Evolution, Function and Structure of the Splicing Factor PRPF39

Inaugural-Dissertation

to obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat)

submitted to the Department of Biology, Chemistry and Pharmacy

of Freie Universität Berlin

by

FRANCESCA DANIELLE DE BORTOLI

2019

This work was carried out in the period of January 2015 to January 2019 under the supervision of Prof. Dr. Florian Heyd at the Institute of Chemistry and Biochemistry, Freie Universität Berlin, Germany.

First Reviewer: Prof. Dr. Florian Heyd
Institute for Chemistry and Biochemistry
Laboratory of RNA-Biochemistry
Takustraße 6
14195 Berlin, Germany

Second Reviewer: Prof. Dr. Markus C. Wahl
Institute for Chemistry and Biochemistry
Laboratory of Structural Biology
Takustraße 6
14195 Berlin, Germany

Date of defense: 14.05.2019

1 Contents

3 Abstract

Beside 5'-capping and 3'-polyadenylation, splicing is one of the essential steps in the processing of most protein-coding genes in higher eukaryotes. It is catalyzed by the spliceosome, a large and dynamic RNA-protein molecular machine that encompasses five core components, the U1, U2, U4, U5 and U6 snRNPs. For each splicing reaction, a spliceosome is assembled anew in a stepwise manner. Spliceosomes must accurately recognize each splice site, as a single mistake can result in the production of a non-functional and potentially toxic protein. The crucial step of exon definition is facilitated by the U1 and U2 snRNP during early splicing. Cryo electron microscopy structures of early spliceosomal complexes in yeast have shown that the Prp39/Prp42 heterodimer is a crucial scaffolding subcomplex. It acts as a hub for multiple protein-protein interactions for example the contact between the U1 and the U2 snRNP, indicating that the Prp39/Prp42 heterodimer is important for the precise spatial positioning of the U1 and U2 snRNPs relative to each other. Interestingly there is no homolog for Prp42 in higher eukaryotes.

PRPF39 is largely unstudied in higher eukaryotes and came to our attention because it is alternatively spliced in a differential manner in murine naïve vs. memory T-cells. I could show that an alternative exon is included in a differential manner. This can control PRPF39 expression by NMD in a tissue- and activation-dependent manner in mice and human, suggesting a role in adapting splicing efficiency to cell type specific requirements. Furthermore, I solved the crystal structure of murine PRPF39 at 3.3 Å resolution. The protein is largely α-helical and the structure shows the protein to be organized as a homodimer. Dimerization in solution could be confirmed with further biophysical assays. The mode of PRPF39 homodimerization is strikingly similar to heterodimerization of Prp39 and Prp42 in yeast. Structure guided point mutations could completely abolish dimerization and by using the monomeric PRPF39 mutants I could show that the monomer has a detrimental effect on splicing *in vitro*. Based on a structural comparison of murine PRPF39 and the yeast heterodimer, we performed a phylogenetic analysis showing, that organisms with a Prp39 homodimer have a substantially shortened U1 snRNA compared to organisms with a Prp39/Prp42 heterodimer. Our analysis indicates that a shortened U1 snRNA accompanied by a PRPF39 homodimer was crucial in the evolutionary development of more complex splicing. This observation is unexpected, as fewer splicing factors usually mirror lower splicing complexity and not the opposite.

Taken together, my results reveal the structural and functional implications of murine PRPF39 on splicing. The data suggests, that a PRPF39 homodimer acts to substitute the Prp39/Prp42 heterodimer observed in yeast. Additionally, the reduction in RNA and protein

complexity of the U1 snRNP may have been crucial in allowing highly complex and sophisticated splicing regulation across species.

4 Zusammenfassung

Spleißen ist neben 5‘-Capping und 3’-Polyadenylierung einer der essenziellen Schritte bei der Prozessierung von prä-mRNA in höheren Eukaryoten. Das Spleißen wird vom Spleißosom katalysiert, einer großen und dynamischen RNA-Protein-Maschinerie. Das Spleißosom umfasst fünf Kernbestandteile: Die U1, U2, U4, U5 und U6 snRNPs. Bei jedem Spleißzyklus wird das Spleißosom schrittweise auf dem prä-mRNA Substrat von Grund auf neu gebildet. Spleißosome müssen dabei jede Spleißstelle genau erkennen, da ein einzelner Fehler zur Produktion von nicht-funktionellem und möglicherweise toxischem Protein führen könnte. Während des frühen Spleißens wird der elementare Schritt der Exon-Definition durch die U1 und U2 snRNPs ermöglicht. Studien mit Kryoelektronenmikroskopie haben bei frühen spleißosomalen Komplexen in Hefe gezeigt, dass das Prp39/Prp42 Heterodimer ein essenzieller Faktor in der räumlichen Anordnung des U1 snRNP zum U2 snRNPs ist. Des Weiteren ist das Prp39/Prp42 Heterodimer eine Interaktionsschnittstelle für viele weitere Proteine, welche das Spleißen beeinflussen können. Dabei gibt es interessanterweise in höheren Eukaryoten kein homolog für Prp42.

PRPF39 ist in höheren Eukaryoten bisher kaum erforscht. Wir sind auf PRPF39 aufmerksam geworden, weil es in naïven T-zellen und T-Gedächtniszellen aus Mäusen unterschiedlich alternativ gespleißt wird. Ich konnte zeigen, dass ein alternatives Exon unterschiedlich stark inkludiert wird. Dadurch wird über NMD die PRPF39 Expression in einer gewebe- und differenzierungs-spezifischen Art kontrolliert. Dies impliziert eine Rolle dieses Spleißereignisses bei der Regulierung der Spleißeffizienz gemäß den Anforderungen der Zelle. Des Weiteren habe ich die Struktur von PRPF39 bei einer Auflösung von 3.3 A gelöst. Das Protein ist größtenteils α-helikal und liegt als Homodimer vor. Weitere biophysikalische Experimente zeigten auch die Dimerisierung in Lösung auf. Die Dimerisierungsart ist sehr ähnlich zu der des Hefe Pr39/Prp42 Heterodimers. Durch strukturbasierte Punktmutationen konnte ich die Dimerisierung komplett auflösen und einen inhibierenden Effekt des Monomers auf das Spleißen zeigen. Basierend auf einem Vergleich des Maus-Homodimers und des Hefe-Heterodimers haben wir eine phylogenetische Analyse durchgeführt. Diese zeigt, dass Organismen mit einem Homodimer im Vergleich zu Organismen mit einem Heterodimer eine stark verkürzte U1 snRNA haben. Korreliert man diese Daten mit der Spleißkomplexität von Organismen, wird deutlich, dass eine kürzere U1 snRNA und ein PRPF39 Homodimer vorteilhaft für komplexeres Spleißen sind. Dies entspricht nicht den Erwartungen, da in der Regel weniger Spleißfaktoren einfacheres und nicht komplexeres Spleißen verursachen.

Zusammenfassend zeigen meine Ergebnisse die strukturellen und funktionellen Auswirkungen von PRPF39 auf das Spleißen. Die Daten deuten an, dass das PRPF39 Homodimer in höheren Eukaryoten das Prp39/Prp42 Heterodimer aus der Hefe substituieren kann. Des Weiteren scheint eine reduzierte RNA- und Proteinkomplexität des U1 snRNP über viele Spezies hinweg unverzichtbar gewesen zu sein, um komplexeres hoch reguliertes Spleißen zu ermöglichen.

5 Introduction

The cell has a wide molecular repertoire, of which three types of macromolecules stand out especially, due to their functions: Deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins. The interplay between these macromolecules is essential for life and has been described as the central dogma of molecular biology. According to this dogma, our genetic information is encoded on our DNA containing the information required to build, maintain and propagate organisms. The DNA is then transcribed into RNA which acts as a coding intermediate that can then be translated into proteins, the functional final product (Crick 1958). This flow of information was originally believed to be unidirectional. However, with new scientific findings like the discovery of retroviruses that use RNA as a template for DNA synthesis (Temin and Mizutani 1970) or the multiple functions of non-coding RNAs this simple picture is no longer enough to satisfy the ever complex interplay of these key macromolecules.

A mature messenger RNA (mRNA) consists of a sequence of nucleotides (nt) that correspond to the amino acid sequence in the resulting protein. This is called the coding region. The additional sequences on either side of the coding region are called 5' untranslated region (UTR) and 3' UTR. The UTRs frequently contain regulatory sequences that can control the stability, cellular location and translational activity of the transcripts (Mayr 2017, Leppek 2018).

However, a precursor mRNA (pre-mRNA) must undergo several steps of processing before it can be exported into the cytoplasm as mature mRNA. Amongst the processing steps is the addition of a 7-methylguanosine cap at the 5' end. The m7G cap (m7GpppN) is important for mRNA export into the cytoplasm, initiation of protein synthesis and the stabilization of the mRNA (McCracken 1997). In the step of polyadenylation, the transcript is cleaved at a specific site and a tail comprised of 100-200 adenosine monophosphate residues is added by the poly(A) polymerase (Colgan and Manley 1997).

The third important step in pre-mRNA processing is called splicing. Most genes in higher Eukarya contain additional noncoding sequences interspersed between the coding sequences. The noncoding parts are called introns and the coding segments are called exons. To allow proper translation the intronic regions need to be excised from the pre-mRNA under subsequent ligation of the exons (**Figure 1**). Splicing was first observed in 1977 (Berget 1977, Chow 1977).

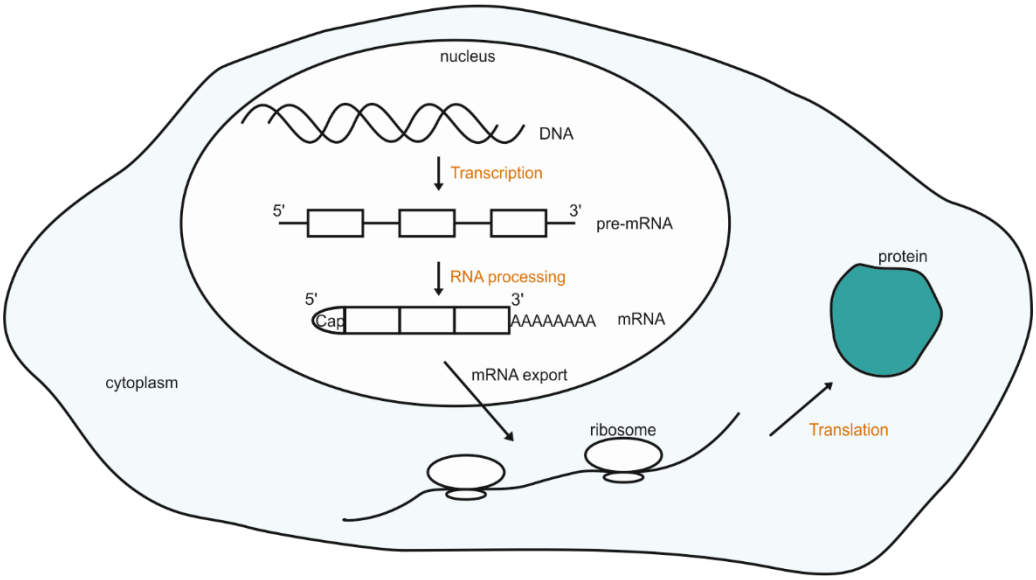

Figure 1 Gene expression in eukaryotes. DNA is transcribed to pre-mRNA containing exons (boxes and introns (lines). After processing (capping, polyadenylation and splicing) it can be exported into the cytoplasm. There translation and protein production is performed by the ribosomes.

The process of nuclear pre-mRNA splicing is highly conserved from yeast to human and can be catalyzed by the spliceosome, a multi megadalton protein RNA machinery. However, the level of splicing complexity can vastly vary between different organisms. Almost all human genes contain introns. In contrast, *Saccharomyces cerevisiae* only has introns in 4.5 % of its genes (Ivashchenko 2009), furthermore, the yeast introns are generally of relatively short size (100-400 nt) with a rather fixed length of approximately 120 nt.compared to the human introns varying from 100 to 500000 nt (Rowen 2002, Ast 2004).

After processing the now mature RNA can be exported into the cytoplasm. Once the mRNA has reached the cytoplasm, translation can take place catalyzed by the ribosome, a complex apparatus made up of proteins and RNA (Figure 1).

5.1 Alternative splicing

When analyzing the human genome, it becomes apparent that the number of protein coding genes (20,000-25,000) cannot account for the much larger amount of 80,000 to 120,000 translated protein products (Liang 2000, Yura 2006, Nilsen and Graveley 2010), assuming a linear relationship in which one protein results from one gene. This discrepancy could possibly be explained by alternative splicing, which occurs in up to 95 % of all human genes (Nilsen and Graveley 2010) emphasizing the importance of this mechanism to the expansion of the proteome. During alternative splicing, exons can be

incorporated in a differential manner, to generate several variable mRNA isoforms from a single pre-mRNA species (Black 2003).

With the help of RNA sequencing technologies, we are able to detect a profusion of splice variants on RNA level. But up until now, it is only poorly studied to which extent these transcripts are actually translated into proteins. Even though we can now find numerous alternative splicing events with RNA sequencing, it remains challenging to functionally characterize the events, as up until now it is still mostly done individually for each case. However, a recent large scale study was performed that showed that isoforms generated by alternative splicing have different interaction profiles, with less than 50 % of their interaction partners in common (Yang 2016).

The most common mode of alternative splicing is the exclusion or inclusion of an entire exon. In some cases, alternative 3' or 5' splice site within exons or introns can be used, which can either shorten or lengthen an exon respectively. More rarely, a complete intron can be included in the final mRNA. Sometimes a special case of exon skipping is present, in which two adjacent exons are included only in a mutually exclusive manner (Keren 2010). Importantly, not only one alternative splice event can occur in a single transcript, but multiple different combinations are possible, even further expanding the plethora of possible different isoforms from one gene (**Figure 2**).

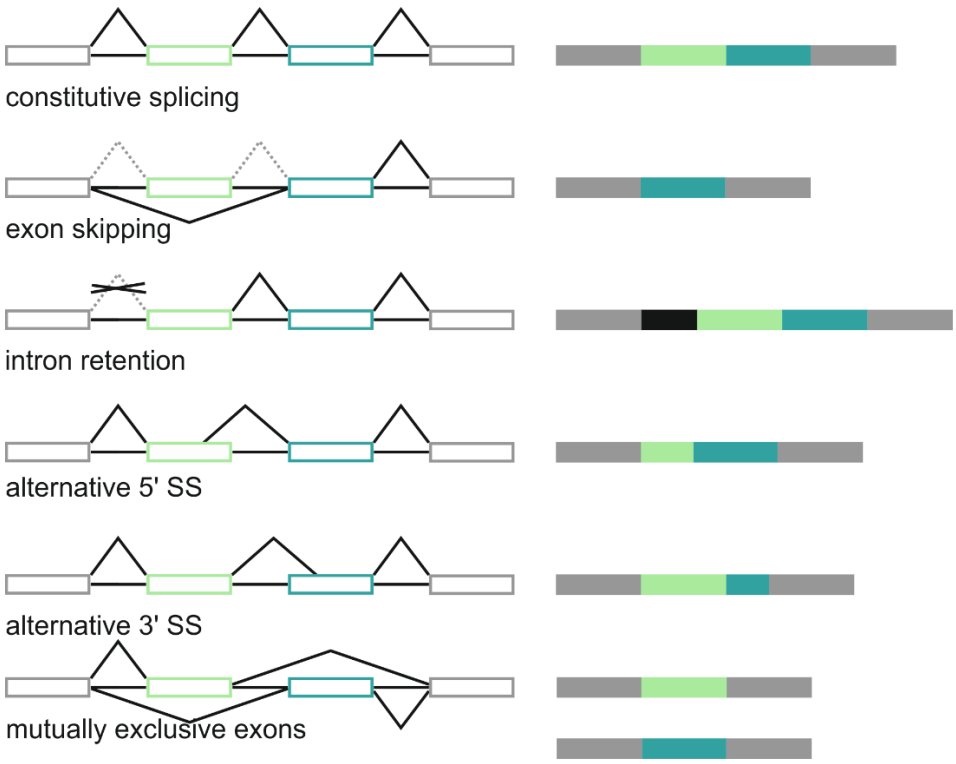

Figure 2 Alternative splicing of pre-mRNA. Five basic types of alternative splicing are included. The exons are indicated as boxes and the introns are show as thick lines. The possible splice site usage is delineated by thin lines with the spliced mRNA variants depicted on the right side.

The different mRNA isoforms can give rise to proteins that differ from each other in their amino acid sequence, and thus also in their chemical and biological properties. Alternative mRNA isoforms however do not always result in different proteins, in some cases inclusion of an exon or intron with a premature stop codon can lead to nonsense mediated decay (NMD). This is one of the many RNA degradation machineries that ensures that mis-spliced mRNA variants are efficiently degraded before they can be translated into proteins that could potentially be toxic to the organism. In some cases, an alternative exon that harbors a premature stop codon can act as a poison exon controlling the amount of protein produced by regulating the alternative splicing of this particular exon (Kervestin and Jacobson 2012) (**Figure 3**).

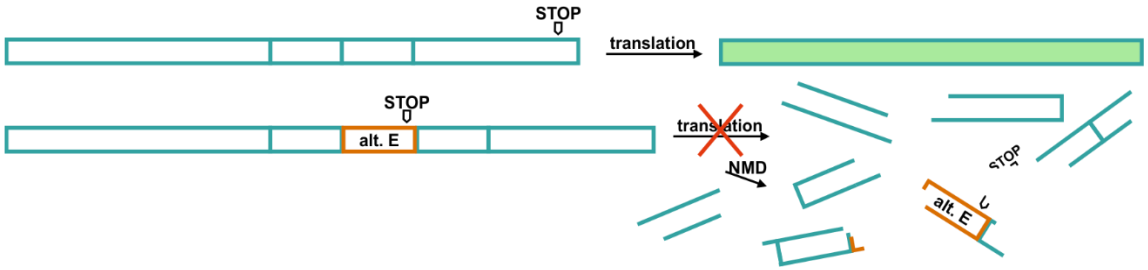

Figure 3 Nonsense mediated decay. After constitutive splicing RNA is translated into protein (filled box). Upon inclusion of an alternative exon with a premature stop codon, the NMD machinery is recruited which results in degradation of the pre-mRNA.

Alternative splicing is known to be an essential switch in gene expression that can influence many cellular and developmental processes like sex determination, apoptosis, axon guidance, cell excitation and contraction (Black 2003). Because of its importance in so many cellular processes, a tight regulation of alternative splicing is necessary to ensure the well-being of the organism.

5.2 Regulation of Alternative Splicing

Splicing can be regulated by *cis*- and *trans*-acting factors. *Cis*-acting factors are sequences in the RNA that can bind *trans*-acting proteins thus regulating splicing by enhancing or decreasing specific splice site usage (**Figure 4**).

Cis-acting elements can be categorized into four different types depending on their location in the pre-mRNA and their effect on splice site usage: exonic splicing silencers (ESS), exonic splicing enhancers (ESE), intronic splicing silencers (ISS) and intronic splicing enhancers (ISE) (Black 2003).

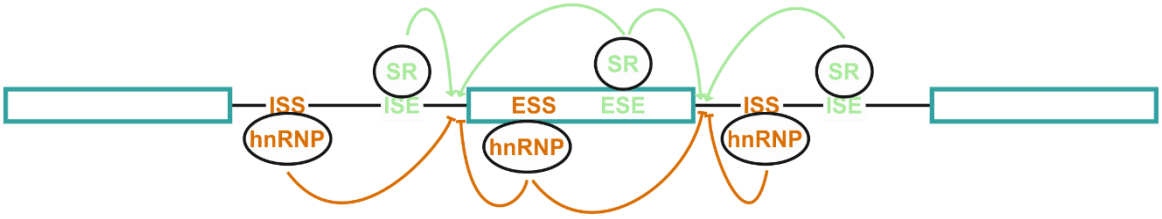

Figure 4 Influence of *cis*- and *trans*-acting factors during splicing. *Trans*-acting factors recognize and bind to *cis*-acting elements. Orange T-bars indicate suppression of a particular splice site, and pale green arrows promotion.

The best characterized protein families of the *trans*-acting factors are the serine/arginine-rich (SR) proteins and the heterogenous nuclear ribonucleoprotein particles (hnRNPs). SR-proteins mostly act as positive regulators of splicing, while hnRNPs are mostly negative regulators of splicing (Busch and Hertel 2012, Fu and Ares 2014).

Alternative splicing is regulated by many different cellular processes and in a development and tissue specific manner. For example, cell stage specific splicing was characterized during neuronal development (Raj and Blencowe 2015), and in differentiated cells in response to cellular signaling and cellular stresses (Biamonti and Caceres 2009, Martinez and Lynch 2013) like temperature changes (Preussner 2017), UV-light (Sharma and Lou 2011) or infections (Shin and Manley 2004).

5.2.1 Tissue Specific Alternative Splicing

As it is of special relevance to this study, I will focus on regulation of alternative splicing in a tissue specific manner. Alternative splicing has been widely implicated in specifying tissues. As previously described (chapter 5.2), *cis*- and *trans*-acting elements are largely responsible for regulation of splicing. However, when looking at splicing in a tissue specific manner, it is clear, that the differences of splicing in different tissues of the same organism have to be based on *trans*-acting elements, as the *cis*-acting environment is the same in every cell of an organism. Each cell-type has its own specific composition of SR proteins and hnRNPs, with their relative abundance in each tissue determining the respective alternative splicing patterns (Chen and Manley 2009).

In addition, there are some well-known examples for tissue specific expression of proteins, most prominently NOVA, RBFOX and ESRPs (Ule 2005, Warzecha 2010, Lovci 2013). A genome wide study compared RNA sequencing data from nine different tissues and ten different species. (Barbosa-Morais 2012). Overall splicing complexity increased with the complexity of the organism. When comparing alternative splicing in different organs and species, the splicing patterns showed more conservation between different organs in one

species, than between the same organ in different species (Barbosa-Morais 2012). This is supported by the fact that the *cis*-acting environment remains the same in all tissues of an organism with only differential expression of *trans*-acting factors influencing alternative splicing.

5.3 The Mechanism of Nuclear pre-mRNA Splicing

Introns can be classified into four major classes with different splicing mechanisms. Transfer RNA (tRNA)/archaeal introns, autocatalytic group I and group II introns and spliceosomal introns (Haugen 2005).

tRNA introns are generally very short (14-160 nt) and can be found in tRNA genes of Archaea and Eukarya. The introns do not share any sequence homology, but have very conserved locations, mostly at the anticodon stem. tRNA introns are enzymatically excised by a cut-and-rejoin mechanism that requires adenosine triphosphate (ATP) an endonuclease and a ligase (Calvin and Li 2008).

Group I splicing introns are also small (250-500 nt) and self-splice using a distinctive two-step transesterification pathway with a guanosine as a cofactor. This type of splicing can be found in bacteria and eukaryotes. Most of the identified type I introns are encountered in Eukarya with a special enrichment in plants, algae and fungi (Haugen 2005).

Like group I introns, group II introns catalyze their own excision, but are phylogenetically unrelated to group I introns. Group II introns can be found in some bacteria and organellar genomes of plants, fungi, protists and some animals and self-splice in a mechanism different from group I introns but similar to the one used by the spliceosome. (Toor 2008, Chan 2012).

Spliceosomal or pre-mRNA introns are excised through a mechanism that is very similar to that of the group II introns and are likely to be evolutionarily related to each other (Irimia and Roy 2014). The introns are removed, and the exons are ligated in two sequential S_N2-type transesterification reactions (Query 1994, Will and Luhrmann 2011). Introns in pre-mRNAs are highly variable in size with minimal conservation in their secondary and tertiary structure. One of the major challenges during splicing is splice site definition, especially in higher eukaryotes, considering the highly variable sequences and lengths of the introns. Accurate splice site definition is crucial, as an error in just one nt can be devastating, causing a shift in the reading frame resulting in a completely different or no protein product. Intron and exon definition is mainly facilitated by short consensus sequences, namely the 5’ splice site, the 3’ splice site and the branch point (BP) (**Figure 5**). During spliceosome

assembly, consensus sequences can be recognized with a subsequent assembly of the spliceosome on the pre-mRNA. Comparing yeast and metazoan organisms, some differences in the consensus sequences become apparent (**Figure 5**) (Wahl 2009, Will and Luhrmann 2011). In *S. cerevisiae,* we see a high sequence conservation in the three *cis*-elements with at least 90 % conservation (Spingola 1999). In higher Eukarya, the conservation is lower, but a nt pattern is still identifiable. Furthermore, just upstream of the 3' splice site lies another highly conserved element of 10-12 nt, the polypyrimidine tract (PPT), which is composed of mainly pyrimidines.

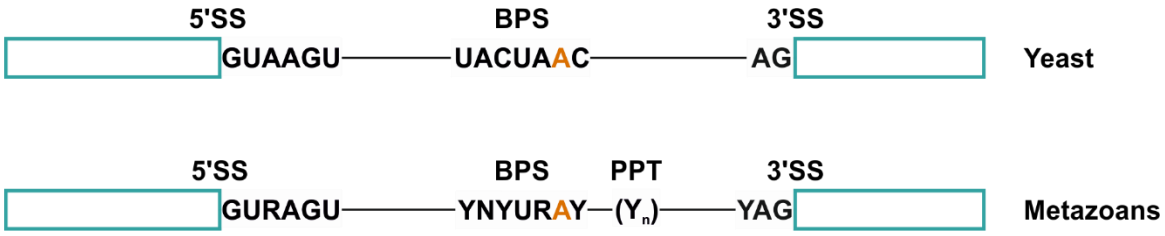

Figure 5 Conserved sequence elements found in introns from metazoans and yeast. Exons and introns are shown as boxes and lines, respectively. The BP adenosine is highlighted in orange. Y and R stand for pyrimidine and purine and the polypyrimidine tract is represented as (Y_n)

In the first step, the oxygen of the 2' OH group of the BP adenosine performs a nucleophilic attack on the phosphodiester bond of the 5' splice site. As a result of the first transesterification reaction, a free 3' OH group at the 3' terminal nt of the 5' exon and a 5'-2' phosphodiester bond between the 5' splice site guanosine and the BP guanosine are formed. In the second step, the 3' OH group of the free 5' exon attacks the phosphodiester bond at the 3' splice site of the lariat intermediate containing the intron and the downstream exon. In this step the 5' and 3' exons are joined, and the intron is excised in form of a lasso-shaped lariat (**Figure 6**).

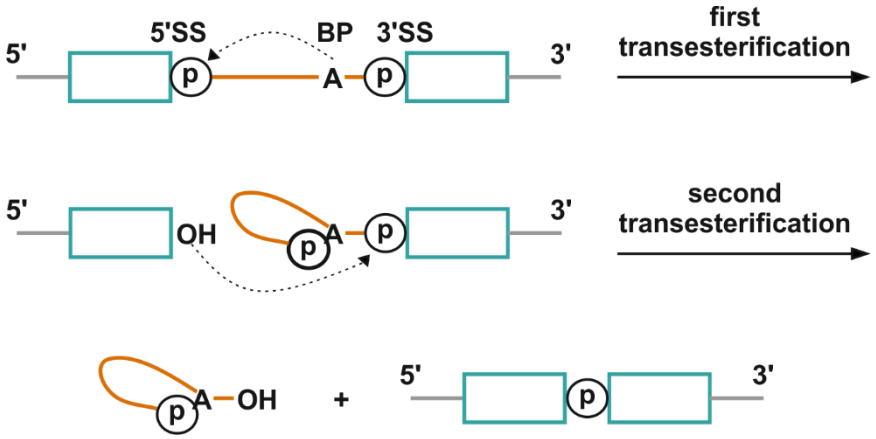

Figure 6 Schematic representation of the two-step mechanism of pre-mRNA splicing. Boxes and solid lines represent the exons and introns, respectively. The branch site adenosine is indicated by the letter "A" and the phosphate groups by the letter "p" inside a circle. The dotted arrows show the nucleophilic attacks at the phosphodiester bond at the 5' and 3' splice site during splicing.

The lariat intron is then debranched and typically degraded but can also be a source of regulatory RNAs (Carthew and Sontheimer 2009, Voinnet 2009), while the mRNA is exported to the cytoplasm for translation (Brow 2002).

Chemically, the process of splicing seems very simple. But due to the scarce information contained in the short consensus sequences, many *trans*-acting elements are necessary for proper splice site selection. Additionally, pre-mRNA folding is required to bring together the splice site to allow for accurate splicing. These *trans*-acting factors are assembled in a stepwise manner and form the highly dynamic macromolecular machine called the spliceosome.

5.4 The Spliceosome

The human spliceosome is a highly complex assembly made up of more than 200 components in humans (Agafonov 2011). In contrast, the yeast system is made up of significantly fewer components than the human spliceosome (Fabrizio 2009) and has been viewed as a simplified model system that lends itself more willingly to studies on the core spliceosome. The major building blocks of the spliceosome are five small nuclear ribonucleoproteins (snRNPs). The core components of snRNPs are small nuclear uridine-rich RNAs (U1, U2, U4, U5 and U6 snRNAs), seven common Sm or Sm-like proteins arranged around the snRNA forming a heteroheptameric ring, and a variable number of snRNP specific proteins (Will and Luhrmann 2011). The five snRNAs are numbered according to their discovery. The U3 snRNP is not part of the spliceosome but involved in ribosomal RNA processing. U4 and U6 form a di-snRNP through extensive base-pairing of their snRNAs.

The spliceosome assembles on the pre-mRNA in a stepwise manner for each splicing event. However, none of the mentioned particles contains a preformed catalytically active center to carry out the two-step splicing reaction. The catalytic center is formed upon complex rearrangements of the spliceosome which happens anew for each splicing cycle (**Figure 7**). The canonical assembly observed in all eukaryotes is based on intron recognition (Berget 1995). However, in metazoans introns can exceed a length of 250 nt by far, and an alternative assembly pathway has been observed in these cases. This pathway mainly differs from the cross- intron assembly in the early steps of splicing, where the splicing complexes gather across the exons first, in a process called exon definition (Fox-Walsh 2005).

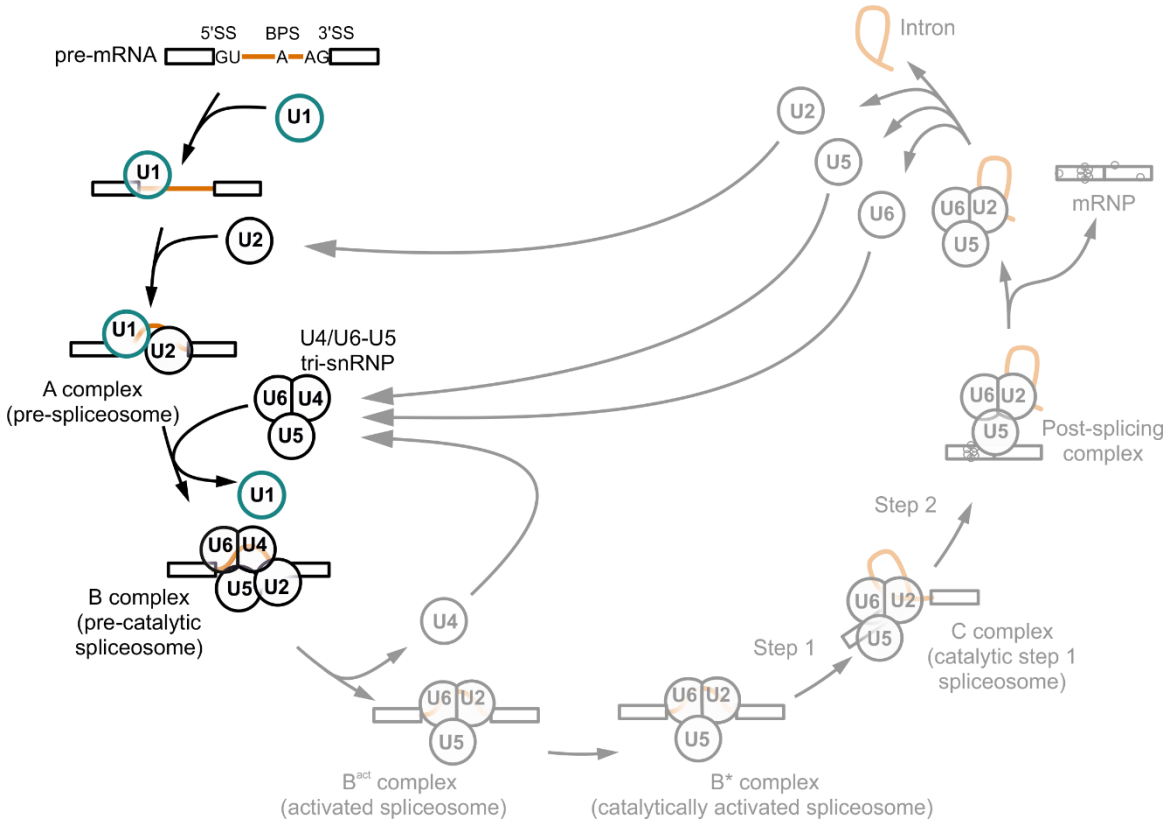

Figure 7 Schematic representation of the splicing cycle. For simplicity the ordered interactions of the snRNPs (indicated as circles), but not those of the non-snRNP proteins are shown. Exon and intron sequences are indicated by boxes and lines respectively. Because the main focus of this thesis lies on the U1 snRNP steps involving this particle are highlighted. Adapted from (Wahl 2009).

