This thesis is submitted to University College London
for the degree of Doctor of Philosophy

September 2014

Target recognition by multi-domain RNA-binding proteins

Katherine Margaret Collins

Andres Ramos group
Division of Molecular Structure
MRC – National Institute for Medical Research
The Ridgeway, Mill Hill
London, NW7 1AA

Declaration

I, Katherine Collins, 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

Multi-functional RNA binding proteins regulate and coordinate the many steps of RNA metabolism. Accurate functioning of these processes is vital in cells and misregulation has been linked to many human diseases. RNA binding proteins contain multiple RNA binding domains. The ability to perform multiple functions depends on the recognition of a diverse range of targets and domains are used combinatorially to achieve this.

In this thesis I ask how the sequence specificity of low affinity RNA-binding domains and the interplay between said domains plays a role in RNA target selectivity. Within this question I focus on three proteins; TUT4, a uridyl transferase involved in the regulation of both non-coding RNAs and histone mRNA; FMRP, a translational repressor whose loss in cells is the cause of Fragile X Syndrome; and RBM10 a regulator of alternative splicing and miRNA biogenesis. I found that through the use of separate RNA binding domains both TUT4 and RBM10 are able to exert flexibility in target recognition; TUT4 by using two CCHC-type zinc fingers, working independently to recognise short RNA stretches; and RBM10 by using different subsets of domains to recognise either specific high affinity splice site sequences or pre-miRNAs. In FMRP the determination of the sequence specificity of KH1 allowed us to isolate its contribution to target selection.

In a secondary objective, looking at methodologies used in RNA-protein interaction, SIA was improved to make it both less laborious and to reduce the sample requirements, and with FMRP a novel mutational strategy was used in combination with SIA to determine the sequence specificity of this low affinity domain.

In summary these data extend our understanding of the RNA binding mechanisms of the three systems studied and introduces improved or novel methodologies to the future study of protein-RNA interactions.

Acknowledgements

I would like to thank my supervisor Andres Ramos for the opportunity to work on this project and who along with all the other members of the lab, past and present, 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 NMR centre Geoff Kelly, Alain Oregioni and Tom Frenkiel, for all their help with NMR, Steve Martin for the work with circular dichroism and Vangelis Christodoulou for help with the many, many protein expression trials.

Thank you to my thesis committee, Paul Driscoll, Steve Martin and Ian Taylor, for their guidance and input into the project.

Finally thank you to my friends and family for their support over the past four years.

Table of Contents

List of Figures

Chapter 1 – Introduction

Chapter 3 – Protein-RNA specificity by high-throughput principal component analysis of NMR spectra

Chapter 4 – Terminal Uridyl Transferase 4 (TUT4)

Chapter 5 – Fragile X Mental Retardation Protein

Chapter 6 – RNA Binding Motif Protein 10

List of Tables

Chapter 1 – Introduction

Chapter 3 – Protein-RNA specificity by high-throughput principal component analysis of NMR spectra

Chapter 4 – Terminal Uridyl Transferase 4

Chapter 5 – Fragile X Mental Retardation Protein

Chapter 6 – RNA-binding protein 10

Abbreviations

4E-BP	4E binding protein
ADARs	adenosine deaminases that act on RNA
ALS	amyotrophic lateral sclerosis
AMPAR	α-amino-3-hydroxyl-4-isoxazole propionic acid receptor
APRA	antibody-positioned RNA amplification
AREs	AU-rich elements
ASF/SF2	serine/arginine splicing factor 1
BC1	brain cytoplasmic 1
Bcl-x	Bcl-2 like protein 1
BLACP	bladder cancer-associated protein
CamKII α	calcium/calmodulin-dependent protein kinase II α
CD	circular dichroism
CF1A	cleavage factor 1A
CLIP	cross-linking immuno precipitation
CYFIP1	cytoplasmic FMR1-interacting protein 1
DMPK	myotonin-protein kinase
dsRBD	double stranded RNA binding domain
dsRBD	double stranded RNA binding domain
EGF	endothelial growth factor
eIF4E	eukaryotic translation initiation factor 4E
elF1A	eukaryotic translation initiation factor 1A
EMSA	electrophoretic mobility shift assay
FMR1	fragile X mental retardation gene1
FMRP	fragile X mental retardation protein
FXS	fragile X syndrome
GABA	gamma-aminobutyric acid receptor
hASH1	human achaete-scute homologue 1
HIF-α	hypoxia-inducible factor 1-alpha
HITS-CLIP	high-throughput sequencing CLIP
HMQC	heteronuclear multiple-quantum correlation
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
IMAC	immobilized metal ion affinity chromatography
IRES	internal ribosomal entry site
ITC	isothermal titration calorimetry
KH	K-homology
KSRP	KH type-splicing regulatory protein
L-CPL	left-handed circularly polarised light
LamB1	laminin subunit beta-1
LTD	long term depression
LTP	long term potentiation
MAP1b	microtubule-associated protein 1B
MBNL1	muscleblind-like protein 1
mGluR	metabotropic glutamate receptor-dependent
MMP-9	matrix metalloproteinase-9
mTOR	mammalian target of rampamycin
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
NOESY	nuclear Overhauser effect spectroscopy
NOVA1 and 2	neuro-oncological ventral antigen 1 and 2
NPC	nuclear pore complex
PABPN1	polyadenylate-binding nuclear protein 1
PAR-CLIP	photoactivatable ribonucleoside CLIP
PCA	Principal component analysis
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
R-CPL	right-handed circularly polarised light
RBDs	RNA binding domains
RBPs	RNA binding proteins

RIP-Chip	RNP immunoprecipitation-microarray
RISC	RNA-induced silencing complex
RNP1	RNP motif 1
RNP2	RNP motif 2
RNPs	ribonucleoprotein particles
RRM	RNA-recognition motif
Sema3F	semaphorin3F
SIA	scaffold independent analysis
SLM2	Sam68-like mammalian protein 2
Sod1	superoxide dismutase 1
SoSLIP	Sod1 stem loops interacting with FMRP
TDP-43	TAR DNA binding protein 43
TEV	tobacco etch virus
TROSY	transverse relaxation optimised spectroscopy
TUT4	terminal uridyl transferase 4
U2A’	U2 small nuclear ribonucleoprotein A’
U2B”	U2 small nuclear ribonucleoprotein B”
UTRs	untranslated regions
Zcchc11	zinc-finger, CCHC domain-containing protein 11
ZnF/ZF	zinc fingers

1. Introduction

1.1 RNA metabolism

In prokaryotes transcription and translation occur in the same compartment of the cell and are physically coupled. In eukaryotes these processes occur in separate compartments and this has allowed for the evolution of a more complex system of regulation. Post translational processes allow the organism to create extra diversity in the set of proteins encoded in the genome and for an additional layer of gene regulation, creating a flexible and rapid way to control local protein expression in response to the changing requirements of the cell.[1]

After being transcribed the nascent mRNA or precursor mRNA (pre-mRNA) undergoes several processing steps and is then, in most cases, exported from the nucleus to be translated, stored or degraded. These steps are mediated by RNA binding proteins (RBPs) and by a subset of trans-acting RNAs (Figure 1.1).

The RBPs and their target RNAs form complexes called ribonucleoprotein particles (RNPs). These protein-RNA complexes are highly dynamic with components associating and disassociating at various stages.[2] The variation in the distinct set of associated RBPs at different time points and locations in the cell allows the fate of the RNA to be temporally and spatially regulated.

The process of mRNA maturation begins as the RNA is being transcribed by RNA polymerase with associated processing enzymes beginning to work on the nascent molecule. A 7-methylguanosine cap is added to the 5' end which is essential for recognition by the ribosome in the initiation of translation. Poly(A) polymerase adds a poly(A) tail of around 200 nucleotides to the 3' end which affects nuclear transport, translation efficiency and stability. The transcript undergoes splicing to remove introns and may also undergo alternative splicing to create different isoforms of a protein. The pre-mRNA may also be subject to RNA modifications including insertions, deletions and deamination.[3]

Once completely processed the mature mRNA is transported through the nuclear pore complex (NPC). This involves numerous adaptor and receptor proteins that target the RNA

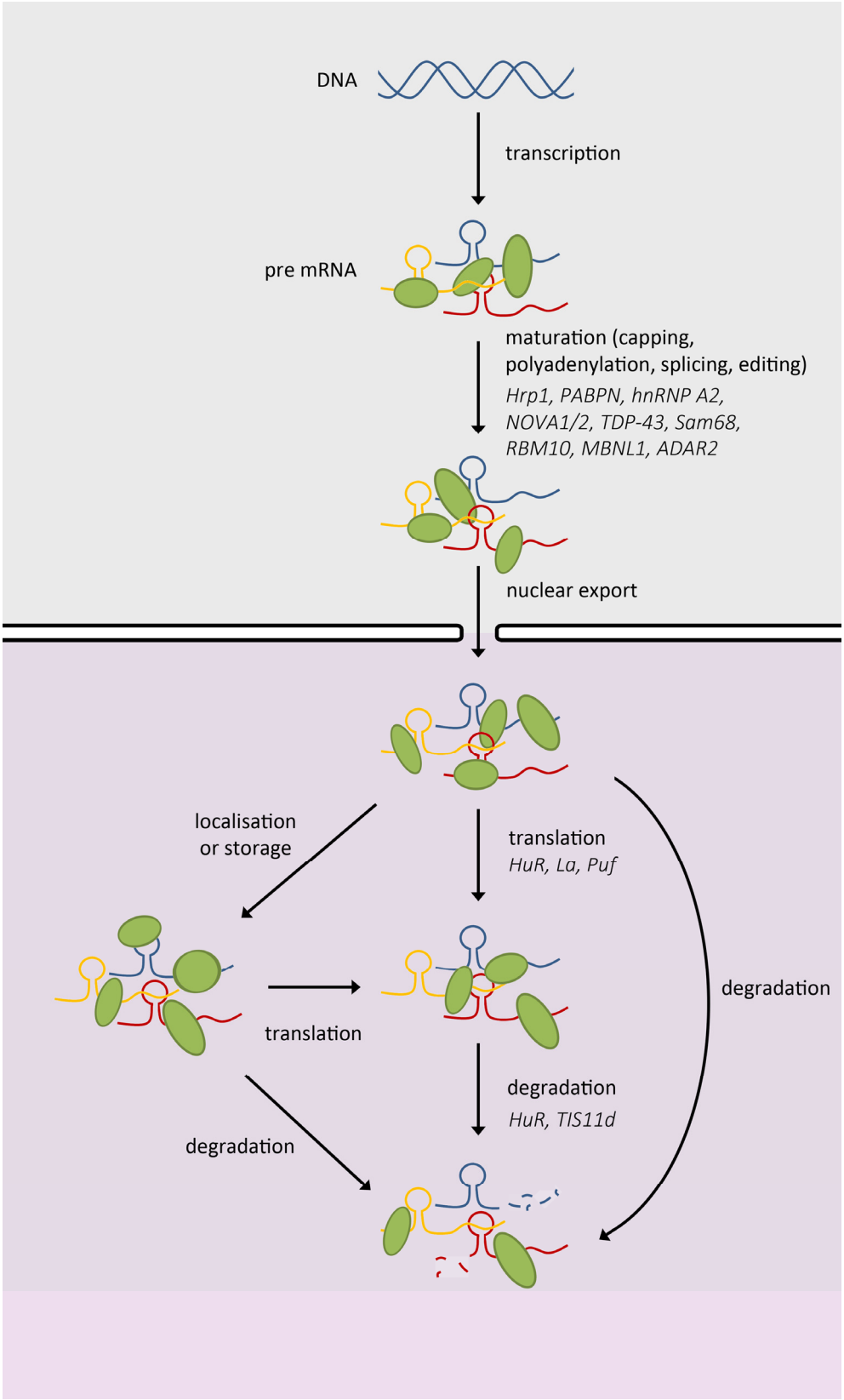


Figure 1.1 | RNA metabolism. Proteins referred to in the text are shown in italics underneath the processes they are involved in.

to the NPC and enable its passage through to the cytoplasm. During or shortly after the translocation process many of the nuclear proteins bound to the mRNA are shed allowing space for the binding of cytoplasmic proteins.[4]

In the cytoplasm the mRNA can perform its role as a template for translation. Various initiation factors bind to the 5' cap and recruit the ribosome. As the ribosome moves along the RNA many of the remaining nuclear proteins are removed. However it may be the case that a cell needs to spatially or temporally regulate expression of a particular mRNA, therefore an mRNA may be held in a translationally quiescent state until it is delivered to the correct location or it is the correct time to produce the encoded protein.[4]

The amount of protein produced by an mRNA molecule is greatly influenced by its stability as this modulates the length of time the transcript is available to the cell as a template for translation. RBPs are greatly influential in this process with proteins binding to elements present in the mRNA in order to increase or decrease stability. Many of these elements are located in the 3’ or 5’ untranslated regions (UTRs). Furthermore a subset of small noncoding RNAs, siRNA or miRNA, in complex with the RISC complex can target mRNAs in order to supress translation or promote degradation.[5]

With such a complex and interconnected system of regulation involving many factors it is unsurprising that changes are associated with a broad range of diseases. A major step found to be misregulated in disease is splicing. As many transcripts in the brain undergo alternative splicing a large number of defects at this level present with neurological symptoms. In paraneoplastic opsoclonus-myoclonus ataxia (POMA) the proteins neuro-oncological ventral antigen 1 and 2 (NOVA1, NOVA2) are targeted by auto-antibodies and thus become depleted. NOVA1 and 2 regulate the alternative splicing of proteins such as mitogen-activated protein kinase 9, neogenin and gephyrin which are involved in inhibitory synaptic transmission and NOVA1 and 2 depletion leads to defects in this pathway.[6] A common feature of frontotemporal dementia, Alzheimer’s disease and amyotrophic lateral sclerosis (ALS) is the aberrant localisation of splicing factor TAR DNA-binding protein 43 (TDP-43). Normally localised in the nucleus, in the aforementioned diseases it is found in cytoplasmic inclusions with the protein ubiquitinated, cleaved or abnormally phosphorylated.[7] Further mutations in the gene encoding for TDP-43 are also associated with ALS.[8]

Along with neuronal disease aberrant levels of RBPs involved in splicing are also commonly found in cancer. The regulator of alternative splicing, Src-associated in mitosis 68 kDa protein (Sam68), is misregulated in cancer. It is involved in regulating the splicing of CD44, cyclin D1, serine/arginine splicing factor 1 (ASF/SF2) and Bcl-2 like protein 1 (Bcl-x) which are involved in cell cycle progression, splicing and apoptosis.[9–12] Other splicing factors implicating in cancer include the RBM family which will be discussed in more detail in Chapter 6.

Likewise errors in the regulation of mRNA stability have been shown to lead to cancer. For example, Hu-antigen R (HuR), a member of the Hu family of proteins, shows increased levels in a number of cancers.[13–15] This family bind to AU-rich elements (AREs) in the 3' end of specific mRNA transcripts and increase their stability.[16] Target mRNAs include cyclins which promote cell proliferation; endothelial growth factor (EGF) and hypoxia-inducible factor 1-alpha (HIF-α) which increase angiogenesis; and matrix metalloproteinase-9 (MMP-9) which facilitates invasion and metastasis.[17–19] HuR has also been shown to repress the translation of cyclin-dependent kinase inhibitor p27, a factor which prevents cell proliferation.[20] While overexpression of Hu family proteins can lead to cancer, the depletion of the same family of proteins by auto antibodies leads to paraneoplastic encephalomyelitis/sensory neuropathy.[21] This illustrates the fine balance required in the expression levels of the RBPs in order to maintain healthy cells.

The La family is another set of RNA-binding proteins whose misregulation is implicated in cancer. Alongside roles in the maturation and subcellular location of RNA polymerase III transcripts such as tRNAs, these proteins regulate internal ribosomal entry site (IRES)-dependent translation through interaction with the IRES.[22,23] La is overexpressed in many cancers and cancer cell lines and upregulates the translation of target mRNAs including laminin subunit beta-1 (LamB1) which drives invasion, angiogenesis and metastasis; cyclin D1 which promotes cell proliferation; and E3 ubiquitin protein ligase Mdm2 which mediates ubiquitination of the tumour suppressor p53.[24–27]

The loss of function of RBPs, rather than aberrant levels of the protein, can also result in disease. In oculopharyngeal muscular dystrophy an expansion in the tri- $(CGC)_6$ repeat to $(CGC)_{8\text{-}13}$ in the coding region of polyadenylate-binding nuclear protein 1 (PABPN1) results in an extended polyalanine tract in the translated product.[28] This promotes the aggregation

of the protein in the nuclei of skeletal muscle fibres and may lead to the symptoms of the disease which include progressive muscle weakness.[29] Expansions in genes can also lead to the loss of function of RBPs by sequestration. A tri- (CTG) nucleotide expansion within the 3' UTR of myotonin-protein kinase (DMPK) or a tetra- (CCTG) nucleotide expansion within intron 1 of the *ZNF9* gene lead to myotonic dystrophy, which is characterised by a wasting of the muscles. The RBP muscleblind-like protein 1(MBNL1) binds to the repeats in the transcribed mRNA and becomes sequestered and unable to perform its function in the regulation of alternative splicing.[30] Similarly a tri- (CGG) nucleotide expansion of between 50 and 200 repeats in the 5' UTR of the *FMR1* gene results in the transcribed mRNA trapping proteins including MBNL1 and heterogeneous nuclear ribonucleoprotein A2 (hnRNP A2). This creates toxic nuclear foci and progressive neurodegeneration.[31] When the number of repeats in this expansion reaches more than 200 the *FMR1* gene is silenced and subsequent loss of the protein product leads to Fragile X Syndrome (FXS).[32] This will be discussed in more detail in Chapter 5.

As previously mentioned siRNAs and miRNAs play a role in the regulation of mRNA stability. Before these molecules can perform their functions they also have to undergo a maturation process, which like for mRNAs is controlled by RBPs. Misregulation of their biogenesis leads to altered levels of mature siRNA and miRNA.[33] This in turn impacts on the levels of their mRNA targets which can lead to disease.[34] Regulation of miRNA biogenesis will be discussed further in the context of Lin28, terminal uridyl transferase 4 (TUT4) and let-7 in Chapter 4 and RBM10 in Chapter 6. Other non-coding RNAs (lncRNAs, piRNAs) also play a role in gene regulation, but will not be reviewed in the context of this thesis.

It is well established that RBPs are essential to the proper functioning of RNA metabolism and so the functioning organism as a whole and that changes to the normal levels or functioning of RBPs can lead to disease states. In order to understand how RBPs perform their functions we must look at the interactions between the proteins and their RNA targets.

1.2 Combinatorial RNA binding

RBPs must be able to recognise a strikingly diverse range of targets, both in sequence and in structure. However one does not find a similarly diverse repertoire of domains capable of recognising RNA, but rather a limited set of basic modular units. RNA binding proteins

often contain multiple copies of these RNA binding domains (RBDs) in various combinations and structural arrangements with target recognition being based on these domain combinations and the positioning of these units within the protein. The benefits of a modular interaction are increased versatility for the proteins in target recognition, and in the assembly and disassembly of the complexes. Each RBD oftenonly recognises a few bases with low-to-intermediate affinity but by using several domains the proteins can build up larger interaction surfaces that recognise RNA targets with higher specificity and affinity. However there are examples of RBDs which bind with high affinity and/or specificity such as the RNA recognition motif (RRM) in Fox-1 and U1A which bind their target sequences with nanomolar affinities.[35,36] Further, individual domains can be augmented by the addition of secondary structural elements or loops.[1,37,38] These additions can be used to extend the nucleic acid binding site and specificity of the domain, to regulate the access to this site and to create additional surfaces to be used in protein-protein interactions in order to add to the functional versatility of the protein.[39–41] Additionally, the modular structure allows catalytic domains to be tethered to RBDs in order to specifically target the catalytic activity of a protein.[37]

The potential of a modular interaction to expand specificity and affinity is exemplified by the PUF family of proteins. These proteins contain an RNA binding region comprised of 8 repeats of a small α-helical structural motif and N- and C-terminal flanking regions, known as the Pumilio homology domain. Each repeat recognises just one nucleotide but the combination of several repeats allows the protein to recognise a sequence of up to 8 nucleotides with high affinity: the *Drosophila* Puf family member Pum binds to its physiological target, the Nanos regulatory element, with a Kd of 0.5nM.[42,43]

Larger units are also involved in combinatorial interactions and RRMs and KH domains are often found in multiple copies allowing various modes of tandem RNA binding (Figure 1.2). These copies are separated by linkers of varying lengths which can play an important role in determining the RNA binding characteristics of a protein. Short linkers may be flexible in the free form, allowing the domains to function as individual units. Upon interaction with RNA these linkers can become structured and help to orient the domains to form an extended and rigid binding surface. Examples of such a conformational transition are observed for proteins including nuclear polyadenylated RNA-binding protein 4 (Hrp1), Sex-lethal and Hu-antigen D (HuD).[44–46] For example, Hrp1 recognises the polyadenylation

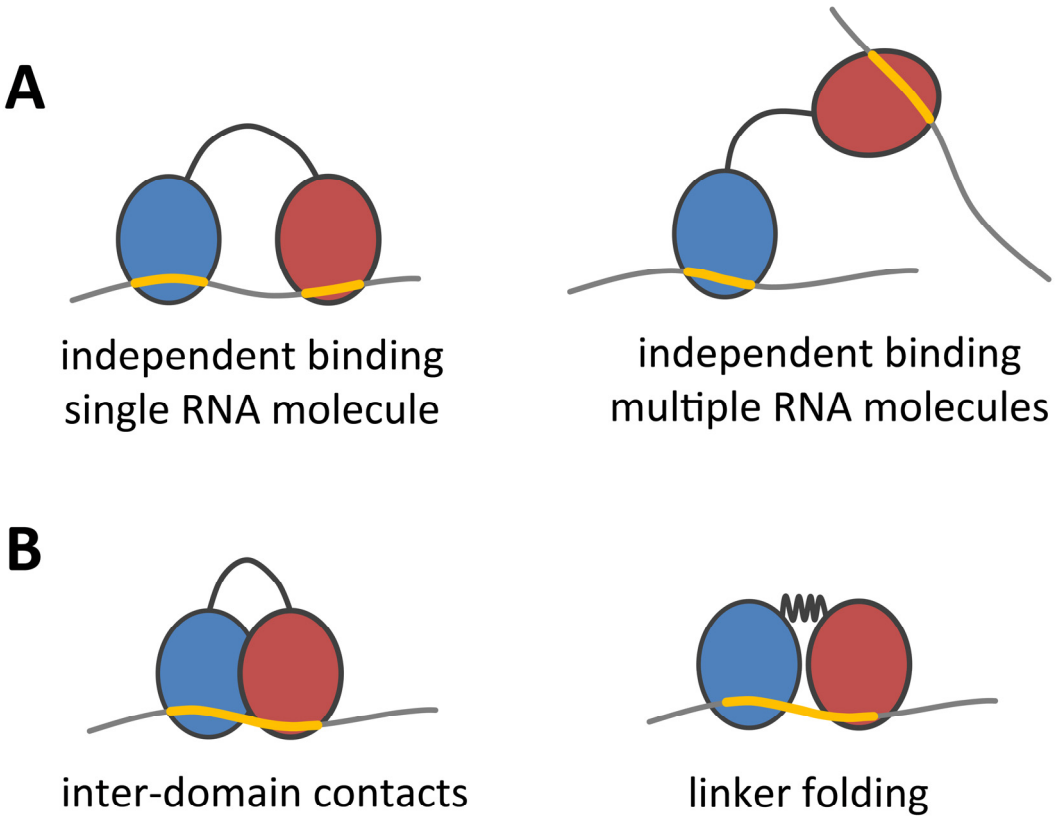


Figure 1.2 | Modes of RNA binding used by tandem RNA binding domains. A) RBDs can bind as individual units to recognise separate RNA sequences on a single RNA molecule or multiple RNA molecules. B) RBDs can also recognise adjacent RNA sequences, here inter-domain contacts or the formation of structured elements in the linker are often observed. Adapted from [38]

enhancement element in the 3' untranslated region of pre-mRNA and promotes recruitment of other polyadenylation factors. The protein uses two tandem RRMs to recognise AU repeats in the RNA with nM affinity. In the free form the RRMs act as two independent bodies but upon addition of RNA residues in the nine amino acid long linker forms a short two turn α helix. The bound RNA stretches across the two β sheets and also makes contacts in the positively charged cleft created between the two domains.[45] Alternatively a short linker can already be structured in the free form thus orienting the two joining domains in the correct position to recognise their RNA target. The TIS11d protein binds to AU-rich elements in target mRNAs and promotes their deadenylation and degradation. The protein contains two CCCH-type zinc fingers separated by an 18 residue linker which forms an extended structure which is relatively rigid in both the free and bound forms.[47]

Longer linkers are unlikely to be structured and therefore introduce an element of flexibility in how the recognised sequences are positioned within the RNA molecule. Such linkers allow the domains to bind separate sequences on the same RNA or even different RNA molecules. This enables the protein to recognise a more diverse range of targets which is particularly important as many RBPs have been shown to be multifunctional with these separate functions depending on the recognition of different targets. Adenosine deaminases that act on RNA (ADARs) selectively modify adenosines to inosines in RNA transcripts. Transcripts edited by ADARs include bladder cancer-associated protein (BLACP), neurotransmitter receptors for glutamate and serotonin and a subunit of the gamma-aminobutyric acid receptor (GABA).[48–50] ADAR2 contains two dsRBM domains linked by an 84 residue stretch found to be unstructured in both free and bound forms. The dsRBM domains work independently and recognise specific secondary elements in the RNA targets thus directing the catalytic activity of the protein.[51] The independence of the RBDs gives flexibility in the distance between the preferred binding sites in the RNA targets.

It has been observed that the affinity of RNA binding proteins containing domains separated by linkers is highly dependent on the length and structure of the amino acid stretch between the domains in question. In a protein where interdomain interactions are observed, indicating the two domains come together to create one RNA binding surface, the expected affinity of binding would be the product of the affinities of the two individual domains. At the other extreme, where two domains are separated by a long flexible linker

and function as individual units the expected affinity of binding would be the sum of the affinities of the individual domains.[52] In many cases a situation between these two extremes is observed. Shamoo et al. produced a model to determine hypothetical affinities for RBDs separated by a linker. In this model the linker restricts the spherical space in which the second domain can move thus increasing the apparent local concentration of the domain and facilitating its binding. They predicted that a linker of just 10 residues results in affinities 350-fold less than the product of the affinities for the individual domains and that to fully segregate the two affinities there must be a linker of 60 residues of more.[52] An example of this is observed in hnRNP A1 which contains two RRMs separated by a 17 amino acid linker. The Kd for the first and second domains individually are 36uM and 2.6uM respectively, while the affinity for RRM1-RRM2 when binding together is 0.11uM. This value is 24-fold higher than RRM2 alone and 325-fold higher than the individual RRM1 but 1000-fold less than the product of the two affinities, which would be expected if the domains interacted with each other.[53] Linker length not only has an effect on the affinity of an RNA binding protein but also on whether separate domains bind in a *cis* or *trans* mode. Once the first domain is bound a shorter linker favours *cis* binding as the second domain is restricted in space and more likely to encounter its preferred binding site on the same RNA molecule. Longer linkers promote *trans* binding as the second domain is just as likely to encounter its preferred binding sequence on a totally separate RNA molecule.[53] 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.

1.3 Types of RNA binding and RNA binding domains

RBDs can recognise both the sequence and/or the structure of RNA. Sequence recognition requires access of the protein moieties to the nucleobases, which can occur in RNA single stranded regions and internal loops and bulges and involves formation of hydrogen bonding, electrostatic interactions, hydrophobic interactions and aromatic stacking interactions between the base and the protein. In addition to these sequence specific contacts the protein also normally interacts with RNA groups in the sugar phosphate backbone. Recognition of RNA structural elements can be performed by individual or multiple RBDs and the recognition of a non-canonical RNA structure is normally associated with a higher binding affinity than sequence specific recognition of a single stranded nucleic

acid or of the double stranded RNA helix by the double stranded RNA binding domain (dsRBD).

Commonly occurring RBDs include RNA-recognition motif (RRM), K-homology (KH) domain, double stranded RNA binding domain (dsRBD), zinc fingers (ZnF) of varying types, PAZ and PIWI domains, among many. Each has a different structure and way of recognising RNA (Table 1.1). Below I discuss in more depth the types of domain which I have studied during the course of my thesis.

1.3.1 RRM

RRM domains interact with single stranded nucleic acids and are also able to mediate protein-protein interactions. They are one of the most abundant protein domains in eukaryotes and are associated with many functions in the cell. The canonical domain is normally between 80 and 90 amino acids in size and arranged in a four stranded anti-parallel β sheet with two α helices packed onto the surface. Each RRM can recognise between four and six nucleotides. The flat β sheet of the RRM domain and the additional elements in the connecting loops and N- and C-termini allow domains of the RRM family to recognise a large number of different RNA sequences and shapes. And while the core fold of the domain is highly conserved, many RRM domains show expansions that include helical and β strand elements. These expansions have defined a number of RRM domain sub-families with non-canonical nucleic acid recognition properties (e.g. UHM domains, quasiRRM and pseudoRRM domains etc.). [54–56]

In the canonical RRM-nucleic acid interaction contacts are made using several conserved amino acids. In the central strands of the β sheet are two conserved motifs, RNP motif 1 (RNP1) with the sequence (R/K)-G-(F/Y)-(G/A)-(F/Y)-(I/L/V)-X-(F/Y) and RNP motif 2 (RNP2) with the sequence (I/L/V)-(F/Y)-(I/L/V)-X-N-L. In 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 and position 3 of RNP1 are generally occupied respectively by a conserved positively charged residue which can interact with the negatively charged phosphate group on the backbone of the nucleic acid and an aromatic residue which interacts hydrophobically with the sugar rings of the stacked bases (Figure 1.3a). The full set of these

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	U1AN-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	Nova-1 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	Fingers 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	Fingers 1 and 2 of TIS11d (1RGO)
Pumilio	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	dsRBD3 from Staufen (1EKZ)

Table 1.1 | Table of commonly occurring RNA binding domains and their properties. Taken from [37]

characteristic contacts is not always observed in RRM-nucleic acid binding and it is more common only to observe one to three of these defined interactions.[54] Specificity is mostly achieved by non-conserved residues in the domain. Typically these interactions involve residues at the top of the β sheet such as the side chains of the amino acids in RNP1 position 7 and two adjacent positions in β strand 1, and the residues C-terminal to β strand 4. The residues are able to make base-specific hydrogen bonds in order to confer specificity. Some RRMs use additional interactions between the two external β strands, loops 1, 3 and 5 and the C- and N-termini to increase the RNA protein interaction network and allow higher affinity binding to be achieved (Figure 1.3a).[54]

There are many examples of proteins containing tandem RRM domains. This enables a higher affinity and/or sequence specificity of binding to be achieved. In the previously discussed example of Hrp1, and in other proteins such as nucleolin and PABP, the two adjacent RRM domains bind synergistically to a continuous stretch of RNA and the interdomain linker may also play a role in recognition of the RNA.[45,57,58] In polypyrimidine tract-binding protein (PTB) and hnRNP A1 tandem RRMs interact in the free form and position the domains such that they are unable to bind a continuous RNA sequence.[39,59] In this case it could be that binding of the domains forces the formation of an RNA loop.

1.3.2 KH domain

KH domains are single stranded, sequence specific nucleic acid binding domains. They are found in archaea, bacteria and eukaryotes and partake in a wide variety of cellular functions including transcriptional and translational regulation. In both bacterial and eukaryotic canonical KH domains, three alpha helices stack on one side of a three-stranded 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 1.3b and d).[60] All KH domains that are known to bind nucleic acids (but not all KH domains) have 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. The length of the variable loop is most often of a few amino acids, but can reach up to 60 residues.[61]

In all known structures of KH-nucleic acid complexes an RNA binding cleft is formed with the GXXG loop and its flanking helices on one side and the β2 strand and variable loop opposing. The nucleic acid molecule is bound in an extended conformation within this

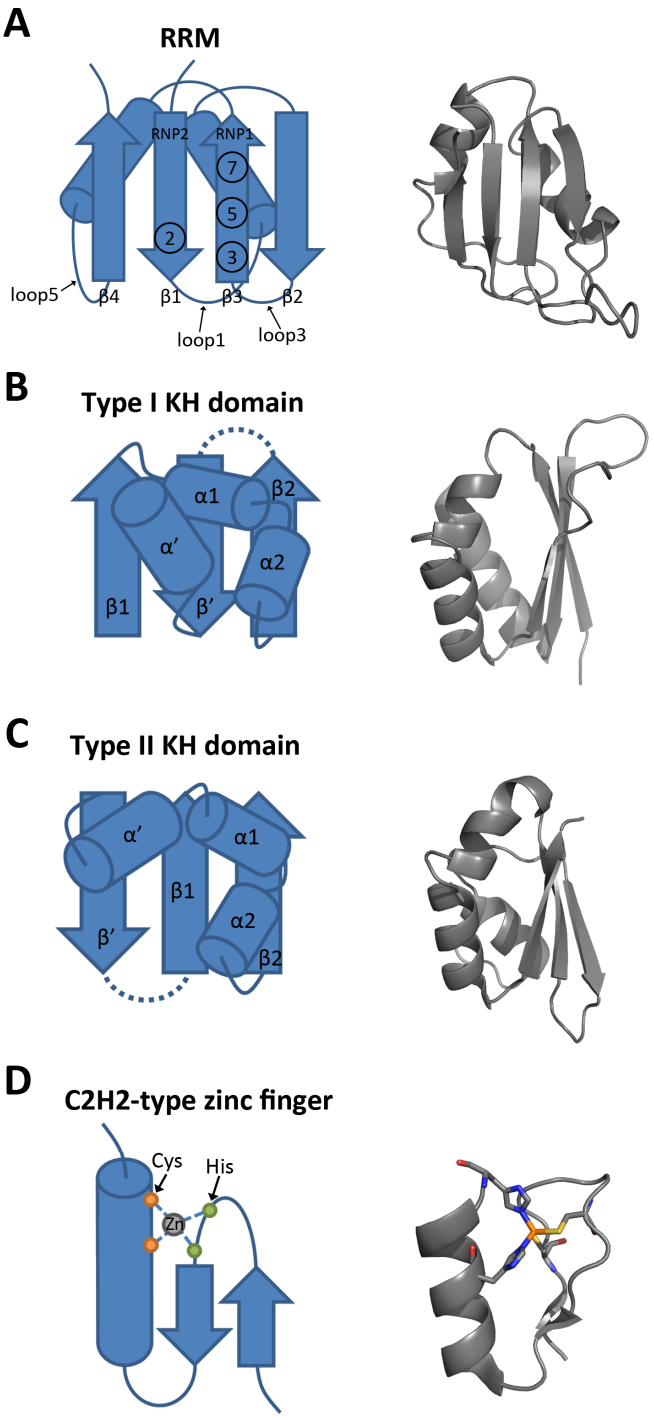


Figure 1.3 | Schematic representations and structures of common RNA binding domains. A) Schematic representation of an RRM (left) The β sheet is annotated with Conserved aromatic residue positions in RNP1 and RNP2 (2, 3, and 5) and variant residue often involved in conferring specificity (7). Second RRM of hnRNPA1 (PDB:1UP1) (right) B) Schematic representation of a Type I KH domain (left). Third KH domain of NOVA1 (PDB:1DT4) (right) C) Schematic representation of a Type II KH domain (left). Second KH domain of NusA (PDB:2ASB) (right)D) Schematic representation of a C2H2-type zinc finger (left). Zinc coordinating cysteine and histidine residues are highlighted. Sixth zinc finger of TFIIIA (PDB:2JFJ) (right) Adapted from [54,61]

mainly hydrophobic groove.[61] The backbone of the first and second nucleobase makes contact with residues in the GXXG loop and orients the Watson-Crick edge of the bases of residues 2 and 3 for specific recognition in the groove.[55] KH domains have been shown to recognise up to four nucleotides specifically using a combination of hydrogen bonding, electrostatic interaction and shape complementarity, although often strong nucleobase discrimination is observed only for one or two positions. In contrast to RRMs, individual KH domains are not know to recognize RNA structures and bind to short single stranded nucleic acids with dissociation constants in the micromolar range. As for RRM domains, the canonical KH core can be expanded by additional elements. For example, in the STAR domain a KH domain is flanked by two regions, QUA1 and QUA2, which can contribute to RNA binding. In the case of SF1, QUA2 forms an α helix which contacts RNA and extends the binding site by up to three nucleotides.[62] While KH domains are normally found in multiple copies in RNA binding proteins, STAR domains are only found in one copy per protein. However the QUA1 extension mediates homodimerisation allowing the concurrent recognition of two RNA recognition elements.[63]

As seen for RRM domains, KH domains can also use the repetition of multiple domains to mediate recognition and binding of target RNA. In the example of NusA two KH domains are separated by a short linker with extensive contacts observed between the domains. The domains form one continuous platform to bind a single stretch of RNA with high affinity.[64] Conversely in KSRP the two central KH domains are oriented such that the classical nucleic acid-binding grooves are on opposite sides of the double domain structure and as such bind two separate stretches of RNA.[65]

1.3.3 Zinc Fingers

Zinc fingers are small domains which contain multiple finger-like protrusions. They can be classified into several different families based on their three dimensional structure or the identity of the residues coordinating the zinc ion. Zinc fingers are very versatile motifs and have been shown to bind RNA, DNA and make protein-protein contacts. They are often found in clusters and the amino acid sequence of the domain, the linkers between fingers, the number of fingers and higher order structures contribute to their binding properties.

Of the different families of zinc finger the most common and well-studied is the C2H2-type zinc finger found in many mammalian transcription factors. The domain consists of an anti-

parallel β sheet which contains a loop formed by two zinc coordinating cysteines, and an α helix containing a his-his loop. The zinc ion holds these two structural elements together (Figure 1.3c). While the best known function of these fingers sees them binding to dsDNA numerous examples of RNA binding have also been shown. TFIIIA contains nine C2H2-type zinc fingers and using different subsets of these domains can bind to both 5S rRNA gene and to the RNA product. A crystal structure shows that recognition of 5S rRNA by fingers 4-6 is mostly driven by recognition of the structure of the RNA but accessible individual bases are also recognised.[66] Another example of RNA binding C2H2-type zinc fingers are in the Tra-1 protein in *C. elegans*. Here they mediate binding to the 3'-UTR of *tra-2* mRNA and regulate its export from the nucleus.[67]

Another class of zinc finger known to bind to RNA are the CCCH-type zinc fingers. Examples are found in both Tis11d and MBNL1 and while they both recognise specific RNA stretches both the sequence specificity and mode of interaction differ between the proteins. Tis11d contains two zinc fingers and both recognise a UAU motif using hydrogen bonds between the protein backbone and the Watson-Crick edges of the bases.[47] MBNL1 contains four CCHC-type zinc fingers with ZF3 and ZF4 being shown to recognise GC and GCU sequences respectively. These interactions are mediated by stacking interactions and hydrogen bonds that involve the main chains of amino acid residues.[68]

Other RNA binding zinc fingers include CCHC-type zinc fingers which are found in HIV-1 NCP, Lin28 and TUT4, and RanBP2-type zinc fingers which are found in RBM10, Ewing's Sarcoma and RBP56. These classes of zinc finger will be discussed in more detail in Chapter 4 and Chapter 6 respectively.

1.4 Study of protein-RNA interactions

Although we have a basic molecular understanding of a small repertoire of protein-RNA interactions a deeper knowledge of the range of physiological RBPs and how they recognise and associate with their target RNAs is required. This is important as misregulation of RNA metabolism has been implicated in many diseases. A better understanding of selectivity is necessary to target the interactions for therapeutic use and diagnostic techniques, but requires a better molecular understanding of RNA recognition by RNA-binding domains.

The application of biochemical approaches, such as RNA Immunoprecipitation, UV Cross-Linking Immuno Precipitation (CLIP), and photoactivatable Ribonucleoside (PAR)-CLIP, has allowed us to define the ensemble of RNAs interacting with a specific protein in a cell.[6,69] More recently researchers have pulled down polyadenylated RNAs in HeLa and HEK 293 cells and identified interacting proteins by mass spectrometry in order to generate a global RBP interactome.[70,71] However 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.

The sequence specificity of a domain is important in target selection in the cell as it has been shown that mutation of the sequence of the RNA target often affects protein function. However, the RNA sequence recognised by a single domain is too short to account for the selectivity of many proteins. Most RNA regulators are multi-functional and in proteins which contain numerous RNA binding domains individual domains can perform different roles in the recognition of different targets. For example, the third KH domain of the protein KH type-splicing regulatory protein (KSRP) binds with micromolar Kd to a G-rich sequence in a pre-miRNA target. The same domain interacts with 100-fold lower affinity with an AU-rich sequence present in the AREs in the 3' UTRs of target mRNA. In the first case the domain dominates the recognition of the pre-miRNA. In the second the contribution to the recognition of AU-rich elements is more even between the four KH domains.[72] By differential use of individual domains the same protein may recognise shorter or longer stretches of RNA, different specific sequences or different structures. Similarly, in the La protein the N-terminal La motif and RRM, in a highly cooperative interaction, recognise the 3' poly(U) tail of primary transcripts produced by RNA polymerase III.[73,74] The recognition of another target, the IRES domain IV of the hepatitis C virus RNA, requires both the N-terminal La motif and RRM and the C-terminal RRM. Furthermore this binding appears to recognise structural features of the RNA rather than a specific sequence.[75]

In this context, to rationalise the binding of a protein to diverse sets of targets it is necessary to possess an account of the full RNA binding capability of a domain and not just identify the best binding sequence. It is also necessary to understand how different domains contribute to the binding, both when coming together to recognise a common RNA sequence or structure, and when acting as semi-independent units. This implies evaluating the contribution of the different domains to the binding of the different targets and understanding the role of the linkers as dynamic entities which tether different binding units and can impose structural constraints on the interactions. Therefore we need to evaluate binding affinities, kinetics and specificity of individual domains and of the multi-domain protein overall and obtain structural and dynamic information on the protein-RNA complexes.

NMR is a useful technique as it gives information on the structure and binding properties of protein-RNA complexes as well as providing information on motions, dynamics and multiple conformations. Along with X-ray crystallography, NMR 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. The atomic resolution detail gained from these structures enables the determination of molecular basis of specificity such as the formation of hydrophobic pockets and precise hydrogen bonding events. However, due to the nature of protein-RNA interactions, regions of the protein may be dynamic and several conformations may exist. In X-ray crystallography the protein is viewed in crystalline form and the packing in the crystal may stabilise just one of these conformations. In contrast, solution NMR the complex is free to sample all conformations and can therefore be used to gain information about dynamics occurring in the picosecond to millisecond timescales. Backbone and side chain motions can be monitored by the acquisition of relaxation data, T1, T2 and heteronuclear NOE experiments, giving information on the flexibility and movement of these elements of the RBD. Comparison of these data for both the free and bound protein is particularly useful as it can show structural rearrangements that are often functionally relevant as in many protein-RNA binding events loops and dynamic regions in the core and/or flanking regions can rearrange upon RNA binding. This is seen in QUAKING which contains a STAR domain. The preformed helix, Qua2, which does not interact with the core KH domain in the free form, is locked into the optimal orientation relative to the

KH domain upon RNA binding by the Qua1 extension creating a high affinity interaction.[76] Furthermore, the flexibility of the linkers between RBDs in the bound form gives information on whether adjacent domains are working cooperatively to bind RNA or as individual entities, or even if the linker itself makes contacts with the RNA to extend the interaction surface.

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, whose overall affinities are often in the nanomolar range. However this high specificity interaction often is created by the binding of several individual domains each with lower affinities. Individual domains more often bind with Kds in the micromolar range and in this case NMR can be used to determine information about binding affinities. These relatively weak Kds often fall in the fast exchange regime and interaction dependent changes in chemical shift can be measured. As the position of the peak moving between free and bound is linked to the fraction of bound protein a binding curve can be produced.

Another advantage of NMR is that while other biophysical techniques such as ITC and CD give an overall view of a binding interaction NMR allows changes occurring to individual residues in the protein to be monitored. Changes in the position or line shape in reporter peaks can allow binding sites to be mapped onto proteins. This information is useful in the design of mutants to alter the binding capabilities of a domain. Furthermore many RBPs are known to interact with RNA and other proteins in order to elicit their function. U2 small nuclear ribonucleoprotein B'' (U2B”) can only bind to its target U2 snRNA in the presence of U2 small nuclear ribonucleoprotein A' (U2A’). The U2A’ protein interacts with U2B” on the surface opposite the RNA-binding surface.[77] In the nuclear cap-binding protein complex the RRM domain requires contacts with CBP80 in order to bind the 5’ cap of the RNA.[78] Monitoring interactions by NMR allows the two separate interaction surfaces to be distinguished with each being tracked individually in the same experiment. This allows for the simplification of the system without having to monitor the two binding events in separate experiments, which could lead to loss of important information of cooperativity.

1.5 Aims

The general aim of this thesis has been to analyse the role played by low affinity protein-RNA interactions in the combinatorial recognition of RNA targets by protein regulators. I have focused on three proteins important for RNA metabolism and whose misregulation is a cause of cancer and neurological disease, TUT4, FMRP and RBM10. I have asked how the sequence specificity of their low RNA-binding affinity domains and the inter-domain interactions could play a role in RNA target selectivity. A secondary objective of the thesis was to assess our current methods on challenging cases and develop new strategies as required.

1.6 References

1. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582,** 1977–86 (2008).

2. 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).

3. Lodish, H. *et al. Molecular Cell Biology (5th Edition)*. (2004).

4. Moore, M. J. From birth to death: the complex lives of eukaryotic mRNAs. *Science* **309,** 1514–8 (2005).

5. 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).

6. Ule, J. *et al.* CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302,** 1212–5 (2003).

7. Neumann, M. *et al.* Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science* **314,** 130–3 (2006).

8. Kabashi, E. *et al.* TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. *Nat. Genet.* **40,** 572–4 (2008).

9. Matter, N., Herrlich, P. & König, H. Signal-dependent regulation of splicing via phosphorylation of Sam68. *Nature* **420,** 691–5 (2002).

10. Paronetto, M. P. *et al.* Alternative splicing of the cyclin D1 proto-oncogene is regulated by the RNA-binding protein Sam68. *Cancer Res.* **70,** 229–39 (2010).

11. Valacca, C. *et al.* Sam68 regulates EMT through alternative splicing-activated nonsense-mediated mRNA decay of the SF2/ASF proto-oncogene. *J. Cell Biol.* **191,** 87–99 (2010).

12. 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,** 929–39 (2007).

13. Bolognani, F., Gallani, A.-I., Sokol, L., Baskin, D. S. & Meisner-Kober, N. mRNA stability alterations mediated by HuR are necessary to sustain the fast growth of glioma cells. *J. Neurooncol.* **106,** 531–42 (2012).

14. Wang, J., Wang, B., Bi, J. & Zhang, C. Cytoplasmic HuR expression correlates with angiogenesis, lymphangiogenesis, and poor outcome in lung cancer. *Med. Oncol.* **28 Suppl 1,** S577–85 (2011).

15. Kakuguchi, W. *et al.* HuR knockdown changes the oncogenic potential of oral cancer cells. *Mol. Cancer Res.* **8,** 520–8 (2010).

16. Brennan, C. M. & Steitz, J. A. HuR and mRNA stability. *Cell. Mol. Life Sci.* **58,** 266–77 (2001).

17. 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,** 2340–50 (2000).

18. Sheflin, L. G., Zou, A.-P. & Spaulding, S. W. Androgens regulate the binding of endogenous HuR to the AU-rich 3'UTRs of HIF-1alpha and EGF mRNA. *Biochem. Biophys. Res. Commun.* **322,** 644–51 (2004).

19. Akool, E. *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,** 4901–16 (2003).

20. Kullmann, M., Göpfert, U., Siewe, B. & Hengst, L. ELAV/Hu proteins inhibit p27 translation via an IRES element in the p27 5'UTR. *Genes Dev.* **16,** 3087–99 (2002).

21. Bolognani, F. & Perrone-Bizzozero, N. I. RNA-protein interactions and control of mRNA stability in neurons. *J. Neurosci. Res.* **86,** 481–9 (2008).

22. Wolin, S. L. & Cedervall, T. The La protein. *Annu. Rev. Biochem.* **71,** 375–403 (2002).

23. Ali, N. & Siddiqui, A. The La antigen binds 5' noncoding region of the hepatitis C virus RNA in the context of the initiator AUG codon and stimulates internal ribosome entry site-mediated translation. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 2249–54 (1997).

24. Sommer, G., Rossa, C., Chi, A. C., Neville, B. W. & Heise, T. Implication of RNA-binding protein La in proliferation, migration and invasion of lymph node-metastasized hypopharyngeal SCC cells. *PLoS One* **6,** e25402 (2011).

25. Petz, M., Them, N., Huber, H., Beug, H. & Mikulits, W. La enhances IRES-mediated translation of laminin B1 during malignant epithelial to mesenchymal transition. *Nucleic Acids Res.* **40,** 290–302 (2012).

26. Sommer, G. *et al.* The RNA-binding protein La contributes to cell proliferation and CCND1 expression. *Oncogene* **30,** 434–44 (2011).

27. Trotta, R. *et al.* BCR/ABL activates mdm2 mRNA translation via the La antigen. *Cancer Cell* **3,** 145–60 (2003).

28. Brais, B. *et al.* Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nat. Genet.* **18,** 164–7 (1998).

29. Calado, A. *et al.* Nuclear inclusions in oculopharyngeal muscular dystrophy consist of poly(A) binding protein 2 aggregates which sequester poly(A) RNA. *Hum. Mol. Genet.* **9,** 2321–8 (2000).

30. Cho, D. H. & Tapscott, S. J. Myotonic dystrophy: emerging mechanisms for DM1 and DM2. *Biochim. Biophys. Acta* **1772,** 195–204 (2007).

31. Iwahashi, C. K. *et al.* Protein composition of the intranuclear inclusions of FXTAS. *Brain* **129,** 256–71 (2006).

32. Fu, Y. H. *et al.* Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* **67,** 1047–58 (1991).

33. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116,** 281–97 (2004).

34. Kloosterman, W. P. & Plasterk, R. H. a. The diverse functions of microRNAs in animal development and disease. *Dev. Cell* **11,** 441–50 (2006).

35. Auweter, S. D. *et al.* Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. *EMBO J.* **25,** 163–73 (2006).

36. Oubridge, C., Ito, N., Evans, P. R., Teo, C. H. & Nagai, K. Crystal structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature* **372,** 432–8 (1994).

37. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8,** 479–90 (2007).

38. Mackereth, C. D. & Sattler, M. Dynamics in multi-domain protein recognition of RNA. *Curr. Opin. Struct. Biol.* **22,** 287–96 (2012).

39. Oberstrass, F. C. *et al.* Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science* **309,** 2054–7 (2005).

40. Mazza, C., Segref, A., Mattaj, I. W. & Cusack, S. Large-scale induced fit recognition of an m(7)GpppG cap analogue by the human nuclear cap-binding complex. *EMBO J.* **21,** 5548–57 (2002).

41. Kielkopf, C. L., Lücke, S. & Green, M. R. U2AF homology motifs: protein recognition in the RRM world. *Genes Dev.* **18,** 1513–26 (2004).

42. Wang, X., McLachlan, J., Zamore, P. D. & Hall, T. M. T. Modular recognition of RNA by a human pumilio-homology domain. *Cell* **110,** 501–12 (2002).

43. 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).

44. Handa, N. *et al.* Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein. *Nature* **398,** 579–85 (1999).

45. Pérez-Cañadillas, J. M. Grabbing the message: structural basis of mRNA 3'UTR recognition by Hrp1. *EMBO J.* **25,** 3167–78 (2006).

46. Wang, X. & Tanaka Hall, T. M. Structural basis for recognition of AU-rich element RNA by the HuD protein. *Nat. Struct. Biol.* **8,** 141–5 (2001).

47. Hudson, B. P., Martinez-Yamout, M. a, Dyson, H. J. & Wright, P. E. Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nat. Struct. Mol. Biol.* **11,** 257–64 (2004).

48. Galeano, F. *et al.* Human BLCAP transcript: new editing events in normal and cancerous tissues. *Int. J. Cancer* **127,** 127–37 (2010).

49. Seeburg, P. H. & Hartner, J. Regulation of ion channel/neurotransmitter receptor function by RNA editing. *Curr. Opin. Neurobiol.* **13,** 279–283 (2003).

50. Ohlson, J., Pedersen, J. S., Haussler, D. & Ohman, M. Editing modifies the GABA(A) receptor subunit alpha3. *RNA* **13,** 698–703 (2007).

51. 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,** 345–55 (2006).

52. Shamoo, Y., Abdul-Manan, N. & Williams, K. R. Multiple RNA binding domains (RBDs) just don't add up. *Nucleic Acids Res.* **23,** 725–728 (1995).

53. Shamoo, Y. *et al.* Both RNA-binding domains in heterogenous nuclear ribonucleoprotein A1 contribute toward single-stranded-RNA binding. *Biochemistry* **33,** 8272–81 (1994).

54. Maris, C., Dominguez, C. & Allain, F. H.-T. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J.* **272,** 2118–31 (2005).

55. Auweter, S. D., Oberstrass, F. C. & Allain, F. H.-T. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res.* **34,** 4943–59 (2006).

56. Cléry, A., Blatter, M. & Allain, F. H.-T. RNA recognition motifs: boring? Not quite. *Curr. Opin. Struct. Biol.* **18,** 290–8 (2008).

57. Allain, F. H.-T., Bouvet, P., Dieckmann, T. & Feigon, J. Molecular basis of sequence-specific recognition of pre-ribosomal RNA by nucleolin. *EMBO J.* **19,** 6870–81 (2000).

58. Deo, R. C., Bonanno, J. B., Sonenberg, N. & Burley, S. K. Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell* **98,** 835–45 (1999).

59. Barraud, P. & Allain, F. H.-T. Solution structure of the two RNA recognition motifs of hnRNP A1 using segmental isotope labeling: how the relative orientation between RRMs influences the nucleic acid binding topology. *J. Biomol. NMR* **55,** 119–38 (2013).

60. Grishin, N. V. KH domain: one motif, two folds. *Nucleic Acids Res.* **29,** 638–43 (2001).

61. Valverde, R., Edwards, L. & Regan, L. Structure and function of KH domains. *FEBS J.* **275,** 2712–26 (2008).

62. Liu, Z. *et al.* Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science* **294,** 1098–102 (2001).

63. Beuck, C. *et al.* Structure of the GLD-1 homodimerization domain: insights into STAR protein-mediated translational regulation. *Structure* **18,** 377–89 (2010).

64. Beuth, B., Pennell, S., Arnvig, K. B., Martin, S. R. & Taylor, I. a. Structure of a Mycobacterium tuberculosis NusA-RNA complex. *EMBO J.* **24,** 3576–87 (2005).

65. 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,** 5193–205 (2010).

66. Lu, D., Searles, M. A. & Klug, A. Crystal structure of a zinc-finger-RNA complex reveals two modes of molecular recognition. *Nature* **426,** 96–100 (2003).

67. Graves, L. E., Segal, S. & Goodwin, E. B. TRA-1 regulates the cellular distribution of the tra-2 mRNA in C. elegans. *Nature* **399,** 802–5 (1999).

68. Teplova, M. & Patel, D. J. Structural insights into RNA recognition by the alternative-splicing regulator muscleblind-like MBNL1. *Nat. Struct. Mol. Biol.* **15,** 1343–51 (2008).

69. Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141,** 129–41 (2010).

70. Baltz, A. G. *et al.* The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* **46,** 674–90 (2012).

71. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149,** 1393–406 (2012).

72. García-Mayoral, M. F., Díaz-Moreno, I., Hollingworth, D. & Ramos, A. The sequence selectivity of KSRP explains its flexibility in the recognition of the RNA targets. *Nucleic Acids Res.* **36,** 5290–6 (2008).

73. Teplova, M. *et al.* Structural basis for recognition and sequestration of UUU(OH) 3' temini of nascent RNA polymerase III transcripts by La, a rheumatic disease autoantigen. *Mol. Cell* **21,** 75–85 (2006).

74. Kotik-Kogan, O., Valentine, E. R., Sanfelice, D., Conte, M. R. & Curry, S. Structural analysis reveals conformational plasticity in the recognition of RNA 3' ends by the human La protein. *Structure* **16,** 852–62 (2008).

75. Martino, L. *et al.* Analysis of the interaction with the hepatitis C virus mRNA reveals an alternative mode of RNA recognition by the human La protein. *Nucleic Acids Res.* **40,** 1381–94 (2012).

76. Teplova, M. *et al.* Structure-function studies of STAR family Quaking proteins bound to their in vivo RNA target sites. *Genes Dev.* **27,** 928–40 (2013).

77. 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,** 645–50 (1998).

78. Calero, G. *et al.* Structural basis of m7GpppG binding to the nuclear cap-binding protein complex. *Nat. Struct. Biol.* **9,** 912–7 (2002).

2. Materials and Methods

Two of the main biophysical techniques I have used throughout this thesis are NMR and CD. In the course of these studies NMR is used extensively to analyse the folding of constructs, assign constructs to enable the mapping of residues onto an existing structure, determine domain interactions and monitor protein-RNA binding. CD has been used to establish the stability of mutants and characterise protein-RNA interactions. In this chapter I give a brief overview of the theory of both techniques and the aspects I have used in the course of these studies.

2.1 NMR

The nuclei of atoms are made up of neutrons and protons which have an intrinsic property known as spin. Quantum mechanics state that the spin quantum number, I, of a nucleus may be zero or some multiple of ½. In NMR we take advantage of nuclei with I=½ such as ^{1}H, ^{13}C, ^{15}N, ^{19}F and ^{31}P. These particles are restricted to two orientations relative to a fixed direction in space due to the rules of space quantisation which state a nucleus with spin number I can be found in one of (2I+1) orientations. In addition to spin, these nuclei also have a magnetic moment and when exposed to an external magnetic field the two allowed orientations have slightly different energies; a lower energy orientation parallel to the applied field and a higher energy orientation anti-parallel to the field. At equilibrium the states are populated according to the Boltzmann distribution. Given that the energy gap between the two states is small the difference in population is also small, however there is a slightly larger population in the lower energy state. This leads to bulk magnetisation along the axis of the applied magnetic field, the z-axis. The equilibrium populations can be perturbed by irradiation at a frequency equal to the energy difference between the two states. As well as perturbation of the population of the energy states the spins of the nuclei are brought into phase. These effects of the irradiation lead to a bulk magnetization in the xy-plane precessing 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 precess 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 environments 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 and NMR spectrum.

2.1.1 Protein-RNA interactions

When using spectroscopic methods to determine information about binding a central dogma is that there must be an observable change in signal for either the protein or the ligand between the free states and the complex. In NMR observables such as the chemical shift, relaxation rate and scalar coupling may all change upon binding of a ligand. Chemical shifts, which are sensitive to changes in the local magnetic field, generally show large perturbation upon the binding of RNA due to the ring current of the bases and the negative charge of the phosphate group.

^{1}H-^{15}N correlational spectroscopy such as HSQC and HMQC are well suited to the study of protein-RNA interactions as the amide reporter resonances are evenly spread around the protein, with one per residue, and are sensitive to changes in the chemical microenvironment. From changes in the spectra, binding sites can be mapped to the protein surface, dissociation constants and stoichiometries determined and it can be observed if the protein becomes aggregated upon addition of RNA.

The weighted average chemical shift of peaks in the ^{1}H and ^{15}N dimensions upon addition of RNA is a useful parameter and can be calculated using Equation 1:

$$\Delta\delta_{\mathrm{av}} = \sqrt{(\Delta\delta_{\mathrm{H}})^2 + (\Delta\delta_{\mathrm{N}}/10)^2} \qquad (1)$$

where $\Delta\delta_H$ and $\Delta\delta_N$ are the change of chemical shift for a crosspeak in the ^{1}H and ^{15}N dimensions respectively. Where the amide crosspeaks are assigned and a structure of the protein is available, peaks showing large chemical shift perturbation can be mapped onto the surface of the protein to show the binding site. Care must be taken as changes in chemical shift can also be due to structural rearrangement of the protein upon binding and not occur from direct interaction with the ligand.

Kds are important parameters for understanding the function and strength of the underlying physiological interaction. The first step to determine the Kd using NMR is to record a series of spectra in which the protein concentration is kept constant in the presence of increasing amounts of RNA. Changes in chemical shift can be used to determine the dissociation constant of an RNA-protein complex formation in thermal equilibrium as described in more detail in the context of peaks in fast exchange below. Upon complex formation molecules are in a state of exchange between the free and bound state and the NMR signal is sensitive to both the difference of chemical shift values of each of the conformations and the interconversion rates between the conformational states. The change in appearance of peaks and chemical shifts upon addition of RNA depends on the rate of exchange relative to the magnitude of the difference of the chemical shift position between the two states (free and RNA-bound protein). Overall three exchange regimes can be identified; slow, intermediate and fast.

Peaks in fast exchange appear as one signal which changes position progressively as more RNA is added. What this peak actually shows is the weighted average chemical shift of the free and bound fractions. One can titrate RNA into a labelled protein sample and acquire spectra at a range of protein:RNA ratios. The chemical shift changes of the backbone amides, in the case of monitoring by ^{1}H-^{15}N correlational spectroscopy, can then be plotted as a function of protein:RNA ratio to produce a binding curve. As the position of this peak reports on the amount of free and bound protein the binding curve can be used to determine the dissociation constant using Equation 2.[1] A wide range of Kd values, between 10 μM and 10 mM, can be determined this way.

$$\Delta\,\delta_{\,\mathrm{av}} = \Delta\,\delta_{\mathrm{max}}\frac{(\mathrm{K_D} + [\mathrm{L}]_0 + [\mathrm{P}]_0) - \sqrt{(\mathrm{K_D} + [\mathrm{L}]_0 + [\mathrm{P}]_0)^2 - (4[\mathrm{P}]_0[\mathrm{L}]_0)}}{2[\mathrm{P}]_0} \quad (2)$$

Where $\Delta\delta_{av}$ is the average chemical shift perturbation of a given resonance at a given titration point, $\Delta\delta_{max}$ is the chemical shift perturbation for a given resonance at saturation, $[P]_0$ is total protein concentration, $[L]_0$ is total RNA concentration, and K_D is the dissociation constant.

Peaks may be in the slow exchange regime which is characterised by two separate chemical shifts, one for the residue in the free state and one for the residue in the bound state,

whose relative intensities change when the equilibrium is shifted by addition of RNA, but whose chemical shifts remain the same. In theory this change in intensity can be used to determine the dissociation constant but in practice numerous factors prevent this. Signals must be sufficiently resolved to allow their accurate measurements and the signal to noise ratio must be such that measurements can be taken at sub-stoichiometric RNA concentrations, where the intensity of one or both peaks may be quite weak, in order to plot a binding curve. Furthermore to accurately measure dissociation constants the protein concentration must be lower than that of the Kd. Slow exchange indicates a slow rate of dissociation and therefore a Kd in the nM range. In these situations the required protein concentrations are often too low to be measured by NMR, a technique that requires relatively high protein concentrations.

These two examples are the extreme examples of exchange regime and many peaks behaviour would fall somewhere in the middle. For intermediate exchange peaks typically shift from the free to bound state but also experience line broadening in the middle of the shift. Also as the timescale is a relative one, for a given equilibrium different resonances in the protein will show different exchange behaviour depending on how large the difference in chemical shift is between the free and the bound state.

2.1.2 Backbone assignment

Backbone assignment is undertaken as the first step towards structure determination or to be able to map the behaviour of peaks to regions of the protein if the structure is already known. In this process each signal in the spectra is attributed to a specific nucleus in the protein. Sequentially connected residues are identified and matched to unique positions in the primary sequences of the protein. This is done is by recording a set of triple resonance experiments on ^{13}C-^{15}N labelled protein. Magnetisation can be efficiently transferred between the three nuclei (^{1}H, ^{13}C and ^{15}N) through heteronuclear scalar couplings.

Triple resonance experiments are often analysed in pairs, for example HNCA and HN(CO)CA experiments. In the HNCA magnetisation is transferred from the HN_i proton via the N_i atom to the $C\alpha_i$ and $C\alpha_{i-1}$ carbon atoms and the frequencies of all three nuclei are recorded. In the HN(CO)CA the magnetisation path goes from the HN_i proton via the N_i atom to the directly attached $C\alpha_{i-1}$ via the CO_{i-1}. In this experiment the CO atom acts only as a relay nucleus and its frequency is not detected. When analysing the spectra the HNCA would

show signals at a given HN and N frequency of the attached $C\alpha_i$ and $C\alpha_{i-1}$. The HN(CO)CA spectra would show a crosspeak only for the attached $C\alpha_{i-1}$ due to the magnetisation passing through the CO atom. Another set of experiments are HNCACB and HN(CO)CACB. While this pair is lower sensitivity than the HNCA/HN(CO)CA they give the extra information of the Cβ frequency.

Using these experiments the ^{1}H, ^{15}N, $^{13}C\alpha$ and $^{13}C\beta$ of each residue are identified. Sequential assignment is made by matching the $C\alpha_i$ and $C\beta_i$ peaks attached to one amide with the $C\alpha_{i-1}$ and $C\beta_{i-1}$ of another amide thus working along the backbone. Several amino acids give distinctive frequencies for the Cα and Cβ, such as the Cα of glycine and the Cβ of alanine, serine and threonine. In this way sequences can be matched to the primary sequence of the protein.

2.1.3 Relaxation experiments

The relaxation properties of a given nuclei can be used to report on its dynamics. From this one can gain information on the flexibility of regions of the protein. The data is acquired by measuring the T1 and T2 parameters. Relaxation is the process by which the bulk magnetisation, perturbed by the RF pulse, returns to equilibrium over time. T1 is the rate at which longitudinal magnetisation recovers to equilibrium and T2 is the rate at which transverse magnetisation recovers to equilibrium.

Relaxation occurs because of fluctuations in the local magnetic field due to molecular reorientation. In the sample each nuclei has its own magnetic moment that generates a local magnetic field thus every molecule is a source of rapidly fluctuating magnetic fields. As a nuclei moves through the sample it experiences these fields of varying magnitude and orientation, some of which will be of a frequency able to induce the transition of the nuclei between quantum states. Transitions tend to dephase the spins (T2) and/or restore equilibrium populations (T1) in the two energy states.

The spectral density function describes the intensity of the magnetic field fluctuations at any given frequency and is related to the rotational correlation time of the protein. Molecules moving with different rates of molecular motion will experience more or less noise of the frequency required to induce transitions and thus will exhibit differing relaxation rates. The rate of T2 relaxation increases as the rotational correlation time of a

molecule increases, where as the rate of T1 relaxation reaches its maximum at a rotational correlation time between 10^{-10} and 10^{-9} seconds in the high field magnets (600, 700 and 800 MHz) used in our studies, with the rate decreasing as molecules tumble faster or slower than this.

In this thesis we use relaxation measurements to determine the rotational correlation time of domains in the context of a construct. Rotational correlation time refers to the time taken for a molecule to rotate by one radian. Assuming a globular fold this parameter is proportional to the molecular weight of the tumbling entity and one can determine the rotational correlation time using Equation 3:

$$\tau_c \approx \frac{1}{4\pi\nu_N}\sqrt{6\frac{T_1}{T_2}-7} \qquad (3)$$

where ν_N is the ^{15}N resonance frequency (in Hz).

2.2 Circular dichroism

Polarised light can be thought of being made up of left-handed circularly polarised light (L-CPL) and right-handed circularly polarised light (R-CPL). When such light passes through a solution containing an optically active molecule, i.e. a molecule which contains a chiral chromophore, the two CPL states are affected differently. Circular dichroism (CD) is the difference in absorption between L-CPL and R-CPL. This difference is plotted as a function of wavelength to give CD spectrum. CD can be measured as the change in absorbance (ΔA) or as ellipticity (θ) most commonly in the units of millidegrees.

In proteins the main chromophores are the aromatic side chains of tryptophan, tyrosine and phenylalanine along with the disulphide bonds between cysteines in the near-UV spectrum (310-255nm) and the peptide bond in the far-UV spectrum (180-260nm). While the chromophores themselves are not chiral they are present in the chiral environment of the protein. Amino acids (apart from glycine) are optically active due to the four different groups attached to their α-carbon. In biological organisms almost all amino acids are found as the L-enantiomer giving rise to the differential absorbance of L-CPL and R-CPL. Similarly in nucleic acids the bases are the chromophores but due to their planarity do not have

intrinsic chiral properties. While a small CD signal comes from the sugar entities in the backbone the large signal from nucleic acid oligomers is due to the close contact and electronic interactions that come with base stacking.

It is important to note that the CD signal is not merely the sum of its constitutive parts but is highly influenced by 3D structure with the aromatic bases reporting on tertiary and occasionally quaternary structure and the peptide bond in proteins giving characteristic spectra depending of the secondary structure. Due to the structural information provided by CD one can monitor secondary structure content, changes in conformation and unfolding and refolding of proteins.

2.2.1 Thermal denaturation

Conformational stability is defined in terms of the free energy of unfolding. The more positive the Gibbs energy the more energy has to be added to the system to enable protein unfolding, therefore the more stable the protein is. The relation between the populations of the folded and unfolded states at a given temperature is described by Equation 4:

$$\Delta G^{o} = -RTlnK \tag{4}$$

where K is the unfolding constant (=[denatured protein]/[folded protein]), R is the universal gas constant, and T is the absolute temperature.

