Mechanistic and Functional Studies of RNA Processing Pathway Components

Mechanistic and Functional Studies of RNA Processing Pathway Components

A dissertation presented

by

Elizabeth Marie Ransey

to

The Committee on Higher Degrees in Chemical Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Chemical Biology

Harvard University

Cambridge, Massachusetts

April 2017

Dissertation Advisor: Piotrek Sliz, Ph.D. **Elizabeth Marie Ransey**

Mechanistic and Functional Studies of RNA Processing Pathway Components

Abstract

The delicate interplay between proteins and the RNAs that they bind or affect forms the bases for numerous, coordinated and interconnected RNA metabolic pathways that maintain all cellular life. The pervasiveness and diversity of protein-RNA interactions, puts RNA binding or processing proteins in integral regulatory positions, the disregulation of which, underlie major human diseases. Here, I have examined three components of major RNA processing pathways, to gather insight on their diverse mechanisms of action.

The Intron degradation pathway is responsible for maintaining cellular free nucleotide levels, which function as sources of energy, signaling molecules and precursors for DNA and RNA synthesis. In this dissertation, I present the highest resolution X-ray crystal structure and cofactor specificity analysis of the Dbr1 enzyme, which cleaves 2'-5' phosphodiester bonds within all excised intronic lariats, thus, initiating intron turnover.

MicroRNAs are small noncoding RNAs that regulate gene expression; they are themselves processed via a multi-step and multi-component biogenesis

pathway. One of the major functions of the oncoprotein LIN28A is to inhibit the biogenesis of let-7 miRNAs by binding and preventing processing by Dicer, and subsequently initiating degradation. Here, I use a comparative approach to understand newly defined interactions between LIN28 and RNAs throughout the transcriptome.

The DEAD-Box Helicase P72, on the other hand, participates as an accessory factor to the Microprocessor complex, and enhances processing of certain classes of miRNA. In this work, I characterize a newly observed behavior of P72 that may define new cellular functions associated with RNA binding.

Table of Contents/ Index

List of Figures Page

List of Tables Page

List of Schemes Page

List of Datasets (available online pending publication) Page

List of Appendix Figures

Acknowledgments

Several years ago, I was asked to deliver an acceptance speech on behalf of my cohort of graduate student recipients of a research fellowship, sponsored by the UNCF-Merck joint initiative. The purpose being to thank the leaders of the program, but also to explain my motivation, my journey, myself and what the award would do for me.

When I delivered the speech at the acceptance gala, I recited the expected appreciation and made declarations of the importance of our network: how the fellowship allowed me and my colleagues to support each other in our minority community, potentially clichéd, (albeit completely sincere) affirmations that we were going to blazing trails and that we were tasked with inspiring generations and making scientific contributions that would literally, tangibly alter the world around us.

Before all of that, however, I described how in high school I thought that ‘science’ was just a class; I thought that everything had been ‘figured out’ already. I took all of the elective science courses offered at my school because I thought they were interesting, but I didn’t realize that I could actually use what I was learning to try to understand the world, the environment, health and disease. Once I became aware that careers as basic or biomedical researchers (not only as physicians) actually existed, I did not connect the dots and see myself in any of those roles… whatever those roles were.

Continuing my speech, I described how in my first summer research program (after freshman year of college), I made and presented my first powerpoint presentation, how I was excited about my 3 x 4 excel table that depicted the 10's of damaged neurons that I counted by hand, the result of my one major experiment. I described how I had been transformed and how my mindset began to change and how the three subsequent (and more challenging) internships I participated in had positively affected me – personally, professionally and scientifically. I then promptly transitioned my comments by saying with emphasis that "Graduate school… is not exactly like a 3-month undergraduate research program" – to laughter, loud cheers and applause. The implication hopefully obvious in that it is more challenging and difficult.

