

# Increased versatility despite reduced molecular complexity: evolution, structure and function of metazoan splicing factor PRPF39

Francesca De Bortoli<sup>1</sup>, Alexander Neumann<sup>1</sup>, Ana Kotte<sup>1</sup>, Bernd Timmermann<sup>2</sup>, Thomas Schüler<sup>3</sup>, Markus C. Wahl<sup>4,5</sup>, Bernhard Loll<sup>4,\*</sup> and Florian Heyd<sup>1,\*</sup>

<sup>1</sup>Institut für Chemie und Biochemie, RNA Biochemie, Freie Universität Berlin, Takustr. 6, 14195 Berlin, Germany, <sup>2</sup>Sequencing Core Facility, Max-Planck-Institute for Molecular Genetics, Ihnestraße 63–73, Berlin 14195, Germany, <sup>3</sup>Institute of Molecular and Clinical Immunology, Medical Faculty, Otto-von-Guericke University, 39120 Magdeburg, Germany, <sup>4</sup>Institut für Chemie und Biochemie, Strukturbiochemie, Freie Universität Berlin, Takustr. 6, 14195 Berlin, Germany and <sup>5</sup>Helmholtz-Zentrum Berlin für Materialien und Energie, Macromolecular Crystallography, Albert-Einstein-Straße 15, D-12489 Berlin, Germany

Received February 06, 2019; Revised March 20, 2019; Editorial Decision March 21, 2019; Accepted March 25, 2019

## ABSTRACT

**In the yeast U1 snRNP the Prp39/Prp42 heterodimer is essential for early steps of spliceosome assembly. In metazoans no Prp42 ortholog exists, raising the question how the heterodimer is functionally substituted. Here we present the crystal structure of murine PRPF39, which forms a homodimer. Structure-guided point mutations disrupt dimer formation and inhibit splicing, manifesting the homodimer as functional unit. PRPF39 expression is controlled by NMD-inducing alternative splicing in mice and human, suggesting a role in adapting splicing efficiency to cell type specific requirements. A phylogenetic analysis reveals coevolution of shortened U1 snRNA and the absence of Prp42, which correlates with overall splicing complexity in different fungi. While current models correlate the diversity of spliceosomal proteins with splicing complexity, our study highlights a contrary case. We find that organisms with higher splicing complexity have substituted the Prp39/Prp42 heterodimer with a PRPF39 homodimer.**

## INTRODUCTION

Splicing is an essential step in precursor messenger RNA (pre-mRNA) processing. It is catalyzed by the spliceosome, a multi-megadalton protein RNA complex (1). This elaborate molecular machine consists of five core components, the U1, U2, U4, U5 and U6 snRNPs, and multiple auxiliary proteins, which are assembled in a stepwise manner for

each splicing event on a pre-mRNA substrate. It is crucial that splice site (ss) recognition is consistently accurate, as a single mistake can result in the production of a nonfunctional protein and in some cases in disease. For 5' ss recognition in the early steps of splicing, the U1 snRNP plays a critical role. During this step a single stranded region of the U1 snRNA base pairs with the 5' ss of the pre-mRNA (2).

Structural work on the spliceosomal machinery has mainly been conducted in the human or *Saccharomyces cerevisiae* (referred to as yeast in the following) systems. The yeast system more readily yields stable complexes for structural analysis and serves as a good model for spliceosomes from higher eukaryotes. The human spliceosome is more complex than the spliceosome in yeast, including around 80 additional, mostly non-snRNP, proteins (3,4), that are considered to yield increased functional versatility such as alternative splicing. However, the U1 snRNP seems to represent an exception. The yeast U1 snRNA alone is already ~3.5 times larger than its human counterpart (5,6). Purified human U1 snRNP contains merely ten proteins (seven Sm proteins, hU1-70K, hU1A, and hU1C). In contrast, seven additional, stably associated proteins (yLuc7, yNam8, yPrp39, yPrp40, yPrp42, ySnu56 and ySnu71) can be found in purified yeast U1 snRNPs. Among the yeast U1 proteins, yPrp42 and ySnu56 have no known human orthologs (7,8).

Earlier studies in yeast showed that yPrp39 associates with the U1 snRNP and is necessary for pre-mRNA splicing (9). A cryo-electron microscopy (cryo-EM) study of the yeast U1 snRNP (10) showed that the core components of the U1 snRNP are connected to its auxiliary proteins through a yPrp39/yPrp42 heterodimer. Both yPrp39 and yPrp42 consist mainly of half- $\alpha$ -tetratricopeptide repeats (HAT repeats). HAT-repeat proteins are mostly involved

\*To whom correspondence should be addressed. Tel: +493083857348; Fax: +4930 838457348; Email: loll@chemie.fu-berlin.de  
Correspondence may also be addressed to Florian Heyd. Tel: +493083862938; Fax: +4930 838457348; Email: florian.heyd@chemie.fu-berlin.de

in protein-protein and protein-RNA interactions and act as scaffolding proteins. They participate in various RNA processing pathways, prominent examples being SART3, which is involved in splicing and translesion DNA synthesis (11) and RNA14, which is important for mRNA 3'-end maturation (12). Furthermore, a cryo-EM structure of a yeast pre-catalytic pre-B complex (13) revealed that the contact between the U1 and U2 snRNP is largely mediated by the yPrp42/yPrp39 heterodimer. In the pre-B complex, yPrp39 contacts the U2-specific protein yLea1 (human homolog U2A') via its N-terminal HAT domain (HAT-NTD), and a positively charged groove in the HAT-NTD of yPrp42 accommodates the U2 snRNA. Moreover, a yeast pre-spliceosome structure (14) provided novel insights into the early spliceosome assembly steps, showing two interfaces between the U1 snRNP and the U2 snRNP. The more stable interaction is mediated by yPrp39 and yLea1. These findings indicate that the yPrp39/yPrp42 heterodimer is crucial for the relative spatial positioning of the U1 and U2 snRNPs during early spliceosome assembly. Together, these studies revealed a central role of the yeast yPrp39/yPrp42 heterodimer as a crucial scaffolding complex involved in protein-protein and protein-RNA interactions at different stages of the splicing cycle.

Notably, in human there is no Prp42 ortholog, leading to the question how the yPrp39/yPrp42-based protein/RNA interactions are established in the human spliceosome. Furthermore, surprisingly little is known about Prp39 in higher eukaryotes, although the protein has been demonstrated to be essential in diverse human cell lines in independent CRISPR-screens (15–17) indicating a similar importance in metazoan species as in yeast. A first indication how the loss of Prp42 may be compensated in mammals came from a recent study. Negative stain EM data revealed the shape of hPRPF39 to resemble the yPrp39/yPrp42 heterodimer and co-immunoprecipitation experiments suggest that human PRPF39 (hPRPF39) forms a homodimer (10).

Murine PRPF39 (mPRPF39, 94% amino acid identity with hPRPF39; Supplementary Figure S1) initially came to our attention because it is alternatively spliced in murine naïve versus memory T-cells. Given the essential role of Prp39 in spliceosome assembly and to further address the question how the loss of Prp42 is compensated in metazoan, we solved the crystal structure of mPRPF39. Structural and biophysical data reveal mPRPF39 to be arranged as a homodimer that can fulfill the function of the yPrp42/yPrp39 heterodimer. Based on our structure, we investigated the effect of point mutations that interfere with dimerization and we show that monomeric mPRPF39 reduces splicing efficiency *in vitro*. Furthermore, a phylogenetic analysis suggests a coevolutionary reduction in U1 snRNA length and the loss Prp42 homologs. Organisms that use a PRPF39 homodimer show higher splicing complexity, providing an example for how reduced diversity of spliceosomal proteins can increase splicing versatility.

## MATERIALS AND METHODS

### Cloning

Open reading frames (ORFs) encoding mPRPF39 were amplified from a murine cDNA sample and cloned into

the pGEX-6P-1 (or pCMV-GFP-N3 or pCMV-Flag-N3) vector using the restriction sites *Bam*HI (*Xho*I) at the 5'-end and *Sal*I (*Bam*HI) at the 3'-end. Point mutations were introduced according to the QuikChange protocol (Agilent). The pGEX-6P-1 vector guides the production of N-terminally glutathione S-transferase (GST)-tagged, PreScission-cleavable fusion proteins. Primer sequences are given in Supplementary Table S1. Constructs were confirmed by sequencing. Nucleotide sequences of all primers used in this study are summarized in Supplementary Table S1.

### Protein expression and purification

Proteins bearing an N-terminal, PreScission-cleavable GST-tag were produced in *Escherichia coli* BL21 pLys cells in LB-medium overnight at 18°C after induction at an OD<sub>600</sub> of ~0.6 with 0.5 mM IPTG. The following steps were performed at 4°C. Cells were resuspended in solubilization buffer (50 mM Tris/HCl, pH 8.4, 1 mM DTT) and lysed by sonication. Cell debris was separated from the soluble fraction by centrifugation for 45 min at 55 900 × g in an Avanti J-26 XP centrifuge (Beckman Coulter). Target proteins were captured on glutathione agarose (Macherey-Nagel) and washed with solubilization buffer. The GST-tag was cleaved with 1:50 PreScission on the column overnight and for the mutant mPRPF39 variants for 4 h. The flow-through was collected and loaded on a 1 ml Q column (GE Healthcare). The protein was eluted with a salt gradient from 0 mM to 500 mM followed by concentration and size exclusion chromatography (SEC) in solubilization buffer using a Superdex 200 column (GE Healthcare). Peak fractions were analyzed by SDS-PAGE. Fractions containing the target proteins were pooled, concentrated, and shock-frozen in liquid nitrogen.

