RNA target recognition by IMP1 – a molecular investigation at the transcriptome level

Christopher Gallagher

University College London

and

The Francis Crick Institute

PhD Supervisor: Andres Ramos

A thesis submitted for the degree of

Doctor of Philosophy

University College London

February 2018

Declaration

I Christopher Gallagher confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

RNA binding proteins orchestrate the assembly of macromolecular RNA-protein complexes that regulate post-transcriptional regulation. Disruption of these finely tuned processes can result in an array of diseases. RNA binding proteins are multi-domain proteins organised into a modular structure. It is through this structure that they can recognise a vast repertoire of RNA targets by employing combinatorial binding. However, this mechanism is poorly understood.

The IMP1 protein provides a model system in which we can investigate how domain sequence specificity and combinatorial binding define *in vivo* RNA selection. IMP1 recognises RNA via six putative RNA binding domains (two N-terminal RNA recognition motifs and four C-terminal K-homology domains). To date, IMP1 is known to bind a diverse range of RNA targets, both in homeostatic cellular events and in cancer. However, detailed information as to how IMP1 recognises these targets, especially at the individual domain level in context of the full-length protein is lacking.

I have implemented a structural driven mutational approach to modify the RNA recognition properties of the individual RNA binding domains of the IMP1 protein. I have successfully generated iCLIP libraries for mutant IMP1 proteins where each KH domain was mutated in turn to inhibit RNA binding. From comparative analysis of these data sets I have begun to explore *in vivo* RNA target selection at the individual domain level and observed an altered RNA binding pattern to the ACTB mRNA.

I have investigated *in vitro,* the RNA sequence specificity of the KH3-KH4 and the RRM1-RRM2 domains. This has led to the design of a mutant which shifts the RNA specificity of the KH3 domain. In addition to the identification that the RRM domains of IMP1 and IMP3 specifically recognise different RNA target sequences with different affinities, and the key residues involved in RNA recognition. These findings will aid in the further understanding of IMP *in vivo* RNA target selection.

Acknowledgements

I would like to thank my supervisor Andres Ramos for the opportunity to work on this project and for the hours and hours of discussion. I would also like to thank Virginia Castilla for the collaboration with the iCLIP part of this thesis. Beppe Nicastro for the help with the structural design of the KH selectivity mutants and the performing of the ITC. In addition to all the other members of the lab, past and present, who have taught me so much along the way and made the lab a pleasant environment to work in.

I would like to thank the members of the MRC NMR centre, especially Geoff Kelly for helping with the setting up and explaining of the NMR experiments in this thesis. Laura Masino for the help with the CD data in this project.

Thank you to my thesis committee, Steve Smerdon, Ian Taylor and Peter Rosenthal for guidance and input into the project.

I would like to say a big thank you to my family, especially Anita Birkett, Paul Gallagher and Samantha Curran

Finally, I would like to say a thank you to all my friends who have supported me throughout my PhD. In particular, Christine Richter and Caterina Alfano

Table of Contents

Table of figures

List of tables

Abbreviations

4E-BP 4E binding protein

ADARs adenosine deaminases that act on RNA

ALS amyotrophic lateral sclerosis

AREs AU-rich elements

Bcl-x Bcl-2 like protein 1

CD circular dichroism

CEF chicken embryo fibroblast

CF1A cleavage factor 1A

CLIP cross-linking immunoprecipitation

CRD-BP CRD-binding protein

dsRBD double stranded RNA binding domain

eCLIP enhanced CLIP

eIF4E eukaryotic translation initiation factor 4E

EMSA electrophoretic mobility shift assay

FTD frontotemporal dementia

FUS fused in sarcoma protein

HITS-CLIP high-throughput sequencing CLIP

hnRNP A2 heterogeneous nuclear ribonucleoprotein A2

Hrp1 nuclear polyadenylated RNA-binding protein 4

HSQC heteronuclear single-quantum correlation

HuD Hu-antigen D

HuR Hu-antigen R

IMP IGF2 mRNA binding protein

ITC isothermal titration calorimetry

KH K-homology

KSRP KH type-splicing regulatory protein

miRNA microRNA (miRNA)

MMP-9 matrix metalloproteinase-9

mTOR mammalian target of rampamycin

NMR nuclear magnetic resonance spectroscopy

NOVA1 and 2 neuro-oncological ventral antigen 1 and 2

NPC nuclear pore complex

PAR-CLIP photoactivatable ribonucleoside CLIP

PI3K phosphatidylinositol 3-kinase

PIKE phosphoinositide 3-kinase enhancer

POMA paraneoplstic opsoclonus-myoclonus ataxia

pre-miRNA precursor miRNA

pri-miRNA primary-miRNA

PTB polypyrimidine tract-binding protein

RBDs RNA binding domains

RBPs RNA binding proteins

Rip-Chip RNP immunoprecipitation-microarray

RISC RNA-induced silencing complex

RNPs ribonucleoprotein particles

RNP1 RNP motif 1

RNP2 RNP motif 2

RRM RNA-recognition motif

Sam68 The Scr-associated in mitosis 68kDa

SIA scaffold independent analysis

Small interfering RNA (siRNA)

SMN1 The Survival of Motor Neurone 1

SRSF1 Serine/arginine-rich splicing factor 1

TDP-43 TAR DNA binding protein 43

TEV tobacco etch virus

U2A' U2 small nuclear ribonucleoprotein A'

U2B" U2 small nuclear ribonucleoprotein B"

UTRs untranslated regions

ZBP1 Zipcode-binding protein 1

ZnF/ZF zinc fingers

Chapter 1. Introduction

1.1 Post-transcriptional gene regulation

During and after transcription RNAs are subjected to multiple processing and regulatory steps that are coordinated by RNA-binding proteins (RBPs).[1,2] These finely tuned regulatory mechanisms are necessary to expand the diversity of the genome and to the specialisation and functioning of a cell. Misregulation of these fine-tuned processes has been linked to cancer, autoimmune, and neurodegenerative diseases.[3–6] A molecular understanding of how RBPs control these mechanisms, and the interconnection with transcriptional and post-translational control networks, is the first step in exploiting RBPs for the potential development of therapeutics.

1.1.1 RNA Metabolism

mRNA transcripts are subject to many regulatory processes that are tightly regulated within the cell. These processes include mRNA splicing, mRNA capping, nuclear export, polyadenylation, localisation, degradation and protection (Figure 1.1). These processing steps are essential for controlling post-transcriptional gene expression and regulating cell differentiation and function. RNA transcripts are rarely in isolation within the cell and it is the RNA binding proteins, in addition to trans-acting RNAs, that regulate and control transcript metabolism.[1]

RNA transcripts associate with numerous RBPs to form complexes called ribonucleoprotein particles (RNP). The formation of RNPs is highly dynamic with components associating and disassociating at various stages. This dynamic system enables distinct sets of RBPs to associate with RNA transcripts at

different time points and cellular locations, thus allowing RNA fate to be temporally and spatially regulated.[7]

At the site of transcription within the nucleus precursor-mRNA (pre-mRNA) undergoes maturation processing before the mature mRNA transcript is exported into the cytoplasm (Figure 1.1). As transcription occurs RNPs form co-transcriptionally on the nascent transcript, they then mediate the nuclear processing of the pre-mRNA to a fully mature mRNA transcript. First, a 7-methylguanosine cap is added to the 5' end of the transcript.[8] This cap protects the transcript from exonuclease degradation and later allows recognition by the ribosome which in turn initiates translation. Polyadenylation introduces a poly(A) tail of around 200 nucleotides to the 3' end of the mRNA.[1,8] This poly(A) tail is important for nuclear export of the transcript to the cytoplasm where it also later aids in the protection of the mRNA molecule from enzymatic degradation as well as playing a role in translation termination. The cleavage and polyadenylation specificity factor (CPSF) complex is essential in mediating the polyadenylation process.[9,10] The mRNA also undergoes splicing to remove the intron sections from the transcript. At this stage the mRNA can undergo alternative splicing where different introns / exons are removed which ultimately results in the formation of different protein isoforms. It is estimated at over 90% of human genes express multiple mRNA isoforms due to alternative splicing,[11] highlighting the importance of this process. Additional modifications can also be made to pre-mRNA in the nucleus, typically within non-coding regions,[12] including insertions, deletions and deamination.[1,8,13] Disruption of the nuclear processing steps can prevent export of the transcript and result in degradation within the nucleus.

Once the pre-mRNA has been fully processed the mature mRNA is transported into the cytoplasm through the nuclear pore complex (NPC) with the aid of several adaptor and receptor proteins, such as the nuclear RNA export factor 1 and the adaptor protein REF.[14,15] Due to the different mRNA processing steps occurring in the nuclear compartment of the cell compared to the cytoplasmic compartment, the associated nuclear proteins are shed from the transcript during nuclear

export. As the nuclear proteins are shed, cytoplasmic RBPs can bind the transcript and control the cytoplasmic fate of the mRNA.[1,8,16] Once in the cytoplasm of the cell mRNA transcripts have multiple end points.[17] The mRNA transcript can be translated into protein via the ribosome. This requires the association of translation initiation factors with the 5' cap on the mRNA which are then recruited to the ribosome complex.[8] However, most cells spatially and temporally regulate protein synthesis. Transcripts can be translocated to specific cellular compartments in a translationally repressed state until the desired site of translation is reached. One such example is the localisation of the ACTB mRNA which is held in a translationally repressed state by the zipcode-binding protein 1 (ZBP1) until located to the desired cellular compartment.[18] Depending on the current requirements of the cell, mRNA transcripts can also be targeted for degradation, or stored in granules for later translation when the protein is required.[16] (Figure 1.1)

The stability of an mRNA transcript dictates its life time within the cell. Highly stable transcripts have the ability to encode for more protein copies compared to less stable and shorter-lived transcripts. RBPs have great influence in altering the stability of transcripts and so ultimately control protein levels within the cell. RBPs control transcript stability by binding to special regulatory elements within the mRNA transcript. These elements are commonly found in the 3' and 5' untranslated regions (UTR).[19,20] As these *cis*-acting RNA regulatory elements reside mainly in untranslated regions, they have greater freedom to diverge due to weaker evolutionary pressure. Furthermore, small noncoding RNAs, such as small interfering RNA (siRNA) and microRNA (miRNA), can work in complex with RBPs to target mRNA transcripts to the RNA-induced silencing complex (RISC) to supress translation or promote degradation.[16,17,20,21] (Figure 1.1)

In conclusion, RNA binding proteins mediate post-transcriptional gene regulation through the control of RNA metabolism, localisation, and degradation. They exert they influence on RNA transcripts by acting via macromolecular RNP complexes. The assembly of RNPs relies on specific protein-RNA recognition events, protein

post-translational modifications, and interactions with additional proteins and non-coding RNAs. Disruption of these recognition events and RNP complex formation can result in a variety of diseases.

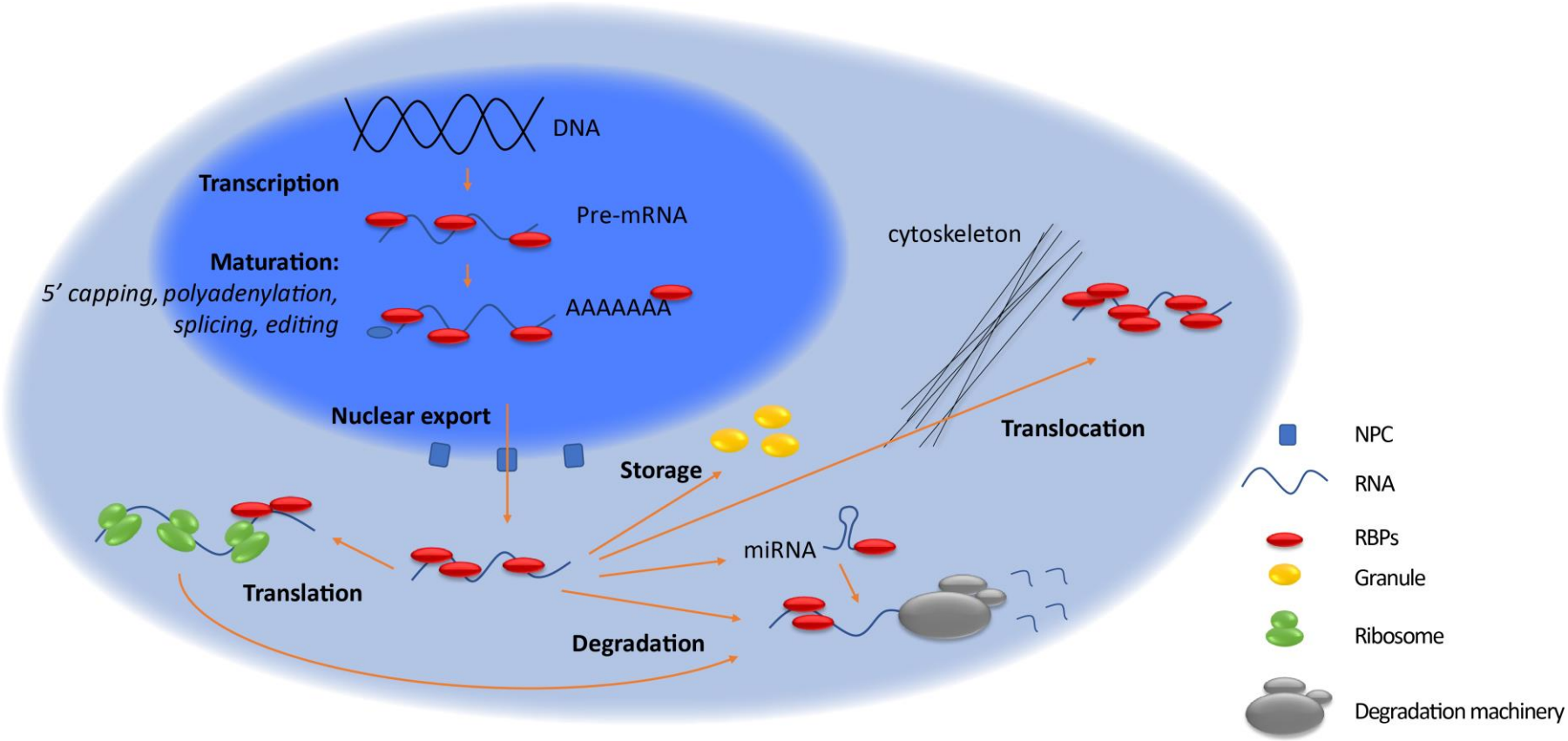


Figure 1.1: Overview of mRNA metabolism
Schematic overview representing the multiple steps in mRNA metabolism.

1.2 RNA binding proteins and disease

RNA binding proteins control all aspects of RNA metabolism and are critically important for cell function. As RBP functionality in gene regulation is dependent on their ability to selectively recognise and bind RNA targets, disruption of these events can cause a spectrum of pathologies and syndromes. Mutations in RBPs, mutations in protein partners that co-regulate with RBPs, aberrant expression of RBPs or protein partners; and alterations of RNA target elements recognised by RBPs, are examples of ways the process controlled by RBPs can be disrupted. Two major disease areas which have been linked to defects in RBP function are neurodegenerative disease and cancer.

1.2.1 RNA binding proteins in neurodegenerative disorders

Neuronal cells are particularly susceptible to defects in RBP function due to the high proportion of RNA transcripts that are alternatively spliced compared to other tissues.[22] Several neurodegenerative diseases result from the disruption of alternative splicing pathways. The survival of motor neuron 1 (SMN1) protein is known to be an essential component for the assembly of small nuclear ribonucleoproteins (snRNPs) that interact with the spliceosome. A homozygous deletion in the SMN1 gene was shown to result in the development of spinal muscular atrophy (SMA). The loss of the SMN1 gene results in extensive splicing defects within motor neurons which leads to SMA.[23] Another RBP dependent neurotological disease is paraneoplastic opsoclonus-myoclonus ataxia (POMA). In this disease system the RBPs neurological ventral antigen 1 and 2 (NOVA1, NOVA2) are targeted by auto-antibodies. The NOVA1 and NOVA2 proteins function to regulate the alternative splicing of proteins that are involved in inhibitory synaptic transmission. Attack of the NOVA proteins via auto-antibodies result in disruption of these transmission pathways and leads to the onset of POMA.[24]

Clinical studies on the neurological disorders, frontotemporal dementia (FTD), and amyotrophic lateral sclerosis (ALS), identified clinical overlap among these diseases with patients initially diagnosed with ALS often also displaying signs of FTD. Genetic screening of these patients identified genetic variability associated with both FTD and ALS.[25] These studies have identified mutations in two RNA binding proteins, The TAR DNA binding protein (TDP-43)[26] and the fused in sarcoma protein (FUS).[27] However, the relationship between the mutations and the development of FTD and ALS is less clear.

TDP-43 contains two RNA recognition motif (RRM) domains and a C-terminal glycine-rich region. The RRM domains have been shown to recognise UG and Poly-A RNA sequences.[28,29] Functionally, the protein is known to be involved in exon skipping during RNA splicing, in addition to a reported involvement in the biogenesis of miRNA via the Drosha and Dicer complexes. Most of the mutations identified in diseased patients reside in the C-terminus of the TDP-43 protein rather than the RNA binding RRM domains.[30] Usually TDP-43 resides in the nucleus of neuronal and glial cells. However, mutated forms of TDP-34 display aberrant localisation as cytoplasmic inclusions with the protein ubiquitinated, cleaved or abnormally phosphorylated.[31,32] Although a strong link between mutations of TDP-43 and the progression of FTD and ALS has been established, the reason behind how these mutations lead to the onset of disease is less well understood. Hypothesis include, disruption to splicing networks due to the loss of the protein from the nucleus; possible cytotoxic effect of accumulated protein in the cytoplasm; or altered biogenesis of RNA targets resulting from modified RNA recognition of the TDP-43 protein.

The FUS protein is a member of the heterogeneous nuclear ribonucleoprotein (hnRNP) complex family which are involved in pre-mRNA splicing and in the export of fully processed mRNA to the cytoplasm. The motif by which FUS recognises RNA is less well defined than that of TDP-43.[33] FUS is also a multifunctional protein, a common feature of most RBPs, and it is not well understood what functions affected by FUS mutations cause FTD or ALS.

1.2.2 RNA binding proteins in cancer

It is known that cancer is a complex genetic disease in which several key regulatory pathways must be altered to result in a cancer phenotype. These include disruption of the cell cycle and growth, cell migration, evasion from normal apoptotic controls, and the ability for cells to grow without correct external stimuli. As RNA binding proteins RBPs are typically multifunctional, due to their ability to post-transcriptionally regulate multiple RNA transcripts, defects in the normal function of RBPs has the potential to modify several regulatory networks within the cell. For this reason, RBPs that regulate many RNA transcripts usually display a highly regulated expression pattern as slight changes in expression levels can result in significant disruption to the post-transcriptional networks. However, disruption of the mechanisms controlling RBP expression is commonly observed in cancers as many tumours display severe upregulation of several RNA binding proteins.

As previously stated RBPs play fundamental roles in alternative splicing, a mechanism that is often disrupted in many neurological disorders. Similarly, disruption of splicing pathways can also result in malignant phenotypes. The Scr-associated in mitosis 68 kDa (Sam68) protein is a member of the signal transduction activator of RNA metabolism (STAR) protein family and is involved in alternative splicing.[34] The Sam68 protein is upregulated in breast, renal, prostate and cervical cancers. Sam68 alternatively spliced transcripts include the proto-oncogene cyclin D1,[35] the cell surface receptor CD44 which is involved in cancer cell proliferation,[36] and Bcl-x.[37] The alternative splicing of Bcl-x is dependent on the phosphorylation state of the Sam68 protein. In its phosphorylated form Sam68 promotes the formation of the Bcl-x(L) splice isoform rather than Bcl-x(S).[31] The Bcl-x(L) protein is anti-apoptotic and thus inhibits cell death contributing to the development of cancer.

Disruption in the translation of specific transcripts can also lead to tumorigenesis. The 5’ cap binding protein eukaryotic initiation factor 4E (eIF4E) binds to the 5’ end of mRNA transcripts. The protein then recruits eIF4G which in turn assembles

with the ribosomal complex resulting in the initiation of translation. Over expression of eIF4E in fibroblasts was shown to result in malignant transformation of these cell types,[38,39] in addition to eIF4E being upregulated in a variety of cancers (gastric, lung, skin, colon and breast).[40] This increased expression of eIF4E was shown not to result in a general increase in protein levels within these cancer cells, but rather a selective increase in expression of oncoproteins. Many mRNA transcripts which encode for oncoproteins contain large 5' UTR regions that often contain stable RNA structures. In normal conditions these transcripts are poorly translated due to the structured nature of the 5' UTR.[41] The binding of eIF4E recruits eIF4F which contains a helicase component that facilitates translation by unwinding structures in the 5' UTR. Elevation in eIF4E levels therefore has a greater effect on oncogene transcripts compared to efficiently translated mRNA transcripts. eIF4F activity is also regulated by the eIF4F binding protein (E4-BP). These proteins bind to eIF4F and prevent formation of the cap-binding complex required to recognise the 5' end of transcripts. Phosphorylation of E4-BP inhibits the interaction with eIF4F and thus enabled eIF4F to initiate translation. The protein kinase mTor is know to phosphorylate the E4-BP. Direct phosphorylation of the eIF4F protein by the MAPK-integrin kinase members MNK1 and MNK2 also enhances cap-depended initiation translation.[39]

Changes in the mechanisms controlling mRNA stability can also result in cancer progression. A member of the Hu family of proteins, Hu-antigen R (HuR), is upregulated in a variety of cancer types. The Hu family of proteins control mRNA stability via association with AU-rich elements (AREs) in the 3' UTR of specific mRNA transcripts.[42] Target transcripts include; cell growth and proliferation controlling cyclins and endothelial growth factor (EGF);[43,44] the hypoxia-inducible factor 1 alpha (HIF-α) which facilities angiogenesis;[45,46] and matrix metalloproteinase 9 (MMP9) which mediates cellular invasion and metastasis.[47] Over expression of the HuR protein increases the abundance of these proteins within the cell which leads to cancer progression. Conversely, the depletion of the Hu family of proteins via auto-antibodies in neuronal cells, results in paraneoplastic encephalomyelitis and sensory neuropathy.[48] Thus, this example

further illustrates the precise control of RBP expression levels required to maintain healthy cells.

These are a selection of examples where modifications in the physiological function of RBPs results in disease phenotypes. The link between alterations in RNA binding proteins and the development of disease is not always well understood. To develop therapeutics designed at treating diseases resulting in dysregulation of RBP controlled pathways we must first understand the mechanisms by which the RBP recognises RNA targets. Only once we understand the system can we then try to manipulate the process to restore normality. This often requires understanding the method by which RBPs interact with specific targets, at both the structural and mechanistic level. Determining the affect the RBP exerts on the target RNA; stability, localisation, or degradation. In addition to understanding the wider signalling or cascade networks affected by such protein-RNA interactions.

1.3 RNA binding proteins and combinatorial RNA recognition

RNA-binding proteins are functionally diverse within cells and can recognise a diverse array of RNA targets. They can bind transcripts in a sequence specific manner or recognise RNA secondary structures. The main site of RNA interactions in canonical RBPs are the RNA binding domains (RBD). There is a variety of different classes of RBDs, with each group sharing a similar domain fold and RNA recognition properties. However, the pool or known RBDs is much more limited than the variety of RNA transcripts RBPs are known to bind. Therefore, RBPs often consist of multiple RBDs, in various combinations and structural arrangements (Figure 1.2). The individual RBD can each recognise a short stretch of RNA, often with low to intermediate affinity with Kds in the µM to mM range. The accumulation of multiple RBD interactions of the same RBP with the RNA target increases the interacting surface, and results in a highly specific and high affinity interaction in the nM binding range. The modularity of RBP-RNA binding also enables the recognition of RNA sequences that are separated by

non-specific stretches of nucleotides. It is therefore, the different domain contributions and positioning of RBDs that determine binding specificity. This method of binding is termed combinatorial recognition and allows RBPs to employ a select repertoire of RBDs to recognise different RNA targets. This mechanism of establishing multiple weaker binding interactions creates a dynamic situation where the assembly and disassembly of protein-RNA complexes can be easily regulated. This reversibility is essential as mRNA transcripts must change its regulatory RBPs according to the needs of the cell.[17,49,16]

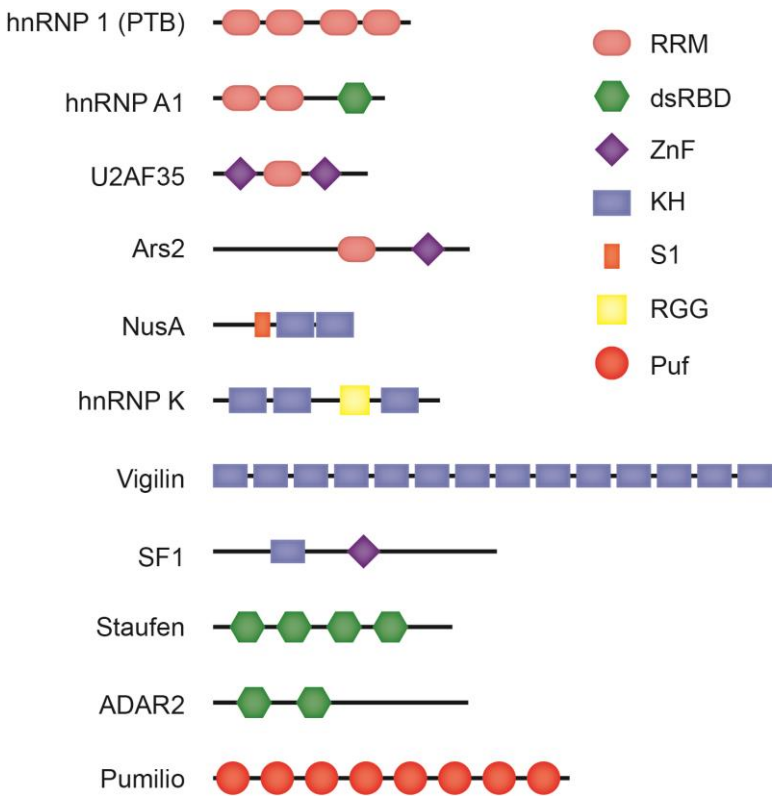


Figure 1.2: Common modular structure of RNA binding proteins
Schematic displaying the different combinations of the most common RNA binding domains (RRM: RNA-recognition motif, KH: K-homology domain, dsRBD: double-stranded RNA binding domain, and ZnF: Zinc-finger) in a selection of well characterised RNA binding proteins. Examples highlight the variability of the number of RBDs (up to 14 in vigilin), and the combinations of different RBDs within the same protein. Different RBDs are displayed by different colours and shapes (key top right). Less common RBDs displayed in figure - S1: splicing factor-1, Puf: Puf RNA-binding repeat, RGG: Arg-Gly-Gly box.

Combinatorial RNA binding is so far not well characterised. Most RNA binding domains bind a short stretch of RNA with low to medium affinity. Often, individual RNA binding domains are able to bind with suboptimal RNA recognition sequences with lower affinity than their preferred RNA target sequences. This in part, makes understanding combinatorial binding of RBPs challenging. However, the RBP Pumilio is a good example that demonstrates the modularity of RBP-RNA interactions, and how this mechanism of binding expands specificity and affinity. Pumilio is a unique example as the proteins RNA binding domains can recognise just a single nucleic acid. Pumilio contains eight Puf RNA binding

repeats (Puf)[50] (Figure 1.3). Each repeat consists of a small α-helical structural with a N- and C-terminal flanking region.[51] The individual repeats alone can recognise just a single nucleobase with low affinity. The combination of 8 Puf repeats allows the protein to recognise eight nucleotides with sequence specificity and high affinity.[52]

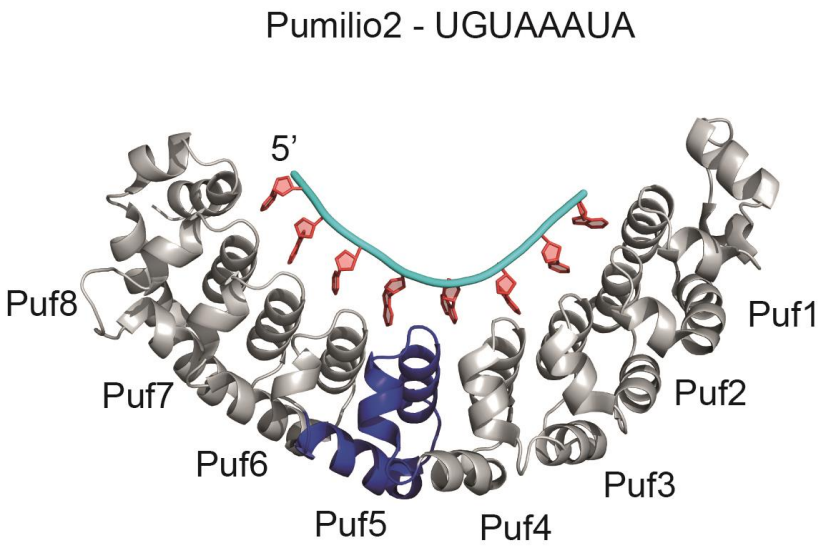


Figure 1.3: Pumilio2 structure in complex with UGUAAAUA RNA
Each Puf domain consists of 3 small α- helices and can recognise one RNA base. Puf5 is coloured in Blue to highlight the domain structure with the remaining Puf domains in Grey. RNA bases are coloured in Red and the phosphate backbone in Cyan. Puf domains specifically recognise each RNA nucleobase via the Watson-Crick edge which can be seen by the orientation of the RNA bases towards the Puf domain fold.

Great understanding of RNA recognition has developed through structural and biochemical studies of single RBDs interacting with RNA targets. This has allowed parallels to be drawn between groups of the same type or RBD that share a common structure and RNA recognition mechanisms. Commonly occurring

RBDs are summarised in Table 1.1. The RNA-recognition motif (RRM), K-homology (KH) domain, double-stranded RNA binding domain (dsRBD), and zinc fingers (ZnF) are the most abundant RNA binding domains in humans. I will briefly discuss the common properties of the RRM and KH domains here, as these are the domains my studies have been focused on, and expand further in later chapters (KH: Chapter 4 and RRM: Chapter 5)

Domain	Topology	RNA-recognition surface	Protein:RNA interactions	Representative structures (PDB ID)
RRM	αβ	Surface of β sheet	Interacts with about four nucleotides of ssRNA through stacking, electrostatics and hydrogen bonding	U1A N-terminal RRM (1URN)
KH	αβ	Hydrophobic cleft formed by variable loop between β2, β3 and GXXG loop. Type II: same as type I except variable loop is between β1 and β2	Recognises about four nucleotides of single stranded RNA through hydrophobic interactions between non-aromatic residues and the bases; sugar-phosphate backbone contacts from the GXXG loop and hydrogen bonding to bases	NOVA1 KH3 (type I) (1EC6) NusA (type II) (2ASB)
ZnF-C2H2	αβ	Primary residues in α helices	Protein side chain contacts to bulged bases in loops and through electrostatic interactions between side chains and the RNA backbone	Finger 4-6 of TFIIIA (1UN6)
ZnF-CCCH	Little regular secondary structure	Aromatic side chains form hydrophobic binding pockets for bases that make direct hydrogen bonds to protein backbone	Stacking interactions between aromatic residues and bases create a kink in the RNA that allows for the direct recognition of Watson-Crick edges of bases by the protein backbone	Finger 1 and 2 of TIS11d (1RGO)
Puf	α	Three conserved amino acid residues positioned in the middle of the repeat make contact with an RNA base	Binding pockets for bases provided by stacking interactions; specificity dictated by hydrogen bonds to the Watson-Crick face of a base by two amino acids in helix α2	Pumilio (1M8Y)
PAZ	αβ	Hydrophobic pocket formed by OB-like β barrel and small αβ motif	Recognises single stranded 3' overhangs of siRNA through stacking interactions and hydrogen bonds	PAZ (1SI3), Argonaute (1U04), Dicer (2FFL)
PIWI	αβ	Highly conserved pocket, including a metal ion that is bound to the exposed C-terminal carboxylate	Recognises the defining 5' phosphate group in the siRNA guide strand with a highly conserved binding pocket that includes a metal ion	PIWI (1YTU), Argonaute (1U04)
dsRBD	αβ	Helix α1, N-terminal portion of helix α2, and loop between β1 and β2	Shape-specific recognition of the minor-major-minor groove pattern of dsRNA through contacts to the sugar-phosphate backbone; specific contacts from the N-terminal α helix to RNA in some proteins	Staufen dsRBD3 (1EKZ)

Table 1.1: Commonly occurring RNA binding domains and their shared properties
Taken from[49]

1.3.1 RNA-recognition motif (RRM)

RRM domains most commonly recognise single stranded nucleic acids, however, some examples have also been shown to recognise structured RNA elements. They are also documented in mediating protein-protein interactions. The RRM domain is the most abundant class of RBD and is conserved in all domains of life. In humans 497 proteins have been identified that contain at least one RRM. This potentially accounts for 2% of all human gene products. Typical to many RBDs, RRMs are often found as multiple copies within a protein. Over 40% of RBPs containing RRMs have between two to six RRMs. The second most abundant RBD found in association with RRM domains are zinc finger domains.[53]

A single RRM typically consists of around 90 amino acid residues, but some RRMs with structured extensions have been identified. For example, the flanking N- and C-terminal regions outside the standard RRM fold can adopt secondary structures. Examples of these RRM extensions include the La C-terminal RRM[54], U1A N-terminal RRM[55] and CstF-64 C-terminal RRM.[56] These expansions to the standard fold have defined several RRM domain sub-families with non-canonical nucleic acid recognition properties.

Currently over 30 RRM structures have been solved either by NMR or X-ray crystallography. Disregarding the variations from the common RRM structure, the RRM fold is a αβ sandwich consisting of two α-helices packed against four antiparallel β-strands that comprise a single β-sheet. The flat β-sheet of the RRM domain, in addition to the connecting loops and N- and C-termini, recognise many different RNA sequences and shapes (Figure 1.4).

All canonical RRM domains contain two conserved motifs termed RNP1 and RNP2. RNP1 is commonly 8 amino acids in length, and RNP2 being 6 residues long. These conserved residues are located in the central strands of the β sheet. They are mostly composed of positively charged or aromatic residues, and these make the primary contact surface for the target RNA. Alignment of canonical RRMs identified the general sequence of RNP1 to be (R/K)-G-(F/Y)-(G/A)-(F/Y)-

(I/L/V)-X-(F/Y) and RNP2 (I/L/V)-(F/Y)-(I/L/V)-X-N-L. The most conserved residues found in RRMs are the four residues that contribute most to RNA binding, namely RNP1 positions 1, 3 and 5 and RNP2 position 2. RNP1 position 5 and RNP2 position 2 are conserved planar residues which stack against two nucleobases of the interacting nucleic acid molecule. These are the most frequently found interactions in RRMs and this characteristic binding results in the nucleic acid lying across the surface of the β sheet. Position 1 of RNP1 is generally a positively charged residues which can interact with the negatively charged phosphate group on the backbone of the nucleic acid, while position 3 is an aromatic residue which interacts hydrophobically with the sugar rings of the stacked bases (Figure 1.4).[53,57,58]

RRMs are highly versatile in their mode of RNA recognition: in canonical RRM-RNA interactions, the bound RNA lies across the β-sheet and contacts one or more key residues in the conserved RNP1 and RNP2 motifs. However, the number of residues within the β-sheet surface directly contacting the RNA varies considerably. The β-sheet surface of a single RRM can contact up to four nucleotides, while engaging the loops or N- / C-terminal regions external to the β-sheet can allow binding of up to six nucleotides. In some cases, the β-sheet surface is not involved in RNA binding.[57,58]

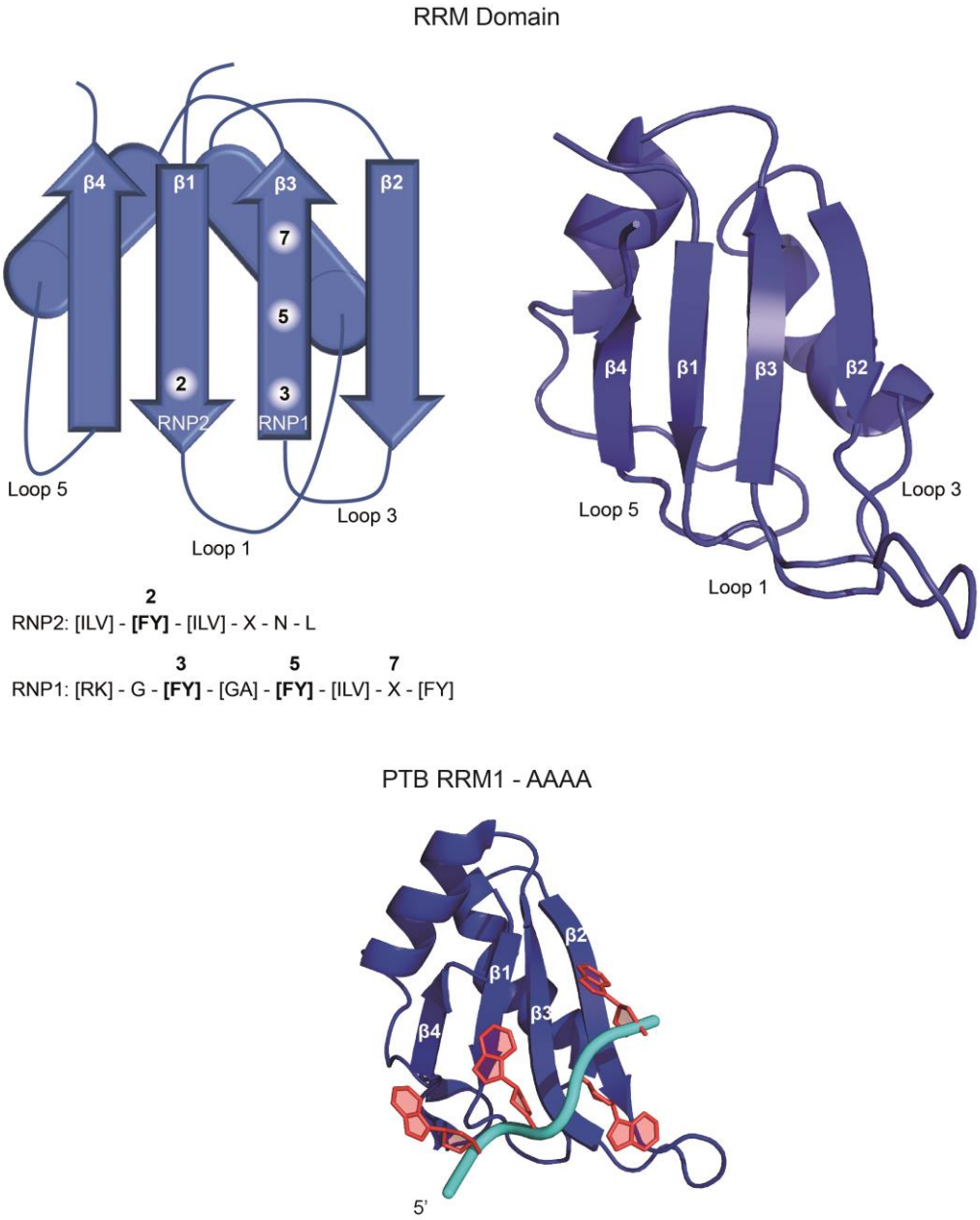


Figure 1.4: Canonical RRM domain fold
Upper: Schematic representation of an RRM domain (left) and structure of the second RRM of hnRNP A1 (right) (PDB:1UP1). The sequence of the conserved RNP1 and RNP2 motifs are displayed below. Figure shows the β-sheet annotated with the conserved aromatic residue positions in RNP1 and RNP2 (2, 3, and 5) and variant residue often involved in conferring specificity (7). Lower: Structure of PTB RRM1 in complex with poly-A RNA target (PDB:1CVJ). RNA bases are coloured in Red and phosphate backbone in Cyan. The structure displays canonical RRM RNA binding with the RNA bases binding across the β-sheet.

The Prp24 protein contains an occluded RRM (oRRM). Prp24 is involved in the assembly of the spliceosome and consists of thee canonical RRM domains, and a fourth C-terminal RRM that lacks the conserved RNP1 and RNP2 motifs and adopts a non-canonical fold. This C-terminal oRRM domain contains two flanking α-helices that occlude the β-sheet preventing canonical RNA recognition. RNA binding is instead mediated via the flanking α-helices that form a large electropositive surface which enables RNA binding.[59]

Another variation of the canonical RRM domain is the di-RRM. The multi-domain splicing factor U2AF65 protein contains a tandem RRM1 and RRM2 domain which exhibit conformational selection of RNA targets. In the absence of RNA, the conformation of the two domains is such that the RNA binding surface of RRM1 is partially occluded via an interaction with the RRM2 domain. Upon RNA binding they display an open conformation where the β-sheets have a parallel arrangement forming an extended RNA binding surface. Therefore, the extension of the RNA binding surface in the open conformation allows for an interaction with a longer stretch of RNA compared to the close form, where only the RRM2 is capable of binding RNA.[60]

The Serine/arginine-rich splicing factor 1 (SRSF1) protein contains a different RRM domain termed a pseudo-RRM. Pseudo-RRMs do not utilise their β-sheet for RNA binding, but instead RNA binding is mediated by the α-helix 1.[61] Interestingly this helix in SRSF1 mediates both protein-protein[62] and protein-RNA[63] interactions. This mode of RNA binding is conserved throughout SF proteins that contain two RRM domains and contain a conserved SWQDLKD.[61] To date, one other structure has been solved that shows a bacterial RRM to interact with interact with RNA using the α-helix 1.[64] However, this RRM is not a pseudo-RRM as it lacks the conserved SWQDLKD motif and does not bind RNA in a sequence specific manner

Finally, the heterogeneous nuclear ribonucleoprotein (hnRNP) F protein is involved in the regulation of mRNA metabolism by associating with G-rich RNA

sequences. Association with these G-rich sequences is mediated through an atypical RRM binding mode mediated via the proteins quasi RRM (qRRM) domains. The qRRM domains lack the canonical RNP motifs and instead specifically interact with RNA through the highly conserved loops that connect the β-strands without involvement of the classical RNA binding surface.[65,66]

RRM domains have also been reported to bind DNA and mediate protein-protein interactions. An elegant example for this is the interactions between the factors involved in the Far-upstream element (FUSE) mediated regulation of the MYC oncogene. FUSE is an AT-rich DNA element located upstream of the MYC oncogene promoter. The FUSE binding protein (FBP) binds the non-coding strand of the FUSE element via its four KH domains. The carboxyl- and amino-terminal regions of the FBP then mediate the activation of the Transcription factor II H (TFIIH), and the recruitment of the FUSE interacting repressor (FIR) respectively. The FIR protein contains an RRM didomain which mediates an interaction between the FUSE element and FBP protein, representing the RRM domain in a non-canonical role as a mediator of DNA-protein and protein-protein interactions.[67]

1.3.2 K-homology (KH) domain

The hnRNP K homology (KH) domain was first identified in the human heterogeneous nuclear ribonucleoprotein K (hnRNP K), and in turn is where the domain obtained its name. KH domains recognise single stranded nucleic acid in a sequence specific manner. Proteins containing KH domains have been identified in archaea, bacteria and eukaryotes, and have been shown to regulate functions including transcriptional and translational regulation. The canonical KH domain structure consists of three α-helices that pack onto the surface of a central antiparallel β-sheet. However, the three-dimensional arrangement of the secondary structural elements is different between the eukaryotic type I KH domain and the bacterial type II domain (Figure 1.5).[68]

Not all identified KH domains have been shown to bind nucleic acid. All the KH domains that have been confirmed as RNA binding domains contain a conserved GXXG loop between α1 and α2 and a variable loop between β2 and β' in type I and β' and β1 in type II domains. KH domains that contain a classical domain fold but are lacking the GXXG motif have shown no nucleic acid-binding activity.[69–72]

The canonical KH domain fold generates a hydrophobic groove. The conserved GXXG motif orients nucleic acids towards this hydrophobic groove, which is the site of RNA recognition. The phosphate backbone of the first two nucleobases interact with the residues in the GXXG loop via electrostatic interactions although the precise details of these structural interactions vary. This interaction helps orientate the Watson-Crick edge of the bases of residues 2 and 3 for specific recognition in the groove. KH domains have been shown to recognise up to four nucleotides specifically using a combination of hydrogen-bonding, electrostatic interaction and shape complementarity. Strong nucleobase discrimination is often observed for one or two nucleobases in the central positions.[71]

Typically, individual KH domains recognise four RNA bases with low-intermediate affinity (10-100 μM).[73] However, structural extensions to the canonical KH domain can increase the hydrophobic interaction surface. This extended surface enables the domain to recognise an extended RNA sequence. The Signal Transduction and Activation of RNA (STAR) fold is currently the best studied example of an expanded KH domain. Proteins that have been shown to contain a STAR domain include, Splicing factor 1 (SF1),[74] Quaking,[75] Sam68, and T-STAR.[76,77]

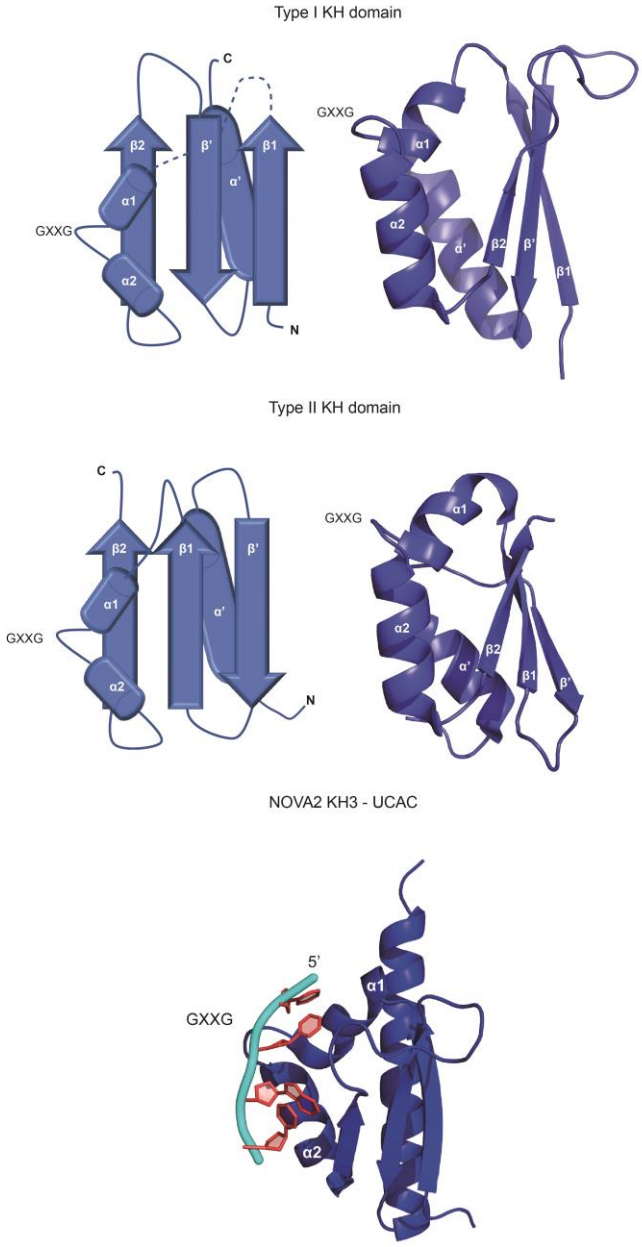


Figure 1.5: Canonical KH domain type I and II fold
Upper: Schematic representation of a Type I KH domain (Left) and the structure of the third KH domain of NOVA1 (Right) (PDB:1DT4). Middle: schematic representation of a Type II KH domain (Left) and the second KH domain of NusA (PDB:2ASB) (Right). Figure labels the sequence of α-helices and β-strands in the folds and highlights the location of the RNA binding, and conserved GxxG motif. Lower: Structure of NOVA2 KH3 in complex with UCAC RNA (PDB:1EC6). RNA bases are coloured in Red and phosphate backbone in Cyan. The structure displays canonical KH domain RNA binding with the RNA bases binding in the hydrophobic groove and the GXXG loop.

The STAR domain contains a central KH domain flanked by two regions, an N-terminal motif (QUA1) and a C-terminal motif (QUA2). the NMR structure of SF1 KH-QUA2 showed the QUA2 to contain a long loop followed by an amphipathic helix which folds back to contact the α1 and α3 helices and the GxxG loop of the KH domain. This leads to an extension of the hydrophobic surface of the KH groove to allow recognition of an additional 3 nucleobases.[78] This enlarged RNA interaction surface involving the QUA2 region was also confirmed by the X-ray structure of the Quaking STAR domain in complex with RNA.[79] Interestingly, NMR studies showed that the KH domains of Sam68 and T-STAR were able to bind RNA in the absence of the QUA2 extension.[80] Further investigation of these domains identified a novel mechanism by which the Sam68 and T-STAR STAR domains recognise RNA. This study showed that the QUA2 extension of these domains was not involved in RNA binding or dimerisation. Rather, the dimerisation of these domains was mediated by the QUA1 extension via a unique mechanism.[76]

As observed for other RNA binding domains, combinatorial binding of multiple KH domains within the same protein is required to establish high affinity and high specificity interactions with RNA targets.[49,81] KH domain combinatorial binding has been evaluated using KH domain deletions, and more recently conservative double mutations in the GxxG loop that eliminate RNA binding of individual domains.[82,83] Due to the presence of multi KH domains within the same protein, there are instances where tandem KH domains establish interdomain contacts resulting in the coupling of RNA recognition.

KH domains separated by flexible linkers can display a low degree of domain coupling allowing the protein to adapt the interdomain arrangement to recognise different RNA targets. Conversely, stable interdomain association of neighbouring KH domains can fix the orientation of the KH domains RNA binding surface so that either an extended RNA binding surface is formed, or an induce rearrangement of the RNA topology is required for binding. Such examples of KH domain coupling include: the KH1 and KH2 domain of NOVA1,[84] the KH3 KH4

domains of ZBP1,[85] the KH2 and KH3 domains of KSRP,[86] and the KH1 KH2 domains of NusA.[87] Examples of how inter-KH domain contacts can orientate the RNA binding surfaces is shown in Figure 1.6.

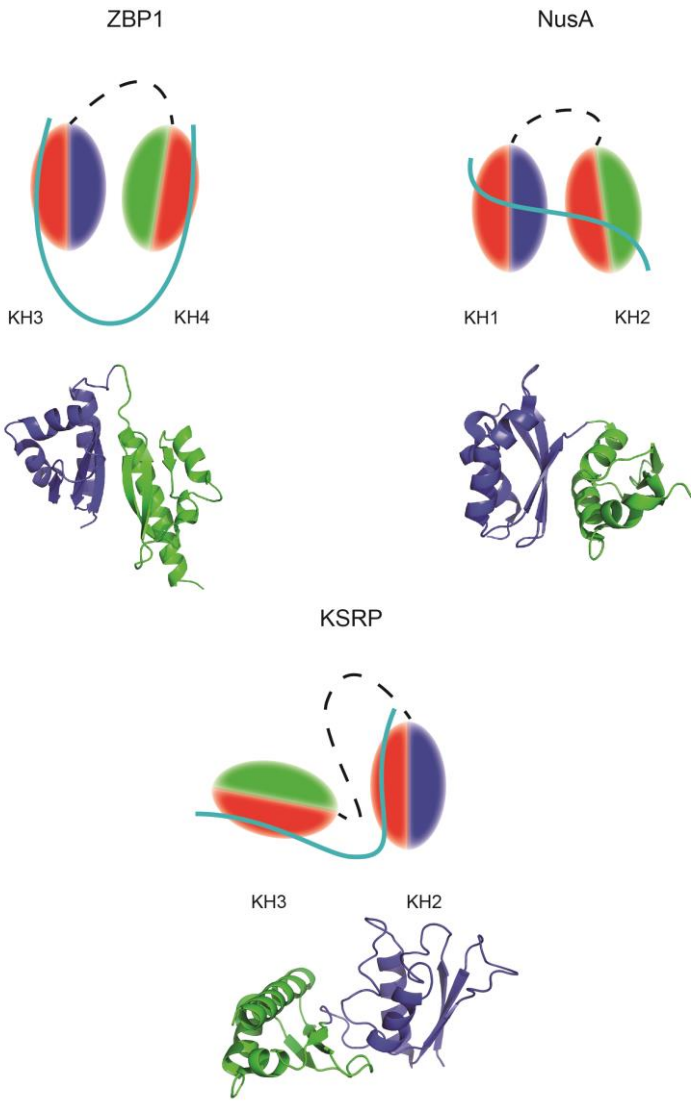


Figure 1.6: Organisation of RNA binding surfaces in tandem KH domains and the effects on RNA topology upon binding
Type I KH domains: ZBP1 (PDB: 2N8M) KH3 (Blue) KH4 (Green), KSRP (PDB: 2JVZ) KH3 (Blue) KH3 (Green). Type II KH domain NusA (PDB: 2ASB) KH1 (Blue) KH2 (Green). Black dotted line represents the orientation of the linker. Red surface in schematic represents the RNA binding surface determined by the structural positioning of the GxxG loop, with a Cyan line depicting RNA upon binding. The position of the RNA binding surface in ZBP1 induces RNA looping upon binding, whereas KSRP induces a ~90º bend in the RNA. The KH di-domains NusA is orientated in a manner that enables the domains to bind a 11 nt continuous stretch of RNA.

1.3.3 The role of RBD linkers in RNA recognition

As many RBP-RNA interactions consist of a surface comprising of multiple RNA binding domain interactions, the protein linker between the different domains can also influence binding (Figure 1.7). Domain linkers can either be directly involved in RNA binding or mediate binding by influencing the orientations of the RBDs. Binding domains that are connected by long protein linkers are usually uncoupled given the high chance that the large linker is unstructured and flexible. Uncoupled domains can recognise separate RNA motifs which are many nucleotides apart within a RNA transcript, or recognise RNA motifs located on different transcripts (Figure 1.7A & B). This kind of binding is observed for the two dsRBDs of ADAR2 which are separated by a 84 amino acid linker,[88] and the two RRM domains of hnRNP A1.[89] Structurally independent RBDs can display a large variability in the distance between the preferred binding motifs. This variability further increases the diversity of RNA targets such proteins can recognise.

Shorter RBD linkers can also be flexible. However, the shorter linker length restricts the distance between potential RNA binding motifs and can greater influence the orientation of the adjoining RNA binding domains. For example, shorter linkers can become structured upon RNA binding, typically in the form of a α-helix structure (Figure 1.7C). This restructuring of the linker can influence orientation of the connected domains so that an extended RNA binding surface is formed. Such conformational transitions are observed for proteins including nuclear polyadenylated RNA-binding protein (Hrp1)[90] and Sex-lethal.[91] The two RBDs can also contain a protein interaction surface where interdomain contacts are formed generating a system in which the two domains function more as a fused di-domain (Figure 1.7D). This can result in the RNA interacting surface being extended and more ridged than with domains that are totally uncoupled. When RBDs have a fixed orientation due to interdomain contacts, or short protein linkers they can influence the topology of the RNA upon binding (Figure 1.7E).

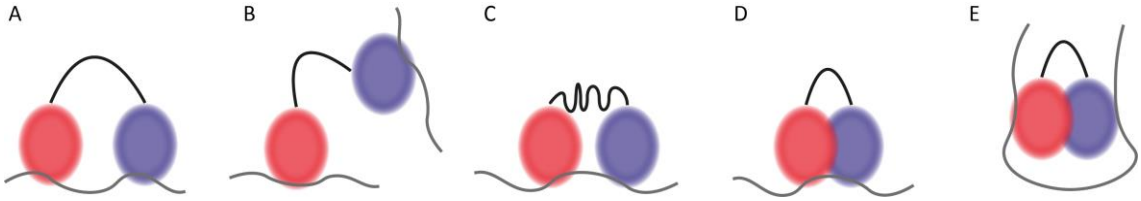


Figure 1.7: Modes of RNA recognition by tandem RNA binding domains
Schematic representation of different orientation of RNA binding surfaces, influenced by interdomain linker.

The relationship between the multiple RBDs influences the affinity towards the RNA target. Therefore, the length and structure of the amino acids in interdomain linkers are highly influential in overall binding affinity. In situations where the domain linker is short and structured, interdomain interactions are typically observed. Often in these examples the two domains come together to create one RNA binding surface. If the linker is larger and fully flexible the two RNA target sequences are independent of each other and the two domains function as individual units. Due to the importance of linker length in defining the RNA binding characteristics of a protein it is unsurprising that in some cases the length of the linker is a conserved feature rather than the specific amino acid sequence. In turn, when investigating the RNA binding of RBPs it is important to study systems which contain the multiple RBDs connected by their protein linkers.

1.4 Study of protein-RNA interactions

Studies on individual RNA binding domains *in vitro* with and without their RNA targets has revealed valuable insight into how RBP recognise RNA transcripts. Similarities across the different types of RBDs brings consensus to the field and enables better understanding of other RBPs containing the same canonical RBDs but that have so far been less well characterised. However, as noted above there are exceptions to the classical mechanisms of binding, resulting from modifications in the typical domain fold or amino acid compositions, or via extensions of the domain or coupling with auxiliary domains. In addition, RBPs recognise transcripts by utilising several of their multiple RBDs. Therefore,

it is critical to study RNA binding in the context of either the full-length RBP or a selection of their RBDs.

There are many experimental techniques that are used to study protein-RNA interactions. The system that is being investigated, and the question that is being asked, determine which method is best for a particular problem. As highlighted above, the RNA sequences individual RNA binding domains recognise is critical for RNA target recognition. However, overall RNA binding is the result of multiple RBD interactions of the same RNA binding protein. To better understand how RBPs select RNA targets we must understand how sequence specificity and combinatorial cooperativity work. A better understanding of these interactions will enable us to develop better therapeutics and diagnostics for diseases that result from RNA binding protein misregulation.

1.4.1 Identifying RNA target sequences of RBPs

The nucleobase specificity of RNA binding domains is critical for RNA recognition. Studies mutating RNA recognition sequences of RBDs have been shown to perturb RNA binding. However, *in vivo* RNA binding proteins do not always associate with the highest affinity RNA binding sequence. In addition, interactions with RNA transcripts are often an accumulation of multiple RBD interactions within the same protein. Therefore, lower affinity sites in combination with higher affinity targets are critical for regulation of RNA metabolism within the cell. This enables RNA binding proteins to associate with a variety of different RNA transcripts. For example, KSRP regulates two alternate RNA metabolic pathways by recognising different RNA target sequences.[92,93]

To begin to understand how RBPs exploit different domains to recognise numerous targets it is necessary to elucidate the full sequence preference of each individual RBD. This can be challenging as individual domains often recognise their short RNA target sequences with low to moderate affinity. Methods therefore need to be sensitive enough to monitor these weaker

interactions. Biophysical techniques such as isothermal titration calorimetry (ITC), electrophoretic mobility shift assay (EMSA) and circular dichroism spectroscopy (CD) can be used to determine the binding affinity of RBPs with different RNA targets, and in turn identify RNA recognition sequences. However, these techniques are typically suited for the study of high affinity interactions. Individual RBDs more often bind RNA with affinity in the micromolar range. NMR can measure lower affinity interactions due to the protein concentrations required for NMR studies. One such NMR technique that has been developed to determine the full sequence specificity of RNA binding domains is scaffold-independent analysis (SIA) and will be explained in Chapter 2.5.[94,95]

1.4.2 Development of high-throughput methods to study RBP-RNA interactions *in vivo*

High-throughput methods focused on understanding RBP-RNA interactions can be split into two categories. The *in vitro* studies investigating RBP-RNA interactions are typically performed independently from other interacting proteins or cellular factors. Where as *in vivo* approaches study RBP-RNA interactions within cells by taking a snapshot of RBPs interacting with available RNAs at a given moment in time. Naturally, given the number of increased influencing variables in such *in vivo* methods, analysis of data sets is challenging, particularly when identifying biologically reverent binding sites from background noise. However, it is vital we develop methods to analyse such data sets in order to fully understand RBPs control of post transcriptional gene regulation.

Three *in vitro* studies used for identifying RNA targets are: Systematic evolution of ligands by exponential enrichment (SELEX), RNAcompete, and RNA Bind-n-Seq. SELEX uses *in vitro* RNA selection to determine RNA binding motifs for RBPs.[96,97] The technique uses a pool of randomised RNA oligos that are incubated with the RBP of interest. After the initial incubation RNA oligos that bound to the RBP are reverse transcribed, PCR amplified and then transcribed

back into RNA. The enriched pool of RNAs are then incubated with the RBP again and the process repeated three to four more times. These multiple rounds of enrichment result in the identification of high-affinity RNA targets. SELEX is a valuable technique for identifying high-affinity RNA targets of RBPs. However, RBPs typically recognise RNA targets with varying affinity and enrichment of only the highest affinity targets may restrict the identification of biologically functional targets with lower affinity.[98]

A method coupling SELEX to high-throughput sequencing, known as SEQRS allows the user to sequence the RNA oligos enriched after each round of RBP incubation. This modification of SELEX enables the identification of suboptimal RNA targets and monitors at which round of enrichment these are lost. In turn, optimal and alternative binding sites for the target RBP can be identified.[99]

RNAcompete requires RBPs to be tagged via an incorporated GST motif. RNA oligos of ~40nt in length are then incubated with the GST-RBP. The RNA oligos are added in vast excess compared to the GST-RBP. This difference in abundance results in the RNA oligos competing for binding with the RBP. RNAcompete requires a single RNA selection round followed by a washing step to remove unbound RNA oligos. Bound targets are then eluted and detected via a microarray analysis. The relative abundance of RNA oligos detected is used to assess relative affinity for that RBP to that RNA sequence.[100]

Bind-n-Seq is based on a similar approach to RNAcompete however, RNA oligos are incubated with varying protein concentrations. Bound RNA is then detected via high-throughput sequencing.[101]

In the past decade there has been great development of techniques designed at studying protein-RNA interactions *in vivo*. This has resulted in a great expansion of our understanding of RNA biology and has led to the discovery of both novel RNAs and RBPs. These methods can be split into two general categories: ‘protein-centric’ and ‘RNA-centric’ methods. Protein-centric methods require

knowledge of the protein or class of protein being investigated. They generally involve purifying proteins form cell lysates followed by sequencing of the associated RNAs. These sequences are then mapped to the transcriptome to identify binding sites. RNA-centric approaches use RNA, or classes of RNA to select protein-RNA complexes. The associated proteins can then be identified via mass spectrometry (MS) analysis.

RNA-centric methods use RNA molecules as 'bait' to selectively purify protein complexes that can recognise the specific RNA molecule being used in the investigation. One example of such methods exploits certain proteins that recognise specific RNAs. For example, the bacteriophage MS2 viral coat protein recognises RNA stem loop structures.[102] RNA transcripts are produced that contain the RNA sequence of interest coupled to several MS2-binding RNA stem loop structures. The MS2 protein is then used to immobilise the RNA.[103] Cell lysate can then be passed over the immobilised RNA and interacting protein complexes captured. This is just one example of several techniques that use RNA to capture interaction proteins.

More recently a study pulled down all polyadenylated RNA transcripts in HeLa and HEK293 cells to generate a global RBP interactome. Mass spectrometry analysis of the purified protein complexes identified many new RNA binding proteins. One interesting outcome of this study was the discovery of several proteins that when investigated do not contain any of the known RNA binding domains. This potentially opens investigation to new mechanisms by with proteins recognise RNA transcripts.[104,105]

To study RNAs associated with specific proteins within cells requires immunoprecipitation of these complexes. Specific antibodies are used to selectively pull down the protein associated complexes. In turn, it is vital that antibodies are specific to the protein being investigated as non-specific interactions have the potential to introduce 'contaminating' RNA molecules in the later stages of the protocols. After immunoprecipitation of the complexes the

associated RNAs are reverse transcribed before PCR amplification and high throughput sequencing or microarray analysis. Bioinformatic analysis of high throughput reads is then used to map reads back to their transcripts of origin to identify protein binding sites. The bioinformatic methods available to analyse these data sets are still developing. Deconvoluting the data obtained from these high throughput screens is particularly challenging and remains a limitation of these methods

The first genomic-wide analysis performed were based on RBP-RNA complexes being immunopurified from cell extract followed by a microarray detection assay. Termed, RNA immunoprecipitation followed by microarray analysis (RIP-chip), or the modified RIP-seq, where the purified RNAs are sequenced via high-throughput sequencing. These studies were applied to dozens of RBPs across several species resulting in the formation of the first database (RBPDB) characterising RNA-binding specificities. It became apparent that one RBP binds to many mRNAs within cells and that one mRNA molecule is regulated by many RBPs, coining the term many-to-many. These RIP studies were the start of unravelling the more complex nature of RBP interactions with cellular RNAs.[106,107]

Complexes can be immunopurified from cells under both native (as in the above RIP studies) and denatured condition. Native purification RIP has advantages over denatured purifications as they preserve the native complexes present in the cell. However, they also suffer from the less rigorous washing procedures. It has been shown that RNAs purified in this way often generally correlate with the abundance of the RNA, in addition to the high presence of contaminating ribosomal RNAs. The consequence being that specific interactions that occur with low abundance transcripts are often masked by non-specific interactions that occur with highly abundant transcripts.[98,107,108]

To overcome these limitations of native purification techniques denaturing methods were developed. The methods utilise the photoreactive nature of nucleotides. Exposing cells or tissues expressing the protein of interest with UV

light results in the crosslinking of protein-RNA complexes present at that moment in time. This crosslinking forms a strong covalent bond between the RNA and proteins and thus forming stable complexes that can undergo stringent washing procedures. These types of studies have been termed crosslink and immunoprecipitation (CLIP) methods.[108] Their major advantage is the ability to distinguish *in vivo* interactions that are crosslinked in the cell from interactions that form subsequently in solution. Since the development of the first CLIP studies there has been several modifications of the original method developed. The different modifications of CLIP solve issues with the original technique and some are better suited for the study of different systems.

UV mediated covalent crosslinking provided a method to overcome the limitations of the original RIP studies.[24,98] Irradiating cells with UV light exploits the photoactive properties of RNA nucleotides and results in the formation of a covalent bond between protein and RNA molecules that are a few Ångstroms apart.[109] The formation of this strong covalent bond enables the user to implement extensive and stringent washing protocols to remove background RNA from immunoprecipitated RBP-RNA complexes. This incorporation of UV crosslinking before immunoprecipitation generated the field of UV induced crosslinking immunoprecipitation (CLIP) studies. CLIP enabled the identification of the positions of protein-RNA interactions with higher resolution and specificity. Since the original CLIP studies, the technique has undergone several modifications and refinements. High-throughput sequencing of RNA isolated by CLIP (HITS-CLIP), photoactivable ribonucleoside-enhanced CLIP (PAR-CLIP),[110] and individual nucleotide resolution CLIP (iCLIP)[111] are the three main variants. Each of these variant uses UV induced crosslinking to identify RNA binding sites. However, the methods for defining the crosslink site differ between protocols.

1.4.3 UV induced crosslinking immunoprecipitation assays

The coupling of CLIP to high-throughput sequencing (HITS-CLIP) was the first step towards studying RNA binding protein interactions on a genome-wide level. Since the discovery of HITS-CLIP there have been several modifications of the technique to help improve sensitivity and resolution of the protein-RNA target sites. The two major modifications of the technique are PAR-CLIP and iCLIP (Figure 1.8).

Traditional CLIP and its recent modification iCLIP use UV-C (254 nm) light to induce formation of covalent crosslinks at the site of protein-RNA contact. PAR-CLIP relies on the use of photoactivatable nucleotides such as 4-thiouridine (4SU) or 6-thioguanosine (6SG). These modified nucleotides must be taken up by the cells the PAR-CLIP is being performed on and incorporated into the transcriptome of the cell. The photoactive nucleotides form crosslinks at the lower energy wavelength of 365 nm in the UV-A range (Figure 1.8).[110]

Protein-RNA complexes are then immunopurified from cell lysate before the RNA and protein are partially digested. Subsequently, adaptors are ligated to the digested RNA fragments. HITS-CLIP and PAR-CLIP require ligation of 3’ and 5’ adaptor sequences to the purified RNA fragments, whereas iCLIP requires only ligation of a 3’ adaptor (Figure 1.8). Reverse transcriptase then reverse transcribes RNA fragments into cDNA molecules. However, digestion of the RBP bound to the RNA is never 100% complete and results in a poly-peptide remaining bound to the RNA transcript. This remaining protein fragment can cause the reverse transcription reaction to stall.[112] HITS-CLIP and PAR-CLIP require the reverse transcription reaction to fully extend from the 3’ adaptor to the 5’ adaptor. RNA transcripts where the reverse transcription stalls before reaching the 5’ adaptor are lost in the later PCR amplification step[113] (Figure 1.8). In HITS-CLIP the poly-peptide bound to the RNA can result in a deletion of the base at this position.[109,114,115] In PAR-CLIP UV crosslinks only occurs with the photoactive nucleotide analogues. The base modifications of these photoactive nucleotides

can cause a base transition during reverse transcription (4SU: T-C and 6GS: G-A transition).[110] After high-throughput sequences of the cDNA transcripts the reads are mapped to either the genome or transcriptome. The point deletions (HITS-CLIP), or base transitions (PAR-CLIP), are used to identify the RNA binding motif as these modifications relate to the site of the protein-RNA crosslink (Figure 1.8).

It is important to highlight two limitations for detecting RNA binding sites in this way. Firstly, studies have shown that up to 80% of reverse transcription reactions stall at the site of the poly-peptide, and thus do not extend to the 5' adaptor sequence.[113] Secondly, the incorporation of deletions or base transitions at the site of the poly-peptide link are crucial in the later analysis steps. Not all reverse transcription across the poly-peptide result in such mutations as some events 'read-through' producing a non-mutated cDNA fragment.[115] These fragments can still be used for genomic mapping, giving indications to gene binding, but the nucleotide resolution of the binding event is lost.

The iCLIP protocol utilises the poly peptide in a different manner by taking advantage of the large number of reverse transcription events that stall at this site. Modified 3' adaptors are ligated onto the RNA fragments, an intramolecular circularisation step followed by a BamH I digestion is introduced. These modifications result in relinearised transcripts to place a 5' adaptor equipped with a unique barcode immediately upstream of the crosslink site (Figure 1.8). This positioning of the adaptor enables the user to identify crosslink sites in the later analysis steps.[111,116]

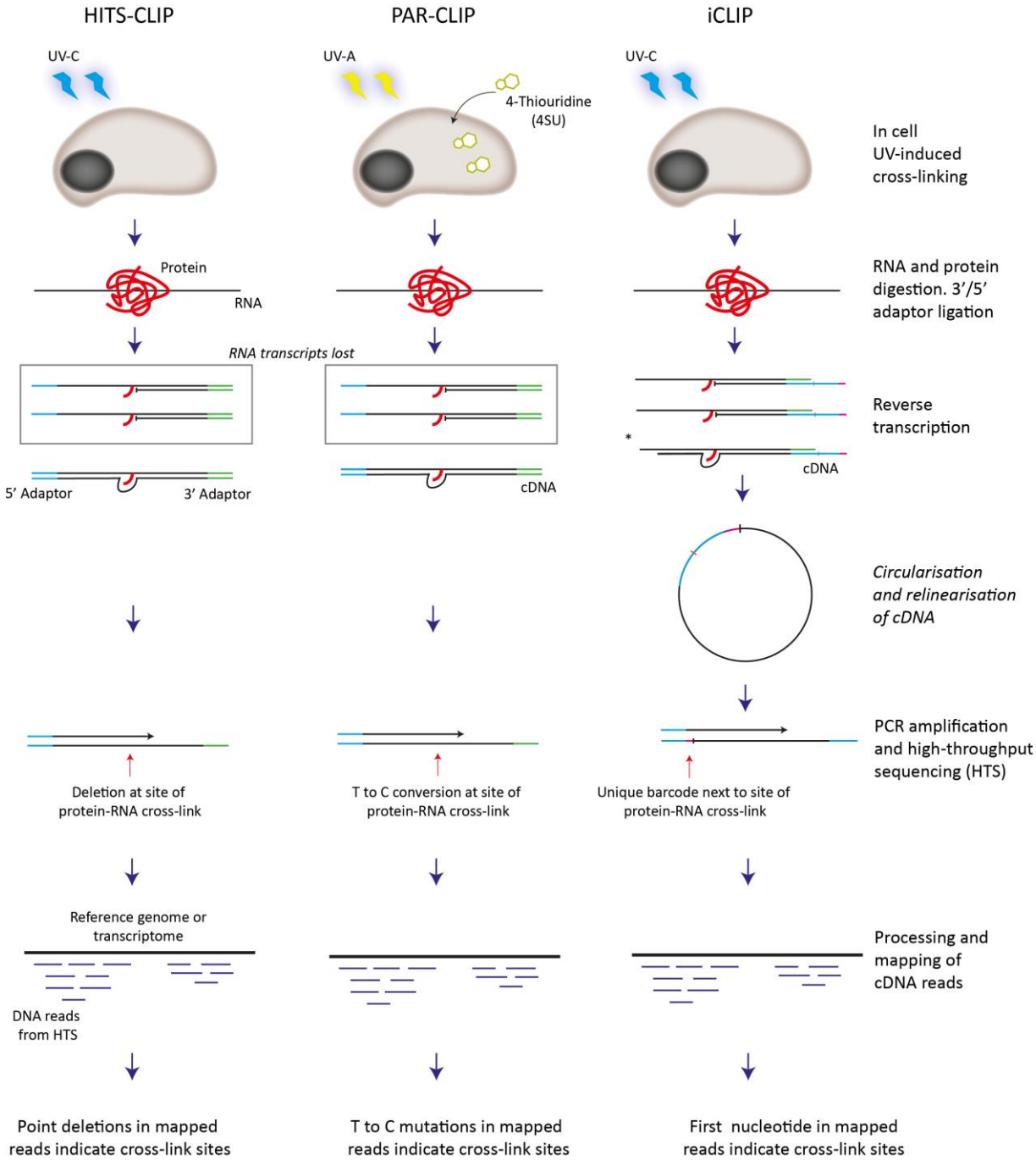


Figure 1.8: Overview of HITS-CLIP, PAR-CLIP and iCLIP protocols
Schematic represents the main steps in the three CLIP protocols. HITS-CLIP and iCLIP use UV-C light to induce protein-RNA crosslinking, while PAR-CLIP requires cell to be pre-incubated with photoactive ribonucleosides to mediate crosslinking in the UV-A spectrum. Grey boxes at reverse transcription stage represent RNA transcripts that are lost in the HITS-CLIP and PAR-CLIP protocols due to reverse transcription stalling. * depicts read-through events in iCLIP protocol that prevent the correct positioning of the unique barcode region to the protein crosslink site. HITS-CLIP and PAR-CLIP identify protein-RNA crosslink events via mutations in mapped cDNA reads that result from reverse transcription

across the covalently bound poly-peptide that remains after proteinase K digestion. iCLIP implements the use of a unique barcode region in the 3' adaptor and a cDNA circularisation step to identify protein crosslink sites.

In most cases a clear correlation between these *in vivo* data and *in vitro* data on the sequence specificity of a protein has not been established. This is not surprising as attempting to determine one specific RNA sequence present in the many RNA targets of a protein is difficult to reconcile with the sequence variability found within the variety of targets and the moderate sequence specificity of many of the RBDs, thus highlighting our still rudimentary understanding of the role of sequence specificity in the cell.

One area of improvement would be the study of weaker protein-RNA interactions. Many of the high-resolution structural RBD-RNA complexes have been obtained with high affinity RNA targets as these interactions are easier to study by both NMR and X-ray crystallography. However, RBP-RNA interactions in cells are extremely dynamic and consist of complexes with a range of affinity, with many physiological interactions being more transient than others. Gaining better understanding of the structural and biochemical nature of these weaker interactions *in vitro* would aid in the analysis of high throughput screens which contain information on both the high and weaker affinity protein-RNA interactions. Another process that needs to be better understood is the combinatorial mechanism by which full-length RNA binding proteins recognise targets. This is a complex process and deconvolution of data collected from in-cell RNA binding with full-length RBPs requires more understanding on RBD interactions in isolation and how binding of one RBD might affect the recognition of a partner RBD. Therefore, we must piece together the understanding we have on individual RDB-RNA interaction into the bigger picture of the full-length protein where RNA target recognition is a combination of several RBD interaction. One approach that has been used to determine individual RBD contribution of RNA recognition is via mutagenesis.

1.4.4 Mutations to investigate RBP-RNA recognition

Modelling the regulatory networks controlled by RBPs requires a molecular understanding of the underlying modular protein-RNA interactions. The biophysical data collected from RBD-RNA interactions needs to be better correlated with the role these individual interactions play in overall protein-RNA recognition in cellular environments[49]. Two common strategies implemented to gain understanding into the contribution individual domains play in RNA target selection are to either delete the RBD *in toto,* or to introduce point mutations to perturb or modify RNA binding. This second approach requires an understanding of the domains structure and RNA-binding properties. Such information can be obtained from crystal structures, NMR studies, or sequence alignments with well-studied RBD examples. However, both strategies have limitations. The deletion of an entire RBD can result in destabilisation of the protein, particularly the neighbouring domains that may form contacts with the domain being deleted. Furthermore, deletion of the domain to remove its RNA binding properties will also remove any additional functions the domain may play in target selection, for example through interactions with additional proteins or via posttranslational modifications that are placed within the domain. A more 'subtle' method is to introduce mutations in the domain to remove RNA binding. Mutation of just a single amino acid may have drastic effects on protein folding and stability. In turn, understanding the contribution the amino acid plays in protein folding removes the potential of mutating out amino acids that make extensive contacts with interdomain residues. Residues involved in RNA recognition are often on the surface of the protein and are solvent accessible, reducing the likelihood of structural disruption upon mutation. However, the folding and stability of mutant proteins should be studied *in vitro* prior to functional testing.

There are several successful examples where mutations have been able to modify RNA recognition of RBDs without altering the domains fold and stability. One example is particularly useful as it can almost be universally applied to KH domains. As noted above, KH domains that can bind RNA require the presence of a GXXG motif in the flexible loop between the α-1 and α-2 helix. The motif acts

to orientate the RNA bases into the hydrophobic groove of the domain. Incorporation of a double negative charge in this motif abolishes the electrostatic interactions the motif makes with the phosphate backbone of the RNA and in turn removes RNA binding. Importantly it has been shown that these mutations do not significantly alter the fold or stability of the domain of several examples such as the KH1, KH2, KH3, and KH4 domains of KSRP and the KH3 and KH4 domains of the ZBP1 protein[82]. These mutations were implemented in functional studies investigating the in-cell recognition of KSRP and its characterised TNFα and β-catenin RNA targets[117]. The study was able to determine that KH1 plays a minor role in recognition of these targets whereas KH2 showed modest recognition and the KH3 and KH4 domains were essential for binding to these transcripts. The study went further to investigate how mutations of the individual domains affect KSRP mediated decay of the β-catenin transcript in KSRP depleted HEK-293 cells. They showed that mutant KSRP with impaired KH1 or KH2 binding was able to mediate decay of the β-catenin transcript, yet loss of KH3 or KH4 binding inhibited this decay process.[82]

Such investigations have also been used to understand RRM binding in multi domain systems. The RNP motifs within RRM domains are the main site of canonical RNA recognition. Loss of aromatic or positively charged side chains in these conserved motifs have been shown to abolish RNA binding.[53] For example, a study was able to abolish the RNA binding of the RRM domain of the RBM38 protein by mutating evolutionary conserved residues involved in canonical RNA binding.[118] The group incorporated the mutations Y77AK103E, residues that reside in the RNP2 and RNP1 motif respectively. NMR analysis showed this mutant protein was folded and non-aggregated. Using this mutation, the group was able to link the RNA binding properties of the RBM38 RRM domain to the proteins role in inhibiting miRNA-150 mediated RNA decay in U2OS cells.[118] Additionally, the *Saccharomyces cerevisiae* protein Cleavage Factor 1A (CFIA), which is involved in 3'-end RNA processing, contains the subunit Rna15 which mediates RNA recognition through a single RRM domain.[119] The group used the crystal structure of Rna15 in complex with its RNA target to identify residues that

mediated binding to RNA bases. Through mutational studies they identified that mutations Y27A and R87K both abolished RNA binding.[120]

Mutating residues within RBDs can also be used to modify RNA interactions rather than totally abolishing binding. For example, mutations could be incorporated to increase the binding affinity of a domain so that the contribution towards target selection is amplified in the full-length protein, or where detailed domain-RNA complex structures are available the specific interacting amino acid side chains make with RNA bases moieties via hydrogen-bonding or electrostatic interactions can be modified by altering the nature of the amino acid side chains. This can alter the network of interactions and may be able to shift the specificity of the domain to favour alternate RNA bases. However, these kinds of mutational approaches are more challenging and require in depth structural and biophysical characterisation.