CD can be used to monitor the unfolding of a protein as in almost all cases there is a large difference in signal between the folded and unfolded states. In our studies we measure the CD signal at 220nm at increasing temperatures. The change in the optical signal as a function of temperature is given by Equation 5:

$$Signal = Opt_{folded}F_{folded} + Opt_{denatured}F_{denatured} \tag{5}$$

where F_{folded} is the fraction of protein in the folded state and $F_{denatured}$ is the fraction of protein in the denatured state. The optical signals are considered to be related to the temperature in a linearly dependent way as described by Equation 6 and Equation 7:

$$Opt_{folded} = I_{folded} + S_{folded}T \tag{6}$$

$$Opt_{denatured} = I_{denatured} + S_{denatured}T \tag{7}$$

where I and S are the intercepts and slopes respectively for the CD signal of the folded and denatured protein. The fraction of protein in the folded and denatured states are described by Equation 8:

$$F_{denatured} = K/(1+K) = \left(e^{-\Delta G^o/RT}\right)/\left(1+e^{-\Delta G^o/RT}\right) = 1 - F_{folded} \qquad (8)$$

and the absolute Gibbs energy for the transition is described by the modified Gibbs-Helmholtz (Equation 9):

$$\Delta G^o(T) = \Delta H_{T_m}\left(1 - \frac{T}{T_m}\right) + \Delta C_p\left\{(T - T_m) - T\ln\left(\frac{T}{T_m}\right)\right\} \qquad (9)$$

where T_m and ΔH_{Tm} are the midpoint melting temperature and the enthalpy at T_m respectively, and ΔC_p is the difference in heat capacity between the folded and the unfolded protein.

A standard non linear least squares fitting procedure written by Stephen Martin (National Institute for Medical Research, UK) was used to fit the data to Equation 8 with T_m, ΔH_{Tm}, ΔC_p, I and S as variables.

As we are measuring thermodynamic parameters the unfolding must be reversible. In our studies we were able to check this by lowering the temperature back to the starting point and checking that the signal that the signal from the folded protein was recovered.

2.2.2 Protein-RNA interactions

Circular dichroism can also be applied to the monitoring of protein-ligand interactions. CD is a quantitative method so can be used to determine dissociation constants as long as the spectrum resulting from the sum of the RNA and protein components of the titration differs from that of the complex. During the titrations CD is monitored at near-UV wavelengths as here the signal from the RNA is significantly greater than that of the protein. Therefore one begins with the RNA in the cuvette and makes additions of the protein. Changes in the CD spectra throughout the titration can result from changes in the environment of the bases and/or changes to the structure of the components. In this case, the CD signal is monitored at a wavelength (or the average over a range of wavelengths) where the RNA signal change is maximised but contribution by the free protein is

minimised. This signal is plotted against protein concentration. The observed CD signal for a mixture of A and B is given by Equation 10:

$$\text{Signal} = \Delta_{\varepsilon_A} A_T + \Delta_{\varepsilon_B} B_T + \left(\Delta_{\varepsilon_{AB}} - \Delta_{\varepsilon_A} - \Delta_{\varepsilon_B}\right)\left(\frac{(A_T+B_T+K_d)-\sqrt{(A_T+B_T+K_d)^2-4A_TB_T}}{2}\right) \quad (10)$$

where A_T and B_T are the total concentrations of A and B. The two unknowns in the equation Kd and $\Delta\varepsilon$(AB) can be determined using a standard non linear least squares fitting procedure.

2.3 References

1. Fielding, L. NMR methods for the determination of protein–ligand dissociation constants. *Prog. Nucl. Magn. Reson. Spectrosc.* **51,** 219–242 (2007).

The other information in this chapter can be found in the following textbooks and reviews:

NMR

Keeler, J. *Understanding NMR Spectroscopy (Second Edition)* (Wiley, 2010)

Manoharan, V., Perez-Canadillas, J. M., and Ramos, A. *Protein-RNA Interactions in NMR of Biomolecules Towards Mechanistic Systems Biology* (Bertini, I., McGreevy, K. S. and Parigi, G.) (Wiley-Blackwell, 2012)

Rattle, H. *An NMR Primer for Life Scientists* (Partnership Press, 1995)

Rule, G. S., and Hitchens, T. K. *Fundamentals of Protein NMR Spectroscopy* (Springer, 2006)

Circular Dichroism

Martin, S. R. & Schilstra, M. J. Circular dichroism and its application to the study of biomolecules. *Methods Cell Biol.* **84,** 263–93 (2008)

3. Protein-RNA specificity by high-throughput principal component analysis of NMR spectra

3. 1 Introduction

RNA binding proteins often use multiple domains to recognise their RNA targets but within this binding there will be recognition of specific sequences by individual domains. Mutation of the RNA target sequence disrupts the protein-RNA interaction proving sequence specific recognition to be important. However, RNA binding proteins are not always bound to the strongest binding sequence in vivo and lower affinity sites are often critical for regulation in RNA metabolism. Furthermore, as RBPs are often involved in multiple functions and need to recognise a wide variety of targets some domains only show moderate or degenerate nucleobase preference. As a first step to understand how the protein can use different domains to recognise a diverse set of targets it is necessary to elucidate the full sequence preference of each individual domain. For many domains defining the sequence specificity is difficult due to low binding affinity: weak protein-RNA interaction is difficult to monitor with many other established techniques such as SELEX that provide the best binding sequence of high affinity binding domains.[1]

In order to overcome this limitation our group has established Scaffold Independent Analysis (SIA), a method tailored to low-to-intermediate affinity domains that provides the nucleobase preference for each position in the binding site.[2] SIA works by comparing titrations of a protein domain with randomised RNA pools in which the base in one position is kept constant, either A, C, G or U, while all other positions are random nucleotides. SIA uses ^{1}H-^{15}N correlation spectroscopy to monitor the binding of RNA to the domain by the shift of resonances in fast exchange. As mentioned before many individual RBDs bind RNA with a Kd in the micromolar range. The chemical shift perturbation caused by the binding of each of the pools is then used to ascertain which bind with the higher average affinity and therefore determine an order of preference for the bases in a position. This allows us to monitor interactions with Kds in the range we are aiming for, 1-1000µM.

Alongside the identification of favoured bases, disallowed nucleotides and degenerate binding are also highlighted. All of these data can be used to rationalise the binding sites determined by in vivo techniques. The additional information can also help in the design of structure-driven mutations that change the specificity of a domain which can then be used to further probe the role of a specific domain in target selection.

3.1.1 Manual SIA

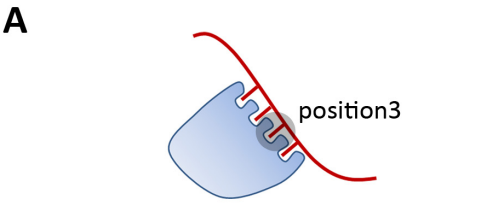
In the classical SIA method ^{1}H-^{15}N correlational spectra are recorded of the free protein and the protein in the presence of two different ratios of each of the randomised pools. The changes in chemical shift (Δδ) of around 10 peaks are measured in each of the titrations. Peaks are chosen on the basis of belonging to a backbone amide, shifting in the fast exchange regime and being able to be clearly followed and distinguishable from other nearby peaks in all of the titrations. The list of peaks is then further refined by removing peaks which exhibit the smallest shifts.

To calculate the scores for position 1 of the binding site the Δδ of each peak in the four titrations (NANNN, NCNNN, NGNNN and NUNNN) is normalized with respect to the largest Δδ of the four. The normalized values across the different peaks are averaged to obtain the final comparative SIA score. These scores reflect the preference of the protein for one nucleobase versus another (Figure 3.1). This procedure averages out the difference in chemical shift that may be cause by the different chemical structure of the nucleobases.[2]

This manual recording of NMR spectra and measurement of peak shifts was laborious and time consuming. In addition, by using only a subset of peaks in the final calculations bias could be introduced into the results. It has been shown that depending on the set of peaks used the end score could change by as much as 0.1.[3] Principal component analysis (PCA) meanwhile undertakes a global analysis of the whole spectrum and includes information provided by the complete set of peaks, including the many that undergo only small chemical shift changes.

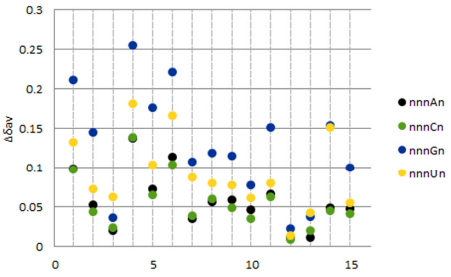
3.1.2 Principal Component Analysis

Principal component analysis allows for the reduction of the variables in a multivariate data set while retaining much of the original variation in the data. It allows for the identification



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**	*nNNNG*



D

	position 1	position 2	**position 3**	position 4
A pools	*0.50*	*0.55*	**0.45**	*0.75*
C pools	*0.48*	*0.62*	**0.44**	*0.55*
G pools	*0.93*	*0.81*	**0.97**	*0.87*
U pools	*0.93*	*0.98*	**0.69**	*0.92*

Figure 3.1 | Scheme of Scaffold Independent Analysis using manual analysis. A) The sequence preference of position 3 of the binding site is to be determined. B) Titrations are performed with the RNA pools in bold and ^{1}H-^{15}N SOFAST-HMQC are recorded for the free protein and each of the titration points. C) Spectra are processed and peak shift perturbations measured. D) The weighted average chemical shift for each residue with the four different RNA pools are normalised with respect to the greatest shift. Normalised values are averaged over the set of residues to give the final scores.

of patterns in the data and displays the data in a way that highlights differences and similarities.[4]

With regards to NMR, PCA has been used to correlate measured observables with conformational transitions, global structural changes and changes in the protein environment.[5,6] PCA has also been used to directly analyse variability between NMR spectra. For example, it is widely used in metabolomics, where the principal components of NMR spectra from patient samples are used to correlate variations in metabolite levels with disease states.[7] Also, PCA has been used in a qualitative fashion to analyse changes in ^{1}H-^{15}N correlation spectra of proteins resulting from high-throughput ligand screening procedures of small molecules.[8] Here we use PCA as a streamlined and unbiased method to provide a semi-quantitative and comparative estimate of RNA binding affinities.

In the analysis of data by PCA the mean is subtracted from each of the data dimensions giving a data set with a mean of zero. The covariance is calculated for the data matrix which indicates how much the dimensions vary from the mean with respect to each other. For this covariance matrix unit eigenvectors and their eigenvalues are calculated. The eigenvectors are placed in order according to the corresponding eigenvalue. The eigenvector with the highest eigenvalue becomes the first principal component and will be along the direction of the largest variation in the data. Next, the orthogonal eigenvector with the highest eigenvalue becomes the second principal component and correlates with the second largest variance in the data. This continues down the list of eigenvectors and their eigenvalues. It is at this point that the number of variables in the data is reduced. As the first few principal components contain the majority of variability in the data components of lesser significance can be discarded. A matrix is formed with the chosen vectors and is called the feature vector. Both the feature vector and the mean subtracted data matrix are transposed and multiplied together with the feature vector on the left. The dimensions of the final data set will be equal to the number of eigenvectors included in the feature vector.[4]

3.1.3 RNA15, T-STAR, TUT4

In order to test the methods we have used examples of the three most common RNA binding domains: RRM, KH domains and ZF. The three protein domains chosen to test the

method were the RRM of RNA15, the KH domain of T-STAR and the third CCHC-type ZF of TUT4.

RNA15 is an mRNA 3' end processing factor found in yeast. It is a core component of the Cleavage Factor 1A (CF1A).[9] The RNA15 interaction with the 3'UTR of the nascent RNA chain is necessary for CF1A recruitment to the RNA, and for RNA cleavage and polyadenylation.[10] The RNA interaction is mediated by an RRM domain located in the N-terminal segment of the protein.[11] Previous studies have shown a preference for G/U-rich sequences.[12]

T-STAR, also referred to as Sam68-like mammalian protein 2 (SLM2), is a member of the STAR family of proteins. These contain a KH domain flanked by two regions, QUA1 and QUA2.[13] In another family member, SF1, GLD-1 and QUAKING, QUA2 which is C-terminal to the canonical KH domain forms an α-helix which contacts RNA and extends the binding site by up to three nucleotides.[14,15] However the QUA2 motif in T-STAR was shown not to be involved in RNA binding.[16] The most well characterised paralogue of T-STAR is Sam68 which regulates the splicing of several exons during neuronal differentiation *in vitro.*[17] T-STAR is expressed in high levels in the brain and testes and due to its high sequence similarity is thought to play a role in similar processes.[18] Indeed a recent paper demonstrated that T-STAR controls alternative splicing of Neurexin pre-mRNAs in the brain by interacting with a UWAA-rich response element immediately downstream of the regulated exon.[19,20]

TUT4 is a non-templated poly-uridylase whose targets include pre-let7 miRNA, histone mRNA and a subset of cytokine targeting miRNAs and will be described in more detail in Chapter 4.[21–23]

3.1.4 Aims

The manual measurement and tabulation of peak shifts in SIA analysis is time consuming and laborious. Furthermore the choice of peaks included in the analysis can introduce bias. We aim to use Principal Component Analysis to analyse the spectra of free and RNA bound protein in a high-throughput and unbiased way to determine the sequence preference of an RNA binding domain. To complement the faster analysis we also aim to automate the spectral acquisition process. The improvements in the method will allow us to apply this technique quickly and easily to the many domains which have contradictory information

about binding specificities and targets in order to get a full readout of the domain RNA binding preferences.

3.2 Methods

3.2.1 Protein preparation

RNA15 RRM was prepared by Laura Robertson at the National Institute for Medical Research. RNA15 RRM was in a final buffer of 40 mM NaCl, 20 mM Tris pH 7.5, 0.5 mM TCEP. TSTAR KH was prepared in the lab of Cyril Dominguez at the University of Leicester. TSTAR KH was in a final buffer of 50 mM NaCl, 20 mM NaH_2PO_4 pH 6.1, 0.1% β-mercaptoethanol. TUT4 CCHC-ZF3 was prepared as described in Section 4.2.4. TUT4 CCHC-ZF3 was in a final buffer of 100 mM NaCl, 10 mM Tris pH 7.4, 0.5 mM TCEP, 10 μM $ZnCl_2$.

3.2.2 Data acquisition

180 μl aliquots of protein sample were prepared and corresponding RNA volumes added where required to produce a free sample and a sample with each of the RNA pools at the chosen protein to RNA ratio. Samples were then transferred into 3mm NMR tubes and loaded into a sample rack. The samples were loaded into the Bruker Avance NMR spectrometer operating at 700 MHz by a Bruker SampleJet. ^{1}H-^{15}N SOFAST-HMQC spectra were recorded for each of the samples.

3.2.3 Data processing and analysis

Manual analysis

A subset of peaks which shift upon addition of RNA was chosen. Peaks which were perturbed upon addition of the RNA pools were selected and used in the analysis if they arose from a backbone amide, were in the fast or fast-to-medium exchange regimes, the position of the peak could be accurately followed in all titrations and did not overlap with other peaks. Peaks were further discounted if they exhibited only minor shift perturbation. Shift perturbation for each of the peaks upon addition of each RNA pool was measured. For each position the shifts are normalised to the largest of the four. Then for each pool the shift values are averaged over all the chosen peaks to give the final score.

Analysis with PCA

Using an NMRpipe-based pipeline spectra were converted and processed in batch. Then the principal components for the group of spectra that define nucleobase preference in each position using the NMRPipe module pcaNMR were calculated. The calculation was performed on the free protein, which is used as reference to define the basis set for the PCA calculation, plus the four spectra of complexes with the RNA pools. Then the PC2 value of each of the RNA bound spectra was subtracted from that of the reference free spectrum. This value represents the distance between the two projections on the second principal component.

3.2.4 RNA binding assays - NMR

30 µM ^{15}N-labelled TSTAR KH in 50 mM NaCl, 20 mM NaH_2PO_4 pH 6.1, 0.1% β-mercaptoethanol, was titrated with unlabelled RNA oligonucleotides up to maximum protein to RNA ratio of 1 to 16. $^{1}H^{15}N$ SOFAST-HMQC spectra were recorded at each titration point at 25^{o}C on Bruker Avance NMR spectrometers operating at 600 or 700 MHz.

3.3 Improvements to data acquisition and analysis

One of the limitations of the SIA was the large amount of time and laborious nature of the work necessary to acquire the spectra. In the current SIA methodology for each of the 16 RNA pools (Figure 3.1b) three ^{1}H-^{15}N SOFAST-HMQC experiments were recorded, one of the free protein and two titration points with increasing amounts of RNA. The exact ratio of protein to RNA in these experiments depended on the affinity of the domain but was kept constant over the 16 pools to allow for direct comparison. Thus in total 48 spectra (16 pools x 3 experiments) were acquired assuming the scanning of four binding positions (Figure 3.2). This approach was taken to ensure that in titrations with all 16 pools chemical shifts chosen to be included in the analysis were shifting in the fast exchange regime and so were reliable indicators of the molar fraction of bound protein and that we were not reaching saturation of binding. However when looking at previously acquired data sets we observed that in general chemical shifts behaved similarly regardless of the RNA pool used. Therefore the behaviour of peaks could be judged by performing a preliminary titration with just one or two of the RNA pools and the information regarding peak choice and an optimal protein:RNA ratio for analysis extrapolated to the other pools. Taking this into account we altered the method so that we now acquire one spectrum of the free protein and one bound spectra for each of the 16 RNA pools at a protein:RNA ratio determined

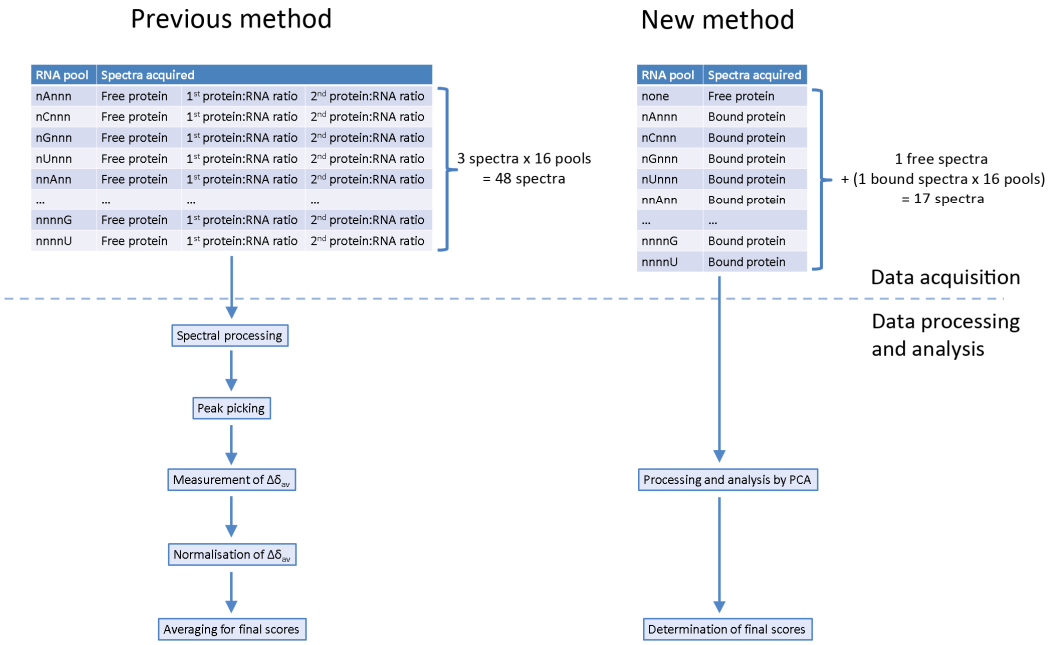


Figure 3.2 | Changes in data acquisition, processing and analysis in SIA methodology. The spectra recorded for analysis and all steps in the processing and analysis of data are shown for the previous methodology (left) and new methodology (right).

from the preliminary titration. This reduced the total number of spectra acquired down to 17 assuming the scanning of four binding positions (Figure 3.2).

Another improvement to the methodology was made possible with the availability of an automatic sample changer. A script was produced by Alain Oregioni (National Institute for Medical Research, UK) using the Bruker Sample Jet to load the samples in turn into the spectrometer and run a ^{1}H-^{15}N SOFAST-HMQC for each. Samples were stored at 4^{o}C and then preheated to 25^{o}C prior to being inserted into the spectrometer. Locking, tune and matching and shimming were performed and then a ^{1}H-^{15}N SOFAST-HMQC was run. The smaller number of spectra recorded and the automation of data acquisition reduces the time required to acquire the data from several days to several hours. The Bruker Sample Jet also facilitated the change from using 5mm Shigemi NMR tubes which require 330µl of sample to 3mm NMR tubes which require just 180 µl of sample. This along with the reduced total number of spectra acquired lowers the total volume of protein and RNA required roughly 2-fold.

Once spectra had been acquired data were processed using a script produced by Geoff Kelly (National Institute for Medical Research, UK) (Appendix I). The script inputs the data of the ^{1}H-^{15}N SOFAST-HMQC of the free protein and that of the four RNA bound samples for a given scanned position. These data are processed using NMRPipe and input into PCAView where PC values are calculated.[24] The output from PCAView gives data for five principal components.

The changes made to the technique add up to make significant sample and time reductions (Table 3.1). The amount of protein and RNA required is reduced by 54% and 45% respectively. In the previous method the amount of time required to acquire the data was 28 hours and this entailed manual changing of the sample so the time had to split over 4 days. With the automation of sample loading and spectral acquisition the amount of spectrometer time required to record the spectra is reduced to 8.5 hours and samples can be prepared in two hours then loaded into the machine to run over night. The change from manual processing and measurement of peak shift for analysis to the use of PCA had reduced the time required at this stage from around 6 hours to roughly 30 minutes. These advances mean that the protocol can quickly and easily be applied to RNA binding domains

	Previous Method	New Method	Absolute reduction	% reduction
Protein sample	20 samples x 330µl =**6600µl**	17 samples x 180µl =**3060µl**	3540µl	54%
RNA sample	16 samples x 330µl =**5280µl**	16 samples x 180µl =**2880µl**	2400µl	45%
Spectrometer time	28 hours split over **4 days**	8.5 hours over **1 day**	3 days	75%
Data acquisition labour time	manual sample changing **~28 hours** split over 4 days	using Bruker SampleJet **2 hours**	26 hours	93%
Data processing and analysis time	by manual analysis **~6 hours**	by PCA **~0.5 hours**	5.5 hours	92%

Table 3.1 | Reduction in sample and time requirements in SIA methodology. Sample volumes and time required to perform and analyse the set of experiments to determine the sequence specificity scanning four positions before and after changes to the methodology were introduced as described in Section 3.3.

from different proteins or allows for the examination of the same protein domain in different binding conditions.

3.4 Comparison of manual analysis and principal component analysis

Previous studies had found a correlation between binding to a small molecular weight compound and the second principal component.[8] The absolute scores of each of the spectra for a given principal component were plotted against the average peak shift observed in the spectrum (a shift of 0 was used for the free spectrum). This showed a correlation between chemical shift and the second principal component score for all binding positions in the T-STAR KH and RNA15 RRM. In most positions the correlation was strong while in others some discrepancies occurred. This is to be expected due to the extra information from the spectra taken into account in PCA as opposed to measuring the shifts of a subset of peaks. Such a correlation did not exist in the first, third, fourth and fifth principal components (Figure 3.3). For TUT4 we saw that while spectra with the highest chemical shifts had a second principal component more distant from that of the free spectrum such a correlation also existed for the first principal component. It could be that the first principal component is dominated by peak intensity as this is another large variation between the free and bound spectra. In the ZF spectra peaks seem to be in the fast to medium exchange regime and so the line width increases and the signal-to-noise decreases as well as peaks shifting. This would mean that along with larger peak shifts upon binding to stronger pools which would dominate PC2 scores the same correlation would also be observed in the PC1 scores due to increasing line broadening upon binding with stronger pools. Overall these results indicate that the second principal component can be used to determine the order of preference of binding of the RNA pools.

We wanted to see if the order of preference calculated was the same for both methods of analysis. SIA scores were calculated for the data sets using either the manual method (Appendix II) or PCA. Due to the very different analysis procedures of the manual and automated methods the absolute values of the scores cannot be compared. However the overall order of preference, which is the output we are interested in, can be compared (Figure 3.4).

For the majority of positions the order of preference determined is identical between the two methods. However in some cases where the difference in preference between two

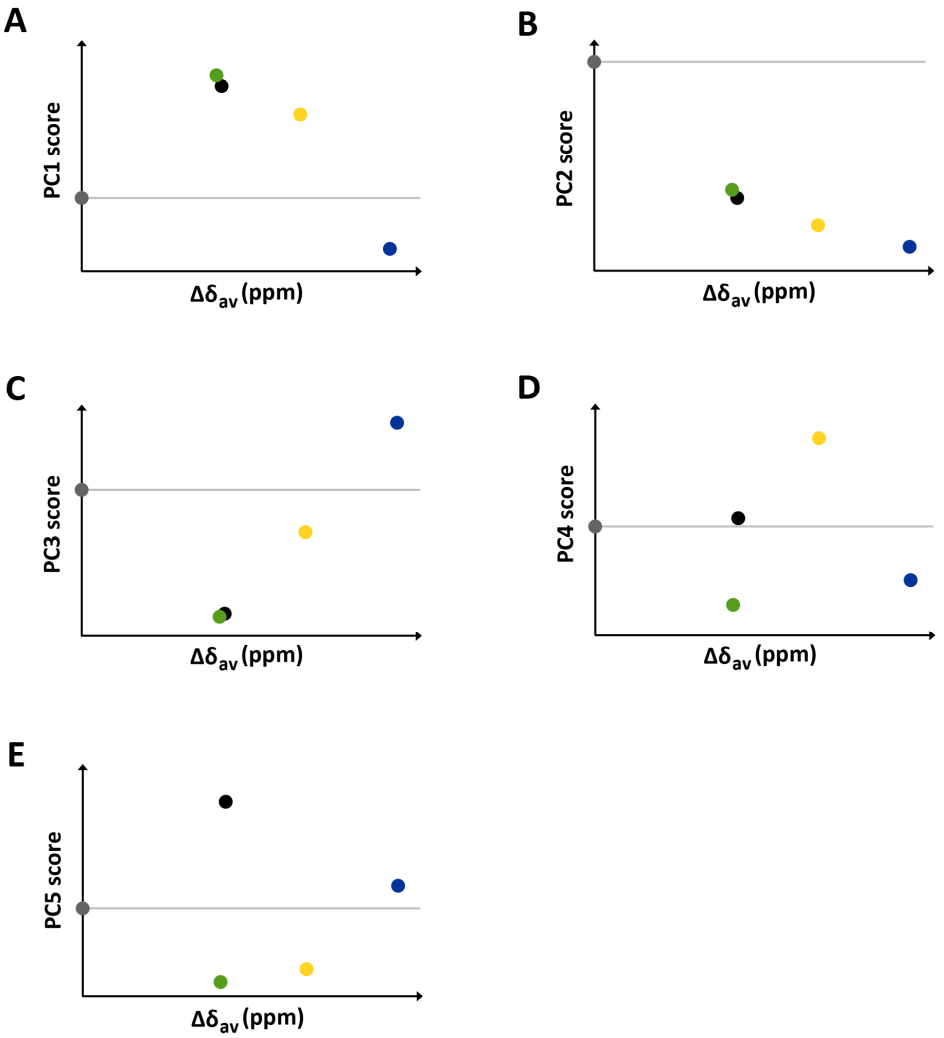


Figure 3.3 | Correlation of principal component scores with chemical shift perturbation. Each plot shows the correlation of free RNA15 RRM (grey), and RNA15 RRM upon addition of NNNAN (black), NNNCN (green), NNNGN (blue), NNNUN (yellow) at protein to RNA ratios of 1:1. The score for the free spectrum is marked with a grey line across the plot for ease of comparison. The x-axis shows the weighted average chemical shift of peaks used in manual analysis. The y-axis shows the calculated score for the (A) first, (B) second, (C) third, (D) fourth, and (E) fifth principal components.

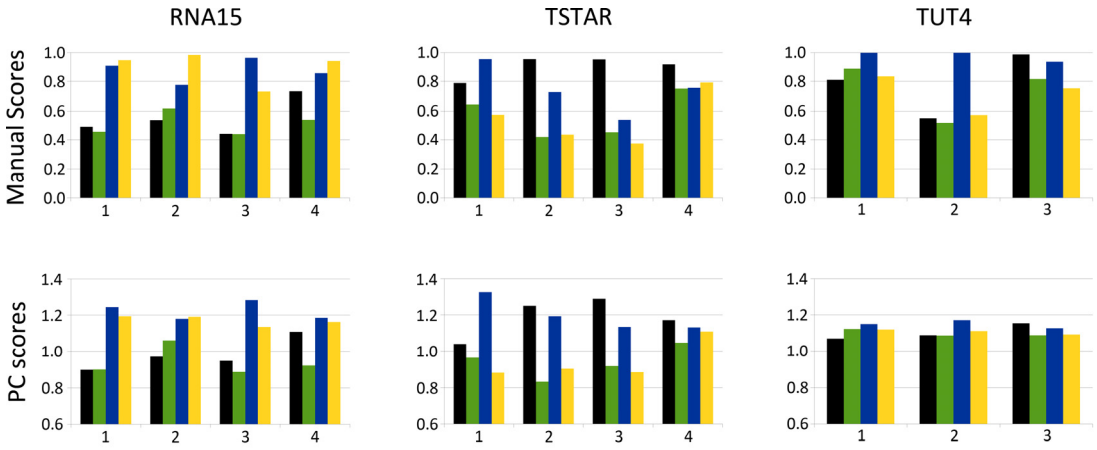


Figure 3.4 | Comparison of the nucleobase preference generated by manual analysis and PCA of RNA15 RRM, T-STAR KH and TUT4 CCHC-ZF3. Histograms display the binding site position on the x-axis with each bar representing the A (black), C (green), G (blue), and U (yellow) RNA pools. The y-axis displays score calculated either by manual measurement or by PCA.

nucleobases is particularly small there are slight discrepancies. For RNA15 positions 1 and 4 manual analysis shows a very small preference for U over G while in the PCA G is slightly above U. However as the difference in preference between the two bases is so slight it could be concluded that the protein does not really distinguish between them and both methods show this.

In the results for TUT4 CCHC-ZF3 while the order of preference matches with the manual scores the range of PCA scores is much smaller than for that of the other domains. Here all scores are within 0.09 of each other while RNA15 and T-STAR average a range of 0.31 and 0.35 respectively. This could be due to the fact that the zinc finger domain is much smaller with only 22 amide peaks in the spectra and only 6 peaks make measurable shifts. Compared to this RNA15 RRM has 74 peaks and the KH domain of T-STAR has 71 peaks and upon RNA binding large numbers of these peaks are affected (Figure 3.5). Because of the small number of peaks in the zinc finger spectrum there are less changes in the spectra overall which could lead to similar PCA scores.

The results show that while there are small differences between the two methods of analysis both give the same general idea for the binding specificity and PCA analysis can be used in order to increase speed in the method and decrease introduced bias.

3.5 Sequence specificity of RNA15 and T-STAR

In order to validate the optimisation of the SIA methodology against functionally relevant results we compared the order of preference given by SIA with the available literature to see if the results were in agreement, and ran further binding tests when they were not. The sequence specificity of TUT4 CCHC-ZF3 is not discussed here as it will be covered in Section 4.3.2.

3.5.1 RNA15

SIA showed that RNA15 is able to bind either a G or a U in positions 1 and 4 and does not discriminate between the two bases. In position 2 a U or a G is preferred and looking at the manual data U is the slightly favoured base. In position 3 a G is preferred over the other bases. In all positions A and C are disfavoured (apart from position 4 which seems able to accommodate an A).

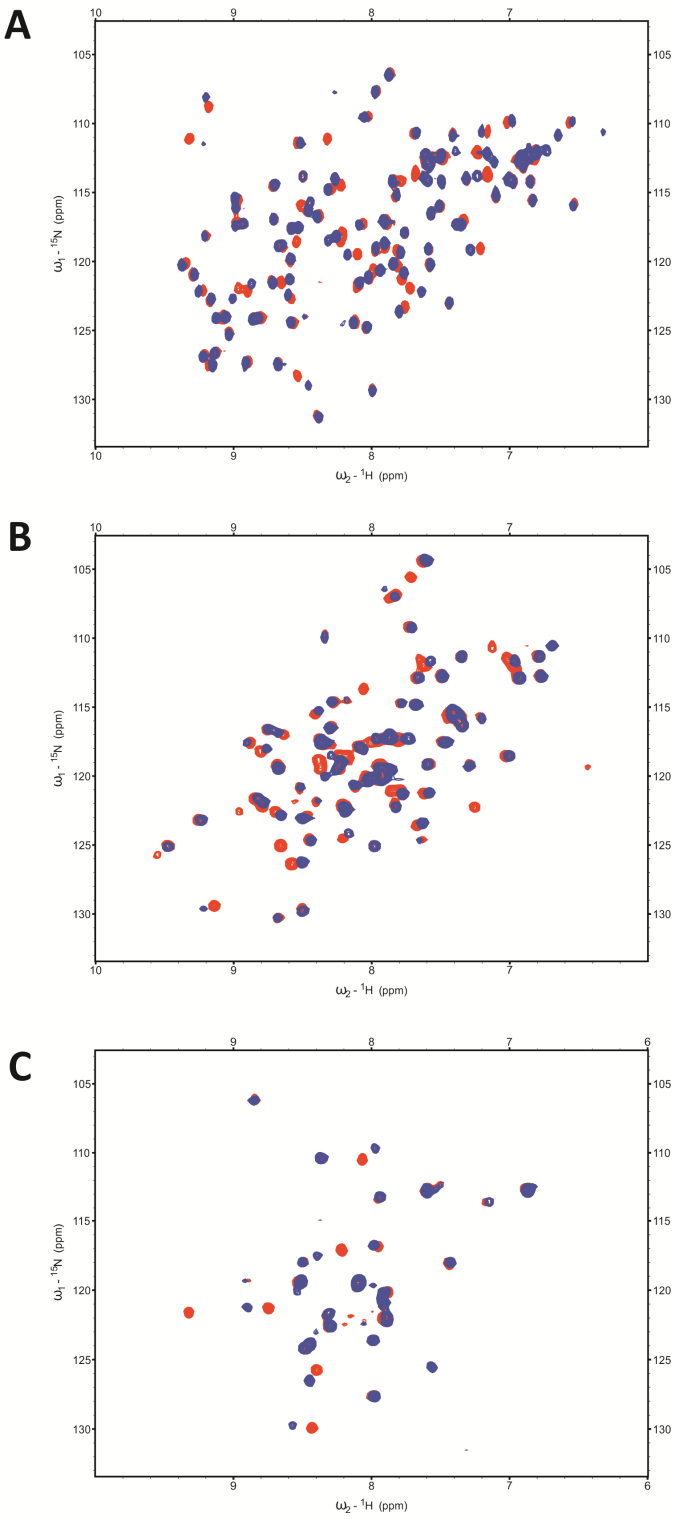


Figure 3.5 | Comparison of spectral changes upon addition of an RNA pool for RNA15 RRM, T-STAR KH and TUT4 CCHC-ZF3. Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of A) free 25µM RNA15 RRM (red) and with NNNGN at protein to RNA ratio 1:1 (blue), B) free 40µM T-STAR KH (red) and with NNNAN at protein to RNA ratio 1:2, C) free 100µM TUT4 CCHC-ZF3 (red) and with NNGN at protein to RNA ratio 1:4.

In the literature fluorescence studies had found the strongest binding sequence to be UGUUGU with a Kd of ~5 μM. Other U or U/G rich sequences (UUUUUU and UGUUUG) had Kds of ~9 μM and ~10 μM respectively. Where U and G bases were replaced individually with C or A bases Kds were 4- to 8-fold lower and a hexamer of ACAACA showed weak binding with a Kd of more than 100 μM (Figure 3.7).[12] Taking into account the SIA data two new hexamers were tested for binding by Laura Robertson, GUGUGU and GUUUGU, and determined to have Kds of 2.2 μM and 1.4 μM respectively (unpublished data).

Crystal structures of RNA15 in complex with GUUGU also show agreement with our SIA data in that guanine and uridine can be accepted whereas adenine and cytosine are excluded. In Site I a U or G base can be specifically recognised. The bases stack against an aromatic side chain and further Watson-Crick like hydrogen bonding occurs between carbonyl and imino functional groups on the edges of U or G and the protein backbone (Figure 3.6a). An A or C would be incompatible with this binding pocket as they lack imino proton and have an exocyclic amino group in place of the carbonyl group in U and G, which can no longer form a hydrogen bond with the protein back bone and could even clash with the side chain of Y27. In a further binding site a guanine was bound stacking against the aromatic side chain on the β sheet surface and hydrogen bonding to a backbone carbonyl in the protein (Figure 3.6b). [12]

3.5.2 T-STAR

Our results on the nucleobase preference of T-STAR KH domain showed a preferred sequence of GAA(A/G/U) for both manual and SIA analysis. Although the ranking was the same there was a slight discrepancy between the strength of preference for bases between the manual and PCA analyses (Figure 3.3). Importantly both C and U are disfavoured in positions 2 and 3.

Previous data on T-STAR binding specificity consistently shows a preference for AU-rich sequences. SELEX data indicates a bipartite motif containing two U(U/A)AA repeats.[20] RNAcompete found the core binding site to be UAA, CLIP showed an AU-rich binding preference and Neurexin mRNA, so far the only identified biological target, has been found to contain 4x(UUAA).[16,19,25] While our data agrees with the specificity of T-STAR for adenine it does not correlate with the presence of uridine in the binding motifs.

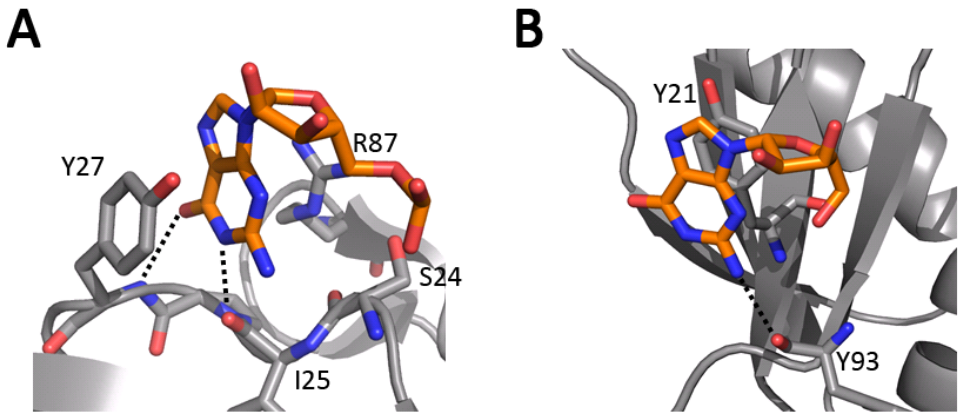


Figure 3.6|**Structural details of RNA15 binding sites.** A) Site I binding pocket. The guanine base (orange) and the side chains of S24, I25, Y27 and R87 are shown in stick representation. B) Site II binding pocket. The guanine base (orange) and the side chains of Y21 and Y93 are shown in stick representation. Hydrogen bonds are shown as black dashed lines

We determined the binding affinities of a set of 5-mer RNAs in order to gain further insight into the discrepancy. Titrations monitored by NMR were run with the short RNAs CGAAA, CAGAA, CAUAA, CAAGA and CAAUA. Weighted average chemical shift perturbation was plotted against RNA to protein ratio and fitted using Equation 2 (Section 2.1.1) in Origin 9.1. Dissociation constants were determined using the Kd value averaged over several peaks. A full list of peaks used in the analysis and binding curve plots can be found in Appendix III. Oligonucleotides containing a G in position 1 or 2 (CGAAA and CAGAA) had the highest affinity of the ones tested with Kds of 54 μM and 58 μM respectively. In comparison when a U was placed in position 2 (CAUAA) the affinity decreased 2-fold to 116 μM. A slight decrease in affinity, from 116 μM to 151 μM was also observed when a G in position 3 (CAAGA) was changed to a U (CAAUA) (Figure 3.7). These results agree with our SIA data that a U is disfavoured in the two central binding positions and that the inclusion of a G in position 1 or 2 increases the affinity.

3.6 Discussion

Many RNA binding proteins are multifunctional and are required to recognise a variety of different targets in the course of performing their functions. In order to do this many proteins contain multiple RBDs and can employ them in different combinations and structural arrangements. This can lead to sets of targets identified *in vivo* being very difficult to interpret. The determination of the full register of binding preference for the individual RNA binding domains in the protein would help to deconvolute this data and identify the mechanism by which different targets are recognised. Previously a method called Scaffold Independent Analysis which determines the nucleobase preference in each position of the binding site of a nucleic acid binding protein was devised to solve this problem.[2] However this method is cumbersome and laborious and the large sample and time requirements make it restrictive. Furthermore the choice of peaks used in the analysis can bias the results.[3] In this chapter I discuss how we have attempted to alter the original SIA methodology to make it more apt to solving the problem of sequence specificity determination for the numerous RBDs present in the cell. We have streamlined the technique in order to get the same, if not more accurate due to lack of bias, information out with less time and effort put in.

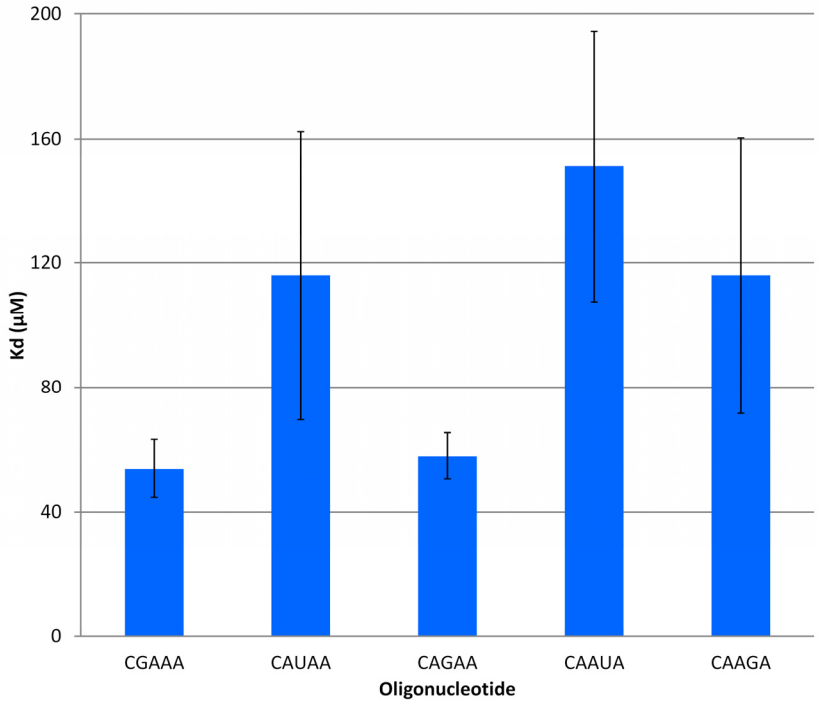


Figure 3.7 | Binding affinities of T-STAR KH with RNA pentamers. The x-axis displays the sequence of the pentamer and the y-axis the Kd in µM.

It must be noted that the technique relies on the assumption that given an equal amount of bound fraction of protein peak shifts will be equal regardless of the RNA pool used. However this may not be the case due to the effect of ring currents of bases on the chemical shifts. Furthermore the technique only looks at domains binding to mostly unstructured 5-mer oligonucleotides and so sequence specificity may not be able to be extrapolated to physiological conditions where the domain contacts structured RNAs.

We first reduced the amount of protein and RNA required by reducing the number of titration points and by changing from 5mm to 3mm NMR tubes. This reduction leads to cost savings and also the time and resources required to make the protein samples. The fewer titration points run, along with the automation of sample loading into the spectrometer and acquiring the spectra drastically reduces the time taken to record a data set. Furthermore as the script can be set up and run overnight this makes much more efficient use of spectrometer time. Overall in the four days previously required to collect a data set for one domain we can now acquire data for eight domains. Time gains were also made in the processing and data analysis of the data. Previously after processing all peaks shifts had to be measured by hand in Sparky before being normalised and averaged to give the final score. In the new methodology processing of the spectra was incorporated into a script which also ran PCA. This makes the analysis much quicker and also less bias is introduced into the final scores as changes in the whole of the spectra are taken into account not just the shift of a subset of peaks.

Overall these improvements make the technique much less cumbersome and will allow, we hope, its use by a larger number of groups and the screening of large numbers of RBDs still lacking a defined specificity. Interestingly we found that for the T-STAR KH domain the higher affinity binder is not necessarily the most common sequence in the functional targets. This is consistent with our growing understanding of protein-RNA interaction in the cell and with the need to provide a full account of nucleobase preference for each of the domains screened.

3.7 References

1. Klug, S. J. & Famulok, M. All you wanted to know about SELEX. *Mol. Biol. Rep.* **20,** 97–107 (1994).

2. Beuth, B., García-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,** 10205–10 (2007).

3. García-Mayoral, M. F., Díaz-Moreno, I., Hollingworth, D. & Ramos, A. The sequence selectivity of KSRP explains its flexibility in the recognition of the RNA targets. *Nucleic Acids Res.* **36,** 5290–6 (2008).

4. Smith, L. I. A tutorial on Principal Components Analysis Introduction. (2002). at <www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf>

5. Sakurai, K. & Goto, Y. Principal component analysis of the pH-dependent conformational transitions of bovine beta-lactoglobulin monitored by heteronuclear NMR. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 15346–51 (2007).

6. Robertson, I. M., Boyko, R. F. & Sykes, B. D. Visualizing the principal component of ^{1}H, ^{15}N-HSQC NMR spectral changes that reflect protein structural or functional properties: application to troponin C. *J. Biomol. NMR* **51,** 115–22 (2011).

7. Tenori, L. in *NMR of Biomolecules: Towards Mechanistic Systems Biology* (eds. Bertini, I., McGreevy, K. S. & Parigi, G.) Chapter 28 (Wiley-VCH Verlag GmbH & Co. KGaA, 2012). doi:10.1002/9783527644506.ch28

8. Ross, A., Schlotterbeck, G., Klaus, W. & Senn, H. Automation of NMR measurements and data evaluation for systematically screening interactions of small molecules with target proteins. *J. Biomol. NMR* **16,** 139–46 (2000).

9. Minvielle-Sebastia, L., Preker, P. J. & Keller, W. RNA14 and RNA15 proteins as components of a yeast pre-mRNA 3'-end processing factor. *Science* **266,** 1702–5 (1994).

10. Gross, S. & Moore, C. L. Rna15 interaction with the A-rich yeast polyadenylation signal is an essential step in mRNA 3'-end formation. *Mol. Cell. Biol.* **21,** 8045–55 (2001).

11. 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,** 334–49 (2010).

12. 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,** 3119–32 (2010).

13. Vernet, C. & Artzt, K. STAR, a gene family involved in signal transduction and activation of RNA. *Trends Genet.* **13,** 479–84 (1997).

14. Liu, Z. *et al.* Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science* **294,** 1098–102 (2001).

15. Teplova, M. *et al.* Structure-function studies of STAR family Quaking proteins bound to their in vivo RNA target sites. *Genes Dev.* **27,** 928–40 (2013).

16. Foot, J. N., Feracci, M. & Dominguez, C. Screening protein--single stranded RNA complexes by NMR spectroscopy for structure determination. *Methods* **65,** 288–301 (2014).

17. Chawla, G. *et al.* Sam68 regulates a set of alternatively spliced exons during neurogenesis. *Mol. Cell. Biol.* **29,** 201–13 (2009).

18. Venables, J. P. *et al.* SIAH1 targets the alternative splicing factor T-STAR for degradation by the proteasome. *Hum. Mol. Genet.* **13,** 1525–34 (2004).

19. Ehrmann, I. *et al.* The tissue-specific RNA binding protein T-STAR controls regional splicing patterns of neurexin pre-mRNAs in the brain. *PLoS Genet.* **9,** e1003474 (2013).

20. Galarneau, A. & Richard, S. The STAR RNA binding proteins GLD-1, QKI, SAM68 and SLM-2 bind bipartite RNA motifs. *BMC Mol. Biol.* **10,** 47 (2009).

21. Heo, I. *et al.* TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. *Cell* **138,** 696–708 (2009).

22. Schmidt, M.-J., West, S. & Norbury, C. J. The human cytoplasmic RNA terminal U-transferase ZCCHC11 targets histone mRNAs for degradation. *RNA* **17,** 39–44 (2011).

23. Jones, M. R. *et al.* Zcchc11-dependent uridylation of microRNA directs cytokine expression. *Nat. Cell Biol.* **11,** 1157–63 (2009).

24. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6,** 277–93 (1995).

25. Ray, D. *et al.* Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.* **27,** 667–70 (2009).

4. Terminal Uridyl Tranferase 4 (TUT4)

4.1 Introduction

4.1.1 Terminal Uridyl Tranferase Family

TUT4/Zcchc11 (zinc-finger, CCHC domain-containing protein 11) is a member of the DNA polymerase β-like nucleotidyltransferase superfamily.[1] Members of this family catalyse the coupling of nucleoside triphosphates to a free hydroxyl group on the nucleic acid primer via elimination of a pyrophosphate. Catalysis takes place in a cleft formed by three globular domains, catalytic, central and RNA-binding domains. Conserved residues involved in the catalysis include three carboxylates critical for activity and a helical turn motif involved in nucleotide binding.[2] The better characterised members of this family are the poly(A) polymerases which catalyse the addition of a poly(A) tail to mRNA in the nucleus.

TUT4 belongs to a different subgroup, the non-canonical poly(A) polymerases. The seven members of this group include TUTase1 which is localised to the mitochondria and polyadenylates mitochondrial RNAs. The crystal structure of TUTase1 shows a more closed conformation in the absence of substrate than canonical PAPs and also that the protein is only active as a dimer.[3] TUTase3 instead is involved in the polyadenylation-mediated degradation of aberrant pre-rRNAs. The yeast homologue of this protein acts in concert with cofactors, one of which contains RNA-binding zinc fingers.[4] However human TUTase3 does not require cofactors for activity and although the protein does not have a typical RNA binding domain it has been shown that a stretch of basic amino acids in the C-terminus of the protein is able to bind RNA.[5] TUTase6 is a highly specific terminal U transferase which adds a uridyl tail to U6 snRNA.[6] TUTase 7 has similar domain architecture to TUT4 and displays redundancy with TUT4 in some of its functions.

TUT4 is a large protein at 185,166Da. It contains a putative C2H2 zinc finger towards the N-terminus followed by an inactive nucleotidyltransferase domain, and a functional nucleotidyltransferase domain surrounded by three canonical CCHC-type zinc fingers (Figure 4.1).

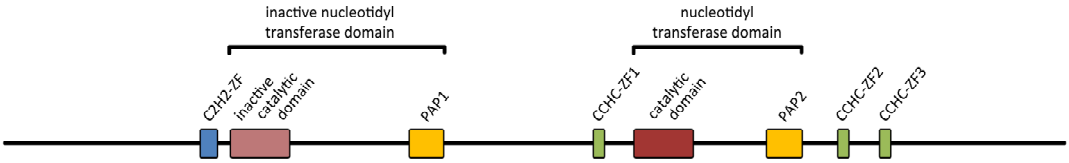


Figure 4.1 | Domain organisation of TUT4. C2H2-ZF, C2H2-type zinc finger; PAP, polyadenylation associated domain; CCHC-ZF, CCHC-type zinc finger.

4.1.2 Roles of TUT4

Three main roles for TUT4 have so far been described which involve the addition of non-templated uridine stretches to the 3' end of a range of diverse RNA targets (Figure 4.2).

Regulation of let-7 biogenesis

miRNAs are short, noncoding single stranded RNAs of around 22 nucleotides in length. They direct the RNA-induced silencing complex (RISC) complex to target mRNA via complementary base pairing where the RISC complex downregulates gene expression by mRNA cleavage or translational repression.[7] miRNAs were first discovered in *C. elegans* but since then have been found in many different organisms and have been shown to play a role in diverse biological functions such as development, life span, cell proliferation, differentiation, signalling pathways, apoptosis and metabolism.[8]

miRNAs are transcribed as part of much longer RNA molecules and can reach several kilobases in length. These primary-miRNA (pri-miRNA) molecules contain one or more ~33nt stems with a terminal loop flanked by ssRNA sequences. The RNAse III Drosha along with its cofactor DGCR8 cleaves at the junction between the dsRNA and ssRNA thus liberating the stem loop. The product of this cleavage is called precursor miRNA (pre-miRNA). The pre-miRNA is exported into the cytoplasm by Exportin-5 where it is further processed by Dicer to form ~22nt dsRNA fragment. This is incorporated into the RISC complex where one strand is chosen for retention and the other is released and degraded. The miRNA then guides RISC to its target mRNA.[7]

The let-7 family of miRNAs were the first to be discovered in humans. One of their major functions is to promote the differentiation of cells and let-7 is absent from human and mouse embryonic stem cells or pluripotent cell populations but is found in differentiated tissues.[9] The family are also tumour repressors with targets of let-7 including the known oncogenes RAS, MYC and HMGA2.[10–12] Furthermore it directly regulates CDC25A and CDK6 which are both key proteins in the cell cycle.[13] A reduction in let-7 levels has been found in several cancer types with reduced let-7 being correlated to increased tumourigenicity and poor patient prognosis.[10,11]

TUT4 is involved in regulating the biogenesis of the let-7 family. Working in concert with another protein, Lin28, TUT4 negatively regulates production of mature miRNAs.[14] Lin28 is expressed in embryonic stem cells and some cancer cells.[15,16] It is also one of the factors

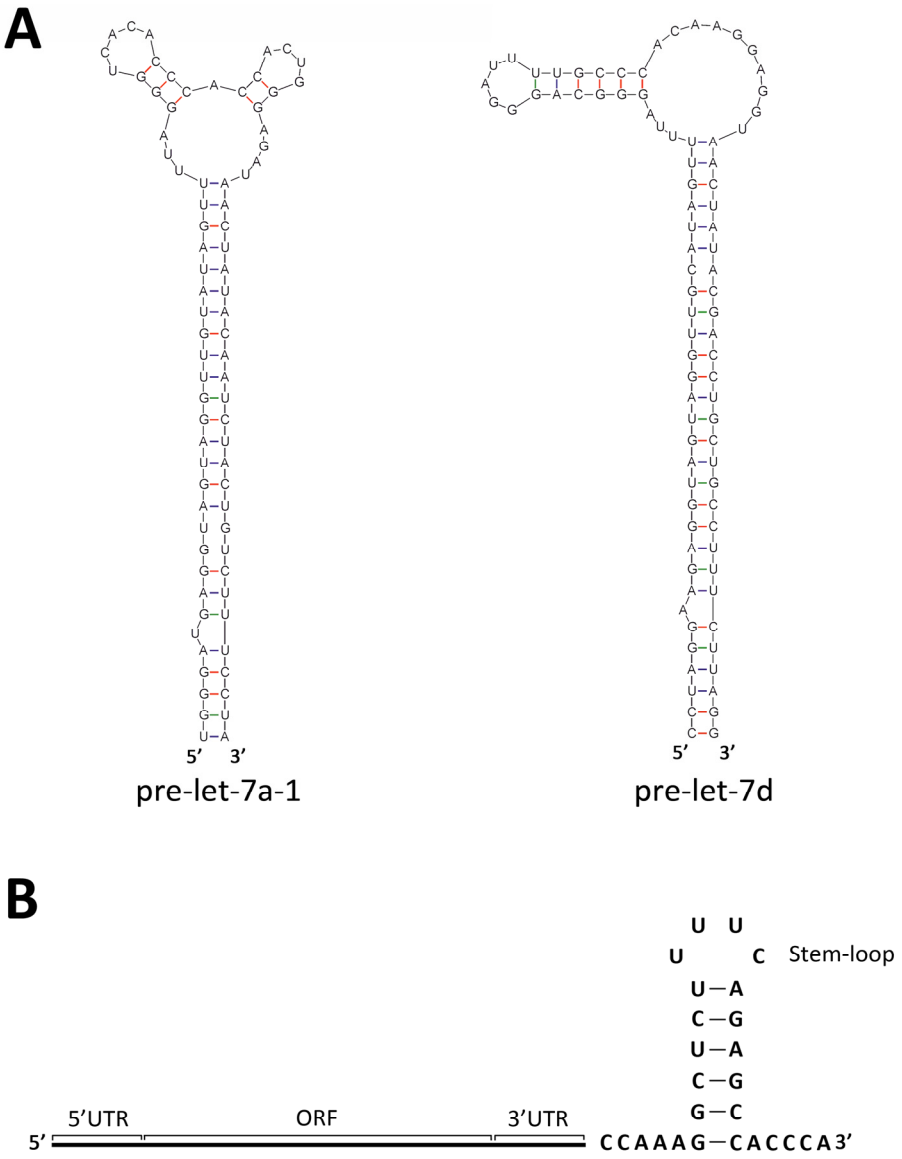


Figure 4.2 | RNA targets of TUT4. A) Predicted secondary structure of pre-let-7a-1 and pre-let-7d using mfold Web Server.[17] B) *Homo sapiens* canonical histone mRNA. C) Mature miR-26a.

that when overexpressed facilitates the formation of induced pluripotent stem cells.[18] Lin28 contains two types of RNA-binding domain, a cold shock domain and two CCHC-type zinc fingers, but does not have a catalytic domain. Binding of Lin28 is thought to directly interfere with Drosha and Dicer processing.[19] Furthermore when Lin28 is bound to pre-let-7 RNA TUT4 can interact and form a stable complex.[20] TUT4 oligo-uridylates the pre-mRNA adding a tail of 10-30 uridines to the 3' end.[14] The long ssRNA tail makes it an unsuitable substrate for Dicer and a target for the exoribonuclease DIS3L2 (Figure 4.3a). [21]

In the absence of Lin28, TUT4 and two other family members, TUT7 and TUT2, promote production of mature let-7. A subset of pri-let-7 miRNAs contains a bulged uridine at the cleavage site of Drosha. The bulged base is not recognised by Drosha so after processing the pre-mRNA is left with a 1nt 3' overhang rather than the canonical 2nt.[22] Dicer inefficiently processes these pre-miRNAs as it prefers substrates with a 2nt 3' overhang.[23] TUT4, TUT7 and TUT2 preferentially bind this subset of pre-miRNA and add a single uridine onto the 3' end. This creates the preferred overhang for Dicer and the pre-miRNAs can now be efficiently processed into the mature form (Figure 4.3a).[22]

Lin28 is described as acting as a molecular switch, converting TUT4 (and TUT7) from a key biogenesis factor to a negative regulator. The characterisation of TUT4 binding with pre-let7 in the presence or absence of Lin28 gives insight into how this switch occurs. TUT4 has a slow catalytic turnover of 0.22nt s^{-1}. In the absence of Lin28 the interaction between TUT4 and pre-let7 is short lived, lasting around one second. In the presence of Lin28 the interaction of TUT4 with the pre-mRNA is more stable, lasting for around 240 seconds.[20] This longer interaction time would allow the oligo-uridylation of the mRNA substrate rather than the mono-uridylation event that occurs in the absence of Lin28.

Targeting histone mRNA for degradation

Histone mRNAs differ from other mRNAs in that they do not contain introns and have a highly conserved stem loop at their 3' ends instead of a poly(A) tail. The stem-loop binding protein (SLBP) is bound to the stem loop throughout the lifetime of the mRNA and is involved in all steps of metabolism. Expression of histones is coordinated with DNA abundance and are synthesised during S-phase in the cell cycle. Synthesis outside of S-phase is toxic to the cell and this requires multiple levels of regulation.[24]

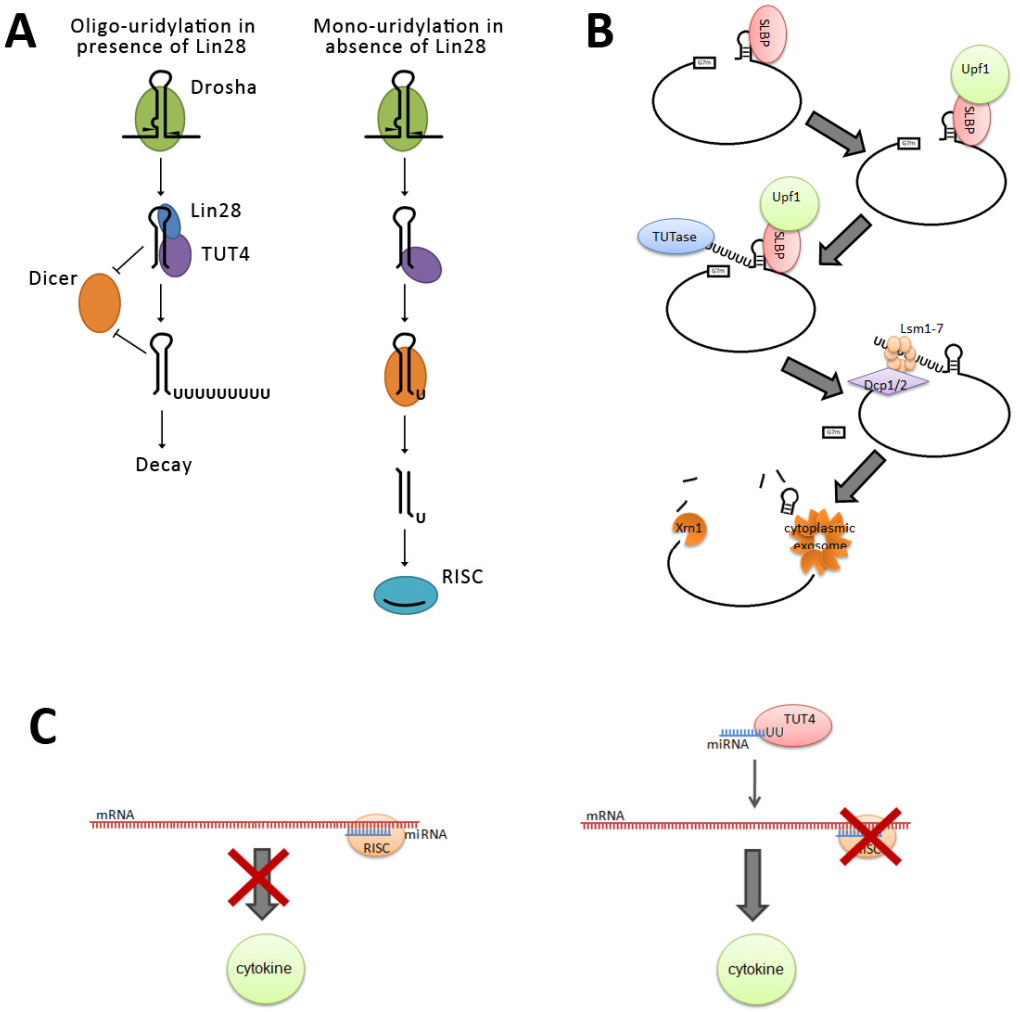


Figure 4.3 | Schematic representation of the roles of TUT4. A) Regulation of the biogenesis of the let-7 family of miRNAs. A bulged nucleotide next to the Drosha cleavage site results in a group of pre-miRNAs with a 1 nt 3' overhang. In the absence of Lin28, TUT4 mono-uridylates these pre-miRNA making them optimal substrates for Dicer cleavage. The mature miRNA is incorporated into the RISC complex. In the presence of Lin28, TUT4 forms a stable complex with the Lin28 bound pre-miRNA and oligouridylates the pre-miRNA. This prevents processing by Dicer and promotes degradation of the product. B) Degradation of histone mRNA. SLBP is bound to the 3' stem loop. Upf1 binds, the TUTase is recruited and oligouridylates the mRNA. Lsm1-7 binds to the poly(U) tail and recruits Dcp1/2 which decaps the mRNA. The mRNA is degraded from both the 3' and 5' ends. C) Regulation of cytokine expression. Mature miRNA targets RISC to the cytokine mRNA resulting in translational repression and/or degradation. TUT4 uridylates the miRNA thus inhibiting its function.

At the end of S-phase a protein complex forms at the stem loop of the histone mRNA. This complex contains the SLBP and Upf1. The TUTase is recruited to the complex and catalyses the addition of a poly(U) tail. The degradation of the mRNA then follows a similar pathway to deadenylation dependent mRNA decay. The Lsm1-7 complex binds to this tail and recruits a decapping complex. The decapped mRNA can then be degraded by Xrn1 exoribonuclease in the 5' to 3' direction and the exosome in the 3' to 5' direction (Figure 4.3b).[25]

While Mullen and Marzluff identified the enzymes involved in the addition of the poly(U) tail as TUTase1 and TUTase3, work by Schmidt et al. states that the required enzyme is TUT4.[26] This seems to be a more viable hypothesis as TUTase1 is located in the mitochondria[27] and TUTase3 uses ATP preferentially as a substrate over UTP.[5]

Attenuation of mature miRNA action

TUT4 has been implicated in the addition of terminal uridines to mature miRNAs.[28] The set of miRNAs affected is diverse and includes those targeting IGF-1, VEGF, TGF-α and IL-6.[29] It is thought that TUT4 adds a short uridine tail onto the 3' end of the mature miRNAs and that this attenuates the repressive effect of the miRNA (Figure 4.3c). Upon uridylation levels of the miRNAs do not decrease indicating the tail is not a marker for degradation as with the histone mRNA.[29] The mechanism by which the addition of the tail prevents the miRNA performing its function is still unknown.

4.1.3 CCHC-type Zinc Finger Domains

TUT4 contains several RNA binding domains, including four zinc fingers: a C2H2-type zinc finger at the N-terminal and three CCHC-type zinc fingers surrounding the catalytic core. The structure of the CCHC-type zinc finger is defined by two short β-strands connected by a turn followed by a short helix or loop. The coordination of zinc by three cysteines and one histidine is important for the formation of the structural motif.

Structural information for the CCHC-type zinc fingers of Lin28 and HIV-1 Nucleocapsid protein in complex with their targets is available. Lin28 contains two CCHC-type zinc fingers separated by a short proline rich linker. The zinc fingers bind to a GGAG sequence in the stem loop of the let-7 family of miRNAs.[19,30] Most of the residues involved in binding to the RNA are in rigid regions such as adjacent to zinc-coordinating residues or in the linker. This

causes a kink to form in the RNA backbone which is matched by a large rearrangement of the four residues which make up the inter domain linker in the protein.[30] As with Lin28, the HIV-1 Nucleocapsid protein contains two CCHC-type zinc fingers separated by a four amino acid linker. The fingers work together to bind specifically to the sequence GGAG found in the RNA tetraloop of the SL3 ψ-RNA recognition element. Upon binding the linker becomes structured and extensive interactions are formed between the two zinc fingers[31].

It is unclear how TUT4 recognises its targets and whether it binds to a specific sequence. Furthermore in contrast to the two examples above the linker between the two C-terminal zinc knuckles is 47 amino acids in length and so the two structural domains could act independently of each other, binding to two separate sequences on the RNA.

4.1.4 Aims

Here we aim to investigate the RNA binding of TUT4 in order to gain more information on how it recognises its varied targets. We focus on the zinc fingers in the protein and try to determine if they are true RNA binding domains and if so their sequence specificity and if they are able to bind as individual units. Furthermore we aim to look at the domain packing in the protein to help in understanding where in the physiological targets the zinc fingers could be binding.

4.2 Methods

4.2.1 Cloning

Expression plasmids for TUT4 zinc finger constructs comprising of C2H2-ZF (301-330), CCHC-ZF1 (903-931), CCHC-ZF2 (1293-1321), CCHC-ZF3 (1354-1382) and CCHC-ZF23 (1293-1382) were made (numbering relates to Homo sapiens Terminal Uridyl Transferase Isoform A PubMed Accession number NP_001009881 and the full amino acid sequence can be found in Appendix IV). The TUT4 gene was cloned from a human cDNA library by David Hollingworth and used as a template for subsequent PCRs. DNA encoding for the sections of interest were PCR-amplified to create the required inserts using primers that introduced 5' NcoI and 3' HindIII restriction sites. Both inserts and pET-M11 vector were digested with NcoI and HindIII restriction enzymes to create compatible overlapping ends. Vector and insert were then ligated together using T4 DNA ligase. Final expression plasmids encoded

for the zinc finger domains of TUT4 with an N-terminal hexa-histidine tag (His-tag) cleavable by tobacco etch virus (TEV) protease all under the control of a T7 promoter.

Construct Name	Start residue	End residue	Forward Primer	Reverse Primer
C2H2-ZF	301	330	C2H2-ZF_301FW	C2H2-ZF_330RV
CCHC-ZF1	902	931	CCHC-ZF1_903FW	CCHC-ZF1_931RV
CCHC-ZF2	1293	1321	CCHC-ZF2_1293FW	CCHC-ZF2_1321RV
CCHC-ZF3	1354	1382	CCHC-ZF3_1354FW	CCHC-ZF3_1382RV
CCHC-ZF23	1293	1382	CCHC-ZF2_1293FW	CCHC-ZF3_1382RV

Table 4.1 | Table of TUT4 zinc finger constructs and primers used in cloning. Sequences of primers can be found in Appendix V.

4.2.2 Site-directed Mutagenesis

Primers were designed to introduce the mutations R1296S/Y1304S and F1360S/V1368S into CCHC-ZF2 and CCHC-ZF3 respectively. Point mutations were introduced into the constructs by amplification of the plasmid using overlapping complementary primers with the mutation of interest inserted at the centre of the oligonucleotides. Following PCR amplification parent DNA was removed by DpnI digestion.

Construct Name	Start residue	End residue	Mutations	Primers
CCHC-ZF2 R1296S/Y1304S	1293	1321	R1296S/Y1304S	CCHC_ZF2_R1296S_FW CCHC_ZF2_R1296S_RV CCHC_ZF2_Y1304S_FW CCHC_ZF2_Y1304S_RV
CCHC-ZF3 F1360S/V1368S	1354	1382	F1360S/V1368S	CCHC_ZF2_F1360S_FW CCHC_ZF2_F1360S_RV CCHC_ZF2_V1368S_FW CCHC_ZF2_V1368S_RV

Table 4.2 | Table of TUT4 ZF mutant constructs and primers used in sire-directed mutagenesis. Sequences of primers can be found in Appendix V.

4.2.3 Protein expression

50 µl of *E. coli* BL21(DE3) cells were transformed with 2µl of plasmid using a standard heat shock protocol. 500 µl of transformed cells were used to inoculate 100 ml of M9 minimal media containing $(^{15}NH_4)_2SO_4$ as the only nitrogen source and/or ^{13}C6-D-glucose as the only carbon source. Cells were grown overnight at 37^{o}C. The overnight culture was used to

inoculate 1500 ml of M9 minimal media to an OD_{600} of 0.1. Cells were then grown to an OD_{600} of 0.6 before the temperature was reduced to 22^{o}C and protein expression induced with IPTG at a final concentration of 0.5 mM. Cells were grown overnight at 22^{o}C, harvested by centrifugation and stored at -80^{o}C.