At that time, I had completed only 2 years of grad school and had completed just rotations, classes and a qualifying exam. I thought I knew what I was talking about – primarily concerning the difficulty. The words were not exactly false, and they were not completely hollow but they were not yet coming from a place of depth, at least not the depth that they would come from had they been delivered today. Now, at the end of this stage, I can confidently say that navigating, growing and ultimately succeeding in my graduate career has been easily the most rewarding and indescribably challenging process of my relatively young life. I owe all of my successes (past and future), strength, survival and my transformation within science to innumerable individuals.

Firstly, I would like to acknowledge my earliest mentor- Dr. L. Paul Rosenberg whose unwavering faith in me as an undergraduate and constant

encouragement and support aided me at every single stage in these endeavors. I credit Dr. Rosenberg with sparking my interest in a research-focused career and being the positive and warm role model that I think every doubtful and insecure young student desperately needs. I thank my first P.I., Dr. Mark Macbeth for taking a chance on me – a student with very little Biochemistry background - and entrusting me with the exciting, but fledgling, Dbr1 project; also for patiently and kindly mentoring me in the early, rougher years of my development.

I am extremely grateful to all of the friends, classmates and colleagues that have supported and guided me throughout this process– especially Karen Kormuth, James Hughes, Shanna Bowersox, Kalin Vasilev, Ardon Shorr, Bianca Jones and Jessica Davis; members of the Blacklow lab – especially Brandon Zimmerman, Sanchez Jarret, Tom Seegar; unofficial mentors that stepped up on their volition including Brandon Ogbuno and Jason Huestis. I am so very grateful to have had support and mentorship from my labmates: Longfei and Chunxiao as well as from the members of SBGrid, especially Jason, Pete, Mick, Carol and Justin. I am appreciative of the environment created by former lab members James Stowell, Adriana Jaimes, Sophia Lee, Cassandra Sunga, Kira Roth and Nozhat Safaee.

I would like to thank the agencies that have financially supported my graduate studies – the National Science Foundation and the UNCF-Merck Graduate Research Science Initiative – for giving me limitless opportunities to explore my interests. I would also like to express my appreciation for the two mentoring relationships that stemmed from the UNCF-Merck program – Dr.

Corey Strickland and Dr. Kafui Dzirasa, both of whom went above and beyond to volunteer their time, guidance and friendship.

I would like to sincerely thank my Dissertation Advisory Committee – Dr. Richard Gregory, Dr. Jack Szostak and Dr. Steven Buratowski for their time and investment in my growth as a scientist. Their guidance, both in committee meetings and informal discussions, has been invaluable in shaping the direction of my work. Their willingness to support and allow me to pursue new scientific interests and directions and to take risks has been hugely beneficial. I am so grateful to Dr. Dan Kahne, Dr. Suzanne Walker and Jason Millberg for accepting me into the Chemical Biology Program and providing a welcoming and productive community in which I could grow personally and professionally. I also want to thank my awesome, talented cohort of graduate students who have significantly enriched my life over the last several years.

I owe a huge debt of gratitude to my P.I., Dr. Piotr Sliz for his enduring patience and goodwill. I am so grateful for having been given the freedom to pursue my own interests and experiences, even in the exciting and nerve wracking times when the path and the story were not always clear. I have truly enjoyed learning from him and I feel very honored to have been part of the lab. I thank my family, especially my parents for their continued support and care.

Finally, in so many ways, at so many times and for so many reasons, I could not and would not have completed this journey without Dawson Rauch.

I thank you all.

Chapter 1

The Roles Of Proteins In RNA Processing Pathways

The transfer of genetic information and the execution if its directives are the two most fundamental, all encompassing processes that maintain cellular life. While the Central Dogma of Biology (Crick, 1958) defined the movement and maintenance of genetic information by presenting a model in which information was transformed and unidirectionally shuffled between the three major classes of macromolecules: DNA to RNA to protein. The interdependent nature of the three components of this system, however, presented a 3-way ‘chicken before the egg’ problem – implying an earlier system evolved into the current one.

