protein interaction domain that interacts with different proteins in the subsequently formed B and B[act] spliceosomal complexes. **c**, Close up of bridge 2 that is formed by intron nucleotides between the U2–BS helix and the 5′-ss. By analogy to the situation in later spliceosomal complexes[12], these intron nucleotides are likely to be chaperoned by Hsh49[RRM1] and Prp11[ZnF]. Intron nucleotides of the *Act* pre-mRNA (but not of other pre-mRNAs such as *Ubc4*) form a hairpin that can be localized adjacent to the U1 snRNP already in the E complex[18]. The resolution is not sufficient to determine the exact intron nucleotides comprising this stem. Note that the intron hairpin is not part of bridge 2. **d**, A lower threshold reveals EM density below the U2–BS helix, adjacent to the open Hsh155[HEAT] domain, that probably corresponds to the Mud2–Msl5 dimer. The EM map is low-pass filtered to 30 Å resolution. Protein crosslinks supporting the localization of Msl5–Mud2 adjacent to the U2–BS helix are shown. Numbers (colour coded to match protein colours) indicate the positions of crosslinked lysine residues, which are connected by black lines. Msl5–Mud2 could not be precisely modelled into the EM density, presumably because of their structural flexibility. However, on the basis of CXMS data, we tentatively position Msl5–Mud2 into weak density directly downstream of the branch site, close to the U2–BS helix, with Mud2 being bound to the 3′-end of the intron. Formation of the U2–BS helix requires that Msl5 hands the branch site over to the U2 snRNA, and thus Msl5 should already be displaced from the branch site in the pre-A complex. Therefore, retention of Msl5–Mud2 close to the U2–BS helix would be consistent with the binding of Mud2 to the intron downstream of the branch site.

a

Data 1 (1.16Å²/pixel)
460,854 particles from 9170 micrographs

Data 2 (1.06Å²/pixel)
869,306 particles from 18,332 micrographs

coarsed to 4.64 Å²/pixel,
reference-free 2D classification

coarsed to 4.64 Å²/pixel,
reference-free 2D classification

218,404 particles

480,488 particles

reference

U1 region filtered to 35Å

3D classification into 4 classes

23.5%   22.1%   20.0%   34.4%

U1 mask

Re-extract, coarsed to 2.32 Å²/pixel,
(240,145 particles)
3D classification with
mask around U1

21.2%   23.4%   23.7%   31.7%

3D Refinement

pre-A complex
10.0 Å resolution

Multibody refinement

3D classification with
mask around U2
(80,853 particles)

U1 region
7.5 Å resolution

22.2%   25.4%   27.7%   24.7%

3D Refinement, multibody refinement
(39,593 particles)

U2 region
13 Å resolution

b

50 nm

c

510 Å

U2 5'   U2 3'

U1

d

Map of U1
Map of U2
Map of pre-A

13 Å   10.4 Å   7.5 Å

FSC / Resolution (1/Å)

e

ΔBS-A pre-A
U257A pre-A

f

U1 snRNP
Prp39
U2 snRNP
Prp9
Prp5
Prp40
Rse1
Hsh155
U257A pre-A

g

U2/BS helix
SLIIa
Cus1
Hsh155
U257A pre-A

h

Prp9
NTR
Prp5
RecA1
Prp5
RecA2
U2/BS helix
U257A pre-A

i

FF3
Bridge 1
FF4
Prp40
Rse1
BPB
U257A pre-A

Extended Data Fig. 7 | See next page for caption.

Article

Extended Data Fig. 8 | Movement of U1 and U2 snRNPs during the transition from the pre-A to the A complex. a, Close-up of the rotation of the U2 3′-region after the release of Prp5. The 3′-region rotates around the indicated axis by roughly 55°. To better show the movement of the 3′-region, the SmD2 protein is in yellow. For simplicity, only the 3′-region of U2 plus U2 SLII and the U2–BS helix are shown in the pre-A complex and the yeast A complex (PDB 6G90). The pre-A and A complexes are aligned via U2 SLIIa and HR 19–20 of Hsh155. **b**, Close-up of the movement of the U1 snRNP and 3′-region of U2. A top view is shown, with the black dot indicating the pivot point of the U2 3′-region, which rotates by roughly 55° in the plane of the paper. For simplicity, only the region of U1 snRNP that contains Prp39 is shown. The U1 snRNP rotates around the

indicated axis by roughly 45°. In the pre-A complex, Prp39 and Lea1 are separated by roughly 130 Å, but the movements of U1 and U2 bring them into close proximity in the A complex. Even though Lea1 is not essential in *S. cerevisiae*, its depletion prevents formation of the A complex, and adding back Lea1 restores A-complex assembly[47]. The Prp39–Lea1 interaction is a structural marker for the formation of a mature A complex, and as such its absence in the pre-A complex is a clear indication that our complex has stalled at an earlier assembly stage. This interaction is also maintained in the pre-B complex[13] and is therefore also a structural marker for the conformation that allows joining of the tri-snRNP.

Article

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	ΔBS-A pre-A complex			U257A pre-A complex		
	U1 snRNP region (EMD-13029) (PDB 7OQC)	U2 snRNP region (EMD-13028) (PDB 7OQB)	Pre-A complex (EMD-13033) (PDB 7OQE)	U1 snRNP region (EMD-13031)	U2 snRNP region (EMD-13032)	Pre-A complex (EMD-13030)
Data collection and processing						
Magnification	120,700	120,700	120,700	Data 1: 120,700 Data 2: 132,000	Data 1: 120,700 Data 2: 132,000	Data 1: 120,700 Data 2: 132,000
Voltage (kV)	300	300	300	300	300	300
Electron exposure (e⁻/Å²)	44	44	44	Data 1: 44 Data 2: 58	Data 1: 44 Data 2: 58	Data 1: 44 Data 2: 58
Defocus range (μm)	1-4	1-4	1-4	1-4	1-4	1-4
Pixel size (Å)	1.16	1.16	1.16	Data 1: 1.16 Data 2: 1.06	Data 1: 1.16 Data 2: 1.06	Data 1: 1.16 Data 2: 1.06
Symmetry imposed	C1	C1	C1	C1	C1	C1
Initial particle images (no.)	5.4 million	5.4 million	5.4 million	1.3 million	1.3 million	1.3 million
Final particle images (no.)	217,460	160,894	217,460	80,853	39,593	80,853
Map resolution (Å)	4.1	9.0	5.9	7.5	13	10.4
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143
Map resolution range (Å)						
U1 region	3.7-6.0	---	4-9	6-9	---	8-10
U2 5' region	---	7-12	10-15	---	10-15	20-30
U2 3' region	---	10-12	20	---	10-15	20-30
Prp40	---	---	15	---	---	20-30
Refinement						
Initial model used (PDB code)	n/a	n/a	n/a	n/a	n/a	n/a
Model resolution (Å)	4.4	---	---	---	---	---
FSC threshold	0.5					
Model resolution range (Å)	4.4	---	---	---	---	---
Map sharpening B factor (Å²)	-240	-600	-300	-300	-600	-600
Model composition				---	---	---
Non-hydrogen atoms	30386	26445	64053			
Protein residues	3078	4555	8039			
Ligands	0	0	0			
B factors (Å²)				---	---	---
Protein	62.8	---	---			
Ligand	---	---	---			
R.m.s. deviations				---	---	---
Bond lengths (Å)	0.004	---	---			
Bond angles (°)	1.0	---	---			
Validation				---	---	---
MolProbity score	1.6	---	---			
Clashscore	5.4	---	---			
Poor rotamers (%)	0.44	---	---			
Ramachandran plot				---	---	---
Favored (%)	95.42	---	---			
Allowed (%)	4.14	---	---			
Disallowed (%)	0.44	---	---			

Extended Data Table 2 | Summary of modelled proteins and RNA in the yeast pre-A structure

Sub-complexes	Protein/RNA	Chain ID	UniProt ID	Total residues	Modeled Residue	Template	modeling approach
U1 snRNP	Mud1	A	P32605	298	2-46; 55-125; 133-148	6N7R	Docked and adjusted
	Snp1	B	Q00916	300	1-91; 94-188	6N7R	Docked and adjusted
	Yhc1	C	Q05900	231	3-197;	6N7R	Docked and adjusted
	Prp39	D	P39682	629	288-553; 561-627	6N7R	Docked and adjusted
	Prp42	E	Q03776	544	1-544	6N7R	Docked and adjusted
	Nam8	F	Q00539	523	161-242; 291-425; 432-449; 492-523	6N7R	Docked and adjusted
	Snu56	G	Q03782	492	43-170; 185-295	6N7R	Docked and adjusted
	Luc7	H	Q07508	261	4-19; 38-140; 172-244	6N7R	Docked and adjusted
	Snu71	J	P53207	620	1-52; 260-311	6N7R	Docked and adjusted
	Sm B	b	P40018	196	2-63; 73-131	6N7R	Docked and adjusted
	Sm D3	d	P43321	101	3-95	6N7R	Docked and adjusted
	Sm D1	h	Q02260	146	1-49; 58-73; 78-119	6N7R	Docked and adjusted
	Sm D2	i	Q06217	110	8-80; 83-108	6N7R	Docked and adjusted
	Sm E	e	Q12330	94	8-63; 73-93	6N7R	Docked and adjusted
	Sm F	f	P54999	86	12-84	6N7R	Docked and adjusted
	Sm G	g	P40204	77	2-48; 53-77	6N7R	Docked and adjusted
	Prp40	K	P33203	583	134-189	2B7E	docked
					203-552	2KFD; 3HFH	predicted model, docked
	U1 snRNA	1		568	1-26; 34-565	6N7R	Docked and adjusted
U2 snRNP	Msl1	Y	P40567	111	28-111	6G90	Docked
	Lea1	W	Q08963	238	1-170	6G90	Docked
	Hsh155	O	P49955	971	161-758	6Y5Q	Docked
					759-971	6G90	Docked
	Rse1	P	Q04693	1361	53-305; 323-571; 581-784; 814-890; 918-1265; 1292-1361	6G90	Docked
	Cus1	Q	Q02554	436	125-213; 239-353; 361-376	6G90	Docked
	Hsh49	R	Q99181	213	9-86; 106-144; 147-185; 189-203	6G90	Docked
	Rds3	S	Q06835	107	2-93	6G90	Docked
	Ysf3	Z	P0C074	85	2-84	6G90	Docked
	Sm B	s	P40018	196	12-54;76-102	6G90	Docked
	Sm D3	v	P43321	101	4-85	6G90	Docked
	Sm D1	t	Q02260	146	1-48;78-101	6G90	Docked
	Sm D2	u	Q06217	110	17-108	6G90	Docked
	Sm E	w	Q12330	94	10-63; 71-93	6G90	Docked
	Sm F	x	P54999	86	12-84	6G90	Docked
	Sm G	y	P40204	77	2-76	6G90	Docked
	Prp5	p	P21372	849	206-698	4LJY	Docked
	Prp9	T	P19736	530	1-97; 112-378; 407-478; 503-528	6G90	Docked and adjusted
	Prp11	U	Q07350	266	34-47; 51-104; 121-136; 149-253	6G90	docked
	Prp21	V	P32524	280	89-206; 220-228	6G90	Docked and adjusted
	U2 snRNA	2		1175	32-46	5NRL	Docked
					47-49; 74; 78		de novo
					50-73; 79-86; 108-122; 139-150;1089-1109; 1115-1130; 1138-1154; 1159-1169	6G90	Docked
	ACT1 pre-mRNA	I		691	-2-8	6N7R	Docked and adjusted
					intron hairpin, unknown sequence	6N7R	Docked
					246-253	5ZWM	Docked
					254-268	5NRL	BS-A deleted, Docked

RNA and protein regions were modelled and fit into the EM density of the ΔBS-A pre-A complex, as indicated.

4 Discussion and Perspectives

4.1 The human 17S U2 snRNP and Prp5-mediated remodelling

The 17S U2 snRNP plays an essential role during the branch site selection and the A complex formation. Stable addition of U2 during the A complex formation is the first ATP-dependent step that requires the DEAD-box ATPase Prp5[578]. Recent advances in cryo-EM have allowed visualization of the fully assembled spliceosome intermediates, namely pre-B, B, B[act], C, C*, and P complexes[31,91,96–98,100,101,114,115,118,121,143,166–171]. In addition, the yeast E and A complexes have also been recently reported[142,172]. Interestingly, these cryo-EM structures have revealed that after Prp5 action (but before Prp2 action), the U2 snRNP does not undergo significant rearrangements from A to B[act] complexes. Therefore, the U2 region in the reported spliceosome A to B[act] structures likely represents the conformation of U2 after Prp5-mediated RNP rearrangements. However, the structure of the U2 snRNP before Prp5 action was unknown for both humans and yeast prior to this study. Here, we determined the first structure of the human 17S U2 snRNP by single-particle cryo-EM, and resolved the more stable U2 5' region to ca. 4 Å resolution on average. However, the resolutions of flexible peripheral parts and the U2 3' domain of the 17S U2 snRNP were limited at ca. 20 Å. To interpret the low-resolution parts of the 17S U2 snRNP, we turned to an integrated structural biology approach. That is, docking known X-ray structures or predicted structures into the EM map on the basis of protein-protein crosslinking mass spectrometry (CXMS). In this way, we built a molecular model of the human 17S U2 snRNP and have revealed: (a) the overall molecular architecture of the human 17S U2 snRNP and insights into U2 snRNP assembly; (b) the first structural evidence of the formation of the BSL; (c) the RNP rearrangements facilitated by Prp5 that are required for stable association of the U2 snRNP with the branch site.

4.1.1 Structural insights into U2 snRNP assembly

During the splicing cycle, the U2 snRNP undergoes a large compositional change and loses most of its components (i.e., Prp5 and TAT-SF1 during the A complex formation; SF3a and SF3b during the B* complex formation), and only the 12S U2
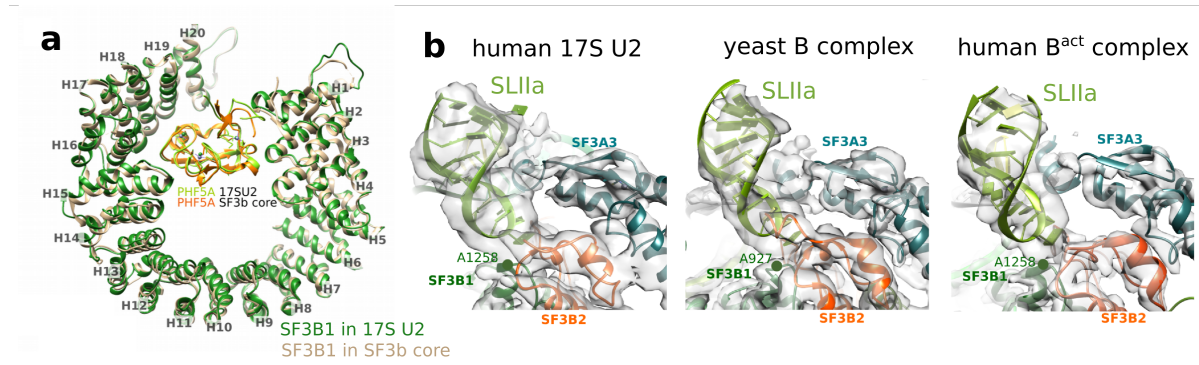
Figure 4.1: **SF3B1 maintains the open conformation in 17S U2 snRNP, and SLIIa is the major anchor to SF3a and SF3b complexes. a.** The SF3B1^{HEAT} in the 17S U2 (green) maintains the same open conformation as in the crystal structure (beige) **b.** The comparison of SLIIa and the surrounding proteins shows that a highly similar SLIIa RNP architecture is established in the 17S U2 and is maintained until the B^{act} complex.

core containing the U2 snRNA, U2 Sm, U2-A', and U2-B" is retained after spliceosome disassembly. The U2 snRNP is recycled by being reassembled into the mature 17S U2 snRNP that is ready for another round of splicing. This assembly step also applies to newly synthesized U2[133]. Since our 17S U2 snRNP structure represents the fully assembled U2 snRNP, it provides structural insights into the U2 snRNP assembly process.

