The University Of Sheffield.

Biochemical characterisation of the nuclear protein hnRNPUL1.

By:

Carmen-Valentina Apostol

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

School of Biosciences

University of Sheffield

Sheffield, United Kingdom

January 2023

Acknowledgements

I would firstly like to thank my supervisors, Prof Stuart Wilson and Dr Dan Bose, for their continued support and guidance throughout this project. I am indebted to Dr Emma Thomson and Dr Ian Sudbery for their advice and insightful conversations. Special thanks to Dr Phil Mitchell for his astute comments and observations and for mentoring me in many biochemistry techniques. I am grateful for having met all the members of the Wilson group, past and present, Vicky, Ivo, Josh, Ang, Jack, Mikayla, Chiara, Charlotte, Llywelyn and Justin, who made the lab a friendly and welcoming environment and helped my sanity during the writing period with lunch breaks, fun conversations, picnics and dog walks. My friends outside of work, Agata and Tymon, have been by my side for many years, helping me through the tough times and sharing many fun trips and memories, for which I am forever grateful. Special thanks to Ivo for all the help and support he has given me at work and outside of it, for all his technical expertise when writing this thesis and for putting up with me for four years and during a pandemic. Le mulțumesc de asemenea părinților si familiei mele pentru sprijinul acordat de-a lungul anilor și pentru susținerea financiară ca să îmi pot urma visul de a studia biologia la cel mai înalt nivel.

Abstract

The nuclear RNA-binding protein hnRNPUL1 has been implicated in essential cellular processes, including gene expression and DNA damage repair. Numerous hnRNPUL1 mutations have been identified in amyotrophic lateral sclerosis (ALS) patients, but the molecular basis for toxicity remains unclear. Similarly, the reported role of hnRNPUL1 in gene expression lacks a mechanistic description, which is paramount for understanding the normal function and disease progression. Structurally, it is distinguished by a central folded domain predicted to serve as an ATP-binding module. Additionally, the C-terminal region (CTD) of hnRNPUL1 encompasses an intrinsically disordered RNA-binding RGG box and a prion-like domain, often found in proteins that undergo liquid-liquid phase separation.

Here, we will characterise the function of the central ATP-binding domain as an inactive polynucleotide kinase (PPNK), based on *in vitro* biochemistry assays informed by structural modelling and homology with mammalian and viral polynucleotide kinases. We present evidence that this domain of hnRNPUL1 recognises nucleoside triphosphates and free nucleic acid 5′ ends and that a single point mutation in the binding pocket efficiently restores kinase activity. We show that the hnRNPUL1 CTD is responsible for interactions with components of the RNA polymerase II transcription machinery and that a functional interplay between the PPNK and CTD modulates protein function and the mode of chromatin association. A mechanism of transcription regulation of snRNA gene targets previously identified in the lab is proposed based on the biochemical evidence. Lastly, we demonstrate phase separation behaviour of the hnRNPUL1 CTD under physiological conditions. This is regulated by post-translational modifications and we propose a mode of interaction with the transcription machinery mediated by protein condensation.

Contents

List of Figures

Abbreviations

AAA+	ATPases Associated with diverse cellular Activities
AID	auxin inducible degron
ALS	amyotrophic lateral sclerosis
AMP-PNP	adenylyl-imidodiphosphate
ATP	adenosine triphosphate
Aux	auxin
CBC	cap binding complex
CE	capping enzyme
ChIP	chromatin immunoprecipitation
CoIP	co-immunoprecipitation
CPSF	cleavage and polyadenylation specificity factor
CstF	cleavage stimulating factor
CTD	C-terminal domain
DBD	DNA-binding domain
DNA	deoxyribonucleic acid
Dox	doxycycline
DSE	distal sequence element
eCLIP	enhanced cross-linking and immunoprecipitation
GMP-PNP	guanylyl-imidodiphosphate
GSH	reduced glutathione
GST	glutathione S-transferase
GTP	guanosine triphosphate
hnRNP	heterogeneous nuclear ribonucleoprotein
IDR	intrinsically disordered region
IP	immunoprecipitation
LLPS	liquid-liquid phase separation
m^7G	7-methylguanosine
mNET-seq	mammalian native elongating transcript sequencing
mRNA	messenger RNA
NTP	nucleoside triphosphate
P-loop	phosphate loop
PCR	polymerase chain reaction
PNK(P)	polynucleotide kinase (phosphatase)
PolII	RNA polymerase II
Poly(A)	polyadenosine
PPNK	pseudo-polynucleotide kinase
pre-mRNA	precursor mRNA
PLD	prion-like domain
PLAAC	prion-like amino acid composition
PSE	proximal sequence element
PTM	post translational modification
RBD	RNA binding domain
RBP	RNA binding protein
RNA	ribonucleic acid
RMSD	root mean square deviation
RIP	RNA immunoprecipitation
rRNA	ribosomal RNA

SAF-A/B	scaffold-attachment factor A/B
SAP	SAF-A/B, Acinus and PIAS
scaRNA	small cajal body-specific RNA
Ser2P	phosphorylated Serine-2
Ser5P	phosphorylated Serine-5
Ser7P	phosphorylated Serine-7
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein
SPRY	splA kinase and ryanodine receptor
TF	transcription factor
TLC	thin-layer chromatography
TSS	transcription start site
WA	Walker A motif
WB	Walker B motif

1 Chapter 1: Introduction

1.1 The RNA Polymerase II transcription cycle

In eukaryotic cells, expression of all protein-coding genes, as well as numerous other classes of non-coding RNAs, is mediated by RNA polymerase II (PolII). The transcription machinery comprises a multitude of protein complexes, which associate with PolII at defined stages during the transcription cycle in a regulated manner to effect processing of the nascent RNA. Rpb1 is the largest subunit of mammalian PolII, which is distinguished by a disordered C-terminal domain (CTD) not present in RNA polymerases I and III. The PolII CTD is crucial for cell viability and consists of 52 heptapeptide repeats with the consensus sequence $Tyr^1Ser^2Pro^3Thr^4Ser^5Pro^6Ser^7$. Ser, Tyr and Thr residues can undergo phosphorylation or glycosylation, while prolines are subject to isomerisation (Harlen & Churchman 2017). The repertoire of post-translational modifications (PTMs) decorating the CTD, known as the CTD code, changes during the transcription cycle and serves as a selective binding platform for co-transcriptional recruitment of RNA processing complexes, including the capping, splicing and cleavage and polyadenylation machineries (Harlen & Churchman 2017).

Hypophosphorylated PolII localises around gene promoters, along with the general transcription factors TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH in a poised pre-initiation complex (PIC). Cyclin-dependent kinases CDK7, CDK9, CDK12 and CDK13 elicit CTD phosphorylation, which is counterbalanced by the activity of phosphatases such as RPAP2 (Egloff et al. 2012). The TFIIH subunit CDK7 promotes phosphorylation at positions Ser5 and Ser7, which accompany PolII exit from the PIC and transcription initiation (Egloff et al. 2007). ChIP-seq analysis confirmed that phosphorylation of Ser5 and Ser7 (Ser5P and Ser7P) peaks around the 5′ end of genes, coinciding with the distribution of the pre-mRNA capping machinery (Guo et al. 2019, Mayer et al. 2010). After synthesis of the first 50 nucleotides, PolII undergoes promoter-proximal pausing through the combined inhibitory effects of DSIF (DRB-sensitivity inducing factor) and NELF (negative elongation factor) complexes. PolII stalling is alleviated by CDK9 phosphorylation of Ser2, mediated by the CDK9

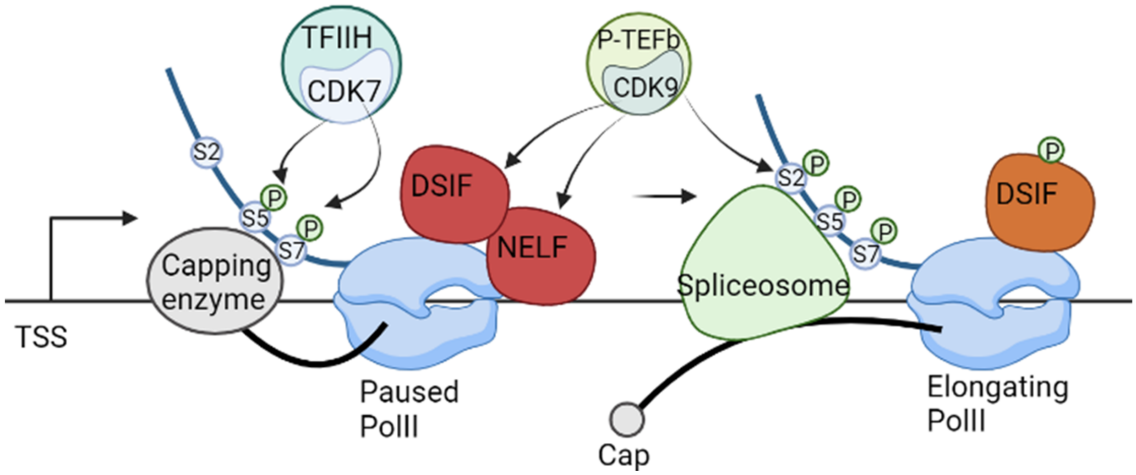

Figure 1.1: The CTD code regulates pre-mRNA processing.
Phosphorylation at Ser5 and Ser7 positions by general transcription factor TFIIH releases PolII from the pre-initiation complex. Ser5P and Ser7P recruit the capping enzyme for co-transcriptional processing of the nascent transcript. After the first 50 nucleotides, PolII is paused by the combined effects of DSIF and NELF. P-TEFb-mediated phosphorylation of Ser2, DSIF and NELF allows transcription to resume. Additionally, Ser5P and Ser2P promote pre-mRNA splicing by co-transcriptional recruitment of spliceosomal subunits. TSS= transcription start site.

subunit of the PTEF-b complex (positive transcription elongation factor b). Additionally, PTEF-b phosphorylates NELF, causing its displacement, and the DSIF subunit Spt5, turning it into a positive transcription elongation factor (Figure 1.1) (Kim & Sharp 2001). Ser2 phosphorylation (Ser2P) persists throughout the gene body (Mayer et al. 2010), mediating pre-mRNA splicing via U2 snRNP recruitment, as well as 3′ end cleavage and polyadenylation through the Pcf11 factor (Gu et al. 2013, Lunde et al. 2010). Mammalian Native Elongating Transcript sequencing (mNET-seq) and mass spectrometry have also implicated Ser5P in promoting stable association with active spliceosome complexes, with a more pronounced role towards the 5′ end (Nojima et al. 2018). From a functional perspective, Guo et al. (2019) observed that CTD phosphorylation promoted eviction of PolII from transcription initiation complexes and incorporation into assemblies formed by splicing factors by enhancing the affinity of CTD phospho-isoforms for protein condensates such as nuclear speckles. Recently, CDK12 has also been recognised as a Ser2 kinase with a distribution along gene bodies and 3′ ends distinct from P-TEFb (Bartkowiak et al. 2010)

PolII elongation rate may control the amount of CTD modifications and promote the 5′-3′ directionality of Ser2P accumulation, according to the reported abnormal Ser2P deposition at promoter-proximal sites following slow PolII progression (Fong et al. 2017). Transcription termination of protein-coding genes correlates with PolII hyperphosphorylation and ensues after synthesis of the polyadenylation signal (pAs). CPSF and CstF complexes recognise conserved sequences around the pAs and promote cleavage and polyadenylation of the pre-mRNA via CPSF-73 and PAP, respectively. Next, the cleavage factor I and II (CFI/II) complexes mediate transcription termination by reducing PolII processivity. This event is coupled with a particular chromatin landscape within an apparent "histone valley", which may promote 3′ PolII pausing ahead of more histone-dense regions (Glover-Cutter et al. 2007). Notably, the CPSF complex has been reported to co-localise with PolII at some gene 5′ regions, suggesting that the 3′ processing fate may be determined at the promoter (Glover-Cutter et al. 2007).

1.1.1 mRNA 5′ capping

In eukaryotes, addition of a 5′ cap onto the nascent transcript is the first co-transcriptional processing step and it is crucial for protecting the pre-mRNA from 5′-3′ exonucleolytic degradation (Figure 1.2A). Moreover, the 5′ cap and the associated cap-binding complex (CBC) have also been linked to promoting pre-mRNA splicing and polyadenylation, as well as cytoplasmic export and translation of the mature mRNA. Capping is a 3-step process that involves hydrolysis of the γ-phosphate from the triphosphorylated 5′ nucleotide by an RNA triphosphatase, followed by 5′-5′ ligation of a guanine monophosphate to the diphosphate group generated in the previous step by an RNA guanylyltransferase. The process is completed by methylation of the inverted G nucleotide at position N7 by RNA guanine-N7 methyltransferase (Figure 1.2B). Some small noncoding RNAs can acquire additional cap methyl marks, as will be discussed later.

In humans, the 5′ phosphatase and guanylyltransferase activities of the capping enzyme (CE) are encoded in one bifunctional polypeptide chain, while yeast possess distinct proteins for each function. In both cases however, recruitment of the CE by Ser5P PolII CTD has been reported very early after transcription initiation (Cho et al. 1997). The RNA guanine-N7 methyltransferase is recruited to the pre-mRNA 5′ end separately after dissociation of the CE from PolII in both humans and yeast. Functional coupling of PolII and CE stimulates the rate of mRNA capping by 2-4 orders of magnitude compared to post-transcriptional capping (Moteki & Price 2002). Consistently, chromatin immunoprecipitation (ChIP) analysis in yeast observed a sharp peak of the CE immediately downstream of the transcription start site (TSS), which closely matched the Ser7P and Ser5P PolII CTD distribution (Mayer et al. 2010). In agreement with these results, Martinez-Rucobo et al. (2015) reported a cryo-EM structure of an *in vitro* assembled complex between transcribing yeast PolII and the CE, placing the RNA triphosphatase active site in close proximity to the PolII exit channel. The structure illustrates how seamless capping of the pre-mRNA can be achieved upon emerging from PolII, thus offering continuous protection from exonucleolytic degradation. Following cap addition, the CE rapidly dissociates from PolII (Martinez-Rucobo et al. 2015).

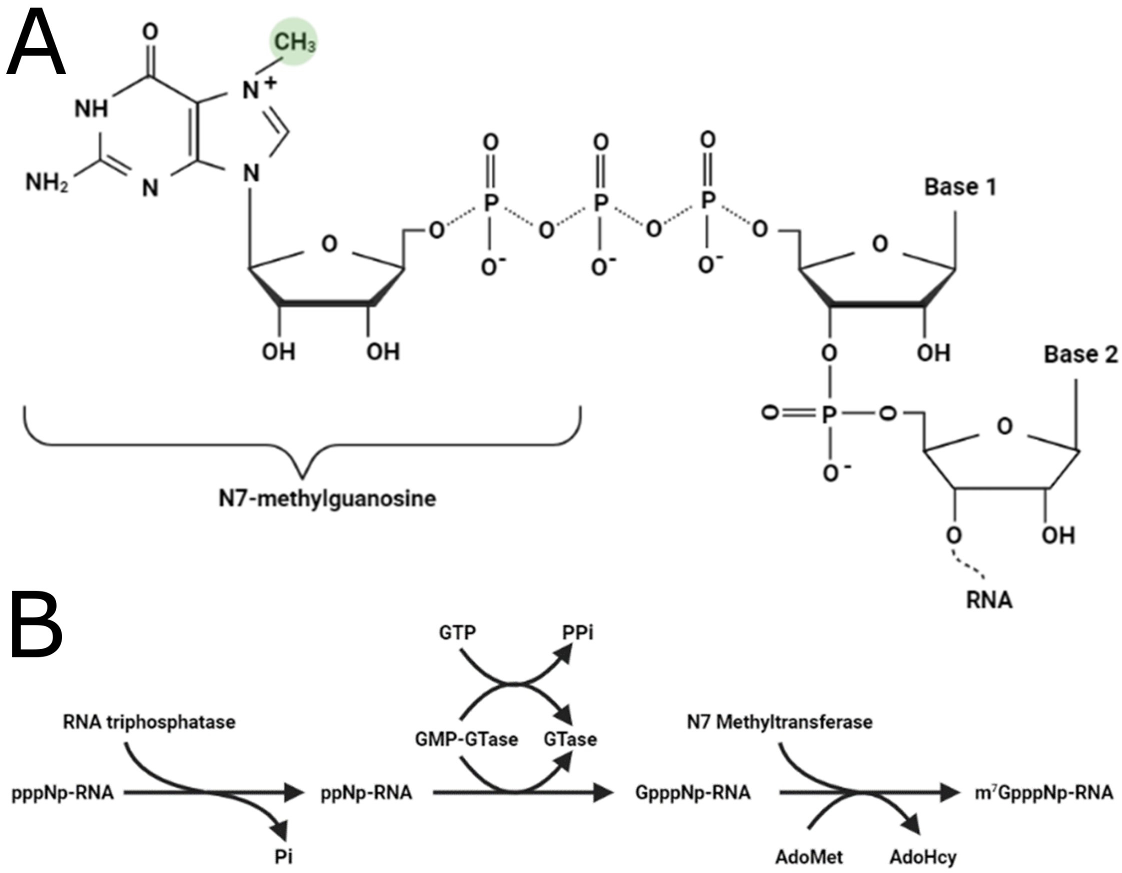

Figure 1.2: mRNA 5′ capping steps.
A) The chemical structure of the inverted 5′ N7-methylated guanosine cap. **B)**. Addition of the m^{7}G cap is the result of three sequential steps. First, RNA triphosphatase removes the γ-phosphate from the first nucleotide, creating a 5′ diphosphorylated product. Next, guanylyltransferase (GTase) reacts with the α-phosphate of a bound GTP, generating a transient GMP-enzyme complex (GMP-GTase). The GMP is then transferred to the 5′ diphosphorylated nucleotide. Finally, N7 methyltransferase adds a methyl group from S-adenosyl-L-methionine (AdoMet) onto the inverted G nucleotide, creating the m^{7}G cap structure and S-adenosyl-L-homocysteine (AdoHcy).

After subsequent cap methylation, the m^7G structure is recognised by the CBC, which consists of CBP20 and CBP80 subunits (Izaurralde et al. 1994, Ohno et al. 1990). While cap recognition is mediated by CBP20, interaction with CBP80 is necessary to induce a favourable conformation for stable binding to the cap and first base (Izaurralde et al. 1994, Mazza et al. 2001). The CBC has a high affinity for the m^7GpppG structure, but not unmethylated (GpppG) or hypermethylated ($m^{2,2,7}GpppG$) sequences (Izaurralde et al. 1992, Ohno et al. 1990). CBC binding to the 5′ cap has important implications for downstream processing, as it was shown to promote splicing through recognition of the first intron (Izaurralde et al. 1994). Additionally, CBC interaction with ARS2 stimulates pre-mRNA cleavage and 3′-end processing, especially for early polyadenylation sites, in a manner dependent on cleavage and polyadenylation factors CLP1 and PCF11 (Hallais et al. 2013). Lastly, CBC has been implicated in the nuclear export of both PolII-transcribed coding and some non-coding RNAs, by providing a binding platform for nuclear export factors such as Alyref and PHAX (Izaurralde et al. 1992, 1995, Ohno et al. 2000, Viphakone et al. 2019).

Once the mRNA has been exported to the cytoplasm, the CBC is exchanged for the translation initiation factor eIF4E on the 5′ m^7G cap, which guides ribosome assembly in preparation for translation. Interestingly, a fraction of eIF4E resides in the nucleus, where it promotes export of a subset of transcripts containing an eIF4E sensitivity element (4ESE), as well as export and subsequent translation of the pre-mRNA capping machinery (Culjkovic-Kraljacic et al. 2020). The study also reported an important proportion of both coding and non-coding RNAs existing in an uncapped form at steady-state, with some transcript pools being up to 80% uncapped. In support of the idea that decapping is not irreversible and does not necessarily lead to degradation, Otsuka et al. (2009) characterised an exclusively cytoplasmic capping enzyme which could repair 5′-monophosphorylated RNAs by serial phosphorylation and GMP ligation. The complex appeared indispensable for cell recovery after oxidative stress, leading the authors to suggest that the cytoplasmic CE may be involved in conditional silencing and reactivation of de-capped mRNAs stored in P-bodies during times of stress. Lastly, a cap quality-control mechanism has been described, which involves the mammalian DXO enzyme or its

yeast homologues Rai1 and Dxo1 (Jiao et al. 2010, 2013). Despite *in vitro* hydrolysis of m^7G and $m^{2,2,7}G$ caps, DXO preferentially acts on unmethylated mRNA caps *in vivo* and its knockdown promotes accumulation of unspliced and improperly polyadenylated transcripts, connecting cap integrity to downstream pre-mRNA processing (Jiao et al. 2013).

1.2 Termination of RNA Polymerase II-dependent transcription

PolII transcription termination of several classes of genes, including protein-coding ones, relies on recognition of the polyadenylation signal (pAs) AAUAAA on the pre-mRNA by the cleavage and polyadenylation specificity factor (CPSF) complex. This initial event promotes endonucleolytic cleavage via the CPSF-73 subunit, aided by the cleavage stimulation factor (CstF) complex binding to a U or GU-rich element (Mandel et al. 2006). In addition to CPSF and CstF complexes, 3′ processing requires the activities of cleavage-polyadenylation factors I and II (CFI/ CFII) and polyA polymerase (PAP) (Shi et al. 2009). CFII is composed of Pcf11 and Clp1 subunits, with Pcf11 being reported to bridge the 3′ cleaved RNA to PolII CTD and reduce its processivity (Zhang & Gilmour 2006, Zhang et al. 2007). PAP catalyses the addition of a protective string of non-templated adenosines at the 3′ end of pre-mRNAs, known as the poly(A) tail, which promotes cytoplasmic export and mRNA translation. The downstream cleavage product of CPSF-73 is subsequently degraded by the 5'-3' exonuclease Xrn2. One model of PolII termination, called the "torpedo" model, proposes that Xrn2 chases the transcribing polymerase, ultimately displacing the elongation complex from chromatin and triggering termination. The competing "allosteric" model argues that transcription of the pAs triggers an alteration in the PolII structure or the composition of the elongating complex, leading to termination.

In support of the torpedo model, increased RNA signal downstream of the 3′ cleavage site was observed for protein-coding and long noncoding primary microRNA (lnc-pri-miRNA) genes upon Xrn2 depletion (Eaton et al. 2018, Sousa-Luís et al. 2021). Some functional redundancy was observed between Xrn2 and the exosome complex

in degrading the downstream RNA cleavage products and promoting termination (Eaton et al. 2018). However, termination of replication-dependent histone mRNAs and snRNAs was unaffected by Xrn2 loss, suggesting that an allosteric mechanism may regulate transcription termination at such loci (Eaton et al. 2018, Sousa-Luís et al. 2021). In contrast to mRNAs, cleavage of pre-snRNAs is mediated by the Integrator complex, via the endonuclease subunit Int11 (Baillat et al. 2005). Int9 and Int11 bear sequence homology to the CPSF subunits CPSF-100 and CPSF-73, respectively, but do not mediate pre-mRNA cleavage. Accumulating evidence suggests that Integrator also participates in 3′-end formation and transcription termination of sense and anti-sense enhancer RNAs (Lai et al. 2015) and replication-dependent histone mRNAs in a DSIF- and NELF-dependent manner (Skaar et al. 2015). Consistently, accumulation of unprocessed eRNAs and impeded PolII recycling coincided with reduced expression from the associated gene promoters (Lai et al. 2015). Recently, it has been suggested that the endonucleolytic activity of Integrator may attenuate PolII transcription of certain protein-coding genes by promoting premature termination of paused polymerases (Beckedorff et al. 2020, Elrod et al. 2019, Tatomer et al. 2019).

Separately, Int6 and Int8 subunits were reported to antagonise CDK9 activity at the transcriptional pause-release checkpoint by recruiting the PP2A phosphatase (Vervoort et al. 2021). The balance between CDK9 phosphorylation of PolII and DSIF and dephosphorylation by PP2A may fine-tune the transcriptional output, as perturbations to this equilibrium resulted in rampant transcription and were linked to several types of cancer (Vervoort et al. 2021).

1.3 Small Nuclear RNA Ribonucleoprotein (snRNA) biogenesis

Co-transcriptional removal of introns is the other important step in PolII-directed expression of protein-coding genes, along with 5′ capping and 3′ polyadenylation. The splicing reaction is mediated to a large extent by the RNA components of the spliceosome in two sequential transesterification reactions (Valadkhan et al. 2009). Two types of spliceosomes have been characterised in eukaryotes, a major and a

minor form. The major spliceosome consists of uracil-rich U1, U2, U4, U5 and U6 snRNPs and catalyses removal of introns flanked by the canonical GU/AG splice sites, while the minor form comprises U11, U12, U4atac, U5 and U6atac snRNPs and recognises the less common AU/AC splice sites. With the exception of U6 and U6atac snRNAs, which are transcribed by PolIII, the remaining snRNAs are products of PolII transcription.

1.3.1 snRNA transcription

Expression of U snRNA genes by PolII bears similarities to that of protein-coding genes, however there are also considerable differences. Both sets of RNAs are co-transcriptionally capped, however PolII-transcribed snRNA genes lack introns and the uracil-rich non-coding snRNAs themselves are devoid of polyA+ tails. Their promoter structure is distinct as well, consisting of a distal sequence element (DSE) and a proximal sequence element (PSE), which serve as enhancer and promoter for these genes, respectively (Figure 1.3A). The DSE and PSE of snRNA genes cannot support synthesis of polyadenylated mRNA (Dahlberg & Schenborn 1988). Moreover, a 3′ box element downstream of the U snRNA encoding region directs 3′ end formation and is intimately connected with transcription from snRNA gene-specific promoters, in contrast to the polyadenylation signal that regulates mRNA transcription termination (Hernandez & Weiner 1986, Vegvar et al. 1986).

The DSE and PSE of PolII-transcribed snRNA genes are similar to and interchangeable with the PolIII-transcribed U6 and 7SK genes, however the presence of a TATA box between the PSE and the transcription start site (TSS) of U6 and 7SK genes confers PolIII-identity (Lobo & Hernandez 1989). Insertion of a TATA box upstream of the U2 snRNA TSS promotes transcription by PolIII, while deletion of the TATA box from U6 snRNA gene directs transcription by PolII (Lobo & Hernandez 1989). The DSE enhancer contains binding sites for the Oct1, Sp1/NF1 and Staf transcription factors (Janson et al. 1989). Oct1 potentiates binding of the PSE transcription factor (PTF) to the PSE promoters of both PolII and PolIII-transcribed U snRNA genes (Murphy et al. 1992). In turn, PTF helps to recruit the TATA box-binding protein (TBP) and a subset of TBP-associated factors (TAFs), which collectively form part of the general transcription factor TFIID. Together with TAF6 and TAF9,

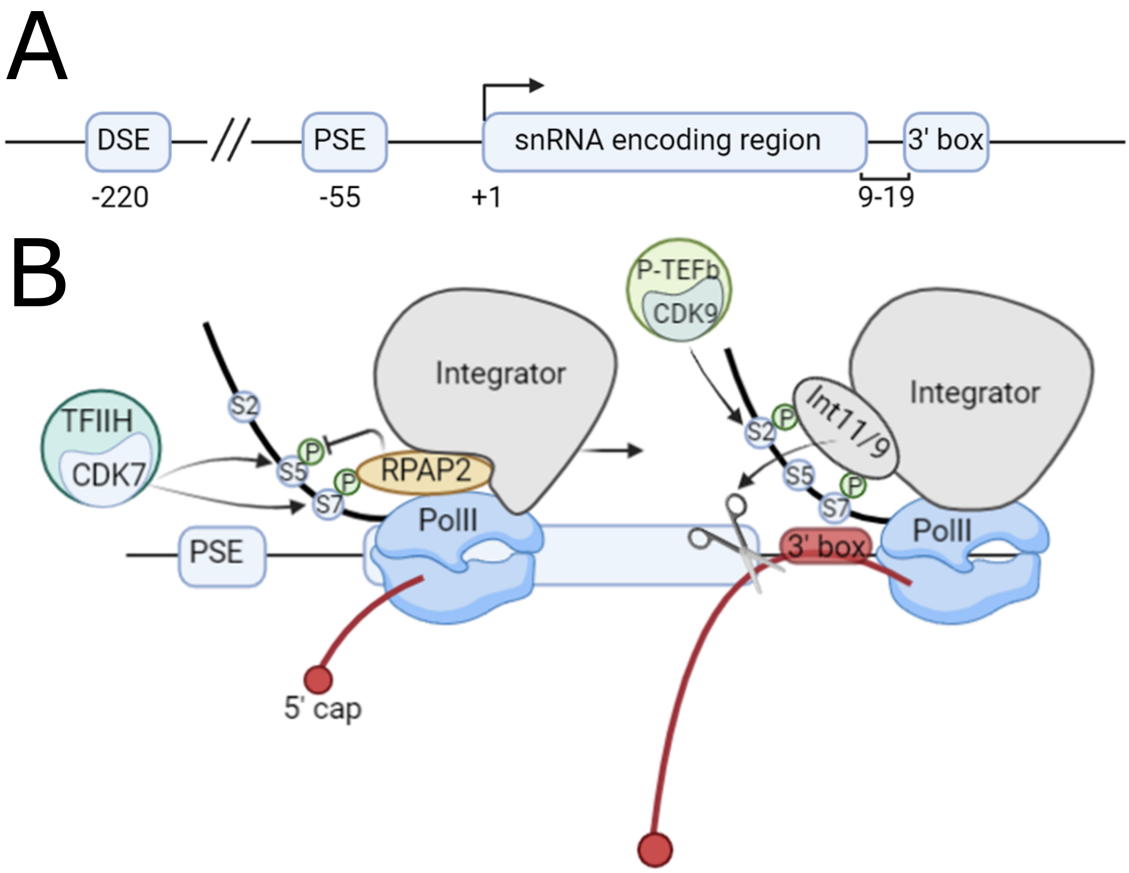

Figure 1.3: PolII-driven transcription of snRNA genes.
A) The structure of PolII-transcribed snRNA genes includes distal and proximal sequence elements 220 and 55 bases upstream of the first transcribed nucleotide, which serve as enhancer and promoter, respectively. A 3′ box is located up to 19 nucleotides downstream of the snRNA-encoding region and directs cleavage of the pre-snRNA by the Integrator complex. **B)**. PolII CTD phosphorylation events orchestrate co-transcriptional recruitment of processing complexes. Ser7P and Ser2P marks produced by CDK7 and CDK9 recruit the catalytic subunits Int11/9 of the Integrator, which cleaves the nascent transcript upstream of the 3′ box. Endonucleolytic cleavage by the Integrator is coupled with snRNA transcription termination via a little understood mechanism.

TBP clusters at the PSE, while TAFs 5, 8, 11 and 13 accumulate broadly over the PSE and coding region of U2 snRNA gene (Zaborowska et al. 2012). The U1, U4, U5 and U11 snRNA genes are distinguished by their use of TAF7, which recruits the Mediator complex via the Med26 subunit. In a handover mechanism, Med26-containing Mediator then exchanges TAF7 binding for the little elongation complex (LEC), promoting snRNA transcription elongation (Takahashi et al. 2015).

In addition to components of TFIID , other general TFs contribute to snRNA gene expression, including TFIIA, TFIIB, TFIIF, TFIIE and TFIIH (Kuhlman et al. 1999). Through its kinase activity mediated by the CDK7 component, TFIIH phosphorylates PolII CTD at positions Ser5 and Ser7. Co-transcriptional capping of snRNAs occurs later than pre-mRNA capping, raising the possibility of other factors protecting the nascent 5′ ends. ChIP analysis reported CE and CBC peaks close to the 3′ end of the snRNA coding region, possibly resulting from the lack of early PolII pausing (Glover-Cutter et al. 2007). Egloff et al. (2007) demonstrated that Ser7 phosphorylation is involved in 3′ end processing of snRNA transcripts by recruiting the Integrator complex early in the transcription initiation phase. More specifically, Ser7P is an important mark recognised by RPAP2 (RNA polymerase-associated protein 2), which mediates recruitment of ten out of the twelve subunits of the Integrator complex on PolII at snRNA gene promoters (Egloff et al. 2012). This provides a mechanistic link between transcription initiation from an snRNA gene promoter and 3′ end formation (Hernandez & Weiner 1986). Additionally, RPAP2 serves as a Ser5P phosphatase (Egloff et al. 2012). Subsequently, phosphorylation of Ser2 by the CDK9 subunit of P-TEFb, together with Ser7P, recruit Int9 and the catalytic Int11 subunits of Integrator (Figure 1.3B), which drive 3′ end recognition and correct processing of the primary snRNA transcript at the 3′ box (Egloff et al. 2010). Disruption of Integrator leads to defects in snRNA termination and acquisition of abnormal polyadenylated tails (Skaar et al. 2015).

Unlike their inhibitory roles on PolII elongation during mRNA expression, DSIF and NELF are involved in proper transcription termination and 3′ end processing at snRNA genes. Both proteins interact with Integrator subunits towards the 3′ end of the coding region and prevent recruitment of the cleavage stimulation factor (CstF)

and aberrant polyadenylation of snRNAs (Yamamoto et al. 2014). Importantly, transcription at snRNA loci terminates beyond the 3′ box, although this sequence is essential for 3′ processing and cleavage by Integrator (Cuello et al. 1999). Recently, other Integrator-independent termination pathways have been proposed for snRNA transcription, including intrinsic PolII termination at runs of Ts in the non-template strand akin to PolIII (Davidson et al. 2020).

1.3.2 Small Nuclear Ribonucleoprotein (snRNP) maturation

Nuclear synthesis of snRNAs is intimately coupled with co-transcriptional processing and cytoplasmic export for further rounds of maturation. Like mRNAs, U snRNAs acquire an inverted monomethylated G cap at the 5′ end co-transcriptionally, which is subsequently recognised by the cap-binding complex (CBC). In contrast to mRNAs, which commit to the NXF1 nuclear export pathway after being sampled by hnRNPC tetramers, snRNA transcripts fail to form a stable assembly with hnRNPC due to their short size (<250 nt) and are subjected to a separate export pathway (Masuyama et al. 2004, McCloskey et al. 2012). As such, the CBC engages with ARS2 and with one of two mutually-exclusive complexes that determine the fate of the transcript. A CBC-ARS2 interaction with the phosphorylated adaptor for RNA export (PHAX) commits the RNA for nuclear export, while interaction of CBC-ARS2 with ZC3H18 recruits the nuclear exosome targeting complex (NEXT) and leads to degradation of the transcript (Giacometti et al. 2017, Hallais et al. 2013). If the choice for snRNA export is made, PHAX mediates transport through the nuclear pore complex in conjunction with the RanGTP-bound export receptor chromosome region maintenance 1 (CRM1) (Ohno et al. 2000). En route to the cytoplasm, the snRNA cargoes are trafficked through specialised nuclear structures known as Cajal bodies, which are enriched in mature snRNP-associated splicing factors and are the site of extensive snRNP remodelling and modification. After translocation to the cytoplasm, PHAX is dephosphorylated, which facilitates disassembly of CBC and PHAX from the snRNA and recycling of the proteins into the nucleus (Ohno et al. 2000).

In the cytoplasm, the survival of motor neuron (SMN) complex assists snRNP assembly and maturation. SMN chaperones association of a heteroheptameric Sm

ring around the conserved Sm-binding site on PolII-transcribed snRNAs, flanked by two stem loops. The Sm ring consists of proteins B/B′, D1, D2, D3, E, F, G, while U6 and U6atac receive an Lsm ring (like-Sm) in the nucleoplasm consisting of proteins Lsm2-8. Additionally, the SMN-guided maturation process includes 5′ cap hypermethylation and 3′-end trimming and is concluded with snRNP re-import into the nucleus. Cap hypermethylation is the product of trimethylguanosine synthase 1 (TGS1) enzymatic activity and relies on a pre-assembled Sm ring, with deletion of the Sm-binding site leading to a block in cap trimethylation (Mattaj 1986, Mouaikel et al. 2003). The trimethyl cap and Sm ring act as nuclear localisation signals, recruiting the import adaptor protein Snurportin-1, which then facilitates nuclear re-import via the import receptor Importin-β (Huber et al. 1998, Palacios et al. 1997)

The final stages of snRNP maturation and formation or recycling of splicing-competent U4/U6·U5 di- and tri-snRNP complexes take place in the Cajal bodies, where SMN disassembles from snRNP particles and the small RNA components receive modifications, including pseudouridylation and 2'-O-methylation. The nucleotide modifications are essential for the splicing function and are elicited by another group of small non-coding RNAs known as small Cajal body-specific RNA (scaRNA) (Jády et al. 2003). Assembly of the complete and functional U snRNPs in Cajal bodies far away from their sites of activity on pre-mRNA serve to sequester the catalytic potential and prevent erroneous splicing, as well as providing a highly concentrated hub with all the prerequisite components. Subsequently, mature U snRNPs accumulate in nuclear speckles, which are often located in the vicinity of highly transcribed loci (Matera & Wang 2014).

In the context of the U4/U6·U5 tri-snRNP, the structure assumes a roughly triangular shape with the vertices formed by the two U4 and U5 Sm rings and one U6 Lsm ring. Extensive base pairing between U4 and U6 snRNAs via the 5′-terminal region of U4 and a central region of U6 maintain the latter in a catalytically-repressed state, while the overall structure is stabilised by an array of protein components (Agafonov et al. 2016). The region of U4 snRNA base paired with U6 is interrupted by a 5′ stem loop (5′ SL) containing a kink, where a small U4/U6 nucleation fac-

tor, called Snu13 or 15.5K, binds. Snu13 constitutes a platform on U4 snRNA for assembling additional U4- and U4/U6-specific protein factors, such as Prp31 and CypH/Prp3/Prp4, respectively, in a strictly hierarchical manner (Nottrott et al. 2002).

Recruitment of U4/U6·U5 tri-snRNP to the pre-mRNA for splicing is closely followed by ejection of U4 snRNP through the activity of Brr2 RNA helicase, which unwinds the U4-U6 snRNA base pairing (Matera & Wang 2014). Numerous protein and RNA rearrangements orchestrate the two steps of the splicing reaction, including displacement of U1 snRNP and extensive base pairing between U6 and U2 snRNAs, U6 and the 5′ splice site, U2 and the branch point and U5 snRNA and the 5′ exon. Eventually, the snRNP components are recycled for additional rounds of splicing, which includes re-assembly of the U4/U6·U5 tri-snRNP. SART3 is an important tri-snRNP recycling factor involved in re-assembly of splicing-competent U4/U6·U5 complexes in two ways. Firstly, the U6-specific factor SART3 promotes di-snRNP formation from individual U4 and U6 snRNPs inside Cajal bodies (Bell et al. 2002). Subsequently, ubiquitin-ligase Prp19 catalyses nonproteolytic ubiquitylation of U4 snRNP subunit Prp3, which increases its affinity to the U5 snRNP factor Prp8 and promotes tri-snRNP assembly. Following U4/U6·U5 recruitment to the spliceosome, SART3 promotes removal of the ubiquitin marks through the activity of deubiquitinating enzyme Usp4, reducing the U4-U5 stability and leading to U4 dissociation from the tri-snRNP (Song et al. 2010). Importantly, SART3 functions exclusively as a recycling factor and has not been detected in association with processing spliceosomes (Bell et al. 2002).

1.4 Liquid-liquid phase separation (LLPS)

In recent years, the process of liquid-liquid phase separation (LLPS) has come into focus as a prevalent mechanism underpinning the formation of biomolecular condensates and membraneless organelles. The range of functions where LLPS plays a role includes the formation of nucleoli (Chong et al. 2022), nuclear speckles (Maharana et al. 2018), stress granules and P bodies (Jain et al. 2016), as well as cellular signalling (Li et al. 2012), transcription hubs (Boija et al. 2018) and centrosomes

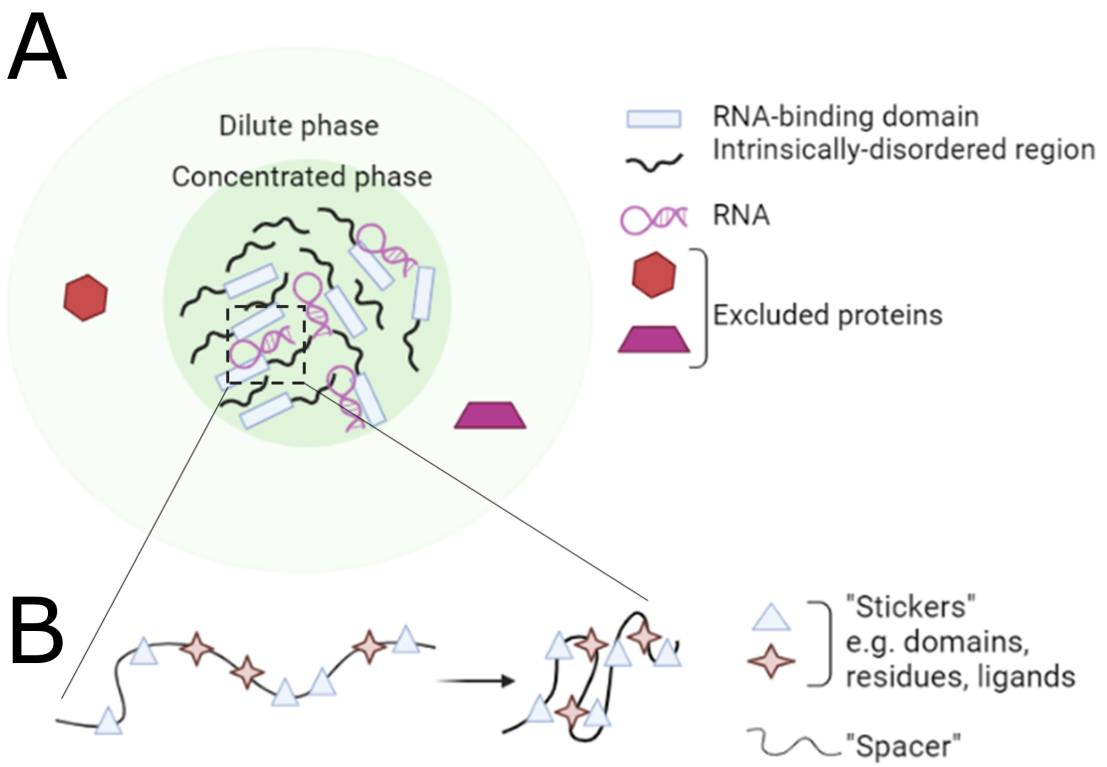

Figure 1.4: Overview of liquid-liquid phase separation.
A) Typical LLPS droplet forms by multiple protein-protein and protein-RNA interactions. This gives rise to a spherical region known as "the concentrated phase", where certain proteins accumulate by virtue of their RNA-binding domains and disordered regions, while other proteins are excluded and remain in the surrounding solute or "dilute phase". **B)** Schematic representation of the "stickers and spacers" model. Certain protein regions, such as domains, motifs and amino acids, can interact with each other or with ligands to form local hubs of highly-concentrated components. Examples of stickers include SH3 domains and the PRM motif ligand, RBDs and RNA molecules, residues with complementary electrostatic charge or which form hydrogen binds, cation-π, π-π and hydrophobic interactions. Several "stickers" can be located on the same molecule, i.e. peptide or nucleic acid chain, and are separated by non-reactive regions known as "spacers".

(Woodruff et al. 2017). Mutations that alter the (dis)assembly and localization of phase separated condensates have been linked to neurodegenerative disorders and cancer (Bouchard et al. 2018, Murakami et al. 2015, Patel et al. 2015).

Phase separation relies on multiple protein-protein and protein-RNA interactions, which drive the formation of spherical droplets enriched in their protein constituents over the surrounding solute (Lin et al. 2015). The droplets are known as the "concentrated phase", while the medium around them is called the "dilute phase" (Figure 1.4A). With time, the liquid-like droplets become stiffer and more solid-like, giving rise to assemblies known as hydrogels. These entities are formed by system-wide protein polymers (Kato et al. 2012) and are generated *in vitro* by long incubations of concentrated protein solutions at low temperatures (Harmon et al. 2017, Kato et al. 2017). LLPS condensates *in vivo* are highly complex entities comprising a multitude of proteins and mRNA molecules (Han et al. 2012, Kato et al. 2012) and are believed to accelerate biological processes by increasing the local concentration of factors and by bringing the necessary components into close proximity. Because of this layer of complexity, LLPS is often studied *in vitro* using model proteins like Fused in Sarcoma (FUS), Ewing Sarcoma (EWSR1), TATA-binding protein-associated factor 15 (TAF15), trans-activating response region DNA-binding protein 43 (TDP43) and heterogeneous nuclear ribonucleoproteins hnRNPA1 and hnRNPA2. Even when LLPS studies are carried out *in vivo*, the focus is on a small subset of protein factors and known disease mutations. However, *in silico* modelling (such as coarse grain simulations) has been tremendously insightful for understanding the submolecular events that may occur within LLPS condensates.

A useful conceptual framework for the types of interactions underpinning LLPS introduces the notions of "stickers" and "spacers" (Borcherds et al. 2021). "Stickers" represent any residues or domains that interact with other protein or RNA elements (such as amino acids, sequence motifs or nucleotides) and are linked together by "spacers", which refer to protein regions that are not involved in any intermolecular interactions conducive to LLPS (Figure 1.4B). Such arrangements are reminiscent of beads on a string. The spacing of "stickers" and the length and amino acid composition of "spacers" determine the mechanical properties of the condensate, such as

the fluidity or stiffness (Borcherds et al. 2021, Harmon et al. 2017). Moreover, the number of "stickers" in a system, also known as the valency, is inversely proportional to the protein concentration at which phase separated droplets begin to form. The "stickers and spacers" framework can be applied to the two main modes of achieving phase separation: multivalent interactions between folded domains and their ligand sequences or multivalent interactions between unfolded intrinsically-disordered regions (IDRs). Both types of interactions (involving folded domains and IDRs) can co-exist in a phase-separated droplet (Wang et al. 2018a). The molecular basis of folded domain-ligand interactions is quite well understood due to the availability of structural information. Such multivalent interactions between repeat domains and their ligand motifs have been directly linked to partitioning of both protein partners to phase separated droplets. For example, Li et al. (2012) demonstrated that increasing the number of repeats of the SH3 domain and its proline-rich motif (PRM) ligand within synthetic constructs lowered the saturation concentration at which protein droplets began to form, both *in vitro* and *in vivo*.

On the other hand, the molecular interactions between proteins containing IDRs are less well characterised, due to the lack of structural information on low complexity protein regions, which do not adopt a discrete three-dimensional fold. This is an active area of research and there are currently two (partially) competing models for how IDR-IDR interactions lead to protein phase separation.

1.4.1 Molecular basis for phase separation by intrinsically-disordered regions (IDRs)

A large proportion of the components identified in membraneless organelles are composed of RNA-binding proteins harbouring intrinsically-disordered regions (Castello et al. 2012). Evidence is emerging in support of two different models for interactions governing phase separation between IDR-containing proteins. The first model involves the formation of reversible cross-β sheet fibrils between short sequence motifs located within IDRs. In direct opposition to this, the second model relies on multiple weak interactions between residues of IDRs that remain unfolded and lacking any organised structure.

Cross-β sheet filaments are formed by ordered stacking of two tandem repeats with the sequence (Ser/Gly)Tyr(Ser/Gly) (or variations thereof) located within IDRs and have been characterised by X-ray crystallography and electron microscopy (Hughes et al. 2018, Murray et al. 2017). Such structures are reminiscent of amyloid fibrils formed by disease-associated proteins like Aβ, Tau, α-synuclein and prion Sup35 (Sawaya et al. 2007). The β-amyloid fibrils mentioned above are all formed by two mating β-sheets with interdigitating side chains which create a dry, solvent-depleted interface known as a dry steric zipper (Sawaya et al. 2007). Unlike these highly stable pathogenic amyloid formations, the cross-β filaments assembled from short IDR segments of FUS, TDP43 and hnRNPA2 contain two mating β-sheets with fewer side chain contacts and a smaller buried interface caused by a kink in the peptide backbone at Gly or aromatic residues (Guenther et al. 2018, Hughes et al. 2018, Lu et al. 2020, Luo et al. 2018). The weaker interactions between the two sheets permit rapid disassembly of the cross-β filaments. The short amino acid motifs that constitute the building blocks of cross-β filaments were termed LARKS, for low-complexity aromatic-rich kinked segments (Hughes et al. 2018) or RACs, for reversible amyloid cores (Luo et al. 2018).

The reason why cross-β filaments are proposed to underpin phase separation is because the short sequence motifs (LARKS or RACs) mentioned before are necessary for droplet formation (Luo et al. 2018). These motifs are also sites of post-translational modifications, such as Ser, Tyr and Thr phosphorylation, which negatively affect both fibrillization and phase separation (Kwon et al. 2014, Luo et al. 2018, Murray et al. 2017). Moreover, solid-state NMR investigations uncovered a 57-residue fragment within the 214-residue N-terminal low-complexity region of FUS which appeared immobilised and structurally-ordered into β-sheets, while the rest of the sequence remained highly flexible (Murray et al. 2017, Schwartz et al. 2013). Importantly, a footprinting assay of the low-complexity region of hnRNPA2 characterised a similar protein ordering in phase separated droplets, hydrogels and inside cellular nuclei (Xiang et al. 2015). The study reported the co-existence in all three cases of unstructured and ordered protein regions, where the ordering relied upon aromatic residues reminiscent of LARK conformations.

The second competing model for phase separation dictates that IDRs remain completely disordered and that multiple weak interactions between amino acids give rise to biomolecular condensates. This model is supported by studies that fail to observe any structured protein assemblies within LLPS droplets. For instance, Burke et al. (2015) and Ryan et al. (2018) reported that the IDRs of FUS and hnRNPA2 remained completely disordered inside LLPS droplets formed by both isolated IDRs and full-length FUS protein, as examined by NMR. Later, the same group expanded their observations to argue that all FUS and hnRNPA2 IDR residues make contact with all others inside an LLPS droplet based on NMR datasets (Murthy et al. 2019, Ryan et al. 2018). Such a network would only be possible for freely-diffusing proteins lacking secondary structure.

The molecular basis for this second model comprises cation-π interactions between Arg-Tyr, as well as Tyr-Tyr and Tyr-Gln hydrophobic interactions, which characterise FUS droplet formation (Burke et al. 2015, Wang et al. 2018b). Interactions involving Arg are also important for hnRNPA2 phase separation (Ryan et al. 2018). Additionally, hydrogen bonding (Murthy et al. 2019), electrostatic interactions (Nott et al. 2015) and π-π stacking (Vernon et al. 2018) involving sidechains and the protein backbone all play a role in phase separation for different IDRs .

Currently, it is uncertain to what extent each model contributes to biomolecular condensation inside the cell. The experimental approaches used to obtain microcrystals and hydrogels containing cross-β filaments (low temperatures, high protein concentrations, prolonged incubations on the order of days) would argue against cellular relevance for these assemblies. However, it is possible that cross-β filaments may appear or play a more important role later in the maturation stages of LLPS droplets (Lin et al. 2015), in pathogenic backgrounds or with older age (Lu et al. 2020). Centrosomes display characteristics of hydrogels early on after formation, which may be important for withstanding the pulling forces exhibited by the microtubule spindle (Woodruff et al. 2017). Some disease-associated mutations have been shown to stabilize cross-β filaments in a dry steric zipper conformation *in vitro*, suggesting that these fibrils could play a significant role in pathogenesis for some patients by stabilizing otherwise transient protein interactions (Guenther et al. 2018,

Lu et al. 2020).