For production of selenomethionine-derivatized protein, transformed *E. coli* Rosetta2 DE3 cells were cultured in defined medium containing selenomethionine (18) and at an OD<sub>600</sub> of ~0.6, protein expression was induced by addition of 0.5 mM IPTG. Purification was carried out as described above for wild type mPrp39 (mPrp39<sup>wt</sup>), except that all buffers contained 2 mM DTT. Integrity of the purified proteins was confirmed by MALDI-TOF mass spectrometry.

### Crystallographic procedures

Crystals of the native protein were obtained by the sitting-drop vapor-diffusion method at 4°C with a reservoir solution composed of 0.1 M Bis-Tris-propane/HCl, pH 6.5, 1.8 M sodium acetate. Crystals of selenomethionine-labeled protein were obtained by the hanging-drop method at 18°C with a reservoir solution composed of 0.1 M Bis-Tris-propane/HCl, pH 6.5, 1.6 M sodium acetate, 0.75 mM TCEP. Crystals were cryo-protected with a solution composed of 80% mother liquor and 20% (v/v) propylene glycol and subsequently flash-cooled in liquid nitrogen.

Synchrotron diffraction data were collected at the beamlines of the MX beamlines of the BESSY II (Berlin, Germany) and at beamline P14-2 of PETRA III (DESY Hamburg, Germany). Diffraction data were processed with XDS

(Table 1) (19). Experimental phases were determined by the selenomethionine single-wavelength anomalous diffraction method with the AUTOSOL routine in PHENIX (20) using PHASER (21) and SOLVE/RESOLVE (22), using diffraction data collected from a selenomethionine-labeled mPRPF39 crystal (Table 1). An initial, partial model of mPRPF39 was built with the program AUTOSOL in PHENIX (20) and completed manually by cycles of maximum-likelihood restrained refinement using PHENIX (23,24) including TLS refinement (25) model building with COOT (26). Model quality was evaluated with MolProbity (27) and the JCSG validation server. (28) Secondary structure elements were assigned with DSSP (29) and ALSCRIPT (30) was used for displaying sequence alignments generated by ClustalOmega (31). Structure figures were prepared using PyMOL.

### Size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS)

SEC-MALS experiments were performed at 18°C with buffer containing 50 mM Tris/HCl, pH 8.4, 300 mM NaCl, 0.02% Na<sub>3</sub>N. 80 µg of mPRPF39<sup>wt</sup> or of mPRPF39 variants were loaded onto a Superdex 200 increase 10/300 column (GE Healthcare) that was coupled to a miniDAWN TREOS three-angle light scattering detector (Wyatt Technology) in combination with a RefractoMax520 refractive index detector. For calculation of the molecular mass, protein concentrations were determined from the differential refractive index with a specific refractive index increment ( $dn/dc$ ) of 0.185 ml/g. Data were analyzed with the ASTRA 6.1.4.25 software (Wyatt Technology).

### Immuno-biochemical assays

For immunoprecipitations (IPs), HEK293 cells were lysed in RIPA buffer (10 mM Tris/HCl, pH7.5, 100 mM NaCl, 2 mM EDTA, 1% (v/v) NP-40, with proteinase inhibitors). 100 mg of lysate were diluted in 400 ml RIPA buffer containing 2% (w/v) BSA and pre-cleaned for 1 h using protein A/G Sepharose (Invitrogen). Pre-cleaned lysates were incubated with protein  $\alpha$ Flag-beads, rotating at 4°C overnight. Beads were washed four times in RIPA buffer. After the last wash, SDS-sample buffer was added, samples were boiled and analyzed by SDS-PAGE and Western blot.

### T cells and RNA sequencing

Naïve and memory CD8<sup>+</sup> T cells were generated using the OT-I system as described previously (32,33). RNA-Seq was done essentially as described (34). Briefly, total RNAs were prepared using RNA-Tri (Bio&SELL) and further purified using the RNeasy mini kit (Qiagen) in combination with a DNase I (Qiagen) treatment. RNA sequencing libraries were prepared by using the TruSeq mRNA Library Preparation kit (Illumina). 125-bp paired-end reads were generated by using a HiSeq 2500 sequencer (Illumina) with V4 sequencing chemistry. Triplicate samples from naïve and memory T cells were sequenced (around  $40 \times 10^6$  reads per sample) and analyzed using a MISO-based pipeline (35).

### In vitro splicing

Splicing-active nuclear extracts were prepared according to (36), except HEK cells harvested at ca. 80% confluency were used. For splicing analysis, m7G-capped RNAs were produced by *in vitro* transcription using linearized plasmid as template. *In vitro* splicing assays were performed in 52% HEK nuclear extract, incubating 1 fmol of pre-mRNA and 5 µg of protein or equivalent volume of purification buffer per 25-µl reaction mixture under splicing conditions for 30 min. The reaction was stopped by proteinase K treatment, followed by RNA extraction. Twenty percent of the RNA was reverse-transcribed and analyzed by RT-PCR.

### RT-PCR

RT-PCRs were done as previously described (37) Briefly, RNA was extracted using RNATri (Bio&Sell) and 1 mg RNA was used in a gene-specific RT-reaction. Low-cycle PCR with a <sup>32</sup>P-labeled forward primer was performed, products were separated by denaturing PAGE and quantified using a Phosphoimager and ImageQuantTL software.

### Bioinformatics

U1 snRNA sequences were extracted from Rfam (38) and NCBI (39). Protein sequences of Prp39 and Prp42 orthologs were extracted from NCBI and the canonical isoform was defined using UniProt (40). RNA secondary structure predictions were performed using the RNAstructure Web Server (41). ClustalOmega was used for multiple sequence alignments (31). The common tree taxonomy tools from NCBI was used to create a phylogenetic tree using selected species. The tree was visualized using Interactive Tree Of Life (42).

## RESULTS

### mPRPF39 forms a homodimer

Full-length *M. musculus* mPRPF39 was purified, crystallized and the structure was determined at 3.3 Å. The electron density was of excellent quality that readily allowed to trace the amino acid sequence (Supplementary Figure S2). The asymmetric unit contains two mPRPF39 polypeptide chains that are practically indistinguishable, with a root mean square deviation (rmsd) of 0.39 Å. The protein is purely  $\alpha$ -helical, consisting of 12 pairs of anti-parallel  $\alpha$ -helices arranged as half-a-tetratricopeptide (HAT) repeats. The overall architecture of mPRPF39 can be divided into three distinct domains/regions: An N-terminal domain (HAT-NTD; residues 74–251) comprised of five HAT repeats and a C-terminal domain (HAT-CTD; residues 348–602) with seven HAT repeats. The HAT-NTD and -CTD are connected by a curved, mainly  $\alpha$ -helical linker region (residues 252–347; Figure 1A and B).

The linker region is composed of two short  $\alpha$ -helices ( $L\alpha 1$  and  $L\alpha 2$ ) and an unstructured region (residues 275–294), which connects the short  $\alpha$ -helices to a long, curved helix ( $L\alpha 3$ ). We did not observe interpretable electron density for the 73 N-terminal as well as the 63 C-terminal residues, most likely due to their flexibility.

**Table 1.** Crystallographic data collection and model refinement statistics

Data collection	SeMet	Native
Wavelength [Å]	0.9763	0.9184
Temperature [K]	100	100
Space group	C2	C2
Unit cell parameters		
<i>a</i> , <i>b</i> , <i>c</i> [Å]; β [°]	189.2, 73.0, 206.7; 112.4	189.5, 72.8, 207.1; 112.5
Resolution range [Å] <sup>a</sup>	50.00–3.80 (3.90–3.80)	50.00–3.30 (3.45–3.30)
Reflections <sup>a</sup>		
Unique	49 231 (3671)	39 170 (4809)
Completeness [%]	98.1 (97.6)	98.5 (97.9)
Multiplicity	9.9 (9.9)	4.7 (4.8)
Data quality <sup>a</sup>		
Intensity [ <i>I</i> /σ( <i>I</i> )]	11.8 (1.5)	9.4 (0.9)
R <sub>meas</sub> [%]	17.7 (168.3)	14.9 (260.3)
CC <sub>1/2</sub>	99.9 (83.7)	99.8 (63.7)
Wilson <i>B</i> value [Å <sup>2</sup> ]	120.0	102.1
Number of selenium atoms	25	-
FOM	0.28	-
BAYES-CC	2.9	-
<b>Refinement</b>		
Resolution range [Å] <sup>a</sup>		50.00–3.30 (3.42–3.30)
Reflections <sup>a</sup>		38 901 (3770)
<i>R</i> <sub>work</sub> <sup>a</sup>		0.246 (0.443)
<i>R</i> <sub>free</sub> <sup>a</sup>		0.293 (0.443)
Contents of an asymmetric unit		
Residues, Atoms		1019, 8592
Mean <i>B</i> -factor [Å <sup>2</sup> ]		154.5
RMSD from target geometry		
Bond lengths [Å]		0.005
Bond angles [°]		1.03
<b>Validation statistics</b>		
Ramachandran plot		
Residues in allowed regions [%]		4.6
Residues in favored regions [%]		95.7
MOLPROBITY score <sup>b</sup>		2.1
MOLPROBITY Clashscore <sup>b,c</sup>		16.2

<sup>a</sup>Data for the highest resolution shell in parentheses.