First, the SF3b core does not undergo major structural remodelling during U2 assembly. Previously, the crystal structure of the isolated SF3b core complex, containing SF3B1, SF3B3, PHF5A, and SF3B5, was resolved by X-ray crystallography and cryo-EM[12,13]. The two structures reported by different methods are almost identical, ruling out the possibility of artificial conformations stabilized by crystal packing. In particular, both structures show that in the isolated SF3b core, the SF3B1^{HEAT} adopts an open conformation, being distinct from the closed conformation that is observed in fully assembled spliceosome structures[31,121]. The closed SF3B1^{HEAT} in the spliceosome context is essential for specifying and shielding the BS-A, as well as stabilizing the U2-BS helix. However, it was previously unclear whether this functionally important conformational change happens during the 17S U2 snRNP assembly or during the A complex formation. Here, by comparing our 17S U2 structure with the SF3b core structure, we show that SF3B1^{HEAT} is in the same open conformation, and therefore the closure of SF3B1^{HEAT} most likely takes place during/after the formation of the U2-BS helix (see also 4.2.1). In addition, SF3B3, SF3B5, and PHF5A interact with

SF3B1$^{\text{HEAT}}$ in the same manner, and the protein pocket formed by SF3B1 and PHF5A is maintained in the 17S U2 snRNP as in the SF3b crystal structure[12]. The highly similar structural organization of the SF3b core in the 17S U2 snRNP context indicates that like the isolated SF3b core, splicing inhibitors such as Pladienolide B or E7107 can interact with the binding pocket formed by residues of HEAT repeats H15-17 and of PHF5A, also in the 17S U2 snRNP[13,173].

Second, the formation of SLIIa is required for U2 snRNP assembly. Previous spliceosome structures of A to B$^{\text{act}}$ complexes show that in both human and yeast, SLIIa interacts extensively with the SF3a and SF3b complexes[31,96,98,167 142,146]. Our structure shows that the interactions between SLIIa and SF3a/SF3b are already established in the 17S U2 snRNP. Therefore SLIIa is the major anchor of the U2 snRNA to SF3a and SF3b during U2 snRNP assembly, which is maintained until the B$^{\text{act}}$ complex. Upon Prp2 mediated dissociation of SF3a and SF3b, the U2 snRNA undergoes the SLIIa to SLIIc transition concomitantly, and the SLIIc conformation is maintained until the ILS complex[120,121,143]. Therefore, to recycle the U2 snRNP for a new round of splicing cycle, the U2 snRNA has to switch from SLIIc back to SLIIa conformation for SF3a/SF3b binding, which is consistent with that the SLIIa conformation facilitates branch site recognition, while SLIIc inhibits this step[159,174]. Although the structural interconversion of SLIIa and SLIIc can be spontaneous and does not require energy *in vitro*, Cus2 (yeast TAT-SF1 homologue) is able to drive the equilibrium toward SLIIa through directly binding to the U2 snRNA[141]. *In vivo*, certain mutations at Cus2 have been shown to rescue the defects resulting from the destabilization of SLIIa by U2 G53A mutation, possibly through stabilizing SLIIa[141,158]. Therefore, Cus2/TAT-SF1 may play a role in facilitating the U2 snRNP assembly by promoting the SLIIa conformation. Nonetheless, since our 17S U2 snRNP structure does not show direct interactions between TAT-SF1 and SLIIa, it remains unclear how Cus2/TAT-SF1 achieves this role mechanistically.

4.1.2 SF3B1$^{\text{HEAT}}$ is bound and stabilized by TAT-SF1 and Prp5

Previous studies in yeast have shown that the N-terminal region of Prp5 binds to H1-6 and H9-12 of Hsh155 (the yeast homologue of SF3B1)[155]. Consistently, our

17S U2 snRNP structure shows that H9-12 of SF3B1[HEAT] are bound to an extended α-helix of Prp5 (aa 146-196). In addition, CXMS data suggest that residues of both the N-terminal and C-terminal of the extended α-helix of Prp5 interact with H12-15 and H1-6, respectively (Figure 4.2a,b). Therefore, the N-terminal region of Prp5 (aa 113-316) seems to encompass SF3B1[HEAT]. Interestingly, based on the CXMS data, the conserved Prp5 DPLD motif (Figure 4.2b), which is essential for stable Prp5 binding to the U2 snRNP, localizes close to the SF3B1 hotspot mutations region. It is consistent with studies in yeast that show Hsh155 mutations at analogous regions directly destabilize Prp5 binding[155]. As Prp5 plays a role in proofreading the branch site, at least in yeast (see 4.2.5), destabilized Prp5 binding may result in aberrant branch-site usage, explaining why Hsh155 mutations lead to an increased usage of a mutated branch site.

Another U2 component, TAT-SF1 (Cus2 in yeast), has been shown to directly interact with SF3B1/Hsh155 via binding to the SF3B1/Hsh155 N-terminal ULM (U2AF ligand motif) by its UHM (U2AF homology motif)[161,162]. Based on CXMS data, we located TAT-SF1[UHM] close to the N-terminal HEAT repeats of SF3B1, consistent with the fact that it binds to SF3B1[ULM] and has the highest affinity to ULM5, which is the closest ULM to SF3B1[HEAT][161]. Surprisingly, our EM map, together with CXMS data, additionally reveals a previously unknown interaction that the RNA recognition motif (RRM) of TAT-SF1 binds directly to H15-16 of SF3B1[HEAT] (Figure 4.2a). During the structural transition of SF3B1[HEAT] from an open to a closed conformation, local twist and curvature changes are clustered at H1-3, H7-H12, H15-H16[152]. As a result, by binding to one of the hinge regions at H15-16, TAT-SF1[RRM] may inhibit the conformational change of SF3B1[HEAT] and stabilizes its open conformation. Similarly, the binding of the N-terminal region of Prp5, including the extended α-helix, to H1-15 of SF3B1[HEAT] may also stabilize the open conformation of the latter.

4.1.3 The BSL is formed in humans and sequestered by U2 proteins

Previous mutagenesis studies in yeast suggest that BSiR is initially sequestered in the branchpoint-interacting stem-loop (BSL), and that the nucleotides at the BSL loop
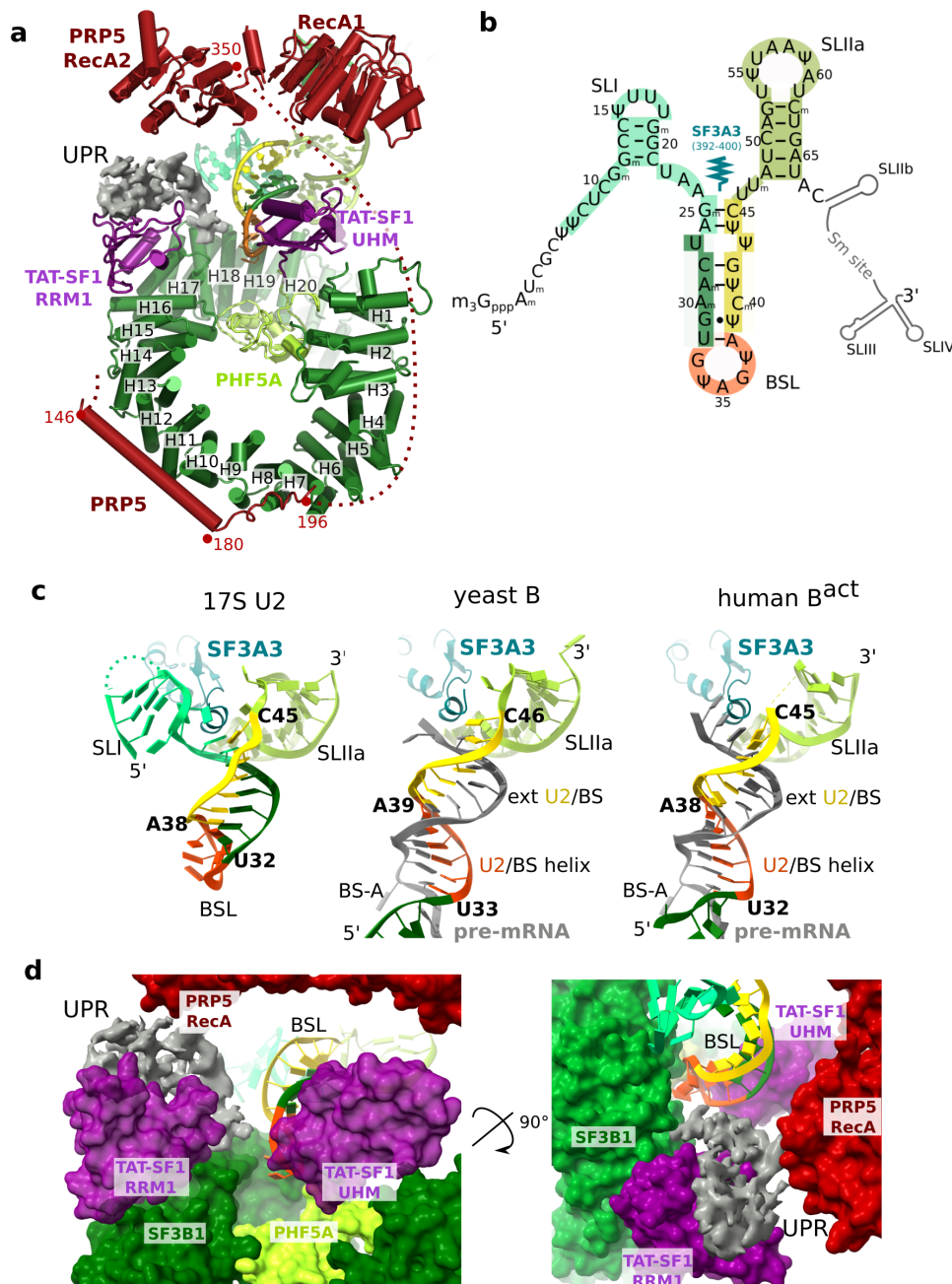
Figure 4.2: **Prp5 and TAT-SF1 bind SF3B1 and sequester the BSL in the 17S U2 snRNP. a.** Location of the Prp5 RecA domains. The interaction of Prp5 and TAT-SF1 with SF3B1. The unassigned density is shown in grey. The dotted line shows the potential path of the unmodelled Prp5 N-terminal region based on CXMS data. **b.** The position of cancer-related hotspot mutations of SF3B1 is shown. CXMS data suggest the N-terminal region of Prp5 wraps around SF3B1$^{\mathrm{HEAT}}$, and the conserved DPLD motif is located close to the hotspot mutations. **c.** Left: 2D and 3D representations of U2 SLI, BSL, and SLIIa in the human 17S U2 snRNP. The BSL is eight base-pairs in length due to the presence of the short alpha-helix of SF3A3 (amino acids Y392 to H400). The BSL nucleotides that later form U2-BS helix are in orange, the extended part in yellow, the remaining part in dark green. SLIIa is shown in light green and SLI in teal. Right: The BSL unwinds and base-pairs with the branch-site region, forming the U2-BS helix, as shown by the human B$^{\mathrm{act}}$ complex. **d.** Two views showing the BSL is sequestered by Prp5 (red), TAT-SF1 (purple), and SF3B1 (green), as well as the unassigned protein region (UPR, grey).

may form initial base-pairing with the branch site (see 2.3.2). However, the existence of BSL has not been proven structurally, and a similar BSL has never been identified in humans previously. Our 17S U2 snRNP structure, thus for the first time, provides a structural evidence for the existence of the BSL and its formation in humans (Figure 4.2c). Previous studies proposed a nine base-pair length of the BSL in yeast, based on the sequence of the U2 snRNA[4]. Our structure shows that the stem of the BSL is contacted by a short α-helix of SF3A3, forcing the two nucleotides U46 (U47 in yeast) and A24 (A25 in yeast) single-stranded, resulting in the BSL of eight base-pairs in length (Figure 4.2c). As the same α-helix of Prp9 (yeast homologue of SF3A3) is also observed in yeast spliceosome structures at the same position[146], it is likely that the yeast BSL may also be eight base-pairs. Indeed, previous mutagenesis studies have shown that neither U47G nor U47A affects yeast growth, while disruption of G26-C46 base-pairing leads to growth defects[4,175].

Furthermore, our 17S U2 snRNP structure reveals that the BSL is sandwiched by U2 components (Figure 4.2d). The C-terminal HEAT repeats of SF3B1 block the BSL on one side and TAT-SF1 and Prp5 RecA domains on the other side. In addition, an unassigned protein region directly contacts the BSL, likely stabilizing its conformation. This unassigned protein region (UPR) is probably part of TAT-SF1, as the density can be traced continuously to TAT-SF1[RRM]. The loop of the BSL is also sequestered by TAT-SF1 and SF3B1 and is thus inaccessible for branch site recognition. Therefore, substantial RNP remodelling of the U2 snRNP, including displacement of TAT-SF1 and Prp5, is required for the BSL unwinding and the U2-BS helix formation (see 4.1.4). Previous studies in yeast suggest that the displacement of TAT-SF1 (Cus2 in yeast) is mediated by the ATPase activity of Prp5[7,162]. In this way, the U2 snRNP ensures that the reactive BSL is initially sequestered, and the liberation of the BSL is triggered by ATP hydrolysis only when required. This may prevent premature interaction of the U2 snRNA with other RNAs in the wrong context. Interestingly, protecting reactive regions of RNAs by other spliceosomal RNAs or proteins is a common mechanism used by the spliceosome repeatedly. One well-characterized example is that in the U4/U6.U5 tri-snRNP and pre-catalytic B complex, the reactive region of the U6 snRNA that later forms the ISL is sequestered by extensive base-pairing with the U4 snRNA[176]. Another example is that the BS-A, the nucleophile of the first transesterification reaction, is

sequestered and shielded by SF3b complex until B* complex formation[31,96,97,121], which is the precise moment when the reactivity of the BS-A is required. The liberation of U6 snRNA and the BS-A are also coupled with ATPase activity of RNA helicases, namely Brr2 and Prp2, respectively. Thus, by coupling the energy consumption that requires a specific RNA helicase action with the liberation of reactive RNA regions, the spliceosome ensures a defined sequential order of activating reactive RNAs, so that accidental discharge of reactants is prevented.

4.1.4   RNP remodelling of U2 snRNP facilitated by Prp5

A comparison of our 17S U2 snRNP structure with the U2 part of the spliceosomes (A to B^{act} complexes) shows that major RNP remodelling must occur to free the BSL nucleotides that later form the U2-BS helix and to allow the stable U2 snRNP incorporation into the spliceosome (Figure 4.3a). As aforementioned in 4.1.3, in order to free the BSL, both TAT-SF1 and the Prp5 RecA domains need to be displaced. After the displacement of TAT-SF1 triggered by Prp5 ATPase activity[7 162], the displacement of Prp5 RecA domains may be spontaneous, as they lose the major interaction partner. The freed BSL can mediate initial base-pairing with the branch site via presenting its loop region, possibly forming a "kissing loop" intermediate[177]. Formation of the U2-BS helix would require a downward movement of the 5' end of the U2 snRNA, including SLI (Figure 4.3b). However, for topological reasons, such a movement of the U2 5' end is not possible because it has to go through a very narrow gap between the 5' strand of the BSL and the C-terminal HEAT repeats of SF3B1. Thus, more room has to be created between the 5' strand of the BSL and the C-terminal HEAT repeats to allow U2 snRNA remodelling. This may be achieved by a rotational movement of the BSL around single-stranded U2 snRNA nucleotides (U46 or U47) that link the BSL to SLIIa. After the disruption of the BSL, a rotational movement of the U2 5' end around the pre-mRNA intron would then facilitate the formation of the U2-BS helix (Figure 4.3c). The newly formed U2-BS helix swings back and is subsequently embraced by the closure of the SF3B1^{HEAT}. Interestingly, the short α-helix of SF3A3 is situated in the same position in the 17S U2 snRNP as in the spliceosomes. Therefore, it may have similar roles, which is to determine the length of the BSL or the U2-BS helix. Moreover, by stably interacting with the U2 snRNA, this short α-helix of SF3A3 may

facilitate the outward movement of the BSL and the backward swinging of the newly formed U2-BS helix. Finally, the correct formation of the U2-BS helix leads to the closure of the SF3B1$^{\text{HEAT}}$ and the concomitant release of Prp5. The trigger of SF3B1 closure and Prp5 release is the correct insertion of the BS-A into the protein pocket formed by SF3B1 and PHF5A, which will be discussed in more detail in 4.2.1 and 4.2.5.

The exact mechanism of how Prp5 mediates this U2 remodelling cannot be concluded solely from our structure. As a DEAD-box RNA helicase, Prp5 is expected to bind a double-stranded RNA via its RecA2 domain. Upon ATP binding, the RecA domains switch from the extended open conformation to the compact closed conformation, introducing bends in one of the RNA strands that facilitate unwinding of the RNA double strand[108]. Although the nature of the RNA strand bound by Prp5 is not clear, Prp5 ATPase activity has high specificity to the U2 snRNA[5]. It is therefore conceivable that Prp5 activity may lead to displacement of TAT-SF1 by direct binding and inducing the BSL unwinding. Indeed, previous studies in yeast have shown that by disrupting the BSL, the essentiality of Prp5 ATPase activity can be bypassed[4]. On the other hand, Prp5 may also indirectly lead to the displacement of TAT-SF1 without the requirements for a direct interaction with the BSL, as several DEAD-box proteins have been shown to disrupt protein-RNA interactions, and in some cases even without unwinding double-stranded RNA[178]. In the latter scenario, TAT-SF1 displacement may result in a destabilization of the BSL, which may allow the subsequent U2-BS helix to form spontaneously without the requirements for chemical energy in the form of ATP hydrolysis. This is supported by the fact that Prp5 ATPase activity can be bypassed in the absence of Cus2, and hyperstabilized BSL can lead to growth defects in yeast[4].