1.4.2 Ribonucleoprotein (RNP) granule formation

A major driver of membraneless compartmentalisation in cells are interactions between low complexity protein domains (Lin et al. 2015, Ryan et al. 2018). Disordered regions such as those mediating phase separation are enriched in RNA-binding proteins (RBPs), which display Gly, Arg, Lys and Tyr amino acid biases (Castello et al. 2012). Furthermore, proteins involved in membraneless condensation also contain various RNA-binding domains (RBDs) and are associated with mRNA biogenesis and maturation (Kato et al. 2012). In addition, low-complexity segments akin to those forming cross-β fibrils have been predicted in hundreds of human proteins, including in the disordered regions of RNA-, DNA- and nucleotide-binding proteins that had not been previously linked to phase separation (Hughes et al. 2018). Molecular condensates are often called RNP granules, in order to reflect the RNA and protein composition. The import of mRNAs into granules is dependent upon interactions with the RNA-binding domains of phase separating proteins and seems to involve longer 3′-UTR elements compared to mRNAs that are not packaged into RNP condensates (Han et al. 2012). RNA-RBD interactions promote phase separation in a positive feedback loop (Li et al. 2012, Lin et al. 2015, Schwartz et al. 2013). However, Maharana et al. (2018) elegantly demonstrated that cellular droplets are highly sensitive to nuclear RNA levels, which keep in check protein condensation by displaying buffering effects. Moreover, RNA secondary structure appears to choreograph the protein phase behaviour, with highly structured long non-coding (lnc) and polyA+ RNAs abrogating the inhibitory effects of short RNAs and possibly seeding paraspeckle and stress granule assembly in the nucleus and cytoplasm (Maharana et al. 2018). Following enzymatic depletion of nuclear RNA, the FET proteins and TDP43 rapidly coalesced into droplets which quickly aged and developed hydrogel properties, leading the authors to suggest a connection between mutations in RBDs and pathogenic IDR assemblies (Maharana et al. 2018).

Assembly of biomolecular condensates is also crucially sensitive to cellular ATP levels, which exhibit buffering effects at high or near-native concentration, similar to RNA (Kang et al. 2019, Jain et al. 2016, Patel et al. 2017). A clear demonstra-

tion of protein solubilization by ATP was performed by incubating egg whites with increasing ATP levels at >60°C: around 10 mM ATP (which is at the high end of cellular concentrations (Greiner & Glonek 2021)) blocked egg white cooking by stabilizing globular protein folds and preventing entanglement produced by melting of secondary and tertiary structures (Patel et al. 2017).

1.4.3 Phase separation and RNA polymerase II transcription

Over half of the chromatin-associated proteins (excluding histones) are RNA-binding proteins (RBPs), such as hnRNPs and transcription factors, involved with different steps of RNA processing. In addition, binding of the majority of these RBPs to chromatin occurs in a transcription- and RNA-dependent manner (Shao et al. 2022). Transcription factors (TFs) comprise a structured DNA-binding domain (DBD) and an activation domain that is crucial for PolII recruitment (Kadonaga et al. 1988). The activation domain is frequently a disordered region with low amino acid complexity (Boija et al. 2018). Such TFs have the ability to undergo phase separation by virtue of their IDR-containing activation domains and promote PolII-directed transcription at target promoters by interacting with the disordered CTD of PolII (Boehning et al. 2018, Schneider et al. 2021, Zuo et al. 2021). Using live-cell single-molecule imaging, Chong et al. (2018) directly observed clustering of TFs (including the FET proteins: FUS/EWSR1/TAF15) at native protein concentrations into dense hubs through DNA-DBD and IDR-IDR interactions. The transcriptionally-active hubs displayed characteristic LLPS features (highly dynamic, multivalent protein interactions) and also sequence specificity for IDR interacting partners, which is an important mechanism for regulating gene expression (Chong et al. 2018). The highly dynamic transcriptional hubs correlate with transient Mediator and PolII clusters observed to rapidly assemble and disassemble in cells with an average lifetime of 5.1 s (Cisse et al. 2013). Importantly, Burke et al. (2015) reported NMR evidence that PolII CTD preferentially interacts with the transcription factor FUS assembled into LLPS droplets instead of in its monomeric form.

Compelling evidence for the link between transcription regulation and IDR-mediated phase separation of TFs comes from cancer-associated mutations. A well-studied example is Ewing's sarcoma, caused by an oncogenic TF fusion protein between

the IDR of EWSR1 and the DNA-binding domain of FLI1 as a result of genomic translocation (May et al. 1993). Dynamic and multivalent IDR-IDR interactions enrich the TF at FLI1-specific GGAA microsatellite loci beyond the availability for DNA-DBD association (Chong et al. 2018). Moreover, EWS/FLI1 occupancy overlaps with transcriptional output from genes around GGAA microsatellites in a tightly balanced manner, with increases in local TF density past a certain threshold reducing gene expression (Chong et al. 2022).

Phase separation has also been proposed as the mechanism connecting IDR-containing TFs and coactivators more broadly, identifying common residues necessary for both transcriptional activation and TF inclusion into dynamic LLPS droplets at transcriptional hubs (Boija et al. 2018). Liquid-like droplets containing Mediator and BRD4 coactivators have been observed at super-enhancer nuclear loci and chemical disruption of the condensates was accompanied by a loss of chromatin-associated transcriptional machinery and a reduction of transcriptional output from the loci in question (Sabari et al. 2018). The same study also demonstrated that *in vitro* assembled Mediator-IDR droplets accrued the transcriptional apparatus from nuclear lysates, demonstrating that Mediator could compartmentalize transcription inside the cell and in an artificial environment through the process of phase separation.

The role that RNA plays in LLPS droplet assembly has also been investigated in the context of transcriptional hubs. The paraspeckle RNA-binding protein PSPC1 employs its IDR and RNA-binding domain to engage with the pre-initiation complex (PIC) and hypo- or Ser5-phosphorylated PolII at discrete LLPS puncta in embryonic stem cells (ESCs) (Shao et al. 2022). The authors also noted that RNA increased PolII retention within condensates, proposing that RNA byproducts from basal transcription enhance RBP recruitment to chromatin loci and consequently stabilise PolII at the promoter until proper CTD phosphorylation, consistent with the report by Henninger et al. (2021). In agreement with this suggestion, PolII is evicted from LLPS condensates after phosphorylation by CDK7 and CDK9 (Kwon et al. 2014, Shao et al. 2022), presenting a feasible mechanism for PolII escape from transcriptional hubs at promoter-proximal pause sites.

The model of LLPS-mediated transcription control bridges the gap between the intricate DNA-sequence specificity offered by the DNA-binding domains of TFs, a limited set of coactivators (including the Mediator complex, BRD4 and p300) sampled by the majority of TFs and PolII recruitment for highly-controlled gene expression. Phase separation represents a dynamic mechanism to orchestrate tight control over gene regulation in response to changing cellular cues by virtue of intrinsically-disordered regions present in the majority of the actors involved.

1.5 P-loop nucleoside triphosphatases (NTPases)

The phosphate-binding loop (P-loop or Walker A fold) is one of the most ancient and abundant protein motifs, with several P-loop NTPases being present in the last universal common ancestor (LUCA) of all cellular life. P-loop domains are found in nucleoside triphosphate (NTP)-binding proteins and typically consist of a mostly parallel β-sheet sandwiched between two α-helical facets, with the α-helices and β-strands encoded consecutively in the secondary structure (Leipe et al. 2003, Ma et al. 2008). The P-loop or Walker A motif itself encompasses the conserved sequence GlyXXXXGlyLysThr/Ser ("X" denoting any residue) and consists of a flexible loop connecting the first β-strand and α-helix (Figure 1.5). The motif is responsible for binding the γ and β phosphate groups of NTP substrates. Depending on the architecture around this motif, P-loop-containing proteins can hydrolyse nucleoside triphosphates and catalyse the transfer of the γ-phosphate. A 2018 study investigating the evolution of the P-loop motif suggested that it might initially have been involved just in binding phosphate-containing substrates, especially RNA and single-stranded DNA, with phosphate hydrolysis appearing later (Romero et al. 2018).

Another conserved motif found in P-loop proteins is the Walker B, with the sequence hhhhAsp, where "h" denotes a hydrophobic residue. Less commonly, P-loop kinases contain substitutions of the Asp in the Walker B motif for a Glu, however both types of acidic residues stabilise the octahedral coordination of the magnesium ion co-factor, along with the Thr/Ser of the Walker A and other residues in the active site (Walker et al. 1982). Together, Walker A and B motifs play a crucial role

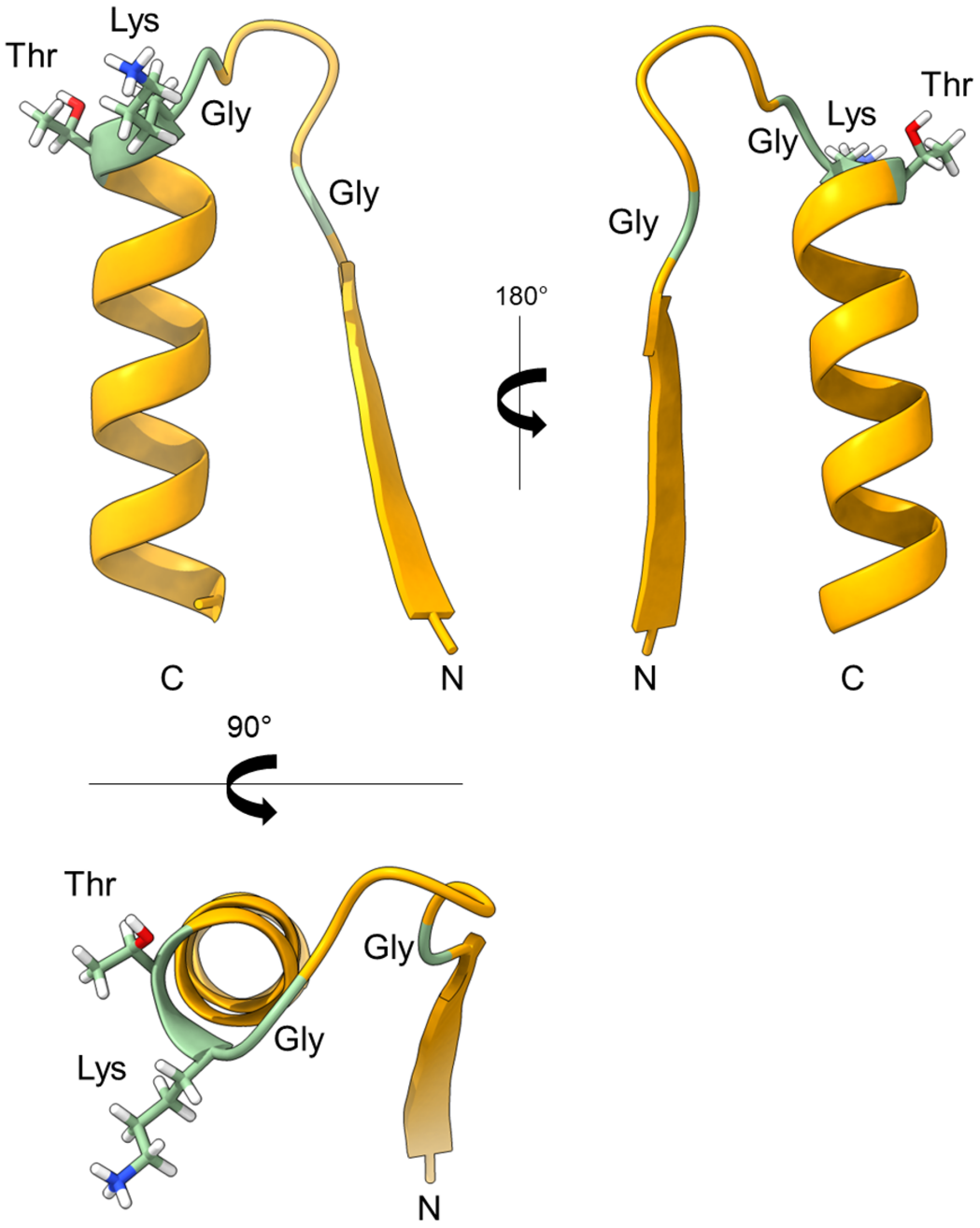

Figure 1.5: Structure of the P-loop/Walker A motif.
Architecture of phosphate loop motif nestled between a β-strand and α-helix. Key residues GxxxxGKT are coloured sage green and their sidechains are shown in stick form. Structure created from PDB: 6C2U. N= amino terminus; C= carboxyl terminus.

in NTP hydrolysis and substrate phosphorylation in NTP hydrolases and kinases, respectively.

Among the members of the P-loop NTPase family, a number of polynucleotide kinases are particularly relevant to this thesis due to their structural homology with hnRNPUL1 and two of them will be discussed below.

1.5.1 Polynucleotide Kinase Phosphatase (PNKP)

Mammalian polynucleotide kinase phosphatase (PNKP) is a dual function enzyme involved in the DNA damage response, generating compatible DNA 5′ and 3′ ends for ligation. PNKP removes 3' phosphate groups through its phosphatase domain and phosphorylates 5′ OH- groups via its kinase domain prior to DNA break repair *in vitro* and *in vivo* (Karimi-Busheri et al. 1998, Chappell et al. 2002, Loizou et al. 2004). The enzyme acts on single-strand breaks in the base-excision repair (BER) pathway (Karimi-Busheri et al. 1998, Loizou et al. 2004) and on double-strand breaks as part of non-homologous end joining (NHEJ) (Chappell et al. 2002). In addition to the 5′ kinase and 3′ phosphatase domains, PNKP also consists of an N-terminal FHA domain which recruits the protein to sites of ss/ds DNA damage via interactions with phosphorylated forms of the XRCC1 and XRCC4 components of BER and NHEJ, respectively (Koch et al. 2004, Loizou et al. 2004, Chappell et al. 2002).

PNKP functions as a monomer (Mani et al. 2003) and the minimal substrate length for phosphorylation measures 8 nucleotides, with the enzyme displaying no preference between 5′ overhangs, blunt ends, nicks or gaps for either its kinase or phosphatase activities (Karimi-Busheri & Weinfeld 1997, Karimi-Busheri et al. 1998). The kinase region of PNKP recognises the 5′ end of the DNA substrate via a positively-charged patch that leads into the active site and helps to position the 5′ OH- in a catalytically-favourable orientation (Bernstein et al. 2009). Importantly, PNKP consists of a truncated α9-loop-α10 region, which accommodates the double-stranded upstream portion on the substrate.

1.5.2 T4 Polynucleotide Kinase (T4 PNK)

Bacteriophage T4 Polynucleotide kinase is also a dual function enzyme that can both phosphorylate a variety of DNA and RNA substrates with 5′ OH- groups and dephosphorylate 3′ ends (including 2′-3′ phosphodiesters) (Richardson 1965). It is active during T4 phage infection to regenerate 5′ Pi and 3′ OH- ends of cleaved tRNAs in preparation for ligation by T4 RNA ligase (Amitsur et al. 1987). The active form of the protein is a tetramer, with no allosteric interactions between the subunits (Lillehaug & Kleppe 1975).

Unlike mammalian PNKP, T4 PNK exhibits a preference for accessible 5′ ends found in substrates such as uncapped RNA, ssDNA or dsDNA with 5′ overhangs and displays limited phosphorylation of DNA with blunt ends, 3′ overhangs, nicks or gaps (Lillehaug et al. 1976). This substrate selectivity can be explained by the presence of an extended helix-loop region above the binding pocket, which serves as a discriminator against double stranded substrates (Eastberg et al. 2004). T4 PNK differs in this respect from mammalian PNKP, which lacks such an extended region, allowing it to accommodate the second strand of a nicked DNA (Bernstein et al. 2009, Garces et al. 2011).

1.6 Heterogeneous Nuclear Ribonucleoprotein (hnRNP) family

The hnRNP family is an abundant group of nuclear RNA-binding proteins, which contribute to every step of RNA maturation. The first members of the hnRNP family were characterised following isolation by glycerol gradient of stable RNA-protein complexes (Beyer et al. 1977). Proteins of the hnRNP group are denoted from A to U, but despite classification into one family, they fulfill distinct roles in nucleic acid metabolism. The unifying features of hnRNPs are the presence of RNA-binding domains (RBDs) and the modular assembly. These proteins contain one or more RBDs from a relatively short list, including the RRM (RNA-recognition motif), KH (hnRNPK-homology), qRRM (quasi-RRM) and RGG box (Arg-Gly-Gly repeats). hnRNPs are also distinguished by a largely disordered structure, consisting predominantly of RBDs, Gly- or Pro-rich regions and acidic portions.

There are few notable exceptions of members with folded domains, namely hnRNPU and hnRNPUL1 (which will be discussed below), hnRNPUL2 and TDP43.

The functional range of hnRNPs spans all stages of RNA biogenesis, including gene expression, (alternative) splicing, polyadenylation, miRNA processing, mRNA packaging and nuclear export, translation and stability. Other nuclear roles have been characterised, such as chromatin remodelling, genome stability, DNA damage response, telomere maintenance and DNA replication (Han et al. 2010). Due to such a vast array of critical functions, mutations in hnRNP members have been associated with motor neuron disease (Matsubara et al. 2019), mental retardation, cardiac and neurodevelopmental disfunction (Yates et al. 2017) and cancer (Zhang et al. 2022).

1.6.1 hnRNPU

hnRNPU was first identified as the largest molecular weight component of the 40S core hnRNP particle, in addition to other hnRNPs and long RNA molecules (Beyer et al. 1977). It is a phosphorylated protein with RNA-binding activity mediated by the C-terminal Arg-Gly-Gly repeat region (RGG box), which was first characterised in this system (Dreyfuss et al. 1984, Kiledjian & Dreyfuss 1992). In parallel, another group identified hnRNPU as a nuclear scaffold protein and initially named it scaffold attachment factor-A (SAF-A) (Romig et al. 1992). The protein bound avidly to and promoted looping of AT-rich DNA sequences termed scaffold-attachment regions or SARs (Romig et al. 1992).

The gene encoding hnRNPU is located on chromosome 1 at position 1q.44 and consists of 14 exons. The main transcript gives rise to a protein measuring 806 amino acids in length and an apparent molecular weight of 120 kDa. The protein comprises an N-terminal SAP (SAF-A/B, Acinus and PIAS) domain, followed by a disordered region rich in acidic residues, a SPRY (splA kinase and ryanodine receptor) domain and a nucleoside triphosphate (NTP)-binding region with putative ^{485}GlyXXXXGlyLysThr^{492} Walker A and ^{557}hhhhAsp^{561} Walker B motifs (Kiledjian & Dreyfuss 1992) (Figure 1.6). The C-terminus is unstructured and harbours the RNA-binding RGG box (Kiledjian & Dreyfuss 1992). The SAP domain facilitates

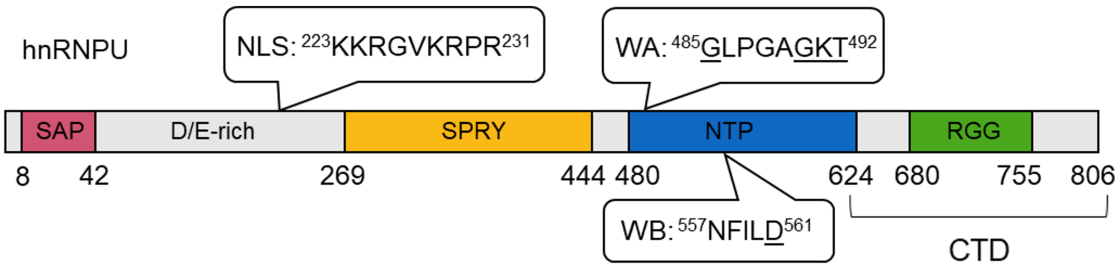

Figure 1.6: Domain architecture of hnRNPU protein.
Schematic diagram of domain organisation of hnRNPU. D/E-rich= acidic linker; NLS= nuclear localisation signal; WA= Walker A motif; WB= Walker B motif; CTD= C-terminal domain encompassing the RGG box.

DNA binding (Kipp et al. 2000), consistent with its role in maintenance of chromatin architecture (Fackelmayer & Richter 1994). A nuclear localisation signal (NLS) with the sequence ^{223}LysLysArgGlyValLysArgProArg^{231} is located at the end of the acidic region, consistent with the protein's nuclear enrichment (Kiledjian & Dreyfuss 1992). The NTP-binding region was reported to have an RNA-mediated ATPase function necessary for protein oligomerization and subsequent chromatin decompaction (Nozawa et al. 2017). The SPRY domain is predicted to mediate protein-protein interactions (D'Cruz et al. 2013).

Over the years, the idea of a stable nuclear matrix has fallen out of favour, being replaced by a responsive "mesh" model. This hypothesis proposes that transcriptionally-active regions are maintained in a more accessible state via the concerted interactions between chromatin-associated RNAs and RNA- and DNA-binding proteins in a feedback system, where reduced transcription promotes chromatin compaction (Nozawa & Gilbert 2019). In support of this model, Hall et al. (2014) observed that intergenic repeat RNAs are stably associated *in cis* with euchromatin during interphase in an hnRNPU-dependent manner and that disruption of this process favours chromatin compaction. Sharp et al. (2020) elegantly dissected the molecular basis for this interaction and reported that hnRNPU directly tethers RNAs to chromatin during interphase via its dual RNA- and DNA-binding arms located at either end of the protein. Moreover, phosphorylation of the SAP domain by Aurora-B kinase during prophase leads to eviction of hnRNPU-RNA complexes from chromatin prior to chromosome condensation and defects in this process lead to aberrant chromosome segregation during mitosis (Sharp et al. 2020).

hnRNPU has been further implicated in maintenance of 3D genome organisation by stabilising inter-chromosome interactions between topologically-associated loci in combination with the long noncoding RNA *Firre* acting *in cis* (Hacisuleyman et al. 2014). Loss of hnRNPU caused nuclear diffusion of the mature lncRNA instead of anchorage at the site of transcription and disruption of topologically-associated loci (Fan et al. 2018, Hacisuleyman et al. 2014). Moreover, Ameyar-Zazoua et al. (2009) implicated hnRNPU in the nuclear retention of the heterochromatin-associated protein HP1α and a contribution to HP1α-mediated gene silencing. An additional

function of hnRNPU related to chromatin architecture involves X chromosome inactivation in females for gene dosage compensation. hnRNPU accumulates at the inactive X chromosome (Xi) during the early stages of inactivation and tethers Xist lncRNA *in cis* by a combination of both its DNA- and RNA-binding domains SAP and RGG box (Hasegawa et al. 2010, Helbig & Fackelmayer 2003). Loss of hnRNPU again caused diffuse nuclear distribution of Xist lncRNA, reminiscent of the *Firre* lncRNA behaviour upon hnRNPU knockdown (Hacisuleyman et al. 2014, Hasegawa et al. 2010).

In addition to maintenance of chromatin architecture, several groups have reported regulatory roles for hnRNPU in PolII transcription. For instance, hnRNPU associated with the pre-initiation complex (PIC) at PolII promoters and was evicted during the early stages of transcription (Kim & Nikodem 1999). Moreover, it interacted with the CTD of PolII directly and inhibited its phosphorylation by the CDK7 subunit of TFIIH, reducing progression into productive transcription by a mechanism dependent on the SPRY-NTP domains (Kim & Nikodem 1999). This function is consistent with our observation that hnRNPU appears to inhibit its own expression when transfected into mammalian cells (Wilson lab, unpublished). Additionally, the NTP domain of hnRNPU is a phosphorylation target of CDK7 *in vivo*, therefore it may compete with PolII for access to TFIIH and be released from the PIC upon phosphorylation (Rimel et al. 2020).

Later, Martens et al. (2002) described an interaction between hnRNPU CTD and the transcription coactivator/histone acetyltransferase (HAT) p300 at SARs upstream of untranscribed genes, accompanied by hyperacetylation of histone H3. The hnRNPU/p300 localisation at SARs was lost during times of active transcription of the associated loci, consistent with the previous report that hnRNPU dissociates from the PIC prior to transcription elongation (Kim & Nikodem 1999).

On the other hand, a complex consisting of hnRNPU, β-actin and the histone acetyltransferase PCAF was reported to promote PolII transcription elongation in cells and *in vitro*. The hnRNPU/actin/PCAF complex recognised Ser2P and Ser5P/Ser2P isoforms of PolII CTD and was specifically removed from hypophosphorylated CTD. The complex promoted transcription elongation via PCAF's HAT activity and was

associated with the nascent pre-mRNA, arguing for a different role for hnRNPU than that involving transcription initiation (Kukalev et al. 2005, Obrdlik et al. 2008).

More evidence implicates hnRNPU as a regulator of splicing. CLIP-seq analysis (crosslinking followed by immunoprecipitation and sequencing) identified a plethora of RNA species associated with hnRNPU, including intronic sequences, lncRNAs and importantly, all the snRNA components of the major and minor spliceosomes (Xiao et al. 2012). Intriguingly, hnRNPU regulated the pool of mature U2 snRNP available in the nucleus and consequently, hnRNPU depletion led to altered splicing patterns globally (Xiao et al. 2012). This observation was reproduced in a more comprehensive analysis by Huelga et al. (2012), who described cooperation between hnRNPs A1, A2, H1, F, M and U in fine tuning alternative splicing genome-wide.

Finally, hnRNPU has been associated with the DNA damage response (DDR). It is a reported substrate of DNA-dependent protein kinase and becomes phosphorylated in response to double strand breaks (DSBs) (Berglund & Clarke 2009). More importantly, transient accumulation of hnRNPU at DSBs mediated by its C-terminus was observed following poly(ADPribosyl)ation (Britton et al. 2014). The transient recruitment at DSBs was accompanied by a lasting exclusion and depended on protein phosphorylation by DDR-associated kinases ATM, ATR and DNA-PK. hnRNPU's behaviour was mirrored by other RNA-processing factors such as FUS, EWSR1 and TAF15 (Britton et al. 2014). The authors argued that RBP exclusion serves as a mechanism to avoid R-loop formation by dampening transcription at sites of DNA damage.

1.6.2 hnRNPUL1

hnRNPUL1 (hnRNPU-like 1) is a nuclear-enriched RNA-binding protein that shares extensive structural homology with hnRNPU. It was first described as an important mediator of nuclear mRNA export and its overexpression overcame the adenovirus-induced block of host mRNA transport to the cytoplasm (Gabler et al. 1998). Based on homology with several other hnRNP proteins, such as hnRNPs A1, G, K, M, the authors suggested that hnRNPUL1 should be included in the hnRNP family.

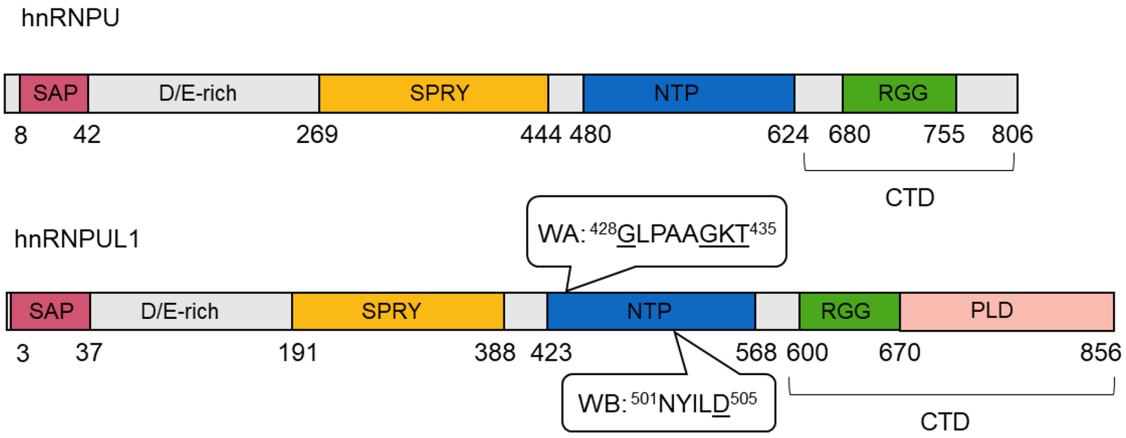

Figure 1.7: Domain architecture of hnRNPUL1 protein.
Schematic diagrams of domain organisation of hnRNPU and hnRNPUL1 highlighting the extensive structural homology. D/E-rich= acidic linker; WA= Walker A motif; WB= Walker B motif; CTD= C-terminal domain encompassing the RGG box and the proline-rich PLD.

hnRNPUL1 is encoded by a gene located on chromosome 19 at position 19q13.2 and consisting of 20 exons. The protein comprises 856 amino acids, giving rise to a predicted size of 95 kDa. Through substantial post-translational modifications (PTMs), the observed molecular weight increases to 120 kDa. The domain architecture is reminiscent of hnRNPU, with the N-terminal SAP and the central SPRY-NTP folded domains separated by a disordered acidic linker (Figure 1.7). An RNA-binding RGG box follows the NTP-binding domain and the C-terminus of hnRNPUL1 is completed by a 170-residue proline-rich region with low amino acid diversity and prion characteristics. This terminal region is termed the prion-like domain (PLD) and is specific to hnRNPUL1, where it mediates several protein-protein interactions. The C-terminus is also the site of extensive PTMs, such as arginine methylation of RGG repeats by protein arginine methyltransferases 1 and 2 (PRMT1/2) (Gurunathan et al. 2015, Kzhyshkowska et al. 2001) and serine/tyrosine phosphorylation by various protein kinases, including CDK7 (Rimel et al. 2020). The SPRY domain is predicted to mediate protein-protein interactions based on its function in other SPRY-containing proteins (D'Cruz et al. 2013). The central NTP-binding domain of hnRNPUL1 harbours the classic Walker A (^{428}GlyXXXXGlyLysThr^{435}) and Walker B (^{501}hhhhAsp^{505}) motifs and was initially proposed to bind GTP based on sequence similarity with the GTPase Ran (Gabler et al. 1998), but an enzymatic function has not been ascribed to it. Unlike the DNA-binding domain (DBD) of hnRNPU, the SAP region of hnRNPUL1 has not been directly observed to interact with DNA and is only labelled as a DBD by virtue of homology. In fact, Shao et al. (2022) reported that hnRNPUL1 associates with chromatin in a transcription- and RNA-dependent manner.

hnRNPUL1 has been studied less extensively than hnRNPU, however some functions have been described in the literature. In support of the mRNA export role suggested initially by Gabler et al. (1998), a later report characterised an interaction between hnRNPUL1 and the N-terminal portion of the nuclear export factor TAP/NXF1, which also mediates RNA binding and the two events are mutually exclusive (Bachi et al. 2000, Hautbergue et al. 2008).

Moreover, hnRNPUL1 has been implicated in the DNA damage response by acting

in the ATR-signalling pathway downstream of the MRN sensor complex to promote DNA-end resection and homologous recombination (Polo et al. 2012). Association with the NBS1 subunit of MRN and subsequent transient recruitment to laser micro-lesions after transcription inhibition were dependent on an intact and methylated RGG domain (Gurunathan et al. 2015, Polo et al. 2012). Additionally, the transient recruitment of hnRNPUL1 to double strand breaks was also contingent on poly(ADPribosyl)ation of the CTD by PARP1 enzyme (Hong et al. 2013). Similarly to hnRNPU, it has been proposed that the soluble and chromatin-associated pools of hnRNPUL1 exhibit a different behaviour at DNA damage sites, with the soluble fraction being transiently recruited by methylation and poly(ADPribosyl)ation, and the insoluble portion being evicted as a result of the transcription arrest that accompanies DNA lesions (Gurunathan et al. 2015, Hong et al. 2013, Polo et al. 2012). Interestingly, hnRNPUL1's involvement in the DNA-damage response appeared co-dependent on hnRNPUL2, which also shares extensive structural homology with hnRNPU and hnRNPUL1 (Polo et al. 2012).

Barral et al. (2005) associated hnRNPUL1 with the DNA damage response through its repressive role on p53-mediated transcriptional activation following UV radiation. On the other hand, hnRNPUL1 was reported to enhance basal transcription from the glucocorticoid-responsive MMTV promoter and to mildly repress its transcription in the presence of the glucocorticoid ligand dexamethasone (Kzhyshkowska et al. 2003). Moreover, transcription enhancement relied on an intact NTP-binding domain, which was shown to also mediate interactions between hnRNPUL1 and the histone-binding bromodomain protein BRD7 (Kzhyshkowska et al. 2003).

1.7 Aims of the study

hnRNPUL1 is an essential protein for mammalian cell viability, and it has been implicated in important cellular processes such as gene expression and maintenance of genome integrity. The central folded domain predicted to mediate ATP binding distinguishes it from most other hnRNPs, while the C-terminal prion-like domain exhibits features characteristic of phase separating proteins, many of which are known hnRNPUL1 interactors. Mutations in the *hnRNPUL1* gene have been associated

with the incurable motor neuron disease amyotrophic lateral sclerosis (ALS), a condition commonly linked to mutated hnRNPs and phase separating proteins. The current study has attempted to elucidate the molecular basis of hnRNPUL1's function in order to improve our understanding of normal and disease regulation.

Biochemical assays were used to characterise the function of the central nucleotide-binding domain, complemented by insights from structural homology with mammalian and viral polynucleotide kinases. The effects of functional and ALS-associated mutations within this domain were quantified *in vitro* and correlated with *in vivo* observations. The types of ligands and manner of binding to the central domain were then characterised and used for developing a protocol to capture *in vivo* targets. The molecular basis of interactions with RNA targets previously identified in the lab was probed with pulldown and immunoprecipitation assays. Lastly, the phase transition behaviour of hnRNPUL1 was explored under physiological conditions, and connected with a transcriptional regulatory role mediated in part by phase separation.

2 Chapter 2: Materials and Methods

2.1 Materials

2.1.1 Bacterial strains

The *E. coli* strain DH5α (Invitrogen) was used for cloning and plasmid propagation, while the *E. coli* strain BL21 DE3 (Invitrogen) was used for overexpression of recombinant proteins.

2.1.2 Growth media and reagent stocks

Luria broth (LB): 10 g/l Tryptone, 10 g/l NaCl, 5 g/l Yeast extract.
Agar plates: LB supplemented with 2% agar.
Terrific broth (TB): 12 g/L Tryptone, 24 g/l Yeast extract, 12.54 g/l K_2HPO_4, 2.31 g/l KH_2PO_4, 4 ml/l Glycerol.

Reagent	[Stock] 1000x
Ampicillin	100 (mg/ml)
Kanamycin	50 (mg/ml)
IPTG	0.5 M

2.1.3 Plasmids

Plasmid	Resistance	Source
pET24b	Kanamycin	Invitrogen
pGEX-6PI	Ampicillin	Amersham
pcDNA5-FRT	Ampicillin	Invitrogen
pET His6 mCherry	Kanamycin	Addgene #29722
pET His6 GFP	Kanamycin	Addgene #29663

2.1.4 Primers

Primer	Sequence	Description
5′ rev pcDNA5-FRT-3XFLAG	5′ TGCGGCCGCAAGCTTGTCATC 3′	Used to amplify pcDNA5-FRT-3XFLAG vector
3′ fw pcDNA5-FRT-3XFLAG	5′ CTCGAGTCTAGAGGGCCCGTTTAAACC 3′	Used to amplify pcDNA5-FRT-3XFLAG vector
UL1 SPRY fw	5′ ATGACAAGCTTGCGGCCGCAATGAGGG GCCGCTCTCCTCAGCCTCCT 3′	Used to clone SPRY-PPNK into pcDNA5-FRT-3XFLAG
UL1 PPNK rev	5′ GGTTTAAACGGGCCCTCTAGACTCGAGC TAGTTGTCAAAGCGCTTTTCAG 3′	Used to clone SPRY-PPNK into pcDNA5-FRT-3XFLAG
GFP-RGG-PLD fw	5′ CTTTAAGAAGGAGATATACATATGGTG AGCAAGGGCGAGGAGCTGTTC 3′	Clone GFP-UL1 CTD into pET24b
GFP-RGG-PLD rev	5′ GATCTCAGTGGTGGTGGTGGTGGTGCT CGAGCTGTGTACTTGTGCCACCCTGTGT 3′	Clone GFP-UL1 CTD into pET24b
mCherry-PolII CTD fw	5′ CAATCCAATATTGGTAGTGGGTATTCTC CTAGTTCTCCACGC 3′	Clone PolII CTD (last 22 repeats) into mCherry pET24b
mCherry-PolII CTD rev	5′ GGATCCGCTGCTGCCGTTGCTAGTCAGA GAGTACGTCGGAGA 3′	Clone PolII CTD (last 22 repeats) into mCherry pET24b
ΔWA fw	5′ GATGGTGGGCCTGCCTGCTGCTGCCGC GGCCACATGGGCCATCAAACATGCAG 3′	Mutate GKT of Walker A motif to AAA
ΔWA rev	5′ CTGCATGTTTGATGGCCCATGTGGCCGC GGCAGCAGCAGGCAGGCCCACCATC 3′	Mutate GKT of Walker A motif to AAA
ΔWB fw	5′ GCCCGCAAGAAACGCAACTATATCCTAG CTCAGACAAATGTTTATGGGTCAGCCCAG 3′	Mutate D of Walker B motif to A
ΔWB rev	5′ CTGGGCTGACCCATAAACATTTGTCTGA GCTAGGATATAGTTGCGTTTCTTGCGGGC 3′	Mutate D of Walker B motif to A
UL1 N456D fw	5′ CAAGAAGTACAACATCCTGGGTACCGAT GCCATCATGGATAAGATGCGGGT 3′	N456D mutation in UL1
UL1 N456D rev	5′ ACCCGCATCTTATCCATGATGGCATCGG TACCCAGGATGTTGTACTTCTTG 3′	N456D mutation in UL1
U N512D fw	5′ AGGGAAATATAACATTCTTGGCACAGAT ACTATTATGGATAAGATGATGGT 3′	N512D mutation in U
U N512D rev	5′ ACCATCATCTTATCCATAATAGTATCTG TGCCAAGAATGTTATATTTCCCT 3′	N512D mutation in U

Primer	Sequence	Description
UL1 T507A fw	5′ GCAACTATATCCTAGATCAGGCAAATGT TTATGGGTCAGCCCA 3′	T507A mutation in UL1
UL1 T507A rev	5′ TGGGCTGACCCATAAACATTTGCCTGAT CTAGGATATAGTTGC 3′	T507A mutation in UL1
UL1 R516A fw	5′ GTTTATGGGTCAGCCCAGAGAGCAAAA ATGAGACCATTTGAAGG 3′	R516A mutation in UL1
UL1 R516A rev	5′ CCTTCAAATGGTCTCATTTTTGCTCTCT GGGCTGACCCATAAAC 3′	R516A mutation in UL1
GFP fw	5′ CTTTAAGAAGGAGATATACATATGGTG AGCAAGGGCGAGGAGCTGTT 3′	PCR amplification of eGFP
GFP rev	5′ TCCACTTCCAATATTGGATTGGAAGTAC AGG 3′	PCR amplification of eGFP
UL1 CTD fw	5′ CCTGTACTTCCAATCCAATATTGGAAGT GGAGAC 3′	PCR amplification of UL1 codon-optimized CTD
UL1 CTD rev	5′ GATCTCAGTGGTGGTGGTGGTGGTGCT CGAGTTGTGTCGACGTTCCTCCCTGAGT 3′	PCR amplification of UL1 codon-optimized CTD
Δacid loop fw	5′ AGGGGCCGCTCTCCTCAGCCTCCTGCTG AA 3′	Exchange WT for mutated SPRY-PPNK into pGEX-6PI vector
Δacid loop rev	5′ AATGAACAGAACCTCATCCAGGAAGTCC CCAAC 3′	Exchange WT for mutated SPRY-PPNK into pGEX-6PI vector
GST-UL1 vector fw	5′ GTTGGGGACTTCCTGGATGAGGTTCTG TTCATT 3′	Exchange WT for mutated SPRY-PPNK into pGEX-6PI vector
GST-UL1 vector rev	5′ TTCAGCAGGAGGCTGAGGAGAGCGGCC CCT 3′	Exchange WT for mutated SPRY-PPNK into pGEX-6PI vector

2.1.5 Mammalian cell lines and growth media

HCT116: Human colorectal carcinoma cell line.

293T: Human embryonic kidney cell line which expresses the SV40 large T antigen.

hnRNPUL1-AID: Human colon cancer with AID tagged hnRNPUL1 and tetracycline inducible TIR1 protein from *Oryza Sativa*. This cell line was generated by Dr Llywelyn Griffith (Griffith 2019).

FLAG-hnRNPUL1 ΔCTD or FLAG-GFP Flp-In 293T: 293T cell line that stably expresses N-terminal FLAG-tagged hnRNPUL1 ΔCTD upon addition of tetracycline. These cell lines were generated by Dr Llywelyn Griffith (Griffith 2019).

Media for HCT116 and 293T cells: Dulbecco's Modified Eagle Medium (DMEM,

Life Technologies) with 10% FCS (v/v) and 1% penicillin/streptomycin (v/v).

Media for hnRNPUL1-AID cells: Dulbecco's Modified Eagle Medium (DMEM, Life Technologies) with 10% FCS (v/v), 1% penicillin/streptomycin (v/v), 10 μg/ml blasticidin, 150 μg/ml hygromycin, 800 μg/ml G418/neomycin and 1 μg/ml puromycin.

Media for FLAG-hnRNPUL1 ΔCTD or FLAG-GFP Flp-In 293T: cells: Dulbecco's Modified Eagle Medium (DMEM, Life Technologies) with 10% Tet-free FCS (v/v), 1% penicillin/streptomycin (v/v), 100 μg/ml zeocin, 15 μg/ml blasticidin and 100 μg/ml hygromycin.

2.1.6 Molecular Biology Kits

Application	Kit
Small scale plasmid extraction	QIAprep Spin Miniprep Kit
Medium scale plasmid extraction	QIAGEN Plasmid Plus Midi Kit
DNA gel extraction	QIAquick Gel Extraction Kit
PCR cleanup	GenElute PCR Clean-Up Kit
Cloning using Gibson assembly	NEB Gibson Assembly Cloning Kit
Site-directed mutagenesis	NEB Q5 Site-Directed Mutagenesis Kit
In vitro transcription (IVT)	ThermoFisher MEGAscript T7 Transcription Kit
In vitro capping	NEB Vaccinia Capping System
IVT cleanup kit	ThermoFisher MEGAclear Transcription Clean-Up Kit

2.1.7 Antibodies

Antibody	Source	Host	Clonality	HRP-conjugate
GST	Abcam #ab3416	Rabbit	Polyclonal	Yes
6xHis	Abcam #ab184607	Mouse	Monoclonal	Yes
FLAG	Sigma #F3165	Mouse	Monoclonal	No
FUS	Santa Cruz #sc-47711	Mouse	Monoclonal	No
EWSR1	Bethyl #A300-417	Rabbit	Polyclonal	No
TAF15	Abcam #ab134916	Rabbit	Monoclonal	No
Total PolII	MBL #MABI0601	Mouse	Monoclonal	No
Ser5 PolII	MBL #MABI0603	Mouse	Monoclonal	No
hnRNPUL1	In-house	Rabbit	Polyclonal	No
m^7G	MBL #RN016M	Mouse	Monoclonal	No
α-mouse	Promega #W4021	Goat	Polyclonal	Yes
α-rabbit	Promega #W4011	Goat	Polyclonal	Yes

2.1.8 Miscellaneous materials

RNA oligonucleotide: 5′ GGGGACUUGUCAUGCACUGACAUCACAGAAGAUGAAUCACAUCGAAUCAGCAUGUAA 3′

IVT template:

Fwd strand: 5′ ACAAGGTTAACGATAATACGACTCACTATAGG GGACTTGTCATGCACTGACATCACAGAAGATGAATCACATCGAATCAGCAT 3′

Rev strand: 5′ ATGCTGATTCGATGTGATTCATCTTCTGTGATGTCAGTGCATGACAAGTCCCCTATAGTGAGTCGTATTATCGTTAACCTTGT 3′

IVT RNA product: 5′ GGGGACUUGUCAUGCACUGACAUCACAGAAGAUGAAUCACAUCGAAUCAGCAU 3′

dsDNA kinase substrate fw 5′ CTTCAGGATGGAAG 3′

dsDNA kinase substrate rv 5′ GTGACCATCTTCCA 3′

2.1.9 DNA and RNA manipulation

5x TBE: 4.4 M Tris, 4.4 M Boric Acid, 0.01 M EDTA, pH 8.0.

6x DNA loading buffer: 0.25% Bromophenol blue, 0.25% Xylene cyanol, 30% Glycerol.

10% Denaturing PAGE: 2.5 ml 40% AccuGel, 6 M urea, 2 ml 5xTBE, 100 μl 10% APS, 20 μl TEMED (for 10 ml).

RNA gel extraction buffer: 1 M $NaCH_3COO^-$, 1 mM EDTA.

2.1.10 SDS-PAGE and western blots

4x Protein loading buffer: 200 mM Tris-HCl pH 6.8, 1% Bromophenol blue, 10% Sodium Dodecyl Sulphate (SDS), 50% Glycerol, 10% β-mercaptoethanol.

4X SDS-PAGE Stacking Gel Buffer: 0.5 M Tris pH 6.8, 0.15% SDS.

4X SDS-PAGE Resolving Gel Buffer: 1.5 M Tris pH 8.8, 0.15% SDS.

5% Stacking Gel: 6.3 ml Water, 1.2 ml 30% Acrylamide/0.8 % Bisacrylamide, 2.5 ml 4x Stacking Gel Buffer, 100 μl 10% APS, 20 μl TEMED (for 10 ml).

10% Resolving Gel: 4.1 ml Water, 3.33 ml 30% Acrylamide/0.8 % Bisacrylamide, 2.5 ml 4x Resolving Gel Buffer, 100 μl 10% APS, 20 μl TEMED (for 10 ml).

5x SDS-PAGE running buffer: 125 mM Tris, 1.25 M Glycine, 17 mM SDS.

10x TG Transfer buffer: 250 mM Tris, 1.9 M Glycine, 35 mM SDS.

1x Transfer buffer: 1x TG buffer, 20% Methanol.

1x TBST: 20 mM Tris pH 7.6, 140 mM NaCl, 0.2 % Tween-20.

5% Blocking solution: 2.5 g Powdered Milk in 50 ml (total) 1x TBST.

ECL1: 100 mM Tris-HCl pH 8.5, 400 μM p-coumaric acid, 2.5 mM Luminol.

ECL2: 100 mM Tris-HCl pH 8.5, 5.3 mM Hydrogen Peroxide.

2.1.11 Purification and reaction buffers

10x PBS (Phosphate-Buffered Saline): 1.4 M NaCl, 27 mM KCl, 100 mM Na_2HPO_4, 18 mM KH_2HPO_4, adjust pH to 7.4 with HCl.

GST lysis buffer: 1X PBS, 0.1% Tween-20.

GST elution buffer: 50 mM Tris pH 8.0, 40 mM GSH, 200 mM NaCl, 10% glycerol.

Co^{2+} Lysis buffer: 50 mM Na_3PO_4 pH 8.0, 300 mM NaCl, 10% Glycerol, 0.5% Triton X-100.

Co^{2+} Wash buffer: Lysis buffer + 6.6% Elution buffer (10 mM Imidazole).

Co^{2+} Elution buffer: 50 mM Na_3PO_4 pH 7.0, 300 mM NaCl, 150 mM Imidazole.

Column recycling buffer: 20 mM MES pH 5.0, 0.1 M NaCl.

IP Lysis buffer: 50 mM HEPES-NaOH pH7.5, 100 mM NaCl, 0.5% Triton X-100, 1 mM EDTA pH 8.0, 10% Glycerol, 1 mM DTT, protease inhibitors.

IP wash buffer: 50 mM HEPES-NaOH pH7.5, 1 M NaCl, 0.5% Triton X-100, 1 mM EDTA pH 8.0, 10% Glycerol.

IP elution buffer: 190 μl IP lysis buffer with 10 μl 5 mg/ml 3XFLAG peptide.

10x NTP binding buffer: 0.5 M Tris-HCl pH 8.0, 1 M NaCl, 0.1 M $MgCl_2$, 10 mM DTT.

5x RNA binding buffer (1): 75 mM Hepes pH 7.9, 0.5 M NaCl, 25 mM $MgCl_2$, 1 mM EDTA, 0.25% Tween-20, 50% Glycerol.

TLC running buffer: 0.4 M phosphate buffer pH 5.5 (4.4 ml 1M Na_2HPO_4, 75.6 ml 1M NaH_2PO_4 in 200 ml total.

10x Buffer for Tryptophan fluorescence measurement: 0.5 M Tris pH 8.0, 1

M NaCl, 0.1 M $MgCl_2$, 50 mM DTT.

Pulldown buffer RB100: 25 mM Hepes pH 7.5, 150 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT, 0.05% Triton-X100, 10% Glycerol.

10x Phase separation reaction buffer: 0.5 M Tris pH 8.0, 250 mM-1 M NaCl.

2.2 Methods

2.2.1 Mammalian cell culture

Cells were incubated at 37°C and 5% CO_2 and passaged twice a week by removal of the old growth medium and washing with 1X PBS. Next, cell monolayers were detached with pre-warmed 0.25% Trypsin/EDTA for 5-10 min, the enzyme was deactivated with fresh growth medium and the cells were aliquoted into new flasks or dishes.

2.2.2 DNA manipulation and cloning

Polymerase Chain Reaction (PCR): Components were mixed in a PCR test tube in the following amounts:

Component	Volume (μl)
5X Q5 Reaction Buffer	10
5X Q5 GC enhancer	10
10 mM dNTPs	2
10 μM Fwd primer	2.5
10 μM Rev primer	2.5
DNA template (10-100 ng/μl)	1
Q5 High-Fidelity Polymerase	0.5
ddH_2O	Up to 50 μl

The reactions were cycled according to the following parameters:

Step	Temperature °C	Duration	Number of cycles
Initial denaturation	98	30 s	1
Denaturation	98	10 s	35
Annealing	55-72	30 s	
Extension	72	30 s/kb	
Final extension	72	2 min	1
Hold	10	Indefinite	/

Q5 site-directed mutagenesis: The PCR and subsequent enzymatic treatment were performed according to the manufacturer's instructions for the NEB Q5 Site-Directed Mutagenesis Kit.

Colony PCR: The following components were mixed in a PCR tube. Subsequently, a small amount of bacteria from a single colony was picked with a pipette tip and resuspended in the PCR mix.