The first step of splicing entails the recruitment of the U1 snRNP to the 5' splice site by base pairing of the U1 snRNA with the 5' splice site. After this, two non-snRNP factors, the splicing factor 1 and the U2 auxiliary factor, bind to the BP and the PPT, respectively. These initial interactors form the center of the spliceosomal complex E. In the second step, the U2 snRNP is incorporated in an ATP-dependent manner. By base pairing of the U2 snRNA with the BP, splicing factor 1 is displaced leading to complex A formation, also called the pre-spliceosome (De Conti 2013). Following this event, the pre-assembled U4/U6.U5 tri-snRNP is recruited, resulting in the pre-catalytic complex B. At this point, all five snRNPs necessary for splicing are present but need to undergo major compositional and conformational rearrangements before the spliceosome becomes catalytically active. During this step, the U1 and U4 snRNP are dissociated from the spliceosome respectively by the Prp28 and Brr2 helicases, to give rise to the activated spliceosome (Laggerbauer 1998, Staley and Guthrie 1999). The 5' end of the U6 snRNA substitutes the U1 snRNA in base pairing with the 5' splice site and an extensive base pairing network is formed between the U2 and U6 snRNA bringing the 5' splice site and BP together for the first transesterification reaction. The resulting C complex must be remodeled again, in this case

by the Prp16 helicase, to reposition the splicing intermediates toward each other (Schwer and Guthrie 1992). Then the U5 contacts exon nts downstream of the 3' splice site to align the 5' and 3' exons to allow the second catalytic step. Finally, the exon junction complex is deposited 20 to 25 nt upstream of the exon-exon junction and the mRNA is released in form of a messenger ribonucleoprotein complex and transported to the cytoplasm. The post-spliceosomal complex is disassembled and all of its components are recycled to take part in a new splicing cycle (Martin 2002).

The mechanism of spliceosome assembly across the exon is slightly different, in which the U1 snRNP binds to the 5' splice site downstream of the exon and assists in U2 auxiliary factor recruitment on the PPT upstream of the exon. After this, the U2 snRNP can be incorporated through binding to the BP upstream of the exon. Additionally, SR-proteins are recruited to ESEs within the exon. All of these factors form a stabilizing network across the exon that mediate the crosstalk. Because splicing nevertheless still occurs across an intron, the exon definition state has to subsequently be converted into an intron recognition state to allow for splicing as previously described (Sharma 2008).

5.5 The U1snRNP

As already mentioned, accurate splice site recognition in early spliceosome assembly is crucial. One of the main players in exon definition is the U1 snRNP, recognizing the 5' splice site via base-pairing of the U1 snRNA with the 5' splice site of the pre-mRNA and additionally the U2 snRNA base-pairs with the BP (**Figure 8**).

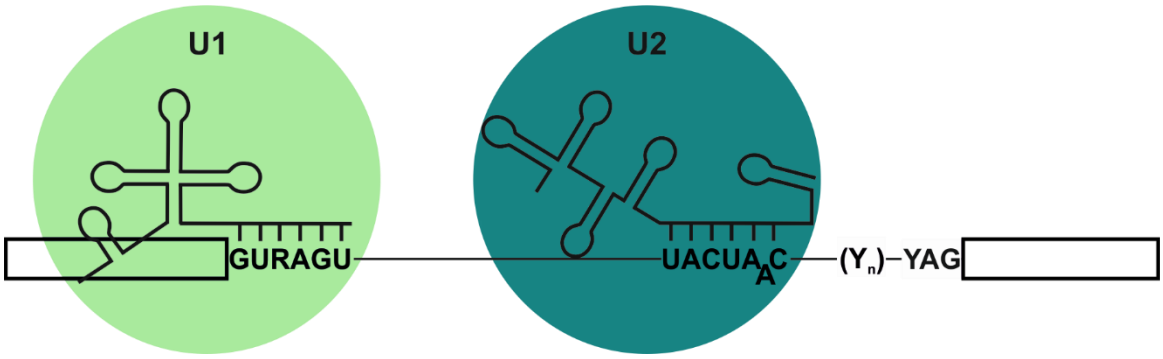

Figure 8 Exon definition during splicing initiation. Exons and introns are shown as boxes and lines. The colored circles indicate snRNPs with schematic representations of their respective snRNAs.

The human U1 snRNP is relatively small, consisting of only a 165 nt long snRNA and ten proteins, including the seven SM proteins common to all snRNPs. The only three proteins specific to the U1 snRNP are U1-60K, U1-A and U1C (Patel and Bellini 2008, Pomeranz Krummel 2009, Buratti and Baralle 2010).

In general, more complex organisms have equally more complex spliceosomes. For example the human spliceosome includes around 80 additional mostly non-snRNP proteins (Fabrizio 2009, Will and Luhrmann 2011) in comparison to yeast, and *Cyanidioschyzon merolae* has an even more dramatically reduced spliceosome missing the U1 snRNP altogether (Stark 2015). However, when comparing the human and yeast U1 snRNP to each other, they seem to pose an exception. The yeast U1 snRNA alone is already approximately 3.5 times larger than its human counterpart (Kretzner 1987, Kretzner 1990). Furthermore, in the yeast system seven additional, stably associated proteins can be found in purified yeast U1 snRNPs, namely Luc7, Nam8, Prp39, Prp40, Prp42, Snu56, and Snu71. Of these, Prp42 and Snu56 have no known human homologs (Gottschalk 1998, Fortes 1999).

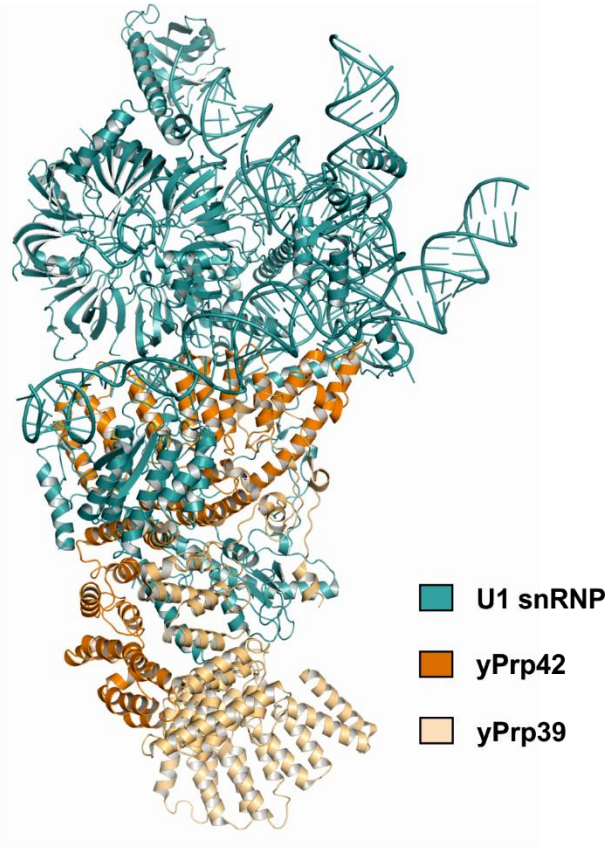

Figure 9 Yeast U1 snRNP structure. The U1 snRNP is shown as cartoon in teal. The heterodimer important for scaffolding is highlighted in orange and pale orange. Modified from PDB entry 5UZ5 (Li 2017).

The first structural information on the U1 snRNP were the crystal structures of the human core U1 snRNP at 5.5 and 4.4 Å resolution (Pomeranz Krummel 2009, Weber 2010). From these structures, even though U1-A and portions of U1C and U1-70K were missing, it was possible to determine that the U1 snRNA forms four stem-loops (SL1 through SL4) and a short α-helix. This confirms the model suggested by Krol et al. (Krol 1990). The heptameric

Sm ring binds to a single-stranded uridine rich site, U1-70K binds to the SL1 with its RNA binding domain and U1C is able to bind the U1 core domain through interactions with the N-terminus of U1-70K. Contact of the U1C to the RNA could be observed at the 5' end of the U1 snRNA where it base-pairs with its pre-mRNA substrate, consistent with its role as a stabilizer in the U1/5' splice site base-pairing (Pomeranz Krummel 2009). X-Ray crystallography can be challenging for large protein RNA assemblies, as it requires highly stable particles that are rigid enough to enable high resolutions. With the recent developments in cryo-EM microscopy however, the structure determination of the spliceosome has progressed quickly.

Recently, the cryo-EM structure of the yeast U1 snRNP at a resolution of 3.6 Å could be solved (Li 2017). This structure reveals the U1 snRNP to be overall arranged in a shape resembling a foot. The U1 snRNP core that forms the ball-and toes region is largely similar to the human core U1 snRNP, with the additionally modeled auxiliary proteins making up the arch and heel region (**Figure 9**).

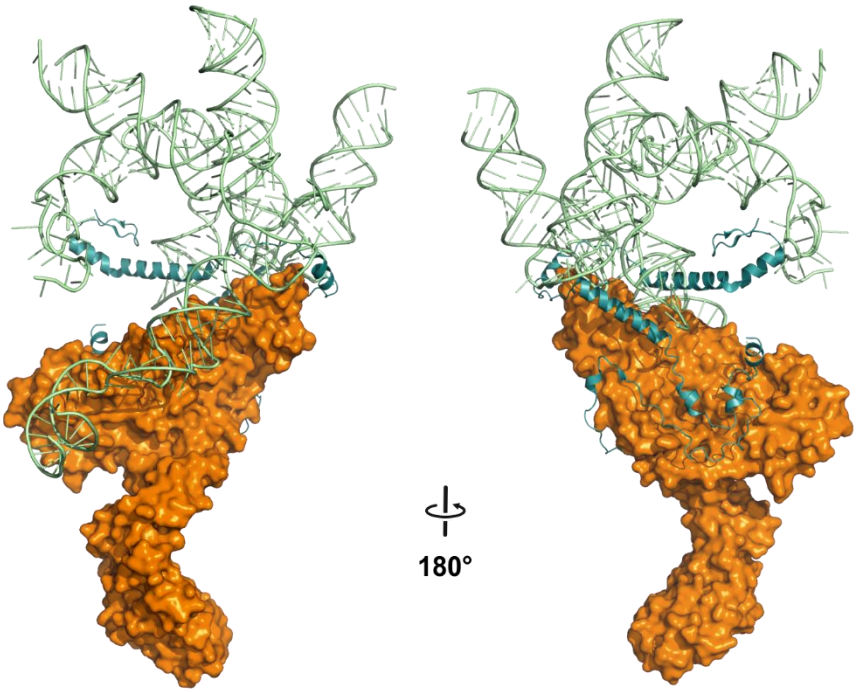

Figure 10 Binding of yPrp42 to the U1 snRNP core. The surface of yPrp42 is shown in orange with a cartoon representation of U1C and the U1 snRNA in teal and pale green, respectively. The main connection between the auxiliary and core U1 snRNP proteins is between the yPrp42 NTD, U1C and the U1 snRNA. Modified from PDB entry 5UZ5 (Li 2017).

The auxiliary proteins yeast proteins Prp39 and Prp42 (yPrp39 and yPrp42) seem to be key players in the U1 snRNP. They both consist of an N-terminal half-a-tetratricopeptide (HAT)-repeat domain. yPrp39 and yPrp42 form a heterodimer mediated over their respective C-termini to form a central scaffold that acts to connect the auxiliary U1 snRNP

proteins to its core. The main interacting regions of core U1 snRNP proteins with the Prp39/Prp42 heterodimer are between the N-terminal domain (NTD) of yPrp42 and the yU1C C-terminal domain (CTD). In addition to this, portions of the much longer U1 snRNA in yeast interact with yPrp42 (**Figure 10**) (Li 2017).

Recently, even more structural information on early splicing emerged. A cryo-EM structure of the yeast pre-spliceosome containing the U1-snRNP, the U2 snRNP and parts of a pre-mRNA substrate could be solved by Plaschka et al. (Plaschka 2018) giving us novel insights into early spliceosome assembly. In this structure, the U1 snRNP and the U2 snRNP respectively bind the 5' splice site and BP sequence of the pre-mRNA and are arranged in a parallel manner to each other forming the A-complex (**Figure 11**).

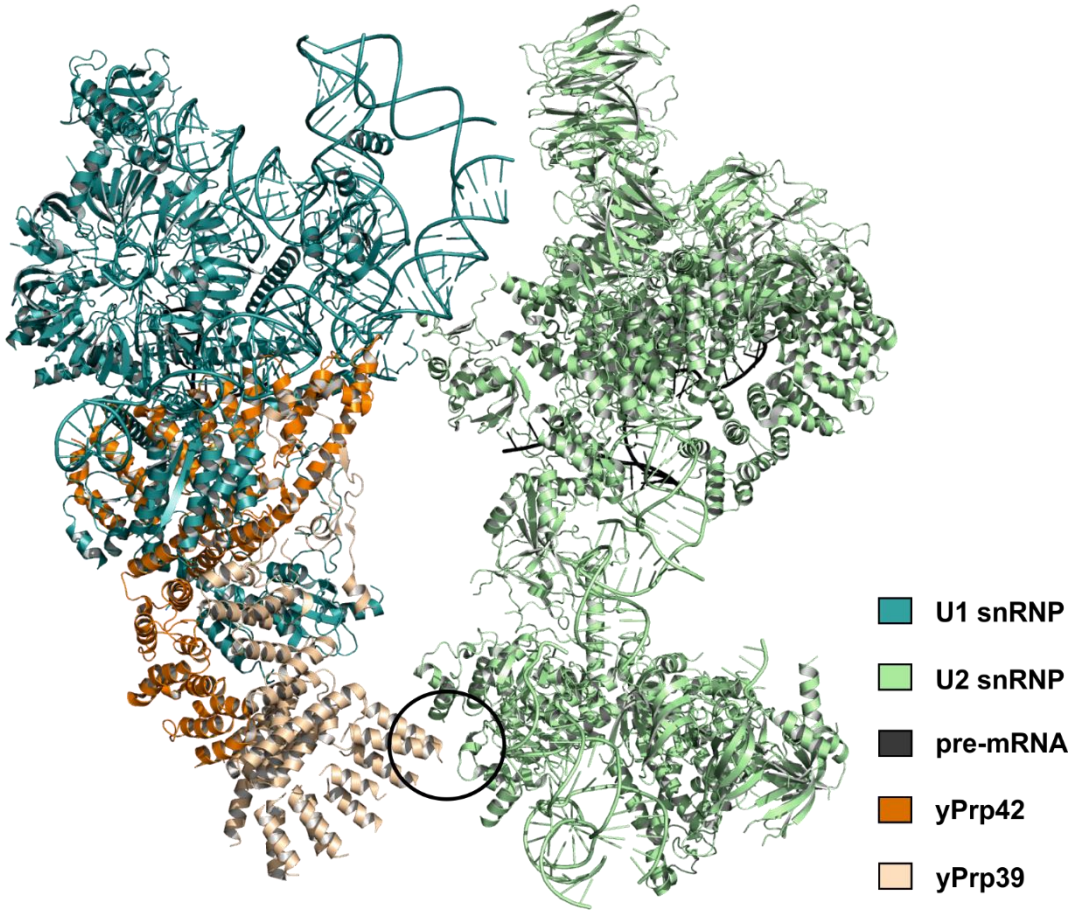

Figure 11 Structure of the yeast pre-spliceosome. U1 and U2 snRNP together with parts of the pre-mRNA are shown in a cartoon representation. The interfacing region between yPrp39 and the U2 snRNP is highlighted with a black circle. For color-coding see inlet. Modified from PDB entry 6G90 (Plaschka 2018).

Two main interfaces can be observed between the two snRNPs. The first one is a more transient interface that can only be found in a subset of cryo-EM images. Here a weak interaction between the yeast specific U1 snRNA stem loop (SL) 3-3 and the U2 SF3b

Rse1 subunit β-propellers B and C and the C-terminus of U2 Sf3a Prp9 can be seen (Plaschka 2018) (**Figure 11**). In the second interface the first two N-terminal α-helices of the yPrp39 stably interact with U2 3' domain subunit of Lea1, the homolog of human U2A' (**Figure 11**).

Another cryo-EM structure of the yeast pre-B complex has been solved, just before the U1 snRNA leaves the spliceosome (Bai 2018).

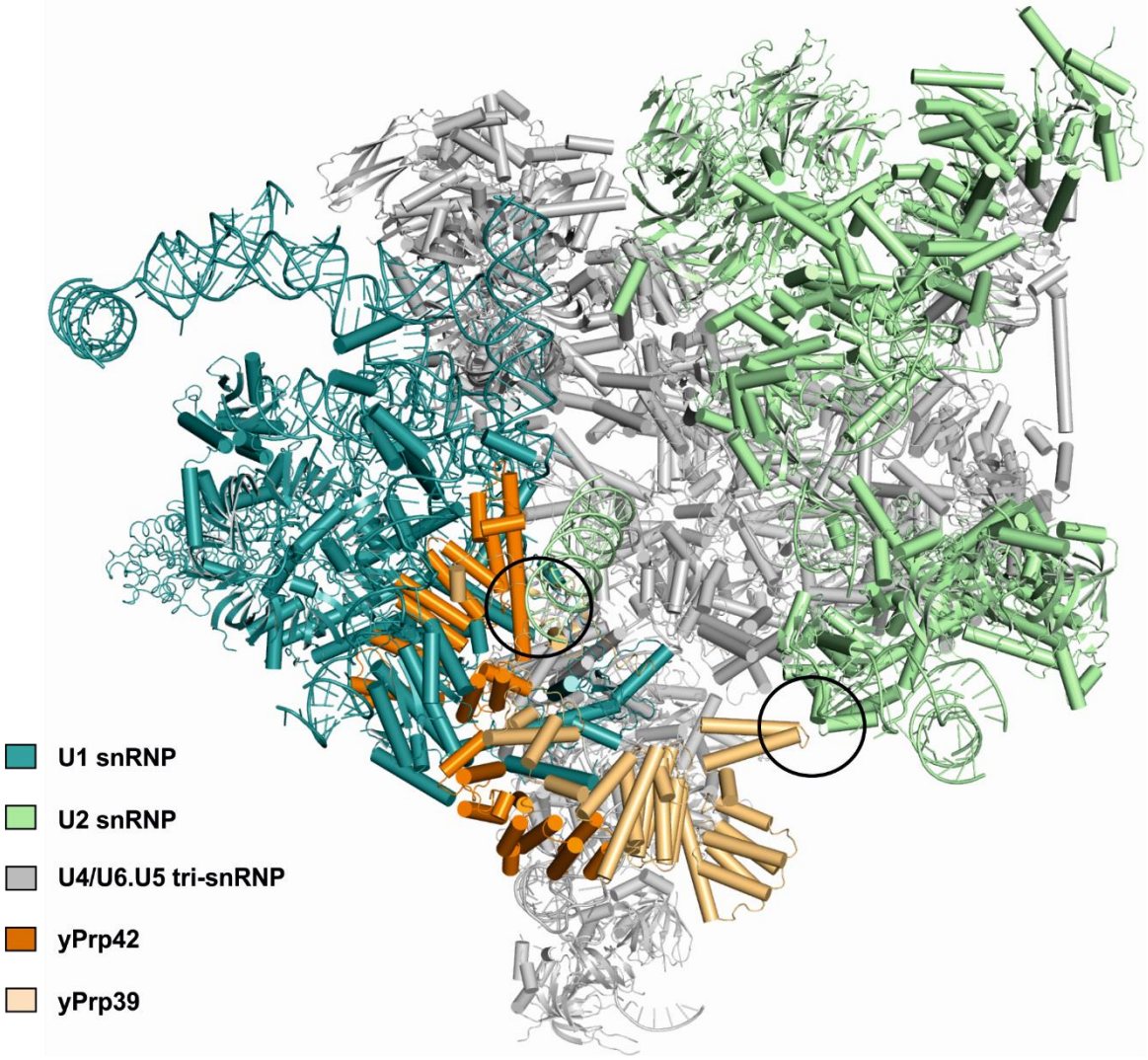

Figure 12 Structure of the yeast pre-B spliceosome. The structure from PDB entries 5ZWN and 5ZWM (Bai 2018) is shown in a cartoon representation. It shows the interaction between yPrp42 and yLea1 already observed (Plaschka 2018) and reveals an additional contact between yPrp39 and the U2 snRNA. For color-coding see inlet.

The yPrp39/yPrp42 heterodimer seems to play an important role in positioning of the U1 and U2 snRNPs toward each other. Here yPrp39 from the U1 snRNP contacts the U2 specific protein yLea1 (human homolog U2A') with its N-terminal domain. In addition to this, a positively charged groove on the surface of the yPrp42 NTD accommodates the U2 snRNA (**Figure 12**)(Bai 2018).

It is highly interesting that the observed contacts in the stable interface between the U1 and the U2 snRNPs in these structures are all involving protein and RNA elements that are either not conserved (yeast U1 snRNA SL 2-2) or have no homolog (yPrp42) in metazoan. Especially the yPrp39/yPrp42 heterodimer seems to play a big role in connecting auxiliary proteins to the U1 snRNP core (Li 2017) and mediating stable contacts between the U1 and U2 snRNP in the A-complex and also the pre-B complex (Bai 2018, Plaschka 2018). Thus, the question of how early spliceosome formation can function in metazoan systems lacking a Prp42 homolog arises.

A first indication how the loss of Prp42 may be compensated in mammals came from Li et al.. 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 (Li 2017).

However, very recently a new structure of the human pre-B spliceosome has been published at 5.7 Å, showing a density lobe assigned as U1 snRNP (Zhan 2018) (**Figure 13**).

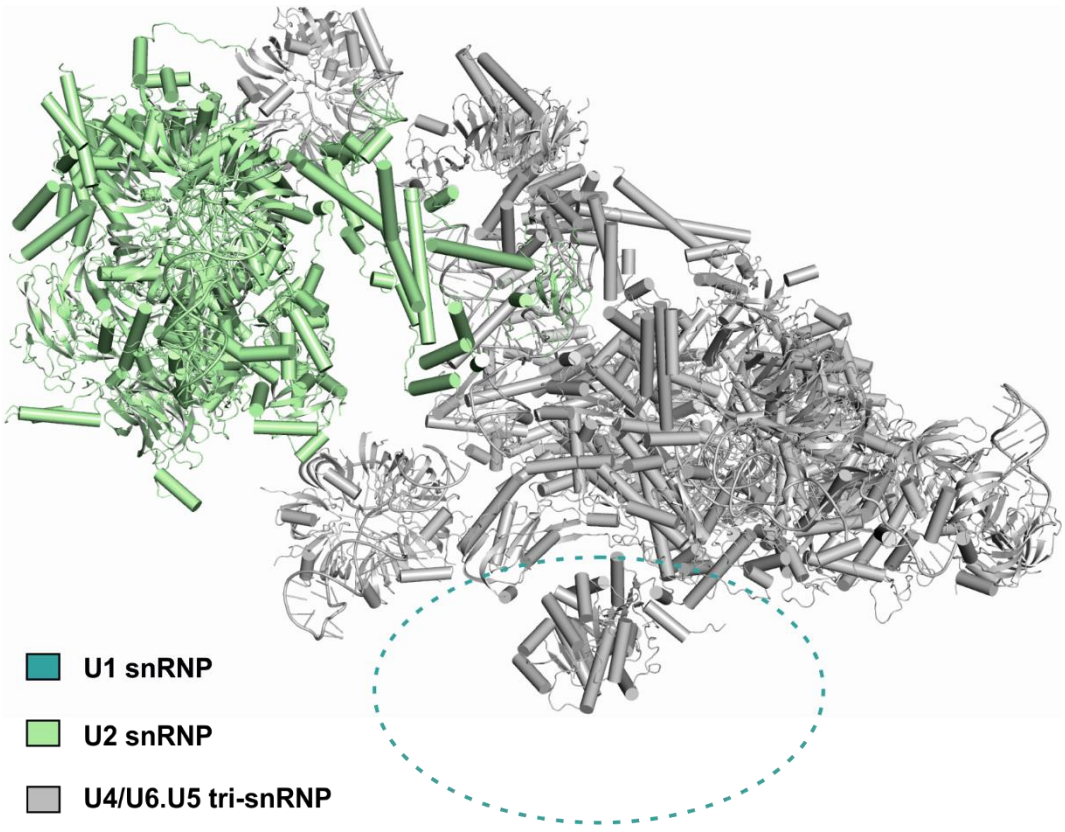

Figure 13 Structure of the human pre-B spliceosome. Cartoon representation of the components and the teal dotted line highlights the area the U1 snRNP is assumed to be. For color-coding see inlet. In the human structure the lobe of density assigned to the U1 snRNP does not contact the U2 snRNP. Modified from PDB entry 6AH0 (Zhan 2018)

In this structure the U1 and U2 snRNP contact the tri-snRNP, with the U1 snRNP binding more flexibly. The U1 snRNP can only be located in the density map without a precise orientation (Zhan 2018). Interestingly here the U1 and U2 snRNPs do not contact each other. There is a gap of approximately 60 to 100 Å, in which density for the U4 Sm ring and the 3'-end sequences of the U4 snRNA can be observed (**Figure 13**) (Zhan 2018).

Even in light of this new structural information, it is unclear how exactly the spatial positioning between the U1 and U2 snRNP occurs during A-complex formation. PRPF39 seems to be a good target to address this question.

5.6 The Spliceosomal Protein PRPF39

Surprisingly very little is known about Prp39 in higher eukaryotes. It has been demonstrated to be essential in diverse human cell lines in independent CRISPR-screens (Blomen 2015, Hart 2015, Wang 2015). This indicates that it has a similar importance in metazoan compared to yeast.

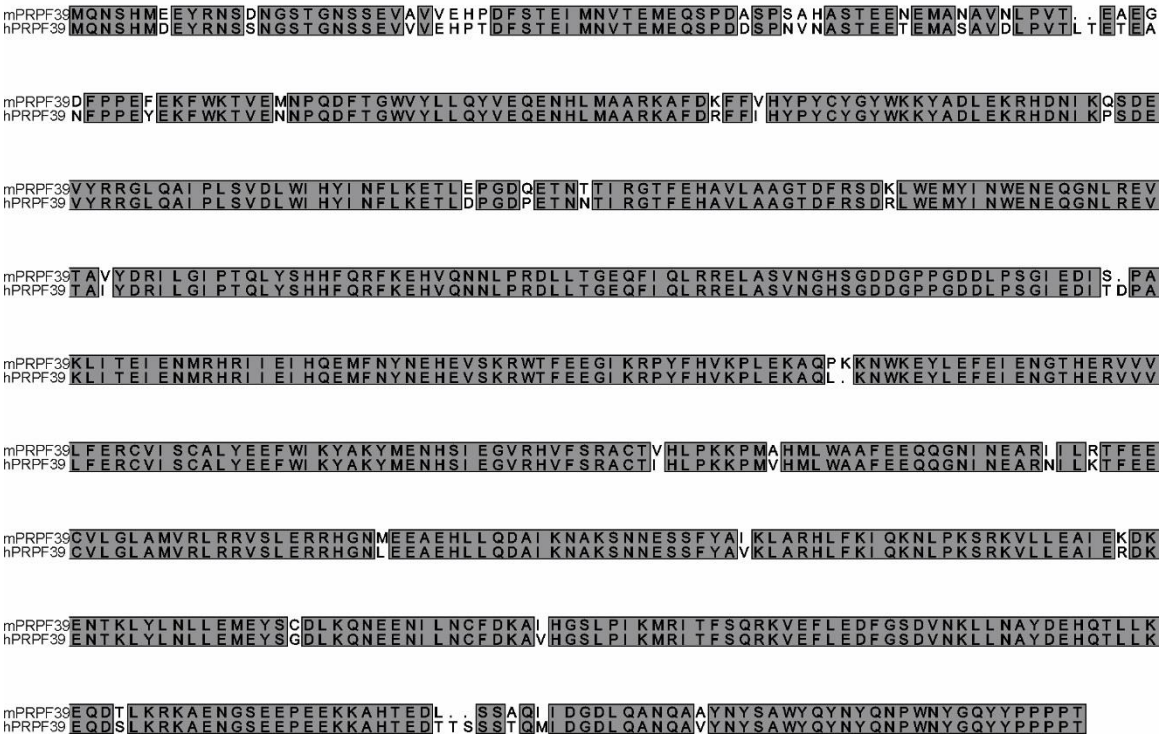

Figure 14 Alignment of human and murine PRPF39 protein sequence. Grey background indicates identical amino acid residues. The proteins are 94% conserved and most differences are in the unstructured N- and C-terminus.

Murine PRPF39 (mPRPF39) initially came to our attention because it is alternatively spliced in a differential manner in murine naïve versus memory T-cells. It shows high similarity to its human counterpart, 94 % amino acid sequence identity and 98 % similarity (**Figure 14**). We were intrigued by the fact, that a splicing factor with such an impact on the structural arrangement of the early-stage spliceosome is so poorly studied.

We therefore decided to analyze mPRPF39 in depth on a structural and functional level. mPRPF39, like both yPrp39 and yPrp42, is predicted to be mostly made up of half a tetratricopeptide (HAT) repeats. HAT repeats are repetitive patterns characterized by three aromatic residues with a conserved spacing. Structurally they are very similar to tetratricopeptide repeats, arranged as two short anti-parallel α-helices connected by a loop (**Figure 15**), however they lack the highly conserved alanine and glycine residues specific to tetratricopeptide repeats.

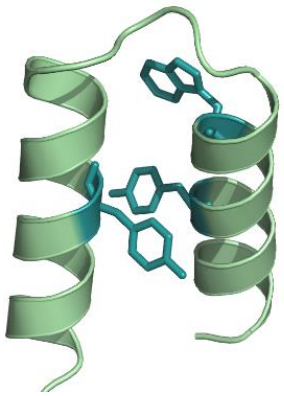

Figure 15 Structure of a HAT repeat. Cartoon representation of a HAT repeat in pale green with the three conserved amino acid residues shown as sticks in teal.

HAT-repeat proteins are mostly involved in protein-protein and protein-RNA interactions and act as scaffolding proteins. Prominent members of the HAT repeat family have been reported to be involved in various RNA processing pathways. For example, SART3 is involved in splicing and translesion DNA synthesis (Huang 2018), RNA14 is important for cleavage during polyadenylation (Paulson and Tong 2012) and yeast Clf1 is part of the NineTeen complex that stabilizes the U6 snRNA in catalytic steps of splicing (Wang 2003).

In most of these cases the HAT repeat proteins act as central interaction hubs for protein-protein interactions and, in some cases, for RNA binding. The protein high chlorophyll fluorescent 107 shows sequence specific binding and remodeling of RNA in *Arabidopsis thaliana* (Hammani 2012). More relevant for this thesis, the already mentioned yPrp42 binds the U1 snRNA SL 2-2 during splicing (Li 2017).

Functionally, Prp39 and Prp42 are very poorly studied. The only studies pertaining to these proteins were performed in the yeast system in the nineties of the last century. In 1994,

yPrp39 was found to be essential for splicing (Lockhart and Rymond 1994). They could show that commitment complex formation only takes place in the presence of yPrp39.

Later, in 1998, yPrp42 was characterized in more detail. Here it was found, that even though yPrp39 and yPrp42 show high sequence homology, they fulfill non-redundant functions. When yeast strains are depleted of yPrp42, splicing is arrested prior to 5' splice site cleavage (McLean and Rymond 1998). Furthermore, without yPrp42, U1 snRNA biogenesis is compromised, resulting in incomplete U1 snRNPs unable to form stable complexes with pre-mRNA *in vitro*. This finding is consistent with yPrp42 scaffolding properties observed in the cryo-EM structures discussed above.

5.7 Aim of This Study

The spliceosome is difficult to analyze as a whole because of its high complexity and dynamic nature. This thesis deals with steps in early spliceosome formation.

At the start of this thesis, very little was known about PRPF39. In particular, very little experimental structural information was available on this protein. The only relevant papers on Prp39 were early studies in the yeast system (Lockhart and Rymond 1994, McLean and Rymond 1998). In the course of my thesis more studies on the early spliceosome have been performed, shedding light on the structural and functional relevance of Prp39 in the yeast system. However, the fundamental question of how splicing could function in a system without a Prp42 homolog remains unclear.

Thus, the overall aim of this work was 1) to gather information, by means of structural and biochemical analyses that will allow us to better understand the role of eukaryotic PRPF39 in early splicing and 2) to determine if Prp39 takes over the function of the lacking Prp42 homolog in higher eukaryotes. To this end, the following specific goals were proposed:

1. Establishment of efficient recombinant production pipeline for murine PRPF39, using different fragments of this protein to increase chances of expression success.
2. Establishment of efficient purification protocols for mPRPF39 to obtain sufficient amounts of highly homogeneous material for further structural and functional studies.
3. Crystallization, structure determination and crystal structure analysis of mPRPF39.
4. Establishment of production of efficiently splicing nuclear extracts from eukaryotic cells for use in functional studies.
5. Structure-based functional analyses, such as targeted mutagenesis followed by *in vitro* splicing assays.
6. Comparison of the yeast system with the metazoan system on a structural and phylogenetic level.