Of course the consensus now is that the “RNA World” (coined by Walter Gilbert, 1986) – a world in which RNA is the primary and progenitor molecule that self-assembled from primitive chemical building blocks in the prebiotic soup – is that system. The hypothesis was solidified with the discovery of catalytic RNA enzymes (ribozymes) in Tetrahymena (Kruger et al., 1982), which was seminal to the theory as it revealed not only that RNA had functional roles outside of the encoding of information, but additionally, this particular function was unnecessary, given the splicing machinery – suggesting the activity was a remnant of a time in which proteins were not present to facilitate these reactions. The subsequent discovery that arguably the most important cellular machine, the ribosome, was in fact a ribozyme, with RNA contributing not only to the structure of the complex but also participating directly in the reactions that generate polypeptide chains catapulted RNA into the spotlight.

Subsequently, the repertoire of characterized activities expanded from the primary genetic encoding role (messenger RNA, mRNA), and components of

protein and ribosomal synthesis (transfer RNA, tRNA; ribosomal RNA, rRNA) to functions previously expected to be carried out exclusively by proteins. This includes roles in the regulation of gene expression in a variety of contexts (microRNA, miRNA; silencing RNA, siRNA; long non-coding RNA, lncRNA riboswitches), cell organization and compartmentalization (mRNAs, rRNA), splicing (small nuclear RNAs, snRNA) and directing post-transcriptional modifications (snoRNA, small nucleolar RNA). Unsurprisingly, RNA is considered one of the most versatile and high-interest subjects in modern biology.

RNA binding proteins (RBPs) play seminal roles in the function of RNAs and often maintain delicate and highly spatially and temporally regulated interactions that inhibit, promote, process or complement RNAs and RNA activity with a variety of cellular consequences. Indeed, functional, noncoding RNAs seldom exist as solitary molecules in the cell but are most often in complexes with proteins as functional ribonucleoprotein (RNP) complexes. As such, RBPs are integral components of virtually all RNA processing pathways – including biogenesis, transcription, pre-mRNA splicing, post-transcriptional editing, translation and degradation. Importantly, mutation or deregulated expression of RBPs has increasingly been found to cause deleterious effects in RNA metabolism leading to a myriad of human diseases and cancer (reviewed in (Cooper et al., 2009; Lukong et al., 2008)). Thus, characterization of proteins that process or regulate RNA is not only of mechanistic value to the RNA biology community but also offers a major avenue towards the advancement of novel therapeutics.

For several decades, the predominant strategy to discerning protein function has centered on the application of the Structure-Function Paradigm – the assumption that the specific activity of a protein is determined by its three-dimensional structure. Indeed, the paradigm is so supported/validated that entire fields – like structural genomics have developed to attempt to indiscriminately determine and computationally predict the structure of every protein encoded in a genome. X-ray crystallography is arguably considered the most powerful technique in structural biology as it is largely responsible for the highest resolution structural information of protein, nucleic acid, small molecule and complexes. This information can generally be coupled with biochemical or other experimentation, to provide functional understanding. Indeed, few could forget how the elucidation of the double helical structure of DNA, combined with the determination of chemically similar and complementary nucleotide bases (among other work), ultimately led to the understanding of the basis of genetic encoding and DNA duplication which, in turn, laid the foundation for the Central Dogma.

A limitation of X-ray crystallography and other structural methods, however, is that the nature of the techniques confine investigators to single-focus studies, which are time consuming and laborious. This limitation is particularly problematic for RBPs that function in RNA processing pathways as they can generally interact with a multitude of RNA targets.The advancement of high throughput, bioinformatic methods, thus, proved invaluable to the expansion and complementation of structural techniques in the elucidation of RBP function and mechanism. Specifically, advances in RNA sequencing technologies allowed

investigators to systematically sequence full transcriptomes (deep sequencing), illuminating tens of thousands of RNA molecules, providing the ability to address numerous biological questions simultaneously and comprehensively.