Component	Volume (μl)
10X Taq reaction buffer	2.5
10 mM dNTPs	0.5
10 μM Fwd primer	0.5
10 μM Rev primer	0.5
Taq DNA Polymerase	0.25
ddH_2O	Up to 25

The reactions were cycled according to the following parameters:

Step	Temperature (°C)	Duration	Number of cycles
Initial denaturation	95	5 min	1
Denaturation	95	30 s	30
Annealing	45-68	30 s	
Extension	68	1 min/kb	
Final extension	68	5 min	1
Hold	10	Indefinite	/

Agarose gel electrophoresis and gel extraction: Gels were prepared by dissolving 0.8-2% agarose in 0.5X TBE, depending on the size of the DNA fragments visualised. Ethidium bromide (BIO-RAD) was added to a final concentration of 10 μg/ml and gels were run in 0.5X TBE buffer at 10 V/cm for 30-40 min. DNA

fragments were visualised by UV light exposure from a BIO-RAD Chemidoc transilluminator. If extraction of DNA from the gel was needed, a thin slice of gel was cut around the DNA band and weighed. The extraction steps were carried out using the QIAquick Gel Extraction kit following the manufacturer's instructions.

Phenol:Chloroform DNA precipitation: The DNA solution to be precipitated was diluted in water to 100 μl and mixed with an equal volume of phenol:chloroform pH 6.7 by vortexing for 30 s. The tube was centrifuged at 16,200x g for 2 min and the aqueous phase was transferred to a new tube. The DNA was then mixed with a 1:10 volume of 3 M sodium acetate pH 5.3, 2.5X volume of 100% ethanol and 5 μg glycogen. The mixture was incubated at -20°C for at least 30 min and then centrifuged at 16,200x g and -20°C for 20 min. The supernatant was removed and the pellet washed with 1 ml 70% ethanol, followed by a 5 min centrifugation as before. Finally, the supernatant was discarded and after 2 min of air drying, the pellet was resuspended in a suitable volume of ultrapure water.

Gibson assembly cloning: For the Gibson assembly reaction, 100 ng of linearised vector and a 7:1 molar excess of insert were mixed with an equal volume of 2X Gibson Assembly Master Mix from the NEB Gibson Assembly Cloning Kit and the manufacturer's instructions were followed for the next steps.

***E. coli* transformation:** An aliquot of 100 μl competent DH5α cells was incubated with around 100 ng of plasmid on ice for 30 min. The cells were heat shocked at 42°C for 30 s and briefly returned on ice for 2 min. The cells were mixed with 900 μl SOC outgrowth medium (NEB) and incubated at 37°C in a thermomixer for 1 hr. The culture was centrifuged at 3,500x g for 1 min and the cell pellet was resuspended in a small volume of medium and spread on agar plates with a suitable antibiotic.

Plasmid propagation: Single colonies were picked and used to inoculate 5 or 50 ml of LB with a suitable antibiotic and incubated at 37°C overnight. Next, the cells were harvested by centrifugation at 4,000x g for 10 min and the QIAGEN kits for small or medium plasmid extraction were used, following the manufacturer's protocol.

Annealing of DNA oligonucleotides: Double stranded DNA substrate for kinase assays was generated by mixing equal volumes of 20 μM oligonucleotides resuspended in water, heating to 95°C for 5 min and gradually cooling down at room temperature over 2 hours.

2.2.3 Protein expression and purification

Bacterial overexpression: BL21 competent cells were transformed with a vector as described above. Two-three colonies were combined and used to inoculate 50 ml LB medium with antibiotic and grown overnight at 37°C. The starter culture was used to inoculate flasks with 750 ml TB medium and antibiotic at 1:100. When the cultures reached an optical density of 0.6 at 600 nm, protein overexpression was induced with IPTG. Next, the cells were either incubated at 37°C for 3 hr or at 18°C overnight. Bacterial pellets were collected by centrifugation at 4,000x g and 4°C for 30 min.

Overexpression in mammalian 293T cells: Cells were seeded on 10 cm plates 24 hr before transfection, in sufficient amounts to give 60-70% confluency. For one plate, 1.8 ml pre-warmed DMEM was thoroughly mixed with 12 μg DNA and polyethylene (PEI) at a final concentration of 3.5 μg/μl. After 15 min, the medium was added drop-wise onto the cells. Cultures were harvested 48 hr post transfection.

FLAG-tagged protein purification: Typically, 10x10 cm dishes were used to overexpress each protein. 293T cell pellets were lysed in 5-10X volume of IP lysis buffer. Nuclei were sheared by aspiration through a 26 gauge needle 5 times. For purifying full-length hnRNPUL1-ΔWA and hnRNPU proteins, 500 U benzonase were added to the lysates and incubated for 30 min on a thermomixer at 37°C and 300x g. Next, the lysates were centrifuged at 16,200x g for 5 min and the supernatants added to 100 μl pre-washed FLAG-agarose beads. The mixture was incubated at 4°C on a rotating platform for 2 hr. The unbound material was then discarded and the beads washed with 2X 1 ml IP lysis buffer, followed by treatment with 4 μg RNAse A at 37°C for 30 min. Next, the beads were washed with 3X 1 ml IP wash buffer (with 5 min incubations on a rotating wheel in between) and finally with 1 ml IP lysis buffer. Proteins were eluted from the beads with 300 μl IP elution

buffer by overnight incubation at 4°C on a rotating platform.

Hexahistidine-tagged protein purification: Typically, 5 g bacterial pellets were resuspended in 50 ml Co^{2+} lysis buffer and sonicated for 10 rounds of 30 s on/30s off at 80% amplitude with a Fisherbrand Model 120 Sonic Dismembrator. The lysates were cleared by centrifugation at 20,000x g and 4°C for 30 min. The supernatant was further treated with 5 μg RNAse A for 20 min at 37°C, then loaded on a 5 ml Talon Superflow Cobalt column. The beads were washed with 1 column volumes (CV) of Co^{2+} lysis buffer, then the column was loaded with 1 CV lysis buffer and 5 μg RNAse A and incubated for 30 min at room temperature. Next, the column was washed with 6 CV Co^{2+} wash buffer and the protein was eluted with 2 CV Co^{2+} elution buffer. The excess imidazole was removed from the column with 1 CV column recycling buffer and 2-3 CV of ultrapure water. The column was stored in 20% ethanol.

SDS-PAGE: Protein samples were prepared by mixing with 4x protein loading buffer up to a total volume of 20 μl. The samples were further denatured by incubating at 95°C for 5 min and loaded on 8-12% gels, depending on the size of the proteins. Samples were run in BIO-RAD electrophoresis chambers at 30 mA per gel for approximately 1 hr. Gels were either stained with InstantBlue Protein Stain (Merk) for 30 min and visualised in a a BIO-RAD Chemidoc transilluminator or processed further for western blotting.

Western blotting: Gels were stacked on top of a nitrocellulose membrane soaked in transfer buffer, and the ensemble was sandwiched between 6 layers of pre-soaked WypAll paper on each side. Proteins were transferred via a BIO-RAD fast transfer machine programmed to run a current of 25 V, 1.3 mA for 20 min. The membrane was stained with Ponceau dye, cut into relevant strips and blocked for 1 hr in 5% blocking solution. Membrane strips were next incubated for 1 hr with primary anitbodies diluted in blocking solution to the specified concentration. Blots were washed in 10-15 ml TBST for 3X 30 s and 3X 5 min, and, if necessary, incubated with secondary antibodies for 1 hr and washed as before. Finally, blots were coated with a 1:1 solution of ECL1 and ECL2 for 30 s and developed in a BIO-RAD Chemidoc transilluminator for an appropriate amount of time.

2.2.4 Biochemistry assays

2.2.4.1 5′ ^{32}P RNA labelling and cleanup: A short (35 nt) RNA oligonucleotide with a 5′ OH- group was 5′-end labelled with T4 PNK. The labelling reaction consisted of 1 μl (10 pmol) RNA, 2 μl T4 PNK 10X buffer (NEB), 2 μl (10 μCi/μl) ^{32}P γ-ATP (Perkin Elmer), 1 μl (10 U) T4 PNK (NEB), 14 μl ultrapure water. The reaction was carried out for 1 hr at 37°C, then the RNA was mixed with 2X RNA loading buffer (NEB) and loaded on a 10% denaturing polyacrylamide gel. The sample was run in 0.5X TBE buffer in a BIO-RAD electrophoresis chamber at 10 V/cm for 3-4 hr, until the lower dye front was close to the bottom of the gel. Next, the lower dye front was cut with a sharp scalpel (to remove free ^{32}P γ-ATP) and Fiducial markers were places in the corners. Afterwards, the gel was wrapped in cling film and exposed on a phosphorimager screen for 10 s, the image was printed and the areas corresponding to the markers and the RNA were cut out. The stencil was placed over the gel and the band corresponding to the RNA was cut out and crushed in a separate tube, then mixed with 400 μl RNA gel extraction buffer and incubated on a rotating platform shaking vigorously overnight. In the morning, the tube was spun briefly, the supernatant was passed through a Spin-X centrifuge tube filter (Costar) and mixed with 1 ml 100% ethanol and 5 μg glycogen. After 30 min-1 hr at -20°C, the RNA was centrifuged at 16,200x g for 30 min. Next, the supernatant was discarded, the pellet was washed with 1 ml 75% ethanol, centrifuged at 16,200x g for 7 min, air dried after removal of the supernatant and finally resuspended in 100 μl nuclease-free water.

2.2.4.2 *In vitro* transcription and cleanup: DNA templates were assembled by annealing two complementary DNA oligonucleotides containing a T7 RNA polymerase promoter after resuspension in water. Equal volumes (4 μl 100 μM) of each oligonucleotide were mixed, heated to 95°C for 2 min and gradually cooled to 25°C over 45 min in a thermocycler. The IVT reactions were carried out with ThermoFisher MEGAscript T7 Transcription Kit, following the kit's instructions. The resulting RNA was cleaned up using the ThermoFisher MEGAclear Transcription Clean-Up Kit.

2.2.4.3 RNA extraction from total cell lysate: Total cell lysates were mixed with 3 volumes of TRIzol-LS (Invitrogen), homogenized and incubated at room temperature for 10 min. Chloroform (200 μl) were added for 750 μl TRIzol-LS, the tubes were shaken vigorously and incubated at room temperature for 10-15 min. Samples were centrifuged at 12,000x g and 4°C for 15 min, then the aqueous phase was transferred to a tube containing 5 μg glycogen. RNA was precipitated with an equal volume of isopropanol, followed by a 10 min room temperature incubation and a 10 min centrifugation at 12,000x g and 4°C. The pellet was washed with 1 ml 75% ethanol and spun at 7,500x g and 4°C for 5 min. Air-dried pellets were resuspended in an appropriate volume.

2.2.4.4 ATP UV crosslinking: A roughly equal concentration of each protein (0.3 μM) was mixed with 1.5 μl 10X NTP binding buffer and 1 μl (5 μCi) ^{32}P γ-ATP in 15 μl total volume. The reactions were incubated on ice for 15 min and then UV-crosslinked on ice for 30 min at 254 nm. Crosslinked proteins were separated on an 8% gel by SDS-PAGE, stained and dried at 80°C for 30 min. Dried gels were exposed overnight on a phosphorimager screen, which was developed on a Typhoon FLA 7000 laser scanner.

2.2.4.5 NTP competition: Full-length hnRNPUL1 (1.6 μM) was mixed with 1.5 μl 10X NTP binding buffer, 1 μl (5 μCi) ^{32}P γ-ATP and AMP-PNP or GMP-PNP non-hydrolysable competitors at final concentrations of 0.55 μM, 5.5 μM and 27.7 μM in a 20 μl reaction volume. Reactions were mixed for 15 min on ice and UV crosslinked for 30 min. Complexes were separated on a 10% gel by SDS-PAGE and the gel was stained, dried and exposed as before.

2.2.4.6 RNA UV crosslinking: Proteins (0.5 μM) were mixed with 4 μl 5X RNA binding buffer, 1 μl 5′-labelled RNA oligonucleotide and 1 μl (8 U) RNAse inhibitor (SLS) in 20 μl total volume and incubated on ice for 10 min, before being UV crosslinked on ice for 15 min. Subsequently, the proteins were separated on a 10% gel by SDS-PAGE and the gel was stained, dried and exposed as before.

2.2.4.7 ATP/ADP competition for RNA binding: hnRNPUL1 ΔCTD (0.5 μM) was incubated with 2 μl 10X NTP binding buffer, 1 μl 5′ ^{32}P-labelled RNA oligonucleotide, 1 μl (8 U) RNAse inhibitor and ATP or ADP competitor at 10, 100 or 1,000 μM. Reactions were incubated at 37°C for 10 min and UV crosslinked on ice for 15 min. Crosslinked complexes were separated on a 10% gel by SDS-PAGE and the gel was stained, dried and exposed as before.

2.2.4.8 GTP/m^{7}G cap competition: hnRNPUL1 ΔCTD (1 μM) was mixed with 2 μl 10X NTP binding buffer, 1 μl (0.2 pmol) ^{32}P α-GTP and GTP or m^{7}G cap analog competitors at 0.1 μM, 1 μM and 10 μM in a 20 μl reaction. Samples were incubated at 37°C for 15 min and UV crosslinked on ice for 30 min. Crosslinked proteins were separated on a 10% gel by SDS-PAGE and the gel was stained, dried and exposed as before.

2.2.4.9 Thin-layer chromatography (TLC): ATP hydrolysis reactions were carried out with 0.5 μM each protein, 2 μl 10X NTP binding buffer, 1 μl (0.1 μCi) ^{32}P γ-ATP and (optionally) 1 μl 10 μM RNA oligonucleotide in 20 μl total volume. The reactions were incubated at 37°C for 30 min and quenched with 20 μl 0.5 M EDTA and 10 μl xylene cyanol. Additionally, 1 μl (20 μg) Proteinase K was added, along with 4 μl 1 M Tris pH 7.5 and the proteins were digested for 30 min at 37°C. T4 PNK and apyrase positive control reactions were carried out differently. The T4 PNK reaction consisted of: 5 μl 10X T4 reaction buffer (NEB), 1 μl 10 μM RNA oligonucleotide, 1 μl 0.1 μCi ^{32}P γ-ATP and 1 μl (10 U) T4 PNK (NEB). The apyrase control consisted of 5 μl apyrase, 5 μl 10X NTP binding buffer and 1 μl 0.1 μCi ^{32}P γ-ATP. Both reactions were mixed in 50 μl total volume and incubated at 37°C for 30 min, then quenched with 50 μl 0.5 M EDTA and 20 μl xylene cyanol. PEI-cellulose plates were pre-run and dried. Next, equal amounts of radioactivity were spotted on the PEI plates 1 cm from the bottom edge. The plates were run in a 0.5 cm layer of TLC running buffer, dried and exposed on a phosphorimager screen for 1 hr and developed as before.

2.2.4.10 Kinase assay: hnRNPU/UL1 proteins (0.5 μM) were mixed with 2 μl 10X NTP binding buffer, 1 μl 10 μM dsDNA/RNA oligonucleotides with 5′ OH-

groups and 0.3 pmol ^{32}P γ-ATP in a 20 μl total volume. The T4 PNK control reaction contained 1 μl (10 U) T4 PNK enzyme and 2 μl 10X T4 PNK buffer (NEB) instead. All reactions were incubated at 37°C for 30 min, then mixed with 2X RNA loading dye (NEB) and loaded on a 12% denaturing gel. Samples were run in 0.5X TBE at 10 V/cm for 2-3 hr until the lower dye front was close to the bottom. The gel was wrapped in cling film and exposed on a phosphorimager screen for 1 hr. Images were collected on a Typhoon FLA 7000 laser scanner.

2.2.4.11 Tryptophan fluorescence quenching measurement: hnRNPUL1 SPRY-PPNK proteins were buffer-exchanged into 50 mM Tris pH 8.0, 100 mM NaCl by serial concentration and dilution with Pierce protein concentrators (10K molecular weight cut off) to dilute the 3XFLAG peptide used for elution to a calculated final concentration of 0.025 ng/μl. Samples were prepared by mixing each hnRNPUL1 construct (0.5 μM final concentration) with 50 μl 10x Buffer for Tryptophan fluorescence measurement in 0.5 ml total volume. A Cary Eclipse fluorometer (Agilent) was used to measure tryptophan fluorescence, with an excitation wavelength of 280 nm and following the emission spectra between 300-400 nm with a spectral resolution of 5 nm and the photomultiplier set to high. First, the reaction buffer with the same concentration of 3XFLAG peptide as in the protein samples was used to determine the background fluorescence, which was minimal, but was subtracted from all subsequent measurements. Each protein was then measured in an apo form, then ATP or 5′-triphosphorylated RNA (generated by *in vitro* transcription) were incrementally titrated into the protein solution. The ligand concentrations tested (accounting for dilution) were: 10, 25, 50, 74, 99, 246, 491, 735, 978 nM and 2.43 and 4.84 μM. After each ligand addition, the protein was incubated for 1 min at 37°C and the emission spectrum recorded in triplicate. Ligand stocks were also prewarmed to 37°C. Five values around the emission peak (347-352 nm) were averaged for each spectrum and the fluorescence quenching (ΔFluorescence) was calculated for each set of triplicate measurements and plotted with GraphPad Prism. Curves were fitted corresponding to a One site– Specific binding equation.

2.2.4.12 Pulldowns: Around 0.5 g cell paste was used to extract bait proteins, while 0.1-0.2 g cells were used for the control baits (GST and GFP). Cells were

resuspended in 1 ml GST lysis buffer or Co^{2+} lysis buffer and sonicated for 5 cycles of 3 s on/3 s off at 25% amplitude with a Fisherbrand Model 120 Sonic Dismembrator. Next, lysates were cleared by centrifugation at 16,200x *g* for 5 min and were incubated with 50-150 μl GSH or Talon beads (depending on the experiment) for 30 min at 4°C. The beads were washed 3X with 1 ml GST lysis buffer or Co^{2+} wash buffer, then split into several tubes (depending on number of conditions), mixed with purified prey proteins or total cell lysates (equalised by expression of transfected hnRNPUL1 constructs) in Pulldown buffer RB100 for a total volume of 200 μl. Reactions were incubated at 37°C for 30 min, then the supernatants were discarded and the beads washed with 2x200 μl PBS and eluted with 2 bead volumes of GST or Co^{2+} elution buffers. Eluates were analysed by SDS-PAGE and western blotting.

2.2.4.13 NO-CAP RIP: Flp-In 293T cell lines expressing FLAG-tagged GFP or hnRNPUL1ΔCTD were induced with 0.01 μg/ml tetracycline for 48 hr. Two 10cm dishes were used for each condition and one dish was transfected with FLAG-CBP20 as a positive control. Cell pellets were lysed with 400 μl IP lysis buffer supplemented with 5 U/ml Turbo DNAse and incubated at 37°C for 30 min. Next, lysates were spun down at 16,200x *g* and 4°C for 5 min. 10% of each clarified supernatant was used for total RNA extraction, while the rest was incubated with 30 μl FLAG agarose at 4°C for 2 hr. The beads were washed with 3x1 ml IP lysis buffer with a 5 min incubation in between. Proteins were eluted with 60 μl IP elution buffer for 2 hr. 10 μl eluate were checked by SDS-PAGE, while the rest of the solution was treated with 1 μl Proteinase K for 20 min at 37°C. The resulting mixture was subjected to RNA extraction. Total RNA pellets were resuspended in 24 μl ultrapure water, while the RIP-ed RNA was resuspended in 12 μl. Yields were measured with a Qubit system and 100 ng RNA from hnRNPUL1 ΔCTD input and RIP, as well as CBP20 RIP, were spotted on a hybond nylon membrane, air dried and UV crosslinked 70 mJ/cm^2. The blot was blocked with 5% blocking solution and processed like a western blot. Additionally, RNA obtained from GFP and hnRNPUL1 ΔCTD total extracts and RIPs were checked with a bioanalyzer, according to the manufacturer's instructions.

2.2.4.14 Phase separation assays: All assays were carried out by mixing the specified protein concentration with the specified reaction buffer and 5% PEG 8,000 (unless stated otherwise). Small aliquots of each condition (2 μl) were spotted on a glass slide, covered with a cover slip and sealed to prevent drying. Slides were incubated at room temperature to allow droplets to settle on the surface of the slide and examined with a Leika fluorescent microscope. 5% 1,6-hexanediol was added either at the beginning of the reaction or to pre-formed droplets as indicated. CDK7 treatment included addition of 0.2 μg CDK7 complex and 0.1 mM ATP or AMP-PNP to the reaction mixture.

2.2.4.15 CDK7 kinase assay: CDK7 complex (0.7 μg) was mixed with 2.5 μl 10X NTP binding buffer, 1 μl 2.5 mM ATP, 1 μl (10 μCi) ^{32}P γ-ATP and protein substrates PolII CTD (20 pmol), hnRNPUL1 CTD (220 pmol) or hnRNPUL1 PLD (440 pmol) in 25 μl total volume. Reactions were incubated at room temperature for 1 hr and separated by SDS-PAGE, followed by gel staining, drying and exposing as before.

2.2.5 Graphics generation

Structural models were analysed and molecular graphics rendered using UCSF ChimeraX (Pettersen et al. 2021). Schematics were drawn using BioRender.com.

3 Chapter 3: hnRNPUL1 is an inactive polynucleotide kinase that recognises free nucleic acid 5′ ends

The initial study that identified hnRNPUL1 reported its involvement in nucleocytoplasmic mRNA export, noting that overexpression of this protein overcame the viral-induced block of mRNA export in adenovirus type 5-infected cells (Gabler et al. 1998). Additionally, the authors described the presence of a proposed central NTP-binding region and suggested that it would serve as a GTP binding module. Bachi et al. (2000) also implicated hnRNPUL1 in mRNA export due to its association with the nuclear export factor NXF1. Further characterised roles for hnRNPUL1 have been in the regulation of gene expression through association with the bromodomain-containing protein BRD7 (Kzhyshkowska et al. 2003) and in contribution to the DNA damage response interdependently with hnRNPUL2 through the homologous recombination pathway (Polo et al. 2012).

With the exception of TDP43, hnRNPU and hnRNPUL2, hnRNPUL1 is distinguished from other hnRNP proteins by virtue of a well-folded region comprising centrally located SPRY and NTP-binding domains. The NTP domain harbors characteristic Walker A and B motifs (Walker et al. 1982). However, there are no studies to date that explore the biochemical function of the central domains of hnRNPUL1, particularly the proposed NTP-binding region. Nozawa et al. (2017) reported that the equivalent region of hnRNPU serves as an ATPase Associated with diverse cellular Activities (AAA+ ATPase) and that it undergoes ATP-dependent oligomerisation. However, there was no clear evidence presented for the protein adopting a fold typical of AAA+ ATPases, rather, the data indicated the NTP domain of hnRNPU adopted a fold similar to the kinase domain from human polynucleotide kinase phosphatase. Consistent with this prediction, a much earlier analysis of P-loop NTPases classified hnRNPU as a P-loop kinase by virtue of a 5-4-1-2-3 strand arrangement in the central β-sheet core and a helical "lid" region at the top of the structure (Leipe et al. 2003).

In this chapter, we set out to probe the function of the central folded SPRY and

NTP-binding domains, aided by structural model predictions generated by Phyre2 and later corroborated by AlphaFold. The models uncovered an extensive interface between the SPRY and NTP-binding regions, which appeared to form a single globular domain. Based on the 3D models, we identified a high degree of structural homology with the kinase domain of the enzyme polynucleotide kinase phosphatase, which guided the subsequent functional assays. UV crosslinking experiments revealed that the NTP-binding region of hnRNPUL1 can bind ATP and certain nucleic acid substrates similarly to the bacteriophage T4 PNK enzyme and that it specifically recognises free RNA 5′ ends and DNA 5′ overhangs.

3.1 AlphaFold predicts structural model for hnRNPUL1

We began investigating the function of hnRNPUL1 by generating model predictions for the SPRY and nucleotide binding domains using the Phyre2 portal, which uses homology recognition to make protein folding predictions. The template-based modelling method, although powerful, can create biases in the output as it models a sequence of unknown structure on a homologous sequence with a known fold (Kelley et al. 2015). Therefore, Phyre2 is restricted to structure prediction where there are pre-determined models available for proteins of a similar sequence. Another disadvantage is that Phyre2 cannot predict tertiary structure interactions well, especially when there are homologous structures available for different domains of a protein with no mutual overlap (Kelley et al. 2015).

Due to these limitations, we favoured the 3D model prediction for hnRNPUL1 which became available after the release of AlphaFold v2.0 (Figure 3.1A) (Jumper et al. 2021, Varadi et al. 2021). We downloaded the model from the AlphaFold Protein Structure Database and generated the molecular graphics using the software UCSF ChimeraX (Pettersen et al. 2021). Unlike the template-based modelling methods, AlphaFold employs a neural network trained on evolutionary constraints, as well as residue and protein physics and geometry constraints. The advantages offered by this approach include near-experimental accuracy in Cα coordinate prediction, improved side chain rotamer prediction and superior performance with tertiary structure elements and very large proteins (Jumper et al. 2021).

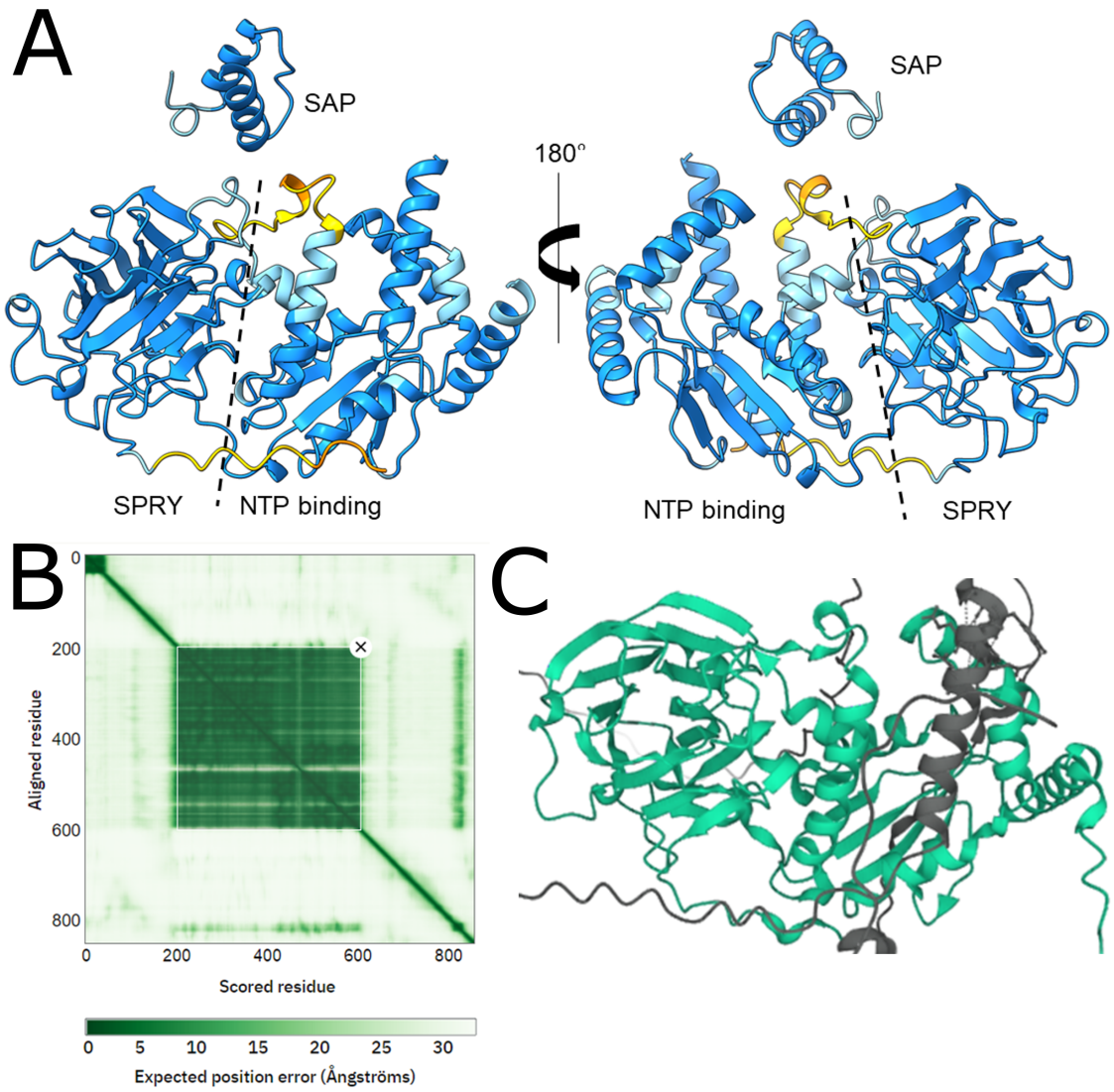

Figure 3.1: AlphaFold model prediction for hnRNPUL1.
A) AlphaFold prediction of the structured parts of human hnRNPUL1 protein, colour-coded based on the confidence score at each residue position (light and dark blue represent >70% and >90% confidence, respectively, while orange and yellow regions represent <70% and <50% confidence predictions, respectively). **B)** Plot of Predicted Aligned Error for the hnRNPUL1 model. Dark green represents a small expected error in the distance between a pair of residues, while light green represents a large expected error in the distance between a pair of residues. The dark central region is selected and the corresponding residue pairs are coloured in cyan in **C)**, with the folded structure overlapping the SPRY-NTP domains shown in **A)** and grey regions representing disordered elements.

Upon investigation of the AlphaFold generated structural prediction of full-length human hnRNPUL1, two aspects were immediately apparent: only the SAP, SPRY and nucleotide binding regions were predicted to adopt a well folded structure, and the SPRY and nucleotide binding regions seemed to form a single domain rather than two separate ones (Figure 3.1A).

An additional feature of AlphaFold is the Predicted Aligned Error (PAE), which informs on the expected error in the distance between a pair of residues (Varadi et al. 2021) and is a useful indication of the confidence in the relative position of different domains within a single model (Figure 3.1B). In the case of hnRNPUL1, there is a central region of low predicted aligned error which corresponds to the SPRY-nucleotide binding domain (Figure 3.1C). This provides further evidence for the close spacial association between the two central SPRY and nucleotide binding domain regions. There is an additional region of low expected position error in the upper left hand corner of the PAE plot (Figure 3.1B), corresponding to the small SAP domain, formed of two helices and two loops.

3.2 hnRNPUL1 contains a folded core formed of the SPRY-NTP binding regions

The novel structural information predicted by Alphafold, suggesting a strong association between the SPRY and nucleotide binding regions, as well as the unusual nature of these well folded regions present within hnRNPUL1 compared to most other hnRNP proteins prompted us to investigate this seeming protein "core" in more detail. A close-up of the core region supports the notion of a single globular domain (Figure 3.2A). Consistent with the earlier Phyre2 homology-based prediction, the AlphaFold model contains a channel spanning the NTP binding region, lined with positively charged residues.

Further analysis of the interface between the two halves of the core revealed an extensive network of residue pairs with complementary electrostatic charge, comprising basic residues on the NTP binding side and acidic residues on the SPRY side (Figure 3.2B). This provides further evidence for the protein core acting as a single functional unit.

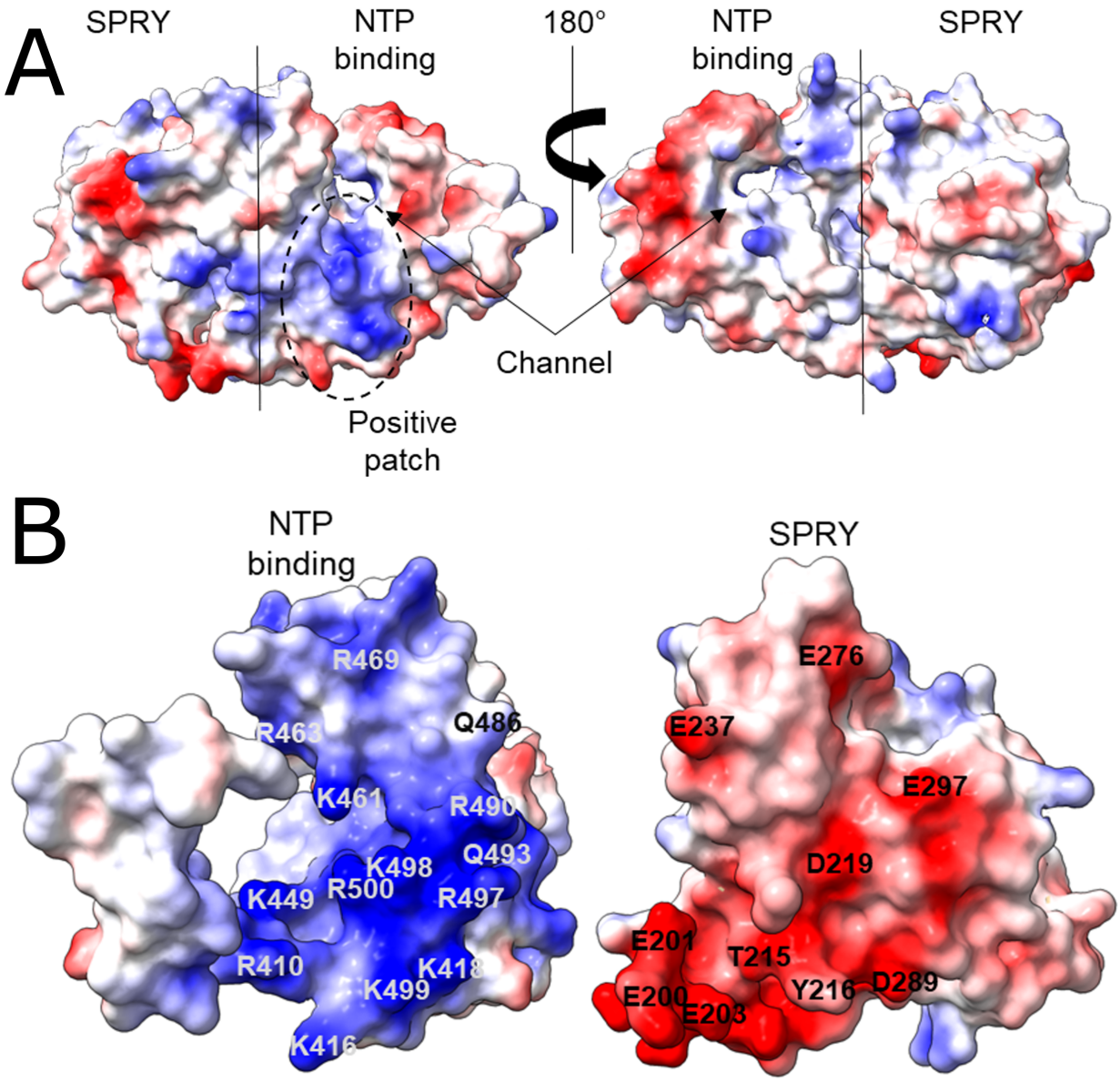

Figure 3.2: The SPRY and NTP binding regions form a tight protein core.
A) Close-up of the well-folded core spanning SPRY and the NTP binding region predicted by AlphaFold to form a single domain. The line divides the core into two halves corresponding to the SPRY and NTP binding regions based on amino acid sequence. **B)** The interface between the SPRY and NTP binding halves comprises complementary electrostatic interactions. Pairs of residues involved in these interactions are labelled. Red and blue colours denote negative and positive surface charges, respectively.

3.3 Structural homology with mammalian PNKP (mPNKP) offers functional clues

In order to understand the function of the protein core, we turned our attention to finding homologues of known structure to all or a part of the core, using both PDBeFold and Phyre2. No homologues were detected for the entire core of hnRNPUL1. However, Phyre2 modelled 78% of the structure of the NTP binding region at >90% accuracy based on structural homology with the kinase domain of mammalian polynucleotide kinase phosphatase (mPNKP) enzyme. Comparison between the Phyre2 prediction and the unbiased AlphaFold model revealed extensive overlap, with a root mean square deviation (RMSD) over 86 Cα atoms of 1.12 Å (Figure 3.3A). The overall RMSD for the two models was 16.14 Å, taking into account the low confidence predictions of Phyre2 for flexible loops and portions lacking a direct counterpart in PNKP, such as regions predicted by AlphaFold to form the interface with SPRY.

The good agreement between the structured regions of the Alphafold and Phyre2 models gave us confidence in the identified homology to the kinase domain of PNKP (Figure 3.3B). Superimposing the AlphaFold model onto the apo form of PNKP revealed an RMSD of 1.1 Å over 82 Cα pairs, corresponding to the structured regions, but not the flexible loops. Key protein features of hnRNPUL1 and PNKP were superimposable, such as the Walker A and B motifs, which are crucial for NTP and co-factor stabilisation. Owing to the extensive structural homology between the two proteins and subsequent biochemical analysis, we chose to name the NTP binding region of hnRNPUL1 pseudo-PNK, or PPNK.

We observed only one region of substantial difference between hnRNPUL1 and PNKP, consisting of an extended helix-loop region in hnRNPUL1 (α2-loop-α3) (Figure 3.3C). The extended feature of hnRNPUL1 is similar to the extended structure found in bacteriophage T4 PNK, where it serves as a double strand discriminator for selecting single stranded DNA substrates (Eastberg et al. 2004).

In light of the structural homology between hnRNPUL1 and mPNKP, we next tested if hnRNPUL1 could bind similar substrates to those of PNKP and whether it bound

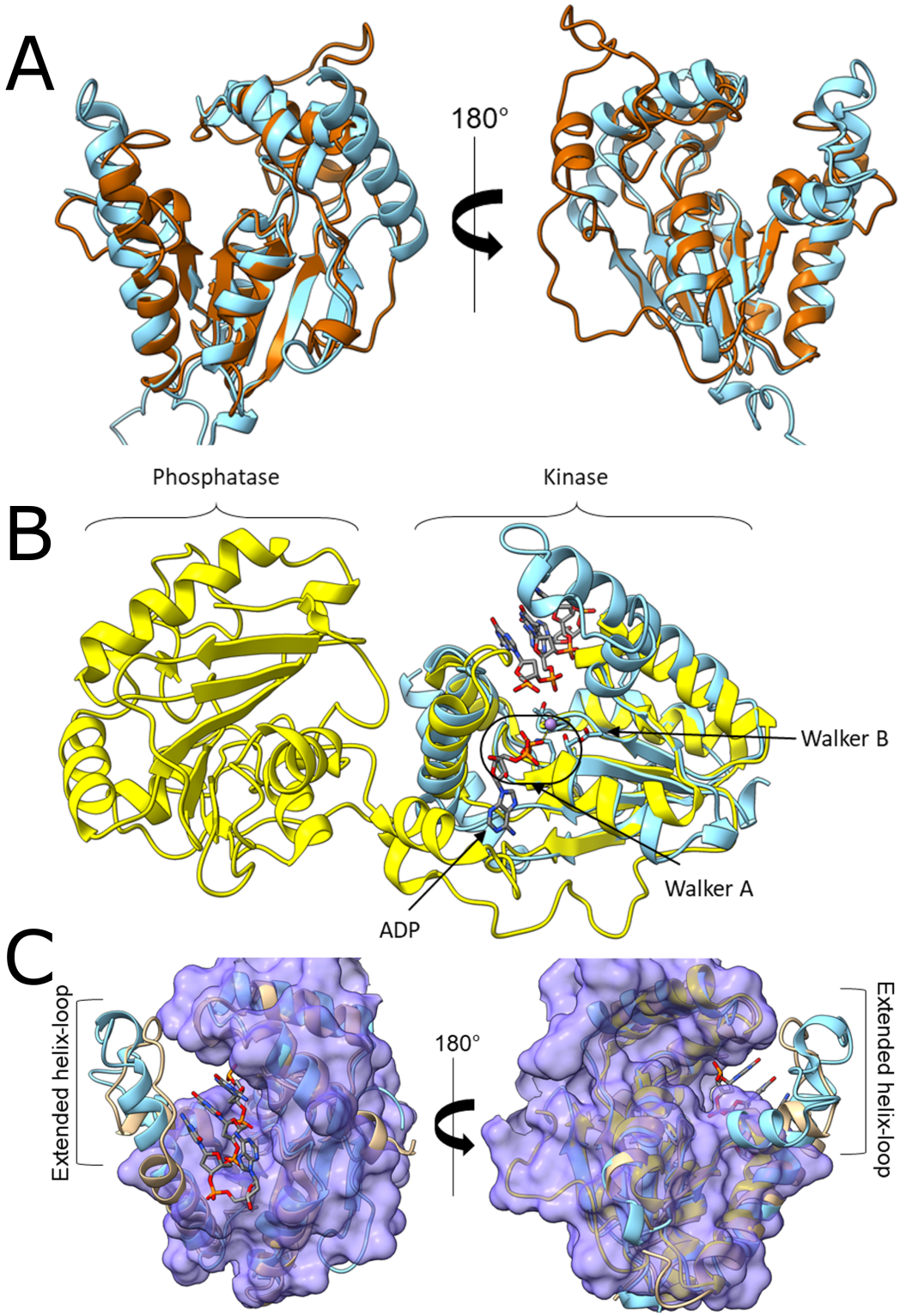

Figure 3.3: hnRNPUL1 is a structural homologue of mammalian PNKP. **A)** Superposition of the AlphaFold (light blue) and Phyre2 (brown) models of hnRNPUL1 PPNK. **B)** Superposition of the AlphaFold model for hnRNPUL1 PPNK (light blue) onto the crystal structure of mPNKP (yellow) bound to catalysis products ADP and 5′P DNA (PDB: 3zvn). The Walker A and Walker B motifs of hnRNPUL1 are highlighted. Mn^{2+} is shown as a purple sphere. **C)** Superposition of the AlphaFold prediction for hnRNPUL1 PPNK (light blue ribbon) onto the crystal structures of bacteriophage T4 PNK (beige ribbon) (PDB: 1rc8) and mPNKP (purple surface contour). The second DNA strand bound to PNKP is omitted for clarity.

them in a similar manner.

3.3.1 hnRNPUL1 binds NTPs, with a preference for ATP

We first sought to confirm whether the PPNK region had NTP binding potential, as it was first predicted to serve as a GTP binder (Gabler et al. 1998). Although GTP binding has not been shown experimentally for hnRNPUL1, interestingly, hnRNPU was recently reported to have a weak ATP hydrolysis activity (0.15 nmol ATP hydrolyzed x $nmol^{-1}$ of protein x min^{-1}), but did not show GTP hydrolysis (Nozawa et al. 2017).

Initial colorimetric assays were carried out to measure ATP or GTP hydrolysis by hnRNPUL1 as described by Nozawa et al. (2017). Both the SPRY-PPNK core expressed in bacteria and the full-length hnRNPUL1 expressed in mammalian cells were tested in these assays with no success. Importantly, bacterial expression of even the structured parts of hnRNPUL1 yielded mostly insoluble material which could not be refolded properly after solubilisation with 6 M guanidine hydrochloride. Addition of solubility tags improved the yield, but the presence of traces of bacterial chaperone contaminants, such as HSP70 and GroEL, produced misleading and variable results in the colorimetric measurements. Complete removal of contaminating ATPases in our hnRNPUL1 preparations yielded no enzymatic activity above background (data not shown). Therefore, we chose a different and more sensitive approach to determine NTP binding using UV crosslinking to radio-labelled ATP. Additionally, we resorted to expressing all protein constructs (unless otherwise specified) in mammalian cells, which produced high quality protein preparations.

For the UV crosslinking assays, we employed versions of hnRNPUL1 mutated in the Walker A and B motifs (ΔWA= ^{428}GlyxxxxAlaAlaAla^{435} and ΔWB= ^{502}hhhAla^{505}), as well as truncations of the protein lacking the C-terminus (ΔCTD) or spanning just the PPNK portion (Figure 3.4A). These proteins, including GFP used as a negative control, were UV crosslinked to ATP labelled with ^{32}P on the α position instead of the γ in order to avoid loss of signal through ATP hydrolysis. The assay confirmed that hnRNPUL1 could bind ATP and that deletion of the Walker A motif severely disrupted the binding, while mutation of the Walker B region enhanced

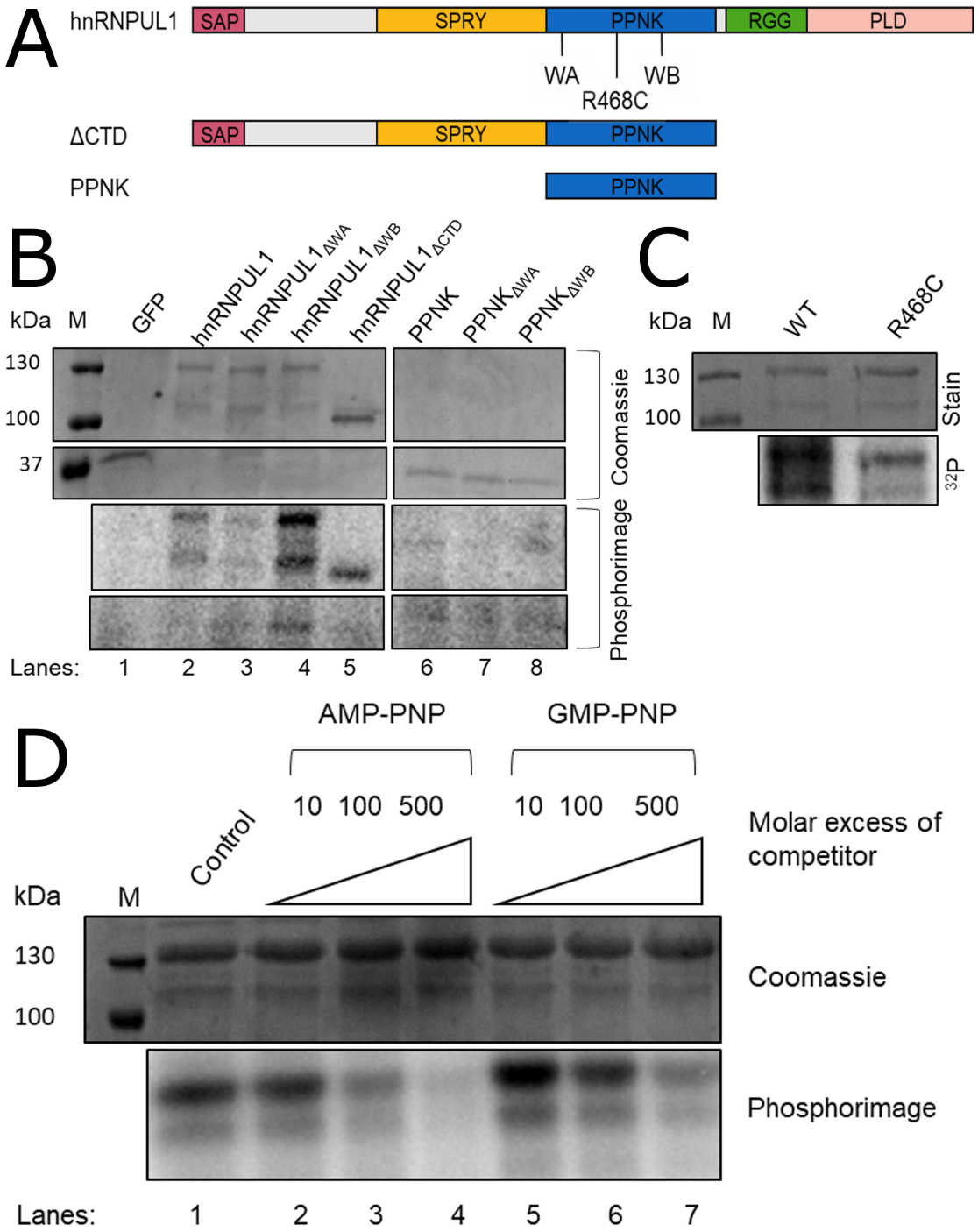

Figure 3.4: Nucleotide triphosphates are binding substrates for hnRNPUL1.

A) Schematic of hnRNPUL1 and the truncations and mutants used in the assays below. The proteins were purified as FLAG-tagged versions after transient transfection in HEK-293T cells. **B)** UV crosslinking of hnRNPUL1 constructs to α-^{32}P ATP, followed by separation by SDS-PAGE and auto radiography. **C)** UV crosslinking of full-length hnRNPUL1 WT and R468C mutant protein associated with ALS disease to α-^{32}P ATP. **D)** Nucleotide competition assay crosslinking wild-type hnRNPUL1 to a mixture of α-^{32}P ATP and unlabelled non-hydrolysable ATP or GTP analogues (AMP-PNP and GMP-PNP, respectively). Protein-NTP complexes were separated by SDS-PAGE and auto radiographed. The small increase in lane 5 can be attributed to pipetting errors.

ATP binding (Figure 3.4B). We excluded the possibility that ATP binding could be achieved non-specifically via electrostatic interactions with the positively charged arginine residues in the RGG box as the ΔCTD mutant bound ATP as avidly as wild-type. However, the PPNK region on its own did not bind ATP, likely due to misfolding caused by the absence of the SPRY half of the core. Additionally, we observed that the ALS-associated mutation R468C also drastically perturbed ATP binding compared to WT protein (Figure 3.4C). The mutation is located at the protein surface, on the ATP-binding side of the PPNK pocket. Therefore, loss of ATP binding could be caused by disruption of stabilising interactions or local misfolding around the ATP site.

We interrogated the nucleotide preference of hnRNPUL1 with a competition assay between ATP and GTP. Full-length hnRNPUL1 was UV crosslinked to a mixture of α-^{32}P ATP and increasing amounts of unlabelled non-hydrolysable ATP or GTP analogues. There was a noticeable preference for ATP binding over GTP, with 50% of the ^{32}P signal being displaced from the gel by ATP and only 10% by GTP at 100-fold molar excess of competitor (Figure 3.4D). At 500-fold excess, ATP displaced almost 75% of the signal while GTP reduced it by 50%. There was very little or no reduction in signal at 10-fold excess.

It is worth noting that we were unable to reproduce the ATPase activity reported for hnRNPU using the same colorimetric approach employed by Nozawa et al. (2017). In our hands, the isolated PPNK region of hnRNPU failed to bind ATP in UV crosslinking experiments in the absence of the SPRY half of the core, consistent with our hnRNPUL1 results (Figure 3.4B). In light of the extensive interface between SPRY and PPNK identified earlier (Figure 3.2), the most likely explanation is that the absence of half of the core causes misfolding of the PPNK region and subsequent loss of function. The only traces of ATPase activity in our protein preparations came from bacterial contaminants such as HSP70 and GroEL and were removed by harsher purification conditions, including buffers supplemented with 6 M guanidine hydrochloride or ATP incubations at 37°C.

Overall, these results indicate that hnRNPUL1 can bind to both ATP and GTP and shows a preference for ATP, but it is uncertain whether it has any NTP hydrolysis

activity. The effect of the ΔWA mutation on ATP binding is consistent with that region's role in stabilising the NTP backbone, but it could also be due to protein misfolding caused by the triple alanine mutation. On the other hand, the enhanced binding of the ΔWB mutant could be due to a more favourable environment in the channel after loss of the ^{505}Glu negative charge (Figure 3.3B) or because of an impaired substrate turnover cycle caused by the mutation.