It is worth mentioning that the interaction between TAT-SF1$^{\text{RRM}}$ and SF3B1$^{\text{HEAT}}$ is governed mainly by the α-helices of the RRM, leaving the β-sheets accessible for RNA binding. Therefore, in principle TAT-SF1$^{\text{RRM}}$ is able to bind RNA while remaining bound to SF3B1$^{\text{HEAT}}$. Mediating both protein-protein and protein-RNA interactions simultaneously is not uncommon for RRMs. For instance, RRM1 of SF3B4 binds to the single-stranded intron upstream of the U2-BS helix while interacting with SF3B2.
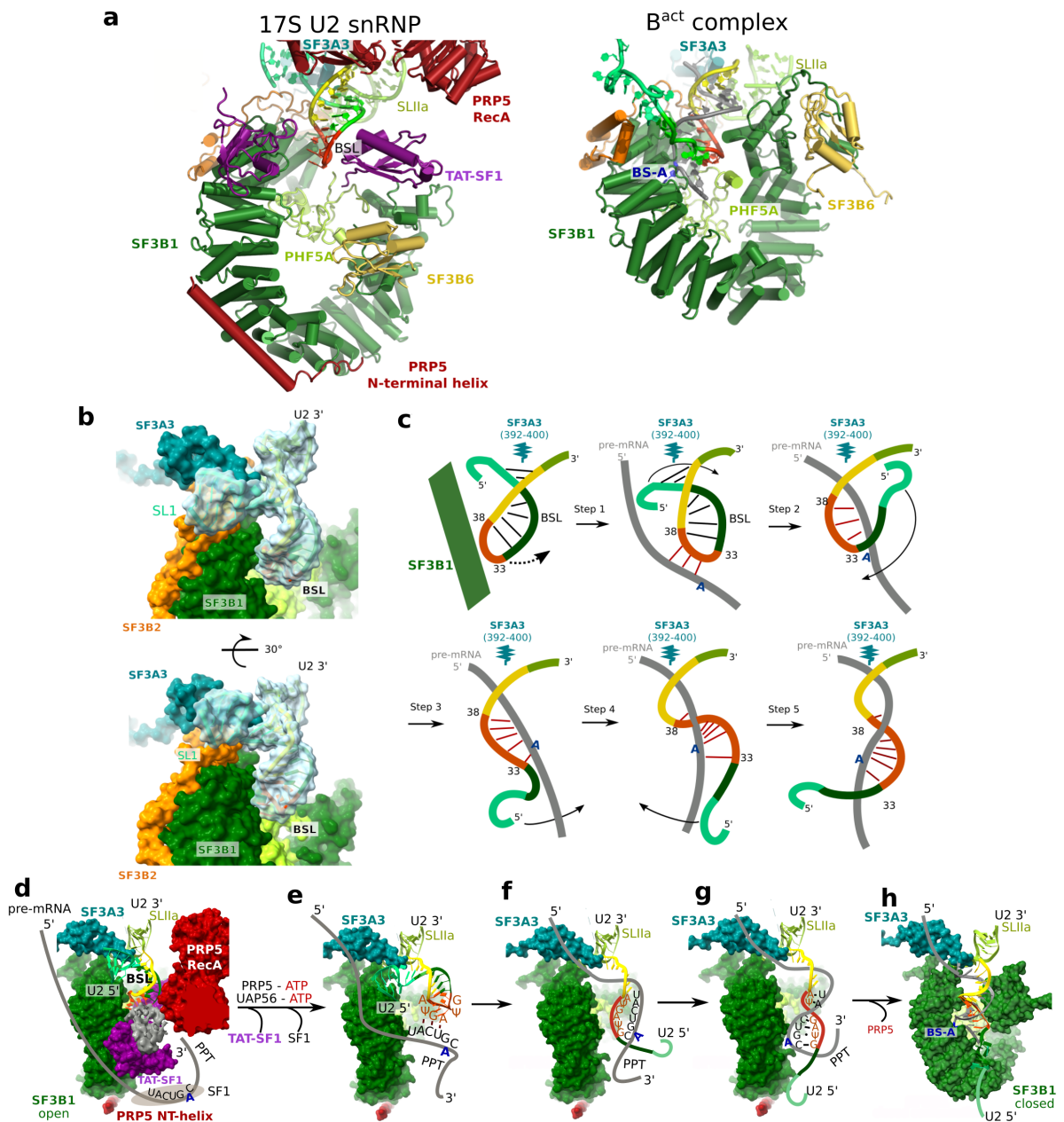
Figure 4.3: **RNP rearrangements facilitated by Prp5 and the movement of the U2 snRNA required for the U2-BS helix formation. a.** A structural comparison of the BSL and the neighboring proteins in the 17S U2 snRNP (left) with the corresponding region after the U2-BS helix formation and stable U2 incorporation into the spliceosome (right). **b.** The formation of the U2-BS helix requires a downward movement of the SLI. Surface model shows that due to topological reasons, the movement of the SLI is not possible because SLI would have to thread through the very narrow gap between the BSL and the C-terminal SF3B1$^{\text{HEAT}}$. As a result, the BSL has to move away from SF3B1, possibly via an outward swinging. **c.** A proposed model of the U2 5' end movement that facilitates the formation of the U2-BS helix. **d-h:** A proposed model of the U2 snRNP rearrangements during A complex formation. **d-e,** The U2 snRNP is recruited to the E complex, and Prp5 ATPase activity displaces TAT-SF1 and Prp5 RecA domains, liberating the BSL for initial base-pairing with the branch-site region. **e-g,** The BSL is disrupted, and the movement of the U2 5' end facilitates the BSL unwinding and the U2-BS helix formation. **g-h,** The closure of SF3B1 clamps the newly formed U2-BS helix, with the BS-A inserted into the protein pocket formed by SF3B1 and PHF5A, leading to the final displacement of Prp5 and the A complex formation.

Similarly, RRM1 of U2-B" binds to U2 snRNA and U2-A' at the same time. In both cases, the protein-protein interactions are mediated by the α-helices, while the β-sheets bind to RNAs. Indeed, *in vitro* studies have shown that Cus2$^{\text{RRM}}$ is able to bind the U2 snRNA directly, and the conserved Y48 is essential for its RNA binding *in vitro* and its function *in vivo*[158]. Therefore, the unassigned density on top of the TAT-SF1$^{\text{RRM}}$ may contain the 5' end of the U2 snRNA (Figure 4.2a,d). By binding directly to U2 snRNA 5' end, the displacement of TAT-SF1 may allow the U2 5' end movement. This would provide another possible scenario of how the displacement of TAT-SF1 may facilitate BSL unwinding.

Taken together, our structure together with decades of biochemical and genetics studies suggests a model for the U2 snRNP remodelling during early spliceosome assembly as follows (Figure 4.3d-h): First, the U2 snRNP is initially recruited to the E complex by protein-protein interactions, bringing U2 to the vicinity of the branch site. Subsequently, Prp5 is activated by binding to ATP and a yet to be defined RNA component, and Prp5 facilitates the displacement of TAT-SF1 via its ATPase activity, thereby freeing BSL nucleotides for base pairing with the branch site. The correct formation of the U2-BS helix finally triggers the closure of SF3B1$^{\text{HEAT}}$ and the release of Prp5 (see 4.2 for more details), leading to stable incorporation of the U2 snRNP.

4.2 Structural insights into early spliceosome assembly and the branch site proofreading

In chapter 4.1, we have revealed the U2 snRNP remodelling that takes place during A complex formation. Mature A complex formation requires the closure of SF3B1/Hsh155$^{\text{HEAT}}$ that embraces the U2-BS helix and the docking of BS-A into the protein pocket formed by SF3B1 (human)/Hsh155 (yeast) and PHF5A (human)/Rds3 (yeast). However, solely from the human 17S U2 snRNP structure, we were unable to discern the trigger to this functionally important conformational change of SF3B1/Hsh155. Moreover, Prp5 is proposed to contribute the branch site proofreading during the A complex formation, at least in yeast. However, it was unclear how Prp5 performs this role, as the published yeast A complex does not contain Prp5 density.

Here, by deleting the BS-A (ΔBS-A) or introducing a U to A mutation at position 257 (U257A), which is two nucleotides upstream of the BS-A, of the actin 1 (ACT1) pre-mRNA, we stall the yeast spliceosome assembly in an intermediate stage, in which Prp5 has performed its ATPase activity but remains stably bound, and resolve their structures by cryo-EM. We show that the ΔBS-A complex and the U257A complex have not only the same protein and RNA compositions but also an almost identical structure at the achieved resolution. Since the U257A complex can progress to later spliceosomal assembly stages[10], we conclude that both structures represent an on-pathway assembly intermediate. Due to the limited resolution at the U2 region (about 8 Å for the ΔBS-A complex and 15Å for the U257A complex), we again turned to the integrative structural biology approach. Based on previously published structures and our CXMS data, we reveal the molecular architecture of this novel spliceosomal intermediate and term this the pre-A complex, as it represents a stage directly before the A complex[14] (also discussed in detail below). By comparing the structures of the 17S U2 snRNP, the pre-A, and the A complex, we have revealed: a) the trigger to HEAT domain closure; b) structural insights into early spliceosome assembly; c) the branch site proofreading mechanism mediated by Prp5.

4.2.1   The BS-A binding triggers SF3B1/Hsh155$^{\text{HEAT}}$ closure

Our Δ BS-A and U257A complexes show an extended helical density located directly upstream of SLIIa. This helical density is significantly longer than the BSL, but can be fit nicely with a modelled U2-BS helix. This indicates that Prp5 has performed its ATPase activity and the U2-BS helix has formed in our complexes. Consistent with the formation of the U2-BS helix, Cus2 is not detected in the mass spectrometry data. Interestingly, despite the formation of the U2-BS helix and the displacement of Cus2, Hsh155$^{\text{HEAT}}$ remains in an open conformation, similar to SF3B1$^{\text{HEAT}}$ in the human 17S U2 snRNP, but in stark contrast to its conformation found in the yeast A to B$^{\text{act}}$ complexes [11 167 96 97 146]. The open conformation of Hsh155$^{\text{HEAT}}$ and the retention of Prp5 (see 4.2.3) indicate that Δ BS-A and U257A complexes are stalled at a pre-A complex stage, which is after Prp5 mediated formation of the U2-BS helix, but during its proofreading role (see 4.2.5). In addition, unlike later spliceosomal complexes, in which U2-BS helix is sequestered by Hsh155$^{\text{HEAT}}$, the U2-BS helix in the pre-A complexes is located further away from the C-terminal Hsh155$^{\text{HEAT}}$. This is consistent with our proposed model that the formation and docking of the U2-BS helix involve an outward movement from the HEAT repeats and a backward swinging (see 4.1.4). Furthermore, as deletion of the BS-A does not affect the stability of the U2-BS helix *per se*, our data indicate that formation of the U2-BS helix alone does not necessarily trigger the closure of Hsh155$^{\text{HEAT}}$. Instead, insertion of the BS-A into the protein pocket formed by Hsh155 and Rds3 is likely the major trigger. In the case of the U257A complex, a destabilized U2-BS helix resulted from the branch-site mutation (e.g., flexible loop instead of a clear bulge may form due to mismatches) may hinder the correct bulging of the BS-A[179] and thereby affects Hsh155$^{\text{HEAT}}$ closure.

However, whether this mechanism of Hsh155$^{\text{HEAT}}$ closure can be extrapolated to the human system is not clear. As the human branch-site sequence is much less conserved, human spliceosomes must use weaker branch sites with mismatches to the U2 snRNA frequently. Therefore, if the BS-A insertion is also the major trigger of SF3B1$^{\text{HEAT}}$ closure, the human spliceosome must have additional mechanisms to compensate for the instability of the U2-BS helix. One possibility is that the human spliceosome uses additional protein components to stabilize the U2-BS helix. For instance, the

human SF3b complex contains an additional protein, SF3B6/p14[6], which may bind and stabilize the U2-BS helix. The human SF3B1 also contains a longer N-terminal region with five ULMs in contrast to only one in yeast[161], which allows the human SF3B1 to interact with more splicing factors that may also play a role in stabilizing the U2-BS helix. Another possibility is that the human SF3b binding pocket is less stringent, allowing insertion of the BS-A that is not perfectly orientated. Previously, it has been shown that by mutating the asparagine (N) 747 and leucine (L) 777 of yeast Hsh155 into the valine (V) and asparagine (N) found in human SF3B1 at the corresponding positions, sensitivity to chemical modulators that specifically target human SF3B1 can be introduced to yeast[180]. More importantly, Hsh155 N747V and L777N mutations also increase the splicing of pre-mRNAs with mismatches at the conserved branch-site sequence in yeast[180]. It is, therefore, tempting to assume that by introducing "human-like" mutations, the binding pocket of Hsh155 becomes more permissive, allowing binding of chemical modulators or a BS-A that is not correctly bulged.

4.2.2 Interactions between the U1 and U2 regions in pre-A complexes

In the pre-A complex structures, the U1 and U2 regions are connected by several bridges. Since our pre-A complexes represent an E-to-A complex intermediate, these bridges may represent the early interactions that play a role in recruiting the U2 snRNP to the E complex.

One major interaction between U1 and U2 snRNPs, termed as bridge 1, is mediated by the U1 protein Prp40 and the U2 protein Rse1. Prp40 is a U1 component that contains six C-terminal FF domains, which are characterized by a conserved N-terminal and a C-terminal phenylalanine (F)[181]. FF domains typically occur in tandem arrays and mediate protein-protein interactions. Recently, another FF domain containing protein, TCERG1, has been identified in human spliceosomes[93]. TCERG1 binds multiple proteins via its six FF domains and likely functions as an interaction mediator that stabilizes a certain conformation of the spliceosome by interacting with multiple spliceosomal components simultaneously. Similarly, our CXMS data together with the EM density suggest that Prp40 also functions as an interaction platform. It binds Luc7 and Snp1 via its FF1 and FF6, respectively, anchoring Prp40 with the

major part of the U1 snRNP, while tethering the U2 snRNP via binding to WD40 β-propeller domain B (BPB) of Rse1 through FF4. Interestingly, the BPB of U2 Rse1 (yeast homologue of SF3B3) is also a known interaction mediator that interacts with multiple components throughout the splicing cycle. Therefore, it is conceivable that the two interaction mediators may play a key role in recruiting the U2 snRNP to the E complex during early spliceosome assembly. In addition to Prp40 and Rse1, CXMS data show that bridge 1 further contains Snu71, as multiple crosslinks are detected between the C-terminal region of Snu71 and Prp40 FF2-3 as well as Rse1$^{\text{BPB}}$. Snu71 also crosslinks extensively with Luc7. The crosslinks are thus consistent with previous data that show Prp40, Luc7, and Snu71 form a stable sub-complex within the U1 snRNP[182]. The Prp40-Luc7-Snu71 trimer bridges U1 to the branch site binding Msl5-Mud2 in the yeast E complex[172]. Our data suggest that this trimer may also play a role in recruiting/tethering the U2 snRNP after the E complex formation. Furthermore, CXMS data indicate Prp5 N-terminus interacts with Snu71 C-terminus and Rse1$^{\text{BPB}}$ at bridge 1. Prp5 has been shown to mediate U1 and U2 interaction by its N-terminal RS domain in *S. pombe*[8]. It is surprising that in *S. cerevisiae*, Prp5 plays a similar role, despite lacking an N-terminal RS domain.

Another bridge between U1 and U2 snRNPs, termed as bridge 2, contains the intron nucleotides upstream of the branch site but downstream of the intron stem found in ACT1 pre-mRNA. This bridge is also found in the yeast A complex assembled on UBC4 pre-mRNA at the analogous position[142]. While the ACT1 intron region is much longer than the UBC4 one, the RNA bridge appears to be of similar length. Therefore, formation of the ACT1 intron stem and its binding to U1 snRNP likely keep the downstream branch site and the U2 snRNP in close vicinity to U1 snRNP. It also explains why disruption of the intron stem of the ACT1 pre-mRNA substantially impedes the splicing efficiency[172]. Furthermore, the intron nucleotides upstream of the U2-BS helix are bound by Hsh49$^{\text{RRM1}}$, which may chaperone the single-stranded RNA. More importantly, the binding of SF3B4 (human Hsh49 homologue) to the intron region directly upstream of the U2-BS helix is essential for the U2 recruitment during A complex formation[183]. Therefore, Hsh49$^{\text{RRM1}}$ may be a key anchor of the U2 snRNP to the pre-mRNA during the initial recruitment of U2 to the E complex.