6 Materials and Methods

6.1 Material

6.1.1 Instruments and Consumables

Instrument	Company
Acupuncture needle	Moxom Medical, Germany
Agarose gel chambers + combs	Peqlab, Germany
Äkta purification systems	GE Healthcare, Germany
Beakers	Schott, Germany
Beamline 14.2	HZB, Berlin, Germany
Beamline 14.3	HZB, Berlin, Germany
Beamline P14, Petra III	DESY, Hamburg, Germany
Cell culture tubes	Greiner Bio-One, Germany
Centrifuge Allegra X-15R	Beckman Coulter, Germany
Centrifuge Avanti J-26 XP	Beckman Coulter, Germany
Centrifuge 5417R	Eppendorf, Germany
Centrifuge 5810R	Eppendorf, Germany
Clean bench	Laminair 1.8 Holton, Danmark
Concentrators	Millipore, USA
Conical flasks	Schott, Germany
Cryo loops	Hampton Research, USA
Crystallisation plates	MRC Molecular Dimensions, UK
Crystallisation robot	Zinsser, Germany
Cylinders	Isolab, Germany
Disposable pipettes	Sarstedt, Germany
Dialysis membranes	Spectra/Por, USA
Electroporation cuvettes	Peqlab, Germany
Electroporator EasyjecT Prima	Equibio, UK

Electrophoresis power supplies	BioRad, Germany
Falcon tubes	Greiner Bio-One, Germany
Glass flasks	Schott, Germany
Glass pipettes	Hirschmann, Germany
Heating block HLC	DITABIS, Germany
HT multitron culture shaker	Infors, Switzerland
Ice machine	Ziegra, UK
Incubator	Heraeus, Germany
Inoculation loop	Sarstedt, Germany
Magnetic stirrers	IKA, Germany
Micro scale XS205 Dual Range	Mettler Toledo, Germany
Microscope	Olympus, Germany
Milli-Q synthesis A10	Millipore, USA
Multitron culture shaker	Infors HT, Switzerland
Nanodrop 2000 spectrophotometer	Peqlab, Germany
Needles	Henke Sass Wolf, Germany
Rocking platform	Biometra, Germany
Parafilm	Pechiney Plastic Packaging, USA
pH meter Professional Meter PP-20	Satorius, Germany
Phosphoimager Typhoon FLA 7000	GE Healthcare, Germany
Pipettes	Abimed, Germany
Pipette tips	Sarstedt + Greiner Bio-One, Germany
Polyvinylidene difluoride membrane	Merck, Germany
Power supplies	Bio-Rad, German
Scale XS4002S Delta Range	Mettler Toledo, Germany
Scanner for gel documentation	Epson, Germany
SDS-PAGE chambers, combs, glass plates	BioRad, Germany
Sonificator Sonoplus	Bandelin, Germany

Speed vac concentrator 5301	Eppendorf, Germany
Sterile Filters 0.22 µm	Sarstedt, Germany
Surgical blades	Martin, Germany
Syringes	Braun, Germany
Table centrifuge 5415R	Eppendorf, Germany
Thermo cycler Star 2X Gradient	Peqlab, Germany
Thermo mixer compact	Eppendorf, Germany
Tubes	Sarstedt, Germany
Tunair flasks	Sigma-Aldrich, Germany
Vortex Genie 2	Scientific Industries, USA
Weighting dishes	Roth, Germany

6.1.2 Chemicals

Chemical	Company
2-(N-morpholino)ethanesulfonic acid (MES)	Roth, Germany
2-Propanol	Roth,Germany
4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES)	Roth, Germany
ABsolute QPCR SYBR Green Mix	Thermo Fisher
Acetic acid	Roth, Germany
Acrylamide : bis-acrylamide 37.5 : 1	Roth, Germany
Agarose	Bio&Sell, Germany
Ammonium chloride	Roth, Germany
Ammonium persulfate (APS)	Merck, Germany
Ampicillin (Amp)	Roth, Germany
Amylose resin	New England Biolabs, USA
Boric acid	Roth, Germany
Bovine serum albumin (BSA)	Roth, Germany

Bromophenol blue	Roth, Germany
Calcium chloride dihydrate	AppliChem, USA
Chloramphenicol (Cm)	Roche, Germany
Cobalt (II) chloride hexahydrate	Sigma-Aldrich, Germany
Coomassie Brilliant Blue G250	Serva, Germany
Coomassie Brilliant Blue R250	Serva, Germany
Copper (II) chloride dihydrate	Roth, Germany
Cycloheximide (CHX)	Roth, Germany
D(+)-glucose	Roth, Germany
Desoxyribnucleotides (dNTPs)	Sigma-Aldrich, Germany
Disodium hydrogen phosphate	Roth, Germany
Dithiothreitol (DTT)	Sigma-Aldrich, Germany
DMEM High Glucose	Biowest, Germany
DNaseI (2mg/ml)	Sigma-Aldrich, Germany
Ethanol	Roth, Germany
Ethidium bromide	Roth, Germany
Ethylene-diamine-tetra-acetic acid (EDTA)	Roth, Germany
Fetal bovine serum (FBS)	Biochrome, Germany
Glycerole	Roth, Germany
Glycine	Roth, Germany
Hydrochloric acid	Merck, Germany
Imidazole	Roth, Germany
Iron (III) chloride hexahydrate	Sigma-Aldrich, Germany
Isopropyl β-D-1-thiogalactopyranoside (IPTG)	Roth, Germany
Kanamycine (Kan)	Roth, Germany
Lactose monohydrate	Roth, Germany
Lysozyme (10mg/ml)	Sigma-Aldrich, Germany
Magnesium acetate tetrahydrate	Fluka, Germany

Magnesium chloride hexahydrate	Merck, Germany
Magnesium sulfate heptahydrate	Roth, Germany
Maltose	Merck, Germany
N,N,N',N'-Tetraethylenediamide (TEMED)	Roth, Germany
Nickel (II) chloride hexahydrate	Merck, Germany
Penicillin/Streptomycin	Biowest, Germany
Peptone	Roth, Germany
Polyethylene glycol monomethyl ether 5000	Fluka, Germany
Potassium dihydrogen phosphate	Roth, Germany
RNA Tri-Flüssig	Bio&Sell, Germany
RPMI-medium-1640	Biowest, Germany
Sodium chloride	Roth, Germany
Sodium dodecylsulfate	Roth, Germany
Sodium molybdate dihydrate	Merck, Germany
Sodium selenite pentahydrate	Roth, Germany
Sodium sulfate	Roth, Germany
Tris-(hydroxylmethyl)-aminomethane (TRIS)	Roth, Germany
Tween-20	Roth, Germany
Xylene cyanol	Sigma-Aldrich, Germany
Yeast extract	AppliChem, USA
Zinc sulfate heptahydrate	Roth, Germany
β-Mercaptoethanol	Roth, Germany

6.1.3 Buffers, Media and Stock Solutions

Solution	Composition
5052 (50x)	25 % Glycerol (v/v)
	10 % Lactose monohydrate (w/v)
	2.5 % Glucose (w/v)

Amp	100 mg/ml
Auto-induction media (1L)	95.8 % ZY medium (v/v) 2 % 5052 (50x) (v/v) 2 % M (50x) (v/v) 0.2 % Magnesium sulfate (1M) (v/v) 0.02 % metals (1000x) (v/v)
Buffer D	20 mM HEPES, pH 8.5 100 mM KCl 0.2 mM EDTA 20 % Glycerol 1 mM DTT 0.5 mM phenylmethylsulfonyl fluoride
Cm	34 mg/ml
Coomassie staining solution	42.5 % Ethanol (v/v) 10 % Acetic acid (v/v) 0.6 % Coomassie Brilliant Blue R250 (w/v) 0.15 % Coomassie Brilliant Blue G250 (w/v)
Coomassie destaining solution I	40 % Ethanol (v/v) 20 % Acetic acid (v/v)
Coomassie destaining solution II	10 % Acetic acid (v/v)

DNA loading dye (5x)	30 % Glycerole (v/v) 0.5 x TRIS-borate-EDTA (TBE) buffer (v/v) 0.033 % bromophenol blue (w/v) 0.033 % xylene cyanol (w/v)
FD Buffer green (10x)	not listed
Fixing solution	10 % methanol (v/v) 10 % acetic acid (v/v)
Gentamicin	7 mg/ml
reduced glutathion (GSH) elution buffer	50 mM Tris/HCl, pH 8.4 2 mM DTT 10 mM GSH
High salt-tris-buffered saline with Tween20 (TBST)	50 mM Tris/HCl, pH 7.5 400 mM NaCl 0.1% Tween-20
Kan	50 mg/ml
Lysogeny broth (LB) medium	1 % Peptone (w/v) 0.5 % Yeast extract (w/v) 0.5 % Sodium chloride (w/v)
Low salt-TBST	50 mM Tris/HCl, pH 7.5 150 mM NaCl 0.1 % Tween-20

Lysis Buffer (protein expression)	50 mM Tris/HCl, pH 8.5
	2 mM DTT
Lysis Buffer (cell culture)	60 mM Tris/HCl, pH 7.5
	30 mM NaCl
	1 mM EDTA
	10 % Glycerol
	1 % Triton X-100
M (50x)	2.5 M Ammonium chloride
	1.25 M Disodium hydrogen phosphate
	1.25 M Monopotassium phosphate
	250 mM Sodium sulfate
Metals (1000x)	50 mM Iron chloride
	40 mM Calcium chloride
	10 mM Manganese (II) chloride tetrahydrate
	10 mM Zinc sulfate heptahydrate
	4 mM Copper (II) chloride dihydrate
	4 mM Sodium molybdate dihydrate
	4 mM Sodium selenite pentahydrate
	4 mM Boric acid
	2 mM Cobalt (II) chloride hexahydrate
	2 mM Nickel (II) chloride hexahydrate
Radioimmunoprecipitation assay	10 mM Tris/HCl, pH 7.5
(RIPA) buffer	100 mM NaCl
	2 mM EDTA

	1% NP-40
	with proteinase inhibitors
Reverse Transcriptase-Buffer	12.50 mM DTT
	12.50 mM Tris/HCl, pH 8.0
	7.50 mM $MgCl_2$
	1.25 µM dNTP
Taq-Polymerase-Buffer	0.5 M KCl
	0.1 M Tris/HCl, pH 8.3
	15 mM $MgCl_2$
	0.01 % Gelatine
Tetracycline	15 mg/ml,
SEC Buffer	20 mM Tris/HCl, pH 8.5
	200 mM NaCl
	2 mM DTT
SDS loading dye (6x)	60 mM Tris/HCl, pH 6.8
	1 mM EDTA-NaOH
	16 % Glycerole (v/v)
	2 % SDS (w/v)
	0.1 % Bromophenol blue (w/v)
SDS-PAGE running buffer (10x)	1920 mM Glycine
	250 mM Tris/HCl
	1% SDS (w/v)

Separation gel buffer	1.5 M Tris/HCl, pH 8.8 0.4 % SDS (w/v)
Stacking gel buffer	500 mM Tris/HCl, pH 6.8 0.4 % SDS (w/v)
T4 DNA ligase buffer (10x)	50 mM Tris/HCl, pH 7.5 10 mM Magnesium chloride 10 mM DTT 1 mM ATP
TRIS-Acetat-EDTA (TAE) buffer	40 mM Tris/HCl, pH 8.0 20 mM Acetic acid 2 mM EDTA
TBE buffer	100 mM Boric acid, pH 8.0 100 mM Tris/HCl 2.5 mM EDTA
Transfer Buffer:	5.82 g/l Tris-Base 2.93 g/l Glycin 3.75 ml SDS (10% Stock) 20 % Methanol
Phusion-Buffer:	Thermo Scientific F-518 5x Phusion HF
Urea-PAGE loading buffer	210.5 µg/ml Bromphenolblau 210.5 µg/ml Xylencyanol

	26.3 mM EDTA in Formamide
Urea-PAGE running buffer	0.5 x TBE
ZY medium	1 % Peptone (w/v) 0.4 % Yeast extract (w/v)

6.1.4 Kits

Kit	Company
Nucleo Spin Gel and PCR Clean-up	Macherey-Nagel, Germany
DNA, RNA, and Protein Purification	Macherey-Nagel, Germany
NucleoBond Xtra Midi	Macherey-Nagel, Germany

6.1.5 Enzymes

Name	Company
Endoproteinase ArgC from mouse submarillary glands (0.5 µg/µl)	Roche, Germany
Carboxypeptidase Y (0.1 µg/µl)	Sigma-Aldrich, Germany
Chymotrypsin (0.5 µg/µl)	Roche, Germany
DNase I	Roche, Germany
Endoproteinase Glu-C from Staphylocossusaereus V8 (0.5 µg/µl)	Sigma-Aldrich, Germany
Elastase (0.13 µg/µl)	Sigma-Aldrich, Germany
Endoproteinase LysC from Lysobacter enzymogenes (0.5 µg/µl)	Sigma-Aldrich, Germany
PreScission protease	Home-made, recombinant

Proteinase K	Sigma-Aldrich, Germany
Restriction endonucleases	New England Biolabs, Germany
Reverse transcriptase (RT)	New England Biolabs, Germany
RNase A, T1	Ambion, Germany
RNasin	Molox, Germany
Subtilisin (0.1 µg/µl)	Sigma-Aldrich, Germany
T4 DNA ligase	New England Biolabs, Germany
T4 polynucleotide kinase	New England Biolabs, Germany
Thermolysin	Sigma-Aldrich, Germany
Trypsin	Roche, Germany

6.1.6 Microorganisms

Strain	Descritption	Use
BL21	deficient in *lon* and *ompT* proteases	general purpose expression host for plasmids
BL21 pLysS	deficient in *lon* and *ompT* proteases	high-stringency expression host
BL21 pRare	deficient in *lon* and *ompT* proteases, contains plasmid encoding *argU*, *ileY*, and *leuW*	expression host; allows expression of genes encoding tRNAs for rare argenine codons AGA and AGG, isoleucine codon AUA, and leucine codon CUA

6.1.7 Vectors

Name	Description	Reference
pGEX6P1	Vector for expression of genes with a PreScission protease-cleavable N-terminal GST-tag in *E. coli*; AmpR	GE Healthcare
pEGFP-N3	Vector for expression of genes with a C-terminal GFP-tag in eukaryotic cell	Clontech
pFlag-N3	Vector for expression of genes with a C-terminal Flag-tag in eukaryotic cells	self made

6.1.8 Cloned Constructs

Vector	Insert
pGEX6P1	PRPF39 full-length
pGEX6P1	PRPF39ΔN
pGEX6P1	PRPF39ΔC
pGEX6P1	PRPF39ΔNΔC
pGEX6P1	PRPF39^{R458D}
pGEX6P1	PRPF39^{R464D}
pGEX6P1	PRPF39^{Y536W}
pGEX6P1	PRPF39$^{E576K/D577K}$
pGEX6P1	PRPF39$^{R458D, R464D}$
pGEX6P1	PRPF39$^{R458D, Y536W}$
pGEX6P1	PRPF39$^{R464D, Y536W}$
pGEX6P1	PRPF39$^{R458D, R464D, Y536W}$
pGEX6P1	PRPF39$^{E576K/D577K, Y536W}$
pEGFP-N3	PRPF39 full-length
pEGFP-N3	PRPF39^{R464D}

pEGFP-N3	PRPF39$^{E576K/D577K}$
pFlag-N3	PRPF39 full-length
pFlag-N3	PRPF39^{R464D}
pFlag-N3	PRPF39$^{E576K/D577K}$

6.1.9 Primers

The nucleotide residues shown in red signify changes in the sequence for mutagenesis.

Primer	Sequence
mPRPF39 fwd for pGEX6P1	GCGGATCCATGCAAAACTCCCACATGGAAGAGT
mPRPF39 rev for pGEX6P1	GCGTCGACTCAAGTTGGAGGTGGAGGATAATACTG
mPRPF39 fwd for GFP/Flag	GCCTCGAGATGCAAAACTCCCACATGGAAGAGT
mPRPF39 rev for GFP/Flag	GCGGATCCAGTTGGAGGTGGAGGATAATACTGTCC
39_R458D_fwd	CAATGGTTCGATTGGACAGAGTAAGTTTAG
39_R458D_rev	CTAAACTTACTCTGTCCAATCGAACCATTG
39_R464D_fwd	GTAAGTTTAGAAGACCGGCATGGAAATATG
39_R464D_rev	CATATTTCCATGCCGGTCTTCTAAACTTAC
39_Y536W_fwd	CTTGAAATGGAATGGAGTTGTGACCTCAAGC
39_ Y536W_rev	GCTTGAGGTCACAACTCCATTCCATTTCAAG
39_ED476/7KK_fwd	GAAAAGTGGAATTCCTTAAAAAGTTTGGTTCAGATG
39_ ED476/7KK _rev	CATCTGAACCAAACTTTTTAAGGAATTCCACTTTTC
PRPF39 fwd splice PCR	ACCATTGGAAAAGGCTCAGC
PRPF39 rev splice PCR	TCTGCTGAAGACATGCCTCA

6.1.10 Crystallization Screens

Screen	Supplier
Classic	Qiagen, Germany
Core I to IV	Qiagen, Germany

Index	Hampton Research, USA
SaltX	Hampton Research, USA
Additive Screen	Hampton Research, USA

Protein	Screen	Concentration
mPRPF39$^{full\text{-}length}$	Classic	7 mg/ml
mPRPF39$^{full\text{-}length}$	Index	7 mg/ml
mPRPF39$^{full\text{-}length}$	SaltX	7 mg/ml
mPRPF39$^{full\text{-}length}$	Additive Screen	7 mg/ml
mPRPF39$^{\Delta C}$	Classic	3.5; 7; 14; 20.4 mg/ml
mPRPF39$^{\Delta C}$	Core I to IV	7; 14 mg/ml
mPRPF39$^{\Delta C}$	Index	3.5; 20.4 mg/ml

6.1.11 Software and Websites

Software	Reference
ApE	M. Wayne Davis, USA
Coot	(Emsley and Cowtan 2004)
CorelDRAW	Corel Corporation, USA
ExPASy - ProtParam tool	http://web.expasy.org/protparam/
ImageQuantTL	GE Healthcare, Germany
iMOSFLM 1.0.7.	(Battye 2011)
Phaser	(McCoy 2007)
PHENIX suite	(Adams 2002)
PyMOL	Schrödinger LLC, USA
PHYRE2 Protein Fold Recognition Server	(Kelley 2015)
XDS	(Kabsch 2010)

6.1.12 Information on Organisms Analyzed in This Study

Table 1 Overview of organisms and the NCBI code for the respective proteins.

Organism	NCBI GeneID Prp39	NCBI GeneID Prp42	NCBI sequence identifier	nt range if applicable
Aspergillus fumigatus	3509441		CM000169.1 ; CM000172.1	1991395-1991543 ; 3279823-3279968
Bos taurus	505547		GK000010.2	74737404-74737241
Callithrix jacchus	100394414		CM000862.1	135695985-135696148
Candida albicans	3635832	3643177	XR_002086426.1	
Candida glabrata	2890874	2890082	CR380957.2	492841-493435
Canis lupus familiaris	480305		CM000001.3	29614860-29614698
Chrysemys picta	101939472		XR_002890345.1	
Danio rerio	368864		KN150351.1	47650-47813
Debaryomyces hansenii	8998180	2904027	CR382133.2	651586-651750
Drosophila melanogaster	43399		X53542.1	2592-2756
Equus caballus	100051463		CM000377.2	171439126-171438963
Eremothecium gossypii	4619435	4618919	NR_149369.1	
Felis catus	101086316		CM001381.2	148196934-148197097
Ficedula albicollis	101815923		KE165361.1	72627-72464
Gallus gallus	100858094		CM000094.4	122405288-122405125

Homo sapiens	55015		CM000679.2	58677737-58677574
Kazachstania africana	13882673	13883163	HE650821.1	1584550-1585071
Kazachstania naganishii	34526819	34523735	HE978319.1	165173-164349
Kluyveromyces lactis	2896694	2892680	CR382124.1	1640297-1640824
Kluyveromyces marxianus	34714086	34715797	AP012217.1	1292327-1292890
Lachancea thermotolerans	8291220	11496297	CU928166.1	664746-665230
Macaca fascicularis	102118141		CM001286.1	53437881-53438044
Meleagris gallopavo	100544498		CM000962.1	32444664-32444822
Monodelphis domestica	100020483		CM000369.1	3371371-3371534
Mus musculus	328110		CM000998.2	85829627-85829787
Naumovozyma dairenensis	11497194		HE580272.1	893934-894500
Nomascus leucogenys	100583605		CM001651.1	15724197-15724359
Ornithorhynchus anatinus	100084128		DS180968.1	711423-711586
Oryctolagus cuniculus	100342733		GL018716.1	1851532-1851695
Ovis aries	101105691		CM001607.2	36635321-36635483
Pan paniscus	100989581		XR_003029256.1	

Pan troglodytes	739795		CM000325.3; CM000314.3	3066181-3066019 ; 7491152-7490990
Papio anubis	101018467		CM001504.2	15159947-15159784
Pongo abelii	100450646		CM000561.1	120637157-120637319
Rattus norvegicus	314171		CM000241.2	30441118-30440958
Saccharomyces cerevisiae	854960	851821	AEEZ01000083.1	68906-68339
Schistosoma mansoni	8352871		HE601624.1	11992667-11992825
Sus scrofa	100153928		GL893043.1	11134-11297
Taeniopygia guttata	100223902		EQ832594.1	24485-24322
Tetrapisispora blattae	14494650	14492777	HE806322.1	39259-38548
Tetrapisispora phaffii	11530654	11531712	HE612870.1	376574-377321
Torulaspora delbrueckii	11501180	11504149	HE616742.1	108192-107653
Xenopus laevis	734636		CM004467.1	160259397-160259560
Zygosaccharomyces rouxii	8206003	8201844	CU928178.1	1248575-1249130

6.2 Methods

6.2.1 Nucleic Acid Methods

6.2.1.1 Determination of Nucleic Acid Concentration

To determine the concentration of nucleic acids, the light absorption of an aqueous solution was measured at the wavelength of 260 nm using a Nanodrop spectrophotometer. The concentration was then calculated using theoretical absorption values at 260 nm (as described in Sambrook, 1989).

double-stranded DNA	1 OD260 = 50 µg/ml
single-stranded DNA	1 OD260 = 33 µg/ml
single-stranded RNA	1 OD260 = 40 µg/ml

6.2.1.2 Agarose Gel Electrophoresis for DNA

Agarose gel electrophoresis was used both for analytical visualization and purification of preparative amounts of DNA. Agarose gels were prepared with agarose concentrations between 1-2 % in 1 x TAE and ethidium bromide in a concentration of 0.05 mg/L, depending on the DNA fragment size. Before loading the samples were mixed with DNA loading dye and a commercial DNA ladder was applied to every gel. The gels were run at a constant voltage of 135 V in 1 x TAE buffer. DNA bands were visualized by UV illumination at 254 nm. DNA bands were cut from the gel with a razor blade and the DNA was extracted from the gel using a kit following the instructions of the manufacturer.

6.2.1.3 Polymerase Chain Reaction (PCR)

PCR was used for amplification of genes or gene fragments. Phusion polymerase was used according to the instructions of the manufacturers. Typical PCR conditions are shown below (**Table 2**).

Table 2 Conditions for PCR

Compound	Amount
DNA Template (plasmid DNA 1µg/ml)	0.5 µl
Forward primer (100ng/ul)	0.5 µl
Reverse primer (100ng/ul)	0.5 µl
5 x PHU Buffer	5.0 µl
dNTPs (20mM)	0.5 µl
PHU DNA polymerase	0.5 µl
H_2O	17.5 µl
Total volume	25 µl

Cycle step	Temperature (°C)	Time	Cycles
Initial denaturation	98	30 s	1
Denaturation	98	10 s	30
Annealing	56-60	20 s	
Extension	72	30-150 s	
Final Extension	72	7 min	1

6.2.1.4 Site-directed Mutagenesis

Desired mutations were introduced according to the QuikChange II XL Site-Directed Mutagenesis Kit manufacturer's instructions. The resulting clones were verified by DNA sequencing.

6.2.1.5 Restriction Digestion and Ligation of DNA

Restriction digestions were carried out to generate compatible ends in vectors and PCR products for subsequent ligation reactions. Buffers and temperatures were chosen according to the manufacturer's instruction. The conditions for the restriction digestion are listed below (**Table 3**).

Table 3 Conditions for restriction digestion

Compound	Amount
Insert/Vector	20.5 µl / 1 µl + 19.5 µl H_2O
Enzyme 1	1 µl
Enzyme 2	1 µl
10 x Buffer	2.5 µl

To decrease the vector background 2.9 µl antarctic phosphatase buffer and 1.1 µl antarctic phosphatase were added to the vectors for 5 minutes. Since phosphatase-treated fragments lack the 5´ phosphoryl termini required by ligases, they cannot self-ligate.

For ligation, the digested DNA was resolved on a preparative agarose gel and the band containing the desired product was excised and extracted as explained in Agarose Gel Electrophoresis for DNA. Ligation reactions typically contained 7 µl digested insert and 1 µl digested vector. Reaction mixtures were incubated at room temperature for 1 h.

6.2.1.6 Plasmid Isolation from *Escherichia coli (E. coli)* Cells

A single colony from an overnight grown LB-agar plate was used to inoculate LB medium. Cells were grown in 3 ml or 100 ml LB medium overnight at 37 °C. Plasmid purification was carried out using Mini- or Midiprep kits, according to the manufacturer's instructions.

6.2.1.7 Plasmid Verification

All cloned plasmids were verified for the presence of the correct insert by analytical restriction digestion. Sequences of the inserts which showed the correct size in agarose gel analysis were verified by DNA sequencing.

6.2.1.8 Isolation of RNAs

Cells for RNA extraction were harvested (7000 rpm, 1 min) and the cell precipitate was solved in 500 µl “RNA-Tri-flüssig”. Next 100 µl chloroform was added and the samples were vortexed and incubated on ice for 10 min. Phase separation was achieved by spinning the tubes for 15 min at 4 °C at full speed. The aqueous phase was transferred to a new tube containing 300 µl ice-cold isopropanol. The tubes were again vortexed and centrifuged for 15 min at 4 °C at full speed to pellet the precipitating RNA. The RNA precipitate was washed three times with 70 % ethanol and then dissolved in 12 µl Milli-Q-water. The concentration and the purity of the RNA were determined with the Nanophotometer P330 from Implen.

6.2.1.9 *In Vitro* Transcription of RNAs for *In Vitro* splicing

A linearized plasmid which includes a T7 promoter as well as the sequence of the pre-mRNA template was used as starting material. After incubation at 37 °C for 2 h in the presence of T7 polymerase, DNase was added for 15 min to degrade the template. The generated RNA transcript was precipitated with PCI and ethanol and resuspended in 40 µl water. The efficiency of the reaction was checked by loading 2 µl on a 10 % denaturing PAGE.

6.2.1.10 RT-PCR and RT-qPCR

RT-PCRs were done as previously described (Preussner 2014). Briefly, RNA was extracted using RNA Tri-Flüssig and 1 mg RNA was used in a gene-specific RT-reaction. Low-cycle PCR with a ^{32}P-labeled forward primer was performed, products were separated by denaturing PAGE and quantified using a phosphoimager and ImageQuantTL software. For qRT-PCR up to 4 gene-specific primers were combined in one RT reaction. The qPCR was then performed in a 96-well format using the ABsolute QPCR SYBR Green Mix on a Stratagene Mx3000P instrument. qPCRs were performed in duplicates, mean values were used to normalize expression to mRNA of GAPDH. Quantifications are given as mean values, error bars represent standard deviation, p values were calculated using Student's unpaired t test. Significance is indicated by asterisks (*p < 0.05; **p < 0.01; ***p < 0.001).

6.2.1.11 T-cells and RNA Sequencing

T-cells generation and RNA sequencing were performed by Thomas Schüler, Otto-von Guericke University Magdeburg and Bernd Timmermann Max-Planck-Institute for Molecular Genetics Berlin, respectively. Naïve and memory CD8+ T cells were generated using the OT-I system as described previously (Stoycheva 2015, Deiser 2016). RNA-Seq was done essentially as described (Herdt 2017). Briefly, total RNAs were prepared using RNA-Tri and further purified using the RNeasy mini kit in combination with a DNase I treatment. RNA sequencing libraries were prepared by using the TruSeq mRNA Library Preparation kit. 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 40x10^6 reads per sample) and analyzed using a MISO-based pipeline (Preussner 2017).

6.2.2 Cell and Cell Culture Methods

6.2.2.1 *E. coli* Strains and Cultivation

E. coli cells were grown in liquid medium or on agar plates supplemented with adequate antibiotics in the following concentrations:

- Amp 100 µg/ml,
- Cm 34 µg/ml,
- Kan 50 µg/ml.

6.2.2.2 Transformation and Selection of *E. coli* Cells

For electroporation, 50-100 ng DNA were mixed with 50 µl electro-competent *E. coli* cells on ice. The mixture was transferred to an ice-cold electroporation cuvette (1-2 mm electrode gap) and subjected to a 4.8 ms pulse of 2.5 kV. Cells were collected in 1 ml of LB medium and incubated for 1 h at 37°C in a shaker. Subsequently the cells were centrifuged and resuspended in 250 µl LB medium and streaked out on an agar-plate containing the selective antibiotics.

For chemical transformation, 100-200 ng of plasmid DNA was mixed with 100 µl of chemically competent *E. coli* cells and incubated for 10 min on ice. Ice-cold cells were then heat-shocked for 1 min at 42°C and cooled on ice for 5 min. Cells were mixed with 200 µl of LB medium and incubated at 37°C for 1 h in a shaker. The cells were collected and selected on LB-agar plates supplemented with appropriate antibiotics.

6.2.2.3 Protein Expression in *E. coli*

Protein expression was conducted using chemically competent *E. coli* cells. The cells were inoculated with 20 ml preculture and grown in LB medium with the appropriate antibiotics at 37°C until an optical density (OD) of 0.6 to 1 was reached. At this point the cultures and the shaker were cooled to 18°C and induced with 1 mM IPTG. Expression was performed overnight.

For auto-induction cells were grown in ZYM-5052 (50 mM phosphate) auto-inducing complex medium (Studier 2005). Complex media containing enzymatic digests of casein and yeast extract are extensively used since they support growth of a wide range of *E. coli* strains, with different nutritional requirements, and cultures typically grow several times faster than in simple mineral salts media supplemented with glucose as the only carbon source. However, due to small amounts of lactose, inducing activity is fairly common in complex media.

Auto-induction depends on mechanisms bacteria use to regulate the use of carbon sources present in the growth medium. Lactose is prevented from inducing production of target protein by compounds that can be depleted during growth, such as glucose, glycerol and amino acids. Consequently, if glucose is present in the medium, it prevents the uptake of lactose. Ideally, the expression strain would grow to saturation in auto-inducing media, when depletion of inhibitory factors would allow cells to take up lactose and convert it to allolactose, the natural inducer of the lac operon. Induction

unblocks both the *lac*UV5 and T7*lac* promoters of T7 RNA polymerase and target protein, respectively, and leads to large scale expression levels in highly saturated cultures.

6.2.2.4 Eukaryotic Cell Lines and Their Cultivation

The cell lines used in this study are listed in **Table 4**:

Table 4 Cell lines used in this study

Name of the cell line	Cell type	Growing system
Jukat (Jsl1)	Human T lymphocytes	Suspension culture
El4	Mouse T lymphocytes	Suspension culture
HEK293T	Human embryonic kidney cells	Adherent culture

Suspension cells were cultivated in RPMI-medium-1640 (1x) + GlutaMAX TMI (+/+) containing 10 % heat-inactivated FBS and Penicillin/Streptomycin 1:100.

Adherent cells were cultivated in DMEM High Glucose containing heat-inactivated 10 % FBS and Penicillin/Streptomycin 1:100. Cell seeding was done in DMEM High Glucose containing only 10 % FBS but no antibiotics.

Both suspension and adherent cells were grown at 37 °C with 5 % CO_2 in a Heraeus incubator from Thermo Scientific.

6.2.2.5 Plasmid and siRNA Transfection

For plasmid transfection HEK293T cells were seeded into 6 well plates in a concentration of 3.5 * 10^5/well in 1.5 ml medium. The next day the cells were transfected with the desired plasmids. To this end a total of 2 µg plasmids were incubated in 250 µl OptiMEM. Then 5 µl RotiFect in a total volume of 250 µl OptiMEM were added to the samples. After 20 minutes of incubation the samples were dripped into the wells.

For siRNA transfection 0.5×10^5 HEK293T cell were seeded in a 12-well plate and transfected with 20 pMol of siRNA.

