CHARACTERIZING THE RNA BINDING PROPERTIES OF THE INTRINSICALLY DISORDERED FUS PROTEIN AND RGG/RG DOMAINS

by

BAGDESER AKDOGAN OZDILEK

B.Sc., Istanbul University, 2009

M.Sc., Fatih University, 2012

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Philosophy

Department of Molecular, Cellular and Developmental Biology

2017

This thesis entitled:
Characterizing the RNA binding properties of the intrinsically disordered
FUS protein and RGG/RG domains
written by Bagdeser Akdogan Ozdilek
has been approved for the department of Molecular, Cellular and Developmental Biology

_____
Ravinder Singh

_____
Robert T. Batey

Date_____

The final copy of this thesis has been examined by the signatories, and we
find that both the content and the form meet acceptable presentation standards
of scholarly work in the above mentioned discipline.

Ozdilek, Bagdeser Akdogan (Ph.D., Molecular, Cellular and Developmental Biology)

Characterizing the RNA binding properties of the intrinsically disordered FUS protein and RGG/RG domains

Thesis directed by Professor Robert T. Batey

Abstract

Recent developments in the comprehensive identification of the RNA-binding protein (RBP) repertoire has accelerated discovery of new RBPs. According to these studies, 20% of both known and novel RNA-binding proteins are highly disordered. Moreover, analysis of the human mRNA interactome revealed most disease related mutations are found within intrinsically disordered RNA-binding domains (RBDs). For most of these proteins, their RNA-binding properties are poorly characterized. Thus, deciphering intrinsically disordered RBD-RNA interactions on a molecular scale is essential to understanding their impact upon human physiology and diseases.

RGG/RG (arginine/glycine) domains are the second most common RNA binding domain in the human genome, yet their RNA-binding properties have not been well understood. Proteins containing RGG/RG domains regulate all aspects of RNA metabolism including transcription, processing, nucleocytoplasmic shuttling and translation. Proteins such as Fused in Sarcoma (FUS), Fragile X mental retardation (FMRP) and hnRNP U, bind a majority of the cellular transcriptome such that their recognition of RNAs has been considered to be non-specific or “promiscuous”.

Here, I report a detailed analysis of the RNA-binding characteristics of the RGG/RG domains from FUS, FMRP and hnRNP U. While previous studies of FUS focused on RNA binding by the RRM and zinc finger (ZnF) domains, my analysis showed RNA binding activity is driven by the RGG/RG domains. Further, I observed a strong synergy between the RRM and adjacent RGG/RG domains to achieve RNA binding affinities of the full-length FUS. To better characterize RNA-binding properties of RGG/RG domains, we have analyzed RGG/RG

domains of FUS, FMRP and hnRNP U *in vitro* against a spectrum of different RNAs with well-defined structural and sequence features. These experiments revealed that RGG/RG domains have different degrees of preference for binding to RNAs but share consistent trends in their selectivity towards RNAs with complex secondary structure. Thus, the binding behavior of RGG domains is best described as “degenerate specificity” reflecting that RGG/RG domains interact with a broad spectrum of RNAs that contain frequently observed sequence/structural elements. This mode of specificity is likely further facilitated by the intrinsically disordered nature of RGG/RG domains that enable them to adopt multiple conformations to adaptively bind RNA.

Dedication

To

My lovely husband, Ahmet Ozdilek

And my entire family

Acknowledgement

First and foremost I would like to take this opportunity to express the deepest appreciation to my advisor, Prof. Robert T. Batey, for his continuous support, excellent guidance, caring, patience and providing me an excellent atmosphere for doing research.

I would like to deeply thank Prof. Jacob C. Schwartz, with whom we collaborated on this thesis project, for his enthusiasm and faith. I also want to thank him for his guidance, valuable suggestions, and great contributions to my academic progress.

This dissertation could not have been completed without the generous help, advice and support of my thesis committee members. I am grateful to Prof. Deborah Wuttke, who has long been an inspiring figure for me. She has been a very valuable source of advice for my thesis project. I would like to deeply thank Prof. Ravinder Singh, Prof. Thomas Perkins and Prof. Shelley Copley for insightful comments and encouragement, but also for the hard questions that motivated me to widen my research from various perspectives. I am very fortunate to have been guided by such great minds.

I am grateful and deeply indebted to my parents, sisters, and brother for their understanding, support and prayers throughout my life.

This journey would not have been possible without the support of my husband, Ahmet Ozdilek. I am grateful for his patience, encouragement and sacrifice. He has also provided insightful discussions about the research.

There are also those who become like a family throughout my Ph.D. I would like to deeply thank Bahar Kavi, Alptekin Kavi, Gulden Kurdoglu and Omer Kurdoglu, with whom I met in Boulder. They were always there when I needed help the most. I was lucky to become close friend with Husna Karaboga, who was my classmate in M.Sc. education. Then, we started doing Ph.D. in different states of the USA. But, the distance between us made our friendship even

stronger. She has become a sister to me for the last five years. I want to thank her for being supportive and listening to me whenever I felt stressed out.

I would also like to give a heartfelt, special thanks to my other friends, Gulsah Kayki, Hulya Kiray, Tuba Evsan, Reham Elhawary, Umida Djakbarova, Esma Turgut and Dilara Batan for their valuable support and motivation that inspires me to do my best.

I thank my colleagues for providing a stimulating and fun filled environment. My thanks go to Jacob Polaski, Zachariah Holmes, Michal Matyjasik, Alex Hopkins, Arden D. Barbour, Meagan Nakamoto, Nicholas Parsonnet and Neil Lloyd who have always been willing to help and give their best suggestions. They have been great resource of scientific wisdom, and critical thinking.

My special gratitude goes to Republic of Turkey Ministry of National Education for their financial support of first two years of my Ph.D. education.

Table of Contents

Chapter 1 : Introduction 1

1.1 Intrinsically disordered proteins and their functions 2

1.1.1 RNA binding activity of intrinsically disordered proteins 8

1.1.1.1 Arginine/Serine Rich disordered RNA binding domain 11

1.1.1.2 RNA binding activity of basic patches 13

1.1.1.3 Arginine/Glycine rich (RGG/RG) RNA binding domain 16

1.1.2 Intrinsically disordered RNA binding proteins and human diseases 19

1.2 Sequence specific, non-specific RNA binding modes and possible binding modes between them 23

1.3 Summary of the thesis 26

Chapter 2 : Materials and Methods 28

2.1 Protein Expression and Purification 28

2.2 *In vitro* transcription and purification of RNAs 29

2.3 RNA Body Labeling Reaction 31

2.4 RNA 5`-End Labeling Reaction 31

2.5 Electrophoretic Mobility Shift Assay 32

2.6 Isothermal Titration Calorimetry (ITC) 33

2.7 CD spectroscopy of proteins 34

2.8 RNase T_1 probing 34

2.9 Composition gradient-multiangle light scattering (CG-MALS) 35

2.10 Dimethyl sulfate (DMS) foot printing 36

2.11 Primer extension 38

2.12 Iodine foot printing 39

2.13 *In vivo* analysis of FUS RNA-binding activity 39

Chapter 3 : RGG/RG domains mediate RNA binding activity of Fused in Sarcoma (FUS) protein 41

3.1 Introduction 41

3.2 Results 43

3.2.1 Individual RRM and ZnF domains of FUS do not bind RNA with high affinity 43

3.2.2 FUS RGG/RG domains promote RNA binding 48

3.2.3 Electrostatic interactions are involved in RGG/RG interactions with RNA 50

3.2.4 Individual RGG domains can bind to RNA 52

3.2.5 RGG/RG domains are important for RNA binding in cells 55

3.2.6 RNA binding mode of FUS protein 56

3.3 Discussion 63

Chapter 4 : Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding 67

4.1 Introduction 67

4.2 Results 69

4.2.1 RGG/RG domains display degenerate specificity 69

4.2.2 G-quadruplex is not requisite for RGG/RG binding 73

4.2 Discussion 75

Chapter 5 : **Conclusions and Future remarks 80**

5.1 Digging deeper into the world of the disordered RNA binding domains 80

5.2 Targeting the intrinsically disordered RNA binding proteins in diseases 83

References 86

Appendix 110

List of Tables

Table 3.1 $K_{D,app}$ (μM) values of ZnF, RRM and RRM+3RGG interactions with different RNA molecules. 48

Table 3.2 RRM-RGG domains binding to DNMT RNA. 50

Table 4.1 Effects of mutations in Sc1 RNA on RNA binding activity of RGG domains. 75

Table A 1 Sequences of RNA substrates used in this thesis. 110

Table A 2 Corresponding biophysical parameters of the ITC graphs in Figure A 1. 112

Table A 3 Mutated arginine residues of SGG mutants. 113

Table A 4 Corresponding $K_{D,app}$ (μM) values of heat-map data in Figure 4.2. 114

List of Figures

Figure 1.1 Coupled folding and binding in disordered proteins. 6

Figure 1.2 Flexible RNA recognition mode of intrinsically disordered RNA binding domains. ...10

Figure 1.3 Comparison of Tat ARM domains and their RNA recognition modes. 15

Figure 1.4 Solution structure of the RGG/RG peptide from FMRP bound to in vitro selected Sc1 RNA. 18

Figure 1.5 Summary of the regulatory non-coding RNAs/mRNA metabolisms and dynamic nature of RNPs in regulation of cellular processes.. 22

Figure 2.1 SDS-PAGE analysis of protein expression and purification. 29

Figure 2.2 Method design for characterization of self- and hetero-association of FUS and Sc1 RNA. 36

Figure 3.1 RRM and ZnF domains of FUS do not bind to the RNA with a high affinity. 45

Figure 3.2 RRM and ZnF domains of FUS are correctly folded, but do not bind to RNA. 47

Figure 3.3 Flanking RGG/RG domains impart the RNA binding activity of the RRM. 49

Figure 3.4 Flanking RGG domains impart the RNA binding activity of the ZnF domain. 50

Figure 3.5 Interaction between RGG1-RRM-RGG2 and DNMT RNA shows a salt dependence. 52

Figure 3.6 Individual RGG/RG domains of FUS bind to RNA. 53

Figure 3.7 RGG/RG domains of FUS mediate high affinity binding to RNA. 54

Figure 3.8 RGG domains mediate RNA binding activity of FUS in vivo. 55

Figure 3.9 FUS protein preferentially binds to structured RNAs. 57

Figure 3.10 Dimethyl sulfate footprinting analysis of FUS and hRRD interaction. 59

Figure 3.12 FUS binding sites on the predicted secondary structure of hRRD. 61

Figure 3.13 Two FUS proteins bind to Sc1 RNA. 62

Figure 4.1 G-quartets formation in Sc1, DNMT and GGUG RNAs. 71

Figure 4.2 RGG/RG domains represent moderate preference for structured RNAs. 72

Figure 4.3 RGG/RG domains and their amino acid sequences are conserved across diverse metazoan species. 73

Figure 4.4 Effects of G-quartet structure on RNA binding activity of different RGG domains. ...74

Figure 4.5 Model for RNA recognition by RGG/RG domains.. 78

Figure A 1 ITC raw data of RGG1-RRM-RGG2 interaction with DNMT RNA at different KCl concentrations.....111

Chapter 1 : Introduction

The traditional sequence-to-structure-to-function paradigm assumes that most proteins need to adopt a defined three-dimensional structure to carry out their function. However, last couple of decades has witnessed an increase in the known number of intrinsically disordered proteins (IDPs) that lack stable tertiary and/or secondary structures under physiological conditions (1-3). Indeed, it has been predicted that more than 30% of eukaryotic proteins contain intrinsically disordered regions (4). These proteins are involved in regulation, signaling and protein-protein interaction networks.

Disordered regions are also common in the RNA binding interface of the proteins. Comprehensive determination of the RNA binding protein (RBP) repertoire of mammalian cells identified that ~20% of RBPs are disordered (5,6). These regions are enriched in disorder promoting (serine (S), glycine (G) and proline (P)) and positively charged (arginine (R) and lysine (K)) amino acids. These amino acids usually form distinct patterns such as RGG and RS repeats, R- and K- rich basic patches. Intrinsically disordered RBPs are essential for the regulation of all steps of gene expression. Thus, mutations in these regions create deleterious affects to the cell and cause different diseases including neuropathies, muscular atrophies and cancer (7).

In this chapter, I review the structure and function of intrinsically disordered proteins. The specifics of RNA recognition mediated by RS, RGG/RG domains and R-, L- basic patches are discussed in detail based on the recent discoveries. I also describe the relation of intrinsically disordered RBP mutations with human disease. Finally, I review current understanding of sequence specific-nonspecific RNA binding modes of proteins and discuss the presence of possible binding modes between them.

1.1 Intrinsically disordered proteins and their functions

Initially, it was thought that protein domains with well-defined secondary and tertiary structures were required for protein function. Thus, proteins were grouped into families according to their “functional” (globular) domains as in Pfam database. However, in the past two decades, it was recognized that disordered regions are very abundant (30-50% of all proteins) in eukaryotic proteins involved in all aspects of cellular regulation, signaling and homeostasis (3,8-10). The term intrinsically disordered protein (IDP) has been used to describe a broad spectrum of protein structures from fully disordered to multi-domain proteins with only short intrinsically disordered regions (IDRs) between globular domains.

Despite being functional comparable to folded regions, IDRs are different from globular domains in terms of their amino acid composition and biophysical properties (1,11). IDRs have a low hydrophobicity, a large net charge and as a result, lack unique three-dimensional structures when they are alone in a solution. IDRs have a tendency to be depleted in order promoting amino acids such as bulky hydrophobic amino acids (Ile, Leu, Val) and aromatic amino acids (Trp, Tyr, Phe) that constitute the hydrophobic core of a globular protein and Cys amino acid that increases the stability of a folded protein via the disulfide bond formation (12,13). IDRs are enriched in disorder promoting, polar amino acids (Ser, Arg, Gly, Gln, Glu, Lys), hydrophobic (Ala) and structure breaking (Pro) amino acids (12,14-16). These amino acids form repetitive amino acid patterns in the protein (for example, R/G/G or R/G) that also give the name to the corresponding domain (for example, RGG/RG domain).

The structure of IDPs or IDRs cannot be described by a single average conformation but instead they exist as dynamic ensembles of interconverting conformations in their unbound states under physiological conditions (Figure 1.1) (17). Each of these conformations is important for the protein function. IDRs show high conformational entropy; therefore, transitioning from a set of conformations to a more restricted arrangement needs payment of entropic cost (18). IDRs can overcome the entropic cost by increasing the gain of enthalpy of an interaction

(achieved via large interaction interfaces of IDRs) (18). Another way to deal with the entropic penalty is minimizing entropy loss upon binding. This can be achieved by limiting the loss of conformational entropy at the IDRs interaction interfaces (19) or by interacting with molecules that have rigid interfaces in which complexes only the IDRs pay an entropic penalty upon binding (20). In addition, the water molecules that bind to the free binding surface of the disordered proteins pay some of the conformational entropic cost of binding (20). Extrinsic factors such as cellular conditions, post-translational modifications and ligand binding can preferentially interact and stabilize certain conformations of the ensemble. They can drive "induced unfolding" that is important for the function of chaperones. Unfolding of certain regions of chaperone by these factors enable folding of the client proteins as described by entropy transfer model (21,22). More common effect of extrinsic factors on IDRs is inducing the folding. "Induced folding" of IDRs requires the payment of entropic cost. Extrinsic factors modulate the free energy landscape of IDRs by providing an enthalpy gain to the system or by increasing the configurational entropy (18). Thus, having high conformational entropy is an advantage for IDRs since controlling of the entropy by extrinsic factors allows IDRs to gain of different conformations therefore facilitates regulation of function of these proteins.

IDRs provide multiple advantages to proteins in their cellular roles. In a recent review paper, a list with 21 advantages was given (23). But, many of them do not have a direct proof yet. If I summarize the advantages that have been studied recently, the list includes the followings: (i) contributing to the flexibility of the proteins as linker domains between structured regions (24-27). This facilitates promiscuous interactions with different targets. (ii) Providing small recognition elements that can also mediate interaction with a large number of targets (28-30). (iii) Enabling the regulation of protein function because of the accessible post-translational modification (PTM) sites within IDRs (31-33). (iv) Modulating cellular functions by affecting the half-life of the proteins (34-36). (v) Tunneling through a narrow pore that is important for signaling activity (37). (vi) Providing kinetic advantages since IDPs have greater on-rates and

off-rates of binding than ordered proteins allowing signals to rapidly turn on and off without an excessive binding strength (38,39). I should note that on average, IDPs form weaker interactions with faster dissociation rate constants than the ordered proteins (40), but the available range of these kinetic properties of IDPs is as wide as of ordered proteins. This might be related to cellular functions of IDPs.

Coupled folding and binding is the primary reason of low affinity of IDPs/IDRs (40-42). As previously explained, folding of disordered regions needs payment of the entropic cost that increases the free energy, however, binding to their targets decreases the free energy. The net free energy change of these reactions becomes smaller than the binding reaction alone. This results in lower affinity that allows IDPs to have high dissociation rates. In addition, this can also allow specific binding without high affinity. It is believed in that combining high specificity with low affinity is an advantage in the context of signaling. However, there are still debates about how IDPs/IDRs can achieve the binding specificity. Specificity in binding is mainly determined by the size and complementarity of the binding interface. But, the interaction interface of the IDPs/IDRs to their binding targets is highly complementary because of the flexible nature of them that can result in “promiscuous” interactions other than the specific ones. IDPs/IDRs can overcome the specificity problem by using the extended structures formed in the complex instead of using only the short disordered regions for the molecular recognition. Moreover, posttranslational modifications of disordered regions can also increase the specificity by changing the charge or flexibility of these regions. High specificity is necessary for the critical functions of IDPs in signal recognition, transduction and regulation. However, further studies are necessary to understand principles of specific recognition mode of IDPs.

Structural disorder information about the IDPs is gained by different biophysical techniques such as nuclear magnetic resonance (NMR), small-angle X-ray scattering (SAXS) and single-molecule fluorescence resonance energy transfer (smFRET) (43,44). Among them, NMR is the most useful technique that gives information about local disorder, folding upon

binding and disorder in complex (45). SAXS is used to determine hydrodynamic parameters and degree of protein globularity (46,47). smFRET provides useful data about the dynamics of individual conformations of unbound ensemble (48,49). All of these experimentally verified structural ensembles of IDPs are deposited in databases such as DisProt, IDEAL and pE-DB (50-52). Since the high abundance and functional importance of IDPs and IDRs, computational tools were developed based on experimental data and biophysical properties of IDRs to efficiently predict the disordered regions and functions of IDPs that lack experimental annotations (3,53). Predictors of function of intrinsic disorder find a subset of IDRs that implement specific functions by binding to DNA, RNA or protein since the most common functions of IDPs are binding related ones.

IDPs have crucial roles in cellular processes such as control of the gene expression, cell cycle and signaling pathways (54-56). In addition to their regulatory activity, they can function as central hubs for protein-protein interaction networks since they have multiple interaction motifs (57-59). For example, High Mobility Group A (HMGA) protein is a well-studied hub protein that is completely disordered and goes to disorder-order transition upon binding to its DNA and protein targets (60-63). HMGA is member of architectural transcription factors protein class and considered as a central hub of nuclear function. It can bind to at least 18 known protein partners and to several specific DNA structures (64). Hence, it participates in modulation of chromosome and chromatin mechanics and regulates the expression of more than 45 different eukaryotic and viral genes. Because of the flexibility of the disordered domains in such proteins, they can specifically interact with multiple protein partners and involve in different cellular processes.

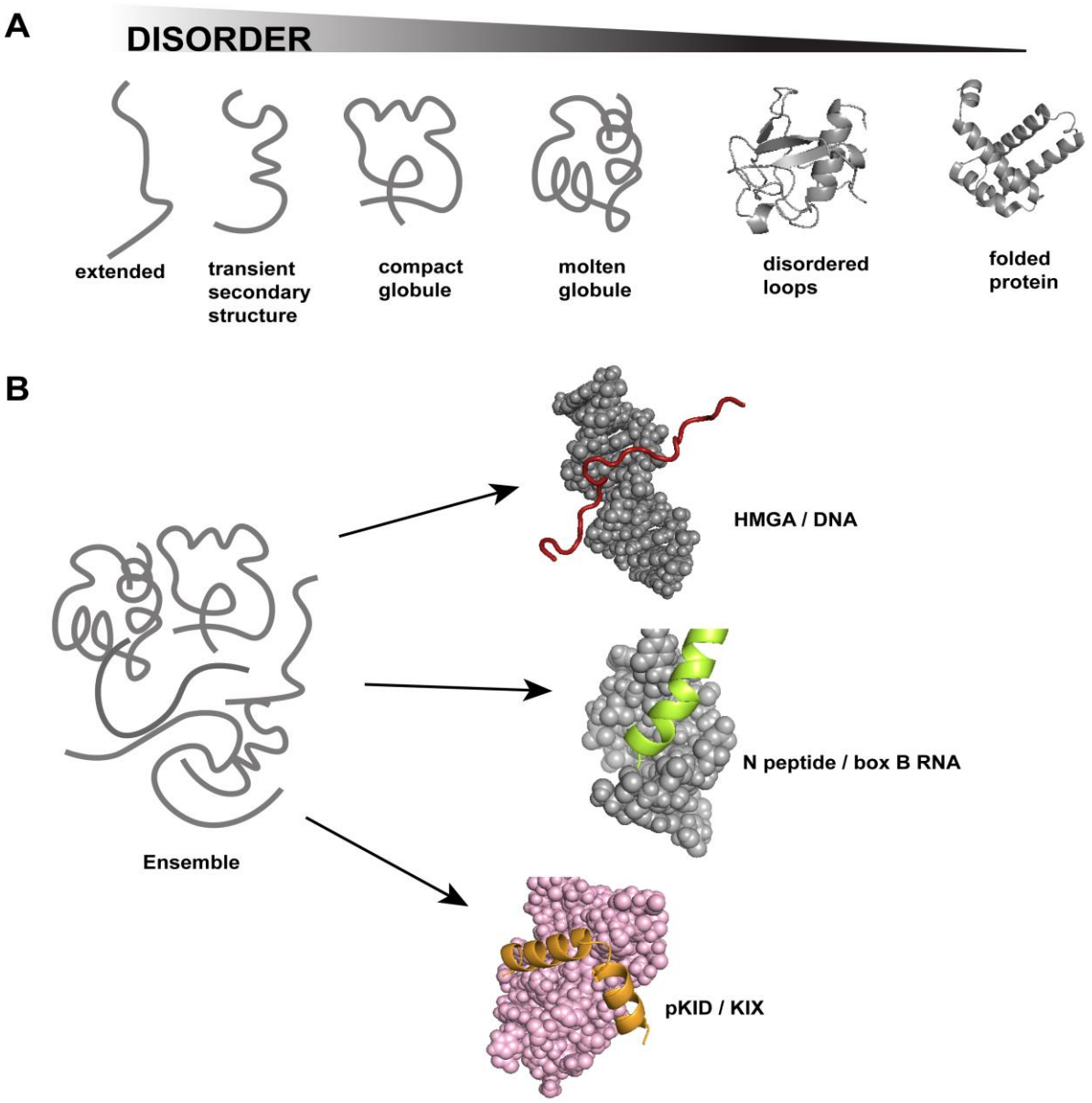

Figure 1.1 Coupled folding and binding in disordered proteins. Coupled folding and binding is a common process for intrinsically disordered proteins, in which they get ordered structure upon binding to their targets. A) Representation of continuum model of protein structure (65). Dynamically disordered states are shown as heavy lines. Cartoons represent folded states. IDPs or IDRs interconvert between these conformations when they are alone in the solution. Adapted from reference (3). B) IDPs or IDRs fluctuate over an ensemble of heterogeneous conformations (66). Binding enthalpy pays the entropic cost of folding into an ordered structure. Completely disordered HMGA (red) protein goes to disorder to order transition upon bind to its target DNA (gray) (63). The first 22 amino acids (green) of the bacteriophage P22 N protein is disordered and only folds when bound to boxB RNA (gray) (67). Phosphorylated kinase inducible domain (pKID, shown in yellow) of CREB protein folds on forming a complex with KID-binding domain (KIX, shown in pink) of CBP protein (68).

Recently, it has been shown that regulatory proteins may assemble into structured, higher-order signaling machines for signal transduction into cellular responses (69,70). Disordered assemblers play important roles in promoting higher-order protein complexes because of their reversible protein-protein interaction properties. FUS is a well-studied IDP that can form reversible higher-order assemblies for its cellular functions (71-74). Its disordered regions are N-term low complexity (LC) domain and RGG domains that are common to proteins involved in assembly formation. Studies revealed that FUS forms higher order assemblies by utilizing LC and RGG domains and the assembly formation is triggered by RNA binding (75). Higher-order structure of FUS is required for its transcription regulation activity. FUS assemblies bind to CTD (C-term domain) of RNA polymerase II and affect the ser2 phosphorylation and transcription. IDPs can also mediate formation of stable amyloid scaffolds such as RIP1/RIP3 necrosome (76). RHIMs (RIP homotypic interaction motifs) embedded within IDRs of RIP1 and RIP3 were shown to mediate heterodimeric amyloid structure formation that is a functional signaling complex mediating programmed necrosis.

Moreover, discovery of non-membrane bounded (membrane-less) bodies or organelles is one of the recent exciting advancements in IDP field (77-81). LC and prion-like disordered domains of IDPs mediate low-affinity, multivalent protein-protein interactions and form these phase-separated droplets in nucleus and cytoplasm. Membrane-less organelles are dynamic structures with liquid-like physical properties and comprised of heterogeneous mixtures of proteins and nucleic acids (81-83). Thus, they can easily adapt to alterations in the cellular environment by changing their composition and morphology. Nuclear bodies such as nuclear speckles, paraspeckles, PML (Promyelocytic leukemia) and cajal bodies are specialized in various aspects of gene regulation and RNA metabolism (84-87). On the other hand, cytoplasmic bodies such as P-bodies, germ granules, mRNP granules and stress granules are involved in specific roles in mRNA metabolism and homeostasis (88-90). Protein components of

the membrane-less organelles contain similar sets of folded and low complexity multivalent domains. Thus, they can promiscuously localize within more than one type of membrane-less organelles. On the other hand, each membrane-less organelle has specific RNA components suggesting crucial roles to RNAs in organelle formation. Studies with membrane-less organelle specific RNAs showed that disruption of transcription of these RNAs result in localization of the protein components in different cytoplasmic or nuclear granules (91,92). Since the IDPs are the main drivers of the membrane-less organelle formation, investigation the RNA binding properties of these proteins will provide a better understanding of membrane-less organelle biology.

1.1.1 RNA binding activity of intrinsically disordered proteins

RNA molecules are crucial partners of IDPs for most of their cellular functions including regulation of signaling pathways, assembly of macromolecular machines such as the ribosome, chromatin organization, regulation of RNA metabolism. Currently, there is heightened interest in understanding the RNA binding properties of IDPs since it will provide better explanations to the molecular mechanisms that are regulated or governed by IDPs.

Recent developments in comprehensive identification of RBP repertoire accelerated discovery of new RBPs. Interestingly, dozens of newly discovered RBPs in yeast cells do not have a classical RNA binding domain (93). In addition, protein-mRNA interactome studies in human cell lines identified hundreds of novel RBPs (5,6). According to these studies, 20% of both known and novel RBPs are highly disordered. Surprisingly, a recent study about the identification of RNA binding domains of RBPs revealed that nearly half of the RNA binding sites correspond to intrinsically disordered regions (5). Moreover, analysis of 1542 RNA binding proteins indicated that the second most common RNA binding domain in the human proteome is an intrinsically disordered domain, RGG (94). All of these studies contributed to the enthusiasm of exploring the RNA binding properties of intrinsically disordered RBPs.

Disordered RNA binding regions are enriched in positively (K and R) and negatively charged amino acids (D and E) as well as aromatic amino acid (Y). Positively charged amino acids and aromatic amino acids are crucial for the interaction with RNA molecules via polar, electrostatic interactions and hydrophobic stacking interactions. However, there is not evidence about the role of D/E amino acids in RNA recognition activity of RBPs. These regions may mimic the negative charge of RNA molecules that may prevent the interaction of other RBPs with RNA molecules during the regulation of RNA processing (95). The enriched amino acids form defined patterns in disordered regions. Until now, three different intrinsically disordered RNA binding domains were identified that are RGG (arginine-glycine rich regions), RS (arginine and serine) repeats (96) and KR (lysine-arginine) basic patches. Disordered RBDs usually exist in a modular manner as with the classical RBDs and provide flexibility to the RBPs (5). Also, they cooperate with other RBDs to achieve a more specific interaction with a better binding affinity.

The major contribution of intrinsically disordered regions to RBPs is flexibility since it facilitates RBPs to gain alternate conformations upon binding to RNA (97) (Figure 1.2). Thus, they can easily interact with different RNA molecules in various cellular processes. One of these processes is the RNA folding that needs RNA chaperones to get a minimum energy conformation. In many cases, functional regions of RNA chaperons consist of disordered residues (98). According to a proposed model for the role of disordered regions in RNA folding, disordered domains of the chaperone proteins undergoes disorder to order transition when they bind to RNAs thereby providing a thermodynamic advantage to the trapped RNA molecule to get a minimum energy conformation (99). Upon binding to the RNA, chaperone molecule

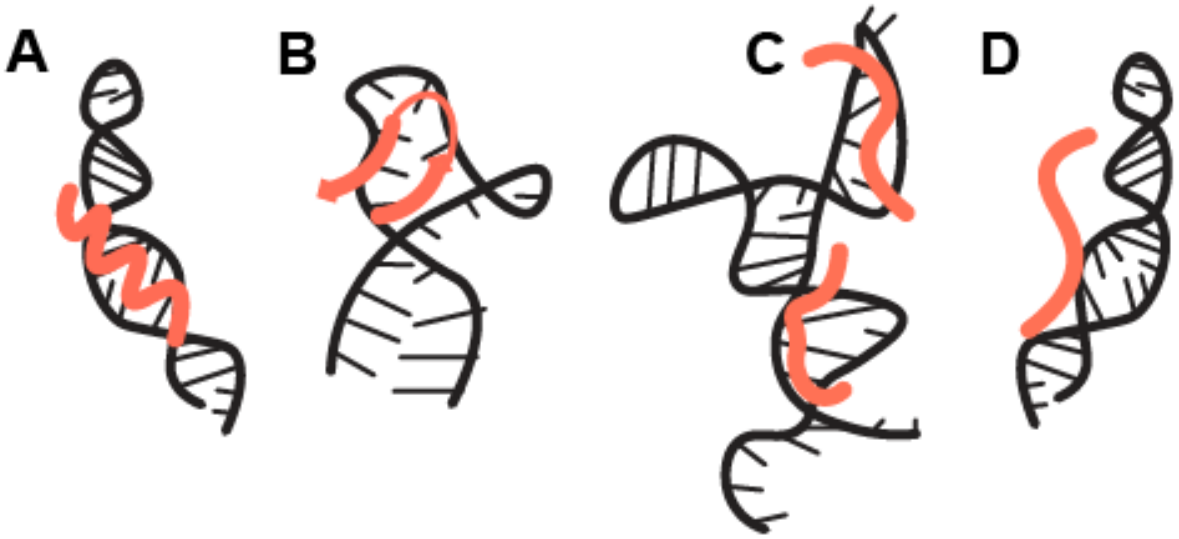

Figure 1.2 Flexible RNA recognition mode of intrinsically disordered RNA binding domains. Intrinsically disordered RNA binding domains can get folded upon binding to RNAs: A) α-helix (e.g.: p22 N-peptide and boxB RNA interaction (67)) B) β-hairpin (e.g.: BIV TAT and BIV TAR (100)) and C) β-turns (e.g.: FMRP-RGG and Sc1 RNA (101)). D) They can also keep the extended form by making interactions via one or two amino acids (e.g.: (102)) .

transfers its entropy to the RNA molecule and this makes the RNA folding process entropy driven. This proves the crucial roles of intrinsically disordered RBPs in the cell.

As will be discussed in the following sections, our knowledge about intrinsically disordered RBP/RNA interactions is limited. Identification of the RBP target sites is an important step toward understanding the mechanism of the cellular processes that they are involved in. Developments of high-throughput techniques such as microarray analysis, RNA immunoprecipitation followed by high-throughput sequencing (RIP-seq), crosslinking immunoprecipitation followed by high-throughput sequencing (CLIP-seq) and its variants revolutionized determination of RNA targets of RBPs (103,104). On the other hand, using only one method to identify the RNA targets is not an efficient way for intrinsically disordered RBPs. For example, multiple studies have attempted to characterize RNA binding sites of FMRP by several methods including cDNA-SELEX, RNAcompete and CLIP-seq methods (105-107).

However, all of them provided a different consensus sequence suggesting that RNA recognition mode of FMRP is completely promiscuous. In contrast to all previous work, a recent CLIP experiment in neurons suggests that FMRP is mostly associated with one unique mRNA (108). Differences between these results might be because of starting material, experimental construction as well as some technical biases. In addition, having high off-rates of binding, interacting with other RNA binding proteins, having less amount of aromatic amino acids (crucial for UV-crosslinking step of CLIP) are disadvantages of intrinsically disordered RBPs that make difficult getting a consistent data from different high throughput technics. Apparently, to overcome some of these problems, comparative analysis of data from different high-throughput techniques is necessary (109) and also these results can be confirmed by using simple *in vitro* technics such as electrophoretic mobility shift assay (EMSA) and foot-printing analysis of RNA-protein interactions.

1.1.1.1 Arginine/Serine Rich disordered RNA binding domain

Arginine/Serine rich domain was first identified in a study that characterizes sequence of Drosophila splicing factors, SWAP (suppressor of white apricot), Tra (transformer) and Tra-2 (transformer) (110-112). After the identification of human splicing factors, it was realized that these proteins (SR protein family) have a common domain structure consisting of one to two amino- terminal RNA recognition motifs (RRMs) and a carboxyl-terminal domain rich in serine and arginine (RS) dipeptide repeats (113). SR proteins are crucial for spliceosome assembly and involved in regulation of alternative splicing (114,115). Early studies suggested that RRM domain is responsible for specific interactions with pre-mRNAs and a variable-length RS domain functions as a protein interaction domain to assemble the spliceosome factors (115-117). Further studies revealed that RS domains directly interact with splicing signal sites of the RNAs at different steps of splicing reaction (118-120). In particular, the analysis of U2AF protein's RS domain provided detailed information about the RNA binding activity of RS domain (118).

Deletion of RS domain killed the RNA splicing activity of U2AF protein. Arginine amino acids substitution with lysines, shortening the RS domain length or substitution of RS region by heterologous RS domains did not severely affect the activity of U2AF protein and suggested that RNA interaction by RS domain is primarily mediated by electrostatic interactions. RS domains can also cooperate with RRM domains to enhance the RNA binding activity of the proteins. This phenomenon was shown in a study with Transformer β1 protein in which RS2 domain positively regulates the RNA binding activity of RRM domain (121). RS2 deletion of Transformer β1 displayed reduced binding affinity, but RS1 deletion did not affect the binding affinity. On the other hand, further studies are necessary to understand whether there is any difference between RNA recognition modes of RS domains, whether all of the RS domains can interact with the RNAs and whether they interact with all RNAs in a non-specific manner.