3.3.2 hnRNPUL1 can bind 5′ mono- and triphosphorylated RNA

Next, we investigated if hnRNPUL1 could bind nucleic acids via its PPNK region independently of the RGG box. To that end, we used the SPRY-PPNK core expressed on its own (Figure 3.5A) and UV crosslinked it to short ^{32}P-labelled RNAs carrying either one or three phosphate groups on the 5′ end. We also employed mutant versions of two conserved residues that help stabilise and position the nucleic acid substrate in PNKP (T423 and R432) (Figure 3.5B). These were mutated to alanines in hnRNPUL1 to test their contribution to RNA binding (T507A and R516A), alongside the aforementioned mutations ΔWA, ΔWB and ALS-associated mutation R468C. Both types of RNA tested had unstructured 5′ end regions spanning at least 7 bases and starting with three guanine nucleotides, as that appeared to be the ideal sequence for a T4 PNK substrate (Lillehaug & Kleppe 1975). The monophosphorylated oligonucleotide was initially employed due to ease of access, despite not being a long-lived RNA species found in the cell, since such RNAs are targets for rapid exonucleolytic degradation. Later, a short triphosphorylated RNA was tested as a potential ligand for hnRNPUL1 PPNK, as such RNAs may be more common in the cell, according to Culjkovic-Kraljacic et al. (2020). The core of hnRNPUL1 was able to bind to both RNA species independently of the putative RNA-binding RGG box (Figure 3.5C). The ΔWA and ΔWB mutations had a similar effect on RNA binding as they did on ATP binding (Figure 3.4), with ΔWA showing reduced RNA binding compared to wild-type, while ΔWB enhanced the binding. On the other hand, T507A and R516A both drastically reduced binding, with the most severe effect visible for R516A. The R468C mutation exhibited stronger binding to RNA than WT, in contrast to its reduced ATP binding potential observed before.

These data support the notion that hnRNPUL1 can bind RNA via its PPNK re-

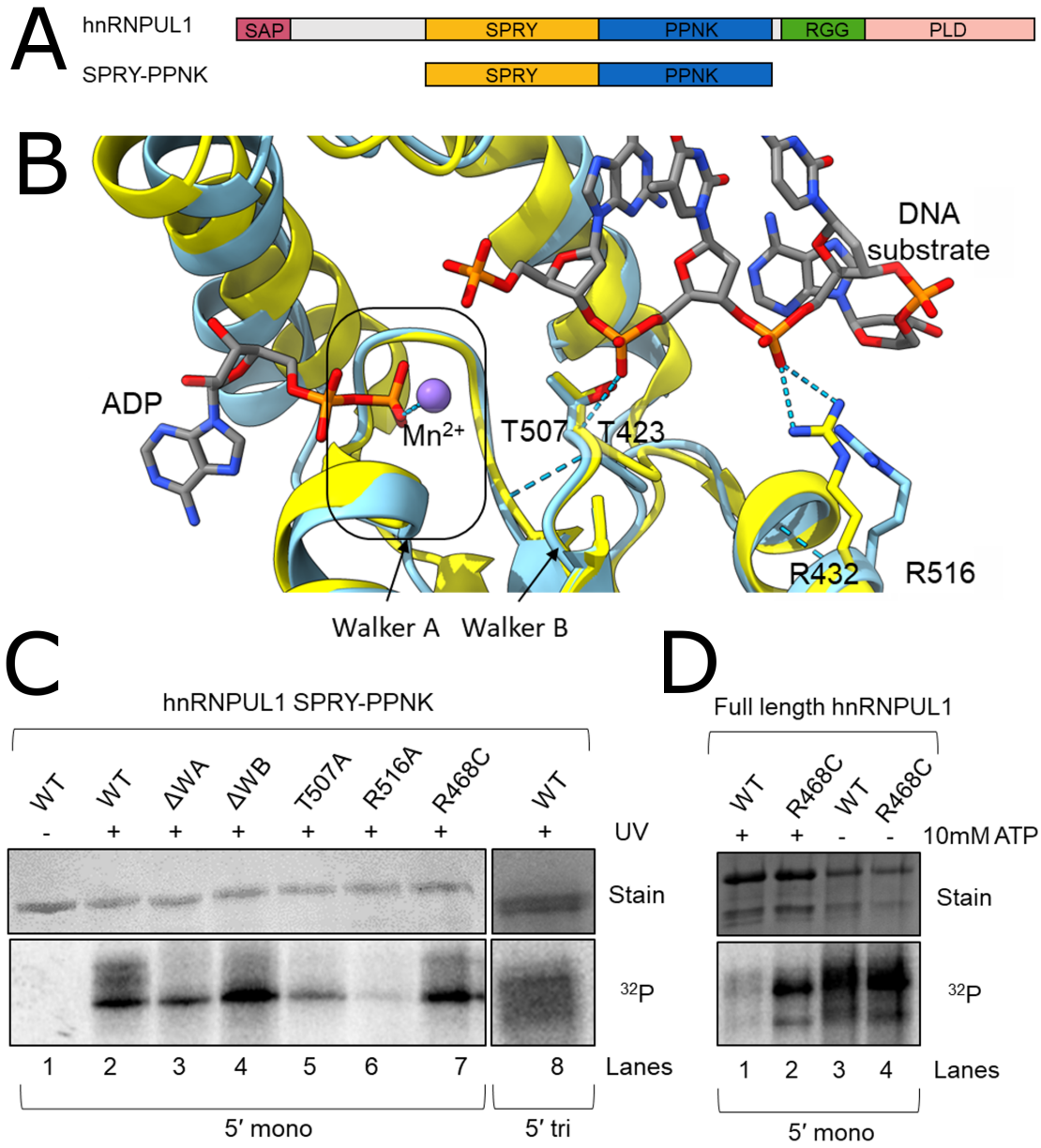

Figure 3.5: Mutational analysis of the RNA binding channel.
A) Schematic of hnRNPUL1 and representation of the SPRY-PPNK truncation used in the assay below. The proteins were purified as FLAG-tagged versions after transient transfection in HEK-293T cells. **B)** Close-up of the AlphaFold-predicted binding pocket of hnRNPUL1 (light blue) superimposed over the crystal structure of mPNKP (yellow) bound to ligands (PDB: 3ZVN). Two important residues of PNKP, T423 and R432, are shown stabilising one strand of the DNA substrate. The equivalent residues of hnRNPUL1, T507 and R516, are also shown in stick form. Dashed lines= hydrogen bonds. **C)** UV crosslinking of hnRNPUL1 SPRY-PPNK truncation containing the indicated binding pocket mutations to ^{32}P-labelled RNA, either 5′ mono or triphosphorylated. Crosslinked complexes were separated by SDS-PAGE (top) and auto radiographed (bottom). **D)** UV crosslinking of full-length hnRNPUL1 WT or R468C to a monophosphorylated ^{32}P-labelled RNA oligonucleotide and ATP competitor. Crosslinked complexes were separated by SDS-PAGE (top) and auto radiographed (bottom).

gion. The deleterious effects of T507A and R516A indicate a mode of nucleic acid threading into the channel and stabilisation via electrostatic interactions between the substrate backbone and positive charges on the channel surface, as identified in Figure 3.2A. The enhanced binding caused by the ΔWB mutation could be due to slower substrate turnover or more avid binding of the RNA to a less charged pocket. Meanwhile, the small reduction in binding to the ΔWA mutant could be due to local misfolding around the NTP-binding portion of the pocket. It is unlikely that the residues in the Walker A motif could stabilise a monophosphorylated RNA substrate directly, as the inter-atomic distances are too great for H-bonding.

We investigated the impact on RNA binding of R468C in the context of the full-length protein, alongside WT hnRNPUL1, in the presence or absence of ATP. We noticed no difference in RNA binding between the two proteins when 5′ monophosphorylated RNA was used as the sole ligand, however co-incubation with ATP displaced the majority of ^{32}P signal from the WT, but not the mutant hnRNPUL1 (Figure 3.5D). This observation suggests an intermolecular competition between the two ligands, which appears to the perturbed in the weaker ATP binder R468C. It is also consistent with a model of ligand binding via opposite sides of the PPNK pocket, as seen in T4 PNK and mPNKP.

3.3.3 hnRNPUL1 can bind ADP and 5′ monophosphorylated RNA simultaneously via PPNK

Having established that NTPs and 5′ mono- or triphosphorylated RNA are ligands for hnRNPUL1, we next wanted to validate the binding mechanism similar to PNKP suggested by the R468C mutant UV crosslinking (Figure 3.5B and D). We employed a ΔCTD truncation of hnRNPUL1 lacking the putative RNA-binding RGG box (Figure 3.6A) and UV crosslinked it to a 5′ ^{32}P monophosphorylated RNA oligonucleotide in the presence of ATP or ADP. Even the smallest amount of ATP competitor displaced over half of the radioactive signal in a dose-dependent manner (Figure 3.6B). On the other hand, ADP appeared to effectively stabilise the RNA in the PPNK pocket and only displaced the signal at the highest concentration. This was likely due to non-specific competition for binding to the positively-charged patch on the channel formed by residues such as T507 and R516 as discussed in the

previous section.

We generated a model showing how the substrates tested in Figure 3.6B likely access the binding pocket of PPNK (Figure 3.6C), which can accommodate up to three phosphate groups contributed either by a 5′ triphosphorylated RNA or by a 5′ monophosphorylated RNA and an ADP. The data are in good agreement with earlier reports of PNKP being able to form a complex with both ATP and an unphosphorylated DNA oligonucleotide simultaneously (Mani et al. 2003). This particular geometry of substrate binding raised the question whether a 5′ capped RNA could access the binding pocket in a similar manner.

3.3.4 hnRNPUL1 cannot bind the inverted 5′ cap structure of an RNA

The results in Figure 3.6 indicated that the substrates bound in the PPNK pocket would adopt a geometry reminiscent of an inverted 5′ cap structure of an RNA. Additionally, it seemed possible that the inverted 5′ base might act to stabilise the rest of the RNA body inside PPNK by inducing a conformation of the pocket more suitable for RNA binding.

In order to test this possibility, we initially used a bandshift assay to look at complex formation between the core of hnRNPUL1 and a capped RNA. However, despite observing protein-RNA complexes, we were unable to prove that the binding occurred via the capped 5′ end, as competition with excess amounts of cap analog did not disrupt complex formation between capped RNA and both the hnRNPUL1 core or the positive control CBP20 (data not shown). Therefore, we opted for an alternative approach where we UV crosslinked the ΔCTD truncated hnRNPUL1 (Figure 3.7A) to a mixture of α ^{32}P-labelled GTP and either unlabelled GTP or m^{7}G cap analog as competitors. CBP20 was used as a positive control due to its avid RNA cap binding. GTP displaced the ^{32}P GTP signal effectively at 1:1 or 10:1 ratios of GTP competitor:protein, while the cap analog showed no disruption (Figure 3.7B). The opposite effect was observed for CBP20, with the cap analog displacing most of the signal at 1:1 or 10:1 ratios of competitor:protein, while GTP proved ineffective. The lack of affinity of CBP20 for GTP compared with cap analog (^{7}GpppG) in this competition assay may be explained by the ∼750 times difference in Kd for the two

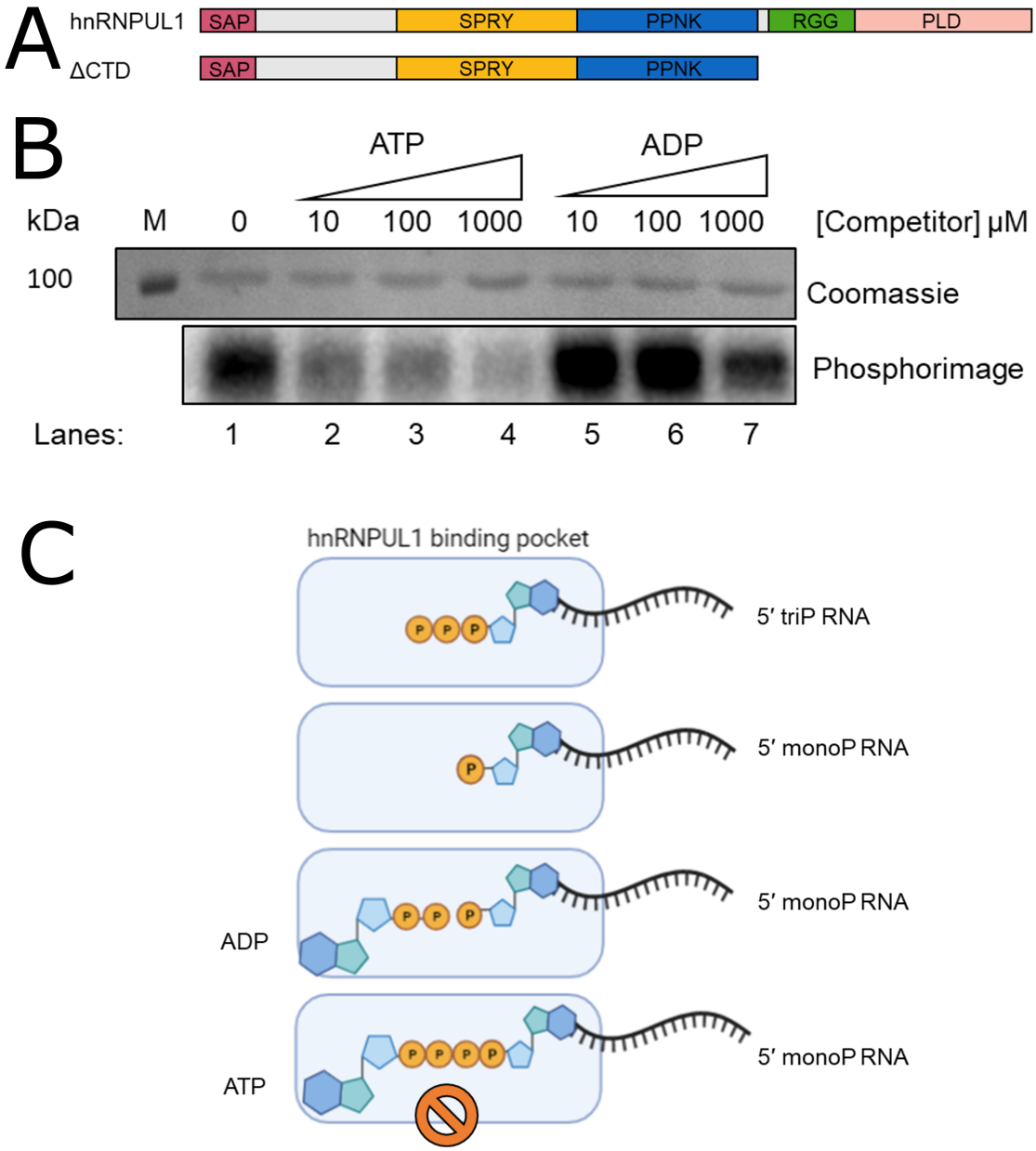

Figure 3.6: Accessibility of hnRNPUL1 binding pocket to substrates. **A)** Schematic of domain organisation of hnRNPUL1 and representation of the ΔCTD truncation used in the assay below. The protein was purified as a FLAG-tag fusion after transient transfection in HEK-293T cells. **B)** UV crosslinking of hnRNPUL1 ΔCTD protein to a mixture of 5′ ^{32}P monophosphorylated RNA oligonucleotide and varying amounts of ATP or ADP. Crosslinked complexes were separated by SDS-PAGE and auto radiographed. **C)** Proposed mode of concomitant substrate binding to the PPNK pocket.

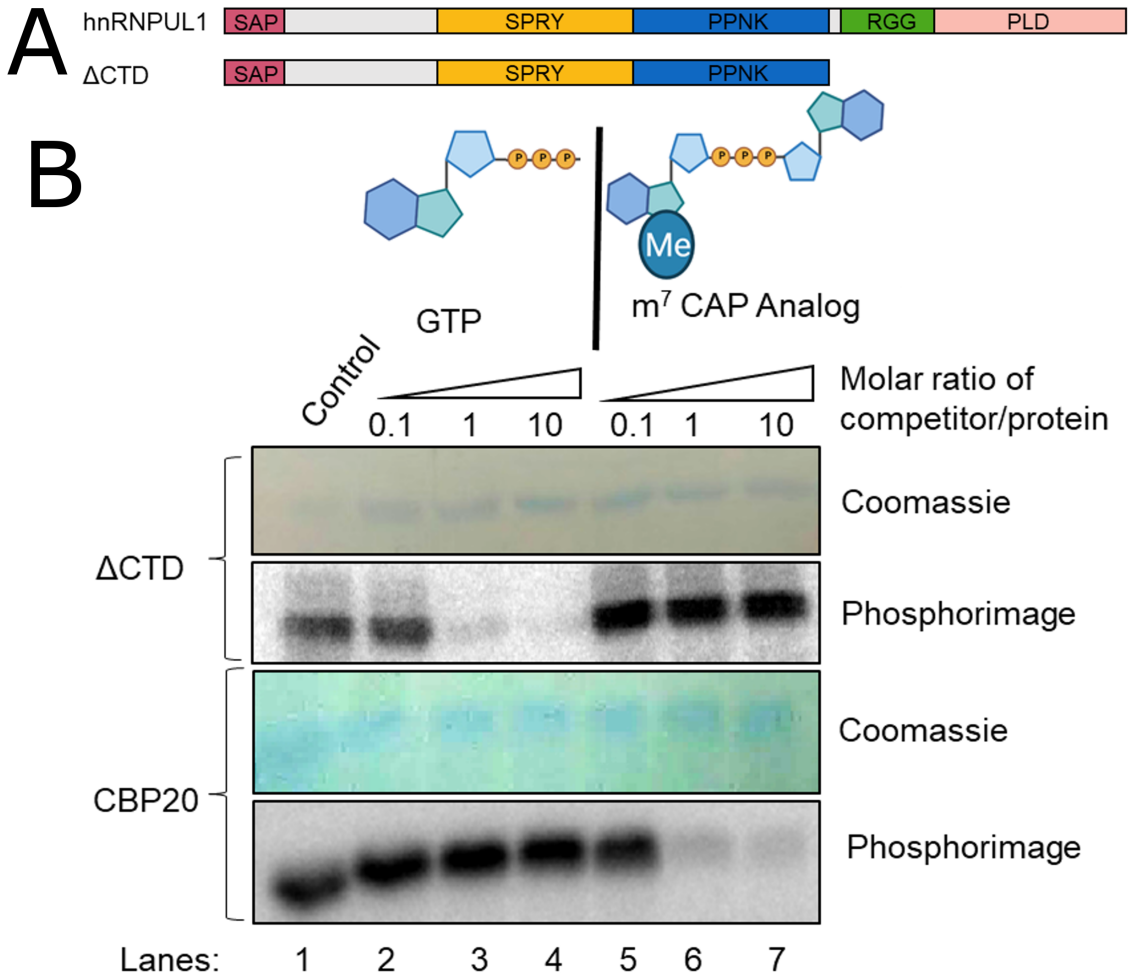

Figure 3.7: 5′ capped RNAs are not suitable binding substrates for hnRNPUL1.
A) Schematic of domain organisation of hnRNPUL1 and representation of the ΔCTD truncation used in the assay below. ΔCTD was purified as a FLAG-tag fusion after transient transfection in HEK-293T cells. CBP20 was purified as a 6x-His-tag fusion from *E. coli* BL21 cells. **B)** UV crosslinking of hnRNPUL1 ΔCTD protein (top two panels) and CBP20 (bottom two panels) to a mixture of α-^{32}P GTP and varying amounts of either unlabelled GTP or m^{7}G cap analog. Crosslinked complexes were separated by SDS-PAGE and auto radiographed.

species: ∼3 μM for GTP and ∼4 nM for m^7GpppG (Worch et al. 2005). For that reason, the competitor concentrations tested may have been too low for GTP to have any effect on signal displacement. The result also suggests that hnRNPUL1 is unlikely to accommodate the inverted 5′ cap structure in the PPNK channel, despite being able to recognise the free 5′ of the RNA.

3.4 A key Asp-Asn change may impair enzymatic activity in hnRNPUL1

In a close-up of the superimposed binding pockets of PNKP kinase and hnRNPUL1 PPNK, we noticed a small but potentially significant difference: residue D396 on PNKP was naturally changed to an asparagine at the equivalent position in hnRNPUL1 (N456) (Figure 3.8). It has been shown that D396 is crucial for kinase activity in PNKP and that mutating the residue to N drastically reduces its enzymatic potential (Bernstein et al. 2009). This is because Glu396 is well placed to activate the 5′ OH- of the DNA substrate, functioning as a general base in the phosphoryl transfer reaction (Garces et al. 2011). Moreover, the general fold of the PPNK region was classified as a P-loop kinase on the basis of a 4-1-2-3 strand configuration within the central β-sheet core and an additional helical "lid" at the top of the structure according to Leipe et al. (2003). Given that we were unable to detect ATP hydrolysis by hnRNPUL1 with colorimetric assays as discussed in section 3.3.1, we hypothesised that it could be due to the Asp→Asn change and that mutating Asn456 to Asp might restore enzymatic activity to hnRNPUL1.

3.4.1 N456D mutation rescues the kinase activity of hnRNPUL1

We introduced the N456D mutation in the full-length hnRNPUL1 protein and tested the ATP hydrolysis activity by thin layer chromatography (TLC) alongside the wild-type and ΔWA proteins. TLC is used to separate charged molecules based on the strength of their interaction with a matrix, such that more highly charged molecules like ATP migrate slower, whereas less charged molecules like ADP or AMP migrate faster as they make fewer contacts with the matrix. T4 PNK and apyrase were used as positive controls to track the migration of ADP and AMP molecules. A further reason T4 PNK was used as a positive control is due to its high structural homology

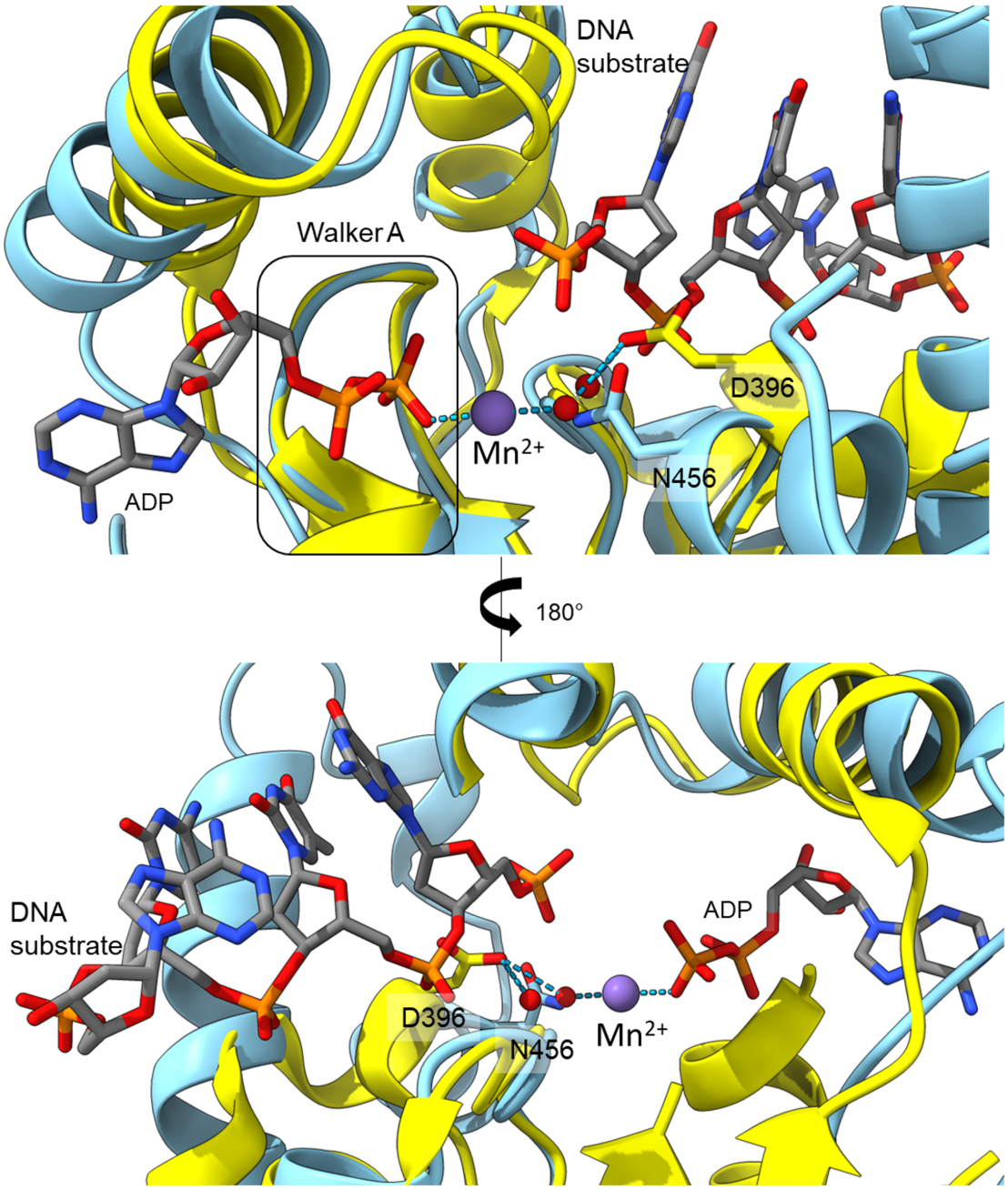

Figure 3.8: Asp-Asn, a key residue difference responsible for catalysis. Orthogonal views showing close-ups of the predicted binding pocket of hnRNPUL1 (light blue) superimposed over the crystal structure of mammalian PNKP (yellow) bound to ligands (PDB: 3ZVN). A key catalytic residue of PNKP, D396, and the equivalent residue of hnRNPUL1, N456, are highlighted and the sidechains represented in stick form. Important motifs and ligands are also highlighted. In both the top and bottom panels, some protein regions are omitted for clarity. Dashed lines= hydrogen bonds; red spheres= solvent molecules.

with the hnRNPUL1 PPNK region (Figure 3.3C) and commercial availability. The ΔWA mutant was used as a measure of background ATP hydrolysis, as the protein previously showed minimal ATP binding in section 3.3.1.

Wild-type hnRNPUL1 was unable to hydrolyse ATP above background, with higher levels of hydrolysis observed by ΔWA hnRNPUL1, likely due to varying amounts of trace contaminants (Figure 3.9A). On the other hand, N456D generated 45% more ADP above background, which was further increased to 120% in the presence of an RNA oligonucleotide with a 5′ OH- group. We introduced an RNA oligonucleotide in the reactions based on the rationale that DNA and RNA kinases such as T4 PNK lack ATP hydrolysis in the absence of a phosphate acceptor (Richardson 1965).

Based on the observation that the 5′ OH- RNA oligonucleotide more than doubled ATP hydrolysis for N456D, we wondered if the protein used the RNA as a substrate for phosphorylation. Therefore, we carried out a kinase assay where we incubated either an RNA oligonucleotide or a double stranded DNA with free 5′ ends and 8 nt 5′ overhangs, both carrying 5′ OH- groups, with the same proteins as above and checked the nucleic acids for ^{32}P signal. We observed that N456D alone transferred the γ phosphate onto both the RNA and DNA substrates, just like T4 PNK, while hnRNPUL1 WT and ΔWA displayed no kinase activity (Figure 3.9B). The dsDNA substrate was included as it is the only phosphorylation substrate suitable for PNKP, unlike T4 PNK, which acts on both DNA and RNA 5′ overhangs (Karimi-Busheri & Weinfeld 1997). Given that the PPNK region of hnRNPUL1 shares extensive structural homology with both PNKP and T4 PNK, we wanted to test the substrates employed by both enzymes.

Lastly, we introduced the same N512D mutation in the PPNK pocket of hnRNPU, which is located within a structurally homologous region of the protein, so we anticipated a similar restorative phenotype. Unfortunately, we did not observe any kinase activity for either the WT or mutated versions of hnRNPU, suggesting that additional amino acid differences within the binding pocket may prevent this restoration (Figure 3.9C). It is also possible that the protein failed to bind to the nucleic acid substrates in the first place, although the experimental conditions were mild enough to allow both hnRNPUL1 and T4 PNK to associate with the RNA and DNA 5′

ends.

The restoration of enzymatic activity to hnRNPUL1 with the N456D point mutation demonstrates several important things. Firstly, it proves that hnRNPUL1 is naturally a nucleic acid pseudo-polynucleotidekinase capable of binding substrates, but not of phosphorylating them. Moreover, the proposed mode of substrate binding identified in section 3.3.3 is correct and both a nucleic acid and ATP can localise to opposite ends of the binding pocket simultaneously as long as they share no more than three phosphate groups between them. Lastly, the experiment shows that hnRNPUL1 is one of very few RNA binding proteins capable of specifically recognising the free 5′ end of an RNA.

3.5 Summary

In this chapter, I have shown how the AI-assisted structural model for hnRNPUL1 has informed our *in vitro* assays to determine its protein function. AlphaFold predicted what appeared to be a well folded protein core formed of the SPRY and nucleotide binding regions, with an extensive interface between the two (Figures 3.1 and 3.2).

We devised experiments to test the substrate preference and mode of binding of hnRNPUL1 based on the structural homology of the nucleotide binding region with mammalian PNKP (Figure 3.3). These revealed hnRNPUL1 to bind to ATP and GTP via its NTP binding region, with a slight preference for ATP (Figure 3.4). hnRNPUL1's ability to bind more than one type of NTP established it as a polynucleotide binding protein. Furthermore, mutating the Walker A motif abolished its ATP binding potential, while mutating the Walker B motif enhanced its ATP binding.

Next, we demonstrated that hnRNPUL1 was able to bind mono- and triphosphorylated RNA via its NTP binding region in the absence of its putative RNA binding RGG box (Figure 3.5). RNA binding via the hnRNPUL1 core was mediated to a large extent by a positively-charged patch on the surface of the binding channel formed by residues T507 and R516. This binding mode is similar to that used by

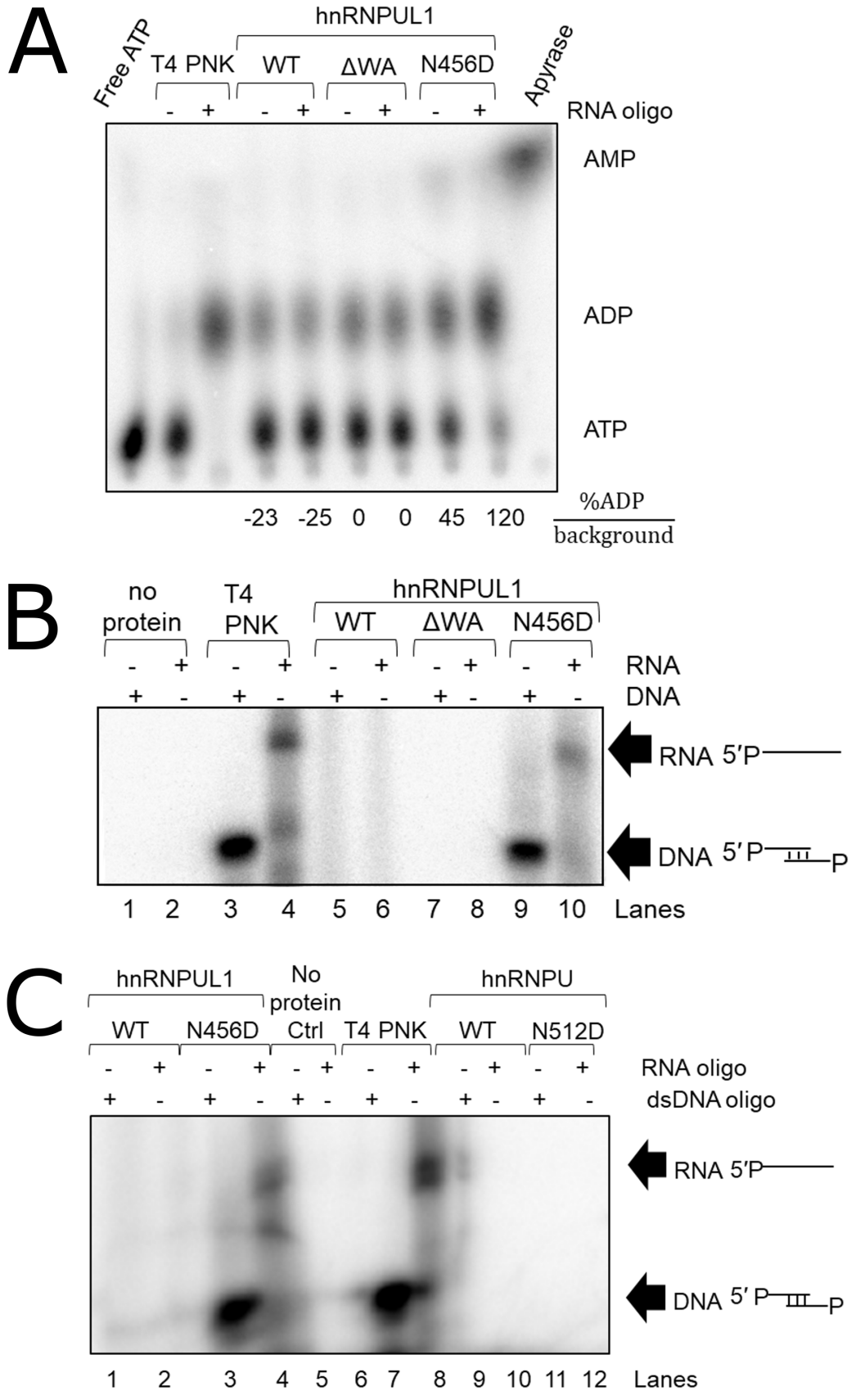

Figure 3.9: Activating a pseudo-kinase.
A) TLC assay tracking α-^{32}P ATP hydrolysis to ADP or AMP. The amount of hydrolysis in each condition is expressed as the % of ADP generated over background. **B) and C)** Kinase assays testing phosphorylation of RNA and DNA fragments with accessible 5′ ends using γ ^{32}P-ATP. The nucleic acids were separated on a denaturing polyacrylamide gel following incubation with the different proteins and auto radiographed. hnRNPUL1 and hnRNPU proteins were purified by FLAG-IP from transfected HEK-293T cells.

PNKP. Further proof of a comparable mode of substrate binding between hnRNPUL1 and PNKP was provided through UV crosslinking experiments which effectively measured the size of the binding pocket and demonstrated that an RNA and ATP or ADP substrates could simultaneously access the pocket as long as there were no more than three phosphate groups shared between them (3.6). Based on these similarities between our protein and PNKP, we renamed the NTP binding region of hnRNPUL1 pseudo-PNK, or PPNK.

Additionally, we tested whether a 5′ capped RNA could access the binding pocket of PPNK in the same manner as ADP and a 5′ monophosphorylated RNA and revealed that hnRNPUL1 would be unlikely to accommodate such a structure (Figure 3.7).

Finally, we noted a small but significant difference in the channels of hnRNPU or hnRNPUL1 and PNKP: a key catalytic residue of PNKP (D396) was naturally changed to N456 at the equivalent position in hnRNPUL1 or N512 in hnRNPU (Figure 3.8. Reversing N456 to D restored ATP hydrolysis and DNA/RNA kinase activity in hnRNPUL1 (Figure 3.9), but not hnRNPU. This experiment demonstrates that hnRNPUL1 is one of only a handful of proteins known to specifically recognise the free 5′ ends of nucleic acids and that the wild-type protein was an inactive, or pseudo-, polynucleotide kinase.

4 Chapter 4: Two separate RNA binding regions coordinate hnRNPUL1's nuclear functions

In this chapter, we will investigate the function of the central SPRY-PPNK domain in the context of the entire protein. First, we will present evidence for a conformational change around the PPNK pocket triggered by ligand binding. Next, we will examine a self-inhibitory intramolecular interaction between SPRY-PPNK and CTD achieved when the protein adopts an apo state. The search for *in vivo* RNA partners bound to the PPNK domain highlighted U4 snRNA as a transcriptional and post-transcriptional hnRNPUL1 target. We uncovered the U4-specificity factor Snu13 as an hnRNPUL1-binding partner, providing a mechanism of post-transcriptional recruitment of hnRNPUL1 to U4 snRNP. Moreover, we will present evidence that hnRNPUL1 is involved in snRNA gene transcription through a CTD-mediated interaction with PolII. We will then analyse a two-state mechanism of hnRNPUL1 chromatin association involving either disordered protein interactions or RNA binding to the PPNK pocket and the RGG box.

4.1 Tryptophan fluorescence quenching upon substrate binding supports a conformational change of hnRNPUL1

Having established hnRNPUL1's ability to bind NTPs and free nucleic acid 5′ ends, we set out to characterise the kinetics of these binding events. To that end, we employed tryptophan fluorescence quenching assays of the core SPRY-PPNK fragment in either wild-type form or carrying ΔWA or ΔWB mutations (Figure 4.1A). ATP and 5′ triphosphorylated RNA were used as ligands in separate assays. Sequence analysis of the hnRNPUL1 core revealed the presence of 5 tryptophans, 3 of them in the SPRY region (W243 and W286 buried in the hydrophobic core and W351 solvent-exposed) and 2 around the NTP/RNA binding pocket: W437 close to the Walker A motif and W477 within the discriminator helix-loop region (Figure 4.2A).

The tryptophan fluorescence quenching protocol was adapted from Mani et al. (2003) and Bernstein et al. (2005). However, the protein concentration used here was

increased to 0.5 μM from 0.15 and 0.4 μM, respectively. This concentration was experimentally determined to be the lowest possible one which allowed fluorescent signal measurement in our hands. Moreover, the ligand concentration range was capped at 5 μM compared to 20 and 70 μM due to the difficulty of producing large amounts of triphosphorylated RNA oligonucleotide by *in vitro* transcription. Finally, the samples were incubated at 37°C instead of 5 and 20°C, to match the conditions used previously for the UV crosslinking experiments.

We observed a 13% quenching of the tryptophan fluorescence signal for the WT and ΔWB proteins upon ATP binding (Figure 4.1B) and 16-18% quenching for the same proteins upon RNA binding (Figure 4.1C). On the other hand, the ΔWA mutant displayed minimal tryptophan fluorescence quenching (3% and 7% for ATP and RNA, respectively), consistent with its impaired substrate binding discussed in sections 3.3.1 and 3.3.2. Overall, the assays revealed that the core of hnRNPUL1 undergoes a change in intrinsic tryptophan fluorescence upon binding to its substrates. This is indicative of a change in the local environment of these residues and suggests there is a conformational change in the PPNK domain upon substrate binding. We calculated the Kd values for ATP and RNA binding to the WT and ΔWB proteins as shown in the table below. Again, the data were in good agreement with the previous UV crosslinking results which indicated the ΔWB mutant was a more avid ATP and RNA binder than the WT protein.

	Kd (nM)	
	ATP	5′ 3·P RNA
WT	164 ±22	787 ±177
ΔWB	19 ±8.5	97 ±16

Residues W437 and W477 are the most likely to contribute to the observed fluorescence quenching due to their proximity to the binding pocket, as we would expect the conformational change of the protein to mainly be localised in this region (Figure 4.2A). A potential conformational change of hnRNPUL1 modeled on the mPNKP apo- and bound structures also suggested that the rearrangement was mainly localised around the discriminator helix-loop region, where ligand binding was accompanied by closing of the protein structure (Figure 4.2B). It is possible that a different conformational change may be assumed, but the exact mechanism may

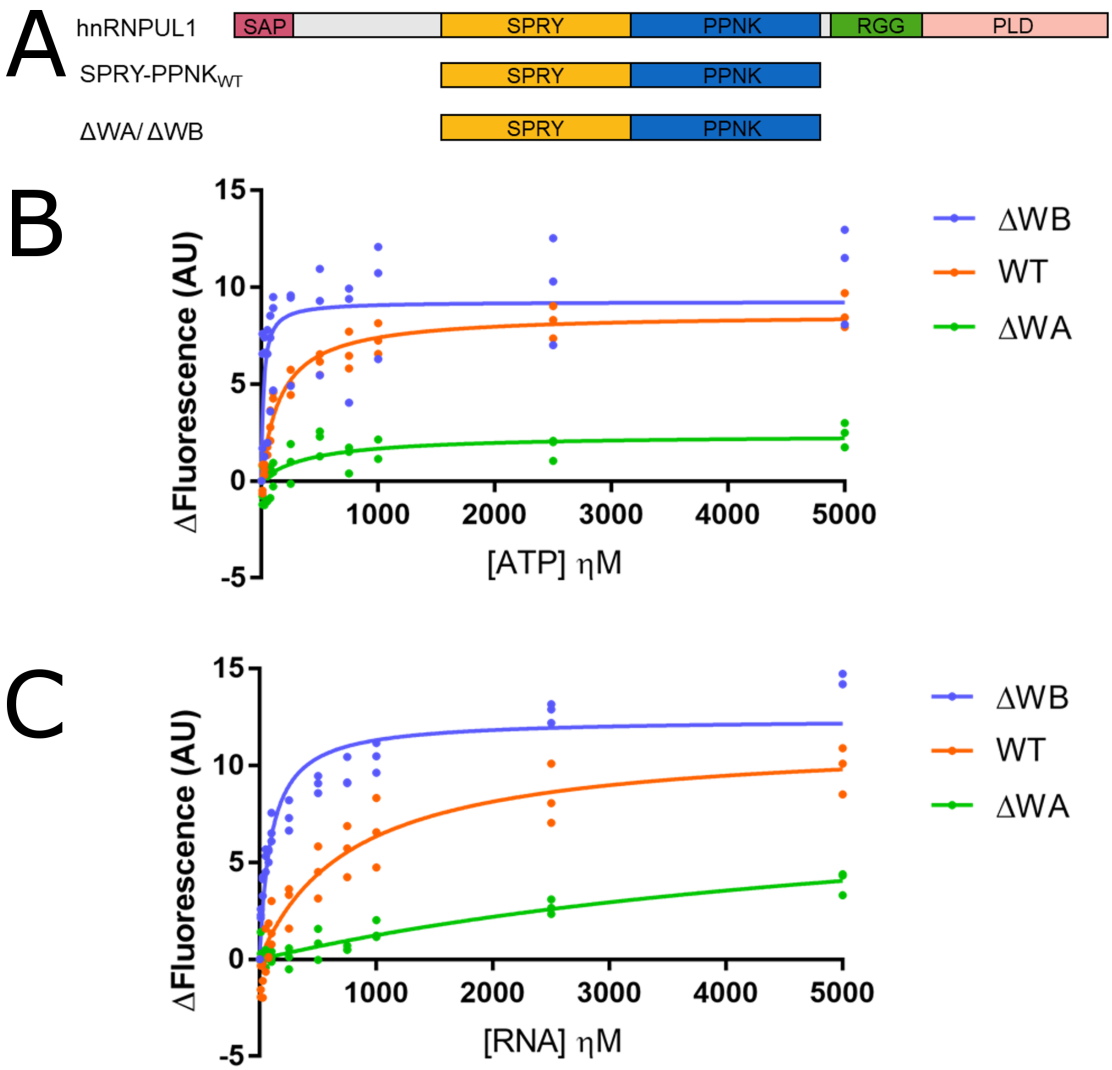

Figure 4.1: Tryptophan fluorescence quenching supports conformational change of hnRNPUL1.
A) Schematic of domain organisation of hnRNPUL1 and representation of the SPRY-PPNK truncations (wild-type and ΔWA/ΔWB mutants) used in the assays below. The proteins were purified as FLAG-tagged versions after transient transfection in HEK-293T cells. **B)** Plot of tryptophan fluorescence quenching of different SPRY-PPNK mutant forms upon titration of ATP. **C)** Plot of tryptophan fluorescence quenching of different SPRY-PPNK mutant forms upon titration of 5′ triphosphorylated RNA. n=3

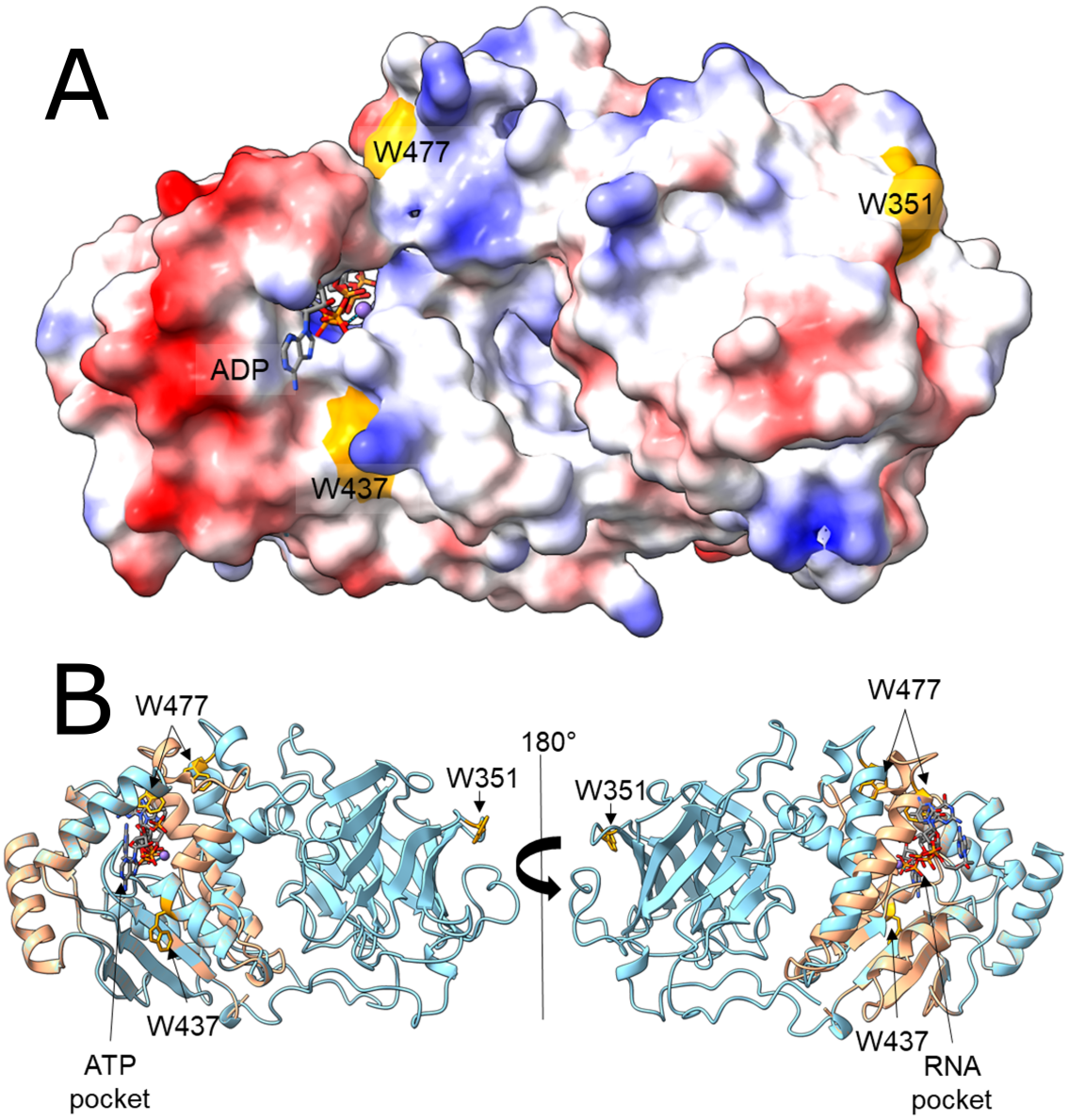

Figure 4.2: Model of PPNK conformational change.
A) Surface charge of the SPRY-PPNK core, highlighting the distribution of solvent-exposed tryptophans (orange). The W243 and W286 residues buried in the SPRY core are omitted. **B)** Model of hnRNPUL1 conformational change from apo- (blue) to liganded (tan) forms, based on the observed conformational change in mPNKP between the apo (PDB: 3ZVL) and liganded (PDB: 3ZVN) states. The model on the left is in the same orientation as the surface charge in **A)** and the sidechains of solvent-exposed tryptophan residues are shown in stick form and coloured in orange.

only be gleaned from direct structural observations. Additionally, the AlphaFold model for the hnRNPUL1 core displayed the lowest confidence (<70%) in the discriminator helix-loop region encompassing W477, indicating that the region may possess a large degree of flexibility. The slightly higher percent of quenching noted for RNA binding could be due to the larger substrate making more contacts with the protein, including non-specific ones away from the binding pocket. The Kd for WT hnRNPUL1 binding to RNA (787 ± 177 nM) was comparable to those reported to PNKP binding to recessed DNA (750 ± 100 nM) and blunt end DNA (1,350 ± 100 nM) (Bernstein et al. 2009). However, the Kd for ATP binding (164 ± 22 nM) was roughly 8 times lower than that reported for PNKP binding to ATP (1,400 ± 200 nM) (Mani et al. 2003). It is possible that the rather limited range of ligand concentrations tested only allows an approximation of the true dissociation constants.

4.2 The ΔWA mutation ablates RNA binding to the full length hnRNPUL1 protein

In the previous chapter, I presented evidence for the SPRY-PPNK core of hnRNPUL1 functioning as an RNA binding module independently from the RGG box-containing C-terminal domain. After studying the SPRY-PPNK core in isolation, we wanted to investigate the effects of the ΔWA and ΔWB mutations in the context of the full-length protein, as they both had profound and opposite effects on ligand binding in the previous chapter.

To that end, full-length hnRNPUL1 proteins harbouring ΔWA or ΔWB mutations were UV-crosslinked to a short ^{32}P-labelled RNA oligonucleotide (Figure 4.3). A ΔCTD mutant, as well as the CTD expressed individually, were included to test the contribution to RNA binding of each segment (Figure 4.3B, lanes 6 and 8). The experiment revealed two important aspects.