6.2.2.6 Harvesting Proteins

2 days post transfection the HEK293T cells were harvested. First the medium was removed, then the cells were washed with 1.5 ml PBS per well. Subsequently 250 µl Flag Lysis Buffer was added to every well, the cells were transferred to eppis and chilled at 4 °C for 10 minutes while occasionally vortexing them. The supernatant is collected after 15 minutes centrifugation at 13,000 rpm

6.2.2.7 Immuno Precipitation (IP)

IPs were performed as previously described with minor adjustments (Heyd and Lynch 2010). For IPs HEK293T cells were lysed in RIPA buffer, 100 ug of lysate were incubated in 500 µl RIPA buffer containing 400 mM NaCl (if not otherwise mentioned) and 3% BSA. After 1h rotation at 4°C, prewashed anti-FlagM2 beads were added and 4°C rotation was continued overnight. Beads were washed 5 times in RIPA buffer and after the last wash SDS-sample buffer was added, samples were boiled and analyzed by SDS-PAGE and Western blot.

The following antibodies were used for Western blotting: Flag (2368, Cell Signaling), GFP (B-2, Santa Cruz).

6.2.3 Protein Methods

6.2.3.1 Determination of Protein Concentration

The protein concentration can be determined by measuring the UV-absorption at 280 nm. The molecular weight and the extinction coefficient were calculated with ProtParam tool (www.expasy.org) and the protein concentrations were automatically calculated according to the Lambert-Beer law:

$$A = \varepsilon * c * d$$

A: absorption at 280nm

ε: molar absorptivity

c: concentration [mg/ml]

d: path length

As reference-solution the appropriate buffer was chosen.

6.2.3.2 Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE)

SDS-PAGE is a method of separating proteins by size in a gel in an electrical field. Proteins bind the detergent sodium dodecyl sulfate (SDS) to form negatively charged SDS-protein-complexes with a constant charge per unit mass ratio. Especially in combination with DTT SDS denatures proteins and inhibits protein-protein interactions. Thus, the proteins differ only in size and have comparable hydrodynamic properties. During the electrophoresis the SDS-protein complexes migrate to the positive pole. The molecular sieve effect of the porous polyacrylamide matrix separates the SDS-protein complexes according to their Stokes radius which in this case means according to their molecular weight.

The samples were diluted if necessary. SDS-6x sample-buffer was added to the samples and then they were heated to 95°C for approximately 5 minutes. 20 µl of every sample was loaded onto a gel. As reference a Protein Ladder was loaded on every SDS-PAGE. Gels were run at 250 V for about 45 minutes. For Western Blot the gels were run at 136 V for approximately 90 minutes.

Coomassie staining solution is applied to the finished gel and excess coloring agent is removed by the Coomassie destaining solution so that only the bands with proteins are colored blue. This method is sensitive in a range of 200 ng to 400 ng protein. The distance travelled by each band is compared with the marker to determine the size of the proteins (Fairbanks 1971).

The SDS-PAGE gels were prepared according to Table 5.

Table 5 SDS-PAGE gel composition

	Stacking gel	Running gel	
Gel concentration	4%	12.5%	17.5%
Stacking gel buffer	2.5 ml	-	-
Running gel buffer	-	5.0 ml	5.0 ml
ddH2O	6.2 ml	6.6 ml	3.4 ml
30% Bis-acrylamide	1.3 ml	8.4 ml	11.6 ml
10% APS	100 µl	100 µl	100 µl
TEMED	10 µl	10 µl	10 µl

6.2.3.3 Western Blot

Western Blot is an analytical method to detect specific proteins after an SDS-PAGE. First the proteins are transferred electrophoretically from the SDS-PAGE gel to the methanol activated blot polyvinylidene difluoride membrane using a semidry-blot-system at 75 mA per gel for 1.5 hours. First three Whatman papers drenched in blot buffer are applied to the electrode bubble free. Then the activated membrane is laid on to the papers. The SDS-PAGE gel is placed on top and an additional layer of three drenched Whatman papers is added onto the stack as bubble free as possible.

After the blot-transfer the free binding sites on the membrane are blocked for one hour at room temperature in 2% BSA in LS-TBST. Then the membrane is incubated with the appropriate primary antibodies diluted in 2% BSA in LS-TBST over night at 4°C and washed 3 times 5 minutes in HS-TBST. After that the membrane was incubated for one hour at room temperature in α-IgG horseradishperoxidase-coupled antibodies from anti-rabbit or mouse in 2% BSA in LS-TBST (1:20,000) it was washed once for 15 minutes and 2 times for 5 minutes. Following this, developing solution was applied to the membrane and a film was exposed. The film was developed in an Amersham pharmacia biotech Hyper processor.

6.2.3.4 GSH Affinity Chromatography

GSH affinity chromatography is based on the high affinity of GST to GSH. Beads coated with GSH were used in a gravity flow column (Biorad). The target protein, fused with a GST-tag, was bound to GSH-beads and equilibrated with Lysis Buffer. After approximately 3 hours of incubation at 4 °C, the sample was poured into a gravity flow column. The column was then washed with 200 ml Lysis Buffer. For on-column cleavage, beads were resuspended in 15 ml Lysis Buffer supplemented with PreScission protease and incubated at 4 °C overnight. The flow-through then contained the target protein cleaved from its affinity tag, which stays bound to the beads, as do uncleaved target protein and PreScission protease (expressed as a GST-fusion protein). Wash fractions, samples of the beads before and after on-column cleavage as well as flow-through fractions were analyzed by SDS-PAGE to track the purification process. Flow-through fractions containing target protein were pooled for subsequent purification steps.

6.2.3.5 Ion Exchange Chromatography

Ion exchange chromatography is a method used for separation of charged particles. To this end charged groups are covalently bound to the column material. In an aqueous solution the charged groups are surrounded by ions of the opposite charge. When the solution has a low ionic strength, molecules can be immobilized on the column that only have weak binding properties. By increasing the ionic strength, the weaker binding molecules are eluted. The mPRPF39 sample was loaded in the minimal buffer containing no salt. After mPRPF39 was bound to the column it was washed with 20 column volumes of minimal buffer and the eluted with a 20 column volume NaCl gradient from 0 to 600 mM NaCl.

6.2.3.6 Size Exclusion Chromatography (SEC)

During a SEC proteins are separated according to their hydrodynamic volume. Under the condition that the protein does not interact with the stationary phase large molecules will run through an SEC column faster than small molecules because they cannot penetrate the beads of the stationary phase and will therefore have to pass a smaller volume and elute earlier than small molecules which penetrate the beads.

A Superdex 200 10/300 increase column was equilibrated with SEC buffer and an injection volume of the protein solution recommended by the manufacturer was loaded onto the column. The protein elution was tracked by measuring the UV absorption at 280 nm and 260 nm. The fractions with the highest protein concentration and purity are selected and concentrated with a filter with an appropriate cut off through ultrafiltration at 4000 g.

6.2.3.7 Protein Purification

Proteins bearing an N-terminal, PreScission-cleavable GST-tag were produced in *E. coli* BL21 pLys cells in LB-medium overnight at 18 °C after induction at OD ~ 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 lyzed by sonication. Cell debris was separated from the soluble fraction by centrifugation for 45 minutes at 55,900 x g in an Avanti J-26 XP centrifuge (Beckman Coulter). Target proteins were captured on glutathione agarose 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, subjected to a 1mlQ column 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 protein were pooled, concentrated, and shock-frozen in liquid nitrogen.

Transformed *E. coli* Rosetta2 DE3 were cultured in minimal media containing selenomethionine (Van Duyne 1993) and at an OD ~ 0.6, protein expression was induced by addition of 0.5 mM IPTG. Purification was carried out as described above for mPrp39^{wt}, except that all buffers contained 2 mM DTT.

6.2.3.8 Differential Scanning Fluorimetry (DSF)

Identifying conditions in which protein samples are stable over long periods of time and are prevented to aggregate or denature is extremely helpful for experiments involving analytical and biophysical techniques that require high protein concentrations.

Many factors can influence protein stability such as: buffers (chemical composition as well as pH), salts and detergents whose interactions with the protein are non-specific. Ligands, which bind proteins at specific sites, may also produce noticeable effects as protein stabilizers. The stability of a protein is characterized by the Gibbs free energy of unfolding, ΔGu, which is temperature-dependent. The stability of most proteins decreases as the temperature increases, meaning that ΔGu decreases and reaches zero at the equilibrium point where the concentrations of folded and unfolded protein are equal. At this equilibrium point, the temperature is called melting temperature (T_m).

DSF monitors the thermal unfolding of proteins in the presence of a fluorescent dye that is highly fluorescent in non-polar environment compared to aqueous solution where the fluorescence is quenched. Practically, DSF is performed using a real-time PCR instrument and monitors the fluorescence intensity of the dye as a function of the temperature. Upon protein unfolding by thermal denaturation, the dye preferentially binds to the now exposed hydrophobic patches of the protein and the fluorescence intensity increases. This generates a sigmoidal curve that can be described as a two-state transition and the T_m can be extracted from the curve by determining the first derivative.

DSF experiments were done in a 96-well plate in a plate reader combined with a thermocycler (Stratagene Mx3005P). In order to determine a suitable protein concentration, purified protein was diluted to series of concentrations varying from 1 µM to 10 µM in the final purification buffer supplemented with 10×SYPRO orange (1:500 dilution of the stock) in a total volume of 20 µl and pipetted into a 96-well plate. The temperature was increased from 25°C to 95°C and the fluorescence emission was

monitored in 1 °C increments with hold steps of 30 second between reads. The fluorescence intensity was then plotted as a function of temperature. The protein concentration that showed a clear sigmoidal curve was chosen for buffer optimization and compound screens. Different buffer compositions and compounds were tested for their stabilizing effect on the protein at the defined protein concentration. The sigmoidal curve from each condition was normalized and corrected for the background signal of the fluorophore in the buffer. The inflection points of the curves, representing the thermal melting temperature of the protein in the respective conditions, were compared.

6.2.3.9 Limited Proteolysis

Limited proteolysis experiments can inform on conformational features of proteins. In a number of studies, it has been demonstrated that the sites of limited proteolysis along the polypeptide chain of a protein are characterized by enhanced backbone flexibility, implying that proteolytic probes can pinpoint the sites of local unfolding in a protein chain (Fontana 2004). This means that only flexible regions of the target protein such as disordered termini, exposed loops, or flexible domain linkers can be cleaved. Removing such flexible parts from a protein generates more compact and conformationally homogeneous molecules or compact single domains. These proteolytic fragments or domains of a protein may crystallize more readily or form better diffracting crystals than the intact protein.

Limited proteolysis has been widely used to define the boundaries of single domains or a set of tightly interacting domains in a molecule by trimming its flexible and unstructured parts and, thus, increasing its propensity to crystallization. The increased likelihood of a domain or a smaller set of domains to yield structural information can be explained, in a way, by the observation that large proteins are often composed of many individual domains. The conformational heterogeneity that results from motion between such domains is a severe impediment to crystallization.

The generation of domains by limited proteolysis relies directly on the tertiary structure of the protein under investigation and provides much firmer evidence for their existence than that provided by sequence homology and secondary structure predictions.

In practice, limited proteolysis is achieved by dilution of the proteases sufficiently so that they will only digest the most accessible and flexible regions of the protein substrate leaving the domains intact. Initially, the protein substrate should be digested with various proteases to establish which conditions are optimal for generating a protease-resistant domain.

To search for stable fragments, full-length mPRPF39 was treated with various proteases. For each reaction, 9 µg of protein were incubated with increasing amounts (0.004, 0.04 and 0.4 µg) of protease at 20 ºC for 30 minutes in buffer containing 10 mM Tris/HCl, pH 7.5, 150 mM NaCl, 2 mM DTT. The reactions were stopped by addition of 10 µl SDS-PAGE loading buffer. Half of each sample was separated by SDS-PAGE and bands were analyzed by tryptic mass spectrometric fingerprinting

6.2.3.10 Band Preparation for Mass Spectrometry

Mass spectrometry was used to analyze the protein fragments produced by limited proteolysis and to test for selenomethionine (SeMet) incorporation. The proteins/protein fragments were separated by SDS-PAGE and stained with Coomassie staining solution Bands of interest were excised and proteins were digested in-gel with trypsin and extracted as previously described by Shevchenko et al., 1996. Tryptic fragments were supplied to Christoph Weise, FU Berlin, for mass spectrometry analysis.

6.2.3.11 Size exclusion with multi-angle light scattering (SEC-MALS)

SEC-MALS experiment was performed at 18 °C with buffer containing 50 mM Tris/HCl, pH 8.4, 300 mM NaCl and 0.02% NaN_3. 80 µg of $mPRPF39^{wt}$ or of mPRPF39 variants was loaded onto a Superdex200 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).

6.2.3.12 *In Vitro* Splicing

Splicing active nuclear extracts were prepared according to (Dignam 1983), except HEK293T 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% HeLa 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, as described previously (Dignam 1983). The reaction

was stopped after 1 h by proteinase K treatment followed by RNA extraction. 20% of the RNA was reverse-transcribed and analyzed by RT-PCR.

6.2.4 Crystallographic methods

6.2.4.1 Basic Principles of X-ray Crystallography

The main goal in X-ray crystallography is to derive the three-dimensional structure of a given protein, other biomacromolecules or complex based on a set of X-ray scattered intensities measured at different directions in space. This process can be divided in three basic steps. The first step is to obtain a protein crystal. The scattering of X-rays by a single protein molecule is extremely weak and still not possible to detect routinely. However, the periodic arrangement of proteins inside a crystal creates interference effects, which greatly enhance the intensity of the scattered X-rays in particular directions allowing them to be measured. In the second step, the crystal is exposed to an intense monochromatic beam of X-rays generating a distribution of scattered radiation in different directions in spaces, known as the diffraction pattern. A detector records the snapshots of the scattered X-rays as the crystal is rotated. The complete pattern is then retrieved based on the measured snapshots and symmetry considerations.

The third step consists of the determination of a structural model from the computational analysis of the diffraction pattern. Firstly, an electron density map is derived from the diffraction pattern. Then, the initially derived electron density is fitted with a structural model, which describes the position of atoms inside a protein. The structural model is refined by repeating this process iteratively until some statistical quantities related to the goodness of fit achieve a certain desired value.

6.2.4.2 General Crystallography Setup

To identify initial crystallization conditions, different commercially available crystallization reagents were screened in 96-well MRC plates by sitting drop vapor diffusion technique. Drops of 200 nl (100 nl protein solution + 100 nl reservoir) were dispensed using a Cartesian liquid dispensing robot with 8 channels.

Initial hits were usually refined by manual setups in a 24-well format. Commercial additive screens were routinely tested to improve crystallization conditions.

6.2.4.3 Crystallization

Native crystals 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 and 1.8 M sodium acetate. Selenomethionine-labeled crystals 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 and 0.75 mM tris(2-carboxyethyl)phosphine. Crystals were cryo-protected with a solution composed of 80% mother liquor and 20% (v/v) propylenglycole and subsequently flash-cooled in liquid nitrogen.

6.2.4.4 Diffraction Data Collection, Structure Determination and Refinement

Synchrotron diffraction data were collected at the beamlines the MX Joint Berlin laboratory at BESSY (Berlin, Germany) and P14-2 of PETRA III (Deutsches Elektronen Synchrotron, Hamburg, Germany). Diffraction data were processed with XDS (Kabsch 2010). Experimental phases were determined by single anomalous dispersion with the AutoSol routine in PHENIX (Terwilliger 2009) using Phaser (McCoy 2004) and SOLVE/RESOLVE (Terwilliger 2000) exploiting selenomethionine-labeled mPRP39. An initial, partial model of mPRPF39 was built with the program AUTOSOL in PHENIX (Terwilliger 2009). The structure was refined by maximum-likelihood restrained refinement using in PHENIX (Adams 2010, Afonine 2012) including Translation/Libration/Screw (TLS) refinement (Winn 2001). Model adjustment was performed with Coot (Emsley 2010). Model quality was evaluated with MolProbity (Chen 2010) and the JCSG validation server (Yang 2004). Secondary structure elements were assigned with DSSP (Kabsch and Sander 1983) and ALSCRIPT (Barton 1993) was used for secondary structure based sequence alignments. Figures were prepared using PyMOL (DeLano 2002).

6.2.5 Bioinformatics

All bioinformatics were performed by Alexander Neumann, FU Berlin. U1 RNA sequences were extracted from Rfam (Kalvari 2018) and NCBI (Geer 2010). Protein sequences of Prp39 and Prp42 homologs were extracted from NCBI and the canonical isoform was defined using UniProt (The UniProt 2017). RNA secondary structure predictions were performed using the RNAstructure Web Server (Reuter and Mathews 2010). ClustalOmega was used for multiple sequence alignments (Sievers 2011). Consensus sequences with identity threshold above 50% were calculated using SeaView v4 (Gouy 2010). The common tree taxonomy tool from NCBI was used to create a phylogenetic tree

using our selected species. The tree was visualized using Interactive Tree Of (Letunic and Bork 2016).

7 Results

7.1 Production of Proteins

One of the most time-consuming steps during high resolution structure-based studies is the production of significant amounts of pure protein suitable for crystallization. For this purpose, the production was optimized by adjusting different parameters in the bacterial expression system. Initial expression test showed that the presence of mPRPF39 seemed to inhibit cell growth, as the OD_{600} stayed constant upon induction with IPTG, showing only poor overexpression.

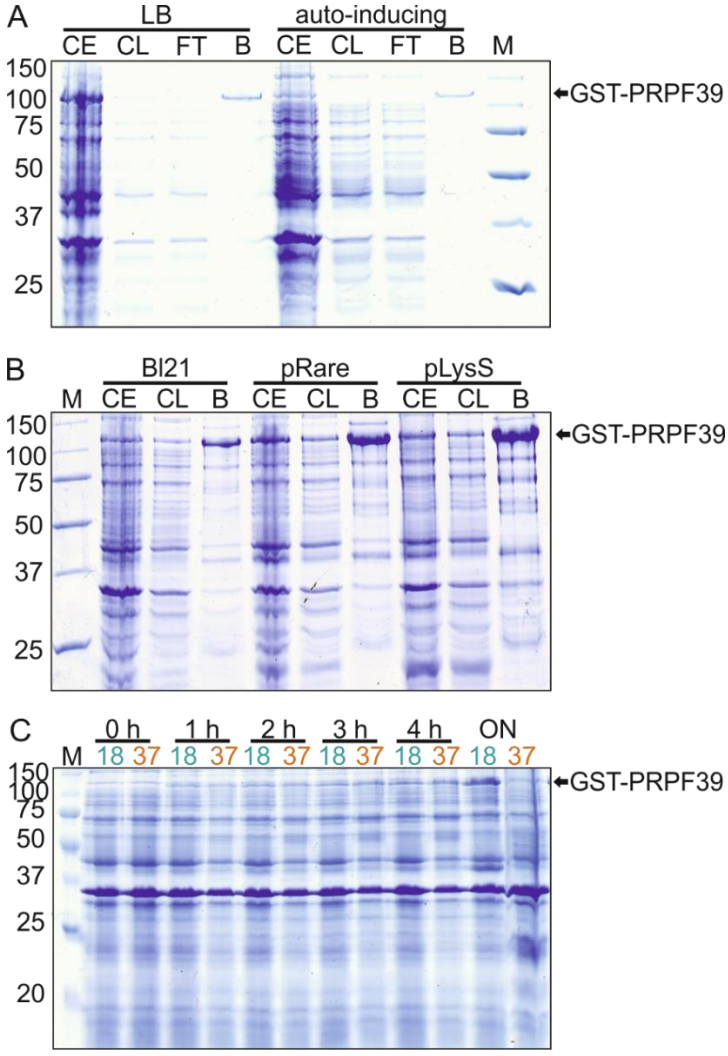

Figure 16 Optimization of mPRPF39 expression. Coomassie-stained SDS-PAGE gels showing a test expression with GST-pulldown: **A** Performed in LB and auto-inducing media. The test expression in LB media shows much higher comparative overexpression of mPRPF39. **B** Performed in different bacterial strains. The pLysS strain showed the highest yield. **C** A time course of the expression of mPRPF39 at 18 °C and 37 °C. Expression overnight at 37 °C leads to near complete protein degradation. The abbreviations CE stands for whole cell extract, CL cleared lysate, FT for flow through, B for beads, M for marker and ON for overnight. The sizes in kDa of the molecular weight marker are given on the left.

Two different media were tested; auto-induction media and LB-Medium with IPTG induction (**Figure 16** A), three different *E.* coli strains; BL21, pRare, and pLysS (**Figure 16** B), and different expression times at two different expression temperatures; 18 °C and 24 °C (**Figure 16** C). The optimal combination proved to be *E. coli* pLysS cells in LB-medium with IPTG induction and expression at 18 °C overnight.

7.2 Purification

To purify the protein after expression, the cell lysate was loaded on to preequilibrated GST-beads and a tag cleavage with PreScission was performed overnight on bead. The elution and wash fragments were then pooled and subjected to size exclusion.

To optimize the amount of soluble protein and encourage homogenous running properties, a differential scanning fluorimetry screen was performed, to find the optimal buffer conditions. The analysis showed, that mPRPF39 was overall relatively stable in the chosen buffer with a T_m of 43 °C and all but one condition having a lower T_m. The condition A10 (50 mM Tris/HCl, pH 8.4) increased the melting temperature slightly to 44 °C (**Figure 17**). This is not a big difference, but because the condition was very similar to the initially chosen buffer (50 mM Tris/HCl, pH 8.5, 400 mM NaCl, 2 mM DTT) and had no additional salt, this minimal buffer was used for all future purifications as a minimal buffer can be beneficial for crystallization.

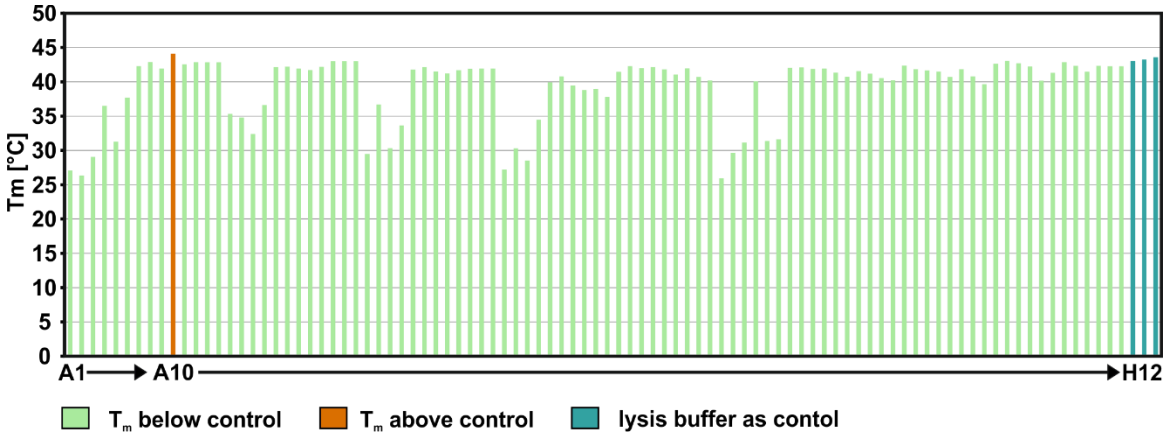

Figure 17 DSF analysis of mPRPF39. The different T_m of the screened conditions are plotted here. Overall the already chosen lysis buffer exhibited a comparatively high T_m. Only one condition, A10, showed a slightly increase T_m. For color-coding see inlet.

After initial purification trials, the protein showed very high 260/280 values (>1.5). To remedy this, the elution and wash fractions after the GST-column were loaded to a resource Q column and eluted with a NaCl gradient going from 0 to 600 mM. The eluted

mPRPF39 was then purified in a final step over a SEC column with the minimal buffer (**Figure 18**).

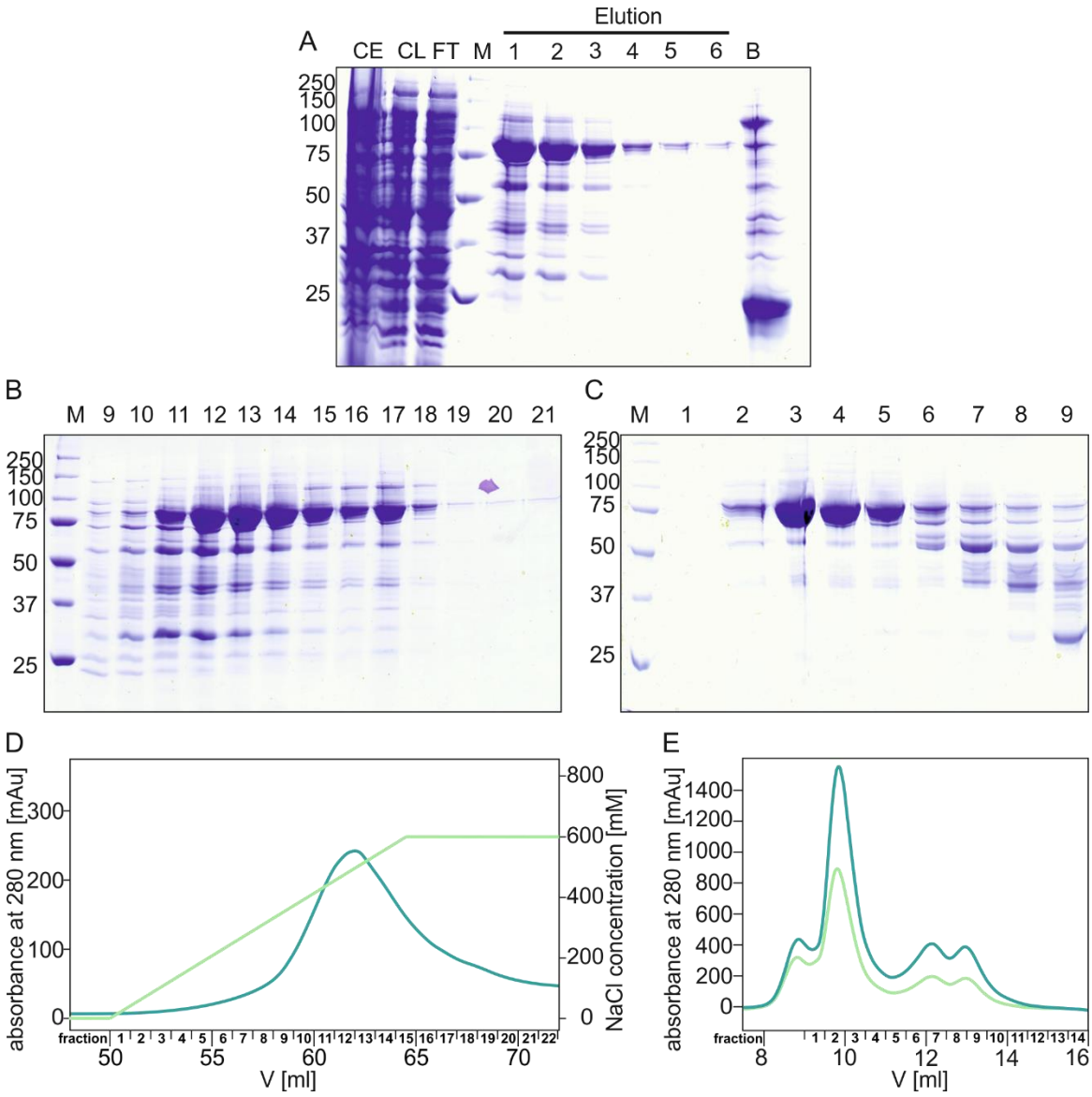

Figure 18 Purification of mPRPF39. Coomassie-stained SDS-PAGE gels showing **A** GSH purification **B** anion exchange and **C** SEC of fractions 11 through 17 from the anion exchange with their corresponding chromatograms D and E. The molecular weight marker is shown on the left (sizes in kDa) and the fraction number is indicated above the lanes. The abbreviations CE stands for whole cell extract and CL cleared lysate, FT for flow through and B for beads.

7.3 Limited proteolysis

Flexible regions in proteins can hinder crystallization. In order to remove such regions, mPRPF39 was treated with different proteases. Upon elastase treatment, a stable fragment appears when a concentration of 0.013 µg/µl is chosen. There is still quite some undigested mPRPF39 present with the lower elastase concentration while with higher

concentrations the protein is completely digested. When mPRPF39 is treated with subtilisin, a stable fragment is observed only with the lowest concentration. If the subtilisin concentration is increased, this leads to complete degradation of the protein, similar to the highest concentration of chymotrypsin. However, chymotrypsin does not seem to significantly degrade mPRPF39 at lower concentrations. When treated with GluC, we observe the stable approximately 60 kDa band of mPRPF39 even at the highest protease concentration (**Figure 19**).

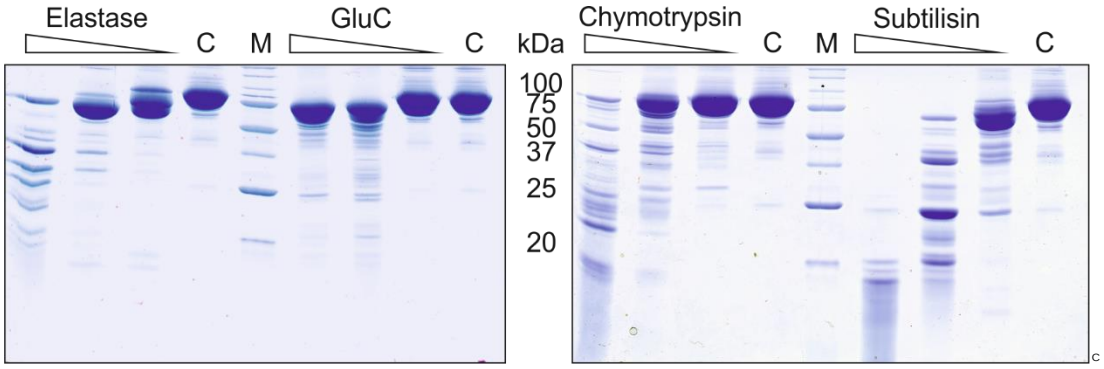

Figure 19 Experimental definition of mPRPF39 stable fragments. Coomassie-stained SDS-PAGE gels showing limited proteolysis of full-length mPRPF39. The decreasing protease concentration is indicated by a triangle. A stable, ca. 60 kDa fragment could be obtained with all proteases but Chymotrypsin. C stands for undigested control protein and M for marker.

After trypsination, the bands were subjected to mass spectrometry analysis. The total mass of the fragment is 66.06 kDa and, after trypsination, peptides in the core region of mPRPF39 could be observed. The region in which peptides were observed coincides nicely with the area of mPRPF39 that is predicted to be structured. Using the results from secondary structure predictions and mass spectrometry three additional constructs missing either the C- or N-terminus or both, were designed (**Figure 20**).

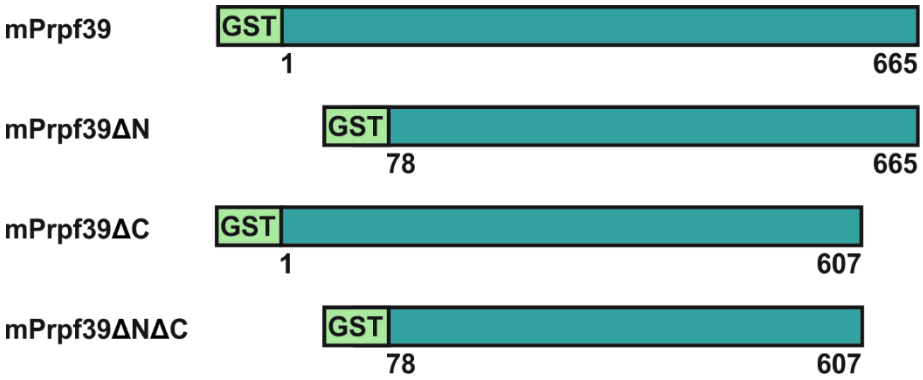

Figure 20 Schematic representation of the cloned mPRPF39 constructs. Teal boxes represent portions of mPRPF39, pale green boxes indicate GST-tag at the appropriate terminus and the amino acid residue borders are indicated under the scheme respectively.

Both constructs missing the N-terminus could not be expressed. mPRPF39$^{\Delta C}$ could be expressed in abundance without the cell growth limitation visible for the full-length mPRPF39. mPRPF39$^{\Delta C}$ could be purified analog and in similar purity to the full-length protein (**Figure 21**).

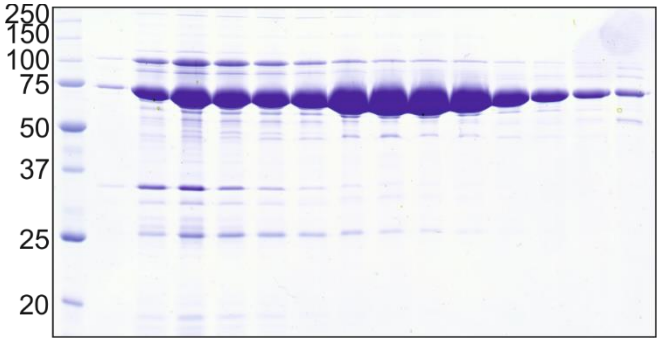

Figure 21 Size Exclusion of mPRPF39$^{\Delta C}$. Coomassie-stained SDS-PAGE gels shows the size exclusion of mPRPF39$^{\Delta C}$ on a SD 200 10/300 column. The purity is comparable to that of the full-length protein.