A genome wide analysis of the RS domain containing proteins in metazoans identified many new SR-related proteins that have an RS domain but not an RRM domain (122). Moreover, this study showed that proteins with RS domains are involved in many cellular processes in addition to splicing, including chromatin structure remodeling, transcription, cell cycle, cell structure and ion homeostasis. For example, Apoptotic Chromatin Condensation Inducer in the Nuclear (Acinus) is one of identified SR-related proteins and is required for chromatin condensation during apoptosis (123). Later studies implicated that Acinus is also involved in gene expression, pre-mRNA processing and is a component of nuclear paraspeckles (124). Indeed, all of the SR proteins are localized in nuclear paraspeckles that is the nuclear storage site of pre-mRNA splicing factors (87). Thus, targeting the nuclear speckles contribute to the identification of novel SR and SR-related proteins. NF-kappa-B activating protein (NKAP) is a newly identified SR-related protein by this way (125). Its role in activation the NF-kappa-B pathway and regulation the Notch signaling pathway and T-cell development was already known (126). After the identification of its involvement in pre-mRNA splicing, it was shown that NKAP binds RNA molecules through it RS domain (125). In addition, transcriptome-

wide analysis of NKAP binding to RNA revealed that it interacts with diverse primary RNA molecules including pre-mRNA and different types of non-coding RNAs.

Splicing activity of SR proteins is regulated by phosphorylation and dephosphorylation of serines at the RS domain since phosphorylation is prerequisite for some of the protein-protein interactions in splicing (127,128). In addition, these modifications also affect the RNA binding activity of RS domains and regulate splice site selection and mRNA transport (129,130). According to these studies, both phosphorylation and dephosphorylation enable to enhance the RNA binding activity of proteins. Further studies are necessary to identify the affect of PTMs on RNA binding activity of different RS domains.

1.1.1.2 RNA binding activity of basic patches

Basic patches consist of lysine (K) or arginine (A) repeats that form positively charged regions at the surface of RBPs (131). These regions are abundant especially in newly discovered RBPs (5,132). Basic patches usually exist as flanking regions near the globular domains suggesting a functional cooperativity with these domains (5).

Lysine amino acid repeats was first identified in transcription factors. Early information about their functional role in the protein activity comes from the homeodomain transcription factors in which they engage in nonspecific interactions with DNA since their positive charge and accelerate the search for a specific target site on DNA (133,134). mRNA interactome data identified 47 proteins that contain a lysine patch, but not any other known RBDs (5,132). This emphasizes the potential role of lysine rich regions in RNA binding activity of RBPs. For example, PAPD5 is a member of the family of noncanonical poly(A) polymerases in human cells and involved in degradation of aberrant rRNAs by polyadenylating them (135). However, It does not have an identified canonical RBD. Studies with PAPD5 protein revealed that a C-terminal K patch mediates the RNA binding activity (135). In addition, expression of eGFP fused K patch domains from SDAD1 and MECP2 indicated that these domains are capable of RNA binding in

mammalian cell lines (132,136). Thinking the role of K repeats in DNA binding activity of transcription factors, they may also contribute to RNA binding activity of RBPs in a nonspecific manner. Further studies are necessary to understand the role of K patches in RNA binding.

Arginine rich motifs (ARM) are better characterized than lysine patches. Indeed, studies to determine ARM peptide-RNA interactions are the basis for our current understanding of induced fit binding model in RNA-protein interactions (137). Characterized ARM peptide-RNA complexes have revealed that ARMs are generally unfolded and can adopt a variety of conformations when bound to RNA. For example, bovine immunodeficiency virus (BIV) TAT protein (trans-activator of transcription, TAT) has an ARM domain that is intrinsically disordered (138). TAT proteins promote the viral gene expression upon binding to 5`-end of the nascent viral RNAs (trans activation response element, TAR). When the ARM peptide of BIV TAT binds to TAR RNA, ARM forms a β-hairpin conformation that promotes high affinity binding through a large sets of specific contacts to the TAR RNA (Figure 1.3A, B) (138,139). Sometimes the conformational change occurs only at the RNA as in HIV TAT-TAR interaction (138,140-143). HIV TAT ARM binding to TAR RNA induces stacking interaction between two helical stems and formation of a base triple. However, ARM remains in an extended conformation and only a single arginine amino acid is required for specific interaction with TAR. BIV TAR and HIV TAR RNAs have similar secondary and tertiary structures, but BIV TAT ARM does not bind to HIV TAR RNA with a high affinity because of the differences between the regions flanking the peptide binding site of these two RNAs (Figure 1.3A) (144). ARM domain of phage λ is another example that exhibits specific RNA binding activity. It binds its own boxB RNA with a 16 fold higher binding affinity than the phage P22 boxB RNA (145). On the other hand, some of the ARM domains have a potential to interact with different RNA molecules since their flexible structures. For example, studies with Jembrana disease virus (JDV) showed that it can bind to its own TAR and BIV TAR via a β-hairpin structure, but it binds to HIV TAR with an extended conformation suggesting that for an efficient binding, an ARM domain can adopt different

conformations depending on the RNA architectures (Figure 1.3A, B) (146). A detailed analysis of different ARMs and their cognate RNAs indicated that ARMs specifically bind to their own targets, however, some cross-recognitions can be also observed with lower binding affinities (147). These examples are inspiring about how other disordered domains can contribute to RNA binding activity of the proteins.

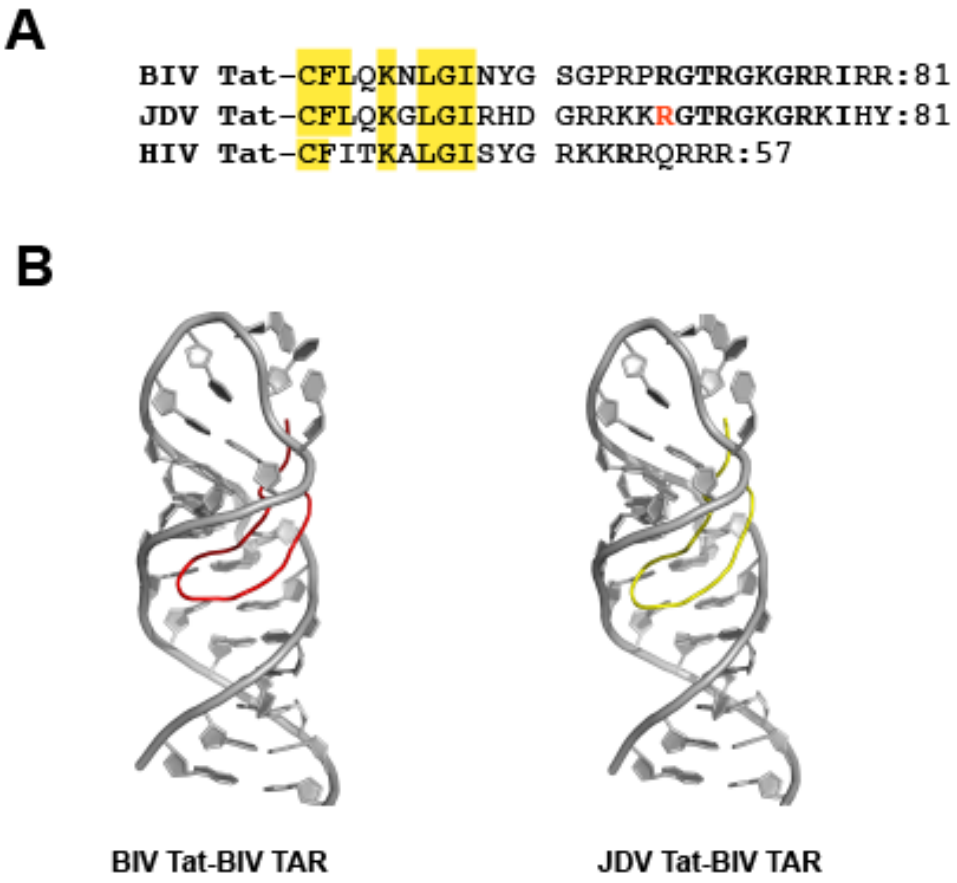

Figure 1.3 Comparison of Tat ARM domains and their RNA recognition modes.
A) Alignment of BIV Tat, JDV Tat and HIV Tat peptide sequences. Amino acids involved in specific RNA recognition are shown in bold. JDV Tat amino acids shown in bold are involved in both for JDV TAR and BIV TAR RNA recognition. Red colored arginine of JDV Tat indicates the only amino acid required for interaction with HIV TAR RNA. B) Resemblance between the solution structures of BIV TAR hairpin interaction with BIV-Tat (shown in red, PDB 1MNB) and JDV Tat (shown in yellow, PDB 1ZBN) ARM peptides.

1.1.1.3 Arginine/Glycine rich (RGG/RG) RNA binding domain

Another intrinsically disordered RNA binding domain with arginine amino acids is RGG/RG domain that consists of repeats of R/G/G (arginine/glycine/glycine) and R/G (arginine/glycine) sequences. Arginine is a positively charged amino acid that can mediate hydrogen bonding, π-stackin and electrostatic interactions with proteins and nucleic acids. There are more than 1000 human proteins with RGG/RG motifs that are involved in a various cellular processes such as transcription, translation, snRNP biogenesis, apoptosis and DNA damage signaling (148). In addition, RGG/RG domains are the second most common RNA binding domain in human proteome as discussed in above (94).

Comparative analysis of RGG/RG domains from different proteins revealed that majority of the sequences consists of glycine amino acids (50%-60%) (149). On the other hand, basic residues (R, H, K) constitute 20%-30% of the total amino acids in RGG/RG domains. These RGG/RG repeats are interspersed with other residues including aromatic amino acids that are crucial to make hydrophobic stacking interactions with RNAs. A recent study classified RGG/RG domains according to repeat numbers of RGG (di-RGG and tri-RGG motifs) or RG (di-RG and tri-RG motifs) sequences (148). They determined tens to thousands of protein members in human genome for each of these motifs. RNA-binding capability is the most enriched function in the Gene Ontology annotation of the proteins in these classes. Nonetheless, any relation between the different motifs and protein functions could not be identified.

RGG/RG domains are one of the most common RNA binding domains in heterogeneous nuclear ribonucleoprotein (hnRNP) family whose members are involved in different aspects of nucleic acid metabolism including alternative splicing, mRNA stabilization, regulation of transcription and translation (150). Thus, it is not surprising that RGG/RG domain was first identified in a hnRNP family member, hnRNP U (151). hnRNP U (also known as scaffold attachment factor A, SAF-A) is an abundant nuclear protein that is able to bind pre-mRNA and ssDNA. In addition to its role in RNA metabolism, hnRNP U is also involved in Xist mediated

transcription silencing, DNA damage response and chromatin organization (152-154). hnRNP U usually functions with different types of noncoding RNAs, including small nuclear noncoding RNAs (snRNAs) involved in RNA splicing, lncRNAs such as Xist for X chromosome inactivation, and chromosome associated RNAs (caRNAs) for the regulation of interphase chromosome structures. RGG/RG domain is the only known RNA binding domain of hnRNP U that was shown to mediate RNA binding activity of the protein (151,152,154). Our knowledge about the RNA binding properties of RGG/RG domain of hnRNP U is limited with early studies suggesting that it binds to G-rich RNA sequences in a nonspecific manner (151).

RNA binding mode of RGG/RG domain of fragile X mental retardation (FMRP) protein is well characterized. FMRP is primarily involved in translational repression. It binds to ~4% of mRNAs in the brain (155) and loss of function in this protein results in Fragile X Syndrome. FMRP also has two different KH domains that contribute to the RNA binding activity. RNA substrates of KH1 are unknown, but KH2 targets the RNAs with double stem loops (156). The RGG/RG domain of FMRP is required for binding to G-rich sequences capable of forming G-quadruplexes. *In vitro* selection experiment with RGG/RG domain of FMRP identified an aptamer with G-quadruplex structure (Sc1 RNA) (157). Structural analysis of RGG/RG and Sc1 RNA interaction revealed that RGG/RG domain is unfolded in unbound state and adopts a turn fold upon binding to Sc1 RNA (101,158). RGG/RG domain was shown to make base-specific interactions with two top bases of Sc1 duplex through its arginine amino acids and two more interactions with the duplex region through the glycine amino acids (159) (Figure 1.4). Further studies showed that arginine amino acids involved in RNA recognition could be different for different RNA targets (160) supporting the flexible interaction mode of RGG/RG domain. Although the *in vitro* evidences suggest that RGG/RG domain of FMRP can mediate binding to G-rich RNA molecules via base-specific interactions, *in vivo* analysis of RNA targets of FMRP did not exhibit enrichment of RNA targets with G-rich secondary structure (106,161). Thus, it is

believed that RGG/RG domain of FMRP just contributes to RNA binding activity of KH domains to achieve an efficient binding.

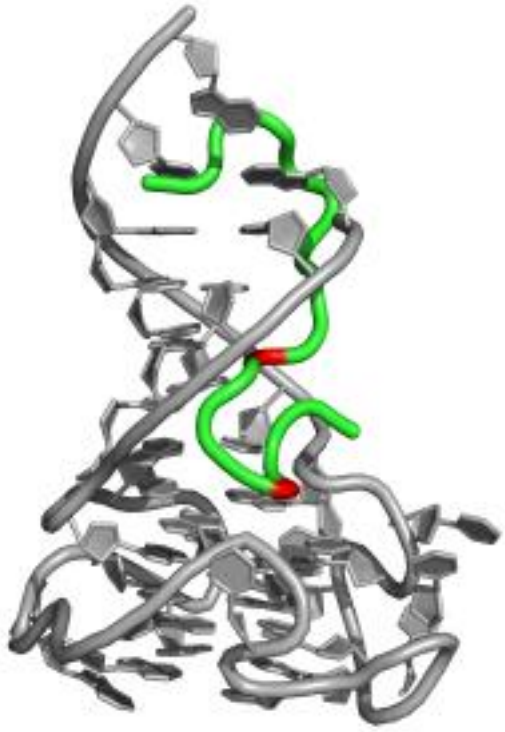

Figure 1.4 Solution structure of the RGG/RG peptide (green) from FMRP bound to in vitro selected Sc1 RNA (gray) (PDB 2LA5). Arginine residues that make hydrogen bonds with nucleotides were shown in red.

Arginine residues of RGG/RG domains are substrate recognition sites for certain protein arginine methyl transferases (PRMTs) (162,163). Arginine methylation of RGG domains was shown to affect protein-protein and protein-RNA interactions. For example, assembly of small nuclear ribonucleoproteins (snRNP) that are the subunits of spliceosome complex is regulated by arginine methylation. Survival of motor neuron complex (SMN) assemble Sm proteins onto snRNPs to produce an active snRNP complex in the cytoplasm. Methylation of arginines at C-terminal RGG/RG domains of Sm proteins facilitates SMN-Sm interaction. Addition of methyl groups can sterically hinder association of arginines with RNAs or remove hydrogens from amino groups that are involved in binding to RNAs. Studies with FMRP and its target RNAs

(Sc1 and AATYK) revealed that arginine methylation of RGG/RG domain inhibits FMRP binding to Sc1, but not to AATYK RNA (160). This suggests that arginine methylation may regulate the interactions of the protein with a group of RNAs. However, further studies are necessary to understand to what extent methylation is involved in the regulation of RNA binding activity of the proteins *in vivo*. It is not clear whether methylation of specific arginine amino acids can regulate the specificity of the protein or whether methylation can create an indirect effect on RNA binding activity of the protein by affecting the interaction with other proteins.

1.1.2 Intrinsically disordered RNA binding proteins and human diseases

Intrinsically disordered RNA-binding proteins are key regulators of the gene expression and homeostasis (10,131) and found within a number of dynamic ribonucleoprotein (RNP) complexes (Figure 1.5). Before an mRNA can be translated, the nascent transcript experiences a series of processing steps. These transcripts are capped at their 5'-end, cleaved and polyadenylated at the 3'-end, spliced, exported to cytoplasm for translation, storage, localization or degradation (164,165). Intrinsically disordered RBPs are pivotal in regulation and coupling of these processes since they can function in more than one step of mRNA maturation. Intrisically disorded RBPs are also involved in regulating the metabolism of the noncoding RNAs by directing and regulating the transcription and post-transcriptional fate of the RNAs in the nucleus and cytoplasm (Figure 1.5). Thus, mutations within intrinsically disordered RBPs or in their RNA targets can severely affect RNP genesis and result in sporadic or hereditary genetic disorders. In particular cancer mutations in these proteins are associated with neurological diseases (amyothropic lateral sclerosis (ALS) and frontotemporal lobar degeneration (FTLD)) and muscular atrophies (7,131).

An analysis of human mRNA-protein interactomes revealed that most disease related mutations are found in non-classical RBDs and intrinsically disordered RBDs (7). These studies indicated that the most frequently mutated amino acids in disordered regions are arginine and

glycine that define the RGG (arginine-glycine rich) domain. One of the extensively studied disease related intrinsically disordered RBPs is Tar DNA binding protein 43 kDa (TDP-43) that is involved in splicing, translation of mRNAs, microRNA processing, apoptosis, cell division as well as stress granule formation (Figure 1.5) (166). It is the major aggregating protein in ALS and FTLD that leads to motor neuron disease and frontotemporal dementia, respectively. TDP-43 comprises two classic RNA-recognition motifs (RRMs) and a glycine-rich C-terminal region that allows it to bind single stranded DNA, RNA and protein. Most of the pathogenic mutations are localized at Glycine-rich domain of TDP-43. Recent studies in *Drosophila* reveal that neurotoxicity is related with RNA binding and/or intracellular mislocalization of TDP-43 (167,168). Moreover, it was shown that one of ALS-related mutations, M337V, within the glycine-rich domain decreases the binding affinity of protein to G-quadruplex RNAs, a putative target of this protein (169).

FUS (Fused in sarcoma), EWS (Ewing sarcoma) and TAF 15 (TATA binding associated factor 15) belong to the FET family of DNA and RNA binding proteins that are known to contain point mutations in their intrinsically disordered RGG/RG domains that are linked to ALS and FTLD. These proteins can interact with thousands of transcripts and affect different steps of RNA metabolism (Figure 1.5) (170). An important function of FET proteins is their participation in the transcription pre-initiation complex. For example, FUS binds the C-terminal domain of RNA pol II in an RNA dependent manner to prevent its inappropriate hyperphosphorylation (171). Moreover, FUS was found to regulate the transcription of *CCND1* by affecting histone acetylation (172). These activities of FUS are dependent on the expression of noncoding RNAs. To date, more than 30 different mutations in RGG/RG domains of FUS detected that are thought to be involved in ALS and FTLD, underscore the physiological importance these domains. In this thesis work, I show that RGG/RG domains of FUS mediate the RNA binding activity of the protein. This work will serve as a foundation for future studies to understand

whether the disease-associated mutations in RGG domains of FUS disrupt the RNA binding activity of the protein.

Most of the intrinsically disordered RNA binding proteins is expressed ubiquitously in a wide range of tissues (94). However, mutations in these proteins affect particular cell types such as motor neurons in ALS disease. This suggests the contributions of cell type specific factors to disease formation such as post-transcriptional gene regulation pathways. To understand the role of intrinsically disordered RBPs in human diseases, their post-transcriptional regulation mechanism, protein-protein interactions and protein-RNA interactions should be well characterized on both molecular and cellular scales. Then, a cell specific target can be identified to develop treatment methods.

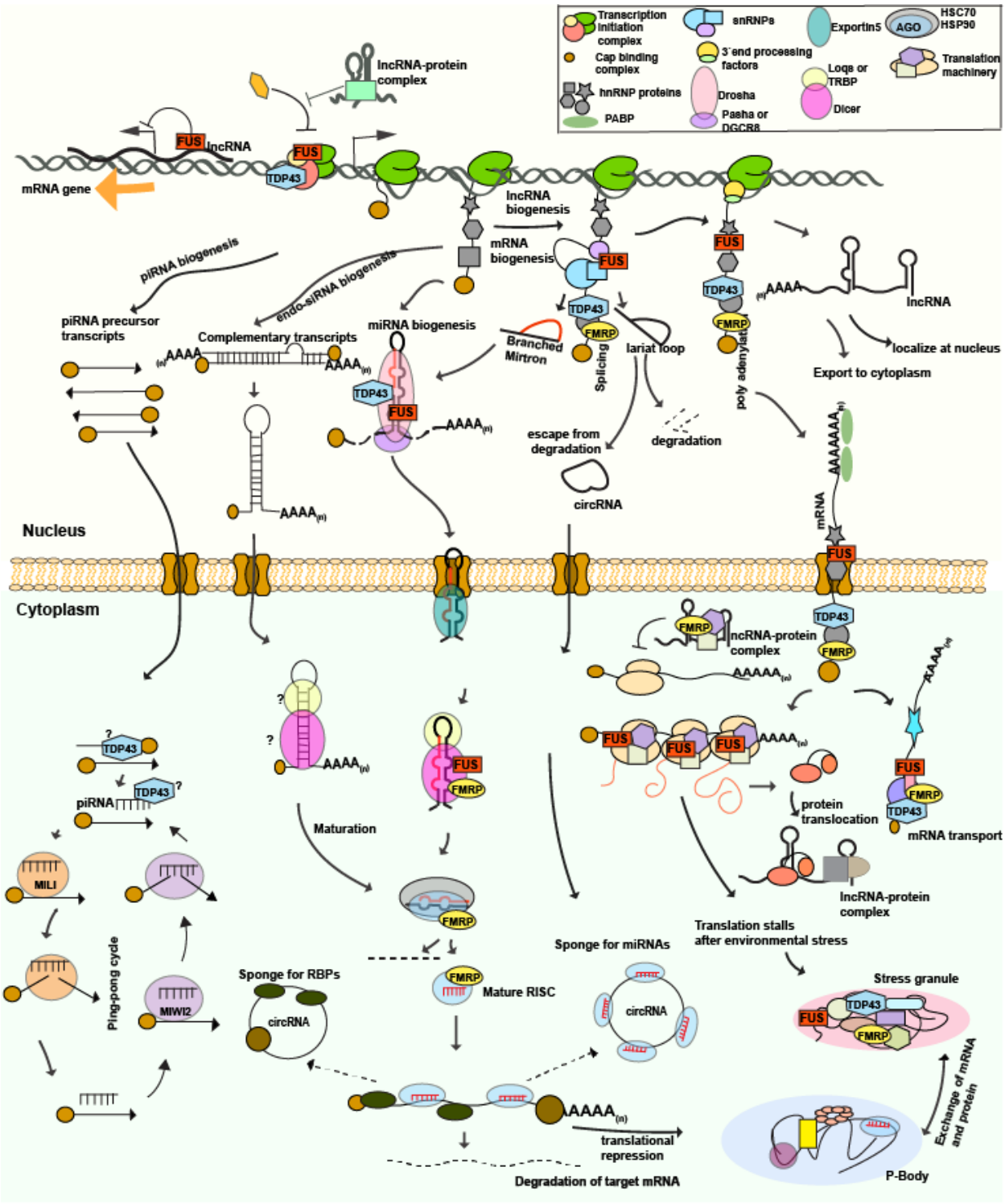

Figure 1.5 Summary of the regulatory non-coding RNAs/mRNA metabolisms and dynamic nature of RNPs in regulation of cellular processes. Figure legend is at the next page.

Figure 1.5. Summary of the regulatory non-coding RNAs/mRNA metabolisms and dynamic nature of RNPs in regulation of cellular processes. Biogenesis of small regulatory non-coding RNAs (miRNA, piRNA and endo-siRNA), long non-coding RNAs (lncRNA, circRNA) and mRNAs were shown. RNA binding proteins have primary functions in regulation of the all steps of RNA metabolism. These proteins bind nascent RNA and form Ribonucleoprotein (RNP) complexes. RNA biogenesis is a highly dynamic process since the protein members of the RNPs change at the every step. RNPs are also formed with the mature regulatory non-coding RNAs to regulate the different stages of the gene expression. piRNAs form RNPs with Piwi proteins and cleave the transposon RNAs, leading to silencing (173-175). miRNAs direct Argonaute proteins (RISC complex) to repress more than half of all human mRNAs (176-178). Endo-siRNAs are derived from long hairpins or complementary transcripts and join to the miRNA pathway after couple of processes (179-181). circRNAs are generally formed by alternative splicing of mRNAs and function as RBP or miRNA sponges(182-184). lncRNAs have similar posttranscriptional processing steps with the mRNAs (5`capping, splicing, polyadenylation), but depending on the sequence/structure of the RNA or the protein partners of the RNP, they may localize in the nucleus or export to the cytoplasm (185-190). lncRNPs can regulate dosage compensation, transcription and translation through several mechanisms (only couple examples were shown for simplicity) (172,191-197). FUS, FMRP and TDP43 proteins were shown as examples to the disease related RBPs. They are involved in all steps of gene expression. They also form RNPs with noncoding RNAs to regulate the cellular processes (172,198-200).

1.2 Sequence specific, non-specific RNA binding modes and possible binding modes between them

RNA-protein interactions are critical for regulation of all steps of gene expression. In a typical mammalian cell, there are at least 1000 proteins that directly interact with RNA molecules; the number of RNA species is much higher. Human cells encode more than 20,000 mRNAs whose diversity substantially increases by alternative splicing and chemical modifications. In addition, there are thousands of different long noncoding RNAs, microRNAs, small nucleolar RNAs and tRNAs that range in length from 22 nucleotides to 10,000 nucleotides. Each of these RNAs interacts with multiple RNA binding proteins during its lifetime. Likewise, a significant number of RNA binding proteins interact with many different RNA molecules. Together, this suggests that the number of possible of combinatorial RNA-protein interactions is extremely large. From this perspective, assembling specific RNPs is a complex but crucial process that is poorly understood.

RNA binding modes of proteins are traditionally classified as either sequence specific and non-specific. Non-specific (or more recently, “promiscuous”) binders associate with RNA sites without preference for a distinct nucleotide sequence. Almost half of the known RNA binding proteins is the “non-specific” and function in nearly all pathways that are related with RNA processing (201-203). In this binding mode, RNA binding is mostly mediated by positively charged amino acid side chains that contact the negatively charged phosphate groups of the RNA backbone (204). Non-specific binding to diverse RNA molecules is essential for the cellular function of some RNA binding proteins. There are many structural studies about the RNA binding proteins that recognize their RNA targets in a sequence-independent manner. A good example of non-specific binding is revealed by the crystal structure of human Argonaute-2 (hAgo2) protein, which is involved in RNA-directed gene silencing, bound to microRNA-20a. This structure revealed that the RNA interacts with all four domains of hAgo2 and also with the linkers connecting these domains (205). The 5’-end of microRNA is tethered to hAgo2 through a number of interactions to form a tight binding pocket composed of residues from Mid and PIWI RNA binding domains. These residues make salt bridges and hydrogen bonds with U1 (the first nucleotide of the microRNA). The following seven nucleotides (5’-AAAGUGCU-3’) that is located in the narrow groove form the seed sequence and interact with main and side chains of amino acids through their backbone phosphates and sugars. These base independent interactions allow hAgo2 protein to associate many small RNA sequences that is vital for its function. However, it should be noted that while hAgo is non-specific for RNA sequence, it is highly selective for length, enabling it to efficiently discriminate between microRNAs and other RNAs in the cell.

Traditionally, specific binding proteins are considered to interact with RNAs through a defined RNA sequence or structural motifs, or a combination of them. This binding mode is generally achieved by shape complementation and hydrogen bond interactions. However, a single RBD usually is not sufficient to achieve specificity for a small set or a single RNA since it

interacts with only 3 to 8 nucleotides and often tolerates sequence variation in these binding sites (206,207). Nature solved this issue by several strategies. First, most of the RNA binding proteins evolved multiple distinct RBDs tethered together to form a much larger binding interface that recognizes a longer sequence to attain sequence-specific binding mode (208). For example, the RRM (RNA recognition motif) is the most widely found RNA binding domain that usually is found as multiple repeats in a protein (209). This class of proteins has a βαββαβ fold that forms a four-stranded-sheet packed against two alpha helices. The first (RNP2) and third β-strands (RNP1) contain characteristic aromatic residues, which form the primary RNA binding surfaces. These regions recognize specific RNA sequences through base stacking, hydrophobic, polar and electrostatic interactions. Three subclasses of RRM using different mechanisms to recognize the RNAs were identified (210). The subclass of quasi-RRMs (qRRM) uses only the loops to interact with the RNAs due to the lack of conserved aromatic amino acids in RNP1 and RNP2. hnRNP F has three qRRM structures that have classical compact RRM fold with three highly conserved loops (211). These tandem qRRM domains are thought to provide hnNRP F sequence specific binding to G-rich single stranded RNA. The observation that three qRRM structures of hnRNP F bound with 5'-AGGGAU-3' RNA indicated that the three guanines of the RNA adopt a compact conformation and three conserved residues from the loops surround the G-tract and stack with each guanine base. The second strategy to achieve the specific interaction is "dimerization" that presents two recognition sites for RNA binding. This strategy can also strengthen the binding affinity by cooperative interactions (208). For example, protein-protein interactions between N-terminal RRM domains of U1A protein only occur in the presence of RNA since the homo dimerization elements is not available when the U1A is free. This interaction forms basis for the cooperative binding and provides high binding affinity (212). Finally, RBDs from different proteins can cooperate to interact with the specific sites of the RNA. During initials steps in spliceosome assembly, U2AF65 recognizes the polyprimidine tract in the pre-mRNA by using two RRM domains. A third non-canonical RRM domain of U2AF65 interacts

with SF1 to further strengthen the binding of U2AF65 to the RNA (213). In addition to these things, local concentrations of protein and the specific RNA targets, competition with other RNAs to interact with the protein and the competition with other proteins to interact with the RNAs also affects the specific interactions between RNA and proteins.

Current studies are increasingly revealing that this binary classification of RNA-protein interactions is not sufficient to define the spectrum of types of interactions. As discussed in previous sections, intrinsically disordered RBDs interact with RNAs through both sequence and structure motifs. However, since these domains are highly flexible, they can recognize different sequences and/or structures that are common in biological RNAs. In addition, analysis of RNA-protein interaction at the transcriptome level revealed that most of the RBPs interact with RNAs containing different sequence and structure properties (214,215). In this manner, these interaction types do not appear to completely fit into the non-specific and the specific binding types. Thus, a new classification approach that considers the RBPs with a binding mode between specific and non-specific is necessary. Measurement of RNA binding affinity of the proteins by more quantitative and high throughput methods such as RNA compete (216) and RNA Bind-n-Seq (217) is crucial for a better determination of the sequence and structure specificity profile for each RBP. This kind of measurements will allow us to create complete RNA affinity distributions of RBPs on a diagram. Then, specificity of a given RBP can be quantified and interpreted comprehensively. This method probably will provide a spectrum of selectivities rather than a discrete set of specificities.

1.3 Summary of the thesis

Current understanding about the RNA recognition mode of RGG/RG domains mostly depend on the studies with hnRNP U and FMRP proteins. It was thought that these domains primarily make nonspecific interactions with G-rich RNA molecules. In addition, RGG/RG domains usually exist with classical RNA binding domains (94). Thus, even a RGG/RG domain

can make specific interactions with a group of RNAs, they are never thought as a primary RNA binding domain of a protein. Instead, they are accepted as an auxiliary domain that contributes to RNA binding activity of other RNA binding domains. To characterize the RNA binding properties of RGG domains, I expressed the RGG/RG domains of Fused in Sarcoma (FUS), Fragile X mental retardation (FMRP), and Heterogeneous nuclear ribonucleoprotein U (hnRNP U) proteins and measured their RNA binding affinities for a spectrum of different RNAs with well-defined structural and sequence features. I also investigated the role of RGG/RG domains of FUS protein in RNA binding activity of the full-length protein. This thesis work exhibited that RGG/RG domains could mediate the RNA binding activity of a protein and their RNA recognition mode is not completely nonspecific.

Chapter 2 : Materials and Methods

2.1 Protein Expression and Purification

All protein constructs were cloned into pET30b vector system with N-terminal His6-MBP tag and transformed into BL21 (DE3) Rosetta *E. coli* cells. 10 mL LB-Kanamycin bacterial culture was grown overnight, and then inoculated into 1 L LB medium. Cultures were incubated at 37°C until OD_{600} reached ~ 0.6. Protein expression was induced by 0.5 mM IPTG and 100 µM $ZnCl_2$ was added to the cultures if the expressed protein included the ZnF domain. The cultures were incubated at 20°C overnight. Bacterial cells were pelleted at 1,500 g and resuspended in lysis buffer (1 M KCl, 1 M Urea, 50 mM Tris-HCl pH 7.4, 10 mM imidazole, with or without 100 µM $ZnCl_2$). Cells were lysed using an Emulsiflex C3 homogenizer. The cell lysates were clarified by centrifugation at 17,000 g for 30 min and the supernatants were incubated with Ni-NTA sepharose beads on an orbital shaker for 1 hour at 4 ºC. Beads were centrifuged at 300 g for 2 min and washed 3 times in lysis buffer and once in lysis buffer supplemented with 100 mM Imidazole. Proteins were eluted in lysis buffer supplemented with 250 mM Imidazole. MBP tag was not cleaved after the purification since it keeps proteins soluble. No RNA binding activity was observed for MBP protein (data not shown). Proteins were dialyzed into FPLC buffer (1 M KCl, 1M Urea, 50mM Tris-HCl, 2 mM DTT, with or without 100 µM $ZnCl_2$) with appropriate molecular weight cutoff dialysis tubing. Size exclusion chromatography was performed using Hiload 16/600 Superdex 200 or Hiload 16/600 Superdex G-75 prep grade columns (GE Life Sciences) (Figure 2.1). All proteins purified as monomers, based on comparison to size standards. Final protein concentration was calculated using molar extinction coefficient as determined using Expasy-Protparam tool and the absorbance at 280 nm. Proteins were stored at 4°C for two weeks.

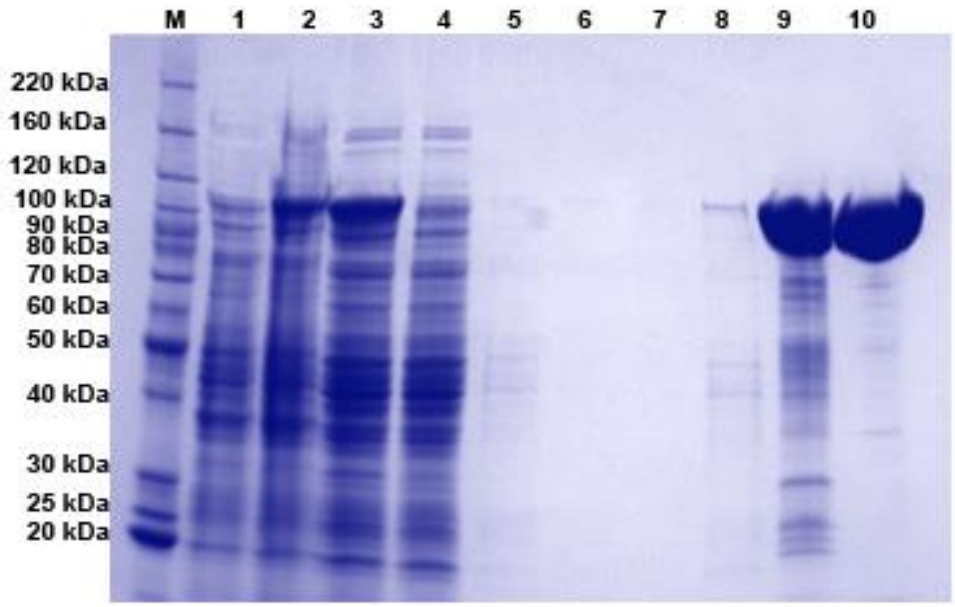

Figure 2.1 SDS-PAGE analysis of protein expression and purification. Representative SDS-PAGE image of FUS protein purification. FUS was expressed and purified as described in methods. M: protein marker (Invitrogen), lane 1: non-induced (BL21 (DE3) Rosetta Plys) cells, lane 2: IPTG-Induced cells, lane 3: supernatant, lane 4: flow through, lane 5, 6 and 7 are washes of the Ni-NTA with lysis buffer, lane 8: wash of the Ni-NTA with 100 mM imidazole in lysis buffer, lane 9: Elution of FUS protein with 250 mM imidazole in lysis buffer, lane 10: purified FUS protein by size exclusion chromatography.