Firstly, deletion of the CTD, harbouring the RGG box, drastically reduced the RNA cross-linking efficiency (compare lanes 4 and 6). Consistent with the ability of the RGG box to bind RNA, we observed cross-linking to RNA with the isolated CTD. However, the CTD fragment tended to produce a smeared signal on the gel due to

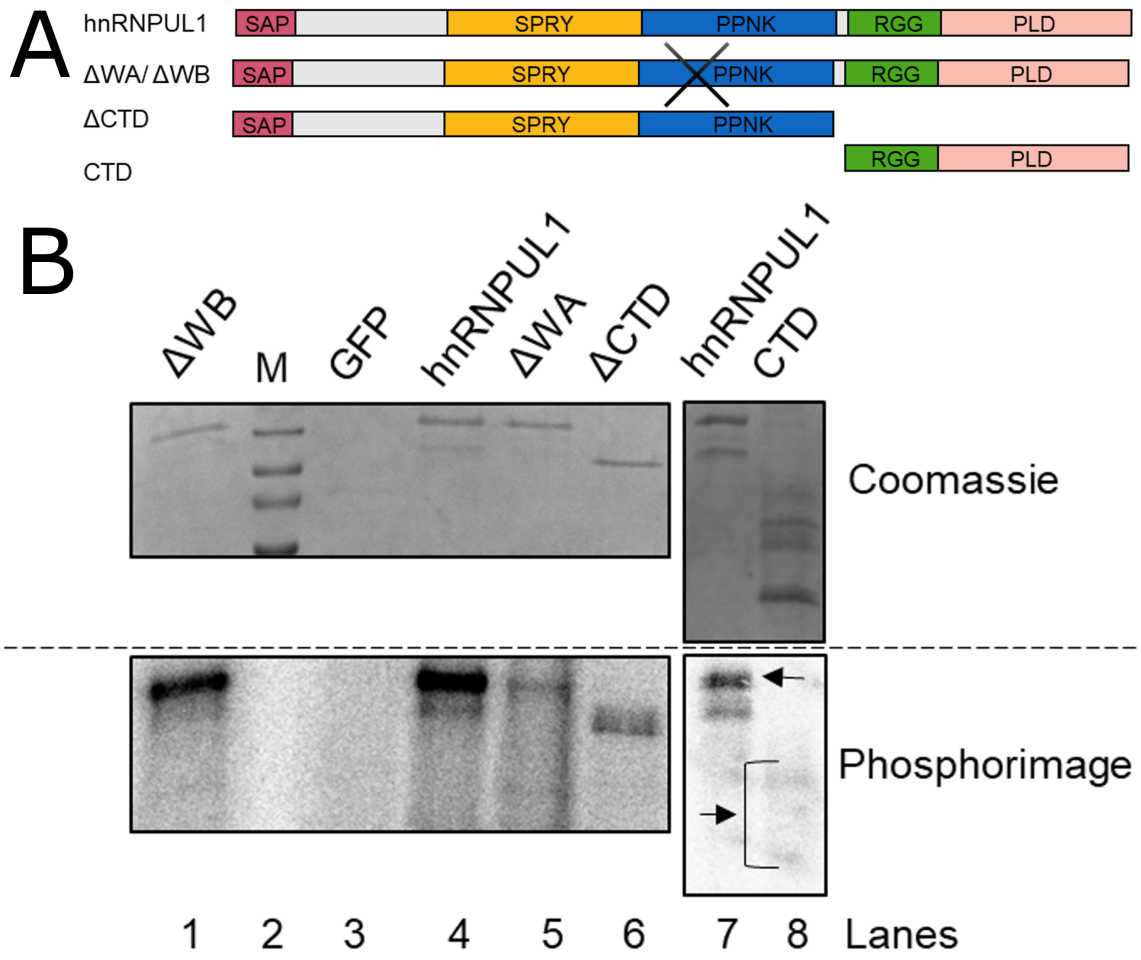

Figure 4.3: RNA binding map of hnRNPUL1.
A) Schematic of hnRNPUL1 domain organisation and illustration of the protein variants used in the assay below. The proteins were purified as FLAG-tagged versions after transient transfection in HEK-293T cells. **B)** UV crosslinking of hnRNPUL1 constructs to ^{32}P-labelled RNA oligonucleotide in a reaction buffer supplemented with 10 mM ATP, followed by separation by SDS-PAGE and auto radiography. GFP was used as a negative control.

protein truncations. It is also possible that RNA contaminants co-purified with the CTD despite extensive RNAse A treatment. We believe this is because the protein tended to self-interact on the beads, creating a less accessible environment for the enzyme.

Secondly, the ΔWA mutation drastically reduced RNA binding to the full-length protein, despite the presence of an intact RGG box. On the other hand, the ΔWB mutation had no adverse effects on RNA binding.

We hypothesised that the observed effect of the ΔWA mutation ablating the RNA binding potential of the whole protein could be due to an intramolecular interaction sequestering the RGG box to the hnRNPUL1 core. We further considered that the conformational change observed upon ligand binding to the PPNK domain might drive ejection of the RGG box, allowing maximal binding to RNA. In this scenario, the inability of the ΔWA mutant to bind to either ligand (as shown in Figures 3.4, 4.1 and 4.3) would not trigger the conformational change necessary to release the RGG domain, preventing optimal RNA binding.

4.3 Intramolecular interaction between the SPRY-PPNK core and CTD of hnRNPUL1

We tested the intramolecular interaction between the SPRY-PPNK core and the CTD discussed in the previous section using a pulldown assay. The core of hnRNPUL1 was fused to GST and used as "bait" to pull down the CTD, which was expressed as a GFP fusion and used as "prey" (Figure 4.4A).

As shown in Figure 4.4B, we observed a specific interaction between the core and the CTD of hnRNPUL1, which was not present in the GST control. Further, we investigated the effect of co-incubation with an excess of total cellular RNA on the amount of CTD being pulled down in these reactions. Surprisingly, the excess of RNA promoted additional binding of CTD to the SPRY-PPNK core instead of displacing it. A possible explanation is that the RNA formed a bridge between the hnRNPUL1 core and the CTD, whereby the 5′ ends of uncapped RNA molecules engaged with the PPNK binding pocket and the RNA bodies interacted with the

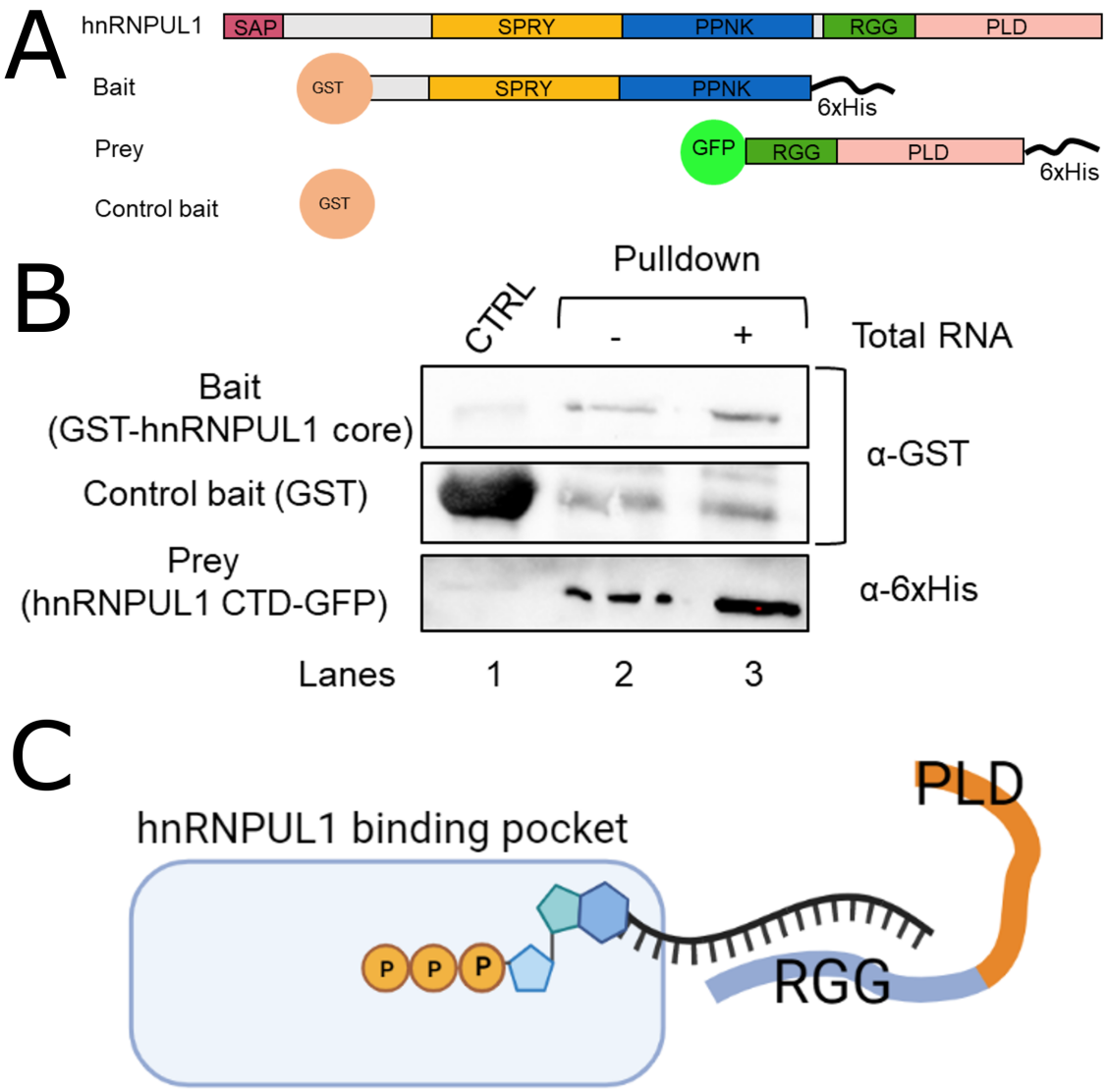

Figure 4.4: Intramolecular switch between the SPRY-PPNK core and CTD of hnRNPUL1.
A) Schematic of hnRNPUL1 domain organisation and of the proteins used as "bait" and "prey" for the pulldown assay. The proteins were expressed in *E. coli* BL21 cells. **B)** Western blot of pulldown reaction of hnRNPUL1 CTD using the GST-tagged SPRY-PPNK core. "Bait" proteins were identified with anti-GST antibody, while the "prey" protein was identified with anti-6xHis antibody. **C)** Proposed model of RNA bridging the core-CTD interaction in **B)**.

CTD via the RGG box, causing additional "prey" to be pulled down (Figure 4.4C). In reality, this scenario may be rather plausible, as it has become apparent that a large proportion of cellular RNA can exist in an uncapped form at steady-state (Culjkovic-Kraljacic et al. 2020). Nevertheless, we realised that this approach was not suitable for testing if an RNA-triggered conformational change could disrupt the core-CTD interaction. Obtaining structural information on apo- and liganded forms of hnRNPUL1 by cryo-EM may support this hypothesis. Alternatively, cysteine residues could be introduced around the PPNK pocket and the RGG box and the protein crosslinked under oxidising conditions in apo- and liganded forms to test the formation and disruption of interdomain interactions by mass spectrometry. The problem with this approach is that the cysteine mutations could cause misfolding, so that possibility would have to be ruled out experimentally.

4.4 The hnRNPUL1 core contains a potential CTD interaction site above the binding pocket

Subsequently, we analysed the AlphaFold model for the SPRY-PPNK core in search of potential interaction regions with the CTD. One obvious feature of the core was an extensive acidic patch above the binding pocket of PPNK (Figure 4.5). There were additional acidic regions on the surface of the core, but none as large as the one above the PPNK pocket. It is possible that this site may engage with the CTD via electrostatic interactions with the positively charged arginine residues in the RGG box. The acidic patch was also an attractive RGG binding target because of its proximity to the PPNK pocket, as it may serve to bring the two RNA-binding regions of hnRNPUL1 into close proximity. However, due to time constraints, we were unable to test this hypothesis.

4.5 U4/U6 di-snRNP is a direct hnRNPUL1 target *in vivo*

Having gained a better understanding of how hnRNPUL1 can bind to RNA in a bimodal manner (to the RNA 5′ end via the PPNK pocket and to the RNA body via the RGG box), we started searching for RNA targets *in vivo* that exhibited one or both binding modes. Using the publicly available ENCODE database, Dr

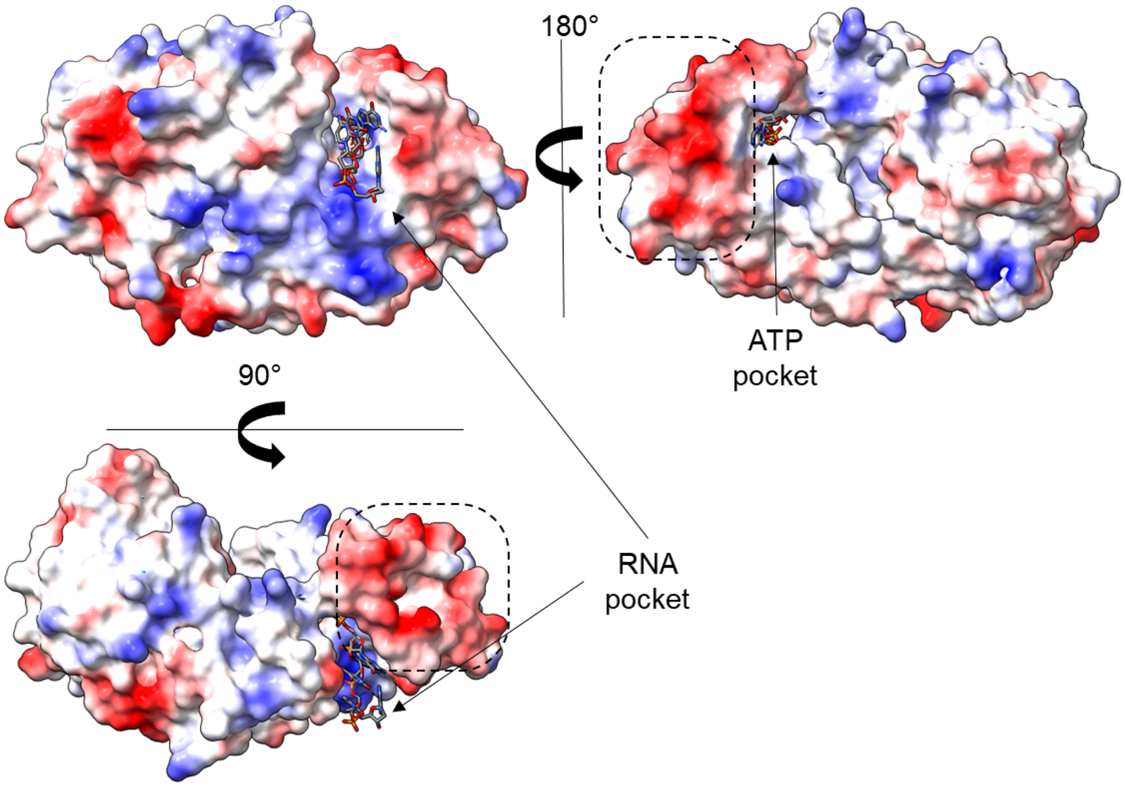

Figure 4.5: Acidic surface patch above the SPRY-PPNK binding pocket. Orthogonal views of hnRNPUL1 core surface charge with the acidic patch above the PPNK binding pocket encircled. RNA and ATP ligands are included and their respective binding sites highlighted.

Ivaylo Yonchev processed the hnRNPUL1 eCLIP dataset published by Nostrand et al. (2020) to single nucleotide resolution (raw data not shown).

One of the identified targets of hnRNPUL1 was U4 snRNA, which displayed eCLIP signal on the first two nucleotides at the 5′ end and also on the 5′ stem loop, as summarised in Figure 4.6. U4 snRNA was an appealing target for further investigation for a number of reasons. Earlier work in the lab performed by Dr Llywelyn Griffith revealed that hnRNPUL1 interacted with all snRNA components of the major spliceosome and that U4 snRNA was the most highly enriched one in RNA immunoprecipitation (RIP) experiments (Griffith 2019). Dr Griffith also showed that knockdown of hnRNPUL1 in an auxin-inducible degron cell line led to a decrease in U4 snRNA levels, which Dr Ivaylo Yonchev was able to complement by overexpressing hnRNPUL1 in the same system (Griffith 2019, Yonchev 2021). Depletion of hnRNPUL1 reduced PolII-mediated transcription of snRNA genes (Griffith 2019) but did not affect the stability of U4 snRNA (Yonchev 2021). The data were in agreement with the ChIP-seq published by Xiao et al. (2019), which identified hnRNPUL1 as the most enriched RBP over snRNA genes (out of a subset of 58 RBPs). Additionally, in the mass spectrometry analysis performed by Dr Llywelyn Griffith, one of the most abundant proteins that co-immunoprecipitated with hnRNPUL1 was SART3, a U6 snRNP recycling factor that remains associated with U6 snRNA until U4/U6·U5 tri-snRNP formation (Bell et al. 2002) (Figure 4.6). However, the SART3 interaction was RNA-dependent and heavily disrupted by mutation of the WalkerA motif of hnRNPUL1 (Griffith 2019).

From the distribution of the hnRNPUL1 eCLIP signal over U4 snRNA, we hypothesised that the first two nucleotides may be bound via the PPNK pocket in the initial stages of transcription, before acquisition of the mono- and later trimethylated 5′ cap (Glover-Cutter et al. 2007, Mattaj 1986). Moreover, we reasoned that the 5′ stem loop binding may occur via the RGG box, possibly at the post-transcriptional level on the mature snRNA, before or after U4/U6 assembly. The hnRNPUL1 eCLIP signal on the 5′ stem loop resided next to the binding site for the highly conserved protein Snu13 (Figure 4.6), known to recruit additional factors to U4 snRNP, such as Prp31 (Liu et al. 2007). For this reason, we wondered if hnRNPUL1 may be one

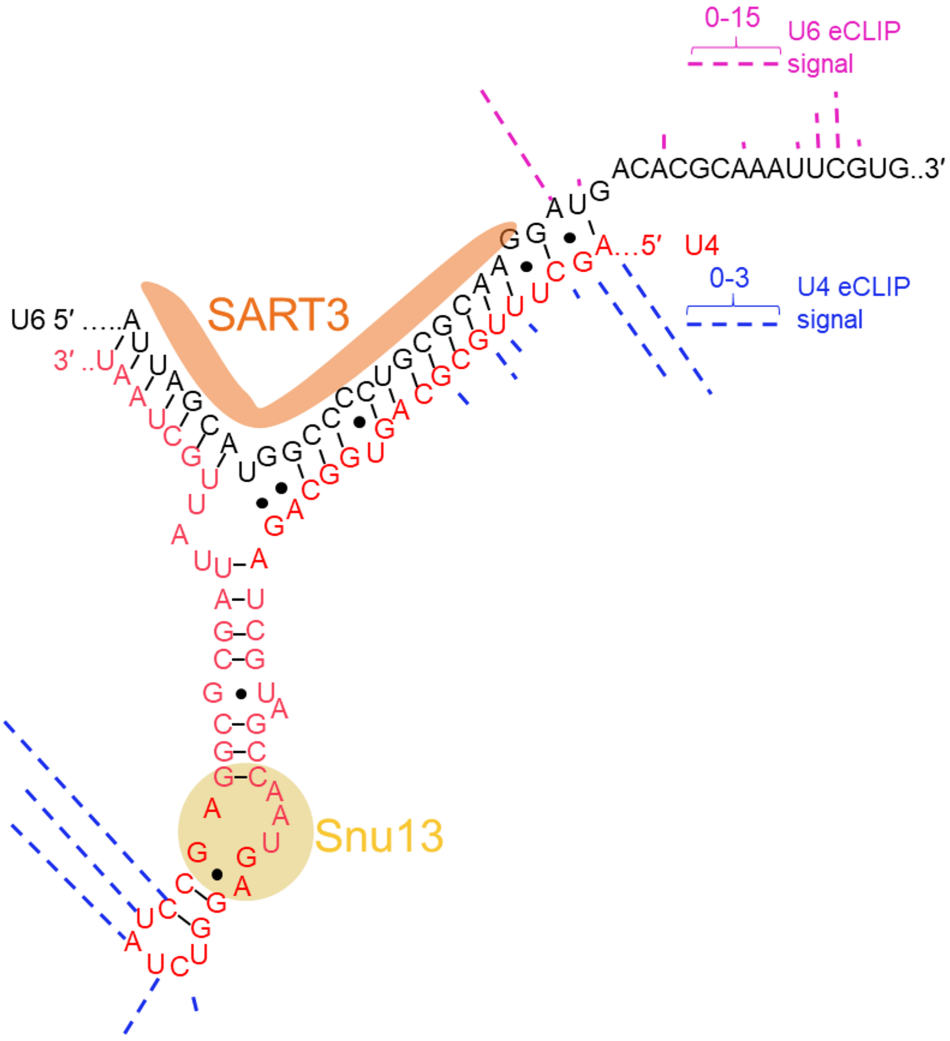

Figure 4.6: U4/U6 di-snRNP model with eCLIP signal for hnRNPUL1. Simplified diagram highlighting the 5′ half of U4 snRNA in contact with U6 snRNA and the position of Snu13 protein around the bulge of the 5′ stem loop. eCLIP signal shown as dashed pink and blue lines indicates individual bases of U4 and U6 snRNAs crosslinked to hnRNPUL1. The size of the dashed lines is proportional to the strength of the eCLIP signal at each position. Solid black lines and dots indicate canonical and non-canonical base pairs, respectively.

of the protein targets recruited by Snu13.

4.5.1 hnRNPUL1 directly interacts with the Snu13 component of U4 snRNP

We investigated if Snu13 could serve as a U4 snRNP specificity factor for hnRNPUL1 using a pulldown assay, where Snu13 served as "bait" and full-length hnRNPUL1 as "prey". GFP was employed as a negative control (Figure 4.7A). We observed a direct protein-protein interaction between Snu13 and hnRNPUL1, whereas no protein was pulled down by GFP (Figure 4.7B).

The direct interaction between the two proteins in the absence of U4 snRNA indicated that Snu13 could be involved in recruiting hnRNPUL1 to the U4 snRNA stem loop. Additionally, the result suggested that hnRNPUL1 binding to U4 snRNP occurred post-transcriptionally during nuclear assembly or recycling of U4/U6 di-snRNP (Nottrott et al. 1999). However, the hnRNPUL1 eCLIP signal on the pentameric loop adjacent to the Snu13 binding site is also known to interact with the Prp31 factor during di-snRNP/tri-snRNP formation, suggesting that hnRNPUL1 might function after U4 snRNP disassembly from U4/U6·U5 tri-snRNP in a recycling mechanism (Liu et al. 2007). In support of a recycling function, Griffith (2019) observed that depletion of hnRNPUL1 reduced U4/U6 and U4/U6·U5 di- and tri-snRNP formation.

While the pulldown assay did not test if hnRNPUL1 could be recruited to the U4 snRNP by other factors and itself serve to engage Snu13, it has been documented that Snu13 can bind the U4 5′ stem loop directly in the absence of other factors. This supports the argument that Snu13 is the specificity factor for hnRNPUL1 and not vice versa, although it was unclear which region of hnRNPUL1 specifically interacted with Snu13.

4.5.2 Interaction with Snu13 occurs via the hnRNPUL1 core

Our next step was to dissect the hnRNPUL1-Snu13 interaction and identify the region of hnRNPUL1 responsible for binding. To that end, we employed a slightly modified pulldown protocol, where instead of purified "prey" proteins, we used whole

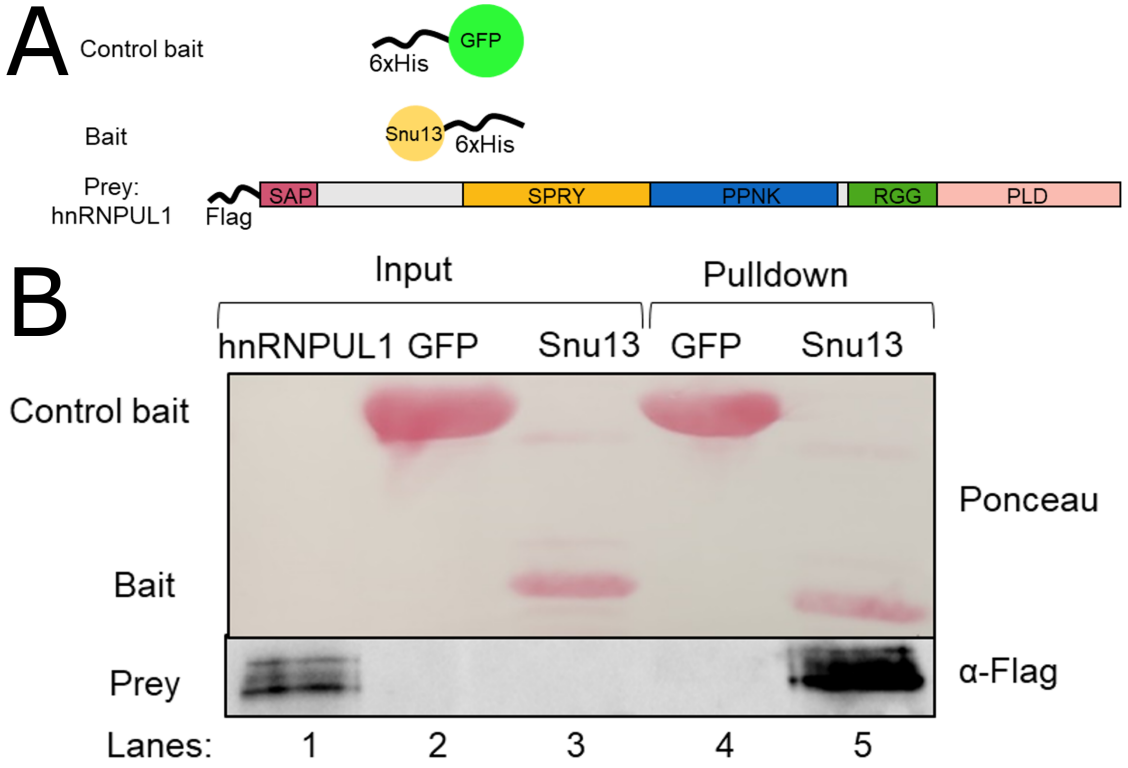

Figure 4.7: hnRNPUL1 directly interacts with Snu13.
A) Schematic of hnRNPUL1 domain organisation and of the proteins used as "bait" and "prey" for the pulldown assay. The "bait" proteins were expressed in *E. coli* BL21 cells, while hnRNPUL1 was overexpressed and purified from HEK 293T cells. **B)** Pulldown of full-length hnRNPUL1 with Snu13 protein. Top: Ponceau stain of the "bait" proteins GFP and Snu13. Bottom: Western blot of hnRNPUL1 "prey" protein using anti-FLAG antibody.

cell extracts of mammalian cells following overexpression of different hnRNPUL1 variants. These were then pulled down with Snu13 or GFP "baits" (Figure 4.8A). The amount of cell lysates added to each reaction was equalised based on the level of hnRNPUL1 variant expressed (Figure 4.8B top).

As shown in the bottom panel of Figure 4.8B, all hnRNPUL1 variants were pulled down by Snu13 to a similar extent, except for the ΔSPRY mutant (lane 4). On the other hand, deleting a part or the entire CTD did not have any impact on the Snu13 interaction (lanes 7 and 8). We also showed that the SAP domain was not involved in Snu13 binding (lane 3) and that a functional PPNK region was not necessary (lanes 5 and 6 compared to 2). We could not rule out the possibility that the PPNK portion of the core contributed to the Snu13 interaction through a different region than the binding pocket and that deletion of the SPRY segment disturbed the fold of that region. However, it was clear that the hnRNPUL1-Snu13 interaction relied on an intact protein core, but not on its ability to bind ATP or RNA ligands. Moreover, the association with Snu13 on the U4 stem loop via the hnRNPUL1 core would allow the RGG box to interact with the RNA bases in the loop region as described in Figure 4.6.

4.6 Developing No Cap RIP to identify *in vivo* 5′ end RNA targets of hnRNPUL1

After having identified the SPRY-PPNK core of hnRNPUL1 as an RNA 5′ end binder necessary for protein activity, we wanted to develop a method to capture RNA targets in an unbiased way. Specifically, we focused on binding events occurring via the SPRY-PPNK core to free RNA 5′ ends. Recently, Culjkovic-Kraljacic et al. (2020) described a protocol for separating RNA molecules into 5′ capped and uncapped populations with an enzymatic treatment prior to sequencing, which was used to measure the capping efficiency for ∼100 transcripts under different conditions.

We wanted to adapt this protocol to characterise the capping status of RNA molecules bound specifically to the PPNK pocket and termed it No Cap RIP. To this end, we developed a modified RIP assay that did not use chemical or UV crosslinking of

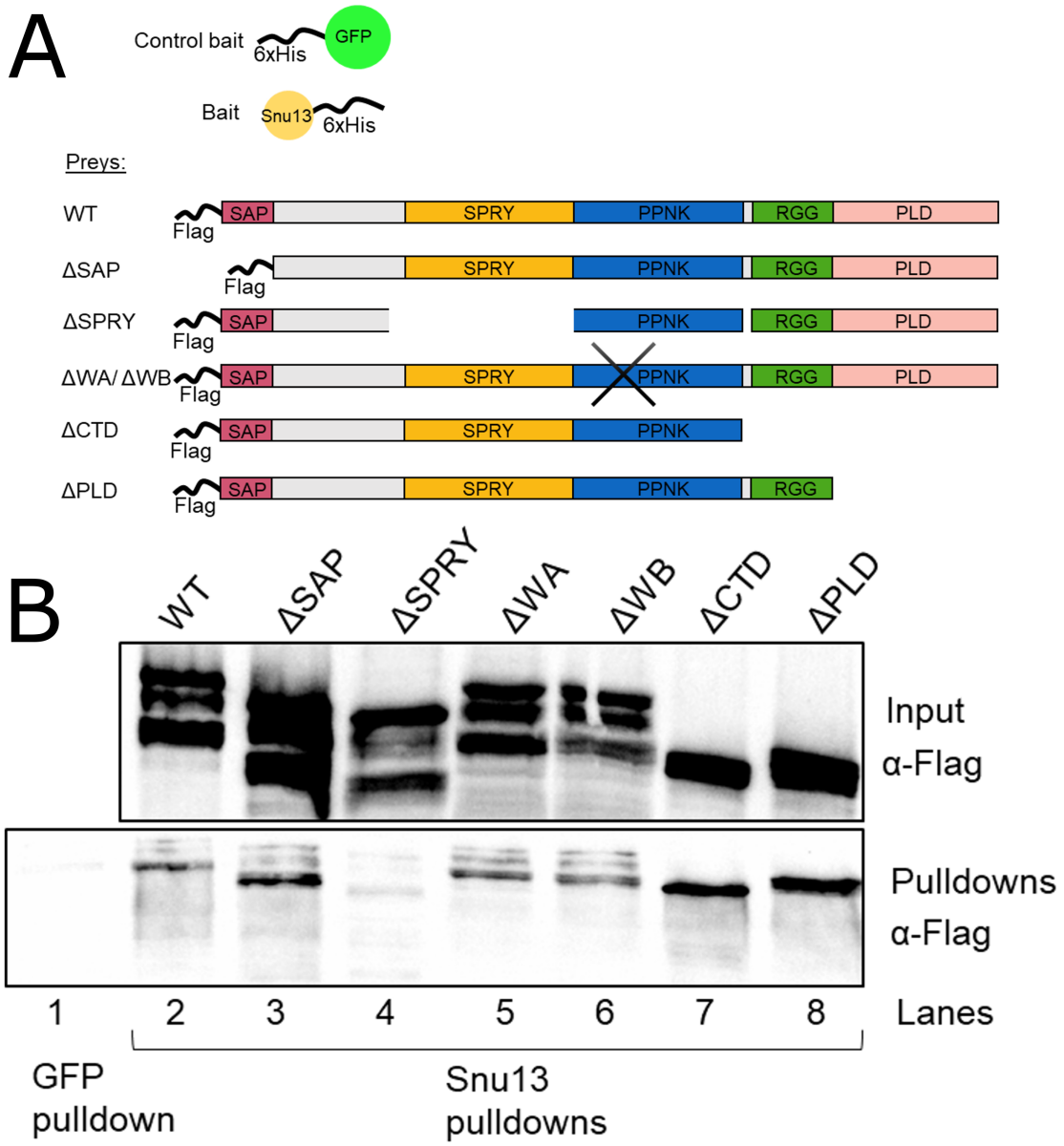

Figure 4.8: Snu13 interaction occurs via the SPRY region of hnRNPUL1. **A)** Schematic of hnRNPUL1 domain organisation and of the proteins used as "bait" and "prey" for the pulldown assay. The "bait" proteins were expressed in *E. coli* BL21 cells, while hnRNPUL1 "prey" variants were overexpressed in HEK 293T cells and whole cell lysates were incubated with "bait" proteins immobilised on beads. **B)** Western blot of pulldown of different hnRNPUL1 variants with Snu13. "Prey" proteins were identified with anti-FLAG antibody.

RNA to proteins in order to avoid disrupting the chemistry of the RNA 5′ ends. Subsequently, the extracted RNA would be processed enzymatically to separate the uncapped population from the total RNA. Treatment with Shrimp Alkaline Phosphatase (rSAP) and T4 Polynucleotide Kinase would generate monophosphorylated 5′ ends only for the uncapped RNA molecules, which could then be ligated to a biotinylated linker and processed to generate an RNA library for sequencing. On the other hand, an additional treatment with RNA 5′ pyrophosphohydrolase (RppH) prior to rSAP and T4 PNK incubation would generate monophosphorylated 5′ ends on all RNA molecules, capped or uncapped, which could be ligated to the biotinylated linker and processed for sequencing in a similar manner.

We employed a ΔCTD truncation of hnRNPUL1 in the RIP assays in order to avoid capturing RNAs associated through the RNA body to the RGG box (Figure 4.9A). GFP was used as a measure of background RNA binding, while CBP20 was used as a positive control for specific RNA 5′ end binding. All three proteins were immunoprecipitated to >90% purity in mild conditions with a single high salt wash (Figure 4.9B). The associated RNA was extracted and first checked by dot blot with an antibody against the m^7G base of the 5′ cap. As shown in Figure 4.9C, while CBP20 enriched for capped RNAs compared to the total RNA input, the hnRNPUL1 ΔCTD truncation did not. There was almost no m^7G signal in the ΔCTD RIP, which indicated an enrichment for uncapped RNA compared to the input, despite equal RNA loading as measured by staining with methylene blue. Before carrying out the enzymatic treatment and library preparation, we analysed the RNA extracted in the GFP and ΔCTD RIPs using a TapeStation, which quantified the abundance of RNA fragments of different lengths. Contrary to our expectations, there was a minimal difference between the ΔCTD RIP and the negative control (Figure 4.9D), with the majority of the signal being ribosomal RNA. This demonstrated that the ΔCTD RIP had predominantly captured background RNA binding due to insufficient washing during immunoprecipitation. Therefore, we decided not to proceed with sequencing of the samples.

To improve the protocol, cellular fractionation could be carried out prior to immunoprecipitation to allow protein extraction only in the nuclear fraction, since

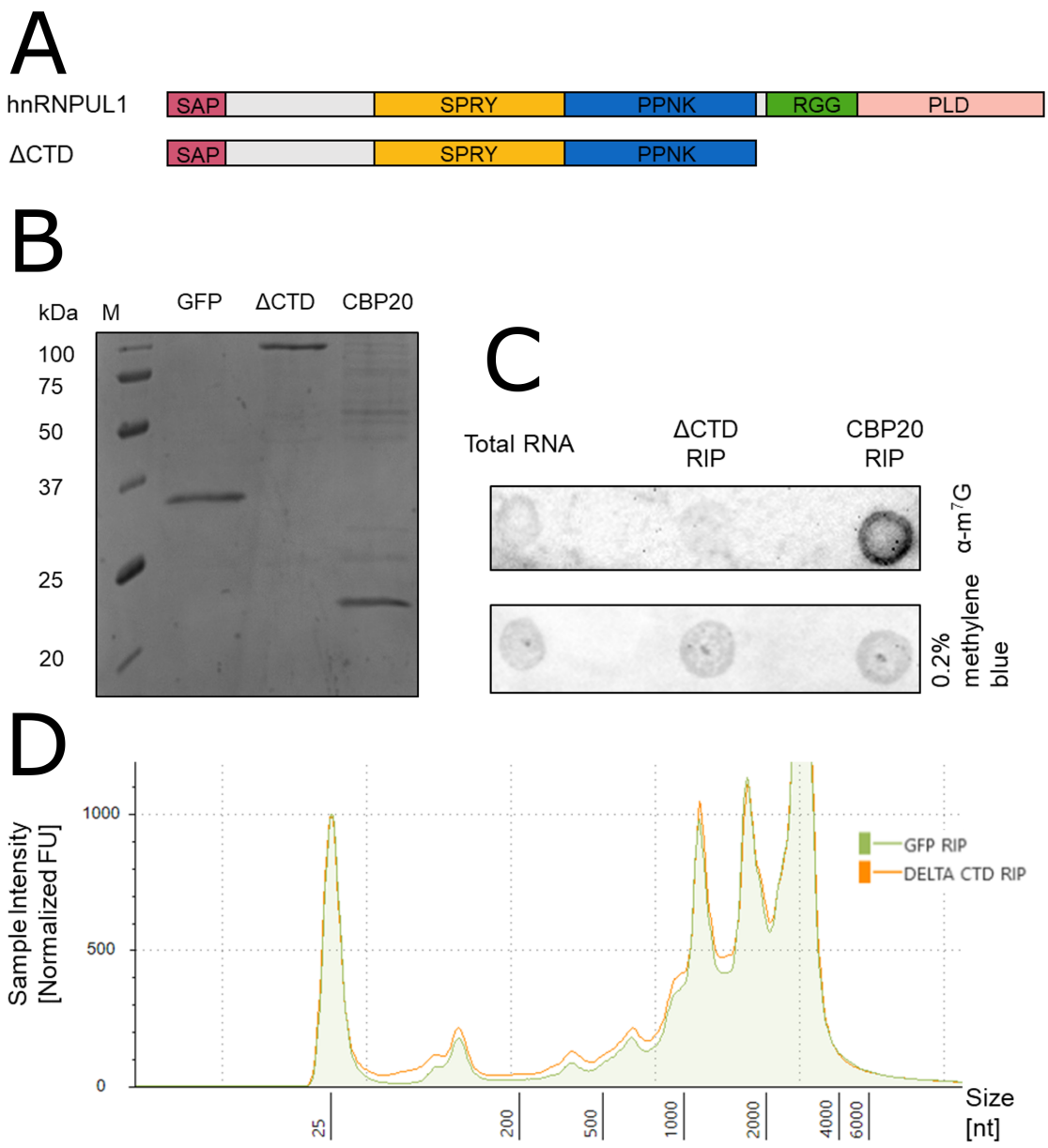

Figure 4.9: No Cap RIP protocol.
A) Schematic of domain organisation of hnRNPUL1 and representation of the ΔCTD truncation used in the No Cap RIP protocol. **B)** Coomassie stain of the proteins immunoprecipitated during No Cap RIP. **C)** Dot blot of RNA obtained from the hnRNPUL1 ΔCTD and CBP20 RIPs, compared with total cellular RNA and probed with m^7G antibody. **D)** TapeStation profile of the RNA obtained from the hnRNPUL1 ΔCTD and GFP RIPs illustrating the abundance of fragments of different sizes. The peaks at ∼ 2,000 and ∼ 4,000 nt correspond to 18S and 28S ribosomal RNA.

hnRNPUL1 is a nuclear protein. This would enable removal of a large amount of the cytoplasmic rRNA, while total depletion could be achieved using commercially available rRNA subtraction kits. However, RNA binding in the PPNK pocket may in reality be a very transient event, so we may be unable to determine the full repertoire of *in vivo* RNA targets of the hnRNPUL1 core with this technique without introducing a cross-linking step.

4.7 hnRNPUL1 interacts with PolII via their respective C-terminal domains

In the previous chapter, we characterised a clear function for the folded SPRY-PPNK core of hnRNPUL1, namely binding triphosphorylated nucleotides and uncapped RNA molecules in a (partially) mutually exclusive manner. Subsequently, we quantified a ligand-dependent conformational change within the core, which may disrupt the intramolecular interaction between the protein core and its C-terminus. In the following two sections, we will try to characterise the function of the CTD of hnRNPUL1 and present a model to encompass the aforementioned observations.

Earlier studies have implicated hnRNPUL1 in PolII transcription regulation and Griffith (2019) identified PolII-transcribed snRNA genes as targets directly affected by hnRNPUL1 depletion (Kzhyshkowska et al. 2003). hnRNPUL1 and PolII both contain disordered C-terminal regions which are often involved in IDR-IDR interactions. Moreover, PolII CTD was observed to form high molecular weight complexes with the unstructured N-terminus of FUS and short RNA molecules (Schwartz et al. 2013). We wondered if hnRNPUL1 could achieve its regulatory role on transcription by directly binding to PolII and more specifically, to its CTD. We also wanted to know if RNA played a role in complex assembly as was reported for FUS.

To answer those questions, we carried out pulldown reactions using the CTD of PolII tagged with GST as "bait" and different truncations of hnRNPUL1 protein as "prey" (Figure 4.10A). Some reactions were incubated with RNAse A to assess the impact of RNA contaminants carried over during protein purification on complex formation. We observed that full-length hnRNPUL1 was directly pulled out of solution by PolII CTD (but not by the GST only control) in an RNA-independent

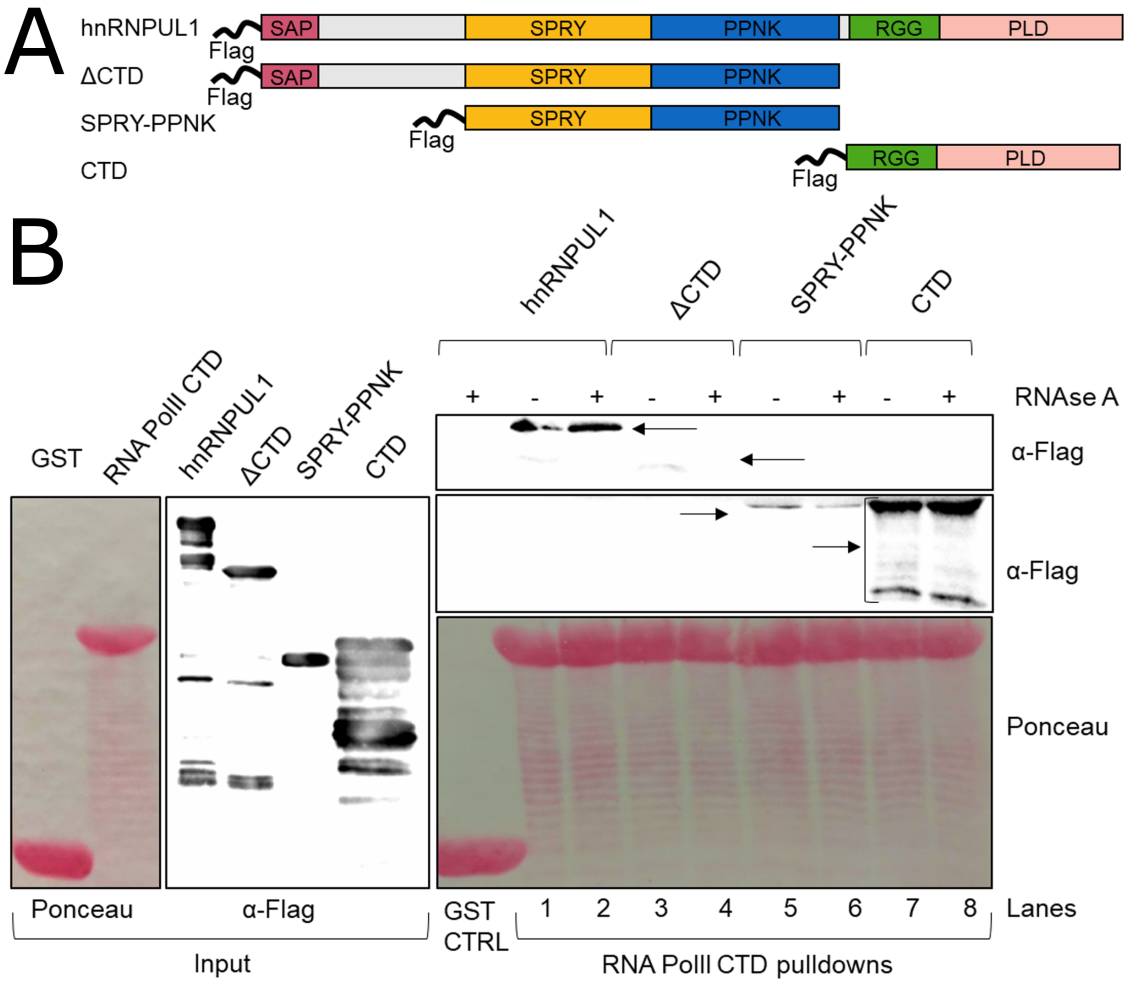

Figure 4.10: hnRNPUL1 interacts with PolII via their respective CTDs. **A)** Schematic of domain organisation of hnRNPUL1 and the FLAG-tagged constructs used as "prey" in the pulldown reactions. All constructs were purified from mammalian 293T cells with a moderate RNAse A treatment step. **B)** Pulldowns using GST-tagged PolII CTD (the first 25 heptad repeats) as "bait" to map the region of hnRNPUL1 involved in PolII binding. RNAse A enzyme was provided in some reactions as indicated. GST was included as a negative control. "Baits" are visualised by Ponceau staining, while "preys" are detected by western blotting and probing with anti-FLAG tag antibody.

manner (Figure 4.10B). However, removal of hnRNPUL1 CTD blocked the protein-protein interaction almost entirely and the same behaviour was observed for the SPRY-PPNK core tested in isolation (lanes 3-6). On the other hand, the CTD of hnRNPUL1 expressed separately was pulled down by PolII CTD in a similar manner to the full-length protein and irrespective of RNAse A treatment. The result demonstrates that hnRNPUL1 and PolII CTD could interact directly and not through other binding partners such as RNA molecules. However, it is possible that other interacting partners of PolII take precedence over hnRNPUL1 in the more complex cellular environment, and that the binding described here was an artifact of the highly controlled experimental setup.

4.8 hnRNPUL1 binds to chromatin via RNA and protein interactions

Following the observation that hnRNPUL1 and PolII CTD interact directly *in vitro* via their disordered regions, we tested the binding in a cellular context. To that end, Dr Llywelyn Griffith designed stable cell lines expressing GFP, hnRNPU or hnRNPUL1 mutants with an N-terminal FLAG tag (Figure 4.11A) and used them to carry out co-immunoprecipitation reactions. The result complemented the *in vitro* observation, with both hnRNPU and -UL1 interacting with PolII and FUS to a similar extent (Figure 4.11B, lanes 7-8). This was consistent with reports that hnRNPU interacted with hypo- and Ser5-phosphorylated PolII at the pre-initiation complex (Kim & Nikodem 1999, Kukalev et al. 2005). Moreover, hnRNPUL1 pulled down TAF15 and EWSR1 more efficiently than hnRNPU, likely due to its longer CTD and the additional PLD region not present in hnRNPU. The experiment demonstrated that hnRNPUL1 binding to the FET proteins and PolII occurred via the disordered C-terminus, because the ΔCTD mutant failed to pull down any interactors, just like the GFP negative control. This was in agreement with the reported RNAse-resistant interaction between hnRNPU/FUS/TAF15/hnRNPUL1 mediated by the hnRNPU CTD and detected by mass spectrometry (Britton et al. 2014). In stark contrast, the ΔWA mutant displayed an enhanced interaction with all the proteins tested. A possible explanation may be that overloading of the ΔWA IP condition caused the apparent enhancement. However, that was not the case because equal levels of

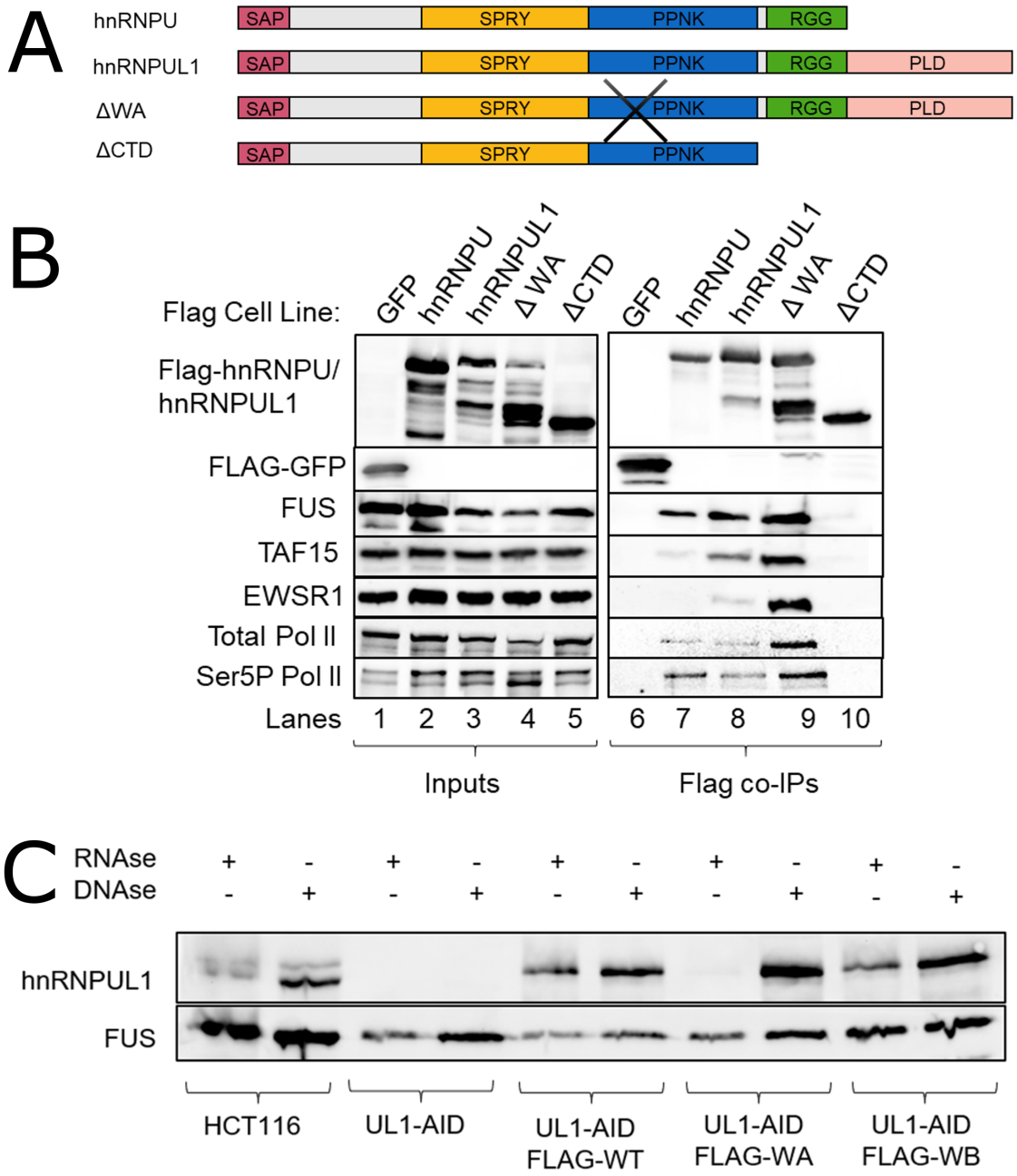

Figure 4.11: Chromatin association of hnRNPUL1 is mediated by protein-RNA and protein-protein interactions.
A) Schematic of domain organisation of hnRNPU and hnRNPUL1 and representation of the mutants used for co-immunoprecipitation. **B)** Western blot of co-IPs of hnRNPU or hnRNPUL1 variants performed by Dr Llywelyn Griffith, probing the interactome with protein-specific antibodies. FLP-In cell lines were used for stable expression of FLAG-tagged versions of GFP, hnRNPU and hnRNPUL1. FLAG-GFP was used as a negative control. **C)** Western blot of hnRNPUL1 protein (WT/ΔWA/ΔWB) released from chromatin by RNAse and subsequent DNAse treatment performed by Dr Ivaylo Yonchev. Chromatin fractions were collected from HCT116 cells (control), hnRNPUL1-AID cells after dox/aux depletion of native protein and hnRNPUL1-AID cells after dox/aux treatment and transfection of hnRNPUL1 WT/ΔWA/ΔWB. Equal volumes from each reaction were used for western blotting and probed with α-hnRNPUL1 antibody. FUS protein was used as a benchmark.

full-length WT and ΔWA were present in both conditions and the dense lower band in the ΔWA lane corresponded to a C-terminal truncation which we had already established to be a null binder.