1.4.5 Studying protein-RNA interactions at high resolution

Much of the molecular understanding of protein-RNA interactions is derived from high resolution protein-RNA complex structures. The majority of complexes have been solved using x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy and a few larger complexes via cryo-electron microscopy (cryo-EM). These techniques can be used to acquire high resolution structural data of protein-RNA complexes with binding affinities ranging from nanomolar to micromolar. This range covers both the micromolar affinities often populated by individual RBDs and the nanomolar affinities which can be reached by multiple RBDs participating in combinatorial binding. From resolving several protein-RNA complexes with atomic resolution scientists have started to understand the molecular basis of specificity. As the number of structures increases, improved computational models can be developed to investigate RNA binding events in other systems. However, the nature of protein-RNA interactions often makes it difficult to obtain such structural information, and in fact the number

or RBP structures in isolation deposited into the PBD far outnumbers those of protein-RNA complexes.[121]

To solve structures via X-ray crystallography the protein-RNA complex must be in a crystalline form. The wavelength of X-rays (0.01 -10 nm) provide the correct resolution required (~0.1 nm) to obtain atomic resolution of the structure. In order to resolve the structure, crystals must contain a regular array of molecules that diffract x-rays in a regular and predictable pattern.[122] This can be challenging due to the conformational flexibility of protein-RNA complexes resulting in non-homogeneous samples that are hard to crystallise. This limitation is amplified when trying to crystallise longer RNA molecules with multiple RBDs. In addition, complexes that do crystallise usually fix the complex into the most stable conformation and so information is obtained only in this orientation.[123]

In contrast, protein-RNA complexes are not fixed into one conformation and so solution NMR can be used to report on the different conformations that occur during protein-RNA binding. This enables the user to gain information on the dynamics of protein-RNA interactions. The flexibility and movement of RBP-RNA interactions can be monitored through relaxation studies which report on the movement of the backbone and side chains of the protein. Comparing the data of the free and bound system, these experiments provide insight into structural rearrangements upon RNA recognition. These kinds of structural dynamics are important in RNA recognition as RBP loops and flexible regions in the core and/or flanking termini regions can rearrange upon RNA binding. For example, the mammalian RBP Quaking (QKI) contains a STAR domain which consists of a KH domain flanked by two conserved Qua1 and Qua2 domains. The KH-Qua2 was known to contain the RNA binding surface but the role of Qua1 was not understood. Studies revealed that upon RNA binding the Qua1 orientates Qua2 into the correct conformation in relation to the KH domain to form a high affinity RNA binding surface.[79] In addition, such studies can be used to study the flexible linkers between domains and how they influence RNA binding. In such cases the linker can directly partake in binding thus elongating the RNA binding surface, or

indirectly by mediating the corporation of adjacent domains cooperatively binding RNA targets.

1.4.6 Nuclear Magnet Resonance (NMR) Spectroscopy

The nuclei of atoms have an intrinsic property known as spin. Quantum mechanics state that the spin quantum number (I) of a nucleus may be zero or a multiple of ½. Nuclear magnetic resonance (NMR) spectroscopy utilises nuclei where I=½, for example ^{1}H, ^{13}C, ^{15}N, ^{19}F and ^{31}P. The rules of space quantisation state a nucleus with spin number I can be found in one of 2I+1 orientations. These nuclei have a magnetic moment, when the nuclei are exposed to an external magnetic field the two allowed orientations have slightly different energies. The orientation parallel to the applied magnetic field is known as the lower energy orientation and the anti-parallel orientation to the field is known as the high energy orientation. At equilibrium the two orientations are populated according to the Boltzmann distribution. This distribution results in a slightly larger population of nuclei in the lower energy state compared to the higher energy state. This leads to a bulk magnetisation along the axis (z-axis) of the applied magnetic field. The populations at equilibrium can be perturbed by employing an electro-magnetic wave, usually radio frequency (RF) pulse. As the population returns to equilibrium, a current is induced in a receiver coil which is $\propto$ to the difference in energy of the two populations. If the sample is irradiated at a frequency that is equal to the difference between the two energy states, the spin of the nuclei is brought into phase. This leads to a bulk magnetization in the xy-plane processing around the z-axis. This signal can be recorded by detecting the varying field from this rotating magnetisation as the system returns to thermal equilibrium by a process called relaxation.

The frequency at which nuclei process is known as the Lamour frequency and is directly proportional to the strength of the magnetic field experienced by the nuclei. As well as the applied magnetic field, nuclei also experience different electromagnetic environments. Therefore, nuclei of the same element but in

different chemical microenvironment within the molecule will have different Lamour frequencies. This effect is called the chemical shift. In NMR spectroscopy the signal recorded will be a mixture of these many different frequencies and presents itself as a complicated waveform. A Fourier transform is employed to transform the NMR-signal from its time domain to the frequency domain to obtain the peaks we see in an NMR spectrum.

^{1}H-^{15}N correlational spectroscopy such as HSQC and HMQC are well suited for the study of protein-RNA interactions. In such spectra each peak reports on the correlation between a proton and a covalently bound nitrogen. In this region we observe the correlations from backbone amide groups (Figure 1.9). Therefore, HSQC/HMQC spectra are commonly referred to as a fingerprint spectrum of the protein where each peak reports a single amino acid residue.

The ^{1}H-^{15}N correlational reporter resonances are sensitive to changes in the microchemical environment. Upon RNA binding it is common to observe large chemical shift perturbations of amino acid residues that are located in the RNA binding surface. A common feature of protein-RNA interfaces is a high proportion of aromatic residues (Phe, Tyr, and Trp). Therefore, protein-RNA interfaces have a high-density of aromatic groups, from both the protein side chains and most prominently the RNA bases, which lead to large ring-current mediated chemical shift perturbations. Where amide cross-peaks have been assigned and the structure of the protein is available, these chemical shift perturbations can be mapped onto the surface of the protein to determine the RNA binding site. However, care must be taken as chemical shift perturbations can also result from structural rearrangement of the protein fold upon RNA binding. Although these changes report on the same binding event, they may or may not map to the RNA binding surface.

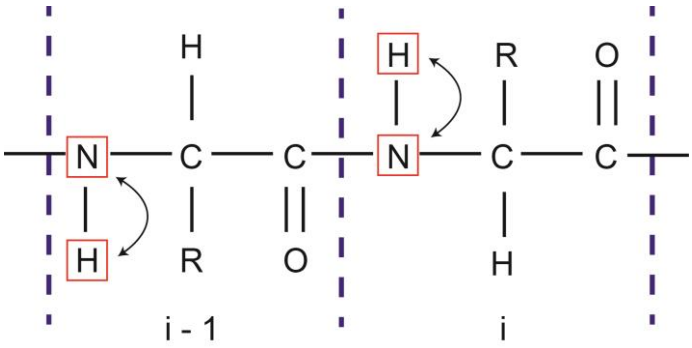


Figure 1.9: Schematic of the nuclei visible in a ^{1}H-^{15}N HSQC/HMQC spectrum
Figure depicts two backbone nuclei of adjacent amino acid residues (i , i-1). The NH pair is emphasised by red boxes

The use of selective isotopic labelling in NMR also allows the user to work in a system where signals are reported from only one protein. This can be useful when studying interactions involving more than one protein. This simplifies the system and reduces the complexity of the data being obtained. This is a useful tool as many RBPs are known to interact with RNA and other proteins in order to elicit their function. For example, the U2 small nuclear ribonucleoprotein B" (U2B") requires the presence of the U2 small nuclear ribonucleoprotein A' (U2A') to bind to its U2 snRNA target.[124] Another being the nuclear cap-binding protein (NCBP) complex. The RRM domain of the NCPB binds the 5' cap of the RNA only once it is in complex with the CBP80 protein which stabilised the N-terminal loop of the domain.[125] NMR can also be used to monitor binding events occurring at two separate interacting surfaces simultaneously. This is an advantage as it prevents the need for separate experiments which could lead to loss of important information of cooperativity.[67,126]

Whilst crystallography and NMR have so far provided many of the key insights into the structure of protein–RNA complexes, NMR is limited by the size of the complex that can be used to studied, and crystallography is hindered by the complexity of crystallising complexes with large unstructured RNA transcripts. This is a fundamental issue in the study of protein-RNA interactions as many

complexes occur within very large macromolecular assemblies such as the ribosome and the spliceosome. For complexes of this size cryo-EM has provided insight into the assembly of such macromolecular structures. However, cryo-EM is currently limited by the resolution that can be achieved, plus the resolution can be highly variable within a single structure.[127] Development of detectors and advancement of software used in cryo-EM are currently seeing a rapid rate of improvement. Once these are achieved the number of atomic resolution protein–RNA complexes solved by cryo-EM will surely increase.[128]

1.5 The IGF2 mRNA binding protein (IMP) family

1.5.1 Discovery of the IMP family

The IGF2 mRNA binding protein (IMP) family was first discovered in multiple experiments performed around the same time. These experiments were investigating three key RNA metabolic processes, RNA stability, localisation and translation.

The first study was investigating the stability of cytoplasmic RNA transcripts. The MYC RNA transcript was known to contain a sequence within its coding region that regulated the transcripts stability.[129] Using complementary RNA fragments to regions within the MYC gene and a cell-free RNA decay system they identified a 182 nt stretch that was responsible for stability and termed the region the MYC coding region stability determinant (CRD).[130] The group discovered that this region was bound by a *trans*-acting factor which controlled the RNAs stability. A gel-shift assay was performed with the CRD region incubated with extract from k562 cells. A migrating complex around 75 kDa was identified that was sensitive to protease digestion. They named the identified protein the CRD-binding protein (CRD-BP).[130]

Early studies exploring the mechanisms by which cells establish polarity identified ACTB mRNA to be actively localised to the leading edge of certain polarised cell

types, including chicken embryo fibroblasts[131] and neuronal cells.[132] It was determined that the 3' UTR region of the ACTB mRNA was necessary and sufficient for this observed localisation. Further analysis of the ACTB 3' UTR identified a 54 nt stretch that contained conserved sequence motifs between β-actin transcripts of other species.[133] This 54 nt stretch was defined as the 'zipcode'. Within the zipcode region was an AC rich sequence comprising of ACACCC. UV crosslinking experiments performed in CEF cells were performed to identify proteins bound to the zipcode RNA region. A 68 kDa protein was found to have the highest affinity for the zipcode region. This protein was termed zipcode binding protein 1 (ZBP1) and is the chicken orthologue of the human IMP1 protein.[134]

Lastly, studies on the translational control of the IGF2 mRNA, the namesake target of the IMP family, in early murine development provided further characterisation of the family.[135] Translation of specific isoforms of the IGF2 transcript were known to be spatially and temporally regulated during development. It was theorised that a *trans*-acting factor recognising the 5' UTR of the transcript may be responsible. Differentially expressed 5′ UTRs of the IGF2 transcripts were incubated with cytoplasmic extract from rhabdomyosarcoma cells and subjected to UV crosslinking. The formed complexes were RNase digested and separated vis SDS-PAGE. Again, a strong band ~69 kD was observed and was identified as being highly conserved with the previously named CDR-BP and ZBP1 proteins.[135] However, they also identified two additional proteins of similar mass within the protein complex band. These were later characterised and were shows to have high sequence conservation, these proteins were identified as the two mammalian isoforms of IMP1, IMP2 and IMP3.[135]

The IMP family has been identified as being highly conserved throughout the animal kingdom (Figure 1.10). There is strong sequence alignment consensus between IMP orthologues, for example the chicken orthologue ZBP1 which shares >94% sequence similarity with IMP1. Between the human IMP family

members, IMP1 and IMP3 share higher sequence similarity. This may in part explain an observed common functionality shared between the IMP1 and IMP3 paralogues. The diversity of organisms in which IMPs are expressed, and sequence similarity between IMP orthologues has so far provided a diverse range of systems in which IMP's function has been studied.[136]

The three isoforms in vertebrates are believed to arise from two gene duplication events that occurred during the evolution from invertebrate species.[137] Studies in invertebrate organisms have shown similarities in the function of the IMP proteins when compared to their documented role in humans. A *Xenopus* orthologue Vg1RBP was discovered to regulate the localisation of the Vg1 RNA towards the vegetal pole of developing oocytes.[138] The *Drosophila* IMP orthologue, dimp, shares a similar expression pattern in neuronal cells during embryonic development as observed for the mammalian IMP members.[139] Evaluation of dimp mutants during development demonstrated a requirement for the dimp protein in synaptogenesis, with *Drosophila* loss-of-function dimp mutations shown to be zygotic lethal.[140]

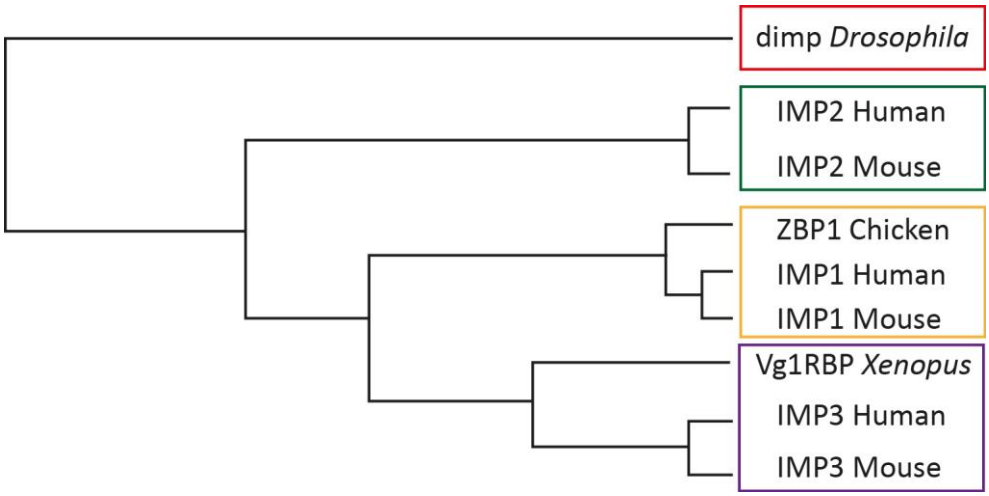


Figure 1.10: Conservation of IMP protein family throughout the animal kingdom
Representation of an evolutionary tree diagram showing divergence of the IMP family of proteins.

1.5.2 Functions of the IMP family

The human IMP family, comprising of three isoforms: IMP1, IMP2 and IMP3, are classified as oncofoetal RNA binding proteins. Their expression is highly regulated both spatially and temporally and are specifically expressed at certain stages of embryonic development.[141] They display multifunctional properties and play an important role during embryonic development regulating cell growth[142,143] and metabolism, cell adhesion and migration,[144,145] and neuronal differentiation.[146] This is likely a result of their ability to control the cytoplasmic fate of mRNAs which typically encode for cytoskeletal, adhesion, and metabolic proteins. In adult tissues IMP1 and IMP3 are expressed at negligible levels. However, *de novo* synthesis or extreme upregulation in expression levels is observed in several cancers and correlate with poor prognosis.[136,141,147]

The IMP family's role in mammalian embryogenesis was first determined by analysing the expression of the isoforms in mouse embryos and human foetal tissues. A series of immunostaining experiments on sections of developing mice embryos identified the IMP family to be expressed as early as embryonic day 3.5 in blastocyst cells. However, the highest level of IMP expression is observed between embryonic development days 12.5-15.5.[141] During this period IMPs were detected in the basal plasma cell membrane of developing epidermis cells, epithelia of the lung and intestine,[135] and in developing muscle tissue.[135,148] Analysis of mRNA collected from human foetal tissue also identified significant levels of IMP mRNAs in the human embryonic liver, lung, kidney, thymus, and placenta. Towards the end of embryogenesis IMP1 and IMP3 are no longer observed to be expressed.[149,150] On the contrary IMP2 expression levels remain moderate in certain tissues, likely due to its documented role in controlling cell metabolism.[151]

IMPs role in development was further cemented from studies using model organisms where IMPs were either mutated or knocked out during embryogenesis. IMP1 deficient mice suffered reduced survival and displayed a dwarfed phenotype and underdeveloped intestinal organs resulting from

hypoplasia. Organ size in the mutant mice was observed to be 14% reduced on average at embryonic day 17.5, increasing to a 45% reduction 1 week after birth. This decrease is likely to result from reduced cell proliferation due to loss of IMP controlled IGF2 translation and MYC stability.[149] IMP1 deficient mice also contained necrotic patches within the intestines, which are likely attributed to intestinal dysfunction. Analysis of mRNA levels during development of these mutant mice reported little change in RNA levels around embryonic day 12.5, the same time that IMP1 expression levels peak.[135,149] They also detected aberrant regulation of mRNAs encoding for extra cellular matrix (ECM) proteins in postnatal intestine, liver and kidney tissues. Similar knock out studies in Drosophila proved lethal. The IMP family also play a fundamental role in neuronal differentiation during early embryogenesis. This function is conserved in the developing nervous systems of zebrafish[152], *Xenopus*[150] and *Drosophila.*[140]

IMP-RNA recognition is key to their function. The three IMP family members contain six common RNA binding domains which orchestrate RNA recognition in a sequence specific manner. These RBDs are split into two N-terminal RNA recognition motifs and four C-terminal K Homology domains arranged as three pairs of di-domains (Figure 1.11). The IMP proteins also contain two nuclear export signals (NES) which aid in their export to the cytoplasm where RNA recognition is reported to take place in the perinuclear region. However, the proteins are reported to be able to shuttle in and out of the nucleus, yet no nuclear localisation signal is found within the proteins. Therefore, it is believed that association with RNA transcripts or auxiliary proteins facilitates re-entry to the nucleus. Cellular localisation of the IMP proteins has been reported to depend on the interaction with RNA via the KH domains. Disruption of IMP RNA binding, or overexpression in transient expression systems is reported to result in mis-localisation of the IMP proteins back to the nucleus, or into the incorrect polyribosomal fraction, yet these findings remain debated.[83,141,153]

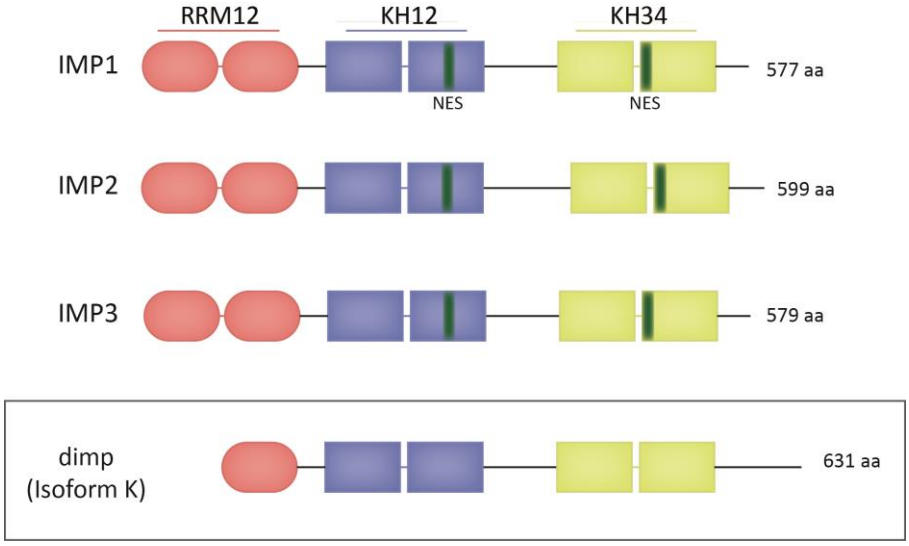


Figure 1.11: IMP family domain organisation
Schematic showing domain organisation of IMP proteins. Upper: Human IMP protein family members, IMP1, IMP2 and IMP3. RNA-binding domains comprising of two N-terminal RNA recognition motifs (Pink ovals) and four C-terminal hnRNP-K homology domains (Blue and Yellow boxes) arranged in di-domain pairs. Location of nuclear export signal within proteins amino acid sequence is represented in green. Lower box: Domain organisation of *Drosophila* orthologue dimp isoform K, highlighting the single N-terminal RRM domains and C-terminal extension.

Typical for most multidomain RNA binding proteins, the multiple RBDs of the IMP protein members enable the selection of different RNA targets by a combinatorial use of the different RBDs. For IMP1, current data points towards the KH domains being the domains that contribute most to RNA specificity,[83,154] with the RRMs playing a role in protein-protein associations.[155] IMPs regulate RNA metabolism transcripts via the formation of ribonucleoprotein (RNP) complexes. Studies to determine the composition of these complexes have given insight into the mechanisms by which IMPs recognise and increase the variety of RNA transcripts IMP proteins can interact with. They have also given insight into the regulation of such transcripts and remain an area of interest in the field.

Identification of exon-junction complex[156] and the nuclear capping protein CBP80[157] protein in IMP1 associated RNPs suggest that IMPs interact with 'virgin' mRNAs transcripts which have not yet undergone their first round of

translation.[156,157] In the cell IMPs were believed to bind to these 'virgin' RNA transcripts at the site of transcription, and association with RNA facilitated nucleocytoplasmic export.[18,158] However, recent findings suggest that IMP proteins associate with RNA transcripts at the perinuclear region.[159] In the cytoplasm the IMP-RNA complex specifically associates with additional proteins and RNAs to form RNP complexes.[141,156,157] Within these complexes IMPs are documented to control the association of the RNP with components of the translational apparatus or decay machinery.[156] Ultimately this enables IMPs to control transcript translation, and in turn, transcript abundance in the cell (Figure 1.12).

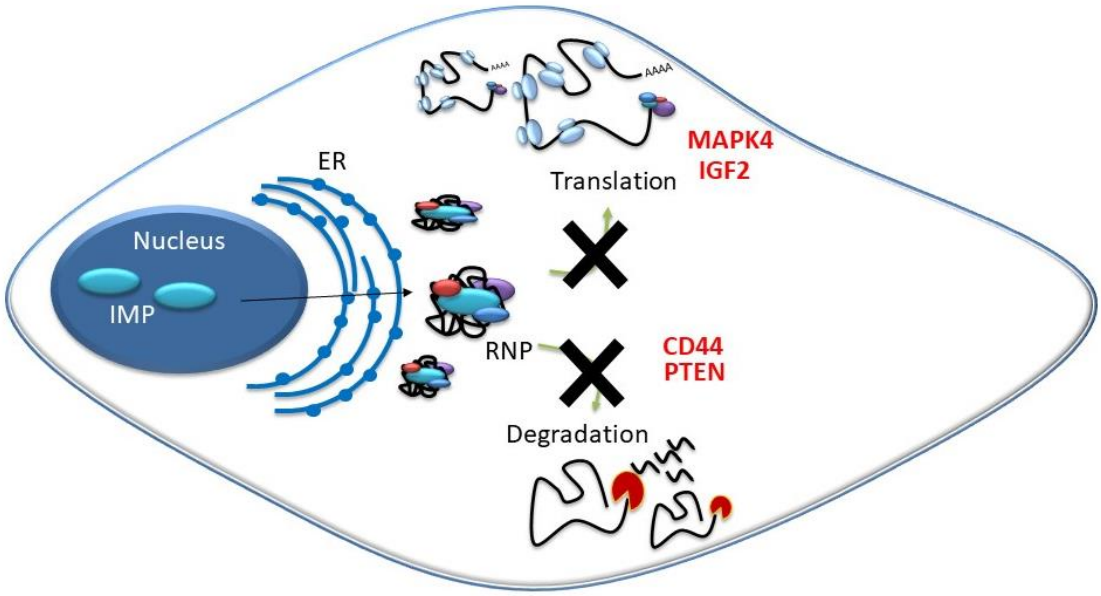


Figure 1.12: Cytoplasmic regulation of specific mRNAs mediated by IMP1
IMP1 (light blue oval) interacts with RNA (black) in the perinuclear region. In the cytoplasm IMP1 and additional RBPs associate specifically with target mRNA to form highly stable mRNA-protein complexes (RNP). IMP1 effectively 'cages' target mRNA to either, 1) Inhibit mRNA translation (MAPK4 and IGF2 mRNA) or 2) Prevent premature mRNA degradation (CD44 and PTEN mRNA)

1.5.3 IMP1 localisation of ACTB mRNA in polarised cells

Follow up investigations from the original finding that ZBP1 regulated ACTB mRNA identified the C-terminal portion of the protein as being required for recognition of the ACTB 3' UTR. However, localisation of the mRNA was lost in the absence of the N-terminal region of the protein.[160] Full-length ZBP1 was shown to bind the zipcode RNA element with a K_d in the nM range. The isolated RRM12 and KH12 domains displayed an affinity for the zipcode that was over 100-fold lower than the full-length protein. In contrast the KH34 domains displayed an affinity for the zipcode that was only slightly reduced compared to the full-length protein suggesting that these domains were the main site for ACTB zipcode recognition.[154] However, it was also noted that inclusion of the KH12 domains in the ZBP1-ACTB complex resulted in an increase in complex stability.[154]

To date ACTB mRNA remains the best characterised RNA target of ZBP1/IMP1. The RNA recognition sequences for KH3 and KH4 domain within the zipcode region have been determined and the structural recognition of these binding motifs characterised. The structure of the KH34 domain revealed the domains form an intramolecular anti-parallel pseudodimer.[85] This arrangement requires the ACTB mRNA to loop around the structure in order for both domains to recognise their RNA recognition motifs (Figure 1.6). Investigation into the effect of the RNA linker between the KH3 and KH4 binding sites revealed that the linker itself does not contribute to RNA binding. However, the length of the RNA linker was identified as being important for recognition. They identified a 10-fold reduction in affinity when the RNA linker was fewer than 10 nt in length. The reduction in binding affinity when the RNA linker was increased showed a less dramatic cut off. Increasing the linker to 25 nt reduced the affinity by 2.5-fold and an increase to 30 nt resulted in an 8-fold decrease. The dramatic reduction in affinity when the linker is fewer than 10 nt is a result of the RNA being too short to loop around the KH34 peusdodimer, and imposes a strict minimum size limit for the RNA linker.[85]

ZBP1/IMP1 association with the ACTB mRNA results in its translational inhibition and long-distance localisation to the growth cone of the neurons. The ACTB remains transcriptionally silenced while in complex with ZBP1/IMP1 until Src-mediated tyrosine phosphorylation at residue Tyr396, which is located in the linker between the KH2 and KH3 domain, triggers the release of the ACTB mRNA.[18] Following mRNA release, translation of the mRNA occurs driving the formation of an elevated concentration of β-actin at the leading edge of the cell. This alters actin dynamics which modulates the cytoskeleton and influences cell morphology. (Figure 1.13). A recent paper has identified an interaction between IMP1 and a cytoskeleton associated motor protein, Kinesin-like protein KIF11, providing a molecular link between IMP1 and the mRNA transport machinery. The study showed KIF11 to interact directly with IMP1 at a site within the RRM domains, and KIF11 association with ACTB mRNA is dependent on the presence of the IMP1 protein. The presence of KIF11 thus mediates association and localisation along the cells cytoskeleton.[155]

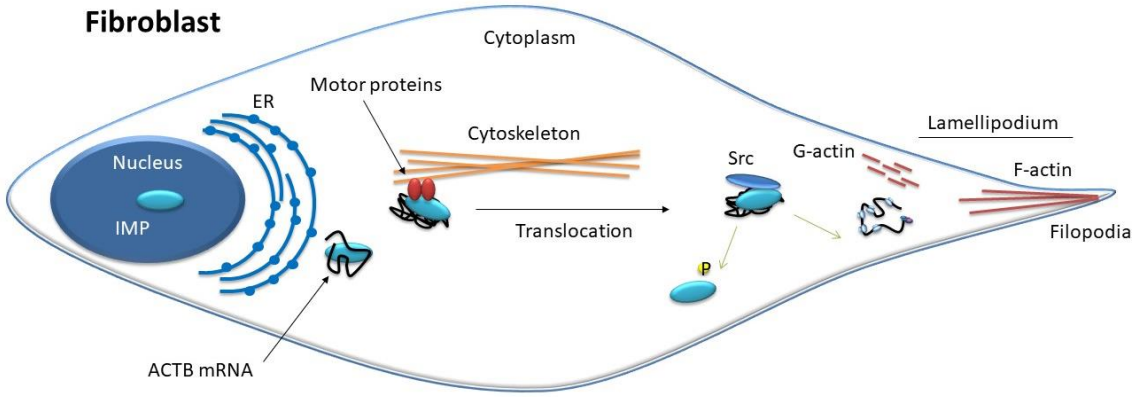


Figure 1.13: IMP1 mediates translational control of ACTB mRNA in polarised cells
IMP1-ACTB RNP complexes associate with motor proteins (KIF11, red ovals) in the cytoplasm to mediate localisation to the leading edge of the cell. ACTB mRNA is transnationally silenced until Src (dark blue oval) mediated phosphorylation of IMP1 triggers transcript release. Polymerisation of actin monomers at lamellipodia influence filopodia formation.

Phosphorylation of the IMP protein family has been reported to mediate functions with additional associated RNA transcripts. Studies have shown that mTORC1 mediated phosphorylation of IMP2 at residues Ser162 and Ser164 promotes binding to the 5' UTR of the IGF2 mRNA transcript.[161] This in turn initiates eIF-4E/5' cap-depended translation. The phosphorylation of these residues was determined to be mediated via mTORC1 as the allosteric inhibitor, rapamycin, was able to inhibit phosphorylation. IMP1 and IMP3 were shown to be phosphorylated at the corresponding residues Ser181 and Ser183. However, this phosphorylation was rapamycin independent and so is not mediated via mTORC1, but instead by mTORC2 which is insensitive to rapamycin treatment.[162] Interestingly the phosphorylation results in the same increase translation of the IGF2 mRNA, as with IMP2, yet mediated via another mechanism. Finally, IMP1 phosphorylation at residue Ser181 is also documented as mediating the translation of the IGF2 mRNA.[162] It is likely that phosphorylation and other post-translational modifications of the IMP protein family regulate interactions with other RNA transcripts, and indeed other protein partners, further investigation into this is required.

1.5.4 IMP proteins in cancer

In most normal adult tissues IMP1 expression is silenced or repressed. However, in a range of tumours and tumour derived cell lines IMP1 and IMP3 protein levels have been shown to be severely upregulated. The most cited malignancies IMP1 and IMP3 have been observed in, are those of the breast, colon, liver, kidney, pancreas, and female reproductive tissues.[141,163–168] There is less convincing evidence thus far for an oncogenic role for IMP2. This is consistent with the observation that IMP1 and to a lesser extent also IMP3 are mainly or even exclusively expressed during embryogenesis but become *de novo* synthesized in various malignancies. In contrast IMP2 is the only paralogue observed to be expressed at moderate levels in adult tissues.[136,151] The level of expression of IMP1 and IMP3 in tumours has been shown to directly correlate with tumour invasiveness and poor prognosis.[143,169]

The mechanisms by which IMP1 and IMP3 expression levels are severely upregulated in cancers in not well understood. One possible explanation is a result of the downregulation of microRNAs (miRNA) often observed in cancer systems.[170] These noncoding RNAs recognise seed sequences residing in the 3' UTR region of certain mRNA transcripts and increase mRNA degradation in association with the RISC complex or reduce translation. The let-7 family of microRNAs is an important family of microRNAs and is documented as controlling the degradation of several oncofoetal genes including the IMP family.[171] During the early stages of tumorigenesis, the let-7 miRNA family can become downregulated, suggesting a mechanism by which IMP levels can become upregulated.[171,172] Specifically, IMP1 which contains six let-7 seed sequences in the 3' UTR. Studies in the cancer derived cell lines k562 and HEPG2 actually showed IMP1 levels to be reduced upon the addition of exogenous let-7 miRNA. Conversely competition of the let-7 seed sequences with antisense oligonucleotides caused an increase in IMP1 expression levels.[171]

Other suggested mechanisms include a direct increase in IMP1 transcription via β-catenin, [173] a positive feedback loop by which MYC regulates the expression of IMP1, [142] and also a link to the WNT signalling pathway promoting IMP1 expression.[174] While the exact mechanism by which IMP expression levels are regulated in cancer remains elusive, the function of these proteins in cancer models is better reported.

IMP1s involvement in cancer progression is likely due to its documented function in controlling the cytoplasmic fate of oncogenic protein mRNAs, proteins which are typically involved in cell-cell adhesion, motility, and cell growth. Misregulation of these mechanisms results in uncontrolled cell growth and migration, which can result in metastatic tumours. For example, IMP1 association with the mRNA coding region instability determinant (CRD) of MYC prevents its premature CDR-dependent mRNA decay.[157] This directly increases MYC expression in tumour-derived cells and thereby promotes cell viability.

A review proposed a novel mechanism in which two separate functions of IMP1 affects two intracellular signalling networks that converge to increase tumour cell invasiveness (Figure 1.14).[175] Firstly, oncofoetal IMP1 enhances migration velocity of tumour-derived cells by influencing actin dynamics via inhibition of MAPK4 mRNA translation[176] and localisation of the ATCB mRNA. Concurrently, IMP1 facilities elevated levels of PTEN and CD44 via IMP1 dependent degradation inhibition.[145,176] The increased PTEN level promotes cell polarity while CD44 drives the formation of invadopodia. This has an overall effect on the directionality of cell migration. In turn, IMP1 promotes both the directionality and velocity of cell migration of tumour-derived cells to create an increased invasive phenotype, even in the absence of defined external guidance cues.

The role of IMP1 in cancer cell invasiveness exemplifies how the multiple RNA regulatory function of the protein enables fine-tuning of several intracellular signalling networks resulting in disease. It highlights the complexity of these disease systems and shows that often multiple genes and signalling networks are altered to create specific disease phenotypes.

IMP1 provides a potential drug target for the development of cancer therapeutics. Both the IMP1-MYC and IMP1-CD44 interactions have been targeted in two studies aiming to develop cancer treatments using antisense oligonucleotides (ASOs). Inhibition of IMP1 recognition of the CRD in MYC mRNA using ASOs in k526 cells resulted in a 70% decrease in cell proliferation. Similarly, ASOs were able to inhibit IMP1 interaction was the 3’ UTR of the CD44 *in vitro*, yet inhibition of this interaction in HeLa cells using ASOs did not result in a detectable functional effect.[177,178]

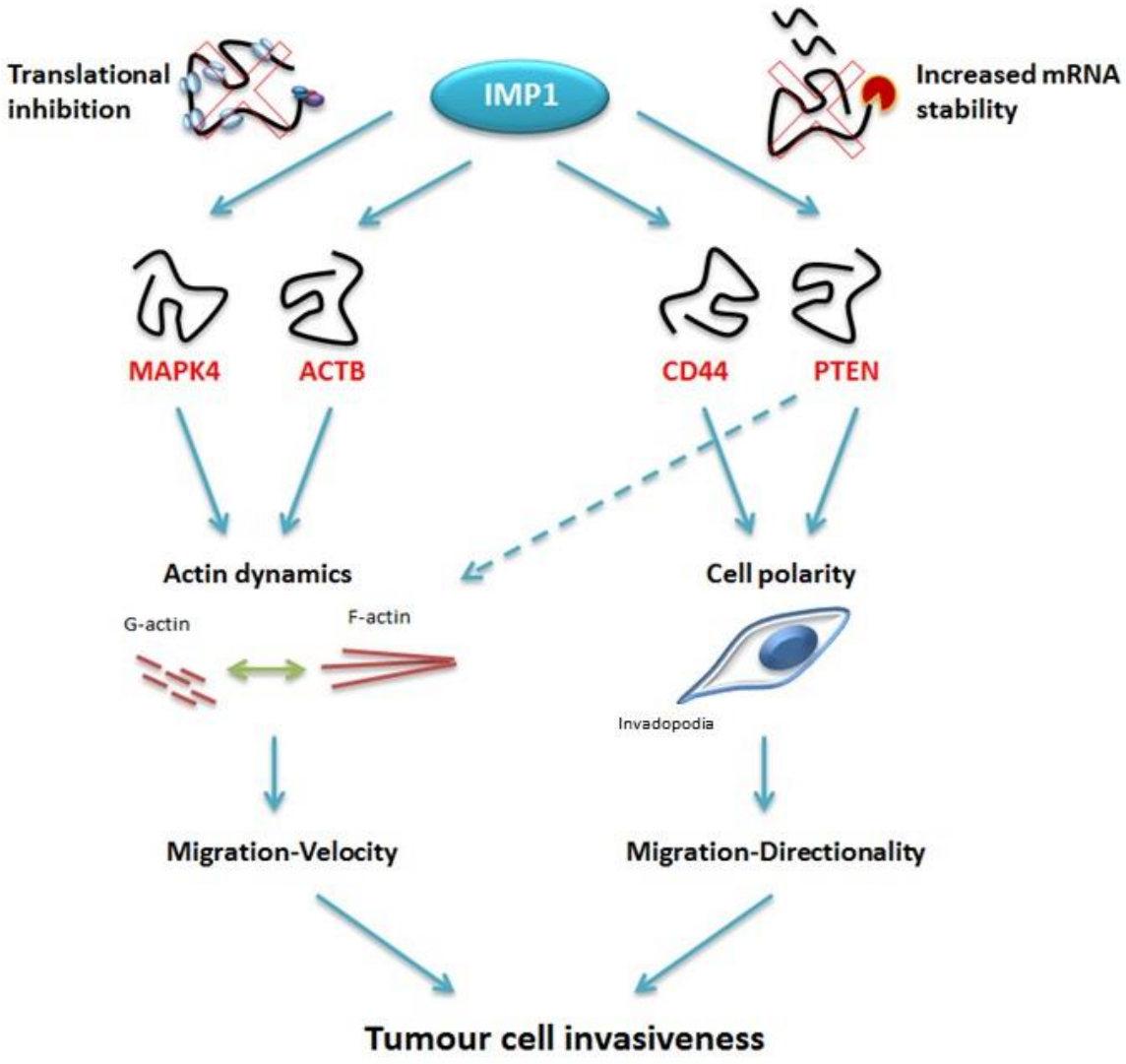


Figure 1.14: How IMP1's control of select RNA targets increases tumour cell invasiveness

Schematic describing IMP1 translational inhibition of MAPK4 and ACTB mRNAs altering actin dynamics to influence migration-velocity converging with IMP1s ability to influence cell polarity and cells migration-directionally via its stabilising effects on CD44 and PTEN mRNAs. These networks combine to result in invasive phenotype. Adapted from[175]

1.5.5 Exploring the sequence specific recognition of IMP RNA targets

To gain better understanding of the regulation of the diverse range of targets regulated by the IMP family a number of attempts have been made to identify the RNA interactome of the protein family and to define RNA recognition sequences. To date there has been two studies investigating the *in vivo* targets of the IMP family. The first was a photoactivatable-ribonucleotide enhanced CLIP (PAR-CLIP) study. The technique uses the basic principles of UV induced crosslinking immunoprecipitation assays but with the addition of photoactivable ribonucleotides to increase crosslinking efficiency and to enable the detection of protein RNA recognition sites. The group transiently expressed FLAG-IMP protein in HEK293 cells and identified RNA targets and recognition sequences of each IMP family member. They concluded all three IMP isoforms recognise a single RNA recognition motif CAUH (H = A, C, or U). This motif was identified in ~75% of identified RNA targets.[110] The study did not provide details on potential alternative RNA recognition motifs. A more recent study was performed in human pluripotent stem cells (hPSCs). This cell system was chosen as a system to greater understand how IMP proteins regulate mechanisms during early human development due to the cell types ability to self-renew and differentiate. Another version of CLIP was implemented, enhanced CLIP (eCLIP), and rather than overexpressing tagged IMP proteins the group immunoprecipitated endogenous IMP proteins with antibodies specific to each isoform.[144]

By comparing the gene regions where binding enrichment was observed the study revealed no enrichment in binding to intron sequences, which is to be expected from the cytoplasmic localisation of the IMP protein family.[83,153] They also identified IMP1 and IMP2 as having similar binding patters but when comparing binding within mRNA coding regions IMP3 biding sites did not correlate well with the sites of the other two isoforms. They also noted an enrichment in IMP1 and IMP2 binding to 3’ UTR regions with a 2.7- and 4.4-fold increase respectively. Enrichment for binding of coding regions was less with a modest 1.2 and 1.7-fold enrichment. Again, IMP3 did not correlate well with these patters as enrichment for 3’ UTR binding was less than that seen for coding

regions (2.3-fold enrichment for 3' UTR compared to 2.9-fold for coding regions). From this comparison they concluded IMP1 and IMP2 to share binding preferences to 3' UTR regions of mature mRNAs, whereas IMP3 favoured binding to coding exons.[144]

The group also performed a parallel RNA Bind-N-seq analysis of the IMP1 and IMP2 proteins to identify sequence specificity motifs of the proteins and compare these with target sites identified in their eCLIP study. Again, the group noted a high correlation between IMP1 and IMP2 6-mer enrichments. They observed enrichment of CA-rich motifs with 52% of IMP1 and 49% of IMP2 enriched 6-mers containing on of the following sequences (CACA, UACA, AACA, CAUA). They then compared these sequences with the IMP binding sites identified in the hESC eCLIP study. They were able to correlate enrichment of the core CACA motif in the eCLIP study but the other enriched sequences from the Bind-N-seq assay (UACA, AACA, CAUA) showed more variable enrichment, thus suggesting a disconnect to the observed IMP binding preferences *in vitro* and *in vivo*.[144]

Additional studies investigating the sequence specificity of the IMP protein include a RIP-Chip study, that lack the site-specific resolution associated with the CLIP studies, which also identified enrichment of multiple CA-dinucleotide-containing motifs.[156] Further to these studies, SELEX was performed on the KH34 di-domain of IMP1 and yielded enrichment of the RNA motif MCAY (M = A or C and Y = U or C) but also suggested enrichment of an additional G-rich element.[84] In addition to these sequences being conflicting, the well-studied KH4 recognition (KH4 – CGGAC) sequence within the zipcode RNA target is not represented.[179,180]

The lack of sequence consistency between the several approaches used to determine IMP RNA recognition motif specificity raises questions as to which method is more reliable. Secondly the results of the PAR-CLIP studies suggest a rather simple RNA-binding motif for all IMPs. Presumably this is oversimplifying the spatial complexity of IMP family RNA association. Moreover, these studies

underestimate the reported variation in RNA-binding properties among the IMP family. The contribution of distinct KH or RRM domains to the specific binding of IMP target RNAs requires further in-depth investigation.

The IMP1 RNA recognition sequence for the majority of IMP1 validated transcripts remains unknown. In most cases the data so far identifies only the RNA region which the IMP1 binds, for example the 3' UTR of the IGF2 mRNA or the CRD of the MYC mRNA. The RNA recognition sequences that have so far been published for the individual KH domains cannot be used to fully explain IMP1 association with the diversity of RNA transcripts IMP1 has been observed to interact with. In order to map the binding sites of IMP1 on a broad ensemble of cellular targets the recognition motifs for all four KH domains need to be identified.

1.6 Conclusions

In this introduction is have briefly reviewed how RNA binding proteins control cellular function through regulating post-transcriptional regulation. I have discussed examples of how misregulation of these networks can lead to disease. In order to develop therapeutics to treat these deficient RBP associated diseases, we need to better understand how RNA binding proteins select their cellular RNA targets. In particular, the role of combinatorial binding, and the roles of individual RBD RNA sequence specificity in *in vivo* RNA target selection, needs to be explored further.

I have discussed the oncofoetal IMP protein family and described their roles in embryonic development and cancer. IMP1 provides a model system in which we can explore *in vivo* RNA target selection of multidomain RNA binding proteins further.

1.7 Aims

The primary aim of this thesis was to develop an *in vivo* system in which we could investigate IMP1 in-cell RNA binding at the individual domain level. We planned to achieve this by using structural knowledge of protein-RNA recognition, in particular the KH3 KH4 domains of IMP1, to introduce mutations into the individual KH domains of the IMP1 protein. In the first instance, we introduced the GDDG mutation to abolish RNA binding of the KH domains. Through the use of stably transfected cell lines, we performed iCLIP on these mutant constructs to investigate, through comparative analysis, if the RNA binding profiles of the mutant proteins shifted, thus enabling us to identify RNA target selection of the IMP1 protein at the individual KH domain level. We then planned to move further and investigate the importance of domain sequence specificity for *in vivo* RNA target selection, and explore the RNA binding properties of the less well characterised RRM12 domains.

Chapter 2. Materials & Methods

2.1 Molecular Biology

2.1.1 Bacterial Strains

E.coli BL21(DE3) cells (Millipore) were used for all recombinant bacterial protein expression. The DE3 lysogen contains the T7 polymerase gene under the *lac*UV5 promotor. Addition of IPTG induces expression of T7 polymerase. Expression vectors used for protein expression contained the T7 *lac* promotor. The polymerase in turn transcribes the mRNA downstream of the T7 promotor at high copy number. Cloning DNA encoding the desired protein constructs downstream of the T7 promotor results in the overexpression of the protein. Plasmid DNA amplification was performed by transforming *E.coli* DH5α cells (Novagen). Site directed mutagenesis cloning implemented the use of ultra-competent *E.coli* XL10-Gold cells (Agilent) due to the low quantity of DNA construct produced in the procedure that is used in the bacterial transformation.

2.1.2 Plasmid vectors and purification

All bacterially expressed recombinant proteins were expressed using the pET-M11 vector (Novagen). The pET-M11 vector contained an N-terminal His-tag for protein purification and a TEV protease digestion site for purification tag removal. The vector also contained resistance to the antibiotic kanamycin which allowed for recombinant bacteria selection after transformation.

Plasmids were amplified and purified from *E.coli* DH5α cells. Cells were transformed and cultured in LB media containing the required selection antibiotic depending on plasmid being amplified. Qiagen Mini (bacterial plasmids) or Gigaprep (mammalian plasmids) kits were used for plasmid purification depending on yield required. Manufactures protocol was followed accordingly and DNA eluted in ddH_2O. Final purified plasmid concentration was determined by UV spectrophotometry at wavelength A_{260}.

2.1.3 Polymerase chain reaction (PCR)

All standard PCR amplification reactions were performed using an Eppendorf mastercycler nexus and Deep Vent DNA polymerase kit (NEB). PCR amplification primers were designed based on the desired protein construct boundaries and contained approximately 21 nt that were complementary to the DNA construct being amplified. Restriction digestion sites were incorporated in the forward and reverse amplification primers that were complementary to the restriction sited used to clone the construct into the corresponding expression vector (Appendix I). All PCR amplification primers were synthesised and desalt purified by Sigma Aldrich. PCR reactions were carried out using the standard guidelines supplied with the polymerase kit in a final reaction volume of 50 ul in a thin walled 200 ul PCR tube.

Reagent	***Volume (µl)***
ddH_2O	40
ThermoPol Buffer (10X)	5
dNTPs	1
Forward Primer (15 µM)	1
Reverse Primer (15 µM)	1
Template DNA Vector (10 ng/µl)	1
Deep Vent DNA polymerase	1

Table 2.1: Standard Deep Vent DNA polymerase PCR amplification reaction composition

Step	***Temperature (ºC)***	***Time (Seconds)***
1	95	180
2	95	30
3	65	20
4	72	60 /kb
5	72	300
6	4	Hold
	Cycles steps 2 – 4 (28 times)	

Table 2.2: Typical thermocycler programme for PCR amplification

Amplified DNA products were purified by electrophoresis on a 1% agarose TBE gel and visualised using ethidium bromide (Sigma) and UV transillumination. DNA product was excised from the agarose gel and purified using QIAquick gel extraction kit (Qiagen) following kit protocol.

2.1.4 Restriction enzymes and DNA ligation reactions

Full-length human IMP1 was cloned from a pCMV6-Entry vector using PCR amplification to incorporate 5' XhoI and 3' BamHI restriction sites. The PCR products were sub-cloned into either an N-terminal FLAG or C-Terminal FLAG pcDNA5 vector (obtained from Ule, The Francis Crick Institute. UK) using BamHI and XhoI restriction enzymes. IMP1 and IMP3 RRM12 domain constructs were also cloned from a pCMV6-Entry vector and PCR amplified with primers to incorporate 5' NcoI and 3' XhoI restriction sites and sub-cloned into pETM-11 expression vector. The chosen restriction enzymes were compatible with each other to perform double digests. All digests were carried out at 10 U restriction enzyme per 1 µg of DNA at 37^{o}C for 1 hour. All restriction enzymes were obtained from NEB and the corresponding protocol for each double digest was followed. Digested PCR amplification products were purified using a PCR clean up kit (Qiagen) to remove 5' and 3' digested nucleotides. Digested vectors were purified using agarose gel electrophoresis in the same manner as amplified PCR DNA fragments. T4 DNA ligase (NEB) was used to ligate digested PCR product and digested vector with complementary sticky ends according to manufacturer's instructions. Ligation reactions were set up to included 1:1 and 4:1 PCR insert:vector ratios.

Reagent	***Volume (µl)***
ddH_2O	to 20 µl final
Reaction buffer (10X)	2
BSA (20X)	1
DNA (PCR insert or vector)	1 µg
5' Restriction Enzyme (10 U/µl)	1
3' Restriction Enzyme (10 U/µl)	1

Table 2.3:Typical double restriction enzyme digestion reaction for PCR insert or expression vectors

Reagent	*Volume (µl)*
ddH_2O	to 20 µl final
T4 DNA ligase buffer (10X)	2
Digested Vector	-
Digested PCR insert	-
T4 DNA ligase (400 U/µ1)	1

Table 2.4: Typical DNA ligation reaction
Adjust insert and vector volumes to obtain a final total DNA concentration of 100 ng using 1:1 and 3:1 PCR insert:vector ratios

2.1.5 Transformations

All bacterial transformations were performed using a standard heat shock protocol according to manufacturer instructions. Cells were incubated in S.O.C media (Invitrogen) for 1h after heat shock to allow for recovery before being plated LB agar plates containing appropriate selection antibiotic and incubated overnight at 37^{o}C. Colonies were selected, and bacterial cultures were grown for either protein expression or plasmid amplification.

2.1.6 Site-directed mutagenesis and DNA sequencing

All site-directed mutations were performed using the QuickChange Lightning Site-Directed mutagenesis kit (Agilent Technologies) according to the manufacturer protocol. Mutagenesis primers (Appendix I) were designed using Agilent Technologies primer design programme and synthesised and HPLC purified by Sigma Aldrich. All DNA constructs were sequences by Beckman Coulter.

Reagent	*Volume (µl)*
ddH_2O	38
Reaction Buffer (10X)	5
Quick Solution	1.5
dNTPs	1
Forward Primer (100 ng)	1.25
Reverse Primer (100 ng)	1.25
Template DNA Vector (10 ng/µl)	1
pfu lightening DNA polymerase	1

Table 2.5: Standard QuickChange Lightning Site-Directed mutagenesis reaction

Step	*Temperature (°C)*	*Time (Seconds)*
1	95	120
2	95	20
3	60	10
4	68	30/kb
5	68	300
6	4	Hold
	Cycles steps 2 – 4 (18 times)	

Table 2.6:Typical thermocycler programme for QuickChange Lightning Site-Directed mutagenesis
Note: Mutagenesis of FLAG-IMP1 KH1-4 GDDG construct required the use of a 'touch down' thermocycle programme due to the similarity of DNA sequence after introducing 2 or more GDDG mutations.

2.2 Protein expression and purification

2.2.1 ^{15}N Labelled protein expression in *E.coli* BL21(DE3)

pET-M11 DNA constructs (2µl) were used to transform 25 µl of *E.coli* BL21(DE3) (Millipore) cells via standard heat shock protocol as described above. Transformed cells were used to inoculate 500 ml of ^{15}N M9 minimal media with 30 µg/ml kanamycin. Cells were grown overnight at 37°C. Overnight culture was used to inoculate 1.5 L (x3) of fresh M9 minimal media and 30 µg/ml kanamycin to achieve an OD_{600} of 0.1. Cultures were incubated at 37°C until they entered mid-log phase (OD_{600} 0.6 – 0.8). Isopropyl β-D-1-thiogalactopyranoside (IPTG) were used to induce protein expression (0.5mM final concentration). Cells were grown for a further 16 h at 22°C before being harvested by centrifugation at 6,000*g* for 20 minutes. Bacterial pellets were then stored at -80°C.

2.2.2 Native protein purification

All recombinantly expressed proteins contained a His-tag for purification. Frozen harvested cell pellets were lysed in ice cold nickel purification lysis buffer (10 mM TRIS pH 8.0, 10 mM Imidazole, 1 M NaCl, 140 µl/L β-Merc, 1 µl/ml TRITON X-100, 1 mg/ml Lysozyme, DNase, Complete Protease inhibitor tablets (Roche)). Cells were sonicated (Branson Sonifier 450) on ice at a 40% power setting with 15 bursts. Burst length was dependent on bacterial lysate volume (~20 sec/30ml). Cell lysate was clarified by centrifugation at 45,000*g* for 1 hour at 4^{o}C.

Clarified supernatant was purified using 5 ml of pre-equilibrated Ni-NTA nickel agarose resin (Qiagen) using gravity flow chromatography. Resin was incubated with bacterial lysate for 1 h at 4^{o}C (batch binding stage). Incubated resin was placed into an extract clean column (GRACE) and washed with 5 x resin volume Ni-NTA wash buffer 1 (10 mM Tris-HCL pH 8.0, 10 mM Imidazole, 1 M NaCl, 140 ul/L β-Merc) followed by a second high-stringent wash of 5 x resin volume Ni-NTA wash buffer 2 (10 mM Tris-HCL pH 8.0, 30 mM Imidazole, 1 M NaCl, 140 ul/L β-Merc). Bound His-tag fusion proteins were eluted in 2 x resin volume Ni-NTA elution buffer (10 mM Tris-HCL pH 8.0, 300 mM Imidazole, 1 M NaCl, 140 ul/L β-Merc). Aliquots of each fraction were taken for SDS-PAGE analysis. Elution fractions were dialysed against 100 x TEV digestion buffer by volume (10 mM Tris-HCL pH 8.0, 10 mM Imidazole, 50 mM NaCl, 140 ul/L β-Merc) using 10,000 MWCO dialysis tubing (Spectrapor) overnight at 4^{o}C with gentle stirring. Dialysed samples were digested with TEV protease for 3 h at 37^{o}C to remove His-purification tag. Digested sample was reverse Ni-NTA nickel resin purified to remove His-tagged TEV protease and cleaved His-tag. SDS-PAGE was used to determine the efficiency of TEV protease digestion before samples were carried through to FPLC purification.

2.2.3 Denatured protein purification

Protein constructs that were purified under denatured conditions were purified using the same protocol as native protein purification but with the addition of 8 M urea to all the purification buffers used above. Proteins were refolded using stepwise dialysis after initial nickel resin purification. Samples were placed into 10,000 MWCO dialysis tubing and dialysed in 100 x dialysis buffer by volume. Samples were first dialysed in buffer (4 M Urea, 10 mM Tris-HCL pH 8.0, 10 mM Imidazole, 50 mM NaCl, 140 ul/L β-Merc) for 5 h at room temperature followed by a second round of dialysis (1 M Urea, 10 mM Tris-HCL pH 8.0, 10 mM Imidazole, 50 mM NaCl, 140 ul/L β-Merc) overnight at 4^{o}C. Finally, samples were dialysed into standard TEV cleavage buffer (as above) for 24 h at 4^{o}C. Refolded samples then proceeded native protein purification protocol as above.

2.2.4 Size exclusion chromatography

Size exclusion (gel filtration) chromatography was used as the final purification step following nickel agarose purification. Proteins that co-purified with nucleic acid (as determined by UV spectroscopy) were purified using cation exchange chromatography (below). TEV digested and purified samples were concentrated using a 10,000 MWCO vivaspin (Sartorius) via centrifugation at 4,000*g* at 4^{o}C. Samples were concentrated to a volume of 5 ml and loaded onto a pre-equilibrated HiLoad 16/60 Superdex 75 column (GE Healthcare) using an AKTA system (Amersham). Gel filtration purifications were performed at a flow rate of 1 ml/min and 3 ml fractions were automatically collected using a Frac-900 (Amersham). Elution fractions were analysed via SDS-PAGE to assess purity. Pure fractions were pooled and dialysed into appropriate buffer before being stored at -20^{o}C. All size exclusion purifications were performed in 10 mM Tris-HCL pH 8.0, 50 mM NaCl, 140 ul/L β-Merc buffer.

2.2.5 Cation exchange chromatography

Cation exchange chromatography was performed to purify proteins from co-purifying nucleic acids. Proteins were dialysed into Buffer A (10 mM Tris-HCL pH 7.3, 10 mM NaCl, 0.5 mM TCEP). A Superloop (GE Healthcare) was used to load the dialysed sample onto a pre-equilibrated HiLoad 26/10 SP Sepharose cation exchange column (GE Healthcare). Column was washed with 3 column volumes of Buffer A to remove unbound species. Buffer B (10 mM Tris-HCL pH 7.3, 1M NaCl, 0.5 mM TCEP) was used to elute bound proteins using a 5-column volume gradient of 0% - 100% Buffer B. Purifications were performed at a flow rate of 8 ml/min. Sample purity was analysed using SDS-PAGE and UV absorbance spectrometry. Pure fractions were pooled and dialysed into appropriate buffer before being stored at -20^{o}C. Figure 2.1 summarises the protein purification strategy we implemented.

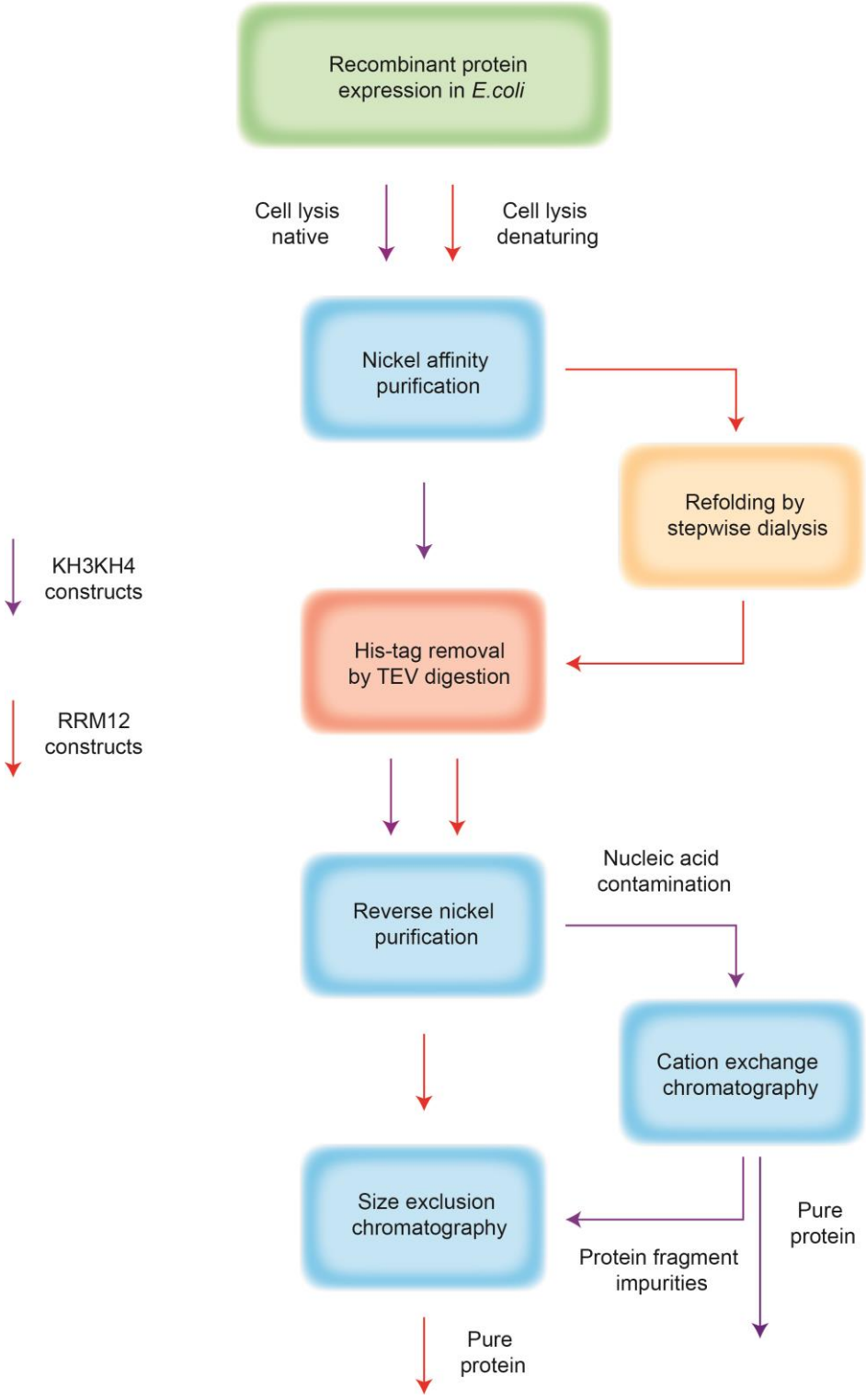


Figure 2.1: Flow chart of protein purification for KH and RRM domain constructs

2.2.6 SDS-PAGE

Protein samples of 8 ul were taken and added to 2 ul of 4x NuPAGE LDS sample buffer (Novex) before being heated at 95^{o}C for 5 mins. Samples were loaded into a 12% Bis-Tris pre cast gel (Invitrogen) placed into a XCell SureLock mini-cell buffered in NuPAGE MES SDS Running buffer (1X) (Novex). SeeBlue Prestained standard marker (Invitrogen) was used to determine molecular weight of protein species. Gels were run at 180 V for 50 min using a Bio-Rad power pack. Gels were developed using InstantBlue (Expedeon) with gentle rocking for ~1 h.

2.2.7 Protein quantification

Protein concentrations were determined by UV spectroscopy using a CE2502 2000 Series (Cecil) spectrophotometer. A spectrum of $\lambda_{210} - \lambda_{320}$ was recorded for each protein sample after blanking with corresponding protein buffer in the same High Precision Quartz cell (Hellma) with 1 cm light path length. Using the Beer-Lambert law (Equation 1) and the calculated extension coefficient of the protein based on the primary amino acid sequence using Protparam of the ExPASy resource portal concentrations were calculated using Equation 2.

Equation 1: Beer-Lambert law

$$A = \varepsilon . C . L$$

Where A corresponds to the absorbance of the protein, ε the extension coefficient, C is the concentration of the protein in mol.L^{-1}, and L is the light path length of the sample in cm.

Equation 2

$$C = \frac{A280}{\varepsilon 280}$$

With the light path length fixed to 1 cm (L = 1) the concentration of the protein (C) is equal to the absorbance value at λ_{280} (A280) divided by the extension coefficient value at λ_{280} (ε280)

2.3 Mammalian cell culture

All mammalian cell experiments were performed with HeLa Flp-In T-REx cells (obtained from Taylor, University of Manchester. UK). HeLa cells were cultured in 10 cm cell culture dishes (Corning) and incubated in a cell culture incubator at 37ºC in 5% CO_2 humidified air. For harvesting, adherent cells were washed with DPBS and detached with 1x trypsin-EDTA (Invitrogen) and incubated for ~3 min at 37ºC. Trypsin was inactivated with the addition of cell culture medium containing 10% FBS. Cell counting was performed by adding Trypan-Blue solution (Invitrogen) to the cells in a 1:1 ratio and cells were counted in a Neubauer counting chamber using an inverted light microscope.

2.3.1 Transfection of HeLa cells

The TransIT-HeLaMONSTER Transfection Kit (Mirus) was used for all HeLa cell transfections. The kit utilises cationic, lipid-based transfection reagents which enable transfection of plasmid DNA into mammalian cells. During the transfection protocol liposomes form which carry a positively charged head group. This head group interacts with negatively charged DNA which results in the formation of DNA-liposome complexes. The positive charge of the DNA-liposome complexe interacts with the cell membrane. The DNA-liposomes are then uptake into the cell vial endocytosis.

2.3.2 Generation of IMP1 construct expressing HeLa Flp-In T-REx cell lines

HeLa Flp-In T-REx cells were first tested for Hygromycin B (Invitrogen) sensitivity prior to cell line generation. Cells (1.0×10^6) were cultured in 10 cm dishes in DMEM + 10% FBS media (Invitrogen) for 12 hours. Media was replaced with DMEM + 10% FBS media containing a range of Hygromycin B concentrations, 50 – 400 µg/ml. Cell death was monitored using an inverted light microscope for two weeks. Cell media was changed every 48 h during the course of the experiment. Establishing HeLa cell sensitivity to Hygromycin B determined the concentration of Hygromycin B required in cell selection medium for stable cell line generation.

Flp-In T-REx system plasmids (pcDNA5/FRT/TO and pOG44) were obtained as a gift from the lab of Jernej Ule (The Francis Crick Institute). The pcDNA5/FRT/TO plasmid was modified to incorporate either a N-terminal or C-terminal FLAG tag. The HeLa Flp-In T-Rex cell line contained a single Flp-In recombination (FRT) site. Cells were co-transfected with pcDNA5/FRT/TO (containing IMP1 gene of interest) and the pOG44 vector (encoding Flp recombinase which catalyses the insertion of the gene of interest into the HeLa cell genome at the Flp recombination site) in a 1:11 ratio by mass. Briefly, 500 µl of Opti-MEM 1 reduced serum medium (Invitrogen) with 25 µl of TransIT-HeLaMONSTER transfection reagents, and 11 µg of plasmid was DNA was incubated at room temperature for a total of 15 min. Transfection reaction was added to 3 x 10 cm plates of confluent HeLa cells. Cells were incubated for 16 h before medium was removed and growth medium changed to selection medium (DMEM + 10% FBS, 15 µg/ml Blasticidin and 200 µg/ml Hygromycin B). Cell culture medium was changed every 2 days. In general, it took between 2 and 3 weeks until stable cell clones were obtained. Correct insertion of the IMP1 gene into the HeLa cell genome removes the cell lines resistance to Zeocin. To confirm correct FTR site integration a selection of cells was tested for Zeocin sensitivity by culturing cells in DMEM + 10% FBS, 15 µg/ml Blasticidin and 100 µg/ml Zeocin for two weeks and observing cell death via an inverted light microscope.

2.3.3 Doxycycline induction of HeLa Flp-In T-Rex lines

Doxycycline induction of incorporated IMP1 genes was tested in 12 well cell culture plates (Corning). Cells were seeded in blank medium (DMEM + 10% FBS) and induced with a range of Doxycycline (Invitrogen) concentrations (0-1000 ng/ml) for 24 h. Cells were harvested and protein expression was determined by western blot analysis (outlined below)

2.3.4 Western blot analysis

Harvested cells were lysed in ice cold RISC buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.5% NP-40, 2mM MgCl2, 1 mM DTT, 1 mM PMSF and EDTA free protease inhibitor (Roche)) for 1 h at 4^{o}C with rigorous shaking. Total protein concentration of cell lysate was determined via Bradford protein assay kit (BioRad) and 12.5 ug of total cell lysates was run on an SDS-PAGE gel (as outlined above) and transferred onto a nitrocellulose 0.4 µm pore size membrane (Millipore) using a XCell SureLock Blot Module (Invitrogen) run at 30 V for 60 min buffered in NuPAGE Transfer Buffer 1X (Invitrogen) and 10% methanol. The membrane was incubated in blocking buffer (1x PBS, 0.1% Tween 20, and 5% milk) on a rocking table at room temperature for 1 h. Blocking buffer was removed and blot was probed with primary antibodies overnight at 4^{o}C. The membrane was washed four times for 10 min with 1x PBS, 0.1% Tween 20 and incubated with the corresponding secondary antibodies for 2 h at room temperature on a rocking table. After secondary antibody removal and additional washing steps as before, blot was exposed to photographic film (SLS) using ECL detection reagent kit (GE Healthcare) and developed. Western blots were also probed with anti-GAPDH antibody with served as a loading control.

Primary antibodies	*Source*	*Working concentration*
M2 clone monoclonal mouse anti-FLAG	Sigma	1/10,000
Polyclonal rabbit anti-IMP1	Siddle Lab, University of Cambridge. UK	1/5,000
Monoclonal mouse anti-GAPDH	Millipore	1/20,000

Table 2.7: Primary antibodies used for western blot analysis and immunoprecipitation assays

2.3.5 Immunoprecipitation of FLAG-IMP1 constructs

Immunoprecipitation of the FLAG IMP1 constructs was optimised for immunoprecipitation with the mFLAG and RbIMP1 primary antibodies. The Dynabead (Novex) system was used in conjunction with a DynaMag-2 (Life Technologies) for the optimisation. Stable cell lines were grown and induced as before, harvested and lysates prepared for IP. Protein G Dynabeads were used for anti-mouse FLAG IP and a 50/50 ratio of Protein A and Protein G Dynabeads for anti-rabbit IMP1 IP. Beads were prepared according to manufactures instructions and incubated at 4^{o}C for 1 h with either mFLAG antibody (1 µg/ml of total lysate) or RbIMP1 antibody (5 µg/ml of total lysate). A mMyc antibody served as a negative control and was incubated with 50/50 Protein A/Protein G Dynabeads at 1 µg/ml for the same length of time. After washing pre-incubated Dynabeads, cell lysate was incubated with the beads for a further 3 h at 4^{o}C. Beads were then washed to remove unbound sample and eluted in 6 x SDS-PAGE loading buffer (Novex) at 90^{o}C for 5 min before running on SDS-PAGE and analysed via mFLAG and RbIMP1 western blot.

2.3.6 Analysis of FLAG-IMP1 cellular localisation via immunofluorescence

The Life Technologies Image-it Fix-Perm kit was used for HeLa cell immunofluorescence studies. HeLa cells expressing FLAG-IMP1 constructs were plated onto BioCoat CultureSlides (Falcon) 24 h prior to staining. HeLa cells were treated with MitoTracker (Invitrogen) for 30 min and fixed in 4% paraformaldehyde in PBS (Invitrogen) at 37°C in humidified incubator for 10 minutes. Cells were permeabilised at room temperature for 15 minutes in 1x PBS and 1.0% Triton X-100. All cells were blocked in 1X PBS plus 3% BSA blocking solution at room temperature for 1 h. The following primary antibodies and dilutions were used: mouse anti-FLAG (Sigma), 1:1,000; and rabbit anti-IMP1 (Siddle lab). Primary antibodies were incubated for 2 hours at room temperature. Cells were washed with washing buffer as described in kit protocol by incubating with secondary antibodies. Fluor-conjugated antibodies Alexa donkey 647 (Invitrogen) and Alexa donkey 488 anti-mouse (Invitrogen), were incubated for 1 h at room temperature at a concentration of 1:200. Cells were incubated with NucBlue stain (Invitrogen) in 1x PBS for 5 min at room temperature and washed with 1x PBS 3 times before slides were mounted onto cover slips using ProLong Gold (Invitrogen). Images were acquired on an Olympus X widefield fluorescence microscope.

2.4 Individual nucleotide resolution crosslink immunoprecipitation (iCLIP)

2.4.1 UV crosslinking and cell harvesting

The iCLIP protocol was performed as previously described.[116] An overview of the steps of the protocol is described in (Figure 2.2). Flp-In T-Rex HeLa cells were plated in 10 cm cell culture dishes and FLAG-IMP1 construct expression was induced with 100 ng/ml doxycycline for 24 h (as described above). Cell media

was removed and 6 ml of ice cold 1x PBS (Invitrogen) was added to each dish of cells. Plates were UV irradiated at 150 mJ/cm2 at 254 nm on ice in a UV Stratalinker 2400 (Stratagene). Immediately cells were harvested using a cell scraper and collected in 3 x 2 ml Eppendorf tubes. Cells were pelleted in a table top centrifuge at 376*g* for 1 min at 4ºC. PBS was aspirated before cell pellets were snap frozen on dry ice and stored at -80ºC until required. (Figure 2.2 - Step 1 & 2)

2.4.2 Partial RNA digestion

Cell pellets were thawed on ice before being resuspended in iCLIP cell lysis buffer. The number of cell pellets required for an individual iCLIP reaction was dependent on the FLAG-IMP1 construct due to altered RNA binding affinities (Chapter 3.12) (WT, KH1DD, KH2DD: 3 pellets and KH3DD, KH4DD: 5 pellets). Pellets were lysed in a total of 1 ml of lysis buffer. RNase I (Ambion) was diluted to the required concentration in 1x PBS (1/10 for high RNase treatment and 1/500 for sample preparation). 2 µl of TURBO DNase (Ambion) and 10 µl of the required RNase I dilution was added to cell lysate. Cells were incubated exactly for 3 min at 37ºC whilst shaking at 1100 rpm. Cells were placed back on ice for 3 min before centrifuging at >18,000x g for 10 min at 4ºC to remove cell debris. Supernatant was transferred to new 1.5 ml Eppendorf tube. (Figure 2.2 - Step 3)

2.4.3 Protein RNA complex Immunoprecipitation

FLAG-IMP1 RNA complexes were immunoprecipitated as described above using Dynabeads and the DynaMag-2 system. For each iCLIP reaction 100 ul of Dynabeads Protein G were pre-incubated with 10 ug of FLAG antibody as described above. For no antibody control reactions, Dynabeads are prepared in the same manner but no FLAG antibody is added to antibody binding mix. Preincubated Dynabeads were added to treated cell lysate and were incubated for 2 h at 4ºC with end-over-end mixing. Dynabeads were separated from cell

lysate using DynaMag. Supernatant was removed, and beads were washed 2 x in high salt wash buffer. Beads were washed 2 x in PNK wash buffer. Washed beads can be resuspend in 1 ml PNK wash buffer and stored at 4^{o}C before proceeding to next step. (Figure 2.2 - Step 4)

2.4.4 RNA adapter ligation

RNA molecules crosslinked to immunoprecipitated FLAG-IMP1 proteins needed to be ligated to a DNA/RNA adaptor required for later reverse transcription. PNK buffer was removed from Dynabeads, to ensure all buffer was removed beads were pulse centrifuged and remaining buffer was aspirated a second time. T4 Polynucleotide Kinase (PNK) enzyme (NEB) was used to remove the 5' phosphate from the RNA molecules. De-phosphorylation buffer is buffered to pH 6.5 which is optimal for PNK enzyme phosphatase activity (Figure 2.2 - Step 5). Dynabeads were resuspend in 20 µl of de-phosphorylation buffer and incubated for exactly 20 min at 37^{o}C whilst shacking at 1100 rpm. Dynabeads were washed with 1 x PNK buffer followed by 1 x high salt wash buffer with end-over-end mixing for 5 min at 4^{o}C. Finally, Dynabeads were washed with 2 x PNK buffer. DNA/RNA adaptor was ligated to RNA fragments using T4 RNA Ligase 1 (NEB) (Figure 2.2 - Step 6). Dynabeads were resuspend in 20 µl of ligation mix and incubated overnight at 16^{o}C whilst shaking at 1100 rpm. After ligation Dynabeads were washed with 1 x PNK buffer followed by 2 x high salt wash buffer. After the final wash Dynabeads were transferred to new 1.5 ml Eppendorf tubes using 1 x PNK buffer. Transferring to new tubes ensures any excess DNA-RNA adaptor is removed.

2.4.5 Protein-RNA complex visualisation

In order to visualise FLAG-IMP1 RNA complexes RNA molecules are radiolabelled with ^{32}P (Figure 2.2 – Step 7). 20% of Dynabeads were collected from the previous adaptor ligation step and preceded to 5' ^{32}P labelling. In fresh

1.5 ml Eppendorf tubes the aliquot of beads were resuspended in 4ul of hot PNK mix. The hot PNK mix contained PNK enzyme (NEB) that catalyses the phosphorylation of the 5' end of the RNA molecules with the ^{32}P isotope (PerkinElmer Health Sciences) also contained in the hot PNK mix. Dynabeads were incubated at 37^{o}C for 5 min with mixing at 1100 rpm. Hot PNK mix was removed from the Dynabeads before hot beads were resuspended in 1 x NuPAGE loading buffer. Hot beads resuspended in loading buffer were added to the remaining beads of the corresponding iCLIP reaction. For SDS-PAGE purification a 4-12% NuPAGE Bis-Tris gel (Novex) was placed in a XCell SureLock mini-cell with 1x MOPS NuPAGE running buffer (Novex). FLAG-IMP1 RNA complexes were eluted from Dyanbeads by heating at 80^{o}C for 5 min. Loading buffer was loaded on the gel along with 5 ul of pre-stained protein ladder. Gel was run at 180 V for 60 min (Figure 2.2 – Step 8). Radiolabelled FLAG-IMP1 RNA complexes were then transferred to a Protran nitrocellulose membrane (Whatman) as described in the western blotting method. The membrane was exposed to photographic film in a shielded cassette at -80^{o}C to amplify radioactive signal (Figure 2.2 – Step 9).

2.4.6 RNA isolation

The radiolabelling of the RNA fragments produces an autoradiograph which is used as a template to identify regions of the membrane containing FLAG-IMP1 RNA complexes that are to be isolated (Figure 2.2 – Step 10). Sections of the film are removed and placed back onto the hot membrane. The desired sections of the membrane are then removed and cut into small fragments using a serial blade. Membrane fragments were placed into a 1.5 ml Eppendorf tube containing 200 µl PK buffer. The PK buffer contains proteinase K (Roche) which digests the FLAG-IMP1 proteins thus releasing the RNA fragments from the nitrocellulose membrane. Membrane fragments were incubated with PK buffer for 20 min at 37^{o}C whilst shaking at 1100 rpm. An additional 200 µl of PK buffer containing 7 M urea was added to the mix and incubated for a further 20 min

(Figure 2.2 – Step 11). The RNA fragments were Phenol/Chloroform (Sigma) purified using a Phase Lock Gel Heavy Tube (VWR). 400 µl of the Phenol/Chloroform was added to the reaction mix before being placed into the Phase Lock tube. Tubes were incubated for 5 min at 30^{o}C whilst shaking at 1100 rpm. Phases were separated via centrifugation at 13,000 rpm at room temperature. The aqueous phase was removed and placed into a fresh tube before RNA precipitation step. RNA is predicated overnight at -20^{o}C by adding 0.75 µl Glycoblue co-precipitant (Ambion), 40 µl 3 M sodium acetate pH 5.5, and 1 ml of ice cold 100% ethanol. The addition of the Glycoblue aids in the precipitation of small quantities of RNA. In addition, it provides a coloured pellet in the later centrifugation step with allows visualisation of the precipitated RNA.