2.2 *In vitro* transcription and purification of RNAs

DNA template for RNA transcription was amplified by using PCR and transcribed by T7 RNA polymerase (218). For the short templates such as Sc1 (36 nucleotide), forward and reverse strands were annealed instead of amplifying by PCR as follows: 1 µl forward strand (1 µM), 1µl reverse strand (1 µM), 10 µL NaCl (1 M) and 88 µl 0.5 x TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA) are combined in a PCR tube. Then, the tubes are heated to 95ºC and cooled down to 16ºC -1ºC/30 sec. For 12.5 ml transcription reaction, 7.5 ml ddH_2O, 1.25 ml transcription buffer (10x), 400 µl $MgCl_2$ (1 M), 500 ul from each rNTPs (100 mM), 800 µl PCR template, 125 µl DTT (1 M), 100 µl inorganic pyrophosphatase (20 U/µl), 200 µl T7 RNA polymerase were assembled in a 50 ml canonical. Reaction was vortexed and incubated at 37ºC for 2 h. For ethanol precipitation, 37.5 ml ethanol was added to the reaction and incubated at -80ºC for at least 30 min or at -20ºC for at least 2 h. Then, the reaction was centrifuged at

4,000 g and at 4ºC for 15 min. The supernatant was discarded and pellet was left for air-drying to get rid of the ethanol at room temperature. The pellet was suspended in 2 ml of 8 M urea, 500 µL 0.5 M EDTA pH 8.0, and 1 mL of formamide load dye. To re-suspend all of the precipitates, tube was vortexed vigorously. To ensure the complete denaturation of the RNA, samples were heated at 65ºC for 5 min and vortexed vigorously (this was repeated couple of times until getting a clear solution. If heating does not help to re-suspend all of the precipitates, more EDTA was added and the samples were vortexed again).

Transcripts were purified using denaturing polyacrylamide gel (6-10% 29:1 acrylamide/bisacrylamide, 1x TBE buffer (0.1 M Tris base, 0.8M boric acid, 1 mM Na_2EDTA), 8 M urea). Gel dimensions that we used depend on the volume of the initial transcription reaction. For a 12.5 ml reaction volume, a 35 cm wide, 23 cm long and 3 mm thick gel works very well. Gel was run with a constant power 30 W until the desired RNA migrated approximately 80% down the gel. This may take 6-7 hours. For a faster run, buffer (1x TBE) exchange can be done for the samples before loading. RNA bands were visualized by putting the gel on a fluorescence TLC plate and shadowing the RNA with short-wave UV in a dark room. Full-length transcripts were excised from the gel and the gel pieces were further crushed into small pieces in a canonical by using a 5 ml serological pipette. 0.5x TE buffer was added into the canonical and left for elution by rotating at 4 ºC for 2 h. RNA was concentrated using centrifugal concentrators with a 10 kDa molecular weight cutoff (Amicon Ultra, 0.5 ml) and buffer (05x TE) exchange was performed by the same method. Final RNA concentration was calculated using the absorbance at 260 nm and the molar extinction coefficient as determined using an extinction coefficient calculator that calculates the extinction coefficients by summing of the individual extinction coefficients for each nucleotide in the RNA. The RNA was stored at -20°C until use.

2.3 RNA Body Labeling Reaction

Consensus sequences of human and mouse RNA repeat domains (RRD; 152 nt and 155 nt respectively) of Firre lncRNA were generated and cloned into pUC19 vector. DNA templates for *in vitro* transcription reactions were amplified from RRD containing plasmids using PCR. 100 µl *in vitro* RNA transcription reaction was prepared with an adenine ribonucleotide concentration that is 10-fold lower than the standard reaction concentration. This facilitates the incorporation of more radioactive ATP into the transcript. 20 µCi ATP [α-32P] was added and the reaction was carried out with T7 RNA polymerase at 37°C for 2 h. MicroSpin G25 columns were used to remove unincorporated nucleotides from the labeling reactions. Labeled transcripts were purified using the appropriate percentage denaturing polyacrylamide gel (6–10% 29:1 acrylamide/bisacrylamide, 1× TBE buffer, 8 M urea). The dimensions of the gels for this kind of purification are 18 cm x 20 cm x 2 mm. The gel was run with a constant power 10 W for around an hour. When the run is complete, the gel was exposed using a phosphoimager for about 10-15 min. Then, the screen was imaged by using a Typhoon PhosphoImager. The image was printed out with actual sizes. The gel was placed on top of the printed image and the corresponding RNA band was excised from the gel. Gel pieces were put into 2 ml eppendorf tubes and crushed into small pieces by using 1 ml pipette tip. 0.5x TE buffer with 0.3 M sodium acetate (pH 5.3) was added into the tube and left for elution by rotating at 4 ºC for 2 h. The RNAs were precipitated with ethanol and glycogen at -80°C for 30 min (or overnight at -20°C) and centrifuged at 17,000 g for 30 min at 4°C. Precipitated RNA was re-suspended in 0.5x TE buffer and quantified by liquid scintillation counting.

2.4 RNA 5`-End Labeling Reaction

DNMT, Sc1, GGUG, ss (single strand), poly-A RNAs were chemically synthesized (Integrated DNA Technologies, IDT and Dharmacon). Following reagents were combined in a RNase-free PCR tube: RNA (1 µM final concentration), 10X T4 PNK buffer, RNAse inhibitor (50

U/μl and 1 μl was added for 20 μl reaction volume), T4 PNK enzyme (10 U/μL, 1 μl was added for 20 μl reaction volume), γ-32P ATP (10 μCi/μl, 1 μl was added for 20 μl reaction volume), x μl ddH_2O to make the desired final volume. Reaction was incubated at 37°C for 30-45 min. Unincorporated nucleotides were removed by MicroSpin G25 columns and labeled RNAs were purified using denaturing polyacrylamide gel. RNAs were visualized and purified from the gel as described in RNA body labeling reaction.

If the RNAs were transcribed *in vitro*, 5` phosphate group should be removed before the labeling with radioactive ATP. For 50 μl reaction volume, RNA molecules (~50 pmol) mixed with 5 μl 10x CutSmart buffer, x μl ddH_2O and 2 μl Calf Intestinal Phosphatase (CIP, 10 U/μl, NEB). Reaction was incubated at 37°C for 1 h. Phenol-chloroform extraction was performed. Then, ethanol precipitation was done. Pellet was re-suspended with 0.5x TE buffer and kept at -20°C until use.

2.5 Electrophoretic Mobility Shift Assay

Appropriate concentrations of each protein and trace amount of radioactive labeled RNAs (heated at 95°C for 3 min and snap cooled) were incubated with binding buffer (250 mM KCI, 250 mM Urea, 50 mM Tris-HCl pH 7.4, 2 mM DTT, 2.5 μM yeast tRNA, 0.1 mg/ml BSA, 10% glycerol) in a 10-μl final reaction volume at room temperature for 30 min. A native polyacrylamide (6–10%, 29:1 acrylamide/bisacrylamide) supplemented with 0.5x TB (45 mM Tris-HCI, 45 mM borate, pH 8.1) buffer was prepared. The dimensions of the gels for EMSA experiments are 18 cm x 20 cm x 2 mm. 2 mm spacers allowed us to load more samples and get sharper bands compared to gels with thinner spacers. Before the samples were loaded, empty gel was run at least 20 min. Each reactions was slowly mixed by pipetting before loaded onto the gel. Loading dye or any additional loading components were not added to the reactions. To trace the running of the samples, xylene cyanol and phenol dyes were loaded to one of the empty lanes. Native polyacrylamide gel electrophoresis was run at constant power 6-

7 W for 1 h at room temperature. Gels were dried and subsequently imaged using a Typhoon PhosphoImager (Molecular Dynamics) and the signals were quantified with ImageQuant software suite. Quantified data was fit to a standard two-state binding isotherm using Igor (Wavemetrics), allowing calculation of both dissociation constants and Hill Coefficients.

2.6 Isothermal Titration Calorimetry (ITC)

DNMT RNA binding affinity to RGG1-RRM-RGG2 or RRM-RGG2 proteins were measured using ITC. DNMT RNAs were synthesized and purified as described above. RNA and protein were dialyzed overnight into 2 L of ITC buffer (appropriate concentrations of KCI, 50mM Tris-HCI, pH 7.4) at 4°C. Briefly, the desired volume of DNMT RNA or protein was mixed with an equal volume of 1x ITC buffer and dispensed into 6-8000 Dalton molecular weight cutoff dialysis tubing (Spectra/Por). The RNA and the protein (in the same beaker) were dialyzed 1 day in 2 L of ITC buffer at 4°C by gentle stirring. Buffer was changed every 8 hours. Appropriate concentrations of RNA and protein were made using centrifugal concentrators. For all of the titrations C-values were between 5-10. Titrations were performed at 25 °C using a MicroCal ITC_{200} microcalorimeter (GE Healthcare) and data were fit using the Origin software suite as previously described (219).

There are 2 critical points of doing ITC for RNA-protein interaction. First of all, buffer of the RNA and protein should match perfectly to get rid off buffer mismatch peaks. To do that, RNA and protein should be dialyzed in the same buffer container. At concentration step after the dialysis, the remaining dialysis buffer should be used to prevent any change in the buffer components. Second one is making sure working with a clean ITC instrument. ITC was cleaned after each experiment by detergent wash protocol. Since the proteins are sticky, regular washing protocols may not clean the sample cell. Dirty cell creates baseline problems and noise data. Before each experiment, buffer-buffer titrations were also performed to be sure about the cleanliness of the instrument and not having a baseline problem.

2.7 CD spectroscopy of proteins

RRM and ZnF domains of FUS were expressed and purified as described above. To get a better idea about the secondary structure of the proteins, His-MBP tags of the domains were cleaved by Prescission protease enzyme at 4°C. Prescission protease fused with Glutathione S-transferase (GST) tag was expressed from pGEX-6P plasmid and purified by using standard GST tagged protein purification protocol (220). Uncleaved ZnF and RRM proteins and cleaved His-MBP were removed using nickel affinity resin, and size exclusion chromatography (using a 16/600 Superdex 75 pg column) was performed to further purify isolated domains. Samples were dialyzed in sodium phosphate buffer (250 mM NaF, 50mM (NaH_2PO_4 and Na_2HPO_4), pH 7) and checked for purifity before the CD analysis. CD spectra of the fragments were measured on a Chirascan Plus CD instrument at 20°C. Samples were placed in a cuvette (0.5 mm path length) and four measurements were taken and averaged. Secondary structure content of the fragments was calculated using the program CDNN (221) using the CD spectra in the wavelength range of 185-260 nm.

2.8 RNase T_1 probing

5'-end-labeled RNA (fresh) in 0.5x TE buffer was heated at 95°C for 3 min and snap-cooled on ice for 10 min. 1 µl of RNA solution was added to 9 µl buffer containing 250 mM KCl or LiCl and Tris-HCl (pH 7.4) for 15 min at room temperature. 0.1 units RNase T_1 (Ambion) was added to the RNA and incubated at room temperature for 5 min. To generate a uniform ladder, 1 µl RNA was incubated with 9 µl alkaline hydrolysis buffer (Ambion) for 3 min at 95 °C. Reactions were quenched with 10 µl phenol. Then, for the phenol-chloroform extraction step, 190 µl ddH20 was added and the sample tube was vortexed for couple seconds. 200 µl buffered phenol solution was added and the tubes were vortexed again until an emulsion forms. The enzyme in the water phase is denatured and partition into the phenol, while the RNA stays in the water. The mixture was centrifuged at 17,000 g at room temperature for 5 min. The tubes

were checked to see whether the organic and aqueous phases are well separated. If not well separated, they were centrifuged for a longer time. Aqueous solution was taken and placed into a clean centrifuge tube. Then, 200 µl chloroform was added and vortexed for couple seconds. After the centrifuge at 17,000 g at room temperature, aqueous part was transferred into a clean centrifuge tube Sodium acetate (final 0.3 M, pH 5.2) was added and the tubes were mixed. 100% ethanol with 3-fold of the sample volume was added. For the precipitation, samples were kept at -80°C for 30 min (or at -20°C) and centrifuged at 17,000 g for 30 min at 4°C. Supernatant was discarded and pellet was dissolved with 10 µl loading buffer (8M urea with xylene cyanol and phenol dye). Products were separated using appropriate percentage of denaturing 29:1 acrylamide;bisacrylamide gel by electrophoresis. Gels were dried and visualized using a Typhoon PhosphoImager (Molecular Dynamics).

2.9 Composition gradient-multiangle light scattering (CG-MALS)

Sc1 RNA was transcribed and FUS protein was purified as described above. The buffer components of the samples are 250 mM KCl, 250 mM Urea, 50 mM Tris.Cl pH 7.4. For CG-MALS analysis, there should not be any aggregates in the samples. Thus, both RNA and protein samples were filtered by a membrane filter with 0.11-micron pore size. Running buffer was also freshly prepared and filtered by 0.11-micron membrane filter. Sample concentrations were measured by Nano drop after the filtering step. Protein concentrations were between 0.04-0.08 mg/ml and the RNA concentrations were ~0.2 mg/ml for each run. ~10 ml sample from both species was necessary for one run. The details of method design are shown in Figure 2.2. Data analysis of hetero association was performed according to the instructions at Wyatt Calypso II Composition-Gradient accessory and Analysis software tutorials.

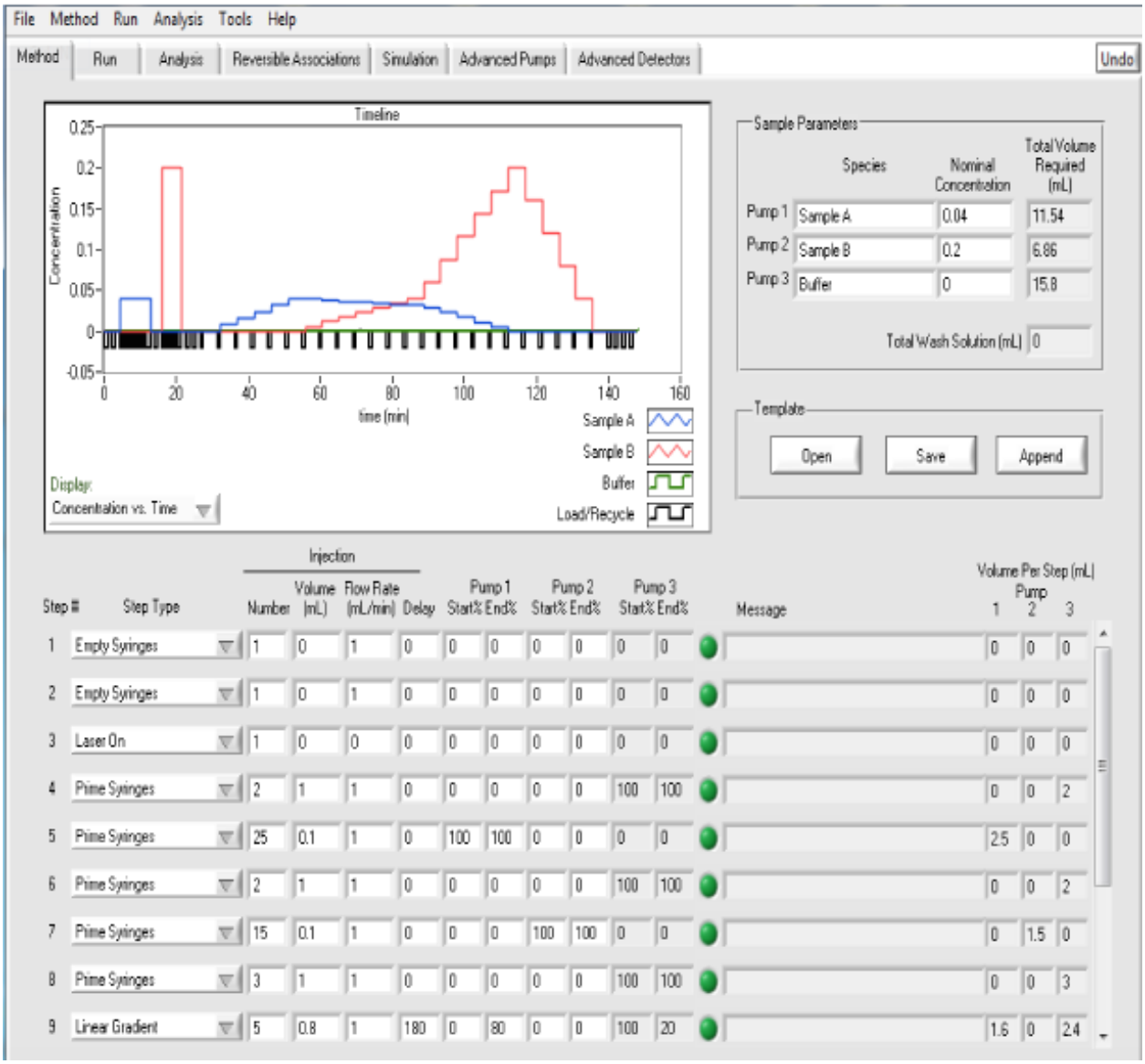

Figure 2.2 Method design for characterization of self- and hetero-association of FUS and Sc1 RNA. Method table (bottom) and the timeline graph are shown.

2.10 Dimethyl sulfate (DMS) foot printing

RNAs with 5` and 3` linkers (shown below) were transcribed and purified as previously described.

5`-TAATACGACTCACTATAGGCCTTCGGGCCAA-RNA of interest-
T7 promoter 5`linker

5`-TCGATCCGGTTCGCCGGATCCAAATCGGGCTTCGGTCCGGTTC
3` linker Reverse transcriptase primer binding site

RNAs (in 0.5 x TE) were heated at 95°C for 3 min and cooled on ice for at least 10 min to ensure that there is one properly folded conformational population. In addition, RNA folding buffer (final concentrations: 100 mM HEPES pH 8.0, 6 mM $MgCl_2$, 100 mM NaCl) was added and the samples were left at room temperature for 15 min. To form protein-RNA complexes, proteins were added to the folded RNAs (final concentration, 1 µM) at different stoichiometric conditions (RNA/protein; 1:1, 1:5, 1:15 etc.). Tubes were incubated at room temperature for 30 min to allow complete complex formation between the RNA and protein.

For dimethyl sulfate (DMS) modification, DMS was carefully diluted (1:15) in ethanol. 1 µl of DMS-ethanol solution was added to 10 µl RNP complex and incubated at room temperature for 2 min. For the non-DMS controls (RNA alone and RNA/protein complex), identical incubation was performed with 1 µl ethanol. RNA/protein complex control is important to understand whether the signals are because of DMS modification or because of RNase contaminants in protein sample that cause the degradation of RNAs. To quench the DMS reaction 190 µl quench solution (0.3 M sodium acetate (pH 5.2) and 25% β-mercaptoethanol (Sigma)) was added to each tube. Then, 2 volumes of ethanol were added and RNA was pelleted by centrifugation at 16,100 x g for 20 min at room temperature. For a more efficient precipitation, samples can be incubated at -80°C for 30 min and then, they can be centrifuged at cold room for 30 min. The supernatant was removed and the RNA pellet was re-suspended in 100 µl sodium acetate (0.3 M, pH 5.2). Phenol-chloroform extraction was performed to get rid of the proteins. Then, additional RNA precipitation was performed and samples were kept at -20°C.

Key points of DMS foot printing: (i) There should not be any degradation at initial RNA samples. After *in vitro* transcription and purification, centrifugal concentrators can be used to get rid of degraded products. Then, one more slab gel purification can be done to ensure about high percentage of the RNAs are in full length. Purified RNA samples should be kept at -20°C with 1

experimental volume aliquots (ex: 5 µl aliquots of 5 µM concentration) to prevent the degradations caused by freeze-thaw process. If the degradation is still seen after all of these precautions, rnase inhibitor can be added to the RNA/protein binding reactions. (ii) In the end of the protocol, there should not be any bound protein to RNA that can prevent the reverse transcription reaction at the next step. Thus, before and at the phenol-chloroform process, vortexing the samples vigorously will help to remove all the proteins.

2.11 Primer extension

After deproteinizing and isolating the RNA, reverse transcription reaction was performed to detect the DMS modified regions of the RNAs. First, RNA pellets were re-suspended in 10 µl 0.5 x TE buffer. 3 µl of 5` end-labeled RT primer (5`-GAACCGGACCGAAGCCCG-3`) was added. Reaction was incubated at 65°C for 5 min and then, at 35°C for 15 min to anneal the primer to the RNA. Reaction was immediately put into slushy ice and kept there for 1min. 6 µl enzyme mix (1.6 mM dNTPs, 16 mM DTT, GSI-II enzyme buffer (10 mM NaCI, 10 mM $MgCl_2$, 1 mM DTT and 20 mM Tris.Cl pH 7.5)) was added. For sequencing reaction 0.7 mM desired ddNTP was added. Tubes were heated to 55°C for 1 min. 0.7 µl GSI-II enzyme was added and then, the reaction was slowly mixed by pipetting 1-2 times. Samples were incubated at 60°C for 15 min for reverse transcription. To quench the reaction, 1 µl 5 M NaOH was added. This degrades the RNA but does not damage DNA. Reaction should be mixed by ~40 times pipetting to get rid of the RNAs. Samples were heated to 95°C for 5 min. Then, 20 µl loading dye (8 M Urea, 20 % formamide, phenol and xylene cyanol dye) was added and the samples were heated to 95°C for 5 min. After this step, samples can be stored at -20°C up to one month, but radioactivity of the samples will decrease during that time. To get a better signal from the radioactive bands, samples were immediately run on the 10% denaturing gels (19:1 acrylamide:bisacrylamide, National Diagnostics, UreaGel system). Gels were dried and visualized using a Typhoon PhosphoImager (Molecular Dynamics).

In this protocol, I did not use Superscript III enzyme that is the most commonly used RT. It could not complete the extension of our RNAs probably because of the secondary structures in them. I decided to use GSI-II enzyme since it can work 60°C and got the full-length bands.

2.12 Iodine foot printing

Foot printing of phosphorothioate-containing RNA transcripts was previously described (222). In this method, I transcribed the RNAs in two different reactions in which two of the NTPs (a/c or g/u) were complemented by 5% of their corresponding nucleoside 5`-*O*-(1-thiotriphosphate). Then, RNAs were purified and labeled with radioactive ATP from 5`end as described above. Purified radioactive RNAs were heated at 95°C for 3 min and cooled at slushy ice for 10 min to get an accurate RNA folding. Desired amount of protein was mixed with trace amount of RNA in binding buffer (250 mM KCl, 250 mM Urea, 2 mM DTT, 50 mM Tris.Cl pH 7.5). Reaction was incubated at room temperature for 30 min. For iodine foot printing, freshly prepared idone (dissolved in ethanol and final concentration is 1 mM) was added to the reaction and incubated at room temperature for 1 min. Reaction was quenched with phenol and sodium acetate. Then, phenol-chloroform extraction and ethanol precipitation were performed.

2.13 *In vivo* analysis of FUS RNA-binding activity

HEK293T cells were cultured in Dulbecco's modified Eagle's medium supplemented with 5% fetal bovine serum and antibiotics. For all transfections cells were seeded at a density of 2.5 x 10^6 cells in T-25 flasks and allowed to grow overnight at 37 °C. Cells were transfected with 10 µg of plasmid xpressing FLAG-tagged FUS constructs using Lipofectamine 2000, Opti-MEM, and media without antibiotics. After 24 hours, each flask was washed with 1x phosphate buffered saline (PBS), trypsinized, resuspended in media with antibiotics, and split into a T-75 flask to grow overnight at 37 °C. Cells were harvested the next day and UV-crosslinked once at 400 mJ/cm^2 to crosslink protein and RNA.

To visualize RNA-protein complexes, the UV-crosslinked cells were lysed in lysis buffer (PBS, 0.1% SDS, 0.5% deoxycholate, 0.5% NP-40) and sonicated. DNase I and RNase T_1 were added to the sonicated lysate, incubated at 37 °C, and centrifuged to collect supernatant. The supernatant was added to Protein A/G agarose beads (Pierce #20421) previously pre-bound to FLAG-antibody (monoclonal M2 anti-FLAG antibody, Sigma-Aldrich, #F1804) and incubated at 4 °C for 2 hours to specifically pull down FLAG-FUS-RNA complexes. Beads were washed to remove nonspecifically bound proteins and the retained RNA was 5'-end labeled by adding polynucleotide kinase (PNK) buffer, T4 PNK, and γ^{32}P-ATP and incubated at 37 °C. After incubation period, labeled beads were washed, 4x NuPage loading buffer added, and the beads incubated at 95 °C to elute ^{32}P-labeled FLAG-FUS-RNA complexes. Samples were electrophoresed on a 4-20% SDS-PAGE gel, transferred to a PDVF membrane, exposed on the Molecular Imager FX Imaging Screen K for 5 minutes, and imaged on a Biorad Pharos FX Plus Molecular Imager. Two technical replicates of westerns and phosphor imaging from three separate pull-downs were quantitated and results averaged together.

Chapter 3 : RGG/RG domains mediate RNA binding activity of Fused in Sarcoma (FUS) protein

Author Contributions: *In vivo* analysis of the FUS mutants was performed by Schwartz lab. I performed all *in vitro* experiments.

3.1 Introduction

RNA-binding proteins (RBPs) are vitally important in cell physiology. In fact, they comprise up to 11% of any organism's proteome (131). The recognition of RNA by these proteins in most cases localize and in many cases active their function. Classically, RBP recognition requires that a specific RNA sequence be presented in the context of a precise structure (223). By this mechanism, RBPs target specific RNA transcripts to accomplish a specific function. Examples include recognition of U1 small nuclear RNA by U1A (224,225), 7SK RNA by HEXIM1 and Larp7 (226,227), or the 4.5S RNA by the signal recognition particle (SRP) (228,229).

Recently, it has become appreciated that a large number of RBPs contain disordered regions of low sequence complexity (94). In most cases, these domains are rich in alternating arginines and glycines, either RG or RGG. These RG- and RGG-rich domains are considerably heterogeneous in their sequence composition and number of repeats (148). RGG/RG domains offer a unique mode of binding for RBPs. The positive charge of arginines provides a non-specific affinity for the phosphate backbone of RNA as well as specific binding through hydrogen bonding via amino group. The flexibility of the glycines allows for multiple conformations of binding to RNA. However, since RGG/RG domains usually occur with other classical RNA binding domains in the proteins, their role in RNA binding activity of a protein was underestimated.

Particularly prominent among the RGG/RG domain proteins are the FET family of RNA binding proteins, comprised of FUS, EWSR1, and TAF15 (170). FET proteins are conserved

throughout metazoans. In human, they are ubiquitously expressed in all tissues and are among the most highly expressed proteins in the cell. FET proteins are nuclear proteins that predominantly associate RNA Pol II and pre-mRNA, affecting both transcription and mRNA processing. Each of these proteins have three RGG/RG domains interspersed between other domains – the low complexity (LC) domain, an RNA recognition motif (RRM), a CCCC-type zinc finger domain (ZnF), and a C-terminal nuclear localization signal (NLS). In addition to have a number of RGG repeat regions, the number of repeats are unusually high – up to nine repeats.

Within the FET family, FUS is the best characterized. FUS is a general transcription factor that binds RNA Pol II near the promoters of two thirds of expressed genes and regulates transcription (171). FUS binds the C-terminal domain of RNA Pol II in an RNA-dependent manner. As a general transcription factor, it is important that FUS not be too specific of an RBP, else it would not have the flexibility to target so many genes. The promiscuity by which FUS recognizes RNA seems unusual; however, a growing appreciation exists that more RBPs possess this property. Finally, FUS is critically important for two human diseases (230). Translocations involving the LC domain of FUS cause a number of sarcomas, including pediatric sarcomas. Also, mutations predominantly found in the RGG domains of FUS cause the neurodegenerative disease, amyotrophic lateral sclerosis (ALS).

LC domain together with the RGG/RG domains give FUS the ability to spontaneously form large phase separated protein assemblies referred to as “droplet” assemblies (72,231). These droplets have very dynamic, liquid-like properties that are also found in certain non-membrane bound organelles within the cell, including nucleoli, p-bodies, and stress granules. Thus, understanding the role of RGG/RG domains in RNA binding activity of FUS will also contribute to the understanding of mechanism of granule formation.

Sequencing studies (iCLIP, CLIP-seq, HITS-CLIP, and PAR-CLIP) have raised more questions than answers about the RNA-binding specificity of the FUS protein (171,232-237). These experimental protocols applied to FUS seek to identify all of the RNA sequences that co-

immunoprecipitate with the protein, to which computational analysis is applied to detect any sequence motif and hypothetical structure that may be enriched. To date, results from these studies have failed to provide agreement regarding an RNA sequence or structure targeted by FUS (170,238). Neither have the sequence-structure motifs proposed stood up to more rigorous biochemical scrutiny (170,238). This may, in part, result from the fact that sequences identified in FUS immunoprecipitations represent the majority of transcribed RNAs in the cell. Also, FUS and other hnRNP proteins are known to oligomerize along RNA, reducing the significance of any target over that of adjacent sequences (75,170,233,234,239,240). Finally, many hnRNPs, like FUS, possess low-complexity (LC) domains, that self-assemble into large RNA-rich granules (70-72,74,75,80,239,241-246). Taken together, the bulk of sequences that immunoprecipitate with promiscuous RNA-binding proteins, like FUS, may not be reflective of RNA sequences or structures with higher affinity for FUS or its RNA-binding domains.

To better define the RNA binding properties of FUS, I have determined the RNA binding characteristics of RRM, ZnF, their fused forms with adjacent RGG/RG domains and RGG/RG domains alone using electrophoretic mobility shift assay (EMSA), isothermal calorimetry (ITC). In addition, I performed mutation analysis RGG domains of FUS to see their role in RNA binding. I found that RNA binding activity of FUS is largely driven by the affinity of the RGG/RG domains *in vitro* and *in vivo*. My data indicated that FUS preferentially binds to RNAs with secondary structure motifs. Overall, I reveal a newly defined specificity of FUS protein that is degenerate and will be a model for future studies about intrinsically disordered RNA binding proteins.

3.2 Results

3.2.1 Individual RRM and ZnF domains of FUS do not bind RNA with high affinity

FET proteins contain the most repeats of RGG/RG and the longest RGG/RG domains. Thus, we used the FET protein Fused in Sarcoma (FUS) as a model to define the RNA binding

properties of these domains. FUS contains an N-terminal low complexity (LC) domain and five putative RNA-binding domains (RBDs): an RNA recognition motif (RRM), a zinc finger domain (ZnF), and three arginine/glycine rich regions (RGG1, 2, and 3) (Figure 3.1A).

Structure of the RRM domain exhibits a similar folding with other RRM domains. However, the amino acid sequence of it significantly differs from the canonical RRM domains (247). It does not have some of the conserved amino acids in RNP motifs (247). It has an unusually long loop 2 that is rich in lysine amino acid (KK loop) and involved in RNA binding. ZnF of FUS forms the C4 type ZnF domain (248). It has 40% sequence homology with ZnF domain of ZNF265 splicing factor, which binds to cyclin mRNA (248,249). To quantify the contributions of the two structured RBDs, RRM and ZnF, to RNA binding, each was tested for binding activity in the absence of other domains. The purified minimal RRM (a.a. 267-373) and ZnF (a.a. 422-453) domains were determined to be folded in a manner consistent with previous circular dichroism (CD) measurements and associated structural studies (Figure 3.2A) (75,247,248). From this, I concluded that my RRM and ZnF domains do not require the flanking RGG/RG domains to fold into a stable form.

Previous studies determined FUS affinity for numerous RNA sequences and found FUS highest affinity to the 48 nucleotide RNA, DNMT RNA (previously referred to as prD), a noncoding RNA identified to be bound to FUS in HEK293T/17 cells by CLIP-seq and ChIP-seq (75,171,238). Therefore, DNMT RNA was chosen as a model RNA to further examine the contribution of individual FUS domains on RNA binding. I observed full length FUS bound DNMT RNA with a $K_{D,app}$ of 0.7±0.2 μM (Figure 3.1B, Table A1). In contrast to previous studies, I did not adjust the K_D values for the relative RNA-binding activity, which is difficult to accurately assess for FUS due to its RNA-dependent oligomerization behavior (74,75,244). Therefore, my reported $K_{D,app}$'s are ~7-fold higher than that previously reported, which had been interpreted to be 15% active (75,238). Unlike full-length FUS, the RRM and ZnF domains had little observed affinity for DNMT RNA as determined by electrophoretic mobility shift assay (EMSA) (Figure

3.1B and C). Isothermal titration calorimetry (ITC) was used to confirm the detectable binding (Figure 3.2B and C). Together, these data reveal that the minimal RRM and ZnF in the absence of other domains contribute little or no affinity for DNMT RNA.

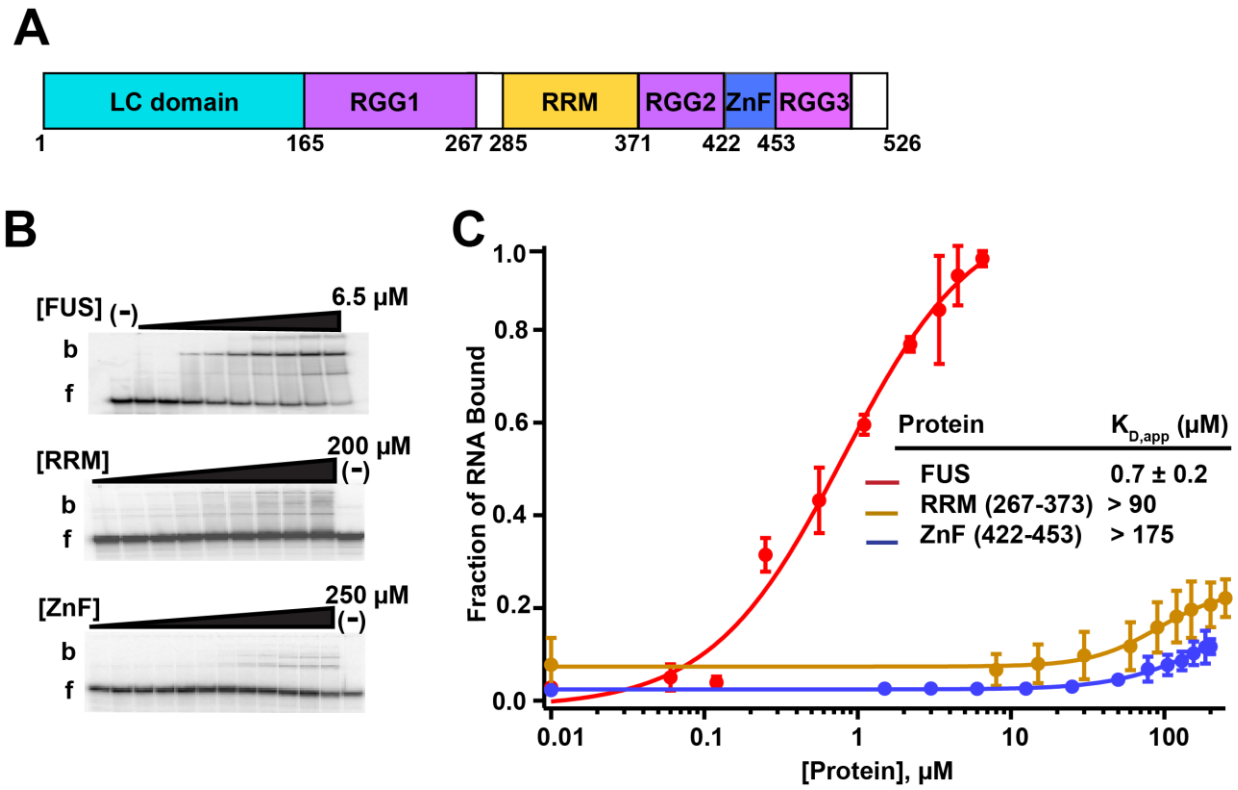

Figure 3.1 RRM and ZnF domains of FUS do not bind to the RNA with a high affinity. (A) Domain structure of FUS. (B) A trace amount of the DNMT RNA was incubated with increasing concentrations of FUS, RRM (267-373) or ZnF (422-453). Binding was analyzed by Electrophoretic Mobility Shift Assay (EMSA). b = bound DNMT RNA and f = free DNMT RNA. '(-)` shows no protein lane. (C) Binding curves of EMSA data. Error bars represent the S.D. of three independent titrations for each construct.