4.2.4 Protein purification

Cells were resuspended in denaturing buffer pH 8.0 (100 mM NaH_2PO_4, 10 mM Tris-HCl pH 8.0, 8 M urea)(20 ml per litre of culture), stirred gently at room temperature for 60 minutes and sonicated (Branson Sonifier 250, power output 50 W, 60% duty cyle, 2x45 seconds). The lysate was centrifuged at 7740g for 30 minutes and the supernatant recovered. The protein was purified by immobilised metal ion affinity chromatography (IMAC) columns. 4 ml of Ni-NTA resin per litre of cell culture was packed into a gravity-driven column and equilibrated with 10 column volumes (CV) denaturing buffer pH 8.0. The supernatant was loaded onto the column and the flow through loaded for a second time. The resin was washed with 5 CV denaturing buffer pH 6.3 (100 mM NaH_2PO_4, 10 mM Tris-HCl pH 6.3, 8 M urea) and a further 5 CV of denaturing buffer pH 5.9 (100 mM NaH_2PO_4, 10 mM Tris-HCl pH 5.9, 8 M urea). Protein was eluted with 5 CV denaturing buffer pH 4.5 (100 mM NaH_2PO_4, 10 mM Tris-HCl pH 4.5, 8 M urea). The eluted protein was dialysed in Spectra/Por Dialysis Membrane 3500 MWCO in 4 litres of equilibration buffer (10 mM Tris pH 8.0, 10 mM Imidazole, 200 mM NaCl, 2 mM β-mercaptoethanol, 10 µM $ZnCl_2$) in order to reduce the urea content in the buffer. TEV protease was then added to a final concentration of 2.5 µM in order to cleave the His-tag. The reaction mixture was dialysed for a further 16 hours for cleavage to take place. Both protease and cleaved tag were separated from the protein by IMAC column (Ni-NTA). The cleavage reaction mixture was loaded into a gravity-driven column packed with Ni-NTA resin equilibrated with 10 CV of equilibration buffer (10 mM Tris pH 8.0, 10 mM imidazole, 200 mM NaCl, 2 mM β-mercaptoethanol, 10 µM $ZnCl_2$) and the flow through loaded for a second time. The resin was washed with 5 CV of equilibration buffer then the tag was eluted with 5CV of elution buffer (10 mM Tris pH 8.0, 250 mM Imidiazole, 1 M NaCl, 2 mM β-mercaptoethanol). The protein containing fractions were then concentrated to 5ml using Vivaspin concentrators in order to be purified further by size exclusion chromatography. Size exclusion chromatography was performed using an ÄKTA purifier system (GE Healthcare) with a Hiload 16/60 Superdex prep grade column equilibrated in equilibration buffer. Fractions of pure protein were pooled and concentrated before being dialysed into a final buffer of 10 mM Tris-HCl pH 7.4, 100 mM

NaCl, 0.5 mM TCEP and 10 µM $ZnCl_2$ and concentrated in a Vivaspin column of appropriate MWCO. Protein concentration was determined from the absorbance at 280 nm and molecular weights confirmed by mass spectrometry.

4.2.5 Coexpression and purification of GST-Lin28 and TUT4 C2H2-ZF

50 µl of *E. coli* BL21(DE3) cells were transformed with 2 µl of full length Lin28 expression plasmid (pGEX-6p3) and 2 µl C2H2-ZF (301-572) or C2H2-ZF (301-701) expression plasmid using a standard heat shock protocol. After 1 hour recovery at 37 ^{o}C cells were plated out onto an agar plate containing both kanamycin and ampicillin to select for cells successfully transformed with both plasmids. Protein expression was performed as described in Section 4.2.3 with harvested cells stored at -80 ^{o}C.

Frozen cells were resuspended in equilibration buffer (10 mM Tris-HCl pH 8.0, 10 mM imidazole, 200 mM NaCl, 2 mM β-mercaptoethanol) (20 ml per 1 L of cell culture) with Triton X-100, DNaseI and lysozyme added, sonicated on ice and centrifuged at 17000 rpm for 60 mins. The recombinant protein was purified by immobilised metal ion affinity chromatography (IMAC) columns. The soluble fraction was incubated with Ni-NTA resin (5ml per litre of culture) at 4^{o}C for 30 mins then poured into a gravity-driven column. The column was washed with 10 CV of wash buffer (10 mM Tris-HCl pH 8.0, 10 mM Imidazole, 1 M NaCl, 2 mM β-mercaptoethanol) and the protein was eluted with 5 CV of elution buffer (10 mM Tris-HCl pH 8.0, 250 mM Imidazole, 1 M NaCl, 2 mM β-mercaptoethanol).

4.2.6 Backbone assignment

Labelled (^{15}N or ^{13}C^{15}N) samples were prepared as described and concentrated to ~200 µM. NMR experiments were conducted at 25^{o}C using a Varian Inova NMR spectrometer operating at 800 MHz. Measurements were made in 90% H2O/10% D2O. A ^{13}C^{15}N labelled TUT4 CCHC-ZF23 sample was used to acquire a HNCA, HN(CO)CA, HNCACB, HN(CO)CACB and ^{15}N-NOESY-HSQC at pH 4.5 and the assignment of CCHC-ZF2 was produced from these data. Unfortunately CCHC-ZF3 was unfolded at pH 4.5 so a HNCA and ^{15}N-NOESY-HSQC on TUT4 CCHC-ZF23 were repeated at pH 7.4. Additionally a ^{13}C^{15}N labelled TUT4 CCHC-ZF3 sample was used to acquire a HNCACB at pH 7.4. These data were used to assign CCHC-ZF3 and the assignment of CCHC-ZF2 at pH 4.5 could be transferred to spectra at pH 7.4. Spectra were processed using NMRPipe/NMRDraw and analysed using XEASY or Sparky in order to determine ^{1}HN, ^{15}N, ^{13}Cα and ^{13}Cβ assignments.

4.2.7 Scaffold Independent Analysis

Scaffold Independent Analysis (SIA) determines the nucleobase preference of a protein domain at each of the positions of a bound RNA molecule. A 70 µM stock protein of CCHC-ZF2 or 100 µM stock protein of CCHC-ZF3 were prepared in the buffer 100 mM NaCl, 10 mM Tris pH 7.4, 0.5 mM TCEP, 10 µM $ZnCl_2$, sodium azide and RNasin Plus. This was aliquoted out into 180 µM samples and RNA pools added where required at a protein to RNA ratio of 1 to 4. Samples were transferred to 3 mm NMR tubes and placed in a rack. Samples were automatically loaded into the spectrometer using the Bruker Sample Jet and ^{1}H-^{15}N SOFAST-HMQC spectra were recorded for each of the samples. NMR data were recorded on a Bruker Avance NMR spectrometer at 700 MHz. Spectra were processed using nmrPipe and data analysed using the manual method described in Section 3.2.3. Briefly the weighted average chemical shift changes (for HN and N) with each RNA pool for a subset of shifting residues is measured and normalised with respect to the highest shift. The average normalised value across the subset of peaks is taken to give a comparative score of binding preference.

4.2.8 RNA binding assays - NMR

60-100 µM ^{15}N-labelled proteins in 10 mM Tris-HCl pH 7.4, 100 mM NaCl, 0.5 mM TCEP and 10 µM $ZnCl_2$ were titrated with unlabelled RNA oligonucleotides up to protein to RNA ratios of 1 to 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 600 or 700 MHz.

4.2.9 NMR relaxation measurements

Standard relaxation experiments were recorded on a ^{15}N-labelled sample to obtain T1, T2 and $^{1}H^{15}N$-NOE values. Experiments were performed on a Varian Inova NMR spectrometer operating at 800 MHz. T1 and T2 values were determined for each residue by fitting an exponential decay to the peak volume over the course of the data collected. Residues were excluded where overlap in the signals prevented accurate measurement on peak volume.

4.2.10 Small scale protein expression and solubility screens

The TUT4 gene cloned from a human cDNA library and used as a template for the cloning of the TUT4 ZF constructs (see section 4.2.1) was found to have a deletion corresponding to the section Arg1108-Glu1145 in the nucleotidyltransferase domain of the protein. Therefore full length TUT4 gene was purchased from OriGene and used as template DNA in

the following PCRs. Primers were designed for use in ligation independent cloning with Crystallisation Construct Designer[32] and used to amplify regions of the TUT4 gene while introducing 5' and 3' extensions complementary to sections of the vectors to produce the inserts (a full list of primers used can be found in Appendix V). Vectors used were pET-47b pET-52b and pET-52-SUMO (provided by Vangelis Christodoulou, National Institute for Medical Research, UK) which contain an N-terminal hexahistidine tag, step tag and SUMO tag respectively and resistance to either kanamycin or ampicillin. Vectors were linearised by digestion with BsaI and then both inserts and linearised vectors were treated with T4 DNA polymerase to produce complementary single stranded overhangs. Insert and vector were mixed and allowed to anneal before 2 µl was used to transform 20 µl of BL21Gold(DE3) cells by a standard heat shock protocol. Transformed cells were allowed to recover in LB then plated out on agar plates containing the relevant antibiotic for selection. Colonies were picked and used to inoculate 1 ml of ZYM-505 media and grown overnight at 37 ^{o}C. 10 µl of overnight culture was used to inoculate 1 ml of fresh ZYM-505 media. The cultures were grown at 30 ^{o}C until OD_{600} was between 6 and 10 then the temperature was reduced as required (trials at 18 ^{o}C, 22 ^{o}C, and 25 ^{o}C) before protein expression was induced with IPTG at a final concentration of 0.5 mM. The cultures were incubated overnight before being harvested by centrifugation and stored at -80 ^{o}C.

Cells were resuspended and lysed, then centrifuged to get rid of cell debris. The protein was purified by affinity chromatography (Ni-sepharose for His-Tagged constructs or Streptactin for GST-tagged constructs). Cleared lysate was loaded onto gravity-driven columns packed with the relevant resin (GE Healthcare). The resin was washed with 0.5 ml of 20mM Tris-HCl pH 7.7, 40mM imidazole, 300 mM NaCl, 1mM TCEP. The wash was repeated three times before protein was eluted with 60 µl of 20 mM Tris-HCl pH 7.7, 500 mM imidazole, 300 mM NaCl, 1 mM TCEP.

Several constructs were also inserted into pET-M41, pET-M60 and Strep-GB1 vector. The DNA sequences encoding for the constructs were PCR-amplified using primers that introduced 5' NcoI and 3' HindIII or XhoI restriction sites. The PCR products were subcloned into the expression vectors and small scale expression tests were run as for vectors produced by ligation independent cloning.

A full list of constructs and expression conditions can be found in Appendix VI.

4.2.11 Small scale protein coexpression with chaperone proteins

Plasmids from the Chaperone Plasmid Set (Takara) containing dnaK-dnaJ-grpE, groES-groEL, tig or combinations of the chaperone sets under the control of an araB or Pzt-1 promoter (Table 4.3) were used to transform 20 µl of BL21Gold(DE3) cells by a standard heat shock protocol. Successfully transformed cells were selected for by resistance to chloramphenicol. These cells were then transformed with plasmids containing the constructs of interest and successfully transformed cells containing both plasmids were selected for by resistance to both kanamycin and chloramphenicol. Growth and expression were performed as described in Section 4.2.10 with the exception of the addition to the media of L-Arabinose and/or tetracycline as required to induce expression of the chaperones.

Plasmid	Chaperone	Promoter	Inducer	Resistant Marker
pG-KJE8	dnaK-dnaJ-grpE groES-groEL	*araB* *Pzt-1*	L-Arabinose Tetracycline	Chloramphenicol
pGro7	groES-groEL	*araB*	L-Arabinose	Chloramphenicol
pKJE7	dnaK-dnaJ-grpE	*araB*	L-Arabinose	Chloramphenicol
pG-Tf2	groES-groEL-tig	*Pzt-1*	Tetracycline	Chloramphenicol
pTf16	tig	*araB*	L-Arabinose	Chloramphenicol

Table 4.3 | Chaperone expression vectors used for coexpression trials.

4.3 RNA binding of TUT4 zinc fingers

TUT4 contains four putative RNA binding domains in the form of one C2H2-type zinc finger towards the N-terminus of the protein and three CCHC-type zinc fingers surrounding the catalytic core. We aimed to characterise the RNA binding of these domains in order to gain insight into how TUT4 recognises its targets.

4.3.1 Expression and purification of ZF constructs

In general CCHC type zinc fingers are small domains which contain only short sections of secondary structural elements. The main fold of the finger is formed by coordination of a combination of four cysteine or histidine residues with a zinc ion and constructs were designed to contain the four coordinating residues with extensions of several amino acids

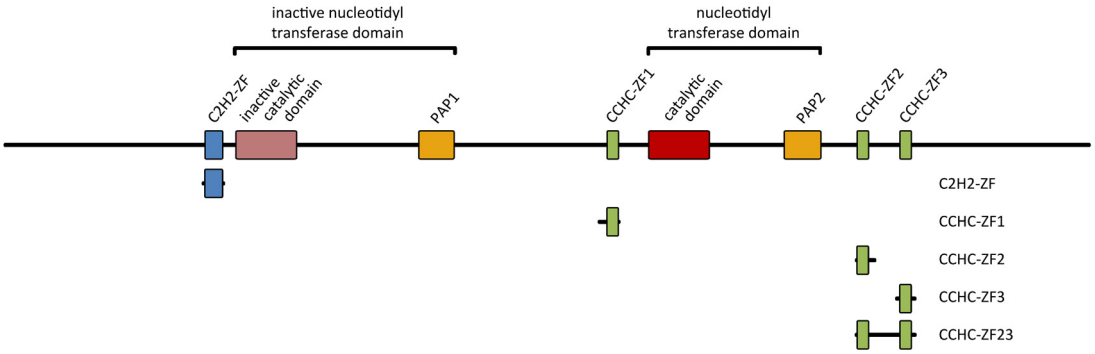


Figure 4.4 | Scheme of TUT4 constructs. Full length TUT4 with domain representations to scale. Constructs shown are C2H2-ZF (301-330), CCHC-ZF1 (903-931), CCHC-ZF2 (1293-1321), CCHC-ZF3 (1354-1382) and CCHC-ZF23 (1293-1382).

at the N- and C-termini (Figure 4.4). Purification under native conditions did not produce sufficient amounts of protein so purification was performed under denaturing conditions as described in Section 4.2.4. and Figure 4.5. This protocol allowed pure samples of reasonable yields for constructs to be obtained.

The three CCHC-type zinc fingers produced $^{1}H^{15}N$ SOFAST-HMQC spectra with the expected number of peaks, 23 amide peaks in each from constructs of 30 (CCHC-ZF1) and 29 (CCHC-ZF2 and CCHCZF3) residues which probably include flexible tail regions (Figure 4.6). A number of these peaks are in the dispersed region of the spectra indicating that the fingers contain secondary structural elements. Upon the addition of EDTA many of these dispersed peaks collapse into the central region of the spectra indicating a loss of this secondary structure and confirming that the fingers coordinate a metal ion to maintain their fold. Lowering the pH of the sample to less than pH 5.5 adversely affected CCHC-ZF1 and CCHC-ZF3 leading to their unfolding. Interestingly CCHC-ZF2 was still folded at acidic pH. Overall the CCHC-type fingers were able to be expressed and purified and are amenable to studies by NMR.

The backbone resonances of TUT4 CCHC-ZF2 and TUT4 CCHC-ZF3 were assigned using the set of experiments detailed in Section 4.2.6 and the strategy described in Section 2.1.2. Assignment data can be found in Appendix VII. Overlap in the central region of the spectra prevented assignment of the linker section of CCHC-ZF23.

The NMR spectra of the N-terminal C2H2-type zinc finger showed the construct to be unfolded so several strategies were followed to attempt to obtain the domain in a folded state. Although purification in denaturing conditions gave a better yield we attempted the purification of C2H2-ZF (301-330) under native conditions (as described for FMRP in Section 5.2.4 except final dialysis buffer used was 10 mM Tris-HCl pH 7.4, 100 mM NaCl, 0.5 mM TCEP and 10 µM $ZnCl_2$) to see if the domain is folded in the cell but is unable to refold outside of the cellular environment. However this did not improve the condition of the construct. In the primary sequence of TUT4 the C2H2-type zinc finger is very close to an inactive nucleotidyl transferase domain. We hypothesised the two domains could interact, stabilising the zinc finger, so we extended the construct at the C-terminus to include both domains. The constructs were also extended at the N-terminus to investigate whether predicted secondary structural elements in this region played a role in folding (a full list of

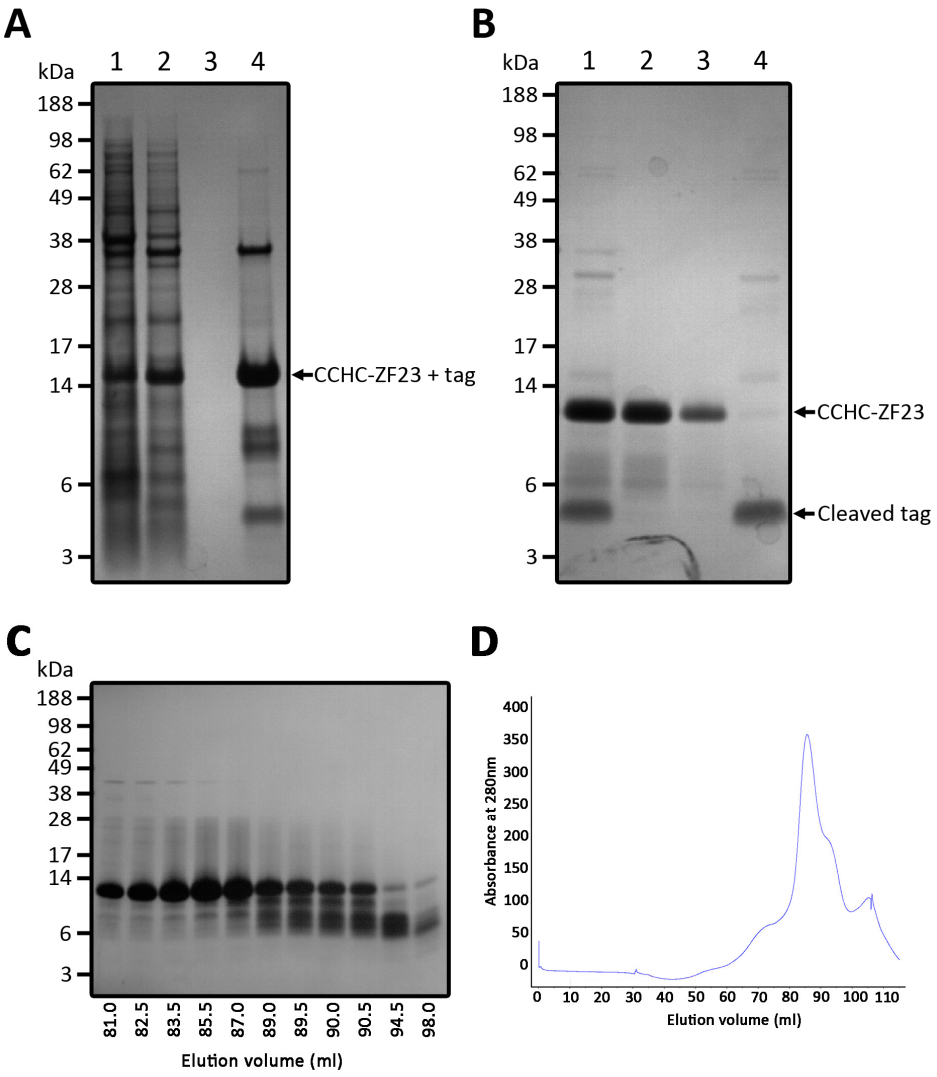


Figure 4.5 | Purification strategy employed for TUT4. A) SDS-page analysis of affinity chromatography fractions of TUT4-ZF23 using nickel-NTA after cell lysis and centrifugation. 1) flow through after column loading with cleared cell lysate, 2) wash with 5 CV denaturing buffer pH 6.3 (100 mM NaH_2PO_4, 10 mM Tris-HCl pH 6.3, 8 M urea), 3) wash with 5 CV of denaturing buffer pH 5.9 (100 mM NaH_2PO_4, 10 mM Tris-HCl pH 5.9, 8 M urea), 4) elution with 5CV denaturing buffer pH 4.5 (100 mM NaH_2PO_4, 10 mM Tris-HCl pH 4.5, 8 M urea). B) SDS-page analysis of affinity chromatography fractions of TUT4-ZF23 using nickel-NTA after His-Tag cleavage with TEV protease. 1) total cleavage reaction, 2) flow through after column loading with total cleavage reaction, 3) wash with 5 CV of equilibration buffer (10 mM Tris pH 8.0, 10 mM imidazole, 200 mM NaCl, 2 mM β-mercaptoethanol, 10 µM $ZnCl_2$), 4) elution with 5 CV of elution buffer (10 mM Tris pH 8.0, 250 mM Imidiazole, 1 M NaCl, 2 mM β-mercaptoethanol). C) SDS-page analysis and D) Chromatogram of size exclusion chromatography on CCHC-ZF23 using a Superdex 75 16/60. Lanes are identified with elution volume in ml.

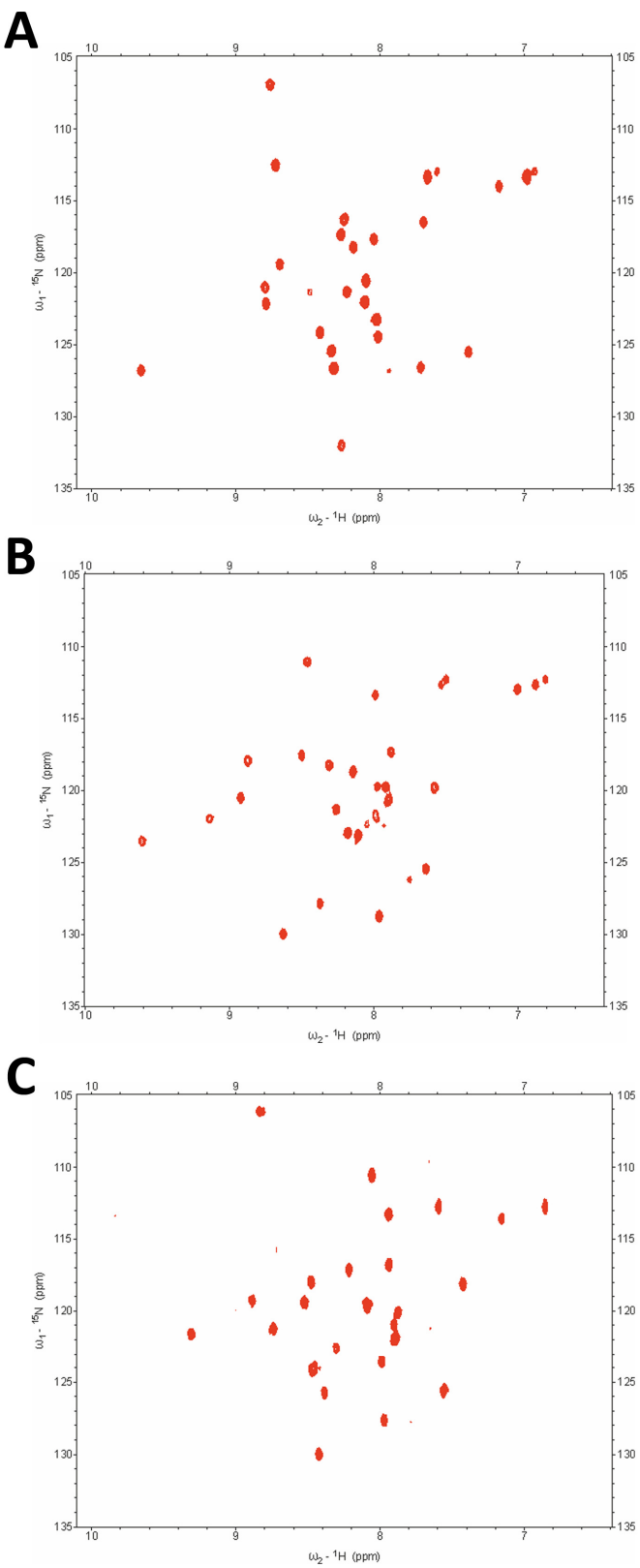


Figure 4.6 | ^{1}H-^{15}N SOFAST-HMQC spectra of TUT4 CCHC type zinc fingers. ^{1}H-^{15}N SOFAST-HMQC spectra of A) TUT4 CCHC-ZF1, B) TUT4 CCHC-ZF2, and C) TUT4 CCHC-ZF3.

constructs can be found in Appendix VI). Constructs were produced as described in Section 4.2.10 however these longer constructs were very poorly soluble and we could not obtain sufficient amounts of product to analyse the fold of the domains (Figure 4.7). Coexpression of C2H2-ZF (301-574) and C2H2-ZF (301-707) with chaperone proteins using the Chaperone Plasmid Set (Takara) and described in detail in Section 4.2.11 also failed to improve the solubility of the constructs.

Truncation and mutation studies have shown the N-terminal zinc finger to be necessary for the interaction with Lin28.[33] In order to see if this interactions could stabilise the zinc finger we mixed C2H2-ZF (301-330) and full length Lin28 at final concentrations of 100 µM each and a final volume of 330 µl and incubated at room temperature for 10 minutes. Upon mixing the two proteins a thick white precipitate formed indicating one or both of the proteins had fallen out of solution and the product could not be analysed. Coexpression of C2H2-ZF (301-572) and C2H2-ZF (301-707) with full length Lin28 as described in Section 4.2.5 failed to produce sufficient amounts of protein product to analyse. As we were unable to obtain a folded construct we could not characterise any potential RNA binding capabilities of the domain or its interaction with Lin28.

The failure of these protocols could be down to the poor quality of the Lin28 construct used. The construct contains an uncleavable GST tag with only a short linker between the tag and the Lin28 protein. This protein is quite unstable and precipitates over time. Furthermore the GST tag is in close proximity to the protein and could occlude potential binding sites. GST is also known to form dimers which could further interfere with potential protein-protein interactions. However newer constructs have now been created in the lab which produce better quality protein with cleavable tags. These constructs could render the interaction studies more successful.

4.3.2 Sequence specificity of CCHC-ZF2 and CCHC-ZF3

We began by studying the zinc finger domains in the C-terminal of the protein as individual units to get a better idea about their RNA binding capabilities.

We first wanted to know how many nucleotides were recognised by each zinc finger so we performed titrations of CCHC-ZF3 with randomised RNA oligonucleotides of increasing lengths, NN, NNN and NNNN. While the same set of peaks were perturbed in all three

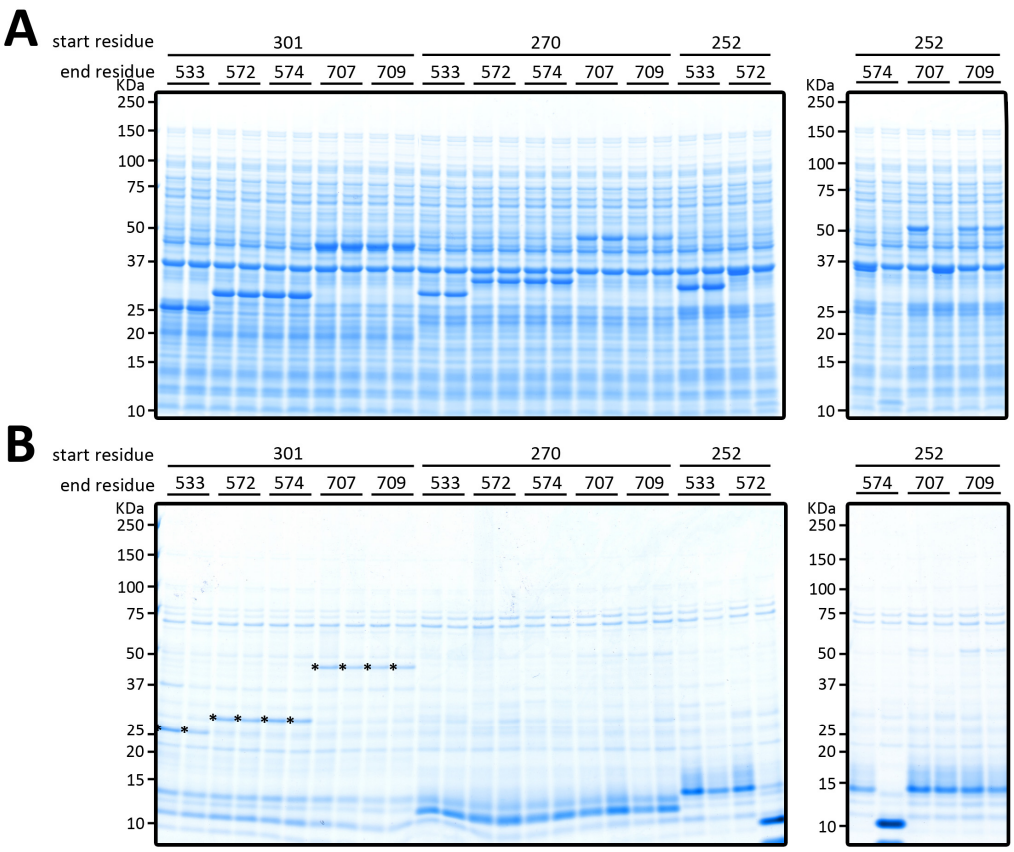


Figure 4.7 | Expression and solubility trials of TUT4 C2H2-ZF constructs in small scale screening. A) SDS-page analysis of whole cell lysate. B) SDS-page analysis of elution fractions after purification with Ni-NTA. The starting and ending amino acid residues of constructs are numbered above the gels. Bands of soluble protein are marked *.

titrations, with two bases only minimal shifts were observed even when the protein to RNA ratio was increased to 1:8. Peaks made measurable shifts with both the 3-mer and 4-mer oligonucleotides with larger shifts observed in the 4-mer titration. The direction of peak shifts were consistent with all the oligonucleotides (Figure 4.8). The increase in affinity is to be expected as increasing the length of a degenerate sequence increases the possible number of binding registers. The peaks also decreased in intensity as upon addition of RNA indicating the binding is in the fast to intermediate exchange regime. The results showed that three bases were required for noticeable binding with affinity increasing when the length of the oligonucleotide is extended to four bases.

In order to determine the sequence specificity of the zinc fingers we performed SIA. From the previous results on the required oligonucleotide length for binding and structural data of the binding of other CCHC-type zinc fingers we decided to perform the SIA scanning three positions for specificity with a random nucleotide at the 5' end to increase affinity.[30]

^{1}H-^{15}N SOFAST-HMQC spectra were recorded on CCHC-ZF2 and CCHC-ZF3 alone and in the presence of the 12 RNA pools at the protein:RNA ratio of 1:4. Analysis was performed by manual measurement of peak shifts with peaks chosen for analysis belonging to a backbone amide, shifting in the fast or fast-to-intermediate exchange regime and being clearly distinguishable from other nearby peaks in all of the titrations. The list of peaks was then further refined by removing peaks which exhibited very small shifts. Peaks used in the analysis are displayed in Figure 4.9. Almost all the peaks used for the analysis of ZF3 had a weighted average chemical shift of more than 0.1 ppm with several shifting up to 0.2 ppm or 0.3 ppm. However weighted average chemical shifts in ZF3 titrations were much smaller with only a small percentage reaching above 0.1 ppm (Figure 4.10). This makes it more difficult to accurately measure shifts and introduces higher errors. The chemical shift perturbations of seven peaks for ZF2 and six peaks for ZF3 were measured. This is less than usual 10-15 peaks measured in this technique due to the small size of the domains and therefore small number of perturbed peaks upon RNA addition. Shifts were normalised and averaged to give final scores (Appendix VII and Appendix II for CCHC-ZF2 and CCHC-ZF3 respectively and Table 4.1).

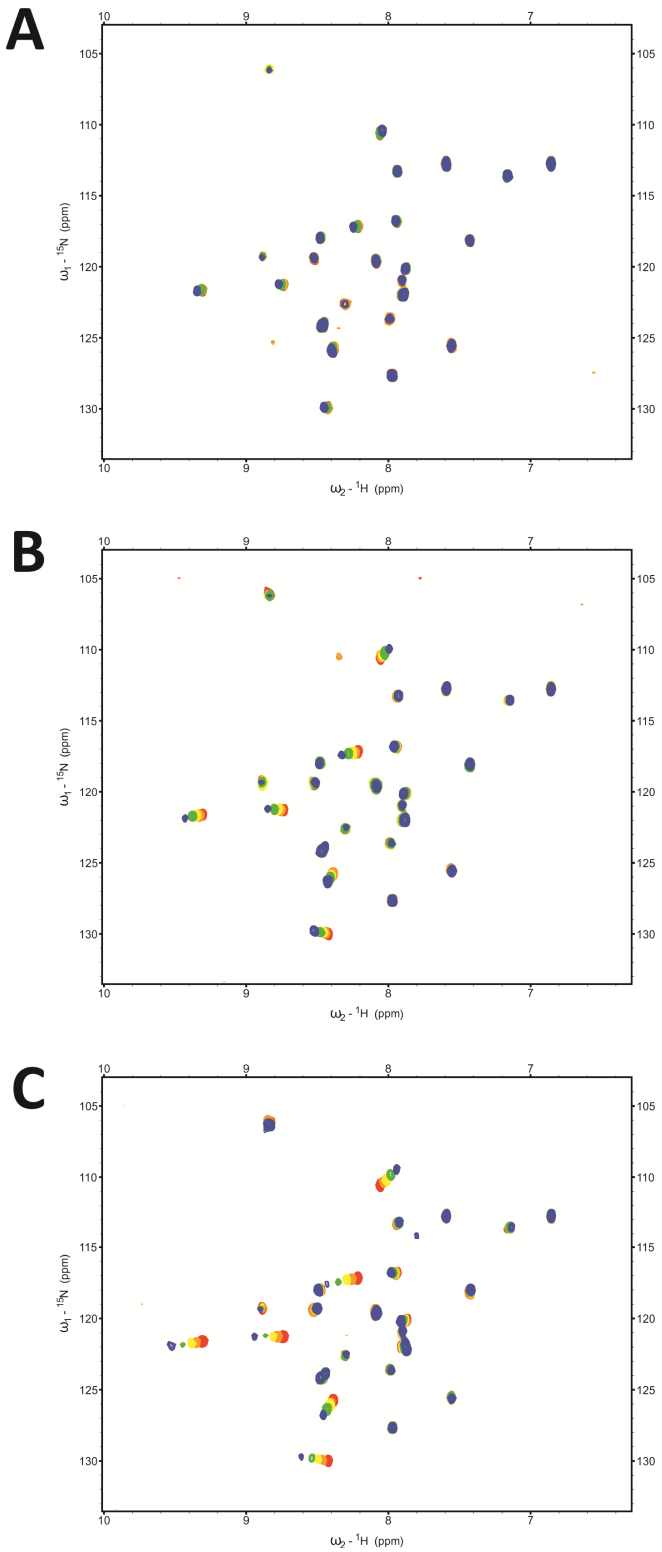


Figure 4.8 | Comparison of TUT4 CCHC-ZF3 binding to pools of random RNA oligonucleotides of increasing length. Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of 50µM TUT4 CCHC-ZF3 with A) NN at protein to RNA ratios of 1:0 (red), 1:0.5 (orange), 1:1 (yellow), 1:2 (green), and 1:8 (blue). B) NNN at protein to RNA ratios of 1:0 (red), 1:1 (orange), 1:2 (yellow), 1:4 (green), and 1:8 (blue), C) NNNN at protein to RNA ratios of 1:0 (red), 1:1 (orange), 1:2 (yellow), 1:4 (green), and 1:8 (blue).

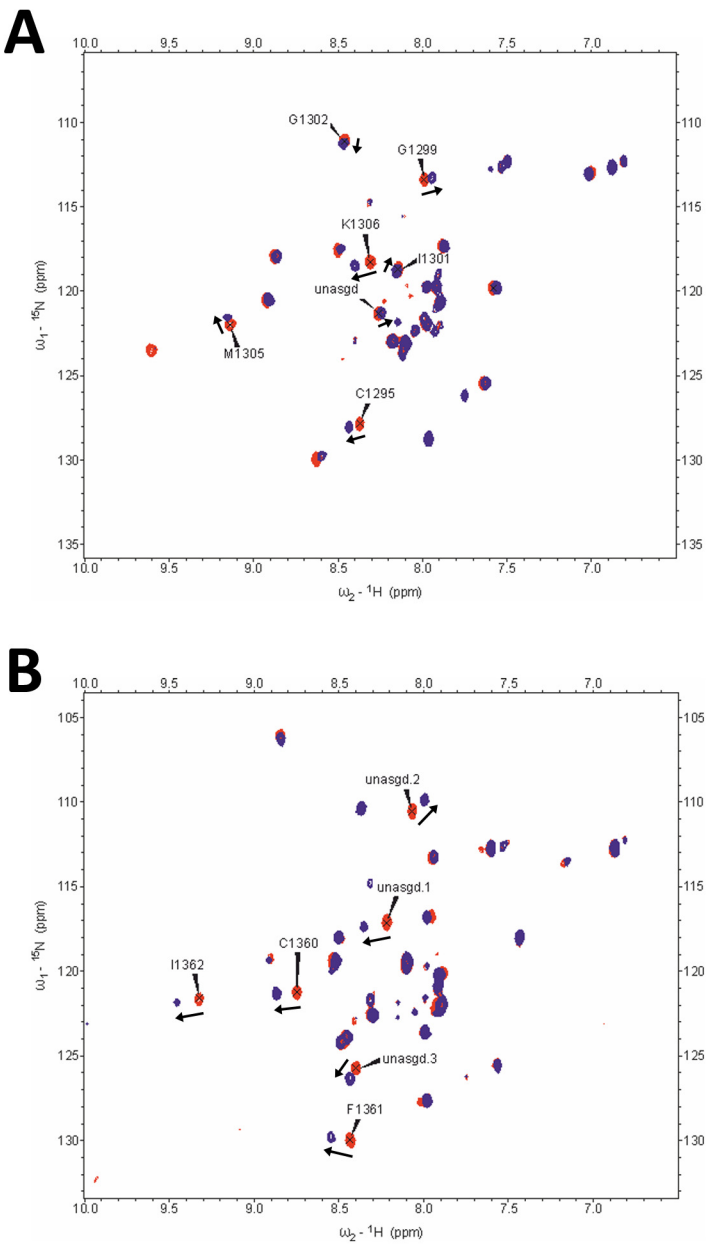


Figure 4.9 | Example spectra showing peaks used in analysis for SIA and typical chemical shift perturbation observed. A) Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of TUT4 CCHC-ZF2 with NGNN at protein to RNA ratios of 1:0 (red) and 1:4 (blue). B) Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of TUT4 CCHC-ZF3 with NGNN at protein to RNA ratios of 1:0 (red) and 1:4 (blue). Peaks used in scaffold independent analysis are labelled and arrows show the direction of peak shift perturbation.

CCHC-ZF2	Position 1	Position 2	Position 3
A	0.73	0.45	0.77
C	0.68	0.54	0.71
G	0.96	0.93	0.82
U	0.66	0.58	0.51

CCHC-ZF3	Position 1	Position 2	Position 3
A	0.81	0.55	0.99
C	0.89	0.52	0.82
G	1.00	1.00	0.94
U	0.83	0.57	0.75

Table 4.4 | SIA scores for TUT4 CCHC-ZF2 and TUT4 CCHC-ZF3.

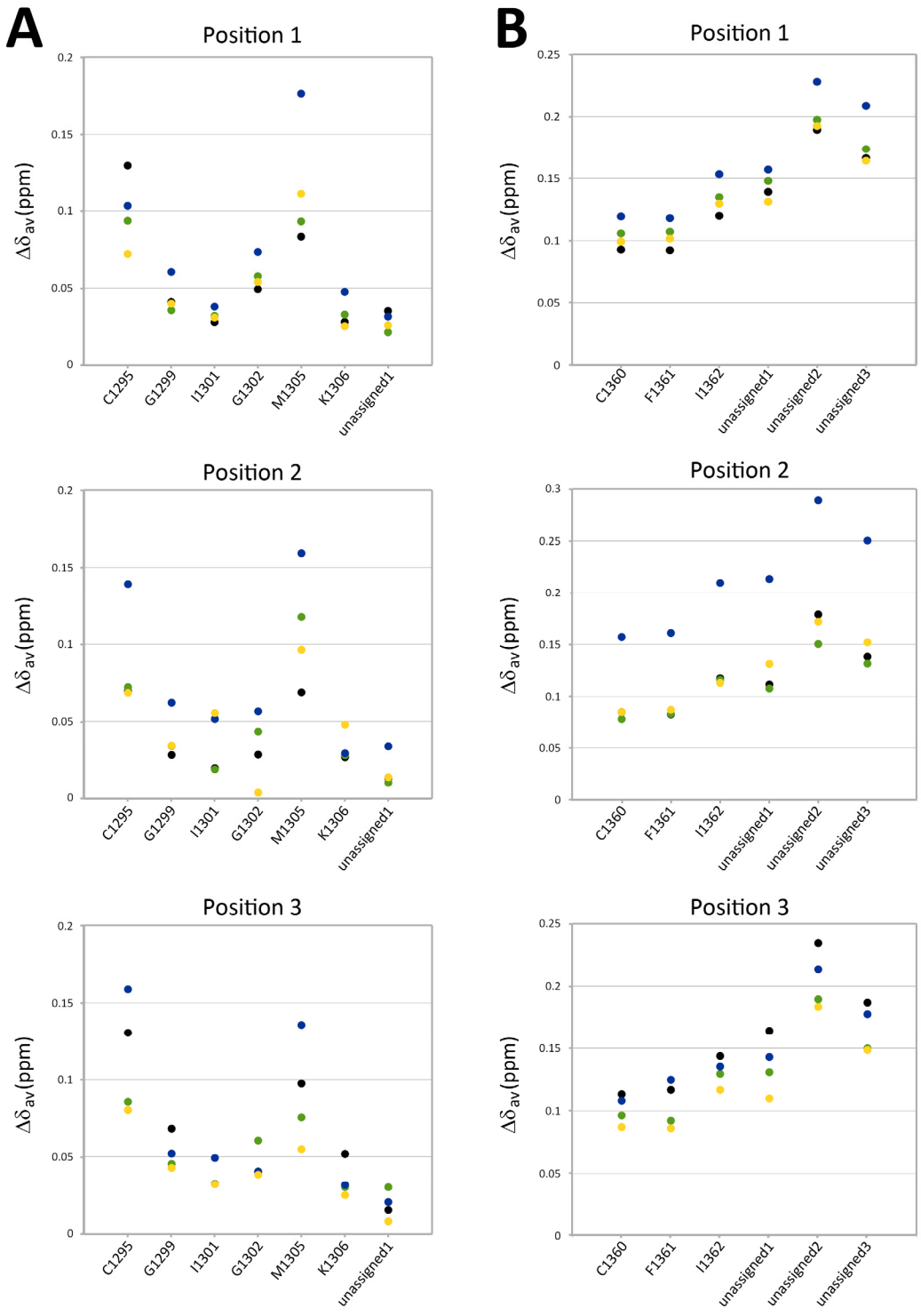


Figure 4.10 | Chemical shift perturbation of residues upon addition of RNA pools. Each plot shows the residues displayed on the x-axis and the weighted chemical shift changes for each residue on the y-axis. A pools (black), C pools (green), G pools (blue), and U pools (yellow). A) TUT4 CCHC-ZF2 for position 1 (top), position 2 (middle), and position 3 (bottom). B) TUT4 CCHC-ZF3 for position 1 (top), position 2 (middle), and position 3 (bottom).

In general both ZF show the same binding preferences. In position 1 both prefer a guanine however ZF2 is more specific than ZF3 as the other bases also have relatively high scores indicating they too can be accommodated in this position. In position 2 again both ZF show specificity for guanine. Position 2 is the most highly discriminatory position as for both fingers guanine receives high scores of 0.93 and 1.00 for ZF2 and ZF3 respectively with all the other bases achieving scores between 0.45 and 0.58. There was less of a clear preference in position 3. ZF2 discriminates slightly against uridine but the other nucleotides all achieved similar scores of between 0.71 and 0.82. ZF3 slightly prefers adenine or guanine but again is able to accommodate other nucleotides. The results show that the majority of specificity of the two fingers comes from a guanine being recognised in position 2.

To verify the nucleobase preference described by SIA we looked at the binding of CCHC-ZF2 with a hexamer containing the top ranking sequence (UGGACA) and a poly(A) oligonucleotide (AAAAAA). Titrations were monitored by NMR ^{1}H-^{15}N correlation spectroscopy. ^{1}H-^{15}N SOFAST-HMQC experiments were recorded at protein to RNA ratios of 1:0, 1:2 and 1:4. Large peak shift perturbations were observed with the UGGACA hexamer while the majority of peaks did not move at all upon addition of the poly(A) oligonucleotide (Figure 4.11). While $\Delta\delta_{av}$ is not directly correlated with affinity as different nucleotide bases may have differing effects on chemical shifts of the protein, by using data from an ensemble of residues across the protein:RNA interface one can minimise this effect and so can compare binding affinities of different oligonucleotides to a domain. This postulation is used at several points in these studies and so these titrations indicate that ZF2 binds more tightly to the sequence containing the GGA motif and so displays a preference for our SIA derived sequence.

Looking at the RNA binding of other CCHC-type zinc fingers in Lin28 and HIV-1 NCP we see that they specifically recognise a guanine[30,31], as is the case with TUT4. The fingers mostly have bulky hydrophobic residues following two of the zinc coordinating cysteines. These form a pocket into which the guanine binds, being sandwiched by the bulky residues and making hydrogen bonds with backbone groups. In Lin28 the mutation of one of these residues dramatically reduces RNA binding.[30] We mutated R1296S/Y1304S and F1360S/V1368S in ZF2 and ZF3 respectively to fully destroy the pocket and tested for protein folding and RNA binding. The domain is folded in our NMR experiments and upon

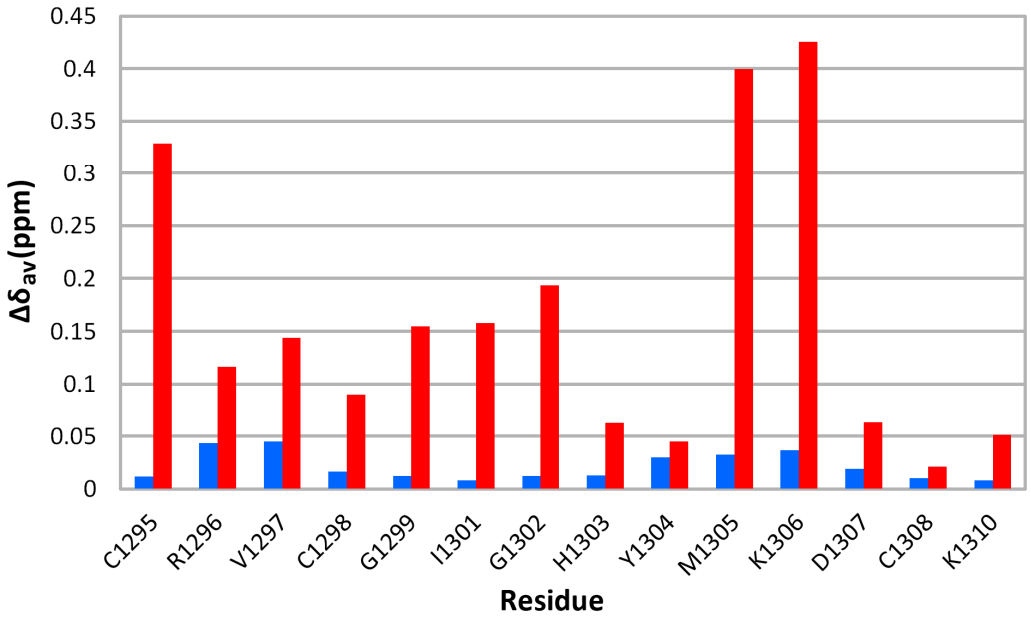


Figure 4.11 | Chemical shift perturbation of TUT4 CCHC-ZF2 residues upon addition of AAAAAA and UGGUCA. A) Chemical shift perturbation of residues at protein to RNA ratio of 1:4 for AAAAAA (blue) and UGGUCA (red). Residues are displayed on the x-axis and weighted chemical shift changes on the y-axis.

addition of NNGN at ratios of 1:2 and 1:4 we did not observe any peak shifts (Figure 4.12), while comparing this with the wild type zinc fingers we see substantial peak shifts at these ratios (Figure 4.9b). This indicates that as expected the mutations have knocked out RNA binding and therefore we expect the CCHC-ZF2 and CCHC-ZF3 of TUT4 to recognise guanine in a similar fashion to those of Lin28 and HIV-1 NCP.

4.3.3 RNA binding of CCHC-ZF2 and CCHC-ZF3

Structures of other proteins containing multiple CCHC-type zinc fingers bound to their target RNA have shown that they often work in tandem to recognise a stretch of bases eg. HIV-1 nucleocapsid protein and Lin28.[30,31] In both of these cases the zinc fingers in question were separated by a short linker of only four amino acids. In TUT4 the two most C-terminal zinc fingers are separated by a much longer linker of 44 amino acids. Therefore these zinc fingers may not be working as the other examples and may function as individual units.

We first wanted to determine if the domains interacted in the free form. The overlay of the spectra from individual domains with that of the domains joined by their physiological linker shows many of the peaks to overlap (Figure 4.13). This indicates a lack of protein-protein interactions between domains in the free form which would cause peak shifts. We recorded T1, T2 and heteronuclear NOE relaxation data CCHC-ZF23. The analysis of the relaxation data was hindered by the lack of full assignment of the construct. Therefore only chemical shifts which could be assigned to ZF2 or ZF3 were used in the analysis. This resulted in 14 chemical shifts for ZF2 and 17 chemical shifts for ZF3. While this is a large majority of the peaks it is clear that some peaks from the ZFs were excluded from the analysis. Despite these limitations, relaxation data yielded an overall picture that helps us to understand whether the two domains interact. Rotational correlation times were determined for ZF2, τ_c=4.36ns±0.63ns (Table 4.5), and for ZF3, τ_c=5.43ns±1.38ns (Table 4.6) which is consistent with the two domains not forming any stable interaction. Representative T1 and T2 plots can be found in Appendix IX.

Next we wanted to determine if the zinc fingers came together upon binding so NMR titrations were performed with RNA oligonucleotides with the sequences GGANGGA, GGANNGGA and GGANNNGGA. These oligonucleotides are made up of the preferred RNA binding sequence with a random linker of increasing length in the middle as we did not know how much space was required for both zinc fingers to fit onto the same stretch of

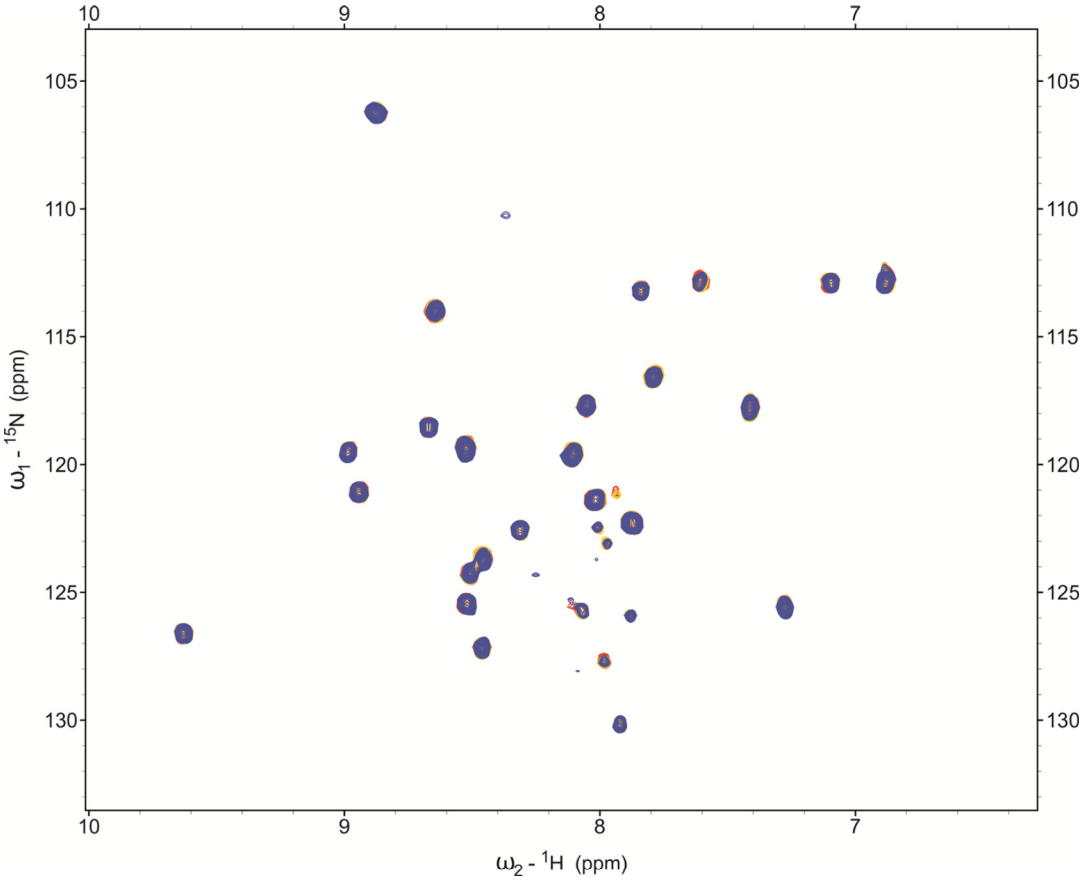


Figure 4.12 | TUT4 CCHC-ZF3 (R1360S/V1368S) upon addition of NNGN. Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of 100 µM TUT4 CCHC-ZF3 (R1360S/V1368S) with NNGN at protein to RNA ratios of 1:0 (red), 1:2 (yellow), and 1:4 (blue).

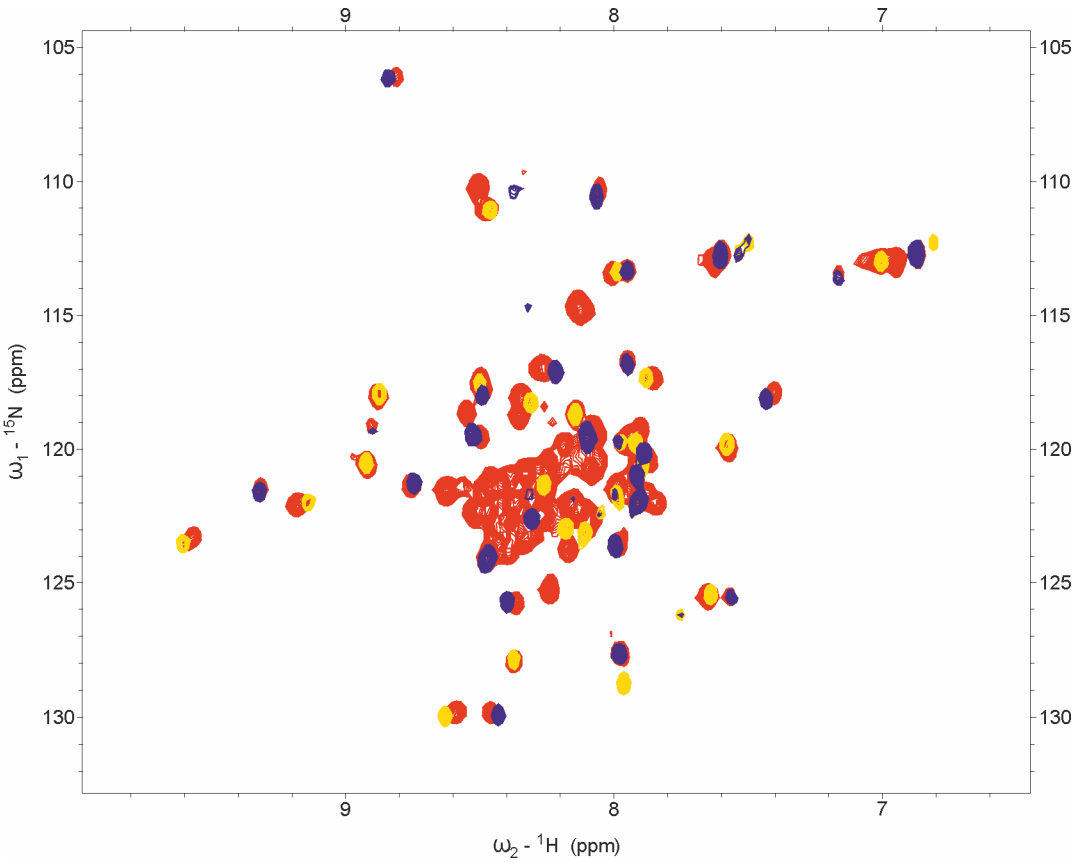


Figure 4.13 | Comparison of TUT4 CCHC-ZF23, TUT4 CCHC-ZF2 and TUT4 CCHC-ZF3. Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of TUT4 CCHC-ZF23 (red), TUT4 CCHC-ZF2 (yellow) and TUT4 CCHC-ZF3 (blue).

	N	HN	T1 (ms)	T2 (ms)	τ_c (ns)
C1295	127.886	8.274	498.889	111.629	4.371
R1296	129.759	8.496	502.034	146.436	3.617
V1297	123.201	9.467	625.851	135.926	4.459
C1298	117.346	8.403	574.544	82.504	5.791
G1299	113.398	7.906	496.204	83.115	5.271
I1301	118.550	8.039	550.336	125.380	4.318
G1302	110.922	8.373	491.013	137.140	3.737
Y1304	117.948	8.781	474.877	151.426	3.375
M1305	122.030	9.076	489.939	126.758	3.951
K1306	117.915	8.244	416.633	91.357	4.431
D1307	117.279	7.751	567.248	111.230	4.770
C1308	125.443	7.549	581.252	139.836	4.159
K1310	120.491	8.821	627.956	137.992	4.424
unsgnd.	119.855	7.472	568.571	126.284	4.393
				Average	**4.362**
				St. Dev.	**0.632**

Table 4.5 | T1, T2 and τ_c values for chemical shifts used in the analysis of CCHC-ZF2.

	N	HN	T1 (ms)	T2 (ms)	τ_c (ns)
C1360	121.261	8.658	609.910	76.720	6.264
F1361	129.759	8.357	561.867	120.686	4.492
I1362	121.395	9.222	666.279	85.092	6.208
C1363	117.781	8.393	575.959	74.489	6.163
G1364	113.230	7.847	565.456	71.947	6.222
D1365	123.737	8.357	562.142	119.744	4.517
A1366	124.071	8.377	651.619	91.438	5.871
G1367	106.037	8.698	545.761	141.700	3.941
H1369	113.465	7.062	544.796	115.531	4.531
unsgnd.	116.643	7.853	628.052	68.881	6.782
unsgnd.	117.848	7.303	508.567	70.268	5.926
unsgnd.	119.052	8.798	720.949	78.046	6.833
unsgnd.	119.454	8.400	468.773	118.664	4.013
unsgnd.	123.436	7.877	539.601	132.195	4.106
unsgnd.	125.477	7.462	1062.542	84.740	8.111
unsgnd.	125.677	8.261	692.109	100.110	5.766
unsgnd.	127.518	7.873	903.208	380.864	2.640
				Average	**5.434**
				St. Dev.	**1.376**

Table 4.6 | T1, T2 and τ_c values for chemical shifts used in the analysis of CCHC-ZF3.

RNA. All the peak shift perturbations we observed could be accounted for by the binding of the zinc fingers to the RNA. Peaks also shifted in the same direction upon RNA addition whether in the individual units or joined by the linker. We did not observe any extra peak shifts which would indicate the two protein domains coming into close proximity upon binding. Furthermore the linewidth of peaks not affected by RNA binding did not increase which would be expected if the domains came together upon binding to form one larger globular domain (Figure 4.14).

4.3.4 Characterisation of CCHC-ZF1

Titrations of CCHC-type ZF1 with a pool of random 4-mer oligonucleotides showed that it lacks the ability to bind to RNA (Figure 4.15b). A further titration with a G-rich 4-mer, assuming similar specificity to the other CCHC-type zinc fingers of the same family, also showed no binding (Figure 4.15c). Looking at an alignment of the zinc fingers of TUT4 with other CCHC-type zinc fingers it is striking that TUT4 CCHC-type ZF1 lacks the two bulky hydrophobic residues which normally create a pocket for the RNA base to slot into (Figure 4.15a). Mutations of these residues in CCHC-type ZF2 and ZF3 of TUT4 and the zinc fingers of Lin28 have shown these residues are important for RNA binding.[30] Instead ZF1 has two serines in these positions. This substitution is also seen in ZF4 of the Air2p protein which is involved in protein-protein interactions so this domain could be involved in a similar role.

4.4 Expression of multidomain TUT4 constructs

In order to investigate how the domains of TUT4 are packed together we needed to express larger constructs containing several of the structural domains. Previous studies in the literature have shown this to be a difficult protein to produce and work with.[20] We attempted to address this problem by creating constructs with many different domain boundaries to see if we could find a soluble construct amenable to structural studies. We did this using a high throughput small scale technique with Vangelis Christodoulou (National Institute for Medical Research, UK). The method allows for up to 48 different constructs and expression conditions to be tested in a week for both expression and solubility.

Domain boundaries of constructs were designed based on secondary structural elements as predicted by the algorithms SOPMA, HNN, MLRC, DPM, DSC, GOR-IV, PHD, PREDATOR and

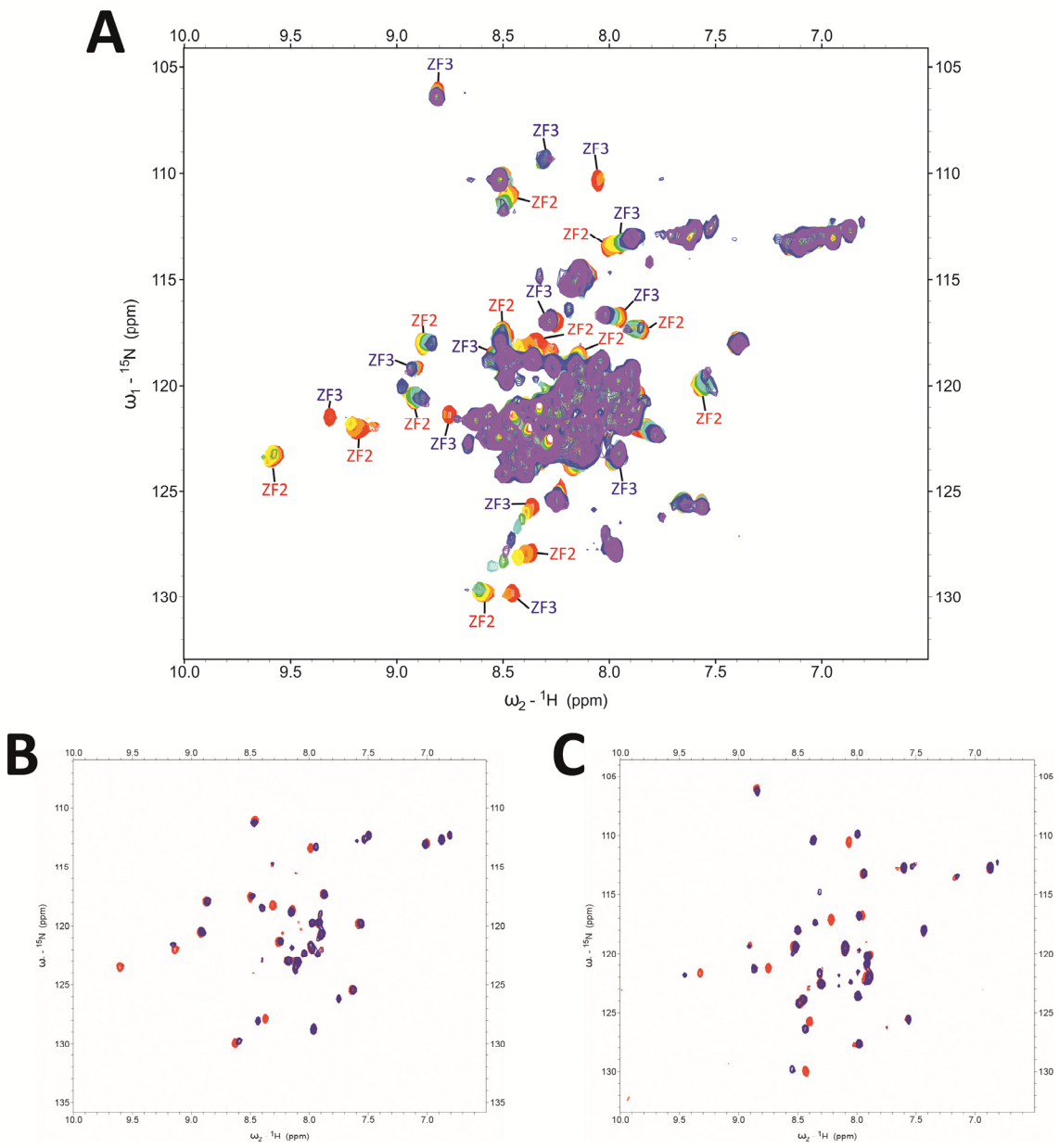


Figure 4.14 | Comparison of chemical shift perturbation of TUT4 CCHC-ZF23 upon binding of GGANNNGGA RNA oligonucleotide and binding of the individual TUT4 CCHC-ZF2 and TUT4 CCHC-ZF3. A) Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of CCHC-ZF23 with GGANNNGGA at protein to RNA ratios of 1:0 (red), 1:0.25 (orange), 1:0.5 (yellow), 1:1 (green), 1:2 (cyan), 1:4 (blue), and 1:8 (purple). Shifting peaks are labelled with the ZF they are assigned to. B) Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of CCHC-ZF2 with NNGN at protein to RNA ratios of 1:0 (red), 1:4 (blue). C) Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of CCHC-ZF3 with NNNN at protein to RNA ratios of 1:0 (red), 1:4 (blue).

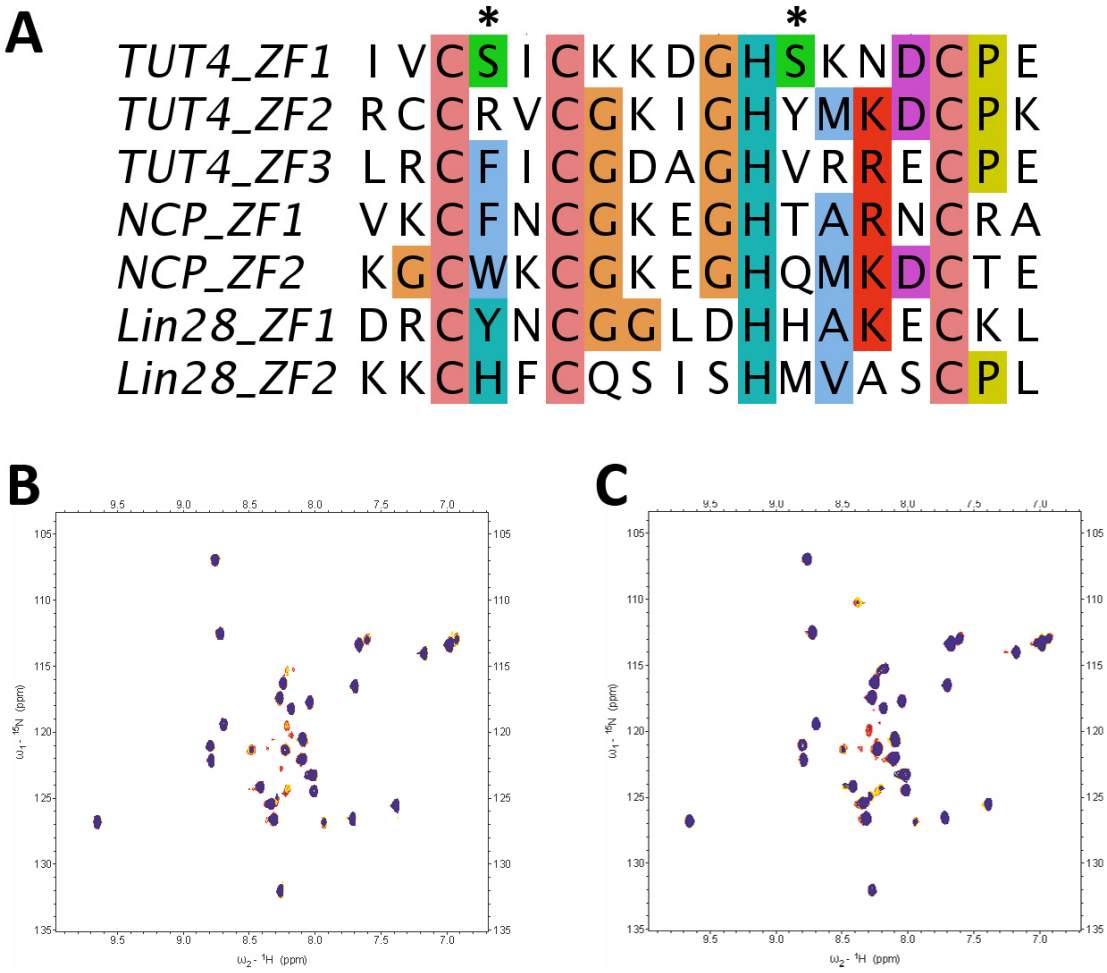


Figure 4.15 | TUT4 CCHC-ZF1 lacks ssRNA binding capabilities and two bulky residues commonly found in RNA-binding CCHC-type zinc fingers. A) Alignment of CCHC-type zinc fingers from TUT4, HIV-1 Nucleocapsid protein and Lin28. The colour scheme is Clustalx. TUT4 CCHC-ZF1 lacks bulky residues in the positions highlighted with asterisks. B) Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of CCHC-ZF1 with NNNN at protein to RNA ratios of 1:0 (red), 1:2 (yellow), and 1:4 (blue). C) Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of CCHC-ZF1 with GGAG at protein to RNA ratios of 1:0 (red), 1:2 (yellow), and 1:4 (blue).