2.4.7 Reverse transcription (RT) and gel purification

To sequence the RNA molecules that crosslinked to the FLAG-IMP1 proteins the RNA needs to be reverse transcribed into DNA. During this stage of the protocol specially designed RT primers are used to incorporate DNA elements that are important for later steps in the protocol. The RT primers are complementary to the adaptor region previously ligated to the RNA molecules, and thus primes reverse transcription of the SuperScript III Reverse Transcriptase (Life Technologies) enzyme used in the RT reaction. Additionally, the RT primers include barcode regions that enable the user to identify which iCLIP experiment the DNA sequences correspond to (de-multiplexing) in addition to a unique barcode region with are used to account for PCR amplification bias before DNA reads are mapped to the reference genome. Finally, the RT primers also contain a BamH I restriction site. The incorporation of the BamH I site along with a later circularisation step of the cDNA molecules, removes the need to ligate both 5’ and 3’ adaptor regions in the previous step of the protocol which is required in the original CLIP and PAR-CLIP protocols (Figure 2.2 – Step 12).

After the overnight incubation the RNA precipitation reaction from the above step was centrifuged at 15000 rpm for 20 min at 4^{o}C. The RNA pellet was washed with

1 ml of ice cold 80% ethanol, the wash buffer was removed, and the RNA pellet resuspended in 5 µl of nuclease free H_2O (Ambion). RNA is then transferred into 0.5 µl PCR tubes ready for reverse transcription. One of the 12 unique RT primers is assigned to each of the different iCLIP reactions. 1 µl of the chosen RT primers (0.5 pmol/ul) was added to the resuspended RNA along with 1 ul of dNTP mix (Promega). The PCR tube was added to the thermocycler pre-set with the RT programme (Table 1.7). At step 2 of the programme 13 µl of the RT mix wasadded to each reaction before the remainder of the programme is completed.

Step	***Temperature (ºC)***	***Time (Min)***
1	70	5
2	25	*Hold*
3	25	5
4	42	20
5	50	40
6	80	5
7	4	Hold

***Hold* at step 2 until RT mix is added**

Table 2.8: Thermocycler programme for iCLIP reverse transcription reaction

After reverse transcription 1.65 µl of 1 M HEPES-NaOH pH 7.3 (Ambion) was added to each reaction and heated at 98ºC for 20 min. This is required to hydrolyse the original RNA templates. 350 µl of TE buffer pH 8.0 (Ambion) was added to each reaction and an overnight ethanol DNA precipitation reaction was performed as above.

Reverse transcribed cDNA was gel purified using a 6% TBE-Urea gel (Invitrogen). The denaturing urea gel insures cDNA fragments are separated according to size only. 6% TBE-Urea gel was placed into a specifically designated XCell SureLock mini-cell to reduce potential contamination. After the precipitated cDNA pellet was ethanol washed (as previously described) the cDNA pellet was resuspended in 6 µl of 2x TBE-Urea loading buffer (Invitrogen). In addition, 6 µl of 2x TBE-Urea loading buffer was added to 1 µl of RNA century size marker (Invitrogen) before all samples were heated at 80ºC for 5 min to

disrupt any potential secondary structure. Samples were loaded on the 6% TBE-Urea gel and run at 180 V for 40 min in 1x TBE running buffer (Novex). The gel lane in which the RNA marker was run was cut from the gel and stained with 2 µl of SYBR Green II (Life Technologies) in 10 ml 1x TBE running buffer. Marker was visualised using UV transillumination (Figure 2.2 – Step 13). The RNA marker was used as a mask to select cDNA fragments from the TBE-Urea gel; 70-80 nt, Low Band; 80-100 nt, Medium Band; and 100-150 nt, High Band. Sections of the gel were collected using a sterile blade and placed into 2 ml Non-stick, RNase-free Microfuge tubes (Ambion) containing 400 ul of TE buffer. Gel pieces were crushed using a 1 ml syringe plunger in order to aid extraction of cDNA molecules. Mixture was incubated for 1 h at 37^{o}C while shaking at 1100 rpm. Gel fragments were removed by passing the mixture through a Costar SpinX Column (Corning) into which 2 x 1 cm glass pre-filters (Whatman) had been placed, via centrifugation at 13000 rpm for 5 min (Figure 2.2 – Step 14). The aqueous solution containing the size purified cDNA fragments was Phenol/Chloroform extracted and ethanol precipitated over night at -20^{o}C as described before.

2.4.8 Circularisation and re-linearisation of cDNA fragments

The final step of the iCLIP protocol before PCR amplification and sequencing requires the cDNA transcripts to be treated with CircLigase II (Cambio). CircLigase is a ssDNA Ligase which catalyses the ligation of the 5’ and 3’ end of individual cDNA transcripts. This is a critical step in the iCLIP protocol as it results in the barcode region designed with the RT primer to be place immediately upstream of the UV induced crosslink site of the RNA transcript (Figure 2.2 – Step 15). After digesting the circularised transcripts with BamH I restriction enzyme, linear transcripts are then produced with adaptor regions placed at both the 3’ and 5’ end of the transcript enabling PCR amplification (Figure 2.2 – Step 16).

The precipitated DNA from the previous step is spun down and washed as before. The DNA pellet was resuspended in 8 µl of circular ligation buffer, placed into

PCR tubes and incubated at 60ºC for 1 h. Before the circular cDNA transcripts can be digested with BamH I the transcripts need to anneal with a specially designed cut oligo which anneals across the BamH I restriction site located in the RT primer sequence. This is due to the BamH I restriction enzyme recognising dsDNA and not ssDNA. 30 µl of the oligo annealing mix is added to the circular cDNA and places in a thermocycler programmed with the annealing programme (Table 2.9)

Step	***Temperature (ºC)***	***Time (Seconds)***
1	95	120
2	*95 - 25*	20
3	25	Hold

***Step 2* repeats successively with the temperature decreasing 1ºC per cycle until 25ºC is reached**

Table 2.9: Thermocycler programme for cut oligo annealing
After oligo annealing, 2 ul of Fast BamH I (Fermentas) is immediately added and incubated for 30 min at 37ºC followed by 80ºC for 5 min which deactivates BamH I enzyme. The cDNA constructs are then ethanol precipitated over night at -20oC before PCR amplification.

2.4.9 PCR amplification

Predicated DNA was washed as before and resuspended in 21 µl of nuclease free water. Before the library can be amplified for sequencing the optimal number of PCR amplification cycles must be determined. Over amplification of libraries produces secondary PCR products that interfere with next generation sequencing as well as an increase in PCR amplification bias of certain transcripts. Additionally, a minimum cDNA concentration of 10 nM Is required for sequencing. In turn, the cycle number between these two extremes needs to be determined (Figure 2.2 – Step 17).

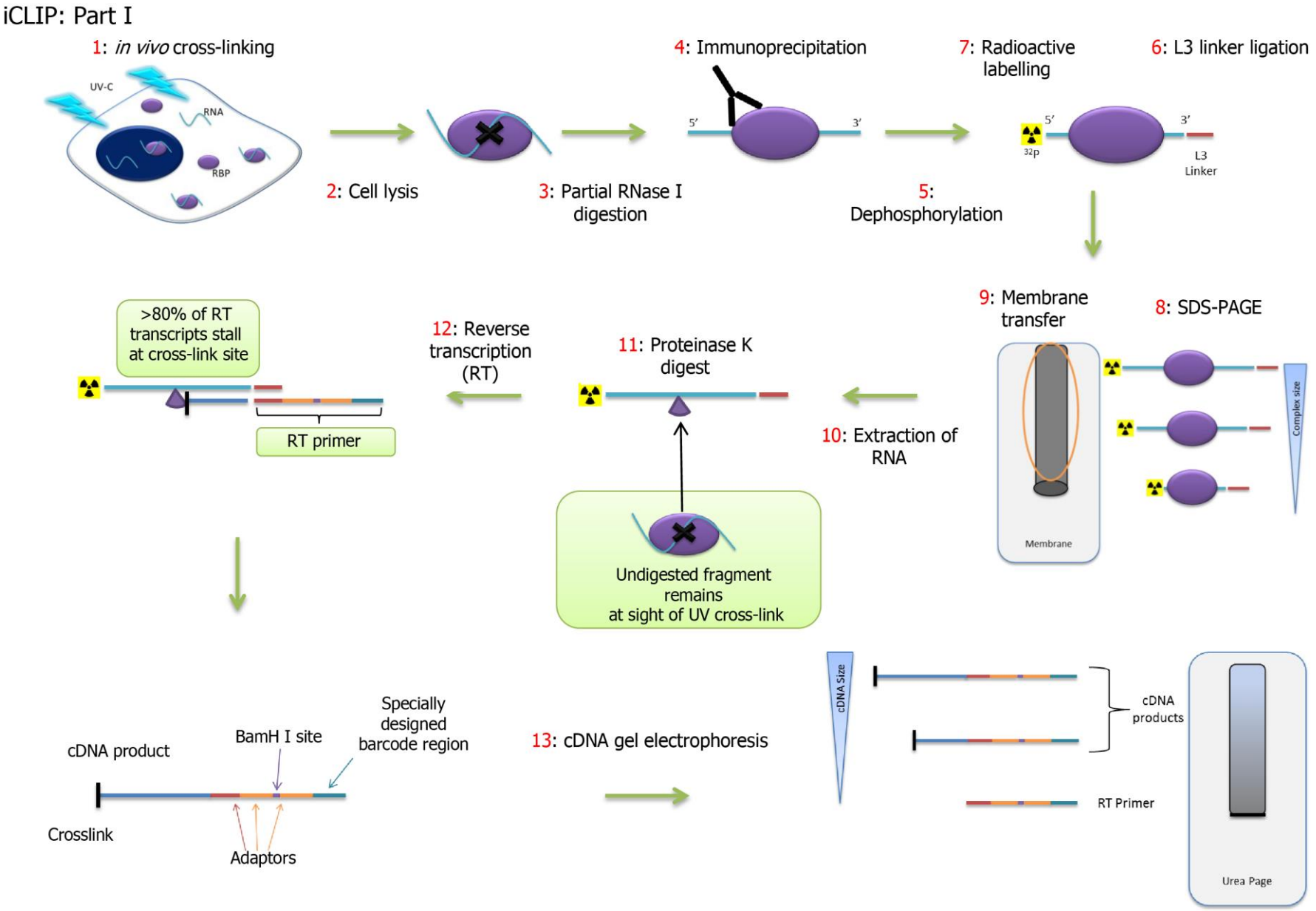
1 µl of the resuspended DNA is added to 9 µl of PCR amplification reaction mix in a PCR tube. The DNA was amplified between 15-20 PCR cycles before being

added to 2 ul of 5x TBE loading buffer (Novex) and loaded on a 6% TBE gel (Invitrogen) and run at 180 V for 30 min buffered in 1x TBE running buffer (Invitrogen). SYBR Green II and UV transillumination is used to visualise amplified cDNA. This process is repeated until over-amplification is observed.

After the optimal number of PCR cycles has been established for the individual iCLIP reactions the sample was amplified for sequencing. 10 µl of the original resuspended DNA was added to 30 µl of the preparative PCR mix in a PCR tube. In this reaction the DNA is 2.25x more concentrated than in the previous preliminary PCR amplification reaction. Therefore, the library is amplified with the cycle number previously determined -1. Amplified DNA was stored at -20ºC. Experiments performed on amplified cDNA libraries was performed in a separate lab due to nucleic acid contaminating the room which can contaminate future iCLIP reactions.

Step	***Temperature (ºC)***	***Time (Seconds)***
1	94	120
2	94	15
3	65	30
4	68	30
5	68	180
6	4	Hold
Cycle steps 2 – 4:	**15 - 20 times**	

Table 2.10: PCR amplification thermocycle for iCLIP cDNA libraries
For optimisation cycle between steps 2 - 4 for 15 – 20 times until optimal cycles number is determined. For preparative PCR cycle between steps 2 - 4 for the optimal number of cycles predetermined -1.


iCLIP: Part I
1: *in vivo* cross-linking
UV-C
RNA
RBP
2: Cell lysis
3: Partial RNase I digestion
4: Immunoprecipitation
5′
3′
5: Dephosphorylation
7: Radioactive labelling
32p
6: L3 linker ligation
L3 Linker
8: SDS-PAGE
9: Membrane transfer
Membrane
Complex size
10: Extraction of RNA
11: Proteinase K digest
Undigested fragment remains at sight of UV cross-link
12: Reverse transcription (RT)
>80% of RT transcripts stall at cross-link site
RT primer
13: cDNA gel electrophoresis
cDNA Size
cDNA products
RT Primer
Urea Page
cDNA product
BamH I site
Specially designed barcode region
Crosslink
Adaptors

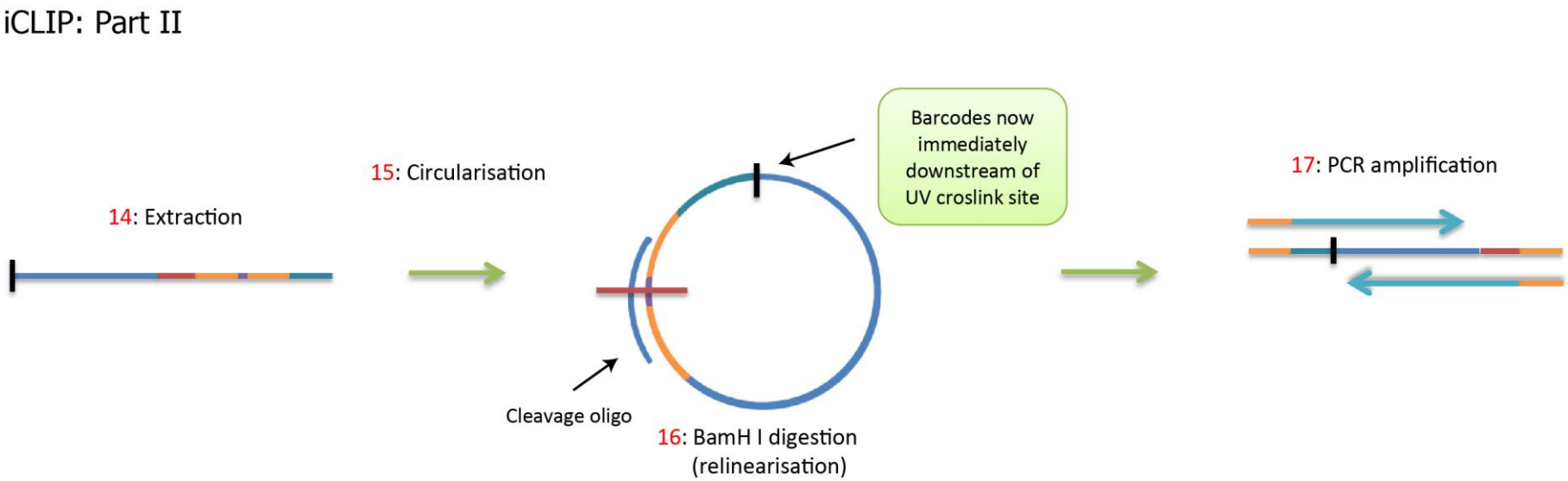


Figure 2.2: Overview of iCLIP protocol
Cells are irradiated with UV-C light. RNA binding proteins (purple ovals) in complex with RNA (blue lines) become covalently bound (black X) at the site of protein-RNA interaction. Crosslinked RNA is partially digested with RNase I enzyme. RBP of interest is immunoprecipitated with a specific antibody (black Y). The RNA is dephosphorylated to allow for later L3 linker ligation. The 5' end is radiolabelled with 32P to enable autoradiograph detection of protein-RNA complexes after SDS-PAGE separation and membrane transfer. Protein-RNA complex are extracted from the nitrocellulose membrane and digested with proteinase K. After digestion a polypeptide fragment remains linked to the RNA transcript (purple triangle). The site of the remaining fragment correlates to the site of UV crosslink. Specially designed reverse transcription primers anneal to the previously ligated L3 linker and prime RT of the RNA fragments to create a cDNA library (dark blue line). Over 80% of the RT reactions stall at the protein-RNA crosslink site due to the presence of the undigested RBP (black vertical line). The RT primers incorporate key elements into the cDNA products; two PCR amplification priming sites (orange lines), a BamH I digestion site (purple line) and a barcode region (turquoise line). The cDNA products are purified by electrophoresis on a urea-PAGE gel. Purified cDNA fragments are circularised by the enzyme CircLigase. Ligation places the protein-RNA crosslink site immediately

upstream of the barcode region. Annealing of a cleavage oligo enabled the ss-cDNA to be digested with BamH I at the incorporated restriction site. The re-linearised cDNA products now have PCR primers at both the 5' and 3' end allowing PCR amplification of libraries for high-throughput sequencing. The 5' primer also serves as a HTS primer. Analysis of sequencing reads enables identification of not just target RNA transcripts, but also RNA recognition motifs due to the orientation of the barcode region and UV crosslink site. Adapted from[116]

2.4.10 qPCR quantification of libraries and next generation sequencing

Prior to next generation sequencing the precise concentration of the amplified cDNA library needed to be determined via qPCR quantification. All iCLIP libraries were quantified using the KAPA Library Quantification Kit KK4824 for Illumina platforms (KAPABiosystems). The KAPA Library Quantification Kit contains P5 and P7 primers which are complementary to the P5 and P3 Illumina adaptors incorporated into the cDNA libraries via the RT primers (Note: P7 sequence resides in the larger P3 adaptor sequence). Serial dilutions of the amplified cDNA library were prepared (1:10, 1:100, 1:1,000, 1:10,000, and 1:100,000) in nuclease free water. qPCR master mix was prepared according to the manufacture's protocol and 16 µl was added to each well of a 96 well qPCR Fast Optical Plate (MicroAmp). 4 µl of the corresponding cDNA dilution was added to the mix. In addition, 4 µl of PCR grade water was used as a negative control. Finally, 4 µl of each of the 6 DNA standards was added to a well containing 16 µl of the master mix. All cDNA dilutions, standards and controls were plated in triplicate.

Reagent	*Volume (µl)*
2x KAPA SYBR FAST qPCR mix + 10x Primer Premix	12.0
50x Low ROX	0.4
PCR grade water	3.6
Total per reaction	16.0

Table 2.11: qPCR reaction mix for KAPA Kit quantification

DNA Standard	*Concentration (pM)*
1	20
2	2
3	0.2
4	0.02
5	0.002
6	0.0002

Table 2.12: Concentration of DNA standards used to generate standard curve for cDNA concentration quantification

The plate was placed into an Applied Biosystems 7700 qPCR machine and quantification programme ran according to KAPA kit protocol.

Step	*Temperature (ºC)*	*Time (Seconds)*
1	95	300
2	95	30
3	60	45
	***Cycle steps 2 – 3:* 30 times**	

Table 2.13: qPCR thermocycle programme for quantification

Results were analysed using the KAPA kit supplies spreadsheet and analysis guidelines were followed according to manufacturer's guidelines.

2.4.11 High throughput sequencing and cDNA mapping

High sensitivity DNA assays were performed on the cDNA libraries prior to sequencing to determine DNA quality. Samples were run on the Bioanalyser 2100 (Agilent) to determine DNA quality and size while Qubit Fluorometric Quantitation (Life Technologies) was performed to determine cDNA concentration. Samples were sequenced on the Illumina Hiseq 2000 by recording single end reads with a read length of 50 nt.

cDNA reads were placed into the iCOUNT pipeline (Curk et al. (2016) iCount: protein-RNA interaction iCLIP data analysis *in preparation*) for genome mapping (human genome h19) and crosslink analysis. Mapped results were viewed in UCSC Genome Browser.

2.5 Nuclear Magnet Resonance (NMR) Spectroscopy

The dissociation constant (K_d) is an important parameter for understanding the function of a physiological interaction between a protein and ligand. To accurately measure dissociation constants, the protein concentration must be in the range of the K_d. High affinity protein-RNA interactions in the nM rage are achieved through multiple RNA binding domains associating with the RNA transcript. However, the binding affinity of the individual RBDs is much weaker and typically falls in the µM to mM range. NMR is uniquely suited to the study of these weaker affinity interactions due to high concentration of sample required because of the insensitivity of the technique. The small energy difference between the higher and lower energy orientations of nuclei when subjected to the magnetic field requires NMR protein samples to be concentrated in the mM to µM range in order to obtain a good signal to noise ratio.

The dissociation constant of a protein-RNA interaction can be determined by recording ^{1}H-^{15}N heteronuclear correlation spectra on a ^{15}N labelled protein during a titration with unlabelled RNA. The titration results in the formation of protein-RNA complexes where the protein is in equilibrium between the free and

bound states. The type of exchange depends on the rate of complex formation. The chemical shifts of nuclei that are involved in RNA recognition experience a change in the microchemical environment and display chemical shift perturbations. The NMR signal from residues that are affected by RNA binding report both on the free and bound state. There are three main exchange regimes that can be observed by NMR and these are dependent on: the relation between the exchange rate of the complex formation (K_{ex}), and the difference in resonance frequency of the nucleus in the free (V_P), and bound states (V_{PL}). The three main exchange regimes are: slow exchange, when K_{ex} is much smaller than $2\pi(V_P-V_{PL})$; fast exchange, when K_{ex} is much larger than $2\pi(V_P-V_{PL})$; and intermediate exchange, when K_{ex} is similar to $2\pi(V_P-V_{PL})$.

The three exchange regimes have characteristic chemical shift perturbation properties (Figure 2.3). In slow exchange two sets of signals are observed, one reporting on the protein in the free state and the second corresponding to the protein in the bound state. In intermediate exchange, the NMR signal of the free protein undergoes line broadening upon addition of the RNA up until more than half of the stoichiometry is reached, from then the linewidth of the signal corresponding to the bound protein state sharpens. In the fast exchange regime only one NMR signal is observed. This signal corresponds to the weighted average of the signals for both the free and the bound protein states. In turn, the signal shifts from the site of the free state towards the site of the bound state as the proportion of bound complexes increases.

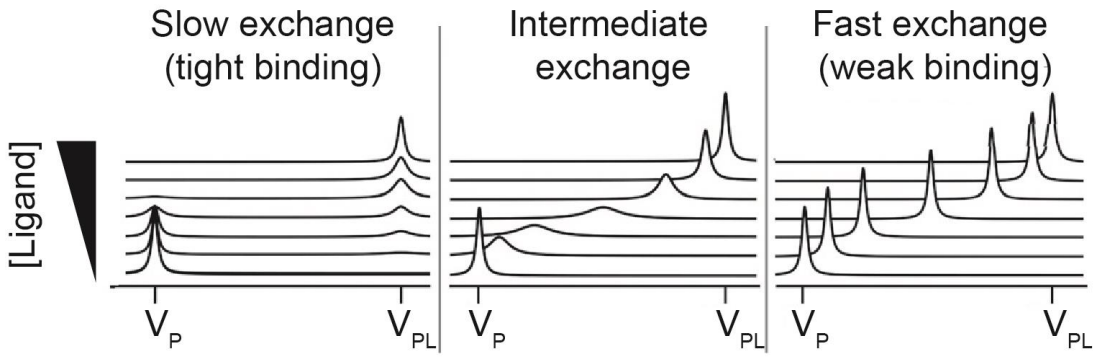


Figure 2.3: Representation of the chemical shift perturbation observed in the three exchange regimes during a protein-ligand titration
Left: Tight protein-ligand binding displays slow exchange due to K_{ex} << 2π(VP-VPL). Peak intensity at the chemical shift of the free protein (V_P) reduces in signal and reappears at the chemical shift position of the bound protein (V_{PL}) as ligand is titrated. Middle: Intermediate exchange results from K_{ex} ~ 2π(VP-VPL). Chemical shift of the free protein (V_P) experiences line broadening up until stoichiometry and then linewidth sharpens towards the position of the chemical shift of the bound complex (V_{PL}). Right: Weak binding displays fast exchange due to K_{ex} >> 2π(VP-VPL). Peak position displays the weighted average of signals for both the free and the bound protein states. This results in the peak shifting from the position of the free protein (V_P) towards the position of the bound complex (V_{PL}) as ligand is titrated. Adapted from[181]

Protein ligand binding that displays a fast exchange regime has a dissociation constant in the 10 μM-10 mM range which corresponds to the K_d of individual RBD-RNA interactions. The chemical shift perturbations observed upon increasing concentrations of RNA can be tracked for individual residues. The weighted average chemical shift ($\Delta\delta av$) of a peak in the ^{1}H and ^{15}N dimension of a HSQC/HMQC spectrum upon binding of RNA can be calculated using Equation 3:

Equation 3: Calculating weighted average chemical shift in the ^{1}H-^{15}N dimensions

$$\Delta\delta_{av} = \sqrt{(\Delta\delta_H)^2 + (\Delta\delta_N/10)^2}$$

Where $\Delta\delta_H$ and $\Delta\delta_N$ are the change of chemical shift for a cross-peak in the ^{1}H and ^{15}N dimensions respectively.

Average chemical shifts of the backbone amides can then be plotted as a function of protein:RNA ratio to produce a binding curve. As the position of this peak is determined by the molar fraction of free to bound protein, the binding curve can be used to determine the dissociation constant using Equation 4.

Equation 4: Calculation of dissociation constant (K_d) of a protein-ligand interaction in fast exchange via NMR

$$\Delta\delta_{av} = \Delta\delta_{max}\frac{(K_d + [L]_o + [P]_o) - \sqrt{(K_d + [L]_O + [P]_O)^2 - (4[P]_O[L]_O)}}{2[P]_o}$$

Where Δδav is the average chemical shift perturbation of a given resonance at a given titration point; Δδmax is the chemical shift perturbation for a given resonance at saturation; $[L]_o$ is total RNA concentration; $[P]_o$ is total protein concentration; and K_d is the dissociation constant.

In this thesis protein-RNA titrations were performed on ^{15}N-labelled protein samples concentrated to 60-100 μM. Proteins were buffered in appropriate protein specific buffer (stated in results). Unlabelled protected RNA oligonucleotides were synthesised and purified by Dharmacon. RNA oligos were deprotected following manufacturer's instructions and resuspended in nuclease free water (Ambion). Concentration was assessed by absorbance at λ_{260} and the RNAs corresponding extension coefficient. RNA oligos were resuspended to an appropriate concentration depending on the required protein:RNA molar ratios required for the particular RNA titration experiment. Titrations used a range of protein:RNA ratios ranging from 1:0.2 to 1:8. ^{1}H-^{15}N SOFAST-HMQC spectra were recorded at each titration point at 25^{o}C on Bruker Avance NMR spectrometers operating at 700 or 800 MHz. Spectra were processed using NMRPipe/NMRDraw and analysed using Sparky. Chemical shift perturbations were manually measured and plotted against protein:RNA ratio to generate a binding isotherm and corresponding K_d values.

2.5.1 Scaffold independent analysis (SIA)

While several *in vivo* methods have been developed to determine the sequence specificity of single-stranded RNA binding domains, such as SELEX, these techniques report high affinity RNA target sequences and cannot efficiently explore the full sequence specificity of RBDs. In turn, suboptimal RNA recognition sequences that maybe biologically relevant are missed. Scaffold independent analysis (SIA) is a method tailored to determine the full nucleobase binding preference of RNA binding domains that bind RNA with low-to-intermediate affinity (K_d 1 µM to 1 mM).[94,95]

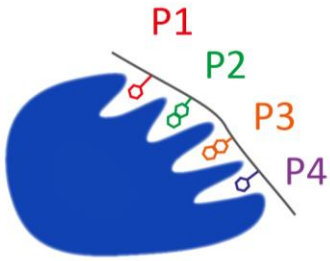
SIA implements an unbiased approach to determine the full sequence specificity of RNA binding domains for each bound RNA position (Figure 2.4A). For each position being scanned four pools of RNA oligos are titrated independently into a ^{15}N labelled protein sample. Each pool contains an equimolar quantity of an ensemble of RNA oligos so that the RNA sequence is randomised in all positions except the position being scanned. In this position the nucleobase remains fixed to either A, C, G or U (Figure 2.4B). ^{1}H-^{15}H correlation spectra are recorded to monitor binding. The induced chemical shift perturbations are then used to ascertain the binding preference for each pool. By relating the affinity of the different pools with the base that was fixed, an order of binding preference for that binding position is obtained.

In a typical SIA analysis, ^{1}H-^{15}N correlational spectra are recorded with a ^{15}N labelled protein in the free form and in the presence of two different molar ratios of each of the randomised pools. The ratio of the RNA chosen depends on the affinity of the protein towards RNA. This parameter is determined by a preliminary titration with a randomised RNA oligo of the same length being used in the SIA study (e.g. NNNNN). Typically, between 10-20 peaks are chosen and the chemical shifts (Δδ) are measured for each of the titrations. When choosing peaks for SIA analysis it is important to consider the following points; peaks need to belong to a backbone amide; shift in the fast exchange regime; and need to be

able to be clearly followed in all four titrations. The raw chemical shifts (Δδ) are processed in a comparative way to obtain SIA scores. The Δδ of each peak measured in the individual titrations is normalised with respect to the largest Δδ. The normalised values are averaged. The final scores ultimately reflect the proteins binding preference for one nucleobase in a fixed position versus another.

As the titrations being compared are performed with different RNA bases, the different chemical structures of the fixed nucleobase could influence the chemical shifts. However, this effect would be localised to the residues recognising the base position being scanned. Selecting between 10-20 peaks samples the whole RNA binding surface and includes chemical shifts occurring from contacts with the randomised RNA bases in the other positions. Normalising and averaging these values minimises the biases due to the chemical nature of the base being fixed.

A



B

	Position 1	Position 2	Position 3	Position 4
A pools	nANNN	nNANN	nNNAN	nNNNA
C pools	nCNNN	nNCNN	nNNCN	nNNNC
G pools	nGNNN	nNGNN	nNNGN	nNNNG
U pools	nUNNN	nNUNN	nNNUN	nNNNU

Figure 2.4: Scheme of scaffold independent analysis probing the sequence specificity of a RDB in four positions
A) Schematic represents a RBD that specifically recognises four RNA bases (P1, P2, P3, and P4). B) Table to represent the four RNA pools used to probe sequence preference in each of the four positions. In each pool the position being scanned contains a fixed RNA base (A, C, G or U) with the remaining positions randomised (N)

The SIA experiments in this thesis were performed by recording ^{1}H-^{15}N SOFAST-HMQC spectra for each titration pool at 25^{o}C on Bruker Avance NMR spectrometers operating at 700 MHz. Samples were prepared in 3mm NMR tubes placed into a NMR tube rack and loaded into a SampleJet to enable automated ^{1}H-^{15}N SOFAST-HMQC recording.

2.5.2 Relaxation experiments

NMR relaxation experiments refer to the timecourse by which the bulk nuclear magnetisation, perturbed by the RF pulse, returns to equilibrium. The rate of return of a spin system to equilibrium is determined by the time dependent magnetic fields experienced by each atomic nucleus. Such experiments relate to the molecular dynamics of a protein given that local magnetic field fluctuations are caused by molecular motions. NMR experiment can be used to extract the longitudinal (spin-lattice) T_1, and transverse magnetisation (spin-spin) T_2, relaxation times. The rate of T_1 relaxation is the decay constant for the recovery of the *z* component of the nuclear spin magnetisation towards the thermal equilibrium. While T_2 corresponds to the decoherence of the transverse nuclear spin magnetisation. It describes the decay constant by which the transverse component of the magnetisation vector exponentially decays towards its equilibrium.

Assuming a globular fold, NMR T_1 and T_2 relaxation experiments can be used to determine overall molecular rotational correlation time: the average time it takes for a molecule to rotate one radian (t_c) using Equation 5.

Equation 5: Correlation time for overall tumbling of a protein can be derived from the ratio of T_1 and T_2

$$t_c = \frac{1}{2\omega_N}\sqrt{\frac{6T_1}{T_2} - 7}$$

Where ω_N is the ^{15}N resonance frequency in Hz.

T_1 and T_2 rate provides information on the molecular tumbling, which is dependent on the shape and size of the molecule. (Figure 2.5) The T_1/T_2 rate provides information about the dynamic behaviour, overall shape, and size of the molecule. Relaxation can also report on the dynamic changes occurring during protein-RNA interactions. T_1 and T_2 can be extracted at the individual residue level. These can be gathered to report on relaxation of the whole domain or structural regions. In this thesis we use relaxation measurements to determine the relationship between two domains in a di-domain construct. Using average rotational correlation values, we can investigate if the domains tumble as independent or fused units. If the domain where to tumble as one unit we would expect a larger t_c than domains that are tumbling individually (As described in Chapter 5.6).

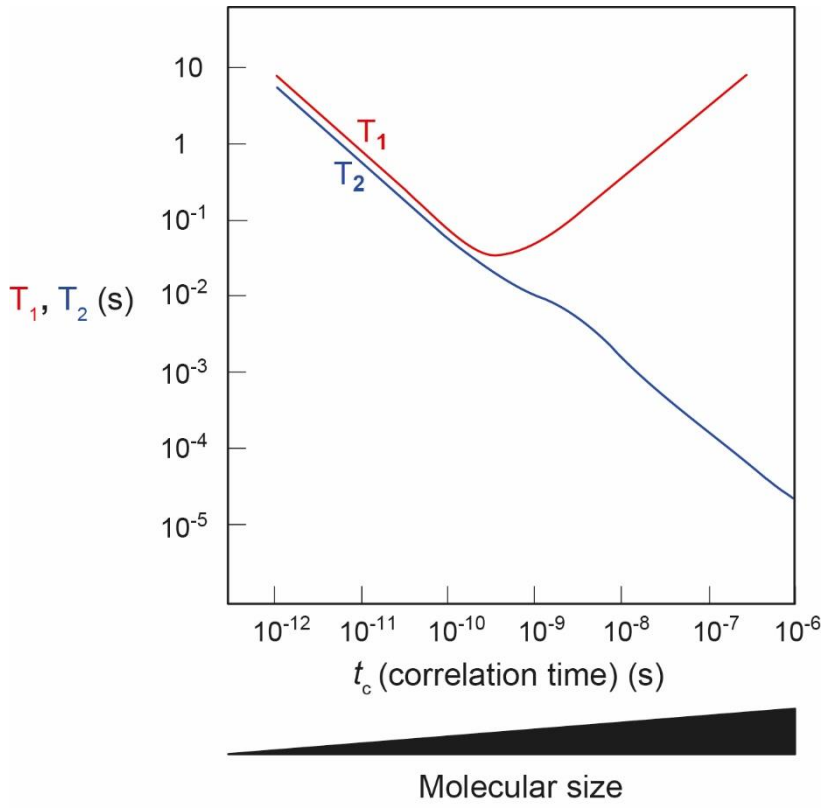


Figure 2.5: Relationship of T_1 and T_2 with respect to correlation time t_c
T_1 values plotted in red and T_2 values in blue. t_c represents the molecular correlation time (average time taken for the molecule to rotate one radian). As a molecule increases in size (black wedge) so does molecular correlation time. Adapted from[182]

Standard relaxation experiments were recorded on ^{15}N-labelled samples to obtain T_1, T_2 values. Experiments were performed on a Bruker Avance NMR spectrometer operating at 800 MHz. T_1 and T_2 values were determined for each residue by fitting an exponential decay of peak volume over the course of the data collected. Residues were excluded where overlap in the signals prevented accurate measurement of peak volume.

2.6 Circular Dichroism

To study the effect of the mutations on secondary structure content we performed Circular Dichroism (CD) on the WT and mutant constructs. CD implements the use of circularised light beams to investigate protein structure. Circularly polarised light is defined as light in which the direction of the electric vector changes with a constant magnitude. A CD machine generates circularly polarised light by producing two beams of plane polarised light differing by 90^{o} in the plane of polarisation. The two beams of light are generated with the same wavelength and magnitude, however, one of the beams is a quarter of a wavelength retarded with respect to the other. Superimposing these two beams of light generates either a left or right handed helix.

When polarised light passes through an optical active solution, the right helical and left helical polarised light are absorbed differently. Proteins in solution display optical activity due to each amino acid (except glycine) containing a chiral centre which acts as a chiral chromophore. This optical activity of proteins in solution can be measured by CD by recording the difference in absorbance of right and left handed circularised light. Absorption of polarised light in a CD machine follows the Beer-Lambert law:

Equation 6

$$A = \varepsilon . C . L$$

A = absorbance. ε = extinction coefficient of protein l = distance of light path and c = molar concentration of protein in solution

As stated, optical density is measured in CD by observing the difference in absorbance of left and right helical light. In turn, we can express the difference as:

Equation 7

$$\Delta A = A_L - A_R = (\varepsilon_L - \varepsilon_R)lc = \Delta \varepsilon . l . c$$

L = left helical and R = right helical light

The difference in absorbance can also be displayed as ellipticity (θ). Ellipticity is related to absorbance by a factor of 32.98:

Equation 8

$$\theta = 32.98\Delta A$$

Molar ellipticity ([θ]) is CD corrected for concentration. In order to study protein structural content from CD absorbance spectra, molar ellipticity must be converted to a normalised value. In order to normalise independent of the polymer length. Mean residue weight used for this purpose, essentially treating the protein as a solution of amino acids. The units of molar ellipticity are historically (deg×cm^2/dmol):

Equation 9

$$[\theta] = 3298A.M/(l.c)$$

M = the mean residue weight

A full CD spectrum ranges from (180-310 nm). This can be split into near and far. The near-UV spectrum reports on the aromatic side chains of tryptophan, tyrosine and phenylalanine residues, in addition to disulphide bonds between cysteine residues. Near-UV spectra are useful for monitoring tertiary structure.

Far-UV spectra report on the peptide bond and reflects secondary structure content. Each secondary structural element possesses a defined absorption pattern. Purely α-helical proteins have minima at ~208 nm and 222 nm and a maximum of ~195 nm. Proteins consisting of only β-sheets have a less pronounced absorption profile with respect to α-helical proteins, but display a minimum of ~215 nm and maximum of ~198 nm. Random coil regions give strong minimum at 195-198 nm. Protein folds often contain a mixture of secondary structural elements and will display far-UV absorption spectra that is a combination of these. Due to the overlapping nature of the individual profiles it is often difficult to deconvolute the exact composition of α-helix, β-sheet or random coil within a protein structure, but informative predictions can be concluded.

The sensitive nature of CD enables the monitoring of minimal structural changes in protein systems on a real time scale. CD can report on secondary structure content, tertiary and occasionally quaternary structures, in addition to following changes during protein unfolding and refolding.

2.6.1 Thermal denaturation

Protein stability can be investigated using either chemical or thermal denaturation. Monitoring thermal denaturation is achieved by monitoring secondary structural content at a fixed wavelength while subjecting the sample to a temperature gradient. The wavelength chosen to monitor the structural changes depends on the secondary structural content of the protein. For proteins containing a high degree of α-helical content 222 nm is commonly used. Additionally, complete CD spectra can be obtained at each temperature interval.

Changes in CD spectrum can be used to determine the thermodynamics of unfolding – van't Hoff enthalpy (ΔH^o) and entropy of unfolding (ΔS^o), the midpoint of the unfolding transition (T_m), and the Gibbs free energy of folding (ΔG^o). For a protein in which thermal denaturation is totally reversible the equilibrium between the folded and unfolded state is determined by the unfolding equilibrium constant (K).

In our investigation I used CD thermal denaturation to determine the T_m unfolding transition point (the temperature at which 50% of the protein is unfolded) of WT and mutant protein constructs. Comparing the T_m of WT and mutant proteins enabled us to evaluate how the incorporated mutations effected the proteins thermal stability.

Thermal unfolding experiments monitored by Circular CD were performed on a Jasco J-815 spectropolarimeter equipped with CDF-426S temperature-control system. Protein samples were diluted to 0.15 mg/ml and placed into a High Precision Quartz Cell (Hellma) with a 1 mm light path length. The solution was heated from 5^{o}C to 95^{o}C at a rate of 2^{o}C per minute and the unfolding of the protein was monitored at either 210 or 220 nm (stated in results).

Chapter 3. iCLIP of FLAG-IMP1 constructs in Flp-In T-REx HeLa cells

3.1 Introduction

Studies exploring the RNA recognition properties of multidomain RNA binding proteins with defined RNA targets *in vitro* identified many proteins that bind RNA transcripts via combinational recognition.[49,92] Characterisation of IMP1 binding to a subset of know RNA targets *in vitro* shows different combinations of KH domain are used to recognise different RNA targets. As previously stated, three well characterised targets of IMP1 are the RNA transcripts of ACTB, MYC, and CD44. Studies of IMP1 binding to the 3' UTR of ACTB have shown that the KH3 and KH4 domain are fundamental for binding, with the KH1 and KH2 domains playing a lesser role in recognition.[83,85,179] However, binding to the CRD region of MYC requires a higher binding contribution of the KH1 and KH2 domains over KH3 and KH4.[83,141] In contrast, equal contributions of all four KH domains is suggested for IMP1 binding to the 3' UTR of CD44 mRNA.[145]

While these studies provide the basis for understanding how IMP1 selects RNA targets *in vitro*, how these mechanisms relate to in-cell RNA recognition is less well understood. Additionally, current data of IMP1 combinatorial RNA recognition is limited to a select few RNA transcripts. The development of high-throughput approaches to study RBP interactions has enabled the investigation of how RBPs recognise multiple RNA transcripts. These studies can be used to determine how RBPs bind RNA targets on a transcriptome-wide level. To date no such study has been performed to understand the individual RNA binding domain contributions in transcriptome-wide RNA recognition. Here I set out an approach to understand how IMP1 utilises its multiple RBDs to select RNA targets *in vivo*.

In our investigation into IMP1 RNA target selection in HeLa cells, I decided to implement the iCLIP technique instead of PAR-CLIP. The rationale for this was that in addition to the loss of RNA transcripts in the PAR-CLIP method, pre-incubation of cells with either 4US or 4GS has been shown to be toxic. In the case of HeLa cells, incubation with 100 µM 4-thiouridine or 25 µM 6-thiogaianosie resulted in cell death.[116] A modified PAR-iCLIP study also showed that the incorporation of photoactive nucleotides and crosslinking in the UV-A range resulted in no increase in crosslink efficiency compared to standard UV-C crosslinking.[116]

3.2 Previous high-throughput RNA binding studies of IMP1

Previous studies on the IMP family have identified specific roles for IMPs in controlling the localisation, translation and turnover of specific mRNA targets. However, identification of a comprehensive list of RNA target transcripts is still outstanding. Additionally, there is no mechanistic understanding of how IMP1 implements its multiple RNA binding domains to select for RNA targets *in vivo*.

To date, two studies have focused on identifying human IMP1 RNA targets on a transcriptome-wide level within a cellular system, and one iCLIP study performed on the *Drosophila* IMP1 homologue dimp.[183] The first was a PAR-CLIP study performed on FLAG-tagged IMP proteins overexpressed in HEK293 cells. This study identified more than 1,000 target mRNAs for IMP1 and suggested a putative RNA binding motif of CAUH (H = A, U or C) for all three members of the IMP family.[110] This suggested that formation of specific IMP-RNA complexes would be defined only by the spacing of this shared binding motif within the target RNA transcripts.

A more recent eCLIP (modified iCLIP) study, aimed specifically at identifying mRNA targets of endogenous IMP proteins during the process of neuronal differentiation was performed in H9-derived human neuronal stem cells (H9 hESC).[144] In this study a family of integrin mRNAs were identified as novel IMP1

targets, in addition to mRNAs encoding for cell adhesion and apoptotic proteins. However, this investigation also included a parallel *in vitro* RNA Bind-N-Seq study which identified RNA binding motifs for the three IMP proteins. The results of the Bind-N-Seq identified enriched Kmers which resemble RNA recognition motifs that have been identified for the IMP proteins in other *in vitro* studies. However, here the author also highlights the issue that RNA recognition motifs identified *in vitro* do not correlate well with binding motifs identified in current CLIP studies performed on the IMP family.[144]

The two previous IMP1 CLIP studies identified enrichment of pentanucleotide binding sequences containing CA di-nucleotides. These studies were performed on full-length IMP1 proteins containing six RNA binding domains potentially capable of recognising RNA. This CA enrichment displays the binding preference of the KH3 domain. However, as the RNA sequence specificity of the KH4 domain has now been characterised, there was not an observed enrichment of the pentanucleotides (CGGAC) displaying KH4 recognition sequence. In addition, the sequence specificities of the KH1 and KH2 domains remain unknown, and so we cannot conclude if this CA enrichment reflects the binding preferences of KH1 and KH2.

3.3 Aims

The four KH domains of IMP1 have been validated as recognising RNA, with the consensus sequences for KH3 and KH4 being characterised in relation to the ACTB zipcode RNA target. In turn, we know that at least these two KH domains recognise distinct RNA motifs with high affinity in a specific manner. However, this specificity has not yet been observed in previous transcriptome-wide binding studies performed on the IMP1 protein. In addition, IMP1 combinatorial recognition of *in vivo* targets is unknown.

Our goal was to develop a system in which I can report IMP1 in-cell RNA target selection at the individual KH domain level. I planned to achieve this by

performing iCLIP on a series of IMP1 mutant proteins where individual KH domains were mutated to inhibit RNA binding. By performing a comparative analysis of the RNA targets identified in the iCLIP study for each mutant protein, we can understand the contribution each KH domain plays in target recognition on a transcriptome-wide level. From this we can begin to build a mechanistic model of IMP1 *in vivo* RNA target selection. An overview of the experimental approach to achieve this is set out in Figure 3.1.

Aims of this chapter were to:

- Generate and validate a cellular system in which iCLIP can be performed on mutant IMP1 proteins;
- Generate high quality iCLIP libraries for each mutant IMP1 protein;
- Identify differences in RNA binding when individual KH domains can no longer recognise RNA;
- Validate if an observable difference in in-cell RNA binding can be seen with a single KH domain knock out mutation;
- Begin to build a binding mechanism for IMP1 in-cell RNA target selection.

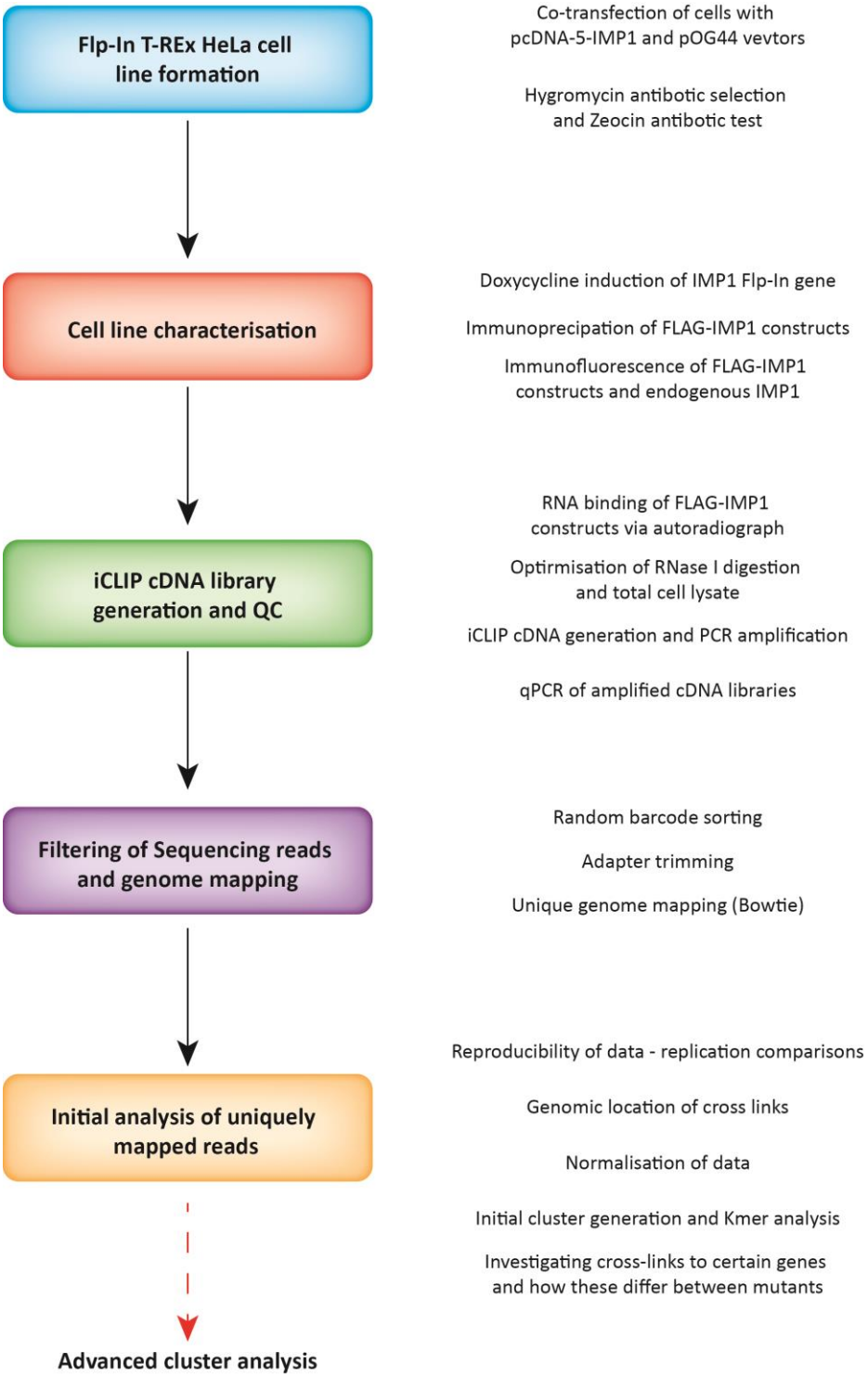


Figure 3.1: Flowchart of the experimental strategy to understand IMP1 *in vivo* RNA selection at the individual domain level

3.4 Mutating the conserved GxxG motif in the KH domain variable loop to GDDG abolishes RNA binding without major structural disruption

Previous studies described how mutating the conserved GxxG motif in the variable loop of KH domains to GDDG abolishes RNA binding without changing the structure of the KH domain.[82] The GxxG motif in the variable loop between the α1- and α2- helices interacts with the first two nucleotides of the RNA that is recognised by the hydrophobic groove of the KH domain. The negatively charged phosphate backbone of the nucleobases interacts with the GxxG motif via electrostatic interactions. Inserting a double negative charge in the GxxG loop in the form of two aspartic acid residues inhibits this interaction and thus prevents the KH domain from binding RNA.

These mutations have previously been well characterised in the KH3 and KH4 domains of IMP1.[82,180] Another member of our research group studied the effects of introducing the same mutations into the KH1 and KH2 domains of IMP1. Their results showed these mutations do not disrupt the structure (Figure 3.2) or the stability of the domains. RNA titrations also confirmed the mutant domains are unable to recognise their RNA targets (data unpublished). In order to determine the contribution of individual KH domain binding in overall IMP1 RNA target recognition I cloned IMP1 mutants in which KH domains were mutated individually to GDDG to abolish RNA binding.

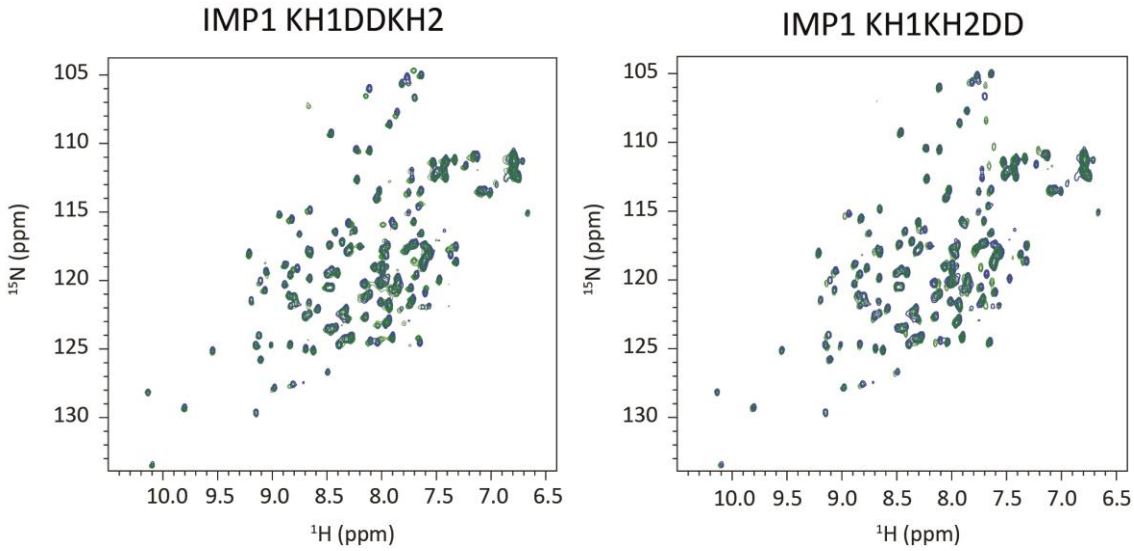


Figure 3.2: GDDG mutations in KH1 and KH2 domain do not cause major structural disruption
^{1}H-^{15}N SOFAST-HMQC overlay of WT KH1KH2 construct in Blue overlaid with KH1DDKH2 construct (Left) and KH1KH2DD construct (Right) in Green.

3.5 Flp-In T-Rex HeLa cells as a model system for investigating IMP1 RNA target selection on a transcriptome-wide level

I decided to use the Flp-In T-Rex HeLa cell lines as our cell system to study in-cell RNA binding of IMP1. These cells were chosen based on three main criteria. Firstly, these cells allow us to recombinantly insert our gene of interest into the genome of the HeLa cells at a single locus. Once inserted into the genome we are able to induce expression of mutant IMP1 proteins with the addition of doxycycline to the cell media. Previous studies using the same HeLa cell system were able to tune expression of their inserted gene so that expression was equal to the expression of the endogenous protein with-in the HeLa cells.[184–186] Secondly, HeLa cells are a cervical cancer-derived immortal cell line. IMP1 is a oncofetal protein, playing roles in neuronal differentiation during embryonic development, while increased expression in cancer cells is linked to a tumours ability to undergo metastasis.[139,143,187] HeLa cells are an invasive cell line in which IMP1 is expressed at detectable levels. The use of the HeLa cell systems enables us to study how IMP1 recognises RNA targets on a transcriptome-wide level, but in a cell system in which cell metabolism has been altered to more resemble a

cancer cell environment in which IMP1 expression has undergone upregulation. Finally, as we planned to use the iCLIP protocol to selectively immunoprecipitate FLAG-IMP1 RNA complexes, HeLa cells provide a cell system that is easy to culture on a large scale so that we would be able to extract enough RNA for iCLIP processing and high throughput sequencing.[111,188]

Our investigation into IMP1 RNA binding is based on the aim of understanding how individual KH domains contribute to RNA recognition. This is the first study implementing an iCLIP study to investigate such a principle on a transcriptome-wide level. Although HeLa cells have limitations, especially regarding their genetic variation as a result of prolonged culturing, they provide an initial platform in which transcriptome-wide RNA protein binding can be studied, in addition to the iCLIP protocol initially being optimised in this cell type.[111,189,190]

3.6 N-terminal FLAG-IMP1 is more stable then C-terminal-FLAG IMP1 in HeLa cells

To selectively immunoprecipitate our IMP1 mutants I inserted a stable tag into our constructs. This approach has been previously used in other CLIP studies. The previous PAR-CLIP study performed on the IMP family used FLAG-tagged proteins transiently expressed in HEK293 cells.[110] Due to the high affinity monoclonal antibodies available for FLAG immunoprecipitation experiments, and this tag previously being used in other CLIP studies,[110,191] I decided to FLAG tag our IMP1 constructs.

After deciding on the use of a FLAG protein tag, I investigated the effects of incorporating that tag at either the N- or C- terminus of IMP1 WT protein (Figure 3.3B). WT IMP1 constructs were cloned into an N- or C-terminal FLAG tag containing pCDNA5 vector. These vectors contain the doxycycline inducible promoter for in the Flp-In T-REx HeLa cell system. The resultant pCDNA5 vectors were transiently transfected into HeLa cells, and a range of doxycycline concentrations was used to induce FLAG-IMP1 expression. Western Blot

analysis of the cell lysate showed C-terminal FLAG-IMP1 is not stable in HeLa cells, compared to the N-terminal construct, as truncated species of the tagged protein were detected (Figure 3.3A). As expected increasing doxycycline concentrations resulted in an increased expression of the IMP1 constructs (Figure 3.3A). Based on these findings I decided to proceed with generating stable transfected cell lines that containing IMP1 constructs with N- terminal FLAG tags.

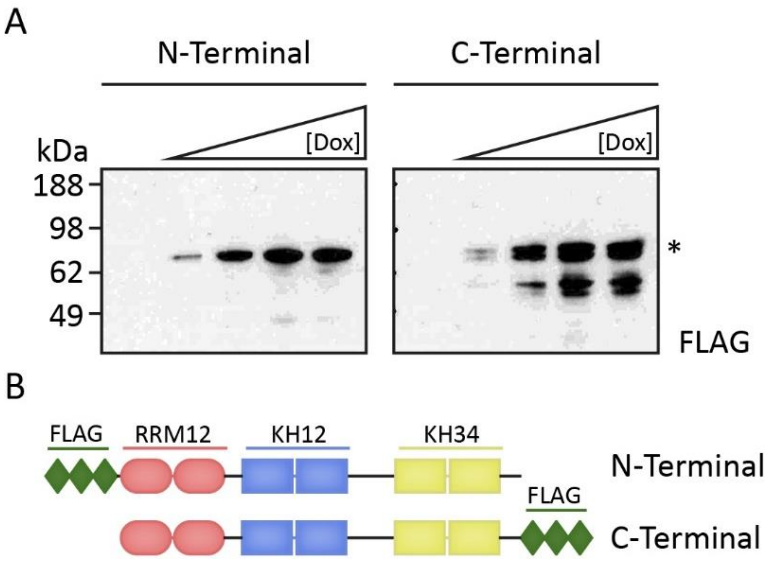


Figure 3.3: Dose-dependent doxycycline-induced transient expression of IMP1 in HeLa cell lines with either N- or C- terminal FLAG tag
A) Doxycycline-inducible pcDNA5 vector containing either N-Terminal FLAG or C-terminal FLAG IMP1 was transfected into Flp-In T-REx-HeLa cells and left untreated or induced with increasing concentrations of doxycycline ([Dox]; 1 ng/ml, 10 ng/ml, 100 ng/ml, 1000 ng/ml) for 24 h. Whole cell lysates were prepared in RISC buffer and total protein lysate concentration determined by Bradford assay kit and subjected to SDS-PAGE and Western Blot analysis. Each lane contains 12.5 µg of total protein lysate. Blots were then probed with mouse anti-FLAG antibody. Full-length FLAG-IMP1 is indicated by * B) Schematic of protein constructs expressed highlighting FLAG tag location.

3.7 Flp-In T-REx-HeLa cells express FLAG-IMP1 constructs at levels equal to endogenous IMP1 when induced with doxycycline

Two antibodies were used to confirm the expression of the FLAG-IMP1 constructs upon doxycycline induction. A mouse monoclonal anti-FLAG antibody was used to determine the expression level of the inserted FLAG-IMP1 genes. A second rabbit polyclonal antibody, raised against a C-terminal peptide of IMP1, was used to evaluate the expression level of the FLAG-IMP1 constructs in relation to the endogenous IMP1 expression (Figure 3.4C).

Other groups using the same HeLa cell system previously optimised gene expression by adjusting the doxycycline concentration used for induction.[184–186] Based on our initial optimisation experiments, we were able to induce FLAG-IMP1 expression to match that of the endogenous IMP1 (Figure 3.4B). We also detected no observable expression of FLAG-IMP1 constructs when cells were not treated with doxycycline (Figure 3.4B).

These findings validated a system which enables us to specifically immunoprecipitate FLAG-IMP1 mutant constructs. We took the decision not to knock out / knock down endogenous IMP1 expression by either CRISPR or siRNA. As our mutant IMP1 constructs are FLAG-tagged I can use the specific FLAG antibody to immunoprecipitate only the constructs we are investigating. Maintaining the endogenous IMP1 prevents further disruption to the HeLa cell RNA metabolism. Also, physiological roles of the endogenous IMP1 would be sustained. By generating a system in with FLAG-IMP1 constructs mimic the expression level of endogenous IMP1, we hope to capture RNA transcripts in our iCLIP study that are also recognised by endogenous IMP1.

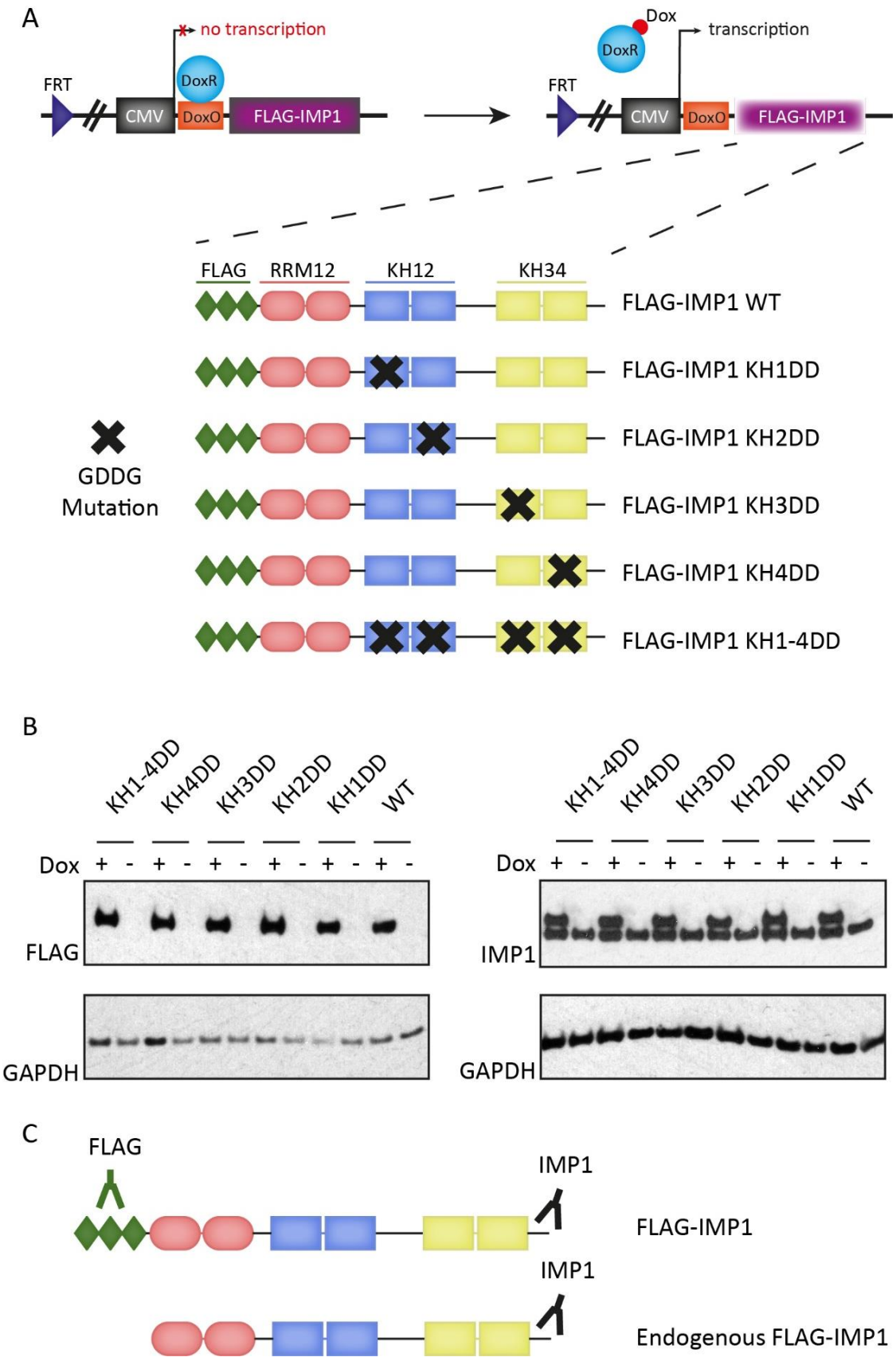

A
no transcription
DoxR
FRT
CMV
DoxO
FLAG-IMP1
Dox
transcription
FLAG
RRM12
KH12
KH34
FLAG-IMP1 WT
FLAG-IMP1 KH1DD
FLAG-IMP1 KH2DD
FLAG-IMP1 KH3DD
FLAG-IMP1 KH4DD
FLAG-IMP1 KH1-4DD
GDDG
Mutation
B
KH1-4DD
KH4DD
KH3DD
KH2DD
KH1DD
WT
Dox
FLAG
IMP1
GAPDH
C
FLAG
IMP1
FLAG-IMP1
Endogenous FLAG-IMP1

Figure 3.4: Characterisation of Flp-In T-REx-HeLa cells expressing a single copy of FLAG-IMP1 WT or KH domain mutations
A) Schematic representation of site-directed recombinase-based Flp-In T-REx-HeLa cell system used to generate stable cell lines. Either FLAG-IMP1 WT or FLAG-IMP1 KHDD mutation genes are integrated into the HeLa genome at a common Flp Recognition Target (FRT) site. A CMV promoter drives the expression of the transgenes through an inducible doxycycline (Dox) repressor. The repressor (DoxR) binds to the Dox operator (DoxO), repressing transcription. Transcription is induced through the addition of doxycycline which binds to DxoR inducing a conformational change resulting in the releases of DoxR. Expansion of the FLAG-IMP1 transgene depicts the proteins expressed in the six different stably transfected cell lines generated. All constructs were N-terminally FLAG-tagged and included WT IMP1, KH1DD, KH2DD, KH3DD, KH4DD and KH1-4DD. X indicates GxxG-GDDG mutation of that KH domain. B) Expression of FLAG-IMP1 proteins with (+) or without (-) doxycycline induction. Induction was performed by incubating cells in media containing 100 ng/µl doxycycline for 24 h. Bradford assay kit was used to determine total protein concentration of cell lysate and 12.5 µg was loaded for each lane. Blots were probed with either mouse anti-FLAG or rabbit anti-IMP1 primary antibodies, and GAPDH used as a loading control. C) Schematic representation of primary FLAG and IMP1 antibody epitope location in FLAG-IMP1 or endogenous IMP1 constructs.

3.8 FLAG-IMP1 does not dimerise with endogenous IMP1 in HeLa cell system

An additional mechanism by which RBPs expand the repertoire of RNAs they can recognise is via association with additional proteins. The simplest example of this is dimerisation, which can take the form of either homodimers with the same RBP or a heterodimer with other RBPs of the same family.[49,76,192] Currently available data as to whether IMP1 can form such dimers are conflicting. Previous *in vitro* work on the IMP1 and IMP3 proteins within our group has not shown these proteins to dimerise, whereas some groups have suggested IMP1 can both homo- and heterodimerise with other IMP family members.[85,154]

As endogenous IMP proteins dimerising with our FLAG-IMP1 mutants could potentially alter the RNA binding of our deficient FLAG-IMP1 KHDD constructs I investigated if our FLAG-IMP1 proteins formed dimers in our HeLa cell system. A FLAG immunoprecipitation was performed on FLAG-IMP1 WT and FLAG-IMP1 KH1DD cell lysate. To account for dimerisation that is RNA dependent I also

immunoprecipitated FLAG-IMP1 WT proteins from cells that had been UV crosslinked. FLAG-IMP1 complexes dimerising with endogenous IMP proteins upon RNA binding would in turn be covalently linked to the RNA molecule and detected as dimers in the Western Blot analysis.

I exploited the different epitopes recognised by the mouse anti-FLAG and rabbit anti-IMP1 antibody to determine if endogenous IMP1 co-immunopurified with FLAG-IMP1 proteins (Figure 3.5B). Our mouse anti-FLAG immunoprecipitation followed by rabbit anti-IMP1 Western Blot analysis detected no dimerisation with endogenous IMP1 (Figure 3.5A). I was able to achieve 100% antibody clearing of FLAG-IMP1 constructs (by comparing cell lysate pre and post immunoprecipitation). I also did not see any detectable reduction in the amount of endogenous IMP1 between pre and post immunoprecipitation lysate (Figure 3.5A). This confirmed our conclusion that endogenous IMP1 does not dimerise with FLAG-IMP1 constructs.

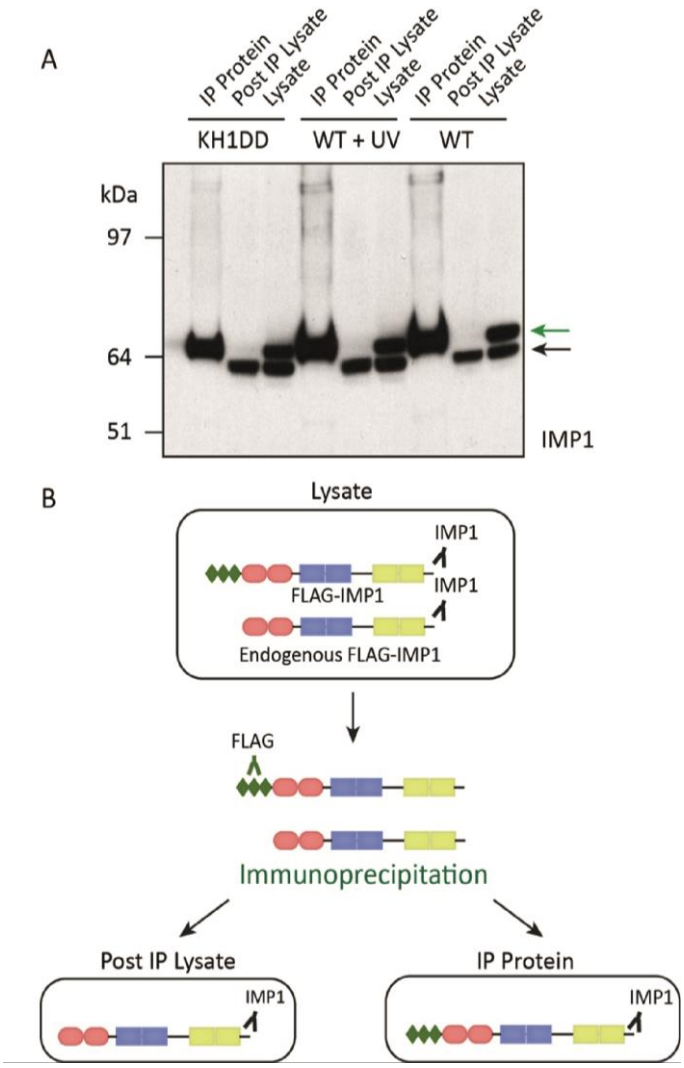


Figure 3.5: Immunoprecipitation of FLAG-IMP1 constructs, with and without UV crosslinking

A) Rabbit anti-IMP1 Western Blot analysis of FLAG immunoprecipitation. Flp-In T-Rex HeLa cells expressing FLAG-IMP1 WT and FLAG-IMP1 KH1DD were incubated with 100 ng/ml doxycycline for 24 h. Cells were either subjected to UV mediated crosslinking or directly lysed in iCLIP lysis buffer. Total protein lysate concentration was then calculated using Bradford assay kit. Lysates were then diluted to 5 mg/ml of total protein and 1 ml collected for immunoprecipitation. 10 µg of mouse anti-FLAG antibody was incubated with 100 µl of protein-G dynabeads and cell lysate for 1 h at 4^{o}C. Lysate post IP was then collected and IP proteins eluted off dynabeads by heating after a series of washes in low salt iCLIP buffer. Green arrow depicts FLAG-IMP1 with black arrow showing endogenous IMP1. B) Schematic representation of the protein components displayed in (A) and the antibodies recognising epitopes in either IP or Western Blot stage. Black boxes represent fractions that were collected for SDS-PAGE and Western Blot analysis. Blot was then probed with rabbit anti-IMP1 antibody as demonstrated in black boxes.

3.9 FLAG-IMP1 proteins have a diffused cytoplasmic distribution consistent with endogenous IMP1 within HeLa cells

To generate a system in which FLAG-IMP1 constructs best represent endogenous IMP1 proteins within our HeLa cell system, I investigated the cellular localisation of both endogenous IMP1 proteins and FLAG-IMP1 constructs. Studies on IMP1 cellular localisation determined that all four KH domains are required for localisation.[83,193] As our FLAG-IMP1 KHDD mutant constructs contain at least one KH domain that is incapable of binding RNA, I investigated if altering the RNA recognition properties of IMP1 drastically changed cellular localisation.

FLAG-IMP1 WT HeLa cells were either left untreated or induced with 100 ng/ml doxycycline. All four FLAG-IMP1 KHDD cell lines were also induced with 100 ng/ml doxycycline for 24 h. Cells were then fixed and stained for mitochondria, nucleus, FLAG and IMP1. Results showed minimal basal expression of FLAG-IMP1 WT when not induced with doxycycline, this is in agreement with our Western Blot analysis of non-induced cells (Figure 3.4). Additionally, endogenous IMP1 localisation can be observed in these cells where FLAG-IMP1 expression is negligible. For endogenous IMP1 and FLAG-IMP1 constructs I observed a diffuse cytoplasmic distribution with both endogenous IMP1 and FLAG constructs being absent in the nuclear compartment. I also observed no enrichment within mitochondria. Comparing localisation of endogenous IMP1 and FLAG-IMP1 constructs revealed similar localisation patterns between all FLAG-IMP1 constructs (Figure 3.6).

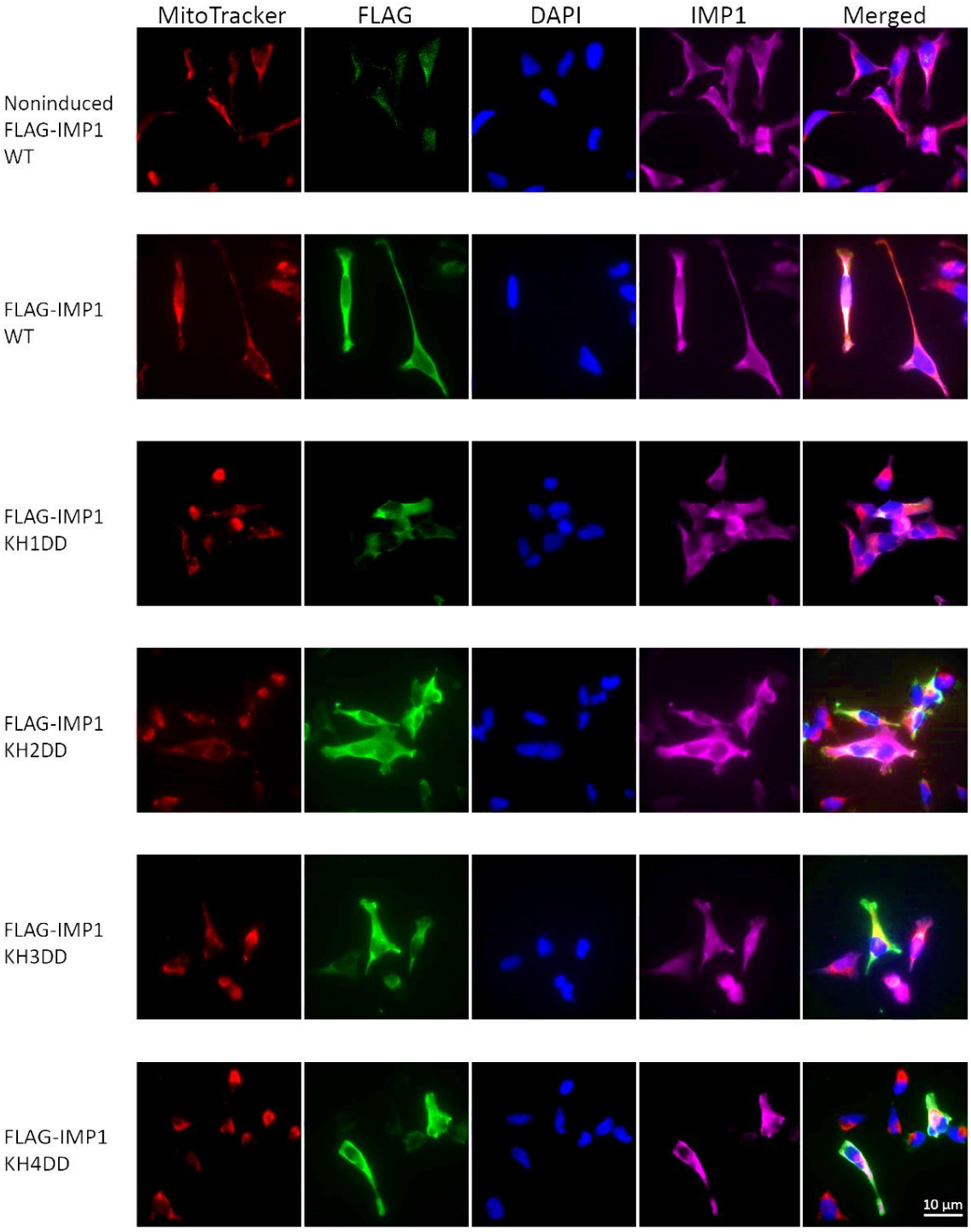


Figure 3.6: Cellular localisation of endogenous and FLAG-IMP1 constructs constructs

Stably transfected HeLa cell lines were either induced with 100 ng/ml doxycycline for 24 h to induce construct expression or left untreated. Cells were then fixed with paraformaldehyde and immunostained. Cell nuclei were visualised with DAPI and mitochondria with MitoTracker™ Red. Additionally, cells were co-stained for FLAG-IMP1 using mouse anti-FLAG and Alexa488 goat anti-mouse,

and IMP1 being detected with rabbit anti-IMP1 and Alexa647-conjugated goat anti-rabbit antibodies. Cells images were takes on a widefield Olympus IX83 fluorescence microscope using a 60x objective and processed in Fiji ImageJ. Merged channel images show colocaliation of endogenous and FLAG-IMP1. Scale bar represents 10 μM.

3.10 Mutant FLAG-IMP1 KHDD constructs have reduced in-cell RNA binding affinity compared to WT FLAG-IMP1

It has previously been reported that the four KH domains of IMP1 have different contributions to overall RNA affinity for specific targets. One research group investigated the RNA binding affinity effects that the insertion of a RNA binding knock out mutation into a single KH domain of the four KH domain IMP1 construct had. The study only investigated the effects on the proteins ability to pull down a select group of RNAs. For the RNAs it was observed that constructs in which KH3 could no longer recognise RNA had the most reduced RNA affinity followed by KH4 and then KH2, with KH1 showing minimal effect.[83]

To determine the effects of the incorporation of a GDDG mutation into a single KH domain on overall RNA affinity I performed UV mediated crosslink immunoprecipitation experiments on cell lysate collected from WT and mutant FLAG-IMP1 proteins. I also included a total RNA binding knock out control in which KH domains 1-4 were mutated with the GDDG mutation. The FLAG-IMP1 RNA complexes were FLAG immunopurified and subjected to partial RNase I digestion and ^{32}P radiolabelling. Protein-RNA complexes were visualised by autoradiography.

Figure 3.7 shows that for all single KHDD mutations I observed a reduction in RNA affinity compared to FLAG-IMP1 WT. For the total RNA knock out control I still observed signal but at an extremely attenuated level. Considering this construct contains two WT RRM domains, the RNA being pulled down could originate from interactions with these domains. I also observe a similar pattern to the previous study in which KH3DD constructs co-immunoprecipitate with less

RNA molecules. However, the reduction in signal is only marginal. These findings also confirm that endogenous IMP proteins do not dimerise with FLAG-IMP1 constructs (Figure 3.5). Such a dimerisation event would rescue RNA binding by associating with a fully functional IMP protein.