Combining traditional biochemical techniques with new high throughput sequencing methods, such as with (RBPs) (RIP) coupled with deep sequencing of captured RNAs has allowed investigators to globally identify conditional RNA targets of RBPs and additionally determine the consensus sequence motifs that underlie molecular recognition. The wide utility of RIP and the related and more advanced CLIP (crosslinking and immunoprecipitation of RBPs) methods has allowed the identification of global RNA targets for several major RBPs such as Argonaute (Hafner et al., 2010), HuR (Kishore et al., 2011), and eIF4AIII (Sauliere et al., 2012). The complementation of structural and high throughput sequencing methods has offered unmatched functional insight into RBPs and the pathways they affect.

There are numerous RBPs, however, that do not fit within the typical bounds of the Structure-Function paradigm. A significant portion of the human (and, broadly, eukaryotic) proteome is predicted to be unstructured or structurally disordered. ~33% of eukaryotic proteins were predicted to contain regions of significant disorder (PONDR score >0.5, spanning at least 30 consecutive residues) (Ward et al., 2004), suggesting regions of protein disorder are functional and not merely artefacts or linkers. Indeed, disordered regions have been shown to contain low-complexity motifs and patches of charge that facilitate multivalent interactions with targets and other proteins (Lin et al., 2015). As

disordered proteins are largely resistant to functional structural determination, biochemical and in cellulo microscopy investigations of such RBPs have offered alternative means to comprehensively study both RBPs and the RNAs they target in these unique, high interest roles.

In this dissertation, I have utilized the aforementioned approaches to illuminate the roles of three components of major cellular RNA processing pathways and address outstanding questions that persist in the RNA biology field.

Specifically, in Chapter 2, I discuss mechanistic evaluations of the RNA Lariat Debranching Enzyme (Dbr1), a component of the Intron Degradation Pathway. As ~90% of nucleotides of RNA Pol II transcripts reside in introns, turnover and subsequent degradation is critical to maintaining nucleotide levels that support transcription as well as generating precursors for the synthesis of dNTPs. The unique 2'-5' nuclease activity of Dbr1 is the required first step of intron turnover, which allows digestion by the exosome and other exonucleases. I used X-ray crystallography, coupled with in vitro activity assays, to interrogate the nature of Dbr1 activity and target recognition. As an aside to the structural and biochemical work outlined in Chapter 2, in Appendix A, I discuss the establishment of an X-ray diffraction data publication and dissemination system, Structural Biology Data Grid (SBDG; data.sbgrid.org), to preserve the primary experimental data sets that support scientific publications.

In Chapters 3 and 4 I explore the oncoprotein LIN28, a major component of the let-7 miRNA biogenesis pathway. MiRNAs are small regulatory molecules

that are critical to the regulation of gene expression with particular implications in development and cancer. A well characterized activity of LIN28 is to bind to and negatively regulate let-7. In recent times, this activity has been expanded by high-throughput methods that determined LIN28 additionally interacts with mRNAs widely throughout the transcriptome. Thus, LIN28 has become a subject of popular CLIP methods, with little or no methodological validation. Therefore, in Chapter 3 I detail my work completing a bioinformatic, comparative analysis of several CLIP high-throughput study results and validate against crystallographic data. In Chapter 4, I additionally discuss the development and implementation of a new technique to help validate and/or relieve ambiguity of precise sites of crosslinked interactions.

Following the binding of let-7 miRNAs, LIN28 completes its inhibition of the let-7 biogenesis pathway by recruiting the terminal uridylyltransferase, Tutase, to oligouridylate the miRNA, thus marking it for degradation by the Dis3l2 exonuclease. In Appendix B, I discuss determination of the basis of LIN28 mediated oligouridylation using in vitro biochemical assays, coupled with the determination and analysis of an X-ray crystal structure of the human LIN28.