SIMPA96. We cloned and tested the expression and solubility of several combinations of domains with each of the domains having several different N- and C-termini for the boundaries (Figure 4.16). Expression temperatures between 18^{o}C and 28^{o}C were screened along with several different tags, coexpression with chaperone proteins and different strains of *E. coli*. A full list of constructs and expression conditions can be found in Appendix VI.

No soluble product was obtained for any of the constructs over 60kDa which was not surprising with expression in the *E. coli* host. Several constructs containing regions between CCHC-ZF1 and CCHC-ZF3 showed limited solubility (Figure 4.17) however upon scale up of the prep to volumes required for analysis by NMR or for crystallography screens no purified product could be obtained. Expression conditions and tags used to improve solubility yielded no better amounts of soluble products.

The difficulty in expressing larger fragments of TUT4 was not unexpected. Other groups have reported difficulty in obtaining protein and even in mammalian cells TUT4 cannot be overexpressed with expression needing to be as close to physiological levels as possible to obtain higher percentages of active protein.[20] More recently in Thornton et al. a section of recombinant mouse TUT4 was expressed in *E. coli* and were able to perform in vitro uridylation assays.[33] However they do not state amounts of protein obtained so it is not known if this method would produce enough for structural studies on the protein. Even so this is a line of protein production it would be interesting to try.

4.5 Discussion

In this Chapter I described how we attempted to characterise the RNA binding of TUT4 focusing on the four zinc finger domains. We optimised expression and purification protocols for the production of the three CCHC-type zinc fingers however were unable to obtain a workable construct of the C2H2-type zinc finger. While this finger was shown to be unnecessary for uridylation activity in *in vitro* assays it is required for the enhanced uridylation activity observed in the presence of Lin28. Previous studies have shown mutation of two of the zinc coordinating residues to alanine abrogates this activity indicating the finger must be folded for this interaction to take place. Attempts using this interaction to help with the folding and solubility of the zinc finger were unsuccessful with the zinc finger and/or Lin28 dropping out of solution. Coexpression of the proteins did not

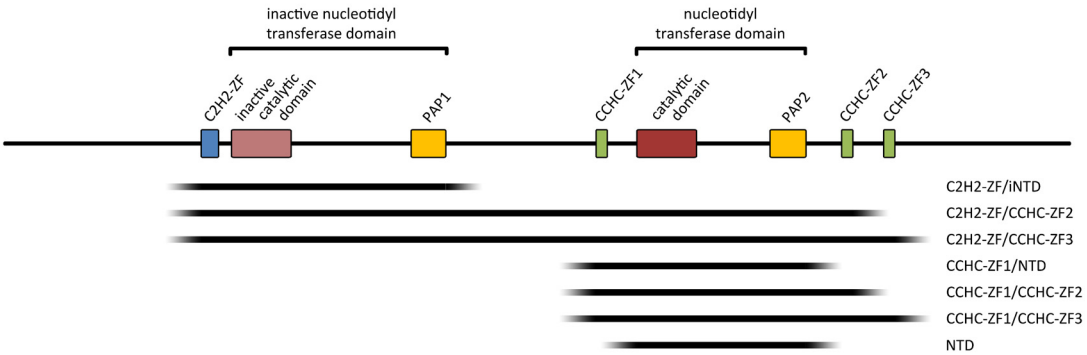


Figure 4.16 | Scheme of TUT4 multidomain constructs in small scale protein expression and solubility screen. N-terminal/C-terminal domains of the constructs are shown on the right hand side.

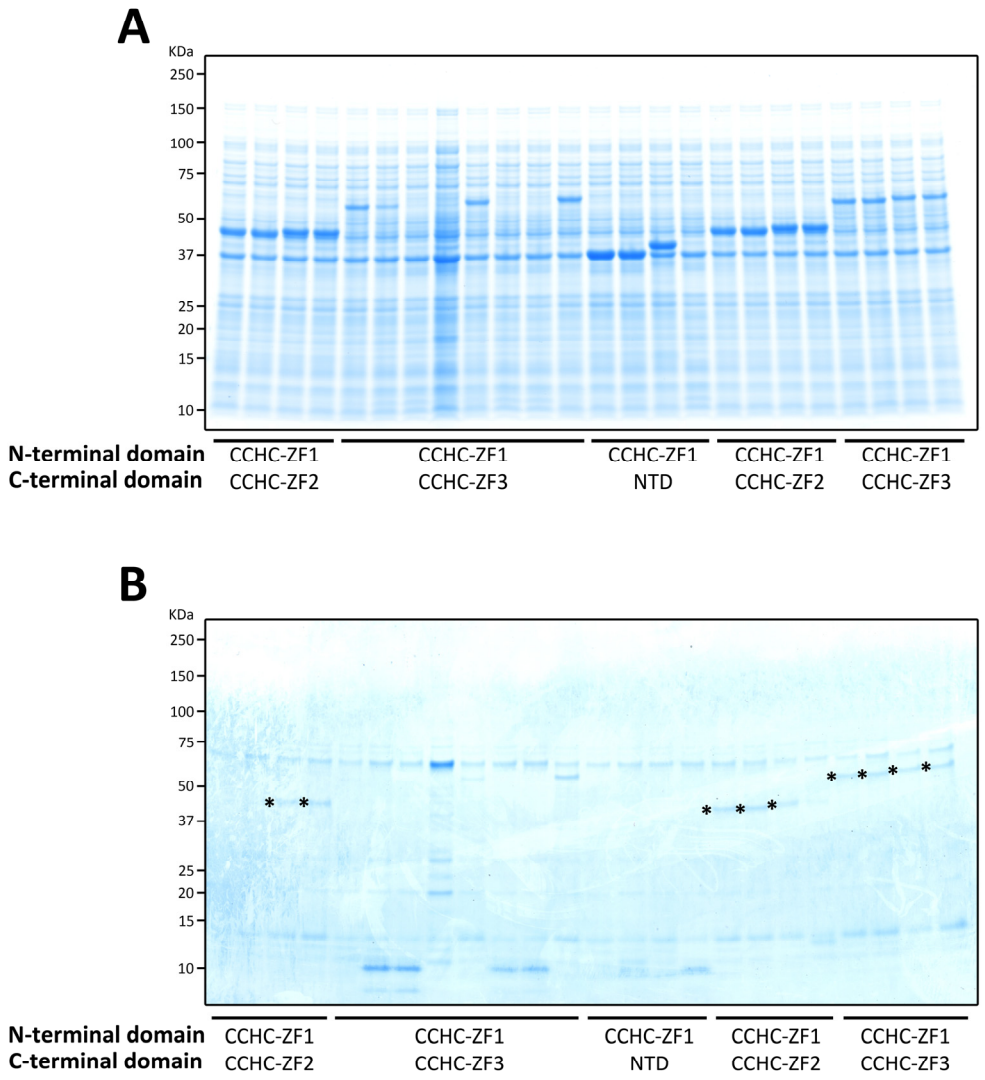


Figure 4.17 | Example of expression and solubility of TUT4 constructs in small scale screening. A) SDS-page analysis of whole cell lysate. B) SDS-page analysis of elution fractions after purification with Ni-NTA. Bands of soluble product are marked *.

increase the solubility achieved by expression of the zinc finger alone. Future work with a better quality Lin28 construct may be more successful in characterising this interaction.

Studies of CCHC-type zinc fingers 2 and 3 determined their preferred binding sequence to be GGA with the majority of the specificity being determined solely by the recognition of a guanine. Mutagenesis studies and structure and sequence alignments with other CCHC-type zinc fingers indicate that this guanine binds in a hydrophobic pocket much like other zinc fingers of this type. However unlike these other examples of zinc fingers, in which two domains bind as one unit, the fingers in TUT4 function as individual units, as seen by correlation times and lack of interaction between the domains upon RNA binding in the context of the ssRNAs we tested. The RNA binding of the zinc fingers is very weak, however truncation studies have shown that at least one of these fingers must be present for uridylation activity on the pre-let7 miRNA indicating their importance for function.[33] As we did not perform any binding experiments with structured RNAs we cannot rule out that in the context of the structured physiological target the binding mechanism may be different from that of ssRNA however the zinc fingers in Lin28 bind the stem loop of pre-let7 miRNAs and a short ssRNA sequence common to the loops in a similar manner indicating that in this case at least binding is sequence specific and not structure specific.

Using our knowledge of the specificity of the zinc fingers we attempted to find potential binding sites in physiological targets however the short preferred binding sequence made this a difficult task with GGA or other preferred sequences appearing very frequently. The short sequence recognised by the fingers and the large range of movement provided by the 44 amino acid linker between the two domains gives the domains much flexibility in the binding. This is particularly important when you look at the high levels of diversity in the RNA targets of TUT4. In Lin28 and HIV-1 NCP the zinc fingers bind to exposed bases in the loop at the end of a double stranded stem. [19,30,31] Several of the targets of TUT4 contain double stranded stretches of RNA with loops or bulges occurring sporadically leading to exposed bases, e.g. pre-let7 miRNA and histone mRNA. The regions with accessible bases occur at varying distances from each other and from the site of catalysis in the different targets and the independent action of the zinc fingers in TUT4 may give it the flexibility necessary to find and bind to such regions. Furthermore the independent action of the two zinc fingers indicates that the use of two domains would not increase affinity much (the long linker means that the overall affinity as would be not much more than sum of the

two). In contrast the linker between NTD and CCHC-ZF2 is only around 15 amino acids thus restricting its movement with regards to the catalytic domain much more than the CCHC-ZF3 which is an extra 44 amino acids away. Therefore it could be the case that where there is an exposed guanine present near to the site of uridine addition CCHC-ZF2 can bind but where this is not the case the protein has the flexibility of CCHC-ZF3 to bind to exposed bases further away.

The low affinity of the zinc fingers for the short ssRNA oligonucleotides we studied makes it likely that other regions of the protein are also involved in the interaction with the target RNA. In *Xenopus* TUT7 a basic stretch of residues in the linker between the two C-terminal CCHC-type zinc fingers binds nucleic acid *in vitro* and that five of the seven human TUTases, including TUT4, contain such regions.[34] When targeting a subset of pre-miRNA for mono-uridylation TUT4 preferentially binds family members with a 1nt 3' overhang at the end of the dsRNA stem[22] indicating the catalytic domain itself confers some specificity in target selection. In the case of oligo-uridylation in concert with Lin28 it could be that the specific binding of Lin28 to the pre-let7 miRNA and subsequent protein-protein interaction with TUT4 confers specificity as well as increases the stability of the complex. Other targets of TUT4 are also often found in protein-RNA complexes, for example mature miRNAs are not commonly found as naked RNA and are instead incorporated into the RISC complex and the 3' stem loop in histone mRNA has numerous proteins bound to it throughout its lifetime.[7,24] Therefore in these systems protein-protein interactions, maybe mediated by C2H2-ZF or CCHC-ZF1, could also aid in the specificity and/or affinity of binding.

We have shown that CCHC-ZF1 is unable to bind to short ssRNA oligonucleotides probably due to the lack of a hydrophobic pocket. This finger could therefore also act to mediate protein-protein interactions in a similar way to the C2H2-zinc finger, which has been shown to interact with Lin28 providing a stable interaction which enables TUT4 to add a long uridyl tail to pre-let7 miRNAs.[20,33] As CCHC-ZF1 is joined to the catalytic domain by approximately 15 amino acids it could help with the positioning of the catalytic domain near to its site of action.

The findings that CCHC-ZF2 and ZF3 recognise a short stretch of RNA specifically with low affinity and act as individual units fits well with our knowledge of the physiological targets of TUT4 potentially giving the protein large flexibility in target recognition and binding. In

the future, if a suitable construct containing the catalytic domain and surrounding zinc fingers could be obtained, it would be interesting to investigate the packing of the domains, how they work in combination to recognise various RNA targets and if the domains function in different ways depending on the target RNA and other bound proteins.

4.6 References

1. Kwak, J. E. & Wickens, M. A family of poly(U) polymerases. *RNA* **13,** 860–7 (2007).

2. Holm, L. & Sander, C. DNA polymerase beta belongs to an ancient nucleotidyltransferase superfamily. *Trends Biochem. Sci.* **20,** 345–7 (1995).

3. Bai, Y., Srivastava, S. K., Chang, J. H., Manley, J. L. & Tong, L. Structural Basis for Dimerization and Activity of Human PAPD1, a Noncanonical Poly(A) Polymerase. *Mol. Cell* **41,** 311–20 (2011).

4. Fasken, M. B. *et al.* Air1 Zinc Knuckles 4 and 5 and a Conserved IWRXY Motif Are Critical for the Function and Integrity of the Trf4/5-Air1/2-Mtr4 Polyadenylation (TRAMP) RNA Quality Control Complex. *J. Biol. Chem.* **286,** 37429–45 (2011).

5. Rammelt, C., Bilen, B., Zavolan, M. & Keller, W. PAPD5, a noncanonical poly(A) polymerase with an unusual RNA-binding motif. *RNA* **17,** 1737–46 (2011).

6. Trippe, R. *et al.* Identification, cloning, and functional analysis of the human U6 snRNA-specific terminal uridylyl transferase. *RNA* **12,** 1494–504 (2006).

7. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116,** 281–97 (2004).

8. Kloosterman, W. P. & Plasterk, R. H. a. The diverse functions of microRNAs in animal development and disease. *Dev. Cell* **11,** 441–50 (2006).

9. Büssing, I., Slack, F. J. & Grosshans, H. let-7 microRNAs in development, stem cells and cancer. *Trends Mol. Med.* **14,** 400–9 (2008).

10. Johnson, S. M. *et al.* RAS is regulated by the let-7 microRNA family. *Cell* **120,** 635–47 (2005).

11. Koscianska, E. *et al.* Prediction and preliminary validation of oncogene regulation by miRNAs. *BMC Mol. Biol.* **8,** 79 (2007).

12. Lee, Y. S. & Dutta, A. The tumor suppressor microRNA let-7 represses the HMGA2 oncogene. *Genes Dev.* **21,** 1025–30 (2007).

13. Johnson, C. D. *et al.* The let-7 microRNA represses cell proliferation pathways in human cells. *Cancer Res.* **67,** 7713–22 (2007).

14. Heo, I. *et al.* TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. *Cell* **138,** 696–708 (2009).

15. Yang, D.-H. & Moss, E. G. Temporally regulated expression of Lin-28 in diverse tissues of the developing mouse. *Gene Expr. Patterns* **3,** 719–726 (2003).

16. Piskounova, E. *et al.* Lin28A and Lin28B Inhibit let-7 MicroRNA Biogenesis by Distinct Mechanisms. *Cell* **147,** 1066–79 (2011).

17. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31,** 3406–15 (2003).

18. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318,** 1917–20 (2007).

19. Nam, Y., Chen, C., Gregory, R. I., Chou, J. J. & Sliz, P. Molecular basis for interaction of let-7 microRNAs with Lin28. *Cell* **147,** 1080–91 (2011).

20. Yeom, K.-H. *et al.* Single-molecule approach to immunoprecipitated protein complexes: insights into miRNA uridylation. *EMBO Rep.* **12,** 690–6 (2011).

21. Ustianenko, D. *et al.* Mammalian DIS3L2 exoribonuclease targets the uridylated precursors of let-7 miRNAs. *RNA* **19,** 1632–8 (2013).

22. Heo, I. *et al.* Mono-Uridylation of Pre-MicroRNA as a Key Step in the Biogenesis of Group II let-7 MicroRNAs. *Cell* **151,** 521–32 (2012).

23. Park, J.-E. *et al.* Dicer recognizes the 5’ end of RNA for efficient and accurate processing. *Nature* **475,** 201–5 (2011).

24. Marzluff, W. F., Wagner, E. J. & Duronio, R. J. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nat. Rev. Genet.* **9,** 843–54 (2008).

25. Mullen, T. E. & Marzluff, W. F. Degradation of histone mRNA requires oligouridylation followed by decapping and simultaneous degradation of the mRNA both 5’ to 3' and 3' to 5'. *Genes Dev.* **22,** 50–65 (2008).

26. Schmidt, M.-J., West, S. & Norbury, C. J. The human cytoplasmic RNA terminal U-transferase ZCCHC11 targets histone mRNAs for degradation. *RNA* **17,** 39–44 (2011).

27. Tomecki, R., Dmochowska, A., Gewartowski, K., Dziembowski, A. & Stepien, P. P. Identification of a novel human nuclear-encoded mitochondrial poly(A) polymerase. *Nucleic Acids Res.* **32,** 6001–14 (2004).

28. Jones, M. R. *et al.* Zcchc11 Uridylates Mature miRNAs to Enhance Neonatal IGF-1 Expression, Growth, and Survival. *PLoS Genet.* **8,** e1003105 (2012).

29. Jones, M. R. *et al.* Zcchc11-dependent uridylation of microRNA directs cytokine expression. *Nat. Cell Biol.* **11,** 1157–63 (2009).

30. Loughlin, F. E. *et al.* Structural basis of pre-let-7 miRNA recognition by the zinc knuckles of pluripotency factor Lin28. *Nat. Struct. Mol. Biol.* **19,** 84–9 (2012).

31. De Guzman, R. N. *et al.* Structure of the HIV-1 nucleocapsid protein bound to the SL3 psi-RNA recognition element. *Science* **279,** 384–8 (1998).

32. Mooij, W. T. M., Mitsiki, E. & Perrakis, A. ProteinCCD: enabling the design of protein truncation constructs for expression and crystallization experiments. *Nucleic Acids Res.* **37,** W402–5 (2009).

33. Thornton, J. E., Chang, H.-M., Piskounova, E. & Gregory, R. I. Lin28-mediated control of let-7 microRNA expression by alternative TUTases Zcchc11 (TUT4) and Zcchc6 (TUT7). *RNA* **18,** 1875–85 (2012).

34. Lapointe, C. P. & Wickens, M. The nucleic acid-binding domain and translational repression activity of a Xenopus terminal uridylyl transferase. *J. Biol. Chem.* **288,** 20723–33 (2013).

5. Fragile X Mental Retardation Protein (FMRP)

5.1 Introduction

5.1.1 Fragile X Syndrome

Fragile X syndrome (FXS) is the most common form of inherited intellectual disability affecting around 1 in 5000 males.[1] The syndrome can cause mild to severe intellectual disability along with autism-like behaviours and increased susceptibility to seizures and post-mortem examinations of individuals with FXS reveal neurons with dense and immature dendritic spines.[2]

In 1991 the gene linked with FXS was identified and named as fragile X mental retardation gene 1 (*FMR1*).[3] Most cases of FXS are caused by expansion of CGG repeats in the 5'UTR of the gene. Unaffected individuals carry between 5 and 50 repeats, the permutation allele contains 50 to 200 repeats, while affected individuals carry more than 200 repeats.[4] The large expansion leads to hypermethylation of the CGG repeats and promoter of the FMR1 gene resulting in transcriptional silencing.[5] This epigenetic silencing is mediated by the hybridisation of the CGG repeats in the FMR1 mRNA itself with the complementary sequence in the gene.[6] The subsequent loss of the fragile X mental retardation protein (FMRP)[7] and the downstream consequences are believed to be at the core of the disease pathophysiology.

5.1.2 Fragile X mental retardation protein

FMRP is an RNA binding protein that is highly expressed in the brain and testes.[8] It is a key regulator of translation, interacting with selective mRNA targets and repressing translation.[9] The main isoform of FMRP present in humans is around 71kDa and contains three RNA binding domains; an RGG box and two KH domains. Other features of the protein are two tandem Agenet domains at the N-terminus which are thought to bind trimethylated lysine residues[10] and nuclear localising and export signals which allow its shuttling in and out of the nucleus.[11]

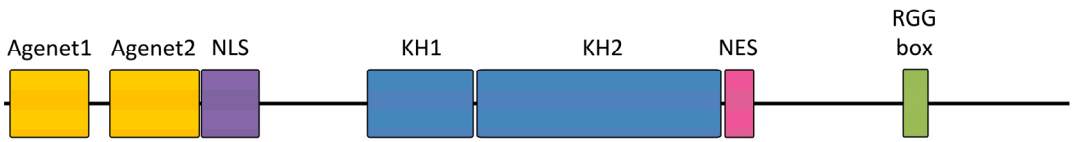


Figure 5.1 | Domain organisation of FMRP. Agenet 1 and Agenet 2, tandem Agenet domains; NLS, nuclear localization signal; KH1 and KH2, K homology domains 1 and 2; NES, nuclear export signal; RGG, arginine-glycine-glycine box.

FMRP plays a critical role in the modulation of synaptic plasticity, which is the ability of synapses to strengthen or weaken over time in response to increases or decreases in activity. Two well characterised mechanisms by which this can occur are long term potentiation (LTP), an increase in synaptic response, and long term depression (LTD), a decrease in synaptic response. These processes are central to learning and memory. As FMRP participates in the regulation of numerous forms of synaptic plasticity the loss of the protein and disruption of these processes could lead to some of the symptoms and the abnormal synapse architecture observed in the syndrome.[9]

Many identified mRNA targets of FMRP are involved in these pathways. One of the most well studied pathways affected by loss of FMRP is metabotropic glutamate receptor-dependent (mGluR) LTD and Bear et al. postulated the mGluR theory of fragile X mental retardation. Stimulation of mGluR leads to rapid activation of protein phosphatase 2 (PP2A), the main phosphatase targeting FMRP. Upon dephosphorylation FMRP no longer inhibits translation and there is a burst of local protein synthesis. These newly synthesised proteins facilitate the internalisation of α-amino-3-hydroxyl-4-isoxazole propionic acid receptors (AMPAR) which leads to LTD. In the absence of FMRP LTD is exaggerated as the proteins required for internalisation are no longer translationally repressed and are constantly present in the cell (Figure 5.2).[12]

FMRP can repress protein synthesis by directly repressing translation of its target mRNA. It is also involved in a more general mechanism of repression of protein synthesis in the synapse through the regulation of components of the Mammalian target of rapamycin (mTOR) pathway. Normally activation of this pathway by mGluR signalling leads to the inhibition of eukaryotic translation initiation factor 4E binding protein (4E-BP) thus relieving the repression of cap-dependent translation. This allows for a burst of protein synthesis at the synapse. FMRP negatively regulates this pathway by targeting the mRNA of two of the upstream kinases, Phosphatidylinositol 3-kinase (PI3K) and Phosphoinositide 3-kinase enhancer (PIKE). Therefore upon loss of FMRP mTOR signalling is at or near to saturation, even in the absence of stimulation of the mGluR, and therefore insensitive to mGluR activation (Figure 5.2).[9]

Several mechanisms have been suggested as to how FMRP performs its function as a repressor of translation. One model is that FMRP directly interacts with the L5 protein in

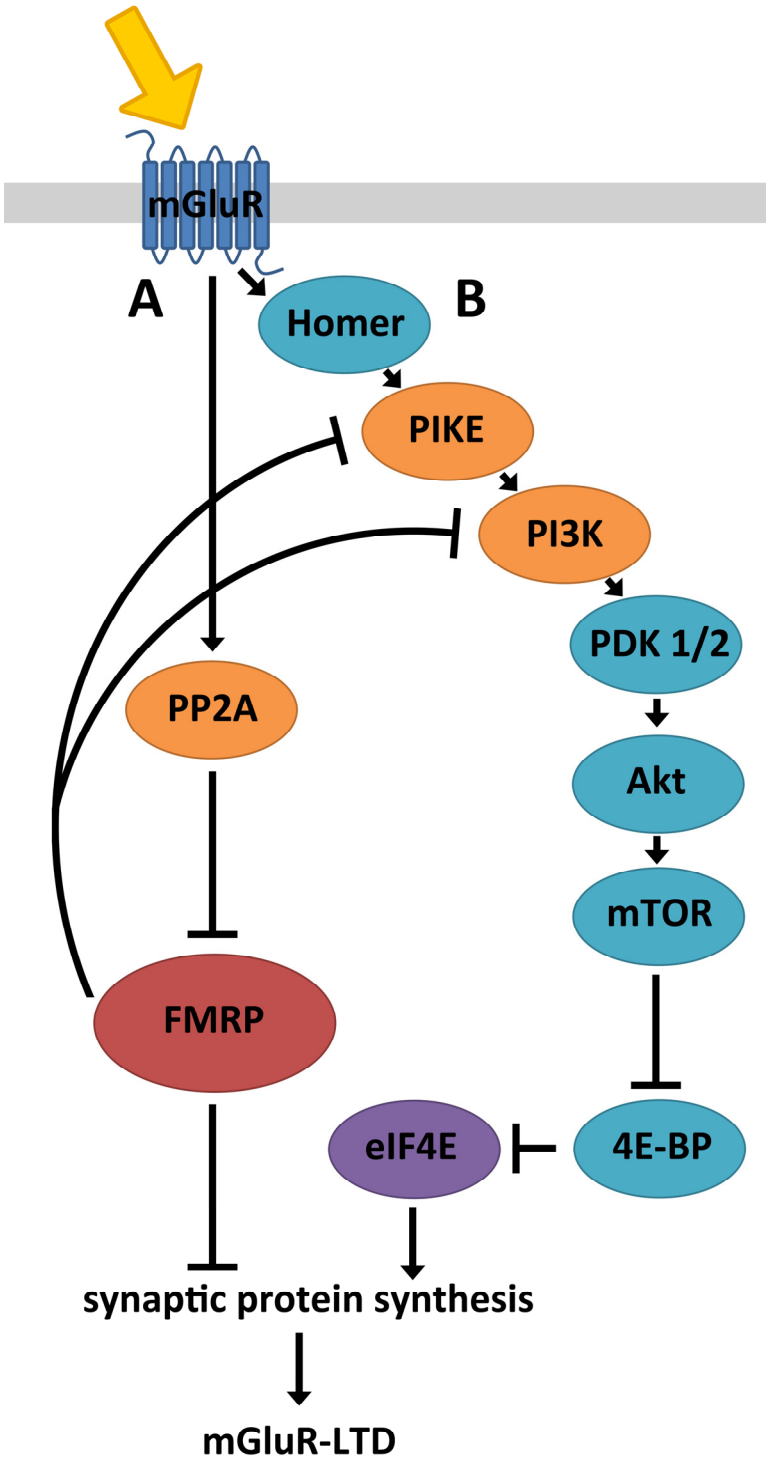


Figure 5.2 | FMRP in mGluR signalling pathways. A) Stimulation of mGlurR activates PP2A which dephosphorylates FMRP leading to its repression. This relieves the repression of FMRPs targets which can now be synthesised leading to mGluR-LTD. B) Stimulation of mGluR activates the mTOR pathway leading to the repression of 4E-BP. This relieves the repression of eIF4E and protein synthesis can proceed leading to mGluR-LTD. Adapted from [9]

the 80S ribosome and blocks the binding of tRNA and translation elongation factors. A cryo-electron microscopy structure suggests the two KH domains bind to the ribosome while the RGG box contacts the mRNA allowing for target selection.[13] However RNA binding data conflicts this suggesting that at least two of the RBDs work cooperatively to recognise mRNA targets.[14] A second model is that FMRP recruits Cytoplasmic FMR1-interacting protein 1 (CYFIP1) to the target mRNA. CYFIP1 subsequently associates with eukaryotic translation initiation factor 4E (eIF4E) and blocks recruitment of the translation initiation machinery.[15]

5.1.3 RNA binding of FMRP

Many research efforts over the years have focused on identifying the RNA targets of FMRP. By elucidating the target genes it is hoped that one can define the pathways affected by the loss of the protein. A better molecular understanding of the disease would allow more focused efforts for designing therapeutics. Secondly as more target gene sets have been described there has been an emerging correlation between the genes targeted by FMRP and those involved in neuropsychiatric diseases such as autism[16] and schizophrenia.[17] It is therefore hoped that the identification of authentic FMRP targets could also be useful in studying the molecular mechanisms underlying these related diseases.

Several large scale studies have tried to identify sets of target genes using techniques such as Photoactivatable-Ribonucleoside-Enhanced Crosslinking Immunoprecipitation (PAR-CLIP),[18] RNP Immunoprecipitation-Microarray (RIP-Chip),[19] High-throughput Sequencing of RNA Isolated by Crosslinking Immunoprecipitation (HITS-CLIP)[20] and Antibody-Positioned RNA Amplification (APRA).[21] Each of these techniques identified hundreds of putative targets and an analysis of the four largest datasets found large overlap and generated high-confidence consensus lists.[14] However from this large pool of potential targets surprisingly few have been verified to show direct biochemical interaction. Verified targets include the mRNA of Calcium/calmodulin-dependent protein kinase IIα (CamKIIα) which is involved in many signalling cascades and is thought to be an important mediator of learning and memory;[22] eukaryotic translation initiation factor 1A (eIF1A);[23] GluR1 and GluR2 which are receptors or subunits of receptors found at neuronal synapses;[22] the signalling molecule Semaphorin3F (Sema3F);[24] and its own transcript Fmr1.[25] Many of the other target genes were also found to be involved in processes such as neuronal and synaptic transmission which are coherent with the predicted role of FMRP.

As well as searching for target genes data from these large scale studies, in conjunction with in vitro biochemical work, have been used to identify potential binding motifs (Summarised in Table 5.1).

G-quadruplex

The RRG box in the C-terminal of FMRP has been reported to bind to RNA structures known as G-quadruplexes. A 'classical' G-quadruplex is formed by two to four G-quartets (four guanine residues arranged in a planar conformation, stabilised by Hoogsteen-type hydrogen bonds) stacking on top of one another.[26] The different runs of guanines are connected by unpaired nucleotides which can be organised in an array of different topologies. Such structures mediate the interaction of FMRP with Fmr1,[25] Microtubule-associated protein 1B (MAP1b)[27] and Sema3F mRNA.[28] Many other putative targets of FMRP have been predicted to contain G-quadruplexes and characteristic G-quadruplex forming motifs are enriched in sets of targets genes.[14] Further, a SELEX study identified a G-quartet which binds to FMRP RGG box with high affinity and a structure of this complex has been solved by NMR, although it is unclear whether this structure is found in physiological targets.[24,29]

SoSLIP

FMRP binds to superoxide dismutase 1 (Sod1) mRNA via a structure of three stem loops separated by stretches ssRNA. The structure has been named Sod1 stem loops interacting with FMRP (SoSLIP). The C-terminal region of FMRP mediates the binding and binding competition between SoSLIP and G-quadruplex containing RNA indicates the RGG box is likely to play a role in the recognition of the structure.[30]

Kissing complex

The kissing complex is a secondary structure of two stem loops with Watson-Crick base pairs between the two loops. It is reported to bind to the KH2 domain of FMRP. It is thought that this interaction by KH2 is important for function as the I304N mutation in the domain abrogates the binding to the kissing complex. Furthermore the association of FMRP with brain polyribosomes, which is important in the mechanism of translational repression, can be competed off with the kissing complex.[31] While this specific RNA secondary structure was formed by a selection of randomly synthesised RNA and has not been observed in endogenous mRNA, similar loop-loop RNA structures are observed in tRNA[32] and the Varkud satellite ribozyme in *Neurospora*.[33]

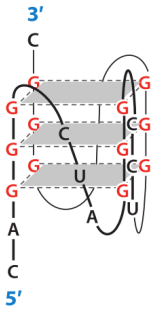
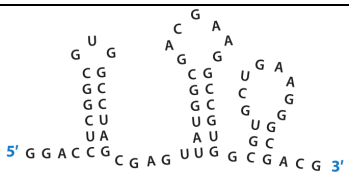
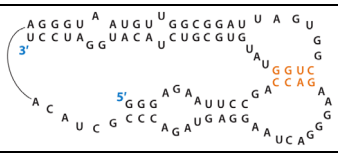
Name	Structure	Domain	Identification method	Found in physiological targets
G-quadruplex		RGG box	SELEX	Fmr1 MAP1b Sema3F
SoSLIP		RGG box	Filter binding assay	Sod1
Kissing complex		KH2	SELEX	none
U-rich pentamer		Unknown	cDNA SELEX	hASH1
ACUK		KH2	Enriched in PAR-CLIP dataset	APC, APP, KDM5C, MAP1B, NF1, UBE3A
WGGA		Unknown	Enriched in PAR-CLIP dataset	APP, FMR1, KDM5C, NF1
GAC		KH1 and/or KH2	RNAcompete	APC, APP, FMR1, NF1

Table 5.1 | Table of identified RNA target motifs of FMRP.

U-rich sequences

cDNA SELEX (Systematic Evolution of Ligands by Exponential Enrichment) identified a class of FMRP target mRNAs containing U-rich pentamers.[34] UV cross-linking and mutagenesis assays have also identified FMRP binding to a U-rich region of one of its targets, human achaete-scute homologue 1 (hASH1) mRNA.[35] However it is not known which domain of FMRP mediates this binding.

ACUK and WGGA

The search for enriched motifs in putative targets identified using PAR-CLIP lead to the identification of two consensus sequences, ACUK and WGGA (K=U or G, W=U or A). Binding assays comparing FMRP WT with FMRP I304N lead to the conclusion that ACUK was the binding motif for KH2.[18] However a meta-analysis of several large scale studies showed the ACUK motif not to be enriched and while the WGGA motif was found to be enriched, this was only when searching for the sequence in clusters.[14] This indicates a role in G-quadruplex formation and therefore RGG box binding.

GAC

A recent study identified the target motifs for hundreds of RNA binding proteins using a technique called RNAcompete. This found a seven nucleotide consensus sequence for FMRP containing a conserved GAC core.[36] The motif is found to be enriched in several of the large scale data sets of targets strengthening its position as a target motif.[14]

Interaction via adapter molecule BC1

To further complicate matters there is evidence that FMRP interacts with target mRNAs via adapter molecules such as brain cytoplasmic 1 (BC1). This is a small non-coding RNA that base pairs with the target mRNA and mediates the interaction with FMRP.[37,38] However other research disputes this, observing that FMRP can bind to its target molecules without the aid of an adaptor and finding no evidence of specific BC1/FMRP interactions.[39]

5.1.4 Aims

The plethora of data acquired from large scale pull down studies, while invaluable, does not answer many of the key questions about FMRP binding. As with numerous other proteins the targets for multicomponent binding systems are difficult to rationalise, in particular with regards to the roles of the separate binding domains. In this study we have started to deconvolute the complex FMRP-RNA recognition by looking at the capability of individual domains to recognise RNA. We have focused on the KH domains and in particular

on KH1, for which a role in RNA binding has not yet been defined. As described earlier, SIA provides a full account of the nucleobase preference of the different domains and is tailored to the analysis of domains that bind RNA with a low-to-intermediate affinity.[40] However, the two KH domains of FMRP are structurally joined and the affinity of the individual KH1 domain is extremely low. In this chapter we describe how we have used a mutagenesis strategy to isolate the RNA binding properties of each individual domain and increase the affinity of KH1, so to be able to analyse its RNA binding properties by SIA.

5.2 Methods

5.2.1 Cloning

The gene encoding FMRP with the deletion mutation Δ331-396, codon optimised for expression in *E. coli*, was purchased from Eurofins. The alignment of the deletion mutation and full length FMRP amino acid sequences can be found in Appendix X and numbering throughout this chapter relates to FMRP Δ331-396 sequence. Primers were designed with Crystallisation Construct Designer[32] and used to amplify the region of DNA encoding for the KH1 and KH2 domains while introducing 5' and 3' extensions complementary to sections of the vector to produce the inserts. The vector used was pET-47b (provided by Vangelis Christodoulou, National Institute for Medical Research) which contains an N-terminal hexa-histidine tag cleavable by Human Rhinovirus 3C protease and a resistance marker to kanamycin. Vectors were linearised by digestion with BsaI and then both inserts and linearised vector were treated with T4 DNA polymerase to produce complementary single stranded overhangs. Insert and vector were mixed and allowed to anneal before 2 µl was used to transform 20 µl of BL21Gold(DE3) cells by a standard heat shock protocol. Transformed cells were allowed to recover in LB then plated out onto agar containing kanamycin for selection of successfully transformed cells. Colonies were picked and grown overnight in LB at 37^{o}C before cells were harvested, plasmid DNA extracted using Quantum Prep Plasmid Miniprep Kit (BioRad) following the manufacturer's instructions and sequenced by Beckman Coulter.

Construct Name	Start residue	End residue	Forward Primer	Reverse Primer
FMRP KH12 (216-359)	216	359	FMRP_216FW	FMRP_259RV
FMRP KH12 (212-383)	216	383	FMRP_212FW	FMRP_383RV
FMRP KH12 (212-405)	212	405	FMRP_212FW	FMRP_405RV

Table 5.2 | Table of FMRP KH domain constructs and primers used in cloning. Sequences of primers can be found in Appendix V.

5.2.2 Site-directed mutagenesis

Primers were designed to introduce the mutations T236D/H237D and T236K/H237K into KH1, and K299D/N300D into KH2 Point mutations were introduced into the constructs by amplification of the plasmid using overlapping complementary primers with the mutation of interest inserted at the centre of the oligonucleotides. Following PCR amplification parent DNA was removed by DpnI digestion.

Construct Name	Start residue	End residue	Mutations	Primers
FMRP KH1DD/KH2WT	216	359	T236D/H237D	FMRP_KH1_gddgFW FMRP_KH1_gddgRV
FMRP KH1WT/KH2DD	216	359	K299D/N300D	FMRP_KH2_gddgFW FMRP_KH2_gddgRV
FMRP KH1WT/KH2DD	216	359	T236K/H237K	FMRP_KH1_gkkgFW FMRP_KH1_gkkgRV

Table 5.3 | Table of FMRP KH mutant constructs and primers used in site-directed mutagenesis. Sequences of primers can be found in Appendix V.

5.2.3 Protein expression

50 µl of *E. coli* BL21(DE3) cells were transformed with 2µl of plasmid using a standard heat shock protocol. 500µl of transformed cells were used to inoculate 100ml of M9 minimal media containing $(^{15}NH_4)_2SO_4$ as the only nitrogen source and/or ^{13}C6-D-glucose as the only carbon source. For $^{2}H^{13}C^{15}N$-labelled protein cells were grown in a $^{2}H_2O$ solution of the above media. Cells were grown overnight at 37^{o}C. The overnight culture was used to inoculate 1000ml of M9 minimal media to an OD_{600} of 0.1. Cells were then grown to an OD_{600} of 0.6 before protein expression was induced with IPTG at a final concentration of 0.5mM. Cells were grown for a further 4 hours at 37^{o}C, harvested by centrifugation and stored at -80^{o}C.

5.2.4 Protein purification

Frozen cells were resuspended in equilibration buffer (10 mM Tris-HCl pH 8.0, 10 mM imidazole, 200 mM NaCl, 2 mM β-mercaptoethanol) (20 ml per 1 L of cell culture) with Triton X-100, DNaseI and lysozyme, sonicated on ice (Branson Sonifier 250, power output 50 W, 60% duty cyle, 2x45 seconds) and centrifuged at 17000 rpm for 60 mins. The recombinant protein was purified by immobilised metal ion affinity chromatography (IMAC) columns. The soluble fraction was incubated with Ni-NTA resin (5ml per litre of culture) at 4^{o}C for 30 mins then poured into a gravity-driven column. The column was washed with 10 CV of wash buffer (10 mM Tris-HCl pH 8.0, 10 mM Imidazole, 1 M NaCl, 2 mM β-mercaptoethanol) and the protein was eluted with 5 CV of elution buffer (10 mM Tris-HCl pH 8.0, 250 mM Imidazole, 1 M NaCl, 2 mM β-mercaptoethanol). HRV 3C protease was used to cleave the HisTag by incubation overnight at 4^{o}C. The sample was dialysed in Spectra/Por Dialysis Membrane of the appropriate MWCO in 4 litres of equilibration buffer then the cleaved tag was separated from the protein by IMAC (Ni-NTA). The cleavage reaction mixture was loaded into a gravity-driven column packed with Ni-NTA resin equilibrated with 10 CV of equilibration buffer and the flow through loaded for a second time. The resin was washed with 5 CV of equilibration buffer, 5 CV of wash buffer, then the tag was eluted with 5 CV of elution buffer. The protein containing fractions were then concentrated to 5ml using Vivaspin concentrators in order to be purified further by size exclusion chromatography. Size exclusion chromatography was performed using an ÄKTA purifier system (GE Healthcare) with a Hiload 16/60 Superdex prep grade column equilibrated in equilibration buffer. Fractions of pure protein were pooled and concentrated before being dialysed into a final buffer of 10mM phosphate pH6.9, 40mM NaCl, 0.5mM TCEP.

5.2.5 Backbone assignment

Labelled (^{15}N, ^{13}C^{15}N or ^{2}H^{13}C^{15}N) samples were prepared as described and concentrated to ~200 μM. NMR experiments were conducted at 25^{o}C or 37^{o}C using Bruker Avance and Varian Inova NMR spectrometers operating at 600, 700 and 800 MHz. Measurements were made in 90% H2O/10% D2O. A ^{13}C^{15}N labelled sample was used to acquire TROSY HNCA and ^{15}N-NOESY-HSQC and a ^{2}H^{13}C^{15}N labelled sample was used to acquire a further TROSY HNCA, TROSY HNCACB and HN(CO)CACB. Spectra were processed using NMRPipe/NMRDraw and analysed using XEASY or Sparky in order to determine ^{1}HN, ^{15}N, ^{13}Cα and ^{13}Cβ assignments.

5.2.6 Scaffold Independent Analysis

SIA was performed as described in Section 3.1.1 on FMRP KH1WT/KH2DD and FMRP KH1KK/KH2DD mutants. 50 µM stock protein was prepared in buffer 40 mM NaCl, 10 mM phosphate pH 6.5, 0.5 mM TCEP, sodium azide and RNasin Plus. This was aliquoted out into 180 µM samples and RNA pools added where required at a protein to RNA ratio of 1 to 4.

5.2.7 RNA binding assays - NMR

85 µM ^{15}N-labelled samples in 10 mM phosphate pH 6.9, 40 mM NaCl, and 0.5 mM TCEP were titrated with unlabelled RNA oligonucleotides up to maximum protein to RNA ratios of 1 to 16 for some titrations. $^{1}H^{15}N$ SOFAST-HMQC spectra were recorded at each titration point at 25^{o}C on Bruker Avance NMR spectrometers operating at 600 or 700 MHz.

5.2.8 Thermal unfolding

Thermal unfolding of FMRP KH12 (216-359), FMRP KH1WT/KH2DD and FMRP KH1KK/KH2DD was monitored by circular dichroism. Experiments were performed on a Jasco J-815 spectropolarimeter equipped with CDF-426S temperature-control system. Protein samples were prepared in 10 mM phosphate pH 6.9, 40 mM NaCl, 0.5 mM TCEP at 0.2 mg/ml. The solution was heated from 20^{o}C to 95^{o}C at a rate of 2^{o}C per minute and the unfolding of the protein was monitored at 220 nm.

5.3 Characterisation of FMRP KH domains

The structure of the KH domains of FMRP has been solved by X-ray crystallography (PDB:2QND). This study provided us with a starting point to design folded and soluble constructs for use in our structural and biophysical analysis. Although the construct used in the determination of the crystal structure is folded, RNA binding data are not available and secondary structure predictions indicated this construct ends in the middle of an α helix. Therefore we designed three constructs: one with the same domain boundaries as used in the crystal structure, FMRP KH12 (216-359); one extended to the end of the predicted α helix, FMRP KH12 (212-383); and one using the C-terminal domain boundary of the construct used in studies by the Darnell laboratory which identified the kissing complex RNA as a target of KH2 and in following filter binding assays, FMRP KH12 (212-405) (Figure 5.3).[31,41]

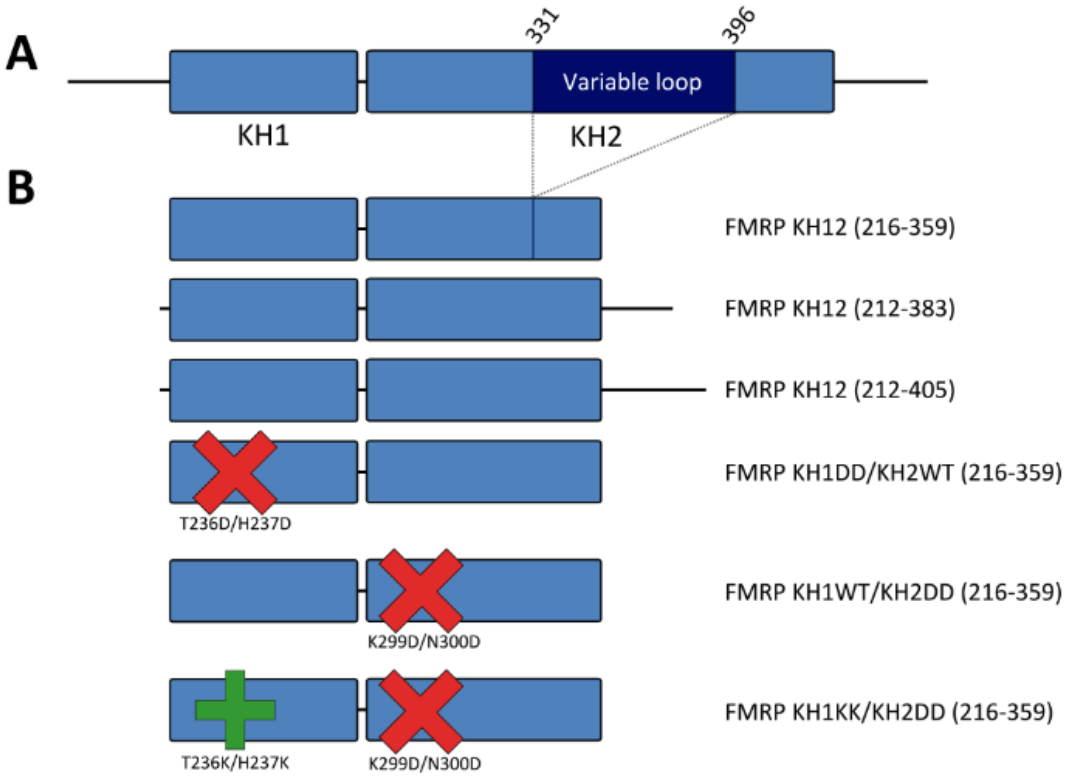


Figure 5.3 | Scheme of FMRP constructs. A) Wild type FMRP KH1-KH2 segment with region of variable loop deleted in FMRP Δ331-396 highlighted. B) Constructs used all lack the variable loop and numbering corresponds to this sequence. Domains with impaired RNA binding, red cross. Domains with enhanced binding, green plus.

The second KH domain contains a lengthy variable loop of 66 amino acids encoded for by exons 11 and 12. Previous attempts to produce constructs containing this loop have shown the protein to be difficult to express and handle. The extended variable loop is not found in the chicken, *Xenopus* or zebrafish orthologues, or close family members FXR1 and FXR2.[41] In the cell exon 12 is alternatively spliced and is lacking in 80% of mature transcripts.[42] Additionally *in vitro* inclusion of the region encoded by exon 12 destabilises the domain leading to a reduction in RNA binding. Filter binding assays have shown a KH2 domain lacking both exons 11 and 12 binds with the same affinity to kissing complex RNA as a construct with exon 11 (the most common naturally occurring isoform).[41] On the basis of this evidence we have used a constructs with exons 11 and 12 regions deleted (FMRP Δ331-396) leaving a variable loop of 12 amino acids.

All three constructs showed soluble expression in *E. coli*. After purification, using affinity chromatography, cleavage of the His-Tag and further purification using size exclusion chromatography pure products were obtained (Figure 5.4). Constructs FMRP KH12 (216-359) and FMRP KH12 (212-383) gave high yields of up to 45mg per litre of culture. Construct FMRP KH12 (212-405) was expressed at lower quantities and exhibited degradation during the purification procedure.

^{1}H-^{15}N correlation NMR experiments allowed us to assess the folding state of the protein and optimise construct boundaries. In NMR spectra from ^{1}H-^{15}N SOFAST-HMQC experiments the chemical shifts of backbone amide groups in unstructured regions are normally clustered in the centre of the spectrum, between 110 ppm and 125 ppm in the ^{15}N dimension and around 8.3ppm in the ^{1}H dimension. Peaks from amides in structured regions of the protein, and in particular in a β sheet arrangement, usually show more dispersion.[43] In construct 216-359 the peaks are well dispersed with few residues in the unstructured region. With the lengthening of the constructs to 212-383 and 212-405 extra dispersed peaks appear in the spectrum indicating some of the additional residues form secondary structural elements (Figure 5.5). However upon addition of RNA in all three constructs chemical shifts in FMRP KH1 were perturbed while chemical shifts in KH2 were not (Figure 5.6). This indicated that the extension of the KH2 domain did not alter the RNA binding capabilities of the KH2 domain, which will be discussed in more detail below. Further experiments in this study were performed on construct lengths of FMRP KH12 (216-359).

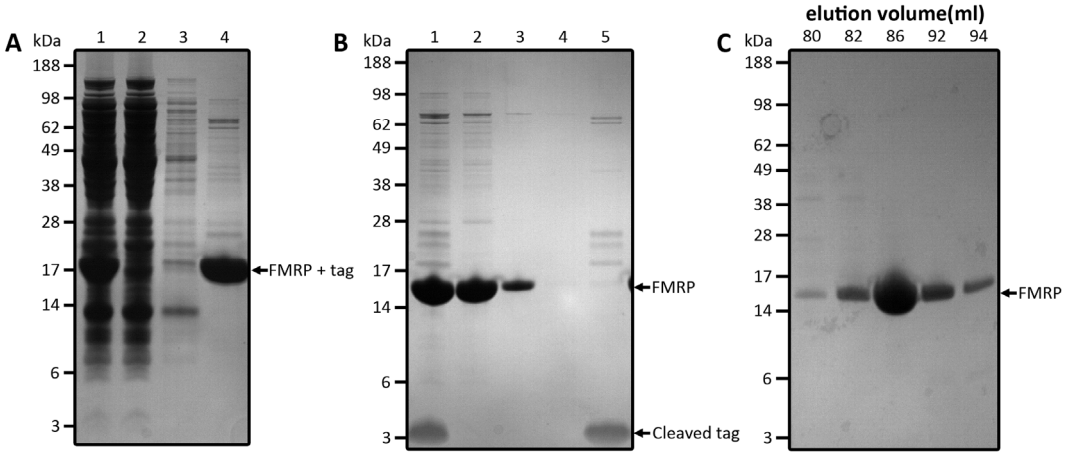


Figure 5.4 | Purification strategy employed for FMRP KH12 (216-359). A) SDS-page analysis of affinity chromatography fractions of FMRP using nickel-NTA after cell lysis and centrifugation. 1) cleared cell lysate, 2) flow through after column loading with cleared cell lysate, 3) wash with 10 CV of wash buffer (10 mM Tris-HCl pH 8.0, 10 mM Imidazole, 1 M NaCl, 2 mM β-mercaptoethanol), 4) elution with 5 CV of elution buffer (10 mM Tris-HCl pH 8.0, 250 mM Imidazole, 1 M NaCl, 2 mM β-mercaptoethanol). B) SDS-page analysis of affinity chromatography fractions of FMRP using nickel-NTA after His-Tag cleavage with HRV 3C protease. 1) total cleavage reaction, 2) flow through after column loading, 3) wash with 5 CV of equilibration buffer (10 mM Tris pH 8.0, 10 mM imidazole, 200 mM NaCl, 2 mM β-mercaptoethanol, 10 µM $ZnCl_2$) 4) wash with 10 CV of wash buffer (10 mM Tris-HCl pH 8.0, 10 mM Imidazole, 1 M NaCl, 2 mM β-mercaptoethanol), 5) elution with 5 CV of elution buffer (10 mM Tris-HCl pH 8.0, 250 mM Imidazole, 1 M NaCl, 2 mM β-mercaptoethanol). C) SDS-page analysis of size exclusion chromatography fractions of FMRP using Superdex 75 16/60. Lanes are identified with elution volume in ml.

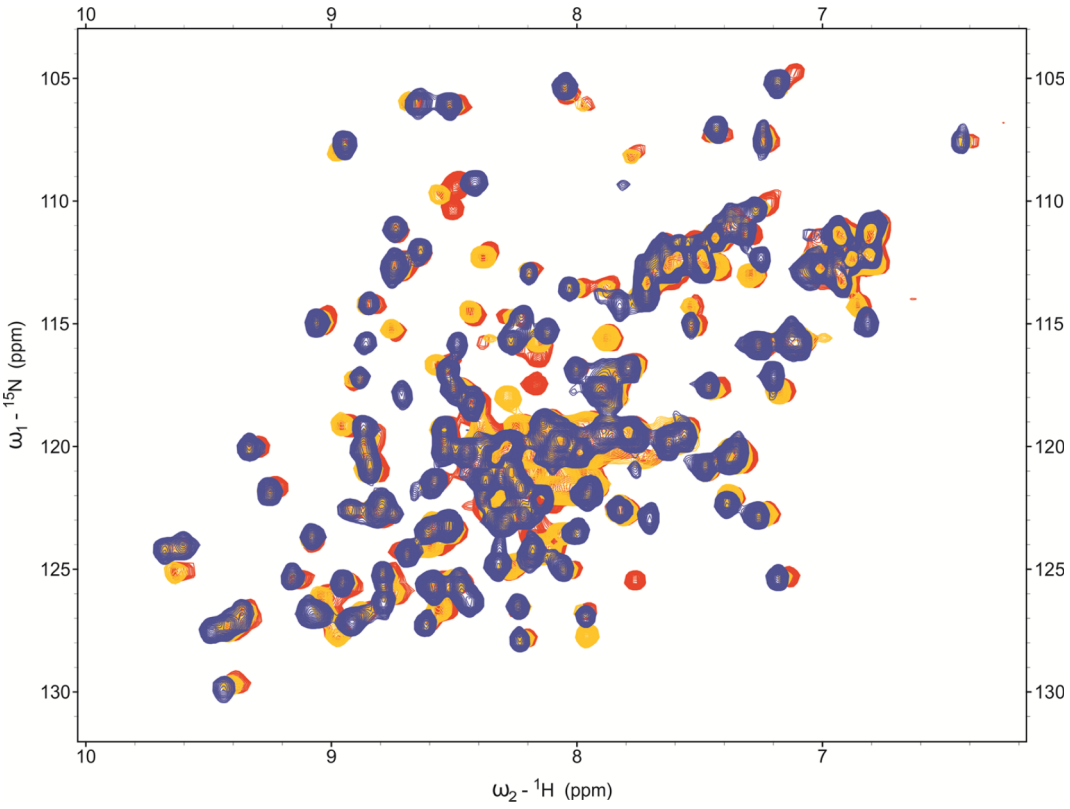
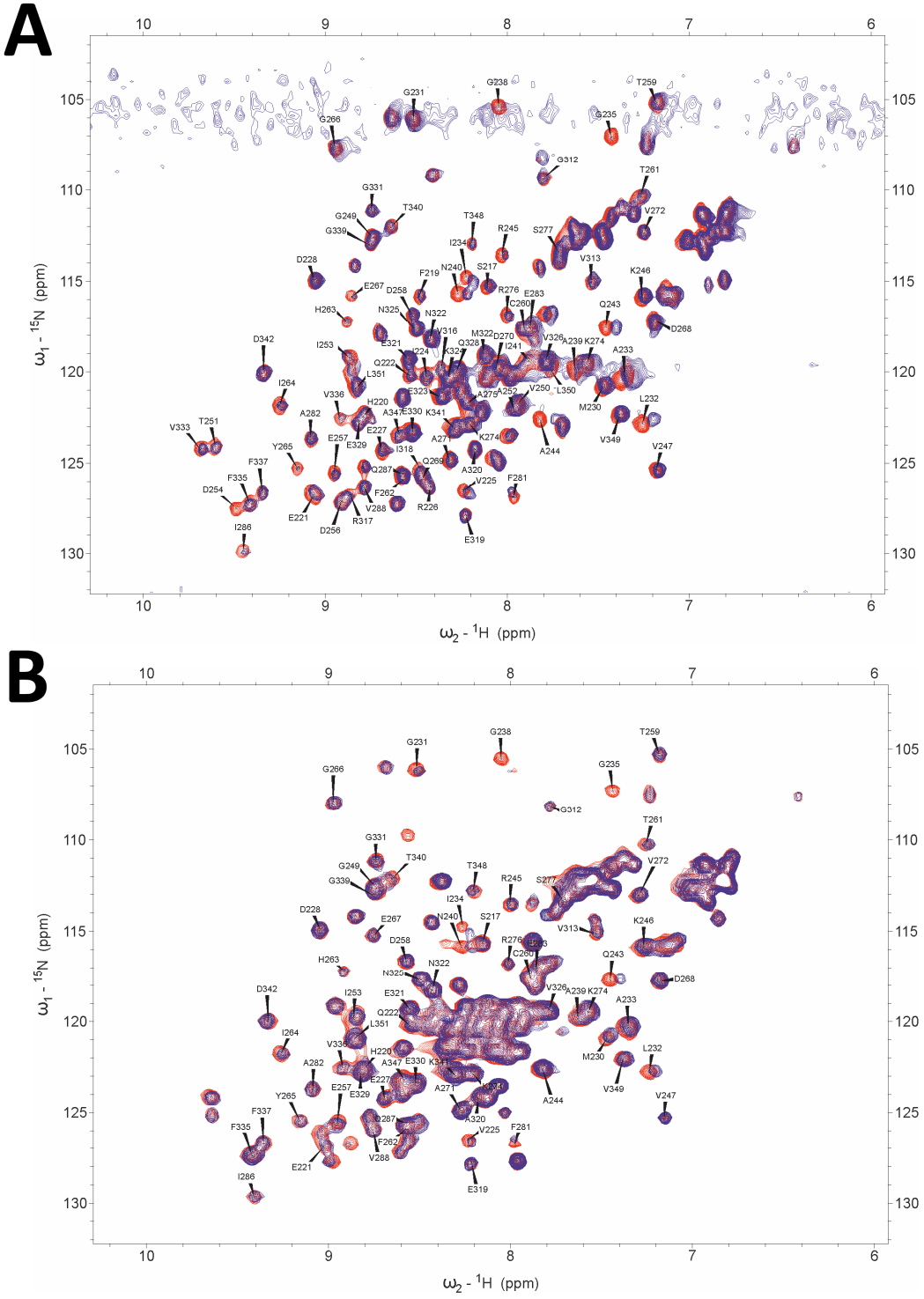


Figure 5.5 | Comparison of constructs of different lengths. Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of 80µM FMRP KH12 (216-359) (blue), 60µM FMRP KH12 (212-383) (yellow), and 60µM FMRP KH12 (212-405) (red).


A
B
ω1 - 15N (ppm)
ω2 - 1H (ppm)
10
9
8
7
6
105
110
115
120
125
130
G231
G238
T259
G235
G266
G312
T261
G331
T340
G249
G339
T348
R245
S277
I234
V272
D228
N240
S217
V313
K246
F219
E267
D258
R276
E283
C260
H263
N325
N322
Q243
D268
V316
M322
D270
V326
A239
K274
D342
I253
E321
I241
Q328
A233
I224
K324
Q222
L351
I264
E323
A275
A252
V250
L350
M230
L232
V336
H220
A347
E330
K341
A282
E227
A271
K274
V333
T251
E257
E329
A244
V349
V247
Y265
I318
Q269
A320
F337
Q287
V225
F281
F335
D254
F262
R226
V288
E221
R317
D256
I286
E319

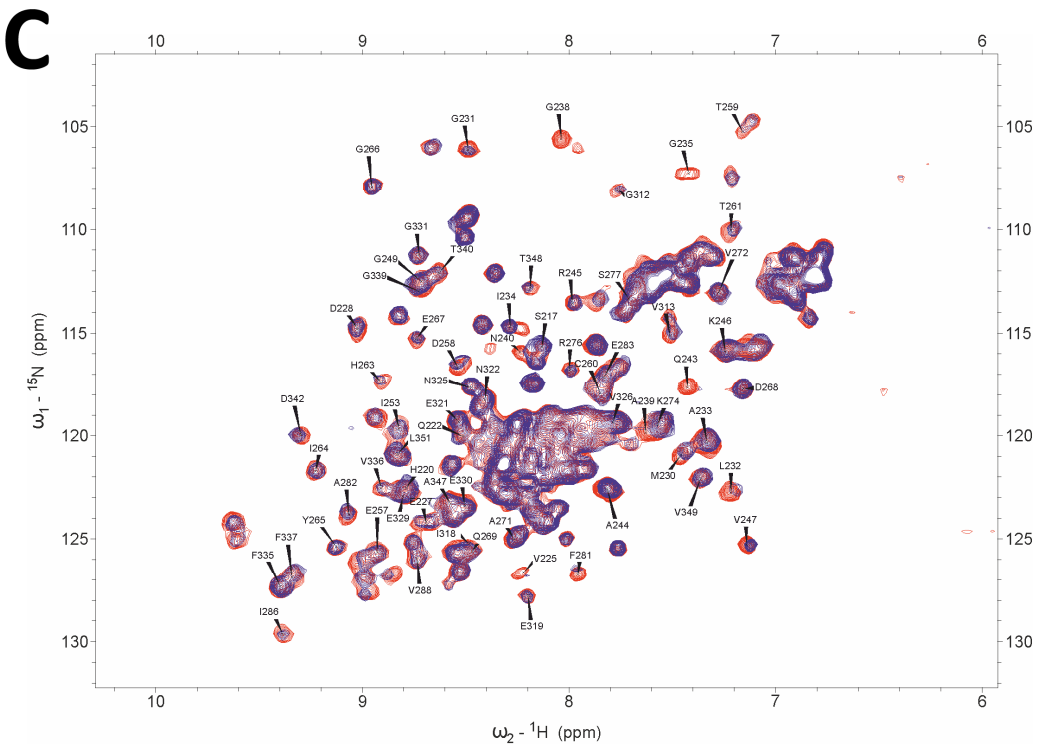


Figure 5.6 | Comparison of peak shift perturbation in FMRP KH12 (216-359), FMRP KH12 (212-383) and FMRP (212-405) upon addition of an RNA pool. Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra with assignment labelled of A) FMRP KH12 (216-359) with UNCNN at protein to RNA ratios of 1:0 (red) and 1:4 (blue). B) FMRP KH12 (212-383) with NNNNN at protein to RNA ratios of 1:0 (red) and 1:8 (blue). C) FMRP KH12 (212-405) with NNNNU at protein to RNA ratios of 1:0 (red) and 1:8 (blue).

The KH domains of FMRP are joined by a short linker of just two amino acids. As they are in such close proximity there could be contacts made between the domains and therefore we wanted to test the RNA binding of each of the KH domains separately but without the complete removal of the other domain. To achieve this we used a mutation previously designed in the lab and shown to eliminate RNA binding of a KH domain without affecting its overall structure.[44] The GXXG loop in KH domains normally contains one or two positively charged residues which interact with the phosphate backbone of the nucleic acid. In these general use mutants the two residues in the GXXG loop are changed to aspartates. This creates a wide negative surface and hinders the contacts with the RNA backbone.[44] We mutated each of the KH domains individually in the context of the two domain construct (Figure 5.3).

The fold and stability of the mutants were assessed by ^{1}H-^{15}N correlation spectroscopy and thermal denaturation monitored by circular dichroism (CD) spectroscopy. An overlay of the wild type ^{1}H-^{15}N HMQC-SOFAST spectrum with that of the mutants shows that the majority of peaks remain unperturbed (Figure 5.7a). Several of the peaks are affected due to the mutation and the effect this has on the chemical environment of the surrounding residues. Thermal denaturation studies showed that the FMRP KH12 (216-359) and FMRP KH1WT/KH2DD mutant have Tm of 68.9^{o}C and 69.6^{o}C respectively (Figure 5.7b). Together the NMR and CD data show that the mutations do not affect the overall fold or stability of the protein.

To test the RNA binding capabilities of the wild type and mutant proteins we performed RNA titrations monitored by NMR. A pool of random pentamer RNA oligonucleotides was added to the free protein with ^{1}H-^{15}N SOFAST-HMQC experiments being recorded at each titration point. The same set of peaks was perturbed upon RNA addition to the FMRP KH12 (216-359) and FMRP KH1WT/KH2DD proteins (Figure 8a and Figure 8b). Upon RNA addition to the FMRP KH1DD/KH2WT mutant peaks did not shift (Figure 8c). This shows that FMRP KH1WT/KH2DD binds in a similar way to the wild type protein while FMRP KH1DD/KH2WT lacks RNA binding capabilities altogether, indicating that only the KH1 domain is capable of binding RNA.

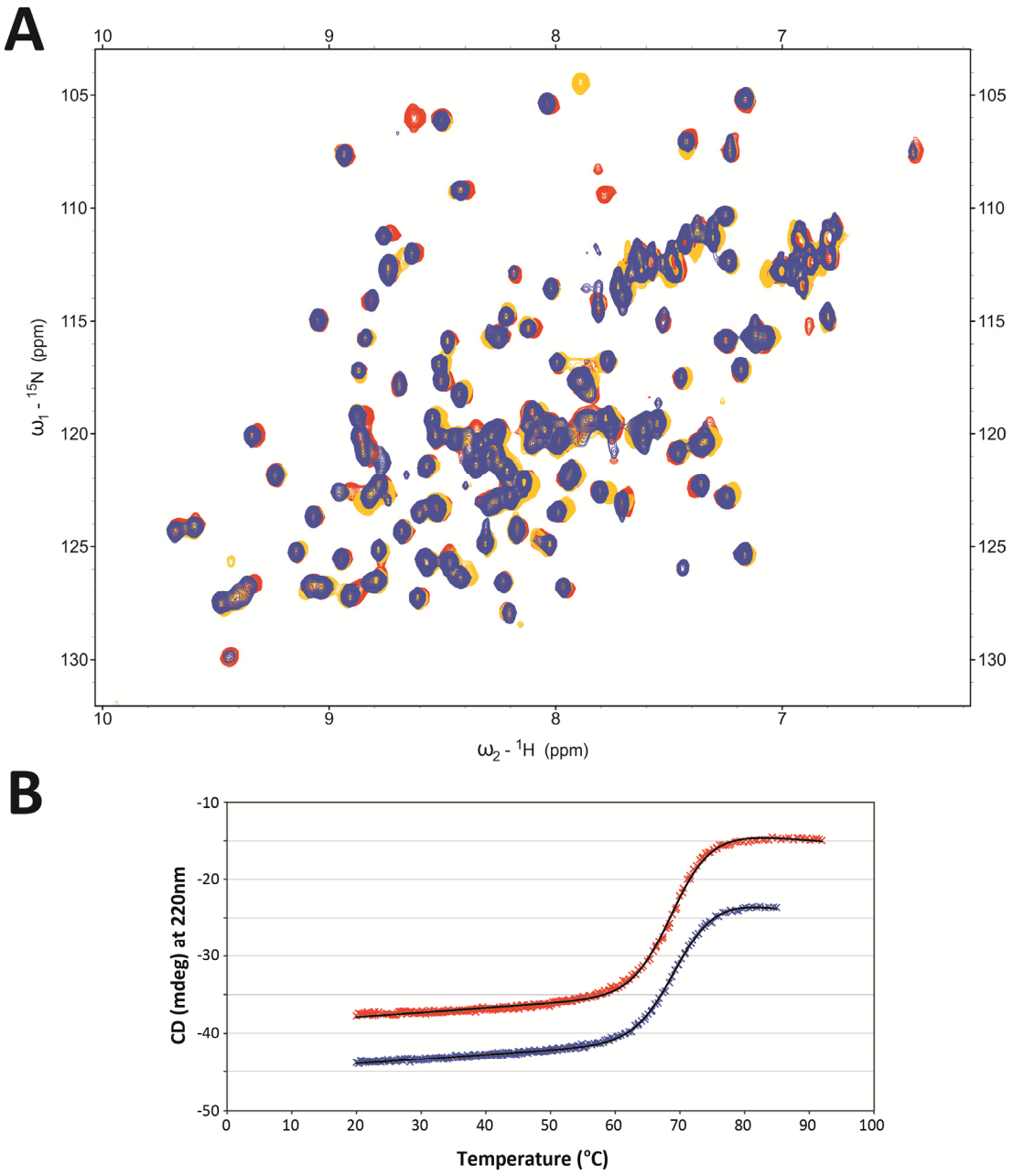
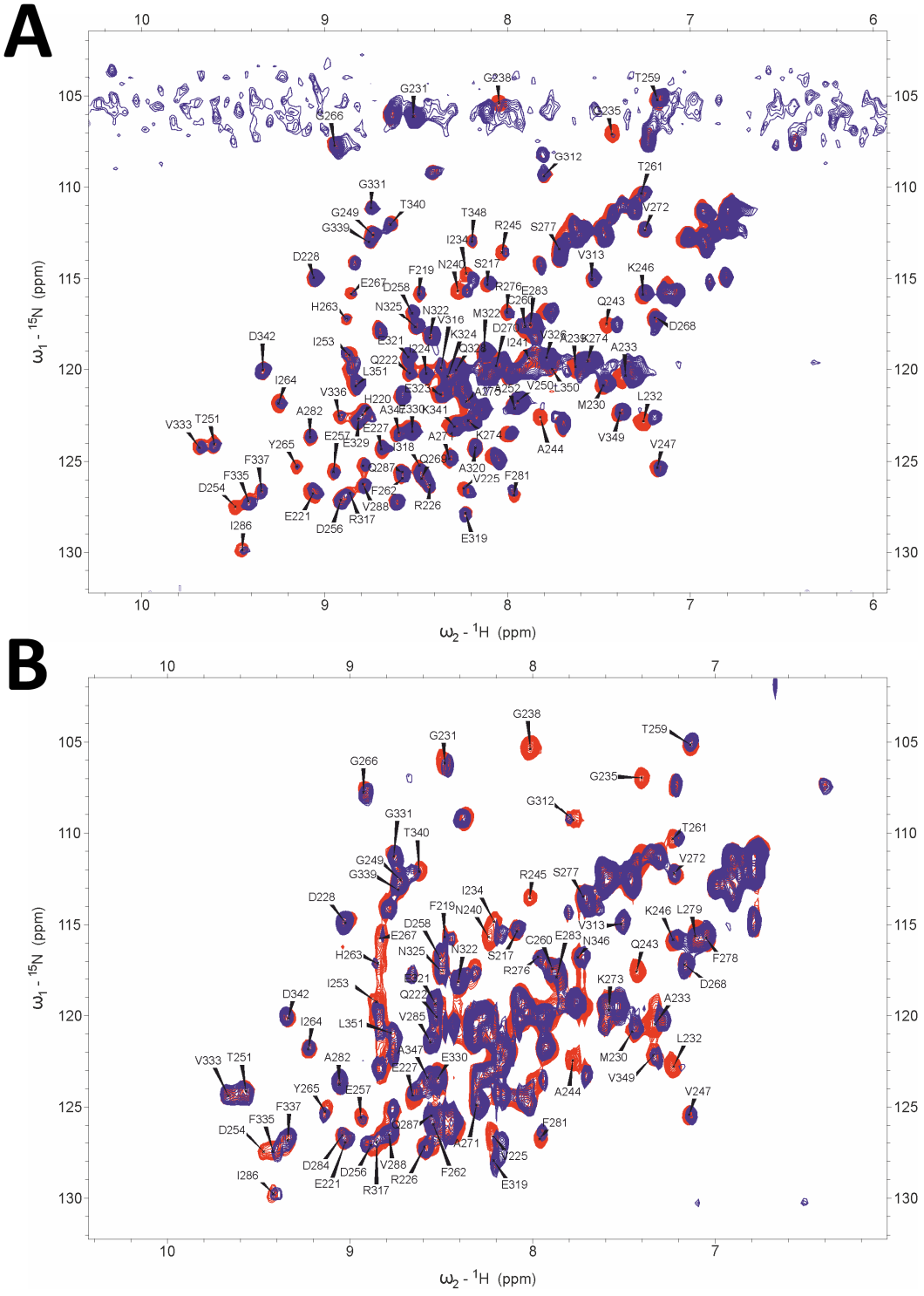


Figure 5.7 | Comparison of the fold and stability of FMRP KH12 (216-359), FMRP KH1DD/KH2WT and FMRP KH1WT/KH2DD. A) Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of 80µM FMRP KH12 (216-359) (red), 50µM FMRP KH1DD/KH2WT (yellow), and 50µM FMRP KH1WT/KH2DD (blue). B) Thermal unfolding of FMRP KH12 (216-359) and FMRP KH1WT/KH2DD monitored at 220nm. The plot shows the full curve and fit for FMRP KH12 (216-359) (red) and FMRP KH1WT/KH2DD (blue). Tm values are 68.9°C and 69.6°C respectively.