The behaviour exhibited by the ΔWA mutant was puzzling. We also knew from experience that in order to purify full-length ΔWA from mammalian cells, a harsh DNAse treatment was necessary to release the protein into the soluble fraction. These observations indicated that the ΔWA mutant might be associated with chromatin more tightly than the WT protein. To elucidate the mode of chromatin binding of our hnRNPUL1 mutants, Dr Ivaylo Yonchev performed a chromatin release assay in an auxin-inducible degron cell line (hnRNPUL1-AID) developed by Dr Llywelyn Griffith, which allowed total depletion of hnRNPUL1 upon treatment with doxycycline (dox) and auxin (aux) (Griffith 2019). In the parental cell line, as well as in the hnRNPUL1-AID system transfected with WT protein, around 20% of the WT hnRNPUL1 protein was solubilised by RNAse digestion and around 80% was released by DNAse treatment (Figure 4.11C). No hnRNPUL1 protein was detected in the untransfected hnRNPUL1-AID cells after dox/aux treatment, consistent with total protein depletion. A similar ratio of exogenous ΔWB protein was released by RNAse or RNAse and DNAse digestions as described for the WT. On the other hand, the ΔWA mutant remained unaffected by the RNAse treatment and was only solubilised following DNAse digestion. For comparison, the FUS protein was invariably released by both single or dual enzyme treatments, but its RNAse sensitivity changed between conditions. As such, equal amounts were released from the chromatin fractions of HCT116 and AID cells transfected with the ΔWB mutant, while around 30-40% of FUS was displaced from the untransfected AID cells or the AID cells transfected with WT and ΔWA hnRNPUL1.

We have not been able to detect binding of hnRNPUL1 to double-stranded DNA in bandshift assays (data not shown) and its SAP domain has not been directly characterised as a DNA-binding module in the literature. These results indicate that while some of the WT and ΔWB hnRNPUL1 is attached to chromatin by protein-RNA interactions, the majority of these proteins and all of the ΔWA mutant are anchored by protein-protein interactions mediated by disordered regions such as those found

in the FET proteins and PolII. One way to test if hnRNPUL1 binding to chromatin is mediated by protein-DNA or protein-protein interactions may be treatment with the aliphatic alcohol 1,6-hexanediol, which is routinely used to disrupt hydrophobic protein-protein interactions.

4.9 Summary

In this chapter, I have shown evidence of an intramolecular interaction which may serve as a self-inhibitory mechanism. Initially, we undertook tryptophan fluorescence quenching assays to measure the tightness of association between hnRNPUL1 and its identified ligands (Figure 4.1). The assays demonstrated that a conformational change occurred within the WT and ΔWB proteins upon ligand binding, but not the ΔWA. This was in agreement with the UV crosslinking results where ΔWA was a poor ATP and RNA binder. Additionally, the Kd measured for WT hnRNPUL1 binding to RNA was similar to that of PNKP, while the Kd for ATP binding was 10 times lower than that of PNKP. On the other hand, the Kds measured for ΔWB binding to either substrate were both 8 times lower than the WT values, supporting the tighter ligand association observed with UV crosslinking. A potential conformational change around the PPNK domain of hnRNPUL1 was modelled based on the apo- and liganded forms of PNKP, which suggested that the conformational change was mostly localised around the binding pocket (Figure 4.2).

Next, we noted that although the core and the CTD of hnRNPUL1 could bind RNA independently, mutating the Walker A motif in the context of the full-length protein severely disrupted its RNA binding potential (Figure 4.3). This observation led us to investigate a potential interaction between the two RNA binding modules of hnRNPUL1 using a pulldown assay, which confirmed the protein-protein interaction hypothesis (Figure 4.4). Subsequently, we identified a potential interaction site on the surface of the SPRY-PPNK core above the binding pocket, consisting of numerous acidic residues which could bind to the RGG box through charge complementarity (Figure 4.5). The negatively-charged area span around 10% of the core's surface and was located in close proximity to the flexible region which undergoes a conformational change upon ligand binding, as demonstrated by the tryptophan

fluorescence quenching. This feature may allow the hnRNPUL1 CTD to associate with the core in the apo form and to be displaced by ligand binding to the PPNK pocket following a conformational change. This mechanism may regulate when the RGG box is available for RNA binding.

Next, we looked for *in vivo* RNA targets of hnRNPUL1 in the publicly available ENCODE database and identified U4 snRNA as a binding substrate (Figure 4.6). Previous work in the lab has already shown that hnRNPUL1 interacts with U4 snRNA and is involved in its transcription. There was eCLIP signal on the first two nucleotides, which could arise from binding to the PPNK pocket, as well as signal on the 5′ stem loop, immediately adjacent to the binding site for the Snu13 protein. We uncovered that hnRNPUL1 could be recruited to the mature U4 snRNP by Snu13 (Figure 4.7) and that this interaction relied upon the presence of an intact but not necessarily functional SPRY-PPNK core (Figure 4.8).

Based on the earlier identification of eCLIP signal on the first two bases of U4 snRNA, which we believed to be bound through the PPNK pocket, we developed a protocol for capturing uncapped RNA targets associated with the hnRNPUL1 core named No Cap RIP. The advantages of the protocol were that it did not involve protein-RNA crosslinking and it would enable separation of RNA molecules into capped and uncapped populations based on enzymatic treatment. Unfortunately, the main RNA species obtained in the hnRNPUL1 No Cap RIP were cytoplasmic rRNA contaminants, which were indistinguishable from the negative control condition (Figure 4.9). Because of time limitations, further optimisation of the protocol was abandoned.

Lastly, we characterised a function for the low-complexity unstructured CTD of hnRNPUL1 in mediating disordered protein interaction with PolII, both *in vitro* by pulldown assays and *in vivo* with co-immunoprecipitations. We observed that the CTD is the site of interactions with other disordered proteins, such as the FET family, and that the ΔWA mutant exhibited a much stronger interaction with both the FET proteins and PolII. Upon further investigation, it was revealed that this mutant was essentially "trapped" on chromatin and it failed to be released by RNAse digestion to any extent, unlike the WT and ΔWB proteins which were partially

sensitive to RNAse treatment. This observation indicated a bi-modal mechanism of chromatin association for hnRNPUL1, with a smaller population linked through protein-RNA interactions, and a larger population associated mainly via protein-protein interactions (which we termed "chromatin-trapped"). Transition between the two states appeared to be linked to the ability of hnRNPUL1 to bind RNA and NTP molecules through its core region, since the ΔWA mutant remained permanently trapped. The functional implications of these observations will be explored further in the Discussion chapter.

5 Chapter 5: hnRNPUL1 undergoes phase separation regulated by post translational modifications

In the previous chapter, I presented evidence for the role of hnRNPUL1 CTD in interacting with proteins with low complexity domains such as PolII and the FET protein family. Numerous studies have analysed proteins with similar unstructured regions, including FUS, EWSR1, TAF15, hnRNPA1, hnRNPA3, hnRNPL and TDP43 from the perspective of liquid-liquid phase separation (LLPS) and have identified hallmarks indicative of LLPS (Burke et al. 2015, Chong et al. 2022, Kwon et al. 2014, Xiang et al. 2015, Wang et al. 2018b).

A common feature of proteins that undergo phase separation is the presence of RNA-binding domains and unfolded regions that contain a limited repertoire of amino acids (called intrinsically-disordered regions, IDRs, or prion-like domains, PLDs) (Hennig et al. 2015, Martin et al. 2020, Wang et al. 2018b). Along with a low diversity of amino acids, PLDs are often distinguished by enrichment in Gln/Asn residues. The online platform PLAAC (Prion-Like Amino Acid Composition) was developed to scan input protein sequences for prion signatures in the amino acid composition (Alberti et al. 2009, Lancaster et al. 2014). Using PLAAC, we investigated the hnRNPUL1 sequence and uncovered a characteristic prion-like signature at the C-terminus (Figure 5.1). The analysis also highlighted the low amino acid diversity in the PLD region, with an over-representation of Pro, Ser, Tyr, Asn and Gln and an under-representation of hydrophobic or charged residues.

Moreover, proteins that undergo IDR-mediated LLPS tend to be enriched in Arg and Tyr residues within their low complexity regions, with Arg distributed mostly in the RNA-binding regions and the Tyr being prevalent in the PLDs (Wang et al. 2018b). Additionally, the sequence motifs (Ser/Gly)Tyr(Ser/Gly) form cross-β sheet or reversible amyloid fibrils and are involved in maintaining the delicate architecture of LLPS droplets (Guenther et al. 2018, Hughes et al. 2018, Lu et al. 2020, Luo et al. 2018). Most often, these motifs are found in prion-like domains (PLD), so-called because of their tendency to form reversible amyloid fibres through

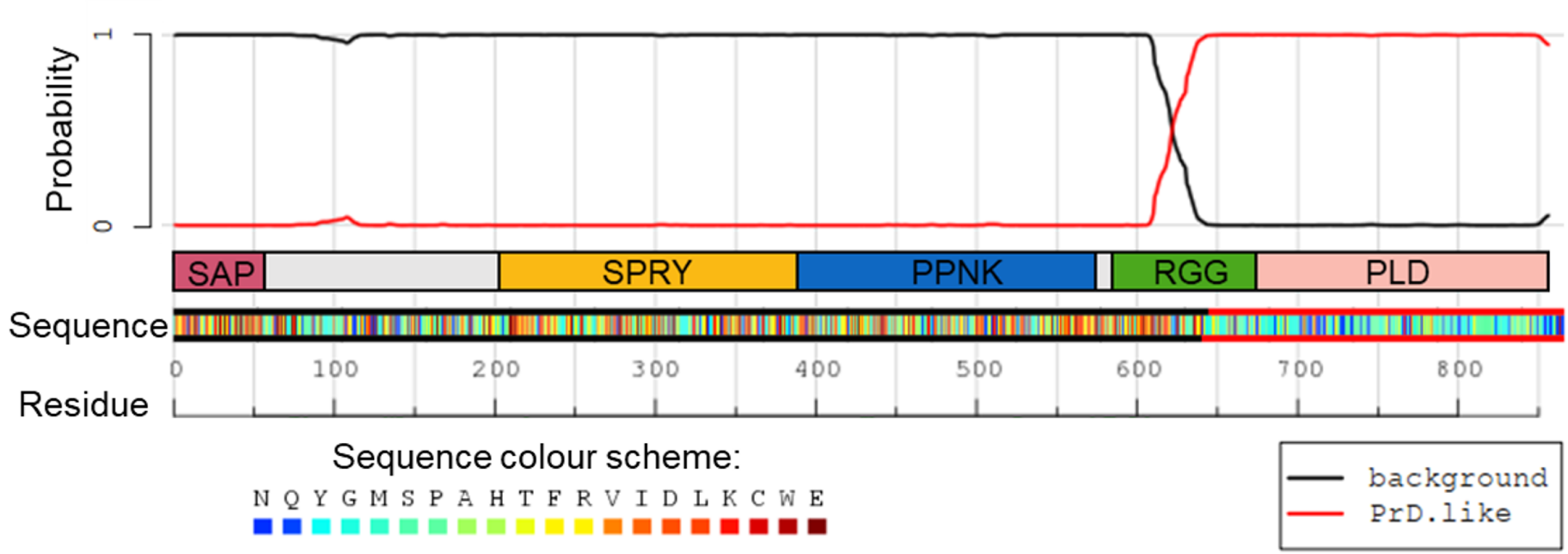

Figure 5.1: PLAAC analysis of the hnRNPUL1 amino acid sequence.
Top: probability of a region of 60 amino acids (termed "core") falling into either the prion-like category (red trace) or the background category (black trace). The probabilities were calculated based on both yeast and human background amino acid frequencies with the same result. Bottom: colour-coded representation of hnRNPUL1 amino acid composition based on the colour scheme provided.

the (Ser/Gly)Tyr(Ser/Gly) regions. For comparison, the FUS PLD contains 21 (Ser/Gly)Tyr(Ser/Gly) motifs (19 singles and 2 SYSGYS repeats), while the hnRNPUL1 PLD contains 11 motifs (9 singles and 2 repeats). Hughes et al. (2018) carried out a proteome-wide search for short segments predicted by Rosetta to adopt a similar orientation as the FUS SYSGYS double motif. hnRNPUL1 was among the proteins identified by the study, with three of these regions residing in the CTD. Moreover, hnRNPUL1 was identified in a mass spectrometry experiment after precipitation from nuclear extracts with the chemical biotinylated isoxazole (b-isox), which is believed to selectively precipitate proteins with unstructured regions that adopt a β-strand conformation (Kwon et al. 2014).

Based on the evidence presented above, including the presence of a prion-like domain adjacent to the RGG box, a skewed amino acid distribution of Arg and Tyr within the RNA-binding domain and the PLD and the presence of (Ser/Gly)Tyr(Ser/Gly) motifs characteristic of phase separation, we decided to investigate the ability of hnRNPUL1 to undergo LLPS *in vitro*.

5.1 The hnRNPUL1 CTD undergoes phase separation

The difficulty of expressing full-length hnRNPUL1 in bacterial cells and of generating sufficient amounts in mammalian systems compelled us to study LLPS using the C-terminal disordered region in isolation. The decision was also based on the fact that phase separation often relies on intra and intermolecular interactions between intrinsically-disordered regions and RNA-binding domains (Borcherds et al. 2021), such as those forming the CTD of hnRNPUL1. Therefore, the relevant protein region was expressed in *E. coli* with an N-terminal GFP tag (Figure 5.2A). The protein purification and LLPS conditions were adapted from Kato et al. (2017), Wang et al. (2018b) and Murakami et al. (2015).

First, we tested different buffer conditions to optimise the LLPS protocol for hnRNPUL1. We selected a set protein concentration of 4 μM, which was in the concentration range where FET proteins underwent LLPS (5 μM for FUS, 2.75 μM for EWSR1 and 2 μM for TAF15). As shown in Figure 5.2B top, only a limited number of droplets were observed during the three-hour incubation, with most of the

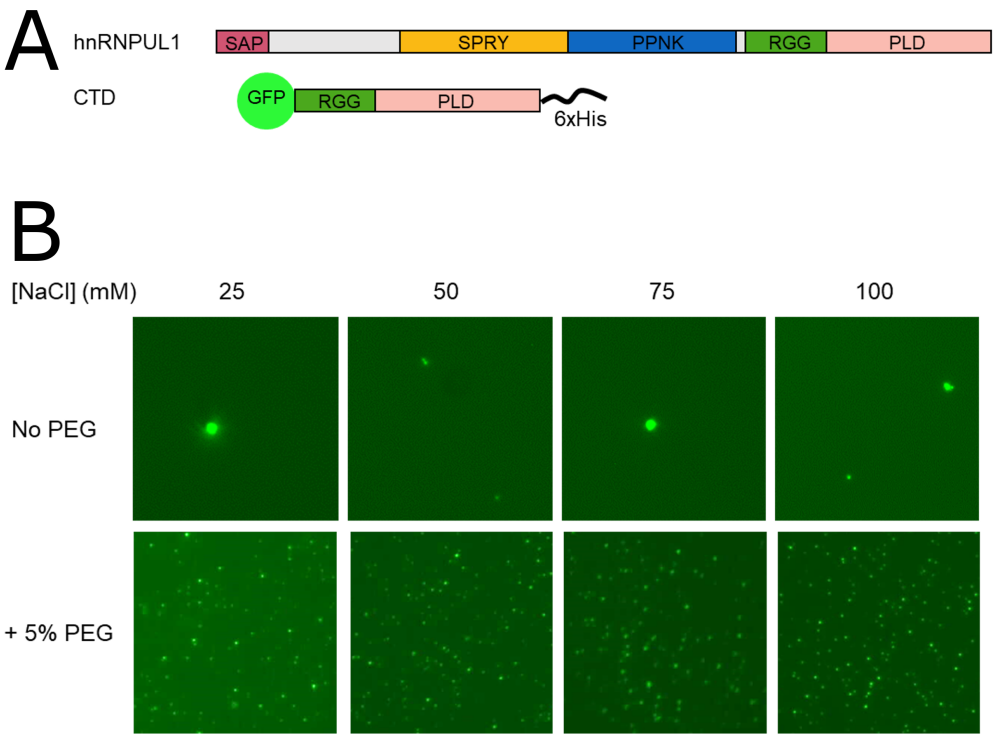

Figure 5.2: Optimisation of phase separation conditions for hnRNPUL1 CTD.
A) Schematic of hnRNPUL1 domain organisation and of the GFP-tagged CTD used in the phase separation assays below. **B)** Top: Fluorescent pictures of phase separation tests at different salt concentrations as indicated. The protein concentration was kept constant at 4 μM. Reactions were incubated for up to 3 hr at room temperature and the droplets allowed to settle on the glass slide before visualising at 63x magnification on a Leika microscope. Bottom: Fluorescent pictures of phase separation tests using the same conditions as before, with the addition of 5% PEG 8,000 as a crowding agent.

green fluorescent signal remaining dispersed in solution rather than concentrated in discrete LLPS droplets.

A common method for inducing LLPS for difficult proteins involves the use of crowding agents, which are inert macromolecular substances that reduce the available volume and effectively increase protein-protein interactions. The most widely used crowding agents include poly(ethylene) glycol (PEG), dextran, bovine serum albumin (BSA), lysozyme, haemoglobin and Ficoll (Kuznetsova et al. 2014). Therefore, we examined the effects of PEG 8,000 to the earlier LLPS conditions (Figure 5.2B bottom). Remarkably, supplementing the LLPS buffers with 5% PEG promoted the formation of numerous droplets in all salt conditions tested. The optimal buffer composition was determined to be 100 mM NaCl, 50 mM Tris pH 8.0 and 5% PEG, which was employed for all subsequent LLPS assays.

5.1.1 Determining the saturation concentration for hnRNPUL1 phase separation

Next, we wanted to determine the minimal protein concentration at which the hnRNPUL1 CTD began to form LLPS droplets, known as the saturation concentration or c_{sat}. Additionally, we analysed LLPS of the prion-like domain expressed separately as a GFP fusion to understand the impact of the RGG-PLD interactions on phase separation (Figure 5.3A). Condensates started to appear around 1 and 5 μM for the CTD and PLD constructs, respectively, corresponding for the c_{sat} for each protein. Increasing the concentration of CTD-GFP up to 6 μM lead the formation of more droplets of the same size and to a reduction of background signal at the highest concentration, indicating that the majority of the protein was sequestered inside LLPS droplets (Figure 5.3B top). On the other hand, the PLD-GFP construct formed larger droplets starting at 20 μM, which grew in size up to the highest concentration tested (50 μM, Figure 5.3B bottom). Importantly, the c_{sat} of hnRNPUL1 CTD (1 μM) was around the estimated nuclear concentration of hnRNPUL1 protein of 1.4 μM (Hein et al. 2015).

It has been reported previously that individual PLD constructs begin phase separating at much higher concentrations in the absence of the associated RNA-binding

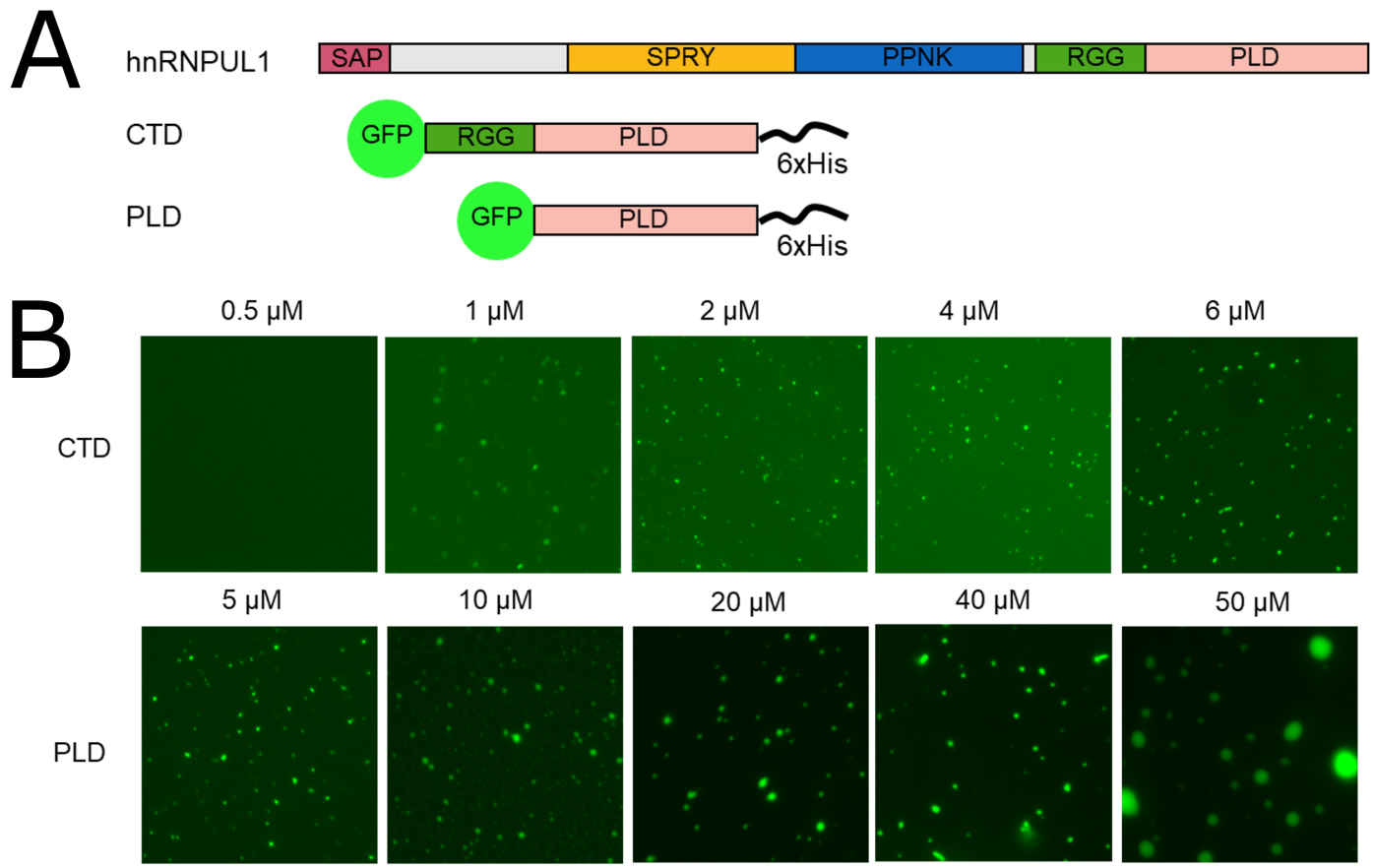

Figure 5.3: Determining the saturation concentration for hnRNPUL1 CTD and PLD.
A) Schematic of hnRNPUL1 domain organisation and of the GFP-tagged proteins used in the phase separation assays below. **B)** Fluorescent pictures of phase separation tests spanning a range of concentrations for GFP-CTD and GFP-PLD to determine the protein concentration at which LLPS droplets began to form. Reactions were carried out in 50 mM Tris pH 8.0 + 100 mM NaCl buffer with 5% PEG.

domain. For instance, the c_{sat} for FUS PLD is over 24 times higher than for full-length FUS (>120 μM compared to 5 μM)(Wang et al. 2018b). Additionally, the c_{sat} could be lowered by mixing the FUS RBD and PLD expressed separately in a 1:1 ratio or even by mixing the RBD and PLD from different proteins (Rawat et al. 2021, Wang et al. 2018b). Our observation that the PLD of hnRNPUL1 had a c_{sat} around 5 times higher than that of the entire CTD was in agreement with the published literature.

The molecular basis for this difference may relate to the different protein-protein interactions that underpin CTD and PLD phase separated droplets. Based on the published literature, we believe that interactions between the Arg residues in the RGG box and Tyr residues in the PLD help to drive phase separation of the CTD, whereas weaker Tyr-Tyr interactions contribute to phase separation of the PLD. This model will be explored in more detail in section 5.2.

5.1.2 hnRNPUL1 CTD droplets are dissolved by 1,6-hexanediol

A feature of many LLPS droplets is solubility in the aliphatic alcohol 1,6-hexanediol, which is often used as a control to delineate droplets formed by phase separation from protein aggregates. For this reason, we screened both CTD and PLD droplets using a 1,6-hexanediol treatment (Figure 5.4A) either at the beginning of the LLPS incubation (*de novo*) or after droplet formation (pre-formed).

1,6-hexanediol prevented droplet formation for both proteins when added to the reaction mixture at the beginning of the incubation (Figure 5.4B, middle panels) and also dissolved the pre-formed CTD droplets (Figure 5.4B, CTD right). On the other hand, more modest effects were observed for pre-formed PLD droplets, which actually appeared to adhere to each other and become slightly enlarged following 1,6-hexanediol treatment (Figure 5.4B, PLD right).

The range of effects of 1,6-hexanediol on the two protein types may be attributed to a different molecular basis for phase separation in each case, as discussed in section 5.2. However, it is noteworthy that there are shortcomings with the 1,6-hexanediol treatment as a tool to characterise LLPS. Firstly, 1,6-hexanediol does not dissolve all types of protein-protein and protein-RNA interactions thought to underlie LLPS,

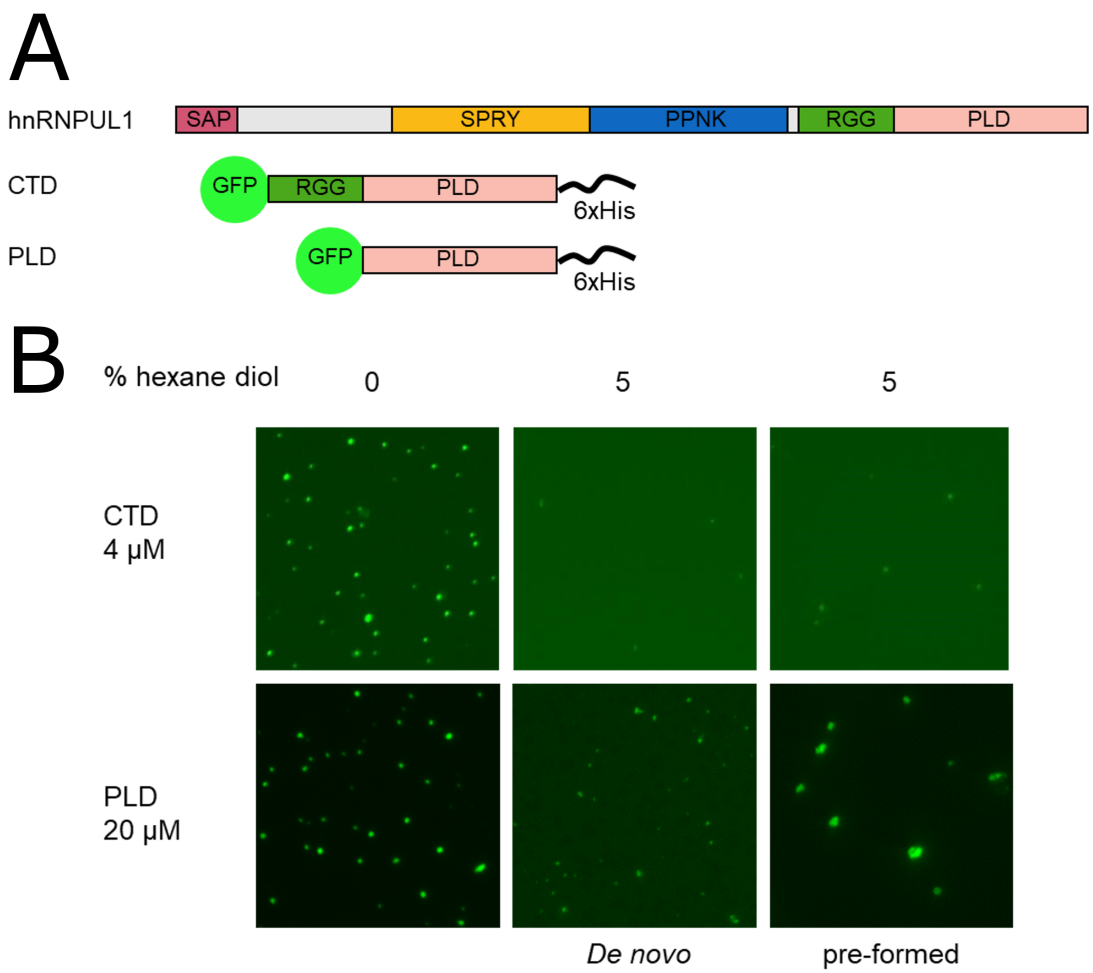

Figure 5.4: 1,6-hexanediol dissolves hnRNPUL1 CTD droplets and prevents LLPS.
A) Schematic of hnRNPUL1 domain organisation and of the GFP-tagged proteins used in the phase separation assays below. **B)** 1,6-hexanediol treatment of preformed CTD and PLD droplets (right panels) and incubation with CTD and PLD from the start of LLPS assays (middle panels). Reactions were carried at 4 μM GFP-CTD and 20 μM GFP-PLD protein concentrations.

which include π-π, cation-π and electrostatic interactions. This was most likely the reason why pre-formed PLD droplets were not dissolved by 1,6-hexanediol. Secondly, this treatment *in vivo* disrupts physiological pathways and may produce misleading results when used in living cells. The reported effects of 1,6-hexanediol treatment included changes in membrane permeability and an impaired kinase and phosphatase function (Kroschwald et al. 2017, Düster et al. 2021).

5.2 Cation-π interactions between Arg-Tyr underpin hnRNPUL1 CTD phase separation

The molecular basis of intrinsically-disordered regions undergoing LLPS is very protein-specific and a topic of active research. A combination of primary amino acid sequence and secondary structure elements mediate the physical and material properties of LLPS droplets. The phase behaviour of the highly studied FUS protein is underpinned by interactions between Tyr residues in the low-complexity N-terminal region and Arg residues in the RNA-binding domains, also known as cation-π interactions (Qamar et al. 2018, Wang et al. 2018b).

The primary sequence of the hnRNPUL1 CTD displayed a similar amino acid distribution pattern, entailing an Arg-rich RNA-binding domain (the RGG box) and a Tyr-rich low-complexity PLD. This provided an appealing avenue for investigating if hnRNPUL1 and FUS shared a common basis for LLPS. Mutant versions of the hnRNPUL1 CTD were designed where either the arginines were changed to lysines (R→K) or the tyrosines to phenylalanine (Y→F) as shown in Figure 5.5A. Consistent with earlier reports for the FUS phase behaviour, we observed a drastic loss of potential for LLPS in both R→K and Y→F mutants compared to the WT protein (Figure 5.5B). The R→K mutant gave rise to a few small droplets of a circular shape and uniform distribution (see insets), while the Y→F protein failed to produce any droplets under similar conditions.

Next, we tested the phase behaviour of the PLD region after introducing the Y→F mutations compared to WT (Figure 5.5C). An equivalent set of mutations in the FUS PLD ablated the region's ability to form LLPS droplets, leading the authors to suggest that Tyr-Tyr interactions underpin phase separation by PLD regions

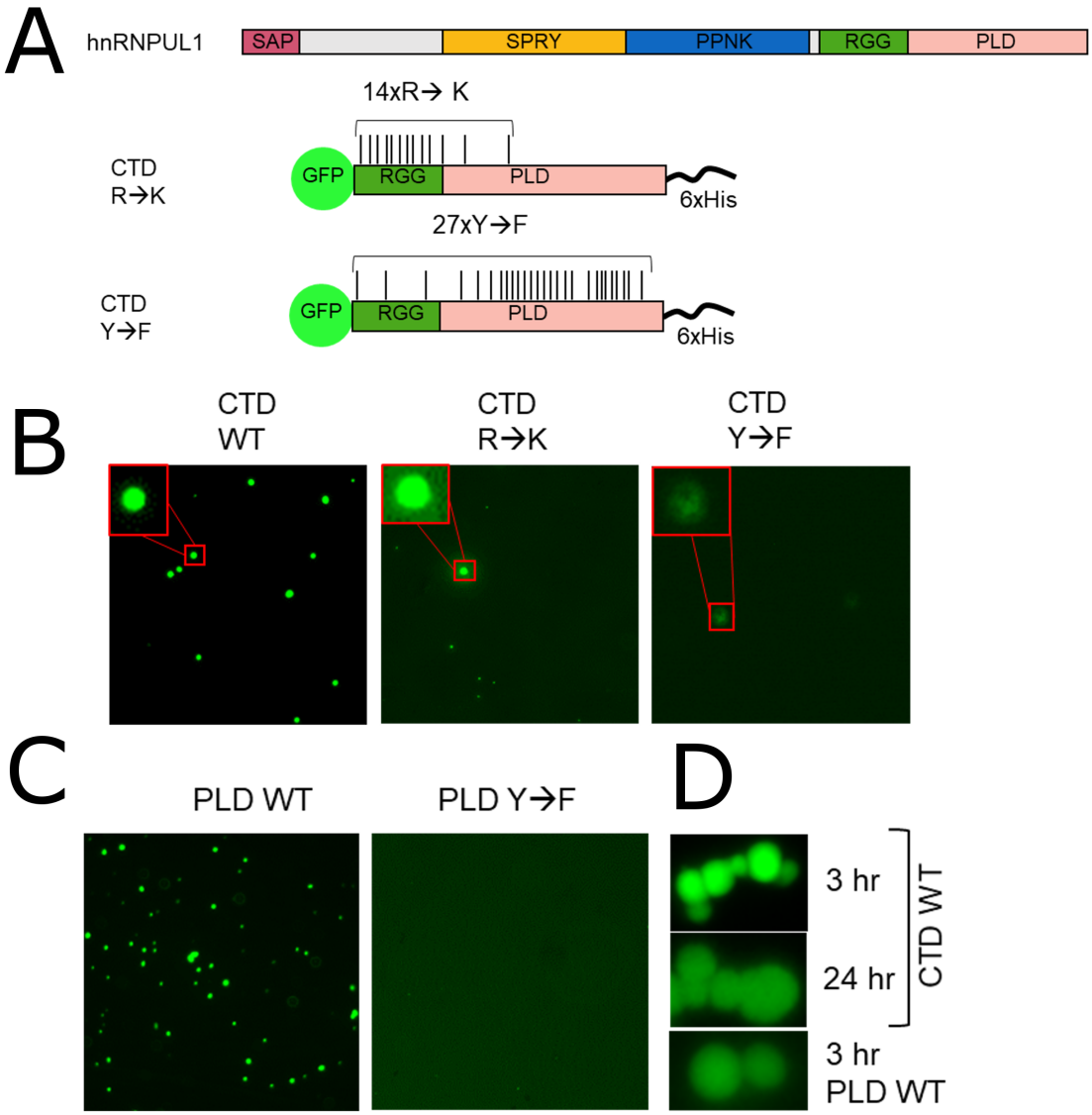

Figure 5.5: Arg-Tyr interactions govern hnRNPUL1 CTD phase separation.
A) Schematic of hnRNPUL1 domain organisation and of the GFP-tagged proteins used in the phase separation assays below. Bars indicate the position of the Arg or Tyr residues mutated to Lys or Phe in the R→K and Y→F mutants, respectively. **B)** Fluorescent pictures of phase separation test of the R→K and Y→F mutants compared to CTD wild-type at 5 μM protein concentration. The insets show representative protein droplets from each condition. **C)** LLPS test of WT or Y→F mutated PLD region at 10 μM protein concentration. **D)** LLPS droplets formed by 20 μM WT hnRNPUL1 CTD and 50 μM hnRNPUL1 PLD adhered to each other but did not coalesce during a 24 hr incubation.

(Wang et al. 2018b). Consistent with this data, the hnRNPUL1 PLD version carrying the Y→F mutations failed to produce any droplets under the same conditions as the WT protein (Figure 5.5C). Our data support the idea that weaker π-π interactions between Tyr residues underpin PLD phase separation, which may explain the 5 times higher c_{sat} of this region alone compared to that of the entire CTD. This may also explain why 1,6-hexanediol failed to dissolve PLD, but not CTD droplets, owing to a different mode of protein self-interaction. We cannot exclude the possibility that mutating the Tyr residues affected the cross-β sheet structure adopted by (G/S)Y(G/S) motifs which have been implicated in promoting phase transition.

Lastly, we investigated the mechanical properties of WT CTD droplets. FRAP assays and measurements of the time taken for two droplets to merge are commonly employed in order to characterise the fluidity of LLPS droplets. However, some proteins form droplets that undergo liquid-to-solid transitions due to experimental conditions and/or the amino acid sequence of the protein (Alberti et al. 2019). At high protein concentrations (20 μM CTD and 70 μM PLD), the hnRNPUL1 droplets adhered to one another but did not merge over the course of 24 hours (Figure 5.5D). Such behaviour was indicative of liquid-to-solid transition or hydrogel conversion and consistently, we failed to detect recovery of fluorescent signal in a later FRAP experiment (data not shown). A similar observation of droplet behaviour was reported for full-length TDP43 protein (Wang et al. 2018a).

The mechanical properties of FUS droplets are governed by Gly, Ser and Gln residues in the PLD region, without affecting the saturation concentration (Wang et al. 2018b). In the table below, the prevalence of constituent amino acids for the PLDs of FUS and hnRNPUL1 are compared. The most striking difference between the two PLDs was the proline content, which accounted for almost 30% of the hnRNPUL1 PLD and only 5% of the FUS PLD. Due to its sidechain configuration, Pro restricts the local flexibility of the polypeptide backbone. Furthermore, there was a marked reduction in the proportions of Gly, Gln and Ser for hnRNPUL1 compared to FUS PLD. The number of Gly residues directly correlates with droplet fluidity, while Gln and Ser are proposed to govern cross-β fibre formation. Overall, the stiffer and less

dynamic appearance of hnRNPUL1 CTD droplets may be caused by the reduced proportion of Gly and the over-representation of Pro.

hnRNPUL1 PLD Percentage	Amino acid	FUS PLD Percentage
27.4	Pro	5.2
14.9	Gln	20.4
13.7	Tyr	12.8
10.7	Ser	24.2
8.3	Gly	24.6
8.3	Asn	3.3
6.5	Thr	4.7
5.4	Ala	1.9
1.2	Arg	0
1.2	Trp	0
0.6	Lys	0
0.6	Ile	0
0.6	Val	0
0.6	Glu	1.9
0	Met	0.9

5.3 ATP modulates the hnRNPUL1 phase separation behaviour

Previous studies have reported that the physiological concentration of ATP had a negative impact on FUS, TAF15 and hnRNPA3 phase separation, while lower concentrations (up to 1 mM) enhanced the size and number of protein droplets (Kang et al. 2018, Patel et al. 2017). ATP concentration varies across tissue types, however it is always measured in the millimolar range, far exceeding the energy needs of a cell. Tissues commonly accumulate between 2.7 and 4 mM, although cardiac and skeletal muscles can store over 7 mM ATP (Greiner & Glonek 2021). It is noteworthy that excess of intracellular ATP is a feature found in eukaryotic, prokaryotic and archaeal cells (Greiner & Glonek 2021). A current line of thought proposed that in addition to being an energy source, the vast excess of ATP stored in the cell may serve to maintain protein solubility and restrict non productive protein aggregation or phase separation (Patel et al. 2017).

The evidence for ATP playing an important role in maintaining protein solubility

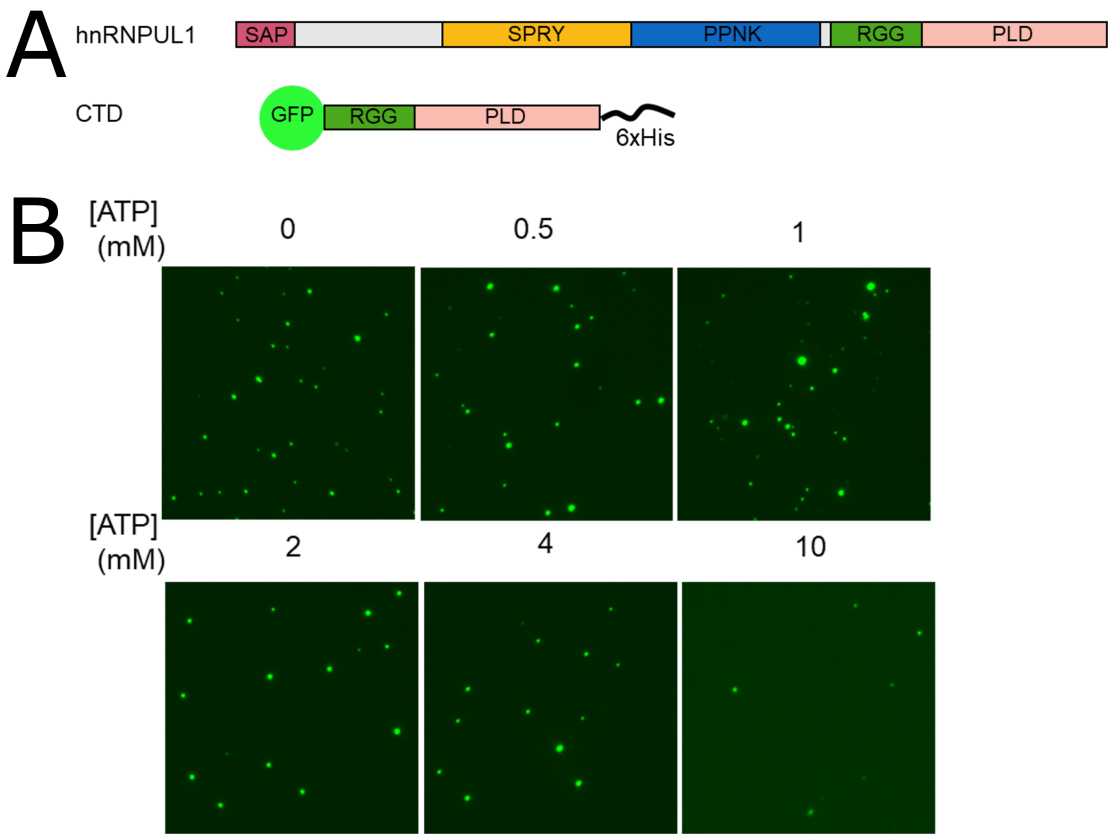

Figure 5.6: ATP concentration modulates the hnRNPUL1 phase separation behaviour.
A) Schematic of hnRNPUL1 domain organisation and of the GFP-tagged construct used in the phase separation assays below. **B)** Fluorescent pictures of phase separation assays testing the effect of different concentrations of ATP on droplet formation. The assays were carried out at 6 μM protein and in the same LLPS buffer as before, supplemented with 10 mM $MgCl_2$ and the specified concentration of ATP.

compelled us to test the effects of different concentrations of ATP in complex with Mg^{2+} on hnRNPUL1 droplet formation (Figure 5.6A). We found that at low concentrations (up to 1 mM), ATP promoted phase separation and increased droplet size (Figure 5.6B top). However, increasing the amount of ATP to physiological concentrations (between 2 and 10 mM) gradually reduced hnRNPUL1 phase separation in a dose-dependent manner, leading to complete solubilization of the protein at 10 mM ATP (Figure 5.6B bottom).

An important caveat is that the actual distribution of ATP inside sub-cellular compartments is unknown, but differences in local ATP concentration may affect protein solubility. Importantly, phase separated assemblies observed *in vivo* contain a highly diverse repertoire of RNPs and proteins, including ATPases. For instance, components of stress granules and P bodies include the protein chaperone CCT complex and helicase complexes MCM and RVB (Jain et al. 2016). These ATPases are believed to regulate granule formation and stability by remodelling RNP structure and accessibility. However, their action may also lead to a local reduction in ATP concentration, increasing the stability of the protein condensates via an additional mechanism.

5.4 RNA Polymerase II CTD co-localises with hnRNPUL1 CTD droplets

Phase separation has been implicated in numerous biological process in recent years, transcription regulation being one of the most prominent (Cho et al. 2018, Chong et al. 2022, Sabari et al. 2018). The CTD of the large Rpb1 subunit of RNA polymerase II (PolII) is intrinsically disordered and comprises 52 heptad repeats with the consensus sequence TyrSerProThrSerProSer. Its interaction with transcription regulators such as NELF (Rawat et al. 2021) and the FET proteins (Burke et al. 2015, Kwon et al. 2014) via LLPS is well documented in the literature. Having already characterised the CTD of hnRNPUL1 as a crucial site for interaction with PolII and the FET proteins in section 4.8, as well as having the ability to undergo phase separation, we wanted to test if it could interact with PolII via LLPS as reported for the FET proteins.

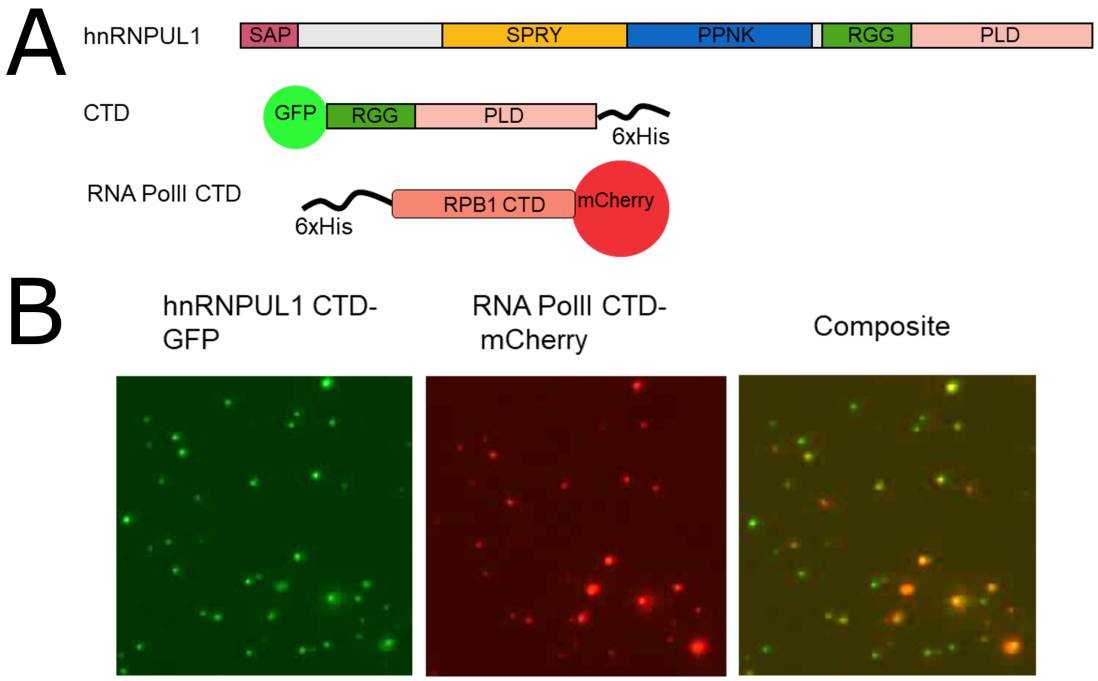

Figure 5.7: PolII CTD co-localises with hnRNPUL1 CTD droplets. **A)** Schematic of hnRNPUL1 domain organisation and of the GFP- and mCherry-tagged proteins used in the phase separation assays below. **B)** Co-localisation test where hnRNPUL1 CTD and PolII CTD were incubated in a ratio of 10:1 (4 μM hnRNPUL1 CTD and 0.4 μM PolII CTD) in the phase separation mixture.

To that end, we generated a C-terminal mCherry fusion of the last 22 degenerate heptad repeats of PolII Rpb1 CTD and expressed it in *E. coli* (Figure 5.7A). That particular region of PolII CTD (and not the first 26 conserved repeats) was identified to interact via LLPS with other proteins with intrinsically-disordered regions (Kwon et al. 2014, Rawat et al. 2021). Next, we incubated the N-terminal GFP fusion of hnRNPUL1 CTD with the C-terminal mCherry fusion of PolII CTD in a 10:1 ratio in the same experimental conditions as before and visualised the droplets in the red and green channels to detect each protein. PolII co-localised with the hnRNPUL1 CTD droplets (Figure 5.7B). In the composite image, it was clear that the hnRNPUL1 droplets recruited PolII CTD in varying amounts, demonstrating that the red signal was not bleed-through from the green channel, as that would cause a uniform mCherry signal across all droplets. A construct of the last 26 degenerate repeats of PolII CTD failed to phase separate in a homotypic mixture at concentrations of up to 500 μM (Burke et al. 2015), suggesting that the mCherry-tagged protein used here is imported into LLPS droplets by virtue of interactions with hnRNPUL1 CTD.

The result supports a mode of interaction between hnRNPUL1 and PolII mediated by LLPS. Burke et al. (2015) provided evidence that PolII CTD preferentially interacts with phase separated droplets of FUS PLD instead of PLD monomers. However, these highly controlled *in vitro* assays are unlikely to illustrate the complexity of protein interactions inside the nucleus. A combination of factors with intrinsically-disordered regions will certainly contribute to recruiting PolII to relevant sites on the chromatin. Our data allow us to propose that hnRNPUL1 may be one of the factors that target PolII to specific promoters such as the PolII-transcribed snRNA genes, whose transcription is downregulated by loss of hnRNPUL1 (Griffith 2019). A further limitation of the co-localisation assay presented here is the lack of post-translational modifications (PTMs) that hnRNPUL1 and PolII would normally acquire, most notably Arg methylation and Ser and Tyr phosphorylation. For this reason, we decided to investigate the effect of PTMs on hnRNPUL1 phase separation next.

5.5 hnRNPUL1 CTD is a phosphorylation substrate for CDK7

Post-translational modifications are known to affect the LLPS behaviour of proteins. For example, phosphorylation of FUS, NELF and TAF15 disrupts the phase separation and perturbs PolII CTD interactions with LLPS droplets (Kwon et al. 2014, Monahan et al. 2017, Ryan et al. 2018, Wang et al. 2018b). Moreover, formation of NELF droplets was promoted by dephosphorylation in conjunction with SUMOylation (Rawat et al. 2021). Additionally, methylation and citrullination of Arg have opposite effects on FUS LLPS and aggregation (Chong et al. 2018b, Qamar et al. 2018, Tanikawa et al. 2018). It is believed that addition and removal of PTMs reversibly regulates protein condensation into phase separated droplets.

The CTD of hnRNPUL1 has been reported to harbour numerous PTM sites, most prominently Arg methylation and Ser/Tyr phosphorylation (Figure 5.8A), as catalogued on Phosphosite Plus (Hornbeck et al. 2015). The platform collates PTM reports from both high and low throughput studies, but it does not offer information on the enzyme(s) responsible for depositing the marks. Previous work carried out by Dr Llywelyn Griffith showed that hnRNPUL1 was a phosphorylation substrate for the enzyme CDK7 *in vitro* (Griffith 2019). More recently, hnRNPUL1 was independently identified as a phosphorylation substrate for CDK7 *in vivo* at position Ser718 within the PLD region in a mass spectrometry study (Rimel et al. 2020).