In the E complex, Msl5 (yeast homologue of SF1) binds the branch site via its KH-QUA2 domain, while Mud2 (yeast homologue of U2AF) binds downstream of the branch site via its two RRM domains[85,105]. Msl5 also binds the WW domains of Prp40 via its N-terminal region[184], thus bridging the U1 snRNP to the branch site in the E complex. For stable U2 incorporation, Msl5 has to be displaced from the branch site. Surprisingly, despite the formation of the U2-BS helix, our mass spectrometry data show a high abundance of Msl5-Mud2 in pre-A complexes. In pre-A complexes, although Msl5 is displaced from the branch site, CXMS data indicate that Msl5-Mud2 remains attached and locates in the poorly resolved region downstream of the U2-BS helix, likely through the interaction of Mud2 with the intron downstream of the branch site, and that Msl5 remains bound to Prp40 WW domains. Thus, in the pre-A complexes, the U1 snRNP remains attached to the intron downstream of the branch site via this Prp40-Msl5-Mud2 interaction, resulting in the third bridge between the U1 and U2 regions, which is highly flexible and cannot be seen from the EM density. For mature A complex formation and U2-BS helix docking into Hsh155, Mud2 has to be detached from the intron. It was previously asserted that Msl5 and Mud2 are simultaneously displaced from the pre-mRNA by the action of the DEAD-box RNA helicase Sub2[105,106]. In contrast, our pre-A complexes indicate that the displacement of Msl5-Mud2 is more likely a two-step process — i.e., Msl5 is displaced from the branch site first for the U2-BS helix formation; next, Mud2 is displaced from intron binding to allow the U2-BS helix docking. However, it is still unclear whether Sub2 ATPase activity is responsible for both steps.

4.2.3   U2 remodelling during the pre-A complex formation

Since the human and yeast U2 snRNPs contain evolutionarily conserved proteins with structurally highly similar domains, it is very likely that the human and yeast isolated U2 snRNPs exhibit a similar overall architecture. By comparing the human 17S U2 snRNP and the yeast pre-A complex structures, we reveal that the U2 snRNP undergoes multiple internal structural rearrangements during pre-A complex formation (Figure 4.4).

First, the N-terminal region of Prp5 appears to be partially dissociated from Hsh155[HEAT] binding (Figure 4.4a-b). Previous biochemical studies have shown
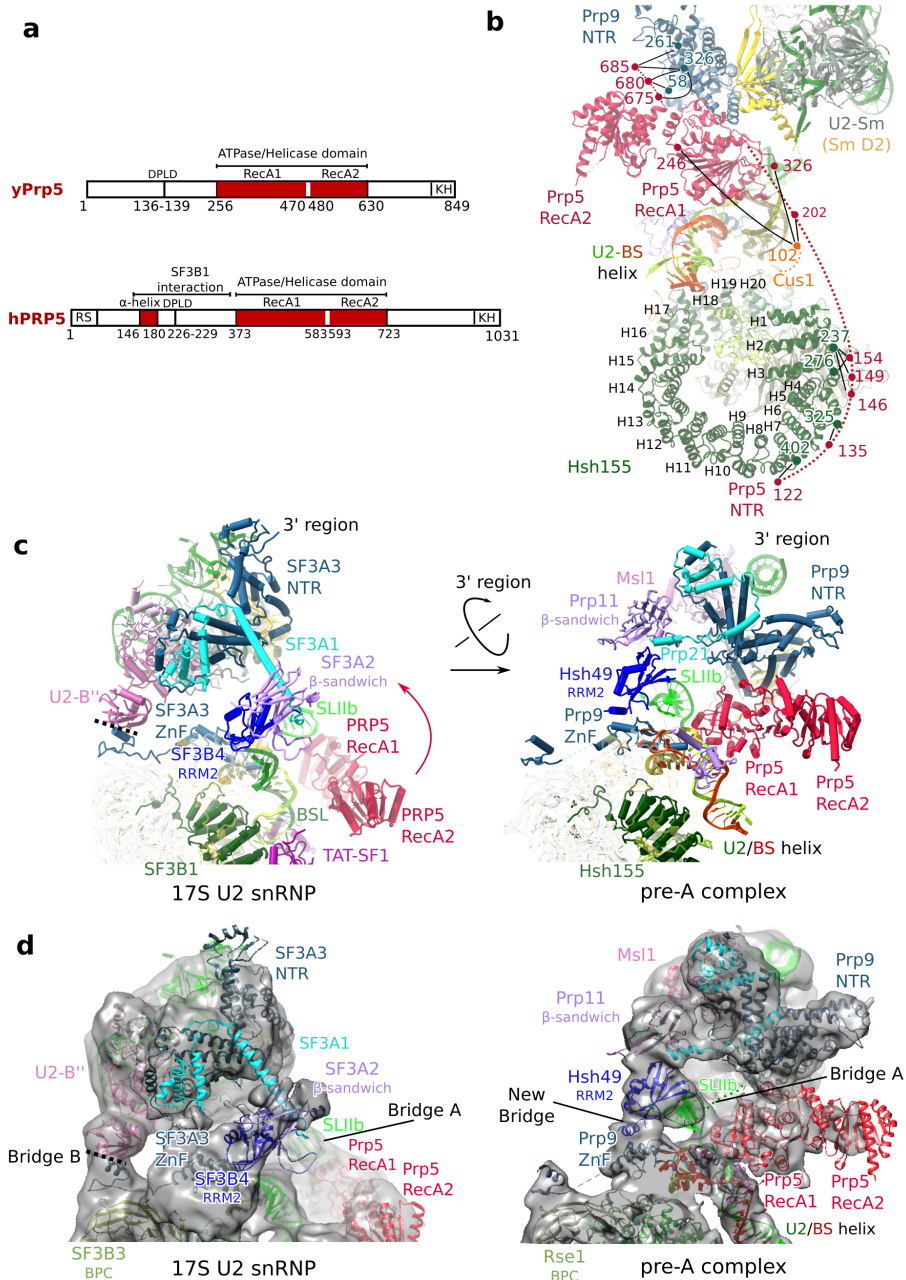
Figure 4.4: **Prp5 N-terminal region remains attached to Hsh155 while the RecA domains and the U2 3' region undergo a repositioning during pre-A formation. a.** Domain representation of the yeast (y) and human (h) DEAD-box RNA helicase Prp5. **b.** CXMS data show the N-terminal region (NTR) of Prp5 remains partially associated with Hsh155. The dotted line represents the potential path based on CXMS. **c-d,** Structural remodelling of the U2 snRNP during pre-A formation. The human 17S U2 snRNP structure is shown, since a structure of the yeast isolated U2 snRNP is lacking. However, due to the high conservation of the yeast and human U2 components and the similar structures of the conserved domains, the yeast U2 snRNP is presumably highly similar to the human one. The 17S U2 snRNP and the pre-A complex are aligned via C-terminal HEAT and SLIIa. The structural comparison shows that the U2 3' region and Prp5 RecA domains undergo a repositioning, forming the new interaction between Prp5[RecA1] and Prp9 N-terminal region (NTR). The movement of Prp5 RecA domains creates a large opening for the U2-BS helix formation. The U2 remodelling also involves dissociation of protein-protein interactions and formation of new interactions.

that the N-terminal region of Prp5 binds to H1-6 and H9-12 of Hsh155 in the U2 snRNP[155]. The human 17S U2 snRNP structure shows that the binding of H9-12 involves an extended α-helix of Prp5[11]. While the structure prediction by AlphaFold 2 indicates that yeast Prp5 also forms an extended α-helix at the corresponding region (https://alphafold.ebi.ac.uk/), we do not see this α-helix density in our pre-A complexes. CXMS data also show that the Prp5 N-terminal region interacts mainly with H1-6 of Hsh155, and no crosslinks between Prp5 and H9-12 can be detected. Thus, the extended α-helix that initially binds H9-12 may have been dissociated during the U2 to pre-A transition.

In addition, Prp5 RecA domains and the U2 3' region (i.e., the U2 Sm core, Lea1, Msl1, and the SF3a core) undergo a repositioning (Figure 4.4c). In the human 17S U2 snRNP, Prp5 RecA domains interact extensively with TAT-SF1 and sequester the BSL. In the pre-A complexes, Cus2 (yeast homologue of TAT-SF1) is displaced by Prp5 ATPase activity for the U2-BS helix formation. Prp5 RecA domains locate further away from Hsh155$^{\text{HEAT}}$ in pre-A, possibly due to loss of interactions with Cus2 that initially binds to Hsh155$^{\text{HEAT}}$. The translocation of Prp5 RecA domains creates a large opening between Prp5 RecA domains and Hsh155$^{\text{HEAT}}$. In addition, the U2 3' region undergoes a large-scale translocation and moves closer to the Prp5 RecA domains. The repositioning of Prp5 RecA domains and the U2 3' region leads to a new interface formation between Prp5$^{\text{RecA1}}$ and Prp9 N-terminal α-helical bundles. As detailed in 4.2.4, this interface prevents a further movement of the U2 3' region towards the U2-BS helix. Thus, the large opening between Hsh155$^{\text{HEAT}}$ and Prp5 RecA domains resulted from the repositioning of the RecA domains is also maintained by the new position of Prp5$^{\text{RecA}}$. This would explain the ATP-independent role of Prp5[4]. That is, Prp5 blocks premature docking of the U2 3' region onto the U2-BS helix, making the room and the time required for unwinding of the BSL and formation of the U2-BS helix. Moreover, the repositioned Prp5$^{\text{RecA1}}$ now contacts the single-stranded U2 snRNA that connects SLIIa and the U2-BS helix, which is consistent with crosslinks of Prp5 to this region (nucleotides 45-49) of yeast U2 snRNA observed previously[10]. That only the Prp5 N-terminal region and RecA1, but not RecA2, are important for Prp5 binding in the pre-A complexes, is also consistent with that the RecA2 domain is dispensable for Prp5 ATP-independent role[4].

Finally, the translocation of the U2 3' region involves disruption of certain protein-protein interactions and formation of new interactions within the U2 snRNP between the 3' and the 5' regions (Figure 4.4c-d). First, the interaction between U2-B" (Msl1 in yeast) RRM2 and SF3A3 (Prp9 in yeast) C-terminal region is dissolved, allowing the movement of the U2 3' region. Meanwhile, the structural rearrangements bring Prp11 β-sandwich domain and Hsh49^{RRM2} towards the Prp9 zinc-finger (ZnF) domain. This allows Hsh49^{RRM2} to dock on top of the Prp9^{ZnF} on one side while interacting with Prp11 β-sandwich domain on the other side, forming a new bridge between the U2 3' and 5' regions. Moreover, in the 17S U2 snRNP, SLIIb is not stabilized and is only poorly resolved. In the pre-A complexes, the position of SLIIb is shifted and stabilized by binding to Hsh49^{RRM2}. Therefore, Hsh49^{RRM2} appears to play a central role in the formation of the new interactions in pre-A complexes by binding to at least three U2 components simultaneously.

4.2.4   Structural remodelling during A complex formation requires Prp5 dissociation

The previously published yeast A complex structure shows a highly similar overall architecture as the corresponding U1 and U2 region found in the yeast pre-B complex[142,167], thus representing a conformation that is primed for tri-snRNP recruitment. On the other hand, our pre-A complex is not compatible with tri-snRNP binding (Figure 4.5a). The comparison between our pre-A complex with the published yeast A complex reveals a large-scale structural remodelling from pre-A to A transition that is required to generate the tri-snRNP docking site, but is blocked by Prp5 binding (Figure 4.5b-c).

First, the U2 3' region rotates about 50° relative to the 5' region during the pre-A to A complex transition. The rotation of the U2 3' region is not possible in the presence of Prp5 RecA domains, as the latter stably bind in a mutually exclusive manner with the SF3a core proteins (Figure 4.5a-b). Second, the U1 snRNP rotates about 45° relative to the U2 snRNP. The rotation of the U1 snRNP together with the rotation of the U2 3' region results in the binding of Prp39 to Lea1, stabilizing the relative

orientation of the U1 and U2 snRNPs observed in the yeast A and pre-B complexes. It is worth mentioning that without the movement of the U2 3' region, Lea1 is shielded behind Prp9 N-terminal region and is not accessible for Prp39 binding (Figure 4.5c). In this way, by blocking the rotation of the U2 3' region, Prp5 RecA domains also prevent premature interaction between Prp39 and Lea1. Furthermore, rotation of the U1 snRNP requires dissolution of the U1-U2 bridge 1. As described in 4.2.2, since bridge 1 involves Prp5 N-terminus binding to Rse1 and Snu71, dissolution of this bridge also may require the dissociation of Prp5. As Prp5 blocks the U2 3' region rotation by its RecA domains and tethers the U1 snRNP by its N-terminus, the release of Prp5 may coordinate both rotations that are required for A complex formation and tri-snRNP joining. Hence, our structures provide a structural explanation of why the tri-snRNP and Prp5 binding are mutually exclusive, as previously revealed by biochemical studies in yeast[10].

4.2.5 A proposed mechanism of how Prp5 plays a role in proofreading the branch site

Taken together, our pre-A structures suggest a model of how Prp5 contributes to proofreading the pre-mRNA branch site (Figure 4.6). First, the U2 snRNP is recruited to the E complex, and the ATPase activity of Prp5 triggers the remodelling of the U2 snRNP, including U2-BS helix formation and the repositioning of the U2 3' region and Prp5 RecA domains. This leads to formation of the pre-A complex, in which the binding of Prp5 transiently blocks further rearrangements that are required for A complex formation and tri-snRNP docking, creating the time and space required for U2-BS helix formation. In the wild-type scenario, the correct formation of the U2-BS helix would orientate the BS-A in a way that allows insertion of the BS-A into the binding pocket formed by Hsh155 and Rds3, which triggers the final closure of Hsh155$^{\text{HEAT}}$. We propose that since the U2-BS helix is flexible, the insertion of BS-A may be achieved by the U2-BS helix moving closer to Hsh155$^{\text{HEAT}}$ intermittently and probing for the binding pocket. Subsequently, the closure of Hsh155$^{\text{HEAT}}$ destabilizes the binding of Prp5 N-terminal region. Concomitantly, the coordinated docking of the U2-BS helix and the resulting movement of the U2 snRNA contacted by Prp5$^{\text{RecA1}}$ destabilize the binding of Prp5 RecA domains. These RNP rearrangements lead to the
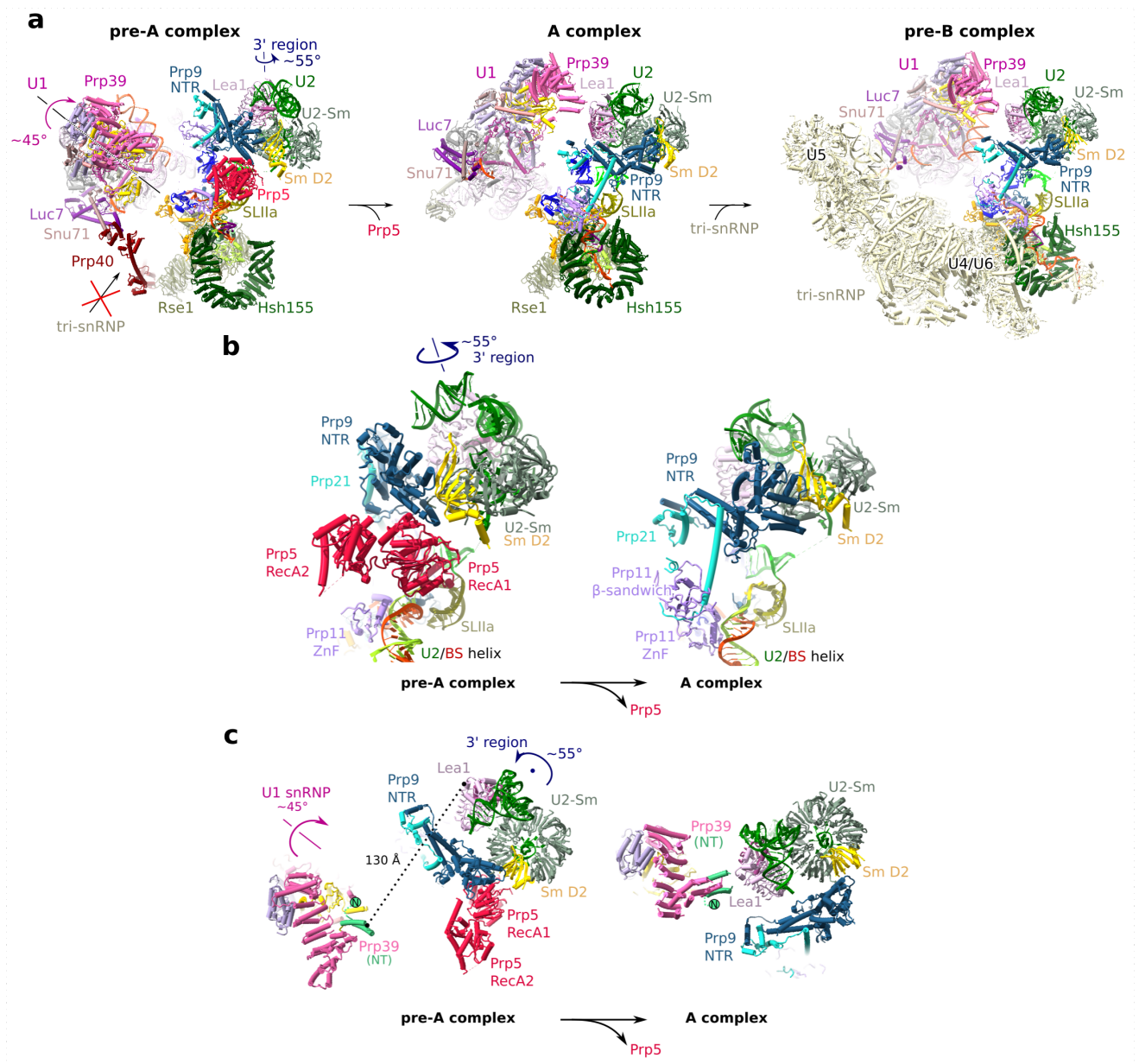
Figure 4.5: **Structural remodelling during pre-A to A transition generates the binding site for the tri-snRNP but is blocked by Prp5 binding. a.** Structural remodelling during pre-A to A and pre-B complex. Structures of the yeast A and pre-B complexes are previously published (PDB: 6G90[142], 5ZWM, 5ZWN[167]) and are aligned with the pre-A complex via Hsh155 C-terminal HEAT and SLIIa. The rotations of the U2 3' region and U1 snRNP are indicated by curved arrows. **b.** Close-up view of the U2 3' region rotation during pre-A to A complex transition. The SmD2 protein is shown in yellow to better follow the rotation. **c.** Close-up "top view" of the U2 3' region and U1 snRNP rotation during pre-A to A complex transition. Prp39 and Lea1 are initially separated by about 130 Å, and Lea1 is initially shielded behind Prp9 N-terminal region and is thus inaccessible for Prp39 binding in the pre-A complex.