7.4 Crystallization and Structure Determination of mPRPF39

To obtain mPRPF39 crystals, initial screens of both constructs purified were prepared using a crystallization robot to facilitate testing a broad spectrum of conditions. Initial hits were observed for both constructs and could be successfully reproduced. The crystals were then tested at a synchrotron beamline (BL 14.2 Bessy, Berlin). The best diffracting crystal of the initial screening was of the full-length mPRPF39 and diffracted to 3.8 Å. The crystals overall showed high variances in diffraction quality with only about one in 20 diffracting under 4 Å.

To further optimize the crystallization conditions, additive screening, micro seeding and different temperatures were tried. Addition of poly(ethylene glycol) methyl ether 5000 to 3% proved to be beneficial to the crystal morphology of mPRPF39$^{\Delta C}$, however the diffraction could not be increased by any of the tested methods. Because the full-length protein seemed to be more promising, the focus was mainly shifted on to mPRPF39$^{\text{full-length}}$. Using the same crystallization conditions as at 18 ° for the full-length mPRPF39 at 4 °C increased the amount of well diffracting crystals with the best dataset collected at 3.3 Å.

Several search models were used, to perform molecular replacement with the goal of solving the phases. We used murine CSTF-77 (PDB entry 2OND (Bai 2007)), *K. lactis* RNA14 (PDB entry 4EBA (Paulson and Tong 2012)) and also the PHYRE prediction model (Kelley 2015) of mPRPF39 itself. None of the models available to us at the time could be placed by PHENIX.

To address the phase problem, a SeMet derivative of mPRPF39 was used. This seemed promising because there are 17 SeMet per molecule spread nicely over the whole protein sequence.

The SeMet mPRPF39 was expressed as described in the Materials and Methods section and could be purified analog to the native mPRPF39. The purity was comparable to the native protein and SeMet incorporation was tested by mass spectrometry (**Figure 22**).

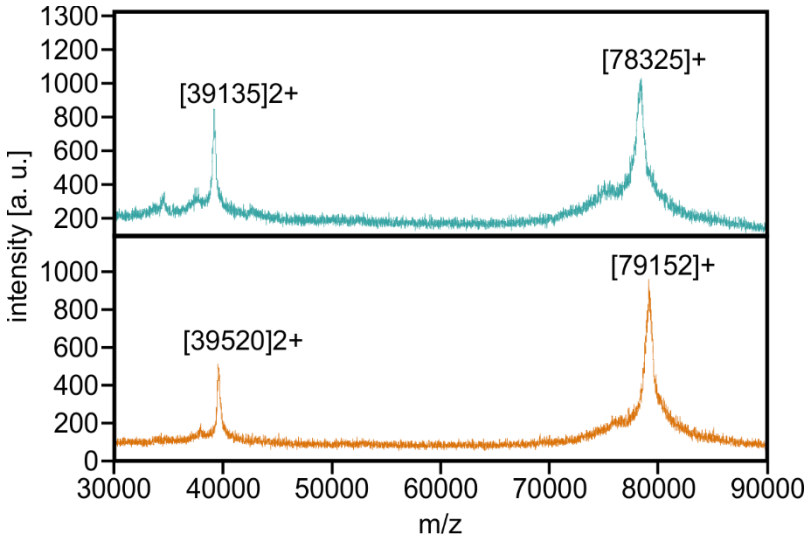

Figure 22 MALDI-MS of whole mPRPF39 and SeMet derivative. The upper panel shows the measurement for native mPRPF39 (teal) and the lower panel for SeMet mPRPF39 (orange). The expected mass for the full-length protein (78,335 kDa) is clearly matched and the increase in mass indicates full SeMet incorporation.

The SeMet mPRPF39 crystallized in the same condition as the native protein and crystal quality could be improved by addition of 7.5 mM tris(2-carboxyethyl)phosphine (**Figure 23**).

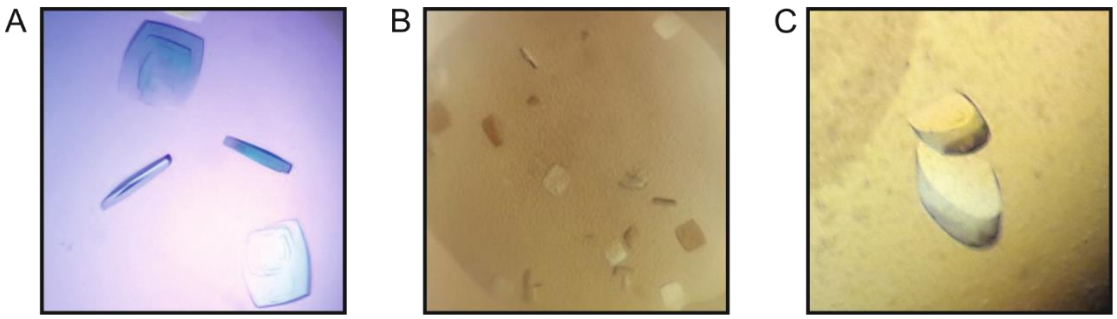

Figure 23 mPRPF39 crystals. **A** Native full-length mPRPF39 crystals. **B** mPRPF39$^{\Delta C}$ crystals. **C** Optimized SeMet full-length mPRPF39 crystals.

A selenomethionyl single-wavelength anomalous diffraction (SAD) dataset diffracting to 3.8 Å could be collected for the improved crystals. The anomalous signal itself was rather weak, at about 6.5 Å resolution. The dataset was used to solve the phase problem with

AutoSol from the PHENIX suite. AutoSol includes HYDD, SOLVE, Phaser and RESOLVE and is optimized for datasets with weak anomalous signal. In the first round of AutoSol 25 anomalous sites could be identified. This was followed by extreme density modification and autobuild which was not successful. The option extreme density modification is recommended for low resolution structures with weak anomalous signal. When this option is chosen fewer cycles of density modification are performed, the thoroughness is set to thorough so more solutions are examined, the density of the first density-modified map is used as a partial model to try and find additional anomalous scatterers, a more extensive search for anomalous scatterers is performed (thoroughness set to thorough) and the minimum non-crystallographic symmetry (NCS) correlation to keep NCS is lowered. The extreme density modification was invaluable in providing an interpretable electron density map.

α-helices were manually fitted into the resulting density to obtain an initial model and then the preliminary model was used as input model for another round of AutoSol. With the additional information of the model, 27 sites could be identified and phase extension to 3.3 Å was performed with the native dataset. After the second iteration, a partial model was obtained by autobuild. However only the placement and the direction of the α-helices was correct, the sequence had to be extensively corrected manually.

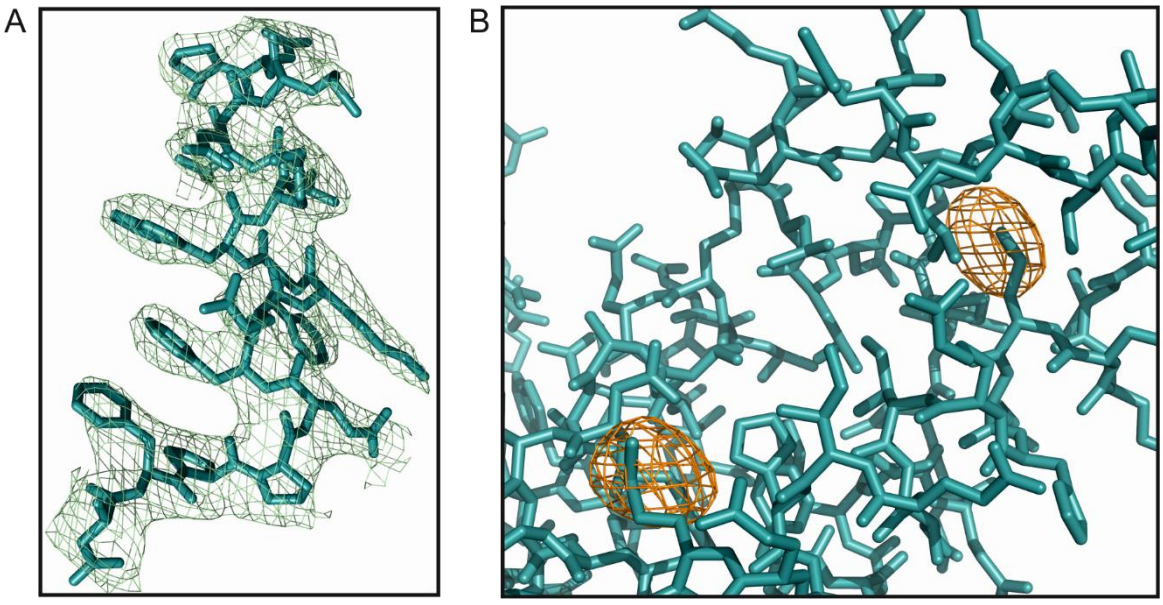

Figure 24 Electron density quality in mPRPF39. **A** 2m*FoFc* electron density drawn as pale-green mesh contoured at 1.0 σ. α-helix α1A of HAT repeat number 1 is shown in teal as stick representation. **B** Anomalous difference Fourier map contoured at the 3σ level. The placement of SeMets allowed sequence assignment.

Overall the electron density was of excellent quality for the resolution and allowed for model building (**Figure 24** A). Two molecules were found to be present in the asymmetric

unit and the model was built independently in both NCS related protein chains. The current structure includes in total 52 α-helices and 1019 residues. After placing the poly-alanine α-helices in the density, the anomalous density was used to place methionines (**Figure 24** B) and, subsequently the sequence in the α-helices harboring a methionine residue. The structure of the RNA 14 dimer (Paulson 2012) was used to verify the sequence according to the HAT repeat motifs, assign residues in non-methionine α-helices and connect the α-helices. For the refinement, strict NCS were used with secondary structure restraints and 3 TLS groups per monomer. For data collection and refinement statistics see **Table 6**.

Table 6 Crystallographic data collection and model refinement statistics.

Data Collection	**SeMet**	**Native**
Wavelength [Å]	0.9763	0.9184
Temperature [K]	100	100
Space group	*C*2	*C*2
Unit Cell Parameters a, b, c [Å]; β [°]	189.2, 73.0, 206.7; 112.4	189.5, 72.8, 207.1; 112.5
Resolution range [Å][a]	50.00 – 3.80 (3.90 – 3.80)	50.00 – 3.30 (3.45 – 3.30)
Reflections[a]		
Unique	49231 (3671)	39170 (4809)
Completeness [%]	98.1 (97.6)	98.5 (97.9)
Multiplicity	9.9 (9.9)	4.7 (4.8)
Data quality[a]		
Intensity [I/σ(I)]	11.8 (1.5)	9.4 (0.9)
R_{meas} [%]	17.7 (168.3)	14.9 (260.3)
$CC_{1/2}$	99.9 (83.7)	99.8 (63.7)
Wilson B value [Å^2]	120.0	102.1
Number of selenium atoms	25	-
FOM	0.28	-
BAYES-CC	2.9	-
Refinement		
Resolution range [Å][a]		50.00 – 3.30 (3.42 – 3.30)
Reflections [a]		38901 (3770)
R_{work}		0.246 (0.443)
R_{free}		0.293 (0.443)
Contents of an Asymmetric Unit		
Residues, Atoms		1019, 8592
Mean B-factor [Å^2][b]		154.5
RMSD from Target Geometry		
Bond Lengths [Å]		0.005
Bond Angles [°]		1.03
Validation Statistics [c]		
Ramachandran Plot		
Residues in Allowed Regions [%]		4.6
Residues in Favored Regions [%]		95.7
MOLPROBITY score[b]		2.1
MOLPROBITY Clashscore [b,c]		16.2

[a] Data for the highest resolution shell in parentheses
[b] Calculated with MOLPROBITY (Chen 2010)
[c] Clashscore is the number of serious steric overlaps (> 0.4) per 1,000 atoms.

7.5 Overall Structure of mPRPF39

The final structure shows mPRPF39 to be arranged as a dimer. Even though the protein is purely α-helical, being made up of 12 pairs of anti-parallel α-helices arranged as HAT repeats, it can be divided into three distinct subdomains (**Figure 25**).

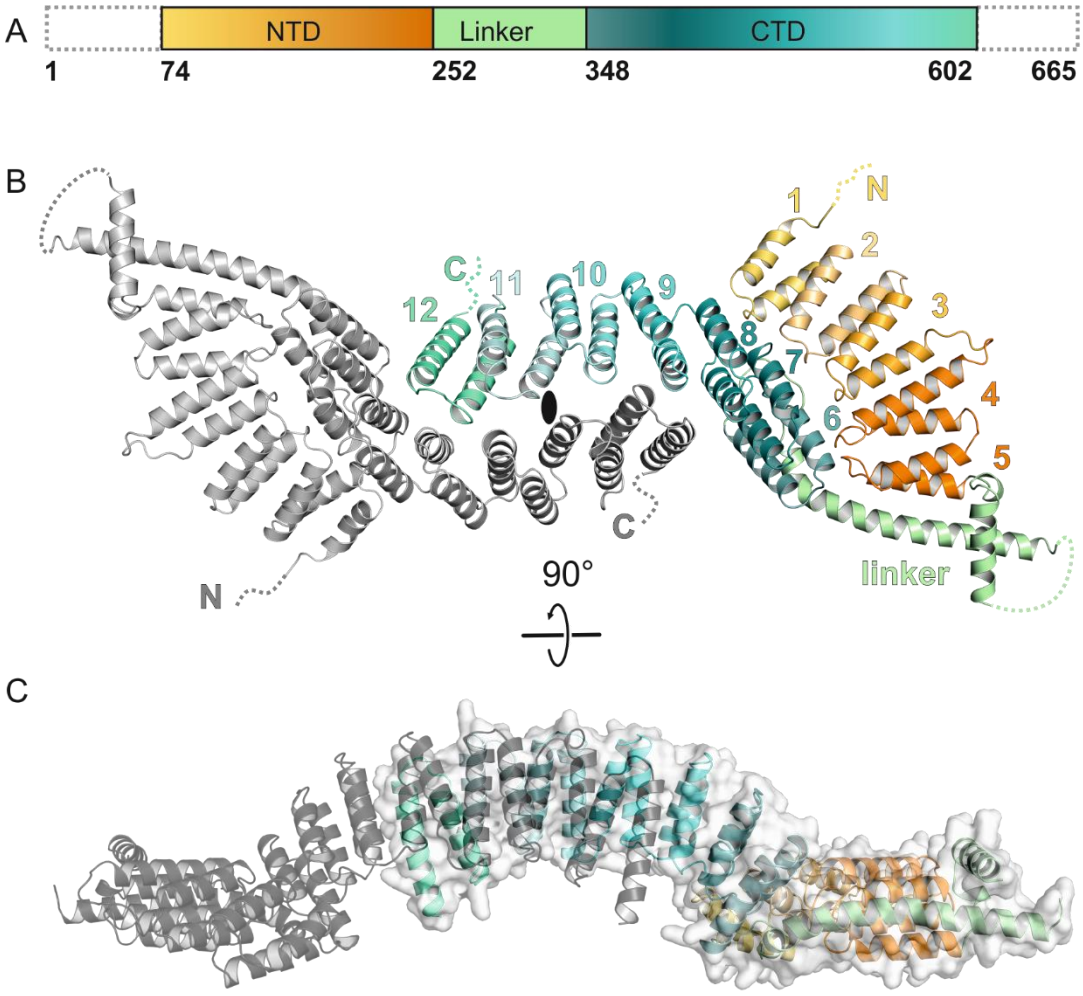

Figure 25 Overall structure of mPRPF39. **A** Schematic representation of the domain architecture of mPRPF39 comprising of 665 amino acid residues. The HAT-NTD is colored in a gradient from yellow to orange, the HAT-CTD in teal to aqua and the linker is colored in pale green. Regions that are unstructured are indicated as dotted lines. **B** Dimer arrangement of mPRPF39 drawn in cartoon representation. One monomer is drawn in gray and the second monomer according to the color gradient as described in panel A. HAT domains are numbered 1 through 12 starting at the N-terminus. **C** View is rotated by 90º and in addition the surface is shown for the second monomer.

An N-terminal domain (HAT-NTD) consisting of 5 HAT repeats, a C-terminal domain (HAT-CTD) comprised of 7 HAT repeats, and a linker region. The linker region is composed of a short and a long curved α-helix (Lα1 and Lα2) and an unstructured region (residues 275 to 294) that connect the α-helices. There was also no interpretable electron density observed for the 73 N-terminal as well as the 63- C-terminal residues. These areas of the

protein were predicted to be unstructured, leading us to the conclusion, that the lack of electron density is due to their flexibility. The areas lacking density were not degraded, as mass spectrometry shows the dissolved crystals to be of the expected size of the full-length protein (**Figure 26**).

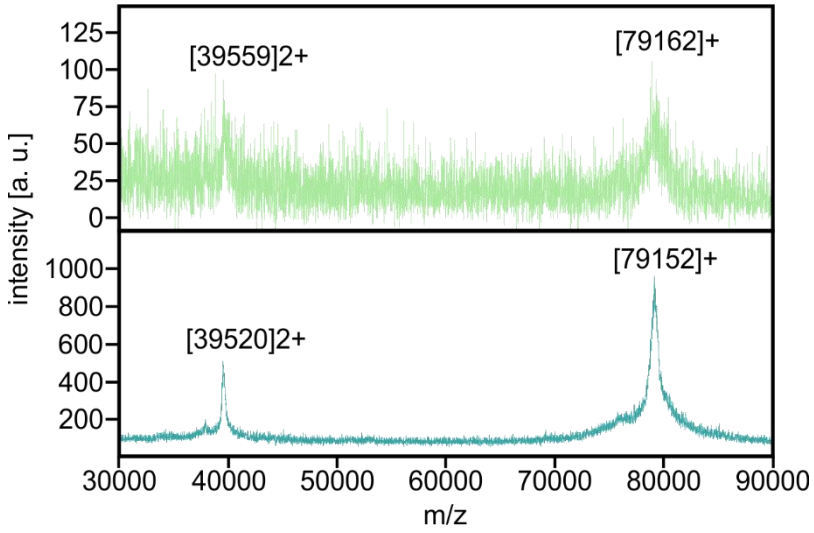

Figure 26 MALDI-MS of mPRPF39 crystals. The upper panel shows the measurement for SeMet full-length mPRPF39 crystals dissolved in water (pale green) and the lower panel for SeMet mPRPF39 (teal). The dissolved crystals show a mass almost identical to the control protein measurement. This indicates that the protein is not degraded in the crystals but that the undefined N- and C-termini and loop region are flexible.

Both mPRPF39 molecules are arranged in an anti-parallel manner as a homodimer in the crystal structure (**Figure 25**). The dimerization is mediated over the concave surfaces of the HAT-CTDs with the HAT-NTDs not playing a role in the dimerization. In the side view, the dimer appears to be bridge shaped with the HAT-NTDs making up the base and the HAT-CTDs the arch (**Figure 25**). The dimerization surface spans approximately 1536 Å^2 of buried surface for each molecule and the dimerization is predicted to be stable in solution by the PISA server (Krissinel and Henrick 2007).

To further investigate whether the dimer is stable in solution and not a crystallographic artefact, SEC-MALS experiments were performed to determine the molecular weight of mPRPF39. The SEC-MALS analysis shows the protein to have a molecular weight of 157 kDa confirming stable dimerization of mPRPF39 in solution (**Figure 27**). The dimer seems to be remarkably stable, still showing dimeric behavior under high salt conditions (900 mM NaCl) (**Figure 27**). The slight difference in retention volume observed between the experiments is most likely due to different hydrodynamic radius in the presence of high salt.

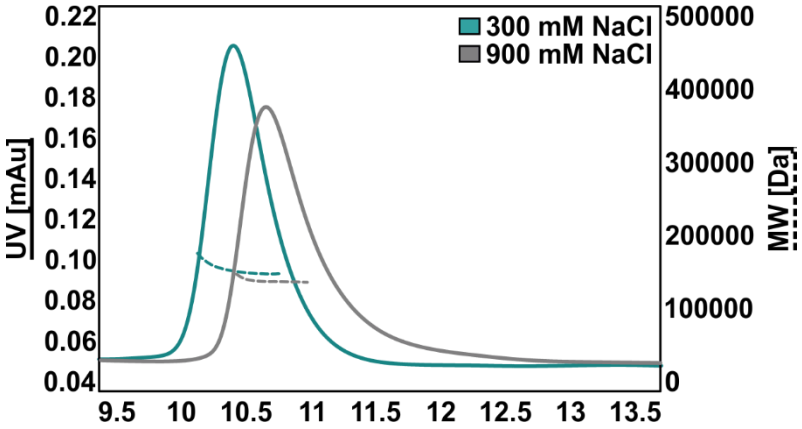

Figure 27 Analysis of the oligomeric state of mPRPF39 by SEC-MALS. Solid lines correspond to the absorbance measured at 280 nm for mPRPF39. For color-coding see inlet. Dashed lines correspond to the average molecular weight (MW) values of mPRPF39. mPRPF39^{wt} shows dimeric behavior. Even at high salt concentrations, the dimer cannot be disrupted, however the elution profile is slightly shifted caused by the high salt concentration.

Interestingly when the C-terminally truncated version of mPRPF39 was analyzed via SEC-MALS, three peaks corresponding to different molecular weights could be observed. The first peak is of inhomogeneous size distributed around 260 kDa, the second peak shows a size of 160 kDa, corresponding to the dimer size and the small third peak has a molecular weight of 84.6 kDa, close to the monomer size (**Figure 28**).

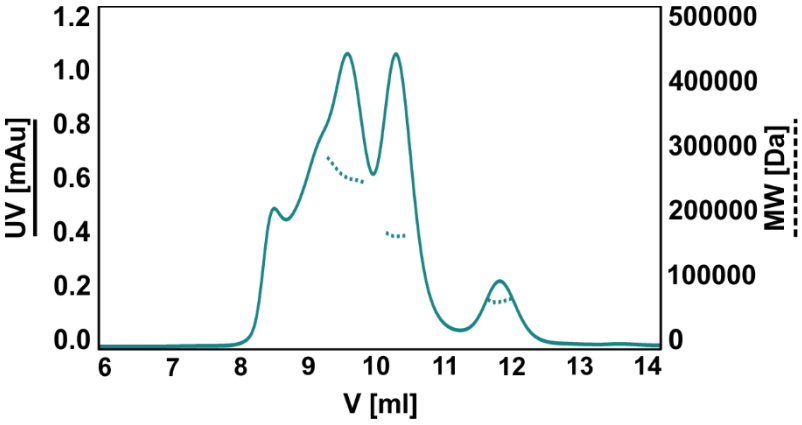

Figure 28 Analysis of the oligomeric state of mPRPF39$^{\Delta C}$ by SEC-MALS. Solid lines correspond to the absorbance measured at 280 nm. Dashed lines correspond to the average MW values of mPRPF39 and its variants in different oligomeric states. mPRPF39$^{\Delta C}$ shows a peak with high molecular weight not clearly assignable to an oligomeric state, a dimeric species and small amounts of monomeric species.

This is a hint, that the dimerization is stabilized over the unstructured C-terminus but truncating the C-terminus is not sufficient to completely abolish dimerization.

In addition to the SEC-MALS experiments, a Flag-IP experiment was also performed. In this experimental setup, HEK293T cells were transfected with GFP-mPRPF39 and either FLAG-mPRPF39 or FLAG empty vector as control. The cell lysates were then subjected

to α-FLAG beads and western blotting. The results demonstrate that GFP-mPRPF39 can only be co-precipitated in the presence of FLAG-mPRPF39 confirming dimerization (**Figure 29**). With this experiment it could be shown that not only *in crystallo* and in solution, but also in cellular environment mPRPF39 forms a dimer. This is especially interesting, because it offers a possible explanation for how the absence of a Prp42 homolog can be compensated in Metazoa.

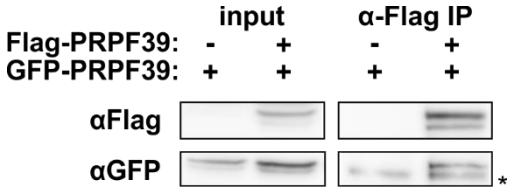

Figure 29 Co-Immunoprecipitation of differently tagged mPRPF39. Western blot of an αFlag-IP of Flag- and GFP-mPRPF39. GFP-mPRPF39 can only be pulled down in the presence of Flag-mPRPF39. Asterisk indicates an unspecific band.

7.6 The Homodimer Interface of mPRPF39

To investigate whether the homodimerization is indeed as crucial as it seems, the aim was to disrupt the dimerization by single point mutation. To find suitable candidates for mutation several properties of the protein were taken into consideration. First, the residues in the dimerization interface were determined and only these were considered for mutation (**Figure 30** A). Then the conservation of the surface residues was analyzed with the help of the ConSurf server (Ashkenazy 2016). Analysis of the mPRPF39 HAT-CTD shows that there is a high degree of conservation in residues located in the dimerization interface of mPRPF39 further cementing the physiological importance of the homodimerization and indicating that the homodimerization is conserved in metazoan (**Figure 30** B). Next, the electrostatic potential of mPRPF39 was calculated and plotted on the surface (**Figure 30** C). A very distinct charge distribution pattern can be observed in the HAT-CTD. There are two distinct patches, one with a strong positive and one with a strong negative charge. Interestingly, in the dimer structure, these patches interlock with each other, forming tow areas where the positive and negative patches interact with each other respectively in the second dimer copy. Overall it can be said that the dimerization is highly conserved among higher eukaryotes and is based largely on charged interactions.

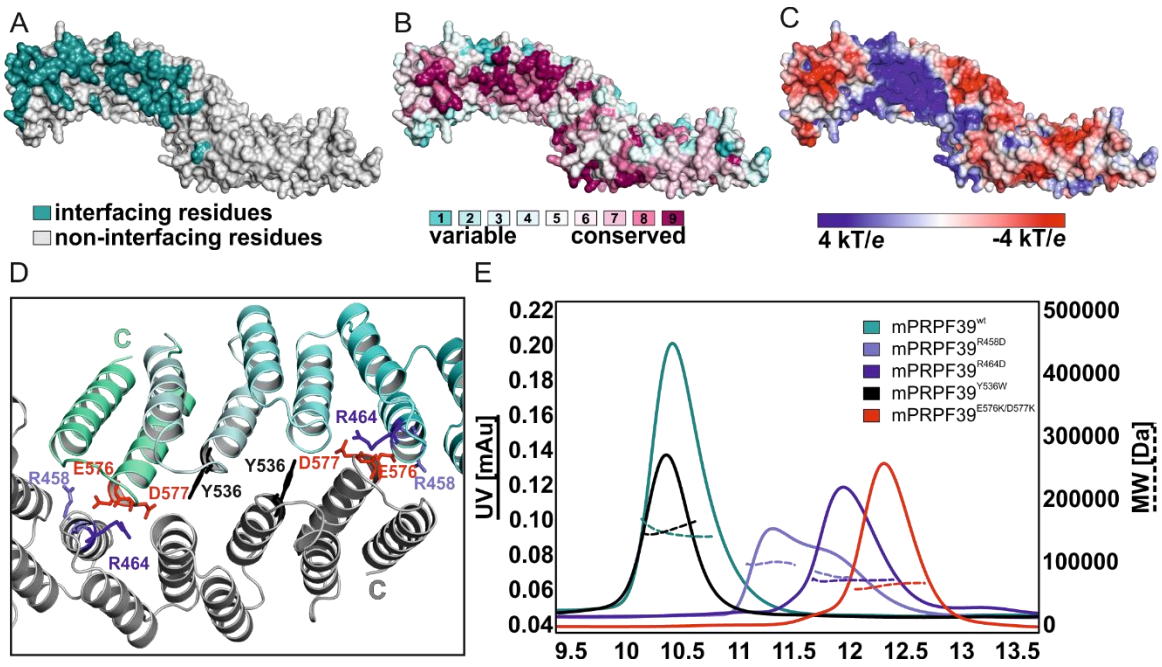

Figure 30 Surface analysis of mPRPF39 to find suitable candidates for mutation. **A** Surface representation of a mPRPF39 monomer. Interfacing and non-interfacing amino acid residues colored in teal and grey, respectively. **B** Amino acid residue conservation projected onto the surface of mPRPF39. Color-coding according to bins: bin 9, colored in magenta, contains the most conserved positions. Whereas bin 1 contains the most variable positions as indicated in cyan. The concave side of the HAT-CTD shows a distinct conserved area. **C** Electrostatic potential mapped on the surface of mPRPF39. There are two distinct positively and negatively charged patches on the concave side of the HAT-CTD. In panels A through C the view is rotated by 90° of the horizontal axis relative to **Figure 25** panel B. **D** Zoom into the dimerization interface. Amino acid residues subjected to site-directed mutagenesis are indicated in light blue (mPRPF39^{R458D}), blue (mPRPF39^{R464D}), red (mPRPF39$^{E576K/D577K}$) and black (mPRPF39^{Y536W}). **E** Analysis of the oligomeric state of mPRPF39 by SEC-MALS. Average MW values as measured by MALS are given by the selected peaks. Solid lines correspond to the absorbance measured at 280 nm for mPRPF39 and the variants. For color-coding see inlet. Dashed lines correspond to the MW values of mPRPF39 and its variants in different oligomeric states. mPRPF39^{wt} and mPRPF39^{Y536W} show dimeric behavior, mPRPF39^{R458D} shows a mixture of monomeric and intermediate species, while dimerization is completely abolished for mPRPF39^{R464D} and mPRPF39$^{E576K/D577K}$.

For mutagenesis the amino acid residues were selected based on their location in the interface, sequence conservation and charge. The single point mutations mPRPF39^{R458D}, mPRPF39^{R464D}, mPRPF39^{Y536W} and the double mutant mPRPF39$^{E576K/D577K}$ were chosen for further analysis. Two highly conserved residues in the positive patch (R458 and R464) were mutated to create a reversal of charge. Y536, as the only aromatic residue to be found in the interface, was mutated to a more bulky tryptophan residue. The idea here was to cause steric hindrance of dimerization. The last mutation performed was the double mutation of E576 and D577. These are highly conserved residues in the negative patch mutated to a positive charge (**Figure 30** D).

The dimerization properties of all four mutants was analyzed with SEC-MALS in comparison to the wild type mPRPF39. Mutation of Y536 to tryptophan seems to have no effect on dimerization. The mutant mPRPF39^{R458D} shows a mixture of monomeric species and species with an intermediate molecular weight. However, mPRPF39^{R464D} in the positively charged patch and mPRPF39$^{E576K/D577K}$ could be clearly identified as monomers compared to mPRPF39^{wt} (**Figure 30** E). These results confirm that the dimerization is mediated over interactions between the charged areas in the HAT-CTD of mPRPF39.

7.7 Functional relevance of mPRPF39 homodimerization

To test whether the homodimerization has a functional relevance, *in vitro* splicing assays were performed with both wild type mPRPF39 and monomeric mPRPF39. Both the mutant mPRPF39^{R464D} and mPRPF39$^{E576K/D577K}$ are monomeric in solution. For further studies mPRPF39^{R464D} was chosen because in contrast to the double mutant mPRPF39$^{E576K/D577K}$ it is a single point mutant and thus has less differences compared to the wild type mPRPF39 on a protein level.

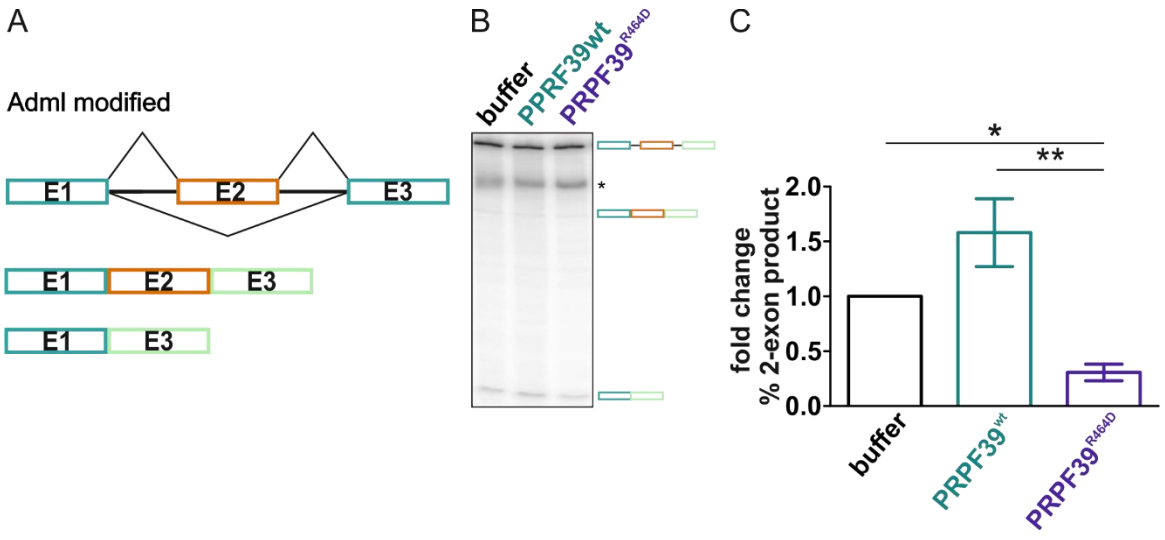

Figure 31 Influence of mPRPF39 dimerization on splicing. **A** Schematic representation of the modified 3 exon 2 intron Adml construct used for *in vitro* splicing. **B** Exemplary gel of RT-PCR of *in vitro* splicing reaction showing buffer mPRPF39^{wt} and mPRPF39^{R464D} treated samples. Samples treated with wild type protein show slightly elevated splicing and mutant treated samples show significantly reduced splicing. The experiment was performed three times with each time freshly prepared splicing active nuclear extract in technical triplicates. The asterisk indicates a non-specific band. **C** Quantification of 4 experiments performed in triplicates on four different days. Errors shown as standard deviation, a one sample T-test was performed, and p values shown in C (*≤0.05, **≤0.01).