A
B
ω1 - 15N (ppm)
ω2 - 1H (ppm)

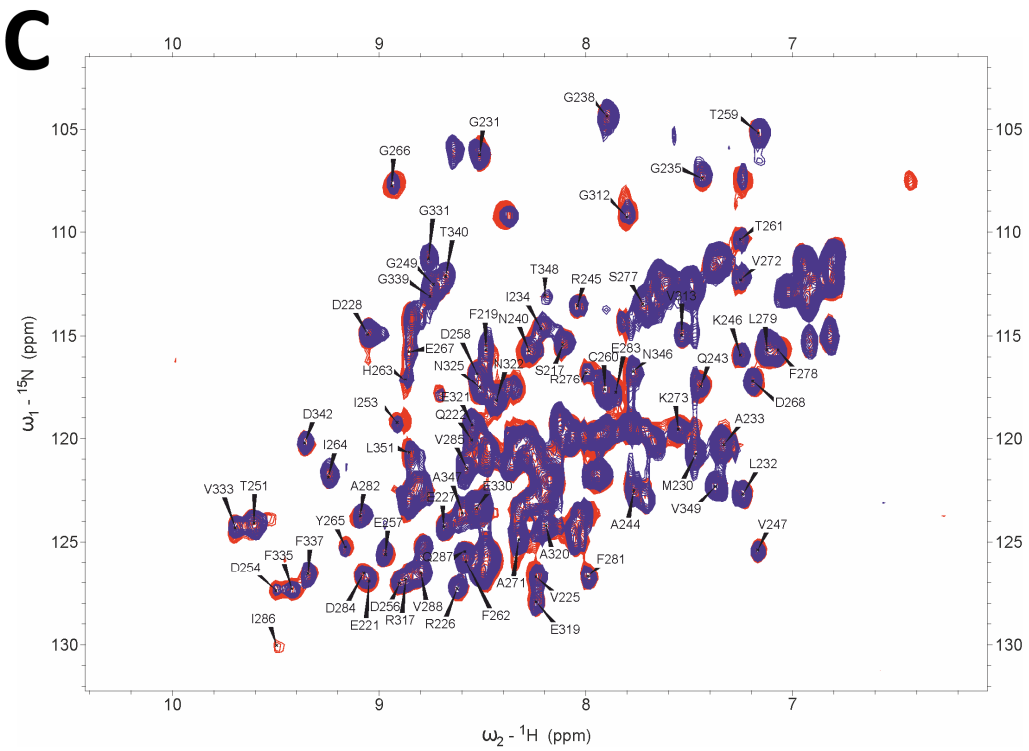


Figure 5.8 | Comparison of peak shift perturbation in FMRP KH12 (216-359), FMRP KH1DD/KH2WT and FMRP KH1WT/KH2DD upon addition of an RNA pool. Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of A) FMRP KH12 (216-359) with UNCNN at protein to RNA ratios of 1:0 (red) and 1:4 (blue). B) FMRP KH1DD/KH2WT with NNNUN at protein to RNA ratios of 1:0 (red) and 1:8 (blue). C) FMRP KH1WT/KH2DD with NNNGN at protein to RNA ratios of 1:0 (red) and 1:8 (blue).

Marino et al. found that the I304N mutation in the second KH domain of FMRP destabilises the hydrophobic core producing a partial unfolding of the domain. Molecular docking with single stranded nucleic acid shows strongly reduced binding.[45] The mutation has been shown to cause the same phenotype as total loss of the protein leading to the hypothesis that the KH2 domain is vital for FMRP function.[46] Therefore it was surprising that in our hands the domain did not bind RNA. This led us to look more closely at the KH2 domain.

In order to do this I first had to identify which peaks belonged to each of the domains and so I assigned the backbone of the protein. A series of 3D NMR experiments were run on FMRP KH1WT/KH2DD at 298K. Initial experiments showed that without deuteration the sensitivity was such that a good quality HNCACB spectrum could not be obtained. Therefore TROSY HNCA and ^{15}N-NOESY-HSQC were run on a non-deuterated sample and TROSY HNCACB and HN(CO)CACB were run on a deuterated sample. A further TROSY HNCA was run at 301K on a deuterated sample to see if any extra signals could be observed.

Using the complement of backbone experiments and the strategy described in Section 2.1.2 I assigned 97 out of 140 backbone amides (69% completeness excluding prolines) (Figure 5.9). A list of chemical shift assignment can be found in Appendix XI. If the domains are analysed separately KH1 is 89% completed (56 out of 63 excluding prolines) while KH2 is just 55% completed (41 out of 75 excluding prolines). The unassigned backbone amides in KH2 are concentrated in two regions, the GXXG loop and flanking helices and the C-terminus. While residues in the GXXG loop are sometimes absent, as is the case for the KH1 domain, it was unexpected that such a large surrounding area should be unassigned. In the ^{1}H-^{15}N HSQC spectrum there are 21 peaks unaccounted for. The majority of these are missing signals, or give very weak signals in the 3D experiments making them difficult to assign but regardless would not account for the full amount of unassigned backbone amides. This lead us to believe that a section of KH2 was not properly structured and in an exchange regime leading to peak broadening. In an effort to move the peaks into a more optimal exchange regime the temperature was raised from 298K to 301K, 305K, 308K and 310K in order to increase the rate of exchange between the potential conformational states however this failed to produce any new signals (Figure 5.10). The multiple conformations predicted to occur in KH2 could account for the lack of RNA binding.

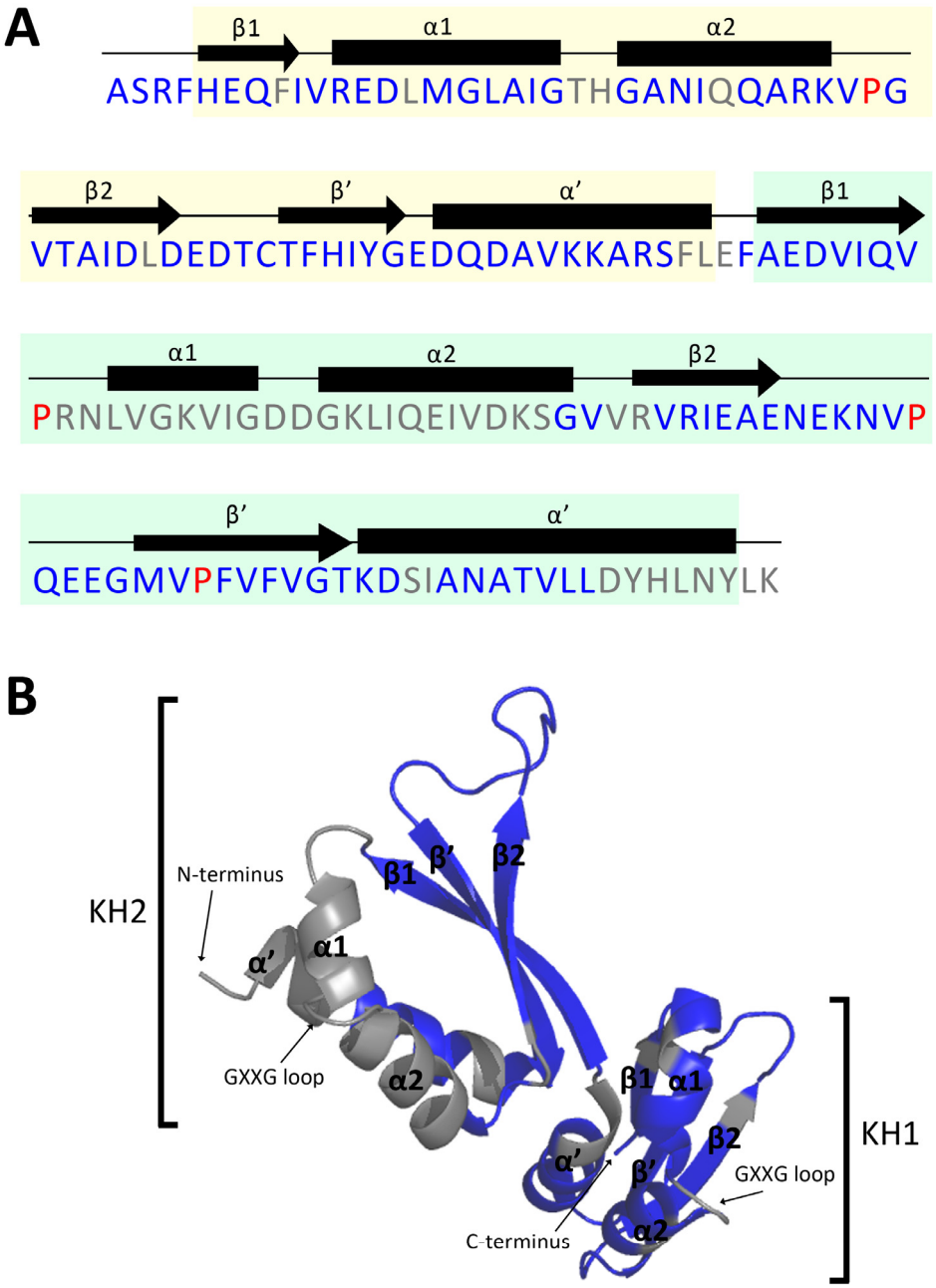


Figure 5.9 | Overview of the backbone assignment of FMRP. A) The residues with assigned and unassigned backbone amides resonances are in blue and grey respectively. Prolines, which are not detectable in ^{1}H-^{15}N correlation spectra, are in red. The three N-terminal residues which are a by-product of cloning are not shown. KH1 is highlighted by a yellow box, KH2 is highlighted by a green box. α helices are represented by rectangles, β strands are represented by arrows. B) Crystal structure of FMRP KH1 and KH2 (PDB:2QND). Residues with unassigned backbone amides are mapped onto the structure in grey.

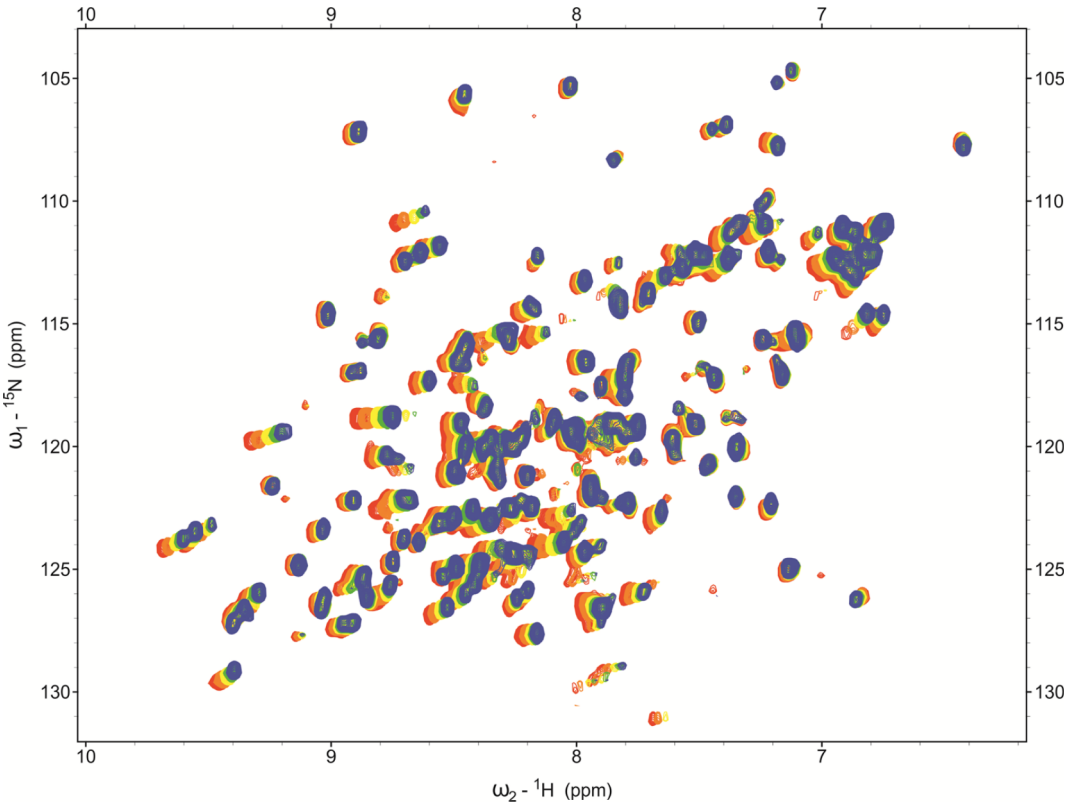


Figure 5.10 | Effect of temperature change on ^{1}H-^{15}N correlation spectra of FMRP. Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of 400µM FMRP recorded at 298K (red), 301K (orange), 305K (yellow), 308K (green), and 310K (blue).

5.4 Determining the specificity of FMRP KH1

We attempted to determine the specificity of FMRP KH1 by performing SIA on the FMRP KH1WT/KH2DD construct (as described in Section 3.1.1) at the protein to RNA ratio 1:4. The peak shifts were measured manually and analysed. The preferred sequence was determined to be A(C/G)CC (Appendix XII), however the chemical shift perturbation upon addition of the RNA pools were much smaller than with other domains examined by this method. The weighted average peak shift measured in FMRP KH1 with the addition of an RNA pool was 0.03 ppm. Weighted average peak shifts for other domains tested by SIA are 0.14 ppm for TUT4 CCHC-ZF3, 0.08 ppm for RNA15 RRM and 0.05 ppm for TSTAR KH domain. The small peak shifts for FMRP KH1 mean that any errors in measurement will be exaggerated and the results may be inaccurate.

As the weak binding of FMRP KH1 was hindering attempts to determine the specificity we designed a general mutation intended to increase affinity for RNA without affecting the nucleobase preferences of the domain. As previously mentioned, in canonical KH-RNA interactions positively charged residues in the GXXG loop make contact with the negatively charged RNA backbone and the GDDG mutation shows that this interaction is critical for RNA binding.[44] FMRP KH1 contains a threonine and a histidine in its GXXG loop and we mutated these residues to lysines in order to increase the positively charged surface available to interact with the RNA backbone. As all known KH domains which bind RNA contain the GXXG loop it is hoped that the GKKG sequence may be a general mutation for increased specificity in the same way GDDG abolishes binding.

As with the FMRP KH1WT/KH2DD and FMRP KH1DD/KH2WT mutations, the FMRP KH1WT/KH2DD mutant expressed and purified as the wild type protein. Analysis of the structure of the construct by ^{1}H-^{15}N correlation spectroscopy showed a largely similar spectrum to the FMRP KH1WT/KH2DD construct and the only perturbed peaks were those to be expected of the mutated residues and those surrounding the mutation (Figure 5.11a). The stability of the construct was tested by thermal denaturation monitored by circular dichroism and found the melting temperature of the FMRP KH1WT/KH2DD and FMRP KH1WT/KH2DD proteins to be roughly the same, at 69.6^{o}C and 68.4^{o}C respectively (Figure 5.11b). Together these show that the GKKG mutation does not affect the structure or stability of the domain.

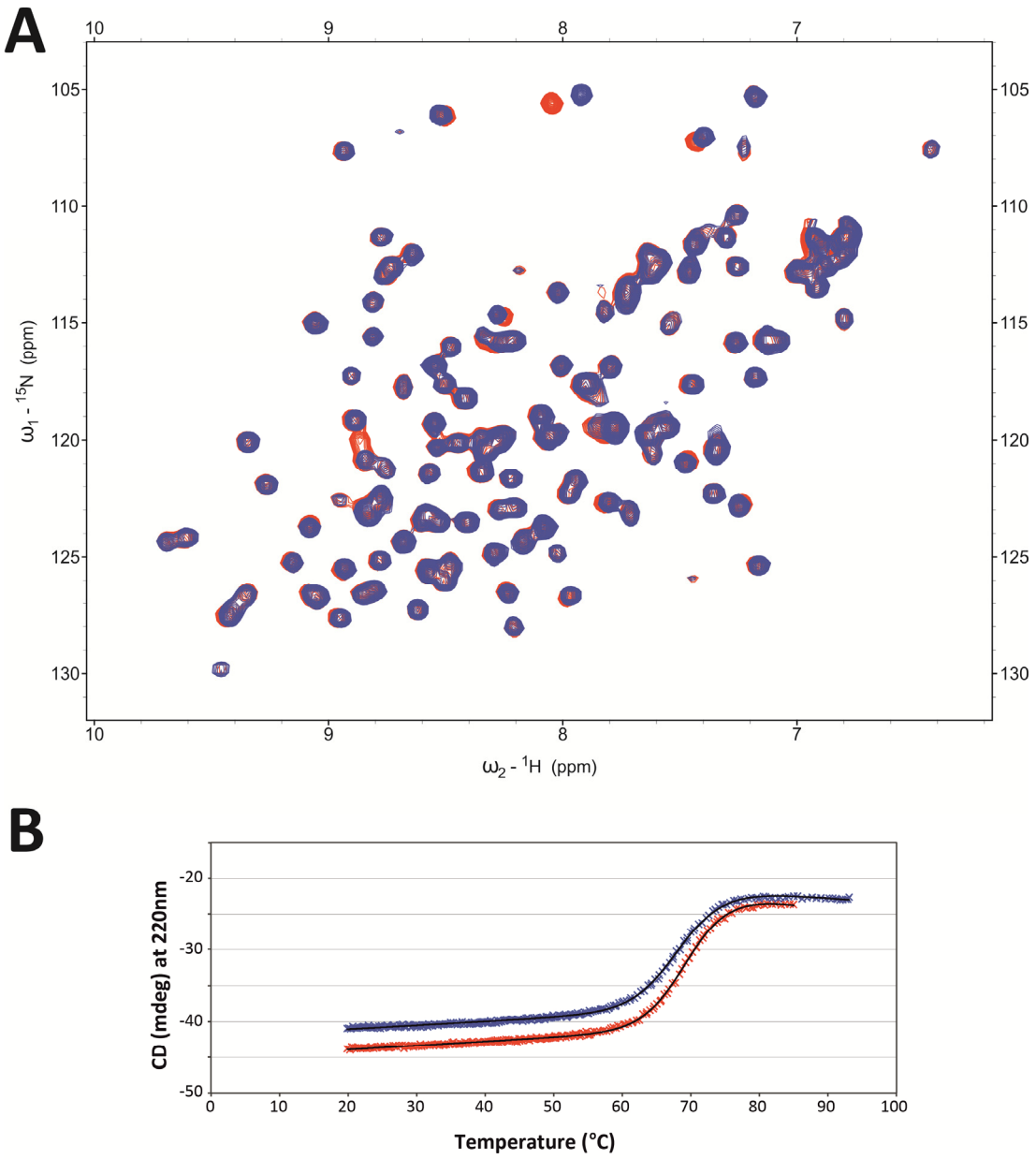


Figure 5.11 | Comparison of the fold and stability of FMRP KH1WT/KH2DD and FMRP KH1WT/KH2DD. A) Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of 65µM FMRP KH1WT/KH2DD (red), and 60µM FMRP KH1WT/KH2DD (blue). B) Thermal unfolding of FMRP KH1WT/KH2DD and FMRP KH1WT/KH2DD monitored at 220nm. The plot shows the full curve and fit for FMRP KH1WT/KH2DD (red) and FMRP KH1WT/KH2DD (blue). Tm values are 69.6°C and 68.4°C respectively.

To test if the mutation was successful in increasing the affinity of the domain for RNA we performed titrations monitored by ^{1}H-^{15}N correlation spectroscopy. A pool of random RNA pentamers was titrated with the FMRP KH1WT/KH2DD or FMRP KH1WT/KH2DD protein and ^{1}H-^{15}N SOFAST-HMQC were recorded at protein to RNA ratios of 1:0, 1:2 and 1:8. We observed much larger chemical shift perturbation upon RNA addition to the FMRP KH1WT/KH2DD mutant than the FMRP KH1WT/KH2DD protein indicating stronger binding (Figure 5.12). This shows that the GKKG mutation has increased the affinity of the KH1 domain for RNA.

With the increased binding affinity we again attempted to determine the specificity of the KH1 domain by SIA using the FMRP KH1KK/KH2DD mutant. ^{1}H-^{15}N SOFAST-HMQC were run for the free protein and the protein in complex with each of the 16 RNA pools at a protein to RNA ratio of 1:4. The analysis to determine the order of preference in each of the binding positions was performed by manual peak shift measurement (Appendix XII).

The SIA results describe a preferred binding sequence of GGCC (Table 5.4). In Position 1 guanine is the most preferred nucleotide with a score of 0.97 this is followed by adenine with a score of 0.83 with cytosine being the least preferred base with a score of 0.56. It is a similar situation in Position 2 for the favoured bases guanine is top with a score of 0.97 and adenine is second with 0.79. Cytosine and uridine are least preferred with scores of 0.65 and 0.60 respectively. In Position 3 the preferred base is cytosine with a score of 0.96 followed by adenine with a score of 0.85. In Position 4 again cytosine is the preferred base at 0.98 followed by guanine and uridine with 0.80 and 0.76 respectively.

In order to verify the SIA results we performed a set of titrations monitored by NMR of FMRP KH1KK/KH2DD with RNA pentamers and determined the binding affinities. Weighted average chemical shift perturbation was plotted against RNA to protein ratio and fitted using Equation 2 (Section 2.1.1) in Origin 9.1. Dissociation constants were determined using the Kd value averaged over several peaks. A full list of peaks used in the analysis and binding curve plots can be found in Appendix XIII. We were unable to perform the titration with the top ranking sequence of GGCC due to its sequence complementarity and propensity to form duplexes. Therefore we probed the first two positions individually comparing a preference for guanine over adenine in position 1 and guanine over cytosine in position 2. To probe the specificity of position 1 the binding of CACCC was compared

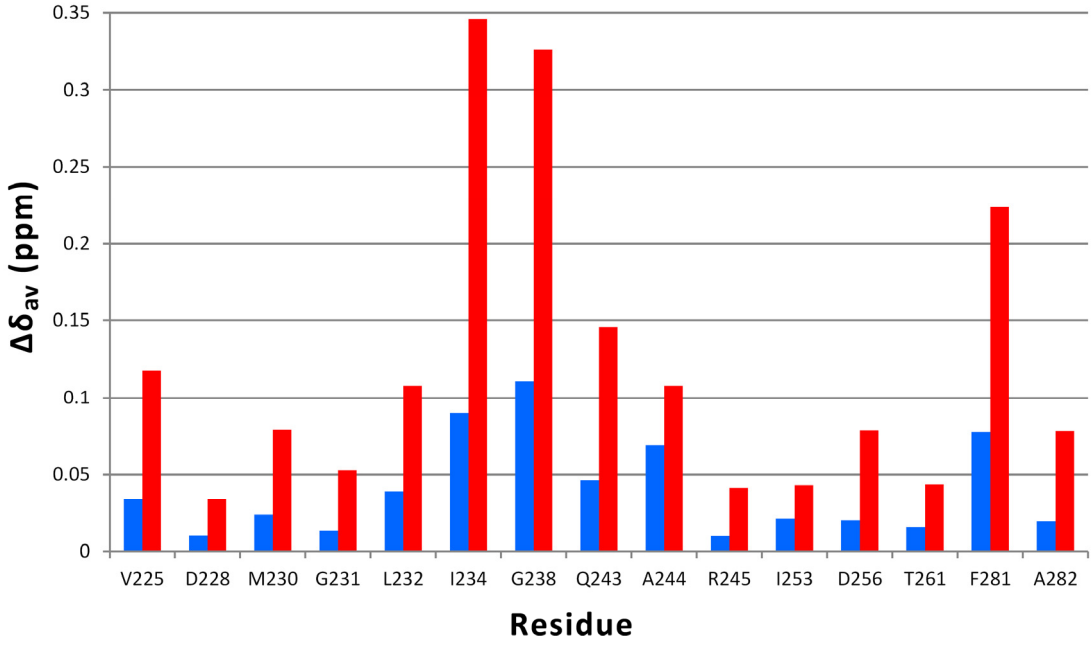


Figure 5.12 | Comparison of FMRP KH1WT/KH2DD and FMRP KH1WT/KH2DD RNA binding. Chemical shift perturbation upon addition of eight molar equivalents of NNNNN to FMRP KH1WT/KH2DD (red) and FMRP KH1WT/KH2DD (blue). The x-axis shows weighted average chemical shift perturbation in ppm and the y-axis displays the residue.

	Position 1	Position 2	Position 3	Position 4
A	0.83	0.79	0.85	0.63
C	0.56	0.65	0.96	0.98
G	0.97	0.97	0.72	0.80
U	0.72	0.60	0.61	0.76

Table 5.4 | SIA scores for FMRP KH1KK/KH2DD.

with CGCCC which differ only by having either an A or G in position 1. The KH1 domain bound CACCC with a Kd=487µM±217µM. Binding with CGCCC was around 3.5-fold tighter at Kd=142µM±35µM thus in agreement with the SIA data which describes a preference for guanine. For position 2 the binding of the pentamer CACCC was compared with CAGCC. Here CAGCC bound with a Kd=212µM±35µM which is 2.3-fold tighter than CACCC. Again this is in agreement for the preference of guanine in the SIA data. For the lowest ranked sequence of CCUUA the binding was too weak for the dissociation constant to be determined (Figure 5.13). Unfortunately the affinity of FMRP KH1WT/KH2DD was so low for the RNA pentamers tested that the specificity of the wild type KH1 domain could not be assessed.

5.5 Discussion

In this chapter I discuss how we have analysed the RNA binding properties of the two KH domains of the Fragile X Mental Retardation Protein. A very diverse set of RNAs has been reported to interact with FMRP *in vivo* and *in vitro* and it is important to fully make sense of these interactions to understand the function of the protein. It is expected that the two classical RNA binding domains, KH1 and KH2 will play a role in target RNA recognition. However, this role has not been clarified, in particular for the *in vivo* targets. We used NMR, in combination with a mutagenesis strategy and Scaffold Independent Analysis, to dissect the KH-domain RNA interactions. The results of this study show that KH1 binds ssRNA specifically, although with very low affinity, and identify the specific sequence targeted by this domain. In our hands the KH2 domain does not interact with short ssRNA. This could indicate the specific recognition of an RNA structure similar to the kissing loop structure identified by SELEX.[31] However, one of the sides of the RNA binding groove is in conformational exchange. Below I discuss our research strategy and findings in the context of the literature as well as future work I think may be useful to clarify some outstanding questions.

The comparison of fingerprint NMR spectra of constructs of increasing length confirmed the boundaries of the KH1/KH2 double domain as reported in the previously published crystal structure.[47] The wild type extended variable loop of KH2 is not present in our construct: this is justified by the reported capability of the loop-deleted domain to recognise a specific RNA target,[41] as well as by the crystal structure of KH1/KH2 that shows both domains adopt a classical KH fold in the absence of the loop.[47] However, in our hands

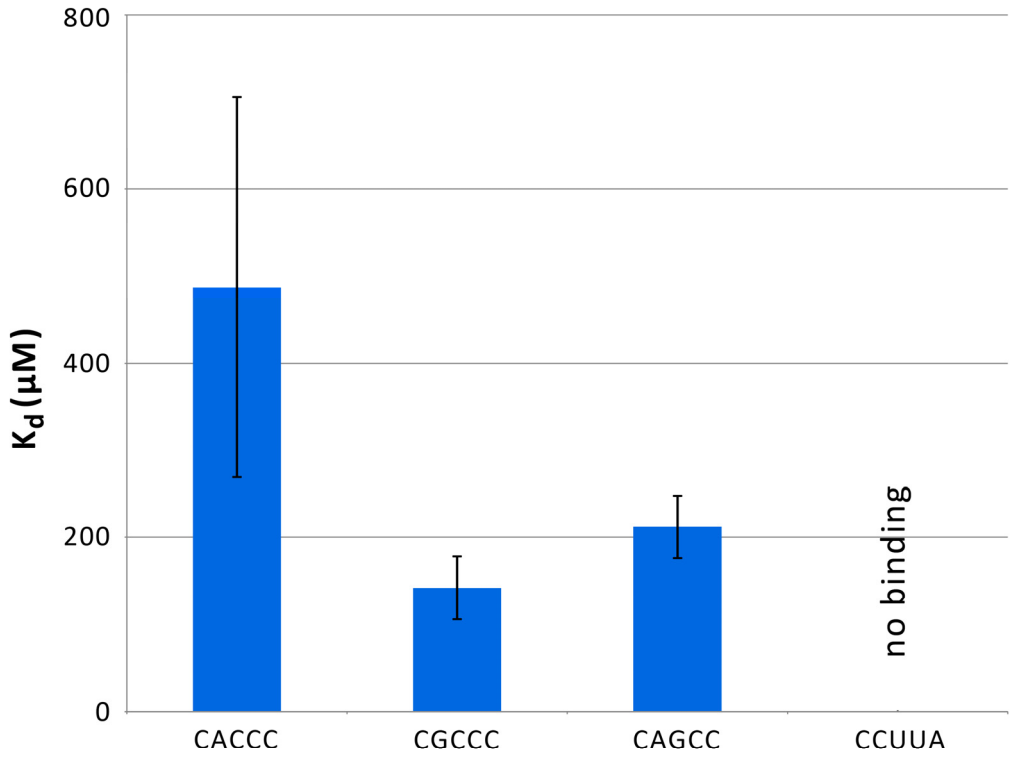


Figure 5.13 | Binding affinities of FMRP KH1KK/KH2DD with SIA derived pentamers. RNA pentamer sequences are displayed along the x-axis and the Kd in µM along the y-axis.

KH2 does not bind ssRNA, even when any possible competition from KH1 is eliminated by the KH1 GDDG mutation. This lack of RNA binding activity could be correlated with the disappearance of the resonances of one side of the RNA binding groove, probably because of exchange between multiple conformations. Changes in temperature were not sufficient to recover the resonances, but this is not surprising if more than two conformations are present. Molecular dynamics calculations performed on KH2 show the region of the GXXG loop and adjacent residues to be more flexible than other regions of the protein supporting our hypothesis of multiple conformations.[45] Interestingly, the GXXG region is visible in the crystal structure possibly because in the crystal the GXXG loop and α1 helices of two protein molecules pack against each other and this may stabilise one conformation.

Whether the absence of the variable loop extension is linked to the more flexible GXXG region and whether either of these is linked to the lack of RNA binding is unclear. A direct connection between variable loop and GXXG loops has not been observed in any KH domain and it is possible that the variable loop helps position the protein in an active complex which stabilises the GXXG loop and is permissive for RNA binding. The groove of KH2 as shown in the crystal structure does not show any clear indication in terms of size and amino acid composition as to why this domain does not bind ssRNA. The variable loop extension of FMRP is not present in the wild type FXR1P and FXR2P paralogues and an analysis of backbone dynamics and KH2 RNA binding in these proteins may shed light on whether the hydrophobic grooves of these domains have different RNA binding properties from the other KH domains. The size of the loop is such that modelling of conformation and dynamic changes is a challenge.

We attempted to determine the sequence preference of KH1 by performing SIA on a KH2 GDDG mutant. However chemical shift changes were very small due to the extremely weak binding of the domain and this did not allow for accurate measurement therefore we decided to manipulate further the affinity of the domain. The existing structure of KH-NA complexes show how the GXXG loop only interacts with the backbone of the RNA and it seemed likely that mutating residues in the GXXG loop to lysines would increase the affinity of the domain for RNA, without affecting specificity. The GKKG mutation increased the affinity of RNA binding and allowed us to determine the specificity of the domain by SIA. However further experiments must be performed in order to determine if the specificity of the domain has been altered by the mutation.

Our SIA assays report that KH1 has a preferred sequence of GG(C/A)C. This excludes KH1 as contributing to the binding of the U-rich sequence identified by SELEX[34] and to the ACUK sequence found enriched in the PAR-CLIP analysis.[18] It is instead consistent with KH1 playing a role in the binding of the WGGA sequence identified in PAR-CLIP data[18] and the GAC sequence identified by RNAcompete.[36] The WGGA sequence has been proposed to be related to the presence of G-quartets as a meta-analysis by Suhl found this motif only to be enriched in clusters in target genes.[14] This would indicate a role for the motif in the binding to the RGG box. Our study supports a different possibility, and agrees with the PAR-CLIP study where in the targets examined, the distance and numbers between GGA repeats are not consistent with G-quartet formation but rather with a high density of G-rich ssRNA sites that increase the apparent affinity. Further, it is interesting to note that a GGAC sequence is present in a single stranded region of the kissing hairpin target identified by SELEX.[31] We wonder whether the KH1-RNA interaction can contribute to the recognition of that structure. Overall, my analysis of KH1-RNA interactions show how the domain is unlikely to engage in the recognition of some of the targets (U-rich, ACUK) but very likely to contribute to the recognition of others and provide an insight into the likely conformation of the domain in the cell. Engineering this domain to change its specificity and test target recognition and function in vitro and in the cell will confirm the role of KH1 in the recognition of the cellular targets of FMRP. Further work would also entail searching the set of physiologically relevant target genes to see if our sequence is enriched in this pool and also its distribution compared to potential KH2 target motifs.

Increasing the RNA binding affinity of a domain in order to be able to analyse its specificity is, after the setting up of appropriate technical tools, the next step in the quest to understand the role that specificity of weakly binding domains has in RNA target recognition. My work indicates that there may be a scope to use such tools, at least in specific cases. We now want to extend this and test whether we can generalize the use of such mutations, starting with the GKKG mutation once it has been determined whether or not the mutation alters the specificity of the domain..

We have shown the KK mutation to increase the affinity of RNA binding for the first KH domain of FMRP. While we predict this mutation not to affect the specificity of the domain as the lysines only make contact with the RNA backbone we cannot rule this out. Due to the extremely weak binding of FMRP KH1 we were unable to determine the sequence

specificity of the WT domain and therefore could not investigate whether the KK mutation had altered specificity. In the future we would compare the specificity of WT KH domains with the KK mutant domains to ensure the mutation does not affect the preferred binding sequence. Once this is established the mutation could be used in an *in vivo* context as a gain of function mutant.

5.6 References

1. Coffee, B. *et al.* Incidence of fragile X syndrome by newborn screening for methylated FMR1 DNA. *Am. J. Hum. Genet.* **85,** 503–14 (2009).

2. Garber, K. B., Visootsak, J. & Warren, S. T. Fragile X syndrome. *Eur. J. Hum. Genet.* **16,** 666–72 (2008).

3. Verkerk, A. J. *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65,** 905–14 (1991).

4. Fu, Y. H. *et al.* Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* **67,** 1047–58 (1991).

5. Pieretti, M. *et al.* Absence of expression of the FMR-1 gene in fragile X syndrome. *Cell* **66,** 817–22 (1991).

6. Colak, D. *et al.* Promoter-bound trinucleotide repeat mRNA drives epigenetic silencing in fragile X syndrome. *Science* **343,** 1002–5 (2014).

7. Verheij, C. *et al.* Characterization and localization of the FMR-1 gene product associated with fragile X syndrome. *Nature* **363,** 722–4 (1993).

8. Hinds, H. L. *et al.* Tissue specific expression of FMR-1 provides evidence for a functional role in fragile X syndrome. *Nat. Genet.* **3,** 36–43 (1993).

9. Santoro, M. R., Bray, S. M. & Warren, S. T. Molecular mechanisms of fragile X syndrome: a twenty-year perspective. *Annu. Rev. Pathol.* **7,** 219–45 (2012).

10. Ramos, A. *et al.* The structure of the N-terminal domain of the fragile X mental retardation protein: a platform for protein-protein interaction. *Structure* **14,** 21–31 (2006).

11. Feng, Y. *et al.* Fragile X mental retardation protein: nucleocytoplasmic shuttling and association with somatodendritic ribosomes. *J. Neurosci.* **17,** 1539–47 (1997).

12. Bear, M. F., Huber, K. M. & Warren, S. T. The mGluR theory of fragile X mental retardation. *Trends Neurosci.* **27,** 370–7 (2004).

13. Chen, E., Sharma, M. R., Shi, X., Agrawal, R. K. & Joseph, S. Fragile X mental retardation protein regulates translation by binding directly to the ribosome. *Mol. Cell* **54,** 407–17 (2014).

14. Suhl, J. A., Chopra, P., Anderson, B. R., Bassell, G. J. & Warren, S. T. Analysis of FMRP mRNA target datasets reveals highly associated mRNAs mediated by G-quadruplex structures formed via clustered WGGA sequences. *Hum. Mol. Genet.* 1–41 (2014). doi:10.1093/hmg/ddu272

15. Napoli, I. *et al.* The fragile X syndrome protein represses activity-dependent translation through CYFIP1, a new 4E-BP. *Cell* **134,** 1042–54 (2008).

16. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74,** 285–99 (2012).

17. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506,** 179–84 (2014).

18. Ascano, M. *et al.* FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature* **492,** 382–6 (2012).

19. Brown, V. *et al.* Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome. *Cell* **107,** 477–87 (2001).

20. Darnell, J. C. *et al.* FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism. *Cell* **146,** 247–61 (2011).

21. Miyashiro, K. Y. *et al.* RNA cargoes associating with FMRP reveal deficits in cellular functioning in Fmr1 null mice. *Neuron* **37,** 417–31 (2003).

22. Muddashetty, R. S., Kelić, S., Gross, C., Xu, M. & Bassell, G. J. Dysregulated metabotropic glutamate receptor-dependent translation of AMPA receptor and postsynaptic density-95 mRNAs at synapses in a mouse model of fragile X syndrome. *J. Neurosci.* **27,** 5338–48 (2007).

23. Sung, Y. J. *et al.* The fragile X mental retardation protein FMRP binds elongation factor 1A mRNA and negatively regulates its translation in vivo. *J. Biol. Chem.* **278,** 15669–78 (2003).

24. Darnell, J. C. *et al.* Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell* **107,** 489–99 (2001).

25. Schaeffer, C. *et al.* The fragile X mental retardation protein binds specifically to its mRNA via a purine quartet motif. *EMBO J.* **20,** 4803–13 (2001).

26. Williamson, J. R., Raghuraman, M. K. & Cech, T. R. Monovalent cation-induced structure of telomeric DNA: the G-quartet model. *Cell* **59,** 871–80 (1989).

27. Menon, L., Mader, S. A. & Mihailescu, M. Fragile X mental retardation protein interactions with the microtubule associated protein 1B RNA. *RNA* **14,** 1644–55 (2008).

28. Menon, L. & Mihailescu, M. Interactions of the G quartet forming semaphorin 3F RNA with the RGG box domain of the fragile X protein family. *Nucleic Acids Res.* **35,** 5379–92 (2007).

29. Phan, A. T. *et al.* Structure-function studies of FMRP RGG peptide recognition of an RNA duplex-quadruplex junction. *Nat. Struct. Mol. Biol.* **18,** 796–804 (2011).

30. Bechara, E. G. *et al.* A novel function for fragile X mental retardation protein in translational activation. *PLoS Biol.* **7,** e16 (2009).

31. Darnell, J. C. *et al.* Kissing complex RNAs mediate interaction between the Fragile-X mental retardation protein KH2 domain and brain polyribosomes. *Genes Dev.* **19,** 903–18 (2005).

32. Kim, S. H. *et al.* The general structure of transfer RNA molecules. *Proc. Natl. Acad. Sci. U. S. A.* **71,** 4970–4 (1974).

33. Rastogi, T., Beattie, T. L., Olive, J. E. & Collins, R. a. A long-range pseudoknot is required for activity of the Neurospora VS ribozyme. *EMBO J.* **15,** 2820–5 (1996).

34. Chen, L., Yun, S. W., Seto, J., Liu, W. & Toth, M. The fragile X mental retardation protein binds and regulates a novel class of mRNAs containing U rich target sequences. *Neuroscience* **120,** 1005–17 (2003).

35. Fähling, M. *et al.* Translational regulation of the human achaete-scute homologue-1 by fragile X mental retardation protein. *J. Biol. Chem.* **284,** 4255–66 (2009).

36. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499,** 172–7 (2013).

37. Zalfa, F. *et al.* The fragile X syndrome protein FMRP associates with BC1 RNA and regulates the translation of specific mRNAs at synapses. *Cell* **112,** 317–27 (2003).

38. Zalfa, F. *et al.* Fragile X mental retardation protein (FMRP) binds specifically to the brain cytoplasmic RNAs BC1/BC200 via a novel RNA-binding motif. *J. Biol. Chem.* **280,** 33403–10 (2005).

39. Iacoangeli, A. *et al.* On BC1 RNA and the fragile X mental retardation protein. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 734–9 (2008).

40. Beuth, B., García-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,** 10205–10 (2007).

41. Darnell, J. C., Fraser, C. E., Mostovetsky, O. & Darnell, R. B. Discrimination of common and unique RNA-binding activities among Fragile X mental retardation protein paralogs. *Hum. Mol. Genet.* **18,** 3164–77 (2009).

42. Verkerk, A. J. *et al.* Alternative splicing in the fragile X gene FMR1. *Hum. Mol. Genet.* **2,** 1348 (1993).

43. Rehm, T., Huber, R. & Holak, T. A. Application of NMR in structural proteomics: screening for proteins amenable to structural analysis. *Structure* **10,** 1613–8 (2002).

44. Hollingworth, D. *et al.* KH domains with impaired nucleic acid binding as a tool for functional analysis. *Nucleic Acids Res.* **40,** 6873–86 (2012).

45. Di Marino, D., Achsel, T., Lacoux, C., Falconi, M. & Bagni, C. Molecular dynamics simulations show how the FMRP Ile304Asn mutation destabilizes the KH2 domain structure and affects its function. *J. Biomol. Struct. Dyn.* **32,** 337–50 (2014).

46. De Boulle, K. *et al.* A point mutation in the FMR-1 gene associated with fragile X mental retardation. *Nat. Genet.* **3,** 31–5 (1993).

47. 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,** 1090–8 (2007).

6. RNA Binding Motif Protein 10 (RBM10)

6.1 Introduction

6.1.1 Medical relevance

TARP syndrome is an X-linked condition with pre or postnatal lethality in affected males. Symptoms in males which come to term include clubfoot, heart defects, a small lower jaw and a cleft palate.[1] From the results of massively parallel sequencing of affected individuals from two families, causative mutations were mapped to the gene encoding RNA binding protein 10 (RBM10).[2] One was a nonsense mutation halfway through the gene and the other a frameshift. These mutations could result in total loss of the protein through nonsense mediated decay or in a misfolded protein no longer able to perform its function. The pattern of RBM10 gene expression was mapped in the mouse embryo and robust staining was observed in the first branchial arch (which gives rise to the mandible), second branchial arch, developing limb buds, and tailbud.[2] This pattern of expression correlates well with the human malformations observed in the jaw and limbs.

RBM10 and two closely related proteins, RBM5 and RBM6, have also been linked to cancer. The genes encoding these proteins are frequently mutated or deleted in lung cancers and levels of the proteins are often altered.[3,4] RBM5 is the most well characterised of the three and was found to be downregulated in ~75% of primary lung cell cancers and other cancer samples.[5,6] Furthermore over expression of the protein induces growth arrest, induction of apoptosis and retarded tumour growth indicating it could play a role as a tumour suppressor.[7,8]

6.1.2 Roles of RBM10

RBM10, RBM5 and RBM6 are paralogues believed to have arisen from gene duplication events and share 30-50% amino acid identity.[9] The best studied role of the RBM family is in the regulation of alternative splicing.

Nearly all pre-mRNAs undergo splicing, a process where regions known as introns are removed from the mRNA and regions known as exons remain and go on to form the mature mRNA. The cell achieves this by joining together the ends of two exons and cutting out the intervening intron in a process called splicing. The process is carried out by a large complex known as the spliceosome which is made up of five snRNPs (U1, U2, U4, U5 and U6) and a large set of auxiliary proteins including SF1 and U2AF (Figure 6.1a).[10]

Alternative splicing is an important mechanism by which the cell can generate proteomic diversity from a limited number of genes. In this process different combinations of exons or coding regions are incorporated into the final mature mRNA and different splice variants can have distinct binding properties, intracellular location, enzymatic activities, stability and post-translational modifications to each other. In addition to differences in the translated protein, the transcript splice variants can have different stabilities or translation profiles leading to different expression levels of the proteins (Figure 6.1b).[10]

The most common mode of alternative splicing in higher eukaryotes is exon skipping, where a particular exon is included under some conditions or tissues but excluded in others.[11] Trans-acting regulatory proteins bind to sites on the pre-mRNA in order to direct this process. Proteins can either enhance the splicing reactions thus leading to inclusion of the exon in question or repress splicing leading to a transcript lacking the exon.[12] These regulatory proteins have been found to work in a number of ways. The serine/arginine-rich proteins (SR proteins) positively regulate alternative splicing. They contain a domain with long repeats of serine and arginine residues, the RS domain.[13] They bind to regions of the pre-mRNA and recruit various components of the spliceosome to the splice site. For example T-cell-restricted intracellular antigen 1 (TIA1) binds to a U-rich sequence downstream of some weak 5' splice sites in order to recruit U1 snRNP,[14] while Sam68 recruits U2AF to the 3' splice site of exon V5 of CD44.[15] The most common mechanism by which repressors of alternative splicing function is to bind elements of the splice site thereby preventing components of the spliceosome machinery from binding. Examples of this include PTB which binds to the polypyrimidine tract in the splice site and occludes the binding site of U2AF,[16] and hnRNP A1 which binds upstream of exon 3 in HIV mRNA for tat to prevent binding of U2 snRNP.[17] Another proposed mechanism of splicing repression is that protein-protein interactions between RNA binding proteins bound to the pre-mRNA could result in looping of the RNA, so while the binding site for the components of the

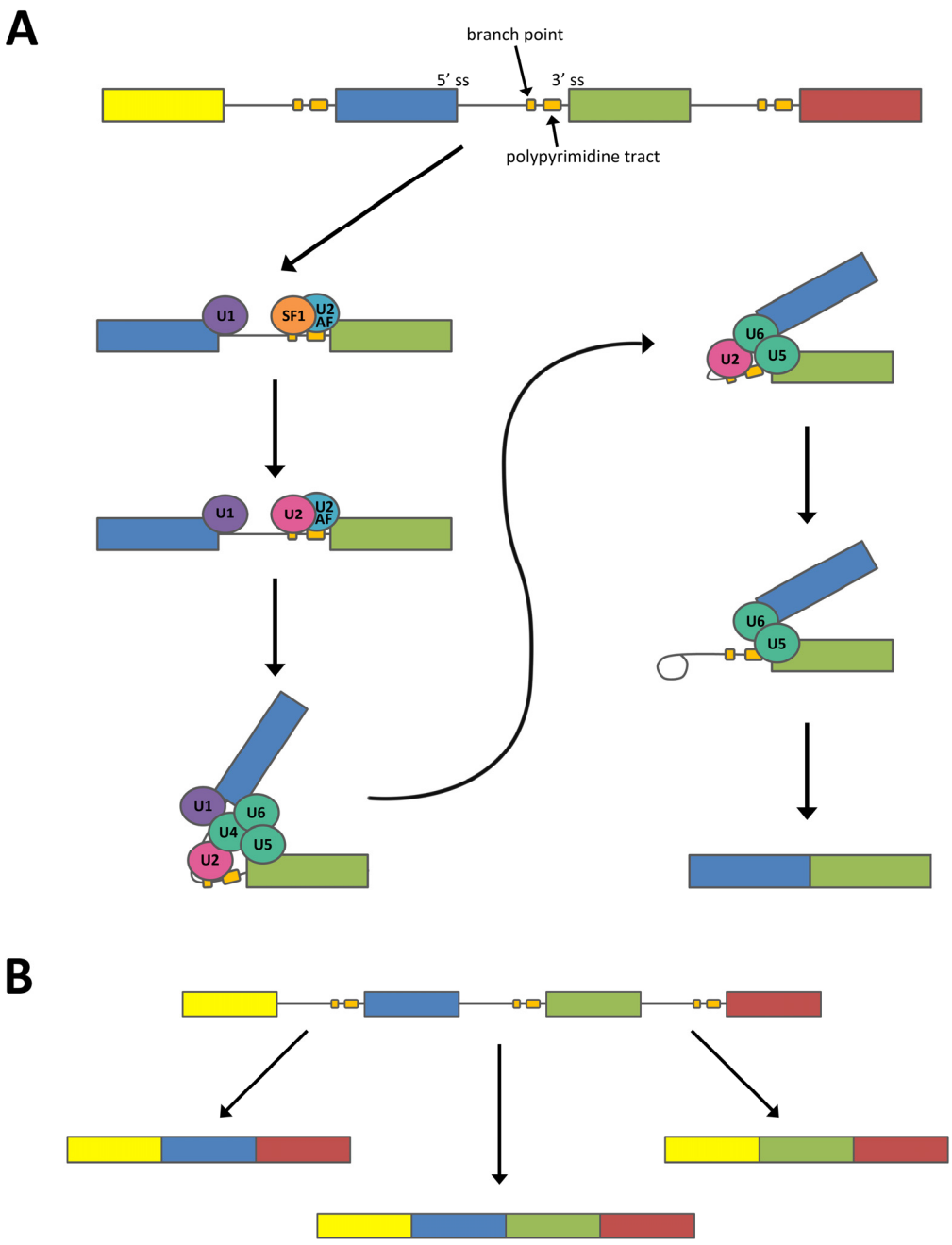


Figure 6.1 | Splicing and alternative splicing. A) Mechanism of splicing. U1 snRNA binds to 5' splice site, splicing factor 1 (SF1) binds to the branch point and the U2 auxilary factor (U2AF) is recruited to the polypyrimidine tract and 3' terminal AG. SF1 is replaced by U2 snRNP at the branch point. The U4/U6-U5 tri-snRNP is recruited. Extensive conformational changes and remodelling, including loss of U1 and U4 snRNPs results in formation of catalytically active spliceosome. B) Alternative splicing allows exons to combine in different configurations to eventually create different isoforms of the protein.

spliceosome are not masked, their positioning on the looped RNA means the complex cannot properly form. PTB is thought to use this mode of repression in the regulation of c-src N1 exon.[18]

Numerous regulators of splicing have been shown to work as both enhancers and repressors, with their function depending on the location of binding. hnRNP H promotes the formation of spliceosome complexes when bound to G-rich regions downstream of the 5' splice site[19] but inhibits this process when bound in exons.[20] Similarly, NOVA1 and NOVA2 inhibit formation of the spliceosomal complex when bound in the exon. They achieve this by altering the position of hnRNPs and inhibiting binding of U1snRNP. However when bound in the intron downstream of the alternatively spliced exon they promote formation of spliceosomal complexes.[21]

Depletion of RBM10 in HeLa cells led to substantial effects on alternative splicing and 97% of binding events identified by CLIP were mapped to alternatively spliced genomic regions.[22] As with many other regulators of alternative splicing the position of RBM10 binding determined whether it played a role as a splicing enhancer or repressor. When bound in the upstream intron or the 3' end of the regulated exon splicing was repressed leading to exon skipping. In contrast binding in the downstream intron was associated with an enhancement in splicing and therefore exon inclusion.[22] This pattern of binding and splicing regulation is similar to that seen for other regulators such as NOVA.[21]

Studies on the impact of RBM10 on one of its target mRNAs, NUMB, found RBM10 binds to the 3' splice site of exon 9 resulting in exon skipping.[22] The NUMB isoform lacking exon 9 is a repressor of the cell proliferation promoting pathway, NOTCH.[22] Changes in the inclusion of exon 9 are commonly found in lung cancer[23] and the NOTCH pathway is an important regulator of cell proliferation, particularly in lung adenocarcinomas,[24] thus providing a link between RBM10 and its observed role in cancer.

Recently RBM10 was identified as having a potential role in the regulation of miRNA biogenesis (unpublished, communication with G. Meister). As described in Chapter 4, pri-miRNAs must undergo rounds of processing by RNases; firstly in the nucleus by Drosha to form precursor miRNA (pre-miRNA) and secondly in the cytoplasm by Dicer to form the mature product.[25] A common factor of pre-miRNAs is a double stranded stem region

ending in a terminal loop.[26] The biogenesis of many pre-miRNAs is regulated by proteins which interact with the terminal loop (Figure 6.2). The proteins up- or downregulate levels of mature miRNA by recruiting RNA enzymes, changing the RNA structure and preventing or enhancing the accessibility and activity of the core processing complexes.[27] For example KSRP interacts with the terminal loop of several let-7 family members and contacts Drosha and Dicer, upregulating their activity.[28] MCPIP1 is a ribonuclease that interacts with and cleaves off the loop in its targets pre-miRNAs. This counteracts Dicer processing and leads to the down regulation of mature miRNA processing.[29] In a large scale screen across 11 cell lines RBM10 bound specifically to a subset of precursor miRNAs, interacting strongly with pre-miR-106b among others. Interestingly it discriminated between two let-7 family members binding strongly to pre-let-7g but only weakly to pre-let-7a. Further experiments verified these interactions and showed that overexpression of RBM10 led to an increase in the processing of pre-let-7g but not pre-let7a (unpublished, communication with G. Meister).

6.1.3 RBM10 domain organisation

RBM10 contains several RNA binding domains; two RRMs, a RanBP2-type zinc finger, a C2H2-type zinc finger and a glycine-rich nucleic binding domain (G-patch)(Figure 6.3). RRMs are well characterised single stranded RNA (ssRNA) binding domains which I have described in detail in Chapter 1. RanBP2-type zinc fingers are found in organisms from plants to mammals. These fingers have been shown to make protein-protein interactions such as between Nup153 and RanGDP,[30] and NPL4 with ubiquitin.[31] However a subset of fingers, including those present in RBM5 and RBM10, has conserved residues shown to be involved in RNA binding. This subset binds with micromolar affinities to ssRNA containing a GGU motif.[32] C2H2-type zinc fingers are the best characterised class of zinc finger and are commonly found in mammalian transcription factors where they bind to dsDNA. Most often they are found in tandem with fingers of the same type and recognize specific bases in the major groove.[33] The G-patch is approximately 48 amino acids with six highly conserved glycine residues. It is often found in tandem with other protein domains including RRMs but so far only one copy of the G-patch has been found per protein.[34] RBM10 also contains functional domains which are not predicted to be involved in RNA binding and so may play other roles important to function. An Octamer Repeat (OCRE) domain, consisting of repeats of eight residues organized around a triplet of aromatic

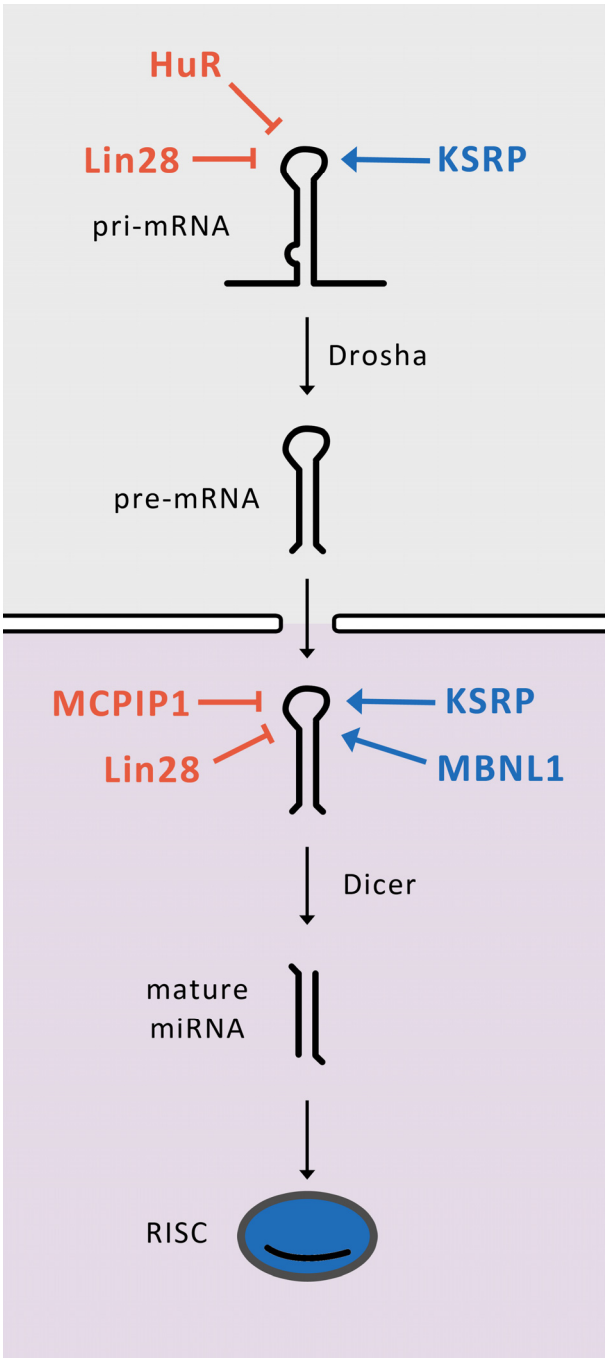


Figure 6.2 | Regulatory proteins functioning in miRNA biogenesis. Pri-miRNA is processed by Drosha and cofactor DGCR8 to form pre-miRNA. Pre-miRNA is exported from the nucleus into the cytoplasm where it is processed by Dicer to form mature miRNA. One strand of RNA is incorporated into the RISC complex. The remaining strand is degraded. Proteins regulate miRNA biogenesis by interacting with the terminal loop of the pri- or pre-miRNA.

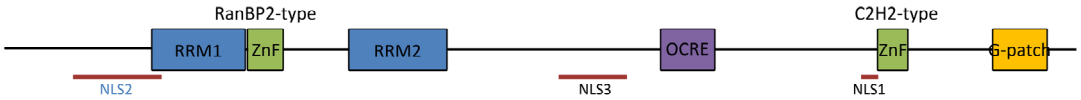


Figure 6.3 | Domain organisation of RBM10. NLS, nuclear localisation signal; RRM, RNA recognition motif; ZnF, zinc finger; OCRE, Octamer repeat domain; G-patch, glycine-rich nucleic binding domain.

amino acids,[35] is present towards the middle of the protein. This domain is also found in RBM5 where it has been shown to make protein-protein contacts with U5 snRNP proteins and play a key role in the alternative splicing regulation of the Fas gene.[36] RBM10 also contains two nuclear localizing signals consistent with its localisation in the nucleus.[37]

6.1.4. Aims

While putative roles have recently been identified for RBM10 as a regulator of alternative splicing and pre-miRNA biogenesis, its regulatory mechanisms are still unknown. Specificity is important in the recognition of its target RNAs, for example in its role in alternative splicing regulation its positioning in long ssRNA regions affects whether RBM10 functions as a repressor or enhancer of splicing. Analysis of CLIP data provided potential binding motifs in relation to its role in alternative splicing and a subset of pre-miRNAs have been identified as targets with RBM10 binding in the apical loop region of the RNAs, but these targets are diverse and little more is known about its RNA binding properties. RBM10 contains several RBDs and a better understanding of how these domains are involved in the recognition of specific targets will allow for more knowledge about the regulatory mechanism of the protein. Furthermore information about different binding modes used to recognise different targets could enable intervention in specific pathways, while leaving others unaffected, to alleviate disease states. Here I aim to determine how RBM10 uses combinations of its RNA binding domains to recognise its targets.

6.2 Methods

6.2.1 Small scale protein expression and solubility screens

The gene encoding human RBM10 was purchased from OriGene. Primers were designed using Crystallisation Construct Designer[38] and used to amplify regions of the RBM10 gene while adding flanking ligation independent cloning sites. A full set of domain boundaries can be found in Figure 6.4 (with numbering relating to RNA-binding protein 10 Isoform 1 NCBI Reference Sequence: NP_005667.2) and the details of constructs used throughout the study are in Table 6.1. PCR products were treated with T4 DNA polymerase to produce single stranded overhangs complementary to those in the pET-52-47 or pET-52-SUMO vectors (provided by Vangelis Christodoulou, National Institute for Medical Research, UK). Ligation independent cloning, protein expression and purification were performed as

previously described in Section 4.2.10. The nucleotide sequence of the plasmids of successfully expressed and soluble constructs was confirmed by DNA sequencing (Beckman Coulter).

Construct Name	Start residue	End residue	Forward Primer	Reverse Primer
RBM10 RRM1	122	207	RBM10_122FW	RBM10_207RV
RBM10 RRM1-ZF	122	248	RBM10_122FW	RBM10_248RV
RBM10 RRM1-RRM2	128	421	RBM10_128FW	RBM10_421RV

Table 6.1 | Table of selected RBM10 constructs and primers used in cloning. Sequences of primers can be found in Appendix V.

6.2.2 Protein expression

50 µl of *E. coli* BL21(DE3) cells were transformed with 2 µl of plasmid using a standard heat shock protocol. 500 µl of transformed cells were used to inoculate 100 ml of M9 minimal media containing $(^{15}NH_4)_2SO_4$ as the only nitrogen source and/or ^{13}C6-D-glucose as the only carbon source. Cells were grown overnight at 37^{o}C. The overnight culture was used to inoculate 1000 ml of M9 minimal media to an OD_{600} of 0.1. Cells were then grown to an OD_{600} of 0.6 before the temperature was reduced to 22^{o}C. Protein expression was induced with IPTG at a final concentration of 0.5 mM. Cells were grown overnight at 22^{o}C, harvested by centrifugation and stored at -80^{o}C.

6.2.3 Protein purification

Frozen cells were resuspended in equilibration buffer (10 mM Tris-HCl pH 8.0, 10 mM imidazole, 200 mM NaCl, 2 mM β-mercaptoethanol) (20 ml per 1 L of cell culture) with Triton X-100, DNaseI and lysozyme, sonicated on ice and centrifuged at 17000 rpm for 60 mins. The recombinant protein was purified by immobilised metal ion affinity chromatography (IMAC) columns. The soluble fraction was incubated with Ni-NTA resin (5ml per litre of culture) at 4^{o}C for 30 mins then poured into a column. The column was washed with 2x5 CV washing buffer (10 mM Tris-HCl pH 8.0, 10 mM Imidazole, 1 M NaCl, 2 mM β-mercaptoethanol) and the protein was eluted with 5 CV elution buffer (10 mM Tris-HCl pH 8.0, 250 mM Imidazole, 1 M NaCl, 2 mM β-mercaptoethanol). HRV 3C protease was used to cleave the His-Tag or His-SUMO-Tag by incubation overnight at 4^{o}C The sample was dialysed in Spectra/Por Dialysis Membrane of the appropriate MWCO in 4 litres of equilibration buffer then the cleaved tag was separated from the protein by IMAC (Ni-NTA). The cleavage reaction mixture was loaded into a gravity-driven column packed with

Ni-NTA resin equilibrated with 10 CV of equilibration and the flow through loaded for a second time. The resin was washed with 5 CV of equilibration buffer, 5 CV of 50mM imidazole wash buffer (10 mM Tris-HCl pH 8.0, 50 mM imidazole, 200 mM NaCl, 2 mM β-mercaptoethanol), 5 CV 10 mM imidazole wash buffer (10 mM Tris-HCl pH 8.0, 100 mM imidazole, 200 mM NaCl, 2 mM β-mercaptoethanol), then the tag was eluted with 5 CV of elution buffer. The protein was purified further on a HiTrap HeparinHP 5ml column (GE Healthcare) to remove nucleic acid contamination. Protein sample was dialysed into Heparin Buffer A (10 mM Phosphate, 50 mM NaCl, 2 mM β-mercaptoethanol) and loaded onto the column via a superloop. After loading the column was washed with 2 CV of Heparin Buffer A before a gradient of increasing NaCl concentration was run over 10 CV up to 1 M NaCl using Heparin Buffer B (10 mM Phosphate, 1 M NaCl, 2 mM β-mercaptoethanol) to elute the protein. The nucleic acid-free protein containing fractions were then concentrated to 5ml using Vivaspin concentrators in order to be purified further by size exclusion chromatography. Size exclusion chromatography was performed using an ÄKTA purifier system (GE Healthcare) with a Hiload 16/60 Superdex prep grade column (GE Healthcare) equilibrated in Heparin Buffer A. Fractions of pure protein were pooled and concentrated before being dialysed into a final buffer of 10 mM phosphate pH 6.9, 50 mM NaCl, 1 mM TCEP, 10 μM $ZnCl_2$ and concentrated in a Vivaspin MWCO 10000. Protein concentration was determined from the absorbance at 280 nm and molecular weights confirmed by mass spectrometry.

6.2.4 RNA binding assays - NMR

50 μM ^{15}N-labelled samples of RBM10 in 10 mM phosphate pH 6.9, 50 mM NaCl, 1 mM TCEP and 10 μM $ZnCl_2$ were titrated with unlabelled RNA oligomers up to protein to RNA ratios of 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 600 or 700 MHz ^{1}H frequencies.

6.2.5 RNA binding assays - Circular dichroism

All CD experiments were performed on a Jasco J-815 spectropolarimeter equipped with CDF-426S temperature-control system. Protein samples were prepared in 600 μl of 2 μM RNA pre-miNRA stem loop RNA or 4-5 μM of 7-mer RNA oligonucleotide was loaded into the cuvette and aliquots of RBM10 added with spectra from 240 nm to 320 nm recorded upon each addition. Titrations were recorded at 5^{o}C or 20^{o}C.

6.2.6 RNA binding assays - Electrophoretic mobility retardation assays

Samples of 5 nM biotinylated pre-miRNA stem loop were incubated with increasing amounts of protein in buffer 50 mM NaCl, 10 mM phosphate pH 6.9 and 0.5 mM TCEP, ranging from 0 nM to 800 nM, for 20 minutes at room temperature. Samples were then run on a 6% DNA retardation gel at 100 V for 1 hour before being transferred onto a nylon membrane. The RNA was visualised using the Pierce Chemiluminescence Kit following the manufacturer's instructions.

6.3 Characterisation of RBM10 RNA binding domains

To characterise the RNA binding capabilities of RBM10 we first limited our studies to the three closely grouped canonical RNA binding domains found towards the C-terminal of the protein; the two RRM domains and the RanBP2-type zinc finger.

NMR structures of the individual RRM domains are available (RRM1 PDB:2LXI and RRM2 PDB:2M2B) and were used in conjunction with secondary structure prediction to determine domain boundaries. We cloned constructs containing RRM1, RRM1-ZF and RRM1-RRM2 with the terminal domains having several different N- and C-termini boundaries (Figure 6.4). Small scale expression trials were run for these constructs with either a His-Tag or His-SUMO-Tag (Figure 6.5). The RRM1 and RRM1-ZF constructs showed increased solubility with the His-SUMO-Tag while the RRM1-RRM2 constructs were more soluble with the His-Tag. Upon large scale production of the most successful constructs the protein showed good expression and solubility. After a purification procedure consisting of nickel affinity chromatography, cleavage and separation of the tag, heparin affinity chromatography and finally size exclusion chromatography, yields of up to 40 mg per ml of purified protein were obtained (Figure 6.6).