final displacement of the Prp5 protein, allowing subsequent transition of the pre-A to the A complex. As the BS-A is correctly orientated in the wild type U2-BS helix, the closure of Hsh155$^{\text{HEAT}}$ may be very efficient, making the pre-A stage highly transient. In contrast, if the U2-BS helix lacks a bulged BS-A or if the BS-A is not correctly bulged due to branch-site mutations, the insertion of the BS-A into the Hsh155 binding pocket is either not possible or significantly impeded. As a result, the Hsh155$^{\text{HEAT}}$ closure and the Prp5 release are hindered, thus blocking further progression of the spliceosome assembly pathway. Nevertheless, unlike the ΔBS-A pre-mRNA that forms a perfect U2-BS helix, pre-mRNAs bearing branch-site mutations (e.g., U257A complex) still contain a potential BS-A. Therefore, although much less efficiently, there is still a chance of the BS-A binding to Hsh155$^{\text{HEAT}}$, explaining why U257A complexes can proceed further along the wild-type assembly pathway and are not dead-end complexes[10].

During the course of this thesis, a cryo-EM structure of a human A-like complex is reported[185]. This complex is stalled by an SF3B1 inhibitor spliceostatin A (SSA) and is believed to be at an E-to-A intermediate. In this human A-like complex, SF3B1$^{\text{HEAT}}$ is kept in an open conformation by SSA, abolishing the docking of the U2-BS helix. Strikingly, the mass spectrometry data indicate that significant amounts of Prp5 are retained in this human A-like complex. Thus, it appears that in the human system, the open conformation of SF3B1$^{\text{HEAT}}$ is correlated with retention of Prp5 as well, indicating that Prp5 may have a similar role in sensing the closure of SF3B1$^{\text{HEAT}}$ in humans.

Previous genetic and biochemical studies have shown that multiple mutations of Prp5 are able to suppress branch-site mutations (i.e., improve the splicing of pre-mRNA containing branch-site mutations)[9 10 163]. These include mutations in Prp5 DPLD motif at Prp5 N-terminal region; the conserved SAT motif at Prp5$^{\text{RecA1}}$ that affects the ATPase activity of Prp5; and mutations at Prp5$^{\text{RecA1}}$ outside of the SAT motif, such as K372E, N399D, and G401E that have no effects on Prp5 ATPase activity. Our pre-A structures provide structural insights into how some of the Prp5 mutations may suppress branch-site mutations by compromising Prp5 binding affinity and hence its
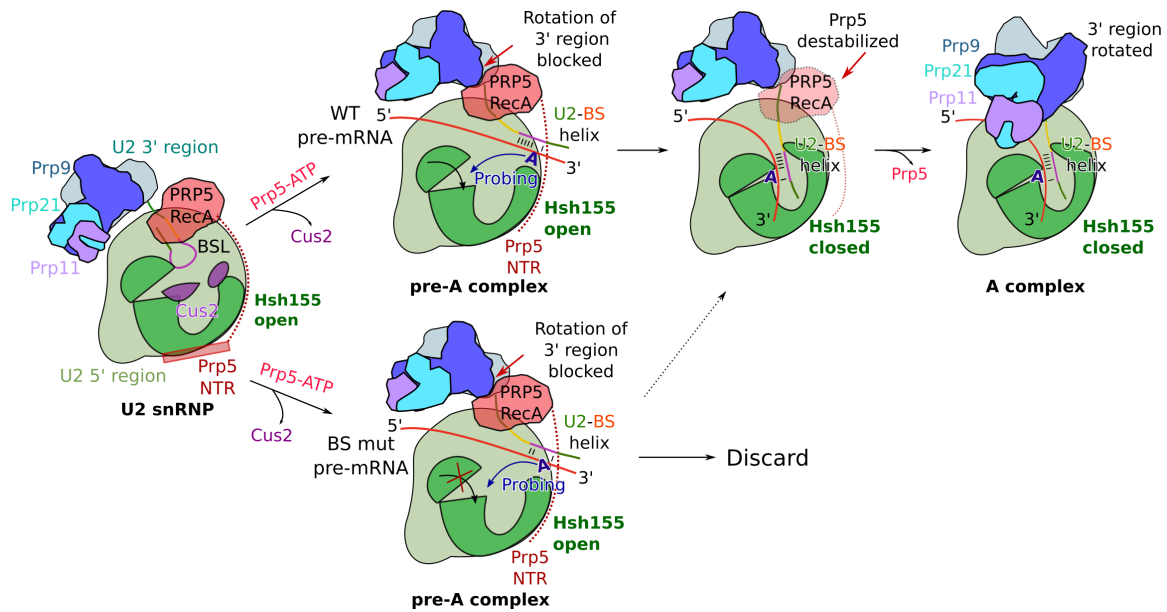
Figure 4.6: **A proposed model of Prp5 proofreading the branch site during A complex formation.** The U1 snRNP is not shown for simplicity. The U2 snRNP depiction is based on the human 17S U2 snRNP. The closure of Hsh155, release of Prp5, and docking of U2-BS helix may occur in a highly coordinated manner concomitantly. The dashed arrow indicates that certain branch-site mutations, including U257A, do not completely block the pathway and are able to undergo further assembly. WT, wild-type; BS mut, mutated branch-site sequence.

proofreading role. In the pre-A complexes, our CXMS data suggest that the Prp5 N-terminal region, including the DPLD motif remains binding to Hsh155^{HEAT}. Therefore, the DPLD mutant destabilizes Prp5 binding by compromising the interaction between Prp5 N-terminal region and Hsh155^{HEAT}. K372, N399, and G401 are located at the corner where Prp5^{RecA1} contacts the U2 snRNA. Thus, K372E, N399D, and G401E mutations that introduce negative charges may significantly affect the binding of Prp5^{RecA1} to the RNA backbone, explaining the reduced affinity observed previously[10]. However, the effects of SAT mutations on Prp5 proofreading cannot be directly explained by our structure, as they are not located at binding interfaces. Previous studies have shown that SAT mutations hinder the ATPase activity and the conformational dynamics of Prp5, and that the reduced ATPase activity and dynamics correlate with improved branch-site mutant usage[9,165]. It is conceivable that the altered dynamics of Prp5 may change its binding affinity to other U2 components (e.g., Prp9 N-terminal region) indirectly. Alternatively, Prp5 may contribute to fidelity control prior to the pre-A complex formation in an ATP-dependent manner, similar to the previously proposed "kinetic proofreading" mechanism (see 2.2.6). Lastly, Prp5 ATPase activity may function again in the stalled complex and play a role in disassembly, preventing

potential further progression. Therefore, the SAT mutants may improve splicing by affecting the disassembly function of Prp5. Nevertheless, the last scenario is less likely, as the role of Prp5 in disassembling stalled spliceosomes has never been reported.

In sum, our pre-A structures show that Prp5 does not proofread the branch-site sequence directly. Instead, it functions as a checkpoint and senses the RNP conformation of the pre-A complex and blocks spliceosomal assembly if mutations in the branch site hinder the final closure of Hsh155$^{\mathrm{HEAT}}$. The stalled complex may then be targeted for disassembly by Prp43 or another RNA helicase.

4.3 Current limitations and perspectives of cryo-EM as a tool to study spliceosomal complexes

Limited resolution due to structural heterogeneity

In this study, we have reported cryo-EM structures of two early spliceosomal complexes. While these structures have provided insightful information of early spliceosome assembly, structural heterogeneity in the data set has largely limited the resolution of both complexes, especially at the peripheral and more dynamic parts. The low resolution resulting from flexible parts of a macromolecular complex is a common issue in SPA studies of macromolecular complexes. In this study, we turned to an integrative structural biology approach, which allowed fitting available structures into the low-resolution regions of the EM map on the basis of CXMS data. Although this allowed us to interpret certain low-resolution regions by docking structures of larger domains as rigid bodies, many smaller domains/motifs and flexible loops could not be modelled, due to lack of interpretable structural features. For instance, in the 17S U2 snRNP structure, only 173 residues out of 895 residues were modelled for SF3B2; and only 182 out of 755 residues modelled for TAT-SF1. Thus, many parts in the EM map cannot be interpreted, including the unmodelled protein region (UPR) near TAT-SF1$^{\text{RRM}}$ that directly contacts the BSL and thus may be a functionally highly important region. Therefore, improving the resolution, especially at peripheral parts, will largely help structural interpretation.

One way to tackle this issue is to improve biochemical sample preparation itself. A gentler sample preparation (e.g., chromatography-free purification, lower salt concentration, chemical fixation) that better preserves the integrity of the complex is useful to limit the structural heterogeneity resulting from damaged complexes, thus improving the overall resolution of the final reconstruction[186,187]. Moreover, flexible parts can be stabilized in an artificial manner, for instance, by binding to small molecules or nano-bodies[13,185,188]. Although conceptually straightforward, screening small molecules or nano-bodies that specifically target and stabilize the flexible regions can sometimes be very challenging.

In addition to improving the sample preparation, another way to improve the reso-
lution, especially at peripheral parts, is to reconstruct multiple parts of the complex
separately *in silico*. For instance, by signal-subtraction coupled with focused 3D
classification, the U2 5' region of the yeast pre-catalytic B complex was improved from
20 Å to about 4 Å[146]. More recently, a non-uniform reconstruction method called
"multibody refinement" has been widely applied, especially for large complexes like
ribosomes and spliceosomes[28]. In contrast to traditional reconstruction that calculates
a 3D structure as one unity, multibody refinement splits the complex into multiple
rigid bodies by user-defined masks. This allows each "body" to translocate and rotate
relative to one another, thus improving the alignment and the resolution of each
user-defined subpart of the complex. In this study, we applied multibody refinement
for the yeast pre-A complex, and improved resolutions of both the 3' and 5' of the U2
snRNP from about 30 Å to 8-9 Å. However, since multibody refinement assumes each
subpart being a rigid body, the intrinsic flexibility within each user-defined "body"
cannot be overcome by this method, as revealed by the limited resolution of 8-9 Å at
the U2 region. In the case of the human 17S U2 snRNP, multibody refinement did not
improve the reconstruction of the low-resolution peripheral regions (e.g., Prp5 RecA
domains and TAT-SF1). This is (partially) due to the intrinsic flexibility of Prp5 RecA
domains and its surrounding regions, and therefore it was not possible to define indi-
vidual rigid bodies. Additionally, each "body" must have sufficient mass for accurate
alignment. The relatively small volume of the peripheral parts of the 17S U2 snRNP
also limited the successful application of multibody refinement. Taken together, to
improve the overall and local resolutions, the biochemical sample preparation should
be improved, and better algorithms that are able to model the continuous nature of
flexible complexes are required (e.g., 3D PCA or deep learning-based methods, see
below).

Hidden information due to low percentage of data used for final recon-structions

For single-particle cryo-EM, another limitation that comes along with the structural
heterogeneity is that only very small subsets of the data that show the highest

structural homogeneity were used for final reconstructions. While this is essential for high-resolution reconstruction, valuable information of rare conformational states and conformational changes remains hidden by discarding large amounts of data during the image processing. In the case of the yeast pre-A complex, only about 7 % of the data were used for the final reconstruction[14]; and 8 % used for the 17S U2 snRNP final reconstruction[11]. The remaining more than 90 % of the data that were considered "bad" particles may provide valuable insights about the conformational states that are less populated, as well as the structural dynamics of the complex.

To tackle this issue, multiple algorithms have been developed to resolve the structural dynamics computationally. In a 3D principal component analysis (PCA)-based method[31], hundreds to thousands of structures are reconstructed from the entire dataset and aligned to a common reference. 3D PCA is applied to describe the structural variations throughout the reconstructions. Subsequently, the first two eigenvectors from PCA can be used as x and y coordinates, to calculate a conformational landscape, and each dot in the landscape represents a 3D reconstruction. By using the known particle number of each 3D reconstruction, an energy landscape can then be calculated by Boltzmann equation. This method has been successfully applied and revealed eight different conformational states of the human B^{act} complex that are more populated. 3D PCA-based methods can also be applied to visualize the conformational dynamics and continuous motions within the dataset[31,187,189]. More recently, a modified method for calculating the aforementioned energy landscape has been proposed[190]. In this method (called AlphaCryo4D), instead of using linear subspace derived from PCA eigenvectors to plot the landscape, non-linear manifold embedding was applied; and instead of using 3D PCA to describe structural variations, a neural network-based method was applied to extract structural features. In addition to AlphaCryo4D, another neural network-based method, cryo-DRGN (Deep Reconstructing Generative Networks), was developed recently[32]. Both neural network-based methods have been shown to successfully uncover previously hidden conformational states in published datasets, as well as reveal continuous motions. In addition, by a better separation of previously hidden conformational states, both methods were shown to improve the overall resolutions of final reconstructions[32,190]. Thus, in future studies, by applying various computational algorithms, multiple structural states

should be reconstructed from the dataset instead of one static reconstruction; and conformational dynamics of the complex should be reported, so that the full potential of the structural information can be extracted from the dataset.

Towards sub-2 Å resolution for spliceosome complexes

Recent cryo-EM studies of spliceosome intermediates have provided detailed structural information, and for the more stable parts of the spliceosomal B^{act} and C complexes, the resolution have achieved 2.5 Å to 3 Å[110,191]. However, the resolution can still be substantially improved, as recent developments of microscope hardware (e.g., better electron detector, monochromator, Cs-corrector, cold field emission gun) were shown to improve the data quality and the achievable resolution significantly[192,193]. Moreover, ultra-stable EM grids with smaller holes (c.a. 200-300nm)[194] have been developed, which were shown to limit the particle movement during imaging, thus providing another layer of improvement in the data quality. Therefore, by combining these recent technical advances, it is conceivable that a much higher resolution (e.g., sub-2 Å) can be achieved for spliceosomal complexes in the near future. At this resolution, the molecular mechanisms of the spliceosome can be studied at a chemical level.