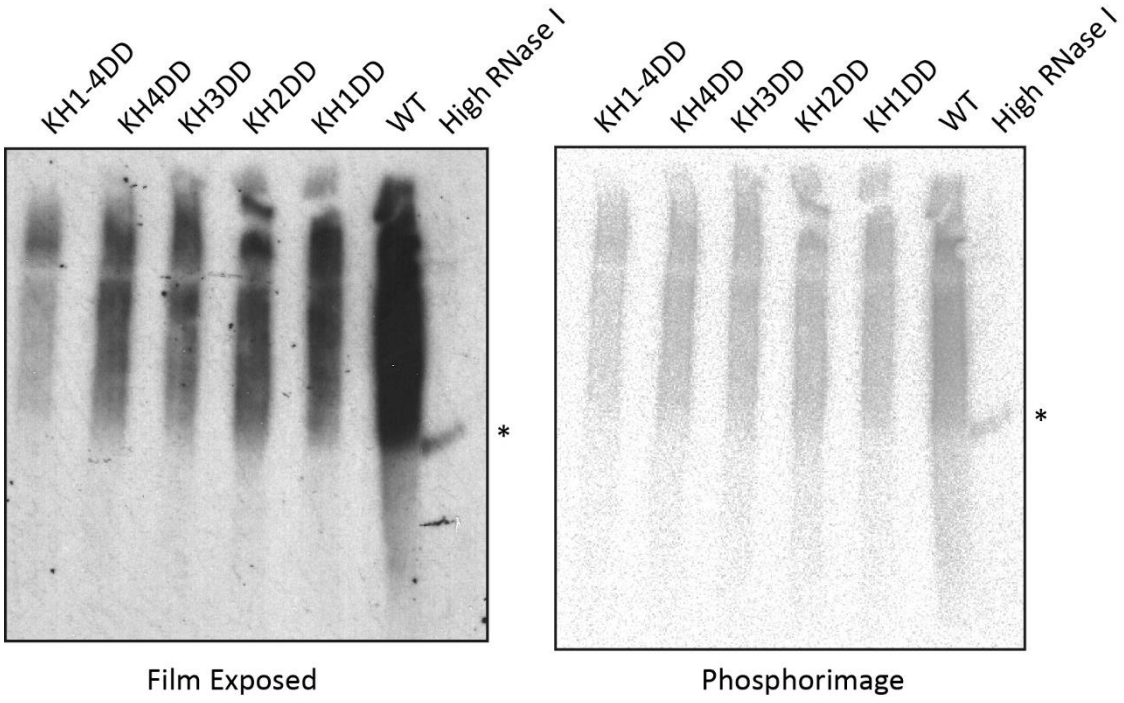


Figure 3.7: Autoradiograph of ^{32}P labelled FLAG-IMP1 RNA complexes comparing WT and mutant in-cell RNA binding
Flp-In TRex HeLa cell lines expressing FLAG-IMP1 constructs were induced with 100 ng/ml doxycycline for 24h in cell incubator. For each cell line ~24 x 10^6 cells were subjected to UV mediated crosslinking and lysed in iCLIP lysis buffer. Bradford assay kit was used to determine total protein lysate concentration and lysates were diluted to a final concentration of 3 mg/ml in 2 ml of lysis buffer. Lysates were then treated with 1/500 dilution of RNase I, except high RNase I control at 1/10 dilution, for 3 minutes at 37^{o}C with 125 rpm shaking. Lysates were then clarified via centrifugation and incubated with protein-G dynabeads pre-bound to mouse anti-FLAG antibody for 1 h at 4^{o}C. Dynabeads were then washed with high salt iCLIP wash buffer and iCLIP PNK buffer before PNK enzyme mediated ^{32}P RNA labelling. Radiolabelled protein-RNA complexes were eluted from dynabeads by heating at 80^{o}C in SDS-PAGE loading buffer. Eluted complexes were then subjected to SDS-PAGE analysis and membrane transfer. A) Membrane exposure to film in amplified cassette at -80^{o}C for 2 hs. B) Exposure to phosphorimager screen for 1 h at RT and read by GE Healthcare Typhoon FLA 7000 phosphorimager. High RNase I control lane represents the migration of FLAG-IMP1 proteins where RNA is digested to fragments that do not contribute to protein molecular weight (*).

3.11 Optimisation of cell lysate and RNase I digestion for iCLIP library generation

An important step in iCLIP cDNA library generation is the optimisation of the partial RNase I digestion (Figure 2.2 Section 8).[116,194,195] I tested a range of RNase I concentrations on FLAG-IMP1 WT RNA complexes pulled down from either ~8 million or ~24 million UV crosslinked HeLa cells. A high RNase I control is also included to determine antibody specificity. A dilution of 1:10 RNase I was used in the high RNase control. Here the RNA fragments covalently bound to FLAG-IMP1 have been digested to a few nucleobases. Accordingly, the migration of the construct in the SDS-PAGE gel is determined by the molecular weight of the protein only. As expected from our previous Western Blots, in these control lanes I see migration of a single band around ~64 kDa. This provided evidence that our FLAG IP pulls down only FLAG-IMP1 proteins in a monomeric form (Figure 3.8).

Ligation of the 3' adaptor to the digested RNA molecules is also a critical step in the protocol (Figure 2.2 Section 6), as only RNA transcripts containing the 3' adaptor can be reverse transcribed in the later stages (Figure 2.2 Section 12). To test the efficiency of our ligation reaction I ligated 3' adaptors to samples that had been treated with high RNase I. Incorporation of the 3' adaptor resulted in an upwards shift in molecular weight. Comparing this lane with the high RNase I treated lane I can conclude successful ligation of the 3' adaptor sequence (Figure 3.8).

To determine the optimal concentration of RNase I I used 1:500, 1:5,000 and 1:40,000 dilutions for IPs from 8 million HeLa cells and 1:500 and 1:5,000 for IPs from 24 million HeLa cells. In general, I observed a weak signal from IPs that were performed on less cells, due to the reduced input of protein-RNA complexes. For the IPs with high number of cells I observed a good signal and optimal digestion at a dilution of 1:500. This concentration resulted in a gradual increase in RNA fragment length (50-300 nt), and the signal started just above the

molecular weight of the protein. The 1:5,000 dilution under-digests the RNA fragments as I observed a reduction in signal and the presence of higher molecular weight complexes (Figure 3.8). The reduced signal is a result of the protein-RNA fragments being too large to migrate through the SDS-PAGE gel. The autoradiograph film was then used as a template to cut the required section from the membrane (depicted via red box).

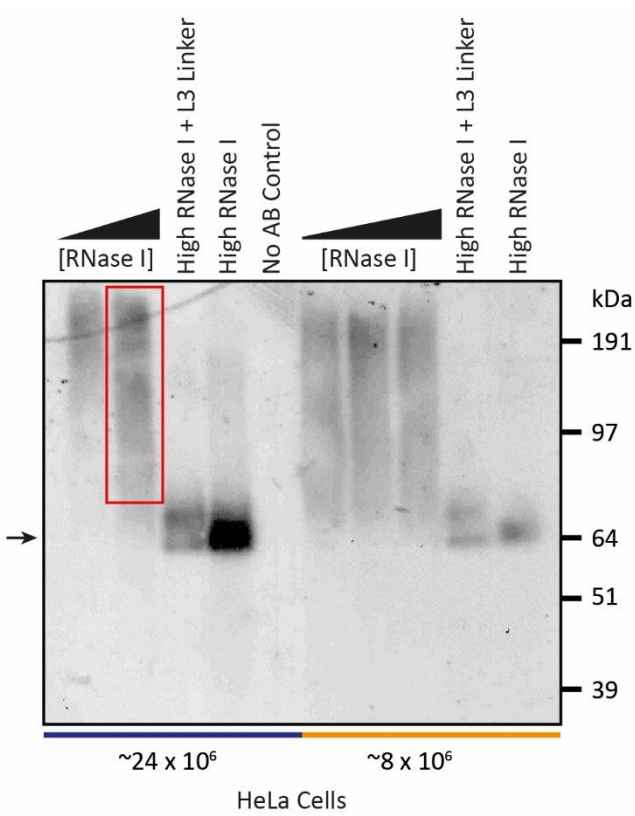


Figure 3.8: Optimisation of RNase I digestion FLAG-IMP1 RNA complexes
Autoradiograph displaying immunopurified FLAG-IMP1 RNA complexes digested with varying concentrations of RNase I. HeLa cell lysates were incubated with either 1:10, 1:500, 1:5,000 or 1:40,000 dilution of RNase I for 3 minutes at 37°C while shaking. Digested samples were then immunopurified using mouse anti-FLAG antibody and Protein G dynabeads. Purified complexes were ^{32}P radiolabelled and subjected to SDS-PAGE. Complexes were then transferred to a nitrocellulose membrane and exposed to film at 80°C for 1 h. Black arrow indicates FLAG-IMP1 protein. Red box depicts region of membrane that was excised and FLAG-IMP1 RNA complexes extracted for continuation of iCLIP protocol.

3.12 iCLIP cDNA PCR amplification for high throughput sequencing

After extracting the protein-RNA complexes from the membrane, the quantity of RNA / transcribed cDNA is too small for us to visualise. The next step at which progress of the iCLIP protocol can be evaluated is after PCR amplification of the relinearised cDNA fragments (Figure 2.2 Section 17).

A cDNA concentration of ~10 nM is required for high throughput sequencing. iCLIP cDNA libraries need to be amplified to this concentration in order to move onto sequencing. It is reported that over amplification of iCLIP cDNA libraries results in the formation of secondary products, which are observed on a native TBE gel as smears above the expected cDNA fragment size.[116] In addition to this, if a high number of PCR amplification cycles is required to achieve a concentration of 10 nM, then the diversity and quality of the iCLIP library is likely to be low.

After reverse transcription cDNA fragments are purified using a UREA-TBE gel (Figure 2.2 Section 13). The denaturing conditions of the gel result in cDNA fragments being separated by size only. This step is important for removing excess reverse transcription primer which can affect later iCLIP protocol steps. After fragment size separation cDNA fragments are purified from the gel according to size. As a quality test of our reverse transcription reaction I excised three bands from the UREA-TBE gel. A lower band between 80-100 nt, a medium band between 100-150 nt and a high band between 150-350 nt. The excised bands were then circularised and relinearised (Figure 2.2 Section 15 and 16) and processed to the PCR amplification stage and amplified with 25 PCR amplification cycles. Figure 3.9 shows the amplified products and how they correspond to the correct size extracted from the gel. This reconfirming the correct RNase I digestion concentration as I generated cDNA fragments of a range of sizes after the reverse transcription stage.

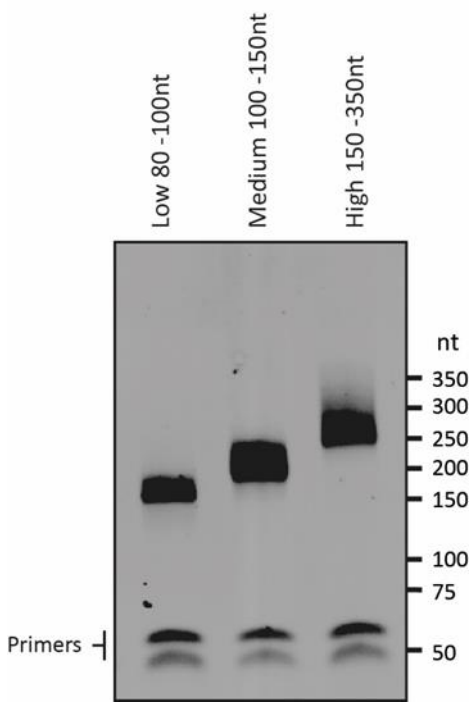


Figure 3.9: iCLIP cDNA libraries after PCR amplification
Amplified libraries were loaded onto a 6% TBE gel and run in TBE running buffer. SyberGreen gel staining and UV transillumination were used to detect cDNA bands.

3.13 Identification of unique cDNA reads and mapping to the human genome

iCLIP libraries were sequenced via hight-throughput sequencing on an Illumina GA2 platform. Sequence reads were then demultiplexed according to the individual experiment by using defined barcode sequences that were inserted into cDNA reads via 3' adaptor. Random barcodes were registered before being removed from sequence reads. Trimmed reads were then mapped to the human genome sequence (version Hg19), allowing one mismatch using Bowtie version 0.10.1 (command line: -a -m 1 -v 1). The reads that mapped to the genome make up the total read number. Logged randomised barcodes were used to remove duplicated sequences that were the result from over-amplification during PCR.

Mapped reads that contained the same unique barcode and the same crosslink nucleotide were discounted. The reaming mapped reads are termed unique reads.

The first iCLIP study published was performed on the RNA binding protein hnRNP C. hnRNP C is one of the most abundant proteins found within the nucleus and is known to regulate RNA splicing. In the original study hnRNP C RNA complexes were immunopurified from HeLa cell lysates using an antibody that specifically recognises hnRNP C. As a negative control the group used a no antibody pull down iCLIP experiment to show the proportion of reads generated from iCLIP samples are several orders of magnitude higher than the control.[111]

In the original hnRNP C study three biological iCLIP repeats were generated. These reads were processed and mapped using the same method as described above. After sequencing the three biological repeats generated 6.5 million sequencing reads, of which 4.2 million reads mapped to the human genome at a single location allowing one base miss match. After PCR amplification bias was accounted for using unique barcode sequences and crosslink nucleotides a total of 641,350 reads were calculated. Each read, therefore, represented a uniquely crosslinked RNA molecule. Crosslink nucleotides were then summarised as ‘cDNA counts’ and give a measure of hnRNP C binding to each crosslink nucleotide.

FLAG-IMP1 iCLIP libraries were processed in the same manner as in the original experiment. As an initial test to determine the quality of sequencing data obtained from our iCLIP libraries I calculated the unique read to total read ratio. Factors such as PCR amplification bias, short DNA reads and low complexity of iCLIP cDNA library all result in a higher total read to unique read ratio, and so these are taken into account in this initial filtering of the data. As the table shows, for each of the constructs, after replicates were summated, I achieved well in excess of 1.5 million unique reads and a unique read to total read ratio in the range of 1.71 to 5.46.

For our endogenous IMP1 and FLAG-IMP1 samples I obtained two biological repeats and three biological repeats for the FLAG-IMP1 KHDD mutant constructs. Previous CLIP studies have typically used between 2-3 biological repeats.[110,111,188,190] As previously stated, the original hnRNP C study used a total of just over 0.6 million unique reads in their analysis of transcriptome-wide binding.[111]

While the number of biological repeats is important for statistical validation of observed RNA binding, the number of unique reads is also a factor that can influence data quality. There is currently no accepted value of what the minimum number of biological repeats or unique reads should be in order to generate a validated conclusion for an analysis of an iCLIP data set. However, the number or repeats and unique reads I have generated for our iCLIP samples matched or exceeded that reported in current iCLIP studies. It is also important to note that a recent article[194] analysed previous iCLIP, PAR-CLIP and HITS-CLIP studies alongside newer repeats of these studies to determine the best method of identifying crosslinked nucleotides. They identified that the number of assigned crosslink clusters can vary greatly between different data sets in a manner that does not necessarily correlate with the number of unique cDNA reads of that library.

Construct	Unique Reads (10^6)	Total Reads (10^7)	Ratio
Endogenous	9.97	2.14	2.14
FLAG-IMP1 WT	6.11	1.96	3.21
FLAG-IMP1 KH1DD	5.18	1.62	3.13
FLAG-IMP1 KH2DD	2.27	1.24	5.46
FLAG-IMP1 KH3DD	9.81	1.56	1.56
FLAG-IMP1 KH4DD	5.92	1.01	1.71

Table 3.1 Unique mapped reads and total reads from grouped iCLIP experiments

3.14 Biological iCLIP repeats display a high degree of reproducibility when comparing sequence composition at crosslink nucleotides

When generating biological repeats of high-throughput data it is important to determine the degree of similarity between the repeats before pooling the data sets together. The statistical power to determine differences between data sets relies on the data generated from biological replicates to be similar. A large variation between repeats of the same condition reduces the confidence of observed differences in data between different data sets.

There are several ways to compare similarity of data sets. Previous studies have determined the similarity of biological repeats by comparing the correlation between region-based fold enrichment of crosslink sites between replicates. Another common method is to compare the occurrence of nucleotide sequences identified at crosslink sites.

I compared the occurrence of pentanucleotides overlapping the crosslink nucleotides. First, I compared the two FLAG-IMP1 WT repeats (Figure 3.10) and found a high degree of similarity with a R^2 of 0.983. In addition, I compared the two endogenous IMP1 iCLIP repeats preformed using the endogenous antibody (Figure 3.10). Comparing the two endogenous repeats showed larger variation of identified pentanucleotide sequences in the two repeats.

Analysis of the mutant iCLIP repeats showed high correlation for each mutant construct, with the exception of the FLAG-IMP1 KH4DD repeats (Figure 3.11). I set a R^2 coefficient of determination threshold of 0.9 as a level of acceptability.[111] This level was not achieved when comparing R1 and R3 of the FLAG-IMP1 KH2DD repeats (R^2 0.822). Considering these repeats contained the lowest number of unique mapped reads (0.53 and 0.58 x 10^6 respectively) I decided to summate these two repeats and re-perform our Kmer analysis. Comparing the new R1 R3 group to the R1 repeat I saw an improvement in correlation to 0.938 (Figure 3.12). I took into consideration that the analysis of FLAG-IMP1 mutant binding will be carried out on iCLIP reads that have been summated for all the biological repeats of that construct. Due to the initial correlation of the IMP1-KH2DD repeats only narrowly falling short of the accepted 0.9 correlation value, and that these repeats had the lowest number of unique reads, I accepted the repeats of this data set. In contrast, R3 of FLAG-IMP1 KH4DD replicates contained 1.19 x 10^6 unique mapped reads. The low R^2 value of 0.318 (R1 vs R3) and 0.580 (R3 vs R2) in addition to the higher unique read count resulted in the R3 replicate of FLAG-IMP1 KH4DD being removed from the data groups that were filtered into the further analysis.

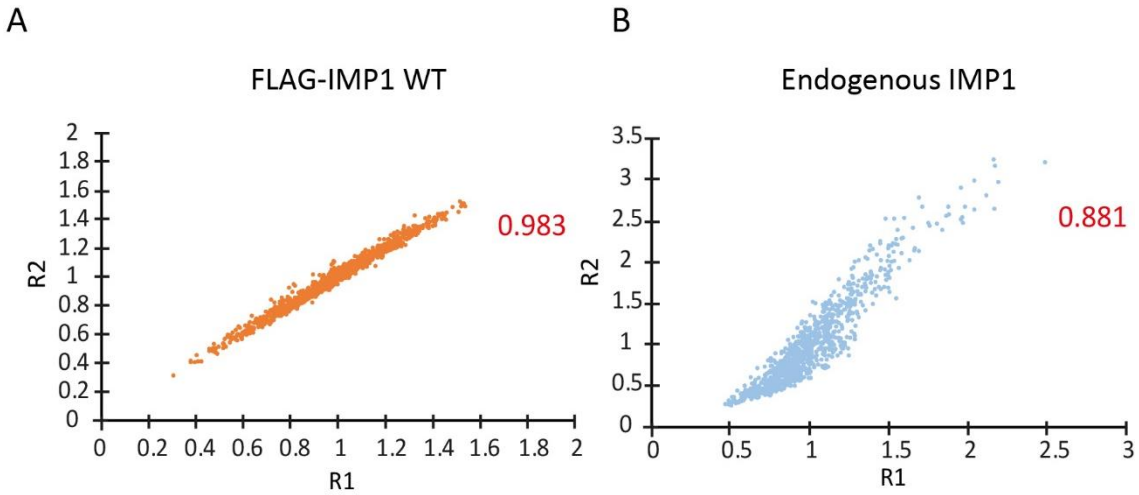


Figure 3.10: Correlation of sequence composition at crosslink nucleotides between endogenous IMP1 and FLAG-IMP1 WT repeats
Frequencies of pentanucleotides overlapping with crosslink nucleotides are shown for FLAG-IMP1 WT and endogenous IMP1 iCLIP repeats. Correlation of enriched pentamers was used to determine reproducibility of biological replicates. Enrichment of pentanucleotides was identified in a window of (-30,-10 nt), (10, 30 nt) relative to each crosslink site to avoid uridine crosslink bias. Each identified pentamer is counted only once in the analysed window. Random occurrence of pentamer sequences was determined by randomly shuffling iCLIP crosslink sites 100 times within corresponding genome segments. To determine pentamer enrichment in iCLIP replicates the occurrence of pentamers in the window of the crosslink sites was divided by the random occurrence of pentamers. Values greater than 1 represent enrichment of a pentanucleotide sequence in iCLIP crosslink site. The R^2 coefficient of determination for each comparison is displayed in red.

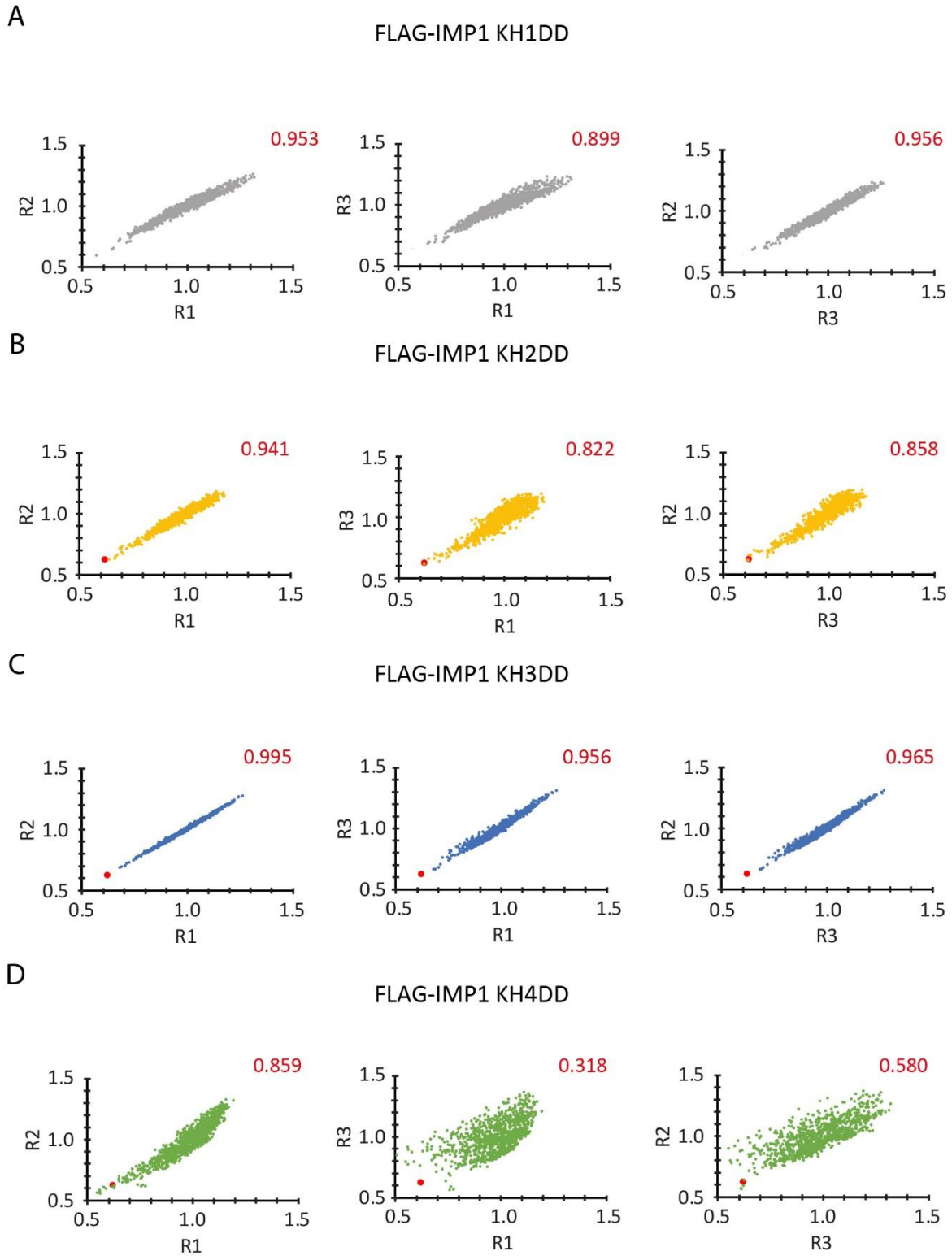

A
FLAG-IMP1 KH1DD
0.953
0.899
0.956
B
FLAG-IMP1 KH2DD
0.941
0.822
0.858
C
FLAG-IMP1 KH3DD
0.995
0.956
0.965
D
FLAG-IMP1 KH4DD
0.859
0.318
0.580
R1
R2
R3
0.5
1.0
1.5

Figure 3.11: Correlation of sequence composition at crosslink nucleotides between FLAG-IMP1 KHDD mutant repeats
Frequencies of pentanucleotides overlapping with crosslink nucleotides are shown for each FLAG-IMP1 KHDD mutant replicate. Correlation of enriched pentamers was used to determine reproducibility of biological replicates. Enrichment of pentanucleotides was identified in a window of (-30,-10 nt), (10, 30 nt) relative to each crosslink site to avoid uridine crosslink bias. Each identified pentamer is counted only once in the analysed window. Random occurrence of pentamer sequences was determined by randomly shuffling iCLIP crosslink sites 100 times within corresponding genome segments. To determine pentamer enrichment in iCLIP replicates the occurrence of pentamers in the window of the crosslink sites was divided by the random occurrence of pentamers. Values greater than 1 represent enrichment of a pentanucleotide sequence in iCLIP crosslink site. The R^2 coefficient of determination for each comparison is displayed in red. A) FLAG-IMP1 KH1DD replicates B) FLAG-IMP1 KH2DD replicates. C) FLAG-IMP1 KH3DD replicates. D) FLAG-IMP1 KH4DD replicates.

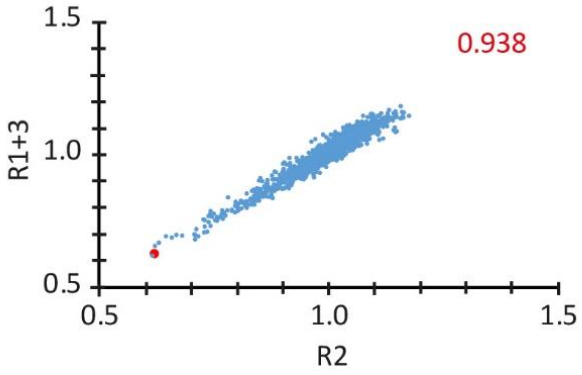


Figure 3.12: Summation of iCLIP repeats improves correlation of sequence composition at crosslink nucleotides
The two iCLIP repeats for FLAG-IMP1 KH2DD with the lowest number of unique reads (R1 and R3) were grouped by summation. Frequencies of pentanucleotides overlapping with crosslink nucleotides were then recalculated, as described above, for the new summated data and correlated with replicate 2. The R^2 coefficient of determination is displayed in red.

3.15 Enrichment of IMP1 binding to 3’ UTR gene region

As an initial analysis I compared the distribution of iCLIP mapped reads across genomic regions of the whole H19 genome. I analysed the percentage of overall total reads located in defined segments and then normalised this to account for segment feature size. I took each repeat individually in order to calculate a group average and corresponding standard error.

I identified widespread enrichment in binding to the 3’ UTR for all FLAG-IMP1 constructs and a depletion in intronic signal globally across all FLAG-IMP1 constructs. All FLAG-IMP1 constructs share a genomic distribution pattern that is within the error of the average measurements (with the exception of the FLAG-IMP1 KH2DD group). I saw a reduction in the 3’ UTR binding for the KH2DD group and an increase in binding to ncRNAs. This could be the result of an altered binding profile due to the loss of KH2 RNA recognition (Figure 3.13).

The distribution of our iCLIP crosslinks across gene segments support previous findings that IMP1 recognises a diverse range of RNA substrates rather than interacting with only a small subset.[110,144] I observed enrichment in the 3’ UTR and also in the ORF across a range of RNA targets. The previous CLIP studies performed on human IMP1 and Drosophila imp, all observed an enrichment in binding to the 3’ UTR region of transcripts.[110,144,183]

This is consistent with previous *in vitro* binding studies in which the location of IMP1 recognition within the RNA transcript was determined. For example I observed 3’ UTR enrichment within the transcripts; CD44,[145] CTNNB1,[196] MAPK4,[176] ACTB,[160] and KRAS,[197] of which 3’ UTR binding has previously been characterised for IMP1 recognition. This binding pattern is expected for IMP1 as the 3’ UTR contains binding sites for regulatory proteins as well as miRNAs. These regulatory elements within the 3’ UTR can influence polyadenylation, translation efficiency, localisation, and stability of the mRNA, functions which IMP1 has been shown to regulate for its target transcripts.

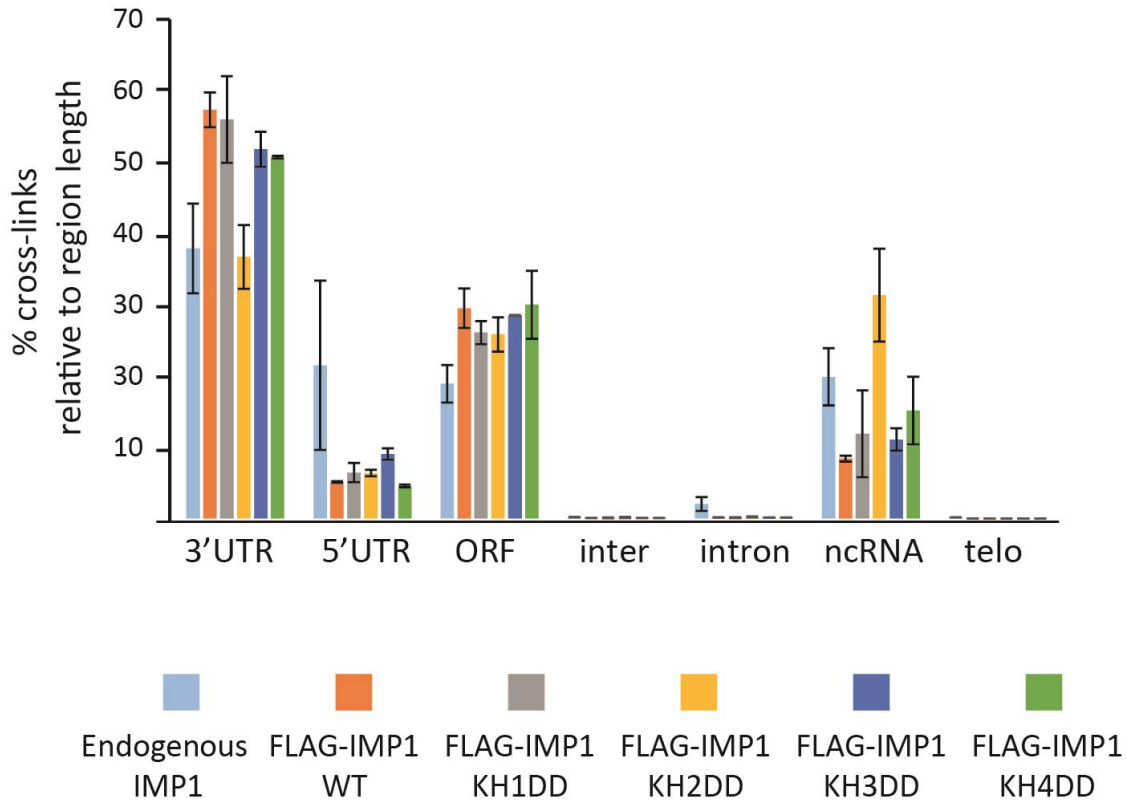


Figure 3.13: Distribution of endogenous and FLAG-IMP1 constructs iCLIP reads among mRNA regions and ncRNAs
Percent of total unique iCLIP reads normalised to feature length for each gene fraction. Percentages were calculated from averaging the iCLIP reads between replicates of that construct group. Standard deviation error bars are included.

3.16 Normalisation of iCLIP data to compare binding sites

The incorporation of random barcodes into the reverse transcription primers used in the iCLIP technique allows for PCR amplification bias to be accounted for. This enables iCLIP crosslink sites to be turned into cDNA counts which results in a semiquantitative measure of the RBP crosslinking to each position.[111] However, when comparing cDNA counts across different genes for the same iCLIP experiment, the number of counts is dependent on both the affinity of the RBP for that transcript and the abundance of the RNA transcript within the cell system used. This has led to some debate regarding the best way

to normalise iCLIP data sets in order to compare RBP binding in a non-bias quantitative manner.[107,114]

One approach to account for difference in transcript abundance is by normalising the enriched crosslink clusters identified in individual transcripts with the overall number of crosslinks to that gene (Figure 3.14). Genes with low expression levels would in theory contain less nucleotide crosslinks overall compared to highly expressed transcripts. This would normalise the enriched binding clusters to allow for a more accurate measure of RBP affinity for those enriched sites. However, this assumes the read coverage across the whole gene is dependent only on the abundance of that gene. Individual transcripts are likely to have multiple binding sites with different degrees of RBP affinity and so total cDNA counts across the whole gene are determined by additional factors. Although this method provides a rough normalisation approach its limitations should be taken into consideration.

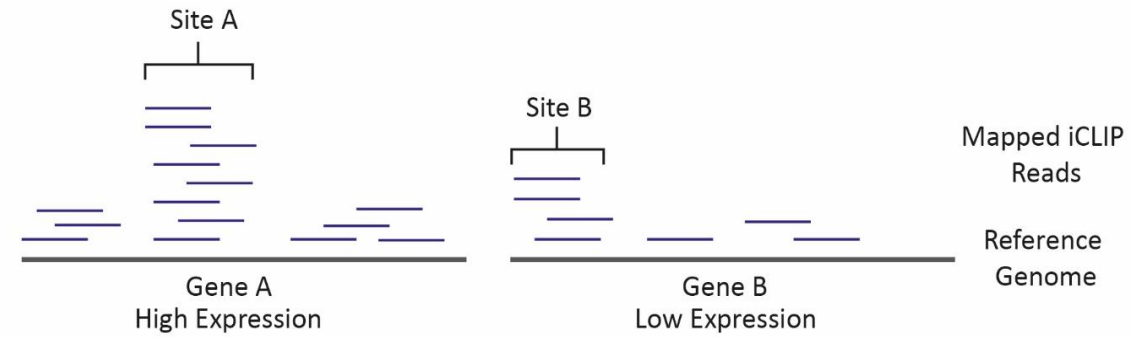


Figure 3.14 Normalising iCLIP clusters to the total cDNA count within that transcript
Illustration demonstrating transcript normalisation principle of crosslink counts. Highly expressed Gene A (left) occurs a higher overall cDNA count across the whole gene (mapped reads in blue) when compared to the lower expressed Gene B (Right). Binding Site A has a higher cDNA count relative to Site B as a result of gene abundance. Accounting for the overall cDNA counts across the transcripts normalises enrichment of Site A to Site B.

Another approach that has been suggested to account for transcript abundance in CLIP studies is to incorporate RNA-seq data sets. This has been implemented

in several studies yet RNA-seq has its own limitations. Read coverage across the genome can vary and different RNA species are sequenced in different ways. In addition, proteins involved in splicing predominantly bind to pre-mRNAs and these RNAs are not efficiently quantified by standard RNA-seq techniques.[107]

Finally, UV induced crosslinking also incorporates bias, with Us having a higher crosslinking potential than other nucleotides.[108] In turn, the efficiency of crosslinking, and therefore number of cDNA counts, is also influenced by the local RNA sequence of binding sites, and RNA structure. Complex bioinformatic approaches are being developed to better account for these biases when identifying significant binding clusters from CLIP data.[114,198–200]

3.17 Normalisation of FLAG-IMP1 iCLIP data

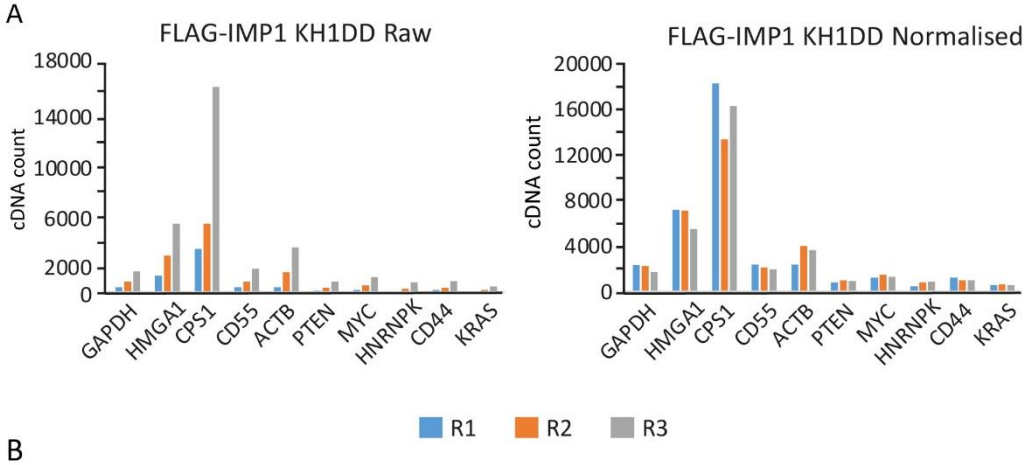
The limitations of CLIP studies explained above relate to the issues of identifying real binding sites within an individual RBP CLIP study. These factors also influence our data when I focus on identifying real RNA targets and RNA recognition motifs for each FLAG-IMP1 construct individually. However, the focus of this study is to observe differences across iCLIP data sets in a comparative analysis between different FLAG-IMP1 mutants. The effects of transcript abundance on IMP1 binding can be accounted for by comparing binding to the same gene across all data sets. This does assume that the mutated IMP1 proteins do not change transcript abundance across the cell lines due to modified binding modes. This issue will need to be confirmed via later validation studies on the effects of altered binding patterns on identified target transcripts. I can also account for any effects local RNA sequences may have on the efficiency of crosslink induced nucleotides by comparing the same crosslink clusters identified in the genes being compared, as the RNA sequences will also be the same across data sets.

However, as reported above, the iCLIP libraries I have generated for each IMP1 construct group vary in the number of total unique reads (cDNA counts). In order

for us to compare binding across groups I needed to implement a normalisation method that would account for this difference. To my knowledge the only other previous CLIP study that directly compared the binding sites and RNA recognition motifs between a WT and mutant protein was a PAR-CLIP study performed on FMRP. In this study the group reported that for their mutant PAR-CLIP library they obtained a much higher unique read count compared to their WT FMRP PAR-CLIP. They accounted for this difference by randomly selecting unique reads in the mutant PAR-CLIP so that the new data set contained a total number of unique reads that was no more than 10% more than that of the WT.[191]

As we are comparing multiple sets of data I initially implemented a simple normalisation approach by applying a normalisation factor that is based on the total unique read count. To investigate the effects of this normalisation process I took the three replicates of the FLAG-IMP1 KH1DD group (Figure 3.15). This was chosen due to replicate 1 and replicate 3 having the largest difference in unique reads (Figure 3.15B). This difference of ~2.63 million reads required a normalisation factor greater than that which would be applied to the grouped FLAG-IMP1 data sets. Figure 3.15A shows that Replicate 3 consistently has a higher cDNA count per gene compared to the replicates with lower overall unique reads. Once normalisation factors are introduced the variation of cDNAs counts across the genes reduces significantly. I used this as conformation that normalising data sets would not introduce additional bias but instead normalise the variation resulting from the difference in total unique read number.

A



B

Construct	Replicate	Total unique reads (10^6)	Correction factor between repeats
FLAG-IMP1 KH1DD	R1	0.61	5.319
	R2	1.33	2.436
	R3	3.24	1

C

Construct	Normalisation Factor
WT	1.61
KH1DD	1.89
KH2DD	4.32
KH3DD	1.00
KH4DD	1.66

Figure 3.15: Normalising iCLIP repeats according to total unique reads standardises data

As a proof of principle, I took the three biological repeats of the FLAG-IMP1 KH1DD group and applied a normalisation factor that was determined by the total of unique reads. A) The total number of raw iCLIP reads identified in a subset of genes for each biological replicate (left). These values were then adjusted using normalisation factor (right). B) The distribution of total unique reads between FLAG-IMP1 KH1DD replicates and corresponding normalisation factor. C) Normalisation factors to standardise between groups. Factors were based on total number of unique reads and implemented as above.

3.18 Endogenous IMP1 and FLAG-IMP1 WT iCLIP data contain conserved crosslink sites but endogenous IMP1 iCLIP has reduced signal to noise ratio

At the time of writing this thesis the data analysis of this investigations is still ongoing. We are yet to report on differences in mutant binding with statistical certainty. Instead here I will report on the binding I observe to the ACTB mRNA.

I observed the binding patterns of endogenous IMP1 and FLAG-IMP1 WT to the whole ACTB gene. I identified crosslink nucleotides across the majority of the ACTB gene for the endogenous IMP1 iCLIP, including in the intronic regions. In contrast, FLAG-IMP1 WT contained crosslink nucleotides only in the 3' UTR and ORF regions. Comparing the total number of unique reads between the endogenous IMP1 and FLAG-IMP1 WT iCLIP groups I obtained ~3.86 million more unique reads for the endogenous IMP1 iCLIP than for the FLAG-IMP1 WT group (with the same number of biological repeats).

Taking into consideration the large difference in unique read counts between the endogenous IMP1 and FLAG-IMP1 WT (Table 3.1), in addition to the unexpected crosslinks observed in the intron region of the ACTB gene for the endogenous IMP1, I concluded that the different antibodies used to immunoprecipitate the different IMP1-RNA complexes were the reason for the difference. The IMP1 antibody is a polyclonal antibody derived from rabbit, whereas the FLAG antibody is a highly specific purified monoclonal mouse antibody. The endogenous IMP1 iCLIP contains a large degree of background cDNA reads that have copurified with the less specific IMP1 antibody. This increase in noise makes it more difficult to identify real binding sites. However, for this example of ACTB, I can still identify the highly specific KH4 binding site (Figure 3.16). In turn, we concluded that the iCLIP data obtained from the FLAG-IMP1 WT iCLIP represent that of the endogenous iCLIP but with an improved signal to noise ratio.

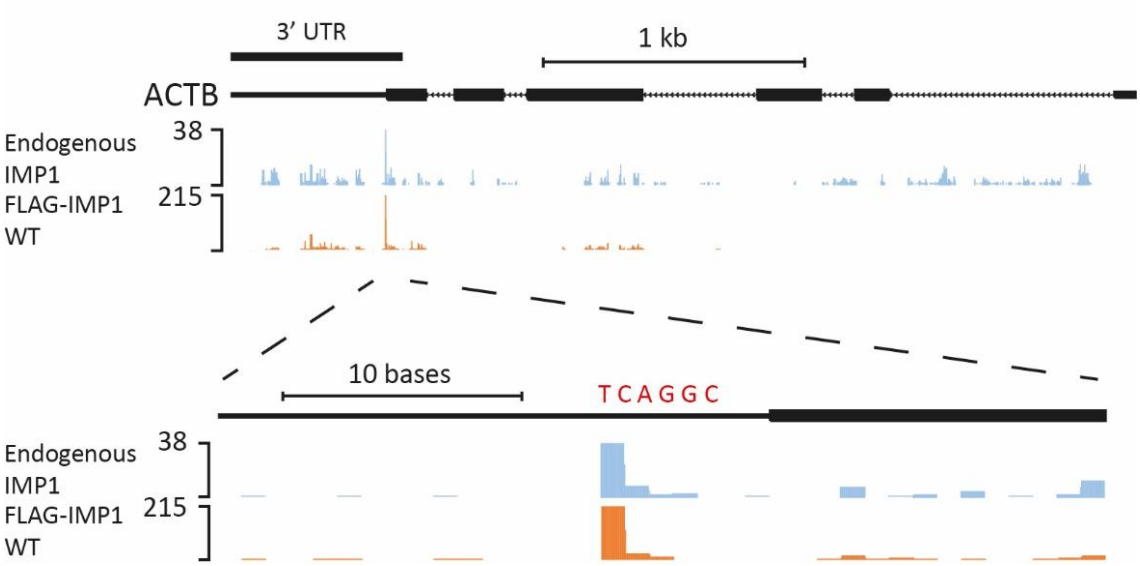


Figure 3.16: Endogenous IMP1 iCLIP displays reduced signal to noise compared to FLAG-IMP1 WT iCLIP but binding patterns are conserved
USCS gene browser view of unique crosslink sites of endogenous IMP1 (light blue) and FLAG-IMP1 WT (orange) identified on the ACTB gene. The number of sequences identified at each crosslink nucleotide are displayed by cDNA count and represented via intensity of bars. Zoom panel represents 3' UTR fraction of ACTB and highlights the KH4 recognition sequence in red (CGGACU). Note IMP1 bound RNA fragments map to the antisense strand of the ACTB gene and in turn the DNA sequence is read 3' to 5' in figure. Numbers displayed represent the most intense crosslink nucleotide via cDNA count in the currently displayed window.

3.19 KH3 and KH4 domain RNA binding knock out mutations result in reduced crosslinks to 3' UTR of ACTB

The zipcode region of the 3' UTR of ACTB is to date the best studied RNA target of IMP1. *In vitro* studies have identified the RNA recognition sequences in the zipcode element and confirmed the KH34 domain as responsible for recognition. Binding studies of the full-length IMP1 protein to ACTB showed that a construct containing only KH34 bound with a similar affinity. Constructs containing just the RRM12 or KH12 domains bound with 100-fold reduced affinity compared to the full-length protein.[83,85] These experiments highlight the importance of KH34 recognition for the high affinity recognition of the ACTB mRNA.

A study comparing the steady state levels of mRNA from isolated embryonic brains of WT and IMP1 KO mice strains showed IMP1 does not regulate the stability of ACTB.[179] This coupled with preliminary mRNA-seq data that I performed on our mutant cell lines enables us to conclude that ACTB abundance is similar across our mutant cell lines. In turn, the differences I observe in cDNA counts across the ACTB gene are a result of the altered RNA binding profiles of our IMP1 KHDD mutants.

Analysis of the crosslink nucleotides in the ACTB mRNA shows that when either KH1 or KH2 binding is knocked out I observe an increased binding to the ACTB gene (Figure 3.17). This suggests a shift in the recognition profile of the mutated proteins. Few previous studies have focused on identifying KH12 RNA targets, with some groups suggesting these domains play a reduced role in RNA recognition compared to the KH34 domains. This increase in ACTB binding for the KH1 and KH2 knock out could suggest there is a reduced competition between the KH12 and KH34 di-domains for RNA targets. Inability of the mutant proteins to recognise KH12 targets due to the binding knock out mutations could result in a shift in binding to the KH34 target subset, as observed for ACTB.

Interestingly, the reduction in crosslinks observed for the KH3 and KH4 domain knock outs is less pronounced when comparing binding across the whole ACTB transcript. When raw cDNA counts are considered, KH3 and KH4DD contain the same number of counts compared to the WT (Figure 3.17A). However, when considering crosslinks that have a FDR of less than 0.05 I see a slight reduction, but still in line with what is observed for the WT protein. If I look at the crosslink cluster over the KH4 recognition site, which represents the most enriched site across the whole gene for all IMP1 constructs (Figure 3.17B). I observe ~50% reduction in binding to this site for KH3 and KH4 knock outs. The increase in binding seen for the KH1 and KH2 domain knock outs is also more pronounced when analysing just this cluster (Figure 3.17A & B).

I do not observe a total loss in crosslinks at the KH4 target sequence for the KH4 or KH3 DD mutation. However, there is a slight decrease in binding for the KH4DD mutation compared to the KH3DD mutation. A previous study that mutated the KH3 recognition sequence form the zipcode sequence of ACTB showed IMP1 was still able to bind to the transcript but with lower affinity.[85,179] Additionally, in the di-domain KH34 construct, where one domain contains the GDDG RNA binding mutation, the partner KH domain is able to bind it's RNA recognition element with a Kd ~1 µM.[180] As we know the KH34 is a di-domain that work as a pair to recognise transcripts, the presence of a functioning KH domain partner can give an explanation as to why we do not see total loss of binding at this site. Additionally, we know KH3 and KH4 recognition are equally important for ACTB binding, yet we do not observe a strongly enriched crosslink cluster at the KH3 recognition site. This could be due to sequence specific crosslink artefacts and so the KH4 recognition site results in more cDNA reads, but recognition of this site is also coupled to KH3 recognition. Accordingly, comparing the same crosslink clusters allows for a fair comparison of binding.

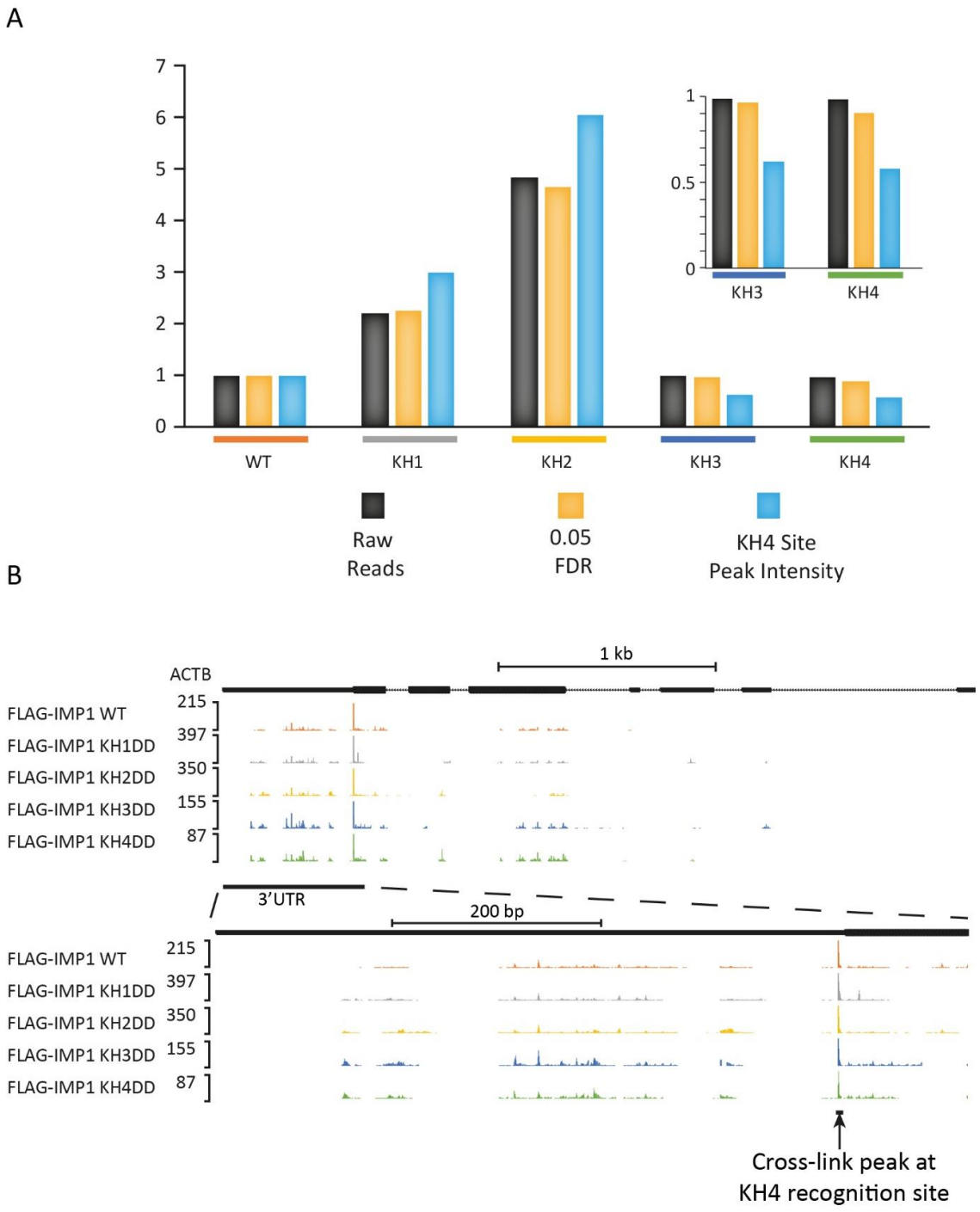


Figure 3.17: Recognition of ACTB mRNA is dependent on KH3 and KH4 domain RNA binding

A) Bar char displaying the cDNA counts identified in the ACTB gene for KHDD constructs as a ratio of cDNA counts identified in FLAG-IMP1 WT iCLIP after normalisation. Black bars: raw cDNA counts across the whole gene. Orange: cDNA counts after FDR of 0.05 is applied. Blue: cDNA count identified at KH4 recognition sequence after FDR filtering. Insert displays the same data for KH3 and KH4 but on a scale that better represents differences in cDNA counts. B) USCS gene browser view of unique crosslink sites of FLAG-IMP1 constructs after

FDR filtering (WT orange; KH1DD grey; KH2DD yellow; KH3DD dark blue; KH4DD green). The number of sequences identified at each crosslink nucleotide are displayed by cDNA counts and represented via intensity of bars. Zoom panel highlights 3' UTR fraction of ACTB and the crosslink cluster identified at KH4 recognition sequence. Numbers displayed represents the most intense crosslink cluster via cDNA count in the current display window.

3.20 Discussion

In this chapter I have described the experimental approach I have implemented to understand how IMP1 selects RNA targets in HeLa cells. I have characterised a Flp-In T-Rex HeLa cell system in which I have generated stably transfected cell lines expressing a series of FLAG-IMP1 KH mutants. This series of FLAG-IMP1 KH mutants contain a GDDG mutation within one of the four KH domains. This mutation has been proven to knock out the RNA binding properties of the KH domain without disruption to the structure or stability of the domain. I have successfully developed a system where these FLAG-IMP1 KH mutants are expressed within the HeLa cells to a level that represents that of the endogenous IMP1 protein. This provided a solution to a potential issue of a previous PAR-CLIP study on FLAG-tag IMP1 proteins in HEK293 cells where constructs were transiently expressed resulting in over expression compared to the endogenous protein. This overexpression was reported to result in the FLAG-IMP1 proteins purifying in a different polysomal fraction compared to the endogenous protein.[141]

From my Flp-In T-REx HeLa system I was able to selectively immunoprecipitate the RNA transcripts bound by our mutant FLAG-IMP1 proteins. We have shown that our constructs do not dimerise with the endogenous IMP1 protein, and that the cellular localisation of the FLAG-IMP1 constructs compares to that of the endogenous IMP1 protein. In addition, I show FLAG-IMP1 KH mutant proteins display a lower in-cell RNA binding affinity compared to the WT FLAG-IMP1 construct.

I successfully produced high quality iCLIP libraries for each of the FLAG-IMP1 constructs. The biological repeats of each construct displayed a high degree of

reproducibility when comparing enriched pentanucleotide sequences identified at crosslink sites. Analysing the overall distribution of the FLAG-IMP1 crosslink sites showed enrichment in the 3' UTR region. This is consistent with previous CLIP studies performed on the IMP1 protein and with the documented *in vitro* studies of IMP1 controlling RNA transcript fate by predominately binding to the 3' UTR region of target transcripts.[110,141,144,183] I observe a depletion of crosslink sites in intronic regions. This again provides validation of our data sets given IMP1's cytoplasmic localisation and RNA regulation in the cytoplasmic cellular compartment.[83,153,160]

Comparing the iCLIP crosslink profiles of endogenous IMP1 and FLAG-IMP1 revealed crosslink sites to previously identified high affinity targets (ACTB) were conserved. However, the endogenous iCLIP contains a lower signal to noise ratio. This was identified by the increased number of crosslinks present in intronic regions and is likely due to the different antibodies used to selectively immunoprecipitate the different IMP1 complexes. Furthermore, comparing our data with data sets obtained from the Ule research group, where iCLIP libraries were produced from a FLAG-tagged IMP1 protein in HeLa cells showed our data sets to be consistent (data not included).

As the zipcode RNA sequence within the 3' UTR of the ACTB gene is the best validated target for the IMP1 protein, I assessed if I could observe different crosslink profiles to this transcript between our mutant FLAG-IMP1 constructs. I observed that mutations in the KH3 and KH4 domain reduced binding to the 3' UTR region of the transcript, specifically at the KH4 recognition site. Conversely, the number of crosslinks identified in this region for the KH1 and KH2 mutant proteins increased. This increase in crosslink could result from a shift in the mutant proteins binding away from RNA targets where the KH1 and KH2 domains play more of a significant role in binding, resulting in an increase in crosslinks at KH3 KH4 specific targets such as ACTB. However, due to the noise associated with CLIP studies, identification of this difference in crosslink number required us to count the crosslinks in the cluster at the KH4 recognition site. This cluster is

the largest cluster observed in the whole ACTB gene, this was also observed in the previous eCLIP study.[144] As the data analysis of our iCLIP investigation is ongoing, one analysis approach I will implement is to analyse the largest crosslink cluster observed in target genes. These enriched clusters could potentially relate to higher affinity RNA targets within the transcript. In turn, it is within these clusters I would expect to observe a difference in binding depending on the specific contributions the individual KH domains play in recognising that target element.

Finally, our initial iCLIP study has employed the use of the HeLa immortalised cell line as an initial system to investigate IMP1 *in vitro* RNA binding. This is a common approach implemented by previous CLIP studies. For example, HEK293 cells were used to identify FMRP binding targets.[191] As with IMP1, FMRP is a protein that is known to mediate the development of neuronal cells.[201] RNA-seq data sets have revealed immortalised cell lines (such as HEK 293) and brain tissues share ~90% of expressed genes.[114,202,203] Therefore, these systems provide a good entry point for the investigation of RNA selection on a transcriptome-wide level. However, the abundance of genes expressed in such cell lines will differ from that of neuronal cells. In addition, specialised regulatory mechanisms such as the axonal localisation of the ACTB mRNA to neuronal growth cones are known to not be represented in immortalised cell systems such as HeLa.[147,204–206] Therefore, we will need to couple our final iCLIP findings with functional validation studies in more specialised cell systems, such as neuronal or fibroblast cells. Additionally, we plan to produce RNA-seq libraries from our stably transfected HeLa cells to determine if mRNA abundance of KH domain specific targets, identified in the iCLIP study, change depending on the mutant proteins altered RNA recognition profile.

Chapter 4. Understanding the RNA specificity of the IMP1 KH3 and KH4 domains

4.1 Introduction

In cells, RNA target selection of RNA binding proteins is dependent on the multiple RNA recognition events of the proteins' individual RNA binding domains. These RBDs recognise RNA in a sequence specific manner, and this sequence specificity is critical for overall RNA target recognition.[40,56,71,165] This is shown when RNA sequences are mutated altering the motif sequence the RBD domain typically recognises. These mutations result in loss or attenuation of RNA binding. However, the RNA recognition motifs identified for RBDs *in vitro* do not always correlate well with corresponding RNA recognition motifs identified *in vivo*.[144] One reason for this is within cells RBDs do not exclusively bind to their highest affinity RNA target sequences. Rather, the domains can interact with multiple RNA sequences with varying affinities. To better understand *in vivo* RNA target recognition we need to fully explore the sequence preference of the individual domains for lower affinity RNA targets. Recognition of these different RNA target sequences can relate to different biological functions of the protein.

A mutational study of the KSRP protein used structural information from the solution complex structure of KSRP KH3-AGGGC to identify amino acids that are involved in specificity. The study identified Lys368 residue to form a hydrogen-bond with the O6 moiety of the G3 base. Mutating the protein to KSRP-K368R resulted in a change in the specificity in base position 2 (G3). The mutation shifted the specificity from a purine to a pyrimidine, but the RNA binding affinity of the domain remained unchanged.[207] This change in specificity also altered the biological function of the KSRP protein. One physiological target of KSRP is the Let-7 miRNA. KSRP association with the RNA promotes the biogenesis of pri–Let-7 to pre–Let-7.[207] The K367R mutation results in a reduction of KSRPs ability to promote this maturation process and shifted the function of KSRP towards

mediating mRNA degradation (another reported function of the protein).[207] This example shows how sequence specificity of KH domains relates to biological function.

I set out to better understand the structural basis of IMP1 KH3 KH4 domain RNA recognition by mutating residues that were predicted to be important for RNA binding from the solution protein:RNA complex structure.[180] This was used in conjunction with comparisons of other well studied KH domains to determine the mutations I would implement. I focused on residues that either mediated hydrogen-bonds with the nucleobases or defined the shape of the KH domains hydrophobic groove. By introducing mutant residues, I aimed to alter the network of hydrogen-bonds that mediate RNA recognition of the preferred RNA sequences and shift the sequence specificity of the domains in a manner similar to the KSRP example.

The identification of such mutants would provide a useful molecular tool to understand the role of RNA binding domain specificity for RNA transcript selection *in vivo*. Successful mutations would be incorporated into an iCLIP study, like the GDDG approach explained in Chapter 3. This would in turn allow us to study how the alerted RNA specificity of the domain shifts RNA target selection in the context of the full-length protein and on a transcriptome-wide scale.

4.2 Determining specificity in KH domains

Previous structural studies that resolved KH-nucleic acid complex structures have helped to establish a set of general rules for how KH domains bind target sequences in a specific manner. The solved complex structures of NOVA1 KH3,[208] SF1,[78] hnRNP K,[209] FBP,[210] and PCBP2[211] indicate that the two central nucleobases (position 2 and position 3) are typically either an adenine or cytosine, whereas nucleobases in position 1 and position 4 are pyrimidines (Figure 4.1).[71] From these structures we can identify key amino acid residues that make contact with these specific bases via hydrogen-bonding, electrostatic,

and hydrophobic interactions. The knowledge of how these interactions determine specificity enable us to better understand RNA target selection.[71]

In the cases of a cytosine in position 2 an arginine side chain located in the central β-sheet of the KH domain creates two direct hydrogen-bonds with the O2 and N3 moieties of the base (Figure 4.1A). In structures in which an adenine was preferred in position 2 the arginine residue is replaced with a lysine (Figure 4.1B). The lysine residue has a smaller side chain which better accommodates the larger purine base.[71]

In position 3 the Watson-Crick edge of the nucleobase is specifically recognised via two hydrogen-bonds formed between the backbone amide and carboxyl moieties of the same amino acid from the second strand of the KH β-sheet. Only adenine or cytosine bases are able to form this double hydrogen-bond and so these bases are preferred. A third water-mediated or direct hydrogen-bond involving the side chain of a residue in the α2-helix is able to generate discrimination between adenine and cytosine in this position (Figure 4.1).[71,68]

Until recently, KH domains were believed to recognise A/C rich sequences and discriminate against guanines in the central nucleotide positions. However, the KSRP protein, consisting of four KH domains, was shown to recognises a G rich sequence within the precursor of the tumour suppressor Let-7 miRNA family.[212] The solution structure of the KH3 domain of KSRP identified a recognition mechanism different from the canonical KH domain recognition. The complex solution structure of KSRP with the RNA sequence AGGGU revealed KH3 of KSRP to have a wider hydrophobic groove (Figure 4.1C).[207,213]

In the KSRP KH3 structure the guanine in position 3 is selected for because of the enlarged hydrophobic groove allowing the edge of the guanine base to shift along the second β-strand which enables the formation of a new set of hydrogen-bonds between the Watson-Crick edge of the base and the carboxyl and amide groups of Ile356 and Phe358. A further fourth hydrogen-bond between the side

chain of Gln349 and the base increases the specificity for a guanine in position 3 further (Figure 4.1C).[207]

The recognition of guanine bases by the KH3 domain of KSRP demonstrates the importance of the overall shape of the hydrophobic groove in directing the base towards the formation of hydrogen-bonds with amino acid backbone and side chain moieties. In turn, KH domain recognition of RNA bases is determined by a complex network of hydrogen-bonds. However, hydrophobic interactions and the overall shape of the hydrophobic groove dictate the orientation of the RNA base which too influences the hydrogen-bonding network.[71] Both of these contributing factors must therefore be taken into consideration when designing mutations to shift RNA base specificity of KH domains.

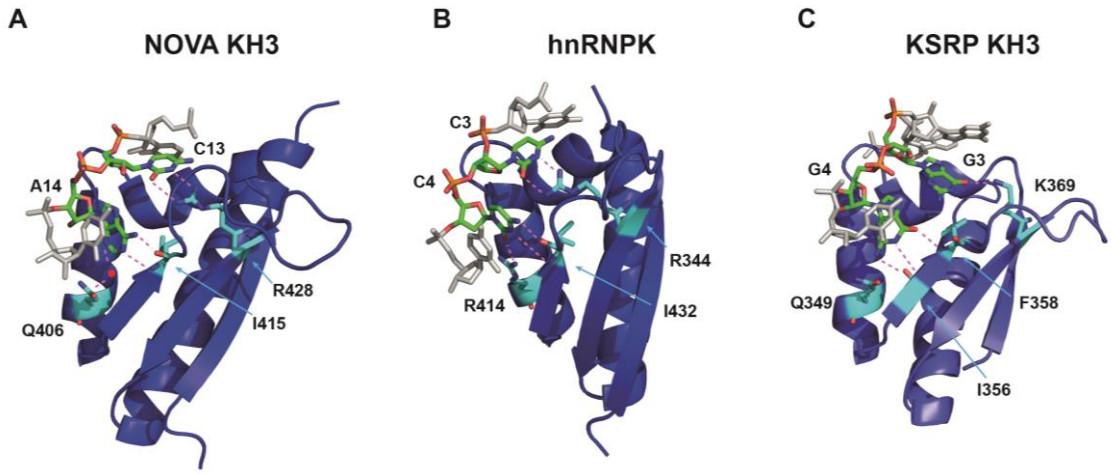


Figure 4.1: KH domain nucleobase sequence recognition of the two central binding positions within the hydrophobic groove
Complex highresolution structures of (A) NOVA2 KH3 with RNA ~UCAC~ (PDB:1EC6) (B) hnRNPK KH with DNA CTCCCC (PDB:1ZZI) and (C) KSRP KH3 with RNA AGGGU (PDB:4B8T). Central bases are numbered according to the sequence of nucleic acid used in the structure, and coloured by atom with carbons in Green, peripheral bases are coloured in Grey. KH domains are coloured in Blue with amino acid residues which are involved in recognition of the central bases are coloured by atom with carbons in Cyan. Hydrogen-bonds are depicted via Pink dashed lines.

4.3 Sequence specific recognition of the IMP1 KH3 and KH4 domain

Previous structural and biochemical studies investigating the RNA recognition properties of the KH3 and KH4 domain of IMP1 have shown the two domains recognise different RNA sequences with similar affinity. The KH3 domains RNA recognition motif is shorter (ACAC) compared to KH4 (CGGAC).[179,180]

IMP1 recognises the 3' UTR region of the ACTB mRNA transcript via the KH3 and KH4 domain binding their RNA recognition motifs located within the zipcode region. These recognition motifs are separated by a RNA linker of 14 nt in length.[85,179] The KH3 and KH4 domains are arranged as a pseudo dimer placing the RNA recognition surfaces on opposite faces of the dimer.[85] Recognition of the zipcode region of RNA causes the RNA to loop around the dimer enabling each domain to bind their target sequences. The linker region of RNA is not directly involved in recognition and does not make contacts with the KH3 KH4 dimer.[85,179] Therefore, I can investigate the binding of the KH3 and KH4 domains by using short stretches of RNA containing their RNA recognition elements.

The structural arrangement of the KH3 and KH4 dimer is such that expression of the domains individually is not possible and so can only be studied as the KH3 KH4 dimer. In order to study the RNA recognition properties of each domain individually, the RNA binding properties of the corresponding KH domain had to be abolished via the previously described GDDG mutation.[82,180]

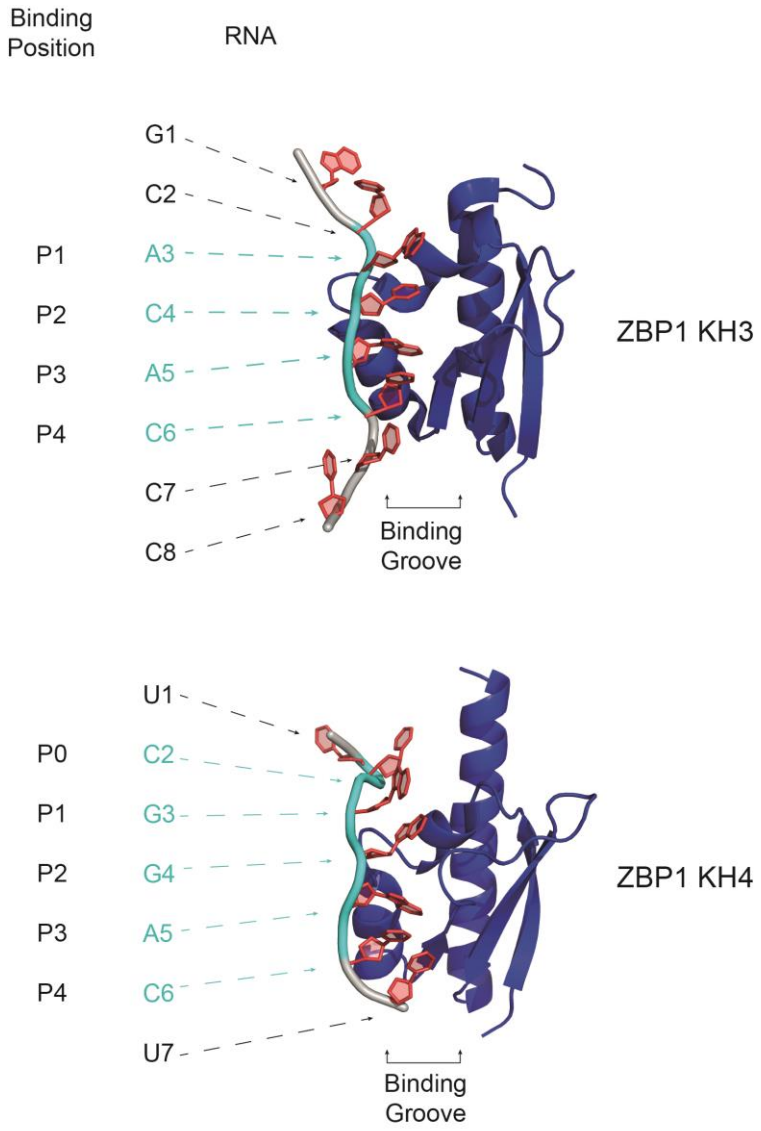


Figure 4.2: NMR complex solution structures of IMP1 KH3 and KH4 in complex full RNA sequence to differentiate between binding position and base number

RNA oligo is numbered by base order alongside structure, with bases that are specifically recognised by the hydrophobic groove coloured in Cyan. Left is the binding position number in relation to the number of bases bound by the groove (note the central bases are position 2 and 3). KH domain is coloured in Blue with RNA bases coloured in Red. The phosphate backbone of bases that are recognised specifically by the hydrophobic groove are coloured in Cyan with peripheral base backbones coloured in Grey.

4.3.1 RNA base specificity and recognition of the KH3 domain

As reported above IMP1 KH3 displays a sequence preference for the CA dinucleotide in the central position of the C/UCAC/A recognition sequence identified via SELEX.[179] The NMR solution structure of KH3 in complex with the RNA oligo GCACACCC shows the two bases in the central position, position 2 (C4) and position 3 (A5) (Figure 4.2), are recognised via multiple hydrogen-bonds and hydrophobic contacts.[180]

To determine the KH3 domain's ability to recognise different RNA bases in the two central positions RNA oligos were generated where the 2 central nucleotides (position 2: C4 and position 3: A5) were mutated individually to the three other possible RNA bases (Table 4.1). Isothermal titration calorimetry (ITC) measurements were then performed with the KH3KH4DD construct to determine the KH3 domains binding affinity towards the mutated RNA sequences.[180]

Specificity RNA oligos

Position 2 (**C**3)	K_d (μM)	Position 3 (**A**4)	Kd (μM)
CA**C**AC	2.0 ± 0.4	CAC**A**C	2.0 ± 0.4
CA**A**AC	3.8 ± 0.4	CAC**C**C	8.2 ± 1.0
CA**G**AC	> 20	CAC**G**C	13.1 ± 2.6
CA**U**AC	> 20	CAC**U**C	7.3 ± 1.4

Table 4.1: **RNA oligos used to probe KH3 domain specificity of the central positions – 2 (C4) and 3 (A5)**
Base mutations from 'WT' RNA are identified in red

Results showed that mutating position 2 (C4) to any other nucleobase reduced the binding affinity by 4-7 fold.[180] In context of other KH domains, this is a lower energy penalty than previously reported for NOVA1 KH3 and other KH domains.[71,208] The specificity of A in position 3 was observed to be high with respect to G or U (20-fold affinity difference). However, with respect to C only a 2- to 3-fold reduction in affinity was observed.[180] This weak A/C discrimination is

much lower than that reported for other canonical KH-RNA interactions in which differences in affinity can reach more than 50-fold.[208,209] These findings indicate IMP1 KH3-RNA binding occurs with lower specificity than other studied KH domains.

The results of the ITC assay are consistent with what is observe in the NMR solution structure and can be explained by the specific KH3-RNA contacts established upon binding. The C4 base in position 2 forms hydrogen-bonds between the Watson-Crick edge of the base and the amino acids, Val417 and Arg452, located in the β-sheet and variable loop respectively. The carbonyl oxygen of Val417 forms a direct hydrogen-bond with the N4 group of C4.[180] Arg452 is a conserved residue and establishes hydrogen-bonds with the C4 base in a fashion similar to NOVA1[208] and hnRNP K[209]. The NH1 and NH2 side-chain groups of Arg452 interact with the N3 and O2 moieties of C4 via hydrogen-bonding (Figure 4.3B). A fourth hydrogen-bond is established between the backbone amide of Lys424 and the phosphate of the C4 base. In addition to hydrogen-bonds, the C4 base also makes van-der-Waals contacts with a hydrophobic patch comprising Ala419, Ile420, and Ile421.[180]

A canonical double hydrogen-bond is observed between the Watson-Crick edge of the A5 base and the backbone amide and carboxy groups of Ile441 located in the β-sheet (Figure 4.3C).[180] KH domains typically recognise the Watson-Crick edge of the nucleobase in this position via a third hydrogen-bond. This third hydrogen-bond typically determines specificity in this position for either A or C. This specificity determining third hydrogen-bond is established through a conserved residue located in the second α-helix of the domain. This residue is conserved throughout KH domains to either a Gln or Arg[71] (Figure 4.4). When a Gln is observed in this position a water-mediated hydrogen-bond forms between the N3 moiety of A5, for example as seen for NOVA1.[208] When Arg is located in this position a direct hydrogen-bond can form between the guanidium group of the Arg side chain and the O2 group of a C base, which is the case for KH3 of hnRNP K.[209] In the case of IMP3 KH3 this residue is not conserved to either a Gln/Arg but rather a Ser residue (Ser432). Due to the reduced side-chain length

of the Ser residue with respect to Gln/Arg, the oxygen in the OH group is too far from the nucleobase to form a water-mediated hydrogen-bond (Figure 4.3C). Therefore, a dramatic reduction in the KH3 domain's ability to discriminate between either a C/A in this position was observed.

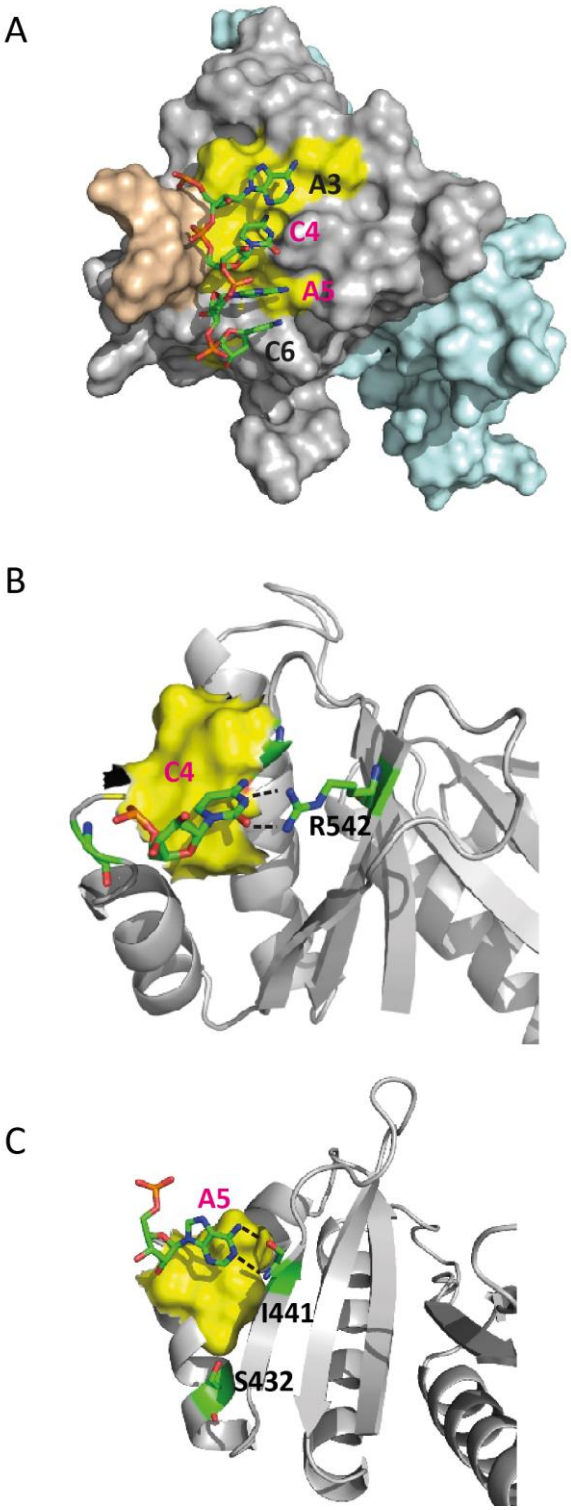


Figure 4.3: NMR complex solution structure of IMP1 KH3KH4DD complex with CACAC highlighting specific contacts which determine specificity
A) KH3KH4DD construct in complex with CACAC oligo highlighting hydrophobic groove in yellow. B) C4 base recognition by contacts with Arg452 residue. C) A5 base recognition via double hydrogen-bond with residue Ile441. Residue Ser432 in α2-helix is highlighted showing distance between amino acid side chain and A5 base.

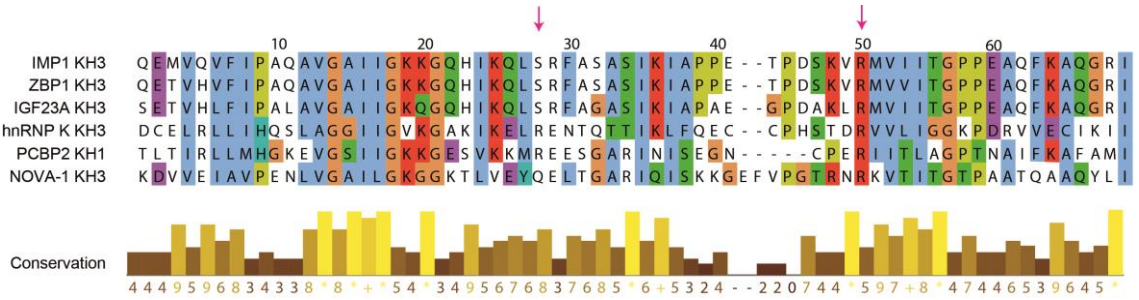


Figure 4.4: Alignment of human, chicken and *Xenopus* IMP KH3 domains with hnRNP K, NOVA1 KH3 and PCBP2 KH1
Sequences were aligned using Clustal Omega and residues coloured according to Clustalx scheme. Conservation scores are displayed below alignments. Conserved R452 and IMP S432 specific residues are highlighted by red arrows.