For the RRM1-RRM2 constructs four different C-termini were trialled. During the purification process it was observed that the shortest construct, RRM1-RRM2 (128-408), was being degraded. The degradation occurs at the C-terminus as the construct is still able to bind to a nickel column indicating an intact His-Tag at the N-terminus. Upon lengthening of the construct the amount of degraded product appears to decrease from roughly 50% for RRM1-RRM2 (106-408) to less than 10% for RRM1-RRM2 (106-421) (Figure 6.7). This pattern was observed regardless of the N-terminal domain boundary. The constructs were analysed by mass spectroscopy (performed by Steve Howell, NIMR). RRM1-RRM2 (128-408)

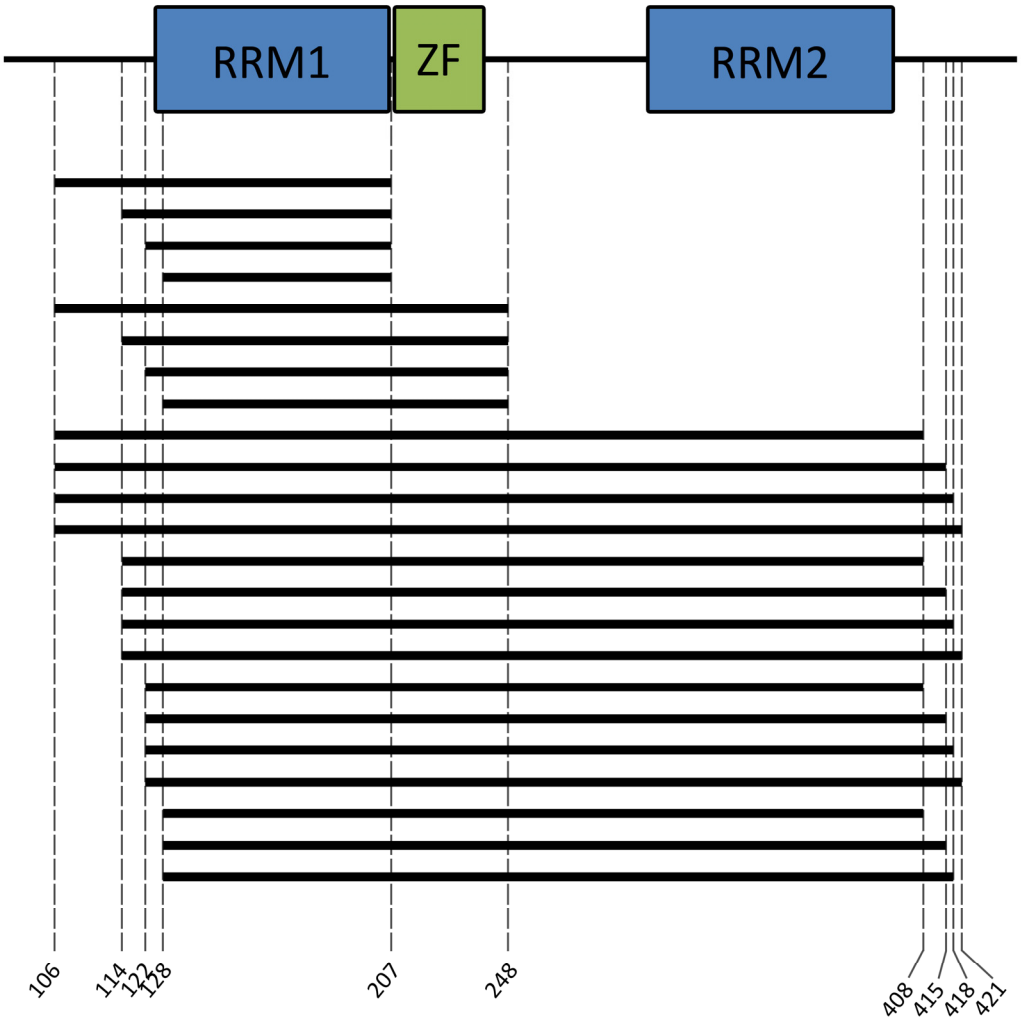


Figure 6.4 | Scheme of RBM10 constructs in small scale protein expression and solubility screens.

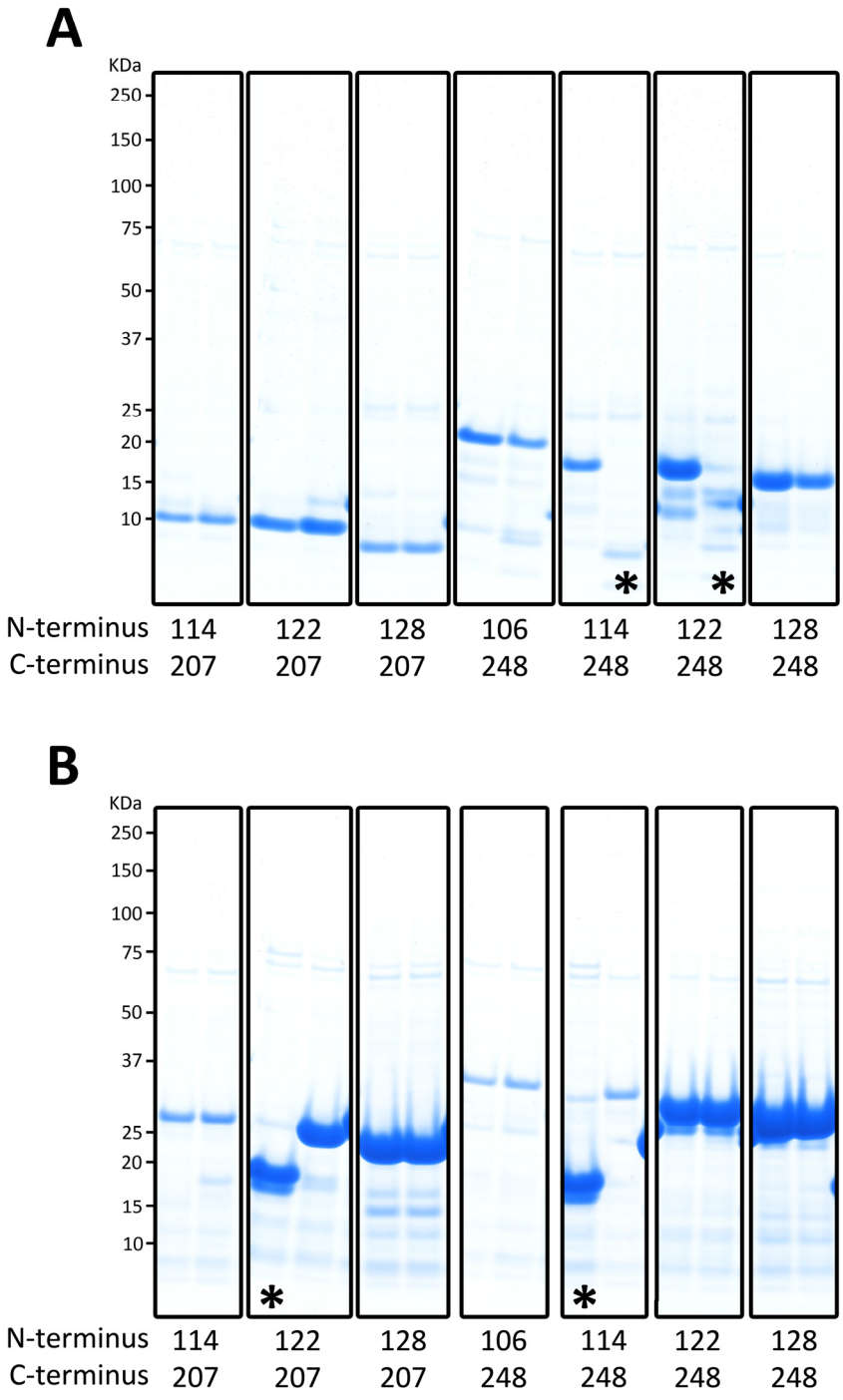


Figure 6.5 | Purified protein from RBM10 small scale expression and solubility screening. Sections from SDS-page gels showing RBM10 protein constructs after purification using nickel-NTA. Lanes where cloning has failed resulting in an empty vector are marked *.

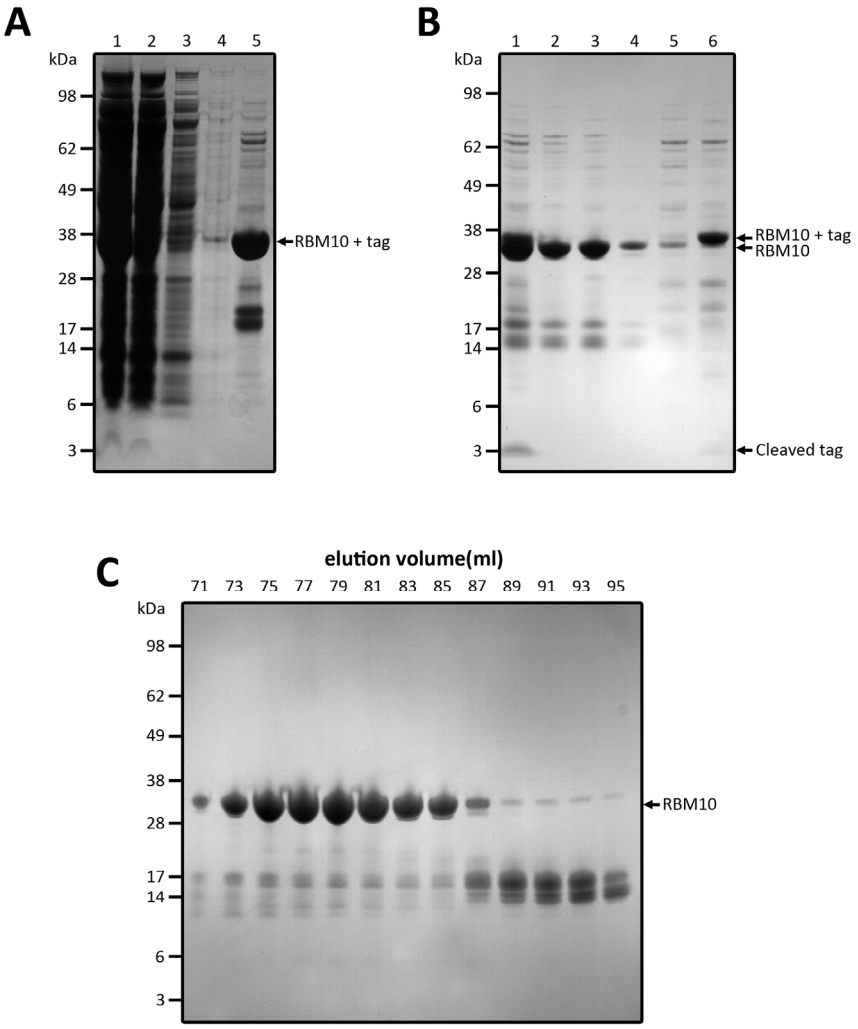


Figure 6.6 | Purification strategy employed for RBM10. A) SDS-page analysis of affinity chromatography fractions of RBM10 using nickel-NTA after cell lysis and centrifugation. 1) cleared cell lysate, 2) flow through after column loading with cleared cell lysate, 3) wash with 5 CV wash buffer (10 mM Tris-HCl pH 8.0, 10 mM Imidazole, 1 M NaCl, 2 mM β-mercaptoethanol) , 4) second wash with 5 CV wash buffer, 5) elution 5 CV of elution buffer (10 mM Tris pH 8.0, 250 mM Imidiazole, 1 M NaCl, 2 mM β-mercaptoethanol). B) SDS-page analysis of affinity chromatography fractions of RBM10 using nickel-NTA after His-Tag cleavage with HRV 3C protease. 1) total cleavage reaction, 2) flow through after column loading, 3) wash with 5 CV equilibration buffer (10 mM Tris pH 8.0, 10 mM imidazole, 200 mM NaCl, 2 mM β-mercaptoethanol), 4) wash with 5 CV 50mM imidazole buffer (10 mM Tris pH 8.0, 50 mM imidazole, 200 mM NaCl, 2 mM β-mercaptoethanol), 5) wash with 5 CV 100 mM imidazole buffer (10 mM Tris pH 8.0, 100 mM imidazole, 200 mM NaCl, 2 mM β-mercaptoethanol), 6) elution with 5 CV elution buffer. C) SDS-page analysis of size exclusion chromatography fractions of RBM10 using a Superdex 75 16/60. Lanes are identified with elution volume in ml.

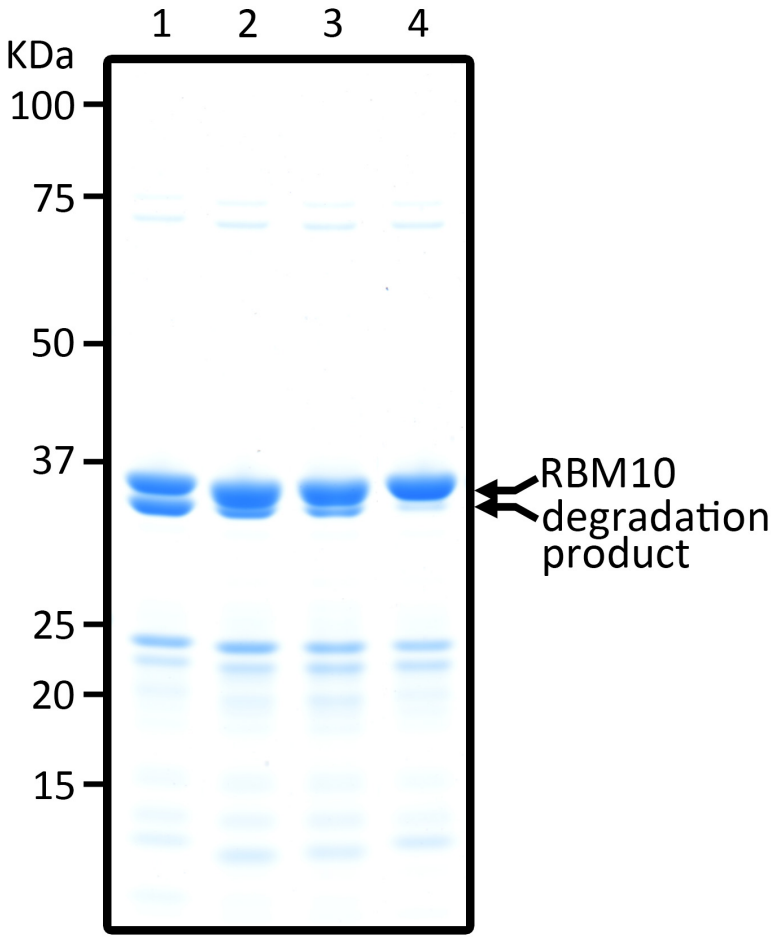


Figure 6.7 | N-terminal degradation of RBM10 RRM1-RRM2. SDS-page analysis of purified protein after His-Tag cleavage and removal. 1) RBM10 RRM1-RRM2 (106-408), 2) RBM10 RRM1-RRM2 (106-415), 3) RBM10 RRM1-RRM2 (106-418), 4) RBM10 RRM1-RRM2 (106-421).

showed only a small peak which could correspond to the full length construct at 31411 Da. Another much larger peak was observed at 29716 Da which would correlate with a product shortened by 18 amino acids at the C-terminus. ^{1}H-^{15}N SOFAST-HMQC experiments were recorded to compare the fold of the constructs. The spectra are mostly identical, with less than 10 peaks shifted or missing, indicating there is little change to the overall fold of the domains (Figure 6.8). The NMR structure of the individual RRM2 shows the region from G384 to G408 to be unstructured therefore it is plausible for part of this tail to be proteolytically cleaved. It is possible that the extension at the N-terminus contains some secondary structure that protects the tail from degradation. However the NMR structure ends at G408 so we are unable to verify this hypothesis until further structural work is performed. All further experiments were performed with RRM1 (122-207), RRM1-ZF (122-248) and RRM1-RRM2 (128-421).

To assess whether the three domains interact we compared ^{1}H-^{15}N SOFAST-HMQC spectra of the different domains. First we assigned peaks to the individual domains by comparison of the spectra from RRM1, RRM1-ZF and RRM1-RRM2. Using this assignment we can see how residues of domains are affected upon removal of other domains. For RRM1-RRM2 the spectra contained over 200 peaks for a construct of 293 residues (279 excluding prolines). The peaks showed good dispersion, indicating regions of secondary structure, however heavy overlap in the centre of the spectra made it difficult to fully distinguish all the peaks. This overlapped region may originate from the 57 amino acid linker between the ZF and RRM2 which is likely to be unstructured. In this construct we were able to assign 35 peaks to RRM1, 13 peaks to the ZF and 32 peaks to RRM2 or the linker region. A full list of peak assignment can be found in Appendix XV. Peaks assigned to all three domains are present in the dispersed region of the spectra indicating all the domains are folded. If RRM2 were interacting with either RRM1 or the ZF the truncation of RRM2 would result in a large change in the chemical environment of residues in the interaction surface causing chemical shift perturbation. Upon truncation of RRM2 we observed that of the 35 peaks we were able to assign to RRM1 only 3 of them shifted significantly. Four more showed medium shifts and the remainder did not shift or shifted only slightly (Figure 6.9). For the 13 peaks assigned to the ZF four displayed small shift perturbations and the others were unaffected. The only slight chemical shift perturbation of peaks belonging to RRM1 and ZF upon removal of RRM2 suggests that these domains do not interact in the free form of the protein which is unsurprising due to the presence of the long linker between the ZF and

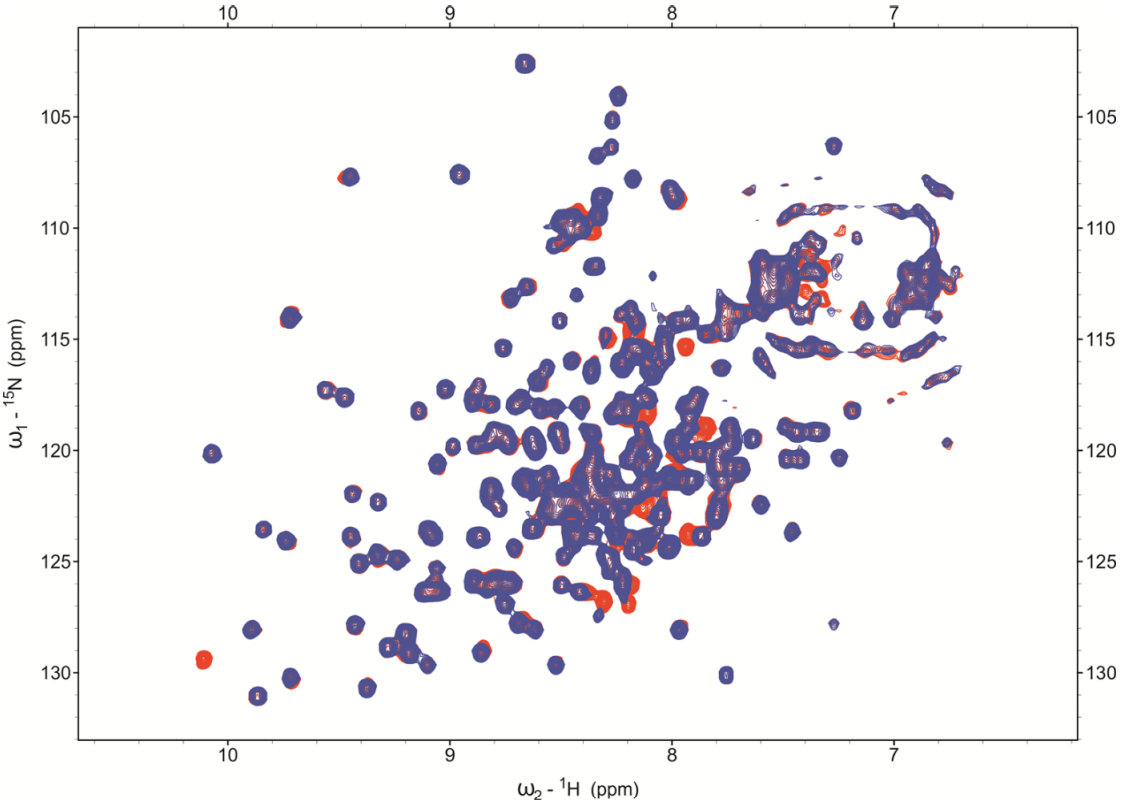


Figure 6.8 | Comparison of constructs of RBM10 RRM1-RRM2 with different N-terminal boundaries. Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of RBM10 RRM1-RRM2 (128-421) (red) and RBM10 RRM1-RRM2 (128-408) (blue).

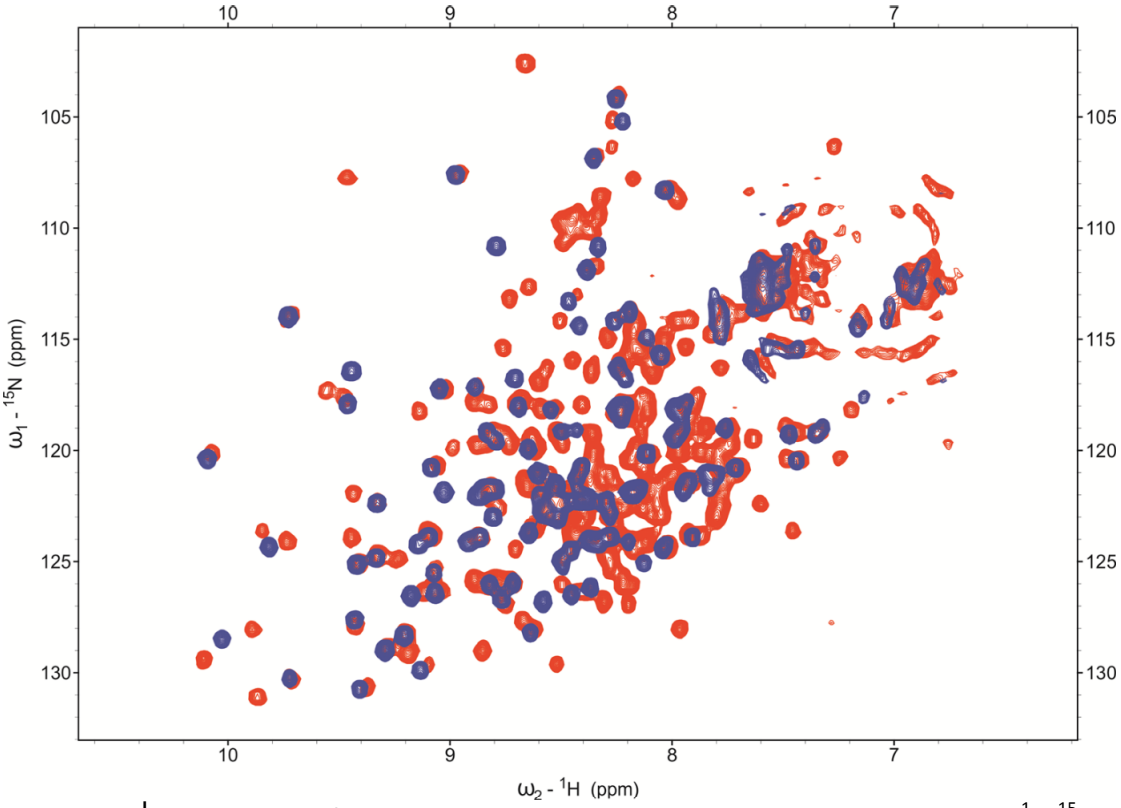


Figure 6.9 | Comparison of RBM10 RRM1-RRM2 and RBM10 RRM1-ZF. Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of RBM10 RRM1-RRM2 (red) and RBM10 RRM1-ZF (blue).

RRM2. This is consistent with the linewidth of peaks in the RRM1-RRM2 construct which is smaller than we would expect for a compact, globular, 300 amino acid protein.

We followed the same logic to investigate the interaction between RRM1 and the ZF. Here the linker between the domains is much shorter so it is more likely an interaction will be present. We compared the spectra of RRM1 and RRM1-ZF and a further spectrum of RRM1-ZF with the addition of EDTA. This stripped the finger of its coordinated zinc ion causing the ZF to unfold. Here we observe that a subset of peaks assigned to RRM1 is perturbed both by the unfolding and truncation of the ZF thus indicating an interaction with the folded region of the finger (Figure 6.10). Another subset of peaks from RRM1 shifts only upon truncation of the finger but is unaffected by the unfolding. This subset could be interacting with the linker between the domains or with other unfolded sections in the construct such as to the N-terminal of the ZF. A further subset is not affected by either modification indicating no contact with the ZF. These results indicate that there is an interaction surface between RRM1 and ZF. If this is the case these domains could be working together to recognise a stretch of continuous RNA.

6.4 RNA binding of RBM10

6.4.1 Splice site sequences

From CLIP-seq experiments enriched sequence motifs for RBM10 could be determined and electrophoretic mobility shift assays (EMSA) showed binding to two of these sequences, CGAUCCC and CUGUGGA, to have apparent affinities of 150 nM.[22] We aimed to determine if our RRM1-RRM2 construct was able to bind to the preferred sequence motifs and investigate the affinity of any interaction observed. Titrations were performed with RRM1-RRM2 and CUCUGGA or CGAUCCC and monitored by circular dichroism (CD) spectroscopy. The changes in the CD spectra of the RNA were recorded upon addition of increasing amounts of protein. For both RNAs the CD was affected upon addition of protein indicating an interaction with RBM10. Values for the dissociation constants of the interactions were obtained by plotting the averaged CD signal between 269 nm and 272 nm against protein concentration and fitting to a one-site binding model. The obtained Kd values are 43.7±25.2 nM for CUCUGGA and 2.0±0.7 μM for CAGUCCC (Figure 6.11).

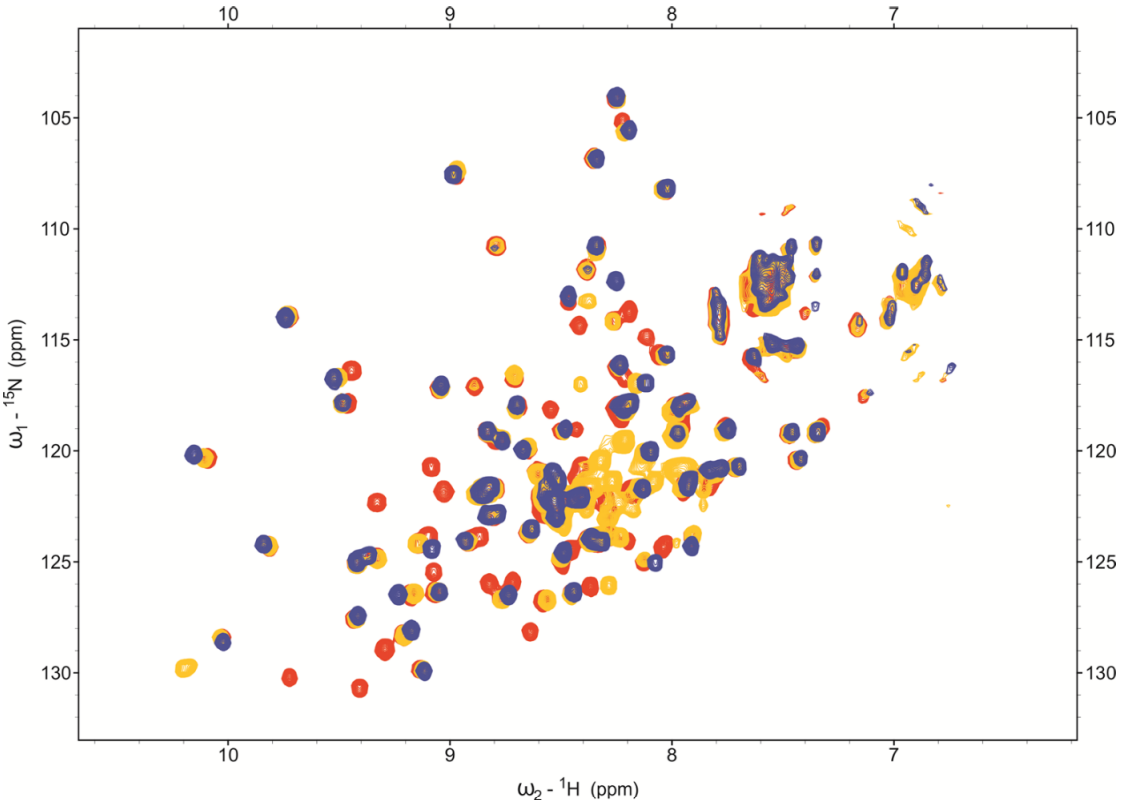


Figure 6.10 | Comparison of RBM10 RRM1-ZF, RBM10 RRM1-ZF plus EDTA, and RBM10 RRM1. Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of RBM10 RRM1-ZF (red), RBM10 RRM1-ZF with 5mM EDTA (yellow), and RBM10 RRM1 (blue).

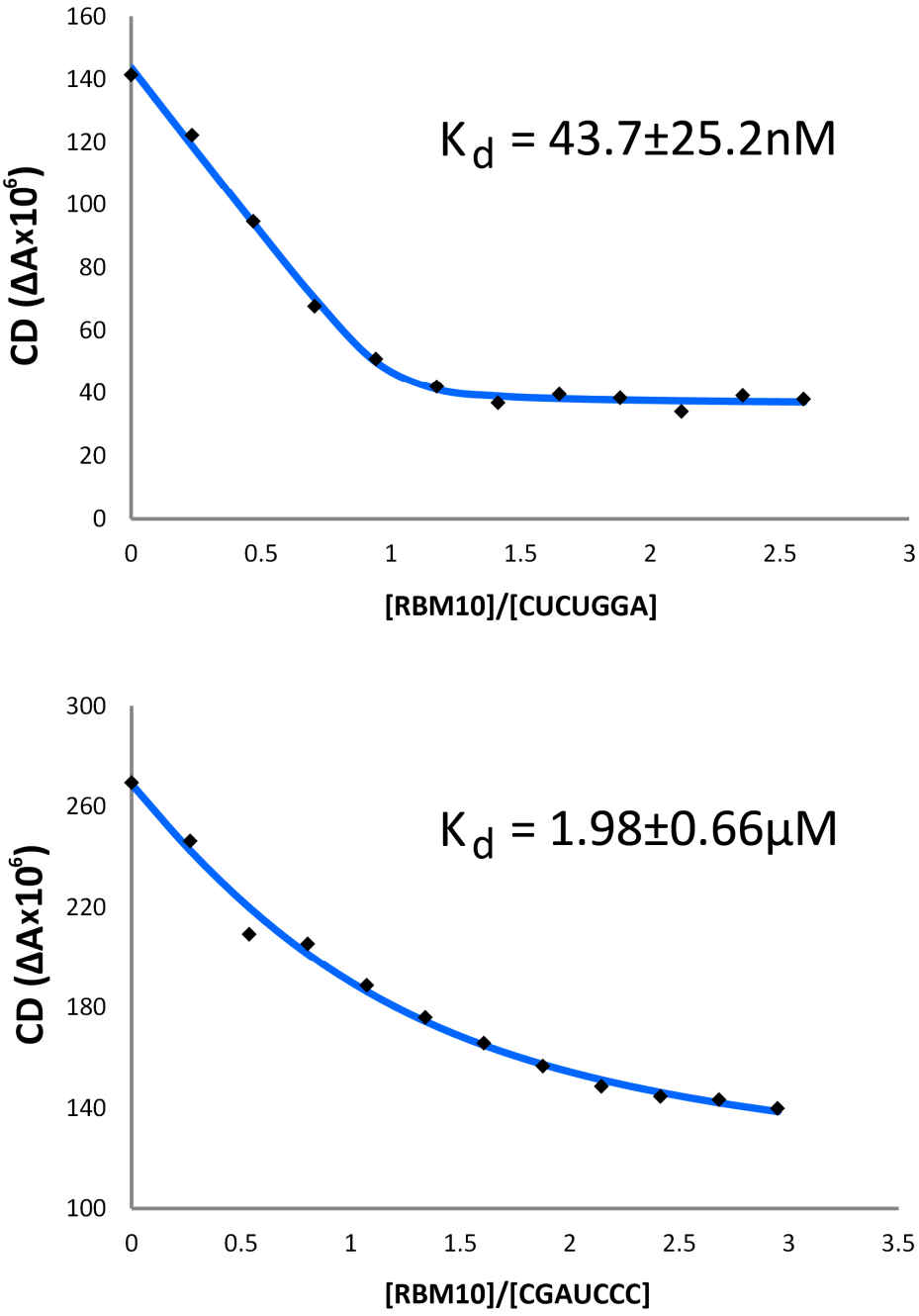


Figure 6.11 | Binding of splice site sequences to RBM10 RRM1-RRM2. A) Binding curve of CUCUGGA. B) Binding curve of CGAUCCC

We wanted to examine the protein-RNA interaction further in order to determine which domains were involved in the binding and if this differed between the two sequences. ^{15}N-labelled RRM1-RRM2 was monitored by ^{1}H-^{15}N correlation spectroscopy upon the addition of increasing amounts of CUCUGGA or CGAUCCC. Upon addition of CUCUGGA a subset of peaks in the intermediate-to-slow or slow exchange regimes shifted and became saturated at around one RNA equivalent. A second subset of peaks began to shift at around one RNA equivalent and these peaks are in the fast exchange regime. The first subset of peaks can be attributed to RRM1 and ZF while the second subset to RRM2. This indicates that CUCUGGA binds preferentially to RRM1 and ZF. Furthermore the intermediate-to-slow and slow exchange regime of the peaks belonging to RRM1 and the ZF confirm that the interaction is high affinity. Once this preferred, high affinity binding surface is saturated CUCUGGA can bind to RRM2 (Figure 6.12). In the titration with CAGUCCC peaks from all three domains beginning to shift at roughly the same time and the majority of these are in the fast exchange regime. This confirms the weaker binding observed by CD. Upon binding to CUCUGGA a large number of peaks from the ZF are greatly perturbed however this is not the case for the CAGUCCC sequence indicating that the ZF only binds to CUCUGGA.

The titrations so far had been performed with ssRNA that was 7nt in length and we aimed to determine if this was sufficient to fill the high affinity binding site or if the protein could accommodate a longer stretch of RNA. We performed titrations of RRM1-ZF with 9-mer RNA oligonucleotides. These were made up of the tight binding sequence CUCUGGA with an additional two random nucleotides added at either the 3' or 5' end giving NNCUCUGGA and CUCUGGANN. Binding was monitored by ^{1}H-^{15}N correlation spectroscopy. By comparison of these titrations with that of the 7-mer RNA (CUCUGGA) we could see that no additional peaks are perturbed upon lengthening of the RNA (Figure 6.13). This indicates that 7nt are sufficient to fill the whole binding site and no additional nucleotides can be accommodated.

We attempted to gain some insight into the specificity of the RRM1-ZF binding site by comparing the binding of RNAs (based on the CLIP derived motifs) with slightly different sequences, CUCUGAA, CUCUGGA and CUGUGGA. Binding was monitored by ^{1}H-^{15}N correlation spectroscopy. Upon addition of the RNA we observed peak shift perturbations in both domains. For the interaction with CUCUGAA the amplitude of peak shifts could be measured and used to determine a Kd of ~50 μM. Accurate Kds could not be determined

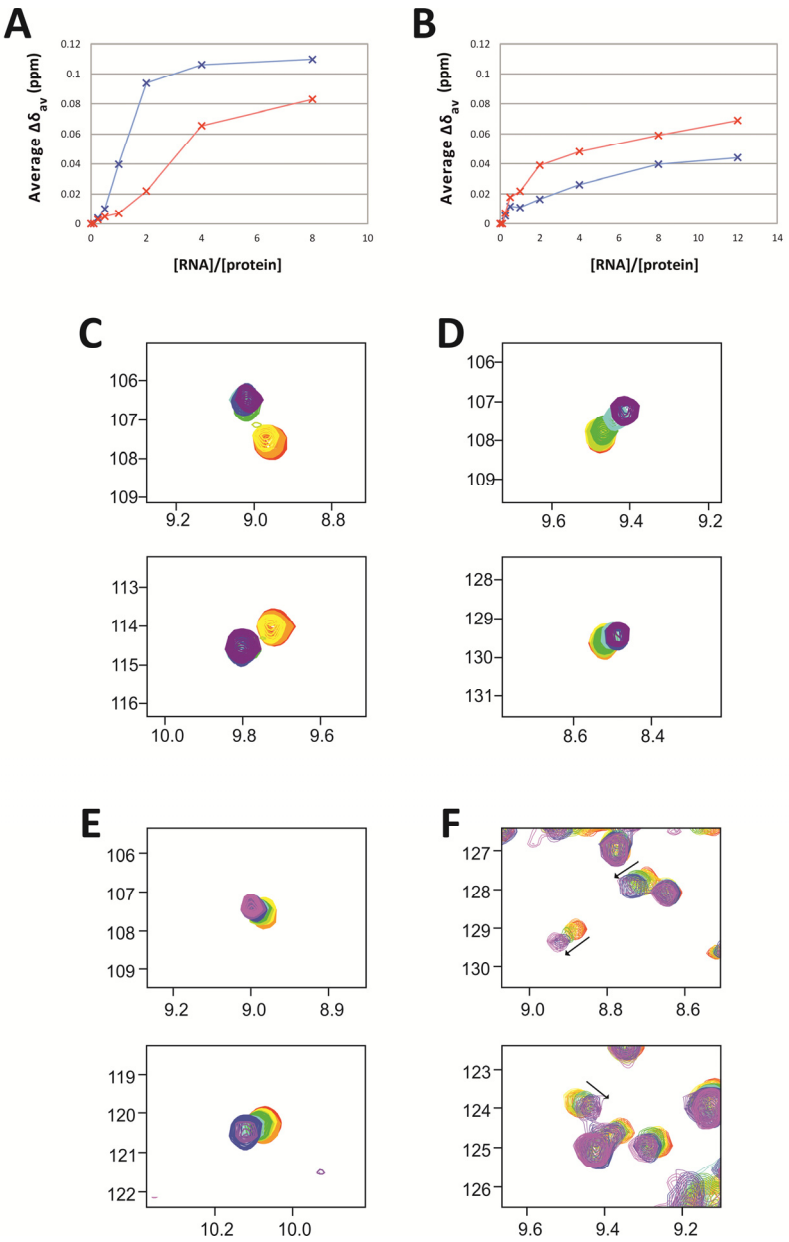


Figure 6.12 | Comparison of binding of splice site sequences to RRM1 and RRM2. A-B) Average chemical shift perturbation upon binding to, A) CUCUGGA, and B) CGAUCCC of residues in RRM1 (red) and RRM2 (blue). The y-axis shows average chemical shift perturbation in ppm and the x-axis shows RBM10 RRM1-RRM2 to CUCUGGA ratio. C-D) Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of 70 µM RBM10 RRM1-RRM2 with CUCUGGA at protein to RNA ratios of 1:0 (red), 1:0.1 (orange), 1:0.25 (yellow), 1:0.5 (light green), 1:1 (green), 1:2 (cyan), 1:4 (blue), and 1:8 (purple). C) Peaks from RRM1. D) Peaks from RRM2. E-F) Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of 65 µM RBM10 RRM1-RRM2 with CGAUCCC at protein to RNA ratios of 1:0 (red), 1:0.1 (orange), 1:0.25 (yellow), 1:0.5 (light green), 1:1 (green), 1:2 (cyan), 1:4 (blue), 1:8 (purple), and 1:12 (magenta). E) Peaks from RRM1. F) Peaks from RRM2.

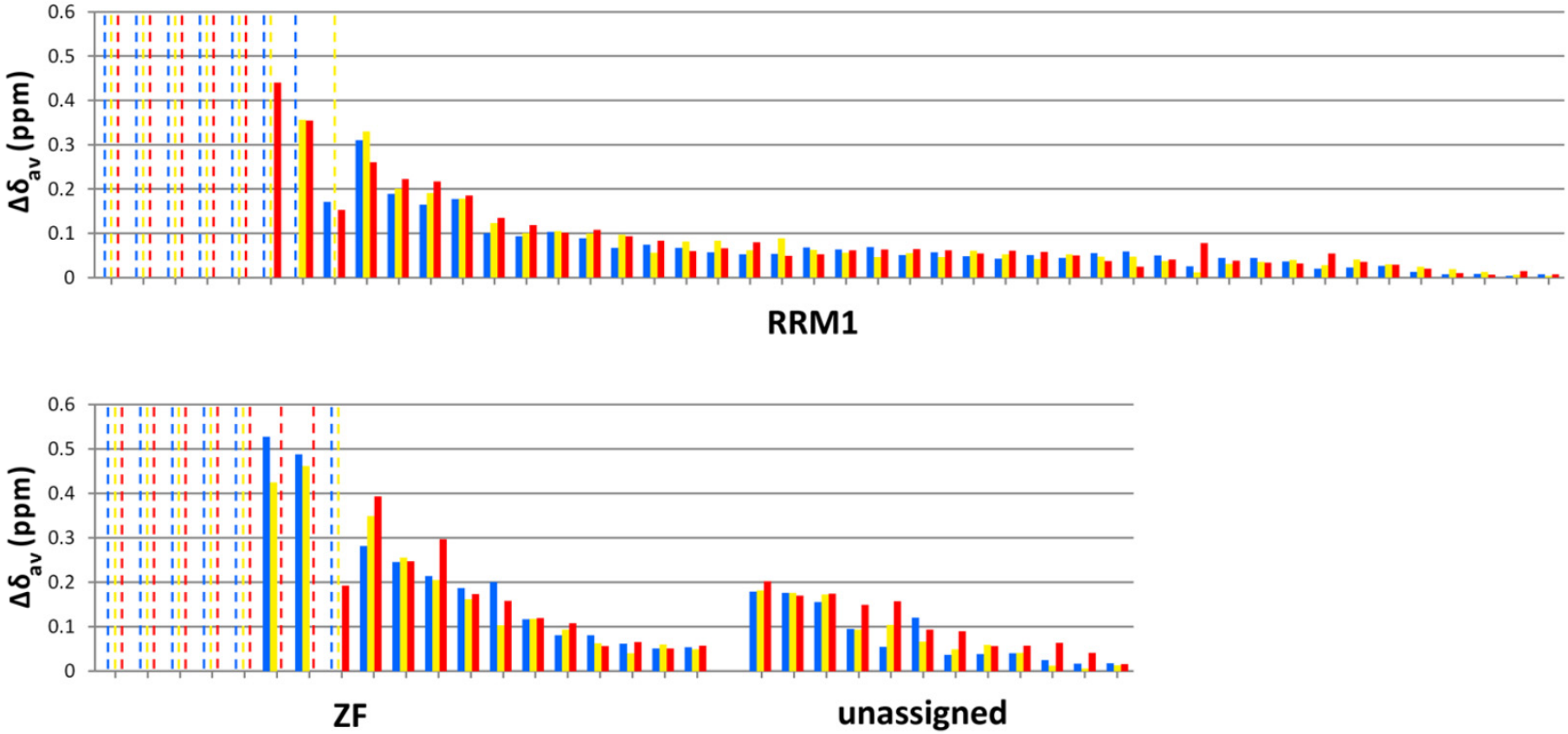


Figure 6.13 | **Chemical shift perturbation of RBM10 RRM1-ZF upon addition of 7-mer and 9-**mer RNA oligonucleotides

Upon addition of two equivalents of CUCUGGA (blue), CUCUGGANN (yellow), and NNCUCUGGA (red). Dashed lines represent peaks which disappear due to line broadening. The y-axis shows weighted average chemical shift perturbation in (ppm).

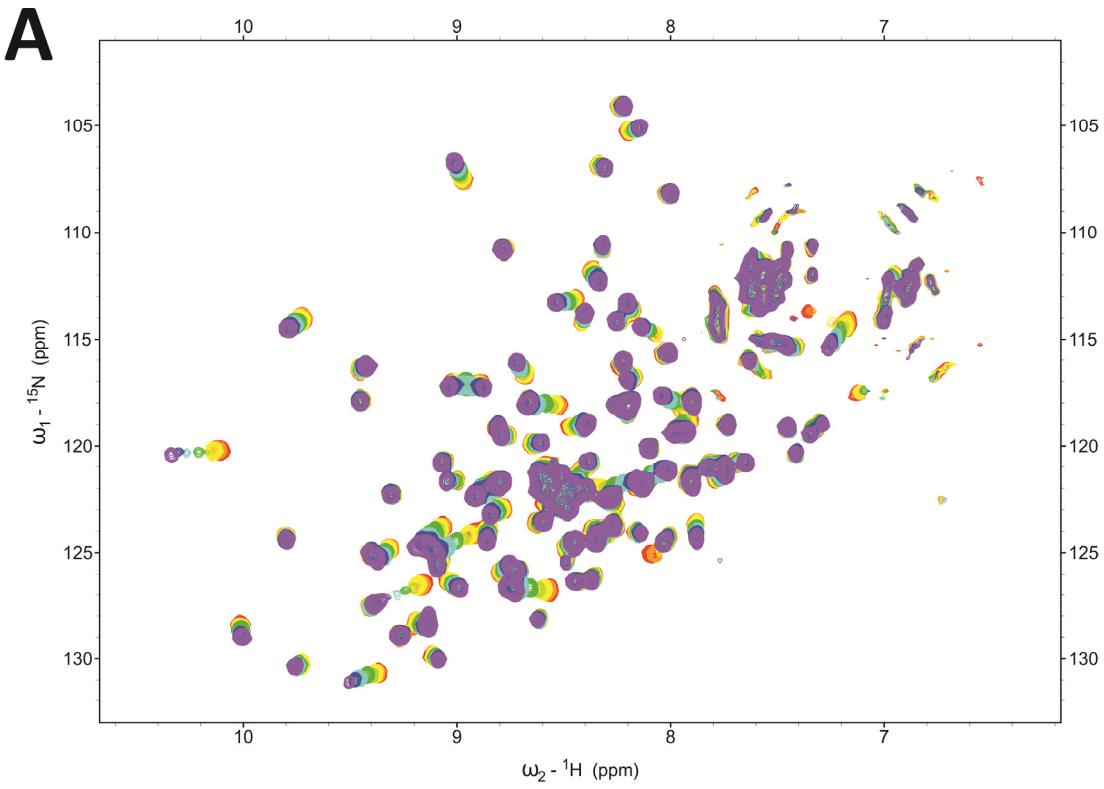
for the interactions with CUCUGGA and CUGUGGA however both titrations reach saturation at protein to RNA ratios of 1:1 while the titration with CUCUGAA is still not fully saturated at 1:8. Additionally, in the titrations with CUCUGGA and CUGUGGA some of the peaks are in the slow exchange regime indicating tight binding while the majority of peaks in the CUCUGAA titration are in fast exchange (Figure 6.14). From these observations we can determine that CUCUGGA and CUGUGGA bind with higher affinity than CUCUGAA.

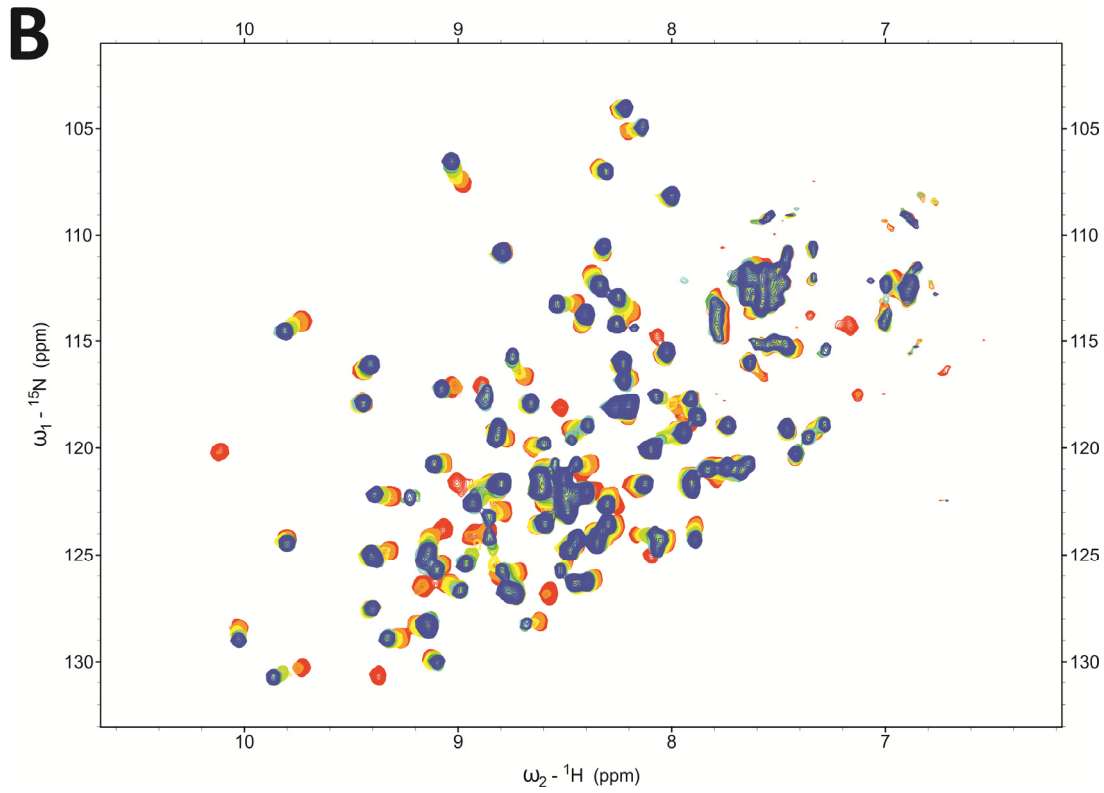
The weakest binding, CUCUGAA only contains one G towards the 3' end. Fluorescence studies on the binding of the RanBP2-type finger of ZRANB2 show that the mutation of one of the guanines in the preferred binding motif decreases the affinity of binding approximately 4-fold.[32] The lack of a strong binding site for ZF would explain the weaker binding of CUCUGAA compared with CUCUGGA and CUGUGGA, both of which contain two consecutive guanines near the 5' end. As CUGUGGA and CUCUGGA both contain the predicted strong ZF binding site but CUGUGGA binds more tightly than CUCUGGA it indicates RRM1 has specificity in its binding as it prefers a guanine over a cytosine in position 3 of the sequence. This is in agreement with early work done on the RBM10 protein which found it binds more strongly to poly(G) and poly(U) sequences than poly(C) or poly(A).[39] Furthermore out of the ten most common sequences predicted from the CLIP data only two of them contain sequences predicted to bind strongly to the ZF and the same situation is observed for RBM10 close family members RBM5 and RBM6.[22] This indicates that RBDs other than the RanBP2-type zinc finger play a role in determining the specificity of binding and in the future it would be interesting to determine the sequence specificity of the other domains in order to see which confer targeting of these sequences.

6.4.2 Pre-miRNA

A role for RBM10 in miRNA biogenesis has been postulated so we aimed to characterise the binding to its putative pre-miRNA targets. Studies by the Meister group found pre-let-7g and pre-miR-106b to be strong targets of RBM10 whereas pre-let-7a was a weaker target or not targeted at all.

We attempted to determine the affinity of binding to the stem loops of pre-let-7g, pre-miR-106b and pre-let-7a. We observed the change in CD signal of the pre-miRNA stem loops upon increasing amounts of RRM1-RRM2. All three RNAs were affected by addition of the protein indicating an interaction was occurring (Figure 6.15). None of the curves could be


A
10
9
8
7
105
110
115
120
125
130
ω1 - 15N (ppm)
ω2 - 1H (ppm)


B
10
9
8
7
105
110
115
120
125
130
ω1 - 15N (ppm)
ω2 - 1H (ppm)

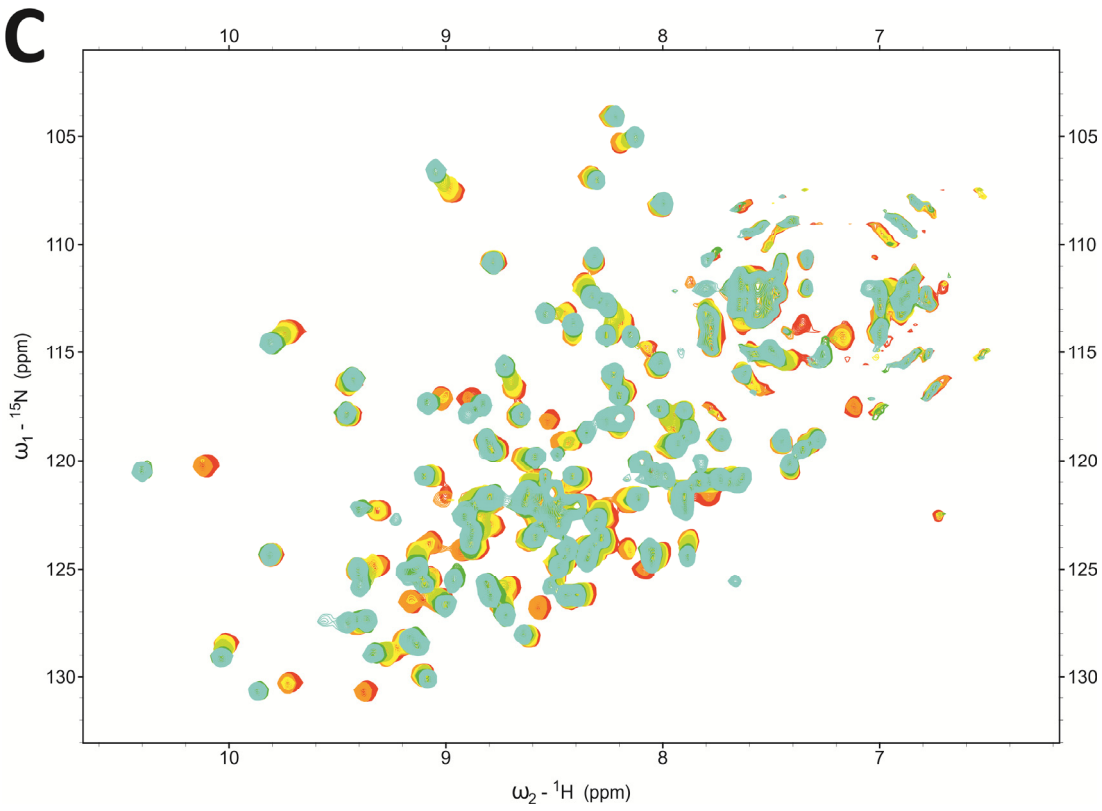


Figure 6.14 | Comparison of RBM10 RRM1-ZF peaks shift perturbation upon binding to CUCUGAA, CUCUGGA and CUGUGGA. Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of RBM10 RRM1-ZF with A) CUCUGAA at protein to RNA ratios of 1:0 (red), 1:0.1 (orange), 1:0.25 (yellow), 1:0.5 (light-green), 1:1 (green), 1:2 (cyan), 1:4 (blue) and 1:8 (purple), B) CUCUGGA at protein to RNA ratios of 1:0 (red), 1:0.1 (orange), 1:0.25 (yellow), 1:0.5 (light-green), 1:1 (green), 1:2 (cyan), and 1:4 (blue) and C) CUGUGGA at protein to RNA ratios of 1:0 (red), 1:0.1 (orange), 1:0.25 (yellow), 1:0.5 (light-green), 1:1 (green), and 1:2 (cyan).

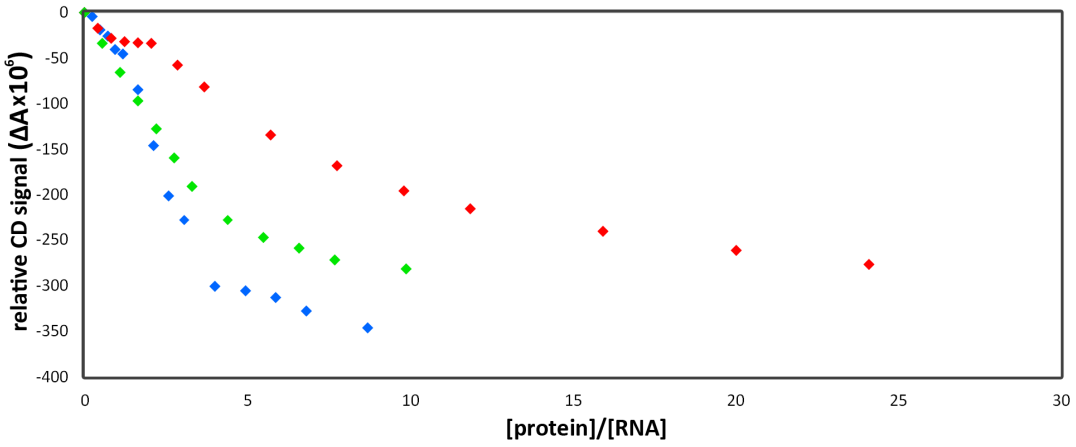


Figure 6.15 | Comparison of pre-miRNA stem loops binding to RBM10 RRM1-RRM2. Plots are of protein to RNA ratio on the x-axis and CD signal relative to a starting point of zero on the y-axis. Plots of pre-let-7g stem loop (blue), pre-miR-106b stem loop (green), and pre-let-7a stem loop (red).

fitted using a one-or two-site binding model and Kds could not be determined. However from visual analysis of the binding curves both pre-let-7g and pre-miR-106b have a steeper slope and begin to reach plateau at a lower RNA to protein ratio than for pre-let-7a. This indicates pre-let-7a binds more weakly than both pre-let-7g and pre-miR-106b.

As our construct contains three RNA binding domains and the RNAs are between 37 and 47nt in length we postulated that several different combinations of interaction could be taking place involving more than one protein binding to a single RNA molecule or a single protein binding to more than one RNA molecule. To gain more information on the affinity and stoichiometry of the interactions we performed electrophoretic mobility shift assays. For pre-miR-106b we saw a clean shift from the unbound RNA to one distinct band of RBM10 bound RNA suggesting a stoichiometry of one to one and is contrary to the fact we were unable to fit the binding data obtained by CD using a one-site binding model. The gels confirmed the tighter binding of pre-miR-106b and pre-let-7g compared to pre-let-7a observed by CD. Pre-miR-106b and pre-let-7g have Kds of 50-100 nM while pre-let-7a has a Kd of 100nM-200 nM. For pre-let-7a and pre-let-7g two bands were observed upon binding to RBM10, one at roughly the same level as bound miR-106b and one much higher up the gel. This indicates that while some RNA molecules have one RBM10 bound other RNAs are interacting with multiple RBM10 molecules. At higher protein concentrations the single bound band weakens in intensity and the higher band disappears completely indicating the formation of large protein-RNA complexes that are unable to migrate through the gel. Furthermore for pre-let-7a the lower band appears split into two which could be due to two slightly different secondary structures of the RNA being present leading to altered migration through the gel (Figure 6.16).

In order to determine which domains were involved in the binding of the pre-miRNA targets we performed a titration RRM1-RRM2 with increasing amounts of pre-miR-106b monitored by ^{1}H-^{15}N correlational spectroscopy. Upon addition of RNA many peaks belonging to the three domains are perturbed which could indicate all three are involved in the interaction with the pre-miRNA (Figure 6.17). However we cannot rule out that peak shift perturbation could be a consequence of domain-domain interactions rather than a direct interaction with the RNA. At two RNA equivalents some of these peaks have reappeared but many are still missing. We also performed a titration monitored by CD of pre-miR-106b with increasing amounts of RRM1-ZF. Upon comparison of this titration with

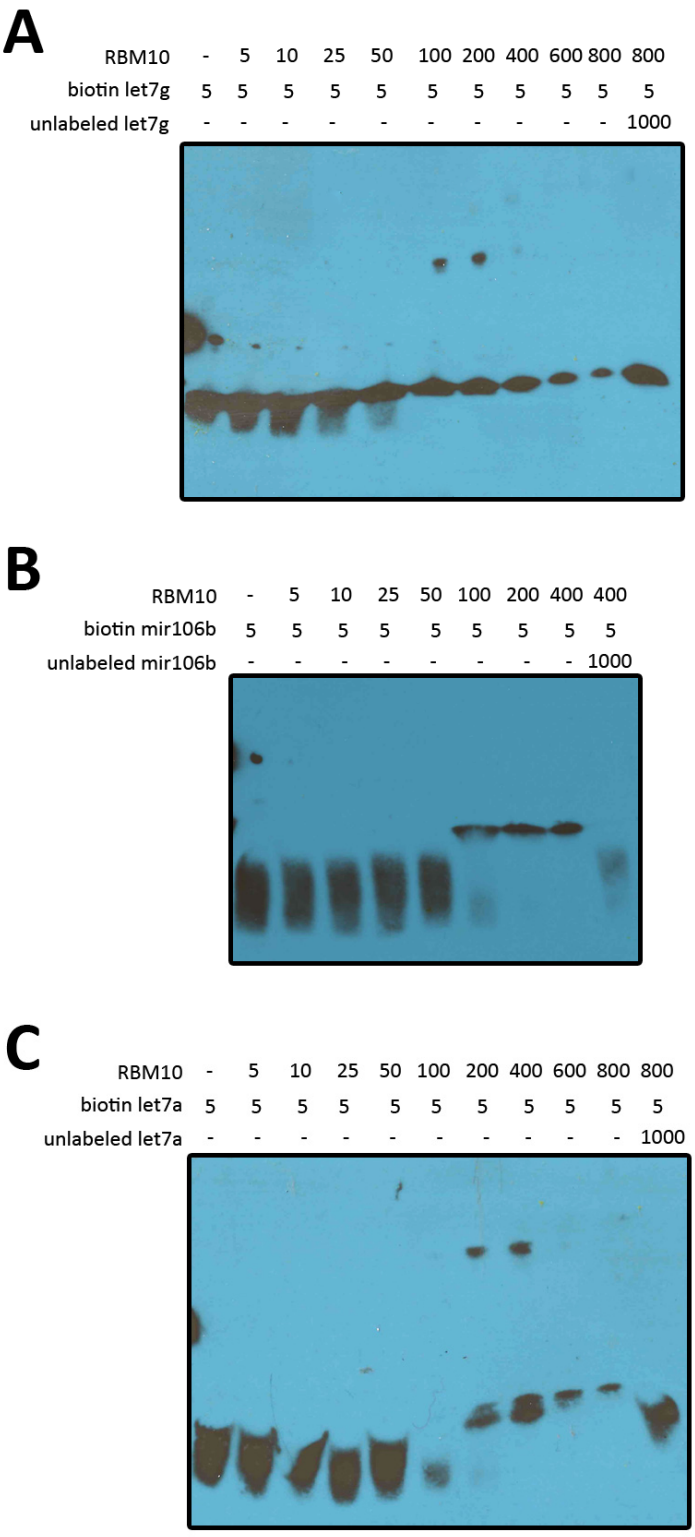


Figure 6.16 | Electropheretic mobility shift assays of pre-miRNA stem loops binding to RBM10 RRM1-RRM2. Samples were run on 6% DNA retardation gels. Concentrations are in nM. A) pre-let-7g stem loop. B) pre-miR-106b stem loop. C) pre-let-7a stem loop.

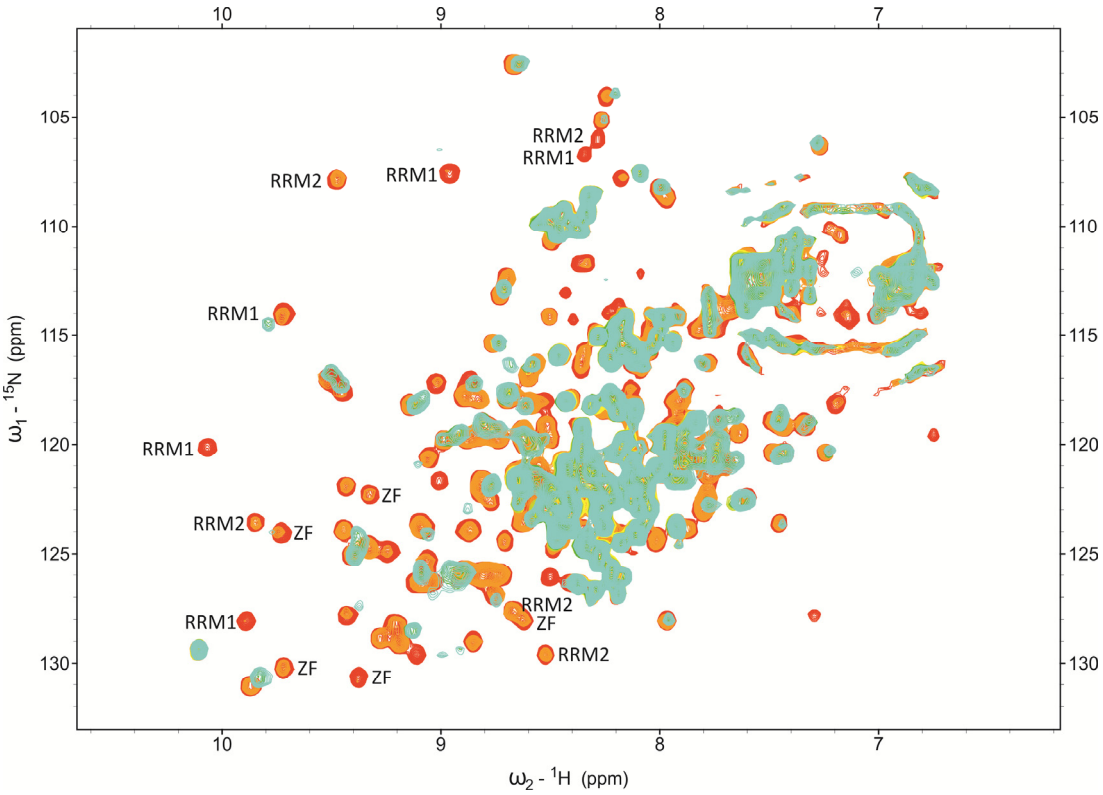


Figure 6.17 | Titration of RBM10 RRM1-RRM2 with pre-miR-106b stem loop. Overlaid ^{1}H-^{15}N SOFAST-HMQC spectra of 70 µM RBM10 RRM1-RRM2 with pre-miR-106b at protein to RNA ratios of 1:0 (red), 1:0.1 (orange), 1:0.25 (yellow), 1:0.5 (light green), 1:1 (green), and 1:2 (cyan). Selected peaks from are labelled with the domain they originate from to demonstrate all three domains participate in binding.

the one of pre-miR-106b with RRM1-RRM2 we observed a much smaller change in the RNA CD signal upon addition of the protein (Figure 6.18). This indicates that in RRM1-RRM2 all three domains are interacting with the RNA and thus causing more change to the base stacking interactions that mostly make up the observed CD signal. When only the RRM1 and ZF domains are present to bind much less disturbance of the stacking takes place and thus a smaller change in CD signal is observed.

6.5 Discussion

In this chapter I discuss how we have analysed the RNA binding of the RanBP2-type zinc finger and the two RRM domains of RBM10 with its predicted RNA targets or target sequences.

By comparison of constructs containing different segments of RBM10 either RRM1, RRM1-ZF or RRM1-RRM2 we could determine that the RRM2 does not interact with either RRM1 or the ZF in the free form of the protein while RRM1 and the ZF make contact, potentially forming one continuous binding surface. The extended binding surface of RRM1 combined with the ZF and contribution of RRM2 could allow the protein to be more specific in its binding by being able to recognise longer stretches of continuous ssRNA. In the case of PAPB two RRM domains interact to form one continuous binding surface which can accommodate eight nucleotides[40] while the second and third RRMs of PTB both contain an extra β strand in the β sheet surface which allows for the recognition of one and two additional nucleotides respectively.[41,42] When binding in the alternatively spliced exon or in the upstream intron RBM10 has been shown to promote exon skipping. Other splicing repressors have been shown to work by competing with components of the spliceosome or splicing enhancers for binding to the RNA.[16,17] The potential extended binding site of RBM10 could aid in rendering larger segments of the RNA inaccessible for binding of other proteins. Another proposed mechanism of action of both splicing enhancers and silencers is to alter the local structure of the mRNA. Enhancers would cause a change in structure that would better present the splice sites of the alternative exon for binding of the spliceosome elements while the change caused by silencers this would impede splice site recognition. If the extended binding surface of RRM1-ZF were rigid this may better enable the remodelling of the mRNA.

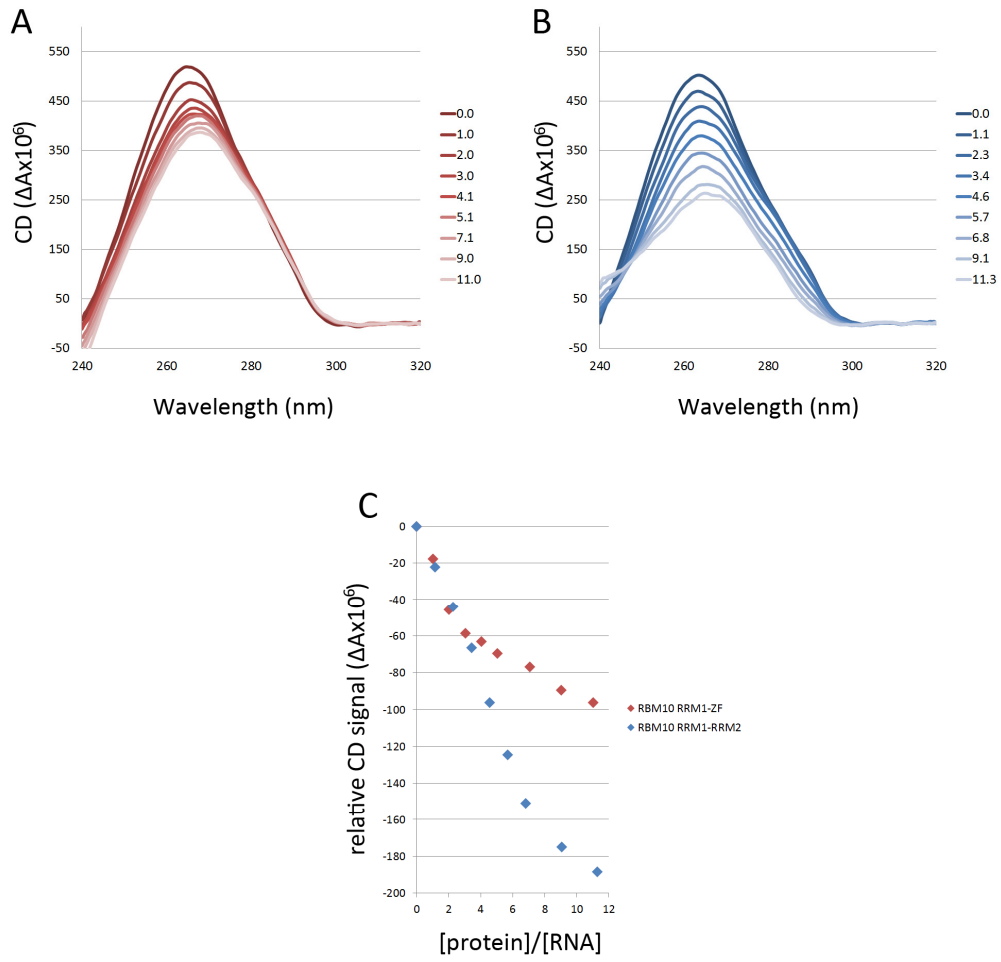


Figure 6.18 | Comparison of pre-miR-106b stem loop interaction with RBM10 RRM1-ZF and RBM10 RRM1-RRM2. A) CD titration spectra of pre-miR-106b upon addition of RBM10 RRM1-ZF. Key shows protein to RNA ratio. B) CD titration spectra of pre-miR-106b upon addition of RBM10 RRM1-ZF. Key shows protein to RNA ratio. C) Relative CD signal change in pre-miR-106b upon addition of RBM10 RRM1-ZF (red) or RBM10 RRM1-RRM2 (blue).