Towards in situ structural determination

Current structural studies of the spliceosome, including this study, largely rely on *in vitro* splicing reaction and purification of distinct states of spliceosomal complexes. During the past six years, most spliceosomal major intermediate states have been purified and structurally characterized by this approach. While these structural studies have provided valuable information, spliceosome structure has never been resolved in their cellular context. Recent development of cryo-electron tomography (cryo-ET) has allowed *in situ* visualization of a 70S ribosome at an average resolution of 3.5 Å[195]. This has provided a path towards *in situ* high-resolution structure determination of other macromolecular complexes, like spliceosomes. *In situ* studies of spliceosomal complexes will allow visualization of spliceosome complexes in a more natural state, and thus may significantly expand our understanding of the splicing

process. For instance, *in situ* studies may allow visualization of previously unknown interactions with other molecular machinery, like RNA polymerase, which will largely promote our understanding of co-transcriptional splicing. However, certain obstacles are still limiting high-resolution *in situ* studies of spliceosomes. Unlike ribosomes that are highly abundant and less dynamic, spliceosomes are relatively sparse and show significant structural dynamics from one state to another. In addition, cryo-ET studies of the human nucleus require a thin sample (i.e., a lamella), which is usually produced by cryo-focused ion beam (cryo-FIB) milling[196]. However, cryo-FIB milling can be time-consuming and thus largely limit the throughput of data acquisition. Moreover, targeting the location of interest for sparse spliceosome complexes can also be challenging during cryo-FIB milling.

4.4   Conclusion

In this thesis work, we have resolved cryo-EM structures of the human 17S U2 snRNP and a yeast early spliceosomal intermediate, the pre-A complex[11][14]. Combining the two structures with previous studies, we not only revealed the molecular architectures of the two individual complexes, but also shed new light on the early spliceosomal assembly pathway and the branch site proofreading mechanism.

Our human 17S U2 snRNP structure reveals the U2 snRNP architecture before Prp5 ATPase activity, in which SF3B1 adopts an open conformation and the BSL is initially sequestered by Prp5, TAT-SF1, and SF3B1$^{\text{HEAT}}$. A comparison between the 17S U2 snRNP and the U2 region in later spliceosome complexes (B$^{\text{act}}$ complex) shows the RNP rearrangements that are required for stable U2 incorporation. Our yeast pre-A complexes further reveal that after the U2 snRNP is recruited to the E complex, Prp5 ATPase activity leads to the displacement of Cus2 and formation of the U2-BS helix, but not necessarily the closure of Hsh155$^{\text{HEAT}}$. The Prp5 mediated U2 remodelling also involves repositioning of Prp5 RecA domains and the U2 3' region. The new "proofreading position" of Prp5 serves as a wedge, transiently preventing further structural rearrangements that are required for tri-snRNP docking. This provides the time and space required for the U2-BS helix formation. The BS-A binding to Hsh155 then triggers the final closure of Hsh155$^{\text{HEAT}}$ as well as the displacement of Prp5. In contrast, branch-site mutations hinder the BS-A bulging and fail to trigger further RNP rearrangements, thus being stalled at the pre-A complex stage. Therefore, we conclude Prp5 senses the overall RNP conformation of the pre-A complex, instead of proofreading the branch-site sequence itself, thus providing the first structural insight into the proofreading mechanism of a DEAD-box RNA helicase.

Nevertheless, while our structures reveal exciting new insights, they also bring new questions. First of all, although we have revealed the end result of Prp5 ATPase activity, the exact mechanism of how Prp5 performs this role is completely unknown. Typically, DEAD-box proteins bind to double-stranded RNA helices[108]. But it is unclear which RNA stimulates Prp5 activity and how Prp5 displaces TAT-SF1/Cus2 mechanistically. Furthermore, we cannot discern the exact conformation of Prp5

RecA domains in the 17S U2 snRNP. Some DEAD-box RNA helicases, like eIF4A, change the RecA domains conformation upon binding to other proteins[197,198]. It is conceivable that due to the interaction with TAT-SF1, Prp5 RecA domains may adopt a yet-to-be-discovered conformation. In addition to Prp5, another DEAD-box RNA helicase, UAP56 (Sub2 in yeast), also plays an important role during A complex formation. UAP56 has been proposed to displace SF1 (Msl5 in yeast) from the branch site binding[105 106], but its mechanism remains poorly understood. Last, although our pre-A complexes provide important insights into the branch site proofreading mechanism mediated by Prp5 in yeast, whether this mechanism can be extended to humans requires further investigations.

Starting from 2015, cryo-EM has revolutionized the understanding of spliceosomes on the basis of decades of biochemical and genetic research. However, due to the flexibility of early spliceosome complexes, much remains to be understood for the early spliceosome assembly. Especially, structural insights into exon definition, alternative splicing, co-transcriptional splicing are still very limited. In the future, different pre-mRNA substrates should be tested to stabilize alternative splicing factors in humans. New biochemical approaches should be developed to further stabilize the flexible complexes, and with new technological advances in the EM field, our understanding of RNA splicing may be revolutionized again.

References

1. Will, C. L. & Lührmann, R. Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology* **3**, 1–2 (2011). URL `http://www.ncbi.nlm.nih.gov/pubmed/21441581http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3119917`.

2. Kastner, B., Will, C. L., Stark, H. & Lührmann, R. Structural insights into nuclear pre-mRNA splicing in higher eukaryotes. *Cold Spring Harbor Perspectives in Biology* **11** (2019). URL `https://pubmed.ncbi.nlm.nih.gov/30765414/`.

3. Plaschka, C., Newman, A. J. & Nagai, K. Structural basis of nuclear pre-mRNA splicing: Lessons from Yeast. *Cold Spring Harbor Perspectives in Biology* **11** (2019). URL `https://pubmed.ncbi.nlm.nih.gov/30765413/`.

4. Perriman, R. & Ares, M. Invariant U2 snRNA Nucleotides Form a Stem Loop to Recognize the Intron Early in Splicing. *Molecular Cell* **38**, 416–427 (2010). URL `https://linkinghub.elsevier.com/retrieve/pii/S1097276510003151`.

5. O'Day, C. L., Dalbadie-McFarland, G. & Abelson, J. The Saccharomyces cerevisiae Prp5 protein has RNA-dependent ATPase activity with specificity for U2 small nuclear RNA. *Journal of Biological Chemistry* **271**, 33261–33267 (1996).

6. Will, C. L. *et al.* Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5p homologue and an SF3b DEAD-box protein. *EMBO Journal* **21**, 4978–4988 (2002). URL `http://www.ncbi.nlm.nih.gov/pubmed/12234937http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC126279`.

7. Perriman, R., Barta, I., Voeltz, G. K., Abelsont, J. & Ares, M. ATP requirement for Prp5p function is determined by Cus2p and the structure of U2 small nuclear RNA. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 13857–13862 (2003).

8. Xu, Y. Z. *et al.* Prp5 bridges U1 and U2 snRNPs and enables stable U2 snRNP association with intron RNA. *EMBO Journal* **23**, 376–385 (2004).

9. Xu, Y. Z. & Query, C. C. Competition between the ATPase Prp5 and Branch Region-U2 snRNA Pairing Modulates the Fidelity of Spliceosome Assembly. *Molecular Cell* **28**, 838–849 (2007). URL `http://www.ncbi.nlm.nih.gov/pubmed/18082608http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2246091`.

10. Liang, W. W. & Cheng, S. C. A novel mechanism for Prp5 function in pre-spliceosome formation and proofreading the branch site sequence. *Genes and Development* **29**, 81–93 (2015).

11. Zhang, Z. *et al.* Molecular architecture of the human 17S U2 snRNP. *Nature* **583**, 310–313 (2020).

12. Cretu, C. *et al.* Molecular Architecture of SF3b and Structural Consequences of Its Cancer-Related Mutations. *Molecular Cell* **64**, 307–319 (2016). URL `https://www.sciencedirect.com/science/article/pii/S1097276516305159?via{%}3Dihub`.

13. Finci, L. I. *et al.* The cryo-EM structure of the SF3b spliceosome complex bound to a splicing modulator reveals a pre-mRNA substrate competitive mechanism of action. *Genes and Development* **32**, 309–320 (2018). URL `http://www.ncbi.nlm.nih.gov/pubmed/29491137http://genesdev.cshlp.org/lookup/doi/10.1101/gad.311043.117`.

14. Zhang, Z. *et al.* Structural insights into how Prp5 proofreads the pre-mRNA branch site. *Nature* **596** (2021). URL `http://dx.doi.org/10.1038/s41586-021-03789-5`.

15. BARLOW, W. Probable Nature of the Internal Symmetry of Crystals 1. *Nature* **29**, 205–207 (1883). URL `http://www.nature.com/articles/029205a0`.

16. Friedrich, W., Knipping, P. & Laue, M. Interferenzerscheinungen bei Röntgenstrahlen. *Annalen der Physik* **346**, 971–988 (1913). URL `https://doi.org/10.1002/andp.19133461004`.

17. De Rosier, D. J. & Klug, A. Reconstruction of three dimensional fiber structures from orthogonal projections. *Nature* **217**, 130–134 (1968).

18. Crowther, R. A., DeRosier, D. J., , & Klug, A. The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* **317**, 319–340 (1970). URL `https://royalsocietypublishing.org/doi/10.1098/rspa.1970.0119`.

19. Crowther, R. A., Amos, L. A., Finch, J. T., De Rosier, D. J. & Klug, A. Three Dimensional Reconstructions of Spherical Viruses by Fourier Synthesis from Electron Micrographs. *Nature* **226**, 421–425 (1970).

20. Dubochet, J. & McDowall, A. W. VITRIFICATION OF PURE WATER FOR ELECTRON MICROSCOPY. *Journal of Microscopy* **124**, 3–4 (1981).

21. Brüggeller, P. & Mayer, E. Complete vitrification in pure liquid water and dilute aqueous solutions. *Nature* **288**, 569–571 (1980). URL `http://www.nature.com/articles/288569a0`.

22. Taylor, K. A. & Glaeser, R. M. Electron microscopy of frozen hydrated biological specimens. *Journal of Ultrasructure Research* **55**, 448–456 (1976). URL `http://www.ncbi.nlm.nih.gov/pubmed/933264`.

23. chen Bai, X., McMullan, G. & Scheres, S. H. How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences* **40**, 49–57 (2015). URL `https://www.sciencedirect.com/science/article/pii/S096800041400187X?via{%}3Dihub`.

24. Wade, R. H. A brief look at imaging and contrast transfer. *Ultramicroscopy* **46**, 145–156 (1992).

25. Heymann, J., Conway, J. F. & Steven, A. C. Molecular dynamics of protein complexes from four-dimensional cryo-electron microscopy. *Journal of Structural Biology* **147**, 291–301 (2004). URL `https://linkinghub.elsevier.com/retrieve/pii/S1047847704000425`.

26. Scheres, S. H. *et al.* Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nature Methods* **4**, 27–29 (2007). URL `http://www.ncbi.nlm.nih.gov/pubmed/17179934http://www.nature.com/articles/nmeth992`.

27. Nogales, E. & Scheres, S. H. Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity. *Molecular Cell* **58**, 677–689 (2015). URL `http://dx.doi.org/10.1016/j.molcel.2015.02.019`.

28. Nakane, T., Kimanius, D., Lindahl, E. & Scheres, S. H. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *eLife* **7**, 1–18 (2018).

29. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science* **289**, 905–920 (2000). URL `https://www.science.org/doi/10.1126/science.289.5481.905`.

30. Yusupov, M. M. *et al.* Crystal Structure of the Ribosome at 5.5 Å Resolution. *Science* **292**, 883–896 (2001). URL `https://www.science.org/doi/10.1126/science.1060089`.

31. Haselbach, D. *et al.* Structure and Conformational Dynamics of the Human Spliceosomal Bact Complex. *Cell* **172**, 454–464.e11 (2018). URL `http://linkinghub.elsevier.com/retrieve/pii/S0092867418300473`.

32. Zhong, E. D., Bepler, T., Berger, B. & Davis, J. H. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. *Nature Methods* **18**, 176–185 (2021). URL `http://dx.doi.org/10.1038/s41592-020-01049-4`.

33. Fica, S. M. Cryo-EM snapshots of the human spliceosome reveal structural adaptions for splicing regulation (2020). URL `https://pubmed.ncbi.nlm.nih.gov/32717639/`.

34. VERNON, H. M. Elemente der exakten Erblichkeitslehre. *Nature* **81**, 424–424 (1909). URL `http://www.nature.com/articles/081424a0`.

35. Avery, O. T., Macleod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *Journal of Experimental Medicine* **79**, 137–158 (1944).

36. CRICK, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970). URL `http://www.nature.com/articles/227561a0`.

37. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5âĂš ends of adenovirus 2 messenger RNA. *Cell* **12**, 1–8 (1977).

38. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 3171–3175 (1977).

39. Gilbert, W. Why genes in pieces. *Nature* **271**, 501 (1978).

40. Lopez, P. J. & Seraphin, B. Genomic-scale quantitative analysis of yeast pre-mRNA splicing: Implications for splice-site recognition [1]. *Rna* **5**, 1135–1137 (1999).

41. Zhang, M. Q. Statistical features of human exons and their flanking regions. *Human Molecular Genetics* **7**, 919–932 (1998).

42. Query, C. C., Moore, M. J. & Sharp, P. A. Branch nucleophile selection in pre-mRNA splicing: Evidence for the bulged duplex model. *Genes and Development* **8**, 587–597 (1994). URL `http://www.ncbi.nlm.nih.gov/pubmed/7926752`.

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

44. Norton, P. A. Polypyrimidine tract sequences direct selection of alternative branch sites and influence protein binding. *Nucleic Acids Research* **22**, 3854–3860 (1994). URL `https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/22.19.3854`.

45. Roscigno, R. F., Weiner, M. & Garcia-Blanco, M. A. A mutational analysis of the polypyrimidine tract of introns. Effects of sequence differences in pyrimidine tracts on splicing. *Journal of Biological Chemistry* **268**, 11222–11229 (1993).

46. Fu, X.-D. & Ares, M. Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics* **15**, 689–701 (2014). URL `http://www.nature.com/articles/nrg3778`.

47. Patel, A. A. & Steitz, J. A. Splicing double: Insights from the second spliceosome. *Nature Reviews Molecular Cell Biology* **4**, 960–970 (2003).

48. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008). URL `http://www.nature.com/articles/nature07509`.

49. Black, D. L. Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annual Review of Biochemistry* **72**, 291–336 (2003). URL `https://www.annualreviews.org/doi/10.1146/annurev.biochem.72.121801.161720`.

50. Guthrie, C. Messenger RNA splicing in yeast: clues to why the spliceosome is a ribonucleoprotein. *Science* **253**, 157–163 (1991). URL `https://www.sciencemag.org/lookup/doi/10.1126/science.1853200`.

51. Moore, M., Query, C. & Sharp, P. Splicing of Precursors to mRNA by the Spliceosome. *The RNA World* 303–345 (1993). URL `www.cshlpress.com/copyright`.

52. Chen, L. L. The expanding regulatory mechanisms and cellular functions of circular RNAs. *Nature Reviews Molecular Cell Biology* **21**, 475–490 (2020). URL `http://dx.doi.org/10.1038/s41580-020-0243-y`.

53. Valadkhan, S. The spliceosome: A ribozyme at heart? *Biological Chemistry* **388**, 693–697 (2007).

54. Pyle, A. M. Group II Intron Self-Splicing. *Annual Review of Biophysics* **45**, 183–205 (2016).

55. Toor, N., Keating, K. S., Taylor, S. D. & Pyle, A. M. Crystal structure of a self-spliced group II intron. *Science* **320**, 77–82 (2008).

56. Daniels, D. L., Michels, W. J. & Pyle, A. M. Two competing pathways for self-splicing by group II introns: A quantitative analysis of in vitro reaction rates and products. *Journal of Molecular Biology* **256**, 31–49 (1996).

57. Jarrell, K. A., Peebles, C. L., Dietrich, R. C., Romiti, S. L. & Perlman, P. S. Group II intron self-splicing. Alternative reaction conditions yield novel products. *Journal of Biological Chemistry* **263**, 3432–3439 (1988).

58. Henry, R. W., Mittal, V., Ma, B., Kobayashi, R. & Hernandez, N. SNAP19 mediates the assembly of a functional core promoter complex (SNAPc) shared by

RNA polymerases II and III. *Genes & Development* **12**, 2664–2672 (1998). URL `http://www.genesdev.org/cgi/doi/10.1101/gad.12.17.2664`.

59. Hamm, J., Darzynkiewicz, E., Tahara, S. M. & Mattaj, I. W. The trimethyl-guanosine cap structure of U1 snRNA is a component of a bipartite nuclear targeting signal. *Cell* **62**, 569–577 (1990). URL `https://doi.org/10.1016/0092-8674(90)90021-6`.