4.3.2 RNA base specificity and recognition of the KH4 domain

The RNA recognition motif of the IMP1 KH4 domain was identified as UCGGACU by Patel et al., 2012. The ability to bind to G nucleotides in position 1: G3 and position 2: G4 results from the non-canonical hydrophobic groove that is larger and more pronounced (Figure 4.2).[180] This increased size allows for the larger G bases to be inserted into the hydrophobic groove at the central RNA recognition positions (Figure 4.3A).[180] This is similar to the RNA recognition of the non-canonical KH3 domain of KSRP.[207] As with the investigation into the KH3 domain specificity, mutated oligos were generated (Table 4.2) to determine the KH3DDKH4 constructs ability to recognise nucleobases other than G in position 2 (G4) and A in position 3 (A5) via binding affinity measured by ITC.[180]

Specificity RNA oligos

Position 2 (**G**4)	K_d (µM)	Position 3 (**A5**)	Kd (µM)
UCG**G**ACU	1.1 ± 0.1	UCGG**A**CU	1.1 ± 0.1
UCG**A**ACU	> 20	UCGG**C**CU	23 ± 2.6
UCG**C**ACU	NA	(UCGG**G**CU)	NA
UCG**U**ACU	> 20	UCGG**U**CU	20.3 ± 2.8

Table 4.2: RNA oligos used to probe KH4 domain specificity in position 2 (G4) and position 3 (A5)
Base mutations from 'WT' RNA are identified in red. Note UCGGGCU oligo is capable of forming G-quadruplex structures and in turn cannot be used in RNA binding titrations.

Mutation of the G4 nucleobase resulted in a dramatic reduction in binding affinity, over 20-fold reduction for either an A or U in this position. Binding in position 3 (A5) in the KH4 domain displays the typical KH domain A/C discrimination. ITC measurements identified a greater than 20-fold decrease in affinity when A5 is mutated to a C. However, KH4 does not show strong discrimination for a U in this position (only a 4-fold reduction in affinity was observed). These data show KH4 to be more specific than the KH3 domain.

From the complex solution structure: The G4 base makes both hydrophobic and hydrogen-bonding contacts with the KH4 domain. A hydrogen-bond is observed between the O6 moiety of the G4 nucleobase and the NH backbone of the highly conserved Gly500 residue. Asp526 is involved in a network of hydrogen-bonds involving G4 and the residue Arg525 (Figure 4.5B). Asp526 interaction with Arg525 is important for determining the orientation of the Arg525 side-chain. This Arg525 in turn makes a hydrogen-bond to the backbone phosphate of the A5. Therefore, Asp526 is important in the recognition of G4, indirectly determining the specificity of A5. Additionally, the conserved Gln residue at position 514 in the α2-helix of the domain, is predicted to form a water-mediated hydrogen-bond with the N3 moiety of the A5 base. In addition to hydrogen-bonds, the central RNA bases are involved in van-der-Waals contacts provided by the hydrophobic patch consisting of Val502, Ile503, and Val523 located in the α1-helix and the β2-strand.[180]

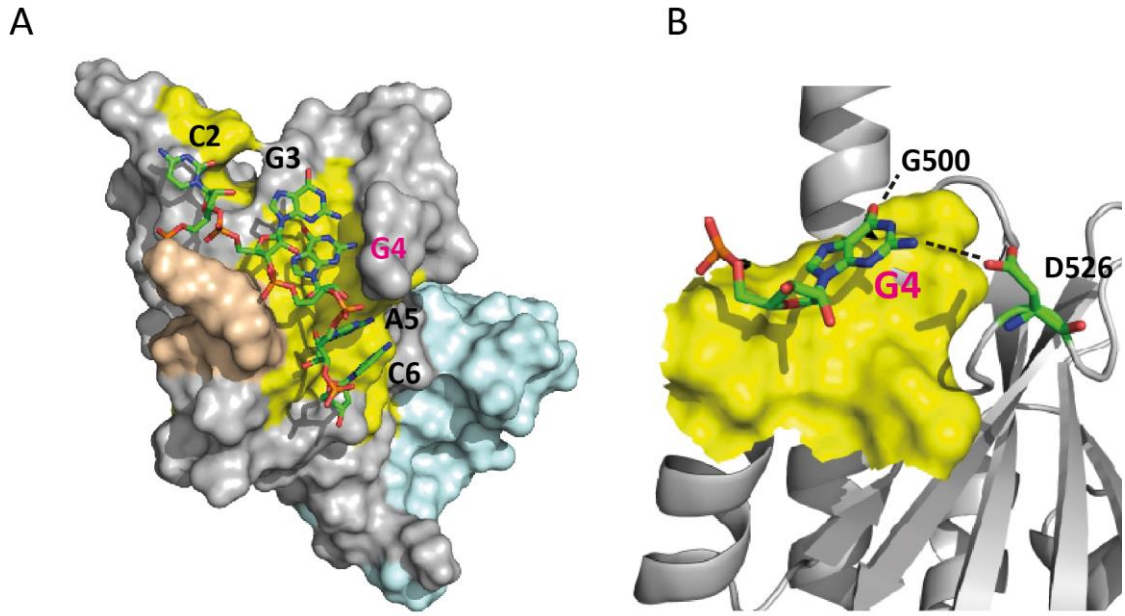


Figure 4.5: NMR complex solution structure of IMP1 KH3DDKH4 complex with UCGGACU highlighting specific contacts which determine specificity
A) KH3DDKH4 construct in complex with UCGGACU oligo with yellow surface indicating the hydrophobic groove. B) Key residues G500 and D526 are highlighted mediating base recognition via hydrogen-bonding. Black dashed lines indicating hydrogen-bonding.

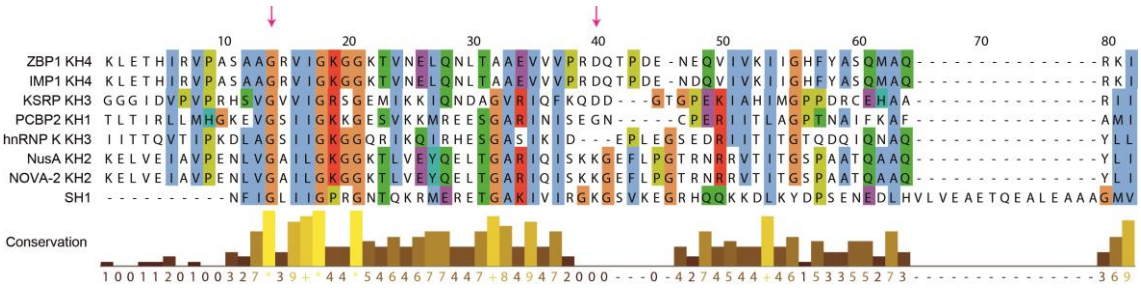


Figure 4.6:Alignment of human and chicken IMP KH3 domains with hnRNP K, KSRP KH3, NusA, NOVA2 KH2, PCBP2 KH1, and SH1 KH domains
Sequences were aligned using Clustal Omega and residues coloured according to Clustalx scheme. Conservation scores are displayed below alignments. Conserved G500 and unconserved D526 residues are highlighted by red arrows.

4.4 IMP1 KH3 and KH4 selectivity mutation rationale

4.4.1 KH3 selectivity mutations S432R and R542G

S342 is located in the α2-helix and the residue does not make contacts with any nucleobase in the KH3 complex structure. In other KH domains this position is typically occupied by an Arg or Gln (Figure 4.4).[71] In IMP1 KH4 for example this residue is a Gln (Q514) (Figure 4.6) and is predicted to form a water-mediated hydrogen-bond with the N3 moiety of the A5 base.[180] This is commonly seen in KH domains with Gln residues in this position. For example, in NOVA1 KH3 a Q406 residue is shown to also form a water-mediated Hydrogen-bond with the A5 base.[208] When an Arg residue occupies this position, the longer side chain results in a direct hydrogen-bond formed between the guanidium group of the Arg side chain and the O2 group of a C base (hnRNPK).[209] The presence of a S432 residue with a shorter side chain compared to Arg or Gln may explain the KH3 domain's lower specificity in this position. Therefore, I plan to generate a S432R mutant to determine if the increased side chain occludes binding of the larger A base and shifts preference to a smaller C nucleobase that would be able to form a direct hydrogen-bond with the Arg side chain (Figure 4.7).

R452 is a conserved residue between KH domains. It aids in the recognition of the C4 nucleobase. The NH1 and NH2 side-chain of the Arg hydrogen-bonded with N3 and O2 of the C4 base. This double hydrogen-bond is also observed in NOVA1 and hnRNPK.[208,209] To investigate the effect of removing this hydrogen-bond with the C4 base I mutated the Arg residue to a Gly to remove the side chain (Figure 4.7). In addition, this permitted us to investigate if reducing the size of the side chain would allow a larger purine base to be accommodated in this position.

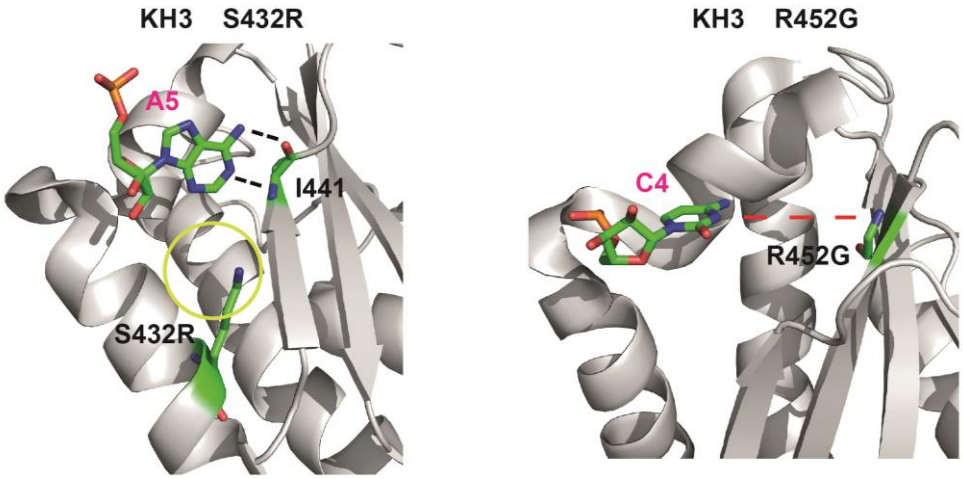


Figure 4.7: IMP1 KH3 domain selectivity mutations (S432R and R452G) modelled in PyMOL
Figure displays a representation of how the incorporated mutations could affect the RNA recognition of the KH3 domain. Left: the increase in the side chain length because of the S432R mutation reduces the hydrophobic space (Yellow circle) in the position the A5 base is recognised. In addition, the longer side chain is placed closer to the A5 base. Right: The R452G mutation removes the Arg side chain that hydrogen-bonds with the base. Red dashed line shows the increased distance between the residue and base.

4.4.2 KH4 selectivity mutations G500A and D526Q

G500 is a highly conserved residue (Figure 4.6) in the α1-helix and is involved in the recognition of position 2 (G4) (Figure 4.5). A hydrogen-bond is observed between the O6 moiety of G4 and the NH of the glycine. The position of the G500 may also restrict binding to C in this position due steric hindrance that would be imposed on the NH2 group of the C base. Mutating G500 to an alanine would introduce an additional CH3 group. This group could occlude the hydrophobic groove in a manner that restricts the binding of larger purine residues (Figure 4.8). The side chain would also provide stronger steric hindrance on the CH3 of a potential C base in this position. However, a hydrogen-bond could potentially form between the O6 of a U, thus shifting preference from a G to a U in this position.

D526 is a nonconserved residue located in the variable loop of the KH4 domain (Figure 4.6). The D526 residue is involved in a network of Hydrogen-bonds

involving the G4 base and the residue R525.[180] The side chain of D526 forms a critical hydrogen-bond with the N2 moiety of G4 (Figure 4.5). Additionally, the hydrogen-bond observed between D526 and R525 establishes the orientation of the R525 residue. This enables the R525 residue to form a hydrogen-bond with the phosphate backbone of the A5 bases.

Introducing a D526Q (Figure 4.8) mutation *in silico* suggests that the side chain of the Q residue would maintain the network of hydrogen-bonds. However, the extended length of the side chain would optimise the hydrogen-bonding distances. Substituting the G4 base for a U or A would maintain these hydrogen-bonds, potentially reducing the specificity for G in this base position.

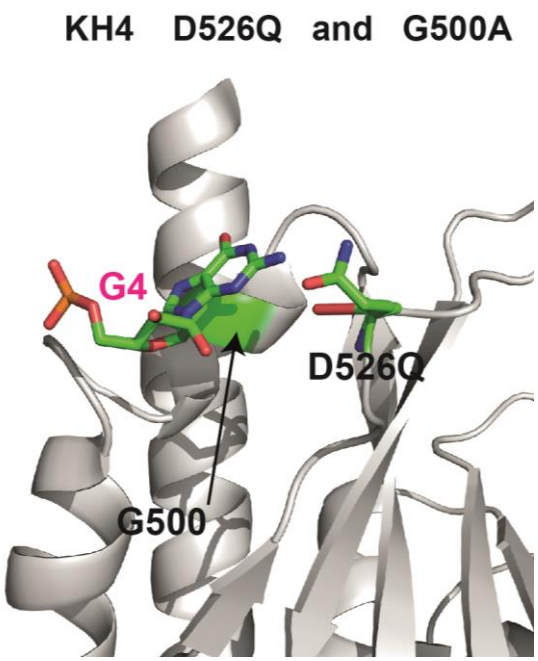


Figure 4.8: IMP1 KH4 domain selectivity mutations (G500A and D526Q) modelled in PyMOL
Figure displays a representation of how the incorporated mutations could affect the RNA recognition of the KH4 domain. Residue G500 is packed close to the G4 base, mutating this residue to an alanine would introduce a side chain that could enforce steric hindrance on the NH2 of the base. D526Q mutation could potentially still hydrogen bond with the G4 base.

4.5 Effects of selectivity mutations on protein structure and stability

To use our KH3 and KH4 domain mutants as molecular tools to study RNA selectivity *in vivo,* the mutations must not significantly perturb the structure and stability of the domain. This would enable the investigation of how KH domain RNA sequence specificity influences overall RNA target selection without disrupting additional functional roles the correctly folded domain may be involved in *in vivo*. I intended to use mutants which alter the domain's specificity in an iCLIP based study similar to what was implemented for the KH domain knock outs. This requires protein expression in HeLa cells at 37^{o}C for up to 48h. Therefore, I first determined if the incorporated selectivity mutations influenced overall protein folding and thermal stability.

4.5.1 Secondary structure of the IMP1 KH3 and KH4 domains is not altered by selectivity mutations

All selectivity mutant proteins were expressed and purified as reported in.[180] Mutants expressed with slightly lower yields than was observed for the KH3KH4DD and KH3DDKH4 proteins. Nickel affinity purification and cation chromatography in combination was sufficient to produce protein constructs that were pure from contaminating proteins and nucleic acid impurities. An example of a SDS-PAGE analysis of fractions collected during the purification protocol is shown for the S432R mutant in Figure 4.9.

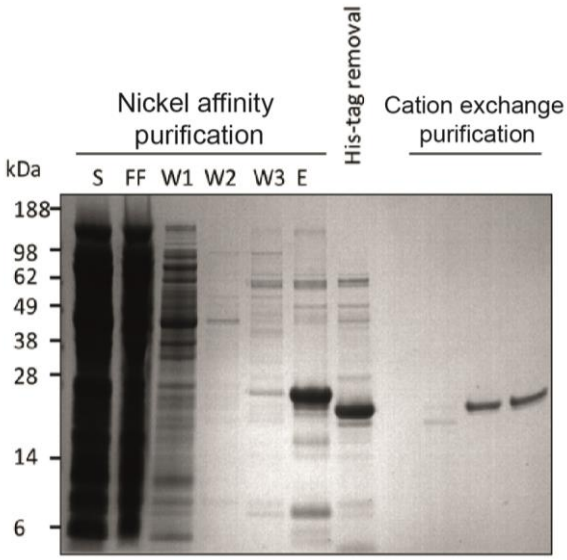


Figure 4.9: Selectivity mutant KH3KH4DD S432R example purification SDS-PAGE gel
First half of gel displays nickel affinity purification (S; bacterial supernatant, FF; nickel resin flow through, W1 & W2; Wash 1 & 2, and E; nickel resin elution). Histag removal lane shows sample after TEV digestion. Second half of gel shows fractions collected from FPLC cation exchange chromatography purification.

Secondary structure content of the selectivity mutants was determined via far-UV CD spectra analysis. As reported, the GDDG mutations required to inhibit RNA binding of the partner KH domain do not alter the structure of the domain.[82] Therefore, I assessed if the selectivity mutations altered protein structure in relation to the KH3DDKH4 and KH3KHDD4 constructs via comparison of far-UV CD spectra.

Far-UV CD spectra were recorded (190 nm – 260 nm at 25^{o}C) on the KH34 constructs buffered in 10 mM NaCl, 10 mM sodium phosphate, 0.5 mM TCEP, pH 6.4 at a concentration of 0.15 mg/ml. I observed strong minima at ~208 nm and a less pronounced minima at ~225 nm (Figure 4.8). This is indictive of α-helical structures. However, the reduction in signal at 225 nm results from the presence of β-sheet structures. In turn, this pattern of far-UV CD spectra shows our constructs to be α-helix / β-sheet mixture. Our findings are consistent with the reported structures of the KH34 constructs.[85,180]

The far UV CD spectra for the KH3DDKH4 and KH3KH4DD constructs resulted in identical overlaying spectra. This is expected as the GDDG mutations have minimal effect on secondary structure.[82] Comparing these CD spectra to the selectivity mutants (S432R, R452G, G500A and D526Q) shows incorporation of these selectivity mutations does not modify secondary structure (Figure 4.10).

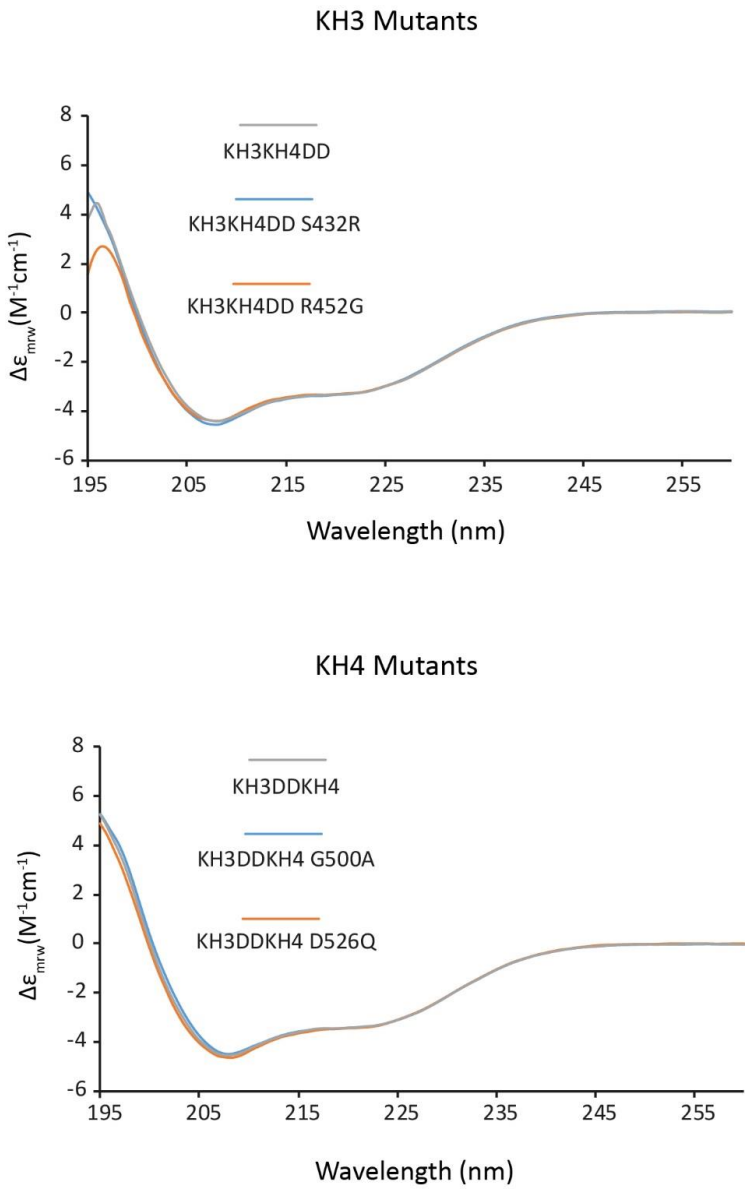


Figure 4.10: Far UV CD spectra of selectivity mutants and WT KH34 constructs
Protein samples were prepared in 10 mM NaCl, 10 mM sodium phosphate, 0.5 mM TCEP, pH 6.4 at a final concentration of 0.15 mg/ml. Spectra are a accumulation of 50 accumulated scans. Top: KH3KH4DD WT spectra overlayed with KH3 specificity mutants, S432R and R452G. Bottom: KH3DDKH4 WT spectra overlayed with KH4 domain specificity mutants, G500A and D526Q. Inserted legend displays colour code for each construct reported.

4.5.2 Selectivity mutants are stable within the range of temperatures typically used for *in vivo* assays

The effects of the selectivity mutations on thermal stability were determined by performing CD thermal denaturation. Proteins were buffered and diluted to the same conditions as used for the far-UV CD analysis. Proteins were heated from 10ºC to 90ºC with a 2ºC/min temperature gradient. Protein folding was reported by measuring the CD signal at 222 nm.

I observed a varying degree of cold denaturation between the constructs, with R452G being particularly extreme (Figure 4.11 and 4.12). Previous thermal stability studies on IMP1 KH3KH4 reported the construct to be sensitive to cold denaturation.[82] For our investigation I was interested as to whether the mutations I incorporated would destabilise the proteins at temperatures typically used for *in vivo* (e.g. iCLIP) studies. Therefore, I discounted the region of the spectra that showed cold denaturation and did not investigate this property of the KH34 domains further. After discounting this region, I calculated T_m values for each construct and compared them to their associated WT proteins (Figure 4.11 and 4.12). For all constructs there was an observed minimal change in signal between RT and 37ºC. Previous studies showed the T_m for the KH3DDKH4 and KH3KH4DD constructs to be 60ºC and 54ºC, respectively.[82] The stabilising effect of the KH3DD mutation on thermal stability by ~5ºC could not rationally be explained. I observed similar results for our thermal stability analysis obtaining T_m vales of 59.6ºC and 57.0ºC for KH3DDKH4 and KH3KH4DD respectively. As the KH3DD mutation increased the thermal stability of the construct, rather than destabilising the domain, I concluded this effect should not be an issue for later studies.

When comparing the T_m of the selectivity mutants in relation to their KH3KH4DD or KH3DDKH4 counterparts, I observed that the mutations in the KH3 domain resulted in a decrease in thermal stability with variable degrees. The S432R mutation reported a T_m of 53.3ºC with the R452G mutation resulting in a more dramatic reduction with a T_m of 49.5ºC. In contrast, the G500A mutation within

the KH4 domain resulted in a slight increase in thermal stability with a reported T_m of 61.8ºC. The mutant D526Q had a calculated T_m of 55.6ºC, which is 4ºC lower than the KH3DDKH4 construct. However, the WT KH3KH4 protein has a T_m of ~55ºC and so this slight reduction in relation to the KH3DDKH4 construct is within the range of T_m values previously reported.[82] Although here I report a varying degree of differences in the calculated T_m values for our selectivity mutant proteins, the constructs remain stable in the range of temperatures used for *in vivo* studies (~37ºC),

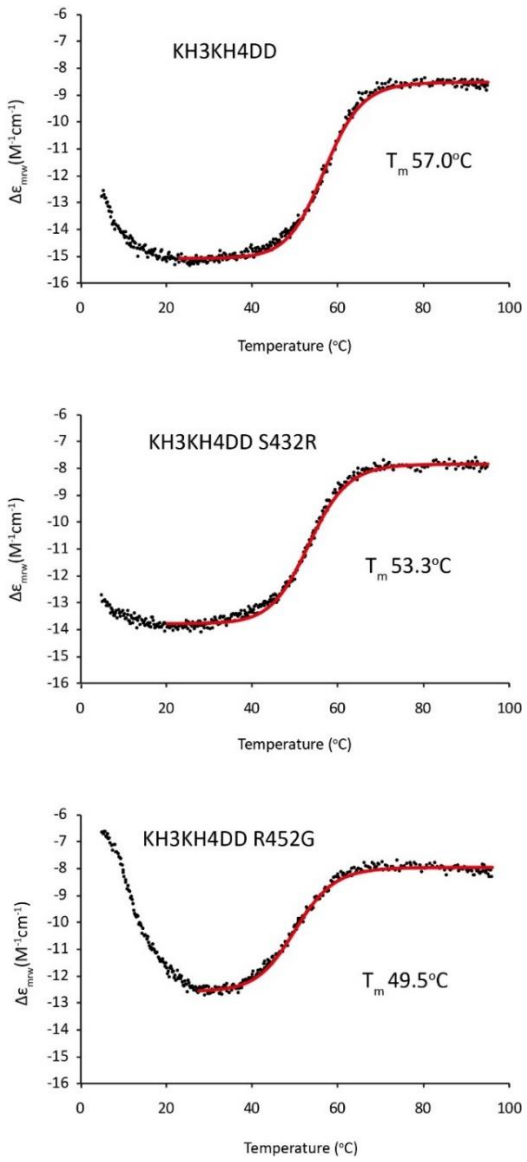


Figure 4.11: CD thermal denaturation of KH3 domain selectivity mutants and KH3KH4DD WT protein

Protein samples were prepared in 10 mM NaCl, 10 mM sodium phosphate, 0.5 mM TCEP, pH 6.4 at a final concentration of 0.15 mg/ml. Samples were heated from 2°C to 95°C at a rate of 2°C/min and monitored at 222 nm. Black dots represent raw data points and red line is the fitted curve. Note red curve was used for T_m calculation and cold denaturation region was discounted. Calculated T_m values are displayed in each graph.

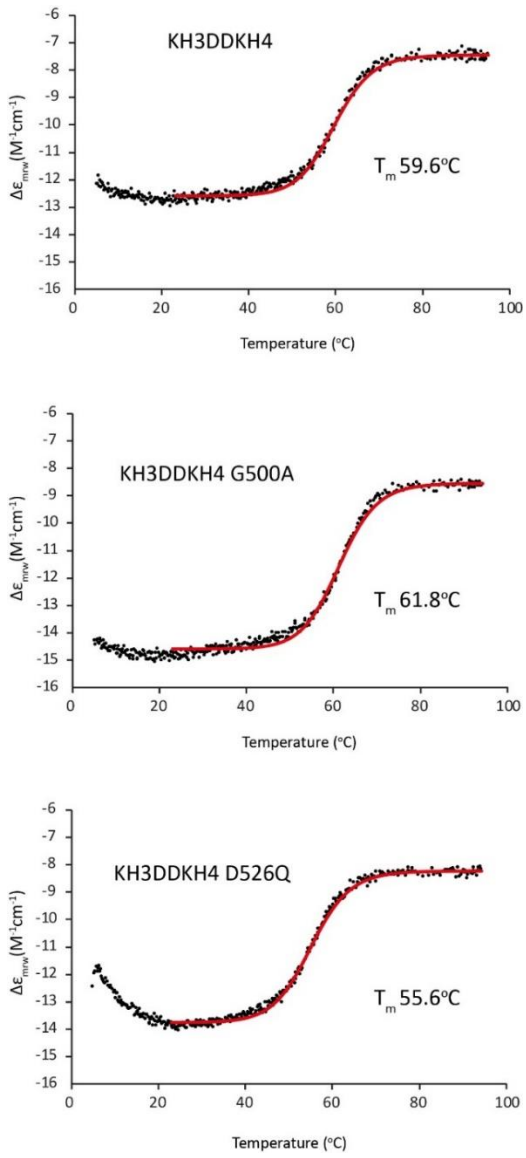


Figure 4.12: CD thermal denaturation of KH4 domain selectivity mutants and KH3DDKH4 WT protein

Protein samples were prepared in 10 mM NaCl, 10 mM sodium phosphate, 0.5 mM TCEP, pH 6.4 at a final concentration of 0.15 mg/ml. Samples were heated from 2°C to 95°C at a rate of 2°C/min and monitored at 222 nm. Black dots represent raw data points and red line is the fitted curve. Note red curve was used for T_m calculation and cold denaturation region was discounted. Calculated T_m values are displayed in each graph

4.5.3 IMP1 KH3 and KH4 selectivity mutants display ^{1}H-^{15}N correlation spectra comparable to WT proteins

To further characterise the effects of the selectivity point mutations on protein structure I performed ^{1}H-^{15}N SOFAST-HMQC experiments on ^{15}N labelled constructs and compared the resultant spectra with the KH3DDKH4, KH3KH4DD constructs. Samples were prepared in 100 mM NaCl, 10 mM sodium phosphate, 0.5 mM TCEP, pH 6.4, with protein concentrations between 50 µM and 100 µM. All NMR experiments were performed at 25^{o}C.

Comparing the spectra of the selectivity mutants with the KH3KH4DD, KH3DDKH4 proteins (Figure 4.13) revealed several peak shifts between spectra. The overall number of peaks remained similar, suggesting the mutations did not result in unfolding or partial unfolding of the domain, as unfolded regions would not produce dispersed peaks.

I did observe a larger number of peaks in the central region of the SOFAST-HMQC spectra for the R452G mutant. This could suggest the presence of an aggregated subspecies of protein. The D526Q mutation resulted in slightly more altered peaks compared to the other three selectivity mutations. This could result from the D526 residue being involved in a network of hydrogen-bonds which involves contacts with residues such as R525. The D526Q mutation results in the incorporation of an amino acid side chain with increased length, but is potentially maintaining the capability of establishing the same hydrogen-bonding network as observed for the WT D526 residue. This may result in a slight change in the local environment of residues connected to this bonding network and in turn, result in the more pronounced differences observed. However, the high degree of similarity between far-UV CD spectra of D526Q with the KH3DDKH4 construct suggests minimal changes in overall secondary structure content (Figure 4.10).

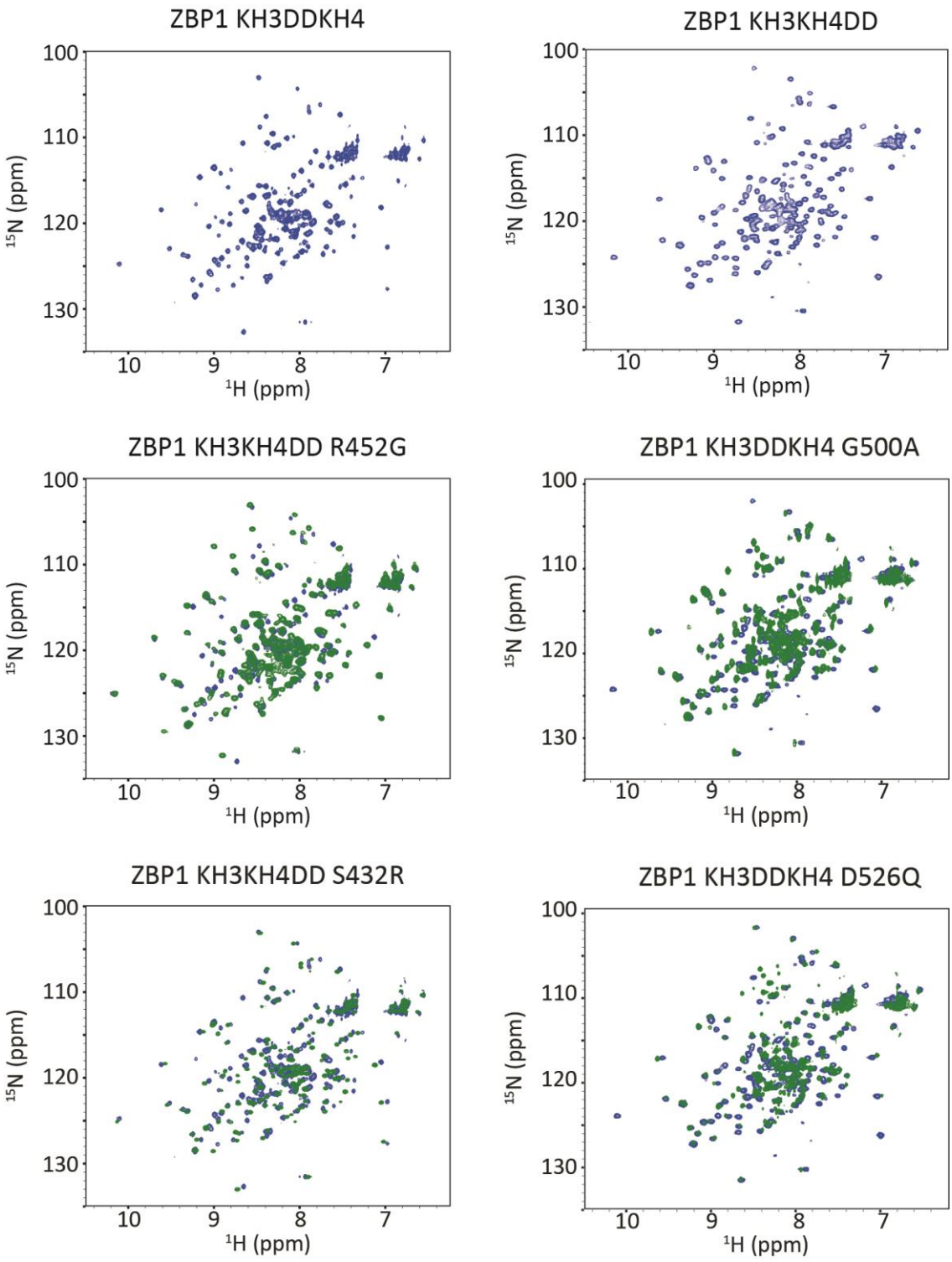


Figure 4.13: Comparison of the structure of selectivity mutants to corresponding WT KH3KH4 constructs
Selectivity mutant ^{1}H-^{15}N SOFAST-HMQC spectra (Green) are overlaid onto corresponding KH3KH4DD or KH3DDKH4 constructs (Blue). Top: KH3DDKH4 and KH3KH4DD ^{1}H-^{15}N SOFAST-HMQC displayed without mutant overlay.

The NMR spectra report on protein structure with a higher degree of resolution than our CD studies. This is due to the fact that protein tertiary structure also influences NMR spectra while only secondary structure is monitored in the CD measurements, and offers an explanation for the differences seen in the NMR compared to the CD measurements. Additionally, the previously published investigation into the design and function of the GDDG mutations reported very minimal changes in ^{1}H-^{15}N correlation spectra of mutated proteins compared to the WT KH3KH4 constructs.[82] Here I observe more substantial changes (Figure 4.13). However, it is worth noting that the mutations in the GXXG loop reside in the variable loop region which has a high degree of flexibility compared to the hydrophobic groove. The mutations I have incorporated for the selectivity mutants are located in the highly structured hydrophobic groove of the protein. In turn, it is expected that mutations in this highly structured region would affect protein structure to a higher degree than the GDDG mutation.

Overall our mutations do not destabilise the KH3KH4 constructs to a degree that would prove problematic for *in vivo* studies (Figure 4.11 and 4.12). The secondary structure content of the mutated proteins was identical to the WT KH3KH4 constructs as shown by the far-UV CD spectra (Figure 4.10). Using NMR to compare protein folding highlighted some differences in the spectra of the mutated proteins. However, due to the nature of the location of these point mutations these changes were expected. From our results I can conclude that all mutants are folded, and the overall protein fold is similar to that of the WT KH3KH4 protein.

4.6 Determining RNA binding specificity of the selectivity mutants

To determine the RNA binding specificity of each of the selectivity mutants I planned to perform ITC titrations in the same fashion as was carried out with the WT KH34 constructs.[180] However, this method required titration of protein into RNA. To achieve this, protein samples needed to be concentrated to ~300 µM. This proved challenging due to the tendency of the constructs to aggregate

at higher concentrations. To overcome this issue, I tried titrating RNA into the protein, and so protein concentrations in the range of 50 – 100 μM would be required. This highlighted another issue which is typically faced when using ITC to measure protein-RNA binding. The heat of dilution of the RNA oligos was, in some cases, several times larger than the energy of binding between the protein and RNA, making it impossible to produce accurate binding curves. However, I was able to produce a full set of reliable ITC titrations with one construct, the S432R mutant, for which protein was injected into RNA.

To compare the binding affinity of the remaining mutants I used NMR titrations to calculate K_d values. For these mutants (R452G, G500A, and D526Q) ^{15}N labelled proteins (~60 μM) were titrated with unlabelled RNA oligos. ^{1}H - ^{15}N SOFAST-HMQC experiments were recorded at 25^{o}C for titration points with increasing molar protein:RNA ratios. Chemical shift perturbations were then measured to calculate a binding K_d. This was repeated for each selectivity mutant with all four RNA oligos where the base position being investigated was changed (explained below).

The binding K_d of the selectivity mutants with each RNA oligo was then compared to the binding affinity of the WT KH3KH4 protein with the same corresponding RNA oligo.[180] In order to account for the different techniques used to determine binding K_d values (NMR and ITC) we implemented an approach to compare relative K_d values. Relative K_d values were calculated in relation to the proteins affinity towards the 'WT' RNA, the affinity towards the mutated oligos were then reported as a factor of this affinity.

4.6.1 S432R mutation shifts specificity of RNA target sequence in position 3 from an A to a C

The rationale for the S432R mutation was based on amino acid sequence conservation, and protein-RNA structures of well studied KH domains. Here we identified a Ser432 residue in the α2-helix of IMP1 KH3 where typically an Arg or Gln is located.[208,209] This residue influences the size of the hydrophobic binding

groove and is involved in either water-mediated or direct hydrogen-bonding to the nucleobase in position 3. We investigated the effect of RNA selectivity by introducing an Arg at this position.

We were able to produce reliable ITC titrations to calculate the binding affinity of the S432R mutant protein with the modified RNA oligos (Figure 4.14). ITC titrations showed the S432R mutation resulted in a reduction in affinity for the preferred CACAC sequence compared to the WT protein (8.7 µM and 2.0 µM respectively). Weak binding to oligos with either G or U in position 3 was observed (>20 µM). However, we see an increase in binding when a C is in position 3 instead of an A (CACCC-2.0 µM, CACAC-8.7 µM) (Figure 4.12).

Comparing the relative K_d values for the modified RNA sequences we observe an increase in the binding affinity towards the CACCCA oligo by a factor of 5 (Figure 4.13). This is due to the incorporation of the large Arg side chain reducing the size of the binding groove in this region and thus hindering the binding of a large purine base in position 3. The C base is preferred due to its smaller structure compared to A and G, in addition to C being able to form a direct hydrogen-bond between its O2 moiety and the guanidium group of the Arg side chain. Furthermore, the actual binding K_d of the S432R mutant with the CACCC sequence is 2 µM. This is the same K_d calculated for the KH3KH4DD protein and the CACAC oligo (Figure 4.15). In turn, the S432R mutant shifts specificity from an A in position 3 to a C by a factor of 5, but the actual affinity is equal to the affinity of the WT protein to the preferred RNA sequence.

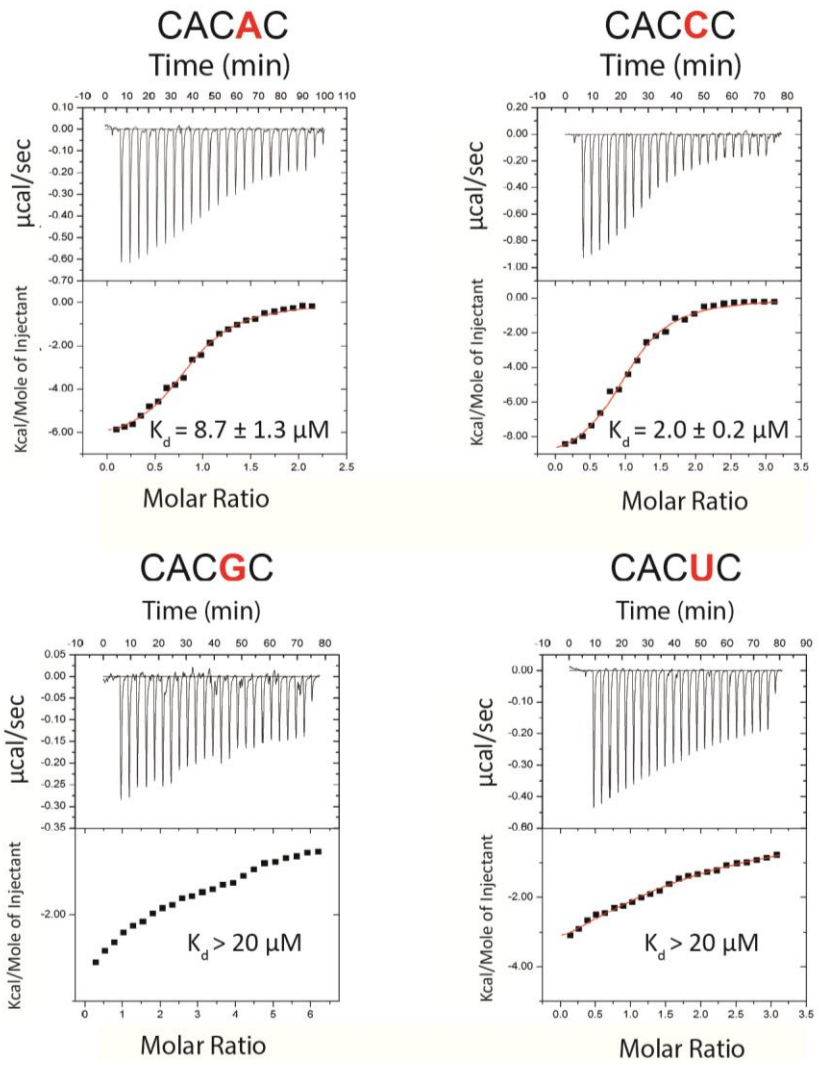


Figure 4.14: ITC titration panels of KH3KH4DD S432R mutant with RNA oligo where base position 3 (A5) is mutated
Proteins were concentrated to 300 µM in 100 mM NaCl, 10 mM sodium phosphate, 0.5 mM TCEP, pH 6.4 buffer and were injected into ITC cell containing 30 µM RNA in the same buffer. Titrations were performed at 25ºC. Raw data are shown above with integration of peaks shown below with respect to protein:RNA molar ratio. Red line shows fit of data points and was used to calculate K_d. Note the affinity of binding for the CACGC and CACUC oligos could not be determined with the range of concentrations used in these titrations. In turn, K_d is reported as greater than 20 µM.

A

	K_d (μM)		Affinity Factor	
Oligo	S432R	WT	S432R	WT
CACAC	8.7 ± 1.3	2.0 ± 0.4	1.0	1.0
CACCC	2.0 ± 0.2	3.8 ± 0.4	0.2	1.9
CACGC	>20	>20	>2.3	>10
CACUC	>20	>20	>2.3	>10

B

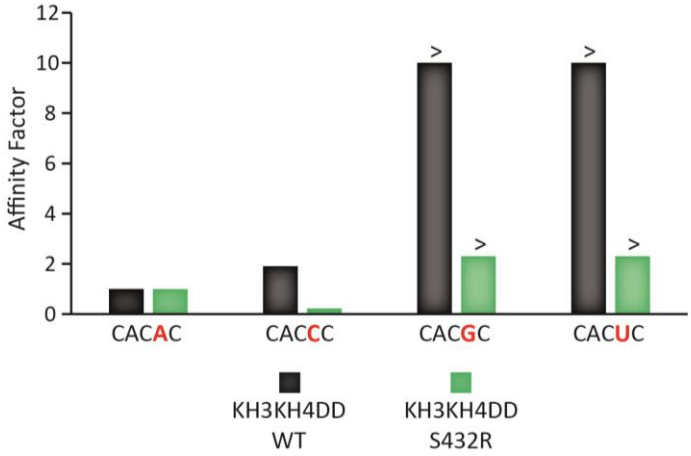


Figure 4.15: Summary of calculated K_d values for KH3KH4DD and S432R constructs and binding preference of KH3DDKH4 S432R relative to KH3KH4DD
A) Table of calculated K_d values with errors (blue columns) and relative binding affinities as a factor of binding relative to CACAC RNA oligo (orange columns). B) Bar chart displays relative binding affinity of KH3KH4DD and S432R mutant.

4.6.2 R452G mutation reduces RNA specificity of the KH3 domain

The effect of the R452G mutation on the recognition of the KH3 domain to bind the C4 nucleotide was investigated using NMR titrations with CA**C**AC and CA**A**AC; CA**G**AC; and CA**U**AC. Protein:RNA titration molar ratios ranged from 0.5 to 8 depending on observed affinity (Figure. 4.16).

I investigated the effect of removing the side chain of the conserved R452 residue by incorporating the R452G mutation. The Arg residue in this position forms a double hydrogen-bond with the C4 nucleobase, similar to what is observed in NOVA1 KH3 and hnRNPK.[208,209] By removing this double hydrogen-bond I

wanted to see if the KH3 domain would have a reduced binding preference for a C nucleobase.

NMR titrations showed the preferred CACAC RNA sequence to bind to the R452G and KH3KH4DD proteins with comparable affinities (3.1 μM, and 2.0 μM respectively). This was unexpected due to the loss of a double hydrogen-bond with the C4. I observed similar affinities also for the CAAAC oligo. Interestingly, the biggest difference in binding was observed with CAGAC where a 2-fold increase in binding preference was observed compared to the WT protein (Figure 4.15). The removal of the long Arg side chain could potentially alter the local shape of the hydrophobic groove in the region which accommodates the RNA base in position 2. In turn, the larger G base could be incorporated into this position more favourably than in the WT protein. I also observe CAUCA binding with an affinity that is relatively equal to that of the WT sequence.

To compare the KH3KH4DD R452G RNA binding preference against the KH3KH4DD protein I calculated relative affinities where binding to the CACAC RNA sequence was set at 1. By comparing affinities in this manner, I account for the fact that the K_d values were calculated using different methods (KH3KH4DD via ITC and R452G via NMR).

The KH3KH4DD protein binds the CACAC sequence with ~4-fold higher affinity than the next best binding sequence CAUAC. The R452G mutation reduces the specificity of the KH3 domain, with the least preferred RNA sequence for this mutation being CAAAC and there being only a 3-fold difference in binding with this sequence compared to the preferred CACAC (Figure 4.17). Although I do not shift the sequence preference of the KH3 domain with the R452G mutation, I reduce overall specificity for all bases in position 3. I also see comparable binding affinity for the R452G mutation towards the RNA oligos which is surprising due to the loss of a double hydrogen-bond.

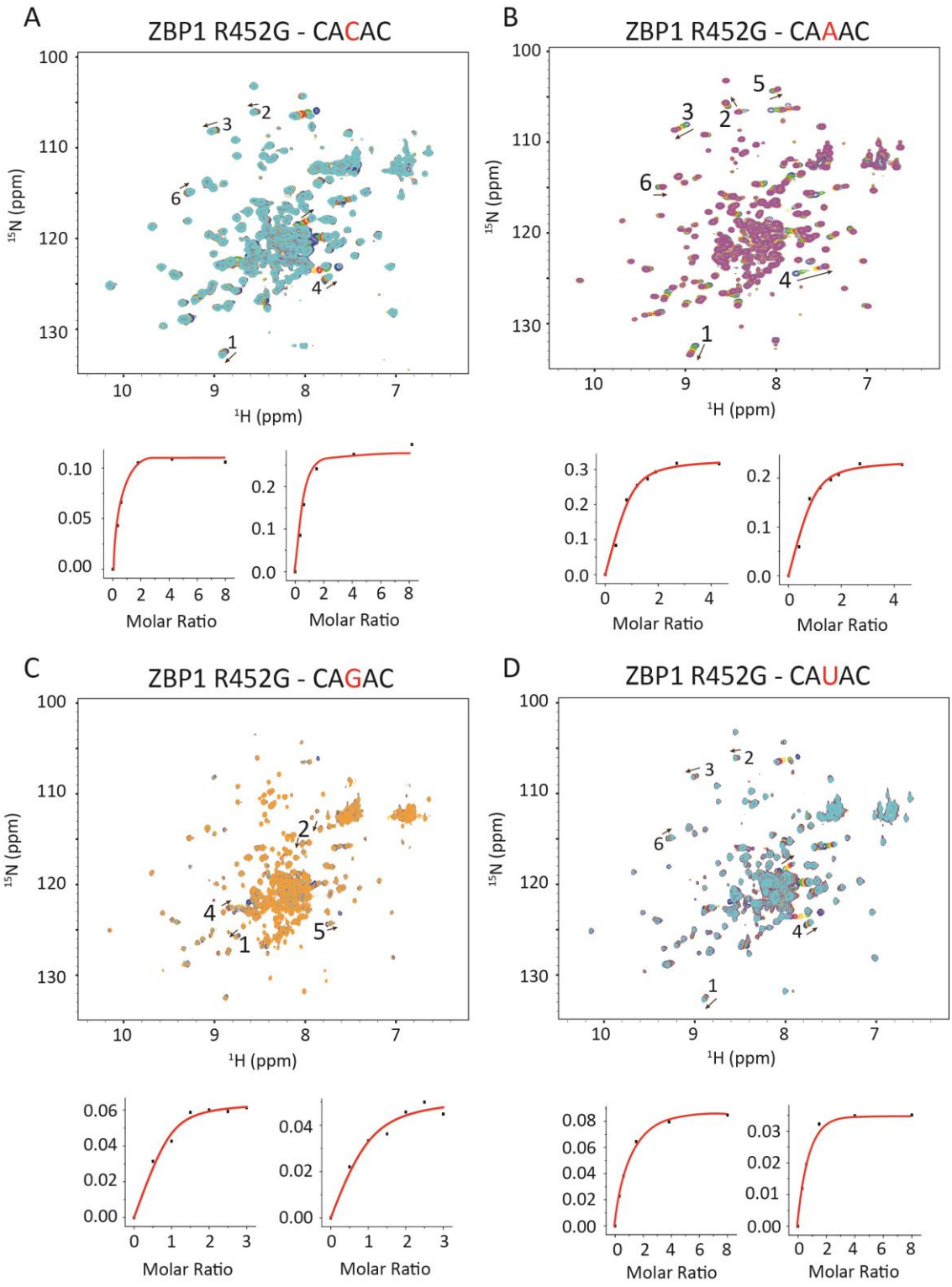


Figure 4.16: NMR ^{1}H-^{15}N SOFAST-HMQC titrations and binding curves for KH3DDKH4 R452G with RNA oligos in which position 2 (C4) is mutated
All proteins were buffered in 100 mM NaCl, 10 mM sodium phosphate, 0.5 mM TCEP, pH 6.4 and were concentrated to between 50 – 60 µM. NMR experiments were performed at 25ºC A) R452G selectivity mutant SOFAST-HMQC overlaid spectra with oligonucleotide CACAC at protein:RNA ratios of 1:0 (Blue), 1:05

(Green), 1:1 (Red), 1:2 (Yellow), 1:4 (Purple) and, 1:8 (Cyan). B) R452G titration with CAAAC at protein:RNA ratios of 1:0 (Blue), 1:0.5 (Green), 1:1 (Red), 1:1.5 (Yellow), 1:1.75 (Purple), 1:2 (Cyan), 1:3 (Orange), and 1:4 (Maroon). C) R452G titration with RNA CAGAC at protein:RNA ratios of 1:0 (Blue), 1:0.5 (Green), 1:1 (Red), 1:1.5 (Yellow), 1:2 (Purple), 1:2.5 (Cyan) and, 1:3 (Orange) D) R452G titration with RNA oligo CAUAC at protein:RNA ratios of 1:0 (Blue), 1:05 (Green), 1:1 (Red), 1:2 (Yellow), 1:4 (Purple) and, 1:8 (Cyan). Each panel also displays chemical shift perturbations upon addition of increasing molar concentrations of RNA.

A

	K_d (μM)		Affinity Factor	
Oligo	R452G	WT	R452G	WT
CACAC	1.8 ± 0.8	2.0 ± 0.4	1.0	1.0
CAAAC	7.7 ± 1.9	8.2 ± 1.0	3.4	4.1
CAGAC	6.4 ± 3.4	13.1 ± 2.6	3.6	6.6
CAUAC	3.9 ± 1.6	7.3 ± 1.4	2.2	3.7

B

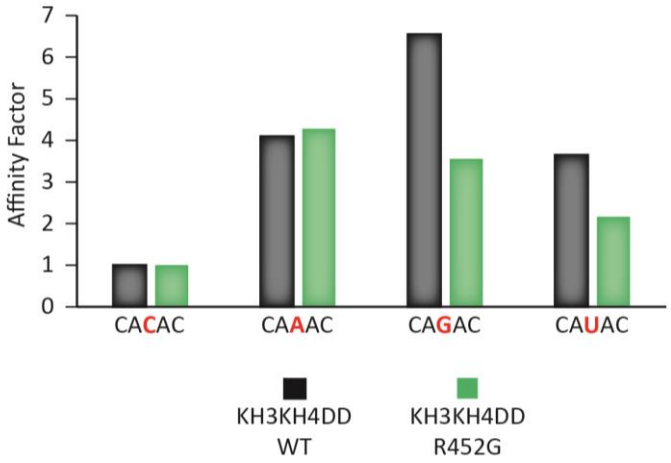


Figure 4.17: Summary of calculated K_d values for KH3KH4DD and R452G constructs and binding preference of KH3DDKH4 R452G relative to KH3KH4DD

A) Table of calculated K_d values with errors (blue columns) and relative binding affinities as a factor of binding relative to CACAC RNA oligo (orange columns). B) Bar chart displays relative binding affinity of KH3KH4DD and R452G mutant.

4.6.3 G500A mutation reduces overall RNA binding affinity of the KH4 domain

I investigated the effect of the G500A mutation on the recognition of the KH4 domain to bind the G4 base position. NMR titrations were performed with UCG**G**ACU and UCG**A**ACU, UCG**C**ACU, UCG**U**ACU oligonucleotides. Protein:RNA titration molar ratios ranged from 0.2 to 8 depending on observed affinity.

The incorporation of the alanine side chain was predicted to potentially occlude the hydrophobic groove in a manner that restricts the binding of larger purine residues. Additionally, the side chain would also provide stronger steric hindrance on the NH2 of a potential C base in this position. However, a hydrogen-bond could form between the O6 moiety of a U, thus shifting preference from a G to a U in this position.

The results of the NMR titrations show that titrations with UCGAACU and UCGCACU do not saturate with protein:RNA molar ratios up to 1:8 as binding curves remain in the linear phase (Figure 4.18). In turn, accurate K_d values could not be measured for these oligonucleotides with the RNA molar ratios used. For these oligos a K_d in excess of 300 µM is reported. The only oligo I observed binding for (other than the UCGGACU sequence) was UCGUACU, with the protein binding approaching saturation around with a 1:6 RNA molar ratio (Figure 4.18). The preferred RNA sequence (UCGGACU) bound with a much higher affinity and binding saturation was observed around a 1:3 RNA molar ratio. In addition, some peaks were in slow to intermediate exchange.

The G500A mutation did not result in an observable change in RNA binding preference. Overall the incorporation of the alanine side chain reduced RNA binding affinity, with a 5-fold reduction in affinity towards the preferred RNA binding sequence compared to the WT protein (5.6 µM and 1.1 µM K_d, respectively) (Figure 4.19). This confirms our prediction of the increased length of the side chain generating greater steric hindrance on the NH2 group of an A

or C bases in this position. The mutation was ineffective in shifting the binding preference from a G to a U in position 3 as both the G500A and KH3DDKH4 protein display a similar binding preference for the preferred UCGGACU sequence.

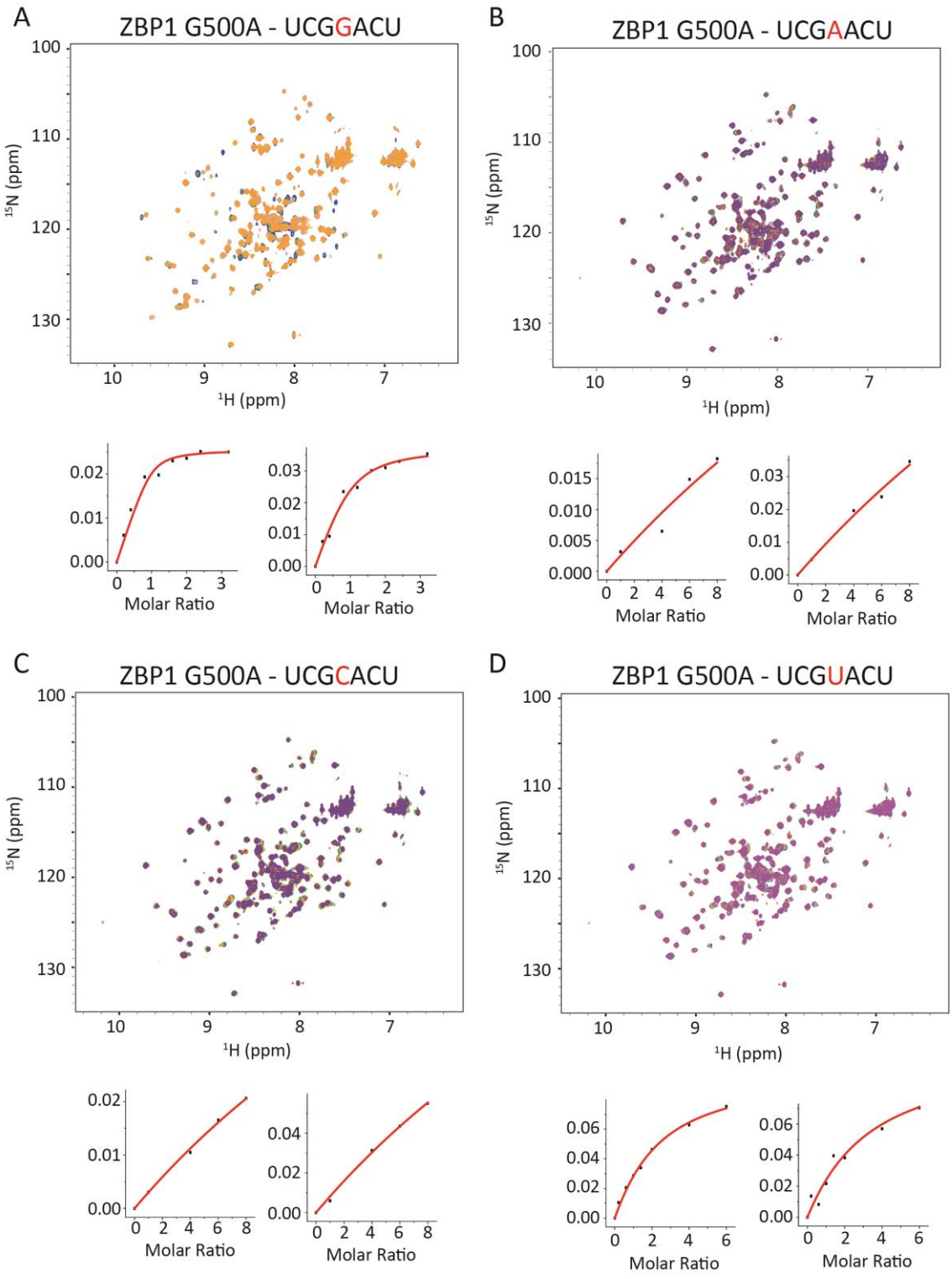


Figure 4.18: NMR ^{1}H-^{15}N SOFAST-HMQC titrations and binding curves for KH3DDKH4 G500A for RNA oligos in which base in position 2 (G4) is mutated

All proteins were buffered in 100 mM NaCl, 10 mM sodium phosphate, 0.5 mM TCEP, pH 6.4 and were concentrated to 50-60 µM. NMR experiments were performed at 25°C A) G500A selectivity mutant SOFAST-HMQC overlaid spectra

with oligonucleotide UCGGACU at protein:RNA ratios of 1:0 (Blue), 1:0.25 (Green), 1:0.5 (Red), 1:1 (Yellow), 1:1.25 (Purple), 1:5 (Cyan), 1:2 (Orange), 1:2.5 (Maroon), and 1:3 (Gold). B) G500A titration with UCGAACU at protein:RNA ratios of 1:0 (Blue), 1:1 (Green), 1:4 (Red), 1:6 (Yellow) and, 1:8 (Purple). C) G500A titration with RNA UCGCACU at protein:RNA ratios of 1:0 (Blue), 1:1 (Green), 1:4 (Red), 1:6 (Yellow) and, 1:8 (Purple). D) G500A titration with RNA oligo UCGUACU at protein:RNA ratios of 1:0 (Blue), 1:0.25 (Green), 1:0.5 (Red), 1:1 (Yellow), 1:1.5 (Purple), 1:2 (Cyan), 1:4 (Orange) and, 1:6 (Maroon). Each panel also displays chemical shift perturbations upon addition of increasing molar concentrations of RNA, with exception of (B) and (C).

A

	K_d (μM)		Affinity Factor	
Oligo	G500A	WT	G500A	WT
UCGGACU	5.6 ± 1.8	1.1 ± 0.1	1.0	1.0
UCGAACU	>300	23 ± 2.6	> 54	20.9
UCGCACU	>>300	N/A	>>54	N/A
UCGUACU	115 ± 58	20.3 ± 2.8	20.5	18.5

B

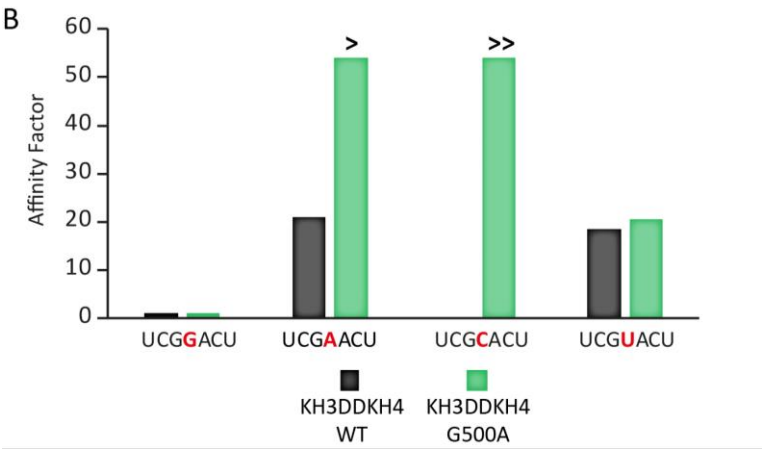


Figure 4.19: Summary of calculated K_d values for KH3DDKH4 and G500A constructs and binding preference of KH3DDKH4 G500A relative to KH3DDKH4 A) Table of calculated K_d values with errors (blue columns) and relative binding affinities as a factor of binding relative to UCGGACU RNA oligo (orange columns). B) Bar chart displays relative binding affinity of KH3DDKH4 and G500A mutant.

4.6.4 D526Q mutation increases RNA binding affinity of the KH4 domain

I investigated the effect of the D526Q mutation on the recognition of the KH4 domain to bind the G4 base. NMR titrations were performed with UCG**G**ACU (preferred sequence) and UCG**A**ACU, UCG**C**ACU, UCG**U**ACU. Protein:RNA titration molar ratios ranged from 0.2 to 8 depending on observed affinity (Figure 4.20).

The D526Q mutation was intended to optimise hydrogen-bonding distances due to the increase in amino acid side chain length. The network of hydrogen-bonds was expected to be maintained, in addition to potential hydrogen-bonds being able to form with U or A nucleotides in position 2.

NMR titrations revealed an increase in affinity for the D526Q mutant with all RNA sequences respect to the KH3DDKH4 protein (Figure 4.21). The preferred sequence UCGGACU bound with a 3-fold higher affinity for the D526Q mutant compared to the KH3DDKH4 (0.6 μM and 1.1 μM affinity, respectively).

Comparing relative affinities with the KH3DDKH4 construct and D526Q I saw an increase in preference for binding A in position 2 (Figure 4.19). However, the D526Q mutant has a slight reduction in preference for U in position 2 (Figure 4.14). Overall, I observe tighter RNA binding for the D526Q mutant. The preferred RNA sequence UCGGACU remains the most preferred sequence by at least 1 order of magnitude compared to the other RNA sequences tested (Figure 4.19). In turn, optimising the hydrogen-bonding network of the Asp526 residue via increasing the amino acid side chain length did not alter specificity, but increased overall RNA binding affinity instead.

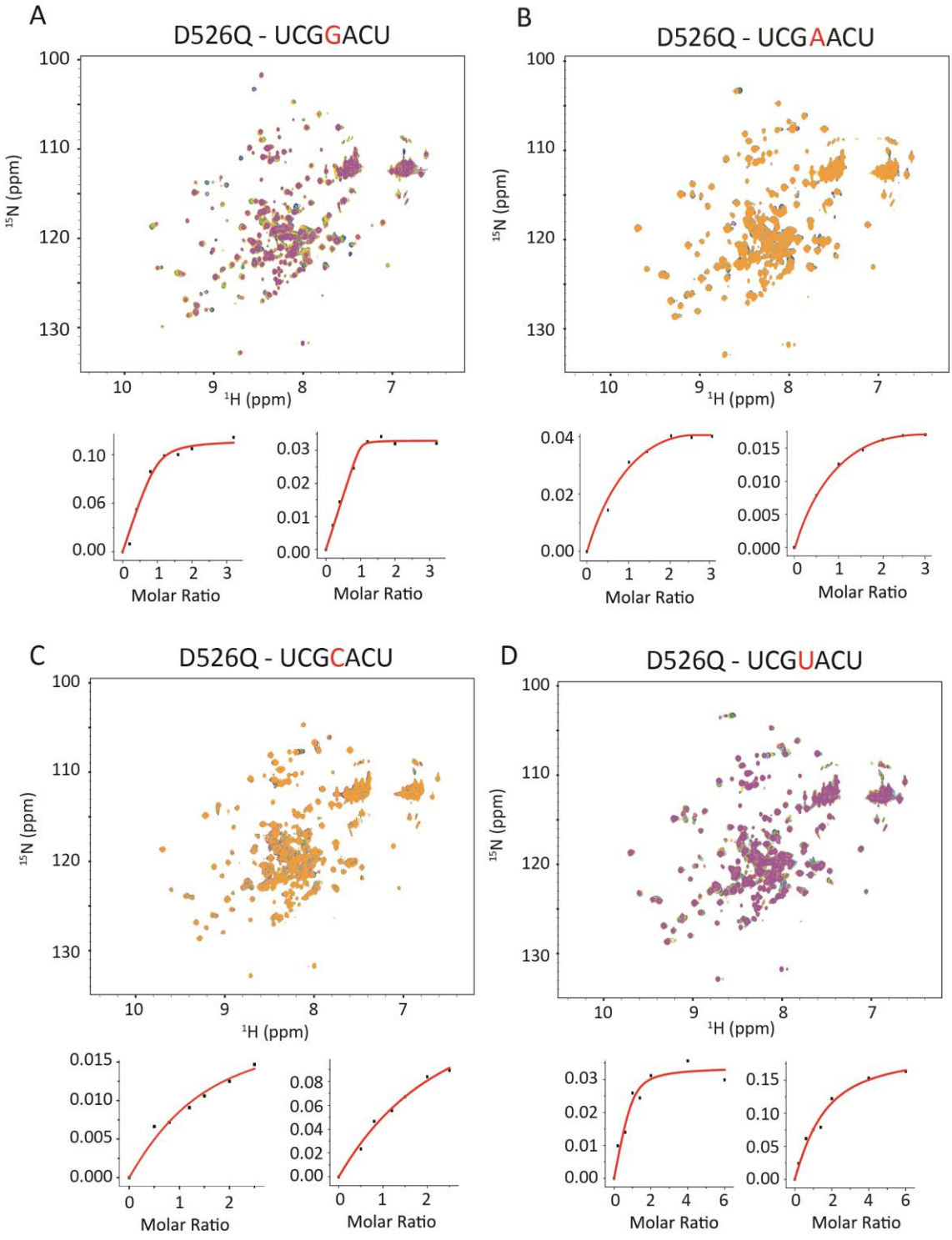


Figure 4.20: NMR ^{1}H-^{15}N SOFAST-HMQC titrations and binding curves for KH3DDKH4 D526Q for RNA oligos in which base in position 2 (G4) is mutated

All proteins were buffered in 100 mM NaCl, 10 mM sodium phosphate, 0.5 mM TCEP, pH 6.4 and were concentrated to 50-60 µM. NMR experiments were performed at 25^{o}C A) D526Q selectivity mutant SOFAST-HMQC overlaid spectra with oligonucleotide UCGGACU at protein:RNA ratios of 1:0 (Blue), 1:0.25

(Green), 1:0.5 (Red), 1:0.75 (Yellow), 1:1 (Purple), 1:5 (Cyan), 1:2 (Orange) and, 1:3 (Maroon). B) D526Q titration with UCGAACU at protein:RNA ratios of 1:0 (Blue), 1:0.5 (Green), 1:1 (Red), 1:1.5 (Yellow), 1:2 (Purple), 1:2.5 (Cyan) and, 1:3 (Orange). C) D526Q titration with RNA UCGCACU at protein:RNA ratios of 1:0 (Blue), 1:0.5 (Green), 1:0.75 (Red), 1:1.25 (Yellow), 1:1.5 (Purple), 1:2 (Cyan) and, 1:2.5 (Orange) D) D526Q titration with RNA oligo UCGUACU at protein:RNA ratios of 1:0 (Blue), 1:0.25 (Green), 1:0.5 (Red), 1:1 (Yellow), 1:1.5 (Purple), 1:2 (Cyan), 1:4 (Orange) and, 1:6 (Maroon). Each panel also displays chemical shift perturbation curves for selected peaks which were used to calculate binding affinity.

A

	K_d (μM)		Affinity Factor	
Oligo	D526Q	WT	D526Q	WT
UCGGACU	0.6 ± 0.3	1.1 ± 0.1	1.0	1.0
UCGAACU	7 ± 3.0	23 ± 2.6	11.7	20.9
UCGCACU	77 ± 38	N/A	128.3	N/A
UCGUACU	14 ± 5.0	20.3 ± 2.8	23.3	18.5

B

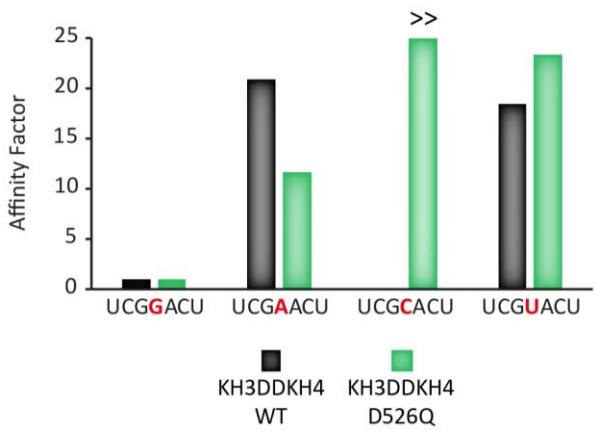


Figure 4.21: Summary of calculated K_d values for KH3DDKH4 and D526Q constructs and binding preference of KH3DDKH4 D526Q relative to KH3DDKH4

A) Table of calculated K_d values with errors (blue columns) and relative binding affinities as a factor of binding relative to UCGGACU RNA oligo (orange columns). B) Bar chart displays relative binding affinity of KH3DDKH4 and D526Q mutant. Binding of UCGCAUC is outside the range of values reported in the chart.

4.7 Discussion

The selectivity mutants tested here had varying effects on selectivity. The loss of a double hydrogen-bond to the RNA base in position 3 (A5), via the removal of the Lys452 side chain by the R452G mutation, within the KH3 domain had an effect on the RNA specificity. However, this change resulted in the domain becoming less specific, with the differences in binding affinities between the oligos with altered RNA bases being less than that observed for the WT protein. Additionally, the binding titrations showed that the mutant, like the WT protein, still preferred to bind an A in position 3 over the other RNA bases.

The incorporation of an enlarged amino acid side chain via the G500A mutation in the KH3 domain resulted in reduced overall RNA affinity with no effect on specificity. Conversely our D526Q mutation in the KH4 domain resulted in an increase in affinity without producing a clear shift in specificity. While this second mutation did not alter the specificity of the domain, an increase in RNA affinity can also serve as a useful molecular tool to investigate the individual contributions of RBDs on RNA selection in the context of the multidomain protein. Therefore, modifying the side chains of amino acid residues within the hydrophobic groove to optimise the distances between RNA bases for the formation of hydrogen-bonds (as in the case of our D526Q mutation) may provide a method to increase RNA affinity.

The S432R mutation was the only mutation tested that resulted in a reduction in affinity for the preferred RNA binding sequence. Interestingly, this mutation was not directly based on the hydrogen-bonding networks formed between the RNA bases and the protein. Instead, it was based on the local hydrophobic shape of the groove that accommodates the nucleobase. The other point mutations were mutated to amino acids that I believed would affect RNA specificity due to observed protein-RNA interactions identified in the NMR solution structure. The residues that were introduced were not amino acids that are typically found in other KH domains and so there was no structural information of how these mutations would interact with RNA bases in reality. Here for the S432R mutation

rationale we based our approach on observed binding interactions between an Arg residue in this position and a C in position 3.[209]

The RNA sequence specificity of KH domains has previously been described as being defined by the specific contacts the amino acids in the hydrophobic groove of the domain make with the accommodated RNA bases. These contacts are predominantly hydrogen-bonds.[71,209] However, the results from our mutational approach into altering the specificity of the KH3 and KH4 domain of IMP1 have shown that hydrophobic contacts, defined by the shape of the groove, may play a greater role in defining RNA specificity. In turn, the shape of the groove, and the contribution of the hydrophobic interactions needs to be taken more into consideration in order to alter the RNA binding specificity of KH domains. This could potentially require the mutation of several amino acids. To design such an approach a detailed structural understanding of the interaction would be required, and successful mutations would likely need to be modelled on examples of other KH domains when residues show a high degree of conservation (as in the case with our S432R mutation). However, the potential structural disruption such a mutation would invoke on the domain could prove dramatic as the hydrophobic groove is a highly structured region. Therefore, the effects of the mutations on the overall structure and stability of the domain must be studied in parallel.

Finally, our investigation into the specificity of the IMP1 KH3 and KH4 domains have reinforced the findings that the KH4 domains displays a higher degree of sequence specificity with respect to the KH3 domain.[85,179,180] The low A/C discrimination displayed by the WT KH3 domain in position 3 may be biologically relevant. As with the KSRP protein,[92,207] the KH3 domain of IMP1 may influence the selection of RNA targets through its ability to recognise different RNA sequences with different affinities. We plan to perform iCLIP on a FLAG-IMP1 S432R mutant to investigate if a shift in RNA target selection is observed as a result in the domain’s altered specificity.

Chapter 5. Investigation into the RNA binding properties of the IMP1 and IMP3 RRM domains

5.1 Introduction: RRM domains of the IMP protein family

In mammals, all IMP family members contain two N-terminal RRM domains,[136,141] while *Drosophila* IMP orthologues lack one or both of these RRMs.[183] To date, little work has been carried out to directly investigate the RNA binding properties of the IMP RRM domains. Currently the consensus in the field is that the KH domains are the main site of RNA binding,[83,85,193] which was our rationale when implementing our iCLIP investigation of IMP1 RNA target selection within HeLa cells.

In the Chapter 3 I reported a dramatic reduction of in-cell RNA binding for our IMP1 KH1-4DD mutant construct. Moreover, a study that knocked out RNA binding of the IMP KH domains, using the same mutational rationale as our GDDG mutations, showed IMP3 (a homologue of IMP1) was able to bind RNA with a 10-fold higher affinity than either IMP1 or IMP2 when all four KH domains contained the RNA binding knock out mutation.[83]. As IMP1 and IMP3 are the most similar in amino acid sequence, and expressed with a similar pattern in human tissues,[136,141] I decided to investigate the RNA binding of the IMP1 and IMP3 RRM domains in isolation to establish if the difference in RNA binding observed in the previous study could be attributed to the RRM domains.

As with the KH domains, the RRM domains are spaced apart by a short linker (five amino acid residues). Furthermore, studies on the KH domains revealed that both KH12 (data unpublished) and KH34 are pseudo-dimers and work in pairs to recognise their RNA targets.[85] I wanted to approach the RRM domains in a similar manner and produce constructs that contained both RRM1 and RRM2. This way I would be able to study the functions of the RRM domains together and see if RNA binding of the two domains is linked. The RRMs of IMP1 are also

reported to be involved in protein–protein interactions, for example they associate with the Kinesin-like protein KIF-11 in the localisation of ACTB mRNA.[155] Investigating the di-domain system enables us to study any effects RNA binding has on protein interactions with other protein partners.

5.2 Defining the IMP1 and IMP3 RRM12 construct boundaries

To deduce the construct boundaries suitable for our RRM12 constructs I examined the primary amino acid sequences of the N-terminal region of the IMP protein family. The RRM2 domain is predicted to end at residue 156. The NMR solution structure of the IMP3 RRM2 domain in isolation has previously been solved. This construct extended to residue 161 and revealed residues 154-161 to be unstructured. As there has been no previous structural investigation studying the RRMs as a di-domain construct, our intention was to extend the construct boundaries to prevent truncation of any potential structural element that may extend beyond the predicted RRM2 boundary. The primary amino acid sequences showed a proline glycine pair at residue numbers 185 and 186 respectively. The combination of a proline followed by a glycine residue typically produces a ‘kink’ in the protein structure due to the side chain combination of the two residues.[214] I chose this as the construct boundary of our RRM12 constructs to be confident that any structural element that may proceed the RRM2 domain would be included (Figure. 5.1).

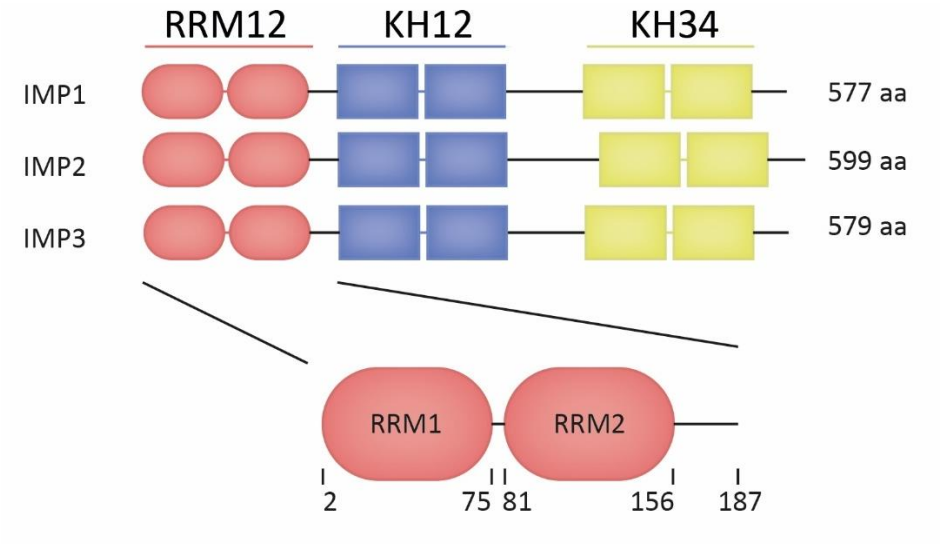


Figure 5.1: Schematic of the human IMP protein family, highlighting N-terminal RRM domain predicted boundaries
All three IMP members share the same domain organisation. The predicted length of the RRM1 and RRM2 domains are the same for each protein and are separated by a five-amino acid linker. The domain boundaries of the cloned RRM12 constructs are shown below.

5.3 Expression and purification of IMP1 and IMP3 RRM12 constructs

Previous studies performed on the full-length IMP1 protein have reported issues in recombinant expression, yet expression of the four KH domains was achievable, suggesting incorporation of the RRMs affects protein expression.[83] Accordingly, I approached this investigation with the anticipation that an expression protocol for the IMP1 RRM domains may need to be optimised.