<sup>b</sup>Calculated with MOLPROBITY (27).

<sup>c</sup>Clashscore is the number of serious steric overlaps (>0.4) per 1000 atoms.

The two mPRPF39 molecules are arranged in an anti-parallel manner, forming a homodimer (Figure 1B). Such a homodimeric organization has been suggested before for hPRPF39 (10). Homodimerization is mediated through the concave surfaces of the HAT-CTDs, with the HAT-NTDs facing away from the dimer interface. The overall architecture resembles the arrangement as observed for other HAT-repeat proteins involved in mRNA processing, such as SART3 (43) and CSTF77 (44). In mPRPF39, ~1536 Å<sup>2</sup> of the surface area of each monomer are buried at the interface of the dimer, and the PISA server (45) suggests the presence of a stable dimer in solution.

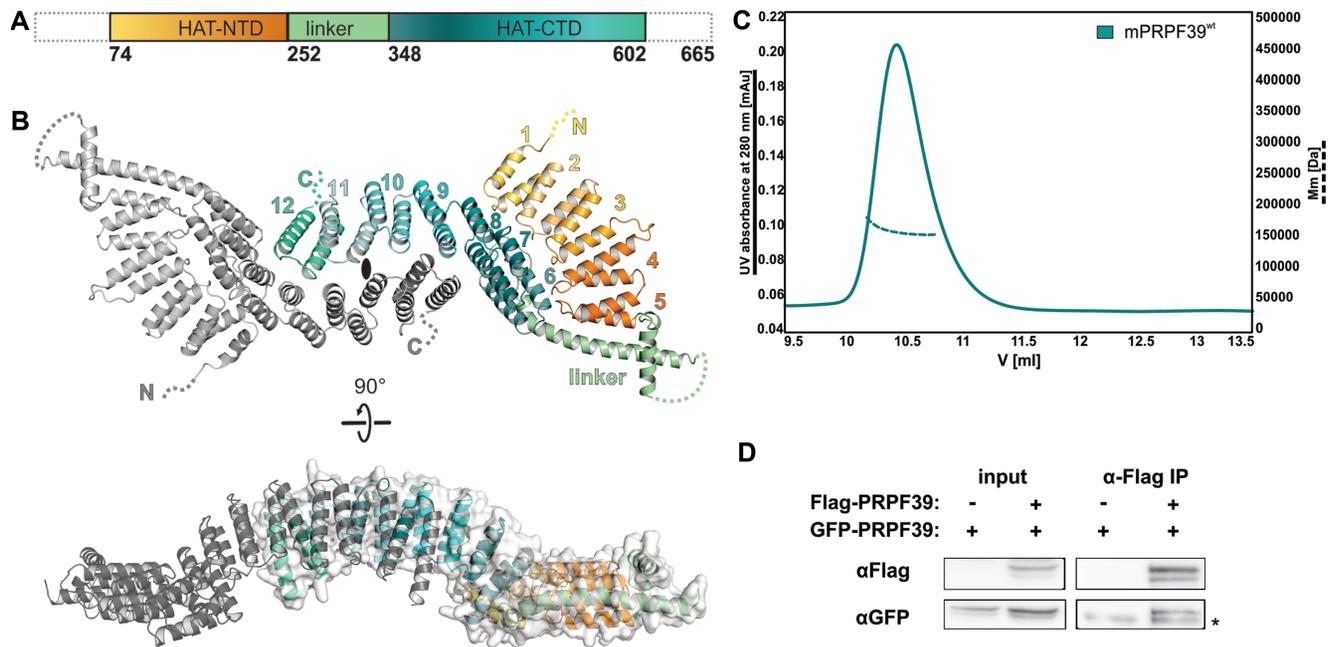
To further investigate whether the dimer exists in solution, we performed analytical size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS) to precisely determine the molecular mass of mPRPF39 in solution. Our SEC-MALS experiments revealed a molecular mass of 157 kDa, confirming the presence of dimeric mPRPF39 in solution (calculated monomeric molecular mass 78.3 kDa; Figure 1C). In addition, FLAG- and GFP-tagged mPRPF39 were co-transfected into HEK 293 cells and a FLAG-IP was performed. GFP-mPRPF39 was co-precipitated in the presence of FLAG-mPRPF39 (Figure 1D). Thus, the mPRPF39 homodimer is also formed in a

cellular environment, in agreement with similar findings for hPRPF39 (10).

### The mPRPF39 homodimer can be disrupted by single point mutations

We mapped the amino acid sequence conservation on the surface of our mPRPF39 structure using The ConSurf Server (46). A highly conserved surface patch is located on the concave side of the HAT-CTD, exactly matching the dimerization interface (Figure 2A and B), indicating that presumably the homodimeric arrangement is conserved in higher eukaryotes. Mapping the electrostatic potential to the surfaces of the mPRPF39 structure revealed two areas in the concave surface of the HAT-CTD with a strongly positive and negative charge, respectively (Figure 2C). Upon dimerization, these positive and negative patches reciprocally interact with each other (Figure 2D). Although homodimerization is thus dominated by charged interactions, the dimer is highly salt-resistant, as even NaCl at a concentration of 900 mM did not lead to disruption of the dimer (Supplementary Figure S3).

To test the importance of interacting residues for homodimerization, we exchanged conserved contact residues



**Figure 1.** mPRPF39 is a homodimer in structure and solution. **(A)** Schematic representation of the domain architecture of mPRPF39. The HAT-NTD and HAT-CTD are colored in gradients from yellow to orange and teal to aqua, respectively; linker – pale green. Regions that are unstructured are indicated as dotted lines. **(B)** Cartoon and combined cartoon/surface representation of the mPRPF39 homodimer in two orientations. Ellipse symbol – non-crystallographic pseudo-twofold axis. One protomer is colored as in panel A, the other protomer is in gray. HAT repeats are numbered 1 through 12 starting at the N-terminus. In this and the following figures, orientations relative to panel A are indicated by rotation symbols. **(C)** Analysis of the oligomeric state of mPRPF39 by SEC-MALS. Molecular mass (Mm) values across the elution peak (solid line; absorbance at 280 nm) are indicated by the dashed line. Average Mm = 157 000 Da; theoretical monomeric mPRPF39<sup>wt</sup> Mm = 78335 Da. **(D)** Western blot of an αFlag-IP of a cellular mixture of Flag- and GFP-tagged mPRPF39. GFP-mPRPF39 can only be pulled down in the presence of Flag-mPRPF39. Asterisk indicates an unspecific band.

(Figure 2D). R458 and R464 are two highly conserved residues in the positive interface patch, which we individually mutated to oppositely charged residues (mPRPF39<sup>R458D</sup>, mPRPF39<sup>R464D</sup>). Likewise, E576 and D577 are highly conserved in the negative interface patch, which we jointly converted to positively charged residues (mPRPF39<sup>E576K/D577K</sup>). Y536 is the only aromatic residue in the interface, which we exchanged for a tryptophan to potentially disrupt the dimer due to steric hindrance (mPRPF39<sup>Y536W</sup>). The dimerization properties of all variants were investigated by SEC-MALS and compared to those of wild type (wt) mPRPF39. Dimerization was unaffected and partially lost in the mPRPF39<sup>Y536W</sup> and mPRPF39<sup>R458D</sup> variants, respectively (Figure 2E). mPRPF39<sup>R464D</sup> and mPRPF39<sup>E576K/D577K</sup> clearly migrated as monomers in SEC. These results confirm that dimerization is based on the interactions between the conserved, oppositely charged interface patches in the HAT-CTD of mPRPF39 and demonstrate that exchange of a single residue (mPRPF39<sup>R464D</sup>) can lead to complete disruption of the dimer.