Finally, DEAD-Box helicases are multifunctional, ATP-dependent RNA remodelers that function in virtually every aspect of RNA metabolism (reviewed in (Janknecht, 2010)). DEAD-Box helicase P72, acts antithetically to LIN28 in the miRNA biogenesis pathway by serving as an accessory factor to the Microprocessor. A significant portion of P72 is predicted to be structurally disordered – the two termini specifically, which is a common trait of DEAD-Box

proteins. It is presumed that sequences in the termini dictate different functions, cellular localizations and associations. Thus, in Chapter 5 I describe my work in the characterization both in vitro and in vivo of a, P72 phase separation mediated by intrinsic disorder.

Overall, the work here presents a broad study, utilizing diverse approaches and techniques to understand the mechanisms and functions of important protein components of RNA processing pathways.

Chapter 2

Crystal structure of the *Entamoeba histolytica* RNA Lariat debranching enzyme, EhDbr1, reveals a catalytic Zn^{2+}/Mn^{2+} heterobinuclear active site

Contributors: Elizabeth M. Ransey, Eduardo Paredes, Sourav K. Dey, Subha R. Das, Annie Heroux, and Mark R. Macbeth

Supplementary Materials for this Chapter are in Appendix C.

At the time of submission, this Chapter is accepted for publication at *FEBS Letters (2017).*

Introduction

Lariats are the natural by-products of Group II and spliceosomal introns (Domdey et al., 1984; Padgett et al., 1984; Ruskin and Green, 1985) and form when the 5' end of an intron is ligated to the 2' hydroxyl of an internal adenosine branchpoint during splicing. The discovery of embedded miRNA and snoRNA sequences within intronic lariats (Okamura et al., 2007; Ooi et al., 1998; Ruby et al., 2007) conferred on them a critical role in the generation of such regulatory molecules with the potential to affect numerous downstream cellular processes.

The 2'-5' phosphoesterase (debranching) activity catalyzed by the RNA lariat debranching enzyme, Dbr1, was first discovered within HeLa cells when excised intronic lariats were linearized upon nuclear extract incubations (Ruskin and Green, 1985). The DBR1 gene was discovered when a genetic screen for cellular factors required for yeast Ty1 retrotransposition revealed a deletion that resulted in a phenotype of reduced retrotransposition and accumulation of intronic RNA lariats (Chapman and Boeke, 1991).

While the DBR1 gene is nonessential in lower eukaryotes (Chapman and Boeke, 1991), it has been speculated that this is due to comparatively low numbers of introns in such organisms. *dbr1*$^{-}$ strains *of Saccharomyces cerevisiae* (which contain ~255 introns (Lopez and Seraphin, 2000)) and *Schizosaccharomyces pombe* (~4,730 introns (Wood et al., 2002)) demonstrate a phenotype of accumulated lariats while mutants of the latter also demonstrate severe growth defects and aberrant cell morphology (Nam et al., 1997). *Caenorhabditis elegans* and *Drosophila Melanogaster dbr1*$^{-}$ mutants

demonstrate partial lethality, slow growth and/or sterility (Conklin et al., 2005) (WormBase). Dbr1 was required for HIV-1 propagation in a human osteosarcoma (HOS) cell line, the replication of which mimics retroelement replication (Galvis et al., 2014; Ye et al., 2005). Finally, siRNA knockdown of Dbr1 rescues TDP-43 toxicity (pertinent to amyotrophic lateral sclerosis (ALS)) in a human M17 neuroblastoma cell line and in rat primary cortical neurons – substantiating Dbr1 as potential therapeutic target in the treatment of neurodegenerative diseases (Armakola et al., 2012).