To ensure that this result is not idiosyncratic to DNMT RNA, I tested other RNAs possessing a range of sequences and structures. These other RNAs include GGUG RNA (from an *in vitro* selection for FUS-binding aptamers (250)), a guanosine-deplete 36-nucleotide (nt) single stranded RNA sequence (derived from cobalamin riboswitch linker, CRL (251)), a 40 nucleotide polyadenosine RNA, the repeating RNA domain (RRD) from human (hRRD) and

mouse (mRRD) Firre lncRNA (190) and the Sc1 RNA aptamer (36 nt) (157) (Table A.1). Each of these RNAs exhibited low micromolar affinity for wild type FUS but not the ZnF domain, while the RRM domain weakly bound Sc1 ($K_{D,app}$ = 48±3 μM) and poly-A RNA ($K_{D,app}$ of 45±2 μM) (Table 3.1). Thus, for the RRM and ZnF domains, often annotated as nucleic acid binding domains, I found no evidence that they alone were responsible for the RNA binding activity of FUS. These results agree well with three prior structural studies of the FUS RRM (247,248,250). A previous study with the ZnF domain of FUS protein exhibited that it binds to GGUG RNA with 10 μM affinity (248). However, in that study, they expressed ZnF domain with flanking amino acids from the adjacent RGG domains (a.a. 398-468). This suggests that the additional amino acids may impart the RNA binding activity of ZnF domain.

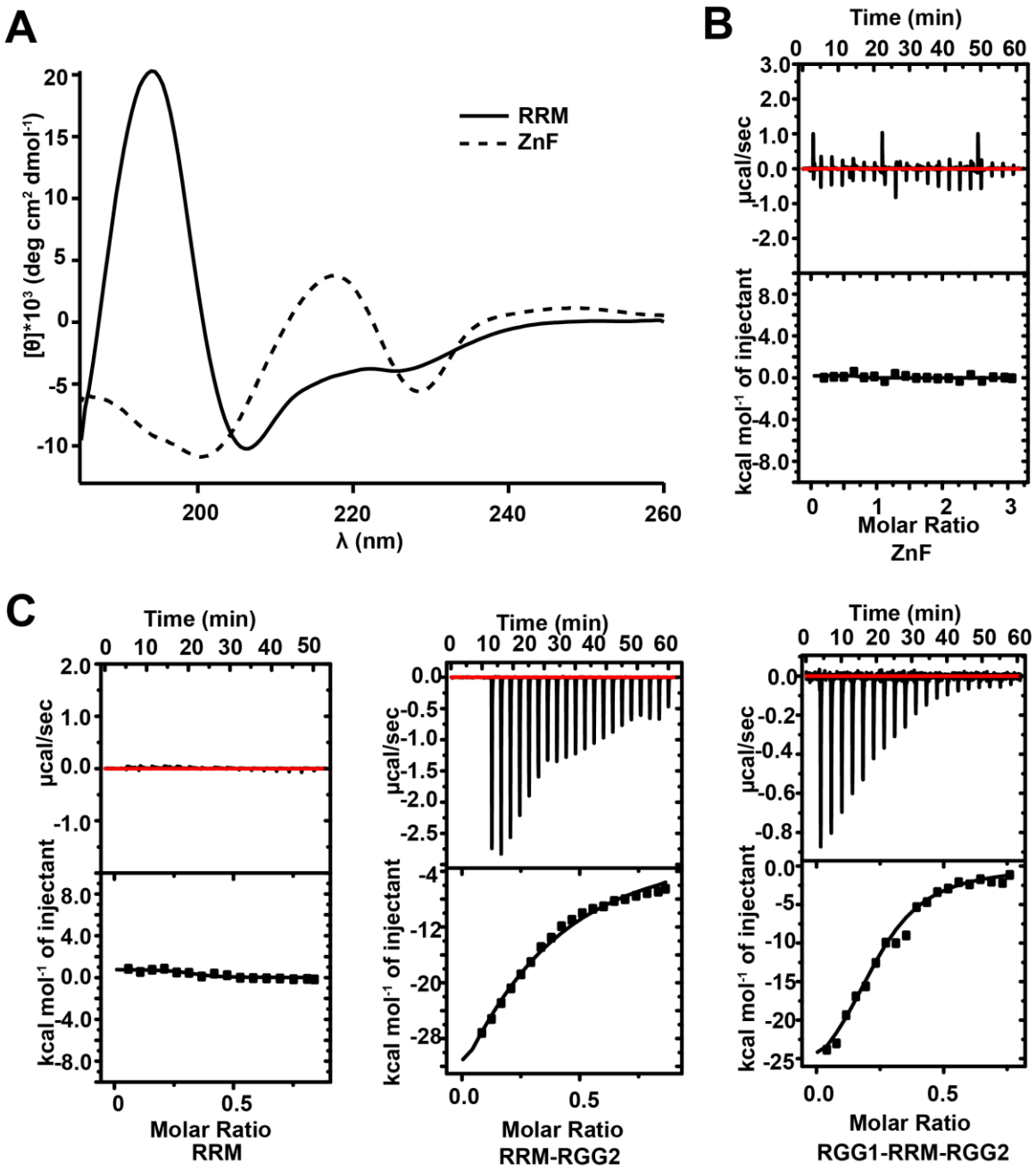

Figure 3.2 RRM and ZnF domains of FUS are correctly folded, but do not bind to RNA. (A) CD spectra of RRM and ZnF domains (no MBP tag) of FUS. The values are 5% alpha-helix, 23% β-strand for ZnF and 28% alpha-helix, 25% β-strand for RRM. Representative binding isotherms and non-linear curve fitting for the titrations of 800 µM DNMT RNA in 200 µM (B) ZnF or (C) RRM and for the titrations of 400 µM DNMT RNA in 100 µM RGG1-RRM-RGG2 and RRM-RGG2. (Top panels) Raw heats of binding obtained by Isothermal titration calorimetry (ITC) when proteins mixed with DNMT RNA. (Bottom panels) Binding isotherms fitted to the raw data using single-state binding model. Titrations were performed in 150 mM KCl, 50 mM tris buffer. RGG1-RRM-RGG2 binds to DNMT RNA with 3 ± 0.4 µM affinity and RRM-RGG2 shows a 5 ± 0.2 µM binding affinity.

Table 3.1 $K_{D,app}$ (μM) values of ZnF, RRM and RRM+3RGG interactions with different RNA molecules.

	RNA						
	Sc1	DNMT	hRRD	mRRD	GGUG	CRL	Poly-A
Protein							
ZnF	n.d[a]	n.d	n.d	n.d	n.d	n.d	n.d
RRM	48 ± 3	n.d	n.d	n.d	n.d	n.d	45 ± 2
RRM+3 RGG	27 ± 2	4 ± 0.5	16 ± 3	10 ± 1	21 ± 1	63 ± 8	40 ± 5

[a]n.d. : not detectable (>100 μM)

3.2.2 FUS RGG/RG domains promote RNA binding

To determine whether the flanking RGG/RG domains can increase the affinity of the RRM or ZnF for RNA, I investigated the RNA-binding activity of each domain with one or both flanking RGG/RG domains. Fusion of RGG1 to the RRM (RGG1-RRM) and RRM to RGG2 (RRM-RGG2) increased the RNA binding similarly (Figure 3.3, Figure 3.2C). RRM with both flanking domains (RGG1-RRM-RGG2) protein showed the same affinity for RNA as RGG1-RRM and RRM-RGG2. These results reveal that modest number of RGG repeats and RRM domain are sufficient to enable RNA binding. In these EMSA gels, there are “bound RNA” bands that look corresponding to different molecular weights. This might be because of multiple proteins binding to one RNA molecule through increasing concentration of the protein. ITC data of RGG1-RRM-RGG2 binding to DNMT RNA in 250 μM KCl revealed a 1 RNA:3 proteins stoichiometry (Table A2). However, it is difficult to tell whether three proteins directly bind to DNMT RNA or the protein itself oligomerize by the increased protein concentration. Indeed, binding stoichiometry for 10 mM KCl (1 RNA : 10 protein) support the second scenario since the direct interaction of all proteins with the RNA may not be possible due to steric problems. For the EMSA analysis of these binding experiments, I assumed the stoichiometry is 1 RNA : 1 protein and calculated the dissociation constant according to this assumption. The RGG/RG domains flanking the ZnF domain also restored RNA binding activity (Figure 3.4A and B).

However, the ZnF required both RGG2 and RGG3 for binding affinities near to that of full-length FUS, which would add a large RNA interaction surface area.

To establish the minimum number of RGG/RG repeats necessary to enhance the RNA binding activity of RRM, I fused to the RRM domain with increasing numbers of RGG repeats from the RGG2 domain. RRM with one or two RGGs did not bind RNA, but three RGGs recovered binding close to that of the full length FUS protein ($K_{D,app}$ = 4±0.5 μM, Table 3.2). Additional RGG repeats increased binding to within 3-fold of wild type FUS.

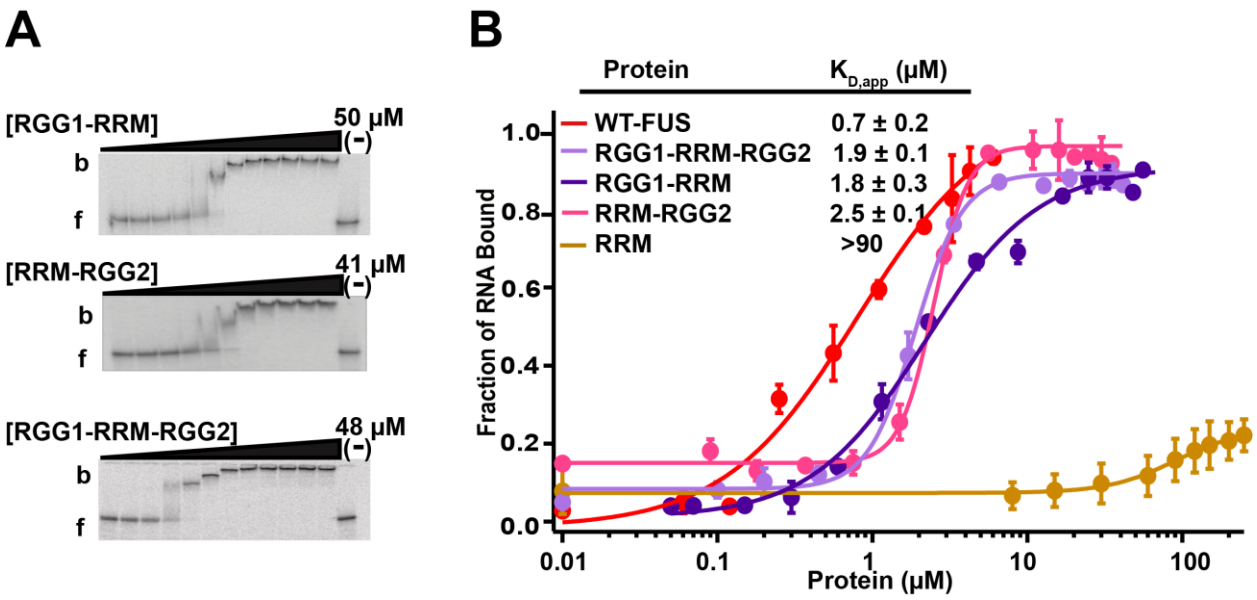

Figure 3.3 Flanking RGG/RG domains impart the RNA binding activity of the RRM.
(**A**) Representative EMSAs and (**B**) the corresponding binding curves showing binding of DNMT RNA to RGG1-RRM (165-373), RRM-RGG2 (267-422), RGG1-RRM-RGG2 (165-422) proteins. b = bound and f = free RNA. '(-)' shows no protein lane. Error bars represent the S.D. of three independent titrations for each construct.

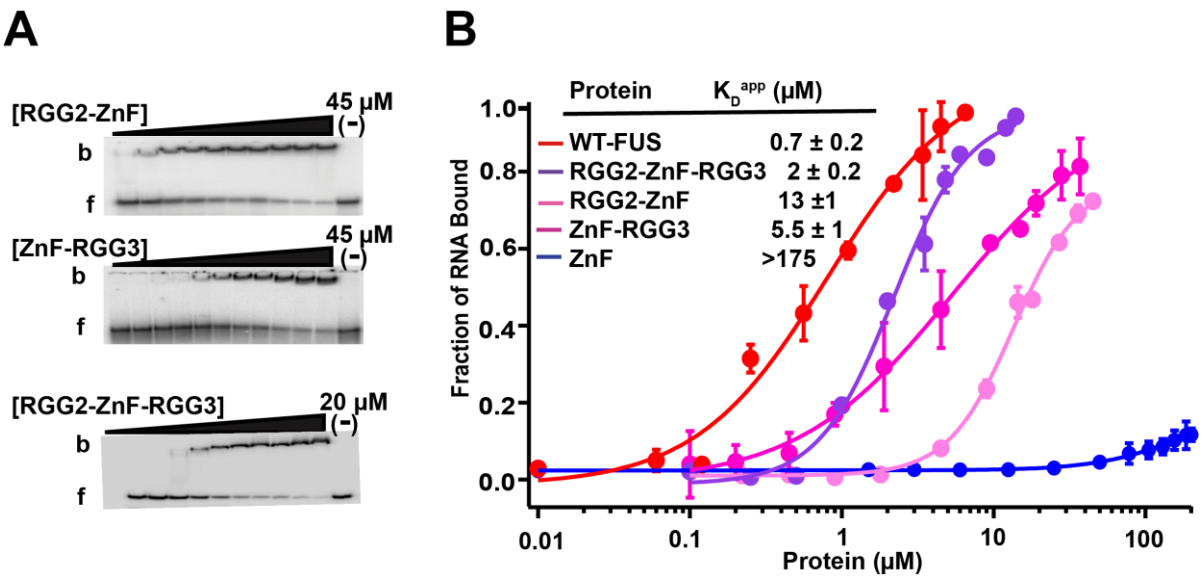

Figure 3.4 Flanking RGG domains impart the RNA binding activity of the ZnF domain. (**A**) Representative EMSAs and the (**B**) corresponding binding curves showing binding of DNMT RNA to RGG2-ZnF (372-453), ZnF-RGG3 (423-501) and RGG2-ZnF-RGG3 (372-501). b = bound DNMT RNA and f = free DNMT RNA. '(-)' shows no protein lane. Error bars represent the S.D. of three independent titrations for each construct.

Table 3.2 RRM-RGG domains binding to DNMT RNA.

FUS domain	$K_{D,app}$, (μM)[a]
RRM (267-373)	> 90
RRM + 1 RGG (267-380)	> 60
RRM + 2 RGG (267-385)	> 60
RRM + 3 RGG (267-390)	4.1 ± 0.4
RRM + 4 RGG (267-397)	2.6 ± 0.8
RRM + 5 RGG (267-421)	2.5 ± 0.1

[a] $K_{D,app}$ values represent range of two or more independent experiments.

3.2.3 Electrostatic interactions are involved in RGG/RG interactions with RNA

I hypothesized that the binding for FUS is dominated by electrostatic interactions between the phosphate backbone of RNA and arginine residues. To examine this, I measured the affinity of the RGG1-RRM-RGG2 protein for DNMT RNA as a function of salt concentration using ITC (Figure 3.5A, Figure A1). This analysis revealed two distinct linear phases. At lower salt concentrations (50 mM – 150 mM), the shallow slope (-0.84) indicated a very modest salt dependence of the RNA-protein interaction, suggesting a small electrostatic component to

binding. Thus, binding at physiological salt concentrations may be driven by hydrogen bonding interactions (101,159). At higher salt concentrations (200 mM – 300 mM), the interaction was found to be more affected by salt concentration with a steeper slope (-2.5). This suggests that electrostatic interactions are a required component of RNA binding, in addition to hydrogen bonding. Alternatively, a structural change in the either the protein or the RNA at higher salt concentration may increase reliance on the electrostatic component.

Typical monovalent ion concentrations in mammalian cells is ~150 mM (139 mM K^+ and 12 mM Na^+). My ITC data showed the contribution of the electrostatic interactions on RNA-protein binding at the salt concentrations higher than the physiological level. Since these are *in vitro* experiments, the data can only be an approximation for a more complex *in vivo* system. In addition, there are evidences about intracellular salt concentration changes in dividing cells and cancer cells (252). Yeast cells can also have 50-300 mM K^+ depending on the growth phase (253,254). Moreover, studies with *Bufo Bufo* oocytes revealed that nuclear K^+ concentration is much more higher than the cytoplasmic K^+ concentration (266 mM and 70 mM respectively) (255). These evidences suggest that the high salt concentration that is necessary for the electrostatic interaction of RGG1-RRM-RGG2 with the RNA molecule can also be physiologically relevant.

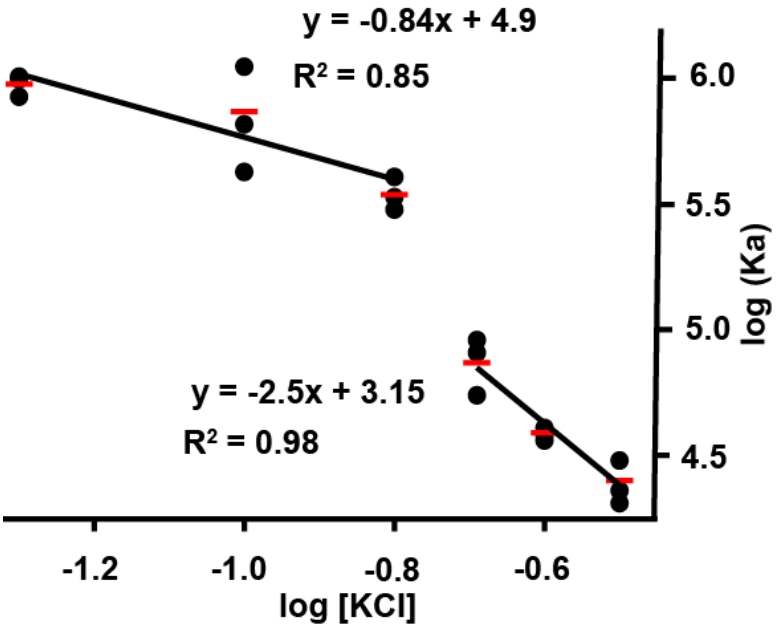

Figure 3.5 Interaction between RGG1-RRM-RGG2 and DNMT RNA shows a salt dependence. ITC titrations were performed for the interaction of RGG1-RRM-RGG2 and DNMT RNA in different KCl concentrations (50 mM, 100 mM, 150 mM, 200 mM, 250 mM, 300 mM). The graph shows linear correlation between log (Ka) and log (KCl concentration in M) for three independent repeats. Values of each repeat are shown with black circles. Red line indicates mean of the three repeats.

3.2.4 Individual RGG domains can bind to RNA

The above results suggested that the RGG/RG domains principally drive FUS affinity for the RNAs tested. I then examined each RGG/RG domain fused to the C-terminus of maltose binding protein (MBP) and tested for RNA binding by EMSA. Of the three RGG/RG domains, RGG1 and RGG3 domains bound the DNMT RNA with significant affinity, 3-fold and 10-fold weaker than full length FUS respectively (Figure 3.6A and B). On the other hand, RGG2 bound RNA 100-fold weaker than full length FUS. The fact that RGG2 fused with either RRM or ZnF domains bound the RNA with markedly higher affinity ($K_{D,app}$ = 2.5±0.1 μM for RRM-RGG2 and 13±1 μM for RGG2-ZnF, Figure 3.3A, B, and Figure 3.4A and B) than either domain alone indicates a synergistic interaction between the RNA binding domains.

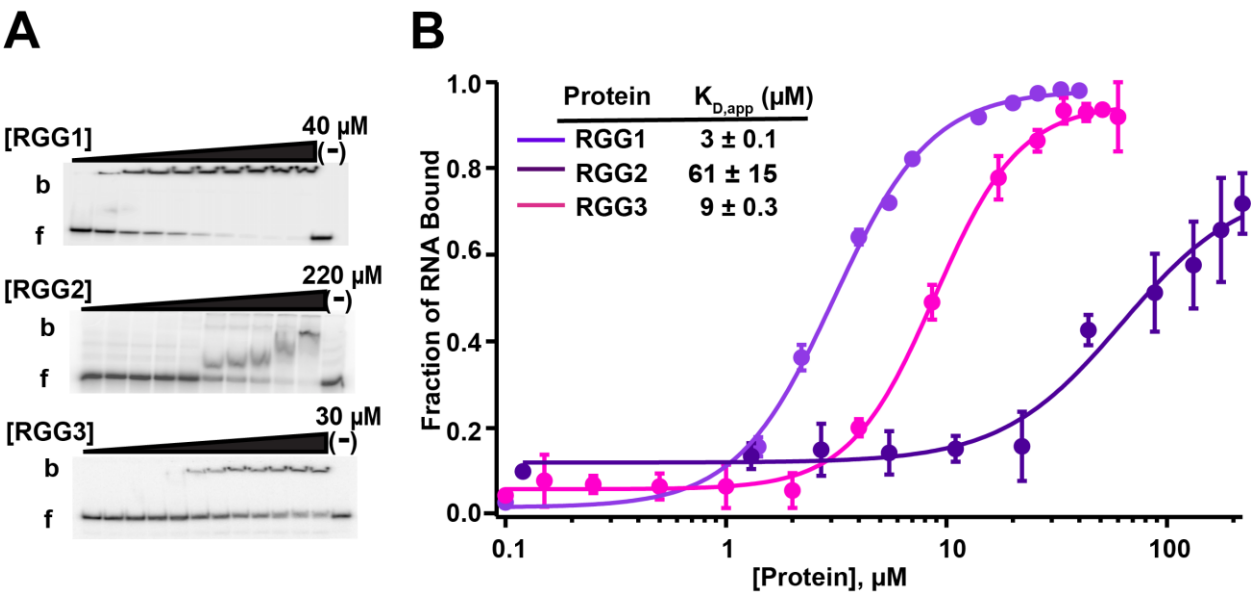

Figure 3.6 Individual RGG/RG domains of FUS bind to RNA. (A) Representative EMSAs and (B) corresponding binding curves of individual MBP-RGG domains of FUS; RGG1 (165-267), RGG2 (372-422), RGG3 (454-501), in the presence of DNMT. b = bound and f = free RNA. '(-)' shows no protein lane. Error bars represent the S.D. of three independent titrations for each construct.

To assess the role of RGG/RG domains in the full-length protein, four mutants were created with arginine residues in either RGG1, 2, or 3 converted to serine (SGG1, 2, and 3, respectively), and a fourth with all arginine residues in these domains mutated to serine (SGG4) (Figure 3.7A, Table A3**)**. SGG1 and SGG2 bound the DNMT RNA with affinity near that of wild type FUS ($K_{D,app}$ = 0.6±0.1 μM and 1.2±0.2 μM respectively, Figure 3.7B and C). On the other hand, SGG3 mutant bound the DNMT RNA with ~5-fold lower affinity than wild type FUS ($K_{D,app}$ = 4.3±2.3 μM). This data suggested that each RGG/RG domain provides affinity for RNA and, in the case of DNMT RNA, these domains may substitute for each other. SGG4 showed a strong reduction in binding; with 30-fold lower affinity than wild type FUS. It should be noted that the nature of the shift induced by the SGG4 mutant is different from the individual SGG mutants, suggesting a different binding mode. The relative affinity of SGG4 for RNA is suggestive that multiple weak interactions synergize to enhance the apparent binding affinity that is known as

"avidity effect". While the isolated RRM and ZnF domains show little individual affinity for RNA, together and in the context of this high affinity RNA partner, some binding can be observed.

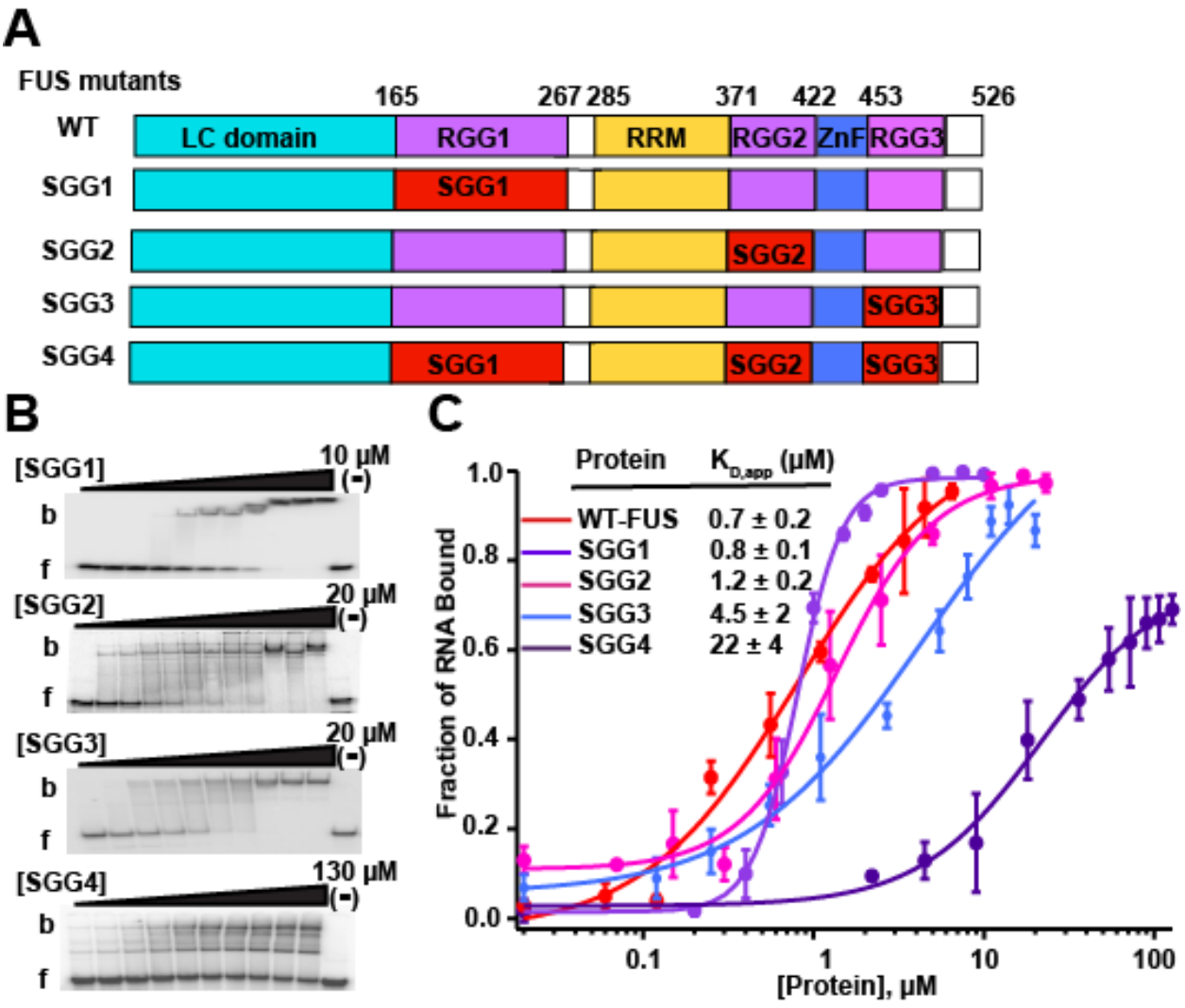

Figure 3.7 RGG/RG domains of FUS mediate high affinity binding to RNA. (A) Schematic illustration of RGG/RG domain mutations. Arginine amino acids of individual RGG/RG domains were converted to serine amino acids in SGG1, SGG2, and SGG3 mutants. In SGG4 mutant, arginine amino acids in all RGG/RG domains were converted to serines. (B) Representative EMSAs of mutant FUS proteins with the DNMT RNA and (C) corresponding binding curves. b = bound and f = free. '(-)' shows no protein lane. Error bars represent the standard deviation of three independent titrations for each construct.

3.2.5 RGG/RG domains are important for RNA binding in cells

Because DNMT RNA represents the highest affinity RNA target was found for FUS, I hypothesized that RGG/RG domains may have varied affinities across the range of bound RNAs found in cells. To examine this, Schwartz`s group expressed FLAG-tagged SGG mutants in HEK293T/17 cells to quantitate the amount of UV-crosslinked RNA recovered by immunoprecipitation with anti-FLAG IgG. Western blots revealed expression levels between the exogenous and endogenous FUS to be comparable. Electrophoretic mobility was affected by SGG mutations, presumably due to altered charge with fewer arginines (Figure 3.8A). Recovered FUS-RNA was fragmented by RNase T_1 digestion and the RNA fragments protected by cross-linked protein were radiolabeled and resolved by SDS-PAGE (Figure 3.8B). Based on mobility, crosslinked RNA fragments were estimated to be less than 20 nts in length. Notably, each of the three individual RGG mutants (SGG1, 2, and 3) bound less RNA than WT FUS. Removal of all three domains in SGG4 yielded no detectable levels of bound RNA (Figure 3.8C). Thus, the avidity effect of RRM and ZnF was sufficient to produce measurable RNA binding *in vitro*, this affinity seems not to be significant *in vivo*.

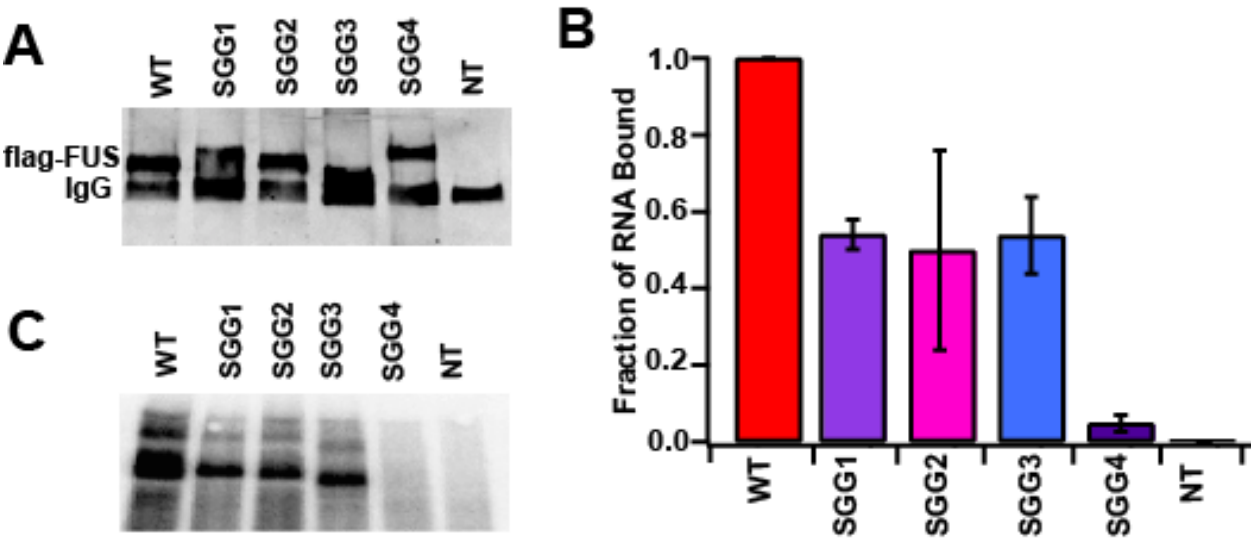

Figure 3.8 RGG domains mediate RNA binding activity of FUS *in vivo*. Figure legend is at the next page.

Figure 3.8 RGG domains mediate RNA binding activity of FUS in vivo. (A) Western blot data of Flag-tagged, wild type, and mutant FUS constructs expressed in HEK293T cells. (B) SDS-PAGE of radiolabeled RNA fragments cross-linked to flag-tagged FUS or SGG mutants of FUS. (C) Two technical replicates of three separate pull-downs were quantitated and average together. Error bars represent standard deviation.

3.2.6 RNA binding mode of FUS protein

To characterize the RNA target preference of FUS, I performed EMSA experiments with a set of RNA molecules with different characteristics (Table A1): an RNA aptamer (Sc1) with a hybrid G-quadruplex/A-form duplex structure (101,157,159), three different single stranded RNAs (a 40 nucleotide polyadenosine homopolymer (poly-A), a 25 nucleotide guanosine-rich RNA that was obtained in an *in vitro* selection against FUS (GGUG) (250), and a heterogeneous 36 nucleotide RNA (CRL) containing five repeats of the sequence (AUACAAC) (251)), and two simple hairpins (one containing a hairpin exclusively composed of 12 A-U pairs (dsAU) and a second with a tract of six G-C pairs in the middle (dsGC) and each capped with a UUCG tetraloop). Finally, the repeating RNA domains (RRD) from human (hRRD) and mouse (mRRD) Firre lncRNA were chosen since they include the mixture of both single strands and secondary structure elements (190). FUS binds to structured RNA molecules with K_D ranging from 0.3 to 0.7 μM (Figure 3.9, Table A4). However, it shows ~30-fold lower binding affinity against the fully single stranded CRL RNA suggesting that FUS preferentially interacts with RNAs that contain some duplex character. LC-RGG1 and RRM-RGG2 constructs of FUS also follow similar binding pattern, indicating that RNA binding affinity of FUS is a result of cumulative binding activity of all domains.

To further characterize RNA binding sites of FUS, dimethyl sulfate (DMS) footprinting analysis was performed with FUS and hRRD RNA. DMS reacts with RNA to form stable, covalent adducts with Watson-Crick edge of unpaired adenosine and cytosine residues. These adducts can be visualized as they block cDNA synthesis by reverse transcriptase (RT). In theory, FUS binding to DMS target sites should protect those nucleotides from DMS reagent.