For the *in vitro* splicing experiments a small-scale production of splicing active nuclear extracts was optimized for HEK293T cells. These nuclear extracts were then incubated

under splicing conditions with pre-mRNA and with either purification buffer as a control, mPRPF39^{wt} or mPRPF39^{R464D}. A modified Adml pre-mRNA construct was used containing three exons (**Figure 31**). After 30 minutes of splicing, the RNA was purified and an RT-PCR was performed. There are three bands visible, the unspliced pre-mRNA, a spliced three-exon product and a shorter spliced product with the second exon skipped. Only trace amounts of the three-exon product are visible on the gel, however, the two-exon product shows clear differences in abundance (**Figure 31**).

Addition of mPRPF39^{wt} to the splicing reaction shows a trend toward higher splicing efficiency when compared to samples treated with the purification buffer. However, this effect is not significant. This suggests that the prepared nuclear extracts may already contain sufficient amounts of PRPF39 and that the addition of more PRPF39 does not affect splicing so strongly in this setup. The samples treated with mPRPF39^{R464D} show significantly reduced splicing levels compared to both samples incubated with mPRPF39^{wt} and buffer (**Figure 31**). This indicates, that monomeric mPRPF39 is not functional in splicing and is able to partially displace the already present and functional PRPF39 in nuclear extracts.

7.8 *mprpf39* is an NMD Target

The previous experiment showed, that reducing or increasing the level of functional mPRPF39 homodimer can be used to control splicing efficiency. This is especially interesting, because mPRPF39 has an alternative exon which contains a premature stop codon (**Figure 32**).

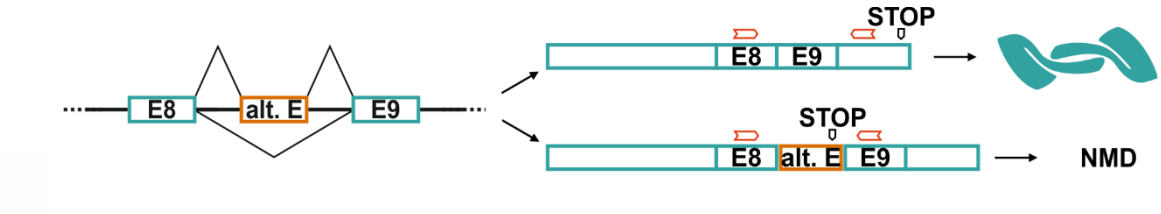

Figure 32 Alternative splicing in *mprpf39*. Schematic representation of alternative splicing in mPRPF39 with resulting mRNA and protein isoforms. An alternative exon (orange) can either be included or excluded. When it is excluded the mRNA can be translated to mPRPF39. If the alternative exon is included a premature stop codon is introduced which may lead to NMD. Red arrows indicate location of primers used for RT-PCR

Premature stop codons usually lead to NMD. This means that the inclusion of the alternative exon could act as a regulator for the mPRPF39 levels. The alternative exon and its flanking region are strongly conserved between mouse and human (**Figure 33**) indicating that the alternative exon could be used as a regulator across species.

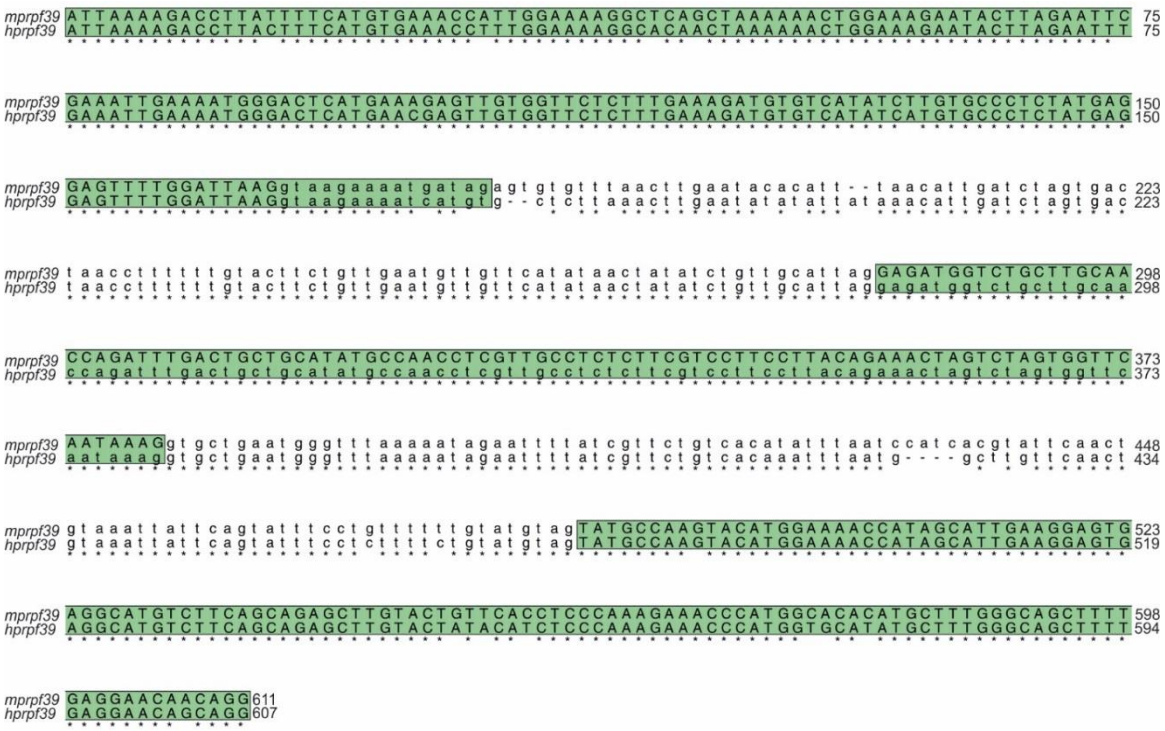

Figure 33 Conservation of the alternative exon and flanking exons between mouse and human. Alignment of the alternatively spliced region of mPRPF39 mRNA to hPRPF39 mRNA. The region is almost completely conserved with no sequence changes in the splice site. Identical nt are marked with an asterisk. Annotated exons are shown in capitals highlighted in pale-green; the middle exon is alternative.

RNA sequencing in murine T-cells shows, that the alternative exon is included to 19% in naïve and to 50% in memory T-cells. This could be validated by RT-PCR (**Figure 34**). Analysis of different tissues shows that the alternative exon is included in a differential manner as well, with especially high inclusion levels in testis and low levels in lymph node (**Figure 34**).

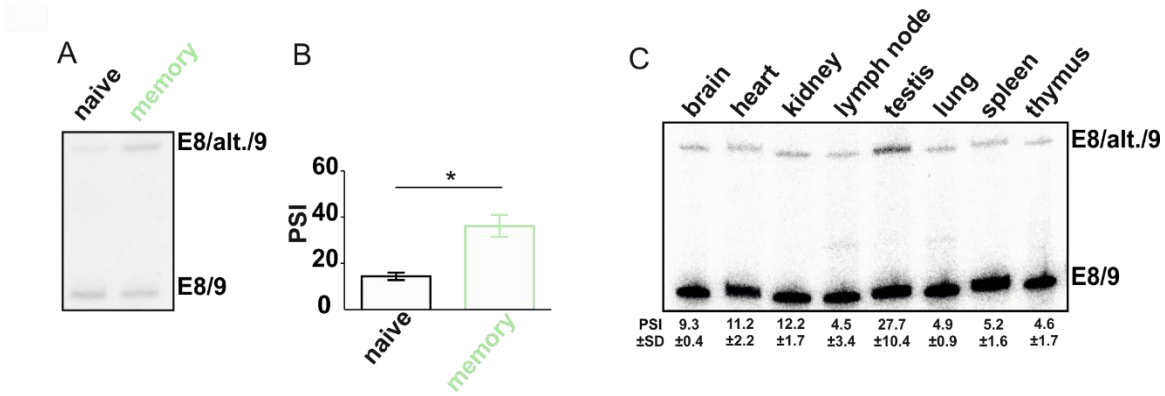

Figure 34 *prpf39* alternative exon inclusion is regulated in a cell type and tissue specific manner. **A** Representative validation of RNA-seq results by RT-PCR. **B** Quantification of validation as shown in A. PSI values represent means ± standard deviations for four independent experiments. **C** Exemplary gel of RT-PCR in different murine tissues. Results of quantification of experiments performed with samples from three individual mice are shown below the respective lanes. Percent spliced in (PSI) values represent means ± standard deviations.

The fact that the alternative exon could be detected at all in mRNA suggests, that it is not a strong NMD target. To determine whether or not inclusion of the alternative exon leads to degradation, murine and human T-cell lines, EL4 and Jsl1 respectively, were treated with Cycloheximide (CHX). CHX inhibits translation and, because NMD is strongly coupled to translation, CHX treatment also acts to block NMD. Indeed, RT-PCR shows that cells treated with CHX show higher levels of isoform containing the alternative exon (**Figure 35**).

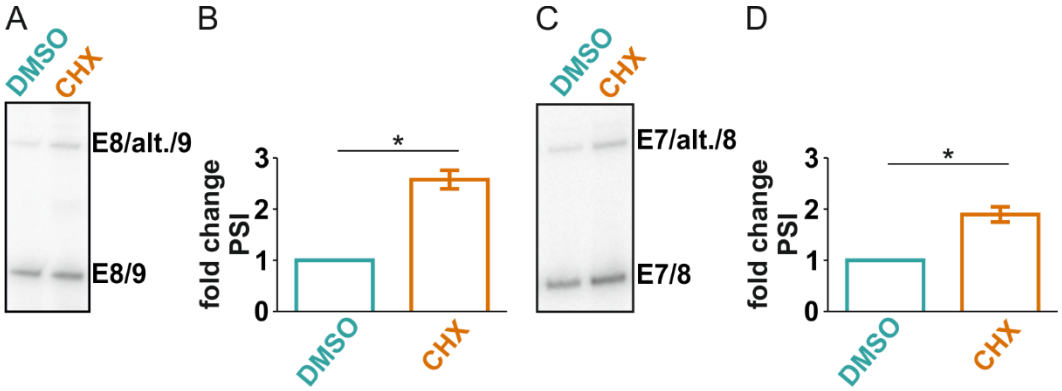

Figure 35 *prpf39* is an NMD target. Analysis of *mprpf39* (pre-)mRNAs in mouse EL4 cells (**A**, **B**) and human Jsl1 cells (**C**, **D**). Left – representative RT-PCR analyses; right – quantification. DMSO/CHX – DMSO/cycloheximide-treated samples. In the CHX treated cells, an accumulation of the transcript containing the alternative exon can be seen, indicating that this isoform is an NMD-target. Values represent means ± standard deviations of three independent experiments. Significance was estimated via a one-sample T-test (* – p ≤0.05, ** – p ≤0.01).

7.9 Comparison of mPRPF39 to yPrp39 and yPrp42

Monomeric mPRPF39 has an inhibiting effect on splicing and thus demonstrates that the mPRPF39 homodimer might substitute the yPrp39/yPrp42 heterodimer observed in yeast. To further investigate this, the mPRPF39 homodimer was compared with the heterodimer as seen in the cryo-EM structure of the yeast U1 snRNP (Li 2017). The cryo-EM structure shows yPrp39 and yPrp42 to be modeled into the electron density as a heterodimer. This heterodimer acts as an interaction hub for various alternative splicing factors, connecting them to the core U1 snRNP over the heterodimer. On a primary sequence level mPRPF39 shows 22% identity to yPrp39 and 23% identity to yPrp42 (Li 2017).

A DALI search (Holm and Rosenstrom 2010) was performed with the mPRPF39 protein coordinates for structural comparison. Tentatively, it was expected to find the yPrp39 homolog as the most similar structure, but surprisingly the structurally closest neighbors were the murine CSTF-77 (PDB entry 2OOE, (Bai 2007)), its *Kluyveromyces lactis*

homolog RNA14 (PDB entry 4EBA, (Paulson and Tong 2012)) and human SART3 (PDB entry 5JPZ (Grazette 2016)) (**Table 7**).

Table 7 Most closely related structures of mPRPF39 identified by Dali search (Holm and Rosenstrom 2010).

PDB	RMSD [Å]	sequence identity [%]	Z-score	Protein	Lit.
2OOE	4.9	18	29.9	CSTF-77	(Bai 2007)
5JPZ	5.5	17	27.5	SART3	(Grazette 2016)
4EBA	5.7	15	27.3	RNA14	(Paulson and Tong 2012)
5UZ5	4.5	19	25.7	Prp42	(Li 2017)
5UZ5	7.6	6	20.3	Prp39	(Li 2017)

Interestingly, the structurally closest neighbor among the snRNP associated proteins was not yPrp39 (PDB entry 5UZ5, (Li 2017)) with an RMSD of 7.6 Å, but yPrp42 from the same structure with an RMSD of 4.5 Å. To help understand the relation of all three proteins to each other, especially under the aspect of how the homodimer could functionally substitute the heterodimer, these proteins were analyzed in more detail.

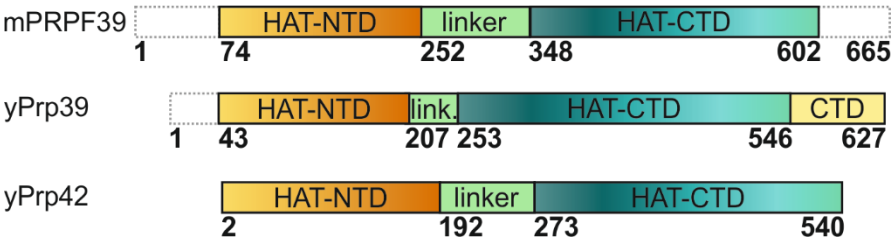

Figure 36 Schematic representation of the domain architecture of mPRPF39, yPrp39 and yPrp42. The HAT-NTD is colored in a gradient from yellow to orange, the HAT-CTD in teal to aqua, the linker is colored in pale green and the CTD in pale yellow. Regions that are unstructured are indicated as dotted lines.

All three proteins have the same overall composition. They all have a HAT-NTD and a HAT-CTD connected by a linker α-helix (**Figure 36**).

Figure 37 Structure-based alignment of mPRPF39 and yPrp42 (PDB entry 5UZ5(Li 2017)). The secondary structure elements of mPRPF39 and yPrp42 are shown above and below the alignment, respectively. α-helices are depicted as cylinders, β-strands as arrows. Slanted lines indicate sections of mPRPF39 and yPrp42, which are not included in the crystal structures. Secondary structure elements are colored by domain affiliation; mPRPF39 – HAT-NTD, orange; linker, pale-green; HAT-CTD, teal; yPrp42 – HAT-NTD, light-gray; linker, gray; HAT-CTD, dark-gray. Colored triangles indicate the positions of residues mutated in the homodimer interface of mPrp39. Color-coding: mPRPF39^{R458D} and mPRPF39^{R464D}, blue; mPRPF39^{Y536W}, black; mPRPF39$^{E576K/D577K}$, red.

At a first glance, the most distinct differences are to be seen in the shortened linker region and C-terminal extension of yPrp39. While the HAT-NTDs of mPRPF39 and yPrp42 consist of 10 α-helices arranged as HAT-repeats, the HAT-NTD of yPrp39 is made up of only 9 α-helices, lacking the N-terminal α-helix of the first HAT-repeat (**Figure 37**).

Figure 38 Structure-based alignment of mPRPF39 and yPrp39 (PDB entry 5UZ5(Li 2017)). The secondary structure elements of mPRPF39 and yPrp39 are shown above and below the alignment, respectively. α-helices are depicted as cylinders, β-strands as arrows. Slanted lines indicate sections of mPRPF39 and yPrp39, which are not defined in the structures. Secondary structure elements are colored by domain affiliation; mPRPF39 – HAT-NTD, orange; linker, pale-green; HAT-CTD, teal; yPrp39 – HAT-NTD, light-gray; linker, gray; HAT-CTD, dark-gray. Colored triangles indicate the positions of residues mutated in the homodimer interface of mPRPF39. Color-coding: mPRPF39^{R458D} and mPRPF39^{R464D}, blue; mPRPF39^{Y536W}, black; mPRPF39$^{E576K/D577K}$, red.

There is no observable density for the first 42 amino acid residues in the cryo-EM structure. However, when performing secondary structure predictions, the algorithms predict an α-

helix in the N-terminus corresponding to the missing α-helix. Most likely the local resolution did not allow modeling the α-helix.

Large differences can be observed in the linker region of mPRPF39, yPrp39 and yPrp42. The murine structure shows the linker region to be an arrangement of three α-helices with a dominating 28 amino acid residue long α-helix (Lα3 in **Figure 37**). The long linker α-helix is present in all three proteins but is vastly shortened in both yPrp39 and yPrp42. Moreover, the yPrp42 linker consists of only two α-helices and in the yPrp39 linker region no perpendicular α-helix whatsoever is to be found (**Figure 37** and **Figure 38**). The HAT-CTDs of all three proteins are composed of 14 α-helices each, arranged in 7 HAT repeats.

Even though the overall architecture of the HAT domains is similar, on average the loop regions connecting the α-helices are longer on both yeast proteins. For example, the linker that connects the α-helices of the 9th HAT repeat in both yPrp39 and yPrp42 is highly elongated and creates an additional interaction platform for dimerization (**Figure 37**, **Figure 38** and **Figure 39**).The linker between HAT repeat 9 and 10 in yPrp39 is also longer compared to mPRPF39. As a result of the elongation, it extends into the convex side of the CTD.

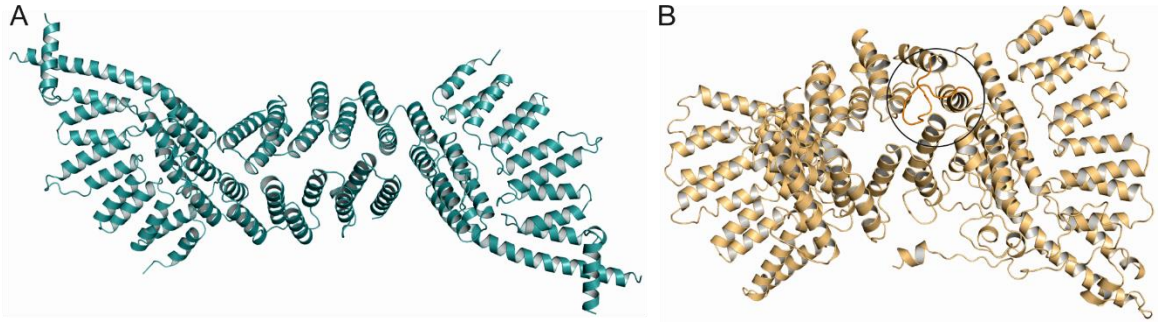

Figure 39 Interaction surface between Prp(f)39 variants and yPrp42. Cartoon representation of the mPRPF39 homodimer (**A**) and the yPrp39/yPrp42 heterodimer (**B**). yPrp42 (light orange) and yPrp39 (light orange) show elongated linkers between the HAT α-helices compared to mPRPF39 (teal). This creates additional interfacing surfaces in the yeast heterodimer highlighted in orange and by a circle.

Beside the central HAT repeat containing domains, both mPrp39 and yPrp39 have N-terminal as well as C-terminal extensions in their amino acid sequence. While in the crystal structure of mPrp39 these extensions could not be resolved in the electron density, the C-terminal domain (CTD) of yPrp39 is structured (**Figure 36** and **Figure 37**). However, we cannot exclude that the C-terminal extension of mPrp39 gets structured in the context of U1snRNP assembly and hence might act in protein-protein interactions.

The similar domain architecture of mPRPF39 to yPrp39 as well as yPrp42 makes the homodimer arrangement of mPRPF39 very similar to that of the heterodimer of

yPrp39/yPrp42. However, there are notable differences. In both systems the dimerization is mediated by their respective HAT-CTDs. The mPRPF39 homodimer shows lower curvature in its HAT-CTD leading to an overall more elongated dimer compared to the yPrp39/yPrp42 heterodimer (**Figure 40**). When superimposing mPRPF39 with either yPrp39 or yPrp42, the HAT-NTDs align better compared to the HAT-CTDs (**Figure 40**). The lower degree of curvature observed for the murine homodimer causes the HAT-NTD of the second monomer to slap out and this leads to a spatial location far away from its corresonding HAT-NTD in the superimposition.

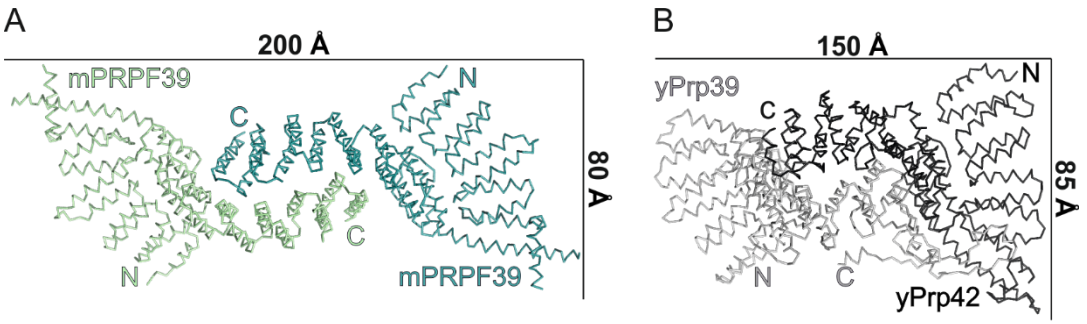

Figure 40 Structural comparison of mPRPF39 and yPrp39 as well as yPrp42 structures. **A** mPRPF39 homodimer shown in ribbon representation. mPRPF39 monomers are colored in deep teal and pale green. The dimensions of the dimer are indicated in Å. **B** yPrp42/yPrp39 heterodimer shown in ribbon representation. yPrp42 is colored in dark grey and yPrp39 in light grey. The dimensions of the dimer are indicated in Å.

The interface of yPrp39 and yPrp42 is dominated by hydrophobic interactions. This is in contrast to the mPRPF39 homodimer, where salt-bridges form the most important dimerization contacts. The only salt-bridge observed in the yeast heterodimer is between the yPrp39 CTD and the yPrp42 HAT-NTD.

Interestingly, mPRPF39 does not only show the patch of conservation on the concave side of the HAT-CTD but also on the concave side of the HAT-NTD (**Figure 41**). Incidentally this region is also of high relevance in the yeast system. The conserved patch in the HAT-NTD corresponds to the area that serves as a binding module for core U1 snRNP proteins to the yPrp39/yPrp42 heterodimer, indicating similar binding of the homodimer to the U1 snRNP in metazoans. This is supported by IP-experiments showing that hPRPF39 can bind to U1C, which binds to the HAT-NTD of yPrp42 (Li 2017).

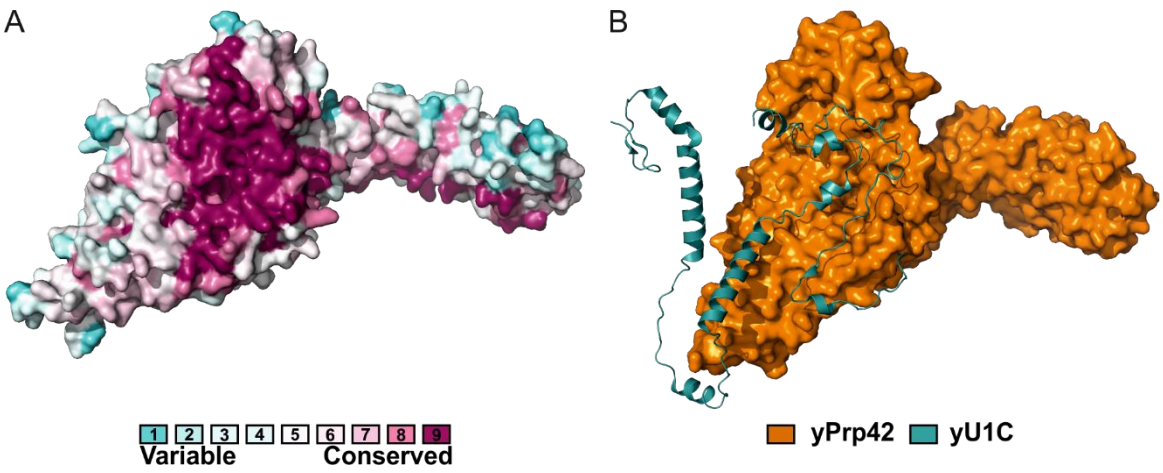

Figure 41 mPRPF39 HAT-NTD as a possible binding interface for the U1 snRNP core. **A** Sequence conservation projected onto the surface of mPRPF39. Color-coding according to bins: bin 9 (magenta) contains the most conserved positions, bin 1 (cyan) contains the most variable positions. The concave side of the HAT-CTD is highly conserved. Same view as in **Figure 25**B rotated by 180°. **B** Surface and cartoon representation of yPrp42 yU1C (PDB entry 5UZ5 (Li 2017)), respectively. In panels A and B the view is rotated by 180° of the vertical axis relative to **Figure 25** panel B.

7.10 Coevolutionary Connection Between PRPF39 and the U1 snRNA Length

In the yeast cryo-EM structure, the convex side of yPrp42 has a striking positively charged groove that can accommodate the U1 snRNA. A kink in the SL 2 of the U1 snRNA causes it to fold over and bind the yPrp42 HAT-CTD (**Figure 42**). Because the hypothesis that mPRPF39 can not only substitute yPrp39 but also yPrp42 in metazoan was suggested (Li 2017), it was interesting to investigate whether one can find a corresponding groove in mPRPF39 to bind the U1 snRNA.

Upon analysis, it becomes apparent, that mPRPF39 is completely lacking a binding groove for RNA (**Figure 42**). This led to the investigation of the U1 snRNA structures in more detail. The yeast U1 snRNA appears to be much more complex than the murine one. Most notably, the SL 2 is much shorter in the murine U1snRNA lacking the part that folds over to bind yPrp42 in the yeast system (**Figure 42**).The SL 3 is strongly elongated and has a more complex branched structure in parts compensating for the lack of a SL 4. Interestingly an evolutionary study focusing on the SL 3 (Mitrovich and Guthrie 2007) suggested that the duplication of prp39 went along with an increasing length of SL3. However, the buried surface area between yPrp42 and SL 3 is much smaller than between yPrp42 and SL 2, indicating a larger influence of SL 2 on the dimeric system.

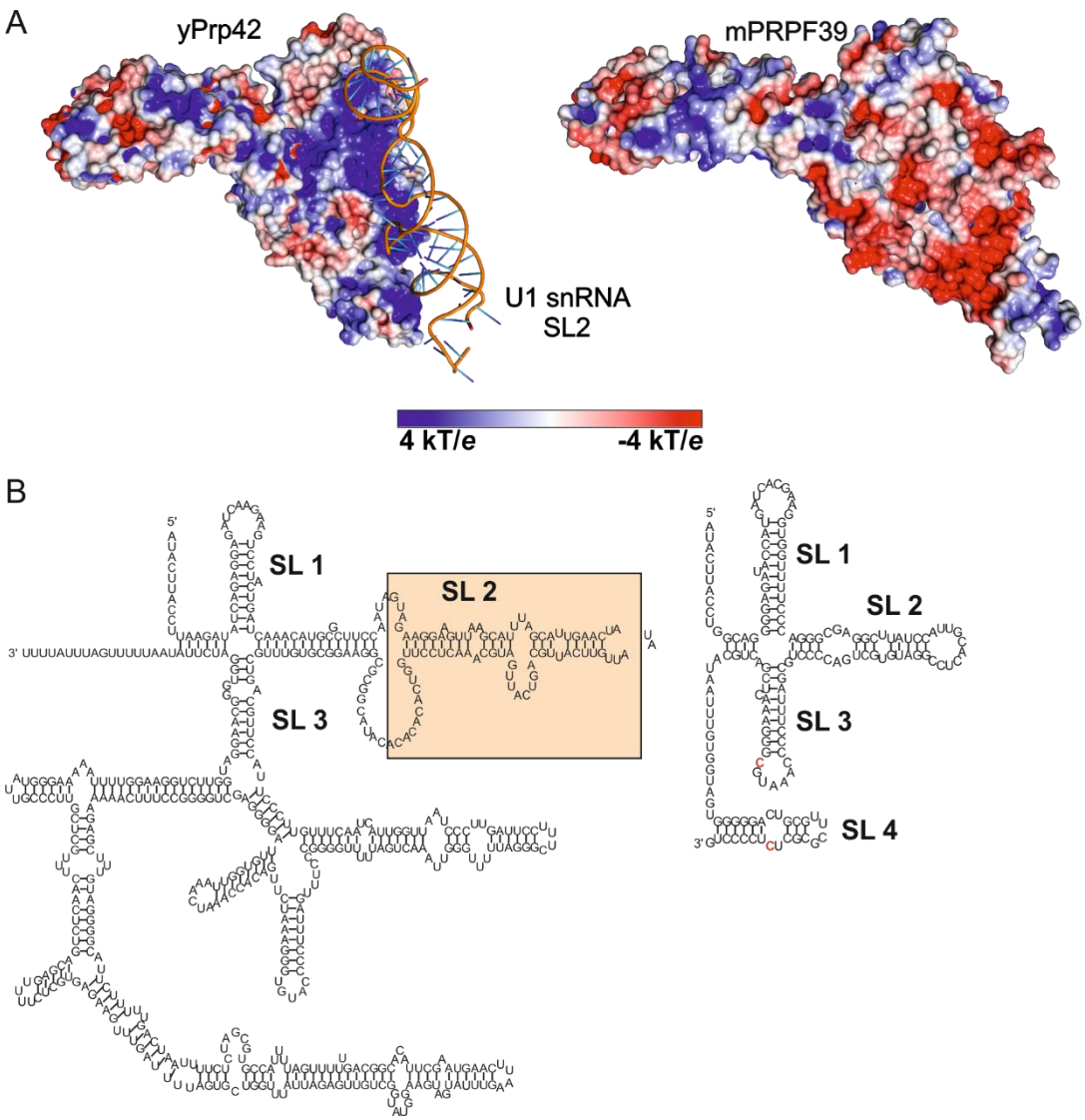

Figure 42 Analysis of surface charge in connection with U1 snRNA binding. **A** Electrostatic potential mapped on the surface of yPrp42 (left) and mPRPF39 (right) in the same orientation as in **Figure 25** B, with the U1 snRNA SL 2 shown as cartoon bound to yPrp42. Only yPrp42 shows the positively charged grove needed to accommodate the RNA. **B** Schematic representations of yeast U1 snRNA (left) and of murine U1 snRNA (right). Characters in red indicate nts that are not conserved between human and murine U1 snRNA. The pale orange box highlights the elongated SL 2 in yeast

This led us to address the question of whether the evolution from heterodimerization in yeast versus homodimerization in metazoan might be evolutionarily linked to the shortening of the U1 snRNA SL 2 and SL 3. To address this hypothesis, we decided to analyze the length of U1 snRNAs in different organisms. We obtained all annotated U1 snRNA sequences from Rfam (Kalvari 2018) and calculated the median sequence length per organism and plotted these against each other (**Figure 43**).

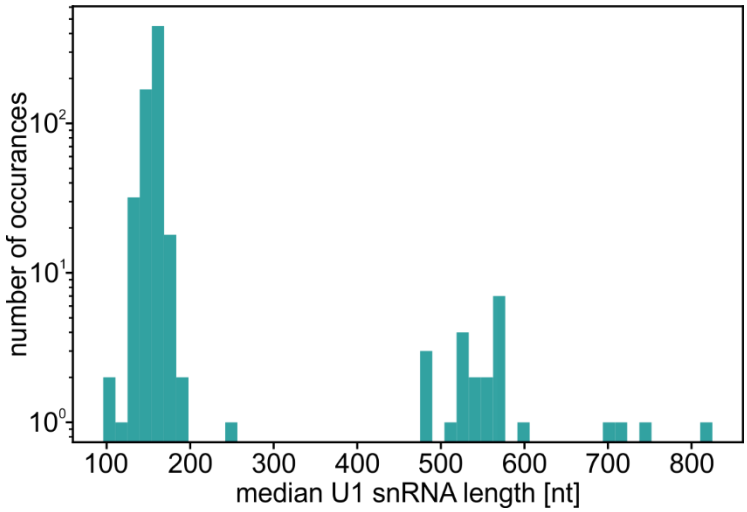

Figure 43 Analysis of the U1 snRNA lengths in different organisms. The number of organisms plotted against the median U1 snRNA length in nt. There are two populations clustering around 160 nt and 550 nt

The sequence lengths range between 96 and 825 nt, but most of the lengths cluster around two peaks. All higher eukaryotes lie on the lower end of the range clustering around a size of 164 nt. Members of the fungal family tend to have longer U1 snRNAs, most of them longer than 450 nt (**Figure 43**). We assume that a longer U1 snRNA corresponds to an elongated SL 2 and SL3.