Based on these insights, we wanted to test the effect of phosphorylation marks upon phase separation by hnRNPUL1 CTD. First, we established a protocol for phosphorylating hnRNPUL1 CTD with CDK7 *in vitro*. In parallel, we used the mCherry-tagged PolII CTD construct as a positive control for CDK7 phosphorylation (Figure 5.8B). The result confirmed that hnRNPUL1 CTD could be phosphorylated by CDK7 in our *in vitro* assay, but with a substantially lower efficiency than PolII CTD (Figure 5.8C). The difference in substrate preference may be attributed to the PolII CTD construct containing 22 heptapeptide repeats, each with up to two CDK7 phosphorylation sites, whereas hnRNPUL1 harboured a single experimentally-determined CDK7 site and up to four possible CDK7 targets in total, based on sequence analysis.

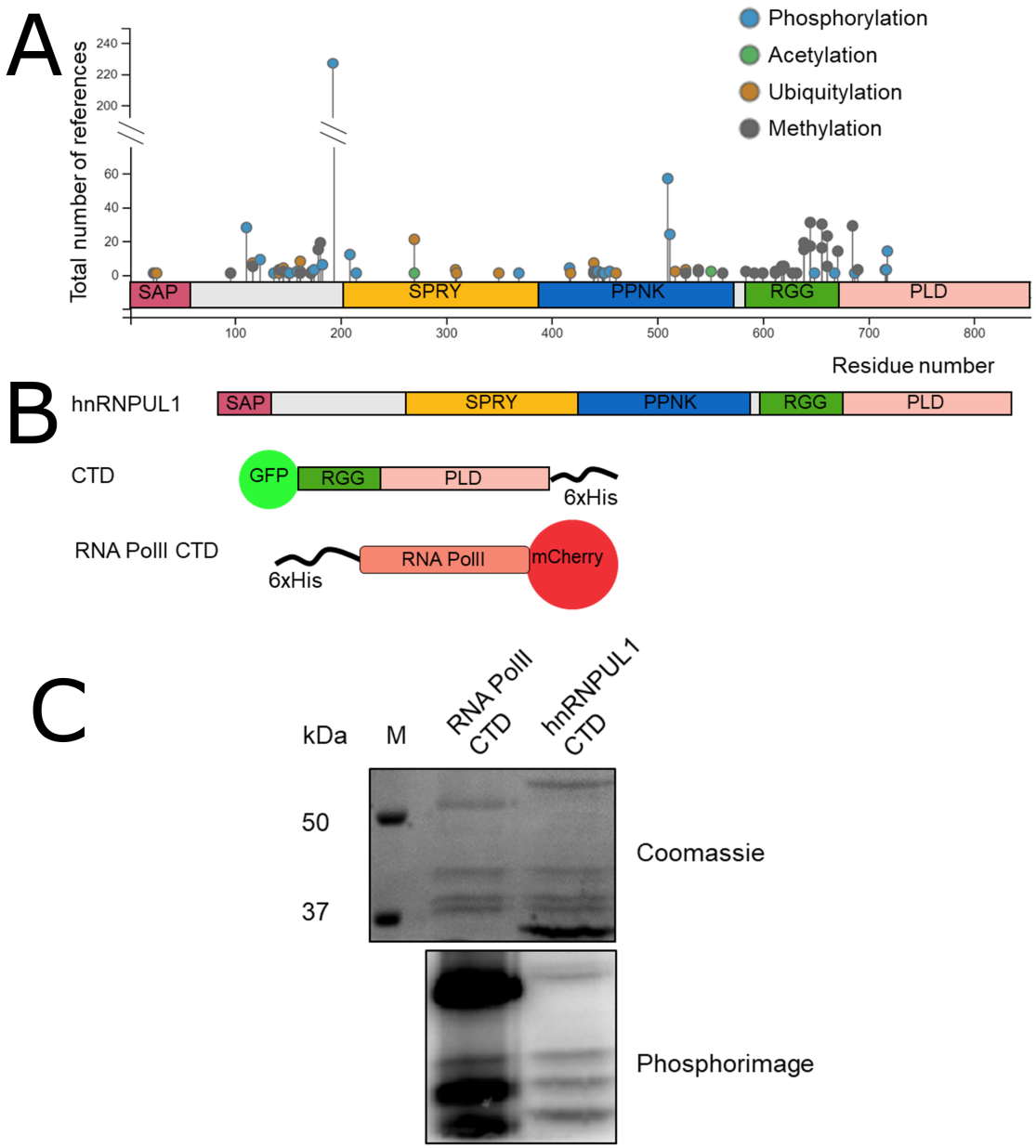

Figure 5.8: hnRNPUL1 CTD is a phosphorylation substrate for CDK7. **A)** Map of PTM distribution along the primary sequence of hnRNPUL1 generated with PhosphoSite Plus. The height of the bars corresponds to the number of reports that identified each one. **B)** Schematic of hnRNPUL1 domain organisation and of the GFP- and mCherry-tagged proteins used for the kinase assay below. **C)** Kinase assay testing the effect of CDK7 on hnRNPUL1 CTD phosphorylation status. PolII was used as a positive control. Substrate proteins were incubated at 37°C with trimeric Cdk-activating kinase (CAK) complex formed of CDK7, Cyclin H and MNAT1. Trace amounts of ^{32}P γ-ATP were provided in the reaction to track protein labelling by auto radiography.

5.5.1 CDK7 phosphorylation disrupts hnRNPUL1 droplet formation

In the previous section, we established a method for phosphorylating the CTD of hnRNPUL1 *in vitro* using CDK7 as part of the trimeric CAK complex. We used that approach to study the effects of protein phosphorylation on hnRNPUL1 phase separation. The CTD of hnRNPUL1 (Figure 5.9A) was incubated in the same LLPS conditions as before, with the addition of ATP or the non-hydrolysable analogue AMP-PNP and CDK7 (Figure 5.9B). Addition of ATP or CDK7 complex separately did not affect hnRNPUL1 droplet formation (panels 1 and 2), however the combination of the two severely disrupted phase separation (panel 3). We demonstrated that the direct action of CDK7 phosphorylating hnRNPUL1 CTD impacted LLPS by incubating the enzyme complex with non-hydrolysable AMP-PNP, which did not affect droplet formation. The control conditions in panels 1, 2 and 4 proved that the effects observed in panel 3 were not due to over dilution of the protein or electrostatic interactions between the CTD and ATP solubilising the droplets, but instead the direct protein phosphorylation prevented phase separation.

5.6 Summary

In this chapter, I have presented evidence for the ability of the C-terminal domain of hnRNPUL1 to undergo liquid-liquid phase separation (LLPS) under physiological conditions *in vitro*.

We determined experimentally that the CTD of hnRNPUL1 could undergo phase transition *in vitro* at a range of salt concentrations both below and around the physiological level, but only in the presence of the crowding agent PEG 8,000 (Figure 5.1). We measured the protein concentration at which LLPS droplets formed (known as the saturation concentration or c_{sat}) to be around 1 μM for the entire hnRNPUL1 CTD and below 5 μM for the PLD region alone, which were comparable to other phase separating proteins (Figure 5.3). Droplets formed by the hnRNPUL1 CTD and PLD regions were next treated with the commonly used aliphatic alcohol 1,6-hexanediol to test their susceptibility. The compound severely impaired new droplet formation for both proteins and completely dissolved pre-formed hnRNPUL1 CTD droplets, but had a more moderate effect on pre-formed PLD droplets (Figure 5.4).

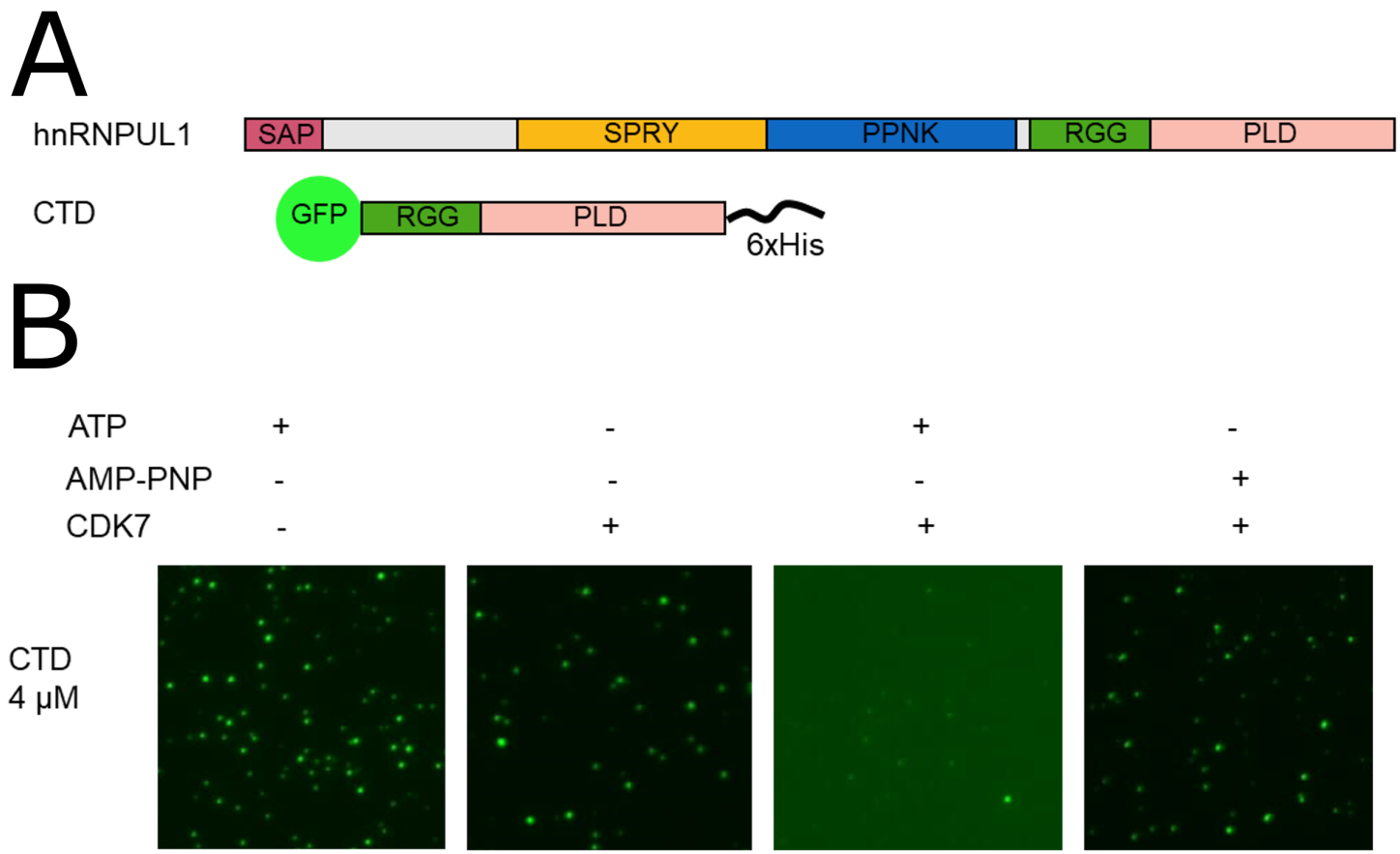

Figure 5.9: hnRNPUL1 droplets are dissolved by CDK7 phosphorylation.
A) Schematic of hnRNPUL1 domain organisation and of the GFP-tagged CTD used in the phase separation assays below. **B)** Phase separation assays of hnRNPUL1 CTD in buffer supplemented with 10 mM $MgCl_2$, 0.1 mM ATP or AMP-PNP and CDK7.

These findings supported the idea that the observed protein droplets were caused by phase separation, rather than aggregation.

Next, we probed the molecular basis for hnRNPUL1 phase separation. We employed two different mutants, one with the Arg residues changed to Lys (R→K) and the other with the Tyr residues changed to Phe (Y→F). In both cases, the mutations severely restricted droplet formation, supporting the cation-π model between Arg-Tyr for hnRNPUL1 CTD phase separation. Additionally, we observed a complete block of phase separation for the PLD region tested separately when Tyr were changed to Phe (Figure 5.5). The observation suggested that Tyr residues contribute to droplet formation either by Tyr-Tyr interactions or by forming cross-β sheets as part of (Ser/Gly)Tyr(Ser/Gly) motifs, although the two models are not mutually exclusive.

Further on, we tested the effects of different levels of ATP on hnRNPUL1 phase separation. We observed that sub-physiological concentrations (up to 1 mM) enhanced the number and size of hnRNPUL1 CTD droplets, while cellular levels of ATP impaired (at 2 and 4 mM) or completely prevented LLPS (at 10 mM) as shown in Figure 5.6.

Subsequently, we examined the interaction between hnRNPUL1 and RNA Polymerase II discussed in the previous chapter from the perspective of LLPS. We observed co-localisation of hnRNPUL1 and PolII to phase separated droplets (Figure 5.7) mediated by hnRNPUL1 recruitment of PolII CTD to the droplets.

Lastly, we investigated the effects of post-translational modifications on hnRNPUL1 phase separation. The CTD of hnRNPUL1 hosts numerous PTM sites, including Ser phosphorylation by CDK7. Initially, we confirmed that the experimental setup used in the LLPS assays was favourable for CDK7 activity on hnRNPUL1 CTD (Figure 5.8). Next, we carried out a CDK7 treatment at the beginning of the phase separation incubation and observed a solubilising effect by the phosphorylation marks on hnRNPUL1 CTD (Figure 5.9).

The work presented here forms the initial steps in understanding the role that phase separation plays for the function of hnRNPUL1. The complexity of cellular protein

condensates cannot be recreated in a test tube. It is likely that hnRNPUL1 would interact with numerous other partners to form LLPS droplets *in vivo* and for a better understanding of this phenomenon, the full-length protein would have to be studied. Another useful approach would be to analyse the phase behaviour of hnRNPUL1 *in vivo* by using a GFP-tagged hnRNPUL1 cell line. It would also be interesting to examine phase transition for hnRNPUL1 mutations located in the CTD which are associated with the neurodegenerative disease ALS, such as R618W, R639C, Y655H, P755L and G847R. All of them are found either in (Ser/Gly)Tyr(Ser/Gly) motifs or in regions predicted to adopt similar cross-β sheet conformation and most of the mutations mentioned introduce residues otherwise absent from the CTD (Cys, His, Leu) or very rare (Trp). This could disrupt the cross-β sheet structures and impair hnRNPUL1 assembly into droplets, such as those formed at gene promoters, leading to detrimental results for patients. Conversely, the mutations (particularly P696L and P755L) may stabilize the filaments by promoting dry steric zipper assemblies and toxic amyloid-like deposits, as prolines have been suggested to limit or interrupt protein regions with amyloidogenic potential (Ryan et al. 2018).

6 Chapter 6: Discussion

The aim of this project was to investigate the function of nuclear protein hnRNPUL1 from a biochemical perspective. We first characterised the structured central domain of the protein with UV crosslinking and enzymatic assays informed by the extensive structural homology with mammalian polynucleotide kinase phosphatase and bacteriophage T4 polynucleotide kinase. Subsequently, we investigated the role of the structured core of hnRNPUL1 in the context of the full-length protein and observed that it mediates an intramolecular interaction with the disordered C-terminal end and also orchestrates the transition of hnRNPUL1 between two modes of chromatin association. In the final chapter, we studied the phase transition behaviour of hnRNPUL1 mediated by its unstructured C-terminus and reported that the protein can undergo condensation into supramolecular assemblies by self-interaction in physiological conditions. Moreover, the phase-separated droplets could recruit the CTD of RNA polymerase II and be dissolved by treatment with an aliphatic alcohol or the cyclin-dependent enzyme CDK7, consistent with the literature on similar factors involved in gene expression. Overall, we have presented evidence for the function and inter-connectivity of two distinguishing features of hnRNPUL1, its folded core and disordered CTD, with implications in gene expression and disease. Next, the importance of these results will be considered in more depth.

6.1 hnRNPUL1 is an inactive polynucleotide kinase

Following the public release of the AI-assisted structural prediction algorithm AlphaFold, we were able to analyse a more unbiased model for the conformation of hnRNPUL1 than previously possible using template-based programmes like Phyre2 (Jumper et al. 2021, Varadi et al. 2021). The AlphaFold prediction confirmed the presence of two unstructured N- and C-terminal arms, but revealed that the central SPRY and PPNK regions, which had hitherto been considered separate domains, might in fact form a single folded unit with an extensive interface stabilised by hydrophobic interactions and electrostatic complementarity. The PPNK half of the core displayed extensive structural homology to the kinase regions of mammalian PNKP and viral T4 PNK, which are both involved in repairing nucleic acid 5′-

ends by phosphorylation in preparation for DNA or RNA ligation. In both the AlphaFold and Phyre2 models, the hnRNPUL1 PPNK contained a channel lined with basic residues which led into a region similar to the active sites of PNKP and T4 PNK. On this basis, we analysed hnRNPUL1's ability to bind similar types of substrates as the other polynucleotide kinases. Mutations in the putative P-loop NTPase motifs Walker A and Walker B of hnRNPUL1 were introduced to test the importance of an intact PPNK region for the protein's function.

UV crosslinking assays revealed that hnRNPUL1 was capable of binding both ATP and GTP via its PPNK pocket with only a slight preference for ATP. Moreover, deletion of the Walker A motif drastically reduced the protein's binding to ATP, while mutation of the Walker B had the opposite effect. Binding of a short mono- or triphosphorylated oligonucleotide to PPNK followed a similar suit, with the ΔWA mutation lessening the interaction and the ΔWB enhancing it. We confirmed that the protein-RNA interaction was specifically mediated by the PPNK pocket because mutation of basic residues lining the channel surface to alanine severely disrupted RNA crosslinking. We proposed that stronger binding of RNA to the ΔWB mutant could be due to either a reduced negative charge on the pocket surface after mutation of the WB aspartate or to an impaired substrate turnover. The possibility remains that the negative effects on RNA binding observed with point mutations was due to local misfolding of the proteins, however this could be tested by using circular dichroism to analyse the secondary structure of the mutants or by obtaining 3D structural information using NMR or X-ray crystallography. In reality, the stimulatory contribution of the ΔWB change argues in favour of those effects being functionally relevant. Additionally, the R468C point mutation associated with the neurodegenerative disease amyotrophic lateral sclerosis (ALS) exhibited an increased affinity for RNA and a reduced binding to ATP compared to WT, supporting a functional link between the two events.

Importantly, both structural models of the hnRNPUL1 PPNK include an extended helix and loop region guarding the top of the channel. The configuration resembles a structure found over the T4 PNK active site, which discriminates against double stranded nucleic acids. On the other hand, such a region is absent from PNKP, allow-

ing the enzyme to phosphorylate both single and double stranded DNA substrates. The presence of this extended feature in hnRNPUL1's structure indicates that the protein would be unlikely to interact with double stranded substrates directly via the PPNK pocket.

Next, we established that ATP and RNA binding in the PPNK pocket occurred from opposite sides of the channel, as observed from PNKP. While increasing amounts of ATP effectively displaced a 5′ monophosphorylated RNA, ADP had no influence on RNA binding except for a minor inhibition at the highest concentration tested, likely due to non-specific interactions. Another result corroborating this model was the RNA crosslinking pattern of the ALS-associated mutant R468C. While the core exhibited increased affinity for RNA, the full-length protein bound RNA as avidly as the WT, however co-incubation of an excess of ATP reduced RNA interaction with the WT, but not the R468C mutant protein. This illustrates how an excess of ATP could displace RNA from the PPNK pocket, while a mutant defective in ATP binding remained unaffected.

The arrangement of ADP and 5′ mono-P RNA within the PPNK pocket was reminiscent of the 5′ inverted cap structure that most PolII transcripts acquire early in transcription. For that reason, we tested if capped RNA molecules were suitable binding substrates. Direct binding to capped RNAs could not be tested rigorously because of technical difficulties with the experimental setup, prompting us to use a competition assay between ^{32}P-GTP and unlabelled GTP or an m^{7}G cap analog instead. The result demonstrated that hnRNPUL1 would be unlikely to accommodate an inverted 5′ cap within its binding pocket, either due to clashes with the methyl group or an inability to thread the entire structure from one end of the pocket to the other.

Lastly, we tested if a single amino acid change from aspartate to asparagine within the binding pocket was responsible for the lack of detectable ATP hydrolysis for hnRNPUL1 and its homologue, hnRNPU (Bernstein et al. 2009). To our surprise, an Asn-Asp reversal was sufficient to activate hnRNPUL1's ATPase function, which was further stimulated by addition of the protein's other ligand, an uncapped RNA, to hydrolyse over 50% more ATP. The observation indicated that the nucleic acid

might be providing its 5′ OH- group to receive the γ-phosphate removed from ATP, in the same way that T4 PNK is only active in the presence of both phosphate donor and acceptor. A kinase assay confirmed this possibility, with the Asn-Asp mutant, but not the WT or ΔWA proteins, phosphorylating both an RNA and a dsDNA substrate with free 5′ ends containing OH- groups (and overhangs, in the case of the dsDNA). The equivalent mutation in hnRNPU was not sufficient to restore ATP hydrolysis or kinase function, suggesting that perhaps there have not been the same evolutionary pressures to maintain a binding pocket architecture and composition capable of interacting with ligands the way that hnRNPUL1 does. Nonetheless, the enzymatic reactivation achieved by hnRNPUL1 demonstrate that it is one of only a handful of RNA binding proteins that specifically recognises the free 5′ end of an RNA and not (just) the RNA body. Other relevant examples are mammalian Xrn1, Xrn2 and DXO enzymes, which degrade uncapped or improperly capped RNAs in a 5′-3′ exonucleolytic manner (Jiao et al. 2013, Jinek et al. 2011). While hnRNPUL1's core does not sustain any enzymatic functions, its ability to recognise the 5′ end of nucleic acids has been preserved and appears crucial for protein function *in vivo*, as will be discussed later. It would be useful to characterise the effects of this gain-of-function Asn-Asp mutation in a cellular context, by either transient transfection or stable expression, although it may be lethal *in vivo*.

We could not reconcile hnRNPU's lack of enzymatic activity in our hands with its reported ATPase activity (Nozawa et al. 2017). However, the rate of hydrolysis quoted in the study (0.15 nmol ATP hydrolyzed x $nmol^{-1}$ of protein x min^{-1}) would make it an extremely inefficient enzyme and it is noteworthy that we observed similar catalytic rates with protein preparations containing trace amounts of *E. coli* chaperone contaminants. We also find it unlikely that the PPNK region expressed in the absence of the SPRY half of the core that was used by Nozawa et al. (2017) would allow proper folding of the protein, given the high degree of inter-connectivity between those two portions predicted by AlphaFold. Lastly, it is improbable that the Asn residue found in lieu of Asp could sustain hydrolysis, as this mutation was demonstrated to abolish PNKP's ATPase activity and the two proteins share extensive structural homology.

6.2 The central domain orchestrates hnRNPUL1's cellular functions

The caveat of using UV crosslinking is that it is a very inefficient process (1-3% of protein-RNA complexes are crosslinked) and it does not allow measurement of binding and dissociation rates. It is possible that in the course of these assays, we used sub-optimal ligand-to-protein concentrations which might have skewed the results or our interpretation. It may be suitable to repeat the RNA binding assays with a greater ratio of ligand/mutant proteins by using non-radiolabelled substrates, although that may reduce the ^{32}P signal. Therefore, we used an orthogonal approach to study the binding the SPRY-PPNK core to ATP and triphosphorylated RNA in solution by quantifying the tryptophan fluorescence quenching after ligand addition.

Analysis of the AlphaFold model revealed that out of five Trp residues found in the core, only two were placed in close proximity to the NTP/RNA binding pocket. The first, W437, was located adjacent to the Walker A motif, while the second, W477, was included in the extended helix-loop feature believed to discriminate against double stranded nucleic acids. It was revealed that the hnRNPUL1 core underwent a similar conformational change after binding ATP or triphosphorylated RNA, as quantified by the intrinsic fluorescence quenching that accompanies changes in the local protein environment. There was a roughly 13% quenching of WT and ΔWB mutant at the highest ligand concentrations tested, while the ΔWA exhibited marginal changes. All results were consistent with the ligand binding patterns determined by UV crosslinking, but additionally allowed estimation of the dissociation constants for protein/ATP or protein/RNA complexes. The calculated values for WT protein/RNA complexes were comparable to Kds for PNKP binding to DNA, however the Kd for WT/ATP were around 8 times lower than those reported for PNKP. For the ΔWB mutant, the calculated Kds were even lower than the WT protein. This allowed the possibility that some of the measurements may have been carried out under conditions resembling the "titration regime". This phenomenon describes situations where the protein concentration is much greater than the Kd, meaning that the ligand concentration at which half of the protein appears to be

in complex does not reflect the true Kd. In such cases, the free ligand added in solution is immediately bound to the protein and the ligand concentration that gives half binding is rather an illustration of the active protein population (Jarmoskaite et al. 2020). In this context, it is possible that the Kd values reported here are an underestimation of the real Kds, which may in fact be substantially smaller. Altogether, this complementary approach supported the observations of the previous chapter and indicated that a conformational change of the protein accompanied ligand binding.

Next, we probed the contribution of the hnRNPUL1 core and RGG box to RNA binding. The UV crosslinking experiment illustrated that while each of those regions could bind RNA separately, there was a cooperativity in binding when both were present at the same time, such that the full-length protein bound more RNA than the hnRNPUL1 ΔCTD and CTD constructs combined. A similar binding cooperativity between RNA binding modules was described for the numerous binding motifs of FUS protein, where some of the regions were suggested to enhance binding by melting RNA secondary structures (Loughlin et al. 2018). In hnRNPUL1's case, we propose that the cooperativity could stem from the RGG box presenting RNA molecules to PPNK by bringing them into close proximity and stabilising the substrates long enough for PPNK to recognise the 5′ ends. A more striking observation of the experiment was that the full-length ΔWA mutant retained a drastically reduced ability to bind RNA, despite the presence of an intact RGG box. Importantly, this phenotype was repeatedly observed only after stringent RNAse treatment of the proteins during purification, which removed all RNA contaminants from the protein solutions. Our explanation for this behaviour was that the purification strategy forced proteins to adopt unnatural apo conformations, whereby the PPNK pockets are devoid of either ATP or RNA ligands. Presumably, this apo state was also accompanied by interactions between RGG and the SPRY-PPNK core, explaining the drastic reduction in RNA binding to the ΔWA protein. However, during the UV crosslinking assay, proteins were presented with both ATP and RNA ligands, allowing them to bind and subsequently undergo a conformational change around the PPNK pocket, except for the ΔWA mutant, which remains trapped in the apo conformation. Such a scenario is unlikely to occur in the highly concentrated cellular

milieu, where (at least) ATP is readily available at up to 10 mM in some cases.

We tested this intramolecular interaction model with a pulldown assay, which confirmed that a direct binding between the apo PPNK core and the CTD could occur. However, addition of RNA ligands to the reaction stimulated, rather than disrupted, the interaction, possibly by forming a bridge between the core and the CTD of hnRNPUL1. Inspection of the PPNK surface charge revealed an extensive acidic patch adjacent to the ligand pocket, which may mediate the CTD interaction via electrostatic complementarity with the basic arginines of the RGG box. It is not clear if this interaction is relevant *in vivo*, since hnRNPUL1 may never adopt the necessary apo state in the cell. However, there are precedents for similar intramolecular interactions between structured and disordered domains even within the hnRNP family. Both the globular domain and the PLD of TDP43 interact and contribute to phase separation, while the folded region additionally undergoes oligomerisation controlled by phosphorylation (Wang et al. 2018a).

Having established a bimodal mechanism of RNA binding for hnRNPUL1, we attempted to establish a protocol for capturing cellular RNA targets bound via the PPNK pocket. Unfortunately, the NO CAP RIP assay needed further refinement to remove the bulk non-specific ribosomal RNA contaminants and to stabilise the relevant interactions without affecting the structure of the RNA 5′ end through covalent modifications. Instead, we searched publicly available databases for potential targets that might be sampled by PPNK separately from the RGG box. An obvious candidate for us was U4 snRNA, which displayed hnRNPUL1 eCLIP signal both at the 5′ end and along the 5′ stem loop, around the binding site for Snu13 protein. Additional lines of evidence from our lab implicated hnRNPUL1 in transcription of U snRNA genes by PolII, as well as in post-spliceosomal recycling of U4 and U4/U6 snRNPs.

The post-transcriptional recycling function is supported by a strong RNA-dependent interaction with the U6 snRNP recycling factor SART3, detected by mass spectrometry (Griffith 2019). Moreover, the levels of U4/U6 and U4/U6·U5 di- and tri-snRNPs were reduced following hnRNPUL1 depletion in the auxin-inducible degron system (Griffith 2019). Snu13 is an integral part of U4 snRNP and serves as a recruitment

platform for additional splicing factors. The hnRNPUL1 eCLIP signal around the Snu13 binding site suggested that hnRNPUL1 might be one of the factors recruited by this protein. A direct protein interaction between the two was detected by pull-down and confirmed by capture of hnRNPUL1 from cell lysates with Snu13-loaded beads. Specifically, the interaction with Snu13 was mediated by the SPRY-PPNK core, although the molecular details are not evident from a simple inspection of the Snu13-U4 snRNA structure.

The position of the hnRNPUL1 eCLIP signal adjoins the binding site for another Snu13-recruited protein, Prp31, which contacts U4 snRNA through a Nop domain (Liu et al. 2007). This raised the possibility that the two factors may be exchanged at a certain point during U4 snRNP recycling. However, Griffith (2019) observed an interaction between hnRNPUL1 and Prp31 after co-immunoprecipitation, which was completely disrupted by deletion of hnRNPUL1 CTD. Therefore, it is possible that hnRNPUL1 and Prp31 exist in different and not mutually exclusive configurations as part of the U4 snRNP. Interestingly, Snu13 and Nop56/58 also interact with box C/D snoRNAs, which are a class of small non-coding RNAs located in the nucleolus (Szewczak et al. 2005). hnRNPUL1 eCLIP analysis revealed that out of a subset of strong RBPs, this protein was enriched over snoRNAs and in particular, over a subset of them localised to Cajal bodies, called scaRNAs. The Snu13 interaction may therefore represent a more general mode of targeting hnRNPUL1 to specific RNA classes, so it will be important to obtain a better understanding of the molecular details. Additionally, the pronounced interaction with scaRNAs could be a result of hnRNPUL1's distribution within Cajal bodies, where it appears to regulate snRNP recycling. hnRNPUL1 is not enriched within Cajal bodies, but rather it displays a diffuse nuclear distribution as visualised by immunostaining, suggesting that it could carry out other roles aside from gene expression and snRNP recycling (Wilson lab, unpublished data).

Additionally, hnRNPUL1 is implicated in snRNA gene expression based on the reduction in nascent snRNA levels after hnRNPUL1 depletion, accompanied by a lower occupancy of PolII over snRNA genes measured by ChIP. A combination of *in vitro* pulldown and *in vivo* co-IP experiments have mapped a direct protein-protein inter-

action between hnRNPUL1 and PolII mediated by their respective disordered CTDs, although a sub-population of hnRNPUL1 associates with PolII exclusively through RNA interactions and is released with RNAse digestion (Wilson lab, unpublished). The unstructured CTD of hnRNPUL1 interacts with numerous other RBPs, including the FET proteins, and is involved in chromatin attachment in conjunction with the SPRY-PPNK region. Chromatin release assays identified two populations of hnRNPUL1 on chromatin after cell fractionation: about 20% of the WT and ΔWB protein was solubilised by RNAse treatment, while the rest was only released by DNAse digestion. In contrast, the ΔWA mutant was exclusively solubilised by chromatin digestion, appearing to be trapped on chromatin. This non-stop chromatin residency could explain ΔWA's enhanced interaction with transcription factors and PolII. The data indicate that ΔWA cannot exchange its mechanism of chromatin association from protein-protein to protein-RNA interactions, in agreement with the observed reduction in RNA binding of the full-length ΔWA mutant. Although we cannot disprove hnRNPUL1 association to chromatin via DNA binding, we could not observe discrete DNA-protein complexes using the SAP domain in mobility shift assays or an interaction with histones in pulldown experiments.

ChIP-seq analysis performed by Dr Yonchev of the K562 dataset published by Xiao et al. (2019) revealed an enrichment of hnRNPUL1 over snRNA genes and accumulation toward the 3′ end and beyond the coding region. Such a distribution suggests a role for hnRNPUL1 in snRNA 3′ end formation. Indeed, a study investigating snRNA 3′ processing factors identified hnRNPUL1 as one of three most enriched proteins associated with the CBC-ARS2 complex, which promotes 3′-end maturation (Hallais et al. 2013). Furthermore, hnRNPUL1 was captured by co-IP with the Integrator subunit Int6, along with PolII and the rest of the Integrator complex (Vervoort et al. 2021). This 14-subunit protein assembly participates in 3′ cleavage and processing at snRNA loci through the catalytic activity of Int11/Int9. Similarly to hnRNPUL1, Integrator's ChIP distribution is enriched toward the 3′ end and downstream of the snRNA encoding region (Xiao et al. 2019, Baillat et al. 2005). Integrator has also been implicated in the 3′-end cleavage and subsequent termination of enhancer RNA (eRNA) transcription (Lai et al. 2015), which are another class of RNAs affected by hnRNPUL1 depletion. Sequencing of chromatin-associated RNAs

indicated that the reduction in eRNA transcription in the hnRNPUL1-AID system was accompanied by improper termination and accumulation of 3′ elongated forms of these transcripts in both the sense and anti-sense directions (Yonchev 2021). These data suggest that hnRNPUL1 is involved in snRNA and eRNA expression by promoting transcription termination, likely in conjunction with Integrator.

We propose a model for hnRNPUL1's function in promoting snRNA termination which takes into account the 5′ monophosphorylated downstream cleavage products generated by Integrator's endonucleolytic activity. These are not subjected to 5′-3′ degradation by Xrn2 (Sousa-Luís et al. 2021) and extend for over half a kilobase past the Integrator cleavage site (Cuello et al. 1999), mirroring hnRNPUL1's ChIP profile. In our model, hnRNPUL1 associates with the transcribing polymerase either at the promoter or very soon after snRNA transcription initiation and may protect the uncapped nascent U4 snRNA 5′ end until later cap acquisition (Glover-Cutter et al. 2007). hnRNPUL1 travels along the snRNA genes through a combination of protein-protein and protein-pre-snRNA interactions. After cleavage of the nascent transcript upstream of the 3′ box by Integrator, hnRNPUL1 binds the free 5′ end of the downstream cleavage product via PPNK and possibly aided by the RGG box (Figure 6.1). This series of event elicits transition of hnRNPUL1 on PolII from a protein-bound state to a (mainly) RNA-bound state and induces transcription termination by an unknown mechanism. hnRNPUL1 found in the RNA-bound state may exert an inhibitory effect on PolII akin to the transcription termination factor Pcf11, which bridges the nascent RNA to PolII CTD, triggering disassembly of the elongation complex by reducing PolII processivity (Zhang & Gilmour 2006, Zhang et al. 2007). Conceptually, hnRNPUL1 may reduce the PolII transcription rate by concomitantly binding the nascent RNA at the 5′ end via PPNK and along the RNA body via RGG, as well as PolII CTD via the PLD. This hypothesis could be tested *in vitro* with pulldown experiments analysing if hnRNPUL1 could bridge the nascent RNA to the PolII CTD. Additionally, hnRNPUL1 could be introduced to pre-assembled transcribing complexes with free nascent RNA 5′ ends to investigate if hnRNPUL1 promotes dismantling of the PolII elongation complex. Lastly, we could test the ChIP or RIP distribution of hnRNPUL1 and PolII on snRNA genes after knockdown of the Integrator catalytic subunit, as we would expect an enhancement

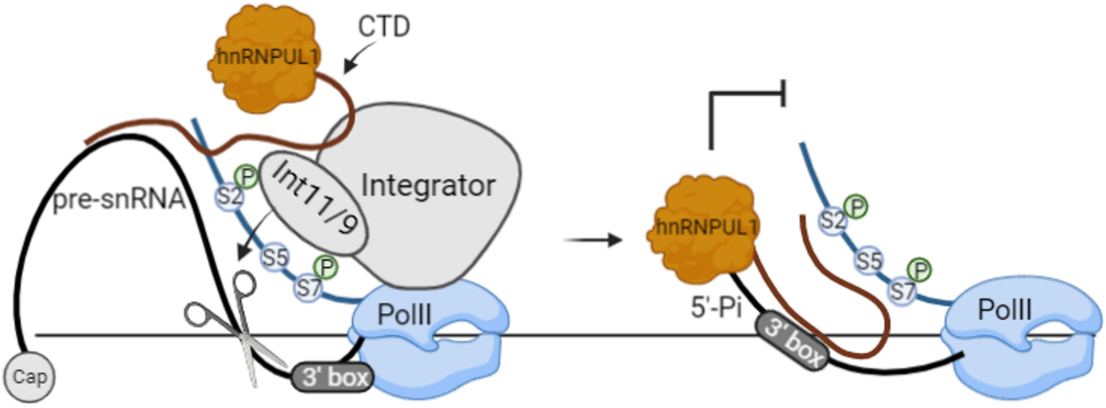

Figure 6.1: hnRNPUL1 promotes PolII transcription termination downstream of Integrator.
hnRNPUL1 CTD is stably associated with the elongating complex via protein-RNA and protein-protein interactions with factors such as PolII and Integrator at snRNA loci. pre-snRNA cleavage upstream of the 3′ box by Integrator generates a 5′ monophosphorylated RNA attached to the transcribing polymerase. The hnRNPUL1 core binds the RNA 5′ end via the PPNK pocket, while the RGG associates with the RNA body. The transfer of hnRNPUL1 from PolII onto the RNA exerts an inhibitory effect on PolII elongation and promotes termination. A similar mechanism may contribute to enhancer RNA expression.

downstream of these loci.

While the identity of the specificity factor(s) targeting hnRNPUL1 to snRNA genes and enhancers remains unknown, we may speculate that a combination of protein interactions such as PolII, the FET family and Integrator complex play a role. We have also considered that hnRNPUL1 could act as positive elongation factor while associated with PolII or that its transfer onto the nascent transcript could allow binding of a negative elongation factor. However, these models are not in agreement with the observed increase in 3′ transcription downstream of snRNA and enhancer loci upon hnRNPUL1 depletion. For these reasons, we favour the mechanism of allosteric PolII inhibition following loading of hnRNPUL1 onto the 5′ end of the cleaved transcript outlined above. The model is also consistent with the inability of the ΔWA mutant to transfer onto RNA as shown *in vitro* and *in vivo* with the chromatin release assay. Lastly, free ATP could bind to PPNK to release the RNA and allow its degradation. When not engaged with RNA 5′ ends, the PPNK pocket may be loaded with readily accessible ATP to prevent formation of the apo state and association of the CTD with the SPRY-PPNK domain.

6.3 hnRNPUL1 interaction with the transcription machinery may be promoted by phase separation

In the final chapter, we investigated the potential of hnRNPUL1 to undergo liquid-liquid phase separation through its unstructured CTD. Sequence analysis of this region had uncovered both characteristics of prion-like domains and a specific amino acid bias for Pro, Ser, Tyr, Asn and Gln, which are a signature of many phase separating proteins. Furthermore, an enrichment of Arg and Tyr in the RGG box and the PLD, respectively, had been established as major drivers of FUS protein condensation (Wang et al. 2018b). Lastly, a proteome-wide search had predicted numerous regions within the hnRNPUL1 CTD to undergo self-interaction akin to those generating protein fibrils observed at later stages of protein condensation or in pathogenesis (Hughes et al. 2018).

The CTD was tested in phase separation reactions, varying parameters around the

physiological concentrations of protein, salt and ATP. First, we observed that the CTD of hnRNPUL1 formed spherical droplets in the presence of 5% PEG 8,000 inert crowding agent, which simulates the concentrated cellular environment. Moreover, the droplets persisted at physiological salt concentrations (100-150 mM) and were dissolved by the aliphatic alcohol 1,6-hexanediol, which is a routine test often used to distinguish between phase separated condensates and protein aggregates. Importantly, the minimal protein concentration necessary for droplet formation (1 μM) was around the 1.4 μM nuclear concentration of hnRNPUL1 reported by Hein et al. (2015). Moreover, we observed that physiological ATP concentrations supported hnRNPUL1 phase separation, with concentrations near the lower limit (1 mM) promoting droplet formation and concentrations above 2 mM reducing the number of droplets, while 10 mM ATP almost completely solubilised the protein. We also determined that Arg-Tyr interactions play an important role in hnRNPUL1 phase transition, since mutating either Arg to Lys or Tyr to Phe abrogated droplet formation, as had been described for FUS (Wang et al. 2018b).

Many components of the transcription machinery have been observed to phase separate *in vitro* and in the cell, including the Mediator complex, the negative elongation factor NELF and the co-activators BRD4 and p300, with the most studied factors being the FET proteins, hnRNPA1/2 and TDP43. PolII does not phase separate on its own, but can be recruited to condensates formed by other proteins. We observed that hnRNPUL1 could be one such protein, since co-incubation with PolII resulted in co-localisation of the two proteins to LLPS droplets. We also examined if post-translational modifications had an impact on the hnRNPUL1 phase behaviour, as phosphorylation and Arg methylation or citrullination had been repeatedly linked to droplet dissolution. The CTD of hnRNPUL1 harbours numerous Arg methylation and Ser/Tyr phosphorylation sites and is a substrate for the transcriptional CDK7 kinase *in vivo*. In agreement with the published literature, treatment of hnRNPUL1 with CDK7 prevented protein condensation, while incubation with CDK7 and non-hydrolysable AMP-PNP had no adverse effects, demonstrating that the specific phosphorylation of hnRNPUL1 CTD, rather than electrostatic effects or protein dilution, inhibited phase separation. Protein phosphorylation has also been shown to promote eviction of PolII from pre-formed TAF15 droplets (Kwon et al.

2014). This suggests a potential mechanism for PolII progression through the transcription cycle after CTD phosphorylation, which promotes exchanging of protein complexes associated with PolII by phase separation.

The limitations of these assays revolve around the failure to recapitulate the diverse nuclear environment, which may produce different phase behaviours for hnRNPUL1. In the future, it will be necessary to study the effects of cellular concentrations of RNA and with different degrees of secondary structure on hnRNPUL1 droplet formation, as well as the phase behaviour of the full-length protein. It is possible that the acidic N-terminal arm may promote additional electrostatic interactions to lower the minimal concentration needed for LLPS. The *in vitro* limitations could be overcome by studying hnRNPUL1 LLPS in a stable cell line expressing the protein as a GFP fusion for easy tracking. It is important to mention that the hnRNPUL1 droplets described here assumed a stiff, solid-like appearance within less than an hour after formation, characterised by failure to coalesce into larger droplets and no detectable fluorescence recovery after photobleaching as monitored for up to three minutes. These features could be a result of the Gln-rich amino acid composition, which affects droplet fluidity, as well as the lack of protein diversity within the droplets, which would inherently moderate the physico-chemical properties (Wang et al. 2018b). Additional PTMs may regulate hnRNPUL1's functions, specifically Pro isomerisation. The PLD and the N-terminal acidic arm contain 80% of the total hnRNPUL1 Pro residues and the CTD composition is made up of 20% Pro. The solubility and assembly into hnRNPs of TDP43 were shown to be modulated by Pro isomerisation through the activity of PPIA, with TDP43 pathogenic mutations causing reduced interaction with PPIA (Lauranzano et al. 2015). Furthermore, three ALS-linked mutations of hnRNPUL1 involve Pro residues located within these disordered regions (P54Q, P696L and P755L), while others are found in regions predicted to modulate fibril formation (R618W, R639C, Y655H and G847R), so it will be informative to assess the impact on protein solubility of these mutations.

Overall, we have characterised the folded SPRY-PPNK core of hnRNPUL1 as being an inactive polynucleotide kinase capable of binding NTPs and free nucleic acid 5′ ends via opposite entrance points to its binding pocket, in a similar orientation to

mammalian polynucleotide kinase. We have shown that ligand binding to this pocket induces a conformational change and that the protein core and CTD work in concert to guide different modes of hnRNPUL1 association with chromatin. A soluble pool of hnRNPUL1 appears to contribute to U4 snRNP recycling in the nucleoplasm, aided by the U4 specificity factor Snu13. An insoluble chromatin-associated fraction of hnRNPUL1 promotes transcription termination of U snRNA and enhancer loci. We propose that the PPNK core and CTD mediate loading of hnRNPUL1 onto the downstream Integrator cleavage product and induce a reduction in PolII processivity akin to the mRNA transcription termination factor Pcf11. Lastly, *in vitro* and cellular evidence suggest that hnRNPUL1 may be associated with components of the transcriptional machinery on chromatin through reversible phase separated protein condensates, which may be dis-regulated by certain ALS-associated mutations of hnRNPUL1.

References

Agafonov, D. E., Kastner, B., Dybkov, O., Hofele, R. V., Liu, W.-T., Urlaub, H., Lührmann, R. & Stark, H. (2016), 'Molecular architecture of the human U4/U6.U5 tri-snRNP', *Science* **351**, 1416–1420.

Alberti, S., Gladfelter, A. & Mittag, T. (2019), 'Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates', *Cell* **176**, 419–434.

Alberti, S., Halfmann, R., King, O., Kapila, A. & Lindquist, S. (2009), 'A systematic survey identifies prions and illuminates sequence features of prionogenic proteins', *Cell* **137**, 146–158.

Ameyar-Zazoua, M., Souidi, M., Fritsch, L., Robin, P., Thomas, A., Hamiche, A., Percipalle, P., Ait-Si-Ali, S. & Harel-Bellan, A. (2009), 'Physical and functional interaction between heterochromatin protein 1 and the RNA-binding protein heterogeneous nuclear ribonucleoprotein U*', *Journal of Biological Chemistry* **284**, 27974–27979.

Amitsur, M., Levitz, R. & Kaufmann, G. (1987), 'Bacteriophage T4 anticodon nuclease, polynucleotide kinase and RNA ligase reprocess the host lysine tRNA.', *The EMBO Journal* **6**, 2499–2503.

Bachi, A., Braun, I. C., Rodrigues, J. P., Panté, N., Ribbeck, K., Kobbe, C. v., Kutay, U., Wilm, M., Görlich, D., Carmo-Fonseca, M. & Izaurralde, E. (2000), 'The C-terminal domain of TAP interacts with the nuclear pore complex and promotes export of specific CTE-bearing RNA substrates.', *RNA (New York, N.Y.)* **6**, 136–58.

Baillat, D., Hakimi, M.-A., Näär, A. M., Shilatifard, A., Cooch, N. & Shiekhattar, R. (2005), 'Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA Polymerase II', *Cell* **123**, 265–276.

Barral, P. M., Rusch, A., Turnell, A. S., Gallimore, P. H., Byrd, P. J., Dobner, T. & Grand, R. J. (2005), 'The interaction of the hnRNP family member E1B-AP5 with p53', *FEBS Letters* **579**, 2752–2758.

Bartkowiak, B., Liu, P., Phatnani, H. P., Fuda, N. J., Cooper, J. J., Price, D. H., Adelman, K., Lis, J. T. & Greenleaf, A. L. (2010), 'CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1', *Genes & Development* **24**, 2303–2316.

Beckedorff, F., Blumenthal, E., daSilva, L. F., Aoi, Y., Cingaram, P. R., Yue, J., Zhang, A., Dokaneheifard, S., Valencia, M. G., Gaidosh, G., Shilatifard, A. & Shiekhattar, R. (2020), 'The human Integrator complex facilitates transcriptional elongation by endonucleolytic cleavage of nascent transcripts', *Cell Reports* **32**, 107917.

Bell, M., Schreiner, S., Damianov, A., Reddy, R. & Bindereif, A. (2002), 'p110, a novel human U6 snRNP protein and U4/U6 snRNP recycling factor', *The EMBO Journal* **21**, 2724–2735.

Berglund, F. M. & Clarke, P. R. (2009), 'hnRNP-U is a specific DNA-dependent protein kinase substrate phosphorylated in response to DNA double-strand breaks', *Biochemical and Biophysical Research Communications* **381**, 59–64.

Bernstein, N. K., Hammel, M., Mani, R. S., Weinfeld, M., Pelikan, M., Tainer, J. A. & Glover, J. N. M. (2009), 'Mechanism of DNA substrate recognition by the mammalian DNA repair enzyme, Polynucleotide Kinase', *Nucleic Acids Research* **37**, 6161–6173.

Bernstein, N. K., Williams, R. S., Rakovszky, M. L., Cui, D., Green, R., Karimi-Busheri, F., Mani, R. S., Galicia, S., Koch, C. A., Cass, C. E., Durocher, D., Weinfeld, M. & Glover, J. M. (2005), 'The molecular architecture of the mammalian DNA repair enzyme, polynucleotide kinase', *Molecular Cell* **17**, 657–670.

Beyer, A. L., Christensen, M. E., Walker, B. W. & LeStourgeon, W. M. (1977), 'Identification and characterization of the packaging proteins of core 40S hnRNP particles', *Cell* **11**, 127–138.

Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A. S., Yu, T., Marie-Nelly, H., McSwiggen, D. T., Kokic, G., Dailey, G. M., Cramer, P., Darzacq, X. & Zweckstetter, M. (2018), 'RNA polymerase II clustering through carboxy-terminal domain phase separation', *Nature Structural & Molecular Biology* **25**, 833–840.

Boija, A., Klein, I. A., Sabari, B. R., Dall'Agnese, A., Coffey, E. L., Zamudio, A. V., Li, C. H., Shrinivas, K., Manteiga, J. C., Hannett, N. M., Abraham, B. J., Afeyan, L. K., Guo, Y. E., Rimel, J. K., Fant, C. B., Schuijers, J., Lee, T. I., Taatjes, D. J. & Young, R. A. (2018), 'Transcription factors activate genes through the phase-separation capacity of their activation domains', *Cell* **175**, 1842–1855.

Borcherds, W., Bremer, A., Borgia, M. B. & Mittag, T. (2021), 'How do intrinsically disordered protein regions encode a driving force for liquid–liquid phase separation?', *Current Opinion in Structural Biology* **67**, 41–50.

Bouchard, J. J., Otero, J. H., Scott, D. C., Szulc, E., Martin, E. W., Sabri, N., Granata, D., Marzahn, M. R., Lindorff-Larsen, K., Salvatella, X., Schulman, B. A. & Mittag, T. (2018), 'Cancer mutations of the tumor suppressor SPOP disrupt the formation of active, phase-separated compartments', *Molecular Cell* **72**, 19–36.

Britton, S., Dernoncourt, E., Delteil, C., Froment, C., Schiltz, O., Salles, B., Frit, P. & Calsou, P. (2014), 'DNA damage triggers SAF-A and RNA biogenesis factors exclusion from chromatin coupled to R-loops removal', *Nucleic Acids Research* **42**, 9047–9062.

Burke, K., Janke, A., Rhine, C. & Fawzi, N. (2015), 'Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II', *Molecular Cell* **60**, 231–241.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B., Strein, C., Davey, N., Humphreys, D., Preiss, T., Steinmetz, L., Krijgsveld, J. & Hentze, M. (2012), 'Insights into RNA biology from an atlas of mammalian mRNA-binding proteins', *Cell* **149**, 1393–1406.