The RRM12 constructs of IMP1 and IMP3 were cloned into a pETM-11 vector. Cloning sites were chosen to place a His-tag and TEV protease cleavage site N-terminal to the start of the RRM1 domain. The incorporation of the His-tag enables purification of the construct and the TEV digestion site allows removal of the His-tag for further analysis of the RRM12 construct.

Initially I performed recombinant protein expression in Bl21 DE3 *E. coli* cells at a reduced temperature of 22ºC for 16 h. Lowering the temperature during protein expression reduces the rate of protein synthesis and typically improves protein folding in *E. coli* and prevents recombinant proteins being packaged into insoluble inclusion bodies. Performing a standard His-tag protein purification protocol on the IMP1 and IMP3 RRM12 constructs, the IMP1 RRM12 construct was found to be largely insoluble and remained mainly in the pellet fraction, with a small proportion in the soluble supernatant fraction after cell lysis. In contrast the IMP3 RRM12 construct expressed to a reasonable yield in a soluble form.

To tackle the issue of the insoluble IMP1 RRM12 construct I optimised a denaturing protein purification protocol (Figure. 5.2). I used 8 M urea as a denaturing agent in the cell lysis buffer to solubilise all the protein species in the *E. coli* lysate. 8 M urea was added to all the buffers used in our typical His-tag purification protocol. I performed stringent washes during the initial protein purification stages by maintaining the concentration of urea at 8 M in addition to 1 M NaCl. I also increased the concentration of imidazole in the later washes from 10 mM to 30 mM.

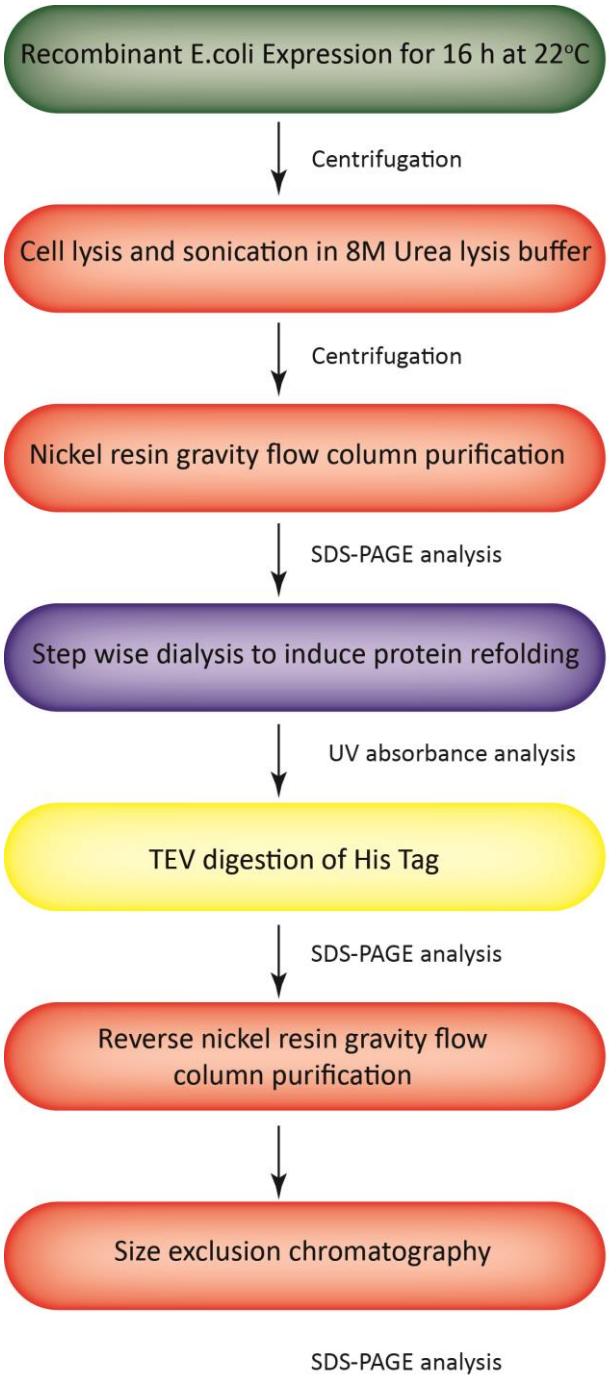


Figure 5.2: Flow-chart depicting the main steps of IMP RRM12 construct expression and purification
Boxes display the main steps during construct expression and purification. Colours represent process of each step; Green: expression, Red: purification, Purple: refolding, Yellow: enzymatic digestion. In order to monitor the efficiency of the purification protocol SDS-PAGE (to determine yield and purity) and UV absorbance analysis (to give an indication of protein folding) are recommended at certain points in the protocol.

The denatured protein purification protocol enabled us to efficiently purify the IMP1 RRM12 construct. I therefore applied the same method to the IMP3 RRM12 construct which resulted in purifying a greater yield than under native purification conditions.

I then implemented a method for the refolding of the RRM12 constructs. I determined that stepwise dialysis performed at 4^{o}C allowed the proteins to refold without aggregation. First, samples were dialysed from 8 M urea to 4 M urea over a time course of 8 h. Samples were then dialysed from 4 M to 1 M urea, again over 8 h, before finally being dialysed into buffer containing no urea overnight. For all the dialysis steps the sample was placed in dialysis buffer that was 10x the volume of the sample being refolded. To monitor potential protein aggregation during refolding I monitored the sample both visually and via UV absorbance (210 – 320 nm). Aggregated protein causes scattering of the UV light path affecting the baseline absorption (between 285 and 320 nm). This was used as an initial check that the refolding was progressing successfully.

Once samples had been returned to non-denaturing conditions, TEV protease was added to remove the N-terminal His-Tag. After His-tag removal size exclusion chromatography on a Superdex 75 column was performed. RNA binding proteins commonly co-purify with nucleic acids. This can be analysed by looking at the 280 to 260 nm absorbance ratio when performing UV absorbance spectroscopy. In these cases, an ion exchange purification step would be required to remove the nucleic acid. However, I did not observe any nucleic acid contamination in our preparations and so this step was not required (Figure. 5.3).

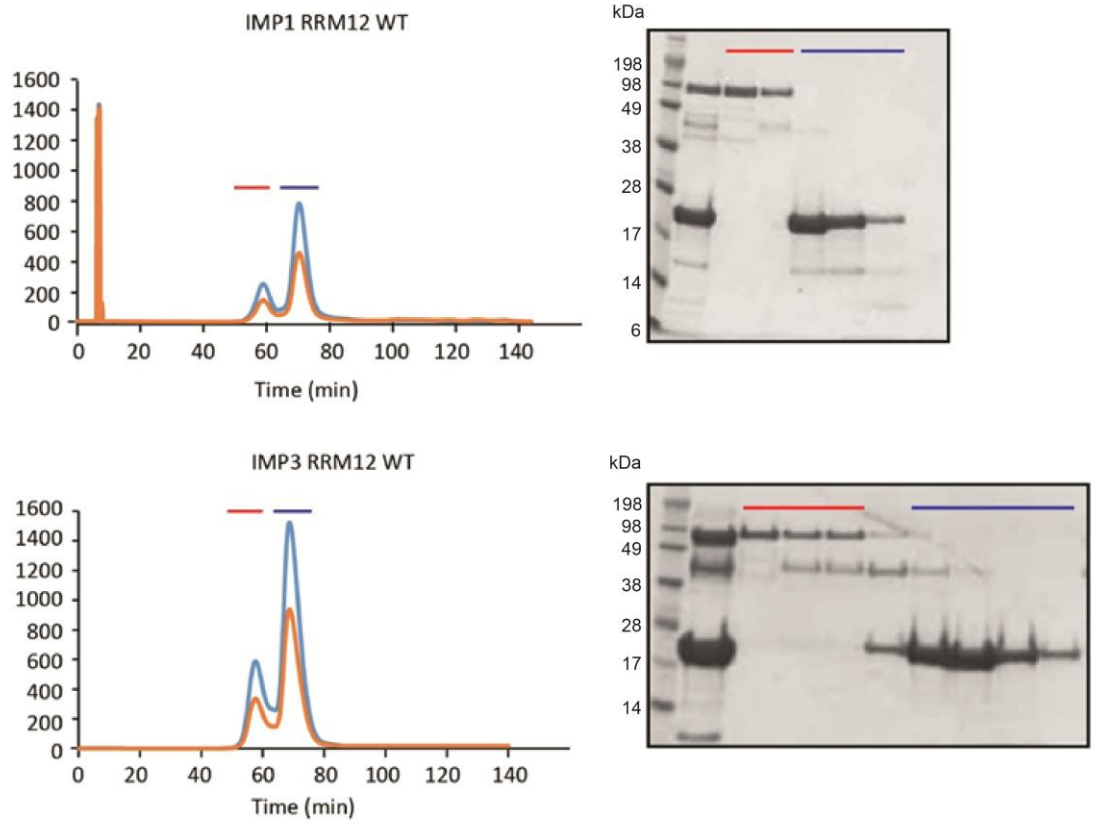


Figure 5.3: Size exclusion chromatography purification of IMP1 and IMP3 RRM12 domains
Chromatograms of size exclusion chromatography are displayed on the left for IMP1 and IMP3 RRM12. Y-axis displays absorbance at 280 nm (displayed as the light blue trace) and at 260 nm (displayed as the orange trace). The X-axis displays column retention time displayed in minutes. Fractions collected between 50 min and 80 min were collected and run on an SDS-PAGE gel and are displayed on the right. The first sample lane on both gels represents the input sample loaded onto the size exclusion column. The red line indicates elution fractions collected in the first purification peak and the blue line represents elution fractions in the second, and main, purification peak which contains the RRM12 constructs.

To confirm the proteins were refolded I used a combination of far UV CD analysis (Figure. 5.4) and ^{1}H-^{15}N HSQC spectra (Figure. 5.5). I expected the secondary structure content of the two constructs to be similar. A secondary structure prediction using JPred4 predicted the IMP1 RRM12 construct to be 22% β-sheet and 20% α-helix, whereas IMP3 was predicted to be 23% β-sheet and 20% α-helix. Therefore, I used far UV CD absorption spectrum of IMP3 RRM12 purified under native conditions as a benchmark to assess the refolding of the IMP1 and

IMP3 RRM12 constructs. The far UV CD spectrum showed both proteins to display a spectrum that is a mixture of β-strand and α-helix, with the predominant signal coming from the α-helical structures (Figure. 5.4). This is in agreement with the predicted secondary structure content and expected due to α-helical structures absorbing stronger then β-strands. There was a slight difference in the absorbance spectra of the two proteins. One possible explanation for this may be that the IMP1 RRM12 di-domain contains a greater proportion of β-strands then IMP3 RRM12. The samples were buffered in a high concentration of NaCl (100 mM) to improve stability. This resulted in the spectra between 190 and 195 nm being cut due to increased noise coming from an increase in voltage resulting from the high salt concentration (Figure. 5.4). Without this region of the spectrum it was not possible to perform an accurate secondary structure content deconvolution from the proteins absorption pattern.

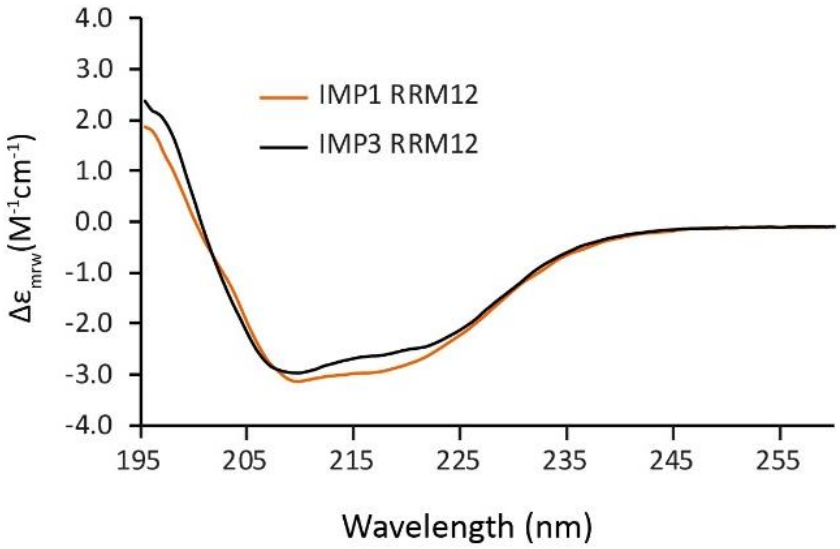


Figure 5.4: Far UV CD analysis of IMP1 and IMP3 RRM12 constructs
0.15 mg/ml protein concentration for each protein. Proteins were buffered in sodium phosphate buffer pH 7.4 100 mM NaCl, 0.5 mM TCEP. Spectra were the result of 50 accumulation scans and then fitted following a smoothing process.

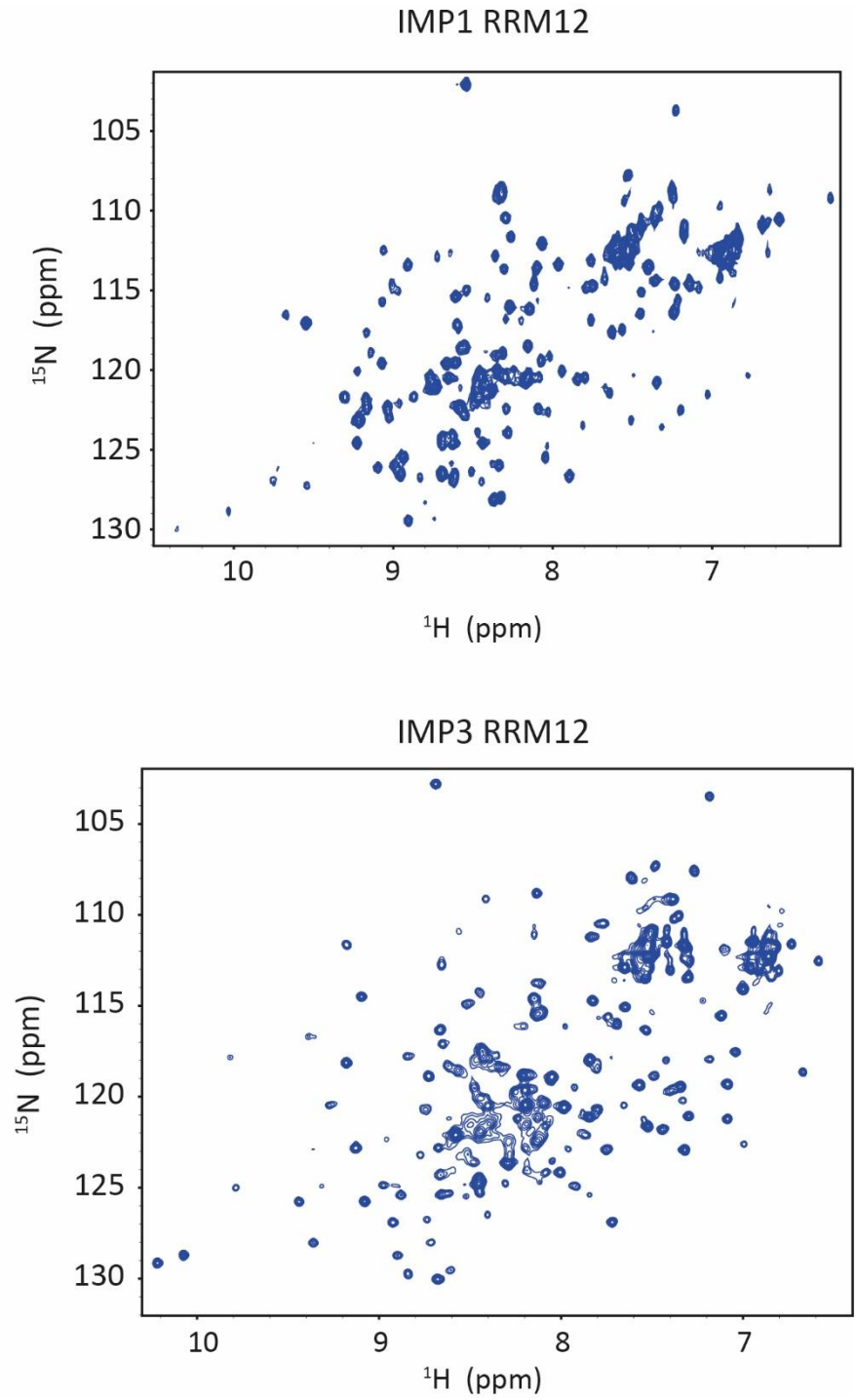


Figure 5.5: ^{1}H-^{15}N SOFAST-HMQC spectra of IMP1 RRM12 and IMP3 RRM12 di-domain constructs
Recorded at 25^{o}C with proteins concentrated to 60 µM and buffered in sodium phosphate buffer pH 7.4 100 mM NaCl, 0.5 mM TCEP.

5.4 Investigating the RNA binding properties of the RRM12 constructs

Currently there is no direct information available regarding IMP RRM12 RNA binding interactions. I first set about to investigate any potential RNA binding capabilities by studying the amino acid composition of the RRM1 and RRM2 domains to see if the canonical RNP motifs were present. Previously other groups have solved the NMR solution structures of the RRM1 domain of IMP2 (PDB:2CQH) and the RRM2 domain of IMP3 (PDB:2E44) in isolation. Therefore, I also compared the fold of these structures with well characterised RRM domains to gain insight into on how these RRMs could potentially recognise RNA.

As previously stated, canonical RRM-RNA recognition is mediated by specific aromatic and positively charged residues with in the RNP motifs in the β-strands that make up the domain β-sheet.[53] Sequence alignment performed on the RRM1 and RRM2 domain sequences of all three human IMP isoforms showed RRM1 contains the conserved canonical RNP1 and RNP2 motifs (Figure 5.6). The NMR solution structure of IMP2 RRM1 confirms that the four β-strands fold together in the canonical topology of a classic RRM domain. The key aromatic residue from RNP1 (Y5) is located in the β1-strand and key aromatic residues of RNP2 (Y39 and F41) are located on the β3-strand (Figure 5.7). Additionally, residue K36 in the RNP2 motif is located towards the end of the β2-strand and it capable of potentially interacting electrostatically with the phosphate backbone of an RNA oligonucleotide accommodated by the β-sheet.

In contrast, the RRM2 domain primary amino acid sequence alignment showed an absence of the canonical RNP motifs (Figure. 5.6). This suggests that the RRM2 domain cannot recognise RNA in a classical fashion. In addition, the RRM2 structure shows the β2-strand to fold across the β3-strand, and so potentially occluding the β-sheet and preventing RNA binding (Figure. 5.7). However, some RRM domains have been shown to bind RNA using alternate mechanisms. Residues within the protein loops connecting the β-strands and α-helices, typically the strands on the ‘south’ side of the β-sheet, have been shown

to mediate RNA interactions. I investigated the amino acid composition of these protein loops to determine if residues shown to bind RNA in other RRM domains were conserved in the RRM2 domain of the IMP protein family. Previously solved RRM-RNA structures have shown that either one, two or all three loops can be involved in RNA recognition. For example, RNA recognition of RBMY,[215] TcUBP1,[216] RRM2 of SF2[217] and Hrp1[90] all have one loop that is involved in RNA binding. Fox-1[218] and REF2-I[219] have residues in two loops, whereas hnRNP F has residues in all three loops that are involved in RNA recognition.[65] Even though these examples utilise residues within their protein loops to bind RNA, they also contain the canonical RNP residues, with the exception of hnRNP F.[65]

hnRNP F contains three RRMs which are termed quasi RRMs (qRRM) due to their atypical mode of RNA binding and lack of RNP motifs.[65] I examined if there were any similarities between the composition of these qRRM domains and RRM2 of the IMPs. qRRM1 and qRRM2 of hnRNP F contain an aromatic residue in both loop 1 (β1/α1) and the β-hairpin motif in loop 5 (α2/β4). The aromatic reside within loop 1 (F120) and within the β-hairpin (Y180) are critical for RNA binding in the qRRM2 domain.[65,66] However, RRM2 of the IMP family lacks any aromatic residues within loop 1, in addition to lacking the β-hairpin in loop 5 (note that RRM1 contains a β-hairpin in loop 5). Additionally, the α-helix 1 of the RRM2 domain does not contain the SWQDLKD motif observed in pseudo-RRM domains which mediated RNA binding.[61] In conclusion, by comparing the RRM2 domain in isolation with known examples of RRM-RNA interactions, I was unable to identify a potential RNA binding surface.

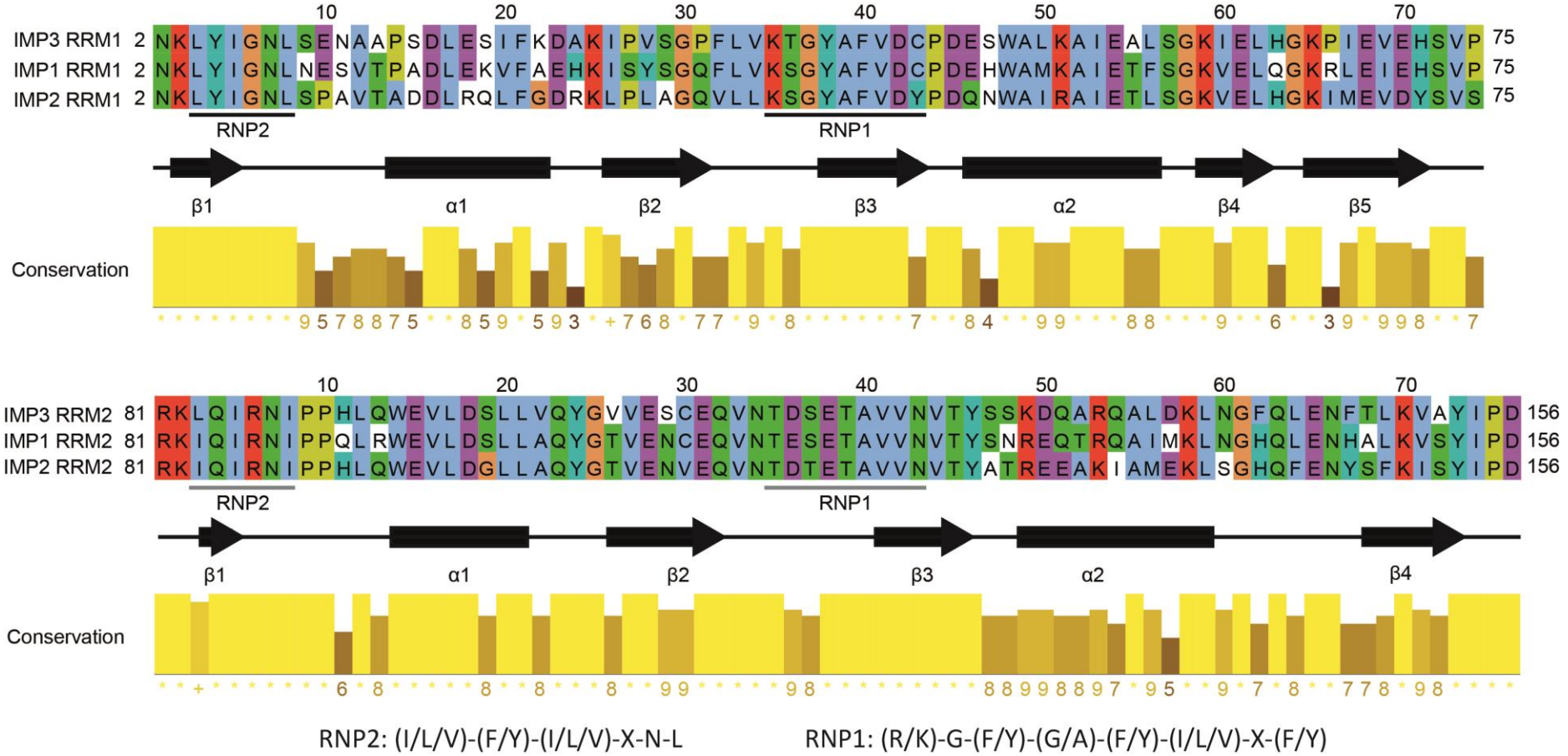
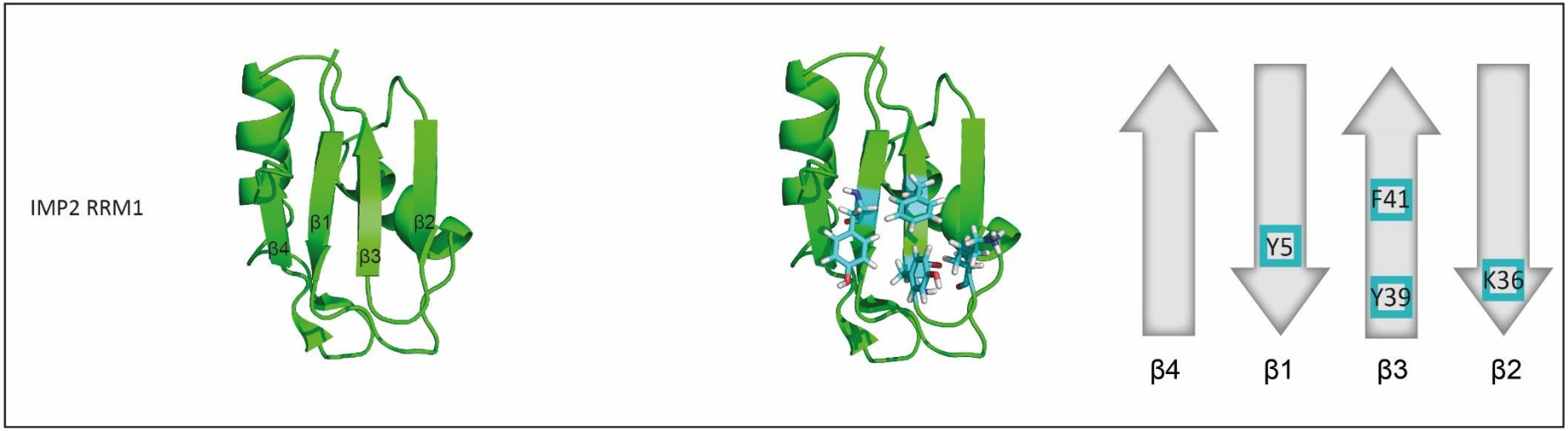


Figure 5.6: Amino acid sequence alignment of RRM1 and RRM2 of IMP1, 2 and 3, with conservation scores and secondary structure predictions

Sequences were aligned using Clustal Omega and residues coloured according to Clustalx scheme. Predicted secondary structure and conservation scores are displayed below the alignments. Canonical RNP sequence motifs are displayed (Bottom) and are highlighted in RRM1 (Black lines) but absent in the RRM2 sequences (Grey lines).


IMP2 RRM1
β1
β2
β4
β3
Y5
F41
Y39
K36
β4
β1
β3
β2

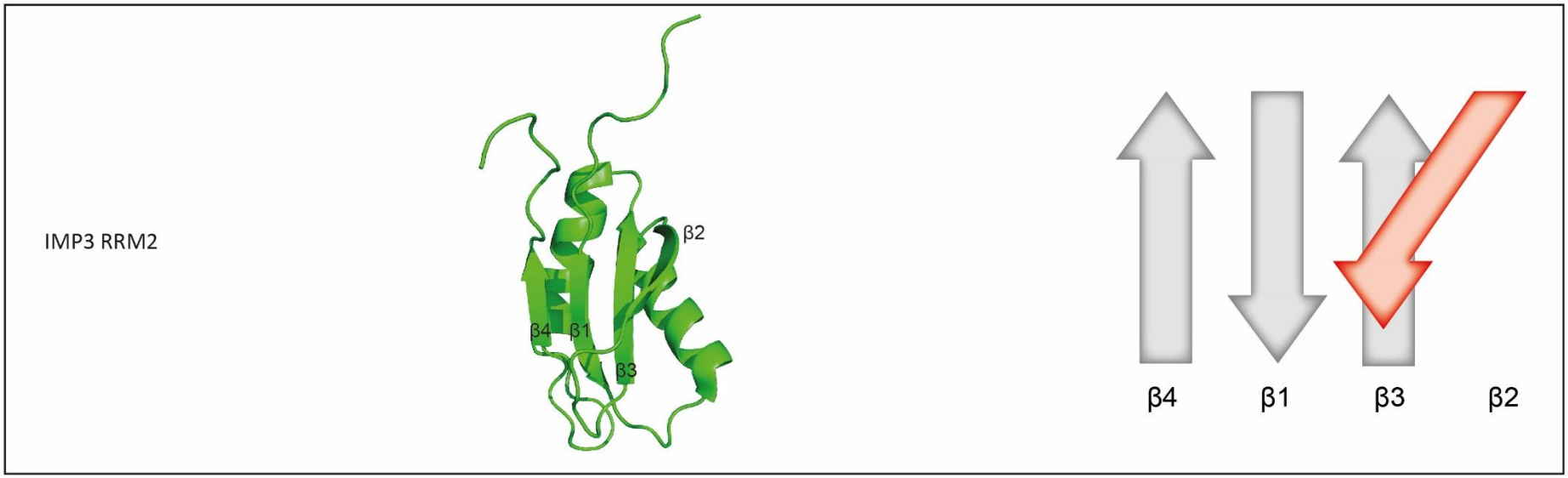

IMP3 RRM2
β2
β4
β1
β3
β4
β1
β3
β2

Figure 5.7: NMR solution structures of IMP2 RRM1 domain and IMP3 RRM2 domain with schematic of β-strand topology in the fold of the RRM domains
Top: Solution structure of IMP2 RRM1. Residues within the RNP1 and RNP2 motif that are reported to be critical for canonical RRM RNA recgnition are identified in the structure and displayed in blue. Location of these residues are further highlighted by the schematic of the β-strands of the IMP2 RRM1 fold. Bottom: NMR solution structure of IMP3 RRM2. Cononical RNP motifs are absent from the RRM2 domain and so are not highlighted. Atypical β2-strand (red) is highlighted in β-sheet schematic and direction across the β-sheet depticted.

As there was no information on the potential RNA binding specificity of the RRM domains poly-N-randomised RNA oligos were initially used for RNA binding titrations. To directly investigate the potential RNA binding capabilities of the IMP1 and IMP3 RRM12 domains I compared ^{1}H-^{15}N SOFAST-HMQC spectra of the proteins with and without RNA. Using a 1:1 protein:RNA molar ratio with a 5N RNA oligonucleotide, I observed chemical shifts for both the IMP1 and IMP3 RRM12 constructs (Figure. 5.9 & 5.10). However, larger shifts were observed for the IMP3 RRM12 construct, in addition to several peaks reducing in intensity upon addition of RNA (Figure 5.10). This is indicative of a fast to intermediate exchange regime.

Next, I determined the number of RNA residues the RRM di-domain constructs could bind. Typically, canonical RRM domains can specifically recognise three or four RNA bases in a sequence specific manner. However, when binding is extended beyond the canonical β-sheet surface to include the loop regions a single RRM domain can accommodate six nucleotides. As our constructs contain two RRM domains I was unsure if the construct contains either one or two RNA binding surfaces, or if the two RRM domains come together to produce an extended binding interface. In turn, I increased the length of the random RNA oligo from 5 bases to 6 and compared the chemical shifts I observed. For both IMP1 and IMP3 I observed the same number of peaks shifting upon binding of the 5N and 6N RNA oligo (~19 peaks for IMP1 and ~ 25 peaks for IMP3). The direction of the peak shifts was consistent for both oligos, as was the intensity of the shifting peaks. For both the IMP1 and IMP3 constructs the size of the peak shifts for the 6N oligo was greater than that of the 5N oligo (Figure. 5.9 & 5.10).

This was to be expected as increasing the length of degenerate RNA sequence increases the number of possible binding registries. In turn, a 1:1 molar ratio of a 6N oligo contains more binding registers then with the 5N oligo, and so an increase in affinity is observed. Based on these findings I decided to probe the RNA binding preference of the domains using RNA oligos 5 nucleotides in length.

I concluded that the RRM12 di-domains contained only one RNA binding surface. I based this conclusion on the results of the comparison of RRM1 and RRM2 domain amino acid sequence and structures with RRM domains known to bind RNA. In addition, our preliminary RNA binding assay with randomised RNA yielded ~ 20 chemical shift perturbations for each construct (Figure 5.9 & 5.10). I would expect to observe more residues shifting upon addition of RNA if both the RRM1 and RRM2 domains contained two separate RNA binding surfaces. Therefore, I continued our investigation into the RNA binding properties of the RRM12 di-domain assuming the presence of a single RNA interacting surface within the constructs.

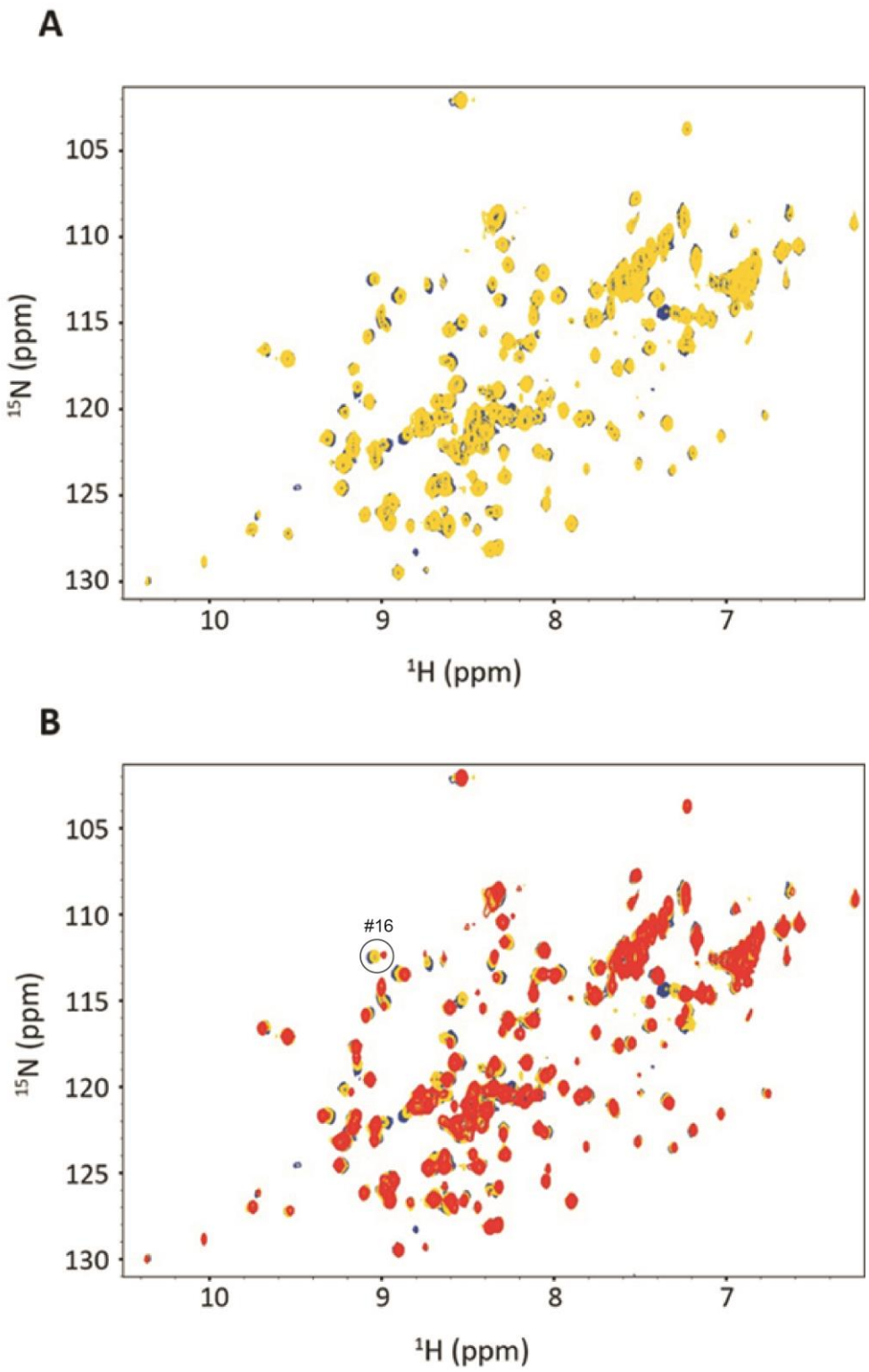


Figure 5.8: IMP1 RRM12 RNA binding to pools of random RNA oligomers
A) Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of 50 µM IMP1 RRM12 in free form (Blue) and with 1:1 molar ratio of randomised 5N RNA oligo (Yellow) at 25^{o}C. B) Overlaid spectra as above in (A) with the addition of the overlaid 1H-15N SOFAST-HMQC spectra of IMP1 RRM12 with 1:1 molar ratio of randomised 6N RNA oligo (Red). Peak #16, which is used in the SIA analysis (explained below) is highlighted.

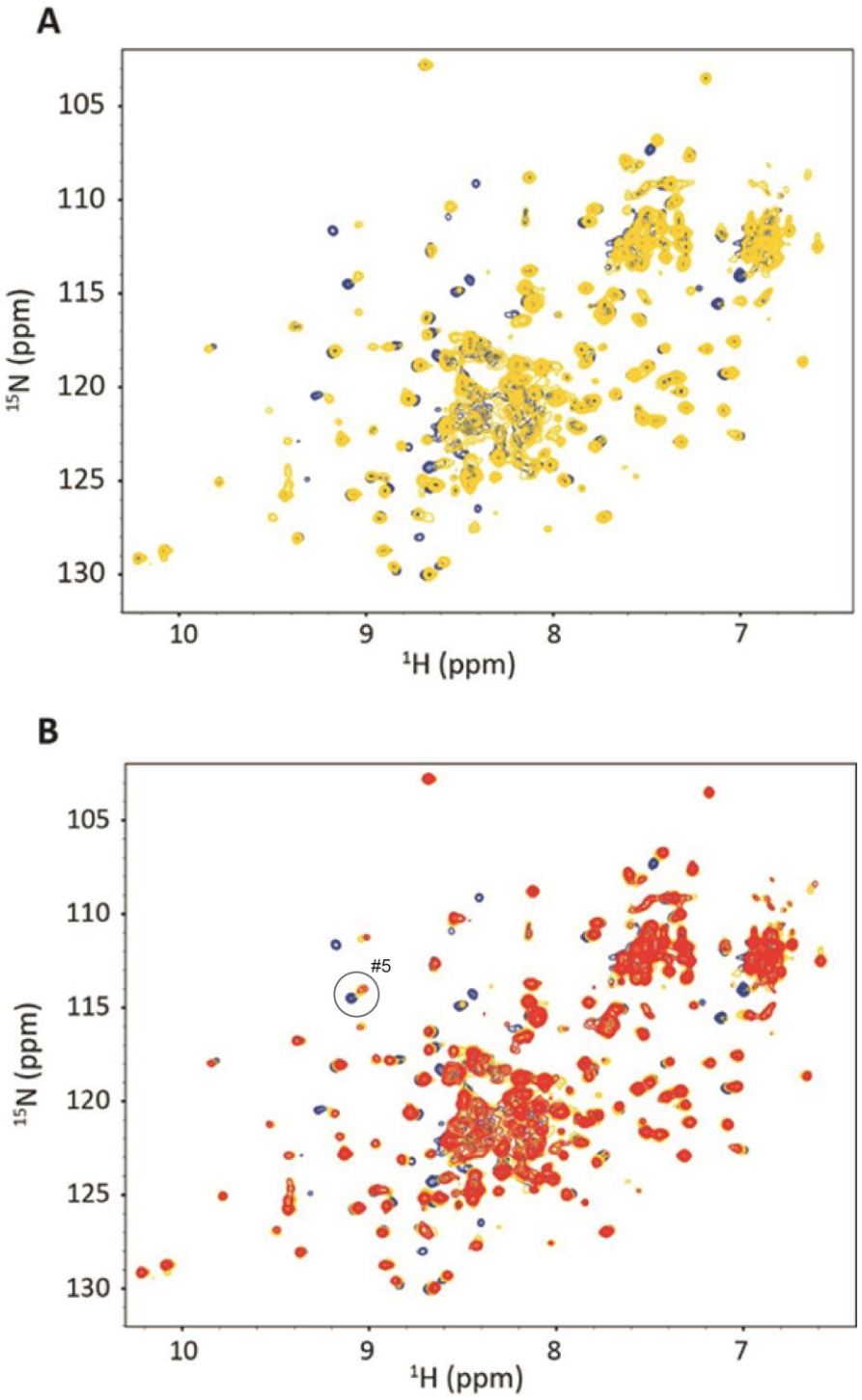


Figure 5.9: IMP3 RRM12 RNA binding to pools of random RNA oligomers
A) Overlaid 1H-15N SOFAST-HMQC spectra of 50 µM IMP3 RRM12 in free form (Blue) and with 1:1 molar ratio of randomised 5N RNA oligo (Yellow) at 25^{o}C. B) Overlaid spectra as above in A with the addition of the overlaid 1H-15N SOFAST-HMQC spectra of IMP3 RRM12 with 1:1 molar ratio of randomised 6N RNA oligo (Red). Peak #4 is highlighted to represent the peak with the largest chemical shifts in later SIA analysis.

5.5 Sequence specificity of IMP1 and IMP3 RRM12 di-domain

To define the nucleobase preference of the RRM12 di-domain I implemented scaffold independent analysis (SIA). Pools of randomised RNA oligos where one base position remains fixed (either A, C, G or U) were used to determine binding preference. A total of 16 randomised RNA pools were required to scan nucleobase preference in four positions (nXNNN, nNXNN, nNNXN, and nNNNX where X is fixed in turn with all four possible nucleotides and N is randomised). The pools were then titrated with either IMP1 or IMP3 RRM12 and ^{1}H-^{15}H correlation spectroscopy were recorded to monitor binding.

Due to the lower affinity IMP1 RRM12 displayed towards the randomised RNA oligos in the previous study compared to the IMP3 RRM12 domain, the IMP1 SIA was performed at a 1:2 protein:RNA molar ratio and the IMP3 SIA performed at a 1:1. For analysis the chemical shift changes of 16 peaks were measured, using the same peaks for all RNA pools (Figure. 5.10 & 5.11). Typically, 10-15 peaks are used during SIA analysis and so I observed enough chemical shifts to fulfil this requirement. Shifts were normalised and averaged to give final scores as described in the methods section 2.5.1 and[95,213] (Figure. 5.10 & 5.11).

As previously seen with the randomised RNA oligonucleotides, I observed smaller peak shifts in the IMP1 SIA compared to the SIA of IMP3. These smaller shifts are more difficult to accurately measure and in turn, introduced a greater error. However, I can compare the overall SIA analysis with the chemical shifts of peak #16 (Figure 5.8B). Peak 16 was observed to have the greatest chemical shift of all the peaks tracked (>0.15 ppm) (Figure. 5.10). While it is not accurate to base binding preference on a single peak, the chemical shifts of this peak were sufficient to measure accurately. The base that gained the highest SIA score in each position also resulted in the greatest shift in peak number 16, adding confidence in our SIA results. The SIA with the IMP3 RRM12 di-domain produced larger shifts with at least half of the peaks analysed shifting by 0.1 ppm or greater (Figure. 5.11).

SIA revealed IMP1 and IMP3 to display different RNA binding preferences. In general, IMP1 displays a negative preference for U in all four positions, particularly in the 1st and 2nd position. IMP3 displayed a similar bias against U but with weaker discrimination overall. For both IMP1 and IMP3 stronger sequence specificity was observed in the 1st position with IMP1 preferring C or G and IMP3 preferring C. In the 2nd position IMP1 did not display a strong sequence preference except against U. In contrast IMP3 displayed sequence preference in the 2nd position with C being preferred with a difference of ~ 0.2 ppm against all other bases in this position. For the 3rd position IMP1 displayed a similar preference as with the 2nd position, yet the discrimination against U was less pronounced. IMP3 showed weaker sequence preference in both position 3 and 4 compared to the first two positions, with a slight preference being observed for A in both positions, yet the difference between A and the second most preferred base C only being ~ 0.1 ppm. IMP1 displayed the strongest sequence preference in the 4th position with a clear preference for G.

In summary, the IMP1 RRM12 sequence preference was determined to be C – A/C/G – A/C/G – G and IMP3 RRM12 to be C – C – A – A, with both proteins displaying a negative bias towards poly U sequences.

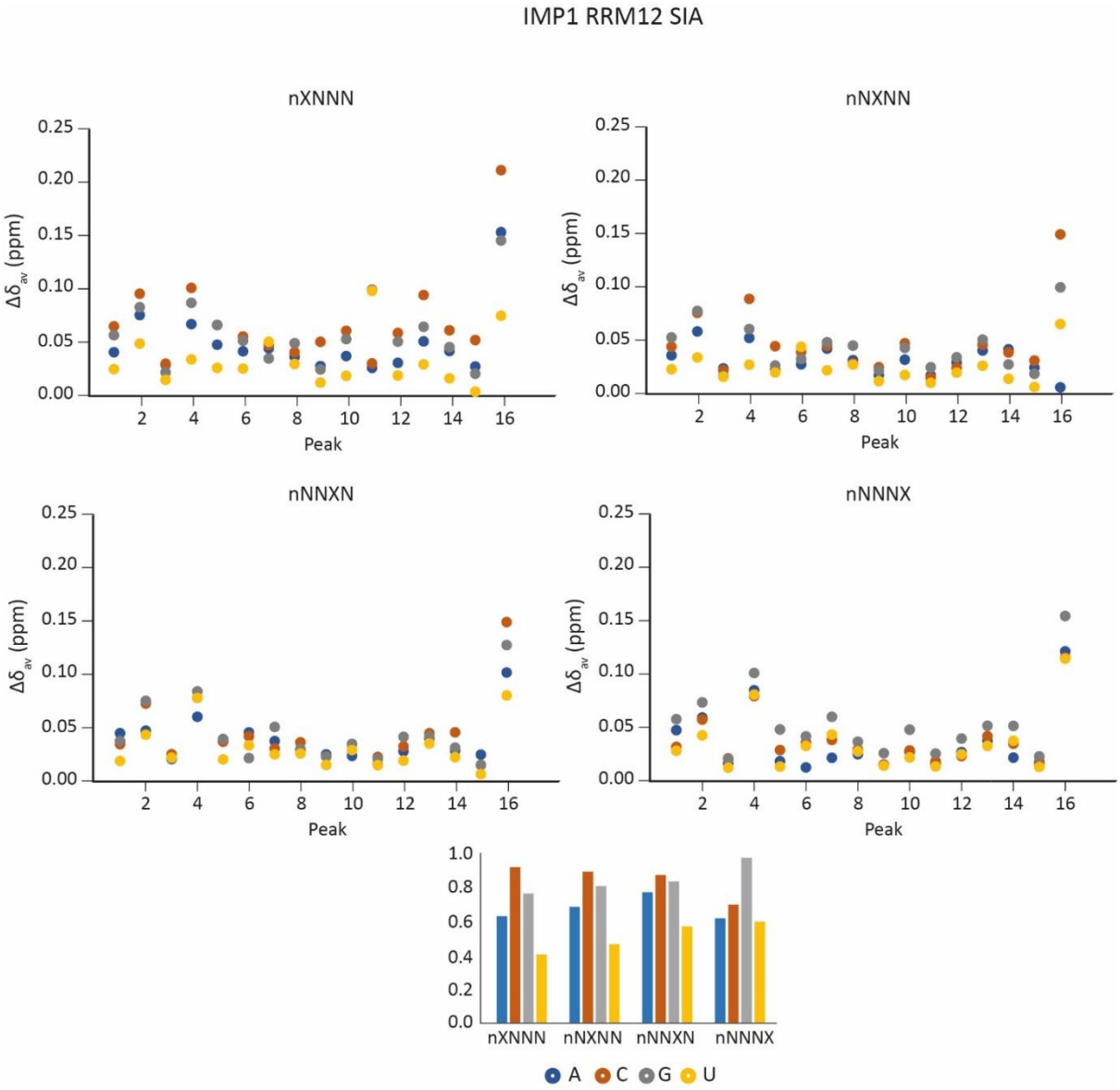


Figure 5.10: Chemical shift perturbation of IMP1 RRM12 peaks upon addition of SIA RNA pools and result average
Top graphs show peak number displayed on x-axis and the weighted chemical shifts on the y-axis. A fixed position (Blue), C fixed position (Red), G fixed position (Grey) and U fixed position (Yellow). Each graph represents one fixed RNA position with 'X' defining that position in the 5 nucleobase oligo pools. Bottom graph shows average normalised SIA scores for each fixed RNA base in each position.

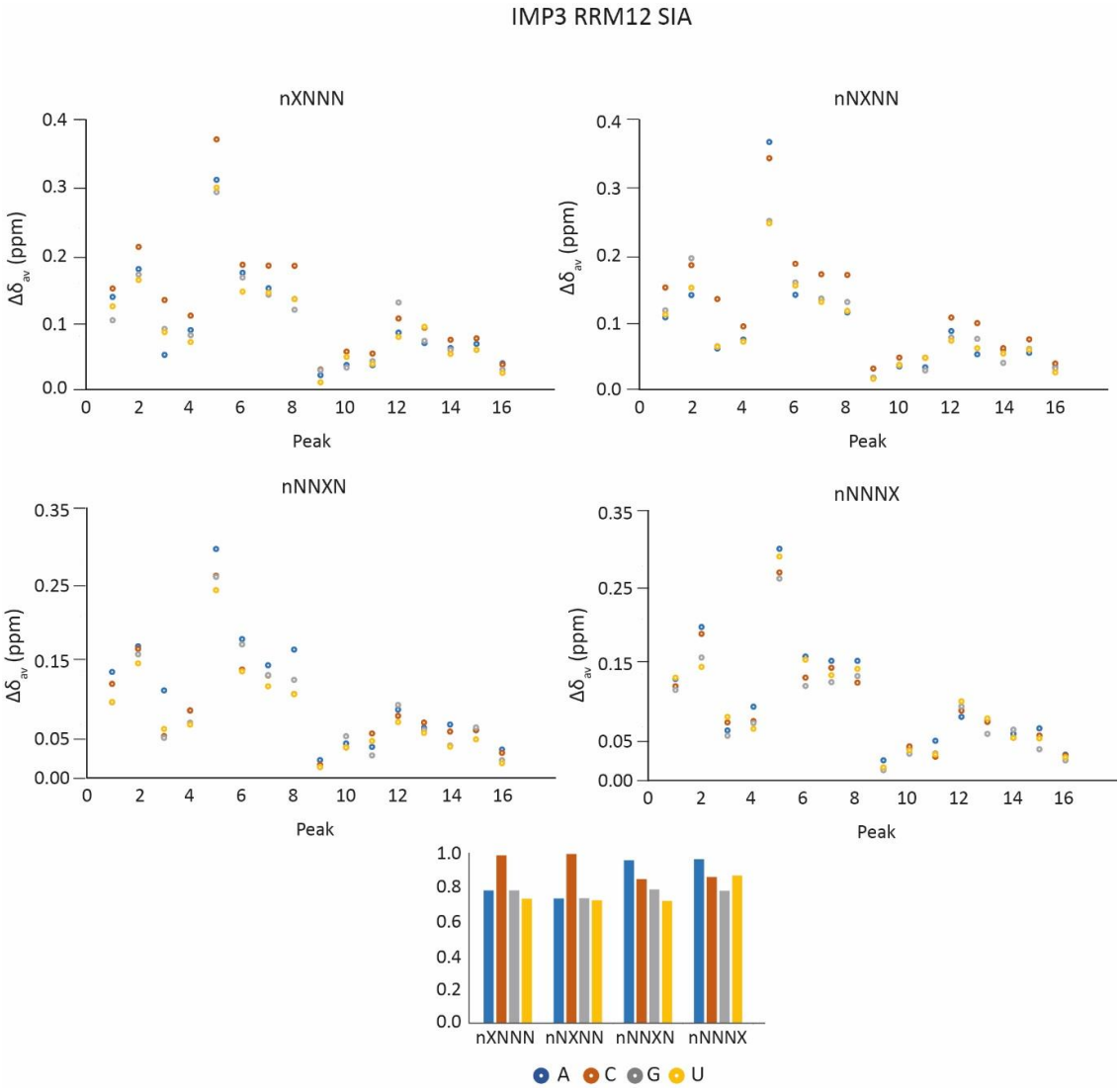


Figure 5.11: Chemical shift perturbation of IMP3 RRM12 peaks upon addition of SIA RNA pools and result average
Top graphs show peak number displayed on x-axis and the weighted chemical shifts on the y-axis. A fixed position (Blue), C fixed position (Red), G fixed position (Grey) and U fixed position (Yellow). Each graph represents one fixed RNA position with 'X' defining that position in the 5 nucleobase oligo pools. Bottom graph shows average normalised SIA scores for each fixed RNA base in each position.

To verify the nucleobase preference determined via our SIA analysis I performed binding titrations with the top and lowest ranking RNA sequences for both proteins. UCCCG was used as the top-ranking sequence for IMP1 and UUUUU as the lowest. For IMP3 UCCAA and UUUUG were chosen for the top and lowest-ranking sequences respectively. Titrations were monitored by NMR ^{1}H-^{15}N correlation spectroscopy.

IMP1 RRM12 di-domain ^{1}H-^{15}N SOFAST-HMQC spectra were recorded with the preferred RNA sequence of UCCCG at protein:RNA ratios of 1:0.5, 1:1, 1:1.5, 1:2, 1:2.5, 1:3 and 1:4 (Figure. 5.12). Analysis was performed by manual measurement of peak shifts. Again, peak shifts observed for IMP1 RRM12 were small and so errors in peak shift measurements were high (Figure 5.12B). To minimise error, peaks that displayed the largest shifts were picked to be measured. Additionally, only peaks that were in fast exchange and could be followed accurately were chosen. In total 15 shifting peaks were chosen, of these 15 peaks four peaks remained in the linear proportion of the binding curve when average chemical shifts were plotted against RNA concentration. From the remaining 11 peaks, six peaks produced chemical shifts that could be fitted to a binding curve with confidence (Figure 5.12C). From these six peaks an average K_d of 148 ± 44 μM was calculated. In contrast, the least preferred RNA sequence for IMP1 RRM12 di-domain (UUUUU) did not result in visible chemical shifts in the ^{1}H-^{15}N SOFAST-HMQC spectra at a 1:6 protein:RNA ratio (Figure 5.13).

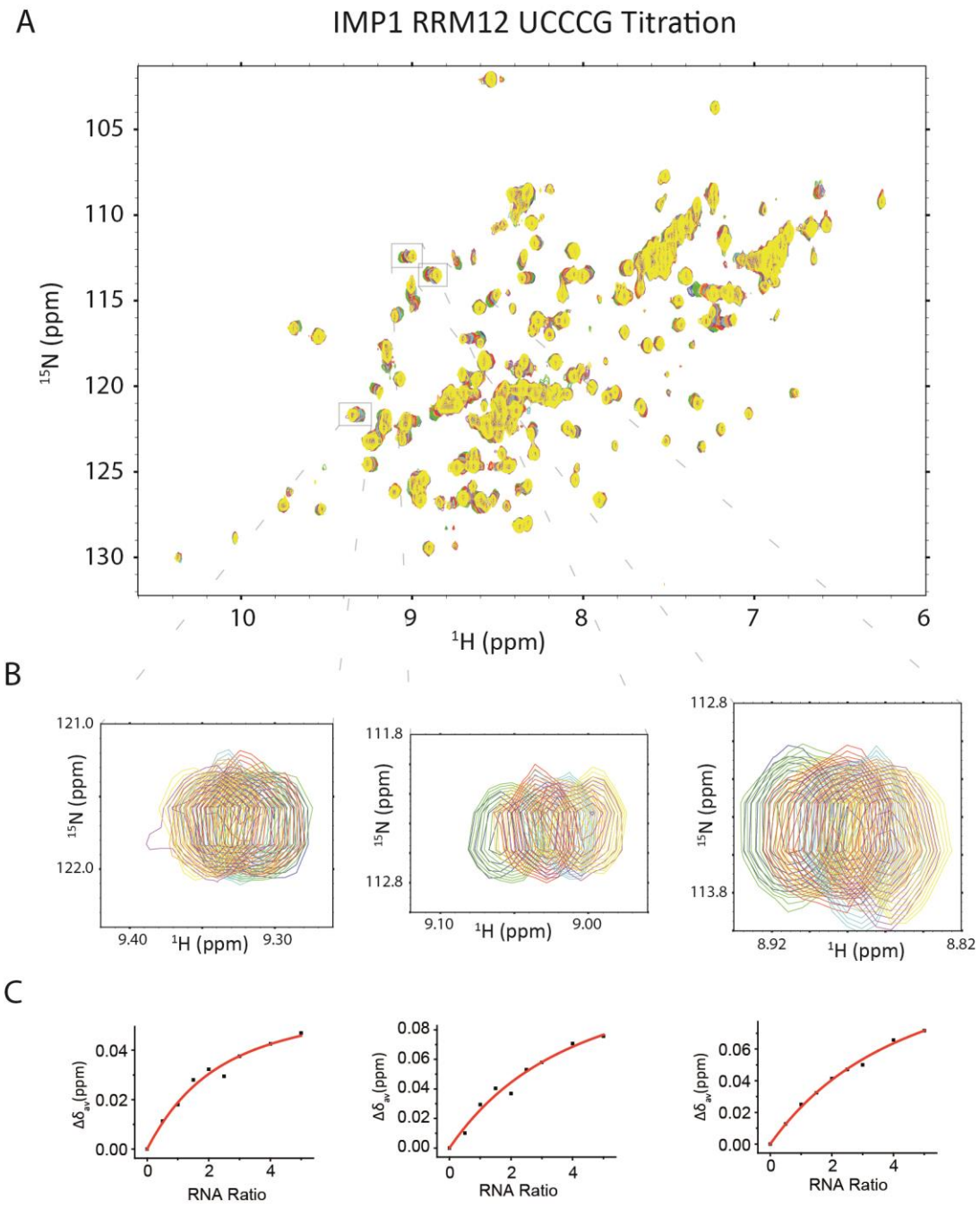


Figure 5.12: Titration of IMP1 RRM12 with UCCCG oligonucleotide
A) ^{1}H-^{15}N SOFAST-HMQC overlaid spectra of 60 µM IMP1 RRM12 at protein:RNA ratios of 1:0 (Blue), 1:0.5 (Green), 1:1 (Gold), 1:1.5 (Red), 1:2 (Purple), 1:2.5 (Cyan), 1:3 (Orange), 1:4 (Magenta), and 1:5 (Yellow). B) Zoomed in areas of spectra in A to highlight peak shifts used to calculate the K_d of IMP1 towards UCCCG RNA oligo. Grey boxes in A indicate the zoomed peaks. C) Chemical shift perturbations upon addition of increasing molar concentrations of RNA. Curves represent the peaks highlighted above in panel B and were used to determine binding affinity.

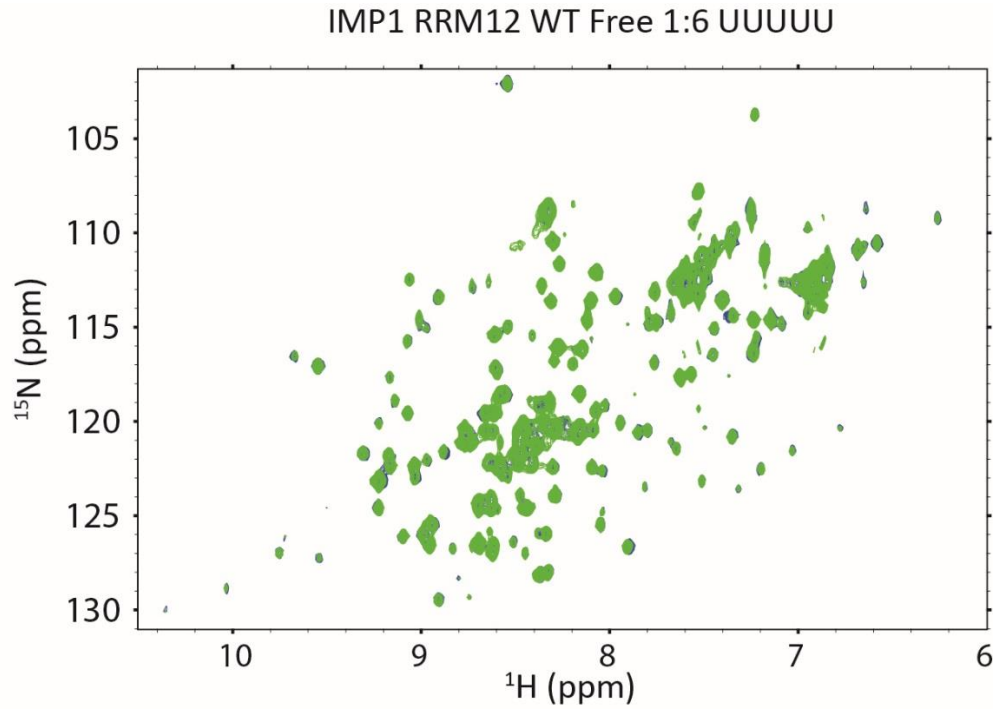


Figure 5.13: IMP1 RRM12 di-domain upon addition of UUUUU
Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of 60 µM IMP1 RRM12 di-domain (Blue) with 1:6 protein:RNA molar ratio UUUUU (Green).

I previously hypothesised that the IMP3 RRM12 di-domain has a higher affinity towards RNA than IMP1. In turn, the NMR titration of IMP3 and its preferred RNA sequence of UCCAA was performed with protein:RNA ratios of 1:0.2; 1:0.4; 1:0.6; 1:0.8; 1:1; 1:1.5 and 1:2. I observed several peaks in slow to intermediate exchange suggesting a tighter interaction (Figure. 5.14). From the peaks that were in fast exchange I monitored chemical shifts by manual analysis, selecting peaks based on the same criteria as for IMP1 titrations. In total I was able to follow 11 peaks accurately. All peaks that were chosen where seen to saturate around a 1:1 protein:RNA molar ratio (Figure 5.14B & C). From the 11 peaks an average K_d of 1.0 ± 0.2 µM was calculated.

Titrating the IMP3 RRM12 di-domain with the lowest scoring RNA sequence showed several peak shifts. However, all shifts were observed to be in the fast exchange regime (Figure 5.15). This was in contrast to the preferred RNA sequence titration where some peaks were in intermediate exchange. Titration points were recorded at 1:0.4; 1:1; 1:1.5; 1:2 and 1:4 protein to UUUUG molar

ratios. The same 11 peaks chosen in the IMP3 UCCAA RNA titration were picked and their chemical shifts were measured (Figure 5.15B and 5.15B). To calculate an estimate of the binding K_d RNA chemical shift binding curves were plotted. No peaks were observed to be in binding saturation at a 1:4 protein:RNA ratio, yet the curving of the points suggests saturation was being approached (Figure 5.15C). In turn, I was unable to calculate a reliable K_d as the affinity was outside the range of molar RNA ratios used in the titration. From these findings I estimated a K_d over 0.5 mM

Rather than defining an accurate K_d for the IMP3 RRM12 di-domain towards the lowest scoring RNA sequence determined by the SIA, I was able to sufficiently demonstrate IMP3 to show a strong sequence specificity in RNA binding when comparing the best and worst binding RNA sequences. However, I do observe RNA binding with the lest preferred sequence at molar ratios below 1:4, whereas the IMP1 RRM12 di-domain did not bind its corresponding least preferred sequence at a ratio of 1:6. This suggests that IMP1 has an overall lower affinity towards RNA and that both IMP1 and IMP3 display sequence specificity.

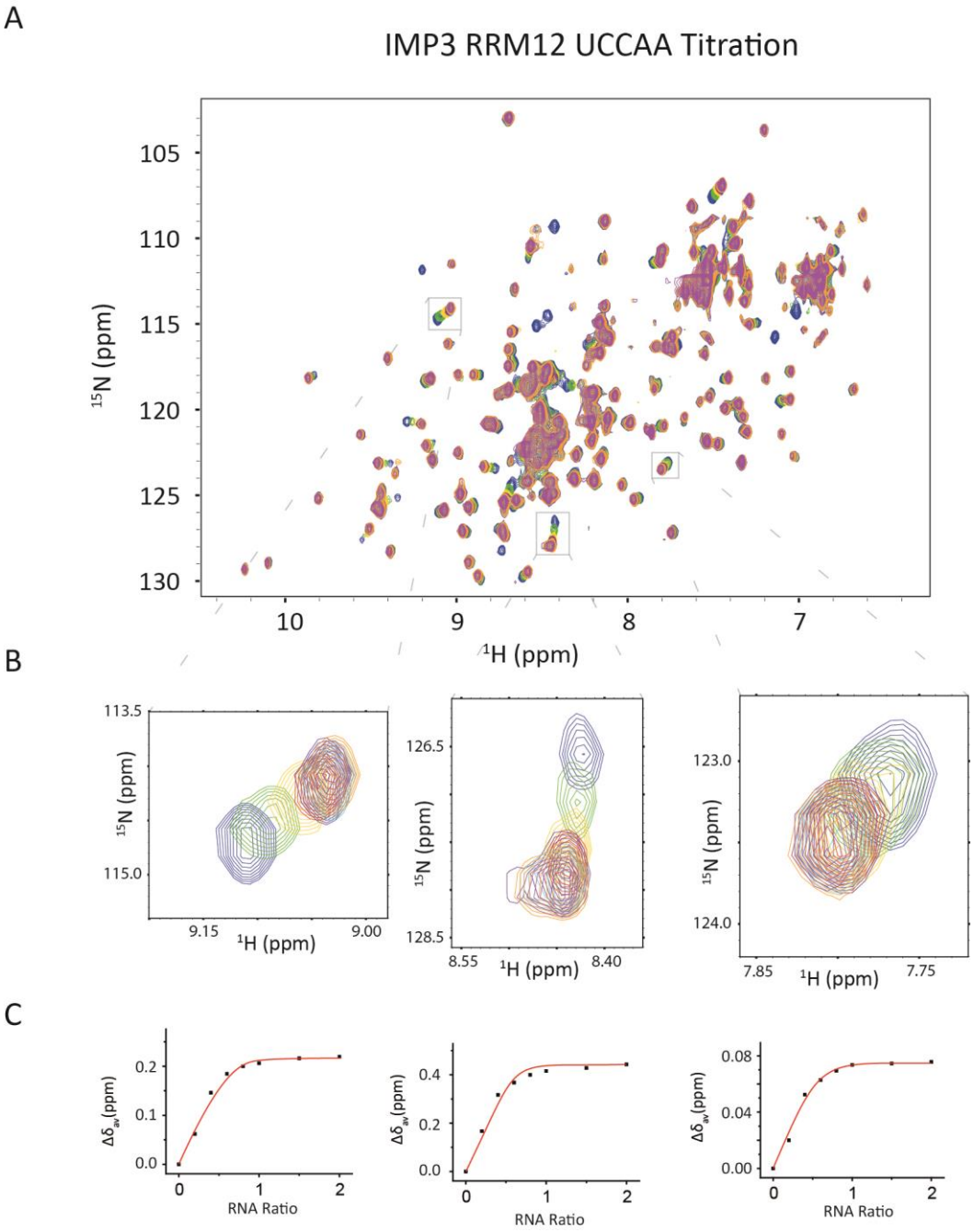


Figure 5.14: Titration of IMP3 RRM12 with UCCCAU oligonucleotide
A) 1H-15N SOFAST-HMQC overlaid spectra of 60 µM IMP3 RRM12 at protein:RNA ratios of 1:0 (Blue), 1:0.2 (Green), 1:0.4 (Gold), 1:0.6 (Red), 1:0.8 (Purple), 1:1 (Cyan), 1:1.5 (Orange), and 1:2 (Magenta). B) Zoomed in areas of spectra in (A) to highlight peak shifts used to calculate the K_d of IMP3 towards UCCAA RNA oligo. Grey boxes in A indicate the zoomed peaks. C) Chemical shift perturbations upon addition of increasing molar ratios of RNA. Curves represent the peaks highlighted above in panel (B) and were used to determine binding affinity.

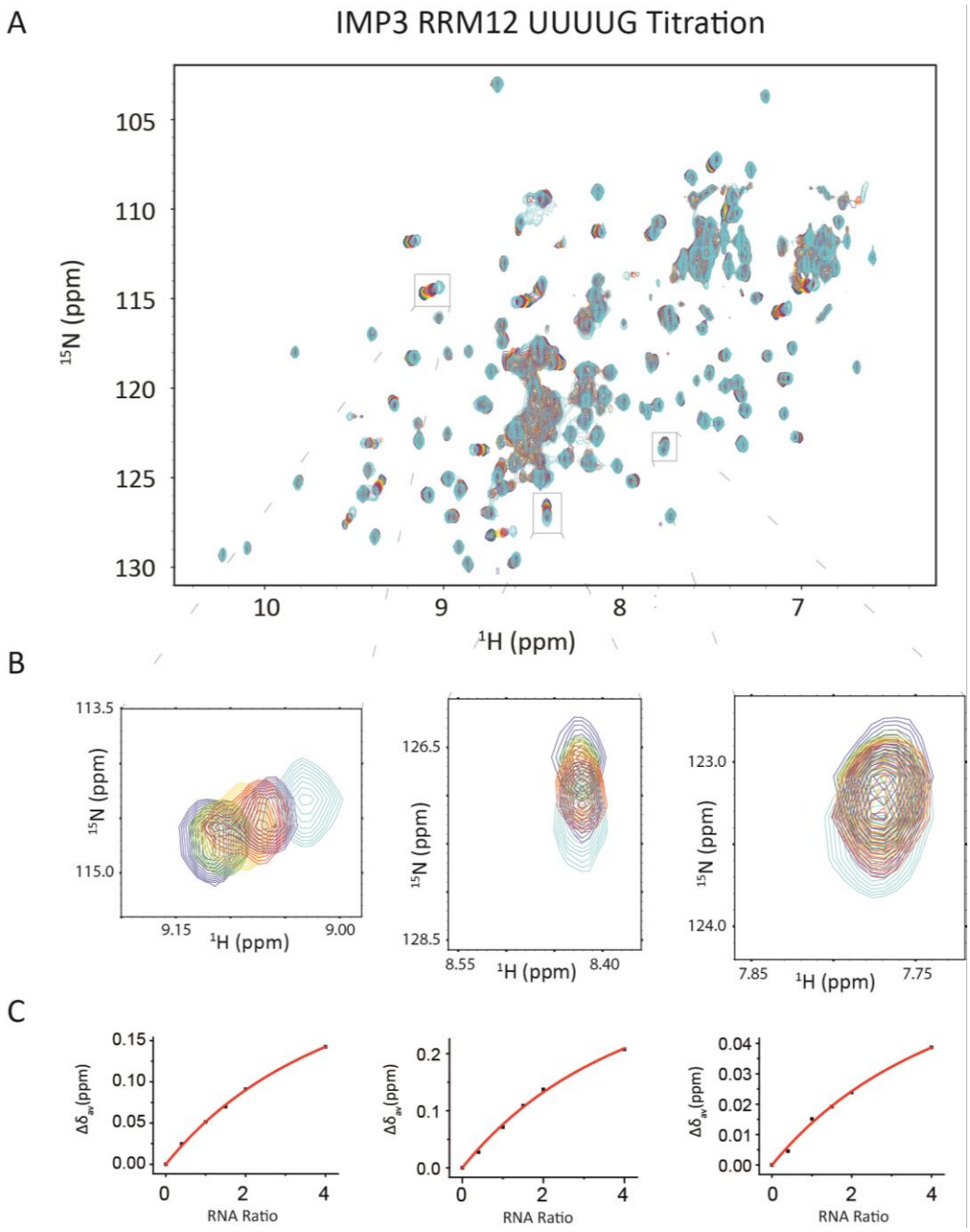


Figure 5.15: Titration of IMP3 RRM12 with UUUUG oligonucleotide
A) ^{1}H-^{15}N SOFAST-HMQC overlaid spectra of 60 µM IMP3 RRM12 at protein:RNA ratios of 1:0 (Blue), 1:0.5 (Green), 1:1 (Gold), 1:1.5 (Red), 1:2 (Purple), and 1:4 (Cyan). B) Zoomed in areas of spectra in (A) to highlight peak shifts used to calculate the K_d of IMP3 towards UUUUG RNA oligo. Grey boxes in A indicate the zoomed peaks. C) Chemical shift perturbations upon addition of increasing molar ratios of RNA. Curves represent the peaks highlighted above in panel (B) and were used to estimate binding affinity.

5.6 Relation between the RRM1 and RRM2 domains

The RRM1 and RRM2 domains of the IMP protein family are separated by a protein linker of just five amino acids. Previously solved RRM structures have shown that two consecutive RRMs that are separated by a short linker of 10–20 residues can interact with each other to form a compact fold.[220] These interdomain interactions can sometimes be induced in the presence of the domains' target RNA but can also occur in the free protein.[59,91,221–223]

The solution structure of IMP2 RRM1 and IMP3 RRM2 have previously been solved and deposited into the PDB (2CQH and 2E44, respectively). However, there is no direct structural information on the RRM12 di-domain and how these domains may interact is unclear. To explore the relationship between the two domains I performed T1 and T2 relaxation experiments to understand if the domains tumble independently or as a fixed unit. Ideally, I would analyse the relaxation of each domain individually within the two-domain construct. However, as we do not yet have assigned spectra for the constructs I determined the average rotational correlation times of the RRM12 di-domain.

Standard relaxation experiments were recorded on a ^{15}N-labelled sample to obtain T1, T2 values. Experiments were performed on a Bruker NMR spectrometer operating at 800 MHz. T1 and T2 values were determined for each peak by fitting an exponential decay to the peak volume over the course of the data collected. Peaks were excluded where overlap in the signals prevented accurate measurement of peak volume. In addition, to all peaks resulting from side chain N-H bonds, and peaks in the crowded central region of the spectra were discounted. Values were then averaged and standard errors calculated (Table. 5.1).

	IMP1 RRM12	IMP3 RRM12
T1 (s)	1.00 ± 0.11	0.94 ± 0.18
T2 (s)	0.065 ± 0.010	0.059 ± 0.014
T1/T2	16.3 ± 2.8	17.8 ± 5.2
App t_c (ns)	10.6 ± 1.0	11.1 ± 1.8

Table 5.1: Average relaxation values for IMP1 and IMP3 RRM12 constructs

Our relaxation studies for IMP1 and IMP3 RRM12 produced similar rotational correlation times of ~ 11 ns. A similar study performed on the 2 N-terminal qRRM domains of hnRNP F, which are also separated by a short protein linker, showed the two domains tumble inderpedently.[65] In this study a shorter average overall correlation time for qRRM1–qRRM2 of 8.3 ± 0.6 ns was reported. In contrast the two C-terminal RRM domains of PTB are known to make contact via a large interdomain interface consisting of 27 residues which form a large hydrophobic core.[224] Relaxation studies performed on the coupled RRM3-RRM4 domain pair resulted in an average overall correlation time of 10.4 ± 0.85 ns which is similar to the values I report for our RRM1-RRM2 domains. Furthermore, the PTB study investigated the effects of disrupting the RRM-RRM interface by incorporating mutations. Relaxation studies performed on mutated constructs that were no longer coupled produced an estimated overall correlation time of 8.0 and 6.9 ns for RRM3 and RRM4 domains respectively.[224]

Comparing our relaxation data to the above examples where two RRM domains are separated by a protein linker of similar size shows it is highly likely that our RRM1 and RRM2 domains are coupled in a fashion that results in a compact RRM12 unit. Interestingly, the RRM1 domain of our RRM12 construct contains a β-hairpin in the protein loop connecting α2-helix and β4-strand.[53,57] The presence of a β-hairpin is not uncommon in RRM domains, and it has been shown to mediate protein-protein interactions. For example, in the case of Sxl[91] and HuD[223] the β-hairpin of the N-terminal RRM domain interacts with the β-sheet of their respective RRM2 domain. This further supports a potential interdomain contact between the RRM1 and RRM2 domains of the IMP proteins.

5.7 Rational design of mutations to abolish RNA binding of the RRM12 di-domain

Our NMR RNA binding investigation showed both IMP1 and IMP3 RRM12 di-domains can recognise RNA in a sequence specific manner. The next stage in our investigation was to determine amino acid residues that were involved in RNA recognition. Identifying key amino acid residues involved in RNA recognition provides useful insight to better characterise the RRM-RNA interaction and help to identify the RNA binding interface. The aim was to mutate RNP residues within the RRM1 domain and investigate the effect of the mutations within our RRM12 constructs on RNA binding. I hoped to produce RRM12 mutants that could no longer recognise RNA, similar to the KH domain GDDG mutants. These mutants could then be used as molecular tools to determine the contribution of the RRM domains in RNA target recognition in the context of the full-length proteins.

As previously stated, I was unable to identify a possible RNA binding surface within the RRM2 domain of the IMP family. Therefore, as an initial approach I focused only on residues within the RRM1 domain. For this purpose, examples of RRM mutational approaches that had abolished RNA binding were used. Our group has previous experience with abolishing the RNA binding properties of the RRM domains of RNA15,[120] RMB38[118] and the RRM2 domain of FIR (data unpublished). I used these as templates and performed a primary amino acid sequence alignment of these RRM domain examples with the RRM1 domain of the IMP family (Figure 5.16).

This multiple sequence alignment identified four residues that were conserved in the RRM1 domain and had previously been successfully mutated to abolish the RNA binding of the RRM domain examples. These residues were Y5, K36, Y39 and K66 (Figure 5.16). Published studies and in group experience showed that typically two amino acid mutations are required to inhibit RNA recognition of a RRM domain. This involves mutating a positively-charged residue to a negatively-charged amino acid in combination with removing an aromatic side chain. These mutations incorporate large differences in the composition of the

WT amino acid side chains which can result in major structural disruption. Therefore, I first mutated only the Y39 residue, as in the canonical RNP recognition this residue provides a fundamental contact with the RNA. If this was unsuccessful I planned to produce double mutants; Y5AK36E, Y5AK66E, K36EY39A, and Y39AK66E. Due to time constrictions not all of these mutants were generated as part of my PhD thesis.

Site directed mutagenesis was used to produce all mutant constructs in the same pET-M11 vectors. After successful cloning, the constructs were expressed and purified using the same denaturing urea purification protocol that was set up for the WT proteins (Chapter 2.2). For simplicity of this report here only data for mutant constructs that provided the most useful insight into the RNA binding properties of the RRM12 di-domain are presented.

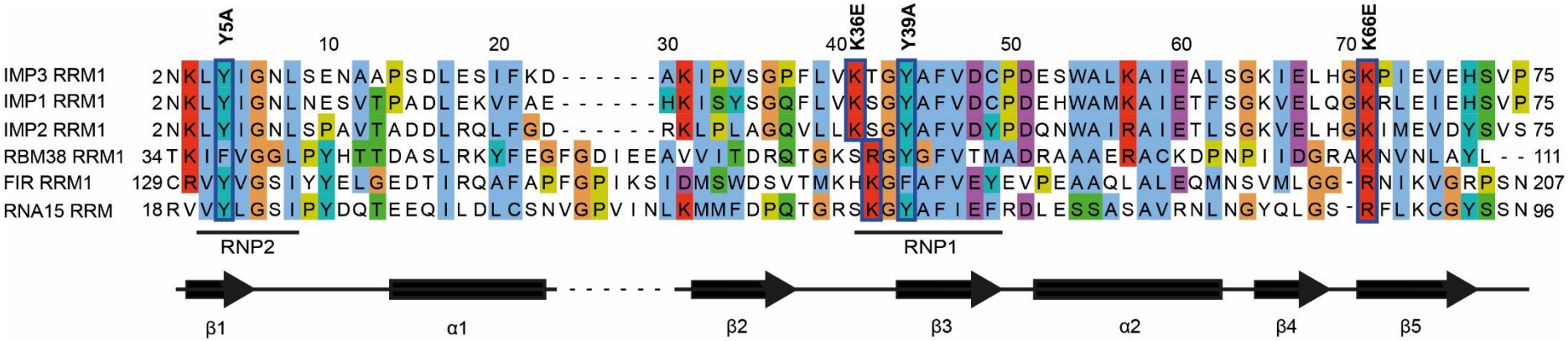


Figure 5.16: Multiple sequence alignment of RRM1 of the IMP family with RRM domains of RBM 38, FIR and RNA 15
Residues previously mutated in RNA 15, FIR and RBM 38 that abolished RNA binding are highlighted in blue boxes with the corresponding residues located in the IMP RRM1 domain via sequence alignment. Mutation to be incorporated in the IMP RRM12 di-domain constructs are displayed above the alignment, from N-terminal to C-terminal: Y5A, K36E, Y39A and K66E. Secondary structural prediction is displayed below the alignment in order to locate the position of the mutated residues within the structure of the RRM1 fold.