My data revealed very weakly protected sites for DMS footprinting data (Figure 3.10 and Figure

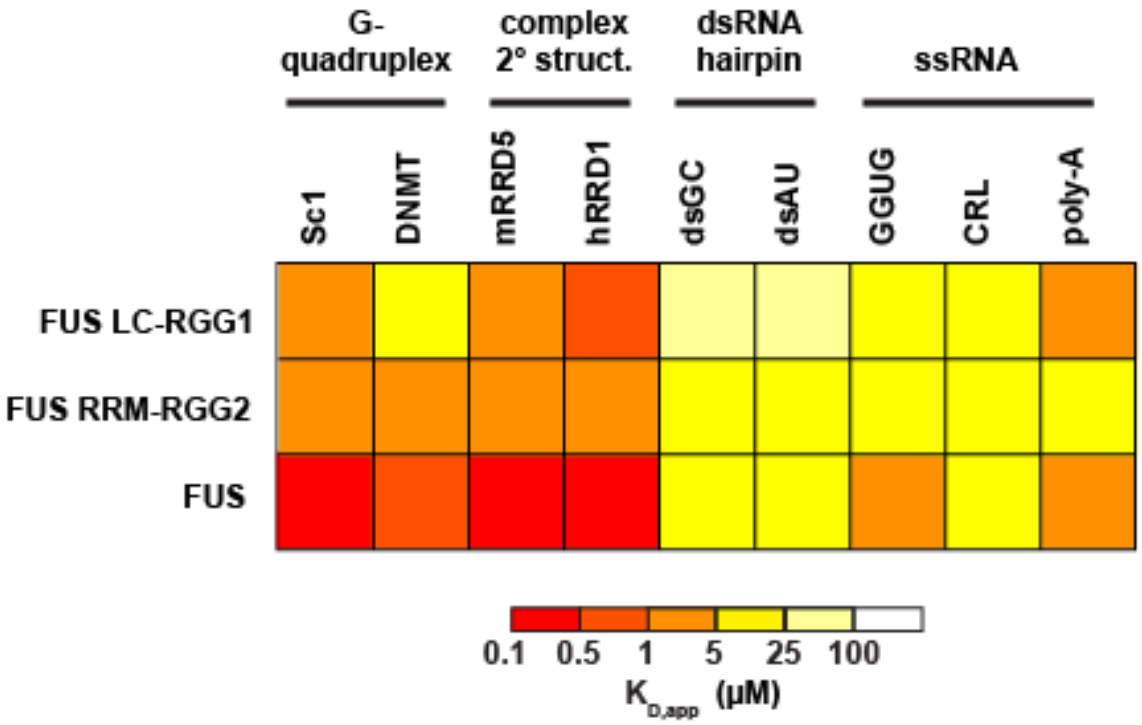

Figure 3.9 FUS protein preferentially binds to structured RNAs. Heat-map for affinity of LC-RGG1, RRM-RGG2 and full length FUS constructs to different RNAs. EMSA experiments were performed with DNMT and Sc1 RNAs containing G-quadruplexes, hRRD and mRRD RNAs with complex secondary structures, dsAU and dsGC simple double stranded RNAs, and GGUG, CRL, and poly-A single stranded RNAs. $K_{D,app}$ (μM) was represented with a color code for all combinations of RNA-protein with a data range from two or more independent experiment.

3.12). To complement DMS, I also tried iodine footprinting for the same samples. hRRD RNA was transcribed with adenosine and cytosine 5`-*O*-(1-thiotriphosphate) as described in the methods. Iodine reacts with the phosphorothioate groups of these nucleotides and the RNA molecules are cleaved at these sites. But, they can be protected if protein binds to those regions. Iodine footprinting gave better-protected sites than DMS foot printing (Figure 3.11 and Figure 3.12), but still these are weak protections compared to what is observed in highly sequence and/or structure specific RBPs (222). The observed weak footprints might be because of the intrinsically disordered nature of FUS protein. Since those proteins have high off-rates, this can prevent an efficient protection on the RNAs. Another reason might be that there is no

single site that is highly protected; the weak sites reveal preferential binding sites but that the selectivity for this site is sufficiently low that only a fraction of RNAs are occupied by FUS at this site. DMS and iodine footprinting results revealed that FUS protein binds to the multiple sites on the hRRD RNA.

Domain structure of FUS is like “beads on string” since there are two globular domains (RRM and ZnF) separated by intrinsically disordered RGG2. RGG1 and RGG3 are placed at the N- and C- termini of the protein. I wanted to test whether FUS could bind to two RNAs at the same time that might provide FUS a crosslinking activity during the cellular processes such as membrane-less organelle formation. To test this hypothesis, I analyzed the stoichiometry of FUS binding to Sc1 RNA that is the shortest and tightest binder that I have. I used composition gradient-multiangle light scattering (CG-MALS) that employs static light scattering to determine stoichiometry and equilibrium constants of self and hetero associations. The intensity of the light scattered from macromolecules in solution is proportional to the product of the concentration (3.13 B) and the weighed average (3.13 C). I should note that there is not any repeat of Figure 3.13, and thus, following results are not certain. According to the CG-MALS analysis, FUS is mostly monomer at low protein concentrations; however, it forms dimer structures at higher protein concentrations (Figure 3.13). Interestingly, Sc1 also forms dimer structures that were not reported in any previous studies. This might be a technical artifact since FUS protein is sticky and any protein contaminant will affect the light scattering of RNA titrations. Overall, CG-MALS data suggest that two FUS proteins interact with one Sc1 RNA and each protein binds to Sc1 with 1 μM binding affinity. Indeed, this data is consisted with a previous study performed with hnRNPA1 protein (256). In that study, they showed that two RNA binding to one-protein phenomena is almost impossible since the K_D of second RNA binding to hnRNPA1-RNA complex is around molar range due to the entropic penalty.

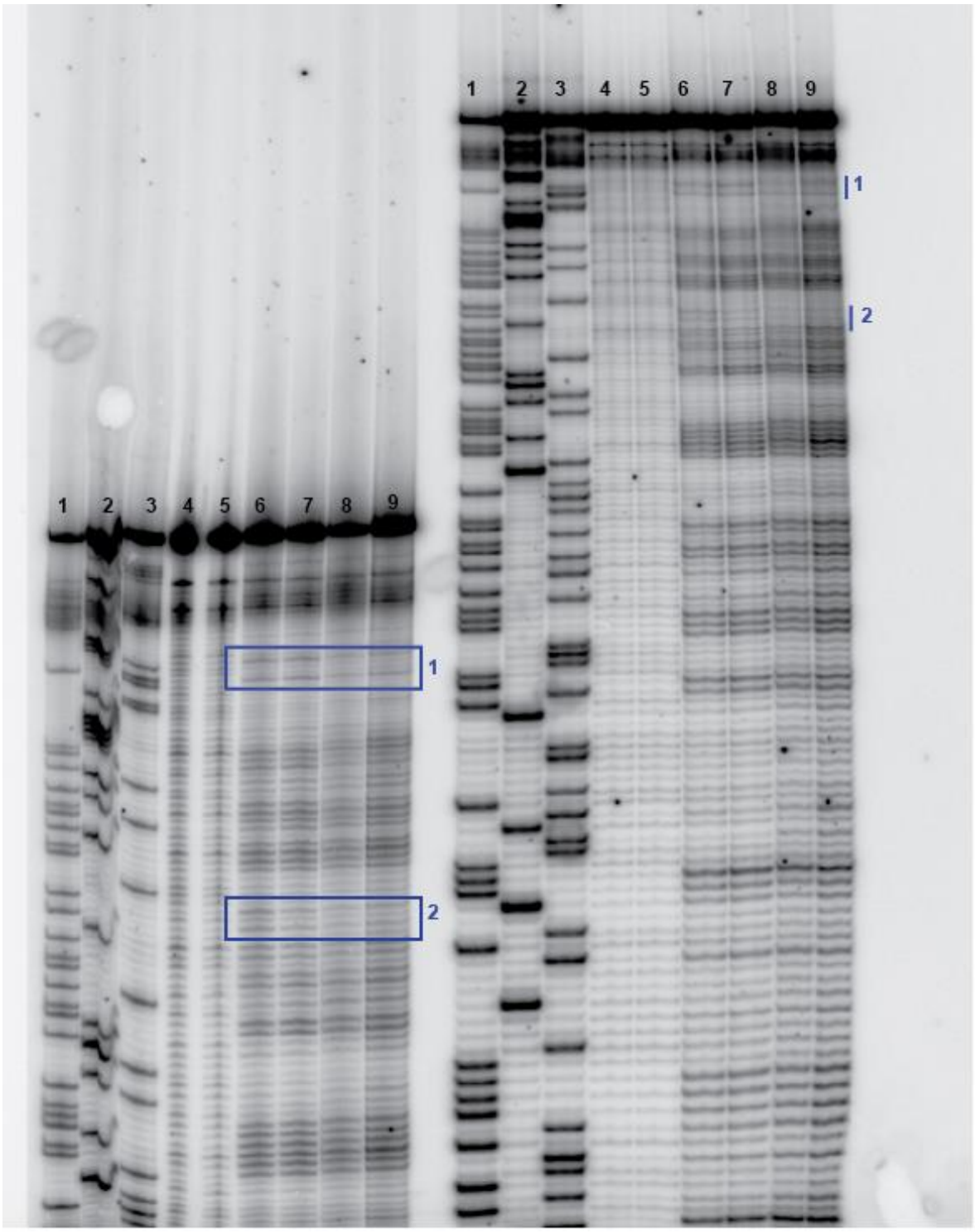

Figure 3.10 Dimethyl sulfate footprinting analysis of FUS and hRRD interaction. 7 hours (left) and 3.5 hours (right) gel runs. Lane 1: ddTTP sequencing, lane 2: ddGTP sequencing, lane 3: ddCTP sequencing, lane 4: hRRD alone, lane 5: hRRD+FUS alone, lane 6: hRRD+DMS, lane 7: DMS footprinting of hRRD+FUS (1:1), lane 8: DMS foot printing of hRRD+FUS (1:5), lane 9: DMS footprinting of hRRD+FUS (1:15). Potential protected sites were shown with blue boxes.

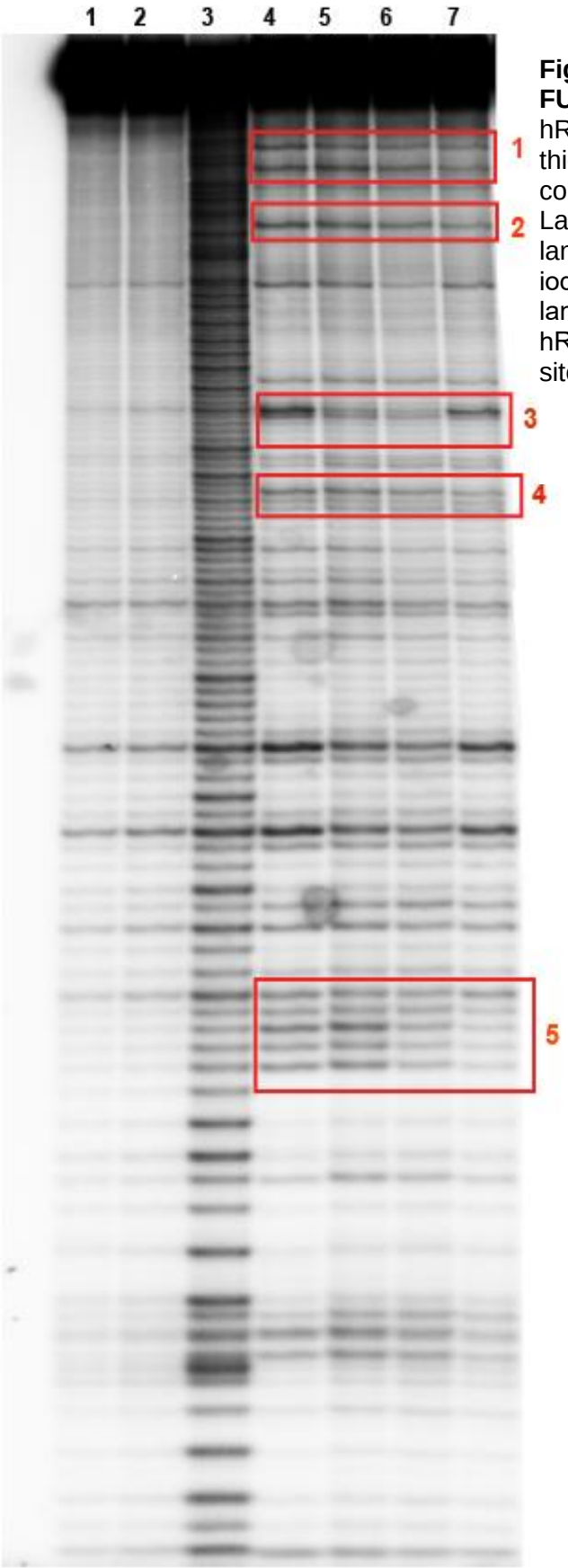

Figure 3.11 Iodine footprinting of hRRD and FUS interaction. Trace amount of 5'-end labeled hRRD (containing adenosine and cytosine thiotriphosphates) was incubated with increasing concentrations of FUS and probed with iodine. Lane 1: hRRD alone, lane 2: hRRD+FUS alone, lane 3: Alkaline hydrolysis, lane 4: hRRD+1 mM iodine, lane 5: hRRD+ 1 μM FUS+ 1 mM iodine, lane 6: hRRD+ 5 μM FUS + 1 mM iodine, lane 7: hRRD+ 15 μM FUS + 1 mM iodine. Protected sites were shown with red boxes.

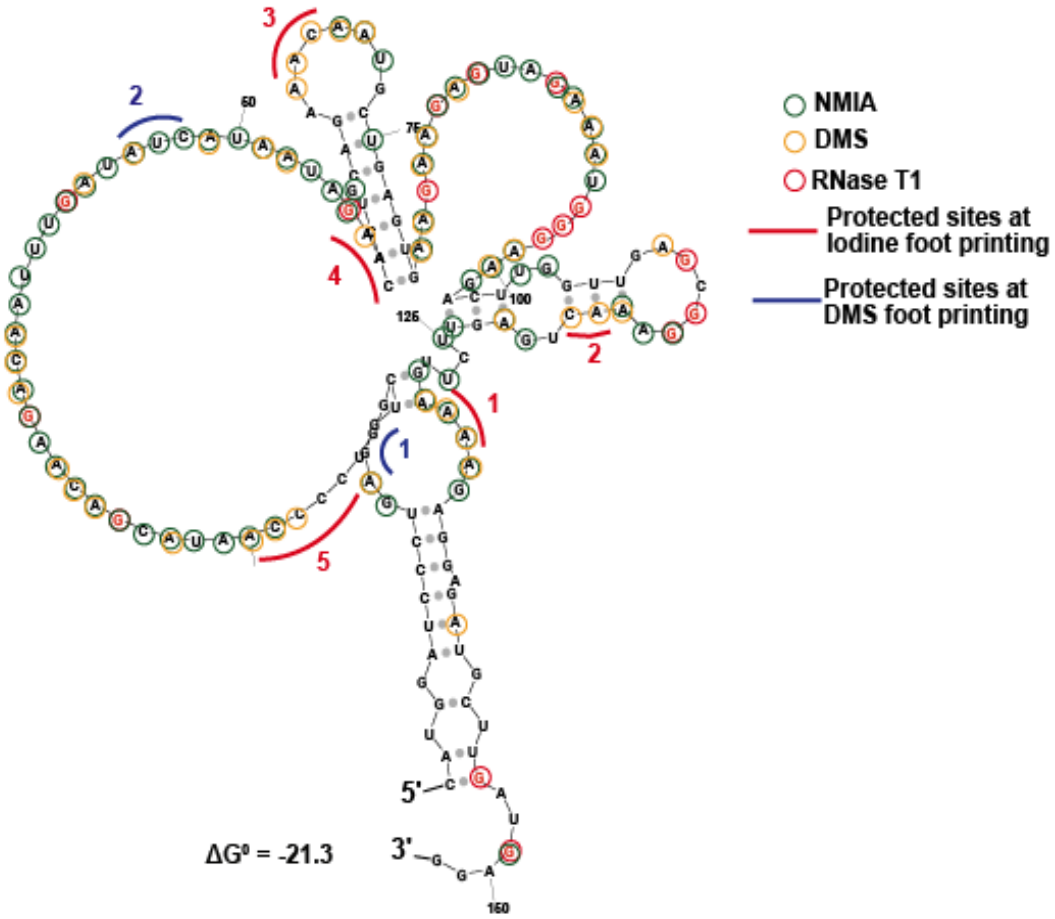

Figure 3.12 FUS binding sites on the predicted secondary structure of hRRD. hRRD secondary structure was predicted by mfold. Probed sites of hRRD RNA by DMS, RNase T1 (gel data not shown) and NMIA (gel data not shown) were shown as colored circles. Corresponding sequences to the protected sites at iodine foot printing (red line) and at DMS footprinting (blue line) were represented by the same numbers with the gel data.

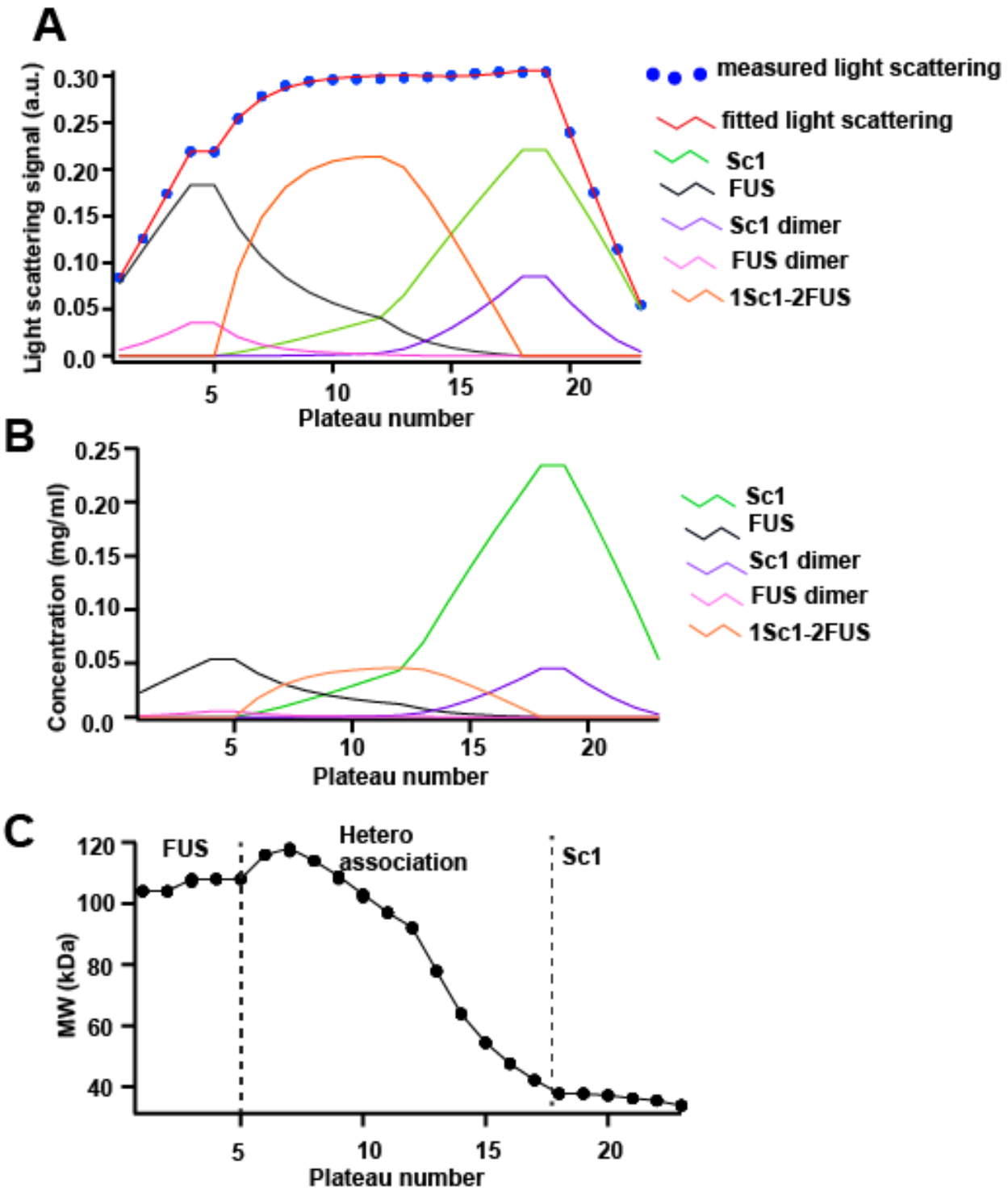

Figure 3.13 Two FUS proteins bind to Sc1 RNA. (A) Light scattering of FUS-Sc1 interaction, corresponding (B) concentration and (C) molecular weight values were represented.

3.3 Discussion

In the last decade, a number of research studies were performed to characterize the RNA binding properties of FUS (234,237,238,247,248,257). However, these studies did not provide consistent data about the RNA binding domains and RNA targets of FUS. Here, my detailed domain analysis indicated that RGG domains of FUS are crucial for the RNA binding activity both *in vitro* and *in vivo*. In addition, the data suggested the presence of synergistic interactions between the domains of FUS. Moreover, the RRM domain that was previously shown not to have a high binding affinity to RNAs can recover RNA binding activity upon addition of three RGG repeats to the C-term of it. I also investigated the RNA binding preferences of FUS by comparing its binding affinity against different sets of RNA molecules. My data revealed that FUS binds to RNAs with secondary structure elements with a better binding affinity than the single stranded RNAs. This study's observations suggest that the RNA-binding behavior of FUS is more accurately described as "degenerate specificity", as we define features important to binding but those can be exhibited by a large number of RNA sequences and structural contexts throughout the transcriptome.

Degenerate specificity was first used to describe flexible (but not non-specific) binding behavior of T cell receptors, since they recognize a large number of different peptides with no amino acid homology (258). There are many examples for protein-protein interactions including kinases, phosphatases, and even glutathione peroxidases that have also been defined as having degenerate specificity (259,260). I propose that the current view of "promiscuous binding" of many RNA binding proteins may be better described by the term "degenerate specificity". This speaks to the differences between individual domains with individual binding characteristics and the complex interactions between domains. These properties can only be fully revealed when the isolated activities of each domain are compared to that of the full-length protein.

Comparison of analogous interactions between intrinsically disordered domains and well-folded protein targets show parallels to these data. As one example, the disordered proline-rich domain (PRD) of NEF weakly binds to the Hck SH3 (K_d = 91 μM) but full length NEF binds the Hck SH3 domain with >300-fold greater affinity due to a reduction in the disorder of the PRD, thereby paying part of the entropic penalty of binding (261,262). Analogous to this, we observe that RGG2 bound RNA only weakly ($K_{D,app}$ = 61 μM), but with the added structural stability of the folded RRM the affinity was greatly enhanced ($K_{D,app}$ = 2.5 μM) beyond that of the RGG1 domain alone. It should be noted that the FUS protein is driven into phase-separation upon RNA binding and these data suggests that a conformation change upon binding may provide an underlying mechanism to this model (74,75,170,244,246,263,264).

The RGG/RG domain is the second most common recurrent RNA binding domain (94,265). This investigation of the putative RNA binding domains of FUS has revealed these domains to be the principal drivers of RNA binding *in vitro* and critical for RNA binding in cells. My results provide new details regarding degenerate RNA-binding that contrasts from or compares favorably to conclusions drawn by recent comprehensive studies. A proteome-wide map of RNA-binding sites, RBDmap, recovered several novel examples of RGG/RG peptides bound to cellular RNAs and RGG3 of FUS was chosen for validation as representative of these disordered RNA-binding domains (132). Published iCLIP and CLIP-seq studies for FUS report enriched GU-rich motifs (233,234), which was also a consensus found by a SELEX study (250), but a third HITS-CLIP study found an AU-rich consensus (237), an *in vitro* competitive binding assay, RNAcompete, found a GC-rich motif bound to FUS RRM (107), and remaining studies report no consensus (171,235,236). While reported sequence motifs have no agreed sequence identity between them, our results support that FUS does show preference for G-rich RNAs, which may partly result from arginine hydrogen bonding to guanine.

In sharp contrast with our findings, structure-predictive analysis of iCLIP data concluded

that sites bound by FUS were predominantly single-stranded (234). Two other studies conclude that FUS targets tend to lie in structured regions of RNA (235,237). Recently improved *in vivo* SHAPE techniques have been able to correlate evidence of complex secondary structures to specific FUS binding sites revealed through CLIP approaches (266,267). Indeed, the agreement of our work with the conclusions inferred by SHAPE-MaP would suggest that this approach currently holds the lead in resolving demonstrable binding specificity for this protein.

Taken together, there are no clear winners among comprehensive approaches to resolving the complex, degenerate binding specificity of a disordered RNA-binding protein, such as the FUS described in this work. Caution remains the key when interpreting transcriptome-wide binding profiles. My results clearly show that validation of structure-function relationships within and without cells remains essential to provide context for interpreting global RNA-binding data. Future work should likely focus on revising overly simplified assumptions employed to interpret and extrapolate meaning from such sequencing-based studies, which do not necessary hold true for novel or underappreciated modes of specificity employed by many RNA-binding proteins.

In this work, I provided detailed analysis of RNA binding properties of FUS. However, there are many questions remain to be answered for a detailed understanding of RNA binding behavior of FUS. My attempts to show RNA binding sites of FUS on hRRD RNA was inconclusive since we could not get a well protected sites by using traditional foot printing methods. *In vitro* foot printing with covalent crosslink might be a solution to see the protected sites if the reason is FUS binding kinetics. Moreover, using other probing reagents that target different nucleotides might be another solution if the problem is probing the nucleotides that are not involved in binding. Another question that needs to be answered with future experiments is whether individual RGG/RG domains of FUS have any specific RNA targets. This can be easily answered by the characterization of the RNA targets of individual SGG mutants. Besides, we do not know whether RGG/RG domains also mediate the degenerate specificity of FUS. RNA

binding mode of RGG/RG domains should be investigated in detail to get a better sense about RNA targets of these domains.

Chapter 4 : Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding

4.1 Introduction

The arginine/glycine-rich (RGG/RG) domain is prevalent throughout eukaryotes and the second most common RNA-binding domain (RBD) in the human genome (94,131,148,265,268,269). The importance of decoding RGG-mediated RNA recognition is underscored by the observation that these RNPs regulate all levels of RNA metabolism including transcription, RNA processing, nucleocytoplasmic shuttling and translation (148,170,269). Furthermore, mutations in RGG/RG proteins are implicated in several neurodegenerative diseases, including amyotrophic lateral sclerosis (ALS), fragile X syndrome, and spinal muscular atrophy (SMA) (148). Unlike the most common RBD, the RNA recognition motif (RRM), the RNA binding properties of RGG/RG domains are still poorly defined. A key challenge for understanding the cellular functions of the rapidly expanding set of RNAs comprising the transcriptome is to identify and characterize their interactions with RNA-binding proteins. This is particularly difficult for RNA-binding proteins that engage a large number of transcripts, suggesting “promiscuous” or non-specific binding (170,171,238,270-272).

RGG/RG domains are intrinsically disordered (131), and thus do not adopt a single, stable structure in the absence of RNA but instead have conformational plasticity and adaptability. This feature may facilitate flexible targeting to a variety of RNAs because their own conformational flexibility provides a larger interaction surface area. While for some proteins, such as hnRNP U, the RGG/RG domain is the only identified RBD (151), this motif is most often found in proteins possessing other RBDs such as RRM and KH (148,265,269).

The prevalence of RGG/RG domains among RNA-binding proteins has only recently been appreciated (101). Thus far, data suggest that RGG/RG domains may display some selectivity in RNA binding (101,157-159,273,274). For example, the RGG/RG domain of FMRP

was found to bind tightly to an *in vitro* selected aptamer, Sc1, containing a G-quadruplex structure (101,157-159). While the G-quadruplex is the dominant feature of the Sc1 aptamer, the RGG/RG peptide sits in the interface between the duplex and quadruplex through a shape complementarity interaction, with the majority of protein-base contacts mediated by two Watson-Crick G-C pairs in the duplex (101,159).

If the preference of RGG/RG domains were actually directed to a particular conformation involving double-stranded RNA, this interaction would not be completely non-specific. Instead, the structure/sequence requirements for binding could appear frequently throughout the transcriptome (170,238,270-272). This behavior may explain reports of binding specificity for well-ordered RBDs that are inconsistent with those of a protein harboring both RBDs and RGG/RG domains. The RBDs of these RGG/RG proteins, such as the RRMs of hnRNPA1 or hnRNPA2/B1, display a robust specificity *in vitro*, but these target motifs are found only in a minor fraction of experimentally identified target sites within the transcriptome (240,275,276). These studies, however, do not shed light on how RGG/RG domains facilitate the molecular recognition of a subset of cellular RNAs (240).

RGG/RG domains are prevalent in many if not most heterogeneous ribonuclear particle (hnRNP) proteins. A particularly prominent family of RGG/RG domain proteins and subset of the hnRNP family is the FET family, comprised of FUS, EWSR1, and TAF15 (170,237,277). Each FET protein has three RGG/RG domains interspersed between structured domains. FET proteins are conserved throughout metazoans. In humans, they are ubiquitously expressed in all tissues and are among the most highly expressed proteins in the cell (74,170,171). These three nuclear proteins predominantly associate with RNA Pol II and pre-mRNA, which affects both transcription and mRNA processing (171,233,278). In addition to having multiple RGG/RG domains, the number of individual RGG or RG repeat are high, with up to nine repeats. My previous study with FUS revealed that FUS has a degenerate specificity to RNA molecules and RGG domains mediate RNA binding activity.

To further define the RNA binding characteristics of RGG/RG domains, those of FUS were contrasted with those of two additional proteins FMRP and hnRNPU, which showed varied degrees of preference in binding to RNAs of different sequence and structure composition. Taken together, I revealed that specificity of RGG/RG domains is also degenerate like FUS protein, and provide a model for the significance of intrinsic disorder to RNA binding.

4.2 Results

4.2.1 RGG/RG domains display degenerate specificity

In the previous chapter, I showed that mutation of any one RGG/RG in FUS did not appreciably reduce affinity for the DNMT RNA and the same mutants lost nearly half of their binding to bulk, cellular RNA, I reasoned that individual RGG/RG motifs may have different binding selectivity, resulting from their binding plasticity (Figure 3.7 and 3.8). To characterize RNA binding properties of different RGG/RG domains, I expressed RGG/RG domains of FMRP, hnRNP-U and FUS as C-terminal MBP fusions to compare their binding to a set of RNAs representing different sequence and structural properties (Table A1). As a benchmark for high affinity binding, I included in this panel the only structurally characterized RGG/RG domain-RNA interaction: an RNA aptamer (Sc1) containing a hybrid G-quadruplex/A-form duplex structure that was obtained in an *in vitro* selection for FMRP (101,157,159). As representative single stranded RNAs (ssRNAs), I employed three different RNAs: a 40 nt polyadenosine homopolymer (poly-A), a 25 nucleotide guanosine-rich RNA that was obtained in an *in vitro* selection against FUS (GGUG) (250), and a heterogeneous 36 nucleotide RNA (CRL) containing five repeats of the sequence (AUACAAC) (251). To represent dsRNA, two simple hairpins were used, one containing a hairpin exclusively composed of 12 A-U pairs (dsAU) and a second with a tract of six G-C pairs in the middle (dsGC) and each capped with a UUCG tetraloop. Finally, as a representative of an RNA with mixed secondary structure containing multiple helices and single stranded (ss-) elements, the repeating RNA domain (RRD) from

human (hRRD) and mouse (mRRD) Firre lncRNA, reported targets of hnRNPU, was chosen (190).

To validate the secondary structures, RNase T_1 nuclease footprinting was performed on selected RNAs (Figure 4.1A-C). G-quadruplexes can be revealed by comparison of the cleavage pattern in lithium and potassium buffers, as lithium does not support quadruplex formation (279,280). These experiments confirmed the structure of most model RNAs used in this study. Unexpectedly, these experiments revealed that the DNMT RNA, which was proposed to have no stable secondary structural features, forms a G-quadruplex at its 3'-end, likely through the association of multiple RNAs. Thus, DNMT RNA was considered by my analysis as a second example of G-quadruplex RNA, with some duplex character as well.

I measured the $K_{D,app}$ by EMSA analysis, revealing that each RGG/RG domain possessed different RNA-binding characteristics but had an overall similar pattern of behavior (Figure 4.2, Table A4). The highest affinity interaction I observed was between FMRP-RGG and Sc1 RNA (0.09±0.02 μM). FMRP-RGG bound all of the RNA sequences with $K_{D,app}$ values spanning a 500-fold range (Figure 4.2, Table A4), showed a strong preference for G-quadruplex containing and complex RNAs, with the lowest affinities for ssRNA. This is consistent with the solved structures that show direct interactions with several G-C pairs adjacent to the G-quadruplexes (101,159). Similarly, the hnRNPU-RGG, FUS-RGG1 and FUS-RGG3 displayed a stronger preference for G-quadruplex and complex RNAs than simple hairpins or ssRNAs. The FUS-RGG2 domain had the lowest affinity for any of the RNAs tested, despite having as many RGG repeats as FMRP or hnRNP-U (Figure 4.3). I concluded that these domains have more complex functional requirements than just a critical amount of RGG repeats to impart RNA binding.

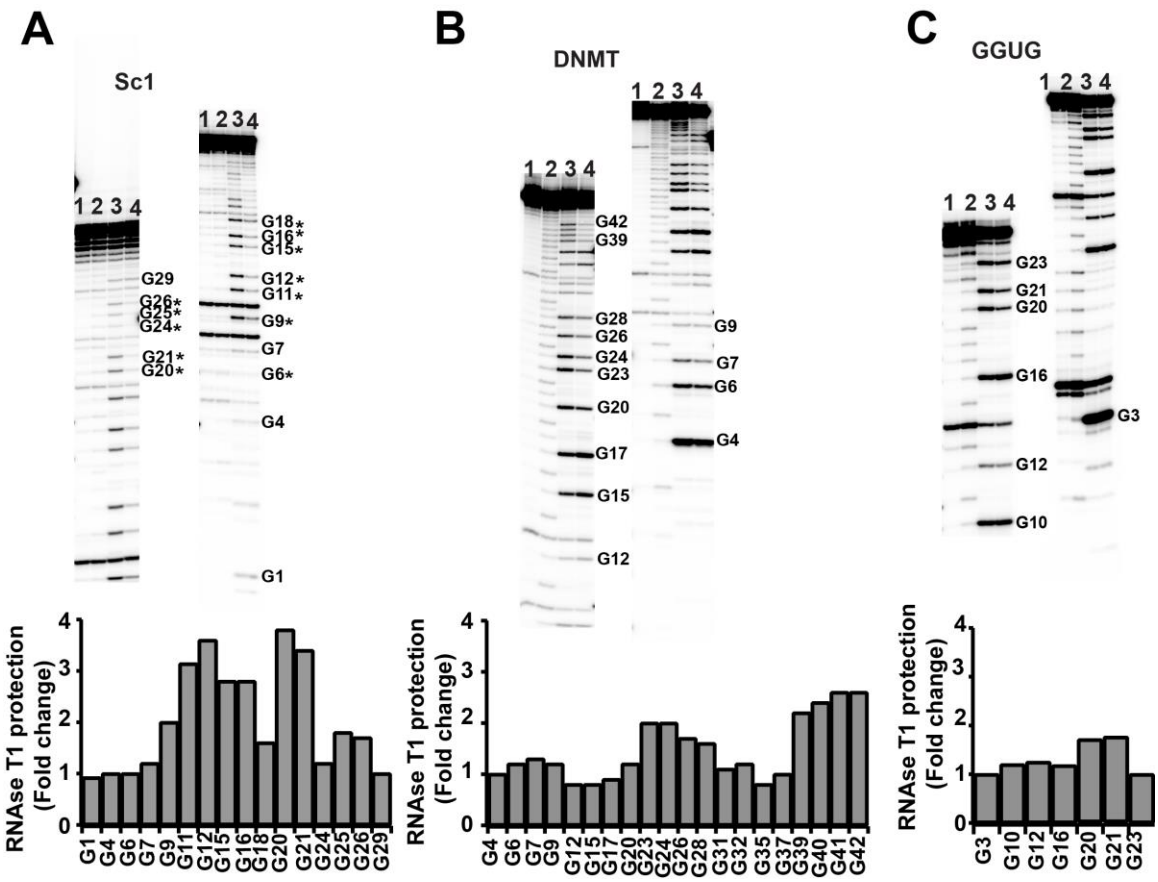

Figure 4.1 G-quartets formation in Sc1, DNMT and GGUG RNAs. (Top) Representative gels with RNAseT1 digestion of (A) Sc1, (B) DNMT and (C) GGUG RNAs in KCI and LiCI buffers. (Bottom) Bands were quantified by Image Quant and the amount of protection for each band (fold change in band intensity, KCI/LiCI) was calculated. The Guanines involved in G-quadruplex formation in Sc1 RNA were shown with '*' symbol. Lane 1 is RNA alone (No RNase T1), lane 2 shows alkaline hydrolysis of corresponding RNA, lane 3 and lane 4 is RNAse T1 reaction in LiCl and KCl, respectively.