60. Hamm, J. & Mattaj, I. W. Monomethylated cap structures facilitate RNA export from the nucleus. *Cell* **63**, 109–18 (1990). URL `http://www.ncbi.nlm.nih.gov/pubmed/2208274`.

61. Izaurralde, E. *et al.* A cap-binding protein complex mediating U snRNA export. *Nature* **376**, 709–12 (1995). URL `http://www.ncbi.nlm.nih.gov/pubmed/7651522`.

62. Kunkel, G. R., Maser, R. L., Calvet, J. P. & Pederson, T. U6 small nuclear RNA is transcribed by RNA polymerase III. *Proceedings of the National Academy of Sciences* **83**, 8575–8579 (1986). URL `http://www.pnas.org/cgi/doi/10.1073/pnas.83.22.8575`.

63. Kambach, C. *et al.* Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell* **96**, 375–387 (1999).

64. Epstein, P., Reddy, R., Henning, D. & Busch, H. The nucleotide sequence of nuclear U6 (4.7 S) RNA. *Journal of Biological Chemistry* **255**, 8901–8906 (1980). URL `http://dx.doi.org/10.1016/S0021-9258(18)43587-9`.

65. Séraphin, B. Sm and Sm-like proteins belong to a large family: Identification of proteins of the U6 as well as the U1, U2, U4 and U5 snRNPs. *EMBO Journal* **14**, 2089–2098 (1995).

66. Vidal, V. P., Verdone, L., Mayes, A. E. & Beggs, J. D. Characterization of U6 snRNA-protein interactions. *Rna* **5**, 1470–1481 (1999).

67. Pomeranz Krummel, D. A., Oubridge, C., Leung, A. K., Li, J. & Nagai, K. Crystal structure of human spliceosomal U1 snRNP at 5.5 resolution. *Nature* **458**, 475–480 (2009).

68. Weber, G., Trowitzsch, S., Kastner, B., Lührmann, R. & Wahl, M. C. Functional organization of the Sm core in the crystal structure of human U1 snRNP. *EMBO Journal* **29**, 4172–4184 (2010).

69. Li, X. *et al.* CryoEM structure of Saccharomyces cerevisiae U1 snRNP offers insight into alternative splicing. *Nature Communications* **8** (2017). URL `http://dx.doi.org/10.1038/s41467-017-01241-9`.

70. De Bortoli, F. *et al.* Increased versatility despite reduced molecular complexity: evolution, structure and function of metazoan splicing factor PRPF39. *Nucleic acids research* **47**, 5867–5879 (2019). URL `http://www.ncbi.nlm.nih.gov/pubmed/30949712http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6582350`.

71. Agafonov, D. E. *et al.* Semiquantitative Proteomic Analysis of the Human Spliceosome via a Novel Two-Dimensional Gel Electrophoresis Method. *Molecular and Cellular Biology* **31**, 2667–2682 (2011).

72. Krämer, A., Grüter, P., Gröning, K. & Kastner, B. Combined biochemical and electron microscopic analyses reveal the architecture of the mammalian U2 snRNP. *Journal of Cell Biology* **145**, 1355–1368 (1999). URL `http://www.ncbi.nlm.nih.gov/pubmed/10385517http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2133165`.

73. Galej, W. P., Oubridge, C., Newman, A. J. & Nagai, K. Crystal structure of Prp8 reveals active site cavity of the spliceosome. *Nature* **493**, 638–643 (2013).

74. Jurica, M. S. & Moore, M. J. Pre-mRNA splicing: awash in a sea of proteins. *Molecular cell* **12**, 5–14 (2003). URL `https://linkinghub.elsevier.com/retrieve/pii/S1097276503002703http://www.ncbi.nlm.nih.gov/pubmed/12887888`.

75. Michaud, S. & Reed, R. An ATP-independent complex commits pre-mRNA to the mammalian spliceosome assembly pathway. *Genes and Development* **5**, 2534–2546 (1991).

76. Zamore, P. D. & Green, M. R. Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. *Proc Natl*

Acad Sci U S A **86**, 9243–9247 (1989). URL `http://www.ncbi.nlm.nih.gov/pubmed/2531895`.

77. Bennett, M., Michaud, S., Kingston, J. & Reed, R. Protein components specifically associated with prespliceosome and spliceosome complexes. *Genes and Development* **6**, 1986–2000 (1992).

78. Berglund, J. A., Abovich, N. & Rosbash, M. A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. *Genes and Development* **12**, 858–867 (1998). URL `http://www.ncbi.nlm.nih.gov/pubmed/9512519http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC316625`.

79. Liu, Z. *et al.* Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science* **294**, 1098–1102 (2001). URL `http://www.ncbi.nlm.nih.gov/pubmed/11691992http://www.sciencemag.org/cgi/doi/10.1126/science.1064719`.

80. Thickman, K. R., Sickmier, E. A. & Kielkopf, C. L. Alternative Conformations at the RNA-binding Surface of the N-terminal U2AF65 RNA Recognition Motif. *Journal of Molecular Biology* **366**, 703–710 (2007). URL `https://linkinghub.elsevier.com/retrieve/pii/S0022283606016366`.

81. Berglund, J. A., Chua, K., Abovich, N., Reed, R. & Rosbash, M. The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell* **89**, 781–787 (1997). URL `http://www.ncbi.nlm.nih.gov/pubmed/9182766`.

82. Jacewicz, A., Chico, L., Smith, P., Schwer, B. & Shuman, S. 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**, 401–414 (2015).

83. Chang, J. H. *et al.* Crystal structure of the eIF4A-PDCD4 complex. *Proceedings of the National Academy of Sciences* **106**, 3148–3153 (2009). URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2651268/pdf/zpq3148.pdf`.

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

85. Kistler, A. L. Deletion of MUD2, the yeast homolog of U2AF65, can bypass the requirement for Sub2, an essential spliceosomal ATPase. *Genes & Development* **15**, 42–49 (2001). URL `http://www.genesdev.org/cgi/doi/10.1101/gad.851301`.

86. Hong, W. *et al.* Association of U2 snRNP with the spliceosomal complex E. *Nucleic Acids Research* **25**, 354–361 (1997). URL `http://www.ncbi.nlm.nih.gov/pubmed/9016565http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC146436`.

87. Das, R., Zhou, Z. & Reed, R. Functional association of U2 snRNP with the ATP-independent splicecosomal complex E. *Molecular Cell* **5**, 779–787 (2000).

88. Gozani, O., Potashkin, J. & Reed, R. A Potential Role for U2AF-SAP 155 Interactions in Recruiting U2 snRNP to the Branch Site. *Molecular and Cellular Biology* **18**, 4752–4760 (1998). URL `http://www.ncbi.nlm.nih.gov/pubmed/9671485http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC109061https://mcb.asm.org/content/18/8/4752`.

89. Boesler, C. *et al.* A spliceosome intermediate with loosely associated tri-snRNP accumulates in the absence of Prp28 ATPase activity. *Nature Communications* **7** (2016).

90. Staley, J. P. & Guthrie, C. An RNA switch at the 5' splice site requires ATP and the DEAD box protein Prp28p. *Molecular Cell* **3**, 55–64 (1999).

91. Bertram, K. *et al.* Cryo-EM Structure of a Pre-catalytic Human Spliceosome Primed for Activation. *Cell* **170**, 701–713.e11 (2017). URL `https://www.sciencedirect.com/science/article/pii/S0092867417308188?via{%}3Dihub`.

92. Laggerbauer, B., Achsel, T. & Lührmann, R. The human U5-200kD DEXH-box protein unwinds U4/U6 RNA duplices in vitro. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 4188–4192 (1998).

93. Townsend, C. *et al.* Mechanism of protein-guided folding of the active site U2/U6 RNA during spliceosome activation. *Science* **370** (2020).

94. Fica, S. M. *et al.* RNA catalyses nuclear pre-mRNA splicing. *Nature* **503**, 229–234 (2013). URL `http://dx.doi.org/10.1038/nature12734`.

95. Fica, S. M., Mefford, M. A., Piccirilli, J. A. & Staley, J. P. Evidence for a group II intron-like catalytic triplex in the spliceosome. *Nature Structural and Molecular Biology* **21**, 464–471 (2014). URL `http://www.ncbi.nlm.nih.gov/pubmed/24747940http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4257784`.

96. Rauhut, R. *et al.* Molecular architecture of the Saccharomyces cerevisiae activated spliceosome. *Science* **353**, 1399–1405 (2016).

97. Yan, C., Wan, R., Bai, R., Huang, G. & Shi, Y. Structure of a yeast activated spliceosome at 3.5 Å resolution. *Science* **353**, 904–912 (2016). URL `http://www.ncbi.nlm.nih.gov/pubmed/27445306http://www.sciencemag.org/cgi/doi/10.1126/science.aag0291`.

98. Zhang, X. *et al.* Structure of the human activated spliceosome in three conformational states. *Cell Research* **28**, 307–322 (2018). URL `http://dx.doi.org/10.1038/cr.2018.14`.

99. Horowitz, D. S. The mechanism of the second step of pre-mRNA splicing. *Wiley Interdisciplinary Reviews: RNA* **3**, 331–350 (2012).

100. Charenton, C., Wilkinson, M. E. & Nagai, K. Mechanism of 5' splice site transfer for human spliceosome activation. *Science* **364**, 362–367 (2019).

101. Zhan, X., Yan, C., Zhang, X., Lei, J. & Shi, Y. Structure of a human catalytic step I spliceosome. *Science* **359**, 537–545 (2018). URL `http://science.sciencemag.org/content/sci/359/6375/537.full.pdf`.

102. Cordin, O. & Beggs, J. D. RNA helicases in splicing. *RNA Biology* **10**, 83–95 (2013).

103. Fairman-Williams, M. E., Guenther, U.-P. & Jankowsky, E. SF1 and SF2 helicases: family matters. *Current opinion in structural biology* **20**, 313–24 (2010). URL `http://www.ncbi.nlm.nih.gov/pubmed/20456941`.

104. Perriman, R. & Manuel, A. ATP can be dispensable for prespliceosome formation in yeast. *Genes and Development* **14**, 97–107 (2000).

105. Wang, Q., Zhang, L., Lynn, B. & Rymond, B. C. A BBP-Mud2p heterodimer mediates branchpoint recognition and influences splicing substrate abundance in budding yeast. *Nucleic Acids Research* **36**, 2787–2798 (2008).

106. Shen, H. *et al.* Distinct activities of the DExD/H-box splicing factor hUAP56 facilitate stepwise assembly of the spliceosome. *Genes and Development* **22**, 1796–1803 (2008).

107. Sengoku, T., Nureki, O., Nakamura, A., Kobayashi, S. & Yokoyama, S. Structural Basis for RNA Unwinding by the DEAD-Box Protein Drosophila Vasa. *Cell* **125**, 287–300 (2006).

108. Mallam, A. L., Del Campo, M., Gilman, B., Sidote, D. J. & Lambowitz, A. M. Structural basis for RNA-duplex recognition and unwinding by the DEAD-box helicase Mss116p. *Nature* **490**, 121–125 (2012). URL `http://dx.doi.org/10.1038/nature11402`.

109. Agafonov, D. E. *et al.* Molecular architecture of the human U4/U6.U5 tri-snRNP. *Science* **351**, 1416–1420 (2016). URL `http://www.ncbi.nlm.nih.gov/pubmed/26912367http://www.sciencemag.org/cgi/doi/10.1126/science.aad2085`.

110. Bai, R. *et al.* Mechanism of spliceosome remodeling by the ATPase/helicase Prp2 and its coactivator Spp2. *Science* **371** (2021).

111. Zhang, L. *et al.* Structural evidence for consecutive Hel308-like modules in the spliceosomal ATPase Brr2. *Nature Structural & Molecular Biology* **16**, 731–739 (2009). URL `http://www.nature.com/articles/nsmb.1625`.

112. Liu, H.-L. & Cheng, S.-C. The Interaction of Prp2 with a Defined Region of the Intron Is Required for the First Splicing Reaction. *Molecular and Cellular Biology* **32**, 5056–5066 (2012). URL `https://journals.asm.org/doi/10.1128/MCB.01109-12`.

113. Hamann, F. *et al.* Structural analysis of the intrinsically disordered splicing factor Spp2 and its binding to the DEAH-box ATPase Prp2. *Proceedings of the National Academy of Sciences* **117**, 2948–2956 (2020). URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1907960117`.

114. Bertram, K. *et al.* Cryo-EM structure of a human spliceosome activated for step 2 of splicing. *Nature* **542**, 318–323 (2017). URL `http://www.ncbi.nlm.nih.gov/pubmed/28076346http://www.nature.com/articles/nature21079`.

115. Zhang, X. *et al.* An Atomic Structure of the Human Spliceosome. *Cell* **169**, 918–929.e14 (2017). URL `http://dx.doi.org/10.1016/j.cell.2017.04.033`.

116. Semlow, D. R., Blanco, M. R., Walter, N. G. & Staley, J. P. Spliceosomal DEAH-Box ATPases Remodel Pre-mRNA to Activate Alternative Splice Sites. *Cell* **164**, 985–998 (2016). URL `http://dx.doi.org/10.1016/j.cell.2016.01.025`.

117. Fica, S. M. *et al.* Structure of a spliceosome remodelled for exon ligation. *Nature* **542**, 377–380 (2017).

118. Wilkinson, M. E. *et al.* Postcatalytic spliceosome structure reveals mechanism of 3âĂšâĂŞsplice site selection. *Science* **358**, 1283–1288 (2017). URL `http://www.ncbi.nlm.nih.gov/pubmed/29146871http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5808836http://www.sciencemag.org/lookup/doi/10.1126/science.aar3729`.

119. Toroney, R., Nielsen, K. H. & Staley, J. P. Termination of pre-mRNA splicing requires that the ATPase and RNA unwindase Prp43p acts on the catalytic snRNA U6. *Genes & development* **33**, 1555–1574 (2019). URL `http://www.ncbi.nlm.nih.gov/pubmed/31558568http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6824469`.

120. Wan, R., Yan, C., Bai, R., Lei, J. & Shi, Y. Structure of an Intron Lariat Spliceosome from Saccharomyces cerevisiae. *Cell* **171**, 120–132.e12 (2017). URL `https://doi.org/10.1016/j.cell.2017.08.029https://linkinghub.elsevier.com/retrieve/pii/S0092867417309546`.

121. Zhang, X. *et al.* Structures of the human spliceosomes before and after release of the ligated exon. *Cell Research* **29**, 274–285 (2019). URL `http://dx.doi.org/10.1038/s41422-019-0143-x`.

122. Boon, K.-L. *et al.* Yeast Ntr1/Spp382 Mediates Prp43 Function in Postspliceosomes. *Molecular and Cellular Biology* **26**, 6016–6023 (2006). URL `https://journals.asm.org/doi/10.1128/MCB.02347-05`.

123. Tanaka, N., Aronova, A. & Schwer, B. Ntr1 activates the Prp43 helicase to trigger release of lariat-intron from the spliceosome. *Genes & Development* **21**, 2312–2325 (2007). URL `http://www.genesdev.org/cgi/doi/10.1101/gad.1580507`.

124. Studer, M. K., Ivanović, L., Weber, M. E., Marti, S. & Jonas, S. Structural basis for DEAH-helicase activation by G-patch proteins. *Proceedings of the National Academy of Sciences* **117**, 7159–7170 (2020). URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1913880117`.

125. Burgess, S. M. & Guthrie, C. A mechanism to enhance mRNA splicing fidelity: The RNA-dependent ATPase Prp16 governs usage of a discard pathway for aberrant lariat intermediates. *Cell* **73**, 1377–1391 (1993).

126. Koodathingal, P. & Staley, J. P. Splicing fidelity. *RNA Biology* **10**, 1073–1079 (2013).

127. Mayas, R., Maita, H. & Staley, J. Exon ligation is proofread by the DExD/H-box ATPase Prp22p. *Nature structural & molecular biology* **13**, 482–490 (2006). URL `http://www.nature.com/nsmb/journal/v13/n6/abs/nsmb1093.html`.

128. Raker, V. A., Plessel, G. & Lührmann, R. The snRNP core assembly pathway: identification of stable core protein heteromeric complexes and an snRNP subcore particle in vitro. *The EMBO journal* **15**, 2256–69 (1996). URL `http://www.ncbi.nlm.nih.gov/pubmed/8641291http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC450151`.

129. Seipelt, R. L., Zheng, B., Asuru, A. & Rymond, B. C. U1 snRNA is cleaved by RNase III and processed through an Sm site-dependent pathway. *Nucleic acids research* **27**, 587–95 (1999). URL `http://www.ncbi.nlm.nih.gov/pubmed/9862984http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC148219`.