5.7.1 IMP1 RRM12 Mutagenesis

The two mutant constructs for the IMP1 RRM12 construct I will report on are the IMP1 RRM12 Y39A and IMP1 RRM12 Y39AK66E mutants. The two constructs expressed and purified to a similar yield as the WT IMP1 RRM12 construct. Size exclusion chromatography was used as before to purify the constructs once the purification His-tag had been removed (Figure 5.17). To determine the effects of the mutations on protein structure far UV CD spectra and ^{1}H-^{15}N SOFAST-HMQC NMR experiments were performed (as above) and the corresponding spectra compared to the WT protein (Figure 5.18).

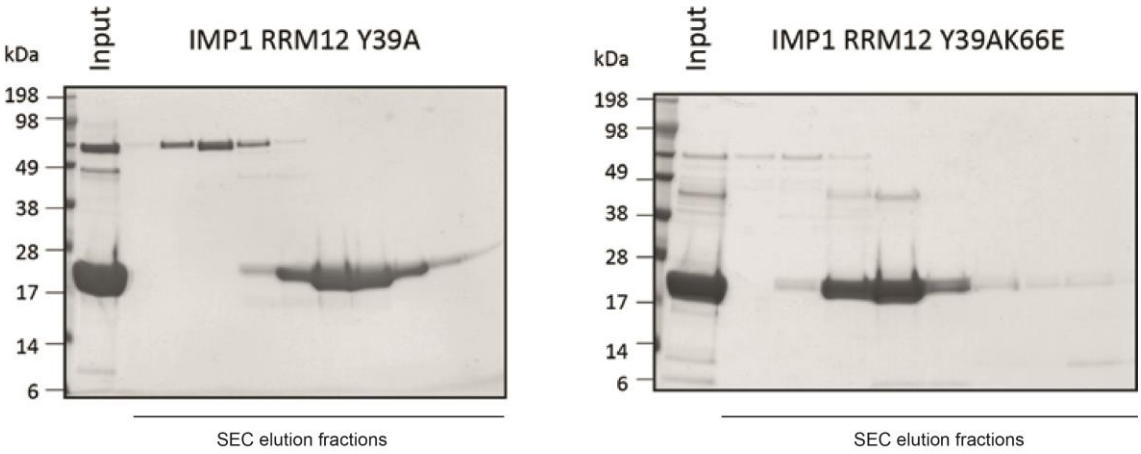


Figure 5.17: SDS-PAGE gel analysis of purification fractions collected during size exclusion chromatography of IMP1 RRM12 mutant constructs
The first lane in both gels represents the input samples loaded onto the column after protein refolding and His-tag removal via TEV protease digestion. Left gel shows the purification fractions for the IMP1 RRM12 Y39A mutant construct and the Right for the double Y39AK66E mutant.

Comparing the far UV CD spectra of the three proteins revealed the IMP1 RRM12 Y39AK66E mutant contained a high proportion of random coil with respect to the WT IMP1 RRM12 construct. This is characterised by the shift of the maxima from 215 nm towards the upper end of the spectrum where random coli structures absorb more strongly. There was also a shift in the maxima at 215 nm for the Y39A mutant, however this was less pronounced than the shift of the double

mutant. Comparing the region around 222 nm, which is the typical second maximum produced by α-helical structures, I observed a reduction in signal intensity for the IMP1 Y39A mutant, suggesting a loss of α-helix content. In turn, both mutations have had an effect on the secondary structure content of the IMP1 RRM12 construct, with the effects of Y39AK66E being more pronounced (Figure 5.18A).

Comparing the ^{1}H-^{15}N SOFAST-HMQC spectra of the mutants to the WT revealed a dramatic loss of structured residues in the IMP1 Y39AK66E double mutant (Figure 5.18). This is shown by the loss of peaks in the dispersed region of the spectra up field from ~ 8 ppm. This region typically reports residues in β-sheet structures due to the nature of the amino acid chemical shifts. Counting the number of peaks in the WT spectrum compared to the mutant spectrum in this region (discounting the crowded central region) revealed a 43% reduction in the number of peaks. In contrast, the NMR spectrum of the IMP1 RRM12 Y39A mutant was more comparable to the WT IMP1 RRM12 spectrum (Figure 5.18B). However, there was a higher number of peaks in the central region, suggesting a certain degree of aggregation had occurred in the mutant protein. This could suggest the presence of two species in the mutant IMP1 RRM12 Y39A spectra. A folded species that resembles that of the WT protein, and an aggregated from. Alternatively, the mutation could potentially have affected the stability of the protein.

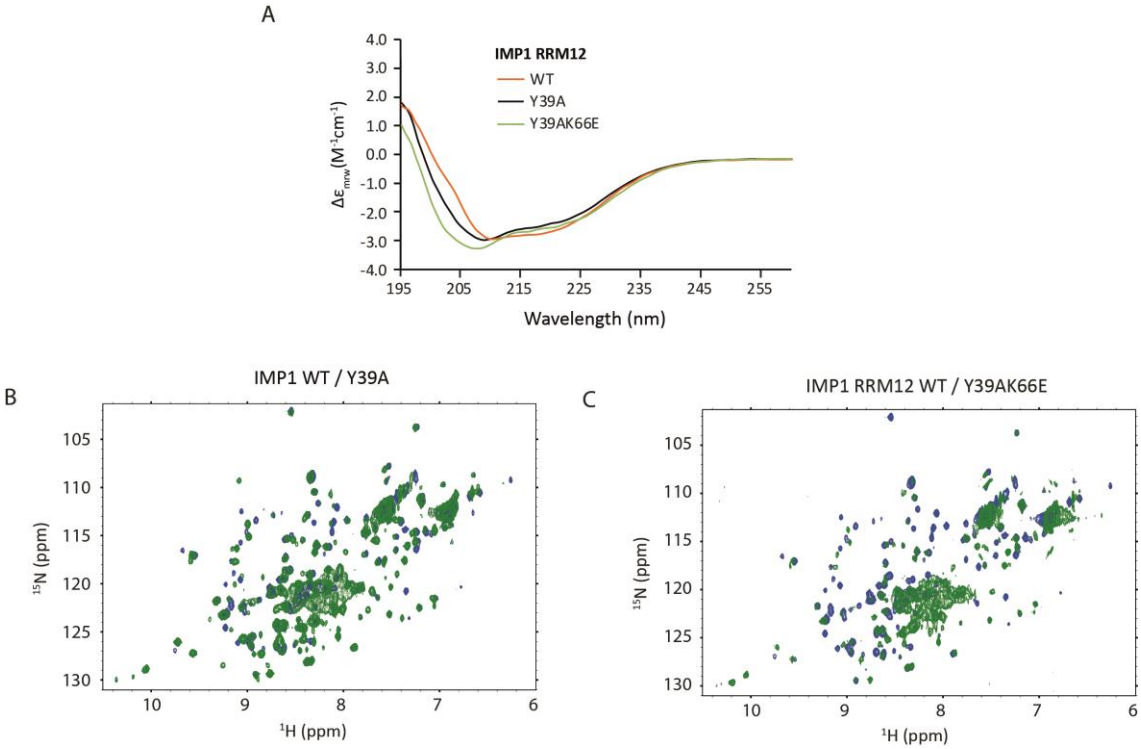


Figure 5.18: Effect of mutations of predicted RNA binding residues Y39A and Y39AK66E on IMP1 RRM12 protein structure at 25ºC
Far UV CD spectra (195 -260 nm) of WT and mutant IMP1 RRM12 di-domain constructs are displayed in (A). Proteins were buffered in the same buffer of sodium phosphate pH 7.4, 100 mM NaCl, 0.5 mM TCEP and diluted to a concentration of 0.15 mg/ml. B) Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of mutant IMP1 RRM12 Y39A (Green) and WT IMP1 RRM12 (Blue). Both proteins were concentrated to 50 µM and buffered in the same buffer as the far UV CD samples. C) Same as B but mutant spectrum in Green is IMP1 RRM12 Y39AK66E.

5.7.2 IMP3 RRM12 Mutagenesis

Two IMP3 RRM12 mutant constructs that will be reported on here are the two double mutants IMP3 RRM12 K36EY39A and Y39AK66E. The same rationale was followed as with the IMP1 mutants. The IMP3 mutants again expressed to similar yields as the IMP3 RRM12 WT protein (Figure 5.19).

Figure 5.19: SDS-PAGE gel analysis of purification fractions collected during size exclusion chromatography of IMP3 RRM12 K36EY39A mutant construct
The first lane represents the input sample loaded onto the column after protein refolding and His-tag removal via TEV protease digestion, the latter are the purification fractions collected during size exclusion chromatography.

The far UV CD spectra analysis showed the IMP3 RRM12 Y39AK66E mutant to have a similar spectrum as the corresponding IMP1 RRM12 Y39AK66E mutant but with a slightly greater shift towards random coil. In contrast, the IMP3 RRM12 K36EY39A mutant exhibited a minimal shift towards random coil compared to the IMP1 Y39A mutant. However, the reduction in signal intensity at 222 nm of the IMP3 RRM12 K36EY39A mutant compared to the WT was similar to the one observed for the IMP1 RRM1 Y39A mutation and its WT counterpart (Figure 5.20A).

Comparing the ^{1}H-^{15}N SOFAST-HMQC spectra of the IMP3 RRM12 mutants and the WT IMP3 RRM12 domain showed a similar trend for the Y39AK66E mutant as seen for the same mutation in IMP1. I observed a reduction in the number of peaks in the dispersed region up field of ~ 8 ppm. Counting the differences in the peaks in this region showed a 29% reduction in the number of mutant peaks compared to the WT. As with the IMP1 Y39AK66E mutation I also observed a large number of peaks in the central region of the spectra suggesting unfolded / aggregated protein (Figure 5.20C). In contrast the IMP3 RRM12 K36EY39A

mutant protein produced a proton-nitrogen correlation spectrum that was comparable with the WT IMP3 RRM12 protein with a minimal number of peak shifts between the two spectra, and no increase in the number of peaks in the central region (Figure 5.20B).

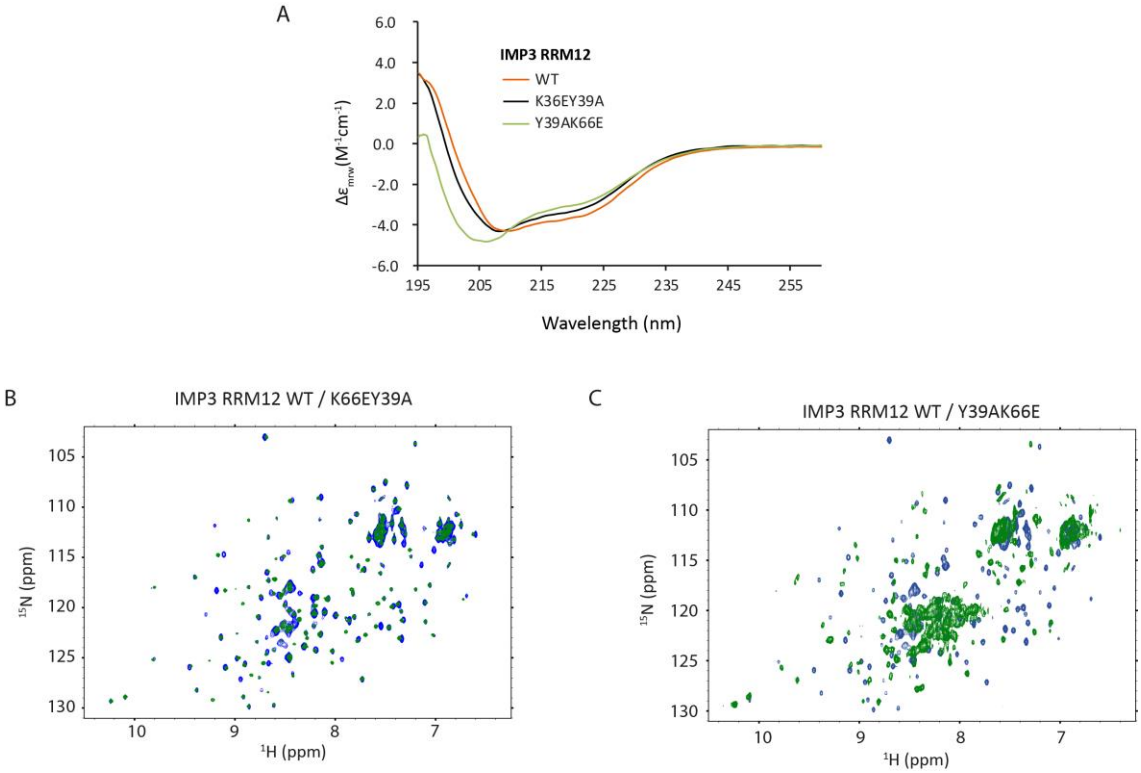


Figure 5.20: Effect of mutations of predicted RNA binding residues K36EY39A and Y39AK66E on IMP3 RRM12 protein structure at 25ºC
Far UV CD spectra (195 -260 nm) of WT and mutant IMP3 RRM12 di-domain constructs are displayed in (A). Proteins were buffered in the same buffer of sodium phosphate pH 7.4, 100 mM NaCl, 0.5 mM TCEP and diluted to a concentration of 0.15 mg/ml. B) Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of mutant IMP3 RRM12 K36EY39A (Green) and WT IMP3 RRM12 (Blue). Both proteins were concentrated to 50 µM and buffered in the same buffer as the far UV CD samples C) Same as (B) but mutant spectrum in Green is IMP3 RRM12 Y39AK66E.

In conclusion, the Y39AK66E mutation pair in both the IMP1 and IMP3 RRM12 constructs appears to result in the unfolding of a proportion of the construct. The far UV CD spectra and SOFAST-HMQC spectra of the Y39AK66E mutant for both IMP1 and IMP3 show a proportion of the protein remains folded. As the mutations I have introduced reside within in the RRM1 domain, I can speculate that the RRM1 domain has been affected more by the mutation then the RRM2 and the folded protein that remains is reporting from the RRM2 domain of the di-domain construct.

Using this as our rationale, I investigated if the partially unfolded Y39AK66E IMP1 and IMP3 mutants were able to bind RNA. This would give us an indication if the RRM2 domain was capable of binding RNA in the absence of the RRM1 domain. I added a 1:5 molar ratio of RNA with each IMP RRM Y39AK66E mutant and performed ^{1}H-^{15}N SOFAST-HMQC analysis to determine if any of the peaks shifted. As I was unsure how the unfolding of the RRM1 domain could potentially alter the RNA recognition properties of the RRM2 domain (if the domain possessed such properties) I used randomised RNA oligos of five nucleotides in length rather than the RNA sequences determined by the SIA study performed on the WT proteins.

For both IMP proteins, the Y39AK66E mutant protein was unable to recognise RNA as no chemical shifts were observed at a 1:5 protein:RNA molar ratio (Figure 5.21 & 5.22). I then went back to our original RNA binding NMR spectra of the WT proteins with the randomised RNA oligos (Figure 5.8 & 5.9). For the peaks I could compare between the WT IMP proteins and their Y39AK66E counterparts, none of the peaks that shifted upon addition of RNA in the WT proteins were present in the mutant protein spectrum. Thus, confirming the part of the RRM12 di-domain that remains folded in the Y39AK66E mutation cannot bind RNA.

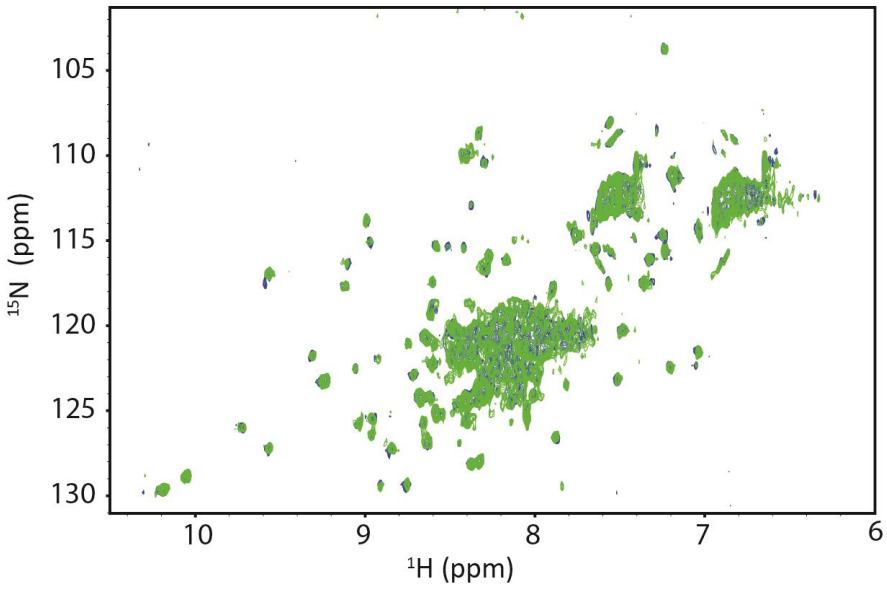


Figure 5.21: IMP1 RRM12 Y39AK66E 1H-15N SOFAST-MHQC spectra of free protein and upon addition of 1:5 molar ratio of 5N oligonucleotide
^{1}H-^{15}N SOFAST-HMQC overlaid spectra of 50 µM IMP1 RRM12 Y39AK66E at protein:RNA ratios of 1:0 (Blue) and 1:5 (Green)

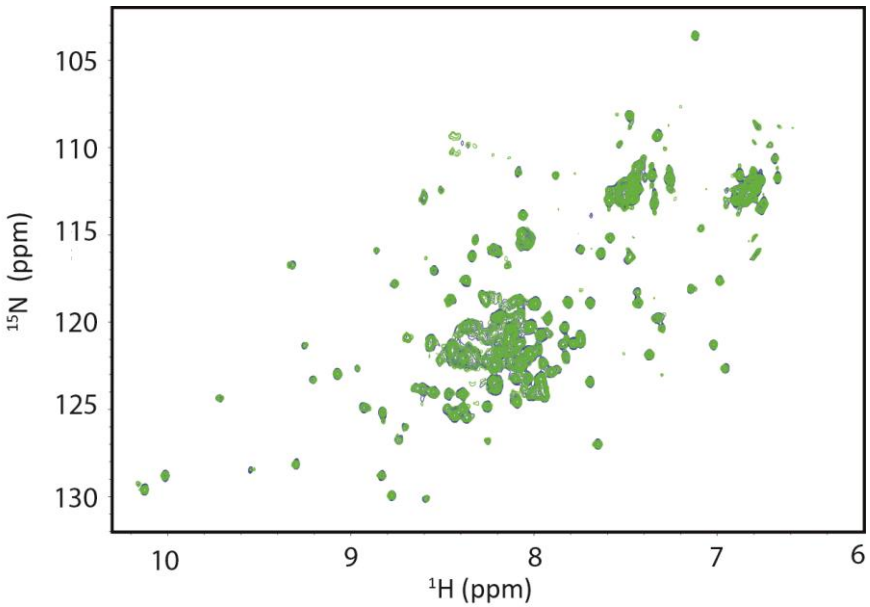


Figure 5.22: IMP3 RRM12 Y39AK66E 1H-15N SOFAST-MHQC spectra of free protein and upon addition of 1:5 molar ratio of 5N oligonucleotide
^{1}H-^{15}N SOFAST-HMQC overlaid spectra of 60 µM IMP3 RRM12 Y39AK66E at protein:RNA ratios of 1:0 (Blue) and 1:5 (Green)

I then investigated the RNA binding properties of the other two mutants, the IMP1 RRM12 Y39A and IMP3 RRM12 K36EY39A proteins. For these mutants the preferred RNA sequences from the SIA study was used to investigate RNA binding properties (IMP1 UCCCG and IMP3 UCCAA). The IMP1 RRM12 Y39A mutant did not produce any chemical shifts upon addition of a 1:5 protein to UCCCG molar ratio (Figure 5.23). This suggests that the loss of a single aromatic residue within the RNP motif (located on the β3-strand of the β-sheet) is capable of inhibiting RNA recognition. RNA titrations of the IMP1 RRM12 WT protein revealed that the WT protein has a weak RNA binding affinity, ~ 150 µM. In contrast, the IMP3 RRM12 WT protein displayed an RNA binding affinity of ~ 1 µM. The single point mutations I tested for the IMP3 RRM12 protein, K36E and Y39A, on their own attenuated RNA binding but were not sufficient to abolish RNA binding individually (data not shown).

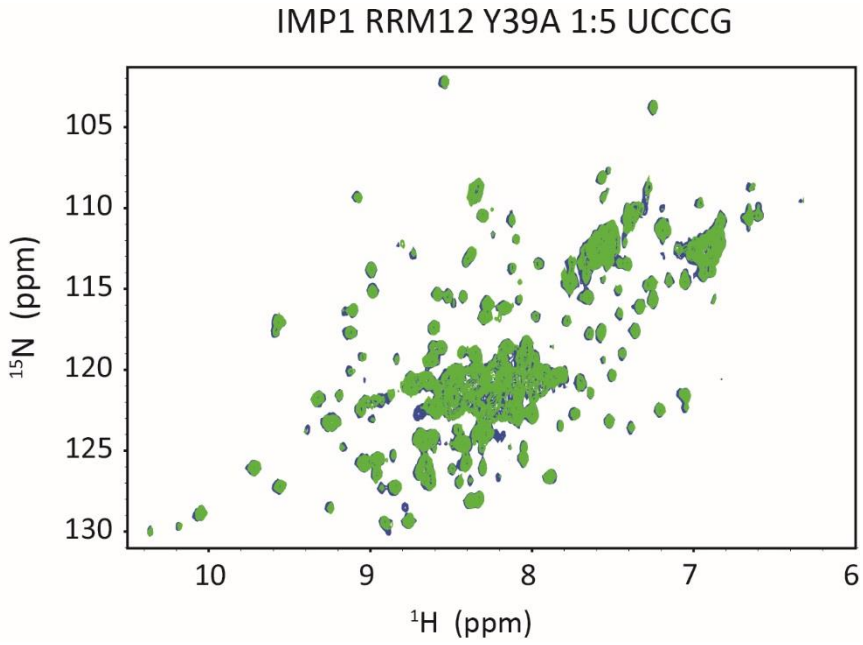


Figure 5.23: IMP1 RRM12 Y39A upon addition of UCCCG RNA oligonucleotide
Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of 60 µM IMP1 RRM12 Y39A in free from (Blue) and 1:6 protein to UCCCG molar ratio (Green). Recorded at 25^{o}C.

I performed a RNA titration with IMP3 RRM12 K36EY39A mutant di-domain and preferred IMP3 RNA sequence UCCAA. ^{1}H-^{15}N SOFAST-HMQC spectra were recorded at protein:RNA ratios of 1:0.5, 1:1, 1:2, 1:4, and 1:6 (Figure 5.24). Analysis was performed by manual measurement of peak shifts. The observed peak shifts for IMP3 RRM12 K36EY39A mutant were extremely small and so error measuring peak shifts was high. In total 15 shifting peaks were chosen, of these 15 peaks only three peaks were able to be tracked with accuracy. The size of the chemical shift perturbations was extremely reduced compared to the titration of the WT IMP1 RRM12 protein and only two peaks appeared to reach saturation at 1:6 protein:RNA ratios (Figure 5.24B & C). From our RNA titration I concluded the IMP3 RRM12 K36EY39A mutation was successful at inhibiting the RNA binding properties of the IMP3 RRM12 di-domain.

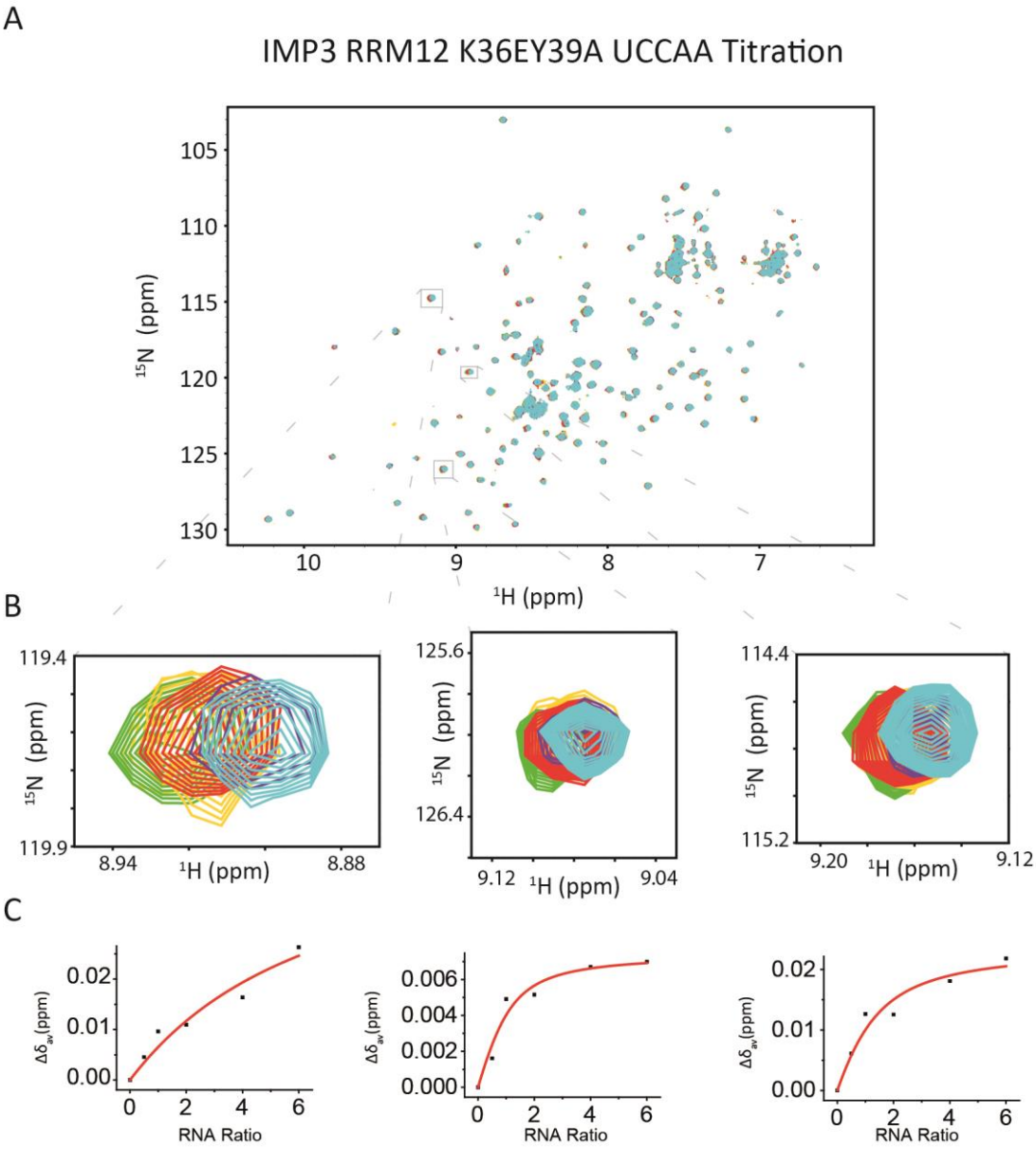


Figure 5.24: Titration of IMP3 RRM12 K36EY39A with UCCAA oligonucleotide

A) ^{1}H-^{15}N SOFAST-HMQC overlaid spectra of 60 µM IMP3 RRM12 K36EY39A at protein:RNA ratios of 1:0 (Blue), 1:0.5 (Green), 1:1 (Gold), 1:2 (Red), 1:4 (Purple), and 1:6 (Cyan). B) Zoomed-in spectra to highlight peak shifts used to calculate K_d. Grey boxes in A indicate the zoomed peaks. C) Chemical shift perturbations upon addition of increasing molar concentrations of RNA.

5.8 Thermal stability of the IMP1 and IMP3 RRM12 di-domains

Finally, I investigated the thermal stability of the IMP1 and IMP3 RRM12 di-domain constructs and compared it to the thermal stability of the RRM12 mutant constructs. As with the KH domains, to use RRM12 RNA binding mutants for in-cell RNA binding studies, our incorporated mutations should abolish RNA binding without major structural changes or altering the proteins' stability. This is to enable us to develop a system where only the RNA binding properties of the system are lost, and any other functionality of the domain is preserved due to the correct folding of the domain. Additionally, the stability of the mutated domain is important given the longevity and temperature of *in vivo* studies these mutant constructs could potentially be used for. Given the reported findings that the RRM domains of IMP1 are involved in protein-protein interactions[155] it is critical that the fold of the protein is maintained when the binding mutations are incorporated. This enables us to study the effects the loss of RNA binding has on the IMP RRM domains ability to bind to protein partners.

Comparing the RNA binding affinity of IMP1 RRM12 to that of IMP3 RRM12, it is likely that the RRM domains of IMP1 play only a minor role in recognising RNA compared to those of IMP3. However, it would be insightful to determine how the loss of IMP1 RRM12 RNA binding affects RNA target selection and translocalisation in the cell. Therefore, I performed CD thermal denaturation studies on the WT IMP RRM12 proteins and the four RNA binding mutants described above. Considering the Y39AK36E mutation partially unfolded a proportion of the RRM12 constructs, I included these mutants in our thermal denaturation to see if I could determine the thermal stability of the RRM2 domains without its folded RRM1 partner.

Based on the far UV CD spectra, I decided to perform the thermal denaturation while monitoring at a wavelength of 210 nm. This was due to there being a large difference in signal this region between the IMP RRM12 WT proteins and their corresponding partially unfolded Y39AK66E mutants. Proteins were diluted to a

protein concentration of 0.15 mg/ml and melted from 2°C to 95°C with a 2°C / min gradient. Thermal denaturation curves were converted from millidegrees into Δε per mean residue weight and curves of best fit were plotted. Apparent melting T_m was then calculated from the fitted curves (Figure 5.25 and 5.26).

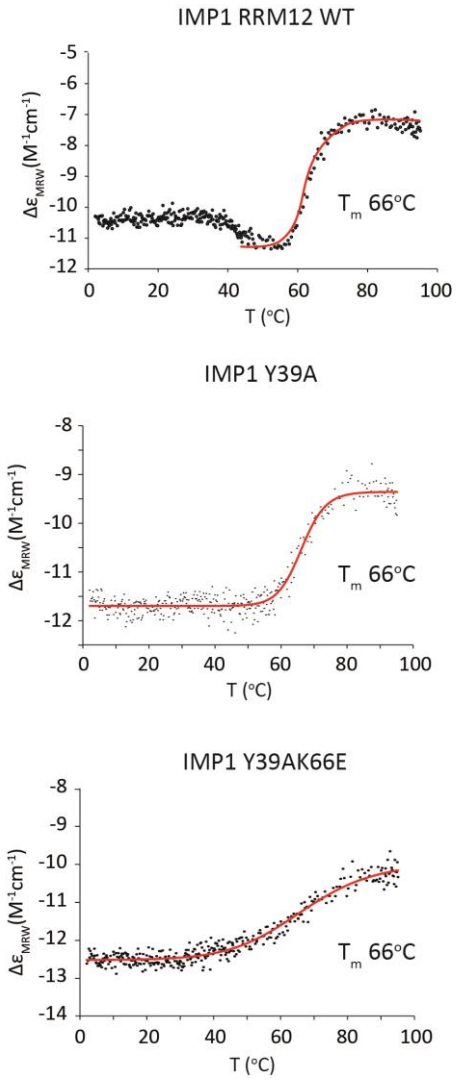


Figure 5.25: Comparing the thermal stability of IMP1 RRM12 WT construct and RNA binding mutants
Thermal unfolding of IMP1 RRM12 WT (Top) Y39A mutant (Middle) and K36EY39A double mutant (Bottom). Unfolding was monitored at 210 nm. Plots show full raw data points in black and fitted curve in red. Note: fitted curve was used to estimate apparent T_m values and only the parts of the curve used for T_m estimation have a fitted curve. Apparent T_m values are displayed for each protein.

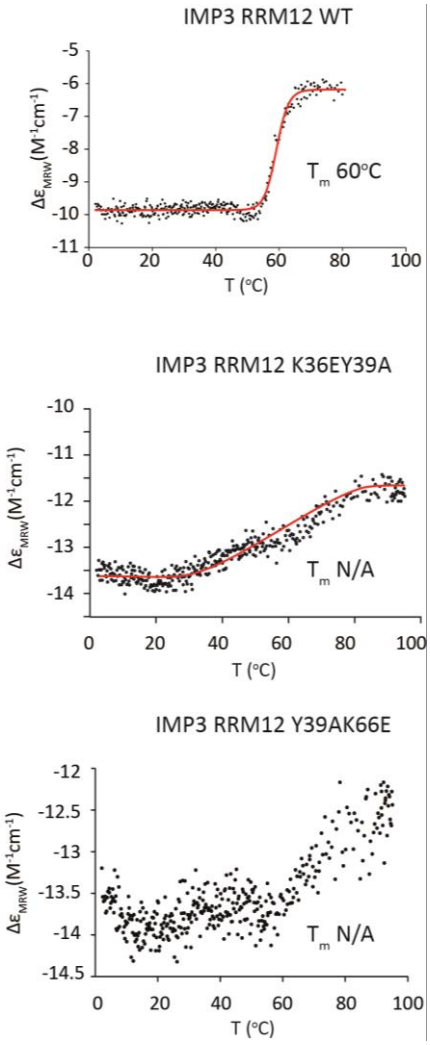


Figure 5.26: Comparing the thermal stability of IMP3 RRM12 WT construct and RNA binding mutants
Thermal unfolding of IMP3 RRM12 WT (Top) K36EY39A mutant (Middle) and K36EY39A mutant (Bottom). Unfolding was monitored at 210 nm. Plots show full raw data points in black and fitted curve in red. Note: fitted curve was used to estimate apparent T_m values. Not all denaturation curves could be fitted with accuracy or used to calculate an apparent T_m. Apparent Tm values are displayed for each protein where estimation was possible.

The IMP1 and IMP3 RRM12 WT proteins produced thermal denaturation curves that were able to generate a fit that could be used to calculate T_m values (Figure 5.25 & 5.26). This revealed the IMP1 RRM12 construct to be slightly more stable then IMP3 with T_m values of 66°C and 60°C respectively. Both curves appear to represent only one single unfolding event rather than displaying two separate

unfolding events for each individual RRM domain. For the IMP3 RRM12 construct mutants (K36EY39A and Y39AK66E) the difference in the signal at 210 nm between the fully folded and fully unfolded state was extremely small (Figure 5.26). This in turn, increased the level of noise in the raw data points. Due to this fact a fit could not be calculated for the IMP3 RRM12 Y39AK66E mutant accurately. Apparent T_m values could also not be estimated.

This small difference in signal between the folded and unfolded state of the IMP1 RRM12 mutants (Y39A and Y39AK66E) was also reduced compared to the IMP1 RRM12 WT protein (Figure 5.25), but to a lesser extent than the IMP3 RRM12 mutants. The thermal denaturation curve fits for the IMP1 RRM12 mutants generated an apparent T_m that was comparable to the IMP1 RRM12 WT protein (60^{o}C). However, due to small signal difference between folded and unfolded protein states of the mutant constructs at 210 nm, I am unable to draw conclusions regarding the effect of the mutations on the thermal stability of the proteins. Further investigation is required.

5.9 Discussion

The multifunctionality of RNA binding proteins relies on their ability to recognise a variety of different RNA transcripts. In order to do this, RBPs utilise different combinations of RBDs to bind different RNA targets via a process known as combinatorial recognition.[49,92] The IMP protein family contain six such canonical RNA binding domains, yet current investigations into understanding the direct RNA binding properties of these domains have largely been focused on the four C-terminal KH domains. Here I have completed an initial investigation into the RNA binding properties of the N-terminal RRM domains and compared RNA recognition of the IMP1 RRM12 domains with that of its homologue IMP3.

I was successful in expressing and refolding IMP1 and IMP3 RRM12 constructs. This provided us with a di-domain system in which I could assess the RNA binding properties of the domains. I identified that both IMP1 and IMP3 RRM12

domains are capable of recognising RNA in a sequence specific manner. The specific RNA motifs the two N-terminal RRM domains recognise are different for IMP1 (C – A/C/G – A/C/G – G) and IMP3 (C – C – A – A) but both proteins displayed a negative bias towards poly-U sequences. The two proteins also differ substantially in the RNA binding affinity displayed by their RRM domains with IMP3 RRM12 possessing ~100-fold higher binding affinity than IMP1. This suggests the IMP3 RRM domains may play a more fundamental role in RNA selection than the RRMs of IMP1, which is consistent with current studies[83,141,154]. Without greater structural understanding of the protein-RNA interactions, we are unable to conclude why I observe such differences in the RNA binding properties. However, given the high sequence similarities within the RNP motifs between the two proteins, the difference is likely to be the result of different amino acids residing in the loops between the β-strands and α-helices. Indeed, it is often interactions with residues in these regions that determine nucleobase specificity of RRM domains.

I also explored the relationship between the RRM1 and RRM2 domains of both proteins by investigating their rotational correlation times using standard NMR relaxation experiments. I could show that IMP1 and IMP3 RRM12 domains tumble with the same overall average correlation time (~11 ns). Comparing this with other examples of studied RRM domains in isolation and as di-domains[65,220] I concluded the RRM domains of IMP1 and IMP3 interact with each other and tumble as a fused unit. The nature of the interaction between the domains could also account for differences observed in RNA binding.

Both our RNA binding and relaxation experiments would benefit from assigned spectra of both protein constructs. This would enable us to identify the residues responsible for RNA binding and to begin building a model of the protein RNA interaction surface. It would give insight into whether just the RRM1 domain provides the interaction surface for RNA recognition or if the surface extends across both domains. I would also be able to assign individual relaxation values

to the amino acids within the RRM1 and RRM2 domain and identify any differences between the two or linker region.

Our rational based mutagenesis approach to abolish RNA binding of the two IMP RRM12 constructs was successful. I determined that mutation of a single residue within the IMP1 construct (Y39A) was sufficient to inhibit RNA binding. The higher affinity of the IMP3 construct towards RNA resulted in the need to mutate two residues to observe sufficient RNA binding inhibition (K36EY39A). However, due to the small signal difference observed between the folded and unfolded protein states in our CD thermal denaturation study, I was unable to accurately determine how these mutations affect the proteins' stability in relation to the WT proteins. These mutants require further characterisation to determine how the mutations have affected the protein fold before they can reliably be used in further studies. Such characterisation could include using NMR to assess thermal unfolding by recording ^{1}H-^{15}N SOFAST-HMCQ spectra at increasing temperature intervals.

Interestingly, a recent study exploring IMP3 RNA target selection in pancreatic ductal adenocarcinoma cells performed both iCLIP and RIP-seq experiments to identify RNA targets bound to the endogenous IMP3 protein.[168] They identified an enrichment of binding sites within the 3' UTR region of transcripts. These binding clusters were observed to peak within a 25 nucleotide window centred on predicted miRNA target sequences. One such site that was observed to have enriched IMP3 binding was the seed sequence the mir-9 miRNA 'ACCAAAG'. This target sequence contains the RNA recognition motif I have identified for the RRM12 domain of IMP3 via SIA. Furthermore, the group showed that IMP3 was able to protect RNA transcripts from mir-9 mediated decay by using a luciferase reporter assay[168]. The function of IMP3 protecting RNA transcripts in pancreatic ductal adenocarcinoma cells from miRNA degradation, specifically in the case of mir-9, could result from RNA recognition via the RRM12 domains. This is potentially an important mechanism by which IMP3 promotes tumorigenesis in cancer. In fact, a recent review covering IMP3 upregulation in cancers quoted IMP3 as being identified in more cancer cells and tissues than any of the other

IMP family members.[167] Additionally, in certain cancer types such as pancreatic cancers, IMP3 is observed to be the highest upregulated RBP and present in over 90% of all invasive pancreatic ductal adenocarcinomas.[168] These are strong links suggesting IMP3 is a fundamental driver of metastasis in these cancer forms. Therefore, further investigation into the IMP3 RRM mutant constructs I developed to abolish RNA binding is needed. Mutations that abolished the RNA binding of the RRM domain without altering the protein fold and stability would provide a useful tool to investigate if IMP3 mediated protection of mir-9 degradation is mediated via the RRM domains. Such a connection could provide a potential target for the development of therapeutics aimed at treating IMP3 dependent cancers.

Chapter 6. General Discussion

RNA metabolism is a finely tuned multi-stepped process that is controlled by RNA binding proteins and other RNA molecules. This extensive regulatory system is made up of hundreds of RNA binding proteins using a diverse range of RNA selection mechanisms to elicit their function. As post-transcriptional control of gene expression is vital for normal cell function, it is unsurprising that small changes in RBP function can result in a wide array of human pathologies.

RNA binding proteins by nature are multifunctional. Their ability to select multiple RNA transcripts is key to their multifunctionally and enables them to control several different post-transcriptional regulatory pathways. RNA binding proteins need to be able to recognise a diverse range of RNA transcripts due to the wide variety of different RNA sequences and structures within the cell. This class of proteins can recognise such a diversity of transcripts because of their common modular structure. This structure, of often multiple RBDs which are generally small structured units, enables the full-length protein to recognise different RNA transcripts via combinatorial binding.

Combinatorial binding involves the accumulation of multiple weak interactions of the individual RBDs with short RNA stretches, to achieve a highly specific and high affinity interactions. However, this process is not well understood. Additionally, RNA binding domains are able to bind multiple RNA sequences with different affinities. The biological reverence of these weaker interactions is poorly understood, in part due to *in vitro* methods mainly reporting on only the high affinity RNA targets of RNA binding domains.

Much work has been done to understand how RNA binding proteins recognise RNA targets in a sequence specific manner *in vitro,* and a general understanding of the structural interactions established between RBDs and their RNA targets is developing. However, recently there has been great expansion into high

throughput techniques that are designed to investigated RBP binding on a large scale with multiple RNA targets. Such studies also enable the investigation of RNA target selection *in vivo,* with UV mediated cross link and immunoprecipitation protocols identifying RNA targets for an array of RNA binding proteins on the transcriptome wide level. However, these techniques remain inherently noisy, and the RNA binding targets identified *in vitro* do not always correlate well with those identified in CLIP studies.

To date, no such CLIP study has been performed with the aim of understanding how individual RNA binding domains contribute to overall RNA recognition of a full length RBP protein. In my thesis I have focused on the multi-RNA binding domain protein IMP1.

The IMP1 protein is as a model system to begin to explore how combinatorial RNA recognition is mediated *in vivo*. IMP1 is an important oncofoetal RNA binding protein which plays a fundamental role in embryonic development, but also mediates metastasis in a variety of cancers. It is highly conserved throughout the animal kingdom and mediates fundamental steps in RNA metabolism such as, controlling the stability of the MYC oncogene mRNA, the localisation of the ACTB mRNA in neuronal and other polarised cells such as fibroblasts, in addition to controlling the translation of the IGF2 mRNA.

Current findings suggest that the IMP1 protein mediates RNA recognition through its four C-terminal KH domains. *In vitro* studies have shown that the contribution of the four KH domains in recognition of a few select RNA targets is not equal. For example, the KH3 and KH4 domains have been shown to be vital for the binding of the ACTB mRNA, whereas the KH1 and KH2 domains play a role in mediating the complexes stability.

Previous work has been done to develop our understanding of IMP1s *in vivo* RNA target selection. However, reports form these findings are incoherent, with conflicting RNA recognition motifs being identified across the different studies.

In my thesis I have used a mutational approach based on structural information of IMP1-RNA complexes to introduce mutations into individual KH domains of the RNA binding protein. Our goal was to modify the RNA recognition properties of the domains. Enabling us to uncouple the RNA binding properties of the domain from the other regulatory functions the folded domain may be regulating in the cell.

By performing iCLIP on IMP1 KH domain RNA binding knock out mutants, I have identified an altered RNA binding pattern between the mutants towards the high affinity ACTB target. This is the first step in understanding how RNA binding proteins utilise their multiple RNA domains to recognise RNA targets *in vivo*. However, initial analysis of our data has proved challenging, and identifying real binding sites from background noise is difficult. Due to this study being the first of its kind, there are no analysis procedures aimed to perform the comparative analysis we are implementing. This is one example as to why IMP1 is a model system to begin this kind of study. As several high affinity RNA targets are known for the protein, I can use these examples to begin our analysis, as I have with the ACTB target. Other high affinity targets I will analyse are the MYC, CD44 and IFG2 mRNAs.

Moving further I have manipulated the RNA sequence specificity of the KH3 and KH4 domains of IMP1. It is known that the RNA sequence specificity of RNA binding domains is important for transcript recognition, and mutating RNA sequences perturbeds or inhibits RNA binding proteins from recognising their targets. This can lead to diseases such as cancer and neurological disorders. By mutating the KH3 and KH4 domains, we aimed to alter the RNA sequence specificity of these domains in order to investigate how this altered specificity would affect RNA target selection in HeLa cells by implementing our iCLIP method.

Unfortunately, changing the RNA specificity of the C-terminal KH domains proved challenging. As the amino acid residues which determine RNA specificity reside

in the highly structured hydrophobic groove, it was difficult to incorporate mutations that did not alter the fold or thermal stability of the protein. One of the more successful mutants was the KH3 S432R mutation. The KH3 domain of IMP1 naturally displays a less specific RNA binding sequence then other canonical KH domains. KH3 shows preference for an A in position 3 yet the energy penalty to bind a C in this position is only ~3-fold, compared to 50-fold for the KH3 domain of NOVA1. My S432R mutant was successful in shifting the RNA binding preference from ACAC to ACCC. Given the IMP1 KH3 domains reduced specificity compared to other KH domains, it will be insightful to investigate how changing the specificity of the domain affects in-cell RNA target selection. A common feature of RNA binding domains is their ability to associate with suboptimal RNA recognition sequences (as explained in the Introduction). iCLIP performed on a FLAG-IMP1 S432R mutant construct may reveal a suboptimal set of RNA targets that are biologically relevant.

In my thesis I have also investigated the RNA binding properties of the N-terminal RRM domains of both IMP1 and IMP3. Currently, the research literature on the IMP proteins identify the KH domains as the main site of RNA target recognition. Our investigation identified that the RRM1 domain of both IMP1 and IMP3 is able to bind RNA with a K_d in the µM to mM affinity range.

By investigating the RNA binding properties of both the IMP1 and IMP3 RRMs I was able to identify differences between the proteins. RRM1 of both IMP1 and IMP3 recognise RNA in a sequence specific manner, but with different specificity (IMP1 RRM1: CCCG and IMP3 RRM1: CCAA). In addition, the binding affinity of the two proteins is different, with IMP3 RRM1 having ~100-fold higher RNA binding affinity than IMP1. In light of these findings, it is likely that the RRM domains of IMP1 play a minor role in RNA target selection, compared to IMP3. Further investigation into these differences could

Currently IMP1 is the most extensively studies member of the IMP protein family. However, some recent research studies suggest that IMP3 is upregulated in more

cancers than IMP1. In addition, a recent iCLIP study performed on IMP3 identified binding sites that overlap with the seed sequences of a set of miRNAs, in particular the mir-9 seed sequence. This sequence contains the IMP3 RRM1 binding site CCAA which I identified in my studies. The mutant IMP3 RRM12 construct I have characterised (K36EY39A), which abolishes RNA binding, could potentially be used as a molecular tool to investigate the relationship between IMP3s ability to occupy RNA sites that are targeted for degradation by miRNAs, and how protection of these transcripts relates to cancer progression.

In my thesis I have studied the KH and RRM domains of IMP1 and IMP3. These RNA binding domains are highly abundant in mammalian RNA binding proteins. As discussed, there are several commonalities between RNA binding domains of the same class. Therefore, the approach we have implemented here can be used for other RNA binding proteins containing the same RNA binding domains, with minimal modification. This is more so true for the KH domain knock out mutation as this mutation has been previously tested in several other KH domains.

The major hurdle in this thesis and in this kind of study, is the tools available to analyse CLIP data sets in such a comparative way, while maintaining the real RNA targets and binding sites and discounting background noise. In developing a system where we can use predictions of what contribution the individual KH domains should play in recognising these targets (for example with the KH3 and KH4 domain binding to the ACTB 3'UTR), we can trial our analysis approach to develop a better pipeline. Refining of this method will enable us to dissect how IMP1 uses its KH domains to select novel IMP1 targets, with a future potential of applying this method to different RNA binding proteins, including IMP3.

Reference List

1. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters* **582,** 1977–1986 (2008).
2. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15,** 829–845 (2014).
3. Castello, A., Fischer, B., Hentze, M. W. & Preiss, T. RNA-binding proteins in Mendelian disease. *Trends in Genetics* **29,** 318–327 (2013).
4. Lukong, K. E., Chang, K., Khandjian, E. W. & Richard, S. RNA-binding proteins in human genetic disease. *Trends Genet.* **24,** 416–425 (2008).
5. Pereira, B., Billaud, M. & Almeida, R. RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends in Cancer* **3,** 506–528 (2017).
6. Nussbacher, J. K., Batra, R., Lagier-Tourenne, C. & Yeo, G. W. RNA-binding proteins in neurodegeneration: Seq and you shall receive. *Trends in Neurosciences* **38,** 226–236 (2015).
7. Dreyfuss, G., Kim, V. N. & Kataoka, N. Messenger-RNA-binding proteins and the messages they carry. *Nat. Rev. Mol. Cell Biol.* **3,** 195–205 (2002).
8. Alberts, B. *et al. Molecular Biology of the Cell 6e*. *Garland Science* **6,** (2014).
9. Mandel, C. R. *et al.* Polyadenylation factor CPSF-73 is the pre-mRNA 3′-end-processing endonuclease. *Nature* **444,** 953–956 (2006).
10. Bienroth, S., Wahle, E., Suter-Crazzolara, C. & Keller, W. Purification of the cleavage and polyadenylation factor involved in the 3’-processing of messenger RNA precursors. *J. Biol. Chem.* **266,** 19768–19776 (1991).
11. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* (2008). doi:10.1038/ng.259
12. Nishikura, K. Editor meets silencer: Crosstalk between RNA editing and RNA interference. *Nature Reviews Molecular Cell Biology* **7,** 919–931 (2006).

13. Valente, L. & Nishikura, K. ADAR Gene Family and A-to-I RNA Editing: Diverse Roles in Posttranscriptional Gene Regulation. *Progress in Nucleic Acid Research and Molecular Biology* **79,** 299–338 (2005).
14. Grüter, P. *et al.* TAP, the human homolog of Mex67p, mediates CTE-dependent RNA export from the nucleus. *Mol. Cell* **1,** 649–659 (1998).
15. Rodrigues, J. P. *et al.* REF proteins mediate the export of spliced and unspliced mRNAs from the nucleus. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 1030–1035 (2001).
16. Moore, M. J. From birth to death: The complex lives of eukaryotic mRNAs. *Science* **309,** 1514–1518 (2005).
17. Denti, M. a, Viero, G., Provenzani, A., Quattrone, A. & Macchi, P. mRNA fate: Life and death of the mRNA in the cytoplasm. *RNA Biol.* **10,** 37–41 (2013).
18. Hüttelmaier, S. *et al.* Spatial regulation of beta-actin translation by Src-dependent phosphorylation of ZBP1. *Nature* **438,** 512–5 (2005).
19. Saini, K. S., Summerhayes, I. C. & Thomas, P. Molecular events regulating messenger RNA stability in eukaryotes. *Mol. Cell. Biochem.* **96,** 15–23 (1990).
20. Knapinska, A., Irizarry-Barreto, P., Adusumalli, S., Androulakis, I. & Brewer, G. Molecular Mechanisms Regulating mRNA Stability: Physiological and Pathological Significance. *Curr. Genomics* **6,** 471–486 (2005).
21. Houseley, J. & Tollervey, D. The Many Pathways of RNA Degradation. *Cell* **136,** 763–776 (2009).
22. Yeo, G., Holste, D., Kreiman, G. & Burge, C. B. Variation in alternative splicing across human tissues. *Genome Biol.* **5,** R74 (2004).
23. Wirth, B. An update of the mutation spectrum of the survival motor neuron gene (SMN1) in autosomal recessive spinal muscular atrophy (SMA). *Human Mutation* **15,** 228–237 (2000).
24. Ule, J. *et al.* CLIP Identifies Nova-Regulated RNA Networks in the Brain. *Science (80-.).* **302,** 1212–1215 (2003).
25. Ferrari, R., Kapogiannis, D., D. Huey, E. & Momeni, P. FTD and ALS: A

Tale of Two Diseases. *Curr. Alzheimer Res.* (2011). doi:10.2174/156720511795563700

26. Scotter, E. L., Chen, H. J. & Shaw, C. E. TDP-43 Proteinopathy and ALS: Insights into Disease Mechanisms and Therapeutic Targets. *Neurotherapeutics* (2015). doi:10.1007/s13311-015-0338-x
27. Vance, C. *et al.* Mutations in FUS, an RNA Processing Protein, Cause Familial Amyotrophic Lateral Sclerosis Type 6. *Science (80-.).* (2009). doi:10.1126/science.1165942
28. Xiao, S. *et al.* RNA targets of TDP-43 identified by UV-CLIP are deregulated in ALS. *Mol. Cell. Neurosci.* **47,** 167–180 (2011).
29. Bhardwaj, A., Myers, M. P., Buratti, E. & Baralle, F. E. Characterizing TDP-43 interaction with its RNA targets. *Nucleic Acids Res.* **41,** 5062–5074 (2013).
30. Tollervey, J. R. *et al.* Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat. Neurosci.* **14,** 452–8 (2011).
31. Neumann, M. *et al.* Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science (80-.).* **314,** 130–133 (2006).
32. Geser, F., Martinez-Lage, M., Kwong, L. K., Lee, V. M. Y. & Trojanowski, J. Q. Amyotrophic lateral sclerosis, frontotemporal dementia and beyond: The TDP-43 diseases. *Journal of Neurology* **256,** 1205–1214 (2009).
33. Rogelj, B. *et al.* Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. *Sci. Rep.* **2,** (2012).
34. Huot, M. étienne *et al.* The Sam68 STAR RNA-Binding Protein Regulates mTOR Alternative Splicing during Adipogenesis. *Mol. Cell* **46,** 187–199 (2012).
35. Paronetto, M. P. *et al.* Alternative splicing of the cyclin D1 proto-oncogene is regulated by the RNA-binding protein Sam68. *Cancer Res.* **70,** 229–239 (2010).
36. Matter, N., Herrlich, P. & König, H. Signal-dependent regulation of splicing via phosphorylation of Sam68. *Nature* **420,** 691–695 (2002).

37. Paronetto, M. P., Achsel, T., Massiello, A., Chalfant, C. E. & Sette, C. The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x. *J. Cell Biol.* **176,** 929–939 (2007).
38. Lazaris-Karatzas, A., Montine, K. S. & Sonenberg, N. Malignant transformation by a eukaryotic initiation factor subunit that binds to mRNA 5′ cap. *Nature* (1990). doi:10.1038/345544a0
39. Furic, L. *et al.* eIF4E phosphorylation promotes tumorigenesis and is associated with prostate cancer progression. *Proc. Natl. Acad. Sci. U. S. A.* (2010). doi:10.1073/pnas.1005320107
40. Wendel, H. G. *et al.* Survival signalling by Akt and eIF4E in oncogenesis and cancer therapy. *Nature* (2004). doi:10.1038/nature02369
41. Leppek, K., Das, R. & Barna, M. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nature Reviews Molecular Cell Biology* (2018). doi:10.1038/nrm.2017.103
42. Brennan, C. M. & Steitz, J. a. HuR and mRNA stability. *Cell. Mol. Life Sci.* **58,** 266–277 (2001).
43. Guo, X. & Hartley, R. S. HuR contributes to cyclin E1 deregulation in MCF-7 breast cancer cells. *Cancer Res.* **66,** 7948–7956 (2006).
44. Wang, W., Caldwell, M. C., Lin, S., Furneaux, H. & Gorospe, M. HuR regulates cyclin A and cyclin B1 mRNA stability during cell proliferation. *EMBO J.* **19,** 2340–2350 (2000).
45. Wang, J., Wang, B., Bi, J. & Zhang, C. Cytoplasmic HuR expression correlates with angiogenesis, lymphangiogenesis, and poor outcome in lung cancer. *Med. Oncol.* **28,** (2011).
46. Sheflin, L. G., Zou, A. P. & Spaulding, S. W. Androgens regulate the binding of endogenous HuR to the AU-rich 3′UTRs of HIF-1α and EGF mRNA. *Biochem. Biophys. Res. Commun.* **322,** 644–651 (2004).
47. Akool, E.-S. *et al.* Nitric oxide increases the decay of matrix metalloproteinase 9 mRNA by inhibiting the expression of mRNA-stabilizing factor HuR. *Mol. Cell. Biol.* **23,** 4901–4916 (2003).
48. Bolognani, F. & Perrone-Bizzozero, N. I. RNA-protein interactions and control of mRNA stability in neurons. *Journal of Neuroscience Research*

86, 481–489 (2008).

49. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8,** 479–490 (2007).
50. Zamore, P. D., Williamson, J. R. & Lehmann, R. The Pumilio protein binds RNA through a conserved domain that defines a new class of RNA-binding proteins. *RNA* **3,** 1421–1433 (1997).
51. Wang, X., McLachlan, J., Zamore, P. D. & Hall, T. M. T. Modular recognition of RNA by a human Pumilio-homology domain. *Cell* **110,** 501–512 (2002).
52. Zamore, P. D., Bartel, D. P., Lehmann, R. & Williamson, J. R. The PUMILIO-RNA interaction: A single RNA-binding domain monomer recognizes a bipartite target sequence. *Biochemistry* **38,** 596–604 (1999).
53. Maris, C., Dominguez, C. & Allain, F. H. T. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS Journal* **272,** 2118–2131 (2005).
54. Jacks, A. *et al.* Structure of the C-terminal domain of human La protein reveals a novel RNA recognition motif coupled to a helical nuclear retention element. *Structure* **11,** 833–843 (2003).
55. Netter, C., Weber, G., Benecke, H. & Wahl, M. C. Functional stabilization of an RNA recognition motif by a noncanonical N-terminal expansion. *RNA* **15,** 1305–13 (2009).
56. Qu, X. *et al.* The C-terminal domains of vertebrate CstF-64 and its yeast orthologue Rna15 form a new structure critical for mRNA 3′-end processing. *J. Biol. Chem.* **282,** 2101–2115 (2007).
57. Cléry, A., Blatter, M. & Allain, F. H. T. RNA recognition motifs: boring? Not quite. *Current Opinion in Structural Biology* **18,** 290–298 (2008).
58. Daubner, G. M., Cléry, A. & Allain, F. H. T. RRM-RNA recognition: NMR or crystallography...and new findings. *Current Opinion in Structural Biology* **23,** 100–108 (2013).
59. Martin-Tumasz, S., Richie, A. C., Clos, L. J., Brow, D. A. & Butcher, S. E. A novel occluded RNA recognition motif in Prp24 unwinds the U6 RNA internal stem loop. *Nucleic Acids Res.* **39,** 7837–7847 (2011).

60. MacKereth, C. D. *et al.* Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. *Nature* **475,** 408–413 (2011).
61. Clery, A. *et al.* Isolated pseudo-RNA-recognition motifs of SR proteins can regulate splicing using a noncanonical mode of RNA recognition. *Proc. Natl. Acad. Sci.* **110,** E2802–E2811 (2013).
62. Ngo, J. C. K. *et al.* A Sliding Docking Interaction Is Essential for Sequential and Processive Phosphorylation of an SR Protein by SRPK1. *Mol. Cell* (2008). doi:10.1016/j.molcel.2007.12.017
63. Chiodi, I. *et al.* RNA recognition motif 2 directs the recruitment of SF2/ASF to nuclear stress bodies. *Nucleic Acids Res.* (2004). doi:10.1093/nar/gkh759
64. Hardin, J. W., Hu, Y. X. & McKay, D. B. Structure of the RNA binding domain of a DEAD-box helicase bound to its ribosomal RNA target reveals a novel mode of recognition by an RNA recognition motif. *J. Mol. Biol.* (2010). doi:10.1016/j.jmb.2010.07.040
65. Dominguez, C. & Allain, F. H. T. NMR structure of the three quasi RNA recognition motifs (qRRMs) of human hnRNP F and interaction studies with Bcl-x G-tract RNA: A novel mode of RNA recognition. *Nucleic Acids Res.* **34,** 3634–3645 (2006).
66. Dominguez, C., Fisette, J. F., Chabot, B. & Allain, F. H. T. Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs. *Nat. Struct. Mol. Biol.* **17,** 853–861 (2010).
67. Cukier, C. D. *et al.* Molecular basis of FIR-mediated c-myc transcriptional control. *Nat. Struct. Mol. Biol.* **17,** 1058–1064 (2010).
68. Nicastro, G., Taylor, I. A. & Ramos, A. ScienceDirect KH – RNA interactions : back in the groove. *Curr. Opin. Struct. Biol.* **30,** 63–70 (2015).
69. Mir-Montazeri, B., Ammelburg, M., Forouzan, D., Lupas, A. N. & Hartmann, M. D. Crystal structure of a dimeric archaeal Cleavage and Polyadenylation Specificity Factor. *J. Struct. Biol.* **173,** 191–195 (2011).
70. Silva, A. P. G. *et al.* Structure and activity of a novel archaeal β-CASP protein with N-terminal KH domains. *Structure* **19,** 622–632 (2011).

71. Nicastro, G., Taylor, I. A. & Ramos, A. KH-RNA interactions: Back in the groove. *Current Opinion in Structural Biology* **30,** 63–70 (2015).
72. Oddone, A. *et al.* Structural and biochemical characterization of the yeast exosome component Rrp40. *EMBO Rep.* **8,** 63–69 (2007).
73. Valverde, R., Edwards, L. & Regan, L. Structure and function of KH domains. *FEBS Journal* **275,** 2712–2726 (2008).
74. Vernet, C. & Artzt, K. STAR, a gene family involved in signal transduction and activation of RNA. *Trends in Genetics* (1997). doi:10.1016/S0168-9525(97)01269-9
75. Beuck, C., Qu, S., Fagg, W. S., Ares, M. & Williamson, J. R. Structural analysis of the quaking homodimerization interface. *J. Mol. Biol.* **423,** 766–781 (2012).
76. Feracci, M. *et al.* Structural basis of RNA recognition and dimerization by the STAR proteins T-STAR and Sam68. *Nat. Commun.* **7,** (2016).
77. Feracci, M., Foot, J. & Dominguez, C. Structural investigations of the RNA-binding properties of STAR proteins. *Biochem. Soc. Trans.* **42,** 1141–1146 (2014).
78. Liu, Z. *et al.* Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science* **294,** 1098–1102 (2001).
79. Teplova, M. *et al.* Structure-function studies of STAR family quaking proteins bound to their in vivo RNA target sites. *Genes Dev.* **27,** 928–940 (2013).
80. Foot, J. N., Feracci, M. & Dominguez, C. Screening protein - Single stranded RNA complexes by NMR spectroscopy for structure determination. *Methods* **65,** 288–301 (2014).
81. Zhang, C., Lee, K. Y., Swanson, M. S. & Darnell, R. B. Prediction of clustered RNA-binding protein motif sites in the mammalian genome. *Nucleic Acids Res.* **41,** 6793–6807 (2013).
82. Hollingworth, D. *et al.* KH domains with impaired nucleic acid binding as a tool for functional analysis. *Nucleic Acids Res.* **40,** 6873–86 (2012).
83. Wächter, K., Köhn, M., Stöhr, N. & Hüttelmaier, S. Subcellular localization and RNP formation of IGF2BPs (IGF2 mRNA-binding proteins) is

modulated by distinct RNA-binding domains. *Biol. Chem.* **394,** 1077–90 (2013).

84. Teplova, M. *et al.* Protein-RNA and protein-protein recognition by dual KH1/2 domains of the neuronal splicing factor Nova-1. *Structure* **19,** 930–944 (2011).
85. Chao, J. a *et al.* ZBP1 recognition of beta-actin zipcode induces RNA looping. *Genes Dev.* **24,** 148–58 (2010).
86. Díaz-Moreno, I. *et al.* Orientation of the central domains of KSRP and its implications for the interaction with the RNA targets. *Nucleic Acids Res.* **38,** 5193–5205 (2010).
87. Beuth, B., Pennell, S., Arnvig, K. B., Martin, S. R. & Taylor, I. A. Structure of a Mycobacterium tuberculosis NusA-RNA complex. *EMBO J.* **24,** 3576–3587 (2005).
88. Stefl, R., Xu, M., Skrisovska, L., Emeson, R. B. & Allain, F. H. T. Structure and specific RNA binding of ADAR2 double-stranded RNA binding motifs. *Structure* **14,** 345–355 (2006).
89. Shamoo, Y. *et al.* Both RNA-Binding Domains in Heterogenous Nuclear Ribonucleoprotein A1 Contribute toward Single-Stranded-RNA Binding. *Biochemistry* **33,** 8272–8281 (1994).
90. Pérez-Cãadillas, J. M. Grabbing the message: Structural basis of mRNA 3′UTR recognition by Hrp1. *EMBO J.* **25,** 3167–3178 (2006).
91. Handa, N. *et al.* Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein. *Nature* **398,** 579–585 (1999).
92. Briata, P. *et al.* KSRP, many functions for a single protein. *Front. Biosci. (Landmark Ed.))* **16,** 1787–1796 (2011).
93. Briata, P., Chen, C.-Y., Ramos, A. & Gherzi, R. Functional and molecular insights into KSRP function in mRNA decay. *Biochim. Biophys. Acta* **1829,** 689–94 (2013).
94. Collins, K. M., Oregioni, A., Robertson, L. E., Kelly, G. & Ramos, A. Protein-RNA specificity by high-throughput principal component analysis of NMR spectra. *Nucleic Acids Res.* **43,** e41–e41 (2015).
95. Beuth, B., Garcia-Mayoral, M. F., Taylor, I. A. & Ramos, A. Scaffold-

independent analysis of RNA-protein interactions: The Nova-1 KH3-RNA complex. *J. Am. Chem. Soc.* **129,** 10205–10210 (2007).

96. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346,** 818–822 (1990).
97. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment:RNA ligands to bacteriophage T4 DNA polymerase. *Science (80-.).* **249,** 505–510 (1990).
98. Cook, K. B., Hughes, T. R. & Morris, Q. D. High-throughput characterization of protein-RNA interactions. *Brief. Funct. Genomics* **14,** 74–89 (2015).
99. Campbell, Z. T. *et al.* Cooperativity in RNA-Protein Interactions: Global Analysis of RNA Binding Specificity. *Cell Rep.* **1,** 570–581 (2012).
100. Ray, D. *et al.* Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.* **27,** 667–670 (2009).
101. Lambert, N. *et al.* RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Mol. Cell* **54,** 887–900 (2014).
102. Carey, J., Uhlenbeck, O. C., Cameron, V. & de Haseth, P. L. Sequence-Specific Interaction of R17 Coat Protein with Its Ribonucleic Acid Binding Site. *Biochemistry* **22,** 2601–2610 (1983).
103. Peabody, D. S. The RNA binding site of bacteriophage MS2 coat protein. *EMBO J.* (1993).
104. Castello, A. *et al.* Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell* **149,** 1393–1406 (2012).
105. Baltz, A. G. *et al.* The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Mol. Cell* **46,** 674–690 (2012).
106. Tenenbaum, S. A., Carson, C. C., Lager, P. J. & Keene, J. D. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc. Natl. Acad. Sci.* **97,** 14085–14090 (2000).
107. König, J., Zarnack, K., Luscombe, N. M. & Ule, J. Protein–RNA interactions: new genomic technologies and perspectives. *Nat. Rev.*

Genet. **13,** 77–83 (2012).

108. Ule, J., Jensen, K., Mele, A. & Darnell, R. B. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* **37,** 376–86 (2005).
109. Granneman, S., Kudla, G., Petfalski, E. & Tollervey, D. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc. Natl. Acad. Sci.* **106,** 9613–9618 (2009).
110. Hafner, M. *et al.* Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell* **141,** 129–141 (2010).
111. König, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **17,** 909–15 (2010).
112. Urlaub, H., Hartmuth, K. & Lührmann, R. A two-tracked approach to analyze RNA-protein crosslinking sites in native, nonlabeled small nuclear ribonucleoprotein particles. *Methods* **26,** 170–181 (2002).
113. Sugimoto, Y. *et al.* Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol.* **13,** R67 (2012).
114. Kishore, S. *et al.* A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods* **8,** 559–64 (2011).
115. Zhang, C. & Darnell, R. B. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat. Biotechnol.* **29,** 607–614 (2011).
116. Huppertz, I. *et al.* iCLIP: Protein-RNA interactions at nucleotide resolution. *Methods* **65,** 274–287 (2014).
117. Gherzi, R. *et al.* A KH Domain RNA Binding Protein, KSRP, Promotes ARE-Directed mRNA Turnover by Recruiting the Degradation Machinery. *Mol Cell* **14,** 571–83 (2004).
118. Léveillé, N. *et al.* Selective inhibition of microRNA accessibility by RBM38 is required for p53 activity. *Nat. Commun.* **2,** (2011).
119. Leeper, T. C., Qu, X., Lu, C., Moore, C. & Varani, G. Novel protein-protein contacts facilitate mRNA 3'-processing signal recognition by Rna15 and

Hrp1. *J. Mol. Biol.* **401,** 334–349 (2010).

120. Pancevac, C., Goldstone, D. C., Ramos, A. & Taylor, I. A. Structure of the Rna15 RRM-RNA complex reveals the molecular basis of GU specificity in transcriptional 3'-end processing factors. *Nucleic Acids Res.* **38,** 3119–3132 (2010).
121. Blake, J. A. *et al.* Gene ontology consortium: Going forward. *Nucleic Acids Res.* **43,** D1049–D1056 (2015).
122. Shi, Y. A glimpse of structural biology through X-ray crystallography. *Cell* **159,** 995–1014 (2014).
123. Ke, A. & Doudna, J. A. Crystallization of RNA and RNA-protein complexes. *Methods* **34,** 408–414 (2004).
124. Price, S. R., Evans, P. R. & Nagai, K. Crystal structure of the spliceosomal U2B'-U2A' protein complex bound to a fragment of U2 small nuclear RNA. *Nature* **394,** 645–650 (1998).
125. Calero, G. *et al.* Structural basis of m(7)GpppG binding to the nuclear cap- binding protein complex. *Nat.Struct.Biol.* **9,** 912–917 (2002).
126. Cukier, C. D. & Ramos, A. Modular protein-RNA interactions regulating mRNA metabolism: a role for NMR. *Eur. Biophys. J.* **40,** 1317–25 (2011).
127. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11,** 63–65 (2014).
128. Callaway, E. The revolution will not be crystallized: A new method sweeps through structural biology. *Nature* **525,** 172–174 (2015).
129. Wisdom, R. & Lee, W. The protein-coding region of c-myc mRNA contains a sequence that specifies rapid mRNA turnover and induction by protein synthesis inhibitors. *Genes Dev.* **5,** 232–243 (1991).
130. Bernstein, P. L., Herrick, D. J., Prokipcak, R. D. & Ross, J. Control of c-myc mRNA half-life in vitro by a protein capable of binding to a coding region stability determinant. *Genes Dev.* **6,** 642–654 (1992).
131. Lawrence, J. B. & Singer, R. H. Intracellular localization of messenger RNAs for cytoskeletal proteins. *Cell* **45,** 407–15 (1986).
132. Zhang, H. L., Singer, R. H. & Bassell, G. J. Neurotrophin regulation of beta-actin mRNA and protein localization within growth cones. *J. Cell*

Biol. **147,** 59–70 (1999).

133. Yaffe, D., Nudel, U., Mayer, Y. & Neuman, S. Highly conserved sequences in the 3' untranslated region of mRNAs coding for homologous proteins in distantly related species. *Nucleic Acids Res.* **13,** 3723–37 (1985).
134. Ross, A. F., Oleynikov, Y., Kislauskis, E. H., Taneja, K. L. & Singer, R. H. Characterization of a beta-actin mRNA zipcode-binding protein. *Mol. Cell. Biol.* **17,** 2158–2165 (1997).
135. Nielsen, J. *et al.* A family of insulin-like growth factor II mRNA-binding proteins represses translation in late development. *Mol. Cell. Biol.* **19,** 1262–1270 (1999).
136. Yisraeli, J. K. VICKZ proteins: a multi-talented family of regulatory RNA-binding proteins. *Biol. Cell* **97,** 87–96 (2005).
137. Runge, S. *et al.* H19 RNA binds four molecules of insulin-like growth factor II mRNA-binding protein. *J. Biol. Chem.* **275,** 29562–29569 (2000).
138. Schwartz, S. P., Aisenthal, L., Elisha, Z., Oberman, F. & Yisraeli, J. K. A 69-kDa RNA-binding protein from Xenopus oocytes recognizes a common motif in two vegetally localized maternal mRNAs. *Proc. Natl. Acad. Sci.* **89,** 11895–11899 (1992).
139. Adolph, S. K., DeLotto, R., Nielsen, F. C. & Christiansen, J. Embryonic expression of Drosophila IMP in the developing CNS and PNS. *Gene Expr. Patterns* **9,** 138–143 (2009).
140. Boylan, K. L. M. *et al.* Motility screen identifies Drosophila IGF-II mRNA-binding protein - Zipcode-binding protein acting in oogenesis and synaptogenesis. *PLoS Genet.* **4,** (2008).
141. Bell, J. L. *et al.* Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): Post-transcriptional drivers of cancer progression? *Cell. Mol. Life Sci.* **70,** 2657–2675 (2013).
142. Noubissi, F. K., Nikiforov, M. A., Colburn, N. & Spiegelman, V. S. Transcriptional Regulation of CRD-BP by c-myc: Implications for c-myc Functions. *Genes Cancer* **1,** 1074–1082 (2010).
143. Hamilton, K. E. *et al.* IMP1 promotes tumor growth, dissemination and a

tumor-initiating cell phenotype in colorectal cancer cell xenografts. *Carcinogenesis* **34,** 2647–2654 (2013).

144. Conway, A. E. *et al.* Enhanced CLIP Uncovers IMP Protein-RNA Targets in Human Pluripotent Stem Cells Important for Cell Adhesion and Survival. *Cell Rep.* **15,** 666–679 (2016).
145. Vikesaa, J. *et al.* RNA-binding IMPs promote cell adhesion and invadopodia formation. *EMBO J.* **25,** 1456–1468 (2006).
146. Atlas, R., Behar, L., Elliott, E. & Ginzburg, I. The insulin-like growth factor mRNA binding-protein IMP-1 and the Ras-regulatory protein G3BP associate with tau mRNA and HuD protein in differentiated P19 neuronal cells. *J. Neurochem.* (2004). doi:10.1111/j.1471-4159.2004.02371.x
147. Gu, W. *et al.* Regulation of local expression of cell adhesion and motility-related mRNAs in breast cancer cells by IMP1/ZBP1. *J. Cell Sci.* **125,** 81–91 (2012).
148. Polesskaya, A. *et al.* Lin-28 binds IGF-2 mRNA and participates in skeletal myogenesis by increasing translation efficiency. *Genes Dev.* **21,** 1125–1138 (2007).
149. Hansen, T. V. O. *et al.* Dwarfism and impaired gut development in insulin-like growth factor II mRNA-binding protein 1-deficient mice. *Mol. Cell. Biol.* **24,** 4448–4464 (2004).
150. Mueller-Pillasch, F. *et al.* Expression of the highly conserved RNA binding protein KOC in embryogenesis. *Mech. Dev.* **88,** 95–99 (1999).
151. Christiansen, J., Kolte, A. M., Hansen, T. V. O. & Nielsen, F. C. IGF2 mRNA-binding protein 2: Biological function and putative role in type 2 diabetes. *J. Mol. Endocrinol.* **43,** 187–195 (2009).
152. Zhang, Q. *et al.* Vg1 RBP intracellular distribution and evolutionarily conserved expression at multiple stages during development. *Mech. Dev.* **88,** 101–106 (1999).
153. NIELSEN, J. *et al.* Nuclear transit of human zipcode-binding protein IMP1. *Biochem. J.* **376,** 383–391 (2003).
154. Nielsen, J., Kristensen, M. A., Willemoës, M., Nielsen, F. C. & Christiansen, J. Sequential dimerization of human zipcode-binding protein

IMP1 on RNA: A cooperative mechanism providing RNP stability. *Nucleic Acids Res.* **32,** 4368–4376 (2004).

155. Song, T. *et al.* Specific interaction of KIF11 with ZBP1 regulates the transport of -actin mRNA and cell motility. *J. Cell Sci.* **128,** 1001–1010 (2015).
156. Jønson, L. *et al.* Molecular composition of IMP1 ribonucleoprotein granules. *Mol. Cell. Proteomics* **6,** 798–811 (2007).
157. Weidensdorfer, D. *et al.* Control of c-myc mRNA stability by IGF2BP1-associated cytoplasmic RNPs. *RNA* **15,** 104–115 (2008).
158. Oleynikov, Y. & Singer, R. H. Real-time visualization of ZBP1 association with β-actin mRNA during transcription and localization. *Curr. Biol.* **13,** 199–207 (2003).
159. Wu, B., Buxbaum, A. R., Katz, Z. B., Yoon, Y. J. & Singer, R. H. Quantifying Protein-mRNA Interactions in Single Live Cells. *Cell* **162,** 211–220 (2015).
160. Farina, K. L. *et al.* Two ZBP1 KH domains facilitate β-actin mRNA localization, granule formation, and cytoskeletal attachment. *J. Cell Biol.* **160,** 77–87 (2003).
161. Dai, N. *et al.* mTOR phosphorylates IMP2 to promote IGF2 mRNA translation by internal ribosomal entry. *Genes Dev.* **25,** 1159–1172 (2011).
162. Dai, N., Christiansen, J., Nielsen, F. C. & Avruch, J. mTOR complex 2 phosphorylates IMP1 cotranslationally to promote IGF2 production and the proliferation of mouse embryonic fibroblasts. *Genes Dev.* **27,** 301–312 (2013).
163. Ross, J., Lemm, I. & Berberet, B. Overexpression of an mRNA-binding protein in human colorectal cancer. *Oncogene* **20,** 6544–6550 (2001).
164. Hammer, N. A. *et al.* Expression of IGF-II mRNA-binding proteins (IMPs) in gonads and testicular cancer. *Reproduction* **130,** 203–212 (2005).
165. Dimitriadis, E. *et al.* Expression of oncofetal RNA-binding protein CRD-BP/IMP1 predicts clinical outcome in colon cancer. *Int. J. Cancer* **121,** 486–494 (2007).