However, there were a few exceptions of organisms with short U1 snRNAs in fungi. *Aspergillus fumigatus* and *Debaryomyces hansenii* have sequence lengths of 148 and 137 nt, respectively. When performing secondary structure predictions (Reuter and Mathews 2010), we observed that their U1 snRNA structures resemble the human structure more closely with a short SL 2 and SL 3 (**Figure 44** A). The *Candida albicans* snRNA has a comparatively short U1 snRNA comprised of 244 nt. It shows a surprisingly shortened SL 3, but the yeast like SL 2 remains elongated (**Figure 44** B).

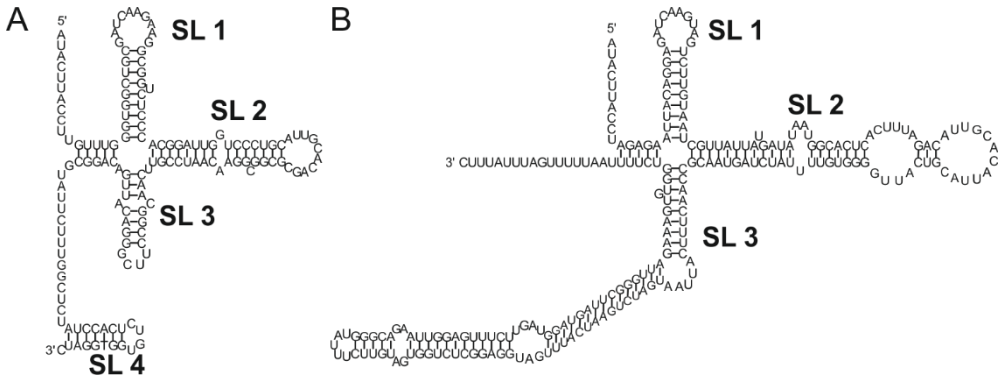

Figure 44 Schematic representation of U1 snRNA. SL structures are labelled. **A** U1 snRNA of *A. fumigatus*. **B** U1 snRNA of *C. albicans*.

To correlate the length of the U1 snRNP with the presence or absence of a Prp42 homolog, we queried NCBI for hPRPF39 as well as yPrp39 and yPrp42 homologs. The overlap between the organisms included in this query and the organisms with annotated U1 snRNAs were plotted in a phylogenetic tree against the median U1 snRNA length and existence of a Prp42 homolog (**Figure 45**).

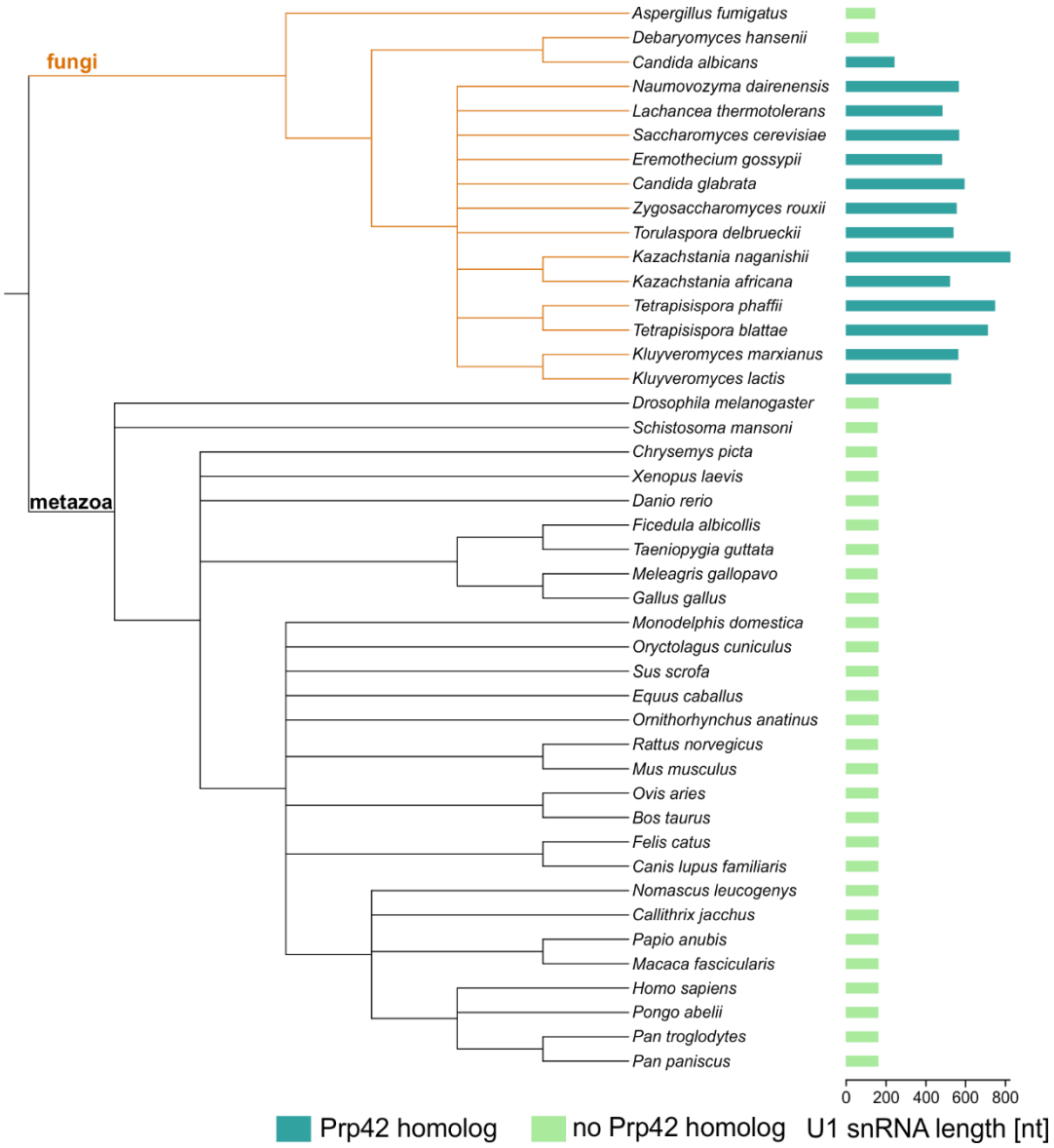

Figure 45 Phylogenetic analysis of the U1 snRNA and dimerization. Phylogenetic tree of the organisms included in the evolutionary analysis. The fungal clade is colored orange and the metazoan one black. On the right side the respective median U1 snRNA nt lengths are plotted as bar charts. The organisms with an annotated Prp42 are colored in teal and the organisms without a Prp42 homolog are colored pale green. Most fungal organisms except for *A. fumigatus* and *D. hansenii* have a Prp42 homolog. These two organisms also have short U1 snRNAs. A list of all NCBI GeneIDs and NCBI sequence identifiers can be found in **Table 1**.

Notably, we only find fungal organisms with Prp42 homologs (**Figure 45**). Interestingly the only fungi without an annotated Prp42 homolog are *Aspergillus fumigatus* and

Debaryomyces hansenii (**Figure 45**). The same fungi mentioned above that have short U1 snRNA lengths. This indicates that only in the absence of a Prp42 homolog, and thus the presence of a Prp39 homodimer, a short U1 snRNA SL 2 and SL 3 is observed. In contrast, organisms with a Prp39 and Prp42 homolog present as a heterodimer have an elongated SL 2 and SL 3. This strongly supports a co-evolutionary mechanism on protein and RNA level.

The exon-intron structure of genes in fungal genomes was analyzed in a previous study (Ivashchenko 2009). This study is especially interesting, as several of the organisms found in the overlap of the queries performed in the course of this thesis are included in it. The authors find a large variance in percentage on intron containing genes spanning from as low as 0.7 % to as high as 97 %. The maximum number of introns contained in a gene is also widely distributed between 2 and 18 (**Table 8**). This striking difference in splicing complexity between the fungal organisms does not seem to correlate with either the genome size, total amount of chromosomes or total amount of genes. Thus, the question arises as to what the cause for such vast differences could be.

Table 8 Characteristics of sequenced fungal genomes (Ivashchenko 2009).

species	Genome size, [kb]	Total chromosomes	Total genes]	Portion of intron containing genes [%]
C. glabrata	12201	13	5083	1.5
K. lactis	10690	6	5216	2.4
E. gossypii	8743	7	4707	4.5
S. cervisiae	12070	16	5495	4.5
D. hansenii	12221	7	6693	5.0
Y. lipolytica	20502	6	5970	10.6
S. pombe	12535	3	4909	45.6
A. fumigatus	28810	8	9811	78.0
N. crassa	23277	7	6483	79.6

When one puts this data in relation to our analysis of the U1 snRNA length and homo- or heterodimerization of the Prp39/Prp42 system, it becomes apparent, that fungi with less complex splicing have long U1 snRNAs und a Prp39/Prp42 heterodimer, whereas fungi with more complex splicing have short U1 snRNAs and a Prp39 homodimer (**Table 9**)

Table 9 Analysis of splicing complexity in connection with U1snRNP 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 (Li 2017). Intron containing genes are given in [%] and the maximal number of introns per gene is given according to Ivashchenko *et al*. (Ivashchenko 2009)

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
C. glabrata	x		595	147	396	1.5	2
K. lactis	x		528	132	330	2.4	2
E. gossypii	x		483	134	276	4.5	2
S. cervisiae	x		568	122	377	4.5	2
D. hansenii	x		165	45	28	5.0	4
Y. lipolytica	x		150	43	14	10.6	4
S. pombe		x	149	44	22	45.6	15
A. fumigatus		x	149	46	18	78.0	18
N. crassa		x	130	43	22	79.6	11

Two irregularities to this hypothesis can be observed. *D. hansenii* and *Y. lipolytica* show relatively low splicing complexity but have short U1snRNAs. In the previously shown analysis, we only found an annotated Prp39 homolog for both of these organisms (**Figure 45**). However, as they do not fit the established pattern of low complex splicing and long U1 snRNAs with a heterodimeric system, these two cases were investigated further. The yeast protein sequence of yPrp39 and yPrp42 were used to blast them both against the complete organism proteome of *D. hansenii* or *Y. lipolytica.* In both cases not only the annotated Prp39 homolog is found, but an additional second protein is found to have high similarity to both yPrp39 and yPrp42. This indicates a heterodimeric system in these two organisms, even though only a Prp39 homolog is annotated. This can be seen as implication that these two organisms are intermediates in which the U1 snRNA is already shortened, but that still have a heterodimeric Prp39/Prp42 system. This leads to slightly more complex splicing compared to the fungi with long U1snRNAs, with the large elevation in complexity only visible in fungi with a Prp39 homodimer. This is also reflected in the fact that, when looking at the portion of intron containing genes and the maximum number of introns per gene, these two organisms lie exactly between the less complex and the higher

complex organisms (**Figure 45**). These findings show the importance of the homodimeric Prp39 arrangement in conjuncture with the U1 snRNA to the development of more complex splicing.

8 Discussion

In this thesis I established and optimized the expression of murine mPRPF39. Choice of the optimal bacterial strain, expression media and lysis buffer proved to be instrumental in this goal. Using limited proteolysis and mass spectrometry, a 66 kDa stable fragment of mPRPF39 comprising of the HAT-NTD, HAT-CTD and linker could be determined. In addition to the full-length protein only a C-terminally truncated mPRPF39 could be expressed and purified. Both of these constructs yielded crystals. However, the mPRPF39$^{\Delta C}$ variant never diffracted beyond 6 Å and only the full-length construct yielded a final dataset at 3.3 Å.

I solved the crystal structure of the full-length mPRPF39 protein by SeMet SAD, showing the protein to be arranged in a distinct three domain manner with a HAT-NTD that is connected via an extended linker to the HAT-CTD. In the crystal structure two mPRPF39 molecules are arranged in the asymmetric unit as a homodimer. With SEC-MALS and Co-IP experiments I could show that the dimerization is not only present *in crystallo*, but also in solution and in a cellular environment.

Guided by the structure I have determined which structural elements and organizational aspects are fundamental for the dimerization using mutational analyses. The results demonstrated that the dimerization is mediated by conserved charged residues in the HAT-CTD. Moreover with *in vitro* splicing experiments I could show, that adding non-functional monomeric mPRPF39 to the experimental setup has a detrimental effect on splicing.

In conjecture with these results I analyzed the alternative splicing event of mPRPF39 which causes the inclusion or exclusion of an alternative exon containing a stop codon. I could show that the inclusion and exclusion is regulated in an activation dependent and tissue specific manner. When the alternative exon is included, either the *mprpf39* pre-mRNA is degraded via the NMD pathway, or a non-functional mPRPF39 is translated.

Furthermore, my work reveals an evolutionary link between the length of the U1 snRNA and the presence or absence of a Prp42 copy. This can be correlated to the complexity of splicing in an organism.

While my work still does not clarify all the details of the precise role of PRPF39 in spicing in metazoans, together with the recent cryo-EM structures (Li 2017, Bai 2018, Plaschka 2018, Zhan 2018) it provides a framework that helps us understand how a spliceosome without a Prp42 homolog could function.

8.1 The Functional Subunit of mPRPF39 is a Homodimer

The crystal structure of mPRPF39 shows the protein to have three distinct subdomains arranged as a dimer. This overall protein architecture of mPRPF39 resembles the arrangement of other HAT repeat proteins involved in mRNA processing such as SART3 (Grazette 2016) and RNA14 (Paulson and Tong 2012).

SART3 plays an important role during translesion DNA synthesis and spliceosome recycling. SART3 is essential for formation of the U4/U6 di-snRNP (Rader and Guthrie 2002) and as a homodimer it is essential for key interactions during this process (Huang 2018).

RNA14 is part of the cleavage factor IA, which is required for 3'-end processing together with the cleavage factor II and the polyadenylation factor I. This process, including cleavage and polyadenylation is an essential step in pre-mRNA processing and defects in the 3'-end processing can have a detrimental effect on mRNA export, stability and translation (Paulson and Tong 2012). Like in mPRPF39 and SART3, RNA14 has two distinct HAT domains, that are connected by a linker α-helix. The structure also shows dimerization mediated over its HAT-CTDs. Interestingly however, it could be shown that only one copy RNA 15, one of the components of the cleavage factor IA, binds to the RNA14 dimer (Paulson and Tong 2012).

Dimerization is not only observed for HAT repeat proteins but is an important part of many biological processes providing, for example, diversity and specificity, regulating gene expression and activity of enzymes (Hashimoto and Panchenko 2010). Oligomerization can help reducing the genome size, still allowing for large structure assemblies through modular complex formation. Additionally small protein subunits will fold more readily than a single large protein would (Jones and Thornton 1995, Goodsell and Olson 2000, Marianayagam 2004). The reduced surface area of oligomers compared to its monomers can also have a beneficial effect on protein stability, by protection against denaturation (Miller 1987). Homodimers are especially interesting cases. More than half of the oligomers observed are dimers, the bulk of the dimers being homodimers (Mei 2005). They occur more frequently than could be explained by random evolution and have approximately double as many interaction partners as proteins that do not interact with themselves (Ispolatov 2005).

The simplest functionality of dimerization is as a general sensor for protein concentrations. If the required concentration is exceeded, dimerization can occur and can be a stimulus for enzymatic activation. Dimerization can also provide new intermolecular interfaces that allow proteins to bind that would otherwise not bind to the monomer (Marianayagam 2004).

In some cases asymmetry can be induced in homodimers by binding of an additional component (Swapna 2012).

It is very interesting, that the crystal structure of mPRPF39 shows the protein to be arranged as a homodimer, because in the yeast system, the structurally relevant heterodimer of yPrp39 and yPrp42 observed in cryo-EM studies (Li 2017, Bai 2018, Plaschka 2018) shows a similar mode of dimerization over its C-terminus.

The yPrp39/yPrp42 heterodimer is a crucial scaffolding subcomplex. It acts as a hub for multiple protein-protein interactions for example the contact between the U1 and the U2 snRNP, indicating that the yPrp39/yPrp42 heterodimer is important for the precise spatial positioning of the snRNPs to each other. As a Prp42 homolog is unknown in metazoan, this suggests, that the PRPF39 homodimer could maybe functionally substitute the heterodimer observed in yeast.

In crystal structures, proteins are arranged in a multimeric fashion very often, due to the nature of crystals. The smallest unit of the crystal is the asymmetric unit (ASU). The whole crystal is made up of the same ASU which can be rotated and translated according to the symmetry operators specific to the space group of the crystal. The crystal packing helps stabilize the crystal, leading to interactions within and between ASUs. These interactions are not always biologically relevant, as the ASU is not defined in a biological context (Valdar and Thornton 2001). Consequently, verifying the dimerization with other methods was indispensable. The fact that the dimerization can be seen in solution as well as in a cellular environment was a promising hint, that mPRPF39 could indeed substitute the heterodimer observed in yeast.

The unstructured C-terminal domain of mPRPF39 seems to have stabilizing properties towards the homodimerization but is not sufficient to disrupt the dimer. Removing the CTD gives rise to a small amount of monomeric mPRPF39, to the dimeric protein and, in addition, to a peak with a disperse molecular weight distribution, most likely corresponding to multimeric aggregated mPRPF39. Most likely the mPRPF39 CTD is necessary for overall stabilization of the protein, fixing it in its preferred dimeric state. When looking at the structure of the yPrp39/yPrp42 the yPrp39 CTD binds across the concave side of the yPrp42 HAT-NTD. This supports the hypothesis of the CTD being important to stabilize the mPRPF39 homodimer with additional contacts if the mPRPF39 CTD assumes a similar fold.

This thesis shows that the dimerization in mPRPF39 is mainly mediated by charged residues located in the concave face of the HAT-CTD. Mutation of just a single residue in the positively charged patch and a double mutation in the negative patch are sufficient to

completely abolish dimerization. Other homooligomers with complementarily charged patches mediating dimerization have been previously described in literature, e.g. the ryanodine receptor 1 (Lee and Allen 2007) and the ribonucleotide reductase small subunit M2 (Chen 2014).

However, the yPrp39/yPrp42 dimerization is not based on charged residues at all. The only salt bridge is between the yPrp39 CTD and the yPrp42 HAT-NTD. To further compare the homodimer of mPrp39 and the heterodimer of yPrp39/ yPrp42 we used the cryo-EM structure of the U1 snRNP (Li 2017) and inspected protein-protein interactions as well as snRNA interactions.

A detailed analysis based on the available structures as well as sequence conservation, revealed that mPRPF39 shows higher similarity both at the structural level as well as in amino acid conservation to yPrp42 than to yPrp39, indicating a misidentification of Prp42 homologs as Prp39 homologs in higher eukaryotes. It is likely that Prp42 and Prp39 have emerged from the same ancestral gene by gene duplication rendering all Prp39 homologs at the same time also Prp42 homologs. Incidentally, an additional patch of sequence conservation in the HAT-CTD can be observed on the concave surface of the HAT-NTD of mPrp39. This is the same area on yPrp42 that mediates protein-protein contacts connecting the core U1 snRNP to its auxiliary proteins. The fact that this analog area is conserved in mPRPF39 supports the hypothesis that mPRPF39 is also a yPrp42 homolog and can substitute it in binding to the U1 snRNP via U1C.

There are many proteins in the human spliceosome that do not have an obvious counterpart in yeast (Fabrizio 2009, Will and Luhrmann 2011). However, this is the opposite for PRPF39. In yeast the Prp39/Prp42 heterodimer is stably associated with the U1 snRNP. This is not the case in metazoan, with PRPF39 only transiently associated to the spliceosome. One of the main factors that allow for more stable binding of the heterodimer is the length of the U1 snRNA SL 2. The elongated SL that is absent in metazoa folds over and binds the charged surface of yPrp42. This interaction acts as a clamp that functions as an additional stabilizer to the binding of the heterodimer to the U1 snRNP core which is otherwise mediated over interactions with the concave side of the yPrp42 HAT-NTD and yU1C. Not only is the elongated SL 2 missing in the metazoan system, but also the positively charged groove at the equivalent position to yPrp42 is missing in mPRPF39. This only leaves the interaction with U1C as a connector between the metazoan auxiliary U1 snRNP proteins to its core, posing a possible explanation for why the homodimer is only associated with the U1 snRNP in a transient manner.

The more transient association of the homodimer to the spliceosome in the metazoan system could also serve as an explanation for the finding that the human pre-B complex structure shows no contacts between the U1 and U2 snRNP (Zhan 2018). Possibly the loosely associated homodimer can mediate inter-snRNP contacts during complex A formation. Due to its loose association it might allow for separation of the particles in the pre-B state, which may be important to encourage the more complex alternative splicing patterns observed in organisms with a Prp39 homodimer. Other splicing factors have also been reported to associate more transiently with the spliceosome in humans. For example, Prp38 and Snu23 are stable snRNP components in yeast but in human they enter the spliceosome individually after the tri-snRNP (Ulrich and Wahl 2017). SF3a and SF3b are another case of proteins that show weak binding to the U2 snRNP. Even though they are phylogenetically conserved they were initially not identified due to their weak association to the spliceosome (Kramer 1996). Both, this as well as the data presented here demonstrate a lower level of fixed pre-organization of metazoan spliceosomes versus yeast spliceosomes.

In vitro splicing experiments with monomeric mPRPF39 show, that disruption of the dimer has a detrimental effect on splicing efficiency. In addition, supplementing splicing reactions with wild type mPRPF39, and thus changing the levels of functional PRPF39, show a trend toward higher splicing efficiency. This trend is not significant, the p value of 0.09 being just over the threshold of 0.05. It is not surprising that the effect of adding mPRPF39 to the reaction does not have a robust effect, because there is still endogenous PRPF39 in the used splicing nuclear extracts. mPRPF39 positively influences splicing efficiency, while monomeric mPRPF39 has an inhibitory effect. This is a further piece of evidence, that the mPRPF39 homodimer is the functional subunit and can substitute the yPrp39/yPrp42 heterodimer observed in yeast in metazoans.

8.2 Activation Dependent and Tissue Specific Regulation of mPRPF39 by a Combination of NMD and Production of Non-functionally Truncated Protein Isoforms

After seeing direct effects on splicing efficiency *in vitro* depending on the levels of functional mPRPF39, it was especially interesting to find, that the mPRPF39 RNA is alternatively spliced in a differential manner. The inclusion of the alternative exon is highly regulated by the T-cell differentiation state and different levels of inclusion can also be observed in different murine tissues. The stop codon in the alternative exon is a classical

NMD target, making this exon a poison exon. Poison exons are a widespread mechanism used by organisms to control expression levels of proteins. For example SRSF3 regulates the inclusion of poison exons, strongly conserved among species (Lareau 2007), in itself and SRSF2, SRSF5 and SRSF7 (Anko 2012).

Our experiments indicate, that the *mprpf39* pre-mRNA is at best a weak NMD target. However, regardless of whether the inclusion isoform is degraded or translated, this isoform will lead to a reduced level of functional protein because the translated product would be missing the part of the HAT-CTD crucial for dimerization (**Figure 46**).

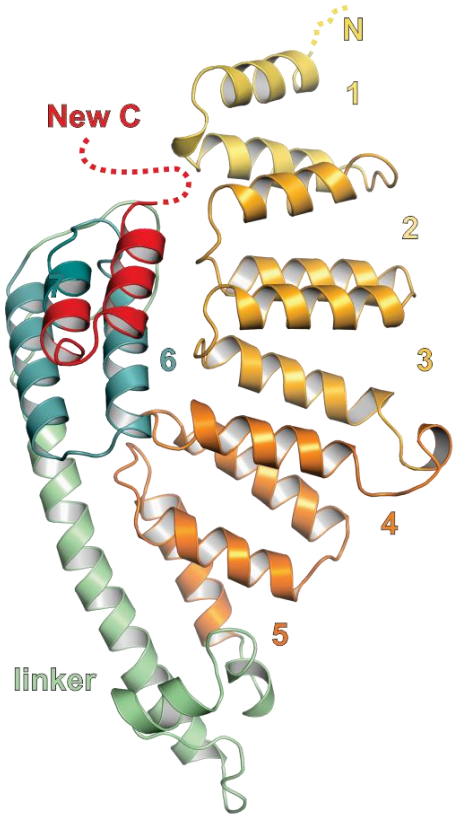

Figure 46 PHYRE Structure prediction (Kelley 2015) of short mPRPF39 fragment arising from alternative splicing. The protein is shown in cartoon representation with coloring analog to **Figure 25**. The short version of the mPRPF39 protein stops with an alternative amino acid sequence deriving from the alternative exon containing a stop codon that is included shown in red. The alternative sequence is also predicted to form α-helices.

The resulting protein would have an intact HAT-NTD with a severely shortened HAT-CTD. This would still allow it to bind to the spliceosome, but act to suppress splicing, because it does now allow dimer formation. Interestingly the new additional amino acid residues stemming from the segment of the alternative exon before the stop codon, are predicted to form HAT repeat similar α-helices. This supports the hypothesis of a truncated mPRPF39 isoform (**Figure 47**).

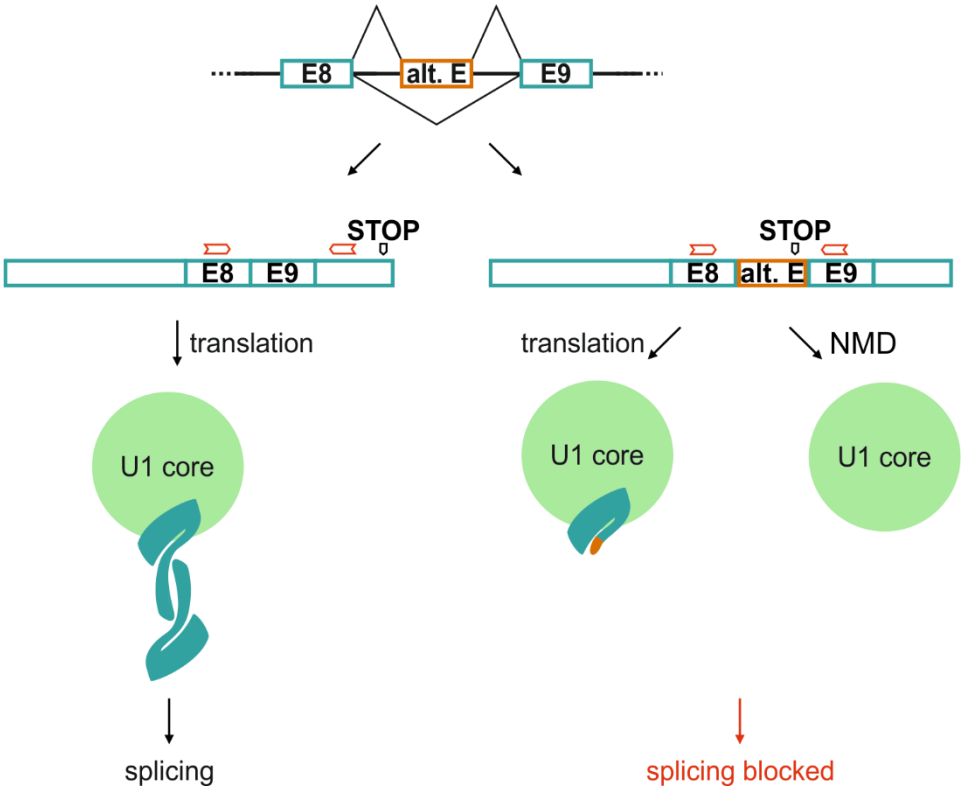

Figure 47 Non-productive splicing of PRPF39. When constitutive splicing occurs, PRPF39 is generated and the homodimer allows splicing. When the alternative exon is included the *prpf39* pre-mRNA is either degraded leading to depletion of PRPF39 or to a truncated PRPF39. Both scenarios would lead to blocking of splicing.

The fact that the function of the premature stop codon in the alternative exon is conserved in the human system cements the importance of this system to fine tune the Prpf39 levels in a tissue specific and differentiation dependent manner and suggests this as a mechanism across species. Together with its impact on splicing *in vitro*, this suggests a role of PRPF39 in adapting splicing efficiency to cell type specific requirements. The tightly controlled regulation of mPRPF39 protein levels is especially interesting, when seen in the context of dimerization. As already mentioned, one of the more basic mechanisms to regulate activity of dimeric proteins, is the concentration. This could explain, why maintaining exact levels of PRPF39 is crucial, because the dimerization could be dependent on the PRPF39 level rising above a certain threshold. Our data suggest a role of PRPF39 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 (Ni 2016). 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 naïve and memory T cells.

8.3 Higher Splicing Complexity is Evolutionarily Connected to Reduced Amounts of Spliceosomal Components

The investigation of the length of U1 snRNA in different organisms revealed that SL 2 and SL 3 are significantly shorter in metazoans compared to its yeast counterpart. This shortened U1 snRNA goes hand in hand with altered structural properties of mPRPF39. yPrp42 has a strongly positively charged groove on the convex side of the HAT-NTD that can accommodate the U1 snRNA. This positively charged groove is not detectable in the structure of mPRPF39. Strikingly, all of the organisms analyzed with a short U1 snRNA are also lacking a Prp42 homolog and consequently a homodimer of PRPF39 is the functional subunit in these organisms. Hence, I propose that a co-evolutionary mechanism based on the protein as well as on the RNA level has occurred.

The correlation between the length of the U1snRNA and the loss of Prp42 with splicing complexity show that this evolutionary development was important to allow for more complex splicing. This could be because a system with lower pre-organization is more susceptible to regulation. In yeast, the yPrp39/yPrp42 heterodimer acts as a binding platform for multiple alternative splicing factors (Li 2017, Bai 2018, Plaschka 2018). In metazoan the transient association could enable an additional level of regulation, which suggests, that homodimeric PRPF39 could also have a role in enabling and regulating more complex alternative splicing.

The analysis implies, that a short U1 snRNA SL 2 in a system without a Prp42 homolog acts to loosen the association of the dimer to the U1 snRNP core. This makes way for more complex regulation of association and disassociation of alternative splicing factors.

8.4 mPRPF39 substitutes for the Yeast Heterodimer and Allows for More Complex Splicing in Metazoans in Conjuncture with a Short U1 snRNA

Because the yPrp39/yPrp42 heterodimer seems to be so crucial in organizing the spliceosome during early splicing (Bai 2018, Plaschka 2018) it would be very surprising if there were no unit that could functionally substitute the heterodimer in the metazoan system.

PRPF39 appears to be a likely candidate for this. It is present in a homodimeric form that shows high similarity to the heterodimer observed in yeast and monomerization has a detrimental effect on splicing *in vitro*. The level of functional mPRPF39 homodimer is tightly

controlled in a differentiation and tissue specific manner, highlighting the importance that mPRPF39 levels, especially in a dimeric form, have on splice efficiency.

Furthermore, our analysis of the U1 snRNA length versus presence of a homodimer or a heterodimer, led us to an interesting conclusion: The reduced U1 snRNA length and the reduced protein content, achieved by replacing two different proteins with a single homodimeric protein, does not in fact reduce slicing complexity, but rather increases it (**Figure 48**).