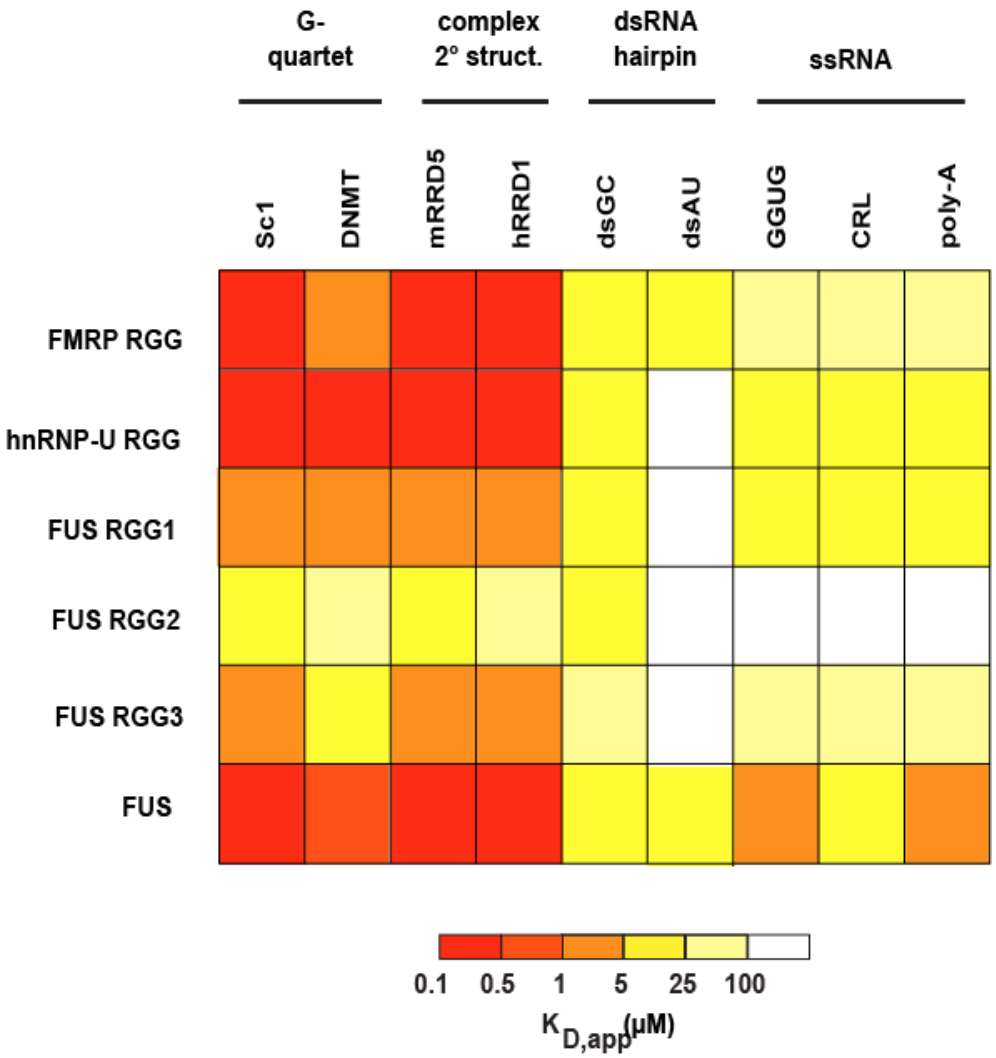

Figure 4.2 RGG/RG domains represent moderate preference for structured RNAs. Heat-map for affinity of RGG/RG domains and FUS for each RNA. EMSA experiments were performed with DNMT and Sc1 RNAs containing G-quadruplexes, hRRD and mRRD RNAs with complex secondary structures, dsAU and dsGC simple double stranded RNAs, and GGUG, CRL, and poly-A single stranded RNAs. $K_{D,app}$ (μM) was represented with a color code for all combinations of RNA-protein with a data range from two or more independent experiments.

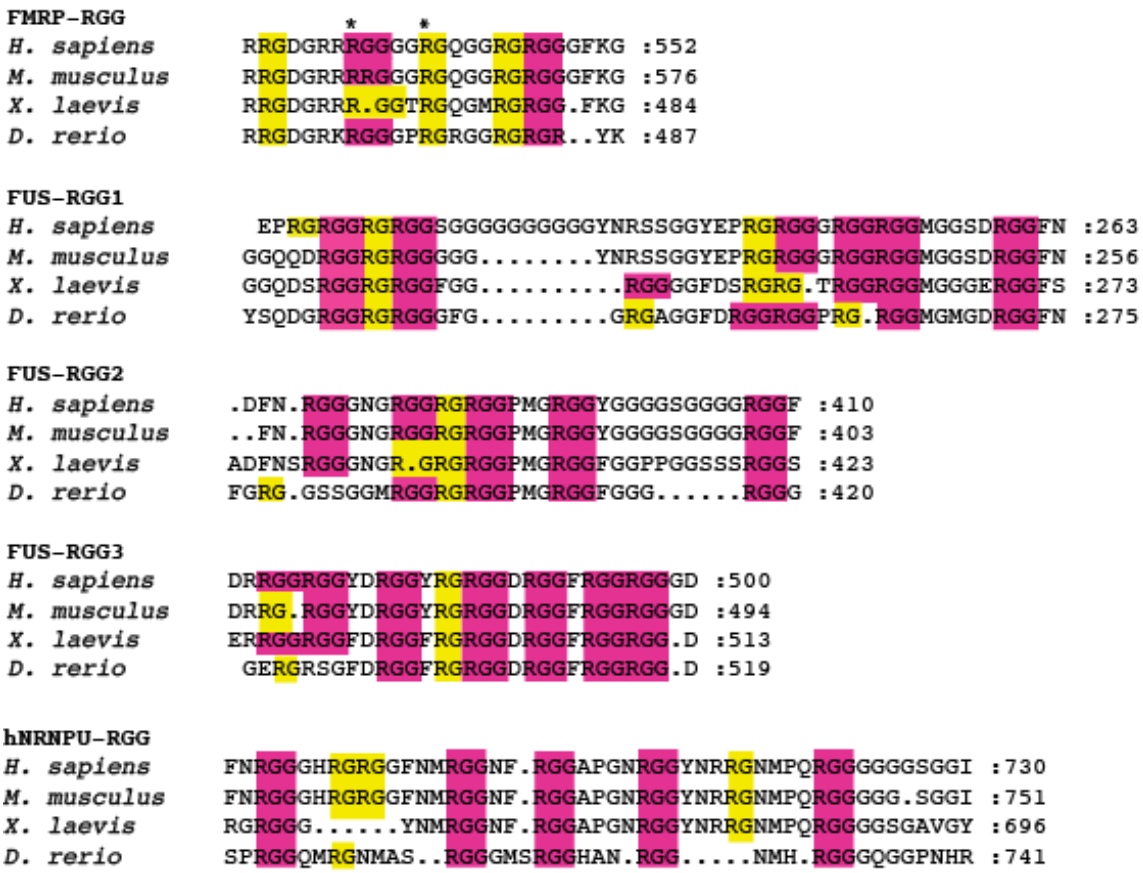

Figure 4.3 RGG/RG domains and their amino acid sequences are conserved across diverse metazoan species. The amino acid sequences of RGG domains of FUS, FMRP and hnRNPU from human, mouse, xenopus and zebrafish were aligned. Pink and yellow colors show conserved RGG and RG repeats, respectively. "*" symbol shows the arginine amino acids of FMRP-RGG that directly make hydrogen bonds with guanines in Sc1 RNA.

4.2.2 G-quadruplex is not requisite for RGG/RG binding

G-quadruplexes have been proposed to be a major biological target of both FMRP and FUS (157,281,282). This perspective is supported by our observation that Sc1 RNA bound strongly to all RGG/RG domains tested. To evaluate RGG/RG domains affinity for G-quadruplex structures, I compared binding affinities of Sc1 RNA to RGG/RG peptides in buffers with either KCl or G-quadruplex disrupting LiCl. In LiCl, FMRP-RGG bound to Sc1 with ~30 fold lower affinity, consistent with previous studies (Figure 4.4) (101,157,159). Similarly, hnRNP-U RGG binding decreased ~8 fold in the presence of LiCl. The RGG1 and RGG2 domains of FUS did not exhibit any difference in the binding affinity, but the RGG3 domain exhibited a ~3 fold lower

affinity. The binding affinity of full length FUS to Sc1 did not change in LiCl containing buffer, indicating that the G-quadruplex of Sc1 is not required for binding by the full-length protein.

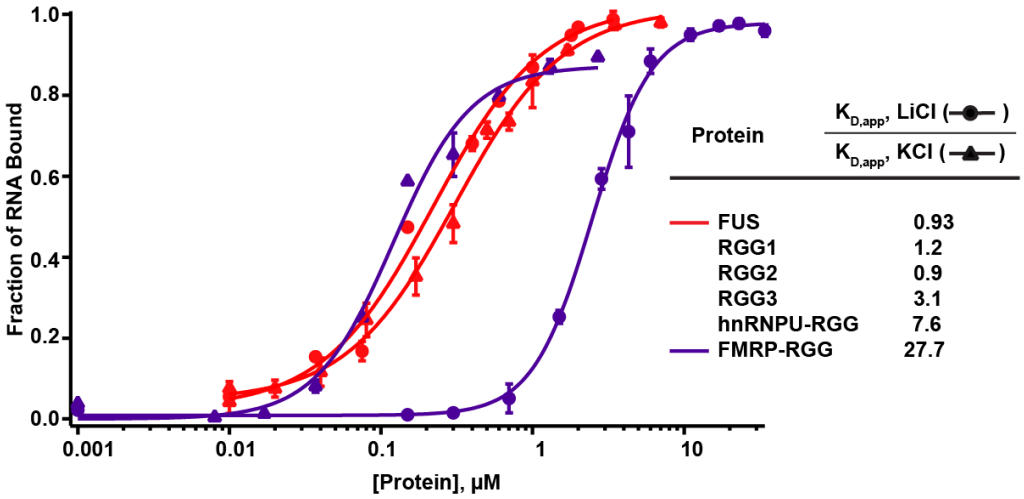

Figure 4.4 Effects of G-quartet structure on RNA binding activity of different RGG domains. Binding of RGG domains to Sc1 RNA was measured in the presence of KCl and LiCl by EMSA. Representative binding curves of FUS and FMRP-RGG binding to Sc1 RNA was shown and fold change in the dissociation constant of RGG domains for LiCl and KCl buffers were presented. Error bars represent the S.D. of at least two independent titrations for each construct.

To determine whether other RGG domains have the same base-specific contacts with Sc1 RNA as observed in the bound structures with FMRP-RGG, I created individual mutants of the nucleotides (G7A-C30U and C5U-G31A) that disrupt hydrogen bonding with FMRP-RGG without disrupting RNA structure (101,159). Results of binding to mutant Sc1 RNAs were diverse for different RGG domains (Table 4.1). Consistent with previous work, FMRP-RGG binding affinity decreased around 25-fold for each mutant compared to the wild type (101,159). Other RGG domains showed a lesser response from a 10-fold lower affinity to almost no change. Wild type FUS displayed a moderate sensitivity to the G-C to A-U pair mutations, with a ~5.5 fold decrease in binding affinity.

Table 4.1 Effects of mutations in Sc1 RNA on RNA binding activity of RGG domains.

	$K_{D,relative}$ ($K_{D,app,\ mutant}$ / $K_{D,app,\ Sc1}$)	
Protein	Mutant C5U-G31A	Mutant G7A-C30U
FMRP-RGG	28	24
hnRNPU-RGG	11	10
FUS-RGG1	4	2
FUS-RGG2	1.5	1.5
FUS-RGG3	4.5	3.8
FUS (WT)	5.5	5

4.2 Discussion

An RGG/RG domain is the second most common RNA binding domain whose RNA binding activity is assumed to be nonspecific. However, my previous studies with FUS indicated that it has a degenerate specificity to RNAs and RGG/RG domains mediate the RNA binding activity. Thus, we investigated the RNA binding modes of RGG/RG domains of FUS, hnRNP U and FMRP proteins to see whether RGG/RG domains alone have any RNA binding preferences. This study has illustrated that RGG/RG domains present an unexpected degree of flexibility in their recognition of RNA sequences and structures. However, these domains are not wholly indiscriminate, as they prefer GC-rich sequences and complex RNA structures featuring double-stranded helices. Similarly, the behavior of these RGG/RG proteins in cells has been shown to not be wholly nonspecific (107,170,171,234-238). This study revealed that RGG/RG domains have degenerate specificity to RNAs that supports our previous conclusions about RNA binding mode of FUS protein.

A broadly important question concerning these data is why the second most common RNA-binding domain should be an intrinsically disordered domain. These domains lack sufficient numbers of hydrophobic residues to allow a compacted core to form; thus, the peptide chain is largely solvent exposed and allowed freedom to sample many conformational states over short time scales (283-285). I find that RGG/RG preferences in binding are strongly biased against single-stranded sequences, simple A-form RNA helices, or dsDNA (238,286,287).

Instead, the preference of RGG/RG domains to bind to the complex FIRRE elements with higher affinity than simple hairpins and ssRNAs suggests that these domains interact with a limited and relatively common core element, likely tandem G-C pairs, in the context of more complicated or heterogeneous secondary structure. Taken together, these data suggest that the intrinsically disordered property of the RGG/RG domains imparts a great degree of plasticity providing numerous complex conformations with which to associate with helical RNA targets in a variety of structural contexts (99,101,159,288). This property provides a mechanism by which the FUS protein can bind a wide variety of RNAs as demonstrated by a number of CLIP-seq experiments (75,170,171,232-238,289). Taken together, the RGG/RG interactions with RNA described here are examples of intrinsically disordered domains that preferentially bind well-structured partners over less structure (97,131,159,259,288,290,291).

My establishment of a degenerate mechanism of binding may be more broadly representative of RGG/RG domain interactions with RNA. Structures of FRMP-RGG bound to Sc1 RNA reveal a β-hairpin inserted into the major grove, as well as hydrogen bonding with guanine bases (101,159). A similar type of recognition was also observed in an arginine rich peptide derived from bovine immunodeficiency virus (BIV) tat protein bound to the transactivation domain (TAR) of the genomic RNA (138). In my findings, the combination of hydrogen and ionic bonds as shown by FMRP-RGG presents an attractive explanation for the salt titration data. In light of a recent study indicating that the G-quadruplex forming regions in eukaryotic cells are overwhelmingly unfolded *in vivo* (292), the role of the G-quadruplex in Sc1 is likely to provide an unusual structure that opens the major groove for recognition. Thus, it is in fact the perturbation of the A-form helix that is particularly accommodating for RGG/RG insertion into the deep major grove. Consistent with this paradigm of selectivity for helices perturbed by local features, a recent study correlating protein-RNA binding to secondary structure within cells has concluded that FUS prefers dsRNA features adjacent to non-canonical regions (266,267). Considering the study that suggests the mRNAs were less structured or more dynamic *in vivo*

(293), RGG/RG domains interaction with structured RNAs might be regulated by dynamic changes in structures of the RNA molecules during RNA processing.

Taken together along with previously published models, I propose that conformational flexibility combined with degenerate specificity of RGG/RG domains can confer new RNA-binding activity to RBPs in at least three ways. First, degeneracy in RNA sequence recognition, particularly in the context of higher order homogeneous or heterogeneous complexes of RBPs, may allow repeated binding of RGG/RG domains in tandem along an RNA molecule (Figure 4.5A and B). CLIP-seq studies identifying cellular targets have suggested oligomerization of RGG/RG proteins along pre-mRNAs, particularly for FUS as well as by RGG/RG containing hnRNPs, including hnRNPA1, A2/B1, and hnRNP-U (75,170,171,233,238,240). Second, the lack of robust folding for RGG/RG domains can afford the protein flexibility to bind a variety of structural conformations of RNA (Figure 4.5C). It is particularly well supported because all RGG/RG domains tested here show their highest affinity for perturbed dsRNA elements as well as highly structured RNAs comprised of multiple dsRNA and ssRNA features. Third and finally, while not demonstrated in this study, the possibility has not been ruled out that long, flexible RGG/RG domains, particularly those interspersed through the RBP, can allow multivalent interactions with more than one RNA (Figure 4.5D). Such interactions might form a crosslinked RNP matrix such as those suggested to comprise non-membrane bound organelles, including p-bodies, stress granules, and nucleoli (207,243,269,274,294).

The exceptionally broad functions of many proteins containing RGG/RG domains can be considered to be consistent with a degenerate specificity in their RNA recognition. Many hnRNP proteins possess RGG/RG domains and their function as members of ribonuclear particles is to broadly coat pre-mRNAs while they are processed within the nucleus (295,296). FUS protein regulates transcription for thousands of genes in the cell while also requiring RNA binding to

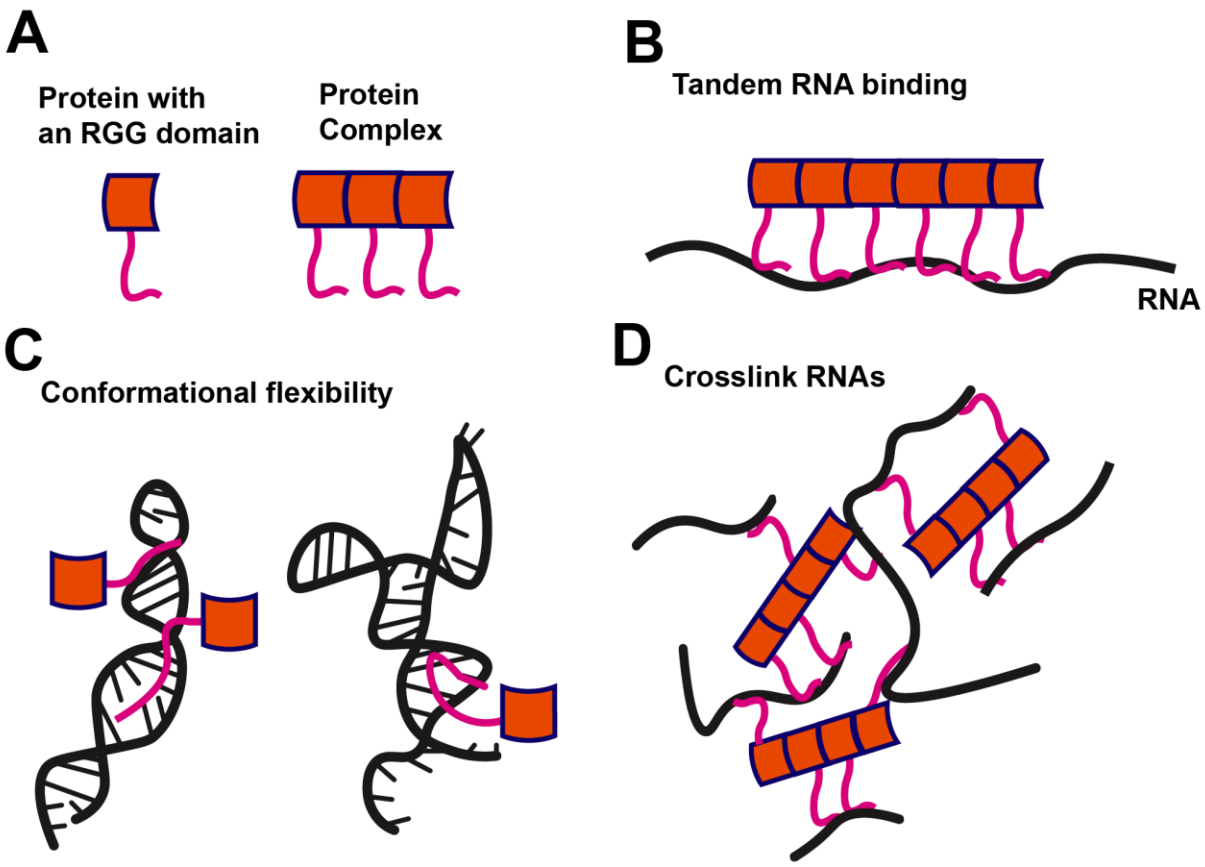

Figure 4.5 Model for RNA recognition by RGG/RG domains. (A) RGG/RG proteins may bind RNA in a 1:1 interaction or as a member of a higher order complex. (B) Multiple RGG/RG domains as part of a larger protein complex may recognize long RNA sequences. (C) RGG/RG proteins have the flexibility to accommodate and bind tightly to a variety of complex RNA structures. (D) Higher order RGG/RG proteins may allow binding to multiple RNAs at the same time.

trigger binding to the C-terminal domain of RNA Pol II (75,170,171). Other examples of RGG/RG domain proteins include Scd1, Sbp1, Npl3 and Ded1, which broadly affect translation (269,274). In each case, if the protein were considerably more selective in binding to RNA, its ability to act on a large number of RNAs throughout the transcriptome would become limited. RGG/RG domains might, in conjunction with other domains, broaden RNA-binding specificity, like the RGG/RG domain in FMRP broadens RNA-binding specificity beyond that typical of KH

domains (157). Future research should reveal how subtle differences in the degenerate specificity of RGG/RG domains might confer broad shifts in populations of RNAs targeted, a property that might be inferred from the high conservation of each RGG/RG domain throughout vertebrates.

Chapter 5 : Conclusions and Future remarks

5.1 Digging deeper into the world of the disordered RNA binding domains

Until now, many experimental data showed that intrinsically disordered regions of proteins are functional in various cellular processes such as RNA metabolism, cell cycle regulation, assembly formation, and virus infection. These regions are also abundant at the RNA binding surfaces of the proteins since they provide flexibility and plasticity that are crucial properties for high affinity and specificity in the RNA interaction. The most detailed information about the RNA binding properties of intrinsically disordered proteins comes from the studies with arginine rich disordered domains. These domains transition from disordered to ordered structure upon binding to target RNAs. They have very simple amino acid sequence but they can target RNAs quite diversely and often specifically that suggests the existence of vital roles of intrinsically disordered RBDs in the RNA binding activity of RBPs.

In this thesis study, we investigated the RNA binding properties of RGG domains, the second most common RNA binding domain in human genome. Previous studies with these domains suggested that they contribute to the RNA binding activity of other RNA binding domains in a non-specific manner. However, I have shown that RGG domains have degenerate specificity in RNA binding by targeting structured RNA molecules using five different RGG domains from three different well-characterized RNA binding proteins. In addition, nine RNA species that are diverse in both sequence and structures were used to test the RNA binding specificity of the RGG domains. All of the RGG domains showed better binding affinity against structured RNA molecules. The sample size that we used in this study gave us a notion about the RNA binding behavior of RGG domains. However, future studies are necessary to understand whether this binding preference is also the same for large sets of RNA targets and whether each RGG domain has its own specific RNA targets.

Recent studies indicate that deep sequencing (pull down-seq) methods alone are not sufficient to determine the RNA targets of intrinsically disordered proteins, as discussed previously. A complementary approach is the application of high-throughput *in vitro* methods to provide clear information about RNA targets of RGG domains. Systematic evolution of ligands by exponential enrichment (SELEX, also known as *in vitro* selection) is a commonly used method to determine RNA binding site of a protein. Indeed, this method was already applied for RGG domain of FMRP protein (157). An aptamer with G-quadruplex structure was determined as a RNA target of FMRP. Detailed studies indicated that RGG domain of FMRP folds upon binding to Sc1 RNA and makes specific hydrogen bond via its arginine and glycine amino acids. On the other hand, a recent study with RNA G-quadruplexes revealed that RNA regions that can fold into G-quadruplexes *in vitro* are unfolded *in vivo* (292). This suggests that *in vivo* validation of the results of *in vitro* is essential for a reliable RNA target determination. In addition, mimicking the *in vivo* conditions and using RNA aptamers derived from human genome library will help getting more accurate data from SELEX. An *in vivo* selection strategy can also be applied to RGG proteins, but these methods also have their disadvantages (297).

My data indicates that each of the RGG domains work within different K_D range suggesting that they have different RNA recognition modes. It would be interesting to understand how specificity or degenerate specificity is generated for RNA sequence or structures by different RGG domains. First of all, smallest RNA molecules that bind with the tightest affinities should be determined by an appropriate method such as SELEX. Then, solution NMR spectroscopy can be used to identify the structure of the RNA-protein complex. As discussed in the first chapter, this method can also provide information about local disorder, folding upon binding and disorder in complex that will further help to characterize different RGG domains.

In this thesis work, I also investigated the RNA binding properties of RGG domains in a protein context. Previous studies suggested that RGG domains function by contributing to the

RNA binding activity of globular RNA binding domains instead of being a primary RNA binding domain of a protein. However, our mutation analysis revealed that RGG domains of FUS protein are the main drivers of RNA binding activity of the protein both *in vitro* and *in vivo*. Despite mutation of individual RGG domains did not affect the RNA binding activity of FUS *in vitro,* these mutations significantly decreased the bound RNA amount *in vivo*. This suggests the effect of other regulators on the RNA binding activity of the protein. Reduction in the bound RNA amount was almost similar for all three RGG mutants. However, we do not know whether the bound RNA species are also similar or whether there are specific RNA targets of individual RGG domains. Identification the cellular targets of each RGG mutant would contribute to the understanding of how RBPs with modular structure function in the cell: whether each domain binds to RNAs independently or all domains together create an interaction surface to achieve high affinity and specificity.

Deletion analysis of domains of FUS gave some interesting results about the RNA binding behavior of RRM and RGG domains. The RRM domain of FUS does not bind to RNA by itself, but addition of three RGG repeats from RGG2 domain to the C-term of RRM recovered binding to the RNA with an affinity close to the full length FUS. Since the RRM domain of FUS has a non-canonical RNP motif, understanding the effect of three RGG repeats on the RNA binding activity of RRM may reveal a new RNA binding mode for this family.

Cellular activities of intrinsically disordered proteins are usually regulated by posttranslational modifications and alternative splicing of disordered. There is a number of examples about the effect of alternative splicing on RNA binding activity of intrinsically disordered proteins (131). For example, *FMR1* pre-mRNA is alternatively spliced at the 3' end to give multiple FMRP isoforms (298). Isoform 3 (the isoform with the largest N-term deletion, but with full length RGG box) was shown to perturb side chain conformations of arginine residues of the RGG domain (299) and influence the RNA binding activity of its RGG domain and its cellular functions (300,301). Intrinsically disordered domains are structurally open

regions for posttranslational modifications (PTMs). RGG domains are the target sites of protein methyltransferases (PRMT) and, in some cases, kinases depending on the amino acid contents. Methylation of arginine amino acids and their effects on the function of the proteins have been reported in a number of studies (148). For example, several studies reveal a regulatory role of methylation in RNA binding activity of the proteins (160,302,303). Methylation of arginine amino acids is known to regulate interactions of RGG domains with both RNAs and proteins. Considering the diverse functions of these proteins in cellular processes, methylation seems to have a high potential as a main regulatory factor of RBPs in cell. There are many questions that remain to be answered about the regulation of RGG-RNA interaction in the cell. Are there different methylation patterns that regulate interactions with different molecules? Can methylation regulate specific interactions with RNAs? What are the effects of other PTMs? Can protein-protein interactions of RGG domain be an indirect regulatory mechanism for RGG-RNA interactions? Future studies are necessary to identify the details of the regulatory mechanisms that are involved in the control of RGG domain activity during various cellular processes.

5.2 Targeting the intrinsically disordered RNA binding proteins in diseases

Mutations in RBPs are known to affect the functions of the proteins in a fashion that causes sporadic or hereditary genetic disorders in particular cancers, neurological diseases and muscular atrophies. A recent study revealed that disease related mutations are often found in disordered regions of human RNA binding proteins as discussed in chapter 1 (7). These mutations usually lead to loss of normal function, gain of toxic function, aggregation and misfolding of disordered RBPs (304,305) that is associated with the pathogenesis of several human diseases, particularly neurodegenerative diseases.

Since the common involvement of IDPs in the pathogenesis of many human diseases, they have been considered as novel drug targets (306-312). Traditionally, protein targeted drugs

are designed to modulate the function of ordered proteins such as enzymes, cell surface receptors, nuclear hormone receptors etc. Thus, new approaches are necessary to develop drugs targeting the interactions of disordered proteins that get structured only upon binding to partners.

There are several strategies that are followed to develop drugs affecting functions of IDPs (308). (i) Disorder-based rational drug design in which the drug molecules mimic a critical region of the disordered partner and compete with this region to bind to its structured partner. For example, Nutlins are p53-Mdm2 interaction inhibitors and developed according to this principle (313). p53 is an important transcription factor that regulates the transcription of the genes involved in cell cycle regulation and apoptosis. Its loss of function results in cancer development. Mdm2 is an E3 ubiquitin ligase and prevents p53 from activating its target genes. p53 interacts with Mdm2 through its N-terminal unstructured region that obtains helical structure upon binding to Mdm2. Nutlins target the p53-Mdm2 interaction interface and helps accumulation of the active p53 that decrease the viability of the cells. (ii) Direct targeting of IDPs/IDRPs in which small molecules directly bind to disordered regions and prevent the activity of the IDPs. For example, c-Myc is overexpressed in different cancers and functions by associating with its partner c-Max (311,314). Monomers of these proteins are disordered, but undergo coupled folding and binding upon dimerization. Screening different drug libraries to target the c-Myc in cancers resulted in various small molecules. Some of them also bound to other transcription factors and further optimization helped to select a c-Myc binding small molecule (311). (iii) Another strategy is to target functionally misfolded proteins. Functionally misfolded concept is used to describe a mechanism preventing IDPs from unwanted interactions with non-native partners. IDPs are highly dynamic and fluctuate between different unfolded conformations when they are unbound. Analysis of IDP behaviors revealed that preformed binding elements might be involved in a set of non-native intramolecular interactions. Small molecules can be designed to target the misfolded IDPs in the ensemble to prevent

further biological interactions of them. (iv) The final strategy is targeting aggregating IDPs. Pathological aggregation of misfolded IDPs in the cell triggers a cascade of events eventually cause to the neurodegenerative diseases. Suppression of these aggregations by small molecules is the aim of this strategy. Recent studies identified a series of small molecules called “molecular tweezers” that target the toxic oligomers and aggregates of IDPs by specifically interacting the lysine amino acids (306,315). However, they are not effective if there is not a lysine in the protein.

Aggregation is also a commonly seen disease-causing mechanism among intrinsically disordered RBPs. For example, mutations at the RGG domains of FUS cause ALS and FTLD disease by forming aggregates in the cells. Recent studies implicate that the alterations in RNA metabolism can be detrimental to the neuron cells in ALS. This suggests that the loss of function of FUS may cause ALS. However, there is not concrete evidence about its exact role in the disease formation yet. In this work, I showed that RGG domains primarily mediate the RNA binding activity of FUS. This suggests that disease related mutations within the RGG domains might deleteriously affect the cellular function of FUS by altering RNA metabolism. RGG mutations can cause conformational selection that result in loss of function of FUS or without any conformational change, they can result in loss or gain of binding sites of FUS binding partners. Thus, further characterization of the effects of these mutations on protein activity can provide a target for the drug design.

Targeting the dysfunctional IDPs by drugs is a newly developing research area. For that reason, there is not an example for a drug targeted intrinsically disordered RBP yet. However, existing evidences are highly promising about the potential effect of this treatment strategy to many diseases.

References

1. Wright, P.E. and Dyson, H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol*, **293**, 321-331.

2. Uversky, V.N. (2015) Functional roles of transiently and intrinsically disordered regions within proteins. *FEBS J*, **282**, 1182-1189.

3. van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T. *et al.* (2014) Classification of intrinsically disordered regions and proteins. *Chem Rev*, **114**, 6589-6631.

4. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. and Jones, D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*, **337**, 635-645.

5. Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M. *et al.* (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, **149**, 1393-1406.

6. Kwon, S.C., Yi, H., Eichelbaum, K., Fohr, S., Fischer, B., You, K.T., Castello, A., Krijgsveld, J., Hentze, M.W. and Kim, V.N. (2013) The RNA-binding protein repertoire of embryonic stem cells. *Nature structural & molecular biology*, **20**, 1122-1130.

7. Castello, A., Fischer, B., Hentze, M.W. and Preiss, T. (2013) RNA-binding proteins in Mendelian disease. *Trends Genet*, **29**, 318-327.

8. Tompa, P. (2012) Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci*, **37**, 509-516.

9. Berlow, R.B., Dyson, H.J. and Wright, P.E. (2015) Functional advantages of dynamic protein disorder. *FEBS Lett*, **589**, 2433-2440.

10. Wright, P.E. and Dyson, H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol*, **16**, 18-29.

11. Uversky, V.N. (2002) What does it mean to be natively unfolded? *Eur J Biochem*, **269**, 2-12.

12. Campen, A., Williams, R.M., Brown, C.J., Meng, J., Uversky, V.N. and Dunker, A.K. (2008) TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept Lett*, **15**, 956-963.

13. Garner, E., Cannon, P., Romero, P., Obradovic, Z. and Dunker, A.K. (1998) Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization. *Genome Inform Ser Workshop Genome Inform*, **9**, 201-213.

14. Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J. and Dunker, A.K. (2001) Sequence complexity of disordered protein. *Proteins*, **42**, 38-48.

15. Weathers, E.A., Paulaitis, M.E., Woolf, T.B. and Hoh, J.H. (2007) Insights into protein structure and function from disorder-complexity space. *Proteins*, **66**, 16-28.

16. Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W. *et al.* (2001) Intrinsically disordered protein. *J Mol Graph Model*, **19**, 26-59.

17. Rauscher, S., Gapsys, V., Gajda, M.J., Zweckstetter, M., de Groot, B.L. and Grubmuller, H. (2015) Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J Chem Theory Comput*, **11**, 5513-5524.

18. Flock, T., Weatheritt, R.J., Latysheva, N.S. and Babu, M.M. (2014) Controlling entropy to tune the functions of intrinsically disordered regions. *Curr Opin Struct Biol*, **26**, 62-72.

19. Davey, N.E., Cowan, J.L., Shields, D.C., Gibson, T.J., Coldwell, M.J. and Edwards, R.J. (2012) SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. *Nucleic acids research*, **40**, 10628-10641.

20. London, N., Movshovitz-Attias, D. and Schueler-Furman, O. (2010) The structural basis of peptide-protein binding strategies. *Structure*, **18**, 188-199.

21. Hong, W., Jiao, W., Hu, J., Zhang, J., Liu, C., Fu, X., Shen, D., Xia, B. and Chang, Z. (2005) Periplasmic protein HdeA exhibits chaperone-like activity exclusively within stomach pH range by transforming into disordered conformation. *The Journal of biological chemistry*, **280**, 27029-27034.

22. Uversky, V.N. (2013) Unusual biophysics of intrinsically disordered proteins. *Biochimica et biophysica acta*, **1834**, 932-951.

23. Uversky, V.N. (2013) A decade and a half of protein intrinsic disorder: biology still waits for physics. *Protein science : a publication of the Protein Society*, **22**, 693-724.

24. Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M. and Obradovic, Z. (2002) Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573-6582.

25. Gardner, K.A., Moore, D.A. and Erickson, H.P. (2013) The C-terminal linker of Escherichia coli FtsZ functions as an intrinsically disordered peptide. *Mol Microbiol*, **89**, 264-275.