Several interaction motifs were identified by analysis of CLIP data. The enrichment of the sequences around splice sites indicates these are functional binding sites for RBM10 in the regulation of alternative splicing.[22] By CD we determined the dissociation constants of two of these sequences to be 43 nM and 2 μM. This 40-fold difference in affinity between the

two varies greatly from published EMSA data which found both sequences to bind with a Kd of 150 nM.[22] Unlike our CD titrations the EMSA experiments by Bechara et al. were carried out in an excess of a non-specific RNA competitor which could lead to the lower affinity observed for CUGUGGA. In addition the EMSA experiments used a construct which while containing the same three RNA binding domains also contains the whole of the protein N-terminal of these domains. There are no predicted RNA binding domains in this N-terminal region but it may be that sections are able to interact with the RNA thus accounting for the higher affinity towards the CAGUCCC sequence. A basic region in TUT7 containing six arginine residues interacts with RNA with a Kd of 40nM.[43] The N-terminal of RBM10 contains a stretch of seven arginine residues (interspersed with one histidine) which could act in a similar manner.

Comparison of the binding monitored by NMR showed that CUCUGGA binds firstly to RRM1 and the ZF with peaks shifting in the intermediate to slow regime indicating tight binding. Once this high affinity site is saturated the RNA begins to interact with RRM2. Here the peaks are mostly in the fast exchange regime indicating a weaker interaction. For CAGUCCC we do not observe this high affinity interaction with RRM1-ZF and it binds weakly to all three domains. It has previously been determined that RanBP2-type zinc fingers ZnF Ran binding domain-containing protein 2 (ZRANB2) bind to a core motif of GGU.[44] In the enriched binding motifs for RBM10 the sequence GGA appears more frequently than GGU. In the binding of the second zinc finger of ZRANB2 to GGU the uridine forms hydrogen bonds with two asparagine side chains.[44] Sequence alignment shows that in RBM10 these asparagine residues are changed to a valine and a phenylalanine which could account for the difference in specificity. The high affinity binding sequence of CUCUGGA contains a putative binding site for the ZF. It could be that the ZF binds to GGA at the 3' end of the RNA thus positioning the rest of the RNA for optimal binding to the RRM1. Also if the domains are interacting as we suggest and bind as one unit the overall affinity would be the product of that of the two domains which could lead to the strong binding we observe. The weaker binding sequence CAGUCCC does not contain a binding site for the ZF and this

could lead to the lower affinity we observe. We also determined the binding site of RRM1-ZF to be seven nucleotides or less. RRM domains have been found to bind from two to eight nucleotides however more typically an RRM domain with no additional secondary structural elements binds between four and six nucleotides. This would fit with our assumptions of the GG element binding to the ZF thus leaving four nucleotides to interact with the RRM.

Characterisation of the binding of RBM10 to pre-miRNA showed stronger binding to its putative targets pre-let-7g and pre-miR-106b than to pre-let-7a which it is predicted not to regulate. Future work with the pre-miRNA targets will provide more information about target recognition. We showed that all three domains are involved in the binding to pre-miR-106b. As both RRMs and RanBP2-type zinc fingers are single stranded RNA binding domains we would predict that they would interact with single stranded elements in the pre-miRNA, such as the stem loop and/or bulges in the stem. However as we do not have a clear idea of the specificity of the RNA binding domains of RBM10 it is difficult to determine what elements in the pre-miRNA RBM10 are being recognised. This presents difficulties in understanding how RBM10 discriminates between pre-miRNAs. Furthermore the long linker between RRM1-ZF and RRM2 could give the protein flexibility in binding to its pre-miRNA with regards to the distance between specific binding sequences and the structural arrangement of said sequences. Lin28 and KSRP are also regulators of the biogenesis of a subset of miRNAs, including the let-7 family. They both contain multiple domains that recognise short ssRNA sequences and use them in a coordinated fashion to recognise their targets. In Lin28 two CCHC-type zinc fingers of the protein bind specifically to a GGAG sequence in the stem loop of the target pre-miRNA and a CSD makes contacts with a different section of the loop. The CCHC-type zinc fingers confer most of the specificity of the binding with the CSD contributing mostly to affinity and not specificity. A flexible linker between the CSD and zinc fingers allows Lin28 to recognise different members of the let-7 family which all have slightly different stem loops and an unconserved distance between the binding sites for the two domains.[45]

We have determined how the RRM1-ZF domains work together as a unit to recognise a subset of splice site sequences containing a preferred binding site for the RanBP2-type ZF binding site while in the absence of this site RRM1 appears to bind alone. Furthermore we have shown that all three RNA binding domains are involved in the interaction with pre-

miR-106b. In the future I plan to gain more structural insight into the organisation of the RNA binding domains upon interaction with these different sets of targets. Another avenue of research would be to characterise the binding of the C2H2-zinc finger in the C-terminal of the protein. Furthermore knowledge of the specificity of the RRM domains and insight into the molecular basis of this specificity would aid in the understanding of RBM10 target recognition.

6.6 References

1. Gripp, K. W. *et al.* Long-term survival in TARP syndrome and confirmation of RBM10 as the disease-causing gene. *Am. J. Med. Genet. A* **155A,** 2516–20 (2011).

2. Johnston, J. J. *et al.* Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am. J. Hum. Genet.* **86,** 743–8 (2010).

3. Angeloni, D. Molecular analysis of deletions in human chromosome 3p21 and the role of resident cancer genes in disease. *Brief. Funct. Genomic. Proteomic.* **6,** 19–39 (2007).

4. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150,** 1107–20 (2012).

5. Oh, J. J., West, A. R., Fishbein, M. C. & Slamon, D. J. A candidate tumor suppressor gene, H37, from the human lung cancer tumor suppressor locus 3p21.3. *Cancer Res.* **62,** 3207–13 (2002).

6. Zhao, L. *et al.* 3p21.3 tumor suppressor gene RBM5 inhibits growth of human prostate cancer PC-3 cells through apoptosis. *World J. Surg. Oncol.* **10,** 247 (2012).

7. Oh, J. J. *et al.* 3p21.3 tumor suppressor gene H37/Luca15/RBM5 inhibits growth of human lung cancer cells through cell cycle arrest and apoptosis. *Cancer Res.* **66,** 3419–27 (2006).

8. Mourtada-Maarabouni, M., Keen, J., Clark, J., Cooper, C. S. & Williams, G. T. Candidate tumor suppressor LUCA-15/RBM5/H37 modulates expression of apoptosis and cell cycle genes. *Exp. Cell Res.* **312,** 1745–52 (2006).

9. Timmer, T. *et al.* An evolutionary rearrangement of the Xp11.3-11.23 region in 3p21.3, a region frequently deleted in a variety of cancers. *Genomics* **60,** 238–40 (1999).

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

11. Sugnet, C. W., Kent, W. J., Ares, M. & Haussler, D. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.* **77,** 66–77 (2004).

12. Chen, M. & Manley, J. L. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell Biol.* **10,** 741–54 (2009).

13. Long, J. C. & Caceres, J. F. The SR protein family of splicing factors: master regulators of gene expression. *Biochem. J.* **417,** 15–27 (2009).

14. Förch, P., Puig, O., Martínez, C., Séraphin, B. & Valcárcel, J. The splicing regulator TIA-1 interacts with U1-C to promote U1 snRNP recruitment to 5' splice sites. *EMBO J.* **21,** 6882–92 (2002).

15. Tisserant, A. & König, H. Signal-regulated Pre-mRNA occupancy by the general splicing factor U2AF. *PLoS One* **3,** e1418 (2008).

16. Saulière, J., Sureau, A., Expert-Bezançon, A. & Marie, J. The polypyrimidine tract binding protein (PTB) represses splicing of exon 6B from the beta-tropomyosin pre-mRNA by directly interfering with the binding of the U2AF65 subunit. *Mol. Cell. Biol.* **26,** 8755–69 (2006).

17. Tange, T. O., Damgaard, C. K., Guth, S., Valcárcel, J. & Kjems, J. The hnRNP A1 protein regulates HIV-1 tat splicing via a novel intron silencer element. *EMBO J.* **20,** 5748–58 (2001).

18. Sharma, S., Falick, A. M. & Black, D. L. Polypyrimidine tract binding protein blocks the 5' splice site-dependent assembly of U2AF and the prespliceosomal E complex. *Mol. Cell* **19,** 485–96 (2005).

19. Schaub, M. C., Lopez, S. R. & Caputi, M. Members of the heterogeneous nuclear ribonucleoprotein H family activate splicing of an HIV-1 splicing substrate by promoting formation of ATP-dependent spliceosomal complexes. *J. Biol. Chem.* **282,** 13617–26 (2007).

20. Caputi, M. & Zahler, a M. Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family. *J. Biol. Chem.* **276,** 43850–9 (2001).

21. Ule, J. *et al.* An RNA map predicting Nova-dependent splicing regulation. *Nature* **444,** 580–6 (2006).

22. Bechara, E. G., Sebestyén, E., Bernardis, I., Eyras, E. & Valcárcel, J. RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. *Mol. Cell* **52,** 720–33 (2013).

23. Misquitta-Ali, C. M. *et al.* Global profiling and molecular characterization of alternative splicing events misregulated in lung cancer. *Mol. Cell. Biol.* **31,** 138–50 (2011).

24. Maraver, A. *et al.* Therapeutic effect of γ-secretase inhibition in KrasG12V-driven non-small cell lung carcinoma by derepression of DUSP1 and inhibition of ERK. *Cancer Cell* **22,** 222–34 (2012).

25. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116,** 281–97 (2004).

26. Griffiths-Jones, S. The microRNA Registry. *Nucleic Acids Res.* **32,** D109–11 (2004).

27. Castilla-Llorente, V., Nicastro, G. & Ramos, A. Terminal loop-mediated regulation of miRNA biogenesis: selectivity and mechanisms. *Biochem. Soc. Trans.* **41,** 861–5 (2013).

28. Trabucchi, M. *et al.* The RNA-binding protein KSRP promotes the biogenesis of a subset of microRNAs. *Nature* **459,** 1010–4 (2009).

29. Suzuki, H. I. *et al.* MCPIP1 ribonuclease antagonizes dicer and terminates microRNA biogenesis through precursor microRNA degradation. *Mol. Cell* **44,** 424–36 (2011).

30. Higa, M. M., Alam, S. L., Sundquist, W. I. & Ullman, K. S. Molecular characterization of the Ran-binding zinc finger domain of Nup153. *J. Biol. Chem.* **282,** 17090–100 (2007).

31. Wang, B. *et al.* Structure and ubiquitin interactions of the conserved zinc finger domain of Npl4. *J. Biol. Chem.* **278,** 20225–34 (2003).

32. Nguyen, C. D. *et al.* Characterization of a family of RanBP2-type zinc fingers that can recognize single-stranded RNA. *J. Mol. Biol.* **407,** 273–83 (2011).

33. Wolfe, S. A., Nekludova, L. & Pabo, C. O. DNA recognition by Cys2His2 zinc finger proteins. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 183–212 (2000).

34. Aravind, L. & Koonin, E. V. G-patch: a new conserved domain in eukaryotic RNA-processing proteins and type D retroviral polyproteins. *Trends Biochem. Sci.* **24,** 342–4 (1999).

35. Callebaut, I. & Mornon, J.-P. OCRE: a novel domain made of imperfect, aromatic-rich octamer repeats. *Bioinformatics* **21,** 699–702 (2005).

36. Bonnal, S. *et al.* RBM5/Luca-15/H37 regulates Fas alternative splice site pairing after exon definition. *Mol. Cell* **32,** 81–95 (2008).

37. Inoue, A. *et al.* S1-1 nuclear domains: characterization and dynamics as a function of transcriptional activity. *Biol. Cell* **100,** 523–35 (2008).

38. Mooij, W. T. M., Mitsiki, E. & Perrakis, A. ProteinCCD: enabling the design of protein truncation constructs for expression and crystallization experiments. *Nucleic Acids Res.* **37,** W402–5 (2009).

39. Inoue, A., Takahashi, K. P., Kimura, M., Watanabe, T. & Morisawa, S. Molecular cloning of a RNA binding protein, S1-1. *Nucleic Acids Res.* **24,** 2990–7 (1996).

40. Deo, R. C., Bonanno, J. B., Sonenberg, N. & Burley, S. K. Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell* **98,** 835–45 (1999).

41. Conte, M. R. *et al.* Structure of tandem RNA recognition motifs from polypyrimidine tract binding protein reveals novel features of the RRM fold. *EMBO J.* **19,** 3132–41 (2000).

42. Oberstrass, F. C. *et al.* Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science* **309,** 2054–7 (2005).

43. Lapointe, C. P. & Wickens, M. The nucleic acid-binding domain and translational repression activity of a Xenopus terminal uridylyl transferase. *J. Biol. Chem.* **288,** 20723–33 (2013).

44. Loughlin, F. E. *et al.* The zinc fingers of the SR-like protein ZRANB2 are single-stranded RNA-binding domains that recognize 5' splice site-like sequences. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 5581–6 (2009).

45. Nam, Y., Chen, C., Gregory, R. I., Chou, J. J. & Sliz, P. Molecular basis for interaction of let-7 microRNAs with Lin28. *Cell* **147,** 1080–91 (2011).

7. Conclusions

RNA metabolism is a multi-step process with almost every stage being subject to regulation in an RNA-specific manner. This elaborate and extensive regulatory system is made up of hundreds of RNA binding proteins using a diverse range of regulatory mechanisms. With the post-transcriptional control of gene expression being essential for normal cell function it is unsurprising that defects in RBP function results in a wide array of human pathologies.

The RNA binding proteins involved in this network are often multifunctional and are therefore required to recognise a diverse range of targets within their roles. A commonality of RBPs is that they often contain multiple RBDs, which are generally small, modular units. Structural and biophysically information available on several of these proteins, shows how through this characteristic modularity they are able to achieve versatility in target recognition by using several lower affinity interactions recognising only a short RNA sequence to build up a larger, higher affinity interaction. However our information is limited to domains and proteins that recognise their targets with a relatively high affinity and well-defined specificity, and these are a minority of the total. In many cases how an RBP actually recognises its subset of target RNAs is unclear and how the many intermediate-to-low affinity domains contribute to this recognition is still not well understood. A key question in RNA biology is how the interplay between high affinity and low affinity domains defines the subsets of RNA targets of a protein.

Answering this question is not only necessary for a general understanding of gene regulation networks but because the misregulation or misfunction of RBPs has been shown to underlie a broad spectrum of human disease. This makes them attractive therapeutical targets to manipulate the expression of gene expression programs linked to, for example, cancer or neurological diseases. However, the multifunctional nature of these proteins raises questions of off-target effects. A thorough understanding of how the domains of an RBP are used differentially to recognise a specific group of targets may be used to target specific interactions and reduce the side effects caused by the unintended misregulation of other pathways.

In my thesis I have focused on a small group of proteins regulating key RNAs and pathways linked to disease. The function of these proteins in different steps of RNA metabolism has

been described, but a molecular characterisation of the principles and contributions to RNA target recognition is still missing. My aim was to provide an insight into the contribution made by the different domains to the recognition of the proteins' RNA targets that can be linked to protein function thanks to the information already available.

Both TUT4 and FMRP have been studied for a number of years and although some of the regulatory mechanisms mediated by these proteins are known the ensemble of targets is very diverse and the principle of recognition unclear. The functional characterisation of RBM10 is more recent, and while a role in alternative splicing regulation has been shown its mechanism of action is still unkown. As for the two proteins above, how RBM10 selects its different RNA targets is unclear.

Characterising protein-RNA interactions in the systems we have chosen is challenging but representative of the difficulties we face when studying RNA recognition by sequence specific (as opposite to structure specific) individual RNA binding domains. During the course of the work we have developed experimental strategies and methods that can be applied to the broader pool of RNA binding domains.

The work on TUT4 showed, for the first time, that individual CCHC ZF domains have clear sequence specificity and in particular maintain a strong preference for guanine in the middle position of the bound RNA sequence. The fingers were found to bind to RNA as independent units and upon analysis of the targets of TUT4 one can see how this could aid protein functionality. Both the pre-miRNAs and histone mRNA contain a large double stranded section and so flexible domains which only recognise an extremely short region are much more adept to hook onto small single stranded regions which may only occur in bulges within the double stranded stem. Furthermore the target RNAs will most likely be associated with other RBPs which would occlude some of the RNA. The flexibility of the domains could allow them to seek out unoccupied stretches of RNA with which to bind. We also determined that the first CCHC-type zinc finger lacked the capability to bind to RNA and so it could be involved in functionally relevant protein-protein interactions. With the plethora of proteins associated with the varied targets of TUT4 one can easily imagine this to be the case, and such interactions could serve to help specially targets the protein or to orient the catalytic domain for optimal activity. In the future the successful expression of a construct containing the multiple domains of TUT4 required for its function would enable

this system to be better characterised and answer some of the open questions about target recognition and the interplay between domains.

FMRP has been studied for more than 20 years but we are still far from a global understanding of its regulatory role and its target RNAs. I have showed that, unexpectedly, the second KH domain of the protein does not bind to ssRNA in the absence of the wild type expanded loop sequence. It is possible that, unusually for KH domains, KH2 binds only to structured targets, or that the loop is necessary to stabilise the canonical RNA binding groove, that we report be dynamic. However, the RNA target recognition takes place in the FXR1P and FXR2P paralogues, that do not have an expanded loop. Further we found that the KH1 domain, reputed to be less important, was able to bind to RNA and we defined the target sequence. Using this sequence we were able to define the role of the domain with respect to previously identified RNA targets and determined that G-rich sequences which are not optimally spaced to form G-quartets, such as those found in the Ascano dataset, may in fact be interacting with the KH1 domain. Overall our studies identified an new role in RNA target recognition for the KH1 domain.

RBM10 was the least well characterised of the proteins we investigated. While some information about its roles and targets has been published the target sequences identified by CLIP are very divergent and little is known about the RNA binding capabilities of the individual domains. We determined that the protein is able to use different sets of RNA binding domains in order to recognise different RNA targets; RRM1 and the RanBP2-type zinc finger create a high affinity binding site for the splice site sequence CUGUGGA while both RRMs and the RanBP2-type zinc finger are used in the recognition of pre-miR-106b. Relaxation analysis of the free protein and of the protein in complex with its RNA targets would provide more information on how the domains function as units to recognise targets and a high resolution structure would allow for the molecular basis of the sequence specificity of domains to be determined. We analysed the set of sequences determined by CLIP and observed differences in the contribution of the domains to binding and the discrimination between high and low affinity binding sites. We showed that the high affinity binding sequences identified contain a specific motif for the RanBP2-type zinc finger, while the lower affinity sequence tested lacked this specific site and bind both RRM1 and RRM2 relatively equally but with lower affinity. We hypothesise the binding of the zinc finger helps to orient the binding of RRM1. We also show how all three domains are used

to recognise structured pre-miRNA targets, but not short splice sites. This study exemplifies that combinatorial target recognition such as this is lost in a standard analysis of CLIP data.

The first general conclusion that can be taken from these studies is that although individual RBDs may interact with RNA with low affinity they are still able to exhibit sequence specificity that may be functionally relevant. Second, different combinations of domains can be used to recognise different RNA targets with some of the putative RNA binding domains actually engaging in protein-protein interactions – something that cannot always be extrapolated based only on the amino acid sequence of the domain. The novel findings in our studies validate the value of our bottom-up approach in deconvoluting the biology of these complex regulators. To determine sequence specificity in these systems we combined a mutational strategy of domain knock out and/or gain of function with a method to determine nucleobase preference described in this thesis and again the results show the effectiveness of this hybrid approach.

While this work adds several pieces of knowledge about the systems investigated it also raises some further questions. One such question is how to better interpret data on multi-domain proteins from *in vivo* binding techniques such as CLIP. While these data are indispensable to understanding the functioning of a protein when one dominant binding domain is not present the data can provide a wealth of varied motifs preferred by the protein, as is the case with RBM10. These data are difficult to interpret without further knowledge about the RNA binding characteristics of the individual domains determined from detailed *in vitro* work and one possibility is to use mutational analysis to relate the domain-based information on RNA binding to the CLIP assay. In general multiple sets of data will be necessary to draw functionally relevant conclusions.

RNA binding proteins and the systems they are involved in are often complex and challenging to study. However the ubiquitous nature of these proteins and their close link to many diseases make this an important area of research. Gains in this field will hopefully lead to a better understanding of both the normal functioning of cells and how disease states differ from this norm thus providing an avenue for research into potential therapeutic agents.

Appendices

Appendix I – Script for processing and analysing SIA data by Principal Component Analysis

```
#!/bin/csh -f

set spectra = (1 2 3 4 5)

set PROCESSING = y
set PCAVIEW = y

if ($PROCESSING == 'y') then

if (!(-d fid)) then
   mkdir fid
endif

if (!(-d ft)) then
   mkdir ft
endif

set n = 0

foreach id ($spectra)
   @ n++
   set inName = None

   if (-e $id/ser) then
      set inName = $id/ser
   endif

   if (-e $id/ser.Z) then
      set inName = $id/ser.Z
   endif

   if (-e $id/fid) then
      set inName = $id/fid
   endif

   if (-e $id/fid.Z) then
      set inName = $id/fid.Z
   endif

   if ($inName == None) then
      echo Error: missing input file from $id
      exit
   endif

   set fidName = (`printf fid/test%03d.fid $n`)
   set ftName  = (`printf ft/test%03d.dat  $n`)

   echo FID: $inName $fidName Spectrum: $ftName


   cat $inName | bruk2pipe -u1 $n -title $id -bad 0.0 -aswap -AMX -
decim 1792 -dspfvs 20 -grpdly 67.9841766357422  \
  -xN                  1024       -yN                   128  \
  -xT                   512       -yT                    64  \
  -xMODE                DQD       -yMODE        States-TPPI  \
  -xSW            11160.714       -ySW             1419.044  \
  -xOBS             700.133       -yOBS              70.952  \
  -xCAR               4.752       -yCAR             118.060  \
  -xLAB                  HN       -yLAB                 15N  \
  -ndim                   2       -aq2D              States  \
```

```
   -out $fidName -verb -ov



   nmrPipe -in $fidName                    \
   |   nmrPipe -fn POLY -time              \
   |   nmrPipe -fn SP -off 0.5 -end 0.98 -c 0.5\
   |   nmrPipe -fn ZF -auto                \
   |   nmrPipe -fn FT -auto                \
   |   nmrPipe -fn PS -p0 115 -p1 0 -di    \
   |   nmrPipe -fn EXT -x1 10.5ppm -xn 6.0ppm -sw    \
   |   nmrPipe -fn TP                      \
   |   nmrPipe -fn SP -off 0.4 -end 0.98 -c 1.0\
   |   nmrPipe -fn ZF -auto                \
   |   nmrPipe -fn FT -auto                \
   |   nmrPipe -fn PS -p0 -90 -p1 180.0 -di     \
   |   nmrPipe -fn TP                      \
   |   nmrPipe -fn POLY -auto              \
      -verb -ov -out $ftName
end

foreach i (ft/*)

  sethdr $i \
    -zN         $n  \
    -zT         $n  \
    -zMODE    Real  \
    -zFT      Freq  \
    -zSW        $n  \
    -zOBS      1.0  \
    -zCAR      0.0  \
    -zLAB       ID  \
    -ndim        3

end

endif



if ($PCAVIEW == 'y') then

if (!(-d pca)) then
   mkdir pca
endif

pcaNMR -in ft/test%03d.dat -noavg -nomask -autoScale \
       -nc 5 -pca pca/pca%03d.dat -load load.dat -score score.dat


pcaView.tcl -hi 2.4e+5 -spec ft/test%03d.dat -load load.dat

pipe2txt.tcl load.dat > pca_values.txt

endif
```

Appendix II - Peak lists, individual Δδ, weighted average Δδ, normalised weighted average Δδ, and final scores for RNA15 RRM, T-STAR KH, and TUT4 CCHC-ZF3.

Peak list for SIA analysis - RNA15 RRM

Peak	$\delta^{15}N$	$\delta^{1}H$
1	108.756	9.185
2	111.112	9.324
3	120.183	9.351
4	121.993	8.968
5	122.135	8.898
6	128.317	8.539
7	123.348	7.76
8	121.926	7.726
9	121.197	7.784
10	119.039	7.212
11	119.485	8.104
12	114.417	8.226
13	109.438	8.022
14	121.48	8.66

Individual Δδ and weighted average Δδ - RNA15 RRM

nAnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.282	-0.014	0.090
2	-0.095	0.046	0.055
3	-0.075	-0.013	0.027
4	-0.304	-0.02	0.098
5	0.252	0.015	0.081
6	-0.32	0.039	0.108
7	-0.124	-0.021	0.044
8	-0.119	0.042	0.056
9	0.186	0.009	0.060
10	-0.053	-0.034	0.038
11	-0.048	-0.034	0.037
12	0.156	-0.019	0.053
13	-0.133	-0.018	0.046
14	0.06	0.032	0.037

nnAnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.251	-0.011	0.080
2	-0.065	0.044	0.049
3	-0.05	-0.008	0.018
4	-0.251	-0.01	0.080
5	0.149	0.011	0.048
6	-0.266	0.037	0.092
7	-0.118	-0.018	0.041
8	-0.127	0.041	0.057
9	0.146	0.009	0.047
10	-0.024	-0.03	0.031
11	-0.032	-0.031	0.033
12	0.152	-0.015	0.050
13	-0.083	-0.014	0.030
14	0.069	0.031	0.038

nCnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.287	-0.011	0.091
2	-0.094	0.048	0.056
3	-0.052	-0.01	0.019
4	-0.349	-0.019	0.112
5	0.217	0.012	0.070
6	-0.351	0.04	0.118
7	-0.121	-0.019	0.043
8	-0.156	0.043	0.065
9	0.126	0.01	0.041
10	-0.028	-0.033	0.034
11	-0.037	-0.032	0.034
12	0.153	-0.018	0.052
13	-0.103	-0.018	0.037
14	0.052	0.03	0.034

nnCnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.29	-0.012	0.092
2	-0.058	0.051	0.054
3	-0.064	-0.01	0.023
4	-0.359	-0.018	0.115
5	0.212	0.011	0.068
6	-0.343	0.041	0.116
7	-0.152	-0.021	0.052
8	-0.133	0.041	0.059
9	0.131	0.009	0.042
10	-0.012	-0.032	0.032
11	-0.058	-0.032	0.037
12	0.138	-0.018	0.047
13	-0.084	-0.026	0.037
14	0.083	0.031	0.041

nGnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.617	-0.023	0.196
2	-0.181	0.085	0.102
3	-0.112	-0.021	0.041
4	-0.701	-0.041	0.225
5	0.43	0.02	0.137
6	-0.621	0.067	0.207
7	-0.263	-0.033	0.089
8	-0.22	0.078	0.105
9	0.307	0.016	0.098
10	-0.121	-0.062	0.073
11	-0.118	-0.069	0.078
12	0.38	-0.039	0.126
13	-0.08	-0.022	0.034
14	0.212	0.057	0.088

nnGnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.35	-0.015	0.112
2	-0.134	0.057	0.071
3	-0.075	-0.013	0.027
4	-0.432	-0.026	0.139
5	0.304	0.015	0.097
6	-0.383	0.048	0.130
7	-0.126	-0.023	0.046
8	-0.18	0.056	0.080
9	0.261	0.013	0.084
10	-0.051	-0.044	0.047
11	-0.079	-0.045	0.051
12	0.277	-0.025	0.091
13	-0.093	-0.015	0.033
14	0.083	0.045	0.052

nUnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.471	-0.008	0.149
2	-0.155	0.098	0.110
3	-0.14	-0.022	0.049
4	-0.655	-0.039	0.211
5	0.41	0.019	0.131
6	-0.667	0.075	0.224
7	-0.295	-0.04	0.102
8	-0.201	0.067	0.092
9	0.264	0.013	0.084
10	-0.088	-0.068	0.073
11	-0.102	-0.06	0.068
12	0.38	-0.027	0.123
13	-0.158	-0.038	0.063
14	0.227	0.056	0.091

nnUnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.437	-0.006	0.138
2	-0.078	0.077	0.081
3	-0.129	-0.021	0.046
4	-0.608	-0.038	0.196
5	0.324	0.018	0.104
6	-0.582	0.059	0.193
7	-0.235	-0.03	0.080
8	-0.176	0.06	0.082
9	0.223	0.012	0.072
10	-0.072	-0.053	0.058
11	-0.098	-0.047	0.056
12	0.25	-0.02	0.082
13	-0.086	-0.028	0.039
14	0.138	0.046	0.063

Individual Δδ and weighted average Δδ - RNA15 RRM (…cont)

nnnAn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.303	-0.014	0.097
2	-0.111	0.049	0.060
3	-0.042	-0.01	0.017
4	-0.356	-0.016	0.114
5	0.215	0.013	0.069
6	-0.347	0.045	0.119
7	-0.09	-0.016	0.033
8	0.184	-0.015	0.060
9	0.166	0.011	0.054
10	-0.041	-0.034	0.036
11	-0.077	-0.035	0.043
12	0.167	-0.017	0.055
13	-0.076	-0.011	0.026
14	0.094	0.033	0.044

nnnnA

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.33	-0.015	0.105
2	-0.125	0.057	0.069
3	-0.068	-0.009	0.023
4	-0.34	-0.019	0.109
5	0.258	0.014	0.083
6	-0.328	0.048	0.114
7	-0.129	-0.02	0.045
8	-0.171	0.047	0.072
9	0.198	0.011	0.064
10	-0.043	-0.037	0.039
11	-0.079	-0.035	0.043
12	0.196	-0.019	0.065
13	-0.107	-0.016	0.037
14	0.089	0.037	0.046

nnnCn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.301	-0.011	0.096
2	-0.104	0.041	0.053
3	-0.06	-0.01	0.021
4	-0.361	-0.017	0.115
5	0.192	0.011	0.062
6	-0.328	0.034	0.109
7	-0.099	-0.018	0.036
8	-0.122	0.04	0.056
9	0.134	0.01	0.044
10	-0.041	-0.032	0.035
11	-0.029	-0.03	0.031
12	0.154	-0.017	0.052
13	-0.103	-0.016	0.036
14	0.078	0.03	0.039

nnnnC

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.255	-0.008	0.081
2	-0.089	0.039	0.048
3	-0.061	-0.01	0.022
4	-0.267	-0.019	0.087
5	0.139	0.007	0.045
6	-0.301	0.031	0.100
7	-0.093	-0.015	0.033
8	-0.129	0.034	0.053
9	0.128	0.009	0.041
10	-0.036	-0.029	0.031
11	-0.025	-0.032	0.033
12	0.142	-0.015	0.047
13	-0.055	-0.015	0.023
14	0.034	0.028	0.030

nnnGn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.664	-0.016	0.211
2	-0.355	0.106	0.154
3	-0.073	-0.024	0.033
4	-0.722	-0.04	0.232
5	0.54	0.028	0.173
6	-0.669	0.081	0.227
7	-0.307	-0.041	0.105
8	-0.239	0.088	0.116
9	0.342	0.018	0.110
10	-0.119	-0.073	0.082
11	-0.067	-0.072	0.075
12	0.426	-0.041	0.141
13	-0.128	-0.032	0.052
14	0.224	0.066	0.097

nnnnG

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.401	-0.018	0.128
2	-0.128	0.068	0.079
3	-0.059	-0.014	0.023
4	-0.439	-0.027	0.141
5	0.306	0.017	0.098
6	-0.399	0.049	0.135
7	-0.169	-0.025	0.059
8	-0.2	0.056	0.084
9	0.248	0.012	0.079
10	-0.051	-0.048	0.051
11	-0.005	-0.027	0.027
12	0.279	-0.027	0.092
13	-0.098	-0.019	0.036
14	0.1	0.044	0.054

nnnUn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.41	-0.015	0.131
2	-0.131	0.069	0.080
3	-0.177	-0.023	0.061
4	-0.485	-0.04	0.159
5	0.311	0.017	0.100
6	-0.52	0.049	0.172
7	-0.251	-0.033	0.086
8	-0.166	0.055	0.076
9	0.226	0.011	0.072
10	-0.065	-0.046	0.050
11	-0.093	-0.05	0.058
12	0.205	-0.022	0.068
13	-0.149	-0.033	0.058
14	0.106	0.04	0.052

nnnnU

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.351	-0.006	0.111
2	-0.092	0.07	0.076
3	-0.142	-0.021	0.050
4	-0.5	-0.038	0.163
5	0.283	0.014	0.091
6	-0.418	0.058	0.144
7	-0.182	-0.027	0.064
8	-0.154	0.054	0.073
9	0.201	0.011	0.065
10	-0.059	-0.044	0.048
11	-0.133	-0.064	0.077
12	0.204	-0.02	0.068
13	-0.077	-0.03	0.039
14	0.114	0.041	0.055

Normalised weighted average Δδ - RNA15 RRM

Position 1

Peak	nAnnn	nCnnn	nGnnn	nUnnn
1	0.459	0.465	1.000	0.759
2	0.501	0.515	0.935	1.000
3	0.547	0.389	0.833	1.000
4	0.436	0.497	1.000	0.935
5	0.590	0.507	1.000	0.953
6	0.484	0.527	0.927	1.000
7	0.438	0.421	0.882	1.000
8	0.540	0.626	1.000	0.884
9	0.605	0.418	1.000	0.859
10	0.516	0.465	0.992	1.000
11	0.475	0.434	1.000	0.868
12	0.418	0.409	1.000	0.975
13	0.729	0.593	0.534	1.000
14	0.409	0.376	0.967	1.000
Average	**0.510**	**0.474**	**0.933**	**0.945**

Position 2

Peak	nnAnn	nnCnn	nnGnn	nnUnn
1	0.579	0.669	0.807	1.000
2	0.601	0.670	0.878	1.000
3	0.386	0.492	0.589	1.000
4	0.408	0.586	0.710	1.000
5	0.465	0.653	0.935	1.000
6	0.475	0.600	0.674	1.000
7	0.517	0.655	0.574	1.000
8	0.701	0.718	0.976	1.000
9	0.563	0.507	1.000	0.856
10	0.536	0.559	0.812	1.000
11	0.579	0.655	0.914	1.000
12	0.553	0.518	1.000	0.895
13	0.762	0.952	0.846	1.000
14	0.598	0.641	0.822	1.000
Average	**0.552**	**0.634**	**0.824**	**0.982**

Position 3

Peak	nnnAn	nnnCn	nnnGn	nnnUn
1	0.460	0.455	1.000	0.620
2	0.390	0.340	1.000	0.521
3	0.275	0.354	0.550	1.000
4	0.491	0.498	1.000	0.684
5	0.400	0.357	1.000	0.577
6	0.524	0.482	1.000	0.757
7	0.310	0.343	1.000	0.816
8	0.518	0.479	1.000	0.655
9	0.489	0.397	1.000	0.660
10	0.443	0.420	1.000	0.613
11	0.568	0.418	1.000	0.773
12	0.394	0.366	1.000	0.486
13	0.459	0.631	0.897	1.000
14	0.459	0.401	1.000	0.539
Average	**0.441**	**0.424**	**0.961**	**0.693**

Position 4

Peak	nnnnA	nnnnC	nnnnG	nnnnU
1	0.823	0.633	1.000	0.868
2	0.877	0.608	1.000	0.958
3	0.470	0.438	0.471	1.000
4	0.671	0.532	0.870	1.000
5	0.843	0.453	1.000	0.922
6	0.792	0.693	0.938	1.000
7	0.715	0.519	0.928	1.000
8	0.848	0.629	1.000	0.861
9	0.801	0.523	1.000	0.813
10	0.778	0.615	1.000	0.944
11	0.562	0.430	0.353	1.000
12	0.703	0.513	1.000	0.732
13	0.969	0.594	0.941	1.000
14	0.852	0.549	0.992	1.000
Average	**0.764**	**0.552**	**0.892**	**0.936**

Final Scores - RNA15 RRM

	Position			
	1	**2**	**3**	**4**
A	0.51	0.55	0.44	0.76
C	0.47	0.63	0.42	0.55
G	0.93	0.82	0.96	0.89
U	0.95	0.98	0.69	0.94

Peak list for SIA analysis - T-STAR KH

Peak	$\delta^{15}N$	$\delta^{1}H$
1	105.575	7.716
2	107.139	7.878
3	115.534	8.408
4	117.017	8.644
5	117.526	8.877
6	118.23	8.804
7	122.652	8.692
8	122.936	8.473
9	129.382	9.141
10	126.389	8.578
11	124.543	8.21
12	123.602	7.667
13	121.065	7.796
14	121.216	7.623
15	106.858	7.821

Individual Δδ and weighted average Δδ - T-STAR KH

nAnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.309	-0.005	0.098
2	0.059	0.035	0.040
3	0.256	0.026	0.085
4	0.193	-0.029	0.068
5	0.013	-0.018	0.018
6	0.135	0.021	0.048
7	-0.134	0.02	0.047
8	-0.13	-0.034	0.053
9	-0.053	-0.051	0.054
10	0.144	0.056	0.072
11	0.277	0.027	0.092
12	0.11	0.025	0.043
13	-0.141	0.01	0.046
14	0.024	0.02	0.021
15	0.333	-0.08	0.132

nnAnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.304	-0.028	0.100
2	0.599	-0.025	0.191
3	0.266	0.025	0.088
4	0.204	-0.034	0.073
5	0.007	-0.027	0.027
6	0.151	0.024	0.053
7	-0.133	0.023	0.048
8	-0.127	-0.028	0.049
9	-0.175	-0.069	0.088
10	0.167	0.056	0.077
11	0.334	0.036	0.112
12	0.123	0.028	0.048
13	-0.147	0.01	0.048
14	-0.001	0.023	0.023
15	-0.138	-0.029	0.052

nCnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.151	-0.039	0.062
2	0.027	0.029	0.030
3	0.207	0.014	0.067
4	0.084	-0.026	0.037
5	0.005	-0.017	0.017
6	0.111	0.014	0.038
7	-0.108	0.021	0.040
8	-0.15	-0.035	0.059
9	-0.048	-0.051	0.053
10	0.124	0.042	0.057
11	0.245	0.022	0.081
12	0.125	0.017	0.043
13	-0.123	0.007	0.040
14	-0.002	0.017	0.017
15	0.212	-0.058	0.089

nnCnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.113	-0.01	0.037
2	-0.016	0.015	0.016
3	0.139	0.012	0.046
4	0.12	-0.01	0.039
5	0.016	-0.013	0.014
6	0.025	0	0.008
7	-0.047	0.011	0.018
8	-0.127	-0.015	0.043
9	-0.056	-0.019	0.026
10	0.08	0.024	0.035
11	0.181	0.016	0.059
12	0.034	0.006	0.012
13	-0.078	0.004	0.025
14	-0.003	0.008	0.008
15	0.154	-0.037	0.061

nGnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.493	0.008	0.156
2	0.001	0.042	0.042
3	0.099	0.007	0.032
4	0.18	-0.034	0.066
5	-0.083	-0.032	0.041
6	0.19	0.012	0.061
7	-0.144	0.023	0.051
8	-0.149	-0.047	0.067
9	-0.19	-0.05	0.078
10	0.109	0.073	0.081
11	0.477	0.033	0.154
12	0.197	0.031	0.070
13	-0.165	0.012	0.054
14	0.009	0.023	0.023
15	0.277	-0.152	0.175

nnGnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.257	0.01	0.082
2	0.373	-0.013	0.119
3	0.04	0.002	0.013
4	0.147	-0.025	0.053
5	-0.054	-0.017	0.024
6	0.125	0.007	0.040
7	-0.081	0.016	0.030
8	-0.171	-0.044	0.070
9	-0.11	-0.039	0.052
10	0.084	0.046	0.053
11	0.253	0.022	0.083
12	0.118	0.018	0.041
13	-0.11	0.008	0.036
14	-0.02	0.013	0.014
15	-0.18	-0.036	0.067

nUnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.156	-0.022	0.054
2	0.047	0.025	0.029
3	0.182	0.014	0.059
4	0.061	-0.016	0.025
5	0.025	-0.017	0.019
6	0.115	0.013	0.039
7	-0.098	0.017	0.035
8	-0.14	-0.026	0.051
9	-0.099	-0.045	0.055
10	0.125	0.038	0.055
11	0.272	0.027	0.090
12	0.05	0.007	0.017
13	-0.1	0.006	0.032
14	0.005	0.015	0.015
15	0.235	-0.047	0.088

nnUnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.084	-0.01	0.028
2	0.036	0.019	0.022
3	0.13	0.013	0.043
4	0.105	-0.011	0.035
5	0.024	-0.004	0.009
6	0.044	0.007	0.016
7	-0.056	0.013	0.022
8	-0.087	-0.016	0.032
9	-0.094	-0.024	0.038
10	0.11	0.026	0.043
11	0.182	0.009	0.058
12	0.087	0.008	0.029
13	-0.066	0.005	0.021
14	-0.018	0.01	0.012
15	0.152	-0.037	0.061

Individual Δδ and weighted average Δδ - T-STAR KH (…cont)

nnnAn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.29	-0.033	0.097
2	0.079	0.045	0.051
3	0.262	0.023	0.086
4	0.194	-0.033	0.070
5	0.019	-0.02	0.021
6	0.166	0.037	0.064
7	-0.133	0.03	0.052
8	-0.17	-0.028	0.061
9	-0.211	-0.082	0.106
10	0.16	0.072	0.088
11	0.384	0.04	0.128
12	0.163	0.035	0.062
13	-0.181	0.016	0.059
14	0.002	0.033	0.033
15	0.376	-0.09	0.149

nnnnA

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.261	-0.02	0.085
2	0.008	0.032	0.032
3	0.194	0.021	0.065
4	0.162	-0.02	0.055
5	0.027	-0.017	0.019
6	0.121	0.023	0.045
7	-0.106	0.023	0.041
8	-0.149	-0.017	0.050
9	-0.167	-0.046	0.070
10	0.113	0.048	0.060
11	0.309	0.027	0.101
12	0.093	0.025	0.039
13	-0.121	0.01	0.040
14	-0.005	0.022	0.022
15	0.189	-0.058	0.083

nnnCn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.231	0.003	0.073
2	0.404	0.024	0.130
3	0.089	0.012	0.031
4	0.077	-0.005	0.025
5	-0.021	-0.006	0.009
6	0.088	0.006	0.028
7	-0.064	0.014	0.025
8	-0.073	-0.013	0.026
9	-0.039	-0.012	0.017
10	0.056	0.028	0.033
11	0.179	0.019	0.060
12	0.088	0.01	0.030
13	-0.066	0.008	0.022
14	-0.007	0.01	0.010
15	-0.216	-0.039	0.079

nnnnC

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.217	-0.001	0.069
2	0.495	0.01	0.157
3	0.152	0.017	0.051
4	0.131	-0.013	0.043
5	0.028	-0.008	0.012
6	0.097	0.01	0.032
7	-0.092	0.017	0.034
8	-0.118	-0.026	0.045
9	-0.084	-0.027	0.038
10	0.083	0.043	0.050
11	0.245	0.023	0.081
12	0.064	0.02	0.028
13	-0.086	0.009	0.029
14	-0.026	0.015	0.017
15	-0.198	-0.04	0.074

nnnGn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.28	0.011	0.089
2	0.009	0.028	0.028
3	0.059	0.011	0.022
4	0.094	-0.015	0.033
5	-0.037	-0.01	0.015
6	0.049	0.006	0.017
7	-0.085	0.019	0.033
8	-0.14	-0.044	0.062
9	-0.102	-0.03	0.044
10	0.092	0.043	0.052
11	0.249	0.023	0.082
12	0.099	0.016	0.035
13	-0.093	0.008	0.030
14	0.007	0.015	0.015
15	0.143	-0.059	0.074

nnnnG

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.268	-0.007	0.085
2	0.033	0.029	0.031
3	0.195	0.02	0.065
4	0.115	-0.019	0.041
5	0.012	-0.01	0.011
6	0.099	0.014	0.034
7	-0.047	0.021	0.026
8	-0.151	-0.037	0.060
9	-0.072	-0.032	0.039
10	0.112	0.046	0.058
11	0.202	0.023	0.068
12	0.083	0.02	0.033
13	-0.09	0.01	0.030
14	-0.004	0.017	0.017
15	0.241	-0.05	0.091

nnnUn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.208	0.001	0.066
2	0.043	0.027	0.030
3	0.088	0.013	0.031
4	0.104	-0.01	0.034
5	-0.011	-0.003	0.005
6	0.077	0.005	0.025
7	-0.043	0.012	0.018
8	-0.103	-0.012	0.035
9	-0.118	-0.014	0.040
10	0.063	0.028	0.034
11	0.182	0.014	0.059
12	0.02	0.013	0.014
13	-0.069	0.007	0.023
14	-0.003	0.009	0.009
15	0.078	-0.044	0.050

nnnnU

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
1	0.211	-0.008	0.067
2	0.021	0.036	0.037
3	0.139	0.017	0.047
4	0.186	-0.016	0.061
5	-0.017	-0.011	0.012
6	0.105	0.012	0.035
7	-0.094	0.02	0.036
8	-0.113	-0.019	0.040
9	-0.158	-0.032	0.059
10	0.098	0.043	0.053
11	0.228	0.025	0.076
12	0.096	0.024	0.039
13	-0.111	0.01	0.036
14	-0.01	0.02	0.020
15	0.16	-0.05	0.071

Normalised weighted average Δδ - T-STAR KH

Position 1

Peak	nAnnn	nCnnn	nGnnn	nUnnn
1	0.627	0.395	1.000	0.346
2	0.944	0.720	1.000	0.692
3	1.000	0.787	0.377	0.697
4	1.000	0.550	0.981	0.371
5	0.446	0.413	1.000	0.453
6	0.777	0.617	1.000	0.630
7	0.918	0.786	1.000	0.693
8	0.802	0.886	1.000	0.771
9	0.687	0.681	1.000	0.701
10	0.894	0.712	1.000	0.679
11	0.594	0.522	1.000	0.584
12	0.616	0.618	1.000	0.249
13	0.853	0.738	1.000	0.601
14	0.923	0.734	1.000	0.651
15	0.754	0.505	1.000	0.501
Average	**0.789**	**0.644**	**0.957**	**0.575**

Position 2

Peak	nnAnn	nnCnn	nnGnn	nnUnn
1	1.000	0.371	0.818	0.283
2	1.000	0.083	0.621	0.116
3	1.000	0.519	0.146	0.491
4	1.000	0.538	0.724	0.480
5	1.000	0.515	0.889	0.317
6	1.000	0.148	0.751	0.291
7	1.000	0.386	0.630	0.458
8	0.702	0.615	1.000	0.457
9	1.000	0.294	0.591	0.432
10	1.000	0.453	0.690	0.564
11	1.000	0.533	0.744	0.522
12	1.000	0.257	0.864	0.598
13	1.000	0.526	0.751	0.451
14	1.000	0.350	0.628	0.500
15	0.778	0.908	1.000	0.901
Average	**0.965**	**0.433**	**0.723**	**0.457**

Position 3

Peak	nnnAn	nnnCn	nnnGn	nnnUn
1	1.000	0.750	0.915	0.675
2	0.396	1.000	0.217	0.233
3	1.000	0.356	0.252	0.357
4	1.000	0.357	0.478	0.493
5	1.000	0.429	0.737	0.220
6	1.000	0.443	0.259	0.387
7	1.000	0.476	0.637	0.351
8	0.971	0.424	1.000	0.556
9	1.000	0.163	0.417	0.377
10	1.000	0.376	0.590	0.391
11	1.000	0.467	0.642	0.463
12	1.000	0.475	0.564	0.232
13	1.000	0.376	0.513	0.386
14	1.000	0.310	0.459	0.274
15	1.000	0.527	0.498	0.338
Average	**0.958**	**0.462**	**0.545**	**0.382**

Position 4

Peak	nnnnA	nnnnC	nnnnG	nnnnU
1	0.999	0.807	1.000	0.790
2	0.205	1.000	0.196	0.233
3	1.000	0.786	1.000	0.727
4	0.902	0.712	0.673	1.000
5	1.000	0.627	0.562	0.644
6	1.000	0.723	0.768	0.791
7	1.000	0.829	0.633	0.881
8	0.829	0.753	1.000	0.670
9	1.000	0.541	0.561	0.847
10	1.000	0.842	0.970	0.886
11	1.000	0.797	0.670	0.753
12	0.997	0.735	0.853	1.000
13	1.000	0.724	0.763	0.923
14	1.000	0.776	0.773	0.918
15	0.914	0.815	1.000	0.780
Average	**0.923**	**0.764**	**0.761**	**0.790**

Final Scores - T-STAR KH

	Position			
	1	**2**	**3**	**4**
A	0.79	0.97	0.96	0.92
C	0.64	0.43	0.46	0.76
G	0.96	0.72	0.55	0.76
U	0.57	0.46	0.38	0.79

Peak list for SIA analysis - TUT4 CCHC-ZF3

Peak	$\delta^{15}N$	$\delta^{1}H$
C1360	121.234	8.748
F1361	129.956	8.436
I1362	121.6	9.323
unassigned 1	117.121	8.219
unassigned 2	110.556	8.069
unassigned 3	125.724	8.399

Individual Δδ and weighted average Δδ - TUT4 CCHC-ZF3

nAnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1360	-0.002	-0.093	0.093
F1361	0.052	-0.091	0.092
I1362	-0.171	-0.107	0.120
unasgd. 1	-0.299	-0.102	0.139
unasgd. 2	0.569	0.059	0.189
unasgd. 3	-0.515	-0.033	0.166

nnAn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1360	0.017	-0.085	0.085
F1361	0.069	-0.080	0.083
I1362	-0.243	-0.089	0.118
unasgd. 1	-0.209	-0.090	0.112
unasgd. 2	0.542	0.050	0.179
unasgd. 3	-0.429	-0.027	0.138

nCnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1360	-0.100	-0.101	0.106
F1361	0.137	-0.098	0.107
I1362	-0.242	-0.111	0.135
unasgd. 1	-0.281	-0.118	0.148
unasgd. 2	0.595	0.060	0.197
unasgd. 3	-0.538	-0.035	0.174

nnCn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1360	0.022	-0.078	0.078
F1361	0.062	-0.081	0.083
I1362	-0.272	-0.078	0.116
unasgd. 1	-0.200	-0.087	0.108
unasgd. 2	0.455	0.044	0.150
unasgd. 3	-0.409	-0.024	0.132

nGnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1360	-0.029	-0.119	0.119
F1361	-0.240	-0.133	0.153
I1362	0.143	-0.109	0.118
unasgd. 1	-0.263	-0.133	0.157
unasgd. 2	0.682	0.074	0.228
unasgd. 3	-0.650	-0.036	0.209

nnGn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1360	0.001	-0.157	0.157
F1361	-0.408	-0.165	0.209
I1362	0.214	-0.146	0.161
unasgd. 1	-0.390	-0.174	0.213
unasgd. 2	0.859	0.098	0.289
unasgd. 3	-0.776	-0.049	0.250

nUnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1360	-0.020	-0.099	0.099
F1361	0.129	-0.093	0.102
I1362	-0.226	-0.108	0.130
unasgd. 1	-0.249	-0.105	0.131
unasgd. 2	0.581	0.057	0.192
unasgd. 3	-0.508	-0.033	0.164

nnUn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1360	0.015	-0.085	0.085
F1361	0.102	-0.081	0.087
I1362	-0.208	-0.092	0.113
unasgd. 1	-0.276	-0.098	0.131
unasgd. 2	0.518	0.051	0.172
unasgd. 3	-0.472	-0.029	0.152

Individual Δδ and weighted average Δδ - TUT4 CCHC-ZF3 (…cont)

nnnA

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1360	-0.013	-0.113	0.113
F1361	0.131	-0.109	0.117
I1362	-0.239	-0.122	0.144
unasgd. 1	-0.304	-0.133	0.164
unasgd. 2	0.707	0.069	0.234
unasgd. 3	-0.578	-0.038	0.187

nnnC

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1360	-0.024	-0.096	0.096
F1361	0.045	-0.091	0.092
I1362	-0.256	-0.101	0.129
unasgd. 1	-0.246	-0.105	0.131
unasgd. 2	0.574	0.054	0.189
unasgd. 3	-0.462	-0.032	0.150

nnnG

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1360	-0.001	-0.108	0.108
F1361	-0.224	-0.115	0.135
I1362	0.202	-0.107	0.125
unasgd. 1	-0.219	-0.125	0.143
unasgd. 2	0.638	0.069	0.213
unasgd. 3	-0.550	-0.035	0.177

nnnU

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1360	0.043	-0.086	0.087
F1361	0.080	-0.082	0.086
I1362	-0.230	-0.091	0.116
unasgd. 1	-0.174	-0.095	0.110
unasgd. 2	0.557	0.051	0.183
unasgd. 3	-0.460	-0.029	0.148

Normalised weighted average Δδ - TUT4 CCHC-ZF3

Position 1

Peak	nAnn	nCnn	nGnn	nUnn
C1360	0.783	0.880	1.000	0.846
F1361	0.779	0.887	1.000	0.831
I1362	0.887	0.942	1.000	0.837
unasgd. 1	0.830	0.866	1.000	0.844
unasgd. 2	0.796	0.832	1.000	0.786
unasgd. 3	0.784	0.908	1.000	0.861
Average	**0.810**	**0.886**	**1.000**	**0.834**

Position 2

Peak	nnAn	nnCn	nnGn	nnUn
C1360	0.561	0.554	1.000	0.540
F1361	0.542	0.499	1.000	0.542
I1362	0.524	0.504	1.000	0.615
unasgd. 1	0.618	0.521	1.000	0.594
unasgd. 2	0.553	0.526	1.000	0.608
unasgd. 3	0.515	0.518	1.000	0.542
Average	**0.552**	**0.520**	**1.000**	**0.574**

Position 3

Peak	nnnA	nnnC	nnnG	nnnU
C1360	1.000	0.902	0.941	0.812
F1361	1.000	0.852	0.955	0.770
I1362	1.000	0.796	0.871	0.669
unasgd. 1	1.000	0.809	0.911	0.784
unasgd. 2	1.000	0.801	0.950	0.795
unasgd. 3	0.936	0.739	1.000	0.689
Average	**0.989**	**0.817**	**0.938**	**0.753**

Final Scores - TUT4 CCHC-ZF3

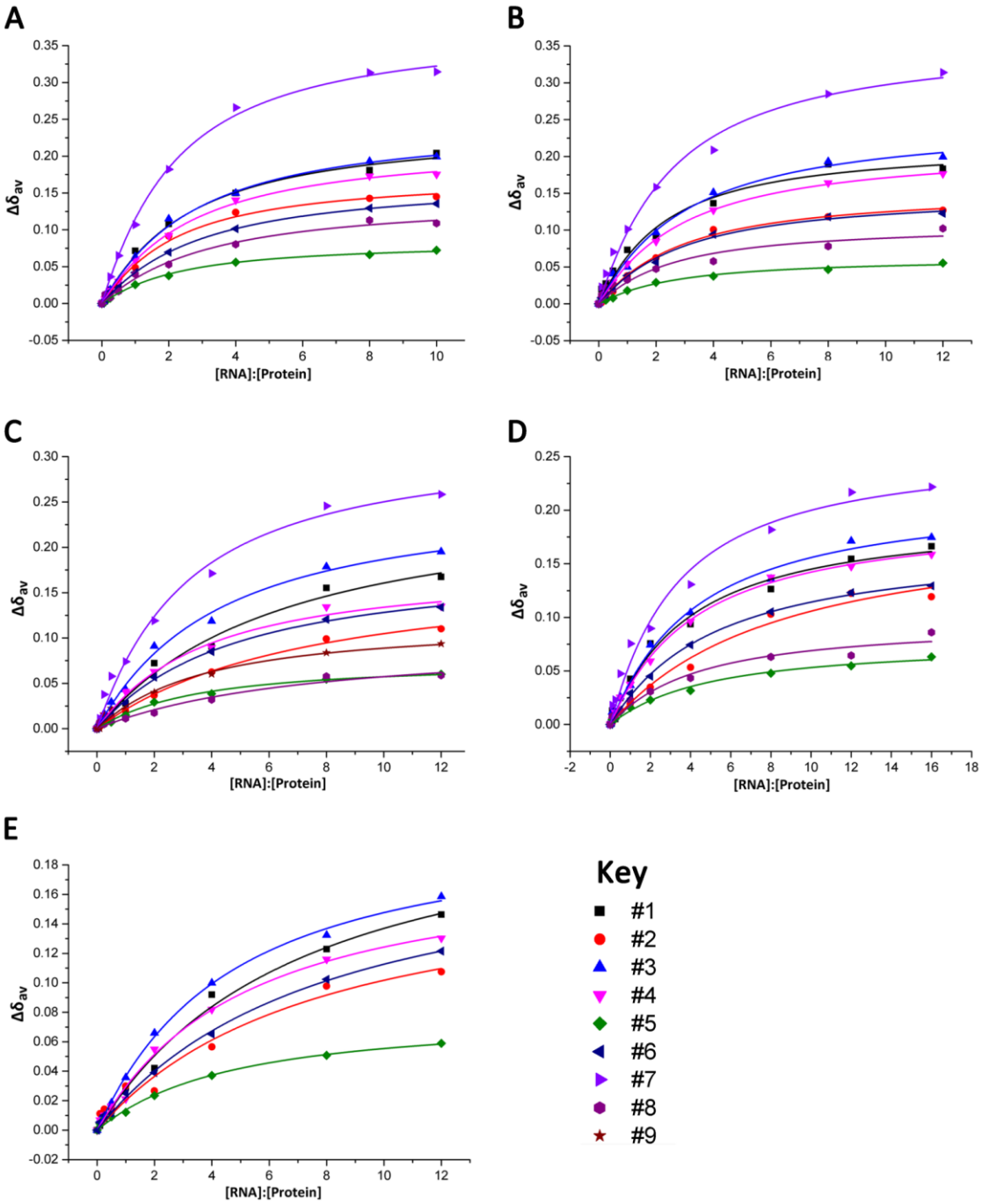
	Position		
	1	**2**	**3**
A	0.81	0.55	0.99
C	0.89	0.52	0.82
G	1.00	1.00	0.94
U	0.83	0.57	0.75

Appendix III – Peak list and binding isotherms for T-STAR KH titrations with CGAAA, CAGAA, CAUAA, CAAGA, and CAAUA

List of peaks used in the determination of the dissociation constants

Peak	$\delta^{15}N$	$\delta^{1}H$
#1	118.305	8.803
#2	122.731	8.69
#3	129.356	9.128
#4	126.417	8.579
#5	121.229	7.586
#6	117.383	7.77
#7	118.917	8.38
#8	117.48	8.849
#9	123.6	7.636

Binding isotherms for T-STAR KH titrations with A)CGAAA. B) CAGAA. C) CAUAA. D)CAAGA. E)CAAUA.



Appendix IV – Amino acid sequence of *Homo sapiens* Terminal Uridyl Transferase Isoform A. PubMed Accession number NP_001009881.

```
   1 meesktlkse nhepkknvic eeskavqvig nqtlkarndk svkeienssp nrnsskknkq
  61 ndiciektev ksckvnaanl pgpkdlglvl rdqshckakk fpnspvkaek atisqaksek
 121 atslqakaek spkspnsvka ekassyqmks ekvpsspaea ekgpslllkd mrqktelqqi
 181 gkkipssfts vdkvnieavg gekcalqnsp rsqkqqtctd ntgdsddsas giedvsddls
 241 kmkndesnke nssemdylen atvidesalt peqrlglkqa eerlerdhif rlekrspeyt
 301 ncrylcklcl ihieniqgah khikekrhkk nilekqeese lrslpppspa hlaalsvavi
 361 elakehgitd ddlrvrqeiv eemskvittf lpecslrlyg ssltrfalks sdvnidikfp
 421 pkmnhpdlli kvlgilkknv lyvdvesdfh akvpvvvcrd rksgllcrvs agndmacltt
 481 dlltalgkie pvfiplvlaf rywaklcyid sqtdggipsy cfalmvmffl qqrkppllpc
 541 llgswiegfd pkrmddfqlk giveekfvkw ecnsssatek nsiaeenkak adqpkddtkk
 601 tetdnqsnam kekhgkspla letpnrvslg qlwlellkfy tldfaleeyv icvriqdilt
 661 renknwpkrr iaiedpfsvk rnvarslnsq lvyeyvverf raayryfacp qtkggnkstv
 721 dfkkrekgki snkkpvksnn matngcillg ettekinaer eqpvqcdemd ctsqrciidn
 781 nnllvneldf adhgqdsssl stsksseiep kldkkqddla psetclkkel sqcncidlsk
 841 spdpdkstgt dcrsnletes shqsvctdts atscnckate dasdlndddn lptqelyyvf
 901 dkfiltsgkp ptivcsickk dghskndcpe dfrkidlkpl ppmtnrfrei ldlvckrcfd
 961 elsppcseqh nreqiligle kfiqkeydek arlclfgssk ngfgfrdsdl dicmtleghe
1021 naeklnckei ienlakilkr hpglrnilpi ttakvpivkf ehrrsglegd islyntlaqh
1081 ntrmlatyaa idprvqylgy tmkvfakrcd igdasrgsls syayilmvly flqqrkppvi
1141 pvlqeifdgk qipqrmvdgw nafffdktee lkkrlpslgk nteslgelwl gllrfyteef
1201 dfkeyvisir qkkllttfek qwtskciaie dpfdlnhnlg agvsrkmtnf imkafingrk
1261 lfgtpfypli greaeyffds rvltdgelap ndrccrvcgk ighymkdcpk rkssllfrlk
1321 kkdseeekeg neeekdsrdv ldprdlhdtr dfrdprdlrc ficgdaghvr recpevklar
1381 qrnssvaaaq lvrnlvnaqq vagsaqqqgd qsirtrqsse csespsyspq pqpfpqnssq
1441 saaitqpssq pgsqpklgpp qqgaqpphqv qmplynfpqs ppaqyspmhn mgllpmhplq
1501 ipapswpihg pvihsapgsa psniglndps iifaqpaarp vaipntshdg hwprtvapns
1561 lvnsgavgns epgfrgltpp ipwehaprph fplvpaswpy glhqnfmhqg narfqpnkpf
1621 ytqdrcatrr crercphppr gnvse
```

Appendix V - List of primers used in cloning.

Primers for cloning of TUT4 zinc fingers

Primer Name	Sequence (5'-3')	Restriction site
C2H2-ZF_301FW	cttgccatgggcaATTGTCGGTATCTATGCAAACTTT	NcoI
C2H2-ZF_330RV	cttgaagcttaTTTCTTATGTCGTTTCTCCTTTATATG	HindIII
CCHC-ZF1_903FW	cttgccatgggcaAGTTTATTTTAACCTCTGGCAAGC	NcoI
CCHC-ZF1_931RV	cttgaagcttaATCCTCTGGGCAATCATTCTTTG	HindIII
CCHC-ZF2_1293FW	cttgccatgggcaGATGTTGCCGTGTGTGTGGAAA	NcoI
CCHC-ZF2_1321RV	gacgtaagcttaCTTCTTTAGTCTAAACAGTAAACTG	HindIII
CCHC-ZF3_1354FW	cttgccatgggcaGACCCGAGAGACCTCAGATG	NcoI
CCHC-ZF3_1382RV	gacgtaagcttaCCTCTGACGGGCCAGCTTG	HindIII

Primers for mutagenesis of TUT4 zinc fingers

Primer Name	Sequence (5'-3')
CCHC-ZF2_R1296S_FW	caatgatagatgttgctctgtgtgtggaaaaatag
CCHC-ZF2_R1296S_RV	ctatttttccacacacagagcaacatctatcattg
CCHC-ZF2_Y1304S_FW	tggaaaaataggccactctatgaaagactgcccta
CCHC-ZF2_Y1304S_RV	tagggcagtctttcatagagtggcctatttttcca
CCHC-ZF3_F1360S_FW	agagacctcagatgttctatatgtggagatgctg
CCHC-ZF3_F1360S_RV	cagcatctccacatatagaacatctgaggtctct
CCHC-ZF3_V1368S_FW	ggagatgctggacattctcgaagggagtgcc
CCHC-ZF3_V1368S_RV	ggcactcccttcgagaatgtccagcatctcc

Primers for cloning of TUT4 multidomain constructs

Primer Name	Sequence (5'-3')	Restriction site or used in ligation independent cloning (LIC)
TUT4_186FW	cagggacccggtAGCTCCTTTACTTCTGTGGAC	LIC
TUT4_189FW	cagggacccggtACTTCTGTGGACAAAGTGAAT	LIC
TUT4_252FW	cagggacccggtTCTTCAGAGATGGACTACTTA	LIC
TUT4_270FW	cagggacccggtACACCTGAGCAGAGGCTGGGG	LIC
TUT4_301FW	cagggacccggtAATTGTCGGTATCTATGCAAA	LIC
TUT4_625FW	cagggacccggtAATCGGGTATCCTTGGGACAG	LIC
TUT4_628FW	cagggacccggtTCCTTGGGACAGTTATGGTTA	LIC
TUT4_630FW	cagggacccggtGGACAGTTATGGTTAGAGCTG	LIC
TUT4_892FW	cagggacccggtCCCACCCAGGAATTATATTAT	LIC
TUT4_894FW	cagggacccggtCAGGAATTATATTATGTGTTT	LIC
TUT4_911FW	cagggacccggtCCAACGATAGTATGCAGCATC	LIC
TUT4_947FW	cagggacccggtTTTCGGGAAATACTTGATTTA	LIC
TUT4_965FW	cagggacccggtCCTTGTTCTGAACAACACAAC	LIC
TUT4_968FW	cagggacccggtGAACAACACAACAGGGAGCAA	LIC
TUT4_971FW	cagggacccggtAACAGGGAGCAAATTTTAATT	LIC
TUT4_1319FW	cagggacccggtGACCTCAGATGTTTTATATGT	LIC
TUT4_1321FW	cagggacccggtAGATGTTTTATATGTGGAGAT	LIC
TUT4_533RV	ggcaccagagcgttaTCTCTGTTGTAGAAAAAACAT	LIC
TUT4_572RV	ggcaccagagcgttaACATTCCCACTTCACAAACTT	LIC
TUT4_574RV	ggcaccagagcgttaTGAATTACATTCCCACTTCAC	LIC
TUT4_707RV	ggcaccagagcgttaAAAATACCGATAAGCTGCCCT	LIC
TUT4_709RV	ggcaccagagcgttaACAGGCAAAATACCGATAAGC	LIC
TUT4_933RV	ggcaccagagcgttaCCTAAAATCCTCTGGGCAATC	LIC
TUT4_1230RV	ggcaccagagcgttaTTCAATTGCAATGCACTTGGA	LIC
TUT4_1257RV	ggcaccagagcgttaATTGATAAATGCTTTCATGAT	LIC
TUT4_1279RV	ggcaccagagcgttaTTGAGCCAACGTATTATATAA	LIC
TUT4_1291RV	ggcaccagagcgttaATTGGGAGCCAGTTCTCCATC	LIC
TUT4_1311RV	ggcaccagagcgttaCCTTTTAGGGCAGTCTTTCAT	LIC
TUT4_1320RV	ggcaccagagcgttaCTTTAGTCTAAACAGTAAACT	LIC
TUT4_1400RV	ggcaccagagcgttaCTGTTGAGCATTTACAAGGTT	LIC
TUT4_1407RV	ggcaccagagcgttaTTGCTGAGCTGAACCAGCCAC	LIC

TUT4_1409RV	ggcaccagagcgttaACCCTGTTGCTGAGCTGAACC	LIC
TUT4_1417RV	ggcaccagagcgttaCTGTCTAGTCCTTATGGACTG	LIC
TUT4_1420RV	ggcaccagagcgttaTTCTGATGACTGTCTAGTCCT	LIC
TUT4_1645RV	ggcaccagagcgttaCTCCGACACGTTTCCTCTTGG	LIC
TUT4_894FW_SUMO	tacttcccaatccatgCAGGAATTATATTATGTGTTT	LIC (pNIC-vector)
TUT4_911FW_SUMO	tacttcccaatccatgCCAACGATAGTATGCAGCAT	LIC (pNIC-vector)
TUT4_947FW_SUMO	tacttcccaatccatgTTTCGGGAAATACTTGATTTA	LIC (pNIC-vector)
TUT4_971FW_SUMO	tacttcccaatccatgAACAGGGAGCAAATTTTAATT	LIC (pNIC-vector)
TUT4_1230RV_SUMO	tatccacctttactgttaTTCAATTGCAATGCACTTGGA	LIC (pNIC-vector)
TUT4_1279RV_SUMO	tatccacctttactgttaATCAAAGAAGTACTCAGCTTC	LIC (pNIC-vector)
TUT4_1291RV_SUMO	tatccacctttactgttaATTGGGAGCCAGTTCTCCATC	LIC (pNIC-vector)
TUT4_1311RV_SUMO	tatccacctttactgttaCCTTTTAGGGCAGTCTTTCAT	LIC (pNIC-vector)
TUT4_1417RV_SUMO	tatccacctttactgttaCTGTCTAGTCCTTATGGACTG	LIC (pNIC-vector)
TUT4_894FW_NcoI	cttgccatgggcCAGGAATTATATTATGTGTTTGATAA	NcoI
TUT4_947FW_NcoI	cttgccatgggcTTTCGGGAAATACTTGATTTAGTAT	NcoI
TUT4_971FW_NcoI	cttgccatgggcAACAGGGAGCAAATTTTAATTGG	NcoI
TUT4_1230RV_XhoI	gaacctcgagttaTTCAATTGCAATGCACTTGGAAG	XhoI
TUT4_1279RV_XhoI	gaacctcgagttaATCAAAGAAGTACTCAGCTTCTC	XhoI
TUT4_1291RV_XhoI	gaacctcgagttaATTGGGAGCCAGTTCTCCATC	XhoI
TUT4_1311RV_XhoI	gaacctcgagttaACCTTTTAGGGCAGTCTTTCATG	XhoI

Primers for ligation independent cloning of FMRP

Primer Name	**Sequence (5'-3')**
FMRP_212FW	cagggacccggtAGTCGTCAACTTGCTTCGCGT
FMRP_216FW	cagggacccggtGCTTCGCGTTTCCACGAGCAG
FMRP_359RV	ggcaccagagcgttaTTTCAGGTAGTTCAGATGATA
FMRP_383RV	ggcaccagagcgttaGCGAGAACTTGCGCCGATTTG
FMRP_405RV	ggcaccagagcgttaGCGACCCATACCTTGGCCGTC

Primers for mutagenesis of FMRP KH domains

Primer Name	**Sequence (5'-3')**
FMRP_KH1_gddgFW	gcttggctatcggcgatgatggcgcgaatattcag
FMRP_KH1_gddgRV	ctgaatattcgcgccatcatcgccgatagccaagc
FMRP_KH2_gddgFW	ggtaaagtcattggggatgatggtaaacttatccagg
FMRP_KH2_gddgRV	cctggataagtttaccatcatccccaatgactttacc
FMRP_KH1_gkkgFW	gcttggctatcggcaagaagggcgcgaatattca
FMRP_KH1_gkkgRV	tgaatattcgcgcccttcttgccgatagccaagc

Primers for ligation independent cloning of RBM10

Primer Name	**Sequence (5'-3')**
RBM10_106FW	cagggacccggtGACTATCGGACCGAGCAAGGGG
RBM10_114FW	cagggacccggtGAGGAGGAGGAGGAGGAGGATG
RBM10_122FW	cagggacccggtGAGGAGGAGGAGAAGGCCAGTA
RBM10_128FW	cagggacccggtAGTAACATCGTCATGCTGAGGA
RBM10_299FW	cagggacccggtAATGACACCATCATTTTGCGC
RBM10_300FW	cagggacccggtGACACCATCATTTTGCGCAAC
RBM10_748FW	cagggacccggtGAGCGGGAGGAGAAGCTCACCG
RBM10_749FW	cagggacccggtCGGGAGGAGAAGCTCACCGACT
RBM10_207RV	ggcaccagagcgttaACTGTAGTGCATCGACACCTT
RBM10_221RV	ggcaccagagcgttaCTTATTGCACAGCCAGTCCTC
RBM10_235RV	ggcaccagagcgttaTTTGAAGCACTTCTCTCGGCG
RBM10_248RV	ggcaccagagcgttaGGGCAGCTTCTGCTCTGCCTC
RBM10_408RV	ggcaccagagcgttaAGCAATGGCAGTGCTGGCCACA
RBM10_415RV	ggcaccagagcgttaTGAGATGGCCCACTGGGCCGCA
RBM10_418RV	ggcaccagagcgttaGGAGGCCTGTGAGATGGCCCAC
RBM10_421RV	ggcaccagagcgttaCCCACCTTGGGAGGCCTGTGAG
RBM10_793RV	ggcaccagagcgttaTCGCCGGTGAATCTCAAGGTTT
RBM10_795RV	ggcaccagagcgttaGTGGGCTCGCCGGTGAATCTCA
RBM10_796RV	ggcaccagagcgttaCAAGTGGGCTCGCCGGTGAATC

Appendix VI - List of TUT4 multidomain constructs and expression conditions.