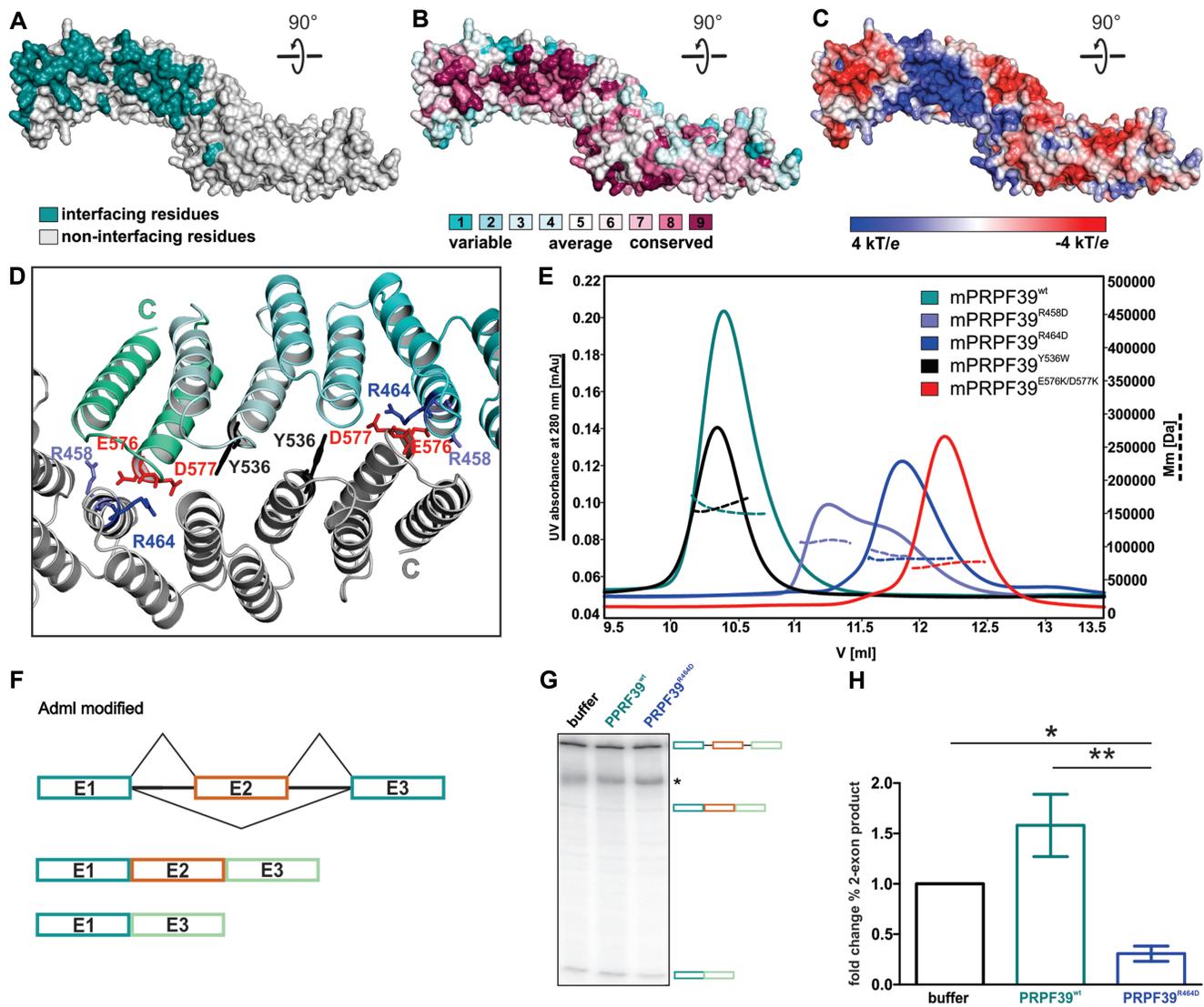
### mPRPF39 homodimerization is functionally important

To test if the observed homodimerization is required for the functionality of mPRPF39, we performed *in vitro* splicing assays with mPRPF39<sup>wt</sup> and a monomeric version of mPRPF39. For the functional assays the R464D exchange was chosen, as it led to complete monomerization with the

least number of residue exchanges. A modified Adml construct (Figure 2F) was used for *in vitro* splicing and the 2-exon product, representing the main product, was quantified. Splicing extracts treated with mPRPF39<sup>wt</sup> showed a trend towards higher splicing efficiency (Figure 2G and H), but the effect was not significant when compared to the samples treated only with buffer. This suggests that PRPF39 is not limiting in nuclear extracts, so that exogenous PRPF39 does not strongly affect splicing. In contrast, samples treated with mPRPF39<sup>R464D</sup> showed significantly reduced splicing levels compared to both, samples treated with only buffer or mPRPF39<sup>wt</sup>. These findings indicate that the mPRPF39 homodimer represents the functional unit of mPRPF39 and is required for splicing.

### mPRPF39 is an NMD target

Our data indicate that reducing the level of the active mPRPF39 homodimer can be used to control splicing efficiency. Interestingly, the *prpf39* gene has an alternative exon which contains a premature stop codon (Figure 3A). The alternative exon and its flanking regions are highly conserved between mouse and human (Supplementary Figure S4). Upon inclusion of the alternative exon, the premature stop codon could potentially lead to NMD and might thus regulate PRPF39 levels. RNA-Seq and RT-PCR validation in independent samples of murine T-cells showed that inclusion of the alternative exon is strongly increased in memory versus naïve T cells (Supplementary Figure S5) Testing

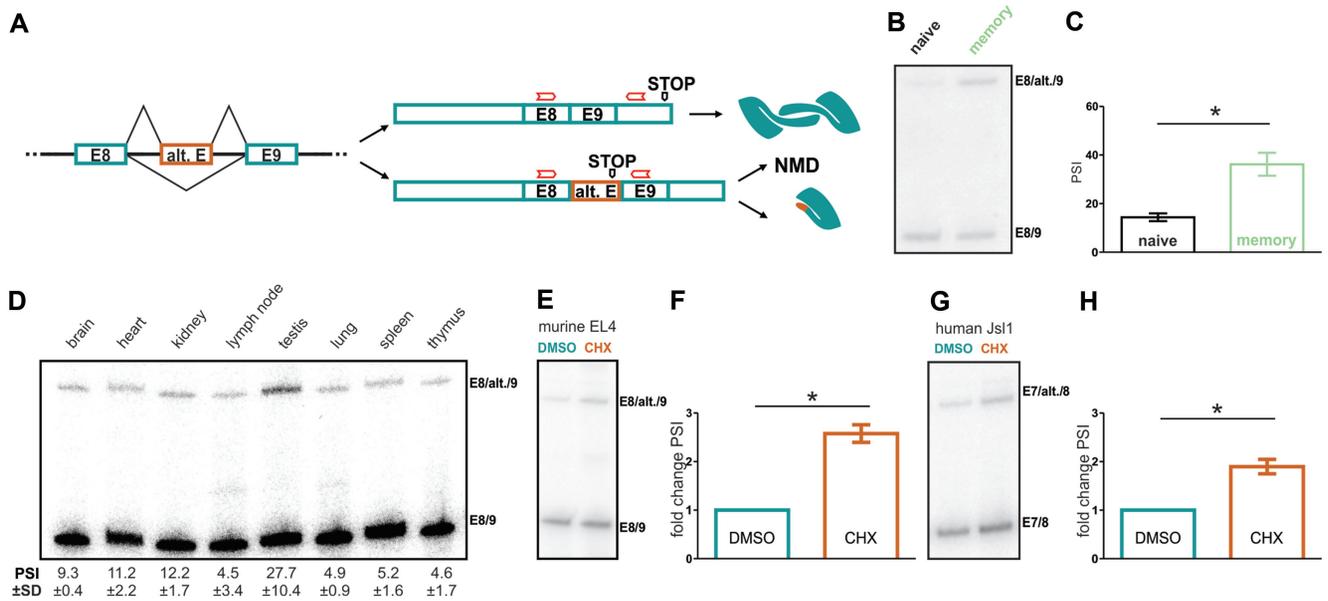


**Figure 2.** Impact of dimerization on splicing. (A) Surface representation of a mPRPF39 monomer. Interface and non-interface residues are colored teal and grey, respectively. View identical to the lower panel of Figure 1B. (B) Sequence conservation projected onto the surface of mPRPF39. Color-coding according to bins: bin 9 (magenta) contains the most conserved positions, bin 1 (cyan) contains the most variable positions. The homodimer interface on the concave side of the HAT-CTD is highly conserved. (C) Electrostatic potential mapped on the surface of mPRPF39. The homodimer interface on the concave side of the HAT-CTD is carpeted by complementary patches of positive and negative surface potential. (D) Zoom into the dimerization interface. Amino acid residues subjected to site-directed mutagenesis are indicated in light blue (mPRPF39<sup>R458D</sup>), blue (mPRPF39<sup>R464D</sup>), red (mPRPF39<sup>E576K/D577K</sup>) and black (mPRPF39<sup>Y536W</sup>). Same view as upper panel in Figure 1B. (E) Analysis of the oligomeric states of mPRPF39 variants in solution by SEC-MALS. Molecular mass (Mm) values across the elution peaks (solid lines; absorbance at 280 nm; legend provided in the inset) are indicated by the dashed lines, average Mm values are indicated next to the respective peaks. Theoretical monomeric mPRPF39<sup>wt</sup> Mm = 78335 Da. mPRPF39<sup>wt</sup> and mPRPF39<sup>Y536W</sup> migrate as dimers, mPRPF39<sup>R458D</sup> forms a mixture of molecular species, mPRPF39<sup>R464D</sup> and mPRPF39<sup>E576K/D577K</sup> migrate as monomers. (F) Schematic representation of the modified 3-exon/2-intron Adml construct used for *in vitro* splicing. (G) Exemplary gel of RT-PCR analysis of an *in vitro* splicing reaction, comparing buffer-, mPRPF39<sup>wt</sup>- and mPRPF39<sup>R464D</sup>-treated samples. mPRPF39<sup>wt</sup> and mPRPF39<sup>R464D</sup> treatments lead to elevated splicing and significantly reduced splicing. The asterisk indicates a non-specific band. (H) Quantification of RT-PCRs of four biologically independent *in vitro* splicing experiments performed in technical triplicates. Values represent means  $\pm$  standard deviations. Significance was estimated via a one-sample *t*-test (\* $P \leq 0.05$ ; \*\* $P \leq 0.01$ ).

different murine tissue types, we observed that inclusion of the alternative exon is particularly high in testis and low in lymph node (Figure 3D).