Dbr1 belongs to the phosphoesterase superfamily of enzymes that includes Mre11, due to the presence of three signature motifs of the class, namely: DIH-(X_{25})-GDYVDR-(X_{27})-GNHE (Hopfner et al., 2000; Koonin, 1994). This assignment was corroborated by structure-function analyses as mutation of several putative catalytic residues within the aforementioned motifs resulted in lariat accumulation *in vivo* and decreased or abolished debranching activity in vitro (Khalid et al., 2005). Based on homology modeling and residue conservation, the binuclear active site observed in Mre11, in which catalytic residues coordinate two Mn^{2+} ions (Hopfner et al., 2000), had been proposed to also exist within Dbr1. Curiously, the first crystal structures of Dbr1 identified Mn^{2+} in only one of the metal-binding sites, while the other site was vacant (Montemayor et al., 2014). Recent co-crystal structures solved by the same group, however, indicated a Zn^{2+}/Fe^{2+} heterobinucleation and in vitro debranching assays determined this metal composition was optimally active (Clark et al., 2016).

Here we present an X-ray crystal structure of the *E. histolytica* Dbr1 (EhDbr1) to a maximum resolution of 1.8 Å. Our structure model indicates a novel heterobinucleation consisting of one Mn^{2+} ion and one Zn^{2+} ion. Inductively coupled plasma atomic emission spectroscopy (ICP-AES) determined that purified EhDbr1 exists in primarily two states – one that is heteronucleated with Zn^{2+} and Mn^{2+} and the other that is homonucleated with Zn^{2+}. In vitro debranching assays determined this mixed sample is comparably active to previously characterized EhDbr1 samples that contain different metal configurations. These results expand upon metal composition characterizations for EhDbr1 and are consistent with the diverse metal configurations that are occasional features of this class of enzymes.

Results

Overview of EhDbr1 structure

X-ray data collection and refinement statistics are reported in Table 2.I. EhDbr1 crystallized in a $P2_12_12_1$ space group with 1 molecule in the asymmetric unit. The maximum resolution of the diffraction data was 1.8 Å. Preliminary ICP-AES analyses determined that the purified protein sample contained physiological quantities of both Mn^{2+} and Zn^{2+} ions (data not shown). Thus, experimental phase information was obtained via single-wavelength anomalous dispersion (SAD) techniques utilizing Zn^{2+} ions present within each protein molecule in the crystal. The structure model was refined to an R_{work}/ R_{free} of 0.17/0.21. The high-resolution structure excludes residues 1-6 of the native protein, and contains

three non-native glycine residues as cloning artifacts before glutamine at position 7.

The X-ray structure we determined is nearly identical to the most recent apo-enzyme structure PDB: 5K73, with an RMSD over 349 residues of 0.424 Å (angstroms) for backbone atoms. Briefly, the EhDbr1 amino-terminal domain (NTD) (residues 7-272), features a globular, metallophosphoesterase (MPE) fold composed of two β-sheets (12 β-strands in a mixed parallel and antiparallel conformation) that creates a central β-sandwich surrounded by 12 α-helices (Figure 2.1A). The carboxy-terminal domain (CTD) (residues 273-354), which varies widely in length and sequence amongst Mre11 proteins, is helical and is located on the posterior of the protein, opposite the active site (Figure 2.1A).

Table 2.I. Data Collection and Refinement Statistics for EhDbr1 structure

PDB ID Code	**5UKI**		
Data Collection			
Wavelength, Å	1.10	1.28	1.30
Space Group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Unit Cell, Å	45.34, 70.92, 109.27	45.60, 71.17, 109.64	45.61, 71.16,109.67
Resolution range, Å	41.88-1.80 (1.85-1.80)	50.00-2.05 (2.09-2.05)	50.00-2.10 (2.13-2.10)
R_{merge}, %	0.074 (0.529)	.065 (0.610)	0.060 (0.762)
I/σI	34.1 (3.81)	45.36 (2.16)	37.5 (1.81)
Completeness, %	95.29 (69.45)	99.2 (93.3)	99.4 (98.5)
Redundancy	12.1 (11.0)	13.0 (6.8)	8.9 (5.2)
Refinement			
Resolution, Å	1.8		
Unique Reflections	30215		
R_{work}/R_{free}	0.174/0.210		
No. atoms	3035		
Protein	2871		
Ligand/ion	2		
Water	162		
B factors (Å^2)	31.42		
Protein	10.54		
Ligand/ion	24.77		
Water	33.43		
R.m.s. deviations			
Bond lengths (Å)	0.018		
Bond angles (°)	1.788		