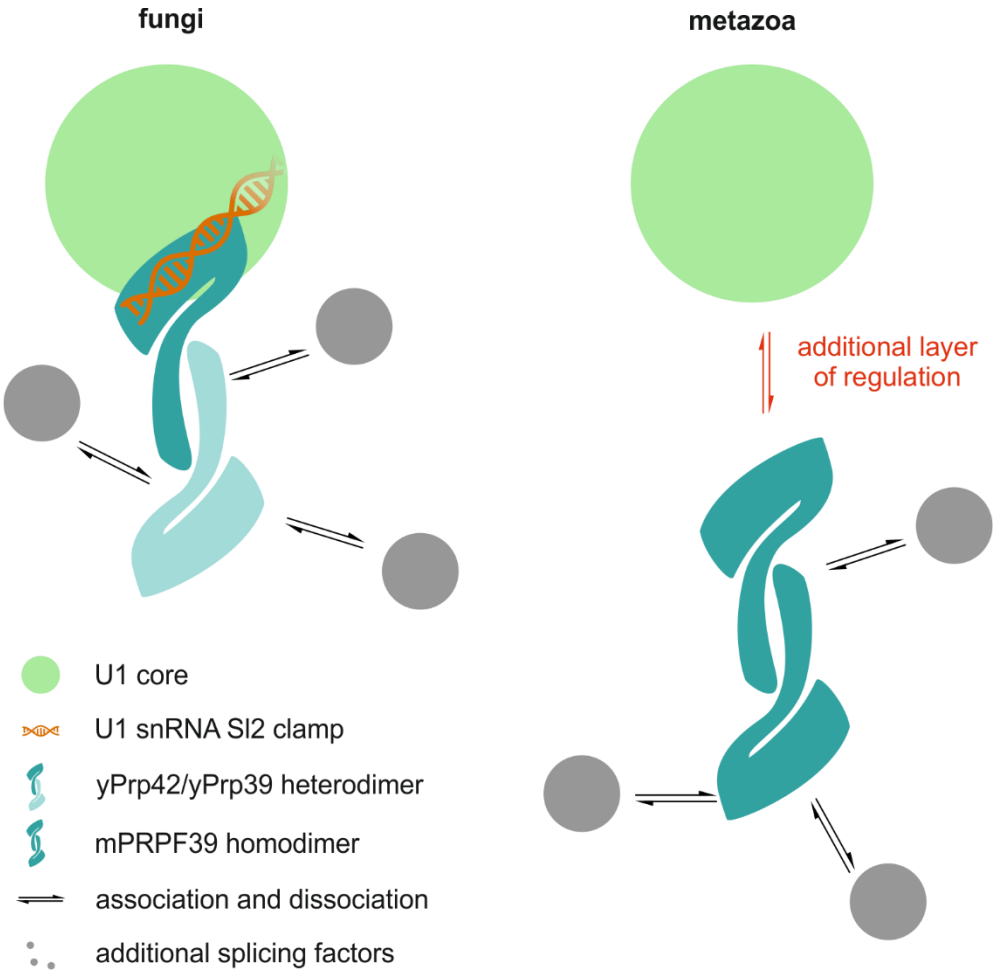

Figure 48 Increased versatility despite reduced molecular complexity. In this suggested model, the U1 snRNA SL 2 acts as a clamp to tightly bind the yPrp39/yPrp42 heterodimer to the U1 snRNP core. In metazoans, the missing extended SL 2 and the missing binding interface in PRPF39 lead to a more transient association of the dimer. This increases the possible layers of regulation. In yeast only the association and dissociation of additional splicing factors can be regulated, but in metazoans the whole homodimer interaction hub can be dissociated.

These findings are in contrast to current models, in which an increasing protein decoration of the eukaryotic spliceosome is correlated with increasing splicing complexity. This study is an example for the opposite. In this case, a reduced spliceosome does in fact not cause

less complex splicing, but might actually be beneficial for highly regulated, sophisticated splicing.

This work presents the first detailed structural and functional analysis of the essential splicing factor Prp39. It seems to be an important modulator of splicing, that is active in early spliceosome formation. It is a new building block that will be of great impact for future structure determination of spliceosomal assemblies by cryo-EM. It is also an interesting target for many more analyses of constitutive splicing and also alternative splicing. We supply an invaluable tool to further characterize PRPF39 and the impact of dimerization on splicing with our monomeric isoforms.

8.5 Future Perspectives

To allow for a broader analysis of the effect of PRPF39 on metazoan splicing, functional assays in a cellular environment would be beneficial. Studying protein functions in cells is greatly helped by loss of function experiments, in which a system is, for example, depleted of PRPF39. However, this is challenging for essential proteins. Completely knocking out PRPF39 would be lethal to the cell. When studying the role of an essential protein it is crucial to use a method that can conditionally deplete the system of this protein. Protein expression can be inhibited at the DNA, RNA or protein level. One method for reducing the amount of protein in a cell is a knock down with specific siRNAs. In our hands the use of six different siRNAs in every possible combination with varying cell seeding concentrations and harvesting after different time points after transfection never led to a substantial downregulation even on the RNA level. This indicates that a more sophisticated approach is necessary.

One of the main drawbacks of conditional technologies working at the level of DNA or RNA is, that the time necessary for depletion of a protein is dependent on the said half-life because they work pre-translationally. To overcome this issue, a knock down on protein level could be employed. In the case of PRPF39 initial experiments hint, that the protein is stable in the cell over a long period of time. This makes a conditional knock down on the protein level especially attractive.

One possible solution would be using a fusion protein with a degron. A degron is a domain which confers instability to a protein of interest which is then degraded by the 16S proteasome. Temperature, small molecules, light, or the expression of other additional proteins can be factors that activate or inhibit conditional degrons (Natsume and Kanemaki

2017). Some of the degron systems allow for exact regulation of the protein levels in the cell.

To use a degron based technology, genetic modification is necessary to attach the degron sequence to the protein of interest. CRISPR-Cas9 based genome editing has opened up many possibilities but insertion of large DNA sequences can still represent a challenging and time-consuming process. An alternative approach would be to create a cell line stably expressing in this case a PRPF39-degron fusion protein. In a second step the endogenous PRPF39 could be knocked out with the CRISPR-Cas9 system, which is a much simpler process that is easier to screen for.

If one had a cell system with an inducible PRPF39-degron fusion protein, a plethora of different assays could be used to analyze phenotypical differences between normal cells and knock down mutants. A very straightforward first experiment would be RNA sequencing of wild type cells or cells depleted of PRPF39. This should present us with information on the effect that PRPF39 has on overall splicing. Using cells depleted of PRPF39 would also allow us to analyze the effect of the monomeric PRPF39 without interference of endogenous wild type protein more readily.

Another interesting open question is, how the alternative splicing event observed in mPRPF39 is regulated. Because the inclusion is differentially regulated in tissue specific manner, the regulation must come from *trans*-acting factors. The degron system could help us find out if PRPF39 self-regulates its poison exon.

The experiments show that different levels of mPRPF39 have an effect on splicing *in vitro* and the protein level is controlled in a tissue and differentiation specific manner. Therefore, using a degron system could be used to tailor PRPF39 levels in cells which could then be harvested and analyzed.

The yeast early spliceosome structures show various alternative splicing factors associated with the yPrp39/yPrp42 heterodimer (Li 2017, Bai 2018, Plaschka 2018). It remains an open question, if these factors will also use the PRPF39 homodimer as an interaction hub in the metazoan system. Furthermore, it would be interesting to analyze whether the dimerization has an influence on binding of the proteins. The monomeric mutants are a powerful tool to use in determining the influence of dimerization on protein association. Mass spectrometry analyses using monomeric and dimeric PRPF39 could be used for this purpose.

In this thesis it is hypothesized, that a short U1 snRNA SL 2 in combination with a homodimeric PRPF39 allows for more complex splicing. An interesting approach to show this would be to use the CRISPR-Cas9 system to introduce a long U1 snRNA in a

metazoan system and a positive patch in metazoan PRPF39 HAT-NTD equivalent to the one seen in yPrp42. The new longer U1 snRNA would then hopefully clamp the otherwise more transiently associated homodimer to the U1 snRNP. If this hypothesis is correct, introducing these yeast specific traits into a metazoan spliceosome would lead to less complex splicing.

9 References

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., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. and Zwart, P. H. (2010). "PHENIX: a comprehensive Python-based system for macromolecular structure solution." Acta crystallographica. Section D, Biological crystallography **66**(Pt 2): 213-221.

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. and Terwilliger, T. C. (2002). "PHENIX: building new software for automated crystallographic structure determination." Acta Crystallogr D Biol Crystallogr **58**(Pt 11): 1948-1954.

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 crystallographica. Section D, Biological crystallography **68**(Pt 4): 352-367.

Agafonov, D. E., Deckert, J., Wolf, E., Odenwalder, P., Bessonov, S., Will, C. L., Urlaub, H. and Luhrmann, R. (2011). "Semiquantitative proteomic analysis of the human spliceosome via a novel two-dimensional gel electrophoresis method." Mol Cell Biol **31**(13): 2667-2682.

Anko, M. L., Muller-McNicoll, M., Brandl, H., Curk, T., Gorup, C., Henry, I., Ule, J. and Neugebauer, K. M. (2012). "The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes." Genome Biol **13**(3): R17.

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**(W1): W344-350.

Ast, G. (2004). "How did alternative splicing evolve?" Nat Rev Genet **5**(10): 773-782.

Bai, R., Wan, R., Yan, C., Lei, J. and Shi, Y. (2018). "Structures of the fully assembled Saccharomyces cerevisiae spliceosome before activation." Science **360**(6396): 1423-1429.

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**(6): 863-875.

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C. M., Wilson, M. D., Kim, P. M., Odom, D. T., Frey, B. J. and Blencowe, B. J. (2012). "The evolutionary landscape of alternative splicing in vertebrate species." Science **338**(6114): 1587-1593.

Barton, G. J. (1993). "ALSCRIPT: a tool to format multiple sequence alignments." Protein engineering **6**(1): 37-40.

Battye, T. G., Kontogiannis, L., Johnson, O., Powell, H. R. and Leslie, A. G. (2011). "iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM." Acta Crystallogr D Biol Crystallogr **67**(Pt 4): 271-281.

Berget, S. M. (1995). "Exon recognition in vertebrate splicing." J Biol Chem **270**(6): 2411-2414.

Berget, S. M., Moore, C. and Sharp, P. A. (1977). "Spliced segments at the 5' terminus of adenovirus 2 late mRNA." Proc Natl Acad Sci U S A **74**(8): 3171-3175.

Biamonti, G. and Caceres, J. F. (2009). "Cellular stress and RNA splicing." Trends Biochem Sci **34**(3): 146-153.

Black, D. L. (2003). "Mechanisms of alternative pre-messenger RNA splicing." Annu Rev Biochem **72**: 291-336.

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., Marceau, C., Janssen, H., Carette, J. E., Bennett, K. L., Colinge, J., Superti-Furga, G. and Brummelkamp, T. R. (2015). "Gene essentiality and synthetic lethality in haploid human cells." Science **350**(6264): 1092-1096.

Brow, D. A. (2002). "Allosteric cascade of spliceosome activation." Annu Rev Genet **36**: 333-360.

Buratti, E. and Baralle, D. (2010). "Novel roles of U1 snRNP in alternative splicing regulation." RNA Biol **7**(4): 412-419.

Busch, A. and Hertel, K. J. (2012). "Evolution of SR protein and hnRNP splicing regulatory factors." Wiley Interdiscip Rev RNA **3**(1): 1-12.

Calvin, K. and Li, H. (2008). "RNA-splicing endonuclease structure and function." Cell Mol Life Sci **65**(7-8): 1176-1185.

Carthew, R. W. and Sontheimer, E. J. (2009). "Origins and Mechanisms of miRNAs and siRNAs." Cell **136**(4): 642-655.

Chan, R. T., Robart, A. R., Rajashankar, K. R., Pyle, A. M. and Toor, N. (2012). "Crystal structure of a group II intron in the pre-catalytic state." Nat Struct Mol Biol **19**(5): 555-557.

Chen, M. and Manley, J. L. (2009). "Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches." Nat Rev Mol Cell Biol **10**(11): 741-754.

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 crystallographica. Section D, Biological crystallography **66**(Pt 1): 12-21.

Chen, X., Xu, Z., Zhang, L., Liu, H., Liu, X., Lou, M., Zhu, L., Huang, B., Yang, C. G., Zhu, W. and Shao, J. (2014). "The conserved Lys-95 charged residue cluster is critical for the homodimerization and enzyme activity of human ribonucleotide reductase small subunit M2." J Biol Chem **289**(2): 909-920.

Chow, L. T., Gelinas, R. E., Broker, T. R. and Roberts, R. J. (1977). "An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA." Cell **12**(1): 1-8.

Colgan, D. F. and Manley, J. L. (1997). "Mechanism and regulation of mRNA polyadenylation." Genes Dev **11**(21): 2755-2766.

Crick, F. H. (1958). "On protein synthesis." Symp Soc Exp Biol **12**: 138-163.

De Conti, L., Baralle, M. and Buratti, E. (2013). "Exon and intron definition in pre-mRNA splicing." Wiley Interdiscip Rev RNA **4**(1): 49-60.

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**(7): e0159690.

DeLano, W. (2002). The PyMOL Molecular Graphics System http://www.pymol.org .

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**(5): 1475-1489.

Emsley, P. and Cowtan, K. (2004). "Coot: model-building tools for molecular graphics." Acta Crystallogr D Biol Crystallogr **60**(Pt 12 Pt 1): 2126-2132.

Emsley, P., Lohkamp, B., Scott, W. G. and Cowtan, K. (2010). "Features and development of Coot." Acta crystallographica. Section D, Biological crystallography **66**(Pt 4): 486-501.

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**(4): 593-608.

Fairbanks, G., Steck, T. L. and Wallach, D. F. (1971). "Electrophoretic analysis of the major polypeptides of the human erythrocyte membrane." Biochemistry **10**(13): 2606-2617.

Fontana, A., de Laureto, P. P., Spolaore, B., Frare, E., Picotti, P. and Zambonin, M. (2004). "Probing protein structure by limited proteolysis." Acta Biochim Pol **51**(2): 299-321.

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**(18): 2425-2438.

Fox-Walsh, K. L., Dou, Y., Lam, B. J., Hung, S. P., Baldi, P. F. and Hertel, K. J. (2005). "The architecture of pre-mRNAs affects mechanisms of splice-site pairing." Proc Natl Acad Sci U S A **102**(45): 16176-16181.

Fu, X. D. and Ares, M., Jr. (2014). "Context-dependent control of alternative splicing by RNA-binding proteins." Nat Rev Genet **15**(10): 689-701.

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**(Database issue): D492-496.

Goodsell, D. S. and Olson, A. J. (2000). "Structural symmetry and protein function." Annu Rev Biophys Biomol Struct **29**: 105-153.

Gottschalk, A., Tang, J., Puig, O., Salgado, J., Neubauer, G., Colot, H. V., Mann, M., Seraphin, B., Rosbash, M., Luhrmann, R. and Fabrizio, P. (1998). "A comprehensive biochemical and genetic analysis of the yeast U1 snRNP reveals five novel proteins." RNA **4**(4): 374-393.

Gouy, M., Guindon, S. and Gascuel, O. (2010). "SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building." Mol Biol Evol **27**(2): 221-224.

Grazette, A. H., S.; Emsley, J.; Layfield, R.; Dreveny, I. (2016). "Crystal structure of HAT domain of human Squamous Cell Carcinoma Antigen Recognized By T Cells 3, SART3 (TIP110)." DOI: 10.2210/pdb5JPZ/pdb.

Hammani, K., Cook, W. B. and Barkan, A. (2012). "RNA binding and RNA remodeling activities of the half-a-tetratricopeptide (HAT) protein HCF107 underlie its effects on gene expression." Proc Natl Acad Sci U S A **109**(15): 5651-5656.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., Mero, P., Dirks, P., Sidhu, S., Roth, F. P., Rissland, O. S., Durocher, D., Angers, S. and Moffat, J. (2015). "High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities." Cell **163**(6): 1515-1526.

Hashimoto, K. and Panchenko, A. R. (2010). "Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states." Proc Natl Acad Sci U S A **107**(47): 20352-20357.

Haugen, P., Simon, D. M. and Bhattacharya, D. (2005). "The natural history of group I introns." Trends Genet **21**(2): 111-119.

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**(12): 1796-1806.

Heyd, F. and Lynch, K. W. (2010). "Phosphorylation-dependent regulation of PSF by GSK3 controls CD45 alternative splicing." Mol Cell **40**(1): 126-137.

Holm, L. and Rosenstrom, P. (2010). "Dali server: conservation mapping in 3D." Nucleic Acids Res **38**(Web Server issue): W545-549.

Huang, M., Zhou, B., Gong, J., Xing, L., Ma, X., Wang, F., Wu, W., Shen, H., Sun, C., Zhu, X., Yang, Y., Sun, Y., Liu, Y., Tang, T. S. and Guo, C. (2018). "RNA-splicing factor SART3 regulates translesion DNA synthesis." Nucleic Acids Res **46**(9): 4560-4574.

Irimia, M. and Roy, S. W. (2014). "Origin of spliceosomal introns and alternative splicing." Cold Spring Harb Perspect Biol **6**(6).

Ispolatov, I., Yuryev, A., Mazo, I. and Maslov, S. (2005). "Binding properties and evolution of homodimers in protein-protein interaction networks." Nucleic Acids Res **33**(11): 3629-3635.

Ivashchenko, A. T., Tauasarova, M. K. and Atambaeva Sh, A. (2009). "Exon-intron structure of genes of fungi genomes." Mol Biol (Mosk) **43**(1): 28-35.

Jones, S. and Thornton, J. M. (1995). "Protein-protein interactions: a review of protein dimer structures." Prog Biophys Mol Biol **63**(1): 31-65.

Kabsch, W. (2010). "XDS." Acta Crystallogr D Biol Crystallogr **66**(Pt 2): 125-132.

Kabsch, W. and Sander, C. (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." Biopolymers **22**(12): 2577-2637.

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**(D1): D335-D342.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. and Sternberg, M. J. (2015). "The Phyre2 web portal for protein modeling, prediction and analysis." Nat Protoc **10**(6): 845-858.

Keren, H., Lev-Maor, G. and Ast, G. (2010). "Alternative splicing and evolution: diversification, exon definition and function." Nat Rev Genet **11**(5): 345-355.

Kervestin, S. and Jacobson, A. (2012). "NMD: a multifaceted response to premature translational termination." Nat Rev Mol Cell Biol **13**(11): 700-712.

Kramer, A. (1996). "The structure and function of proteins involved in mammalian pre-mRNA splicing." Annu Rev Biochem **65**: 367-409.

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**(2): 851-855.

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**(4): 593-602.

Krissinel, E. and Henrick, K. (2007). "Inference of macromolecular assemblies from crystalline state." J Mol Biol **372**(3): 774-797.

Krol, A., Westhof, E., Bach, M., Luhrmann, R., Ebel, J. P. and Carbon, P. (1990). "Solution structure of human U1 snRNA. Derivation of a possible three-dimensional model." Nucleic Acids Res **18**(13): 3803-3811.

Laggerbauer, B., Achsel, T. and Luhrmann, R. (1998). "The human U5-200kD DEXH-box protein unwinds U4/U6 RNA duplices in vitro." Proc Natl Acad Sci U S A **95**(8): 4188-4192.

Lareau, L. F., Inada, M., Green, R. E., Wengrod, J. C. and Brenner, S. E. (2007). "Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements." Nature **446**(7138): 926-929.

Lee, E. H. and Allen, P. D. (2007). "Homo-dimerization of RyR1 C-terminus via charged residues in random coils or in an alpha-helix." Exp Mol Med **39**(5): 594-602.

Leppek, K., Das, R. and Barna, M. (2018). "Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them." Nat Rev Mol Cell Biol **19**(3): 158-174.

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**(W1): W242-245.

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**(1): 1035.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L. and Quackenbush, J. (2000). "Gene index analysis of the human genome estimates approximately 120,000 genes." Nat Genet **25**(2): 239-240.

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**(6): 3623-3633.

Lovci, M. T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T. Y., Stark, T. J., Gehman, L. T., Hoon, S., Massirer, K. B., Pratt, G. A., Black, D. L., Gray, J. W., Conboy, J. G. and Yeo, G. W. (2013). "Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges." Nat Struct Mol Biol **20**(12): 1434-1442.

Marianayagam, N. J., Sunde, M. and Matthews, J. M. (2004). "The power of two: protein dimerization in biology." Trends Biochem Sci **29**(11): 618-625.

Martin, A., Schneider, S. and Schwer, B. (2002). "Prp43 is an essential RNA-dependent ATPase required for release of lariat-intron from the spliceosome." J Biol Chem **277**(20): 17743-17750.

Martinez, N. M. and Lynch, K. W. (2013). "Control of alternative splicing in immune responses: many regulators, many predictions, much still to learn." Immunol Rev **253**(1): 216-236.

Mayr, C. (2017). "Regulation by 3'-Untranslated Regions." Annu Rev Genet **51**: 171-194.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. and Read, R. J. (2007). "Phaser crystallographic software." J Appl Crystallogr **40**(Pt 4): 658-674.

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**(Pt 7): 1220-1228.

McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S. D., Wickens, M. and Bentley, D. L. (1997). "The C-terminal domain of RNA polymerase II couples mRNA processing to transcription." Nature **385**(6614): 357-361.

McLean, M. R. and Rymond, B. C. (1998). "Yeast pre-mRNA splicing requires a pair of U1 snRNP-associated tetratricopeptide repeat proteins." Mol Cell Biol **18**(1): 353-360.

Mei, G., Di Venere, A., Rosato, N. and Finazzi-Agro, A. (2005). "The importance of being dimeric." FEBS J **272**(1): 16-27.

Miller, S., Lesk, A. M., Janin, J. and Chothia, C. (1987). "The accessible surface area and stability of oligomeric proteins." Nature **328**(6133): 834-836.

Mitrovich, Q. M. and Guthrie, C. (2007). "Evolution of small nuclear RNAs in S. cerevisiae, C. albicans, and other hemiascomycetous yeasts." RNA **13**(12): 2066-2080.

Natsume, T. and Kanemaki, M. T. (2017). "Conditional Degrons for Controlling Protein Expression at the Protein Level." Annu Rev Genet **51**: 83-102.

Ni, T., Yang, W., Han, M., Zhang, Y., Shen, T., Nie, H., Zhou, Z., Dai, Y., Yang, Y., Liu, P., Cui, K., Zeng, Z., Tian, Y., Zhou, B., Wei, G., Zhao, K., Peng, W. and Zhu, J. (2016). "Global intron retention mediated gene regulation during CD4+ T cell activation." Nucleic Acids Res **44**(14): 6817-6829.

Nilsen, T. W. and Graveley, B. R. (2010). "Expansion of the eukaryotic proteome by alternative splicing." Nature **463**(7280): 457-463.

Patel, S. B. and Bellini, M. (2008). "The assembly of a spliceosomal small nuclear ribonucleoprotein particle." Nucleic Acids Res **36**(20): 6482-6493.

Paulson, A. R. and Tong, L. (2012). "Crystal structure of the Rna14-Rna15 complex." RNA **18**(6): 1154-1162.

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

Pomeranz Krummel, D. A., Oubridge, C., Leung, A. K., Li, J. and Nagai, K. (2009). "Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution." Nature **458**(7237): 475-480.

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**(3): 433-446 e434.

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**(4): 651-662.

Query, C. C., Moore, M. J. and Sharp, P. A. (1994). "Branch nucleophile selection in pre-mRNA splicing: evidence for the bulged duplex model." Genes Dev **8**(5): 587-597.

Rader, S. D. and Guthrie, C. (2002). "A conserved Lsm-interaction motif in Prp24 required for efficient U4/U6 di-snRNP formation." RNA **8**(11): 1378-1392.

Raj, B. and Blencowe, B. J. (2015). "Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles." Neuron **87**(1): 14-27.

Reuter, J. S. and Mathews, D. H. (2010). "RNAstructure: software for RNA secondary structure prediction and analysis." BMC Bioinformatics **11**: 129.

Rowen, L., Young, J., Birditt, B., Kaur, A., Madan, A., Philipps, D. L., Qin, S., Minx, P., Wilson, R. K., Hood, L. and Graveley, B. R. (2002). "Analysis of the human neurexin genes: alternative splicing and the generation of protein diversity." Genomics **79**(4): 587-597.

Schwer, B. and Guthrie, C. (1992). "A conformational rearrangement in the spliceosome is dependent on PRP16 and ATP hydrolysis." EMBO J **11**(13): 5033-5039.

Sharma, A. and Lou, H. (2011). "Depolarization-mediated regulation of alternative splicing." Front Neurosci **5**: 141.

Sharma, S., Kohlstaedt, L. A., Damianov, A., Rio, D. C. and Black, D. L. (2008). "Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome." Nat Struct Mol Biol **15**(2): 183-191.

Shin, C. and Manley, J. L. (2004). "Cell signalling and the control of pre-mRNA splicing." Nat Rev Mol Cell Biol **5**(9): 727-738.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D. and Higgins, D. G. (2011). "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega." Mol Syst Biol **7**: 539.

Spingola, M., Grate, L., Haussler, D. and Ares, M., Jr. (1999). "Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae." RNA **5**(2): 221-234.

Staley, J. P. and Guthrie, C. (1999). "An RNA switch at the 5' splice site requires ATP and the DEAD box protein Prp28p." Mol Cell **3**(1): 55-64.

Stark, M. R., Dunn, E. A., Dunn, W. S., Grisdale, C. J., Daniele, A. R., Halstead, M. R., Fast, N. M. and Rader, S. D. (2015). "Dramatically reduced spliceosome in Cyanidioschyzon merolae." Proc Natl Acad Sci U S A **112**(11): E1191-1200.

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**(2): 553-559.

Studier, F. W. (2005). "Protein production by auto-induction in high density shaking cultures." Protein Expr Purif **41**(1): 207-234.

Swapna, L. S., Srikeerthana, K. and Srinivasan, N. (2012). "Extent of structural asymmetry in homodimeric proteins: prevalence and relevance." PLoS One **7**(5): e36688.

Temin, H. M. and Mizutani, S. (1970). "RNA-dependent DNA polymerase in virions of Rous sarcoma virus." Nature **226**(5252): 1211-1213.

Terwilliger, T. C. (2000). "Maximum-likelihood density modification." Acta Crystallogr D Biol Crystallogr **56**(Pt 8): 965-972.

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**(Pt 6): 582-601.

The UniProt, C. (2017). "UniProt: the universal protein knowledgebase." Nucleic Acids Res **45**(D1): D158-D169.

Toor, N., Keating, K. S., Taylor, S. D. and Pyle, A. M. (2008). "Crystal structure of a self-spliced group II intron." Science **320**(5872): 77-82.

Ule, J., Ule, A., Spencer, J., Williams, A., Hu, J. S., Cline, M., Wang, H., Clark, T., Fraser, C., Ruggiu, M., Zeeberg, B. R., Kane, D., Weinstein, J. N., Blume, J. and Darnell, R. B. (2005). "Nova regulates brain-specific splicing to shape the synapse." Nat Genet **37**(8): 844-852.

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**(1): 91.

Valdar, W. S. and Thornton, J. M. (2001). "Conservation helps to identify biologically relevant crystal contacts." J Mol Biol **313**(2): 399-416.

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**(1): 105-124.

Voinnet, O. (2009). "Origin, biogenesis, and activity of plant microRNAs." Cell **136**(4): 669-687.

Wahl, M. C., Will, C. L. and Luhrmann, R. (2009). "The spliceosome: design principles of a dynamic RNP machine." Cell **136**(4): 701-718.

Wang, Q., Hobbs, K., Lynn, B. and Rymond, B. C. (2003). "The Clf1p splicing factor promotes spliceosome assembly through N-terminal tetratricopeptide repeat contacts." J Biol Chem **278**(10): 7875-7883.

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**(6264): 1096-1101.

Warzecha, C. C., Jiang, P., Amirikian, K., Dittmar, K. A., Lu, H., Shen, S., Guo, W., Xing, Y. and Carstens, R. P. (2010). "An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition." EMBO J **29**(19): 3286-3300.

Weber, G., Trowitzsch, S., Kastner, B., Luhrmann, R. and Wahl, M. C. (2010). "Functional organization of the Sm core in the crystal structure of human U1 snRNP." EMBO J **29**(24): 4172-4184.

Will, C. L. and Luhrmann, R. (2011). "Spliceosome structure and function." Cold Spring Harb Perspect Biol **3**(7).

Winn, M. D., Isupov, M. N. and Murshudov, G. N. (2001). "Use of TLS parameters to model anisotropic displacements in macromolecular refinement." Acta crystallographica. Section D, Biological crystallography **57**(Pt 1): 122-133.

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**(Pt 10): 1833-1839.

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y. A., Murray, R. R., Spirohn, K., Begg, B. E., Duran-Frigola, M., MacWilliams, A., Pevzner, S. J., Zhong, Q., Trigg, S. A., Tam, S., Ghamsari, L., Sahni, N., Yi, S., Rodriguez, M. D., Balcha, D., Tan, G., Costanzo, M., Andrews, B., Boone, C., Zhou, X. J., Salehi-Ashtiani, K., Charloteaux, B., Chen, A. A., Calderwood, M. A., Aloy, P., Roth, F. P., Hill, D. E., Iakoucheva, L. M., Xia, Y. and Vidal, M. (2016). "Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing." Cell **164**(4): 805-817.

Yura, K., Shionyu, M., Hagino, K., Hijikata, A., Hirashima, Y., Nakahara, T., Eguchi, T., Shinoda, K., Yamaguchi, A., Takahashi, K., Itoh, T., Imanishi, T., Gojobori, T. and Go, M. (2006). "Alternative splicing in human transcriptome: functional and structural influence on proteins." Gene **380**(2): 63-71.

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**(12): 1129-1140.

10 Abbreviations

APS	ammonium persulfate
Amp	ampicillin
ASU	asymmetric unit
ATP	adenosine triphosphate
BP	branch point
BSA	bovine serum albumin
CTD	C-terminal domain
Cm	chloramphenicol
DNA	deoxyribonucleic acid
dNTP	desoxyribonucleotide
DTT	dithiothreitol
EDTA	ethylene-diamine-tetra-acetic acid
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
FBS	fetal bovine serum
GSH	reduced glutathione
GST	glutathione S-transferase
HAT	half-a-tetratricopeptide
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
hnRNP	heterogenous nuclear ribonucleoprotein particle
IP	immuno precipitation
IPTG	isopropyl β-D-1-thiogalactopyranoside
ISE	intronic splicing enhancer
ISS	intronic splicing silencer
Kan	kanamycine
Kb	kilobase (unit of NA molecule length)

kDa	kilodalton (unit of molecular weight)
LB	lysogeny broth
MES	2-(N-morpholino)ethanesulfonic acid
mPRPF39	murine PRPF39
mRNA	messenger RNA
MW	molecular weight
NCS	non-crystallographic symmetry
NMD	nonsense mediated decay
nt	nucleotide
NTD	N-terminal domain
OD	optical density
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PPT	polypyrimidine tract
pre-mRNA	precursor mRNA
PSI	percent spliced in
RIPA	radioimmunoprecipitation assay
RMSD	root mean square deviation
RNA	ribonucleic acid
RT	reverse transcriptase
SAD	single wavelength anomalous dispersion
SDS	sodium dodecylsulfate
SDS-PAGE	SDS-polyacrylamide gel electrophoresis
SEC	size exclusion chromatography
SeMet	selenomethionine
SL	stem loop
snRNA	small nuclear ribonucleic acid
snRNP	small nuclear ribonucleoprotein complex

SR	serine/arginine-rich
SS	splice site
TAE	TRIS-Acetat-EDTA
TBE	TRIS-borate-EDTA
TBST	tris-buffered saline with tween20
TEMED	N,N,N',N'-tetraethylenediamide
TLS	Translation/Libration/Screw
T_m	melting temperature
TRIS	tris-(hydroxymethyl)-aminomethane
tRNA	transfer RNA
UTR	untranslated region
UV	ultra-violet

11 List of Figures

12 List of Tables

13 Acknowledgements

First of all, I would like to thank Florian Heyd for giving me the opportunity to work on this enthralling topic and for his constant support also in difficult times. I also would like to thank Markus Wahl who not only consented to evaluate this thesis as the 2nd referee but together with Florian guided me through my whole PhD with fruitful advice and discussions.

Of course, I would like to thank the current and former members of the Heyd and Wahl and Chakrabarti lab for creating a wonderful working atmosphere. I found my family away from home here: my lab mother, supportive big sisters and brothers to look up to and little siblings to help along in first their baby steps, and of course, the occasional odd uncle and aunt to help me with a particularly confounding problem.

Special thanks go to Bernhard Loll, for helping me solve my very first crystal structure and holding my hand throughout the whole process of bringing the story around the structure to life and publishing it. His delight in structural work was always infectious. I thank the teams of the BESSY II and PETRA III for technical support and access to the beamlines.

Alexander Neumann was invaluable in performing the computational parts of my project and I thank him for his patience in answering every question. Many thanks go to Ana Kotte for helping me with the practical work and making sure the project never had to be stalled.

Furthermore, thanks go to Jan Wollenhaupt for support in SEC-MALS measurements, and Christoph Weise for mass spectrometric analyses.

I would like to thank Claudia Alings for help and advice on crystallography and also life, Karin Hesse for solving a multitude of everyday problems, and Sutapa Chakrabarti for discussion and help with my project. You were all strong pillars for me to lean on.

Many thanks go to Marco Preussner and Eva Absmeier for welcoming me in the labs during my masters thesis. I could always come to both of you with any question.

Karine Santos, Eva Absmeier, Olga Herdt and Ronja Driller, you always were there for me with practical tips to help in lab work and have become wonderful friends.

I would like to thank my family for their valuable support. Specially to my parents, whose determination and great effort to give my siblings and me a good education were fundamental to set the basis that allowed me to be here.

Last, but certainly not least, I would like to thank my beloved husband, Marius. He has been at my side since I began this PhD giving me support, love and unwavering faith in me. There are no words to convey how much I love him.

14 Curriculum Vitae

For reasons of data protection, the curriculum vitae is not published in the electronic version.

For reasons of data protection, the curriculum vitae is not published in the electronic version.

For reasons of data protection, the curriculum vitae is not published in the electronic version.

For reasons of data protection, the curriculum vitae is not published in the electronic version.

15 Statutory Declaration

I, Francesca Danielle De Bortoli, declare that I have developed and written the enclosed dissertation entitled

“Evolution, Function and Structure of the Splicing Factor PRPF39”

by myself. I have not used sources or means without declaration in the text. Any thought from others or literal quotations are clearly marked. The thesis was not used in the same or in a similar version to achieve an academic grading.

Location, Date

Signature