Detection of the exon in the mRNA suggests that the inclusion mRNA isoform is not a strong NMD target. To address whether inclusion leads to degradation at all, cells of a murine and human T-cell line (EL4 and Jsl1, respectively) were treated with cycloheximide (CHX). RT-PCR demon-

strated an increase in the isoform containing the alternative exon in cells treated with CHX (Figure 3E–H), confirming that inclusion of the alternative exon induces NMD. This effect was moderate, indicating that the corresponding mRNA is only a weak NMD target. However, regardless of whether the inclusion isoform is degraded or translated, this isoform will lead to reduced functional protein (Supplementary Figure S6), as the translated product would be missing



**Figure 3.** Alternative splicing of *prpf39* pre-mRNA with resulting mRNA and protein isoforms. (A) An alternative exon (orange) can either be included or excluded. Upon exon exclusion, the mRNA can be translated to full-length mPRPF39; upon exon inclusion, a premature stop codon is introduced, which may lead to NMD or production of a non-functional truncated mPRPF39. Red arrows indicate location of primers used for RT-PCR. (B) Representative validation of RNA-seq results by RT-PCR. (C) Quantification of validation as shown in panel B. Percent spliced in (PSI) values represent means  $\pm$  standard deviations for four independent experiments. (D) Exemplary gel of RT-PCR in different murine tissues. The degree of alternative exon inclusion is tissue specific, with the highest inclusion in testis and the lowest in lymph node. Results of quantification of experiments performed with samples from three individual mice are shown below the respective lanes. Values represent means  $\pm$  standard deviations. (E–H) Analysis of *mprpf39* (pre-)mRNAs in mouse EL4 cells (E, F) and human Jsl1 cells (G, H). Left – representative RT-PCR analyses; right – quantification. DMSO/CHX – DMSO/cycloheximide-treated samples. In the CHX treated cells, an accumulation of the transcript containing the alternative exon can be seen, indicating that this isoform is an NMD-target. Values represent means  $\pm$  standard deviations of three independent experiments. Significance was estimated via a one-sample T-test (\* $P \leq 0.05$ , \*\* $P \leq 0.01$ ).

the C-terminal domain involved in dimerization and will thus represent a dominant negative protein isoform. Furthermore, the importance of the alternative exon to regulate PRPF39 protein levels is supported by its conservation and the conserved effect as weak NMD exon between mouse and human.

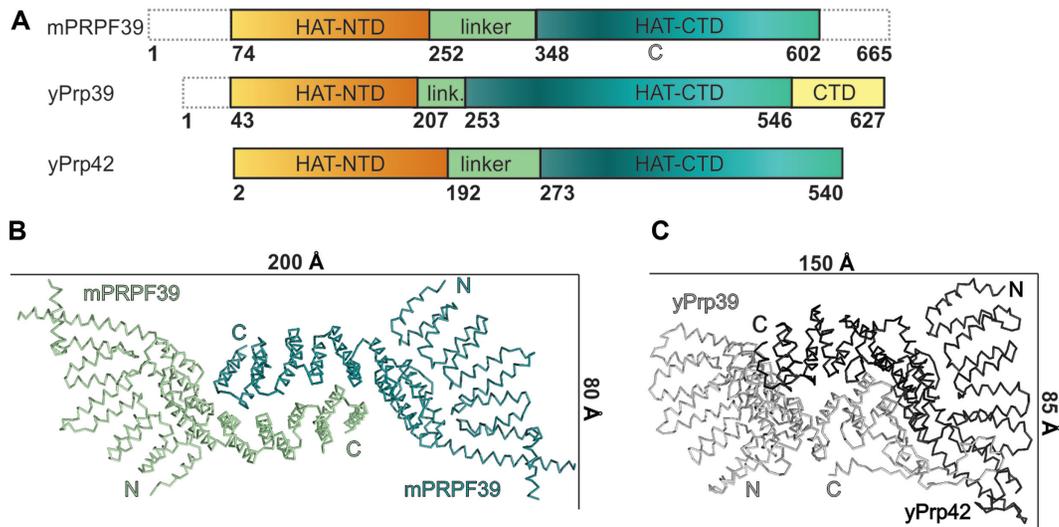
#### A PRPF39 homodimer might functionally substitute for the yPrp39/yPrp42 heterodimer in metazoans

Although a search using the DALI server (47) identified murine CSTF-77 (PDB entry 2OOE (44)), its *Kluyveromyces lactis* ortholog RNA14 (PDB entry 4EBA (12)) and human SART3 (PDB entry 5JPZ, DOI: 10.2210/pdb5JPZ/pdb) as the closest structural neighbors of mPRPF39 (Supplementary Table S2), the structurally most similar snRNP proteins were yPrp39 and yPrp42 (PDB entry 5UZ5 (10); rmsd of 7.6/4.5 Å for 583/539 number of residues, respectively). While all three proteins share the same domain organization, mPRPF39 and yPrp42 contain an additional  $\alpha$ -helix in their HAT-NTDs (Supplementary Figure S7 and Supplementary Discussion). Moreover, while the linker region is the structurally most divergent part in the three proteins (Supplementary Figures S7 and S8 and Supplementary Discussion), the yPrp42 linker with two  $\alpha$ -helices is more similar to the three-helix mPRPF39 linker than the yPrp39 linker that bears only a single  $\alpha$ -helix. Thus, somewhat surprisingly,

mPRPF39 is structurally more similar to yPrp42 than to the formal ortholog, yPrp39. Consistently, mPRPF39 exhibits a slightly higher sequence identity to yPrp42 than to yPrp39 (23 vs. 22%, respectively). But notably in human, hPRPF39 shows a sequence identity of 17% to yPrp42 compared to 18% to yPrp39. Hence, PRPF39 may be considered an ortholog of both, yPrp39 and yPrp42.

Because of the similar central architecture of mPRPF39, yPrp39 and yPrp42, also the mPRPF39 homodimer is globally similar to the heterodimer of yPrp39/yPrp42 (Figure 4A). In both complexes the interaction is established via the respective HAT-CTDs. The mPRPF39 homodimer is more elongated than the yPrp39/yPrp42 heterodimer and it shows less curvature in its bridging HAT-CTDs (Figure 4B and C). In contrast to the mPRPF39 homodimer interface, the interface of yPrp39 and yPrp42 is dominated by hydrophobic interactions. Only one salt-bridge is established between the yPrp39 CTD extension and the yPrp42 HAT-NTD.

As (i) metazoans lack a yPrp42 ortholog, (ii) mPRPF39 shows higher sequence and structural similarity to yPrp42 than to yPrp39, (iii) mPRPF39 forms a homodimer, (iv) the mPRPF39 homodimer resembles the yPrp39/yPrp42 heterodimers and (v) monomeric mPRPF39 has a detrimental effect on splicing, a PRPF39 homodimer is likely to functionally substitute the yPrp39/yPrp42 heterodimer in metazoan species. Consistent with this notion, the area on yPrp42 that mediates protein-protein contacts connecting



**Figure 4.** Comparison of mPRPF39, yPrp39 and yPrp42 structures. (A) Schematic representation of the domain architecture of mPRPF39, yPrp39 and yPrp42. Domain coloring as in Figure 1A. Regions that are unstructured are indicated as dotted lines. (B) Ribbon diagram of the mPRPF39 homodimer (teal and palegreen) (C) Ribbon diagram of yPrp42/yPrp39 heterodimer (dark and light gray). yPrp42 is colored in dark grey and yPrp39 in light gray. The dimensions of the dimer are indicated in Å. Same view as in the lower panel of Figure 1B.

the core U1 snRNP to its auxiliary proteins, is highly conserved in mPRPF39 (Supplementary Figure S9).

#### PRPF39 homodimerization is evolutionarily connected to a short U1 snRNA

The yeast U1 snRNA has a strongly elongated stem-loop (SL) 2 compared to its human counterpart (Figure 5). It consists of two sub-domains, the conserved SL2-1 and elongated SL2-2 (Figure 5). Interestingly an evolutionary study focusing on the SL3 (48) suggested that the duplication of *prp39* went along with an increasing length of SL3. The recent cryo-EM structure of the yeast U1 snRNP indeed revealed an interface between the sub-domain SL3-4 (Figure 5B) and yPrp42 (10). However, the buried surface area between yPrp42 and SL3-4 is much smaller than between yPrp42 and SL2-2, indicating a larger influence of SL2-2 on the evolution of the dimeric system. The convex side of the yPrp42 HAT-NTD has a strongly positively charged groove that accommodates SL2-2 of U1 snRNA (Figure 5A). A similar positive patch is not seen on mPRPF39 (Figure 5A), most likely because the U1 SL2 is much shorter in metazoans than in yeast (Figure 5B) and will thus not reach this surface area of PRPF39. This observation led us to address the question, whether PRPF39 homodimerization in metazoan might be evolutionarily linked to a short U1 SL2. The sequence lengths of all annotated U1 snRNAs from Rfam (38) range between 96 and 825 nucleotides (nts). Notably, higher eukaryotic U1 snRNAs all lie on the lower end of the range, with most of them having a length of 164 nts (Figure 6B). In contrast, fungi tend to have longer U1 snRNA sequences, most of them over 450 nts (5). The longer U1 snRNA likely goes hand in hand with the presence of a yeast-like SL2. However, we found a few fungi with short U1 snRNAs; e.g., *Aspergillus fumigatus* and *Debaryomyces hansenii* have U1 sequence lengths of 149 and 165 nts, respectively. When performing secondary structure pre-