26. Tompa, P. (2002) Intrinsically unstructured proteins. *Trends Biochem Sci*, **27**, 527-533.

27. Magidovich, E., Orr, I., Fass, D., Abdu, U. and Yifrach, O. (2007) Intrinsic disorder in the C-terminal domain of the Shaker voltage-activated K+ channel modulates its interaction with scaffold proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 13022-13027.

28. Van Roey, K., Uyar, B., Weatheritt, R.J., Dinkel, H., Seiler, M., Budd, A., Gibson, T.J. and Davey, N.E. (2014) Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem Rev*, **114**, 6733-6778.

29. Tompa, P., Davey, N.E., Gibson, T.J. and Babu, M.M. (2014) A million peptide motifs for the molecular biologist. *Mol Cell*, **55**, 161-169.

30. Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N.P., Trave, G. and Gibson, T.J. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci*, **13**, 6580-6603.

31. Gsponer, J. and Babu, M.M. (2009) The rules of disorder or why disorder rules. *Prog Biophys Mol Biol*, **99**, 94-103.

32. Lee, C.W., Ferreon, J.C., Ferreon, A.C., Arai, M. and Wright, P.E. (2010) Graded enhancement of p53 binding to CREB-binding protein (CBP) by multisite phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 19290-19295.

33. Van Roey, K., Dinkel, H., Weatheritt, R.J., Gibson, T.J. and Davey, N.E. (2013) The switches.ELM resource: a compendium of conditional regulatory interaction interfaces. *Sci Signal*, **6**, rs7.

34. van der Lee, R., Lang, B., Kruse, K., Gsponer, J., Sanchez de Groot, N., Huynen, M.A., Matouschek, A., Fuxreiter, M. and Babu, M.M. (2014) Intrinsically disordered segments affect protein half-life in the cell and during evolution. *Cell reports*, **8**, 1832-1844.

35. Fishbain, S., Inobe, T., Israeli, E., Chavali, S., Yu, H., Kago, G., Babu, M.M. and Matouschek, A. (2015) Sequence composition of disordered regions fine-tunes protein half-life. *Nature structural & molecular biology*, **22**, 214-221.

36. Guharoy, M., Bhowmick, P. and Tompa, P. (2016) Design Principles Involving Protein Disorder Facilitate Specific Substrate Selection and Degradation by the Ubiquitin-Proteasome System. *The Journal of biological chemistry*, **291**, 6723-6731.

37. Housden, N.G., Hopper, J.T., Lukoyanova, N., Rodriguez-Larrea, D., Wojdyla, J.A., Klein, A., Kaminska, R., Bayley, H., Saibil, H.R., Robinson, C.V. *et al.* (2013) Intrinsically disordered protein threads through the bacterial outer-membrane porin OmpF. *Science*, **340**, 1570-1574.

38. Huang, Y. and Liu, Z. (2009) Kinetic advantage of intrinsically disordered proteins in coupled folding-binding process: a critical assessment of the "fly-casting" mechanism. *J Mol Biol*, **393**, 1143-1159.

39. Pontius, B.W. (1993) Close encounters: why unstructured, polymeric domains can increase rates of specific macromolecular association. *Trends Biochem Sci*, **18**, 181-186.

40. Liu, Z. and Huang, Y. (2014) Advantages of proteins being disordered. *Protein science : a publication of the Protein Society*, **23**, 539-550.

41. Mollica, L., Bessa, L.M., Hanoulle, X., Jensen, M.R., Blackledge, M. and Schneider, R. (2016) Binding Mechanisms of Intrinsically Disordered Proteins: Theory, Simulation, and Experiment. *Front Mol Biosci*, **3**, 52.

42. Shammas, S.L., Crabtree, M.D., Dahal, L., Wicky, B.I. and Clarke, J. (2016) Insights into Coupled Folding and Binding Mechanisms from Kinetic Studies. *The Journal of biological chemistry*, **291**, 6689-6695.

43. Jensen, M.R., Zweckstetter, M., Huang, J.R. and Blackledge, M. (2014) Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. *Chem Rev*, **114**, 6632-6660.

44. Habchi, J., Tompa, P., Longhi, S. and Uversky, V.N. (2014) Introducing protein intrinsic disorder. *Chem Rev*, **114**, 6561-6588.

45. Dyson, H.J. and Wright, P.E. (2004) Unfolded proteins and protein folding studied by NMR. *Chem Rev*, **104**, 3607-3622.

46. Receveur-Brechot, V., Bourhis, J.M., Uversky, V.N., Canard, B. and Longhi, S. (2006) Assessing protein disorder and induced folding. *Proteins*, **62**, 24-45.

47. Sibille, N. and Bernado, P. (2012) Structural characterization of intrinsically disordered proteins by the combined use of NMR and SAXS. *Biochem Soc Trans*, **40**, 955-962.

48. Brucale, M., Schuler, B. and Samori, B. (2014) Single-molecule studies of intrinsically disordered proteins. *Chem Rev*, **114**, 3281-3317.

49. Schuler, B. and Hofmann, H. (2013) Single-molecule spectroscopy of protein folding dynamics--expanding scope and timescales. *Curr Opin Struct Biol*, **23**, 36-47.

50. Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A.K., Felli, I.C., Forman-Kay, J.D., Kriwacki, R.W., Pierattelli, R. *et al.* (2014) pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic acids research*, **42**, D326-335.

51. Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N. *et al.* (2007) DisProt: the Database of Disordered Proteins. *Nucleic acids research*, **35**, D786-793.

52. Fukuchi, S., Amemiya, T., Sakamoto, S., Nobe, Y., Hosoda, K., Kado, Y., Murakami, S.D., Koike, R., Hiroaki, H. and Ota, M. (2014) IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic acids research*, **42**, D320-325.

53. Meng, F., Uversky, V.N. and Kurgan, L. (2017) Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cellular and molecular life sciences : CMLS*.

54. Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z. and Dunker, A.K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol*, **323**, 573-584.

55. Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N. and Dunker, A.K. (2006) Intrinsic disorder in transcription factors. *Biochemistry*, **45**, 6873-6888.

56. Galea, C.A., Nourse, A., Wang, Y., Sivakolundu, S.G., Heller, W.T. and Kriwacki, R.W. (2008) Role of intrinsic flexibility in signal transduction mediated by the cell cycle regulator, p27 Kip1. *J Mol Biol*, **376**, 827-838.

57. Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M. and Uversky, V.N. (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J*, **272**, 5129-5148.

58. Haynes, C., Oldfield, C.J., Ji, F., Klitgord, N., Cusick, M.E., Radivojac, P., Uversky, V.N., Vidal, M. and Iakoucheva, L.M. (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol*, **2**, e100.

59. Kim, P.M., Sboner, A., Xia, Y. and Gerstein, M. (2008) The role of disorder in interaction networks: a structural analysis. *Mol Syst Biol*, **4**, 179.

60. Grosschedl, R., Giese, K. and Pagel, J. (1994) HMG domain proteins: architectural elements in the assembly of nucleoprotein structures. *Trends Genet*, **10**, 94-100.

61. Reeves, R. and Nissen, M.S. (1990) The A.T-DNA-binding domain of mammalian high mobility group I chromosomal proteins. A novel peptide motif for recognizing DNA structure. *The Journal of biological chemistry*, **265**, 8573-8582.

62. Evans, J.N., Zajicek, J., Nissen, M.S., Munske, G., Smith, V. and Reeves, R. (1995) 1H and 13C NMR assignments and molecular modelling of a minor groove DNA-binding peptide from the HMG-I protein. *Int J Pept Protein Res*, **45**, 554-560.

63. Huth, J.R., Bewley, C.A., Nissen, M.S., Evans, J.N., Reeves, R., Gronenborn, A.M. and Clore, G.M. (1997) The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif. *Nat Struct Biol*, **4**, 657-665.

64. Reeves, R. (2001) Molecular biology of HMGA proteins: hubs of nuclear function. *Gene*, **277**, 63-81.

65. Dyson, H.J. and Wright, P.E. (2005) Elucidation of the protein folding landscape by NMR. *Methods Enzymol*, **394**, 299-321.

66. Dyson, H.J. and Wright, P.E. (1998) Equilibrium NMR studies of unfolded and partially folded proteins. *Nat Struct Biol*, **5 Suppl**, 499-503.

67. Cai, Z., Gorin, A., Frederick, R., Ye, X., Hu, W., Majumdar, A., Kettani, A. and Patel, D.J. (1998) Solution structure of P22 transcriptional antitermination N peptide-boxB RNA complex. *Nat Struct Biol*, **5**, 203-212.

68. Radhakrishnan, I., Perez-Alvarado, G.C., Parker, D., Dyson, H.J., Montminy, M.R. and Wright, P.E. (1997) Solution structure of the KIX domain of CBP bound to the transactivation domain of CREB: a model for activator:coactivator interactions. *Cell*, **91**, 741-752.

69. Csizmok, V., Follis, A.V., Kriwacki, R.W. and Forman-Kay, J.D. (2016) Dynamic Protein Interaction Networks and New Structural Paradigms in Signaling. *Chem Rev*, **116**, 6424-6462.

70. Wu, H. and Fuxreiter, M. (2016) The Structure and Dynamics of Higher-Order Assemblies: Amyloids, Signalosomes, and Granules. *Cell*, **165**, 1055-1066.

71. Han, T.W., Kato, M., Xie, S., Wu, L.C., Mirzaei, H., Pei, J., Chen, M., Xie, Y., Allen, J., Xiao, G. *et al.* (2012) Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies. *Cell*, **149**, 768-779.

72. Kato, M., Han, T.W., Xie, S., Shi, K., Du, X., Wu, L.C., Mirzaei, H., Goldsmith, E.J., Longgood, J., Pei, J. *et al.* (2012) Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell*, **149**, 753-767.

73. Murakami, T., Qamar, S., Lin, J.Q., Schierle, G.S., Rees, E., Miyashita, A., Costa, A.R., Dodd, R.B., Chan, F.T., Michel, C.H. *et al.* (2015) ALS/FTD Mutation-Induced Phase Transition of FUS Liquid Droplets and Reversible Hydrogels into Irreversible Hydrogels Impairs RNP Granule Function. *Neuron*, **88**, 678-690.

74. Patel, A., Lee, H.O., Jawerth, L., Maharana, S., Jahnel, M., Hein, M.Y., Stoynov, S., Mahamid, J., Saha, S., Franzmann, T.M. *et al.* (2015) A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell*, **162**, 1066-1077.

75. Schwartz, J.C., Wang, X., Podell, E.R. and Cech, T.R. (2013) RNA seeds higher-order assembly of FUS protein. *Cell reports*, **5**, 918-925.

76. Li, J., McQuade, T., Siemer, A.B., Napetschnig, J., Moriwaki, K., Hsiao, Y.S., Damko, E., Moquin, D., Walz, T., McDermott, A. *et al.* (2012) The RIP1/RIP3 necrosome forms a functional amyloid signaling complex required for programmed necrosis. *Cell*, **150**, 339-350.

77. Banjade, S., Wu, Q., Mittal, A., Peeples, W.B., Pappu, R.V. and Rosen, M.K. (2015) Conserved interdomain linker promotes phase separation of the multivalent adaptor protein Nck. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, E6426-6435.

78. Holehouse, A.S. and Pappu, R.V. (2015) Protein polymers: Encoding phase transitions. *Nature materials*, **14**, 1083-1084.

79. Nott, T.J., Petsalaki, E., Farber, P., Jervis, D., Fussner, E., Plochowietz, A., Craggs, T.D., Bazett-Jones, D.P., Pawson, T., Forman-Kay, J.D. *et al.* (2015) Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol Cell*, **57**, 936-947.

80. Bergeron-Sandoval, L.P., Safaee, N. and Michnick, S.W. (2016) Mechanisms and Consequences of Macromolecular Phase Separation. *Cell*, **165**, 1067-1079.

81. Mitrea, D.M. and Kriwacki, R.W. (2016) Phase separation in biology; functional organization of a higher order. *Cell Commun Signal*, **14**, 1.

82. Brangwynne, C.P., Mitchison, T.J. and Hyman, A.A. (2011) Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 4334-4339.

83. Lu, Y., Xu, W., Zeng, M., Yao, G., Shen, L., Yang, M., Luo, Z., Pan, F., Wu, K., Das, T. *et al.* (2015) Topological properties determined by atomic buckling in self-assembled ultrathin bi(110). *Nano letters*, **15**, 80-87.

84. Cioce, M. and Lamond, A.I. (2005) Cajal bodies: a long history of discovery. *Annu Rev Cell Dev Biol*, **21**, 105-131.

85. Fox, A.H. and Lamond, A.I. (2010) Paraspeckles. *Cold Spring Harb Perspect Biol*, **2**, a000687.

86. Lallemand-Breitenbach, V. and de The, H. (2010) PML nuclear bodies. *Cold Spring Harb Perspect Biol*, **2**, a000661.

87. Spector, D.L. and Lamond, A.I. (2011) Nuclear speckles. *Cold Spring Harb Perspect Biol*, **3**.

88. Voronina, E., Seydoux, G., Sassone-Corsi, P. and Nagamori, I. (2011) RNA granules in germ cells. *Cold Spring Harb Perspect Biol*, **3**.

89. Buchan, J.R. (2014) mRNP granules. Assembly, function, and connections with disease. *RNA Biol*, **11**, 1019-1030.

90. Anderson, P., Kedersha, N. and Ivanov, P. (2015) Stress granules, P-bodies and cancer. *Biochimica et biophysica acta*, **1849**, 861-870.

91. Andersen, J.S., Lam, Y.W., Leung, A.K., Ong, S.E., Lyon, C.E., Lamond, A.I. and Mann, M. (2005) Nucleolar proteome dynamics. *Nature*, **433**, 77-83.

92. Shav-Tal, Y., Blechman, J., Darzacq, X., Montagna, C., Dye, B.T., Patton, J.G., Singer, R.H. and Zipori, D. (2005) Dynamic sorting of nuclear components into distinct nucleolar caps during transcriptional inhibition. *Mol Biol Cell*, **16**, 2395-2413.

93. Tsvetanova, N.G., Klass, D.M., Salzman, J. and Brown, P.O. (2010) Proteome-wide search reveals unexpected RNA-binding proteins in Saccharomyces cerevisiae. *PloS one*, **5**.

94. Gerstberger, S., Hafner, M. and Tuschl, T. (2014) A census of human RNA-binding proteins. *Nat Rev Genet*, **15**, 829-845.

95. Chou, C.C. and Wang, A.H. (2015) Structural D/E-rich repeats play multiple roles especially in gene regulation through DNA/RNA mimicry. *Mol Biosyst*, **11**, 2144-2151.

96. Haynes, C. and Iakoucheva, L.M. (2006) Serine/arginine-rich splicing factors belong to a class of intrinsically disordered proteins. *Nucleic acids research*, **34**, 305-312.

97. Varadi, M., Zsolyomi, F., Guharoy, M. and Tompa, P. (2015) Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. *PloS one*, **10**, e0139731.

98. Dyson, H.J. (2012) Roles of intrinsic disorder in protein-nucleic acid interactions. *Mol Biosyst*, **8**, 97-104.

99. Tompa, P. and Csermely, P. (2004) The role of structural disorder in the function of RNA and protein chaperones. *FASEB J*, **18**, 1169-1175.

100. Moras, D. and Poterszman, A. (1996) Getting into the major groove. Protein-RNA interactions. *Curr Biol*, **6**, 530-532.

101. Phan, A.T., Kuryavyi, V., Darnell, J.C., Serganov, A., Majumdar, A., Ilin, S., Raslin, T., Polonskaia, A., Chen, C., Clain, D. *et al.* (2011) Structure-function studies of FMRP RGG peptide recognition of an RNA duplex-quadruplex junction. *Nat Struct Mol Biol*, **18**, 796-804.

102. Puglisi, J.D., Tan, R., Calnan, B.J., Frankel, A.D. and Williamson, J.R. (1992) Conformation of the TAR RNA-arginine complex by NMR spectroscopy. *Science*, **257**, 76-80.

103. Cook, K.B., Hughes, T.R. and Morris, Q.D. (2015) High-throughput characterization of protein-RNA interactions. *Brief Funct Genomics*, **14**, 74-89.

104. Marchese, D., de Groot, N.S., Lorenzo Gotor, N., Livi, C.M. and Tartaglia, G.G. (2016) Advances in the characterization of RNA-binding proteins. *Wiley Interdiscip Rev RNA*, **7**, 793-810.

105. Chen, L., Yun, S.W., Seto, J., Liu, W. and Toth, M. (2003) The fragile X mental retardation protein binds and regulates a novel class of mRNAs containing U rich target sequences. *Neuroscience*, **120**, 1005-1017.

106. Ascano, M., Jr., Mukherjee, N., Bandaru, P., Miller, J.B., Nusbaum, J.D., Corcoran, D.L., Langlois, C., Munschauer, M., Dewell, S., Hafner, M. *et al.* (2012) FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature*, **492**, 382-386.

107. Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A. *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172-177.

108. Tabet, R., Moutin, E., Becker, J.A., Heintz, D., Fouillen, L., Flatter, E., Krezel, W., Alunni, V., Koebel, P., Dembele, D. *et al.* (2016) Fragile X Mental Retardation Protein (FMRP) controls diacylglycerol kinase activity in neurons. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, E3619-3628.

109. Suhl, J.A., Chopra, P., Anderson, B.R., Bassell, G.J. and Warren, S.T. (2014) Analysis of FMRP mRNA target datasets reveals highly associated mRNAs mediated by G-quadruplex structures formed via clustered WGGA sequences. *Human molecular genetics*, **23**, 5479-5491.

110. Boggs, R.T., Gregor, P., Idriss, S., Belote, J.M. and McKeown, M. (1987) Regulation of sexual differentiation in D. melanogaster via alternative splicing of RNA from the transformer gene. *Cell*, **50**, 739-747.

111. Chou, T.B., Zachar, Z. and Bingham, P.M. (1987) Developmental expression of a regulatory gene is programmed at the level of splicing. *The EMBO journal*, **6**, 4095-4104.

112. Amrein, H., Gorman, M. and Nothiger, R. (1988) The sex-determining gene tra-2 of Drosophila encodes a putative RNA binding protein. *Cell*, **55**, 1025-1035.

113. Manley, J.L. and Krainer, A.R. (2010) A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). *Genes & development*, **24**, 1073-1074.

114. Tacke, R. and Manley, J.L. (1999) Determinants of SR protein specificity. *Curr Opin Cell Biol*, **11**, 358-362.

115. Graveley, B.R. (2000) Sorting out the complexity of SR protein functions. *RNA*, **6**, 1197-1211.

116. Wu, J.Y. and Maniatis, T. (1993) Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell*, **75**, 1061-1070.

117. Kohtz, J.D., Jamison, S.F., Will, C.L., Zuo, P., Luhrmann, R., Garcia-Blanco, M.A. and Manley, J.L. (1994) Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. *Nature*, **368**, 119-124.

118. Valcarcel, J., Gaur, R.K., Singh, R. and Green, M.R. (1996) Interaction of U2AF65 RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA [corrected]. *Science*, **273**, 1706-1709.

119. Shen, H., Kan, J.L. and Green, M.R. (2004) Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. *Mol Cell*, **13**, 367-376.

120. Shen, H. and Green, M.R. (2007) RS domain-splicing signal interactions in splicing of U12-type and U2-type introns. *Nature structural & molecular biology*, **14**, 597-603.

121. Jamros, M.A., Aubol, B.E., Keshwani, M.M., Zhang, Z., Stamm, S. and Adams, J.A. (2015) Intra-domain Cross-talk Regulates Serine-arginine Protein Kinase 1-dependent Phosphorylation and Splicing Function of Transformer 2beta1. *The Journal of biological chemistry*, **290**, 17269-17281.

122. Boucher, L., Ouzounis, C.A., Enright, A.J. and Blencowe, B.J. (2001) A genome-wide survey of RS domain proteins. *RNA*, **7**, 1693-1701.

123. Sahara, S., Aoto, M., Eguchi, Y., Imamoto, N., Yoneda, Y. and Tsujimoto, Y. (1999) Acinus is a caspase-3-activated protein required for apoptotic chromatin condensation. *Nature*, **401**, 168-173.

124. Wang, F., Wendling, K.S., Soprano, K.J. and Soprano, D.R. (2014) The SAP motif and C-terminal RS- and RD/E-rich region influences the sub-nuclear localization of Acinus isoforms. *J Cell Biochem*, **115**, 2165-2174.

125. Burgute, B.D., Peche, V.S., Steckelberg, A.L., Glockner, G., Gassen, B., Gehring, N.H. and Noegel, A.A. (2014) NKAP is a novel RS-related protein that interacts with RNA and RNA binding proteins. *Nucleic acids research*, **42**, 3177-3193.

126. Pajerowski, A.G., Nguyen, C., Aghajanian, H., Shapiro, M.J. and Shapiro, V.S. (2009) NKAP is a transcriptional repressor of notch signaling and is required for T cell development. *Immunity*, **30**, 696-707.

127. Tacke, R., Chen, Y. and Manley, J.L. (1997) Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 1148-1153.

128. Soret, J. and Tazi, J. (2003) Phosphorylation-dependent control of the pre-mRNA splicing machinery. *Prog Mol Subcell Biol*, **31**, 89-126.

129. Xiao, S.H. and Manley, J.L. (1997) Phosphorylation of the ASF/SF2 RS domain affects both protein-protein and protein-RNA interactions and is necessary for splicing. *Genes & development*, **11**, 334-344.

130. Huang, Y., Yario, T.A. and Steitz, J.A. (2004) A molecular link between SR protein dephosphorylation and mRNA export. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 9666-9670.

131. Jarvelin, A.I., Noerenberg, M., Davis, I. and Castello, A. (2016) The new (dis)order in RNA regulation. *Cell Commun Signal*, **14**, 9.

132. Castello, A., Fischer, B., Frese, C.K., Horos, R., Alleaume, A.M., Foehr, S., Curk, T., Krijgsveld, J. and Hentze, M.W. (2016) Comprehensive Identification of RNA-Binding Domains in Human Cells. *Mol Cell*, **63**, 696-710.

133. Iwahara, J., Zweckstetter, M. and Clore, G.M. (2006) NMR structural and kinetic characterization of a homeodomain diffusing and hopping on nonspecific DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 15062-15067.

134. Vuzman, D. and Levy, Y. (2010) DNA search efficiency is modulated by charge composition and distribution in the intrinsically disordered tail. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 21004-21009.

135. Rammelt, C., Bilen, B., Zavolan, M. and Keller, W. (2011) PAPD5, a noncanonical poly(A) polymerase with an unusual RNA-binding motif. *RNA*, **17**, 1737-1746.

136. Strein, C., Alleaume, A.M., Rothbauer, U., Hentze, M.W. and Castello, A. (2014) A versatile assay for RNA-binding proteins in living cells. *RNA*, **20**, 721-731.

137. Calabro, V., Daugherty, M.D. and Frankel, A.D. (2005) A single intermolecular contact mediates intramolecular stabilization of both RNA and protein. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 6849-6854.

138. Puglisi, J.D., Chen, L., Blanchard, S. and Frankel, A.D. (1995) Solution structure of a bovine immunodeficiency virus Tat-TAR peptide-RNA complex. *Science*, **270**, 1200-1203.

139. Chen, L. and Frankel, A.D. (1995) A peptide interaction in the major groove of RNA resembles protein interactions in the minor groove of DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 5077-5081.

140. Weeks, K.M., Ampe, C., Schultz, S.C., Steitz, T.A. and Crothers, D.M. (1990) Fragments of the HIV-1 Tat protein specifically bind TAR RNA. *Science*, **249**, 1281-1285.

141. Calnan, B.J., Tidor, B., Biancalana, S., Hudson, D. and Frankel, A.D. (1991) Arginine-mediated RNA recognition: the arginine fork. *Science*, **252**, 1167-1171.

142. Dayie, K.T., Brodsky, A.S. and Williamson, J.R. (2002) Base flexibility in HIV-2 TAR RNA mapped by solution (15)N, (13)C NMR relaxation. *J Mol Biol*, **317**, 263-278.

143. Pitt, S.W., Majumdar, A., Serganov, A., Patel, D.J. and Al-Hashimi, H.M. (2004) Argininamide binding arrests global motions in HIV-1 TAR RNA: comparison with Mg2+-induced conformational stabilization. *J Mol Biol*, **338**, 7-16.

144. Smith, C.A., Crotty, S., Harada, Y. and Frankel, A.D. (1998) Altering the context of an RNA bulge switches the binding specificities of two viral Tat proteins. *Biochemistry*, **37**, 10808-10814.

145. Tan, R. and Frankel, A.D. (1995) Structural variety of arginine-rich RNA-binding peptides. *Proceedings of the National Academy of Sciences of the United States of America*, **92**, 5282-5286.

146. Smith, C.A., Calabro, V. and Frankel, A.D. (2000) An RNA-binding chameleon. *Mol Cell*, **6**, 1067-1076.

147. Bayer, T.S., Booth, L.N., Knudsen, S.M. and Ellington, A.D. (2005) Arginine-rich motifs present multiple interfaces for specific binding by RNA. *RNA*, **11**, 1848-1857.

148. Thandapani, P., O'Connor, T.R., Bailey, T.L. and Richard, S. (2013) Defining the RGG/RG motif. *Mol Cell*, **50**, 613-623.

149. Corley, S.M. and Gready, J.E. (2008) Identification of the RGG box motif in Shadoo: RNA-binding and signaling roles? *Bioinform Biol Insights*, **2**, 383-400.

150. Geuens, T., Bouhy, D. and Timmerman, V. (2016) The hnRNP family: insights into their role in health and disease. *Hum Genet*, **135**, 851-867.

151. Kiledjian, M. and Dreyfuss, G. (1992) Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box. *EMBO J*, **11**, 2655-2664.

152. Hasegawa, Y., Brockdorff, N., Kawano, S., Tsutui, K., Tsutui, K. and Nakagawa, S. (2010) The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. *Dev Cell*, **19**, 469-476.

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

154. Nozawa, R.S., Boteva, L., Soares, D.C., Naughton, C., Dun, A.R., Buckle, A., Ramsahoye, B., Bruton, P.C., Saleeb, R.S., Arnedo, M. *et al.* (2017) SAF-A Regulates Interphase Chromosome Structure through Oligomerization with Chromatin-Associated RNAs. *Cell*, **169**, 1214-1227 e1218.

155. Brown, V., Jin, P., Ceman, S., Darnell, J.C., O'Donnell, W.T., Tenenbaum, S.A., Jin, X., Feng, Y., Wilkinson, K.D., Keene, J.D. *et al.* (2001) Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome. *Cell*, **107**, 477-487.

156. Darnell, J.C., Fraser, C.E., Mostovetsky, O., Stefani, G., Jones, T.A., Eddy, S.R. and Darnell, R.B. (2005) Kissing complex RNAs mediate interaction between the Fragile-X mental retardation protein KH2 domain and brain polyribosomes. *Genes & development*, **19**, 903-918.

157. Darnell, J.C., Jensen, K.B., Jin, P., Brown, V., Warren, S.T. and Darnell, R.B. (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell*, **107**, 489-499.

158. Ramos, A., Hollingworth, D. and Pastore, A. (2003) G-quartet-dependent recognition between the FMRP RGG box and RNA. *RNA*, **9**, 1198-1207.

159. Vasilyev, N., Polonskaia, A., Darnell, J.C., Darnell, R.B., Patel, D.J. and Serganov, A. (2015) Crystal structure reveals specific recognition of a G-quadruplex RNA by a beta-turn in the RGG motif of FMRP. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, E5391-5400.

160. Blackwell, E., Zhang, X. and Ceman, S. (2010) Arginines of the RGG box regulate FMRP association with polyribosomes and mRNA. *Human molecular genetics*, **19**, 1314-1323.

161. Darnell, J.C., Van Driesche, S.J., Zhang, C., Hung, K.Y., Mele, A., Fraser, C.E., Stone, E.F., Chen, C., Fak, J.J., Chi, S.W. *et al.* (2011) FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell*, **146**, 247-261.

162. Bedford, M.T. and Richard, S. (2005) Arginine methylation an emerging regulator of protein function. *Mol Cell*, **18**, 263-272.

163. Bedford, M.T. and Clarke, S.G. (2009) Protein arginine methylation in mammals: who, what, and why. *Mol Cell*, **33**, 1-13.

164. Wickramasinghe, V.O. and Venkitaraman, A.R. (2016) RNA Processing and Genome Stability: Cause and Consequence. *Mol Cell*, **61**, 496-505.

165. Reed, R. (2003) Coupling transcription, splicing and mRNA export. *Curr Opin Cell Biol*, **15**, 326-331.

166. Mackenzie, I.R. and Rademakers, R. (2008) The role of transactive response DNA-binding protein-43 in amyotrophic lateral sclerosis and frontotemporal dementia. *Curr Opin Neurol*, **21**, 693-700.

167. Ritson, G.P., Custer, S.K., Freibaum, B.D., Guinto, J.B., Geffel, D., Moore, J., Tang, W., Winton, M.J., Neumann, M., Trojanowski, J.Q. *et al.* (2010) TDP-43 mediates degeneration in a novel Drosophila model of disease caused by mutations in VCP/p97. *J Neurosci*, **30**, 7729-7739.

168. Voigt, A., Herholz, D., Fiesel, F.C., Kaur, K., Muller, D., Karsten, P., Weber, S.S., Kahle, P.J., Marquardt, T. and Schulz, J.B. (2010) TDP-43-mediated neuron loss in vivo requires RNA-binding activity. *PloS one*, **5**, e12247.

169. Ishiguro, A., Kimura, N., Watanabe, Y., Watanabe, S. and Ishihama, A. (2016) TDP-43 binds and transports G-quadruplex-containing mRNAs into neurites for local translation. *Genes to cells : devoted to molecular & cellular mechanisms*, **21**, 466-481.

170. Schwartz, J.C., Cech, T.R. and Parker, R.R. (2015) Biochemical Properties and Biological Functions of FET Proteins. *Annu Rev Biochem*, **84**, 355-379.

171. Schwartz, J.C., Ebmeier, C.C., Podell, E.R., Heimiller, J., Taatjes, D.J. and Cech, T.R. (2012) FUS binds the CTD of RNA polymerase II and regulates its phosphorylation at Ser2. *Genes & development*, **26**, 2690-2695.

172. Wang, X., Arai, S., Song, X., Reichart, D., Du, K., Pascual, G., Tempst, P., Rosenfeld, M.G., Glass, C.K. and Kurokawa, R. (2008) Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature*, **454**, 126-130.

173. Siomi, M.C., Sato, K., Pezic, D. and Aravin, A.A. (2011) PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol*, **12**, 246-258.

174. Hartig, J.V., Tomari, Y. and Forstemann, K. (2007) piRNAs--the ancient hunters of genome invaders. *Genes & development*, **21**, 1707-1713.

175. Malone, C.D. and Hannon, G.J. (2009) Small RNAs as guardians of the genome. *Cell*, **136**, 656-668.

176. Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281-297.

177. Chiang, H.R., Schoenfeld, L.W., Ruby, J.G., Auyeung, V.C., Spies, N., Baek, D., Johnston, W.K., Russ, C., Luo, S., Babiarz, J.E. *et al.* (2010) Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes & development*, **24**, 992-1009.

178. Wightman, B., Ha, I. and Ruvkun, G. (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, **75**, 855-862.

179. Tam, O.H., Aravin, A.A., Stein, P., Girard, A., Murchison, E.P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R.M. *et al.* (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, **453**, 534-538.

180. Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T. *et al.* (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, **453**, 539-543.

181. Babiarz, J.E., Ruby, J.G., Wang, Y., Bartel, D.P. and Blelloch, R. (2008) Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes & development*, **22**, 2773-2785.

182. Ashwal-Fluss, R., Meyer, M., Pamudurti, N.R., Ivanov, A., Bartok, O., Hanan, M., Evantal, N., Memczak, S., Rajewsky, N. and Kadener, S. (2014) circRNA biogenesis competes with pre-mRNA splicing. *Mol Cell*, **56**, 55-66.

183. Hansen, T.B., Kjems, J. and Damgaard, C.K. (2013) Circular RNA and miR-7 in cancer. *Cancer Res*, **73**, 5609-5612.

184. Capel, B., Swain, A., Nicolis, S., Hacker, A., Walter, M., Koopman, P., Goodfellow, P. and Lovell-Badge, R. (1993) Circular transcripts of the testis-determining gene Sry in adult mouse testis. *Cell*, **73**, 1019-1030.

185. Tsuiji, H., Yoshimoto, R., Hasegawa, Y., Furuno, M., Yoshida, M. and Nakagawa, S. (2011) Competition between a noncoding exon and introns: Gomafu contains tandem UACUAAC repeats and associates with splicing factor-1. *Genes to cells : devoted to molecular & cellular mechanisms*, **16**, 479-490.

186. Wilusz, J.E. (2016) Long noncoding RNAs: Re-writing dogmas of RNA processing and stability. *Biochimica et biophysica acta*, **1859**, 128-138.

187. Zhang, B., Gunawardane, L., Niazi, F., Jahanbani, F., Chen, X. and Valadkhan, S. (2014) A novel RNA motif mediates the strict nuclear localization of a long noncoding RNA. *Mol Cell Biol*, **34**, 2318-2329.

188. Zhang, Y., Yang, L. and Chen, L.L. (2014) Life without A tail: new formats of long noncoding RNAs. *Int J Biochem Cell Biol*, **54**, 338-349.

189. Coccia, E.M., Cicala, C., Charlesworth, A., Ciccarelli, C., Rossi, G.B., Philipson, L. and Sorrentino, V. (1992) Regulation and expression of a growth arrest-specific gene (gas5) during growth, differentiation, and development. *Mol Cell Biol*, **12**, 3514-3521.

190. Hacisuleyman, E., Goff, L.A., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., McClanahan, P., Hendrickson, D.G., Sauvageau, M., Kelley, D.R. *et al.* (2014) Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nature structural & molecular biology*, **21**, 198-206.

191. Lee, J.T. (2012) Epigenetic regulation by long noncoding RNAs. *Science*, **338**, 1435-1439.

192. Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A. *et al.* (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 11667-11672.

193. Camblong, J., Iglesias, N., Fickentscher, C., Dieppois, G. and Stutz, F. (2007) Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in S. cerevisiae. *Cell*, **131**, 706-717.

194. Carrozza, M.J., Li, B., Florens, L., Suganuma, T., Swanson, S.K., Lee, K.K., Shia, W.J., Anderson, S., Yates, J., Washburn, M.P. *et al.* (2005) Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell*, **123**, 581-592.

195. Bumgarner, S.L., Neuert, G., Voight, B.F., Symbor-Nagrabska, A., Grisafi, P., van Oudenaarden, A. and Fink, G.R. (2012) Single-cell analysis reveals that noncoding RNAs contribute to clonal heterogeneity by modulating transcription factor recruitment. *Mol Cell*, **45**, 470-482.

196. Geisler, S. and Coller, J. (2013) RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat Rev Mol Cell Biol*, **14**, 699-712.

197. Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer, I., Collavin, L., Santoro, C. *et al.* (2012) Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature*, **491**, 454-457.

198. Fan, Z., Chen, X. and Chen, R. (2014) Transcriptome-wide analysis of TDP-43 binding small RNAs identifies miR-NID1 (miR-8485), a novel miRNA that represses NRXN1 expression. *Genomics*, **103**, 76-82.

199. Kim, S.H., Dong, W.K., Weiler, I.J. and Greenough, W.T. (2006) Fragile X mental retardation protein shifts between polyribosomes and stress granules after neuronal injury by arsenite stress or in vivo hippocampal electrode insertion. *J Neurosci*, **26**, 2413-2418.