Some of the constructs are lacking a 43 amino acid segment in the nucleotidyl transferase domain and these are highlighted with an X in the column titled missing segment.

					His (pET-47b)						SUMO (pET-52-SUMO)		GB1-Strep	MBP (pET-M41)	NusA (pET-M60)	Strep (pET-52)
Start residue	**End residue**	**Start Domain**	**End Domain**	**Segment missing**	**25°C**	**18°C**	**28°C**	**Expression in Enbase media**	**Coexpression with chaperones**	**Expression in Rosetta2 (DE3) cells**	**18°C**	**25°C**	**25°C**	**18°C**	**18°C**	**25°C**
252	533	C2H2-ZF	C2H2-ZF	no	x											
252	572	C2H2-ZF	C2H2-ZF	no	x											
252	574	C2H2-ZF	C2H2-ZF	no	x											
270	533	C2H2-ZF	C2H2-ZF	no	x											
270	572	C2H2-ZF	C2H2-ZF	no	x											
270	574	C2H2-ZF	C2H2-ZF	no	x											
301	533	C2H2-ZF	C2H2-ZF	no	x	x	x	x								
301	572	C2H2-ZF	C2H2-ZF	no	x	x	x	x								
301	574	C2H2-ZF	C2H2-ZF	no	x	x	x	x	x	x						
252	707	C2H2-ZF	PAP1	no	x											
252	709	C2H2-ZF	PAP1	no	x											
270	707	C2H2-ZF	PAP1	no	x											
270	709	C2H2-ZF	PAP1	no	x											
301	707	C2H2-ZF	PAP1	no	x	x	x	x	x	x						
301	709	C2H2-ZF	PAP1	no	x	x	x	x								
252	933	C2H2-ZF	CCHC-ZF1	no	x											
270	933	C2H2-ZF	CCHC-ZF1	no	x											
301	933	C2H2-ZF	CCHC-ZF1	no	x											
186	1311	C2H2-ZF	CCHC-ZF2	no	x								x			
186	1320	C2H2-ZF	CCHC-ZF2	no	x								x			
189	1311	C2H2-ZF	CCHC-ZF2	no	x								x			
189	1320	C2H2-ZF	CCHC-ZF2	no	x								x			
252	1311	C2H2-ZF	CCHC-ZF2	no	x											
252	1320	C2H2-ZF	CCHC-ZF2	no	x											
270	1311	C2H2-ZF	CCHC-ZF2	no	x											
270	1320	C2H2-ZF	CCHC-ZF2	no	x											
301	1311	C2H2-ZF	CCHC-ZF2	no	x											
301	1320	C2H2-ZF	CCHC-ZF2	no	x											
186	1420	C2H2-ZF	CCHC-ZF3	no	x								x			
189	1420	C2H2-ZF	CCHC-ZF3	no	x				x	x			x			
252	1400	C2H2-ZF	CCHC-ZF3	yes	x											
252	1400	C2H2-ZF	CCHC-ZF3	no	x											
252	1407	C2H2-ZF	CCHC-ZF3	no	x											
252	1409	C2H2-ZF	CCHC-ZF3	no	x											
252	1417	C2H2-ZF	CCHC-ZF3	no	x											
270	1400	C2H2-ZF	CCHC-ZF3	yes	x											
270	1400	C2H2-ZF	CCHC-ZF3	no	x											
270	1407	C2H2-ZF	CCHC-ZF3	no	x											
270	1409	C2H2-ZF	CCHC-ZF3	yes	x											
270	1409	C2H2-ZF	CCHC-ZF3	no	x											
270	1417	C2H2-ZF	CCHC-ZF3	no	x											
301	1400	C2H2-ZF	CCHC-ZF3	yes	x											
301	1400	C2H2-ZF	CCHC-ZF3	no	x											
301	1407	C2H2-ZF	CCHC-ZF3	no	x											
301	1409	C2H2-ZF	CCHC-ZF3	yes	x											
301	1409	C2H2-ZF	CCHC-ZF3	no	x											

					His (pET-47b)						SUMO (pET-52-SUMO)		GB1-Strep	MBP (pET-M41)	NusA (pET-M60)	Strep (pET-52)
Start residue	**End residue**	**Start Domain**	**End Domain**	**Segment missing**	**25°C**	**18°C**	**28°C**	**Expression in Enbase media**	**Coexpression with chaperones**	**Expression in Rosetta2 (DE3) cells**	**18°C**	**25°C**	**25°C**	**18°C**	**18°C**	**25°C**
301	1417	C2H2-ZF	CCHC-ZF3	no	x											
625	1230	PAP1	PAP2	yes	x											
625	1257	PAP1	PAP2	yes	x											
628	1230	PAP1	PAP2	yes	x											
628	1257	PAP1	PAP2	yes	x											
630	1257	PAP1	PAP2	yes	x											
625	1400	PAP1	CCHC-ZF3	yes	x											
892	1230	CCHC-ZF1	PAP2	yes	x											
892	1230	CCHC-ZF1	PAP2	no	x				x	x						
892	1257	CCHC-ZF1	PAP2	no	x											
894	1230	CCHC-ZF1	PAP2	yes	x											
894	1230	CCHC-ZF1	PAP2	no	x											
894	1257	CCHC-ZF1	PAP2	no	x											
894	1219	CCHC-ZF1	PAP2	no	x								x			
911	1230	CCHC-ZF1	PAP2	yes	x											
911	1230	CCHC-ZF1	PAP2	no	x											
911	1257	CCHC-ZF1	PAP2	yes	x											
911	1257	CCHC-ZF1	PAP2	no	x											
911	1291	CCHC-ZF1	PAP2	no	x								x			
892	1311	CCHC-ZF1	CCHC-ZF2	yes	x											x
892	1311	CCHC-ZF1	CCHC-ZF2	no	x											
892	1320	CCHC-ZF1	CCHC-ZF2	yes	x	x	x	x								x
892	1320	CCHC-ZF1	CCHC-ZF2	no	x											
894	1311	CCHC-ZF1	CCHC-ZF2	yes	x	x	x	x								x
894	1311	CCHC-ZF1	CCHC-ZF2	no	x						x			x	x	
894	1311	CCHC-ZF1	CCHC-ZF2	no	x											
894	1320	CCHC-ZF1	CCHC-ZF2	yes	x	x	x	x								
894	1320	CCHC-ZF1	CCHC-ZF2	no	x											
911	1311	CCHC-ZF1	CCHC-ZF2	yes	x											
911	1311	CCHC-ZF1	CCHC-ZF2	no	x											
911	1320	CCHC-ZF1	CCHC-ZF2	yes	x											
911	1320	CCHC-ZF1	CCHC-ZF2	no	x											
892	1400	CCHC-ZF1	CCHC-ZF3	yes	x	x	x	x								x
892	1400	CCHC-ZF1	CCHC-ZF3	no	x											
892	1407	CCHC-ZF1	CCHC-ZF3	yes	x											x
892	1407	CCHC-ZF1	CCHC-ZF3	no	x											
892	1409	CCHC-ZF1	CCHC-ZF3	no	x											
892	1417	CCHC-ZF1	CCHC-ZF3	yes	x	x	x	x								x
892	1417	CCHC-ZF1	CCHC-ZF3	no	x											
892	1420	CCHC-ZF1	CCHC-ZF3	no	x								x			
894	1400	CCHC-ZF1	CCHC-ZF3	yes	x	x	x	x								x
894	1400	CCHC-ZF1	CCHC-ZF3	no	x				x	x						
894	1407	CCHC-ZF1	CCHC-ZF3	yes	x											x
894	1407	CCHC-ZF1	CCHC-ZF3	no	x											
894	1409	CCHC-ZF1	CCHC-ZF3	yes	x	x	x	x								x
894	1409	CCHC-ZF1	CCHC-ZF3	no	x				x	x						
894	1417	CCHC-ZF1	CCHC-ZF3	yes	x	x	x	x								x

					His (pET-47b)						SUMO (pET-52-SUMO)		GB1-Strep	MBP (pET-M41)	NusA (pET-M60)	Strep (pET-52)
Start residue	**End residue**	**Start Domain**	**End Domain**	**Segment missing**	**25°C**	**18°C**	**28°C**	**Expression in Enbase media**	**Coexpression with chaperones**	**Expression in Rosetta2 (DE3) cells**	**18°C**	**25°C**	**25°C**	**18°C**	**18°C**	**25°C**
894	1417	CCHC-ZF1	CCHC-ZF3	no	x											
894	1420	CCHC-ZF1	CCHC-ZF3	no	x								x			
911	1400	CCHC-ZF1	CCHC-ZF3	yes	x	x	x	x								x
911	1400	CCHC-ZF1	CCHC-ZF3	no	x											
911	1407	CCHC-ZF1	CCHC-ZF3	yes	x											x
911	1407	CCHC-ZF1	CCHC-ZF3	no	x											
911	1409	CCHC-ZF1	CCHC-ZF3	yes	x											x
911	1409	CCHC-ZF1	CCHC-ZF3	no	x											
911	1417	CCHC-ZF1	CCHC-ZF3	yes												x
911	1417	CCHC-ZF1	CCHC-ZF3	no	x						x					
911	1417	CCHC-ZF1	CCHC-ZF3	no	x											
947	1192	NTD	PAP2	no	x								x			
947	1219	NTD	PAP2	no	x								x	x		
947	1279	NTD	PAP2	no							x	x			x	
947	1291	NTD	PAP2	no	x								x		x	
947	1291	NTD	PAP2	no							x	x				
965	1230	NTD	PAP2	no	x											
965	1230	NTD	PAP2	no					x	x						
965	1257	NTD	PAP2	no	x											
968	1230	NTD	PAP2	no	x											
968	1257	NTD	PAP2	no	x											
971	1230	NTD	PAP2	no	x											
971	1230	NTD	PAP2	no							x	x		x	x	
971	1257	NTD	PAP2	no	x											
965	1311	NTD	CCHC-ZF2	yes	x											x
965	1320	NTD	CCHC-ZF2	yes	x											
968	1311	NTD	CCHC-ZF2	yes	x											
971	1311	NTD	CCHC-ZF2	yes	x	x	x	x								x
965	1400	NTD	CCHC-ZF3	yes	x											
965	1407	NTD	CCHC-ZF3	yes	x											
965	1409	NTD	CCHC-ZF3	yes	x	x	x	x								
965	1417	NTD	CCHC-ZF3	yes	x	x	x	x								x
968	1400	NTD	CCHC-ZF3	yes	x	x	x	x								x
968	1407	NTD	CCHC-ZF3	yes	x	x	x	x								x
968	1409	NTD	CCHC-ZF3	yes	x	x	x	x								x
971	1400	NTD	CCHC-ZF3	yes	x	x	x	x								x
971	1407	NTD	CCHC-ZF3	yes	x	x	x	x								x
971	1409	NTD	CCHC-ZF3	yes	x	x	x	x								x
971	1417	NTD	CCHC-ZF3	yes	x	x	x	x								x
1319	1645	CCHC-ZF3	C-term	yes	x											
1321	1645	CCHC-ZF3	C-term	yes	x											

Appendix VII - Assigned chemical shifts of TUT4 CCHC-ZF2 and TUT4 CCHC-ZF3.

Peak list of assigned chemical shifts in TUT4 CCHC-ZF2 (pH 4.5)

Residue	N	HN	Cα	Cβ
C1295	128.46	8.36	60.18	30.12
R1296	129.33	8.46	58.11	30.50
V1297	123.15	9.55	65.65	31.52
C1298	117.49	8.52	58.50	32.27
G1299	113.55	8.00	46.57	---
I1301	118.36	8.09	60.77	39.22
G1302	111.02	8.48	45.38	---
H1303	112.99	7.03	55.09	30.54
Y1304	117.49	8.82	56.62	40.36
M1305	121.99	9.09	60.04	33.12
K1306	117.80	8.25	58.78	31.68
D1307	117.49	7.84	53.28	42.57
C1308	125.37	7.62	58.86	30.76
K1310	120.59	8.91	57.10	32.14

Peak list of assigned chemical shifts in TUT4 CCHC-ZF3 (pH 7.4)

Residue	N	HN	Cα	Cβ
D1354	122.45	8.26	51.99	41.64
C1360	121.16	8.72	60.05	?
F1361	129.85	8.41	59.71	40.29
I1362	121.50	9.29	63.36	?
C1363	117.87	8.45	58.85	32.41
G1364	113.23	7.90	46.19	---
D1365	124.18	8.45	54.81	43.70
A1366	123.82	8.41	52.14	19.60
G1367	106.04	8.79	45.89	---
H1368	113.50	7.12	?	30.33
A1379	123.55	7.95	52.86	19.01
Q1381	121.49	8.26	55.95	?
R1382	127.58	7.93	57.45	31.62

Appendix VIII - Peak lists, individual Δδ, weighted average Δδ, normalised weighted average Δδ, and final scores for TUT4 CCHC-ZF2.

Peak list for SIA analysis - TUT4 CCHC-ZF2

Peak	$\delta^{15}N$	$\delta^{1}H$
C1295	127.75	8.36
G1299	113.415	8.001
I1301	118.693	8.145
G1302	111.028	8.462
M1305	121.988	9.134
K1306	118.187	8.29
unassigned	121.364	8.264

Individual Δδ and weighted average Δδ - TUT4 CCHC-ZF2

nAnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1295	-0.362	-0.06	0.129246
G1299	0.095	0.027	0.040392
I1301	-0.074	-0.014	0.027269
G1302	-0.156	0	0.049332
M1305	0.252	-0.024	0.083225
K1306	-0.087	-0.001	0.02753
unasgd.	0.108	0.005	0.034517

nnAn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1295	-0.101	-0.062	0.06974
G1299	0.02	0.027	0.02773
I1301	-0.044	-0.013	0.01904
G1302	-0.084	-0.009	0.02805
M1305	0.194	-0.031	0.06874
K1306	-0.083	0.001	0.02627
unasgd.	0.011	0.012	0.01249

nCnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1295	-0.219	-0.063	0.093622
G1299	0.057	0.03	0.034999
I1301	-0.089	-0.014	0.031434
G1302	-0.182	-0.002	0.057588
M1305	0.279	-0.03	0.093189
K1306	-0.102	-0.002	0.032317
unasgd.	0.064	0.005	0.020847

nnCn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1295	-0.144	-0.056	0.07218
G1299	0.071	0.025	0.0336
I1301	-0.044	-0.012	0.01837
G1302	-0.137	-0.002	0.04337
M1305	0.368	-0.017	0.11761
K1306	-0.087	-0.001	0.02753
unasgd.	0.029	0.003	0.00965

nGnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1295	-0.206	-0.08	0.103168
G1299	0.085	0.054	0.06032
I1301	-0.114	-0.01	0.037411
G1302	-0.232	-0.003	0.073426
M1305	0.555	-0.019	0.176532
K1306	-0.121	-0.028	0.047414
unasgd.	-0.042	0.028	0.03099

nnGn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1295	-0.334	-0.091	0.13942
G1299	0.061	0.059	0.06207
I1301	-0.159	-0.012	0.05169
G1302	-0.179	-0.002	0.05664
M1305	0.501	-0.017	0.15934
K1306	-0.062	-0.021	0.02873
unasgd.	0.082	0.021	0.03337

nUnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1295	-0.111	-0.063	0.072119
G1299	0.08	0.03	0.039243
I1301	-0.077	-0.018	0.03028
G1302	-0.17	-0.004	0.053907
M1305	0.342	-0.025	0.111002
K1306	-0.078	-0.001	0.024686
unasgd.	0.079	0.004	0.0253

nnUn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1295	-0.07	-0.065	0.06867
G1299	0.023	0.033	0.03379
I1301	-0.166	-0.018	0.05549
G1302	-0.004	-0.003	0.00326
M1305	0.284	-0.035	0.09639
K1306	-0.151	-0.003	0.04784
unasgd.	-0.041	0.002	0.01312

Individual Δδ and weighted average Δδ - TUT4 CCHC-ZF2 (…cont)

nnnA

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1295	-0.329	-0.079	0.130633
G1299	0.172	0.041	0.068113
I1301	-0.08	-0.02	0.032249
G1302	-0.127	-0.004	0.04036
M1305	0.29	-0.033	0.097463
K1306	-0.154	-0.018	0.051919
unasgd.	0.002	0.015	0.015013

nnnC

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1295	-0.176	-0.065	0.085572
G1299	0.091	0.035	0.045311
I1301	-0.082	-0.019	0.032147
G1302	-0.191	0	0.0604
M1305	0.21	-0.036	0.075538
K1306	-0.093	-0.005	0.029831
unasgd.	-0.089	0.01	0.029868

nnnG

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1295	-0.443	-0.075	0.158902
G1299	0.119	0.036	0.052078
I1301	-0.144	-0.019	0.049342
G1302	-0.125	-0.007	0.040143
M1305	0.428	0.011	0.135792
K1306	-0.094	-0.01	0.031362
unasgd.	0.046	0.014	0.020189

nnnU

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
C1295	-0.153	-0.064	0.08023
G1299	0.093	0.031	0.042731
I1301	-0.083	-0.018	0.031826
G1302	-0.121	-0.002	0.038316
M1305	0.104	-0.044	0.054933
K1306	-0.078	0	0.024666
unasgd.	-0.01	0.007	0.007681

Normalised weighted average Δδ - TUT4 CCHC-ZF2

Position 1

Peak	nAnn	nCnn	nGnn	nUnn
C1295	1.000	0.724	0.798	0.558
G1299	0.670	0.580	1.000	0.651
I1301	0.729	0.840	1.000	0.809
G1302	0.672	0.784	1.000	0.734
M1305	0.471	0.528	1.000	0.629
K1306	0.581	0.682	1.000	0.521
unasgd.	1.000	0.604	0.898	0.733
Average	**0.732**	**0.678**	**0.957**	**0.662**

Position 2

Peak	nnAn	nnCn	nnGn	nnUn
C1295	0.500	0.518	1.000	0.493
G1299	0.447	0.541	1.000	0.544
I1301	0.343	0.331	0.931	1.000
G1302	0.495	0.766	1.000	0.057
M1305	0.431	0.738	1.000	0.605
K1306	0.549	0.575	0.600	1.000
unasgd.	0.374	0.289	1.000	0.393
Average	**0.449**	**0.537**	**0.933**	**0.585**

Position 3

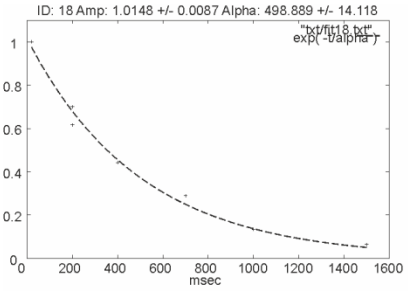
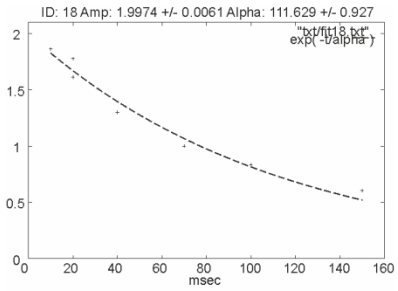
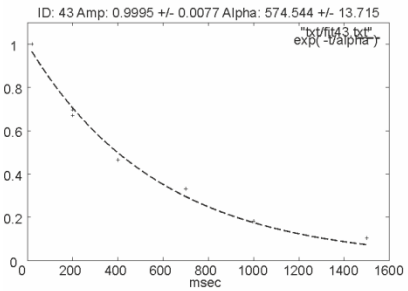
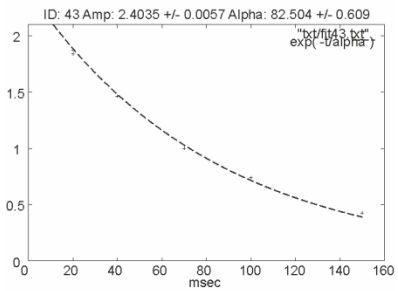
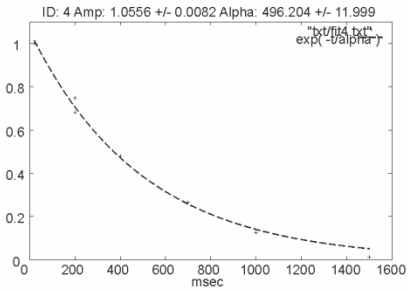
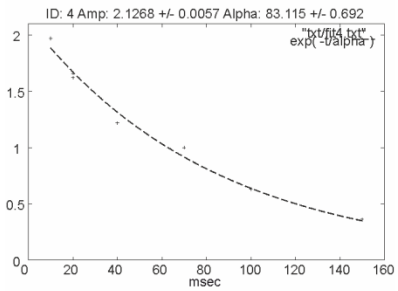
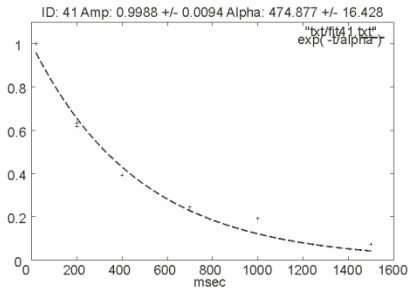
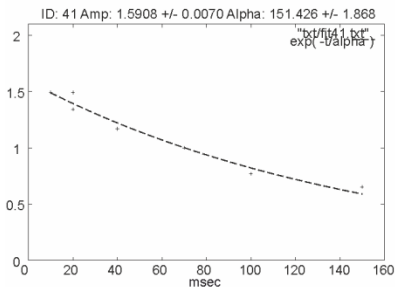
C1295	0.822	0.539	1.000	0.505
G1299	1.000	0.665	0.765	0.627
I1301	0.654	0.652	1.000	0.645
G1302	0.668	1.000	0.665	0.634
M1305	0.718	0.556	1.000	0.405
K1306	1.000	0.575	0.604	0.475
unasgd.	0.503	1.000	0.676	0.257
Average	**0.766**	**0.712**	**0.816**	**0.507**

Final Scores - TUT4 CCHC-ZF2

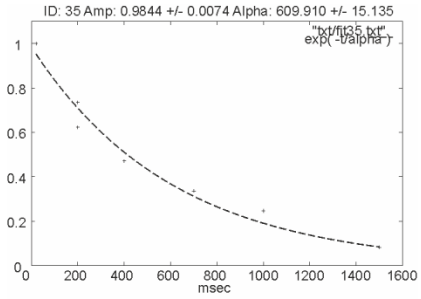
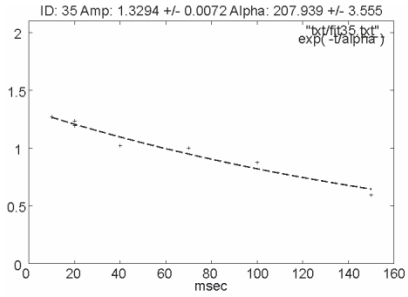
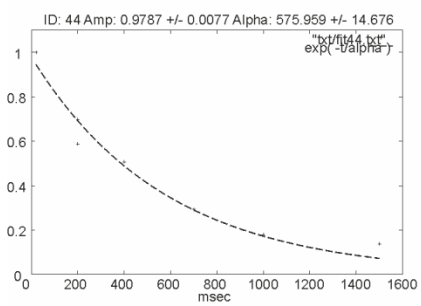
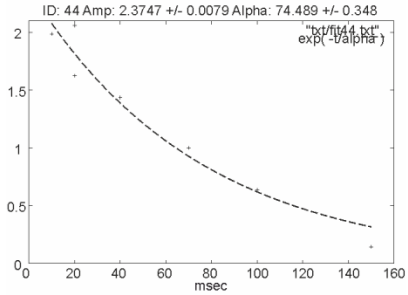
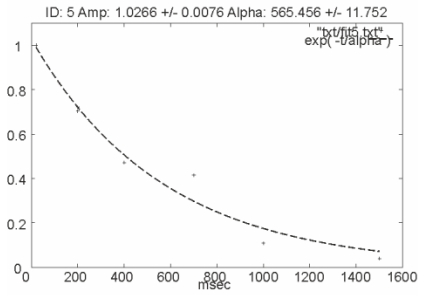
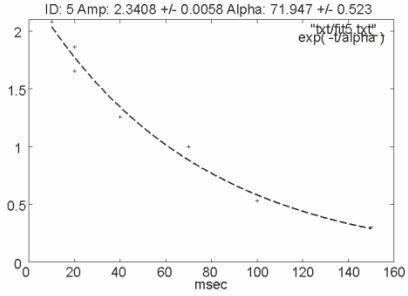
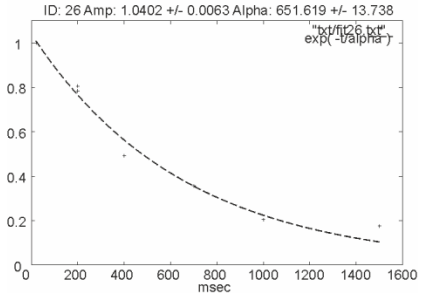
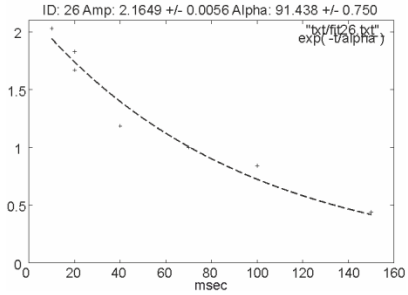
	Position		
	1	**2**	**3**
A	0.73	0.45	0.77
C	0.68	0.54	0.71
G	0.96	0.93	0.82
U	0.66	0.58	0.51

Appendix IX – Representative T1 and T2 plots for CCHC-ZF2 and CCHC-ZF3.

Chemical shifts from CCHC-ZF2

	T1 plots	T2 plots
C1295		
C1298		
G1299		
Y1304		

Chemical shifts from CCHC-ZF3

	T1 plots	T2 plots
C1360		
C1363		
G1364		
A1366		

Appendix X - CLUSTAL 2.0.12 multiple amino acid sequence alignment of full length Fragile X mental retardation protein (UniProtKB/Swiss-Prot: Q06787.1) and FMRP Δ331-396.

```
FMRP             1    MEELVVEVRGSNGAFYKAFVKDVHEDSITVAFENNWQPDRQIPFHDVRFPPPVGYNKDIN
FMRP_Δ331-396    1    MEELVVEVRGSNGAFYKAFVKDVHEDSITVAFENNWQPDRQIPFHDVRFPPPVGYNKDIN

FMRP            61    ESDEVEVYSRANEKEPCCWWLAKVRMIKGEFYVIEYAACDATYNEIVTIERLRSVNPNKP
FMRP_Δ331-396   61    ESDEVEVYSRANEKEPCCWWLAKVRMIKGEFYVIEYAACDATYNEIVTIERLRSVNPNKP

FMRP            121   ATKDTFHKIKLDVPEDLRQMCAKEAAHKDFKKAVGAFSVTYDPENYQLVILSINEVTSKR
FMRP_Δ331-396   121   ATKDTFHKIKLDVPEDLRQMCAKEAAHKDFKKAVGAFSVTYDPENYQLVILSINEVTSKR

FMRP            181   AHMLIDMHFRSLRTKLSLIMRNEEASKQLESSRQLASRFHEQFIVREDLMGLAIGTHGAN
FMRP_Δ331-396   181   AHMLIDMHFRSLRTKLSLIMRNEEASKQLESSRQLASRFHEQFIVREDLMGLAIGTHGAN

FMRP            241   IQQARKVPGVTAIDLDEDTCTFHIYGEDQDAVKKARSFLEFAEDVIQVPRNLVGKVIGKN
FMRP_Δ331-396   241   IQQARKVPGVTAIDLDEDTCTFHIYGEDQDAVKKARSFLEFAEDVIQVPRNLVGKVIGKN

FMRP            301   GKLIQEIVDKSGVVRVRIEAENEKNVPQEEEIMPPNSLPSNNSRVGPNAPEEKKHLDIKE
FMRP_Δ331-396   301   GKLIQEIVDKSGVVRVRIEAENEKNVPQEE----------------------------

FMRP            361   NSTHFSQPNSTKVQRVLVASSVVAGESQKPELKAWQGMVPFVFVGTKDSIANATVLLDYH
FMRP_Δ331-396   331   -----------------------------------GMVPFVFVGTKDSIANATVLLDYH

FMRP            421   LNYLKEVDQLRLERLQIDEQLRQIGASSRPPPNRTDKEKSYVTDDGQGMGRGSRPYRNRG
FMRP_Δ331-396   355   LNYLKEVDQLRLERLQIDEQLRQIGASSRPPPNRTDKEKSYVTDDGQGMGRGSRPYRNRG

FMRP            481   HGRRGPGYTSGTNSEASNASETESDHRDELSDWSLAPTEEERESFLRRGDGRRRGGGGRG
FMRP_Δ331-396   415   HGRRGPGYTSGTNSEASNASETESDHRDELSDWSLAPTEEERESFLRRGDGRRRGGGGRG

FMRP            541   QGGRGRGGGFKGNDDHSRTDNRPRNPREAKGRTTDGSLQIRVDCNNERSVHTKTLQNTS
FMRP_Δ331-396   475   QGGRGRGGGFKGNDDHSRTDNRPRNPREAKGRTTDGSLQIRVDCNNERSVHTKTLQNTS

FMRP            600   SEGSRLRTGKDRNQKKEKPDSVDGQQPLVNGVP
FMRP_Δ331-396   534   SEGSRLRTGKDRNQKKEKPDSVDGQQPLVNGVP
```

Appendix XI - Assigned chemical shifts of FMRP KH1WT/KH2DD.

Residue	N	HN	Cα	Cβ
A216	123.71	8.09	51.12	17.80
S217	115.73	8.20	57.40	63.42
R218	123.59	8.45	?	30.03
F219	115.85	8.47	56.45	40.37
H220	122.36	8.78	52.13	30.84
E221	126.87	9.03	53.55	32.06
Q222	120.10	8.53	52.13	?
I224	120.44	8.44	?	39.56
V225	126.68	8.25	59.98	33.87
R226	126.57	8.53	56.71	30.41
E227	124.27	8.72	59.67	28.52
D228	114.89	9.05	54.02	37.60
M230	121.01	7.47	56.72	30.23
G231	106.11	8.51	45.78	---
L232	122.80	7.24	55.33	38.74
A233	120.22	7.34	54.02	17.73
I234	114.79	8.26	62.97	37.00
G235	107.20	7.43	42.74	---
G238	105.63	8.06	44.97	---
A239	119.56	7.62	54.49	17.87
N240	115.96	8.27	55.31	36.90
I241	119.51	7.86	63.55	?
Q243	117.66	7.45	57.34	26.51
A244	122.80	7.81	53.94	17.47
R245	113.60	8.01	57.33	29.29
K246	115.88	7.26	54.97	31.18
V247	125.33	7.17	60.64	30.84
G249	112.63	8.74	44.08	---
V250	121.73	7.94	62.74	30.17
T251	124.35	9.61	62.74	68.01
A252	122.33	7.98	52.53	21.58
I253	119.15	8.91	61.25	39.83
D254	127.51	9.41	52.26	44.22
D256	127.52	8.98	52.23	39.29
E257	125.51	8.93	59.35	29.00
D258	116.34	8.52	55.51	39.49
T259	105.27	7.18	59.76	69.48
C260	117.67	7.91	58.55	37.87
T261	110.19	7.24	60.71	69.16
F262	125.69	8.55	56.99	40.91
H263	117.29	8.90	53.82	32.53
I264	121.78	9.26	59.00	38.17
Y265	125.20	9.15	54.63	39.42
G266	107.68	8.94	45.57	---
E267	115.57	8.80	54.83	29.76
D268	117.41	7.18	52.59	44.62
Q269	126.01	8.51	58.31	27.66
D270	119.90	8.06	56.72	39.83

Residue	N	HN	Cα	Cβ
A271	124.89	8.29	54.63	19.08
V272	112.68	7.26	65.44	30.43
K274	122.80	8.20	58.88	31.38
K274	119.39	7.55	58.41	30.97
A275	121.60	8.23	55.23	18.67
R276	116.86	8.01	58.95	28.95
S277	113.64	7.73	59.67	61.68
F281	126.70	7.98	58.28	41.59
A282	123.78	9.08	50.24	21.78
E283	117.57	7.88	53.21	31.72
V285	121.64	8.57	59.90	32.40
I286	129.88	9.43		38.68
Q287	125.57	8.58	53.73	26.44
V288	126.52	8.79	57.81	33.47
G312	108.51	7.83	44.56	---
V313	115.19	7.56	61.59	30.10
V316	119.86	8.34	58.28	33.41
R317	126.66	8.84	53.55	31.65
I318	125.22	8.48	59.77	36.86
E319	128.07	8.22	54.74	27.87
A320	124.31	8.16	51.25	18.14
E321	119.34	8.55	55.71	28.68
N322	118.27	8.43	52.40	37.46
E323	121.34	8.36	55.95	28.66
K324	120.38	8.33	55.37	30.91
N325	117.64	8.50	52.78	37.40
V326	119.25	7.78	58.35	31.24
Q328	120.03	8.28	53.94	28.80
E329	122.85	8.83	54.63	29.77
E330	123.38	8.52	56.44	28.24
G331	111.18	8.76	44.49	---
M332	119.04	8.11	52.21	34.42
V333	124.31	9.68	57.81	33.41
F335	127.15	9.40	54.97	38.41
V336	122.56	8.93	60.31	30.91
F337	126.64	9.35	56.86	42.93
V338	118.68	8.37	59.56	33.21
G339	113.01	8.76	44.49	---
T340	112.05	8.64	60.71	68.15
K341	122.98	8.27	59.22	30.33
D342	120.10	9.34	56.32	38.68
A347	123.44	8.59	54.62	16.09
T348	112.89	8.19	66.05	67.53
V349	122.31	7.36	64.98	30.39
L350	120.49	7.77	56.79	41.32
L351	121.03	8.77	57.67	40.44

Appendix XII - Peak lists, individual Δδ, weighted average Δδ, normalised weighted average Δδ, and final scores for FMRP KH1WT/KH2DD and FMRP KH1KK/KH2DD.

Peak list for SIA analysis - FMRP KH1WT/KH2DD

Peak	$\delta^{15}N$	$\delta^{1}H$
UA4	110.320	7.255
V225	126.511	8.239
L232	122.784	7.239
I234	114.712	8.252
G238	105.582	8.047
Q243	117.655	7.459
A244	122.728	7.820
I253	119.113	8.895
F281	126.789	7.980
T340	112.079	8.647

Individual Δδ and weighted average Δδ - FMRP KH1WT/KH2DD

nAnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.002	0.011	0.011
V225	-0.076	0.012	0.027
L232	0.052	0.024	0.029
I234	-0.181	0.017	0.060
G238	-0.108	-0.025	0.042
Q243	0.039	0.029	0.032
A244	0.015	0.033	0.033
I253	-0.016	-0.01	0.011
F281	0.173	-0.004	0.055
T340	-0.054	-0.007	0.018

nnAnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.002	0.007	0.007
V225	-0.044	0.008	0.016
L232	0.028	0.018	0.020
I234	-0.154	0.012	0.050
G238	-0.103	-0.021	0.039
Q243	-0.002	0.021	0.021
A244	0.011	0.024	0.024
I253	0.004	-0.003	0.003
F281	0.162	-0.001	0.051
T340	-0.034	-0.009	0.014

nCnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.014	0.005	0.00668
V225	-0.036	0.008	0.01391
L232	0.013	0.012	0.01268
I234	-0.137	0.009	0.04425
G238	-0.059	-0.014	0.02333
Q243	0.029	0.02	0.022
A244	0	0.023	0.023
I253	-0.033	-0.006	0.01204
F281	0.076	-0.003	0.02422
T340	-0.025	-0.002	0.00815

nnCnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	-0.002	0.008	0.008
V225	-0.064	0.008	0.0218
L232	0.047	0.02	0.0249
I234	-0.189	0.021	0.0633
G238	-0.2	-0.025	0.068
Q243	0.033	0.036	0.0375
A244	-0.005	0.039	0.039
I253	0.031	-0.008	0.0127
F281	0.187	-0.009	0.0598
T340	-0.042	-0.008	0.0155

nGnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.002	0.008	0.00802
V225	-0.049	0.008	0.01744
L232	0.029	0.018	0.0202
I234	-0.174	0.012	0.05632
G238	-0.059	-0.024	0.0304
Q243	0.019	0.027	0.02766
A244	0.005	0.029	0.02904
I253	-0.009	-0.005	0.00575
F281	0.186	-0.004	0.05895
T340	-0.048	-0.006	0.01632

nnGnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.001	0.011	0.011
V225	-0.061	0.01	0.0217
L232	0.048	0.02	0.0251
I234	-0.195	0.01	0.0625
G238	-0.122	-0.029	0.0483
Q243	0.014	0.03	0.0303
A244	-0.017	0.034	0.0344
I253	0.032	-0.002	0.0103
F281	0.208	-0.005	0.066
T340	-0.05	-0.011	0.0193

nUnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.005	0.007	0.00718
V225	-0.041	0.006	0.01429
L232	0.034	0.019	0.02183
I234	-0.194	0.016	0.0634
G238	-0.112	-0.027	0.04454
Q243	0.034	0.027	0.02906
A244	0.007	0.03	0.03008
I253	-0.029	-0.009	0.01285
F281	0.141	-0.006	0.04499
T340	-0.043	-0.006	0.01486

nnUnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.001	0.003	0.003
V225	-0.019	0.005	0.0078
L232	0.024	0.011	0.0134
I234	-0.152	0.013	0.0498
G238	-0.104	-0.022	0.0396
Q243	-0.004	0.02	0.02
A244	-0.009	0.022	0.0222
I253	0.011	-0.006	0.0069
F281	0.133	-0.002	0.0421
T340	-0.032	-0.008	0.0129

Individual Δδ and weighted average Δδ - FMRP KH1WT/KH2DD (…cont)

nnnAn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.016	0.01	0.011
V225	-0.05	0.009	0.018
L232	0.051	0.022	0.027
I234	-0.169	0.015	0.056
G238	-0.148	-0.029	0.055
Q243	0.02	0.023	0.024
A244	0.001	0.027	0.027
I253	0.028	-0.001	0.009
D256	-0.035	-0.012	0.016
T340	-0.05	-0.009	0.018

nnnnA

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.01	0.007	0.008
V225	-0.038	0.007	0.014
L232	0.035	0.016	0.019
I234	-0.14	0.011	0.046
G238	-0.09	-0.017	0.033
Q243	0.002	0.022	0.022
A244	0.011	0.023	0.023
I253	-0.016	-0.001	0.005
F281	0.165	-0.003	0.052
T340	-0.036	-0.007	0.013

nnnCn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.015	0.013	0.01384
V225	-0.069	0.011	0.02444
L232	0.044	0.022	0.02603
I234	-0.195	0.021	0.06514
G238	-0.227	-0.032	0.07859
Q243	0.066	0.031	0.03737
A244	-0.005	0.041	0.04103
I253	0.017	-0.01	0.01135
F281	0.213	-0.008	0.06783
T340	-0.052	-0.01	0.01925

nnnnC

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.019	0.013	0.01432
V225	-0.072	0.016	0.02783
L232	0.06	0.029	0.03466
I234	-0.239	0.025	0.07961
G238	-0.212	-0.043	0.07965
Q243	-0.008	0.04	0.04008
A244	-0.354	0.108	0.15555
I253	0.015	-0.013	0.01384
F281	0.241	-0.007	0.07653
T340	-0.059	-0.015	0.02394

nnnGn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.016	0.008	0.00947
V225	-0.035	0.006	0.01259
L232	0.039	0.018	0.02182
I234	-0.165	0.009	0.05295
G238	-0.092	-0.018	0.03421
Q243	0.008	0.022	0.02214
A244	0.006	0.024	0.02407
I253	0.032	-0.002	0.01032
F281	0.172	-0.003	0.05447
T340	-0.048	-0.006	0.01632

nnnnG

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.008	0.007	0.00744
V225	-0.047	0.007	0.01643
L232	0.047	0.016	0.02184
I234	-0.163	0.012	0.05292
G238	-0.104	-0.022	0.03957
Q243	0.026	0.022	0.02349
A244	-0.372	0.108	0.15969
I253	-0.004	-0.006	0.00613
F281	0.143	-0.005	0.0455
T340	-0.034	-0.009	0.01402

nnnUn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.035	0.008	0.01366
V225	-0.025	0.008	0.01125
L232	0.014	0.016	0.0166
I234	-0.174	0.012	0.05632
G238	-0.087	-0.024	0.03651
Q243	0.029	0.025	0.02663
A244	0.006	0.028	0.02806
I253	0.001	-0.006	0.00601
F281	0.189	-0.003	0.05984
T340	-0.046	-0.007	0.01614

nnnnU

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd.	0.007	0.009	0.00927
V225	-0.046	0.007	0.01614
L232	0.032	0.019	0.02153
I234	-0.152	0.014	0.05006
G238	-0.13	-0.024	0.0476
Q243	0.011	0.026	0.02623
A244	-0.373	0.109	0.1606
I253	0.003	-0.003	0.00315
F281	0.171	-0.004	0.05422
T340	-0.045	-0.009	0.01684

Normalised weighted average Δδ - FMRP KH1WT/KH2DD

Position 1

Peak	nAnnn	nCnnn	nGnnn	nUnnn
unasgd.	1.000	0.606	0.728	0.651
V225	1.000	0.518	0.649	0.532
L232	1.000	0.436	0.694	0.750
I234	0.942	0.698	0.888	1.000
G238	0.950	0.524	0.683	1.000
Q243	1.000	0.698	0.878	0.922
A244	1.000	0.690	0.871	0.902
I253	0.872	0.937	0.448	1.000
D256	0.930	0.411	1.000	0.763
T340	1.000	0.442	0.884	0.805
Average	**0.969**	**0.596**	**0.772**	**0.833**

Position 2

Peak	nnAnn	nnCnn	nnGnn	nnUnn
unasgd.	0.639	0.729	1.000	0.274
V225	0.738	1.000	0.998	0.359
L232	0.799	0.992	1.000	0.532
I234	0.792	1.000	0.986	0.786
G238	0.570	1.000	0.710	0.582
Q243	0.561	1.000	0.809	0.535
A244	0.621	1.000	0.882	0.568
I253	0.257	1.000	0.815	0.548
D256	0.777	0.907	1.000	0.638
T340	0.728	0.805	1.000	0.670
Average	**0.648**	**0.943**	**0.920**	**0.549**

Position 3

Peak	nnnAn	nnnCn	nnnGn	nnnUn
unasgd.	0.810	1.000	0.684	0.987
V225	0.745	1.000	0.515	0.460
L232	1.000	0.954	0.800	0.609
I234	0.852	1.000	0.813	0.865
G238	0.701	1.000	0.435	0.465
Q243	0.638	1.000	0.593	0.713
A244	0.658	1.000	0.587	0.684
I253	0.785	1.000	0.909	0.529
D256	0.804	1.000	0.803	0.882
T340	0.945	1.000	0.848	0.839
Average	**0.794**	**0.995**	**0.699**	**0.703**

Position 4

Peak	nnnnA	nnnnC	nnnnG	nnnnU
unasgd.	0.536	1.000	0.520	0.647
V225	0.500	1.000	0.590	0.580
L232	0.561	1.000	0.630	0.621
I234	0.573	1.000	0.665	0.629
G238	0.416	1.000	0.497	0.598
Q243	0.549	1.000	0.586	0.654
A244	0.145	0.969	0.994	1.000
I253	0.373	1.000	0.443	0.227
D256	0.683	1.000	0.594	0.708
T340	0.558	1.000	0.586	0.703
Average	**0.489**	**0.997**	**0.611**	**0.637**

Final Scores - FMRP KH1WT/KH2DD

	Position			
	1	**2**	**3**	**4**
A	0.97	0.65	0.79	0.49
C	0.60	0.94	1.00	1.00
G	0.77	0.92	0.70	0.61
U	0.83	0.55	0.70	0.64

Peak list for SIA analysis - FMRP KH1KK/KH2DD

Peak	$\delta^{15}N$	$\delta^{1}H$
unasgd. 1	126.738	9.049
unasgd. 2	112.771	8.190
unasgd. 3	110.328	7.255
V225	126.550	8.236
M230	120.947	7.478
G231	106.138	8.522
L232	122.813	7.248
I234	114.648	8.280
G238	105.245	7.920
Q243	117.671	7.447
R245	113.711	8.021
I253	119.125	8.885
D256	127.669	8.948
E257	125.538	8.929
F281	126.685	7.969
T340	112.113	8.639

Individual Δδ and weighted average Δδ - FMRP KH1KK/KH2DD

nAnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.054	0.013	0.021
unasgd. 2	0.028	-0.008	0.012
unasgd. 3	0.038	0.027	0.030
V225	-0.212	0.026	0.072
M230	0.079	0.015	0.029
G231	-0.109	0.016	0.038
L232	0.159	0.053	0.073
I234	-0.368	0.039	0.123
G238	-0.379	-0.050	0.130
Q243	0.100	0.065	0.072
R245	0.011	0.022	0.022
I253	0.034	-0.022	0.024
D256	-0.105	-0.024	0.041
E257	-0.054	-0.004	0.018
F281	0.377	-0.009	0.120
T340	-0.137	-0.025	0.050

nnAnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.033	0.010	0.014
unasgd. 2	-0.008	-0.001	0.003
unasgd. 3	0.028	0.018	0.020
V225	-0.127	0.015	0.043
M230	0.094	0.015	0.033
G231	-0.072	0.001	0.023
L232	0.039	0.031	0.033
I234	-0.325	0.031	0.107
G238	-0.346	-0.046	0.119
Q243	0.016	0.048	0.048
R245	0.020	0.014	0.015
I253	0.015	-0.013	0.014
D256	-0.060	-0.015	0.024
E257	-0.038	-0.001	0.012
F281	0.233	-0.013	0.075
T340	-0.096	-0.023	0.038

nCnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.064	0.014	0.025
unasgd. 2	0.055	-0.018	0.025
unasgd. 3	0.026	0.026	0.027
V225	-0.212	0.026	0.072
M230	0.107	0.015	0.037
G231	-0.125	0.019	0.044
L232	0.119	0.048	0.061
I234	-0.362	0.039	0.121
G238	-0.349	-0.050	0.121
Q243	0.057	0.053	0.056
R245	0.006	0.025	0.025
I253	0.080	-0.019	0.032
D256	-0.095	-0.026	0.040
E257	-0.058	-0.021	0.028
F281	0.401	-0.005	0.127
T340	-0.169	-0.030	0.061

nnCnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.056	0.009	0.020
unasgd. 2	0.038	-0.015	0.019
unasgd. 3	0.026	0.021	0.023
V225	-0.175	0.016	0.058
M230	0.079	0.017	0.030
G231	-0.069	-0.014	0.026
L232	0.068	0.033	0.039
I234	-0.466	0.052	0.156
G238	-0.433	-0.046	0.144
Q243	0.122	0.075	0.084
R245	0.014	0.015	0.016
I253	0.064	-0.024	0.031
D256	-0.068	-0.024	0.032
E257	-0.046	-0.004	0.015
F281	0.297	-0.020	0.096
T340	-0.078	-0.016	0.029

nGnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.079	0.017	0.030
unasgd. 2	0.003	-0.013	0.013
unasgd. 3	0.020	0.025	0.026
V225	-0.236	0.029	0.080
M230	0.098	0.016	0.035
G231	-0.090	0.024	0.037
L232	0.150	0.057	0.074
I234	-0.454	0.045	0.150
G238	-0.452	-0.067	0.158
Q243	-0.022	0.061	0.061
R245	0.011	0.027	0.027
I253	0.060	-0.009	0.021
D256	-0.108	-0.032	0.047
E257	-0.069	-0.010	0.024
F281	0.449	-0.005	0.142
T340	-0.169	-0.032	0.062

nnGnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.133	0.016	0.045
unasgd. 2	-0.019	-0.018	0.019
unasgd. 3	0.043	0.028	0.031
V225	-0.250	0.030	0.085
M230	0.112	0.024	0.043
G231	-0.084	0.006	0.027
L232	0.107	0.049	0.060
I234	-0.652	0.060	0.215
G238	-0.678	-0.078	0.228
Q243	-0.778	0.101	0.266
R245	0.043	0.026	0.029
I253	0.058	-0.039	0.043
D256	-0.104	-0.036	0.049
E257	-0.063	0.002	0.020
F281	0.468	-0.029	0.151
T340	-0.166	-0.033	0.062

nUnnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.049	0.014	0.021
unasgd. 2	0.025	-0.022	0.023
unasgd. 3	0.032	0.019	0.022
V225	-0.200	0.015	0.065
M230	0.077	0.018	0.030
G231	-0.079	0.011	0.027
L232	0.078	0.035	0.043
I234	-0.360	0.034	0.119
G238	-0.317	-0.037	0.107
Q243	0.010	0.051	0.051
R245	0.019	0.015	0.016
I253	0.049	-0.001	0.016
D256	-0.093	-0.024	0.038
E257	-0.037	0.002	0.012
F281	0.312	-0.012	0.099
T340	-0.122	-0.022	0.044

nnUnn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.117	0.016	0.040
unasgd. 2	0.057	-0.024	0.030
unasgd. 3	0.049	0.030	0.034
V225	-0.227	0.036	0.080
M230	0.096	0.020	0.036
G231	-0.095	0.015	0.034
L232	0.143	0.057	0.073
I234	-0.575	0.055	0.190
G238	-0.712	-0.095	0.244
Q243	0.014	0.085	0.085
R245	0.041	0.031	0.034
I253	0.058	-0.033	0.038
D256	-0.099	-0.035	0.047
E257	-0.064	0.006	0.021
F281	0.501	-0.015	0.159
T340	-0.137	-0.034	0.055

Individual Δδ and weighted average Δδ - FMRP KH1KK/KH2DD (…cont)

nnnAn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.087	0.018	0.033
unasgd. 2	0.022	-0.023	0.024
unasgd. 3	0.036	0.030	0.032
V225	-0.213	0.026	0.072
M230	0.091	0.023	0.037
G231	-0.127	0.022	0.046
L232	0.074	0.039	0.045
I234	-0.506	0.050	0.168
G238	-0.466	-0.051	0.156
Q243	0.067	0.075	0.078
R245	0.040	0.026	0.029
I253	0.065	-0.018	0.027
D256	-0.121	-0.033	0.051
E257	-0.058	0.003	0.019
F281	0.460	-0.012	0.146
T340	-0.160	-0.030	0.059

nnnnA

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.060	0.012	0.022
unasgd. 2	-0.009	-0.007	0.008
unasgd. 3	0.037	0.018	0.021
V225	-0.137	0.017	0.047
M230	0.067	0.017	0.027
G231	-0.059	0.000	0.019
L232	0.124	0.049	0.063
I234	-0.450	0.049	0.151
G238	-0.476	-0.060	0.162
Q243	0.057	0.068	0.070
R245	0.018	0.018	0.019
I253	0.048	-0.021	0.026
D256	-0.089	-0.023	0.036
E257	-0.040	0.002	0.013
F281	0.339	-0.018	0.109
T340	-0.104	-0.024	0.041

nnnCn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.097	0.020	0.037
unasgd. 2	0.059	-0.025	0.031
unasgd. 3	0.045	0.033	0.036
V225	-0.274	0.028	0.091
M230	0.102	0.021	0.038
G231	-0.157	0.039	0.063
L232	0.153	0.058	0.076
I234	-0.490	0.044	0.161
G238	-0.577	-0.075	0.197
Q243	-0.001	0.081	0.081
R245	0.025	0.032	0.033
I253	0.057	-0.016	0.024
D256	-0.144	-0.041	0.061
E257	-0.055	0.012	0.021
F281	0.539	-0.013	0.171
T340	-0.139	-0.040	0.059

nnnnC

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.085	0.007	0.028
unasgd. 2	0.055	-0.014	0.022
unasgd. 3	0.023	0.009	0.012
V225	-0.023	0.016	0.018
M230	0.091	0.022	0.036
G231	-0.105	-0.029	0.044
L232	0.034	0.019	0.022
I234	-0.413	0.045	0.138
G238	-0.350	-0.049	0.121
Q243	0.003	0.062	0.062
R245	-0.006	0.013	0.013
I253	0.053	-0.024	0.029
D256	-0.081	-0.019	0.032
E257	-0.026	-0.002	0.008
F281	0.345	-0.012	0.110
T340	-0.105	-0.022	0.040

nnnGn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.069	0.015	0.026
unasgd. 2	0.002	-0.011	0.011
unasgd. 3	0.016	0.022	0.023
V225	-0.213	0.020	0.070
M230	0.082	0.017	0.031
G231	-0.092	0.017	0.034
L232	0.084	0.042	0.050
I234	-0.409	0.036	0.134
G238	-0.319	-0.033	0.106
Q243	0.003	0.063	0.063
R245	0.033	0.025	0.027
I253	0.049	-0.007	0.017
D256	-0.118	-0.030	0.048
E257	-0.040	0.006	0.014
F281	0.412	-0.011	0.131
T340	-0.129	-0.030	0.051

nnnnG

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.088	0.009	0.029
unasgd. 2	0.000	-0.014	0.014
unasgd. 3	0.026	0.014	0.016
V225	-0.088	0.018	0.033
M230	0.077	0.018	0.030
G231	-0.091	-0.006	0.029
L232	0.036	0.029	0.031
I234	-0.382	0.034	0.125
G238	-0.366	-0.054	0.128
Q243	0.017	0.060	0.060
R245	0.016	0.014	0.015
I253	0.036	-0.018	0.021
D256	-0.087	-0.022	0.035
E257	-0.045	0.002	0.014
F281	0.335	-0.012	0.107
T340	-0.096	-0.023	0.038

nnnUn

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.079	0.012	0.028
unasgd. 2	0.077	-0.027	0.036
unasgd. 3	0.014	0.020	0.020
V225	-0.201	0.023	0.068
M230	0.098	0.018	0.036
G231	-0.113	0.007	0.036
L232	0.073	0.040	0.046
I234	-0.437	0.040	0.144
G238	-0.363	-0.044	0.123
Q243	-0.041	0.061	0.062
R245	0.027	0.025	0.026
I253	0.063	-0.017	0.026
D256	-0.122	-0.026	0.047
E257	-0.049	-0.002	0.016
F281	0.417	-0.010	0.132
T340	-0.150	-0.030	0.056

nnnnU

Peak	$\Delta\delta^{15}N$	$\Delta\delta^{1}H$	$\Delta\delta_{av}$
unasgd. 1	-0.060	0.014	0.024
unasgd. 2	0.049	-0.026	0.030
unasgd. 3	0.028	0.023	0.025
V225	-0.213	0.021	0.071
M230	0.090	0.018	0.034
G231	-0.099	0.013	0.034
L232	0.090	0.040	0.049
I234	-0.412	0.039	0.136
G238	-0.344	-0.059	0.124
Q243	-0.032	0.064	0.065
R245	0.030	0.016	0.019
I253	0.058	-0.010	0.021
D256	-0.095	-0.027	0.040
E257	-0.058	0.002	0.018
F281	0.411	-0.017	0.131
T340	-0.141	-0.025	0.051

Normalised weighted average Δδ - FMRP KH1KK/KH2DD

Position 1

Peak	nAnnn	nCnnn	nGnnn	nUnnn
unasgd. 1	0.653	0.440	1.000	0.683
unasgd. 2	0.497	0.113	1.000	0.314
unasgd. 3	0.921	0.625	1.000	0.669
V225	0.996	0.594	1.000	0.645
M230	0.791	0.904	1.000	0.737
G231	0.830	0.498	1.000	0.407
L232	1.000	0.457	0.623	0.859
I234	0.732	0.640	1.000	0.898
G238	0.801	0.732	0.962	1.000
Q243	0.927	0.619	1.000	0.903
R245	0.770	0.531	1.000	0.653
I253	0.896	0.506	1.000	0.948
D256	0.811	0.479	1.000	0.719
E257	0.944	0.649	1.000	0.689
F281	0.819	0.513	1.000	0.745
T340	0.850	0.647	1.000	0.692
Average	**0.827**	**0.559**	**0.974**	**0.723**

Position 2

Peak	nnAnn	nnCnn	nnGnn	nnUnn
unasgd. 1	0.672	0.542	1.000	0.759
unasgd. 2	0.802	0.616	1.000	0.716
unasgd. 3	0.759	0.628	1.000	0.322
V225	0.790	0.633	1.000	0.193
M230	0.962	0.785	1.000	0.941
G231	0.695	0.411	1.000	0.698
L232	0.808	0.521	1.000	0.289
I234	0.751	0.970	1.000	0.858
G238	0.614	0.732	1.000	0.614
Q243	0.664	1.000	0.960	0.735
R245	0.761	0.474	1.000	0.399
I253	1.000	0.992	0.762	0.925
D256	0.648	0.526	1.000	0.520
E257	1.000	0.541	0.758	0.303
F281	0.742	0.562	1.000	0.642
T340	1.000	0.480	0.970	0.650
Average	**0.792**	**0.651**	**0.966**	**0.598**

Position 3

Peak	nnnAn	nnnCn	nnnGn	nnnUn
unasgd. 1	0.672	1.000	0.588	0.650
unasgd. 2	0.687	1.000	0.581	0.738
unasgd. 3	0.828	1.000	0.725	0.522
V225	0.947	1.000	0.831	0.392
M230	0.815	1.000	0.725	0.708
G231	1.000	0.731	0.905	0.790
L232	1.000	0.803	0.670	0.420
I234	0.701	1.000	0.625	0.584
G238	0.692	1.000	0.465	0.560
Q243	0.962	1.000	0.778	0.770
R245	0.928	1.000	0.923	0.507
I253	0.487	1.000	0.395	0.494
D256	0.960	1.000	0.982	0.722
E257	1.000	0.834	0.583	0.599
F281	0.942	1.000	0.867	0.707
T340	1.000	0.995	0.813	0.611
Average	**0.851**	**0.960**	**0.716**	**0.611**

Position 4

Peak	nnnnA	nnnnC	nnnnG	nnnnU
unasgd. 1	0.518	1.000	0.688	0.585
unasgd. 2	0.643	0.826	1.000	0.832
unasgd. 3	0.638	1.000	0.607	0.730
V225	0.809	1.000	0.842	0.879
M230	0.833	1.000	0.986	0.926
G231	0.750	0.922	1.000	0.931
L232	0.588	1.000	0.635	0.675
I234	0.625	1.000	0.757	0.716
G238	0.437	1.000	0.503	0.506
Q243	0.600	1.000	0.733	0.761
R245	0.481	1.000	0.786	0.554
I253	0.411	1.000	0.694	0.553
D256	0.808	1.000	0.991	0.860
E257	0.562	1.000	0.740	0.874
F281	0.625	1.000	0.831	0.824
T340	0.806	1.000	1.019	0.928
Average	**0.633**	**0.984**	**0.801**	**0.758**

Final Scores - FMRP KH1KK/KH2DD

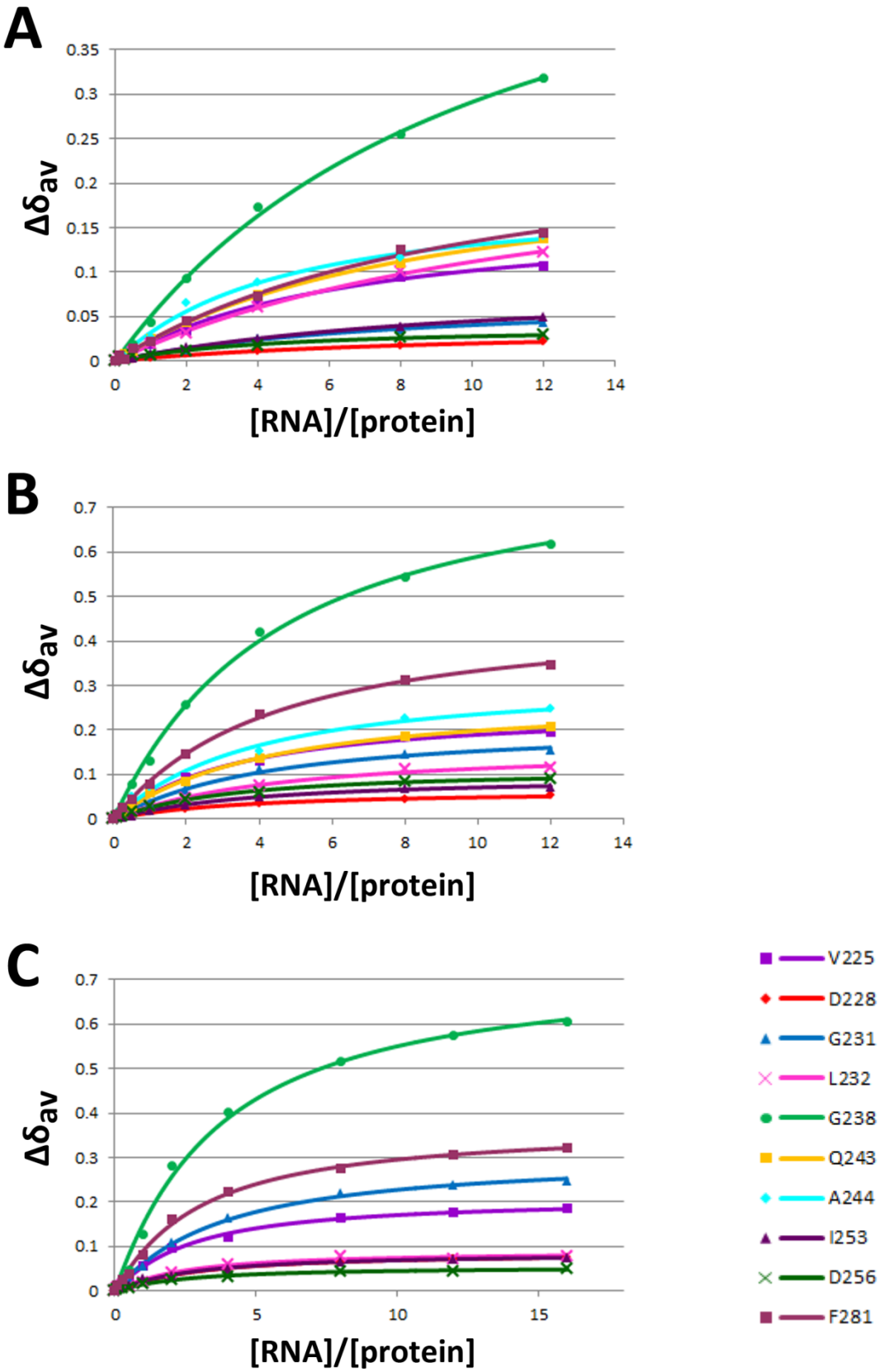
	Position			
	1	**2**	**3**	**4**
A	0.83	0.79	0.85	0.63
C	0.56	0.65	0.96	0.98
G	0.97	0.97	0.72	0.80
U	0.72	0.60	0.61	0.76

Appendix XIII – Peak list and binding isotherms for FMRP KH1KK/KH2DD titrations with CACCC, CGCCC and CAGCC

List of peaks used in the determination of the dissociation constants

Peak	$\delta^{15}N$	$\delta^{1}H$
V225	126.550	8.227
D228	114.994	9.048
G231	106.149	8.512
L232	122.808	7.240
G238	105.260	7.911
Q243	117.642	7.437
A244	122.712	7.794
I253	119.138	8.881
D256	127.682	8.936
F281	126.654	7.962

Binding isotherms for T-STAR KH titrations with A) CACCC. B) CGCCC. C) CAGCC.



Appendix XIV – Amino acid sequence of *Homo sapiens* RNA-binding protein 10 isoform 1. NCBI Reference Sequence NP_005667.2

```
  1 meyerrggrg drtgrygatd rsqddggenr srdhdyrdmd yrsypreygs qegkhdydds
 61 seeqsaedsy easpgsetqr rrrrrhrhsp tgppgfprdg dyrdqdyrte qgeeeeeeed
121 eeeeekasni vmlrmlpqaa teddirgqlq shgvqarevr lmrnkssgqs rgfafvefsh
181 lqdatrwmea nqhslnilgq kvsmhysdpk pkinedwlcn kcgvqnfkrr ekcfkcgvpk
241 seaeqklplg trldqqtlpl ggrelsqgll plpqpyqaqg vlasqalsqg sepssenand
301 tiilrnlnph stmdsilgal apyavlsssn vrvikdkqtq lnrgfafiql stiveaaqll
361 qilqalhppl tidgktinve fakgskrdma snegsrisaa svastaiaaa qwaisqasqg
421 gegtwatsee ppvdysyyqq degygnsqgt esslyahgyl kgtkgpgitg tkgdptgagp
481 easlepgads vsmqafsraq pgaapgiyqq saeasssqgt aansqsytim spavlkselq
541 spthpssalp patsptaqes ysqypvpdvs tyqydetsgy yydpqtglyy dpnsqyyyna
601 qsqqylywdg errtyvpale qsadghketg apskegkekk ekhktktaqq iakdmerwar
661 slnkqkenfk nsfqpisslr dderresata dagyailekk galaerqhts mdlpklasdd
721 rpspprglva aysgesdsee eqerggpere ekltdwqkla cllcrrqfps kealirhqql
781 sglhkqnlei hrrahlsene lealekndme qmkyrdraae rrekygipep pepkrrkygg
841 istasvdfeq ptrdglgsdn igsrmlqamg wkegsglgrk kqgivtpiea qtrvrgsglg
901 argssygvts tesyketlhk tmvtrfneaq
```

Appendix XV - Chemical shifts assigned to RRM1, ZF or RRM2 of RBM10

RRM1

15N	1H
120.179	10.074
113.999	9.723
117.339	9.560
117.680	9.473
107.539	8.960
104.045	8.241
106.791	8.341
113.017	8.427
117.216	9.023
125.050	9.410
127.891	9.431
128.204	9.201
129.643	9.100
126.329	9.059
126.899	8.755
124.757	8.481
121.580	8.815
122.043	8.824
122.573	8.779
120.371	7.421
128.061	9.891
124.053	9.736
123.728	9.092
123.516	8.620
116.169	8.235
115.838	8.033
108.320	8.003
119.139	7.348
119.015	7.745
120.819	7.697
121.332	7.912
126.395	9.125
124.704	9.323
126.362	8.418
105.167	8.269

ZF

15N	1H
130.262	9.721
130.623	9.372
128.903	9.281
125.170	9.064
122.312	9.325
120.618	9.060
128.050	8.621
123.913	8.866
125.948	8.723
125.931	8.805
126.411	8.380
124.355	8.017
113.671	8.180

RRM2

15N	1H
102.578	8.663
107.737	9.465
106.347	8.268
107.771	8.176
108.669	7.977
118.178	9.138
121.914	9.435
123.647	9.847
123.967	9.444
131.077	9.867
129.060	9.189
129.012	8.853
129.633	8.520
128.056	7.966
125.898	8.893
124.392	8.705
126.047	8.491
126.792	8.302
126.924	8.197
127.713	8.666
123.648	7.457
122.433	7.598
122.986	7.798
120.341	7.485
120.308	7.245
119.488	7.632
118.999	7.842
116.235	7.781
115.358	7.940
114.118	7.922
114.200	7.975
106.298	7.271