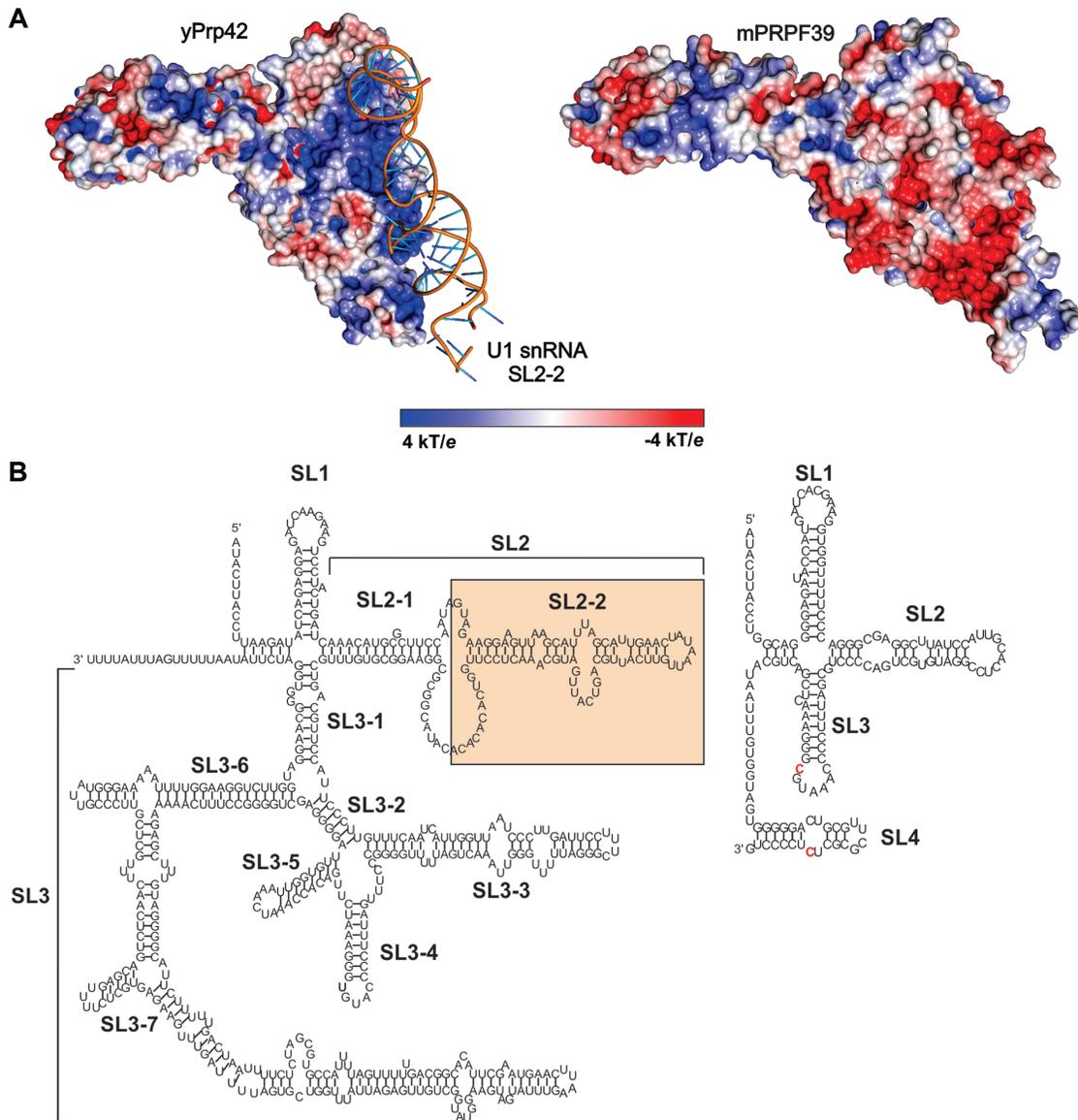
dictions (41), we observed that their U1 snRNA structures resemble the human U1 structure with a short SL2 (Supplementary Figure S10). The *Candida albicans* snRNA with an intermediate length of 244 nts shows a shortened SL3, but a yeast-like SL2.

We then queried NCBI for hPRPF39, yPrp39 and yPrp42 orthologs. Interestingly, only fungi contain Prp42 proteins (Figure 6A and Supplementary Figure S11). Notably, we found that the only fungi with an annotated Prp39 but no annotated Prp42 homolog are *D. hansenii*, *Y. lipolytica*, *S. pombe*, *A. fumigatus* and *N. crassa* (Table 2, Figure 6A and Supplementary Figure S11), which also harbor short U1 snRNAs with a metazoan-like SL2 and SL3. Thus, organisms that lack Prp42, and thus most likely are dependent on Prp39 homodimers, also contain U1 snRNAs with a short SL2 and SL3. In contrast, the U1 snRNAs of organisms that have both Prp39 and Prp42, and that thus most likely use a Prp39/Prp42 heterodimer, exhibit an extended SL2 and SL3.

#### PRPF39 homodimerization and short U1 snRNA correlate with higher splicing complexity

A previous analysis the exon-intron structure of genes in fungal genomes revealed a large variation in the percentage of intron containing genes (from 0.7 to 97) and in the number of introns per gene (2 to 18) as well as a lack of correlation between splicing complexity and genome size, number of chromosomes or number of genes (49). We noticed that fungi with less complex splicing have long U1 snRNAs and a Prp39/42 heterodimer, while fungi with more complex splicing have short U1 snRNAs and a Prp39 homodimer (Table 2).

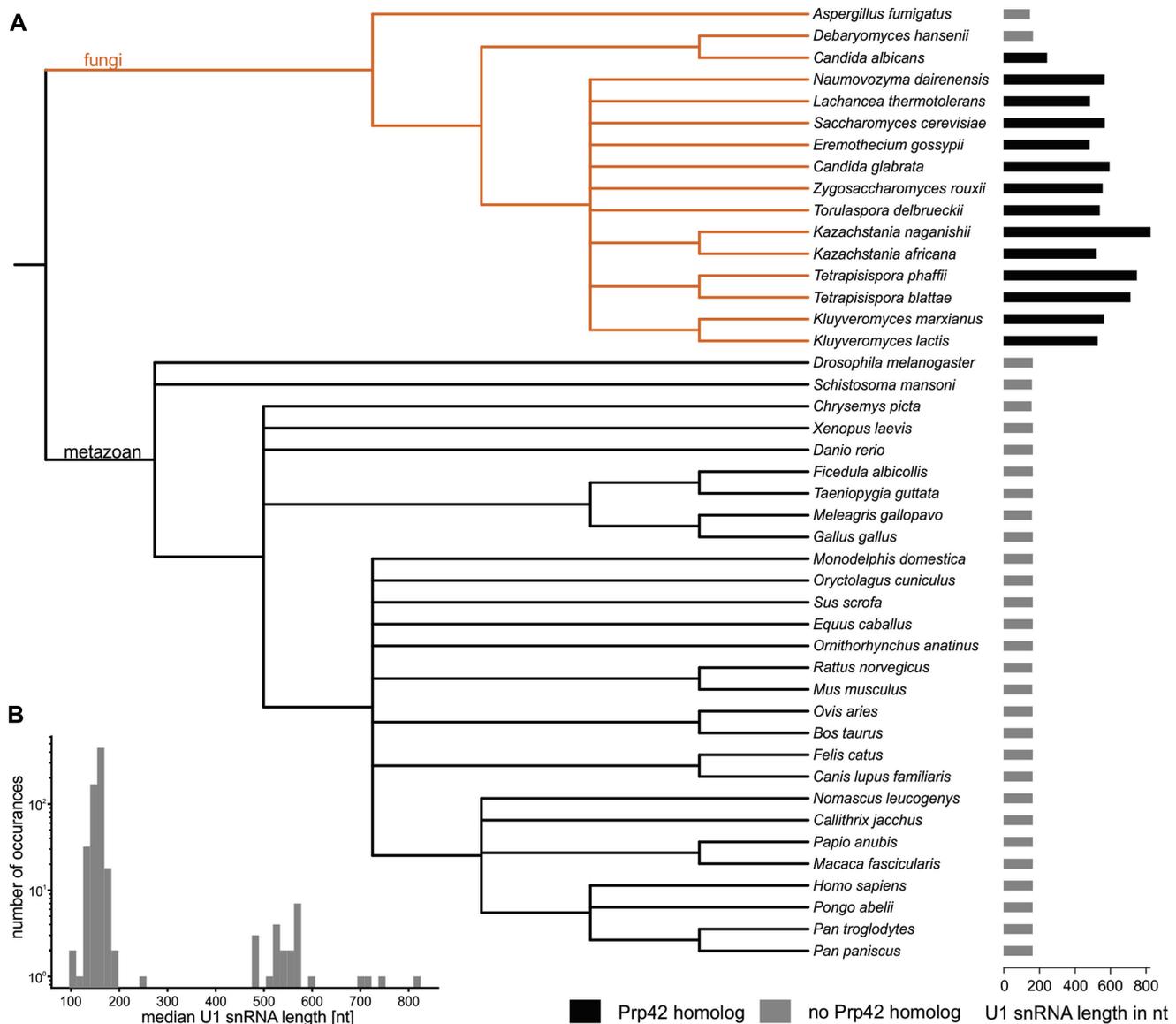
Initially, *D. hansenii* and *Y. lipolytica* seemed to represent exceptions to this trend (relatively low splicing complexity and short U1 snRNAs). However, BLAST searches with the *S. cerevisiae* yPrp39 or yPrp42 orthologs uncovered



**Figure 5.** U1 snRNA binding to yPrp42. (A) Electrostatic potential mapped on the surface of yPrp42 (left) and mPRPF39 (right) in the same orientation as in the upper panel of Figure 1B, with the U1 snRNA SL2-2 shown as cartoon bound to yPrp42. Only yPrp42 shows the positively charged groove needed to accommodate the RNA. Same view as in Figure 1B. (B) Schematic representations of yeast U1 snRNA (left) and of murine U1 snRNA (right). Characters in red indicate nucleotides that are not conserved between human and murine U1 snRNA. The pale orange box highlights the elongated SL2-2 in yeast.