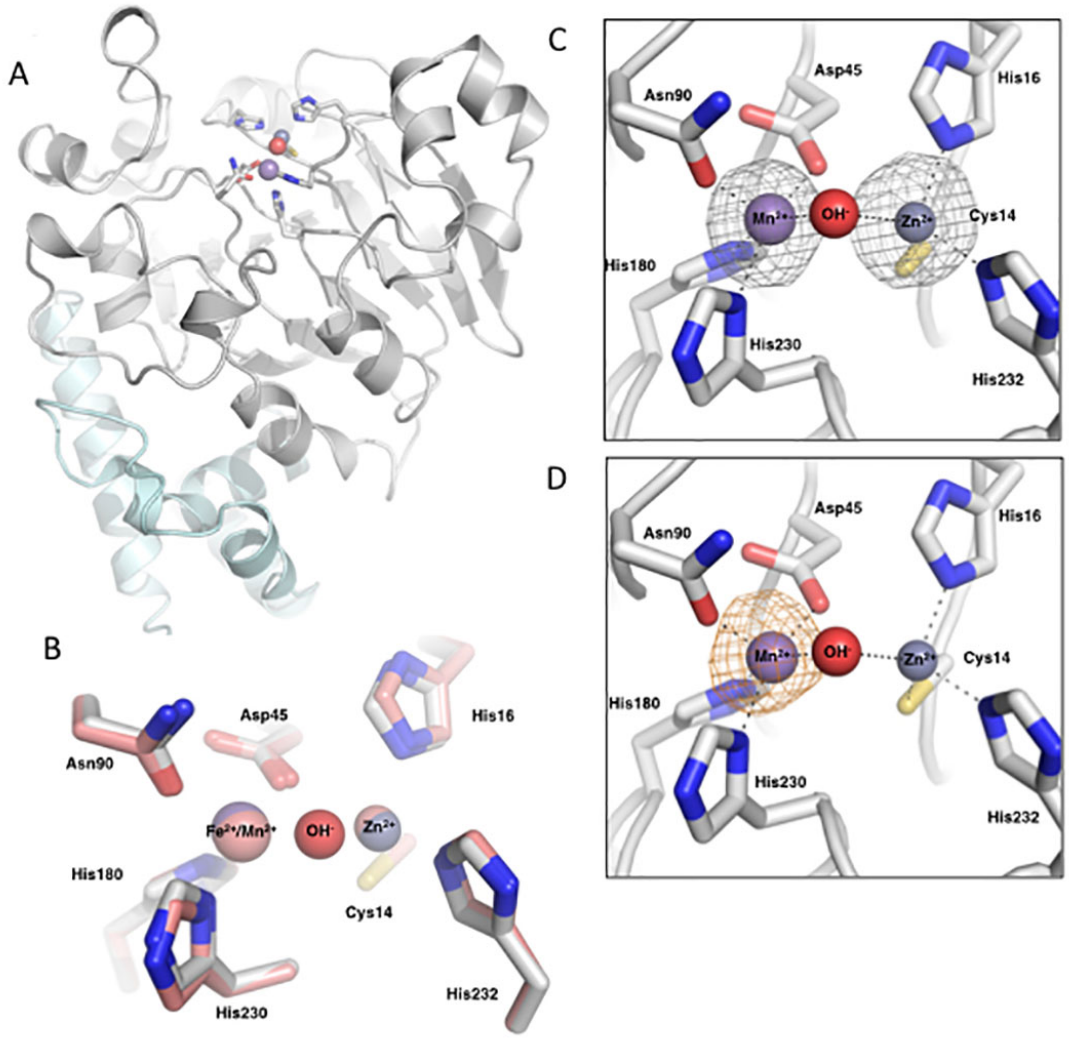


Figure 2.1. Structure model of the EhDbr1. (A) Cartoon representation of Dbr1. The MPE fold characteristic of the Mre11 superfamily is displayed in gray. The carboxy-terminal domain is in pale green. The lariat binding cleft is at the top, showing the Mn^{2+}(purple sphere), Zn^{2+} (gray sphere), the OH^- nucleophile (red sphere) and their coordinating residues (sticks). (B) Overlay of the active site of our model (gray sticks, purple Mn^{2+}, gray Zn^{2+}) with the model of Clark et al., 2016 (Clark et al., 2016) PDB: 5K73 (pink sticks and spheres), which shows the superposition and similar coordination of the Fe^{2+} ion in their structure and the Mn^{2+} ion in ours. (C) Active site residues (sticks) coordinate a binuclear metal cluster consisting of one Mn^{2+} ion (purple sphere) and one Zn^{2+} ion (gray sphere) with a bridging hydroxyl ion (red sphere). Superimposed on the model are anomalous scattering maps of electron density (gray mesh) from data collected using X-rays of 1.28 Å wavelength. The maps are contoured at 5s. (D) Our model with superimposed anomalous scattering maps (orange mesh, contoured at 5s) of diffraction data collected using X-rays of 1.30 Å wavelength (beyond the Zn^{2+} absorption edge), demonstrating differential density at the two metal sites.

Our structure reveals active site features that are consistent with the previously determined structures (Clark et al., 2016; Montemayor et al., 2014). Namely, residues His16, Asp45, Asn90, His180, His230 and His232 directly coordinate two metal ion binding sites and have conserved counterparts with Mre11 proteins. Superposition of the active site of our structure model with the active site of the Clark