166. Gu, L., Shigemasa, K. & Ohama, K. Increased expression of IGF II mRNA-binding protein 1 mRNA is associated with an advanced clinical stage and poor prognosis in patients with ovarian cancer. *Int J Oncol* **24,** 671–678 (2004).
167. Lederer, M., Bley, N., Schleifer, C. & Hüttelmaier, S. The role of the oncofetal IGF2 mRNA-binding protein 3 (IGF2BP3) in cancer. *Seminars in Cancer Biology* **29,** 3–12 (2014).
168. Ennajdaoui, H. *et al.* IGF2BP3 Modulates the Interaction of Invasion-Associated Transcripts with RISC. *Cell Rep.* **15,** 1876–1883 (2016).
169. Köbel, M. *et al.* Expression of the RNA-binding protein IMP1 correlates with poor prognosis in ovarian carcinoma. *Oncogene* **26,** 7584–7589 (2007).
170. Peng, Y. & Croce, C. M. The role of MicroRNAs in human cancer. *Signal Transduct. Target. Ther.* (2016). doi:10.1038/sigtrans.2015.4
171. Boyerinas, B. *et al.* Identification of let-7-regulated oncofetal genes. *Cancer Res.* **68,** 2587–2591 (2008).
172. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318,** 1917–20 (2007).
173. Noubissi, F. K. *et al.* CRD-BP mediates stabilization of βTrCP1 and c-myc mRNA in response to β-catenin signalling. *Nature* **441,** 898–901 (2006).
174. Nishino, J., Kim, S., Zhu, Y., Zhu, H. & Morrison, S. J. A network of heterochronic genes including Imp1 regulates temporal changes in stem cell properties. *Elife* **2013,** 1–30 (2013).
175. Stöhr, N. & Hüttelmaier, S. IGF2BP1: a post-transcriptional 'driver' of tumor cell migration. *Cell Adh. Migr.* **6,** 312–8 (2012).
176. Stöhr, N. *et al.* IGF2BP1 promotes cell migration by regulating MK5 and PTEN signaling. *Genes Dev.* **26,** 176–189 (2012).
177. Coulis, C. M., Lee, C., Nardone, V. & Prokipcak, R. D. Inhibition of c-myc expression in cells by targeting an RNA-protein interaction using antisense oligonucleotides. *Mol Pharmacol* **57,** 485–94. (2000).
178. King, D. T., Barnes, M., Thomsen, D. & Lee, C. H. Assessing specific oligonucleotides and small molecule antibiotics for the ability to inhibit the

CRD-BP-CD44 RNA interaction. *PLoS One* **9,** (2014).

179. Patel, V. L. *et al.* Spatial arrangement of an RNA zipcode identifies mRNAs under post-transcriptional control. *Genes Dev.* **26,** 43–53 (2012).
180. Nicastro, G. *et al.* Mechanism of β-actin mRNA Recognition by ZBP1. *Cell Rep.* **18,** 1187–1199 (2017).
181. Rule, G. S. & Hitchens, T. K. *Fundamentals of Protein NMR Spectroscopy*. *Focus on structural biology* **5,** (2006).
182. Bloembergen, N., Purcell, E. M. & Pound, R. V. Relaxation effects in nuclear magnetic resonance absorption. *Phys. Rev.* **73,** 679–712 (1948).
183. Hansen, H. T. *et al.* Drosophila Imp iCLIP identifies an RNA assemblage coordinating F-actin formation. *Genome Biol.* **16,** 123 (2015).
184. Ling, S.-C. *et al.* ALS-associated mutations in TDP-43 increase its stability and promote TDP-43 complexes with FUS/TLS. *Proc. Natl. Acad. Sci.* **107,** 13318–13323 (2010).
185. Krenn, V., Wehenkel, A., Li, X., Santaguida, S. & Musacchio, A. Structural analysis reveals features of the spindle checkpoint kinase Bub1-kinetochore subunit Knl1 interaction. *J. Cell Biol.* **196,** 451–67 (2012).
186. Gassmann, R. *et al.* Removal of Spindly from microtubule-attached kinetochores controls spindle checkpoint silencing in human cells. *Genes Dev.* **24,** 957–971 (2010).
187. Kato, T. *et al.* Increased expression of insulin-like growth factor-II messenger RNA-binding protein 1 is associated with tumor progression in patients with lung cancer. *Clin. Cancer Res.* **13,** 434–442 (2007).
188. Wang, Z. *et al.* iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol.* **8,** e1000530 (2010).
189. Konig, J. *et al.* iCLIP--transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. *J. Vis. Exp.* 1–7 (2011). doi:10.3791/2638
190. Zarnack, K. *et al.* Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152,** 453–66 (2013).
191. Ascano, M. *et al.* FMRP targets distinct mRNA sequence elements to

regulate protein expression. *Nature* **492,** 382–6 (2012).

192. Ramos, A. *et al.* Role of dimerization in KH/RNA complexes: The example of Nova KH3. *Biochemistry* **41,** 4193–4201 (2002).
193. Nielsen, F. C., Nielsen, J., Kristensen, M. a, Koch, G. & Christiansen, J. Cytoplasmic trafficking of IGF-II mRNA-binding protein by conserved KH domains. *J. Cell Sci.* **115,** 2087–2097 (2002).
194. Haberman, N. *et al.* Insights into the design and interpretation of iCLIP experiments. *Genome Biol.* **18,** 7 (2017).
195. Hauer, C. *et al.* Improved binding site assignment by high-resolution mapping of RNA-protein interactions using iCLIP. *Nat. Commun.* **6,** 7921 (2015).
196. Gu, W., Wells, A. L., Pan, F. & Singer, R. H. Feedback regulation between zipcode binding protein 1 and beta-catenin mRNAs in breast cancer cells. *Mol. Cell. Biol.* **28,** 4963–4974 (2008).
197. Mongroo, P. S. *et al.* IMP-1 displays cross-talk with K-Ras and modulates colon cancer cell survival through the novel proapoptotic protein CYFIP2. *Cancer Res.* **71,** 2172–2182 (2011).
198. Maticzka, D., Lange, S. J., Costa, F. & Backofen, R. GraphProt: Modeling binding preferences of RNA-binding proteins. *Genome Biol.* **15,** (2014).
199. Chen, B., Yun, J., Kim, M. S., Mendell, J. T. & Xie, Y. PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. *Genome Biol.* **15,** R18 (2014).
200. Wang, T., Xie, Y. & Xiao, G. dCLIP: a computational approach for comparative CLIP-seq analyses. *Genome Biol.* **15,** R11 (2014).
201. Valverde, R., Pozdnyakova, I., Kajander, T., Venkatraman, J. & Regan, L. Fragile X Mental Retardation Syndrome: Structure of the KH1-KH2 Domains of Fragile X Mental Retardation Protein. *Structure* **15,** 1090–1098 (2007).
202. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci.* **101,** 6062–6067 (2004).
203. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).

204. Sanchez, M., Galy, B., Muckenthaler, M. U. & Hentze, M. W. Iron-regulatory proteins limit hypoxia-inducible factor-2α expression in iron deficiency. *Nat. Struct. Mol. Biol.* **14,** 420–426 (2007).
205. Katz, Z. B. *et al.* β-actin mRNA compartmentalization enhances focal adhesion stability and directs cell migration. *Genes Dev.* **26,** 1885–1890 (2012).
206. Park, H. Y., Trcek, T., Wells, A. L., Chao, J. A. & Singer, R. H. An Unbiased Analysis Method to Quantify mRNA Localization Reveals Its Correlation with Cell Motility. *Cell Rep.* **1,** 179–184 (2012).
207. Nicastro, G. *et al.* Noncanonical G recognition mediates KSRP regulation of let-7 biogenesis. *Nat. Struct. Mol. Biol.* **19,** 1282–1286 (2012).
208. Lewis, H. A. *et al.* Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. *Cell* **100,** 323–32 (2000).
209. Backe, P. H., Messias, A. C., Ravelli, R. B. G., Sattler, M. & Cusack, S. X-ray crystallographic and NMR studies of the third KH domain of hnRNP K in complex with single-stranded nucleic acids. *Structure* **13,** 1055–67 (2005).
210. Braddock, D. T., Louis, J. M., Baber, J. L., Levens, D. & Clore, G. M. Structure and dynamics of KH domains from FBP bound to single-stranded DNA. *Nature* **415,** 1051–1056 (2002).
211. Fenn, S. *et al.* Crystal structure of the third KH domain of human poly(C)-binding protein-2 in complex with a C-rich strand of human telomeric DNA at 1.6Å resolution. *Nucleic Acids Res.* **35,** 2651–2660 (2007).
212. Trabucchi, M. *et al.* The RNA-binding protein KSRP promotes the biogenesis of a subset of microRNAs. *Nature* **459,** 1010–1014 (2009).
213. García-Mayoral, M. F. *et al.* The sequence selectivity of KSRP explains its flexibility in the recognition of the RNA targets. *Nucleic Acids Res.* **36,** 5290–5296 (2008).
214. Rigoutsos, I., Riek, P., Graham, R. M. & Novotny, J. Structural details (kinks and non-α conformations) in transmembrane helices are intrahelically determined and can be predicted by sequence pattern

descriptors. *Nucleic Acids Res.* **31,** 4625–4631 (2003).

215. Skrisovska, L. *et al.* The testis-specific human protein RBMY recognizes RNA through a novel mode of interaction. *EMBO Rep.* **8,** 372–379 (2007).
216. Volpon, L., D'Orso, I., Young, C. R., Frasch, A. C. & Gehring, K. NMR structural study of TcUBP1, a single RRM domain protein from Trypanosoma cruzi: contribution of a beta hairpin to RNA binding. *Biochemistry* **44,** 3708–17 (2005).
217. Tintaru, A. M. *et al.* Structural and functional analysis of RNA and TAP binding to SF2/ASF. *EMBO Rep.* **8,** 756–762 (2007).
218. Auweter, S. D. *et al.* Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. *EMBO J.* **25,** 163–173 (2006).
219. Golovanov, A. P., Hautbergue, G. M., Tintaru, A. M., Lian, L.-Y. & Wilson, S. a. The solution structure of REF2-I reveals interdomain interactions and regions involved in binding mRNA export factors and RNA. *RNA* **12,** 1933–1948 (2006).
220. Oberstrass, F. C. *et al.* Structural biology - Structure of PTB bound to RNA: Specific binding and implications for splicing regulation. *Science (80-.).* **309,** 2054–2057 (2005).
221. Crichlow, G. V. *et al.* Dimerization of FIR upon FUSE DNA binding suggests a mechanism of c-myc inhibition. *EMBO J.* **27,** 277–289 (2008).
222. Bae, E. *et al.* Structure and Interactions of the First Three RNA Recognition Motifs of Splicing Factor Prp24. *J. Mol. Biol.* **367,** 1447–1458 (2007).
223. Wang, X. & Hall, T. M. Structural basis for recognition of AU-rich element RNA by the HuD protein. *Nat. Struct. Biol.* **8,** 141–145 (2001).
224. Vitali, F. *et al.* Structure of the two most C-terminal RNA recognition motifs of PTB using segmental isotope labeling. *EMBO J.* **25,** 150–162 (2006).

Appendix

Appendix I: List of primers used for cloning and mutagenesis

Primers for IMP1 pcDNA5-TO-FRT vector cloning

Name	Sequence (5'-3')	Restriction Site
IMP1_N-Term_Flag_RV	ATCACTCGAGTCACTTCCTCCGTGC	XhoI
IMP1_C-Term_Flag_RV	ATCACTCGAGCTTCCTCCGTGCCTG	XhoI
IMP1_FLAG_FW	ATCAGGATCCATGAACAAGCTTTACATC	BamHI

Primers for GxxG mutagenesis of FLAG_IMP1

Name	Sequence (5'-3')
IMP1_KH1DD_ K213D/E214D_FW	GGGTGCCATTATTGGCGATGATGGGGCCACCATCCGC
IMP1_KH1DD_ K213D/E214D_RV	GCGGATGGTGGCCCCATCATCGCCAATAATGGCACCC
IMP1_KH2DD_ K294D/E295G_FW	GTAGGGCGTCTCATTGGCGATGATGGACGGAACCTGAAGAAG
IMP1_KH2DD_ K294D/E295G_RV	CTTCTTCAGGTTCCGTCCATCATCGCCAATGAGACGCCCTAC
IMP1_KH3DD_ K423D/K424D_FW	GTGGGCGCCATCATCGGCGATGATGGGCAGCACATCAAACAG
IMP1_KH3DD_ K423D/K424D_RV	CTGTTTGATGTGCTGCCCATCATCGCCGATGATGGCGCCCAC
IMP1_KH4DD_ K505D/G506D_FW	CTGGCCGGGTCATTGGCGATGATGGAAAAACGGTGAACG
IMP1_KH4DD_ K505D/G506D_RV	CGTTCACCGTTTTTCCATCATCGCCAATGACCCGGCCAG

Primers for mutagenesis of ZBP1 KH34

Name	Sequence (5'-3')
ZBP1_D526E_QC_FW	GTGGTTCCACGGGAGCAGACCCCTGATGA
ZBP1_D526N_QC_FW	GTGGTGGTTCCACGGAATCAGACCCCTGATG
ZBP1_D526Q_QC_FW	GTGGTGGTTCCACGGCAGCAGACCCCTGATGAG
ZBP1_G500A_QC_FW	CCTCGGCTGCAGCGAGGGTGATCGG
ZBP1_Q514R_QC_FW	CGTCAATGAGCTGCGGAACCTGACGGCTG
ZBP1_R452C_QC_FW	CCGGACTCCAAAGTGTGCATGGTGGTCATCA
ZBP1_R452G_QC_FW	CGGACTCCAAAGTGGGCATGGTGGTCATC
ZBP1_S432R_QC_FW	GTGTAGTTTGTCGAGGCGGCCAAACGGTCGCG
ZBP1_D526E_QC_RV	TCATCAGGGGTCTGCTCCCGTGGAACCAC

ZBP1_D526N_QC_RV	CATCAGGGGTCTGATTCCGTGGAACCACCAC
ZBP1_D526Q_QC_RV	CTCATCAGGGGTCTGCTGCCGTGGAACCACCAC
ZBP1_G500A_QC_RV	CCGATCACCCTCGCTGCAGCCGAGG
ZBP1_Q514R_QC_RV	CAGCCGTCAGGTTCCGCAGCTCATTGACG
ZBP1_R452C_QC_RV	TGATGACCACCATGCACACTTTGGAGTCCGG
ZBP1_R452G_QC_RV	GATGACCACCATGCCCACTTTGGAGTCCG
ZBP1_S432R_QC_RV	CACATCAAACAGCTCCGCCGGTTTGCCAGCGC

PRIMERS FOR IMP1 AND IMP3 RRM1 PETM-11 VECTOR CLONING

NAME	Sequence (5'-3')	Restriction Site
IMP1_RRM12_PETM11_FW	CTTGCCATGGGCAAGCTTTACATCGGCAACCTCAACG	NcoI
IMP1_RRM12_PETM11_RV	CTTGCTCGAGATTATGGGGCCCCCGCTG	XhoI
IMP3_RRM12_PETM11_FW	CTTGCCATGGGGAAACATCACCATCACCATCACC	NcoI
IMP3_RRM12_PETM11_RV	CTTGCTCGAGATTAAGGCAGGGGTCTCCAGGATCCTAA	XhoI

PRIMERS FOR MUTAGENESIS OF IMP1 AND IMP3 RRM12

NAME	Sequence (5'-3')
IMP1_Y5A_QCML	CGCCATGGGCAAGCTTGCCATCGGCAACCTCAAC
IMP1_K36E_QCML	GGCCAGTTCTTGGTCGAGTCCGGCTACGCCTTC
IMP1_Y39A_QCML	CTTGGTCAAATCCGGCGCCGCCTTCGTGGACTGC
IMP1_K66E_QCML	ACTTTCTCCGGGAAAGTAGAATTACAAGGAGAGCGCTTAGAGATTGAAC
IMP1_K36EY39A_QCML	GGCCAGTTCTTGGTCGAGTCCGGCGCCGCCTTCGTGGACTG
IMP3_Y5A_QCML	CGCGTAGCCAGTCTCCACCAGGAAGGGTC
IMP3_K36E_QCML	GAGTGCTCAACTTCTATGGGCTCCCCGTGCAGTTCTATTTTAC
IMP3_Y39A_QCML	CAGTCCACGAACGCGGCGCCAGTCTTCACCAG
IMP3_K66E_QCML	CGGTACCCGTTGTTTGACCGATAGCCTTTGGAGTCGCT
IMP3_K36EY39A_QC_FW	GTCCACGAACGCGGCGCCAGTCTCCACCAGGAAGGGT
IMP3_K36EY39A_QC_RV	ACCCTTCCTGGTGGAGACTGGCGCCGCGTTCGTGGAC

Appendix II: List of additional ZBP1 KH34 selectivity mutations
ZBP1 KH34 selectivity mutations that were tested in addition to those reported in Chapter 4: Results section that were unsuccessful

Mutant	*Result*
S432Q	Not soluble
R452C	Not soluble
Q514R	No effect on RNA binding
D526E	Unfolded
D526N	Unfolded

Appendix III: Individual Δδ, weighted average Δδav, and normalised SIA values for ^{15}N and ^{1}H resonances of IMP1 RRM12

Free			***nANNNN***						
#	^{15}N	^{1}H	*#*	^{15}N	^{1}H	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	Δδav	Normalised
1	112.478	9.052	1	112.383	9.029	-0.095	-0.023	0.037835	0.610483
2	112.836	8.362	2	112.607	8.36	-0.229	-0.002	0.072444	0.783811
3	111.621	8.265	3	111.589	8.289	-0.032	0.024	0.026046	0.961642
4	108.91	8.329	4	108.708	8.334	-0.202	0.005	0.064073	0.654489
5	120.07	9.22	5	120.201	9.203	0.131	-0.017	0.044778	0.706939
6	117.771	9.18	6	117.654	9.169	-0.117	-0.011	0.038599	0.735257
7	115.709	9.066	7	115.811	9.092	0.102	0.026	0.041429	0.870321
8	117.328	8.598	8	117.429	8.608	0.101	0.01	0.033468	0.723708
9	113.584	8.102	9	113.524	8.086	-0.06	-0.016	0.024819	0.522414
10	113.369	7.966	10	113.452	7.988	0.083	0.022	0.034248	0.592683
11	122.613	8.031	11	122.666	8.047	0.053	0.016	0.023171	0.240734
12	122.331	7.933	12	122.412	7.944	0.081	0.011	0.027877	0.499703
13	122.509	8.279	13	122.66	8.283	0.151	0.004	0.047918	0.525398
14	126.351	8.509	14	126.468	8.521	0.117	0.012	0.038896	0.66771
15	121.851	9.168	15	121.779	9.159	-0.072	-0.009	0.024483	0.498
16	121.736	8.865	16	121.286	8.818	-0.45	-0.047	0.149863	0.721214

Free			***nCNNNN***						
#	^{15}N	^{1}H	*#*	^{15}N	^{1}H	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	Δδav	Normalised
1	112.478	9.052	1	112.358	9.003	-0.12	-0.049	0.061976	1
2	112.836	8.362	2	112.544	8.358	-0.292	-0.004	0.092425	1
3	111.621	8.265	3	111.597	8.291	-0.024	0.026	0.027085	1
4	108.91	8.329	4	108.601	8.335	-0.309	0.006	0.097898	1
5	120.07	9.22	5	120.239	9.186	0.169	-0.034	0.063341	1
6	117.771	9.18	6	117.611	9.166	-0.16	-0.014	0.052498	1
7	115.709	9.066	7	115.821	9.093	0.112	0.027	0.044535	0.935568
8	117.328	8.598	8	117.439	8.612	0.111	0.014	0.03779	0.817174
9	113.584	8.102	9	113.465	8.073	-0.119	-0.029	0.047509	1
10	113.369	7.966	10	113.519	7.999	0.15	0.033	0.057784	1
11	122.613	8.031	11	122.672	8.051	0.059	0.02	0.027351	0.284165
12	122.331	7.933	12	122.5	7.949	0.169	0.016	0.055786	1
13	122.509	8.279	13	122.796	8.288	0.287	0.009	0.091203	1
14	126.351	8.509	14	126.529	8.524	0.178	0.015	0.058253	1
15	121.851	9.168	15	121.704	9.152	-0.147	-0.016	0.049162	1
16	121.736	8.865	16	121.099	8.814	-0.637	-0.051	0.207793	1

Free			***nGNNNN***						
#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	112.478	9.052	1	112.361	9.013	-0.117	-0.039	0.053758	0.8674
2	112.836	8.362	2	112.584	8.357	-0.252	-0.005	0.079846	0.863901
3	111.621	8.265	3	111.602	8.283	-0.019	0.018	0.018976	0.700619
4	108.91	8.329	4	108.645	8.333	-0.265	0.004	0.083896	0.856968
5	120.07	9.22	5	120.261	9.202	0.191	-0.018	0.063025	0.995003
6	117.771	9.18	6	117.619	9.173	-0.152	-0.007	0.048574	0.925254
7	115.709	9.066	7	115.779	9.089	0.07	0.023	0.031922	0.67059
8	117.328	8.598	8	117.464	8.615	0.136	0.017	0.046245	1
9	113.584	8.102	9	113.529	8.089	-0.055	-0.013	0.021714	0.457052
10	113.369	7.966	10	113.502	7.993	0.133	0.027	0.049979	0.864927
11	122.613	8.031	11	122.375	8.091	-0.238	0.06	0.096252	1
12	122.331	7.933	12	122.475	7.947	0.144	0.014	0.04764	0.85398
13	122.509	8.279	13	122.702	8.287	0.193	0.008	0.061554	0.674916
14	126.351	8.509	14	126.479	8.522	0.128	0.013	0.042514	0.72981
15	121.851	9.168	15	121.801	9.16	-0.05	-0.008	0.01772	0.360442
16	121.736	8.865	16	121.305	8.825	-0.431	-0.04	0.142043	0.683578

Free			***nUNNNN***						
#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	112.478	9.052	1	112.42	9.04	-0.058	-0.012	0.021918	0.353655
2	112.836	8.362	2	112.691	8.362	-0.145	0	0.045853	0.49611
3	111.621	8.265	3	111.619	8.277	-0.002	0.012	0.012017	0.443664
4	108.91	8.329	4	108.835	8.349	-0.075	0.02	0.031024	0.316902
5	120.07	9.22	5	120.136	9.21	0.066	-0.01	0.023143	0.365371
6	117.771	9.18	6	117.7	9.179	-0.071	-0.001	0.022474	0.428104
7	115.709	9.066	7	115.839	9.09	0.13	0.024	0.047603	1
8	117.328	8.598	8	117.409	8.606	0.081	0.008	0.026835	0.580272
9	113.584	8.102	9	113.561	8.096	-0.023	-0.006	0.009429	0.198461
10	113.369	7.966	10	113.397	7.979	0.028	0.013	0.015729	0.272202
11	122.613	8.031	11	122.388	8.094	-0.225	0.063	0.095034	0.98735
12	122.331	7.933	12	122.373	7.942	0.042	0.009	0.016044	0.287592
13	122.509	8.279	13	122.589	8.287	0.08	0.008	0.026533	0.290924
14	126.351	8.509	14	126.387	8.516	0.036	0.007	0.013364	0.229416
15	121.851	9.168	15	121.852	9.169	0.001	0.001	0.001049	0.021334
16	121.736	8.865	16	121.523	8.84	-0.213	-0.025	0.071846	0.345759

Free			***nNANN***						
#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	112.478	9.052	1	112.535	9.022	0.057	-0.03	0.034999	0.676486
2	112.836	8.362	2	112.655	8.365	-0.181	0.003	0.057316	0.748705
3	111.621	8.265	3	111.602	8.287	-0.019	0.022	0.022806	1
4	108.91	8.329	4	108.751	8.339	-0.159	0.01	0.051265	0.583882
5	120.07	9.22	5	120.118	9.205	0.048	-0.015	0.02134	0.489769
6	117.771	9.18	6	117.706	9.163	-0.065	-0.017	0.026674	0.620056
7	115.709	9.066	7	115.833	9.079	0.124	0.013	0.041311	0.873518
8	117.328	8.598	8	117.418	8.609	0.09	0.011	0.030512	0.691038
9	113.584	8.102	9	113.548	8.09	-0.036	-0.012	0.016541	0.690402
10	113.369	7.966	10	113.441	7.987	0.072	0.021	0.030974	0.668302
11	122.331	7.933	11	122.407	7.948	0.076	0.015	0.02833	0.853761
12	122.509	8.279	12	122.632	8.286	0.123	0.007	0.039521	0.795767
13	126.351	8.509	13	126.474	8.521	0.123	0.012	0.040705	1
14	121.851	9.168	14	121.778	9.163	-0.073	-0.005	0.02362	0.78282
15	121.736	8.865	15	121.726	8.861	-0.01	-0.004	0.005099	0.034438

Free			***nNCNN***						
#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	112.478	9.052	1	112.408	9.015	-0.07	-0.037	0.043116	0.833389
2	112.836	8.362	2	112.6	8.359	-0.236	-0.003	0.07469	0.975661
3	111.621	8.265	3	111.614	8.286	-0.007	0.021	0.021116	0.925924
4	108.91	8.329	4	108.633	8.323	-0.277	-0.006	0.0878	1
5	120.07	9.22	5	120.185	9.196	0.115	-0.024	0.043572	1
6	117.771	9.18	6	117.655	9.17	-0.116	-0.01	0.038021	0.883828
7	115.709	9.066	7	115.824	9.091	0.115	0.025	0.04413	0.933135
8	117.328	8.598	8	117.407	8.609	0.079	0.011	0.027297	0.618208
9	113.584	8.102	9	113.534	8.084	-0.05	-0.018	0.023958	1
10	113.369	7.966	10	113.498	7.988	0.129	0.022	0.046348	1
11	122.331	7.933	11	122.396	7.944	0.065	0.011	0.023313	0.702565
12	122.509	8.279	12	122.649	8.287	0.14	0.008	0.044989	0.905868
13	126.351	8.509	13	126.466	8.519	0.115	0.01	0.037716	0.926569
14	121.851	9.168	14	121.759	9.16	-0.092	-0.008	0.030173	1
15	121.736	8.865	15	121.292	8.818	-0.444	-0.047	0.148063	1

Free			***nNGNN***						
#	^{15}N	^{1}H	#	^{15}N	^{1}H	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta av$	Normalised
1	112.478	9.052	1	115.496	112.352	9.019	103.3	0.051736	1
2	112.836	8.362	2	110.505	112.594	8.36	104.232	0.076553	1
3	111.621	8.265	3	115.232	111.617	8.281	103.352	0.01605	0.703768
4	108.91	8.329	4	120.686	108.722	8.333	100.393	0.059585	0.678645
5	120.07	9.22	5	127.298	120.13	9.203	110.91	0.025475	0.584678
6	117.771	9.18	6	111.459	117.689	9.162	108.509	0.031566	0.733771
7	115.709	9.066	7	107.02	115.843	9.087	106.777	0.047293	1
8	117.328	8.598	8	114.254	117.464	8.608	108.866	0.044154	1
9	113.584	8.102	9	125.846	113.529	8.09	105.427	0.021131	0.881972
10	113.369	7.966	10	114.948	113.486	7.986	105.52	0.042058	0.907454
11	122.331	7.933	11	117.973	122.43	7.944	114.497	0.033183	1
12	122.509	8.279	12	121.435	122.664	8.287	114.385	0.049664	1
13	126.351	8.509	13	123.306	126.431	8.517	117.922	0.026533	0.651836
14	121.851	9.168	14	121.877	121.798	9.162	112.63	0.017802	0.58999
15	121.736	8.865	15	129.446	121.425	8.871	112.56	0.09853	0.665459

Free			***nNUNN***						
#	^{15}N	^{1}H	#	^{15}N	^{1}H	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta av$	Normalised
1	112.478	9.052	1	112.484	9.03	0.006	-0.022	0.022082	0.426815
2	112.836	8.362	2	112.732	8.365	-0.104	0.003	0.033024	0.431389
3	111.621	8.265	3	111.638	8.279	0.017	0.014	0.014997	0.657584
4	108.91	8.329	4	108.844	8.345	-0.066	0.016	0.026298	0.299524
5	120.07	9.22	5	120.128	9.215	0.058	-0.005	0.019011	0.436304
6	117.771	9.18	6	117.635	9.179	-0.136	-0.001	0.043019	1
7	115.709	9.066	7	115.744	9.084	0.035	0.018	0.021131	0.446804
8	117.328	8.598	8	117.409	8.604	0.081	0.006	0.026308	0.595815
9	113.584	8.102	9	113.555	8.096	-0.029	-0.006	0.010959	0.45742
10	113.369	7.966	10	113.402	7.979	0.033	0.013	0.01667	0.359681
11	122.331	7.933	11	122.385	7.941	0.054	0.008	0.018857	0.568287
12	122.509	8.279	12	122.585	8.287	0.076	0.008	0.02533	0.510025
13	126.351	8.509	13	126.381	8.518	0.03	0.009	0.013077	0.321255
14	121.851	9.168	14	121.834	9.169	-0.017	0.001	0.005468	0.181226
15	121.736	8.865	15	121.541	8.847	-0.195	-0.018	0.064238	0.433855

Free			***nNNAN***						
#	^{15}N	^{1}H	*#*	^{15}N	^{1}H	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta av$	Normalised
1	112.478	9.052	1	112.37	9.028	-0.108	-0.024	0.041742	1
2	112.836	8.362	2	112.697	8.358	-0.139	-0.004	0.044137	0.611639
3	111.621	8.265	3	111.593	8.28	-0.028	0.015	0.017418	0.790439
4	108.91	8.329	4	108.731	8.337	-0.179	0.008	0.057167	0.705307
5	120.07	9.22	5	120.163	9.202	0.093	-0.018	0.03448	0.94854
6	117.771	9.18	6	117.643	9.167	-0.128	-0.013	0.042514	1
7	115.709	9.066	7	115.805	9.082	0.096	0.016	0.034316	0.72089
8	117.328	8.598	8	117.408	8.602	0.08	0.004	0.025612	0.770706
9	113.584	8.102	9	113.537	8.086	-0.047	-0.016	0.021838	1
10	113.369	7.966	10	113.419	7.979	0.05	0.013	0.020469	0.646979
11	122.613	8.031	11	122.64	8.04	0.027	0.009	0.012406	0.638838
12	122.331	7.933	12	122.406	7.94	0.075	0.007	0.024729	0.645849
13	122.509	8.279	13	122.624	8.284	0.115	0.005	0.036708	0.878728
14	126.351	8.509	14	126.429	8.515	0.078	0.006	0.025385	0.596248
15	121.851	9.168	15	121.789	9.159	-0.062	-0.009	0.021573	1
16	121.736	8.865	16	121.444	8.83	-0.292	-0.035	0.098749	0.675827

Free			***nNNCN***						
#	^{15}N	^{1}H	*#*	^{15}N	^{1}H	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta av$	Normalised
1	112.478	9.052	1	112.414	9.028	-0.064	-0.024	0.031394	0.752101
2	112.836	8.362	2	112.616	8.362	-0.22	0	0.06957	0.964078
3	111.621	8.265	3	111.625	8.287	0.004	0.022	0.022036	1
4	108.91	8.329	4	108.674	8.337	-0.236	0.008	0.075057	0.926027
5	120.07	9.22	5	120.167	9.206	0.097	-0.014	0.033718	0.927564
6	117.771	9.18	6	117.651	9.172	-0.12	-0.008	0.038781	0.912214
7	115.709	9.066	7	115.774	9.084	0.065	0.018	0.027322	0.573964
8	117.328	8.598	8	117.426	8.61	0.098	0.012	0.033233	1
9	113.584	8.102	9	113.528	8.093	-0.056	-0.009	0.019865	0.90963
10	113.369	7.966	10	113.44	7.988	0.071	0.022	0.031434	0.993536
11	122.613	8.031	11	122.652	8.046	0.039	0.015	0.019419	1
12	122.331	7.933	12	122.413	7.947	0.082	0.014	0.029469	0.769649
13	122.509	8.279	13	122.638	8.288	0.129	0.009	0.041774	1
14	126.351	8.509	14	126.477	8.524	0.126	0.015	0.042575	1
15	121.851	9.168	15	121.814	9.165	-0.037	-0.003	0.012079	0.559905
16	121.736	8.865	16	121.289	8.828	-0.447	-0.037	0.146116	1

Free			***nNNGN***						
#	^{15}N	^{1}H	*#*	^{15}N	^{1}H	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta av$	Normalised
1	112.478	9.052	1	112.387	9.033	-0.091	-0.019	0.034483	0.826105
2	112.836	8.362	2	112.608	8.359	-0.228	-0.003	0.072162	1
3	111.621	8.265	3	111.601	8.282	-0.02	0.017	0.018138	0.823111
4	108.91	8.329	4	108.654	8.333	-0.256	0.004	0.081053	1
5	120.07	9.22	5	120.168	9.201	0.098	-0.019	0.036351	1
6	117.771	9.18	6	117.716	9.174	-0.055	-0.006	0.018398	0.432765
7	115.709	9.066	7	115.839	9.09	0.13	0.024	0.047603	1
8	117.328	8.598	8	117.402	8.611	0.074	0.013	0.026769	0.805518
9	113.584	8.102	9	113.54	8.088	-0.044	-0.014	0.019738	0.903849
10	113.369	7.966	10	113.449	7.985	0.08	0.019	0.031639	1
11	122.613	8.031	11	122.641	8.046	0.028	0.015	0.017418	0.896973
12	122.331	7.933	12	122.441	7.949	0.11	0.016	0.038288	1
13	122.509	8.279	13	122.629	8.285	0.12	0.006	0.038419	0.919672
14	126.351	8.509	14	126.433	8.52	0.082	0.011	0.028167	0.661599
15	121.851	9.168	15	121.821	9.161	-0.03	-0.007	0.01179	0.546505
16	121.736	8.865	16	121.361	8.827	-0.375	-0.038	0.124525	0.852234

Free			***nNNUN***						
#	^{15}N	^{1}H	*#*	^{15}N	^{1}H	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta av$	Normalised
1	112.478	9.052	1	112.443	9.041	-0.035	-0.011	0.015604	0.373831
2	112.836	8.362	2	112.709	8.365	-0.127	0.003	0.040273	0.558087
3	111.621	8.265	3	111.596	8.282	-0.025	0.017	0.018748	0.850792
4	108.91	8.329	4	108.673	8.329	-0.237	0	0.074946	0.924653
5	120.07	9.22	5	120.109	9.208	0.039	-0.012	0.017208	0.473371
6	117.771	9.18	6	117.677	9.174	-0.094	-0.006	0.030325	0.7133
7	115.709	9.066	7	115.756	9.082	0.047	0.016	0.021838	0.458758
8	117.328	8.598	8	117.395	8.606	0.067	0.008	0.022647	0.68148
9	113.584	8.102	9	113.556	8.094	-0.028	-0.008	0.011933	0.546439
10	113.369	7.966	10	113.436	7.981	0.067	0.015	0.02596	0.820504
11	122.613	8.031	11	122.632	8.041	0.019	0.01	0.011666	0.60076
12	122.331	7.933	12	122.376	7.94	0.045	0.007	0.015859	0.414192
13	122.509	8.279	13	122.607	8.287	0.098	0.008	0.032006	0.766169
14	126.351	8.509	14	126.406	8.517	0.055	0.008	0.019144	0.449662
15	121.851	9.168	15	121.841	9.167	-0.01	-0.001	0.003317	0.153739
16	121.736	8.865	16	121.506	8.839	-0.23	-0.026	0.07724	0.52862

Free ***nNNNA***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	112.478	9.052	1	112.347	9.029	-0.131	-0.023	0.047382	0.823191
2	112.836	8.362	2	112.649	8.364	-0.187	0.002	0.059168	0.80642
3	111.621	8.265	3	111.618	8.281	-0.003	0.016	0.016028	0.767519
4	108.91	8.329	4	108.642	8.331	-0.268	0.002	0.084773	0.839988
5	120.07	9.22	5	120.116	9.209	0.046	-0.011	0.018237	0.380399
6	117.771	9.18	6	117.744	9.171	-0.027	-0.009	0.012406	0.298606
7	115.709	9.066	7	115.746	9.084	0.037	0.018	0.021469	0.359114
8	117.328	8.598	8	117.401	8.607	0.073	0.009	0.024777	0.67886
9	113.584	8.102	9	113.551	8.091	-0.033	-0.011	0.015162	0.589706
10	113.369	7.966	10	113.44	7.981	0.071	0.015	0.027002	0.564366
11	122.613	8.031	11	122.642	8.044	0.029	0.013	0.015909	0.62521
12	122.331	7.933	12	122.408	7.944	0.077	0.011	0.026719	0.677329
13	122.509	8.279	13	122.624	8.288	0.115	0.009	0.037463	0.726916
14	126.351	8.509	14	126.413	8.518	0.062	0.009	0.021573	0.420352
15	121.851	9.168	15	121.797	9.164	-0.054	-0.004	0.017539	0.771642
16	121.736	8.865	16	121.372	8.827	-0.364	-0.038	0.121217	0.785455

Free ***nNNNC***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	112.478	9.052	1	112.421	9.026	-0.057	-0.026	0.031637	0.54964
2	112.836	8.362	2	112.655	8.359	-0.181	-0.003	0.057316	0.781171
3	111.621	8.265	3	111.602	8.285	-0.019	0.02	0.020883	1
4	108.91	8.329	4	108.66	8.335	-0.25	0.006	0.079284	0.785606
5	120.07	9.22	5	120.149	9.206	0.079	-0.014	0.028637	0.597326
6	117.771	9.18	6	117.662	9.17	-0.109	-0.01	0.03589	0.863882
7	115.709	9.066	7	115.803	9.09	0.094	0.024	0.038205	0.639066
8	117.328	8.598	8	117.418	8.605	0.09	0.007	0.029309	0.803023
9	113.584	8.102	9	113.552	8.091	-0.032	-0.011	0.014947	0.58131
10	113.369	7.966	10	113.438	7.984	0.069	0.018	0.028286	0.591207
11	122.613	8.031	11	122.645	8.046	0.032	0.015	0.018094	0.711082
12	122.331	7.933	12	122.398	7.942	0.067	0.009	0.02302	0.58355
13	122.509	8.279	13	122.639	8.286	0.13	0.007	0.041701	0.809147
14	126.351	8.509	14	126.455	8.52	0.104	0.011	0.034679	0.675711
15	121.851	9.168	15	121.801	9.165	-0.05	-0.003	0.016093	0.708064
16	121.736	8.865	16	121.38	8.84	-0.356	-0.025	0.11532	0.74724

Free ***nNNNG***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta av$	Normalised
1	112.478	9.052	1	112.357	9.009	-0.121	-0.043	0.05756	1
2	112.836	8.362	2	112.604	8.361	-0.232	-0.001	0.073372	1
3	111.621	8.265	3	111.587	8.282	-0.034	0.017	0.020115	0.963208
4	108.91	8.329	4	108.591	8.332	-0.319	0.003	0.100921	1
5	120.07	9.22	5	120.215	9.206	0.145	-0.014	0.047943	1
6	117.771	9.18	6	117.641	9.174	-0.13	-0.006	0.041545	1
7	115.709	9.066	7	115.886	9.087	0.177	0.021	0.059782	1
8	117.328	8.598	8	117.437	8.61	0.109	0.012	0.036498	1
9	113.584	8.102	9	113.523	8.085	-0.061	-0.017	0.025712	1
10	113.369	7.966	10	113.498	7.991	0.129	0.025	0.047845	1
11	122.613	8.031	11	122.628	8.056	0.015	0.025	0.025446	1
12	122.331	7.933	12	122.442	7.951	0.111	0.018	0.039447	1
13	122.509	8.279	13	122.67	8.287	0.161	0.008	0.051537	1
14	126.351	8.509	14	126.508	8.522	0.157	0.013	0.051322	1
15	121.851	9.168	15	121.785	9.159	-0.066	-0.009	0.022729	1
16	121.736	8.865	16	121.263	8.827	-0.473	-0.038	0.154327	1

Free ***nNNNU***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta av$	Normalised
1	112.478	9.052	1	112.413	9.033	-0.065	-0.019	0.027991	0.486298
2	112.836	8.362	2	112.702	8.365	-0.134	0.003	0.042481	0.578978
3	111.621	8.265	3	111.615	8.277	-0.006	0.012	0.012149	0.581768
4	108.91	8.329	4	108.655	8.329	-0.255	0	0.080638	0.79902
5	120.07	9.22	5	120.103	9.212	0.033	-0.008	0.013149	0.274268
6	117.771	9.18	6	117.674	9.169	-0.097	-0.011	0.032587	0.784371
7	115.709	9.066	7	115.83	9.086	0.121	0.02	0.043175	0.72221
8	117.328	8.598	8	117.412	8.605	0.084	0.007	0.02747	0.752645
9	113.584	8.102	9	113.55	8.093	-0.034	-0.009	0.014021	0.545328
10	113.369	7.966	10	113.412	7.983	0.043	0.017	0.021769	0.455
11	122.613	8.031	11	122.631	8.043	0.018	0.012	0.013282	0.521951
12	122.331	7.933	12	122.401	7.944	0.07	0.011	0.024718	0.626617
13	122.509	8.279	13	122.607	8.289	0.098	0.01	0.032564	0.631848
14	126.351	8.509	14	126.463	8.521	0.112	0.012	0.037395	0.728645
15	121.851	9.168	15	121.811	9.166	-0.04	-0.002	0.012806	0.563436
16	121.736	8.865	16	121.388	8.834	-0.348	-0.031	0.11433	0.74083

Appendix IV: Individual Δδ, weighted average Δδav, and normalised SIA values for ^{15}N and ^{1}H resonances of IMP3 RRM12

Free			***nANNNN***						
#	^{15}N	^{1}H	#	^{15}N	^{1}H	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	Δδav	Normalised
1	115.661	7.066	1	115.591	6.929	-0.07	-0.137	0.138777	0.917265
2	111.121	8.511	2	110.552	8.5	-0.569	-0.011	0.18027	0.84486
3	115.064	8.462	3	115.184	8.498	0.12	0.036	0.052307	0.39003
4	120.614	9.216	4	120.663	9.128	0.049	-0.088	0.089354	0.806329
5	126.505	8.351	5	127.496	8.361	0.991	0.01	0.313541	0.838573
6	111.777	9.127	6	111.411	8.996	-0.366	-0.131	0.174804	0.935692
7	107.435	7.426	7	106.969	7.39	-0.466	-0.036	0.151696	0.818268
8	114.624	9.041	8	114.228	8.989	-0.396	-0.052	0.135594	0.732383
9	125.874	9.021	9	125.826	9.005	-0.048	-0.016	0.022054	0.722962
10	114.88	7.173	10	114.92	7.208	0.04	0.035	0.037216	0.651214
11	117.93	9.76	11	118.03	9.779	0.1	0.019	0.036892	0.679336
12	121.519	9.371	12	121.49	9.456	-0.029	0.085	0.085493	0.655503
13	123.073	7.699	13	123.288	7.715	0.215	0.016	0.069846	0.741457
14	121.812	9.056	14	121.985	9.086	0.173	0.03	0.062393	0.834165
15	129.625	8.54	15	129.424	8.514	-0.201	-0.026	0.068674	0.892727
16	111.341	7.778	16	111.255	7.749	-0.086	-0.029	0.039757	1

Free			***nCNNNN***						
#	^{15}N	^{1}H	#	^{15}N	^{1}H	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	Δδav	Normalised
1	115.661	7.066	1	115.458	6.929	-0.203	-0.137	0.151294	1
2	111.121	8.511	2	110.447	8.501	-0.674	-0.01	0.213372	1
3	115.064	8.462	3	115.302	8.573	0.238	0.111	0.13411	1
4	120.614	9.216	4	120.835	9.13	0.221	-0.086	0.110816	1
5	126.505	8.351	5	127.686	8.369	1.181	0.018	0.373899	1
6	111.777	9.127	6	111.438	8.974	-0.339	-0.153	0.186818	1
7	107.435	7.426	7	106.864	7.384	-0.571	-0.042	0.185386	1
8	114.624	9.041	8	114.077	8.975	-0.547	-0.066	0.18514	1
9	125.874	9.021	9	125.79	9.006	-0.084	-0.015	0.030506	1
10	114.88	7.173	10	114.973	7.222	0.093	0.049	0.057148	1
11	117.93	9.76	11	118.079	9.787	0.149	0.027	0.054306	1
12	121.519	9.371	12	121.395	9.47	-0.124	0.099	0.106483	0.816437
13	123.073	7.699	13	123.351	7.729	0.278	0.03	0.092889	0.98607
14	121.812	9.056	14	121.996	9.103	0.184	0.047	0.074797	1
15	129.625	8.54	15	129.401	8.51	-0.224	-0.03	0.076926	1
16	111.341	7.778	16	111.251	7.753	-0.09	-0.025	0.037881	0.952829

Free ***nGNNNN***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	115.661	7.066	1	115.615	6.963	-0.046	-0.103	0.104022	0.687549
2	111.121	8.511	2	110.578	8.497	-0.543	-0.014	0.172281	0.807423
3	115.064	8.462	3	115.339	8.489	0.275	0.027	0.091058	0.678979
4	120.614	9.216	4	120.778	9.153	0.164	-0.063	0.0816	0.736361
5	126.505	8.351	5	127.436	8.372	0.931	0.021	0.295156	0.789402
6	111.777	9.127	6	111.438	8.998	-0.339	-0.129	0.167729	0.89782
7	107.435	7.426	7	106.998	7.393	-0.437	-0.033	0.142077	0.766384
8	114.624	9.041	8	114.295	8.982	-0.329	-0.059	0.119604	0.646018
9	125.874	9.021	9	125.788	9.01	-0.086	-0.011	0.029336	0.961655
10	114.88	7.173	10	114.919	7.204	0.039	0.031	0.033363	0.583802
11	117.93	9.76	11	118.065	9.764	0.135	0.004	0.042878	0.789564
12	121.519	9.371	12	121.171	9.441	-0.348	0.07	0.130424	1
13	123.073	7.699	13	123.298	7.716	0.225	0.017	0.073154	0.77657
14	121.812	9.056	14	121.97	9.089	0.158	0.033	0.059878	0.800542
15	129.625	8.54	15	129.45	8.517	-0.175	-0.023	0.059929	0.77905
16	111.341	7.778	16	111.276	7.756	-0.065	-0.022	0.030108	0.757309

Free ***nUNNNN***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	115.661	7.066	1	115.494	6.953	-0.167	-0.113	0.124731	0.82443
2	111.121	8.511	2	110.604	8.498	-0.517	-0.013	0.164006	0.768638
3	115.064	8.462	3	115.266	8.52	0.202	0.058	0.086281	0.643361
4	120.614	9.216	4	120.737	9.156	0.123	-0.06	0.071505	0.645257
5	126.505	8.351	5	127.46	8.356	0.955	0.005	0.302039	0.80781
6	111.777	9.127	6	111.501	9.009	-0.276	-0.118	0.146771	0.785632
7	107.435	7.426	7	106.99	7.391	-0.445	-0.035	0.145009	0.782197
8	114.624	9.041	8	114.228	8.989	-0.396	-0.052	0.135594	0.732383
9	125.874	9.021	9	125.852	9.012	-0.022	-0.009	0.011375	0.372894
10	114.88	7.173	10	114.979	7.211	0.099	0.038	0.049235	0.861537
11	117.93	9.76	11	118.031	9.782	0.101	0.022	0.038783	0.714157
12	121.519	9.371	12	121.44	9.446	-0.079	0.075	0.079051	0.60611
13	123.073	7.699	13	123.366	7.716	0.293	0.017	0.094201	1
14	121.812	9.056	14	121.944	9.09	0.132	0.034	0.053837	0.719771
15	129.625	8.54	15	129.452	8.516	-0.173	-0.024	0.05974	0.776595
16	111.341	7.778	16	111.288	7.759	-0.053	-0.019	0.025336	0.637269

Free ***nNANN***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	115.661	7.066	1	115.52	6.969	-0.141	-0.097	0.106757	0.705627
2	111.121	8.511	2	110.679	8.501	-0.442	-0.01	0.14013	0.71842
3	115.064	8.462	3	115.18	8.51	0.116	0.048	0.060412	0.450467
4	120.614	9.216	4	120.697	9.147	0.083	-0.069	0.073823	0.790807
5	126.505	8.351	5	127.67	8.357	1.165	0.006	0.368454	1
6	111.777	9.127	6	111.522	9.012	-0.255	-0.115	0.140455	0.751824
7	107.435	7.426	7	107.025	7.396	-0.41	-0.03	0.133079	0.777536
8	114.624	9.041	8	114.28	9.006	-0.344	-0.035	0.114274	0.671585
9	125.874	9.021	9	125.834	9.01	-0.04	-0.011	0.016763	0.549505
10	114.88	7.173	10	114.939	7.201	0.059	0.028	0.033647	0.719561
11	117.93	9.76	11	118.017	9.777	0.087	0.017	0.03234	0.694665
12	121.519	9.371	12	121.395	9.448	-0.124	0.077	0.086409	0.811487
13	123.073	7.699	13	123.225	7.718	0.152	0.019	0.051686	0.526162
14	121.812	9.056	14	121.979	9.082	0.167	0.026	0.058863	0.967905
15	129.625	8.54	15	129.468	8.519	-0.157	-0.021	0.053906	0.728106
16	111.341	7.778	16	111.277	7.753	-0.064	-0.025	0.032165	0.849103

Free ***nNCNN***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	115.661	7.066	1	115.458	6.929	-0.203	-0.137	0.151294	1
2	111.121	8.511	2	110.539	8.495	-0.582	-0.016	0.184739	0.947122
3	115.064	8.462	3	115.302	8.573	0.238	0.111	0.13411	1
4	120.614	9.216	4	120.798	9.143	0.184	-0.073	0.093352	1
5	126.505	8.351	5	127.593	8.365	1.088	0.014	0.344341	0.934554
6	111.777	9.127	6	111.438	8.974	-0.339	-0.153	0.186818	1
7	107.435	7.426	7	106.908	7.387	-0.527	-0.039	0.171155	1
8	114.624	9.041	8	114.135	8.97	-0.489	-0.071	0.170156	1
9	125.874	9.021	9	125.79	9.006	-0.084	-0.015	0.030506	1
10	114.88	7.173	10	114.945	7.215	0.065	0.042	0.04676	1
11	117.93	9.76	11	118.058	9.783	0.128	0.023	0.046555	1
12	121.519	9.371	12	121.395	9.47	-0.124	0.099	0.106483	1
13	123.073	7.699	13	123.375	7.722	0.302	0.023	0.098231	1
14	121.812	9.056	14	121.967	9.092	0.155	0.036	0.060815	1
15	129.625	8.54	15	129.407	8.513	-0.218	-0.027	0.074036	1
16	111.341	7.778	16	111.251	7.753	-0.09	-0.025	0.037881	1

Free ***nNGNN***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	115.661	7.066	1	115.496	6.961	-0.165	-0.105	0.11725	0.774979
2	111.121	8.511	2	110.505	8.501	-0.616	-0.01	0.195053	1
3	115.064	8.462	3	115.232	8.497	0.168	0.035	0.063619	0.474382
4	120.614	9.216	4	120.686	9.149	0.072	-0.067	0.070763	0.758023
5	126.505	8.351	5	127.298	8.351	0.793	0	0.250769	0.680596
6	111.777	9.127	6	111.459	9.004	-0.318	-0.123	0.158875	0.850427
7	107.435	7.426	7	107.02	7.395	-0.415	-0.031	0.134846	0.787862
8	114.624	9.041	8	114.254	8.985	-0.37	-0.056	0.129715	0.76233
9	125.874	9.021	9	125.846	9.007	-0.028	-0.014	0.016565	0.543013
10	114.88	7.173	10	114.948	7.202	0.068	0.029	0.036103	0.772083
11	117.93	9.76	11	117.973	9.784	0.043	0.024	0.027584	0.592508
12	121.519	9.371	12	121.435	9.443	-0.084	0.072	0.076744	0.720714
13	123.073	7.699	13	123.306	7.712	0.233	0.013	0.074819	0.761662
14	121.812	9.056	14	121.877	9.089	0.065	0.033	0.038878	0.63928
15	129.625	8.54	15	129.446	8.521	-0.179	-0.019	0.059708	0.806474
16	111.341	7.778	16	111.267	7.758	-0.074	-0.02	0.030783	0.812618

Free ***nNUNN***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	115.661	7.066	1	115.53	6.963	-0.131	-0.103	0.111018	0.733793
2	111.121	8.511	2	110.644	8.504	-0.477	-0.007	0.151003	0.774165
3	115.064	8.462	3	115.186	8.512	0.122	0.05	0.063154	0.470912
4	120.614	9.216	4	120.737	9.157	0.123	-0.059	0.070668	0.757001
5	126.505	8.351	5	127.286	8.357	0.781	0.006	0.247047	0.670495
6	111.777	9.127	6	111.435	9.017	-0.342	-0.11	0.154261	0.825726
7	107.435	7.426	7	107.034	7.398	-0.401	-0.028	0.129862	0.75874
8	114.624	9.041	8	114.275	9.004	-0.349	-0.037	0.116401	0.684081
9	125.874	9.021	9	125.832	9.014	-0.042	-0.007	0.015013	0.492148
10	114.88	7.173	10	114.95	7.201	0.07	0.028	0.035693	0.763326
11	117.93	9.76	11	118.069	9.775	0.139	0.015	0.046445	0.997621
12	121.519	9.371	12	121.445	9.439	-0.074	0.068	0.071914	0.675356
13	123.073	7.699	13	123.252	7.721	0.179	0.022	0.06073	0.618232
14	121.812	9.056	14	121.943	9.089	0.131	0.033	0.052963	0.870886
15	129.625	8.54	15	129.449	8.522	-0.176	-0.018	0.058494	0.790076
16	111.341	7.778	16	111.279	7.763	-0.062	-0.015	0.024686	0.651666

Free ***nNNAN***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	115.661	7.066	1	115.466	6.944	-0.195	-0.122	0.136699	1
2	111.121	8.511	2	110.584	8.5	-0.537	-0.011	0.17017	1
3	115.064	8.462	3	115.175	8.569	0.111	0.107	0.11261	1
4	120.614	9.216	4	120.748	9.141	0.134	-0.075	0.086143	0.996215
5	126.505	8.351	5	127.444	8.363	0.939	0.012	0.29718	1
6	111.777	9.127	6	111.382	8.998	-0.395	-0.129	0.179565	1
7	107.435	7.426	7	106.989	7.391	-0.446	-0.035	0.145316	1
8	114.624	9.041	8	114.125	8.99	-0.499	-0.051	0.165835	1
9	125.874	9.021	9	125.817	9.009	-0.057	-0.012	0.021654	1
10	114.88	7.173	10	114.965	7.207	0.085	0.034	0.043342	0.822702
11	117.93	9.76	11	118.042	9.776	0.112	0.016	0.038864	0.689288
12	121.519	9.371	12	121.394	9.449	-0.125	0.078	0.087444	0.935332
13	123.073	7.699	13	123.261	7.722	0.188	0.023	0.063745	0.903332
14	121.812	9.056	14	121.986	9.096	0.174	0.04	0.068026	1
15	129.625	8.54	15	129.446	8.519	-0.179	-0.021	0.060375	0.940078
16	111.341	7.778	16	111.273	7.75	-0.068	-0.028	0.035304	1

Free ***nNNCN***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	115.661	7.066	1	115.476	6.96	-0.185	-0.106	0.121072	0.885688
2	111.121	8.511	2	110.594	8.503	-0.527	-0.008	0.166844	0.980453
3	115.064	8.462	3	115.161	8.505	0.097	0.043	0.05282	0.469046
4	120.614	9.216	4	120.825	9.161	0.211	-0.055	0.08647	1
5	126.505	8.351	5	127.334	8.361	0.829	0.01	0.262343	0.882776
6	111.777	9.127	6	111.508	9.016	-0.269	-0.111	0.139847	0.778809
7	107.435	7.426	7	107.028	7.396	-0.407	-0.03	0.132155	0.909434
8	114.624	9.041	8	114.314	8.996	-0.31	-0.045	0.107866	0.650441
9	125.874	9.021	9	125.835	9.012	-0.039	-0.009	0.015268	0.705068
10	114.88	7.173	10	114.934	7.207	0.054	0.034	0.038047	0.722207
11	117.93	9.76	11	118.1	9.777	0.17	0.017	0.056383	1
12	121.519	9.371	12	121.474	9.449	-0.045	0.078	0.079287	0.848084
13	123.073	7.699	13	123.287	7.719	0.214	0.02	0.070566	1
14	121.812	9.056	14	121.979	9.082	0.167	0.026	0.058863	0.865301
15	129.625	8.54	15	129.448	8.516	-0.177	-0.024	0.060901	0.948269
16	111.341	7.778	16	111.266	7.758	-0.075	-0.02	0.031024	0.878763

Free ***nNNGN***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	115.661	7.066	1	115.564	6.974	-0.097	-0.092	0.096979	0.709436
2	111.121	8.511	2	110.617	8.501	-0.504	-0.01	0.159692	0.938426
3	115.064	8.462	3	115.137	8.507	0.073	0.045	0.050576	0.449121
4	120.614	9.216	4	120.749	9.16	0.135	-0.056	0.070417	0.814345
5	126.505	8.351	5	127.329	8.36	0.824	0.009	0.260727	0.877336
6	111.777	9.127	6	111.37	9.012	-0.407	-0.115	0.172598	0.961199
7	107.435	7.426	7	107.027	7.395	-0.408	-0.031	0.132693	0.913136
8	114.624	9.041	8	114.257	8.991	-0.367	-0.05	0.126368	0.762013
9	125.874	9.021	9	125.844	9.013	-0.03	-0.008	0.01241	0.573087
10	114.88	7.173	10	114.992	7.212	0.112	0.039	0.052682	1
11	117.93	9.76	11	118.013	9.769	0.083	0.009	0.027747	0.492121
12	121.519	9.371	12	121.377	9.453	-0.142	0.082	0.09349	1
13	123.073	7.699	13	123.259	7.715	0.186	0.016	0.060956	0.863808
14	121.812	9.056	14	121.898	9.086	0.086	0.03	0.040492	0.595239
15	129.625	8.54	15	129.431	8.521	-0.194	-0.019	0.064223	1
16	111.341	7.778	16	111.296	7.762	-0.045	-0.016	0.021413	0.606514

Free ***nNNUN***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	115.661	7.066	1	115.607	6.97	-0.054	-0.096	0.097507	0.713299
2	111.121	8.511	2	110.654	8.503	-0.467	-0.008	0.147895	0.8691
3	115.064	8.462	3	115.181	8.512	0.117	0.05	0.0622	0.552351
4	120.614	9.216	4	120.694	9.153	0.08	-0.063	0.06789	0.785121
5	126.505	8.351	5	127.273	8.368	0.768	0.017	0.243457	0.819224
6	111.777	9.127	6	111.479	9.027	-0.298	-0.1	0.137406	0.765217
7	107.435	7.426	7	107.074	7.396	-0.361	-0.03	0.118034	0.812262
8	114.624	9.041	8	114.307	9.002	-0.317	-0.039	0.107563	0.648619
9	125.874	9.021	9	125.844	9.013	-0.03	-0.008	0.01241	0.573087
10	114.88	7.173	10	114.945	7.205	0.065	0.032	0.038033	0.721932
11	117.93	9.76	11	118.061	9.781	0.131	0.021	0.046445	0.823739
12	121.519	9.371	12	121.501	9.442	-0.018	0.071	0.071228	0.761875
13	123.073	7.699	13	123.245	7.716	0.172	0.017	0.056986	0.807552
14	121.812	9.056	14	121.892	9.086	0.08	0.03	0.039243	0.576876
15	129.625	8.54	15	129.481	8.523	-0.144	-0.017	0.048607	0.75684
16	111.341	7.778	16	111.313	7.763	-0.028	-0.015	0.017418	0.493377

Free ***nNNNA***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	115.661	7.066	1	115.55	6.94	-0.111	-0.126	0.130798	0.986976
2	111.121	8.511	2	110.494	8.499	-0.627	-0.012	0.198638	1
3	115.064	8.462	3	115.18	8.514	0.116	0.052	0.063636	0.782186
4	120.614	9.216	4	120.822	9.148	0.208	-0.068	0.094607	1
5	126.505	8.351	5	127.455	8.365	0.95	0.014	0.300742	1
6	111.777	9.127	6	111.486	8.996	-0.291	-0.131	0.160091	1
7	107.435	7.426	7	106.96	7.39	-0.475	-0.036	0.154462	1
8	114.624	9.041	8	114.164	8.989	-0.46	-0.052	0.15448	1
9	125.874	9.021	9	125.81	9.007	-0.064	-0.014	0.024609	1
10	114.88	7.173	10	114.939	7.208	0.059	0.035	0.039662	0.927561
11	117.93	9.76	11	118.075	9.78	0.145	0.02	0.050025	1
12	121.519	9.371	12	121.565	9.451	0.046	0.08	0.081312	0.799757
13	123.073	7.699	13	123.312	7.716	0.239	0.017	0.077467	0.976191
14	121.812	9.056	14	121.947	9.097	0.135	0.041	0.05919	0.914569
15	129.625	8.54	15	129.431	8.515	-0.194	-0.025	0.066247	1
16	111.341	7.778	16	111.282	7.752	-0.059	-0.026	0.032002	1

Free ***nNNNC***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	115.661	7.066	1	115.499	6.956	-0.162	-0.11	0.121344	0.915639
2	111.121	8.511	2	110.523	8.496	-0.598	-0.015	0.189698	0.954996
3	115.064	8.462	3	115.221	8.517	0.157	0.055	0.074094	0.910722
4	120.614	9.216	4	120.772	9.159	0.158	-0.057	0.075798	0.801196
5	126.505	8.351	5	127.358	8.358	0.853	0.007	0.269833	0.897223
6	111.777	9.127	6	111.563	9.013	-0.214	-0.114	0.132573	0.828111
7	107.435	7.426	7	106.986	7.395	-0.449	-0.031	0.145331	0.940885
8	114.624	9.041	8	114.253	8.995	-0.371	-0.046	0.126016	0.815746
9	125.874	9.021	9	125.843	9.01	-0.031	-0.011	0.014734	0.598738
10	114.88	7.173	10	114.962	7.207	0.082	0.034	0.04276	1
11	117.93	9.76	11	118.003	9.778	0.073	0.018	0.029273	0.585165
12	121.519	9.371	12	121.432	9.456	-0.087	0.085	0.089341	0.878735
13	123.073	7.699	13	123.299	7.721	0.226	0.022	0.074777	0.942296
14	121.812	9.056	14	121.93	9.095	0.118	0.039	0.053976	0.833999
15	129.625	8.54	15	129.465	8.514	-0.16	-0.026	0.056886	0.8587
16	111.341	7.778	16	111.268	7.758	-0.073	-0.02	0.030543	0.954435

Free ***nNNNG***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	115.661	7.066	1	115.467	6.967	-0.194	-0.099	0.116467	0.878839
2	111.121	8.511	2	110.621	8.496	-0.5	-0.015	0.158824	0.799566
3	115.064	8.462	3	115.194	8.501	0.13	0.039	0.056666	0.696505
4	120.614	9.216	4	120.7	9.148	0.086	-0.068	0.073237	0.774118
5	126.505	8.351	5	127.332	8.364	0.827	0.013	0.261843	0.870656
6	111.777	9.127	6	111.606	9.018	-0.171	-0.109	0.121676	0.760044
7	107.435	7.426	7	107.046	7.396	-0.389	-0.03	0.126618	0.819735
8	114.624	9.041	8	114.225	8.994	-0.399	-0.047	0.134644	0.871599
9	125.874	9.021	9	125.847	9.013	-0.027	-0.008	0.0117	0.475454
10	114.88	7.173	10	114.941	7.2	0.061	0.027	0.033183	0.776029
11	117.93	9.76	11	118.018	9.779	0.088	0.019	0.033696	0.673577
12	121.519	9.371	12	121.392	9.457	-0.127	0.086	0.094915	0.933557
13	123.073	7.699	13	123.251	7.717	0.178	0.018	0.059097	0.7447
14	121.812	9.056	14	121.998	9.083	0.186	0.027	0.064719	1
15	129.625	8.54	15	129.519	8.52	-0.106	-0.02	0.039033	0.589213
16	111.341	7.778	16	111.296	7.758	-0.045	-0.02	0.024546	0.767021

Free ***nNNNU***

#	^{15}N	^{1}H	#	^{15}N	^{1}H	Δδ^{15}N	Δδ^{1}H	Δδav	Normalised
1	115.661	7.066	1	115.505	6.943	-0.156	-0.123	0.132524	1
2	111.121	8.511	2	110.658	8.504	-0.463	-0.007	0.146581	0.73793
3	115.064	8.462	3	115.274	8.509	0.21	0.047	0.081357	1
4	120.614	9.216	4	120.662	9.152	0.048	-0.064	0.065775	0.695252
5	126.505	8.351	5	127.423	8.365	0.918	0.014	0.290634	0.96639
6	111.777	9.127	6	111.462	9.007	-0.315	-0.12	0.155957	0.974176
7	107.435	7.426	7	107.016	7.396	-0.419	-0.03	0.135853	0.879525
8	114.624	9.041	8	114.197	8.991	-0.427	-0.05	0.143989	0.932091
9	125.874	9.021	9	125.837	9.011	-0.037	-0.01	0.015392	0.625446
10	114.88	7.173	10	114.937	7.206	0.057	0.033	0.037602	0.879374
11	117.93	9.76	11	118.01	9.78	0.08	0.02	0.032249	0.644658
12	121.519	9.371	12	121.292	9.443	-0.227	0.072	0.101671	1
13	123.073	7.699	13	123.315	7.72	0.242	0.021	0.079356	1
14	121.812	9.056	14	121.939	9.092	0.127	0.036	0.053934	0.833355
15	129.625	8.54	15	129.475	8.516	-0.15	-0.024	0.05316	0.802459
16	111.341	7.778	16	111.288	7.755	-0.053	-0.023	0.028459	0.889292

Appendix V: Individual peak T1, T2, T1/T2 and tc relaxation values for IMP1 RRM12. Average and standard deviation values included at table bottom

IMP1				
Peak #	T1 (ms)	T2 (ms)	T1/T2	tc (ns)
1	1258.12	91.89	13.69	9.73
2	931.21	54.76	17.01	10.94
3	1137.42	66.12	17.20	11.00
4	1095.37	62.40	17.56	11.12
5	1093.95	68.59	15.95	10.56
6	1173.62	75.53	15.54	10.42
7	937.75	66.83	14.03	9.86
8	1083.12	59.90	18.08	11.30
9	979.73	55.58	17.63	11.15
10	1062.58	63.81	16.65	10.81
11	1011.60	68.03	14.87	10.17
12	1040.34	53.53	19.44	11.75
13	1175.99	69.61	16.89	10.90
14	1004.22	73.61	13.64	9.71
15	1102.57	63.36	17.40	11.07
16	1087.02	62.93	17.27	11.03
17	1027.81	69.59	14.77	10.13
18	1005.46	76.34	13.17	9.52
19	865.43	72.22	11.98	9.04
20	1048.10	66.26	15.82	10.52
21	1002.39	67.34	14.88	10.18
22	1119.50	57.04	19.63	11.81
23	1078.48	63.58	16.96	10.92
24	939.47	55.65	16.88	10.89
25	884.22	79.93	11.06	8.64
26	1014.94	77.12	13.16	9.52
27	1112.17	65.42	17.00	10.93
28	1159.73	62.75	18.48	11.43
29	892.34	60.48	14.75	10.13
30	1040.92	61.82	16.84	10.88
31	1090.08	60.76	17.94	11.25
32	1186.48	70.78	16.76	10.85
33	986.23	56.58	17.43	11.08
34	1008.21	68.92	14.63	10.08
35	1076.86	65.50	16.44	10.74
36	982.69	54.64	17.99	11.27
37	1180.48	71.27	16.56	10.78
38	880.89	92.17	9.56	7.96

39	902.45	55.23	16.34	10.70
40	1136.90	60.32	18.85	11.55
41	1057.38	60.12	17.59	11.14
42	924.16	66.94	13.81	9.77
43	1157.45	61.82	18.72	11.51
44	965.48	56.94	16.96	10.92
45	945.26	56.40	16.76	10.85
46	1215.85	54.13	22.46	12.68
47	1016.49	55.91	18.18	11.33
48	1049.19	66.02	15.89	10.55
49	844.79	81.82	10.32	8.32
50	970.00	70.09	13.84	9.78
51	930.17	66.17	14.06	9.87
52	967.29	66.15	14.62	10.08
53	1139.34	57.28	19.89	11.89
54	909.74	55.92	16.27	10.68
55	1064.28	86.31	12.33	9.18
56	946.70	71.73	13.20	9.53
57	920.86	73.57	12.52	9.26
58	999.17	66.39	15.05	10.24
59	916.63	64.91	14.12	9.89
60	1040.17	59.25	17.56	11.12
61	1040.19	66.68	15.60	10.44
62	1114.18	59.07	18.86	11.56
63	983.04	52.95	18.56	11.46
64	875.56	58.43	14.98	10.21
65	1079.79	61.24	17.63	11.15
66	998.13	60.62	16.46	10.75
67	916.22	49.47	18.52	11.45
68	973.42	51.09	19.05	11.62
69	954.79	96.33	9.91	8.13
70	993.89	79.15	12.56	9.27
71	943.72	67.47	13.99	9.84
72	1051.18	63.43	16.57	10.79
73	1058.50	61.92	17.09	10.97
74	1199.41	54.55	21.99	12.54
75	1157.82	66.37	17.44	11.09
76	979.51	58.07	16.87	10.89
77	1048.73	51.79	20.25	12.00
78	1043.03	58.45	17.85	11.22
79	1333.80	65.47	20.37	12.04
80	990.29	63.21	15.67	10.46
81	928.11	60.83	15.26	10.31
82	888.52	68.21	13.03	9.46
83	1133.17	57.86	19.59	11.79
84	890.86	68.52	13.00	9.45

85	930.33	37.89	24.55	13.29
84	1159.05	51.94	22.32	12.64
83	929.16	62.45	14.88	10.18
82	1057.27	69.55	15.20	10.29
81	1050.83	90.56	11.60	8.88
82	983.86	62.31	15.79	10.51
83	1130.33	58.65	19.27	11.69
Average	**1024.33**	**64.62**	**16.26**	**10.63**
Standard deviation	**112.94**	**9.81**	**2.77**	**0.99**

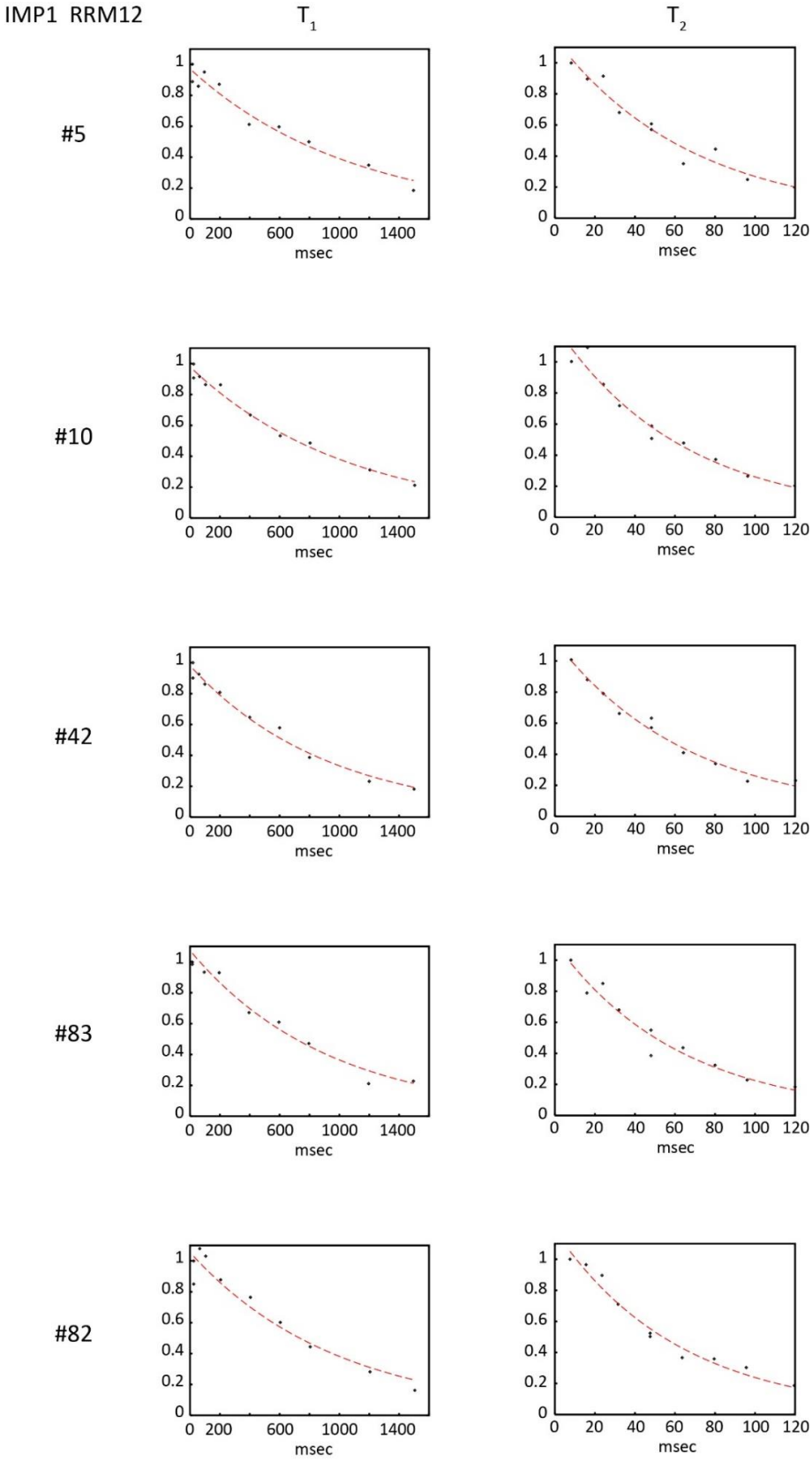
Appendix VI: Individual peak T1, T2, T1/T2 and tc relaxation values for IMP3 RRM12. Average and standard deviation values included at table bottom

IMP3				
Peak #	T1 (ms)	T2 (ms)	T1/T2	tc (ns)
1	1013.38	62.02	16.34	10.70
2	1023.95	44.49	23.02	12.84
3	1203.90	35.61	33.81	15.70
4	1195.45	50.84	23.51	12.99
5	1082.58	47.13	22.97	12.83
6	977.70	62.52	15.64	10.45
7	1065.93	68.79	15.50	10.40
8	1154.37	55.35	20.86	12.19
9	1034.44	50.56	20.46	12.07
10	951.87	34.77	27.38	14.07
11	1059.12	48.25	21.95	12.53
12	720.78	59.58	12.10	9.09
13	975.10	72.60	13.43	9.62
14	1068.31	70.87	15.07	10.25
15	1115.18	36.74	30.36	14.85
16	1021.07	71.10	14.36	9.98
17	1048.36	40.45	25.92	13.67
18	996.39	61.12	16.30	10.69
19	1068.89	70.13	15.24	10.31
20	786.81	37.24	21.13	12.28
21	891.67	51.57	17.29	11.03
22	1072.55	56.94	18.84	11.55
23	976.20	78.00	12.51	9.26
24	1107.49	49.32	22.46	12.68
25	916.52	46.95	19.52	11.77
26	991.47	66.53	14.90	10.18
27	886.68	70.42	12.59	9.29
28	1058.37	49.63	21.32	12.34
29	1084.80	49.27	22.02	12.55
30	722.64	97.87	7.38	6.85
31	1115.47	48.80	22.86	12.80
32	1148.83	54.57	21.05	12.25
33	1035.10	58.27	17.77	11.20
34	1042.39	62.70	16.63	10.80
35	1087.75	54.85	19.83	11.87
36	766.88	52.59	14.58	10.06
37	869.64	65.50	13.28	9.56
38	938.48	90.80	10.34	8.32

39	996.23	45.38	21.95	12.53
40	1118.21	50.29	22.23	12.61
41	837.16	66.30	12.63	9.30
42	829.76	76.20	10.89	8.57
43	986.25	64.95	15.18	10.29
44	879.94	53.81	16.35	10.71
45	1027.87	67.36	15.26	10.32
46	933.83	64.66	14.44	10.01
47	937.87	60.63	15.47	10.39
48	895.74	34.58	25.90	13.67
49	941.85	51.57	18.27	11.36
50	1047.51	58.44	17.92	11.25
51	1036.51	51.61	20.08	11.95
52	786.14	31.46	24.99	13.41
53	1017.00	63.97	15.90	10.55
54	831.70	68.50	12.14	9.10
55	1191.09	67.00	17.78	11.20
56	1074.64	65.76	16.34	10.70
57	1121.85	48.01	23.37	12.95
58	1150.27	54.95	20.93	12.22
59	933.38	70.04	13.33	9.58
60	974.46	66.27	14.71	10.11
61	1075.66	55.93	19.23	11.68
62	1117.89	49.18	22.73	12.76
63	1042.13	59.55	17.50	11.11
64	1035.22	61.84	16.74	10.84
65	910.01	68.21	13.34	9.59
66	835.18	72.70	11.49	8.83
67	911.81	64.95	14.04	9.86
68	1112.20	46.09	24.13	13.17
69	1017.76	44.80	22.72	12.76
70	1010.10	47.45	21.29	12.33
71	788.17	88.46	8.91	7.65
72	1121.48	50.84	22.06	12.56
73	822.27	88.47	9.29	7.83
74	1058.86	45.56	23.24	12.91
75	1160.77	45.20	25.68	13.61
76	912.22	36.62	24.91	13.39
77	518.14	89.03	5.82	5.93
78	924.26	67.45	13.70	9.73
79	1116.17	45.15	24.72	13.34
80	772.89	56.26	13.74	9.74
81	1031.57	48.47	21.28	12.32
82	1113.11	47.61	23.38	12.95
83	1044.32	45.53	22.94	12.82
84	847.96	57.61	14.72	10.12

85	648.44	48.53	13.36	9.60
86	1027.44	55.03	18.67	11.50
87	932.90	74.34	12.55	9.27
88	1077.64	52.23	20.63	12.12
89	939.70	62.56	15.02	10.23
90	1045.22	58.81	17.77	11.20
91	1145.60	46.08	24.86	13.38
92	730.54	71.53	10.21	8.26
93	876.15	71.77	12.21	9.13
94	1093.13	62.39	17.52	11.11
95	900.70	101.33	8.89	7.64
96	1111.95	46.73	23.80	13.07
97	939.81	68.15	13.79	9.76
98	960.11	57.79	16.61	10.80
99	1045.77	53.07	19.71	11.83
100	973.97	70.46	13.82	9.78
101	953.10	63.07	15.11	10.26
102	1059.76	57.65	18.38	11.40
103	887.04	93.00	9.54	7.95
104	1033.16	57.91	17.84	11.22
105	659.07	59.53	11.07	8.65
106	785.06	65.09	12.06	9.07
107	1107.26	60.09	18.43	11.42
108	968.44	57.06	16.97	10.93
109	856.40	31.59	27.11	14.00
110	1169.81	45.94	25.46	13.55
111	820.20	69.77	11.76	8.94
112	1128.05	55.33	20.39	12.05
Average	**935.52**	**58.63**	**17.84**	**11.08**
Standard deviation	**176.12**	**13.73**	**5.21**	**1.79**

Appendix VII: Representative T1 and T2 decay plots for IMP1 RRM12. Peak # corresponds to T1 and T2 values above



Appendix VIII: Representative T1 and T2 decay plots for IMP3 RRM12. Peak # corresponds to T1 and T2 values above