200. Lacoux, C., Di Marino, D., Boyl, P.P., Zalfa, F., Yan, B., Ciotti, M.T., Falconi, M., Urlaub, H., Achsel, T., Mougin, A. *et al.* (2012) BC1-FMRP interaction is modulated by 2'-O-methylation: RNA-binding activity of the tudor domain and translational regulation at synapses. *Nucleic acids research*, **40**, 4086-4096.

201. Parker, R. and Song, H. (2004) The enzymes and control of eukaryotic mRNA turnover. *Nature structural & molecular biology*, **11**, 121-127.

202. Singh, R. and Valcarcel, J. (2005) Building specificity with nonspecific RNA-binding proteins. *Nature structural & molecular biology*, **12**, 645-653.

203. Aitken, C.E. and Lorsch, J.R. (2012) A mechanistic overview of translation initiation in eukaryotes. *Nature structural & molecular biology*, **19**, 568-576.

204. Auweter, S.D., Oberstrass, F.C. and Allain, F.H. (2006) Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic acids research*, **34**, 4943-4959.

205. Elkayam, E., Kuhn, C.D., Tocilj, A., Haase, A.D., Greene, E.M., Hannon, G.J. and Joshua-Tor, L. (2012) The structure of human argonaute-2 in complex with miR-20a. *Cell*, **150**, 100-110.

206. Jankowsky, E. and Harris, M.E. (2015) Specificity and nonspecificity in RNA-protein interactions. *Nat Rev Mol Cell Biol*, **16**, 533-544.

207. Mitchell, S.F. and Parker, R. (2014) Principles and properties of eukaryotic mRNPs. *Mol Cell*, **54**, 547-558.

208. Lunde, B.M., Moore, C. and Varani, G. (2007) RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol*, **8**, 479-490.

209. Maris, C., Dominguez, C. and Allain, F.H. (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J*, **272**, 2118-2131.

210. Clery, A., Jayne, S., Benderska, N., Dominguez, C., Stamm, S. and Allain, F.H. (2011) Molecular basis of purine-rich RNA recognition by the human SR-like protein Tra2-beta1. *Nature structural & molecular biology*, **18**, 443-450.

211. Dominguez, C., Fisette, J.F., Chabot, B. and Allain, F.H. (2010) Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs. *Nature structural & molecular biology*, **17**, 853-861.

212. Klein Gunnewiek, J.M., Hussein, R.I., van Aarssen, Y., Palacios, D., de Jong, R., van Venrooij, W.J. and Gunderson, S.I. (2000) Fourteen residues of the U1 snRNP-specific U1A protein are required for homodimerization, cooperative RNA binding, and inhibition of polyadenylation. *Mol Cell Biol*, **20**, 2209-2217.

213. Selenko, P., Gregorovic, G., Sprangers, R., Stier, G., Rhani, Z., Kramer, A. and Sattler, M. (2003) Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP. *Mol Cell*, **11**, 965-976.

214. Sanford, J.R., Wang, X., Mort, M., Vanduyn, N., Cooper, D.N., Mooney, S.D., Edenberg, H.J. and Liu, Y. (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res*, **19**, 381-394.

215. Milek, M., Wyler, E. and Landthaler, M. (2012) Transcriptome-wide analysis of protein-RNA interactions using high-throughput sequencing. *Semin Cell Dev Biol*, **23**, 206-212.

216. Ray, D., Kazan, H., Chan, E.T., Pena Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q. and Hughes, T.R. (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol*, **27**, 667-670.

217. Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A. and Burge, C.B. (2014) RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell*, **54**, 887-900.

218. Edwards, A.L., Garst, A.D. and Batey, R.T. (2009) Determining structures of RNA aptamers and riboswitches by X-ray crystallography. *Methods Mol Biol*, **535**, 135-163.

219. Gilbert, S.D. and Batey, R.T. (2009) Monitoring RNA-ligand interactions using isothermal titration calorimetry. *Methods Mol Biol*, **540**, 97-114.

220. Harper, S. and Speicher, D.W. (2011) Purification of proteins fused to glutathione S-transferase. *Methods Mol Biol*, **681**, 259-280.

221. Bohm, G., Muhr, R. and Jaenicke, R. (1992) Quantitative analysis of protein far UV circular dichroism spectra by neural networks. *Protein Eng*, **5**, 191-195.

222. Schatz, D., Leberman, R. and Eckstein, F. (1991) Interaction of Escherichia coli tRNA(Ser) with its cognate aminoacyl-tRNA synthetase as determined by footprinting with phosphorothioate-containing tRNA transcripts. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 6132-6136.

223. Stefl, R., Skrisovska, L. and Allain, F.H. (2005) RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. *EMBO reports*, **6**, 33-38.

224. Nagai, K., Oubridge, C., Ito, N., Jessen, T.H., Avis, J. and Evans, P. (1995) Crystal structure of the U1A spliceosomal protein complexed with its cognate RNA hairpin. *Nucleic Acids Symp Ser*, 1-2.

225. Oubridge, C., Ito, N., Evans, P.R., Teo, C.H. and Nagai, K. (1994) Crystal structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature*, **372**, 432-438.

226. Muniz, L., Egloff, S. and Kiss, T. (2013) RNA elements directing in vivo assembly of the 7SK/MePCE/Larp7 transcriptional regulatory snRNP. *Nucleic Acids Res*, **41**, 4686-4698.

227. Alfano, C., Sanfelice, D., Babon, J., Kelly, G., Jacks, A., Curry, S. and Conte, M.R. (2004) Structural analysis of cooperative RNA binding by the La motif and central RRM domain of human La protein. *Nat Struct Mol Biol*, **11**, 323-329.

228. Doudna, J.A. and Batey, R.T. (2004) Structural insights into the signal recognition particle. *Annu Rev Biochem*, **73**, 539-557.

229. Batey, R.T., Rambo, R.P., Lucast, L., Rha, B. and Doudna, J.A. (2000) Crystal structure of the ribonucleoprotein core of the signal recognition particle. *Science*, **287**, 1232-1239.

230. Tan, A.Y. and Manley, J.L. (2012) TLS/FUS: a protein in cancer and ALS. *Cell Cycle*, **11**, 3349-3350.

231. Boeynaems, S., Bogaert, E., Kovacs, D., Konijnenberg, A., Timmerman, E., Volkov, A., Guharoy, M., De Decker, M., Jaspers, T., Ryan, V.H. *et al.* (2017) Phase Separation of C9orf72 Dipeptide Repeats Perturbs Stress Granule Dynamics. *Mol Cell*, **65**, 1044-1055 e1045.

232. Kapeli, K., Pratt, G.A., Vu, A.Q., Hutt, K.R., Martinez, F.J., Sundararaman, B., Batra, R., Freese, P., Lambert, N.J., Huelga, S.C. *et al.* (2016) Distinct and shared functions of ALS-associated proteins TDP-43, FUS and TAF15 revealed by multisystem analyses. *Nat Commun*, **7**, 12143.

233. Lagier-Tourenne, C., Polymenidou, M., Hutt, K.R., Vu, A.Q., Baughn, M., Huelga, S.C., Clutario, K.M., Ling, S.C., Liang, T.Y., Mazur, C. *et al.* (2012) Divergent roles of ALS-

linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nat Neurosci*, **15**, 1488-1497.

234. Rogelj, B., Easton, L.E., Bogu, G.K., Stanton, L.W., Rot, G., Curk, T., Zupan, B., Sugimoto, Y., Modic, M., Haberman, N. *et al.* (2012) Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. *Sci Rep*, **2**, 603.

235. Ishigaki, S., Masuda, A., Fujioka, Y., Iguchi, Y., Katsuno, M., Shibata, A., Urano, F., Sobue, G. and Ohno, K. (2012) Position-dependent FUS-RNA interactions regulate alternative splicing events and transcriptions. *Sci Rep*, **2**, 529.

236. Colombrita, C., Onesto, E., Megiorni, F., Pizzuti, A., Baralle, F.E., Buratti, E., Silani, V. and Ratti, A. (2012) TDP-43 and FUS RNA-binding proteins bind distinct sets of cytoplasmic messenger RNAs and differently regulate their post-transcriptional fate in motoneuron-like cells. *J Biol Chem*.

237. Hoell, J.I., Larsson, E., Runge, S., Nusbaum, J.D., Duggimpudi, S., Farazi, T.A., Hafner, M., Borkhardt, A., Sander, C. and Tuschl, T. (2011) RNA targets of wild-type and mutant FET family proteins. *Nat Struct Mol Biol*.

238. Wang, X., Schwartz, J.C. and Cech, T.R. (2015) Nucleic acid-binding specificity of human FUS protein. *Nucleic acids research*, **43**, 7535-7543.

239. Kwon, I., Kato, M., Xiang, S., Wu, L., Theodoropoulos, P., Mirzaei, H., Han, T., Xie, S., Corden, J.L. and McKnight, S.L. (2013) Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. *Cell*, **155**, 1049-1060.

240. Huelga, S.C., Vu, A.Q., Arnold, J.D., Liang, T.Y., Liu, P.P., Yan, B.Y., Donohue, J.P., Shiue, L., Hoon, S., Brenner, S. *et al.* (2012) Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep*, **1**, 167-178.

241. Xiang, S., Kato, M., Wu, L.C., Lin, Y., Ding, M., Zhang, Y., Yu, Y. and McKnight, S.L. (2015) The LC Domain of hnRNPA2 Adopts Similar Conformations in Hydrogel Polymers, Liquid-like Droplets, and Nuclei. *Cell*, **163**, 829-839.

242. Molliex, A., Temirov, J., Lee, J., Coughlin, M., Kanagaraj, A.P., Kim, H.J., Mittag, T. and Taylor, J.P. (2015) Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell*, **163**, 123-133.

243. Lin, Y., Protter, D.S., Rosen, M.K. and Parker, R. (2015) Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. *Mol Cell*.

244. Burke, K.A., Janke, A.M., Rhine, C.L. and Fawzi, N.L. (2015) Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II. *Mol Cell*, **60**, 231-241.

245. Kim, H.J., Kim, N.C., Wang, Y.D., Scarborough, E.A., Moore, J., Diaz, Z., MacLea, K.S., Freibaum, B., Li, S., Molliex, A. *et al.* (2013) Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature*, **495**, 467-473.

246. Banani, S.F., Rice, A.M., Peeples, W.B., Lin, Y., Jain, S., Parker, R. and Rosen, M.K. (2016) Compositional Control of Phase-Separated Cellular Bodies. *Cell*.

247. Liu, X., Niu, C., Ren, J., Zhang, J., Xie, X., Zhu, H., Feng, W. and Gong, W. (2013) The RRM domain of human fused in sarcoma protein reveals a non-canonical nucleic acid binding site. *Biochimica et biophysica acta*, **1832**, 375-385.

248. Iko, Y., Kodama, T.S., Kasai, N., Oyama, T., Morita, E.H., Muto, T., Okumura, M., Fujii, R., Takumi, T., Tate, S. *et al.* (2004) Domain architectures and characterization of an RNA-binding protein, TLS. *The Journal of biological chemistry*, **279**, 44834-44840.

249. Plambeck, C.A., Kwan, A.H., Adams, D.J., Westman, B.J., van der Weyden, L., Medcalf, R.L., Morris, B.J. and Mackay, J.P. (2003) The structure of the zinc finger domain from human splicing factor ZNF265 fold. *The Journal of biological chemistry*, **278**, 22805-22811.

250. Lerga, A., Hallier, M., Delva, L., Orvain, C., Gallais, I., Marie, J. and Moreau-Gachelin, F. (2001) Identification of an RNA binding specificity for the potential splicing factor TLS. *J Biol Chem*, **276**, 6807-6816.

251. Polaski, J.T., Holmstrom, E.D., Nesbitt, D.J. and Batey, R.T. (2016) Mechanistic Insights into Cofactor-Dependent Coupling of RNA Folding and mRNA Transcription/Translation by a Cobalamin Riboswitch. *Cell reports*, **15**, 1100-1110.

252. Cameron, I.L., Smith, N.K., Pool, T.B. and Sparks, R.L. (1980) Intracellular concentration of sodium and other elements as related to mitogenesis and oncogenesis in vivo. *Cancer Res*, **40**, 1493-1500.

253. Volkov, V. (2015) Quantitative description of ion transport via plasma membrane of yeast and small cells. *Front Plant Sci*, **6**, 425.

254. Arino, J., Ramos, J. and Sychrova, H. (2010) Alkali metal cation transport and homeostasis in yeasts. *Microbiol Mol Biol Rev*, **74**, 95-120.

255. Dick, D.A. (1978) The distribution of sodium, potassium and chloride in the nucleus and cytoplasm of Bufo bufo oocytes measured by electron microprobe analysis. *J Physiol*, **284**, 37-53.

256. Shamoo, Y., Abdul-Manan, N. and Williams, K.R. (1995) Multiple RNA binding domains (RBDs) just don't add up. *Nucleic acids research*, **23**, 725-728.

257. Lagier-Tourenne, C., Baughn, M., Rigo, F., Sun, S., Liu, P., Li, H.R., Jiang, J., Watt, A.T., Chun, S., Katz, M. *et al.* (2013) Targeted degradation of sense and antisense C9orf72 RNA foci as therapy for ALS and frontotemporal degeneration. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, E4530-4539.

258. Bhardwaj, V., Kumar, V., Geysen, H.M. and Sercarz, E.E. (1993) Degenerate recognition of a dissimilar antigenic peptide by myelin basic protein-reactive T cells. Implications for thymic education and autoimmunity. *J Immunol*, **151**, 5000-5010.

259. Roy, J. and Cyert, M.S. (2009) Cracking the phosphatase code: docking interactions determine substrate specificity. *Sci Signal*, **2**, re9.

260. Toppo, S., Flohe, L., Ursini, F., Vanin, S. and Maiorino, M. (2009) Catalytic mechanisms and specificities of glutathione peroxidases: variations of a basic scheme. *Biochim Biophys Acta*, **1790**, 1486-1500.

261. Lee, C.H., Saksela, K., Mirza, U.A., Chait, B.T. and Kuriyan, J. (1996) Crystal structure of the conserved core of HIV-1 Nef complexed with a Src family SH3 domain. *Cell*, **85**, 931-942.

262. Lee, C.H., Leung, B., Lemmon, M.A., Zheng, J., Cowburn, D., Kuriyan, J. and Saksela, K. (1995) A single amino acid in the SH3 domain of Hck determines its high affinity and specificity in binding to HIV-1 Nef protein. *EMBO J*, **14**, 5006-5015.

263. Jain, S., Wheeler, J.R., Walters, R.W., Agrawal, A., Barsic, A. and Parker, R. (2016) ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. *Cell*, **164**, 487-498.

264. Berry, J., Weber, S.C., Vaidya, N., Haataja, M. and Brangwynne, C.P. (2015) RNA transcription modulates phase transition-driven nuclear body assembly. *Proc Natl Acad Sci U S A*, **112**, E5237-5245.

265. Fornerod, M. (2012) RS and RGG repeats as primitive proteins at the transition between the RNA and RNP worlds. *Nucleus*, **3**, 4-5.

266. Smola, M.J., Christy, T.W., Inoue, K., Nicholson, C.O., Friedersdorf, M., Keene, J.D., Lee, D.M., Calabrese, J.M. and Weeks, K.M. (2016) SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. *Proc Natl Acad Sci U S A*, **113**, 10322-10327.

267. Smola, M.J., Calabrese, J.M. and Weeks, K.M. (2015) Detection of RNA-Protein Interactions in Living Cells with SHAPE. *Biochemistry*, **54**, 6867-6875.

268. Beckmann, B.M., Castello, A. and Medenbach, J. (2016) The expanding universe of ribonucleoproteins: of novel RNA-binding proteins and unconventional interactions. *Pflugers Arch*, **468**, 1029-1040.

269. Rajyaguru, P. and Parker, R. (2012) RGG motif proteins: modulators of mRNA functional states. *Cell Cycle*, **11**, 2594-2599.

270. Davidovich, C., Wang, X., Cifuentes-Rojas, C., Goodrich, K.J., Gooding, A.R., Lee, J.T. and Cech, T.R. (2015) Toward a consensus on the binding specificity and promiscuity of PRC2 for RNA. *Mol Cell*, **57**, 552-558.

271. Davidovich, C. and Cech, T.R. (2015) The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2. *RNA*, **21**, 2007-2022.

272. Davidovich, C., Zheng, L., Goodrich, K.J. and Cech, T.R. (2013) Promiscuous RNA binding by Polycomb repressive complex 2. *Nat Struct Mol Biol*, **20**, 1250-1257.

273. Fellows, A., Deng, B., Mierke, D.F., Robey, R.B. and Nichols, R.C. (2013) Peptides modeled on the RGG domain of AUF1/hnRNP-D regulate 3' UTR-dependent gene expression. *Int Immunopharmacol*, **17**, 132-141.

274. Rajyaguru, P., She, M. and Parker, R. (2012) Scd6 targets eIF4G to repress translation: RGG motif proteins as a class of eIF4G-binding proteins. *Mol Cell*, **45**, 244-254.

275. Burd, C.G. and Dreyfuss, G. (1994) Conserved structures and diversity of functions of RNA-binding proteins. *Science*, **265**, 615-621.

276. Swanson, M.S. and Dreyfuss, G. (1988) Classification and purification of proteins of heterogeneous nuclear ribonucleoprotein particles by RNA-binding specificities. *Mol Cell Biol*, **8**, 2237-2241.

277. Tan, A.Y. and Manley, J.L. (2009) The TET family of proteins: functions and roles in disease. *J Mol Cell Biol*, **1**, 82-92.

278. Tan, A.Y., Riley, T.R., Coady, T., Bussemaker, H.J. and Manley, J.L. (2012) TLS/FUS (translocated in liposarcoma/fused in sarcoma) regulates target gene transcription via single-stranded DNA response elements. *Proc Natl Acad Sci U S A*.

279. Campbell, N.H. and Neidle, S. (2012) G-quadruplexes and metal ions. *Met Ions Life Sci*, **10**, 119-134.

280. Bhattacharyya, D., Mirihana Arachchilage, G. and Basu, S. (2016) Metal Cations in G-Quadruplex Folding and Stability. *Front Chem*, **4**, 38.

281. Takahama, K., Miyawaki, A., Shitara, T., Mitsuya, K., Morikawa, M., Hagihara, M., Kino, K., Yamamoto, A. and Oyoshi, T. (2015) G-Quadruplex DNA- and RNA-Specific-Binding Proteins Engineered from the RGG Domain of TLS/FUS. *ACS Chem Biol*.

282. Chen, E., Sharma, M.R., Shi, X., Agrawal, R.K. and Joseph, S. (2014) Fragile X mental retardation protein regulates translation by binding directly to the ribosome. *Mol Cell*, **54**, 407-417.

283. Choy, M.S., Page, R. and Peti, W. (2012) Regulation of protein phosphatase 1 by intrinsically disordered proteins. *Biochem Soc Trans*, **40**, 969-974.

284. Dyson, H.J. (2016) Making Sense of Intrinsically Disordered Proteins. *Biophys J*, **110**, 1013-1016.

285. Espinoza-Fonseca, L.M. (2009) Reconciling binding mechanisms of intrinsically disordered proteins. *Biochem Biophys Res Commun*, **382**, 479-482.

286. Kipp, M., Gohring, F., Ostendorp, T., van Drunen, C.M., van Driel, R., Przybylski, M. and Fackelmayer, F.O. (2000) SAF-Box, a conserved protein domain that specifically recognizes scaffold attachment region DNA. *Mol Cell Biol*, **20**, 7480-7489.

287. Ashley, C.T., Jr., Wilkinson, K.D., Reines, D. and Warren, S.T. (1993) FMR1 protein: conserved RNP family domains and selective RNA binding. *Science*, **262**, 563-566.

288. Leulliot, N. and Varani, G. (2001) Current topics in RNA-protein recognition: control of specificity and biological function through induced fit and conformational capture. *Biochemistry*, **40**, 7947-7956.

289. Masuda, A., Takeda, J., Okuno, T., Okamoto, T., Ohkawara, B., Ito, M., Ishigaki, S., Sobue, G. and Ohno, K. (2015) Position-specific binding of FUS to nascent RNA regulates mRNA length. *Genes Dev*, **29**, 1045-1057.

290. Corbin-Lickfett, K.A., Souki, S.K., Cocco, M.J. and Sandri-Goldin, R.M. (2010) Three arginine residues within the RGG box are crucial for ICP27 binding to herpes simplex virus 1 GC-rich sequences and for efficient viral RNA export. *J Virol*, **84**, 6367-6376.

291. Lu, C.C., Wu, C.W., Chang, S.C., Chen, T.Y., Hu, C.R., Yeh, M.Y., Chen, J.Y. and Chen, M.R. (2004) Epstein-Barr virus nuclear antigen 1 is a DNA-binding protein with strong RNA-binding activity. *J Gen Virol*, **85**, 2755-2765.

292. Guo, J.U. and Bartel, D.P. (2016) RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science*, **353**.

293. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. and Weissman, J.S. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505**, 701-705.

294. Ramaswami, M., Taylor, J.P. and Parker, R. (2013) Altered ribostasis: RNA-protein granules in degenerative disorders. *Cell*, **154**, 727-736.

295. Dreyfuss, G., Kim, V.N. and Kataoka, N. (2002) Messenger-RNA-binding proteins and the messages they carry. *Nat Rev Mol Cell Biol*, **3**, 195-205.

296. Pinol-Roma, S. and Dreyfuss, G. (1992) Shuttling of pre-mRNA binding proteins between nucleus and cytoplasm. *Nature*, **355**, 730-732.

297. Jijakli, K., Khraiwesh, B., Fu, W., Luo, L., Alzahmi, A., Koussa, J., Chaiboonchoe, A., Kirmizialtin, S., Yen, L. and Salehi-Ashtiani, K. (2016) The in vitro selection world. *Methods*, **106**, 3-13.

298. Ashley, C.T., Sutcliffe, J.S., Kunst, C.B., Leiner, H.A., Eichler, E.E., Nelson, D.L. and Warren, S.T. (1993) Human and murine FMR-1: alternative splicing and translational initiation downstream of the CGG-repeat. *Nat Genet*, **4**, 244-251.

299. Dolzhanskaya, N., Bolton, D.C. and Denman, R.B. (2008) Chemical and structural probing of the N-terminal residues encoded by FMR1 exon 15 and their effect on downstream arginine methylation. *Biochemistry*, **47**, 8491-8503.

300. Denman, R.B. and Sung, Y.J. (2002) Species-specific and isoform-specific RNA binding of human and mouse fragile X mental retardation proteins. *Biochemical and biophysical research communications*, **292**, 1063-1069.

301. Blackwell, E. and Ceman, S. (2011) A new regulatory function of the region proximal to the RGG box in the fragile X mental retardation protein. *Journal of cell science*, **124**, 3060-3065.

302. Rajpurohit, R., Lee, S.O., Park, J.O., Paik, W.K. and Kim, S. (1994) Enzymatic methylation of recombinant heterogeneous nuclear RNP protein A1. Dual substrate specificity for S-adenosylmethionine:histone-arginine N-methyltransferase. *The Journal of biological chemistry*, **269**, 1075-1082.

303. Dolzhanskaya, N., Merz, G., Aletta, J.M. and Denman, R.B. (2006) Methylation regulates the intracellular protein-protein and protein-RNA interactions of FMRP. *Journal of cell science*, **119**, 1933-1946.

304. Da Cruz, S. and Cleveland, D.W. (2011) Understanding the role of TDP-43 and FUS/TLS in ALS and beyond. *Curr Opin Neurobiol*, **21**, 904-919.

305. Tsuiji, H., Iguchi, Y., Furuya, A., Kataoka, A., Hatsuta, H., Atsuta, N., Tanaka, F., Hashizume, Y., Akatsu, H., Murayama, S. *et al.* (2013) Spliceosome integrity is defective in the motor neuron diseases ALS and SMA. *EMBO Mol Med*, **5**, 221-234.

306. Sinha, S., Lopes, D.H., Du, Z., Pang, E.S., Shanmugam, A., Lomakin, A., Talbiersky, P., Tennstaedt, A., McDaniel, K., Bakshi, R. *et al.* (2011) Lysine-specific molecular tweezers are broad-spectrum inhibitors of assembly and toxicity of amyloid proteins. *Journal of the American Chemical Society*, **133**, 16958-16969.

307. Cuchillo, R. and Michel, J. (2012) Mechanisms of small-molecule binding to intrinsically disordered proteins. *Biochem Soc Trans*, **40**, 1004-1008.

308. Uversky, V.N. (2012) Intrinsically disordered proteins and novel strategies for drug discovery. *Expert Opin Drug Discov*, **7**, 475-488.

309. Toth, G., Gardai, S.J., Zago, W., Bertoncini, C.W., Cremades, N., Roy, S.L., Tambe, M.A., Rochet, J.C., Galvagnion, C., Skibinski, G. *et al.* (2014) Targeting the intrinsically disordered structural ensemble of alpha-synuclein by small molecules as a potential therapeutic strategy for Parkinson's disease. *PloS one*, **9**, e87133.

310. Cheng, Y., LeGall, T., Oldfield, C.J., Mueller, J.P., Van, Y.Y., Romero, P., Cortese, M.S., Uversky, V.N. and Dunker, A.K. (2006) Rational drug design via intrinsically disordered protein. *Trends Biotechnol*, **24**, 435-442.

311. Metallo, S.J. (2010) Intrinsically disordered proteins are potential drug targets. *Curr Opin Chem Biol*, **14**, 481-488.

312. Wang, J., Cao, Z., Zhao, L. and Li, S. (2011) Novel strategies for drug discovery based on Intrinsically Disordered Proteins (IDPs). *Int J Mol Sci*, **12**, 3205-3219.

313. Vassilev, L.T., Vu, B.T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., Kong, N., Kammlott, U., Lukacs, C., Klein, C. *et al.* (2004) In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science*, **303**, 844-848.

314. Nair, S.K. and Burley, S.K. (2003) X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors. *Cell*, **112**, 193-205.

315. Prabhudesai, S., Sinha, S., Attar, A., Kotagiri, A., Fitzmaurice, A.G., Lakshmanan, R., Ivanova, M.I., Loo, J.A., Klarner, F.G., Schrader, T. *et al.* (2012) A novel "molecular tweezer" inhibitor of alpha-synuclein neurotoxicity in vitro and in vivo. *Neurotherapeutics*, **9**, 464-476.

316. Seol, Y., Skinner, G.M., Visscher, K., Buhot, A. and Halperin, A. (2007) Stretching of homopolymeric RNA reveals single-stranded helices and base-stacking. *Physical review letters*, **98**, 158103.

Appendix

Table A 1 Sequences of RNA substrates used in this thesis.

Name	Length (Nucleotides)	Sequence	Structure
DNMT	48	AUUGAGGAGCAGCAGAGAAGUUGGA GUGAAGGCAGAGAGGGGUUAAGG	N.A
Sc1	36	GCUGCGGUGUGGAAGGAGUGGCUG GGUUGCGCAGG	Stem loop(159)
dsGC	36	AUAUACGCGCGUAUAUUUCGAUAUAC GCGCGUAUAU	Hairpin
dsAU	36	AUAUAUAUAUAUAUAUUUCGAUAUAU AUAUAUAUAU	Hairpin
hRRD	152	CAUGGAUCCCUGAGGUCGGUCCCCAAUA CGACAAGACAAUUUGAUAUCAUAAUAGAA CACUGCAGAAACAAUGCUGAGUGAAGAA GAGUAGAAAUGGGAAGACUUGGUUGAGC GGAAACUGAGUUCUUGAAAAGAGGAGAU GCUUGAUGAGG	N.A[a]
mRRD	155	AACUGGCCCCUGGGAUUUGCUGCUC AGAACCUGAGUUCACUGAGACAUCAG GAGCAAGCACUGGAGGCCGGGUGCU GCUGGACCCAGAUGGGAGCCAUGCA GGACUUGACCAUGGCCUGCACACAC UUCUUCCCAGGAGAAGGGGAAUGAG GAAG	N.A
GGUG	25	UUGUAUUUUGAGCUAGUUUGGUGAU	N.A
CRL	36	AUACAACAUACAACAUACAACAUACAA CAUACAACA	Single stranded(251)
Poly-A	40	AAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAA	Single stranded(316)

N.A.[a]: Not available.

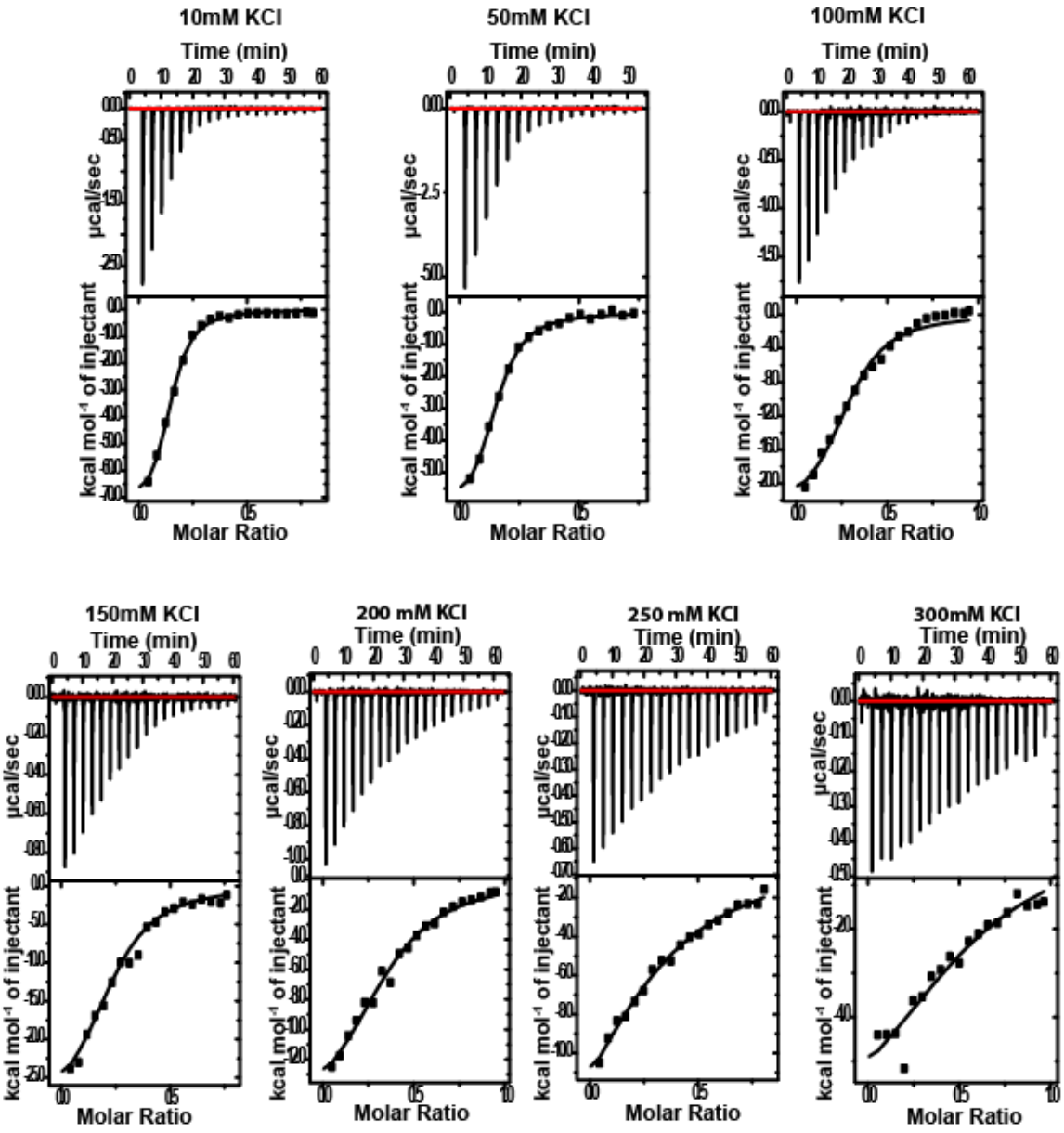

Figure A 1 ITC raw data of RGG1-RRM-RGG2 interaction with DNMT RNA at different KCl concentrations.

Table A 2 Corresponding biophysical parameters of the ITC graphs in Figure A 1.

[KCl], mM	N (RNA:protein)	ΔH (kcal/mol)	ΔS (cal/mol^{-1}/K^{-1}	K_D, (μM)
10	0.13 ± 0.03	-75.5 ± 2	-227	1.5 ± 0.3
50	0.14 ± 0.02	-64.5 ± 2	-192	1 ± 0.1
100	0.2 ± 0.01	-37 ± 3	-97	1.7 ± 0.5
150	0.23 ± 0.01	-30 ± 2	-77.3	3 ± 0.5
200	0.26 ± 0.01	-20.5 ± 6	-47	14 ± 4
250	0.3 ± 0.03	-21 ± 4	-52	26 ± 3
300	0.56 ± 0.07	-7 ± 2	-3	42 ± 10

Table A 3 Mutated arginine residues of SGG mutants.

SGG1	R213S, R216S, R218S, R234S, R242S, R244S, R248S, R251S, R259S
SGG2	R377S, R383S, R386S, R388S, R394S, R407S, R422S
SGG3	R472S, R473S, R476S, R481S, R485S, R487S, R491S, R495S, R498S, R503S
SGG4	R213S, R216S, R218S, R234S, R242S, R244S, R248S, R251S, R259S, R377S, R383S, R386S, R388S, R394S, R407S, R422S, R472S, R473S, R476S, R481S, R485S, R487S, R491S, R495S, R498S, R503S

Table A 4 Corresponding $K_{D,app}$ (μM) values of heat-map data in Figure 4.2.

	RNA								
	Sc1	DMNT	hRRD	mRRD	dsGC	dsAU	GGUG	CRL	poly-A
Protein									
FMRP-RGG	0.09 ± 0.02	2.5 ± 0.4	0.30 ± 0.01	0.35 ± 0.03	8.0 ± 0.1	8.3 ±0.5	27 ± 3	27 ± 1	50±1
hnRNPU-RGG	0.25 ± 0.01	0.5 ± 0.1	0.30 ± 0.02	0.5 ± 0.1	15 ± 1	43 ± 20	8.5 ± 1.0	8.0 ± 0.5	9.0±0.5
FUS-RGG1	2.8 ± 0.1	3.0 ± 0.1	1.6 ± 0.1	1.3 ± 0.1	22 ± 1	29 ± 1	16 ± 1	17 ± 1	25±5
FUS-RGG2	25 ± 1	60 ± 15	14 ± 1	26 ± 4	n.d.[a]	n.d.	n.d.	n.d.	n.d.
FUS-RGG3	3.7 ± 0.2	8.0 ± 0.5	4.5 ± 1.0	4.5 ± 0.5	100 ± 30	n.d.	50±3	65±6	110±30
FUS-LC-RGG1	2.8 ± 0.5	6.2 ± 0.3	3.0 ± 0.3	1.0 ± 0.3	45 ± 2	72 ± 5	12 ± 1	19 ± 1	5 ± 2
FUS-RRM-RGG2	2.8 ± 0.2	2.5 ± 0.1	1.2 ± 0.2	1.8 ± 0.1	15 ± 1	17 ± 2	11 ± 1	9.5 ± 0.5	17 ± 3
FUS (wt)	0.30 ± 0.02	0.7 ± 0.2	0.60 ± 0.05	0.33 ± 0.07	7.3 ± 0.2	8.4 ± 0.5	3.0 ± 0.1	10 ± 1	3.2 ± 0.2

[a]n.d. : not detectable (>100 μM).