**Table 2.** Analysis of splicing complexity in connection with U1 snRNP properties. The length of SL2 and SL3 was estimated based on secondary structure predictions of the respective U1 snRNA. The length of SL2 and SL3 of *S. cerevisiae* is based on the structure of U1 snRNP (10). Intron containing genes are given in [%] and the maximal number of introns per gene is given according to Ivashchenko *et al.* (49)

species	Prp39/Prp42	Prp(f)39/Prp(f)39	U1 snRNA [nts]	SL2 [nts]	SL3 [nts]	Intron-containing genes [%]	max. number of introns per gene
<i>C. glabrata</i>	x		595	147	396	1.5	2
<i>K. lactis</i>	x		528	132	330	2.4	2
<i>E. gossypii</i>	x		483	134	276	4.5	2
<i>S. cerevisiae</i>	x		568	122	377	4.5	2
<i>D. hansenii</i>	x		165	45	28	5.0	4
<i>Y. lipolytica</i>	x		150	43	14	10.6	4
<i>S. pombe</i>		x	149	44	22	45.6	15
<i>A. fumigatus</i>		x	149	46	18	78.0	18
<i>N. crassa</i>		x	130	43	22	79.6	11



**Figure 6.** Phylogenetic analysis of U1 snRNA length and presence of a hetero- or homodimer. (A) Phylogenetic tree of the organisms included in the analysis. Phylogenetic distances are not to scale but are given in Supplementary Figure S11. Fungal clade – orange, metazoan clade – black. On the right side, the respective median U1 snRNA lengths are plotted as a bar chart (organisms with an annotated Prp42 – black; organisms lacking an annotated Prp42 – grey). Most fungi, except *A. fumigatus* and *D. hansenii*, have a Prp42 ortholog. *A. fumigatus* and *D. hansenii* also have short U1 snRNAs. A list of all NCBI GeneIDs and NCBI sequence identifiers can be found in Supplementary Table S3. (B) The number of organisms plotted against the median U1 snRNA length. There are two populations that cluster around 160 nts and 550 nts.

an additional Prp39/Prp42-like protein in these species, in addition to the annotated Prp39 ortholog. These observations suggest that *D. hansenii* and *Y. lipolytica* occupy intermediate phylogenetic positions, with a Prp39/Prp42 heterodimer, short U1 snRNAs yet only slightly elevated splicing complexity (Table 2). Our observations suggest that lack of a Prp42 ortholog, and thus a Prp39 homodimer, in conjunction with a short U1 snRNA enabled the development of more complex splicing patterns.

While we cannot pinpoint a particular sequence of evolutionary events, these findings clearly demonstrate that presence of a Prp39/Prp42 heterodimer phylogenetically correlates with U1 snRNAs bearing a long SL2 and SL3. More-

over, the shortening of the U1 SL2 and SL3 alone appears to allow only slightly higher splicing complexity (*D. hansenii*, *Y. lipolytica*), but it is the absence of a Prp42 homolog that then causes dramatically increased levels of splicing complexity. This indicates that relying on a Prp(f)39 homodimer is the driving force behind diversifying splicing in these organisms.

## DISCUSSION

We have determined a crystal structure of full-length mPRPF39 composed of a HAT-NTD that is connected via an extended linker to the HAT-CTD. This structure in combination with biophysical and CoIP assays demon-

strates that mPRPF39 forms a stable homodimer *in vitro* and *in vivo*. As disruption of the mPRPF39 homodimer has detrimental effects on splicing, we conclude that mPRPF39 functions as a homodimer in splicing, by functionally replacing the Prp39/Prp42 heterodimer of the yeast system. Given the high sequence conservation, in particular at the dimer interface, PRPF39 most likely forms similar homodimers in all eukaryotes lacking a Prp42 ortholog. Prp39 and Prp42 presumably emerged from the same ancestral gene by gene duplication in some fungi. Our detailed structural and phylogenetic comparisons revealed that PRP(F)39 in higher eukaryotes should rather be considered orthologs of yeast Prp42 than Prp39.

There are ~80 proteins in the human spliceosome that lack an obvious counterpart in yeast, but only few examples, such as Prp42, for which a yeast spliceosomal protein lacks an assigned ortholog in human (3,4). We find that the level of functional mPRPF39 has a large impact on splicing *in vitro*. Furthermore, mPRPF39 expression is controlled by alternative splicing leading to either NMD or a dominant negative protein incapable of forming a dimer, in a tissue- and activation-dependent manner. These data suggest a role of PRP(F)39 in adapting splicing efficiency to the requirements of specific cells or tissues. In particular in immune cell differentiation and activation it has been shown that regulated intron retention plays an important role in controlling gene expression and function (50). Controlling PRPF39 levels through inclusion of an exon containing a premature stop codon may contribute to fine tune splicing efficiency and intron retention in such contexts, as we find it strongly regulated between naive and memory T cells.

Our phylogenetic analysis show that lack of a Prp42 ortholog (as presently assigned) correlates with more complex splicing patterns. While it is straightforward to imagine mechanisms, by which splicing complexity may be increased in a molecularly more complex spliceosome, the converse scenario appears counter-intuitive. The *S. cerevisiae* Prp39/Prp42 heterodimer is crucial in organizing the spliceosome during early steps of splicing (13,14). However, while in yeast the Prp39/Prp42 heterodimer is stably associated with the U1 snRNP, hPRPF39 is only transiently associated with the spliceosome. A less stable association of splicing factors with the spliceosome may allow for more flexibility in splicing, e.g. with respect to splice site choice, with respect to enabling other modes of alternative splicing or with respect to tuning the ratio of productive and abortive splicing, by rendering particular splicing events dependent on additional factors or signals.

In the present case, increased splicing complexity correlates not only with the lack of a Prp42 ortholog but also with a U1 snRNA bearing a short SL2 and SL3. These two features seem to allow for more flexibility in the apposition of U1 and U2 snRNPs during the early steps of splicing, as PRPF39 and a short U1 SL2 and SL3 are less tightly associated compared to yPrp42 and a long U1 SL2 and SL3. While the yPrp39/yPrp42 heterodimer mediates a stable connection between the U1 and U2 snRNPs in a pre-spliceosome (A-complex (14)) and the spliceosomal pre-B complex in yeast (13), in the human spliceosome stable inter-U1/U2 snRNP contacts are lacking in the pre-B state (51). Thus, PRPF39-mediated U1-U2 bridging in the

human system seems to depend on additional proteins and this dependency may be one of the molecular principles, by which splicing is additionally regulated. Other splicing factors have also been reported to be 'reprogrammed' from stable to facultative snRNP subunits. For example, Prp38 and Snu23 are stable snRNP components in yeast but in human they enter the spliceosome individually upon integration of the tri-snRNP into the (pre-)B complex (52).

The present example provides a particularly striking example that splicing complexity does not necessarily scale with the number of spliceosomal protein—even though the yPrp39/yPrp42 heterodimer in yeast represents a molecularly more complex spliceosomal subunit, the molecularly more simple PRP(F)39 homodimer in metazoan species correlates with higher flexibility in spliceosome assembly and thus splicing complexity. The increased splicing complexity seems to be provided at least in part by a reduced pre-organization of the U1 snRNP, conferred by an absent PRP(F)39-U1 snRNA interface, and in part by alternative splicing of the *prp(f)39* gene itself, which controls the level of functional PRP(F)39.

## DATA AVAILABILITY

The atomic coordinates and structure factor amplitudes have been deposited in the Protein Data Bank under the accession code 6G70.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We would like to thank Gleb Bourenkov for assistance during data collection at beamline P14 of PETRA III (Deutsches Elektronen Synchrotron, Hamburg, Germany). We accessed beamlines of the BESSY II (Berliner Elektronenspeicherring-Gesellschaft für Synchrotronstrahlung II) storage ring (Berlin, Germany) via the Joint Berlin MX-Laboratory sponsored by the Helmholtz Zentrum Berlin für Materialien und Energie, the Freie Universität Berlin, the Humboldt-Universität zu Berlin, the Max-Delbrück Centrum, and the Leibniz-Institut für Molekulare Pharmakologie. We are grateful to Claudia Alings for support during crystallization, Jan Wollenhaupt for support in SEC-MALS measurements, and Christoph Weise for mass spectrometric analyses. We would like to thank Sutapa Chakrabarti and Marco Preußner for discussions. We thank Cindy Will for splice extracts.

## FUNDING

Deutsche Forschungsgemeinschaft [SFB 958 A21 to F.H. and TRR 186 A16 to F.H. and M.C.W.]; We acknowledge support by the German Research Foundation and the Open Access Publication Fund of the Freie Universität Berlin. *Conflict of interest statement.* None declared.