130. Fischer, U. & Luhrmann, R. An essential signaling role for the m3G cap in the transport of U1 snRNP to the nucleus. *Science* **249**, 786–790 (1990). URL `https://www.sciencemag.org/lookup/doi/10.1126/science.2143847`.

131. Fischer, U., Sumpter, V., Sekine, M., Satoh, T. & Lührmann, R. Nucleocytoplasmic transport of U snRNPs: definition of a nuclear location signal in

the Sm core domain that binds a transport receptor independently of the m3G cap. *The EMBO Journal* **12**, 573–583 (1993). URL `https://onlinelibrary.wiley.com/doi/10.1002/j.1460-2075.1993.tb05689.x`.

132. Hetzer, M. & Mattaj, I. W. An ATP-dependent, Ran-independent mechanism for nuclear import of the U1A and U2B' spliceosome proteins. *Journal of Cell Biology* **148**, 293–303 (2000).

133. Nesic, D., Tanackovic, G. & Krämer, A. A role for Cajal bodies in the final steps of U2 snRNP biogenesis. *Journal of Cell Science* **117**, 4423–4433 (2004).

134. Brosi, R., Hauri, H. P. & Kramer, A. Separation of splicing factor SF3 into two components and purification of SF3a activity. *Journal of Biological Chemistry* **268**, 17640–17646 (1993). URL `http://www.ncbi.nlm.nih.gov/pubmed/8349644`.

135. Ares, M. U2 RNA from yeast is unexpectedly large and contains homology to vertebrate U4, U5, and U6 small nuclear RNAs. *Cell* **47**, 49–59 (1986). URL `http://www.ncbi.nlm.nih.gov/pubmed/3530502`.

136. Igel, A. H. & Ares, M. Internal sequences that distinguish yeast from metazoan U2 snRNA are unnecessary for pre-mRNA splicing. *Nature* **334**, 450–453 (1988). URL `http://www.nature.com/doifinder/10.1038/334450a0`.

137. Shuster, E. O. & Guthrie, C. Human U2 snRNA can function in pre-mRNA splicing in yeast. *Nature* **345**, 270–273 (1990). URL `http://www.nature.com/articles/345270a0`.

138. Ares, M. & Igel, A. H. Lethal and temperature-sensitive mutations and their suppressors identify an essential structural element in U2 small nuclear RNA. *Genes and Development* **4**, 2132–2145 (1990).

139. Hilliker, A. K., Mefford, M. A. & Staley, J. P. U2 toggles iteratively between the stem IIa and stem IIc conformations to promote pre-mRNA splicing. *Genes and Development* **21**, 821–834 (2007).

140. Perriman, R. J. & Ares, M. Rearrangement of competing U2 RNA helices within the spliceosome promotes multiple steps in splicing. *Genes and Development* **21**, 811–820 (2007).

141. Rodgers, M. L. *et al.* Conformational dynamics of stem II of the U2 snRNA. *Rna* **22**, 225–236 (2016).

142. Plaschka, C., Lin, P. C., Charenton, C. & Nagai, K. Prespliceosome structure provides insights into spliceosome assembly and regulation. *Nature* **559**, 419–422 (2018).

143. Wan, R., Bai, R., Yan, C., Lei, J. & Shi, Y. Structures of the Catalytically Activated Yeast Spliceosome Reveal the Mechanism of Branching. *Cell* **177**, 339–351.e13 (2019). URL `https://doi.org/10.1016/j.cell.2019.02.006`.

144. Krämer, A. & Utans, U. Three protein factors (SF1, SF3 and U2AF) function in pre-splicing complex formation in addition to snRNPs. *The EMBO journal* **10**, 1503–9 (1991). URL `http://doi.wiley.com/10.1111/j.1471-4159.1991.tb06421.xhttp://www.ncbi.nlm.nih.gov/pubmed/1827409http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC452814`.

145. Lin, P. C. & Xu, R. M. Structure and assembly of the SF3a splicing factor complex of U2 snRNP. *EMBO Journal* **31**, 1579–1590 (2012). URL `http://www.ncbi.nlm.nih.gov/pubmed/22314233http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3321192http://emboj.embopress.org/cgi/doi/10.1038/emboj.2012.7`.

146. Plaschka, C., Lin, P. C. & Nagai, K. Structure of a pre-catalytic spliceosome. *Nature* **546**, 617–621 (2017). URL `http://www.ncbi.nlm.nih.gov/pubmed/28530653http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5503131http://www.nature.com/doifinder/10.1038/nature22799`.

147. Quesada, V. *et al.* Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nature genetics* **44**, 47–52 (2011). URL `http://www.ncbi.nlm.nih.gov/pubmed/22158541`.

148. Ebert, B. & Bernard, O. A. Mutations in RNA Splicing Machinery in Human Cancers. *New England Journal of Medicine* **365**, 2534–2535 (2011).

149. Papaemmanuil, E. *et al.* Somatic SF3B1 Mutation in Myelodysplasia with Ring Sideroblasts . *New England Journal of Medicine* **365**, 1384–1395 (2011). URL `http://www.nejm.org/doi/abs/10.1056/NEJMoa1103283`.

150. Papaemmanuil, E. *et al.* Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* **122**, 3616–27; quiz 3699 (2013). URL `http://www.ncbi.nlm.nih.gov/pubmed/24030381http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3837510`.

151. Yoshida, K. *et al.* Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64–69 (2011).

152. Maji, D., Grossfield, A. & Kielkopf, C. L. Structures of SF3b1 reveal a dynamic Achilles heel of spliceosome assembly: Implications for cancer-associated abnormalities and drug discovery. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1862**, 194440 (2019). URL `https://doi.org/10.1016/j.bbagrm.2019.194440`.

153. Darman, R. B. *et al.* Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. *Cell Reports* **13**, 1033–1045 (2015). URL `http://www.ncbi.nlm.nih.gov/pubmed/26565915http://linkinghub.elsevier.com/retrieve/pii/S2211124715010785`.

154. Alsafadi, S. *et al.* Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. *Nature Communications* **7**, 10615 (2016). URL `http://www.ncbi.nlm.nih.gov/pubmed/26842708http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4743009`.

155. Tang, Q. *et al.* SF3B1/Hsh155 HEAT motif mutations affect interaction with the spliceosomal ATPase Prp5, resulting in altered branch site selectivity in pre-mRNA splicing. *Genes and Development* **30**, 2710–2723 (2016). URL `http://www.ncbi.nlm.nih.gov/pubmed/28087715http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5238730http://genesdev.cshlp.org/lookup/doi/10.1101/gad.291872.116`.

156. Liu, Z. *et al.* Pan-cancer analysis identifies mutations in SUGP1 that recapitulate mutant SF3B1 splicing dysregulation. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 10305–10312 (2020).

157. Jenkins, J. L. & Kielkopf, C. L. Splicing Factor Mutations in Myelodysplasias: Insights from Spliceosome Structures. *Trends in Genetics* **33**, 336–348 (2017). URL `https://linkinghub.elsevier.com/retrieve/pii/S0168952517300355`.

158. Yan, D. *et al.* CUS2, a Yeast Homolog of Human Tat-SF1, Rescues Function of Misfolded U2 through an Unusual RNA Recognition Motif. *Molecular and Cellular Biology* **18**, 5000–5009 (1998).

159. Zavanelli, M. I. & Ares, M. Efficient association of U2 snRNPs with pre-mRNA requires an essential U2 RNA structural element. *Genes and Development* **5**, 2521–2533 (1991).

160. Loerch, S. & Kielkopf, C. L. Unmasking the U2AF homology motif family: A bona fide protein-protein interaction motif in disguise. *RNA* **22**, 1795–1807 (2016).

161. Loerch, S. *et al.* The pre-mRNA splicing and transcription factor Tat-SF1 is a functional partner of the spliceosome SF3b1 subunit via a U2AF homology motif interface. *Journal of Biological Chemistry* **294**, 2892–2902 (2019).

162. Talkish, J. *et al.* Cus2 enforces the first ATP-dependent step of splicing by binding to yeast SF3b1 through a UHMâĂŞUlm interaction. *Rna* **25**, 1020–1037 (2019).

163. Shao, W., Kim, H.-S., Cao, Y., Xu, Y.-Z. & Query, C. C. A U1-U2 snRNP Interaction Network during Intron Definition. *Molecular and Cellular Biology* **32**, 470–478 (2012).

164. Zhang, Z. M. *et al.* Crystal Structure Of Prp5P Reveals Interdomain Interactions That Impact Spliceosome Assembly. *Cell Reports* **5**, 1269–1278 (2013). URL `http://dx.doi.org/10.1016/j.celrep.2013.10.047`.

165. Beier, D. H. *et al.* Dynamics of the DEAD-box ATPase Prp5 RecA-like domains provide a conformational switch during spliceosome assembly. *Nucleic acids research* **47**, 10842–10851 (2019). URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6846040/?report=classic`.

166. Yan, C. *et al.* Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science* **349**, 1182–1191 (2015).

167. Bai, R., Wan, R., Yan, C., Lei, J. & Shi, Y. Structures of the fully assembled Saccharomyces cerevisiae spliceosome before activation. *Science* **360**, 1423–1429 (2018).

168. Wan, R., Yan, C., Bai, R., Huang, G. & Shi, Y. Structure of a yeast catalytic step i spliceosome at 3.4 Å resolution. *Science* **353**, 895–904 (2016). URL `http://www.sciencemag.org/cgi/doi/10.1126/science.aag2235`.

169. Galej, W. P. *et al.* Cryo-EM structure of the spliceosome immediately after branching. *Nature* **537**, 197–201 (2016).

170. Bertram, K. *et al.* Structural Insights into the Roles of Metazoan-Specific Splicing Factors in the Human Step 1 Spliceosome. *Molecular Cell* **80**, 127–139.e6 (2020).

171. Fica, S. M., Oubridge, C., Wilkinson, M. E., Newman, A. J. & Nagai, K. A human postcatalytic spliceosome structure reveals essential roles of metazoan factors for exon ligation. *Science* **363**, 710–714 (2019).

172. Li, X. *et al.* A unified mechanism for intron and exon definition and back-splicing. *Nature* **573**, 375–380 (2019). URL `http://dx.doi.org/10.1038/s41586-019-1523-6`.

173. Cretu, C. *et al.* Structural Basis of Splicing Modulation by Antitumor Macrolide Compounds. *Molecular Cell* **70**, 265–273.e8 (2018).

174. Zavanelli, M. I., Britton, J. S., Igel, A. H. & Ares, M. Mutations in an essential U2 small nuclear RNA structure cause cold-sensitive U2 small nuclear ribonucleoprotein function by favoring competing alternative U2 RNA structures. *Molecular and Cellular Biology* **14**, 1689–1697 (1994). URL `http://www.ncbi.nlm.nih.gov/pubmed/8114704http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC358527`.

175. Yan, D. & Ares, M. Invariant U2 RNA sequences bordering the branchpoint recognition region are essential for interaction with yeast SF3a and SF3b subunits. *Molecular and Cellular Biology* **16**, 818–828 (1996).

176. Will, C. L. & Lührmann, R. Spliceosomal UsnRNP biogenesis, structure and function. *Current Opinion in Cell Biology* **13**, 290–301 (2001).

177. Paillart, J. C., Marquet, R., Skripkin, E., Ehresmann, C. & Ehresmann, B. Dimerization of retroviral genomic RNAs: Structural and functional implications. *Biochimie* **78**, 639–653 (1996).

178. Fairman, M. E. *et al.* Protein displacement by DExH/D "RNA helicases" without duplex unwinding. *Science (New York, N.Y.)* **304**, 730–4 (2004). URL `https://www.sciencemag.org/lookup/doi/10.1126/science.1095596http://www.ncbi.nlm.nih.gov/pubmed/15118161`.

179. Kennedy, S. D., Bauer, W. J., Wang, W. & Kielkopf, C. L. Dynamic stacking of an expected branch point adenosine in duplexes containing pseudouridine-modified or unmodified U2 snRNA sites. *Biochemical and Biophysical Research Communications* **511**, 416–421 (2019). URL `https://linkinghub.elsevier.com/retrieve/pii/S0006291X19302682`.

180. Hansen, S. R., Nikolai, B. J., Spreacker, P. J., Carrocci, T. J. & Hoskins, A. A. Chemical Inhibition of Pre-mRNA Splicing in Living Saccharomyces cerevisiae. *Cell Chemical Biology* **26**, 443–448.e3 (2019). URL `https://doi.org/10.1016/j.chembiol.2018.11.008`.

181. Bedford, M. T. & Leder, P. The FF domain: A novel motif that often accompanies WW domains. *Trends in Biochemical Sciences* **24**, 264–265 (1999).

182. Ester, C. & Uetz, P. The FF domains of yeast U1 snRNP protein Prp40 mediate interactions with Luc7 and Snu71. *BMC Biochemistry* **9**, 1–11 (2008).

183. Gozani, O., Feld, R. & Reed, R. 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 and Development* **10**, 233–243 (1996). URL `http://www.genesdev.org/cgi/doi/10.1101/gad.10.2.233`.

184. Chang, J., Schwer, B. & Shuman, S. Structure-function analysis and genetic interactions of the yeast branchpoint binding protein Msl5. *Nucleic Acids Research* **40**, 4539–4552 (2012).

185. Cretu, C. *et al.* Structural basis of intron selection by U2 snRNP in the presence of covalent inhibitors. *Nature Communications* **12**, 4491 (2021). URL `http://www.nature.com/articles/s41467-021-24741-1`.

186. Kastner, B. *et al.* GraFix: sample preparation for single-particle electron cry-omicroscopy. *Nature Methods* **5**, 53–55 (2008). URL `http://www.nature.com/articles/nmeth1139`.

187. Singh, K. *et al.* Discovery of a Regulatory Subunit of the Yeast Fatty Acid Synthase. *Cell* **180**, 1130–1143.e20 (2020). URL `https://linkinghub.elsevier.com/retrieve/pii/S0092867420302117`.

188. Maeda, S. *et al.* Development of an antibody fragment that stabilizes GPCR/G-protein complexes. *Nature Communications* **9**, 3712 (2018). URL `http://www.nature.com/articles/s41467-018-06002-w`.

189. Punjani, A. & Fleet, D. J. 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. *Journal of Structural Biology* **213**, 107702 (2021). URL `https://linkinghub.elsevier.com/retrieve/pii/S1047847721000071`.

190. Wu, Z., Zhang, S., Wang, W. L., Dong, Y. & Mao, Y. Hidden dynamics of human proteasome autoregulation during protein degradation. *bioRxiv* 2020.12.22.423932 (2021). URL `http://biorxiv.org/content/early/2021/08/11/2020.12.22.423932.abstract`.

191. Wilkinson, M. E., Fica, S. M., Galej, W. P. & Nagai, K. Structural basis for conformational equilibrium of the catalytic spliceosome. *Molecular Cell* **81**, 1439–1452.e9 (2021). URL `https://linkinghub.elsevier.com/retrieve/pii/S1097276521001313`.

192. Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature* **587**, 157–161 (2020). URL `http://www.nature.com/articles/s41586-020-2833-4`.

193. Nakane, T. *et al.* Single-particle cryo-EM at atomic resolution. *Nature* **587**, 152–156 (2020). URL `http://www.nature.com/articles/s41586-020-2829-0`.

194. Naydenova, K., Jia, P. & Russo, C. J. Cryo-EM with subâĂŞ1 Å specimen movement. *Science* **370**, 223–226 (2020). URL `https://www.science.org/doi/10.1126/science.abb7927`.

195. Tegunov, D., Xue, L., Dienemann, C., Cramer, P. & Mahamid, J. Multi-particle cryo-EM refinement with M visualizes ribosome-antibiotic complex at 3.5 Å in cells. *Nature Methods* **18**, 186–193 (2021). URL http://www.nature.com/articles/s41592-020-01054-7.

196. Rigort, A. *et al.* Micromachining tools and correlative approaches for cellular cryo-electron tomography. *Journal of Structural Biology* **172**, 169–179 (2010). URL https://linkinghub.elsevier.com/retrieve/pii/S1047847710000584.

197. Caruthers, J. M., Johnson, E. R. & McKay, D. B. Crystal structure of yeast initiation factor 4A, a DEAD-box RNA helicase. *Proceedings of the National Academy of Sciences* **97**, 13080–13085 (2000). URL http://www.pnas.org/cgi/doi/10.1073/pnas.97.24.13080.

198. Schütz, P. *et al.* Crystal structure of the yeast eIF4A-eIF4G complex: An RNA-helicase controlled by protein - Protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 9564–9569 (2008).