Chappell, C., Hanakahi, L. A., Karimi-Busheri, F., Weinfeld, M. & West, S. C. (2002), 'Involvement of human polynucleotide kinase in double-strand break repair by non-homologous end joining', *The EMBO Journal* **21**, 2827–2832.

Cho, E.-J., Takagi, T., Moore, C. R. & Buratowski, S. (1997), 'mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminaldomain', *Genes & Development* **11**, 3319–3326.

Cho, W.-K., Spille, J.-H., Hecht, M., Lee, C., Li, C., Grube, V. & Cisse, I. I. (2018), 'Mediator and RNA polymerase II clusters associate in transcription-dependent condensates', *Science* **361**, 412–415.

Chong, P. A., Vernon, R. M. & Forman-Kay, J. D. (2018b), 'RGG/RG motif regions in RNA binding and phase separation', *Journal of Molecular Biology* **430**, 4650–4665.

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G. M., Cattoglio, C., Heckert, A., Banala, S., Lavis, L., Darzacq, X. & Tjian, R. (2018), 'Imaging dynamic and selective low-complexity domain interactions that control gene transcription', *Science* **361**, eaar2555.

Chong, S., Graham, T. G., Dugast-Darzacq, C., Dailey, G. M., Darzacq, X. & Tjian, R. (2022), 'Tuning levels of low-complexity domain interactions to modulate endogenous oncogenic transcription', *Molecular Cell* pp. 2084–2097.e5.

Cisse, I. I., Izeddin, I., Causse, S. Z., Boudarene, L., Senecal, A., Muresan, L., Dugast-Darzacq, C., Hajj, B., Dahan, M. & Darzacq, X. (2013), 'Real-time dynamics of RNA Polymerase II clustering in live human cells', *Science* **341**, 664–667.

Cuello, P., Boyd, D. C., Dye, M. J., Proudfoot, N. J. & Murphy, S. (1999), 'Transcription of the human U2 snRNA genes continues beyond the 3′ box in vivo', *The EMBO Journal* **18**, 2867–2877.

Culjkovic-Kraljacic, B., Skrabanek, L., Revuelta, M. V., Gasiorek, J., Cowling, V. H., Cerchietti, L. & Borden, K. L. B. (2020), 'The eukaryotic translation initiation factor eIF4E elevates steady-state m7G capping of coding and noncoding transcripts', *Proceedings of the National Academy of Sciences* **117**, 26773–26783.

Dahlberg, J. E. & Schenborn, E. T. (1988), 'The human U snRNA promoter and enhancer do not direct synthesis of messenger RNA', *Nucleic Acids Research* **16**, 5827–5840.

Davidson, L., Francis, L., Eaton, J. D. & West, S. (2020), 'Integrator-dependent

and allosteric/intrinsic mechanisms ensure efficient termination of snRNA transcription', *Cell Reports* **33**, 108319.

D'Cruz, A. A., Babon, J. J., Norton, R. S., Nicola, N. A. & Nicholson, S. E. (2013), 'Structure and function of the SPRY/B30.2 domain proteins involved in innate immunity', *Protein Science* **22**, 1–10.

Dreyfuss, G., Choi, Y. D. & Adam, S. A. (1984), 'Characterization of heterogeneous nuclear RNA-protein complexes in vivo with monoclonal antibodies.', *Molecular and Cellular Biology* **4**, 1104–1114.

Düster, R., Kaltheuner, I. H., Schmitz, M. & Geyer, M. (2021), '1,6-Hexanediol, commonly used to dissolve liquid–liquid phase separated condensates, directly impairs kinase and phosphatase activities', *The Journal of Biological Chemistry* **296**, 100260.

Eastberg, J. H., Pelletier, J. & Stoddard, B. L. (2004), 'Recognition of DNA substrates by T4 bacteriophage polynucleotide kinase', *Nucleic Acids Research* **32**, 653–660.

Eaton, J. D., Davidson, L., Bauer, D. L., Natsume, T., Kanemaki, M. T. & West, S. (2018), 'Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73 activity', *Genes & Development* **32**, 127–139.

Egloff, S., O'Reilly, D., Chapman, R. D., Taylor, A., Tanzhaus, K., Pitts, L., Eick, D. & Murphy, S. (2007), 'Serine-7 of the RNA Polymerase II CTD is specifically required for snRNA gene expression', *Science* **318**, 1777–1779.

Egloff, S., Szczepaniak, S. A., Dienstbier, M., Taylor, A., Knight, S. & Murphy, S. (2010), 'The Integrator complex recognizes a new double mark on the RNA Polymerase II carboxyl-terminal domain*', *Journal of Biological Chemistry* **285**, 20564–20569.

Egloff, S., Zaborowska, J., Laitem, C., Kiss, T. & Murphy, S. (2012), 'Ser7 phosphorylation of the CTD recruits the RPAP2 Ser5 phosphatase to snRNA genes', *Molecular Cell* **45**, 111–122.

Elrod, N. D., Henriques, T., Huang, K.-L., Tatomer, D. C., Wilusz, J. E., Wagner, E. J. & Adelman, K. (2019), 'The Integrator complex attenuates promoter-proximal transcription at protein-coding genes', *Molecular Cell* **76**, 738–752.

Fackelmayer, F. O. & Richter, A. (1994), 'Purification of Two Isoforms of hnRNP-U and Characterization of Their Nucleic Acid Binding Activity'', *Biochemistry* pp. 10416–10422.

Fan, H., Lv, P., Huo, X., Wu, J., Wang, Q., Cheng, L., Liu, Y., Tang, Q.-Q., Zhang, L., Zhang, F., Zheng, X., Wu, H. & Wen, B. (2018), 'The nuclear matrix protein HNRNPU maintains 3D genome architecture globally in mouse hepatocytes', *Genome Research* **28**, 192–202.

Fong, N., Saldi, T., Sheridan, R. M., Cortazar, M. A. & Bentley, D. L. (2017), 'RNA Pol II dynamics modulate co-transcriptional chromatin modification, CTD phosphorylation, and transcriptional direction', *Molecular Cell* **66**, 546–557.

Gabler, S., Schütt, H., Groitl, P., Wolf, H., Shenk, T. & Dobner, T. (1998), 'E1B 55-kilodalton-associated protein: a cellular protein with RNA-binding activity implicated in nucleocytoplasmic transport of adenovirus and cellular mRNAs.', *Journal of virology* **72**, 7960–71.

Garces, F., Pearl, L. & Oliver, A. (2011), 'The structural basis for substrate recognition by mammalian polynucleotide kinase 3′ Phosphatase', *Molecular Cell* **44**, 385–396.

Giacometti, S., Benbahouche, N. E. H., Domanski, M., Robert, M.-C., Meola, N., Lubas, M., Bukenborg, J., Andersen, J. S., Schulze, W. M., Verheggen, C., Kudla, G., Jensen, T. H. & Bertrand, E. (2017), 'Mutually exclusive CBC-containing complexes contribute to RNA fate', *Cell Reports* **18**, 2635–2650.

Glover-Cutter, K., Kim, S., Espinosa, J. & Bentley, D. L. (2007), 'RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes', *Nature Structural and Molecular Biology* **15**, 71–78.

Greiner, J. V. & Glonek, T. (2021), 'Intracellular ATP concentration and implication for cellular evolution', *Biology* **10**, 1166.

Griffith, L. (2019), 'An investigation of the cellular functions of hnRNPUL1', *PhD thesis, The University of Sheffield* .

Gu, B., Eick, D. & Bensaude, O. (2013), 'CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo', *Nucleic Acids Research* **41**, 1591–1603.

Guenther, E. L., Cao, Q., Trinh, H., Lu, J., Sawaya, M. R., Cascio, D., Boyer, D. R., Rodriguez, J. A., Hughes, M. P. & Eisenberg, D. S. (2018), 'Atomic structures of TDP-43 LCD segments and insights into reversible or pathogenic aggregation.', *Nature structural & molecular biology* **25**, 463–471.

Guo, Y. E., Manteiga, J. C., Henninger, J. E., Sabari, B. R., Dall'Agnese, A., Hannett, N. M., Spille, J.-H., Afeyan, L. K., Zamudio, A. V., Shrinivas, K., Abraham, B. J., Boija, A., Decker, T.-M., Rimel, J. K., Fant, C. B., Lee, T. I., Cisse, I. I., Sharp, P. A., Taatjes, D. J. & Young, R. A. (2019), 'Pol II phosphorylation regulates a switch between transcriptional and splicing condensates', *Nature* **572**, 543–548.

Gurunathan, G., Yu, Z., Coulombe, Y., Masson, J.-Y. & Richard, S. (2015), 'Arginine methylation of hnRNPUL1 regulates interaction with NBS1 and recruitment to sites of DNA damage', *Scientific Reports* **5**, 10475.

Hacisuleyman, E., Goff, L. A., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., McClanahan, P., Hendrickson, D. G., Sauvageau, M., Kelley, D. R., Morse, M., Engreitz, J., Lander, E. S., Guttman, M., Lodish, H. F., Flavell, R., Raj, A. & Rinn, J. L. (2014), 'Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre', *Nature Structural & Molecular Biology* **21**, 198–206.

Hall, L., Carone, D., Gomez, A., Kolpa, H., Byron, M., Mehta, N., Fackelmayer, F. & Lawrence, J. (2014), 'Stable C0T-1 repeat RNA is abundant and is associated with euchromatic interphase chromosomes', *Cell* **156**, 907–919.

Hallais, M., Pontvianne, F., Andersen, P. R., Clerici, M., Lener, D., Benbahouche, N. E. H., Gostan, T., Vandermoere, F., Robert, M.-C., Cusack, S., Verheggen, C.,

Jensen, T. H. & Bertrand, E. (2013), 'CBC–ARS2 stimulates 3′-end maturation of multiple RNA families and favors cap-proximal processing', *Nature Structural and Molecular Biology* **20**, 1358–1366.

Han, S., Tang, Y. & Smith, R. (2010), 'Functional diversity of the hnRNPs: past, present and perspectives', *Biochemical Journal* **430**, 379–392.

Han, T., Kato, M., Xie, S., Wu, L., Mirzaei, H., Pei, J., Chen, M., Xie, Y., Allen, J., Xiao, G. & McKnight, S. (2012), 'Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies', *Cell* **149**, 768–779.

Harlen, K. M. & Churchman, L. S. (2017), 'The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain', *Nature Reviews Molecular Cell Biology* **18**, 263–273.

Harmon, T. S., Holehouse, A. S., Rosen, M. K. & Pappu, R. V. (2017), 'Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins', *eLife* **6**, e30294.

Hasegawa, Y., Brockdorff, N., Kawano, S., Tsutui, K., Tsutui, K. & Nakagawa, S. (2010), 'The matrix protein hnRNP U is required for chromosomal localization of Xist RNA', *Developmental cell* **19**, 469–76.

Hautbergue, G. M., Hung, M.-L., Golovanov, A. P., Lian, L.-Y. & Wilson, S. A. (2008), 'Mutually exclusive interactions drive handover of mRNA from export adaptors to TAP', *Proceedings of the National Academy of Sciences* **105**, 5154–5159.

Hein, M., Hubner, N., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I., Weisswange, I., Mansfeld, J., Buchholz, F., Hyman, A. & Mann, M. (2015), 'A human interactome in three quantitative dimensions organized by stoichiometries and abundances', *Cell* **163**, 712–723.

Helbig, R. & Fackelmayer, F. O. (2003), 'Scaffold attachment factor A (SAF-A) is concentrated in inactive X chromosome territories through its RGG domain', *Chromosoma* **112**, 173–182.

Hennig, S., Kong, G., Mannen, T., Sadowska, A., Kobelke, S., Blythe, A., Knott, G. J., Iyer, K. S., Ho, D., Newcombe, E. A., Hosoki, K., Goshima, N., Kawaguchi, T., Hatters, D., Trinkle-Mulcahy, L., Hirose, T., Bond, C. S. & Fox, A. H. (2015), 'Prion-like domains in RNA binding proteins are essential for building subnuclear paraspeckles.', *The Journal of cell biology* **210**, 529–39.

Henninger, J. E., Oksuz, O., Shrinivas, K., Sagi, I., LeRoy, G., Zheng, M. M., Andrews, J. O., Zamudio, A. V., Lazaris, C., Hannett, N. M., Lee, T. I., Sharp, P. A., Cissé, I. I., Chakraborty, A. K. & Young, R. A. (2021), 'RNA-mediated feedback control of transcriptional condensates', *Cell* **184**, 207–225.

Hernandez, N. & Weiner, A. M. (1986), 'Formation of the 3′ end of U1 snRNA requires compatible snRNA promoter elements', *Cell* **47**, 249–258.

Hong, Z., Jiang, J., Ma, J., Dai, S., Xu, T., Li, H. & Yasui, A. (2013), 'The role of hnRPUL1 involved in DNA damage response is related to PARP1', *PLoS ONE* **8**, e60208.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V. & Skrzypek, E. (2015), 'PhosphoSitePlus, 2014: mutations, PTMs and recalibrations', *Nucleic Acids Research* **43**, D512–D520.

Huber, J., Cronshagen, U., Kadokura, M., Marshallsay, C., Wada, T., Sekine, M. & Lührmann, R. (1998), 'Snurportin1, an m3G-cap-specific nuclear import receptor with a novel domain structure', *The EMBO Journal* **17**, 4114–4126.

Huelga, S., Vu, A., Arnold, J., Liang, T., Liu, P., Yan, B., Donohue, J., Shiue, L., Hoon, S., Brenner, S., Ares, M. & Yeo, G. (2012), 'Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins', *Cell Reports* **1**, 167–178.

Hughes, M. P., Sawaya, M. R., Boyer, D. R., Goldschmidt, L., Rodriguez, J. A., Cascio, D., Chong, L., Gonen, T. & Eisenberg, D. S. (2018), 'Atomic structures of low-complexity protein segments reveal kinked sheets that assemble networks', *Science* **359**, 698–701.

Izaurralde, E., Lewis, J., Gamberi, C., Jarmolowski, A., McGuigan, C. & Mattaj, I. W. (1995), 'A cap-binding protein complex mediating U snRNA export', *Nature* **376**, 709–712.

Izaurralde, E., Lewis, J., McGuigan, C., Jankowska, M., Darzynkiewicz, E. & Mattaj, I. W. (1994), 'A nuclear cap binding protein complex involved in pre-mRNA splicing', *Cell* **78**, 657–668.

Izaurralde, E., Stepinski, J., Darzynkiewicz, E. & Mattaj, I. W. (1992), 'A cap binding protein that may mediate nuclear export of RNA polymerase II-transcribed RNAs.', *The Journal of Cell Biology* **118**, 1287–1295.

Jain, S., Wheeler, J., Walters, R., Agrawal, A., Barsic, A. & Parker, R. (2016), 'ATPase-modulated stress granules contain a diverse proteome and substructure', *Cell* **164**, 487–498.

Janson, L., Weller, P. & Pettersson, U. (1989), 'Nuclear factor I can functionally replace transcription factor Sp1 in a U2 small nuclear RNA gene enhancer', *Journal of Molecular Biology* **205**, 387–396.

Jarmoskaite, I., AlSadhan, I., Vaidyanathan, P. P. & Herschlag, D. (2020), 'How to measure and evaluate binding affinities', *eLife* **9**, e57264.

Jiao, X., Chang, J., Kilic, T., Tong, L. & Kiledjian, M. (2013), 'A mammalian pre-mRNA 5′ end capping quality control mechanism and an unexpected link of capping to pre-mRNA processing', *Molecular Cell* **50**, 104–115.

Jiao, X., Xiang, S., Oh, C., Martin, C. E., Tong, L. & Kiledjian, M. (2010), 'Identification of a quality-control mechanism for mRNA 5′-end capping', *Nature* **467**, 608–611.

Jinek, M., Coyle, S. & Doudna, J. (2011), 'Coupled 5′ nucleotide recognition and processivity in Xrn1-mediated mRNA decay', *Molecular Cell* **41**, 600–608.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S.,

Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021), 'Highly accurate protein structure prediction with AlphaFold', *Nature* **596**, 583–589.

Jády, B. E., Darzacq, X., Tucker, K. E., Matera, A. G., Bertrand, E. & Kiss, T. (2003), 'Modification of Sm small nuclear RNAs occurs in the nucleoplasmic Cajal body following import from the cytoplasm', *The EMBO Journal* **22**, 1878–1888.

Kadonaga, J. T., Courey, A. J., Ladika, J. & Tjian, R. (1988), 'Distinct regions of Sp1 modulate DNA binding and transcriptional activation', *Science* **242**, 1566–1570.

Kang, J., Lim, L., Lu, Y. & Song, J. (2019), 'A unified mechanism for LLPS of ALS/FTLD-causing FUS as well as its modulation by ATP and oligonucleic acids', *PLOS Biology* **17**, e3000327.

Kang, J., Lim, L. & Song, J. (2018), 'ATP enhances at low concentrations but dissolves at high concentrations liquid-liquid phase separation (LLPS) of ALS/FTD-causing FUS', *Biochemical and Biophysical Research Communications* **504**, 545–551.

Karimi-Busheri, F., Lee, J., Weinfeld, M. & Tomkinson, A. E. (1998), 'Repair of DNA strand gaps and nicks containing 3′-phosphate and 5′-hydroxyl termini by purified mammalian enzymes', *Nucleic Acids Research* **26**, 4395–4400.

Karimi-Busheri, F. & Weinfeld, M. (1997), 'Purification and substrate specificity of polydeoxyribonucleotide kinases isolated from calf thymus and rat liver', *Journal of Cellular Biochemistry* **64**, 258–272.

Kato, M., Han, T., Xie, S., Shi, K., Du, X., Wu, L., Mirzaei, H., Goldsmith, E., Longgood, J., Pei, J., Grishin, N., Frantz, D., Schneider, J., Chen, S., Li, L., Sawaya, M., Eisenberg, D., Tycko, R. & McKnight, S. (2012), 'Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels', *Cell* **149**, 753–767.

Kato, M., Lin, Y. & McKnight, S. L. (2017), 'Cross- polymerization and hydrogel formation by low-complexity sequence proteins', *Methods* **126**, 3–11.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. (2015), 'The Phyre2 web portal for protein modeling, prediction and analysis', *Nature Protocols* **10**, 845–858.

Kiledjian, M. & Dreyfuss, G. (1992), 'Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box.', *The EMBO Journal* **11**, 2655–2664.

Kim, J. B. & Sharp, P. A. (2001), 'Positive transcription elongation factor b phosphorylates hSPT5 and RNA Polymerase II carboxyl-terminal Domain Independently of cyclin-dependent kinase-activating kinase*', *Journal of Biological Chemistry* **276**, 12317–12323.

Kim, M. K. & Nikodem, V. M. (1999), 'hnRNP U inhibits carboxy-terminal domain phosphorylation by TFIIH and represses RNA Polymerase II elongation', *Molecular and Cellular Biology* **19**, 6833–6844.

Kipp, M., Göhring, F., Ostendorp, T., Drunen, C. M. v., Driel, R. v., Przybylski, M. & Fackelmayer, F. O. (2000), 'SAF-Box, a conserved protein domain that specifically recognizes scaffold attachment region DNA', *Molecular and Cellular Biology* **20**, 7480–7489.

Koch, C. A., Agyei, R., Galicia, S., Metalnikov, P., O'Donnell, P., Starostine, A., Weinfeld, M. & Durocher, D. (2004), 'Xrcc4 physically links DNA end processing by polynucleotide kinase to DNA ligation by DNA ligase IV', *The EMBO Journal* **23**, 3874–3885.

Kroschwald, S., Maharana, S. & Simon, A. (2017), 'Hexanediol: a chemical probe to investigate the material properties of membrane-less compartments', *Matters* .

Kuhlman, T. C., Cho, H., Reinberg, D. & Hernandez, N. (1999), 'The general transcription factors IIA, IIB, IIF, and IIE are required for RNA Polymerase II transcription from the human U1 small nuclear RNA promoter', *Molecular and Cellular Biology* **19**, 2130–2141.

Kukalev, A., Nord, Y., Palmberg, C., Bergman, T. & Percipalle, P. (2005), 'Actin and hnRNP U cooperate for productive transcription by RNA polymerase II', *Nature Structural and Molecular Biology* **12**, 238–244.

Kuznetsova, I. M., Turoverov, K. K. & Uversky, V. N. (2014), 'What macromolecular crowding can do to a protein', *International Journal of Molecular Sciences* **15**, 23090–23140.

Kwon, I., Kato, M., Xiang, S., Wu, L., Theodoropoulos, P., Mirzaei, H., Han, T., Xie, S., Corden, J. & McKnight, S. (2014), 'Phosphorylation-Regulated Binding of RNA Polymerase II to Fibrous Polymers of Low-Complexity Domains', *Cell* **156**, 1049–1060.

Kzhyshkowska, J., Rusch, A., Wofl, H. & Dobner, T. (2003), 'Regulation of transcription by the heterogeneous nuclear ribonucleoprotein E1B-AP5 is mediated by complex formation with the novel bromodomain-containing protein BRD7', *Biochemical Journal* **371**, 385–393.

Kzhyshkowska, J., Schütt, H., Liss, M., kremmer, E., Stauber, R., Wolf, H. & Dobner, T. (2001), 'Heterogeneous nuclear ribonucleoprotein E1B-AP5 is methylated in its Arg-Gly-Gly (RGG) box and interacts with human arginine methyltransferase HRMT1L1', *Biochemical Journal* **358**, 305–314.

Lai, F., Gardini, A., Zhang, A. & Shiekhattar, R. (2015), 'Integrator mediates the biogenesis of enhancer RNAs', *Nature* **525**, 399–403.

Lancaster, A. K., Nutter-Upham, A., Lindquist, S. & King, O. D. (2014), 'PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition', *Bioinformatics* **30**, 2501–2502.

Lauranzano, E., Pozzi, S., Pasetto, L., Stucchi, R., Massignan, T., Paolella, K., Mombrini, M., Nardo, G., Lunetta, C., Corbo, M., Mora, G., Bendotti, C. & Bonetto, V. (2015), 'Peptidylprolyl isomerase A governs TARDBP function and assembly in heterogeneous nuclear ribonucleoprotein complexes', *Brain* **138**, 974–991.

Leipe, D. D., Koonin, E. V. & Aravind, L. (2003), 'Evolution and Classification of P-loop Kinases and Related Proteins', *Journal of Molecular Biology* **333**, 781–815.

Li, P., Banjade, S., Cheng, H.-C., Kim, S., Chen, B., Guo, L., Llaguno, M., Hollingsworth, J. V., King, D. S., Banani, S. F., Russo, P. S., Jiang, Q.-X., Nixon, B. T. & Rosen, M. K. (2012), 'Phase transitions in the assembly of multivalent signalling proteins', *Nature* **483**, 336–340.

Lillehaug, J. R. & Kleppe, K. (1975), 'Kinetics and specificity of T 4 polynucleotide kinase', *Biochemistry* **14**, 1221–1225.

Lillehaug, J. R., Kleppe, R. K. & Kleppe, K. (1976), 'Phosphorylation of double-stranded DNAs by T 4 polynucleotide kinase', *Biochemistry* **15**, 1858–1865.

Lin, Y., Protter, D. W., Rosen, M. & Parker, R. (2015), 'Formation and maturation of phase-separated liquid droplets by RNA-binding proteins', *Molecular Cell* **60**, 208–219.

Liu, S., Li, P., Dybkov, O., Nottrott, S., Hartmuth, K., Lührmann, R., Carlomagno, T. & Wahl, M. C. (2007), 'Binding of the human Prp31 Nop domain to a composite RNA-protein platform in U4 snRNP', *Science* **316**, 115–120.

Lobo, S. M. & Hernandez, N. (1989), 'A 7 bp mutation converts a human RNA polymerase II snRNA promoter into an RNA polymerase III promoter.', *Cell* **58**, 55–67.

Loizou, J. I., El-Khamisy, S. F., Zlatanou, A., Moore, D. J., Chan, D. W., Qin, J., Sarno, S., Meggio, F., Pinna, L. A. & Caldecott, K. W. (2004), 'The protein kinase CK2 facilitates repair of chromosomal DNA single-strand breaks', *Cell* **117**, 17–28.

Loughlin, F. E., Lukavsky, P. J., Kazeeva, T., Reber, S., Hock, E.-M., Colombo, M., Schroetter, C. V., Pauli, P., Cléry, A., Mühlemann, O., Polymenidou, M., Ruepp, M.-D. & Allain, F. H.-T. (2018), 'The solution structure of FUS bound to RNA reveals a bipartite mode of RNA recognition with both sequence and shape specificity', *Molecular Cell* **73**, 490–504.

Lu, J., Cao, Q., Hughes, M. P., Sawaya, M. R., Boyer, D. R., Cascio, D. & Eisenberg, D. S. (2020), 'CryoEM structure of the low-complexity domain of hnRNPA2 and its conversion to pathogenic amyloid', *Nature Communications* **11**, 4090.

Lunde, B. M., Reichow, S. L., Kim, M., Suh, H., Leeper, T. C., Yang, F., Mutschler, H., Buratowski, S., Meinhart, A. & Varani, G. (2010), 'Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain', *Nature Structural & Molecular Biology* **17**, 1195–1201.

Luo, F., Gui, X., Zhou, H., Gu, J., Li, Y., Liu, X., Zhao, M., Li, D., Li, X. & Liu, C. (2018), 'Atomic structures of FUS LC domain segments reveal bases for reversible amyloid fibril formation', *Nature Structural & Molecular Biology* **25**, 341–346.

Ma, B.-G., Chen, L., Ji, H.-F., Chen, Z.-H., Yang, F.-R., Wang, L., Qu, G., Jiang, Y.-Y., Ji, C. & Zhang, H.-Y. (2008), 'Characters of very ancient proteins', *Biochemical and Biophysical Research Communications* **366**, 607–611.

Maharana, S., Wang, J., Papadopoulos, D. K., Richter, D., Pozniakovsky, A., Poser, I., Bickle, M., Rizk, S., Guillén-Boixet, J., Franzmann, T., Jahnel, M., Marrone, L., Chang, Y.-T., Sterneckert, J., Tomancak, P., Hyman, A. A. & Alberti, S. (2018), 'RNA buffers the phase separation behavior of prion-like RNA binding proteins', *Science* **360**, 918–921.

Mandel, C. R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J. L. & Tong, L. (2006), 'Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease', *Nature* **444**, 953–956.

Mani, R. S., Karimi-Busheri, F., Fanta, M., Cass, C. E. & Weinfeld, M. (2003), 'Spectroscopic Studies of DNA and ATP Binding to Human Polynucleotide Kinase: Evidence for a Ternary Complex †', *Biochemistry* **42**, 12077–12084.

Martens, J. H. A., Verlaan, M., Kalkhoven, E., Dorsman, J. C. & Zantema, A. (2002), 'Scaffold/matrix attachment region elements interact with a p300-Scaffold Attachment Factor A complex and are bound by acetylated nucleosomes', *Molecular and Cellular Biology* **22**, 2598–2606.

Martin, E. W., Holehouse, A. S., Peran, I., Farag, M., Incicco, J. J., Bremer, A., Grace, C. R., Soranno, A., Pappu, R. V. & Mittag, T. (2020), 'Valence and patterning of aromatic residues determine the phase behavior of prion-like domains', *Science* **367**, 694–699.

Martinez-Rucobo, F., Kohler, R., van de Waterbeemd, M., Heck, A., Hemann, M., Herzog, F., Stark, H. & Cramer, P. (2015), 'Molecular Basis of Transcription-Coupled Pre-mRNA Capping', *Molecular Cell* **58**, 1079–1089.

Masuyama, K., Taniguchi, I., Kataoka, N. & Ohno, M. (2004), 'RNA length defines RNA export pathway', *Genes & Development* **18**, 2074–2085.

Matera, A. G. & Wang, Z. (2014), 'A day in the life of the spliceosome', *Nature Reviews Molecular Cell Biology* **15**, 108–121.

Matsubara, T., Oda, M., Takahashi, T., Watanabe, C., Tachiyama, Y., Morino, H., Kawakami, H., Kaji, R., Maruyama, H., Murayama, S. & Izumi, Y. (2019), 'Amyotrophic lateral sclerosis of long clinical course clinically presenting with progressive muscular atrophy', *Neuropathology* **39**, 47–53.

Mattaj, I. W. (1986), 'Cap trimethylation of U snRNA is cytoplasmic and dependent on U snRNP protein binding', *Cell* **46**, 905–911.

May, W. A., Lessnick, S. L., Braun, B. S., Klemsz, M., Lewis, B. C., Lunsford, L. B., Hromas, R. & Denny, C. T. (1993), 'The Ewing's sarcoma EWS/FLI-1 fusion gene encodes a more potent transcriptional activator and is a more powerful transforming gene than FLI-1.', *Molecular and Cellular Biology* **13**, 7393–7398.

Mayer, A., Lidschreiber, M., Siebert, M., Leike, K., Söding, J. & Cramer, P. (2010), 'Uniform transitions of the general RNA polymerase II transcription complex', *Nature Structural & Molecular Biology* **17**, 1272–1278.

Mazza, C., Ohno, M., Segref, A., Mattaj, I. W. & Cusack, S. (2001), 'Crystal structure of the human nuclear cap binding complex', *Molecular Cell* **8**, 383–396.

McCloskey, A., Taniguchi, I., Shinmyozu, K. & Ohno, M. (2012), 'hnRNP C Tetramer Measures RNA Length to Classify RNA Polymerase II Transcripts for Export', *Science* **335**, 1643–1646.

Monahan, Z., Ryan, V. H., Janke, A. M., Burke, K. A., Rhoads, S. N., Zerze, G. H., O'Meally, R., Dignon, G. L., Conicella, A. E., Zheng, W., Best, R. B., Cole, R. N., Mittal, J., Shewmaker, F. & Fawzi, N. L. (2017), 'Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity', *The EMBO Journal* **36**, 2951–2967.

Moteki, S. & Price, D. (2002), 'Functional coupling of capping and transcription of mRNA', *Molecular Cell* **10**, 599–609.

Mouaikel, J., Narayanan, U., Verheggen, C., Matera, A. G., Bertrand, E., Tazi, J. & Bordonné, R. (2003), 'Interaction between the small-nuclear-RNA cap hypermethylase and the spinal muscular atrophy protein, survival of motor neuron', *EMBO reports* **4**, 616–622.

Murakami, T., Qamar, S., Lin, J., Schierle, G. , Rees, E., Miyashita, A., Costa, A., Dodd, R., Chan, F. S., Michel, C., Kronenberg-Versteeg, D., Li, Y., Yang, S.-P., Wakutani, Y., Meadows, W., Ferry, R., Dong, L., Tartaglia, G., Favrin, G., Lin, W.-L., Dickson, D., Zhen, M., Ron, D., Schmitt-Ulms, G., Fraser, P., Shneider, N., Holt, C., Vendruscolo, M., Kaminski, C. & St George-Hyslop, P. (2015), 'ALS/FTD mutation-induced phase transition of FUS liquid droplets and reversible hydrogels into irreversible hydrogels impairs RNP granule function', *Neuron* **88**, 678–690.

Murphy, S., Yoon, J. B., Gerster, T. & Roeder, R. G. (1992), 'Oct-1 and Oct-2 potentiate functional interactions of a transcription factor with the proximal sequence element of small nuclear RNA genes', *Molecular and Cellular Biology* **12**, 3247–3261.

Murray, D. T., Kato, M., Lin, Y., Thurber, K. R., Hung, I., McKnight, S. L. & Tycko, R. (2017), 'Structure of FUS protein fibrils and its relevance to self-assembly and phase separation of low-complexity domains', *Cell* **171**, 615–627.

Murthy, A. C., Dignon, G. L., Kan, Y., Zerze, G. H., Parekh, S. H., Mittal, J. & Fawzi, N. L. (2019), 'Molecular interactions underlying liquidliquid phase separation of the FUS low-complexity domain', *Nature Structural & Molecular Biology* **26**, 637–648.

Nojima, T., Rebelo, K., Gomes, T., Grosso, A. R., Proudfoot, N. J. & Carmo-Fonseca, M. (2018), 'RNA Polymerase II phosphorylated on CTD serine 5 interacts with the spliceosome during co-transcriptional splicing', *Molecular Cell* **72**, 369–379.

Nostrand, E. L. V., Pratt, G. A., Yee, B. A., Wheeler, E. C., Blue, S. M., Mueller, J., Park, S. S., Garcia, K. E., Gelboin-Burkhart, C., Nguyen, T. B., Rabano, I., Stanton, R., Sundararaman, B., Wang, R., Fu, X.-D., Graveley, B. R. & Yeo, G. W. (2020), 'Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins', *Genome Biology* **21**.

Nott, T., Petsalaki, E., Farber, P., Jervis, D., Fussner, E., Plochowietz, A., Craggs, T. D., Bazett-Jones, D., Pawson, T., Forman-Kay, J. & Baldwin, A. (2015), 'Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles', *Molecular Cell* **57**, 936–947.

Nottrott, S., Hartmuth, K., Fabrizio, P., Urlaub, H., Vidovic, I., Ficner, R. & Lührmann, R. (1999), 'Functional interaction of a novel 15.5kD[U4/U6·U5] tri-snRNP protein with the5′ stem–loop of U4 snRNA', *The EMBO Journal* **18**, 6119–6133.

Nottrott, S., Urlaub, H. & Lührmann, R. (2002), 'Hierarchical, clustered protein interactions with U4/U6 snRNA: a biochemical role for U4/U6 proteins.', *The EMBO journal* **21**, 5527–38.

Nozawa, R.-S., Boteva, L., Soares, D. C., Naughton, C., Dun, A. R., Buckle, A., Ramsahoye, B., Bruton, P. C., Saleeb, R. S., Arnedo, M., Hill, B., Duncan, R. R., Maciver, S. K. & Gilbert, N. (2017), 'SAF-A regulates interphase chromosome structure through oligomerization with chromatin-associated RNAs', *Cell* **169**, 1214–1227.

Nozawa, R.-S. & Gilbert, N. (2019), 'RNA: nuclear glue for folding the genome', *Trends in Cell Biology* pp. 201–211.

Obrdlik, A., Kukalev, A., Louvet, E., Farrants, A.-K. , Caputo, L. & Percipalle, P. (2008), 'The histone acetyltransferase PCAF associates with actin and hnRNP U

for RNA Polymerase II transcription', *Molecular and Cellular Biology* **28**, 6342–6357.

Ohno, M., Kataoka, N. & Shimura, Y. (1990), 'A nuclear cap binding protein from HeLa cells', *Nucleic Acids Research* **18**, 6989–6995.

Ohno, M., Segref, A., Bachi, A., Wilm, M. & Mattaj, I. W. (2000), 'PHAX, a mediator of U snRNA nuclear export whose activity is regulated by phosphorylation', *Cell* **101**, 187–198.

Otsuka, Y., Kedersha, N. L. & Schoenberg, D. R. (2009), 'Identification of a cytoplasmic complex that adds a cap onto 5′-monophosphate RNA', *Molecular and Cellular Biology* **29**, 2155–2167.

Palacios, I., Hetzer, M., Adam, S. A. & Mattaj, I. W. (1997), 'Nuclear import of U snRNPs requires importin ', *The EMBO Journal* **16**, 6783–6792.

Patel, A., Lee, H., Jawerth, L., Maharana, S., Jahnel, M., Hein, M., Stoynov, S., Mahamid, J., Saha, S., Franzmann, T., Pozniakovski, A., Poser, I., Maghelli, N., Royer, L., Weigert, M., Myers, E., Grill, S., Drechsel, D., Hyman, A. & Alberti, S. (2015), 'A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation', *Cell* **162**, 1066–1077.

Patel, A., Malinovska, L., Saha, S., Wang, J., Alberti, S., Krishnan, Y. & Hyman, A. A. (2017), 'ATP as a biological hydrotrope', *Science* **356**, 753–756.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H. & Ferrin, T. E. (2021), 'UCSF ChimeraX: Structure visualization for researchers, educators, and developers', *Protein Science* **30**(1), 70–82.

Polo, S., Blackford, A., Chapman, J., Baskcomb, L., Gravel, S., Rusch, A., Thomas, A., Blundred, R., Smith, P., Kzhyshkowska, J., Dobner, T., Taylor, A., Turnell, A., Stewart, G., Grand, R. & Jackson, S. (2012), 'Regulation of DNA-end resection by hnRNPU-like proteins promotes DNA double-strand break signaling and repair', *Molecular cell* **45**, 505–16.

Qamar, S., Wang, G., Randle, S. J., Ruggeri, F. S., Varela, J. A., Lin, J. Q., Phillips, E. C., Miyashita, A., Williams, D., Ströhl, F., Meadows, W., Ferry, R., Dardov,

V. J., Tartaglia, G. G., Farrer, L. A., Schierle, G. S. K., Kaminski, C. F., Holt, C. E., Fraser, P. E., Schmitt-Ulms, G., Klenerman, D., Knowles, T., Vendruscolo, M. & George-Hyslop, P. S. (2018), 'FUS phase separation is modulated by a molecular chaperone and methylation of arginine cation-π interactions', *Cell* **173**, 720–734.

Rawat, P., Boehning, M., Hummel, B., Aprile-Garcia, F., Pandit, A. S., Eisenhardt, N., Khavaran, A., Niskanen, E., Vos, S. M., Palvimo, J. J., Pichler, A., Cramer, P. & Sawarkar, R. (2021), 'Stress-induced nuclear condensation of NELF drives transcriptional downregulation', *Molecular Cell* pp. 1013–1026.

Richardson, C. C. (1965), 'Phosphorylation of nucleic acid by an enzyme from T4 bacteriophage-infected Escherichia coli.', *Proceedings of the National Academy of Sciences* **54**, 158–165.

Rimel, J. K., Poss, Z. C., Erickson, B., Maas, Z. L., Ebmeier, C. C., Johnson, J. L., Decker, T.-M., Yaron, T. M., Bradley, M. J., Hamman, K. B., Hu, S., Malojcic, G., Marineau, J. J., White, P. W., Brault, M., Tao, L., DeRoy, P., Clavette, C., Nayak, S., Damon, L. J., Kaltheuner, I. H., Bunch, H., Cantley, L. C., Geyer, M., Iwasa, J., Dowell, R. D., Bentley, D. L., Old, W. M. & Taatjes, D. J. (2020), 'Selective inhibition of CDK7 reveals high-confidence targets and new models for TFIIH function in transcription', *Genes & Development* **34**, 1452–1473.

Romero, M. L. R., Yang, F., Lin, Y.-R., Toth-Petroczy, A., Berezovsky, I. N., Goncearenco, A., Yang, W., Wellner, A., Kumar-Deshmukh, F., Sharon, M., Baker, D., Varani, G. & Tawfik, D. S. (2018), 'Simple yet functional phosphate-loop proteins', *Proceedings of the National Academy of Sciences* **115**, E11943–E11950.

Romig, H., Fackelmayer, F., Renz, A., Ramsperger, U. & Richter, A. (1992), 'Characterization of SAF-A, a novel nuclear DNA binding protein from HeLa cells with high affinity for nuclear matrix/scaffold attachment DNA elements.', *The EMBO Journal* **11**, 3431–3440.

Ryan, V. H., Dignon, G. L., Zerze, G. H., Chabata, C. V., Silva, R., Conicella, A. E., Amaya, J., Burke, K. A., Mittal, J. & Fawzi, N. L. (2018), 'Mechanistic view of

hnRNPA2 low-complexity domain structure, interactions, and phase separation altered by mutation and arginine methylation', *Molecular Cell* **69**, 465–479.

Sabari, B. R., Dall'Agnese, A., Boija, A., Klein, I. A., Coffey, E. L., Shrinivas, K., Abraham, B. J., Hannett, N. M., Zamudio, A. V., Manteiga, J. C., Li, C. H., Guo, Y. E., Day, D. S., Schuijers, J., Vasile, E., Malik, S., Hnisz, D., Lee, T. I., Cisse, I. I., Roeder, R. G., Sharp, P. A., Chakraborty, A. K. & Young, R. A. (2018), 'Coactivator condensation at super-enhancers links phase separation and gene control', *Science* **361**, eaar3958.

Sawaya, M. R., Sambashivan, S., Nelson, R., Ivanova, M. I., Sievers, S. A., Apostol, M. I., Thompson, M. J., Balbirnie, M., Wiltzius, J. J. W., McFarlane, H. T., Madsen, A. , Riekel, C. & Eisenberg, D. (2007), 'Atomic structures of amyloid cross- spines reveal varied steric zippers', *Nature* **447**, 453–457.

Schneider, N., Wieland, F.-G., Kong, D., Fischer, A. A. M., Hörner, M., Timmer, J., Ye, H. & Weber, W. (2021), 'Liquid-liquid phase separation of light-inducible transcription factors increases transcription activation in mammalian cells and mice', *Science Advances* **7**, eabd3568.

Schwartz, J., Wang, X., Podell, E. & Cech, T. (2013), 'RNA seeds higher-order assembly of FUS protein', *Cell Reports* **5**, 918–925.

Shao, W., Bi, X., Pan, Y., Gao, B., Wu, J., Yin, Y., Liu, Z., Peng, M., Zhang, W., Jiang, X., Ren, W., Xu, Y., Wu, Z., Wang, K., Zhan, G., Lu, J. Y., Han, X., Li, T., Wang, J., Li, G., Deng, H., Li, B. & Shen, X. (2022), 'Phase separation of RNA-binding protein promotes polymerase binding and transcription', *Nature Chemical Biology* **18**, 70–80.

Sharp, J. A., Perea-Resa, C., Wang, W. & Blower, M. D. (2020), 'Cell division requires RNA eviction from condensing chromosomes', *Journal of Cell Biology* **219**, e201910148.

Shi, Y., Giammartino, D. C. D., Taylor, D., Sarkeshik, A., Rice, W. J., Yates, J. R., Frank, J. & Manley, J. L. (2009), 'Molecular architecture of the human pre-mRNA 3′ processing complex', *Molecular Cell* **33**, 365–376.

Skaar, J. R., Ferris, A. L., Wu, X., Saraf, A., Khanna, K. K., Florens, L., Washburn, M. P., Hughes, S. H. & Pagano, M. (2015), 'The Integrator complex controls the termination of transcription at diverse classes of gene targets', *Cell Research* **25**, 288–305.

Song, E. J., Werner, S. L., Neubauer, J., Stegmeier, F., Aspden, J., Rio, D., Harper, J. W., Elledge, S. J., Kirschner, M. W. & Rape, M. (2010), 'The Prp19 complex and the Usp4Sart3 deubiquitinating enzyme control reversible ubiquitination at the spliceosome', *Genes & Development* **24**, 1434–1447.

Sousa-Luís, R., Dujardin, G., Zukher, I., Kimura, H., Weldon, C., Carmo-Fonseca, M., Proudfoot, N. J. & Nojima, T. (2021), 'POINT technology illuminates the processing of polymerase-associated intact nascent transcripts', *Molecular Cell* **81**, 1935–1950.

Szewczak, L. B. W., Gabrielsen, J. S., Degregorio, S. J., Strobel, S. A. & Steitz, J. A. (2005), 'Molecular basis for RNA kink-turn recognition by the h15.5K small RNP protein', *RNA* **11**, 1407–1419.

Takahashi, H., Takigawa, I., Watanabe, M., Anwar, D., Shibata, M., Tomomori-Sato, C., Sato, S., Ranjan, A., Seidel, C. W., Tsukiyama, T., Mizushima, W., Hayashi, M., Ohkawa, Y., Conaway, J. W., Conaway, R. C. & Hatakeyama, S. (2015), 'MED26 regulates the transcription of snRNA genes through the recruitment of little elongation complex', *Nature Communications* **6**, 5941.

Tanikawa, C., Ueda, K., Suzuki, A., Iida, A., Nakamura, R., Atsuta, N., Tohnai, G., Sobue, G., Saichi, N., Momozawa, Y., Kamatani, Y., Kubo, M., Yamamoto, K., Nakamura, Y. & Matsuda, K. (2018), 'Citrullination of RGG motifs in FET proteins by PAD4 regulates protein aggregation and ALS susceptibility', *Cell Reports* **22**, 1473–1483.

Tatomer, D. C., Elrod, N. D., Liang, D., Xiao, M.-S., Jiang, J. Z., Jonathan, M., Huang, K.-L., Wagner, E. J., Cherry, S. & Wilusz, J. E. (2019), 'The Integrator complex cleaves nascent mRNAs to attenuate transcription', *Genes & Development* **33**, 1525–1538.

Valadkhan, S., Mohammadi, A., Jaladat, Y. & Geisler, S. (2009), 'Protein-free small nuclear RNAs catalyze a two-step splicing reaction', *Proceedings of the National Academy of Sciences* **106**, 11901–11906.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D. & Velankar, S. (2021), 'AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models', *Nucleic Acids Research* **50**, D439–D444.

Vegvar, H. E. N. d., Lund, E. & Dahlberg, J. E. (1986), '3′ end formation of U1 snRNA precursors is coupled to transcription from snRNA promoters', *Cell* **47**, 259–266.

Vernon, R. M., Chong, P. A., Tsang, B., Kim, T. H., Bah, A., Farber, P., Lin, H. & Forman-Kay, J. D. (2018), 'Pi-Pi contacts are an overlooked protein feature relevant to phase separation', *eLife* **7**, e31486.

Vervoort, S. J., Welsh, S. A., Devlin, J. R., Barbieri, E., Knight, D. A., Offley, S., Bjelosevic, S., Costacurta, M., Todorovski, I., Kearney, C. J., Sandow, J. J., Fan, Z., Blyth, B., McLeod, V., Vissers, J. H., Pavic, K., Martin, B. P., Gregory, G., Demosthenous, E., Zethoven, M., Kong, I. Y., Hawkins, E. D., Hogg, S. J., Kelly, M. J., Newbold, A., Simpson, K. J., Kauko, O., Harvey, K. F., Ohlmeyer, M., Westermarck, J., Gray, N., Gardini, A. & Johnstone, R. W. (2021), 'The PP2A-Integrator-CDK9 axis fine-tunes transcription and can be targeted therapeutically in cancer', *Cell* **184**, 3143–3162.

Viphakone, N., Sudbery, I., Griffith, L., Heath, C. G., Sims, D. & Wilson, S. A. (2019), 'Co-transcriptional Loading of RNA Export Factors Shapes the Human Transcriptome', *Molecular Cell* pp. 310–323.

Walker, J., Saraste, M., Runswick, M. & Gay, N. (1982), 'Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-

requiring enzymes and a common nucleotide binding fold.', *The EMBO Journal* **1**, 945–951.

Wang, A., Conicella, A. E., Schmidt, H. B., Martin, E. W., Rhoads, S. N., Reeb, A. N., Nourse, A., Montero, D. R., Ryan, V. H., Rohatgi, R., Shewmaker, F., Naik, M. T., Mittag, T., Ayala, Y. M. & Fawzi, N. L. (2018a), 'A single N-terminal phosphomimic disrupts TDP-43 polymerization, phase separation, and RNA splicing', *The EMBO Journal* **37**, e97452.

Wang, J., Choi, J.-M., Holehouse, A. S., Lee, H. O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D., Poser, I., Pappu, R. V., Alberti, S. & Hyman, A. A. (2018b), 'A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins', *Cell* **174**, 688–699.

Woodruff, J. B., Gomes, B. F., Widlund, P. O., Mahamid, J., Honigmann, A. & Hyman, A. A. (2017), 'The centrosome is a selective condensate that nucleates microtubules by concentrating tubulin', *Cell* **169**, 1066–1077.

Worch, R., Niedzwiecka, A., Stepinski, J., Mazza, C., Jankowska-Anyszka, M., Darzynkiewicz, E., Cusack, S. & Stolarski, R. (2005), 'Specificity of recognition of mRNA 5′ cap by human nuclear cap-binding complex', *RNA* **11**, 1355–1363.

Xiang, S., Kato, M., Wu, L., Lin, Y., Ding, M., Zhang, Y., Yu, Y. & McKnight, S. (2015), 'The LC domain of hnRNPA2 adopts similar conformations in hydrogel polymers, liquid-like droplets, and nuclei', *Cell* **163**, 829–839.

Xiao, R., Chen, J.-Y., Liang, Z., Luo, D., Chen, G., Lu, Z. J., Chen, Y., Zhou, B., Li, H., Du, X., Yang, Y., San, M., Wei, X., Liu, W., Lécuyer, E., Graveley, B. R., Yeo, G. W., Burge, C. B., Zhang, M. Q., Zhou, Y. & Fu, X.-D. (2019), 'Pervasive chromatin-RNA binding protein interactions enable RNA-based regulation of transcription', *Cell* **178**, 107–121.

Xiao, R., Tang, P., Yang, B., Huang, J., Zhou, Y., Shao, C., Li, H., Sun, H., Zhang, Y. & Fu, X.-D. (2012), 'Nuclear matrix factor hnRNP U/SAF-A exerts a global control of alternative splicing by regulating U2 snRNP maturation', *Molecular Cell* **45**, 656–668.

Yamamoto, J., Hagiwara, Y., Chiba, K., Isobe, T., Narita, T., Handa, H. & Yamaguchi, Y. (2014), 'DSIF and NELF interact with Integrator to specify the correct post-transcriptional fate of snRNA genes', *Nature Communications* **5**, 4263.

Yates, T. M., Vasudevan, P. C., Chandler, K. E., Donnelly, D. E., Stark, Z., Sadedin, S., Willoughby, J., Genomics, B. C. f. M., study, D. & Balasubramanian, M. (2017), 'De novo mutations in HNRNPU result in a neurodevelopmental syndrome', *American Journal of Medical Genetics Part A* **173**, 3003–3012.

Yonchev, I. (2021), 'Characterisation of the roles of hnRNPU and hnRNPUL1 in gene expression', *PhD thesis, The University of Sheffield* .

Zaborowska, J., Taylor, A., Roeder, R. G. & Murphy, S. (2012), 'A novel TBP-TAF complex on RNA Polymerase II-transcribed snRNA genes', *Transcription* **3**, 92–104.

Zhang, M., Zhang, H., Li, Z., Bai, L., Wang, Q., Li, J., Jiang, M., Xue, Q., Cheng, N., Zhang, W., Mao, D., Chen, Z., Huang, J., Meng, G., Chen, Z. & Chen, S.-J. (2022), 'Functional, structural and molecular characterizations of leukemogenic driver MEF2D-HNRNPUL1 fusion', *Blood* pp. 1390–1407.

Zhang, Z. & Gilmour, D. S. (2006), 'Pcf11 Is a Termination Factor in Drosophila that Dismantles the Elongation Complex by Bridging the CTD of RNA Polymerase II to the Nascent Transcript', *Molecular Cell* **21**, 65–74.

Zhang, Z., Klatt, A., Henderson, A. J. & Gilmour, D. S. (2007), 'Transcription termination factor Pcf11 limits the processivity of Pol II on an HIV provirus to repress gene expression', *Genes & Development* **21**, 1609–1614.

Zuo, L., Zhang, G., Massett, M., Cheng, J., Guo, Z., Wang, L., Gao, Y., Li, R., Huang, X., Li, P. & Qi, Z. (2021), 'Loci-specific phase separation of FET fusion oncoproteins promotes gene transcription', *Nature Communications* **12**, 1491.