## REFERENCES

1. Wahl, M.C., Will, C.L. and Luhrmann, R. (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell*, **136**, 701–718.

2. Seraphin, B., Kretzner, L. and Rosbash, M. (1988) A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5' cleavage site. *EMBO J.*, **7**, 2533–2538.
3. Fabrizio, P., Dannenberg, J., Dube, P., Kastner, B., Stark, H., Urlaub, H. and Luhrmann, R. (2009) The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. *Mol. Cell*, **36**, 593–608.
4. Will, C.L. and Luhrmann, R. (2011) Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.*, **3**, a003707.
5. Kretzner, L., Krol, A. and Rosbash, M. (1990) *Saccharomyces cerevisiae* U1 small nuclear RNA secondary structure contains both universal and yeast-specific domains. *Proc. Natl. Acad. Sci. U.S.A.*, **87**, 851–855.
6. Kretzner, L., Rymond, B.C. and Rosbash, M. (1987) *S. cerevisiae* U1 RNA is large and has limited primary sequence homology to metazoan U1 snRNA. *Cell*, **50**, 593–602.
7. Fortes, P., Bilbao-Cortes, D., Fornerod, M., Rigaut, G., Raymond, W., Seraphin, B. and Mattaj, I.W. (1999) Luc7p, a novel yeast U1 snRNP protein with a role in 5' splice site recognition. *Genes Dev.*, **13**, 2425–2438.
8. Gottschalk, A., Tang, J., Puig, O., Salgado, J., Neubauer, G., Colot, H.V., Mann, M., Seraphin, B., Rosbash, M., Luhrmann, R. *et al.* (1998) A comprehensive biochemical and genetic analysis of the yeast U1 snRNP reveals five novel proteins. *RNA*, **4**, 374–393.
9. Lockhart, S.R. and Rymond, B.C. (1994) Commitment of yeast pre-mRNA to the splicing pathway requires a novel U1 small nuclear ribonucleoprotein polypeptide, Prp39p. *Mol. Cell Biol.*, **14**, 3623–3633.
10. Li, X., Liu, S., Jiang, J., Zhang, L., Espinosa, S., Hill, R.C., Hansen, K.C., Zhou, Z.H. and Zhao, R. (2017) CryoEM structure of *Saccharomyces cerevisiae* U1 snRNP offers insight into alternative splicing. *Nat. Commun.*, **8**, 1035.
11. Huang, M., Zhou, B., Gong, J., Xing, L., Ma, X., Wang, F., Wu, W., Shen, H., Sun, C., Zhu, X. *et al.* (2018) RNA-splicing factor SART3 regulates translesion DNA synthesis. *Nucleic Acids Res.*, **46**, 4560–4574.
12. Paulson, A.R. and Tong, L. (2012) Crystal structure of the Rna14-Rna15 complex. *RNA*, **18**, 1154–1162.
13. Bai, R., Wan, R., Yan, C., Lei, J. and Shi, Y. (2018) Structures of the fully assembled *Saccharomyces cerevisiae* spliceosome before activation. *Science*, **360**, 1423–1429.
14. Plaschka, C., Lin, P.C., Charenton, C. and Nagai, K. (2018) Prespliceosome structure provides insights into spliceosome assembly and regulation. *Nature*, **559**, 419–422.
15. Blomen, V.A., Majek, P., Jae, L.T., Bigenzahn, J.W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F.R., Olk, N., Stukalov, A. *et al.* (2015) Gene essentiality and synthetic lethality in haploid human cells. *Science*, **350**, 1092–1096.
16. Hart, T., Chandrasekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S. *et al.* (2015) High-Resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell*, **163**, 1515–1526.
17. Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S. and Sabatini, D.M. (2015) Identification and characterization of essential genes in the human genome. *Science*, **350**, 1096–1101.
18. Van Duyne, G.D., Standaert, R.F., Karplus, P.A., Schreiber, S.L. and Clardy, J. (1993) Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. *J. Mol. Biol.*, **229**, 105–124.
19. Kabsch, W. (2010) XDS. *Acta Crystallogr. D. Biol. Crystallogr.*, **66**, 125–132.
20. Terwilliger, T.C., Adams, P.D., Read, R.J., McCoy, A.J., Moriarty, N.W., Grosse-Kunstleve, R.W., Afonine, P.V., Zwart, P.H. and Hung, L.W. (2009) Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallogr. D. Biol. Crystallogr.*, **65**, 582–601.
21. McCoy, A.J., Storoni, L.C. and Read, R.J. (2004) Simple algorithm for a maximum-likelihood SAD function. *Acta Crystallogr. D. Biol. Crystallogr.*, **60**, 1220–1228.
22. Terwilliger, T.C. (2000) Maximum-likelihood density modification. *Acta Crystallogr. D. Biol. Crystallogr.*, **56**, 965–972.
23. Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W. *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D. Biol. Crystallogr.*, **66**, 213–221.
24. Afonine, P.V., Grosse-Kunstleve, R.W., Echols, N., Headd, J.J., Moriarty, N.W., Mustyakimov, M., Terwilliger, T.C., Urzhumtsev, A., Zwart, P.H. and Adams, P.D. (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. D. Biol. Crystallogr.*, **68**, 352–367.
25. Winn, M.D., Isupov, M.N. and Murshudov, G.N. (2001) Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallogr. D. Biol. Crystallogr.*, **57**, 122–133.
26. Emsley, P., Lohkamp, B., Scott, W.G. and Cowtan, K. (2010) Features and development of Coot. *Acta Crystallogr. D. Biol. Crystallogr.*, **66**, 486–501.
27. Chen, V.B., Arendall, W.B. 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S. and Richardson, D.C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D. Biol. Crystallogr.*, **66**, 12–21.
28. Yang, H., Guranovic, V., Dutta, S., Feng, Z., Berman, H.M. and Westbrook, J.D. (2004) Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank. *Acta Crystallogr. D. Biol. Crystallogr.*, **60**, 1833–1839.
29. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
30. Barton, G.J. (1993) ALSCRIPT: a tool to format multiple sequence alignments. *Protein Eng.*, **6**, 37–40.
31. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
32. Stoycheva, D., Deiser, K., Starck, L., Nishanth, G., Schluter, D., Uckert, W. and Schuler, T. (2015) IFN-gamma regulates CD8+ memory T cell differentiation and survival in response to weak, but not strong, TCR signals. *J. Immunol.*, **194**, 553–559.
33. Deiser, K., Stoycheva, D., Bank, U., Blankenstein, T. and Schuler, T. (2016) Interleukin-7 modulates anti-tumor CD8+ T cell responses via its action on host cells. *PLoS One*, **11**, e0159690.
34. Herdt, O., Neumann, A., Timmermann, B. and Heyd, F. (2017) The cancer-associated U2AF35 470A>G (Q157R) mutation creates an in-frame alternative 5' splice site that impacts splicing regulation in Q157R patients. *RNA*, **23**, 1796–1806.
35. Preussner, M., Goldammer, G., Neumann, A., Haltenhof, T., Rautenstrauch, P., Muller-McNicoll, M. and Heyd, F. (2017) Body temperature cycles control rhythmic alternative splicing in mammals. *Mol. Cell*, **67**, 433–446.
36. Dignam, J.D., Lebovitz, R.M. and Roeder, R.G. (1983) Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res.*, **11**, 1475–1489.
37. Preussner, M., Wilhelmi, I., Schultz, A.S., Finkernagel, F., Michel, M., Moroy, T. and Heyd, F. (2014) Rhythmic U2af26 alternative splicing controls PERIOD1 stability and the circadian clock in mice. *Mol. Cell*, **54**, 651–662.
38. Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D. and Petrov, A.I. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335–D342.
39. Geer, L.Y., Marchler-Bauer, A., Geer, R.C., Han, L., He, J., He, S., Liu, C., Shi, W. and Bryant, S.H. (2010) The NCBI BioSystems database. *Nucleic Acids Res.*, **38**, D492–D496.
40. The UniProt, C. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
41. Reuter, J.S. and Mathews, D.H. (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, **11**, 129.
42. Letunic, I. and Bork, P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.*, **44**, W242–W245.
43. Park, J.K., Das, T., Song, E.J. and Kim, E.E. (2016) Structural basis for recruiting and shuttling of the spliceosomal deubiquitinase USP4 by SART3. *Nucleic Acids Res.*, **44**, 5424–5437.

44. Bai, Y., Auperin, T.C., Chou, C.Y., Chang, G.G., Manley, J.L. and Tong, L. (2007) Crystal structure of murine CstF-77: dimeric association and implications for polyadenylation of mRNA precursors. *Mol. Cell*, **25**, 863–875.
45. Krissinel, E. and Henrick, K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774–797.
46. Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T. and Ben-Tal, N. (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.*, **44**, W344–W350.
47. Holm, L. and Rosenstrom, P. (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res.*, **38**, W545–W549.
48. Mitrovich, Q.M. and Guthrie, C. (2007) Evolution of small nuclear RNAs in *S. cerevisiae*, *C. albicans*, and other hemiascomycetous yeasts. *RNA*, **13**, 2066–2080.
49. Ivashchenko, A.T., Tauasarova, M.K. and Atambaeva, S.A. (2009) Exon-intron structure of genes of fungi genomes. *Mol. Biol. (Mosk)*, **43**, 28–35.
50. Ni, T., Yang, W., Han, M., Zhang, Y., Shen, T., Nie, H., Zhou, Z., Dai, Y., Yang, Y., Liu, P. *et al.* (2016) Global intron retention mediated gene regulation during CD4+ T cell activation. *Nucleic Acids Res.*, **44**, 6817–6829.
51. Zhan, X., Yan, C., Zhang, X., Lei, J. and Shi, Y. (2018) Structures of the human pre-catalytic spliceosome and its precursor spliceosome. *Cell Res.*, **28**, 1129–1140.
52. Ulrich, A.K. and Wahl, M.C. (2017) Human MFAP1 is a cryptic ortholog of the *Saccharomyces cerevisiae* Spp381 splicing factor. *BMC Evol. Biol.*, **17**, 91.