et al., 2016 model, demonstrates the similar coordinations of both metal binding sites (Figure 2.1B). An otherwise invariant catalytic aspartic acid residue within the Mre11 superfamily (Matange et al., 2015) has been evolutionarily mutated to a cysteine. This cysteine residue is present in EhDbr1 proteins and participates in the coordination of the α-site metal, in place of the aspartic acid found in Mre11.

High-resolution crystal structure reveals a Zn^{2+}/Mn^{2+} heterobinucleated EhDbr1 active site

To determine the metal ion composition of the EhDbr1 active site, we collected and compared diffraction data generated from irradiation using X-rays tuned to different wavelengths. Anomalous difference maps generated from data collected at λ=1.28 Å, revealed density at both metal binding sites, whereas anomalous difference maps from data collected at λ=1.3 Å, only revealed density within the β-site (Figure 2.1C, D). Given the absorption edges of Mn^{2+} and Zn^{2+} ions are λ=1.8961 Å (keV= 6.5390) and λ=1.2837 Å (keV= 9.6586), respectively, and that the previously mentioned ICP-AES analyses determined stoichiometric amounts of both ions in our sample, we determined that Mn^{2+} resides within the

β-site and Zn^{2+} within the α-site. The data does not preclude a Zn^{2+} from occupying the β-site, however in this instance, Zn^{2+} is not contributing to the observed β-site density. In our model, Asp45, Asn90, His180, His230, and the nucleophilic hydroxyl ion coordinate the Mn^{2+} ion in a trigonal bipyramidal configuration while Cys14, His16, His232 and the same hydroxyl coordinate the Zn^{2+} ion in a tetrahedral geometry (Figure 2.1C, D).

Zn^{2+}/Zn^{2+} and Zn^{2+}/Mn^{2+} nucleated active sites support catalytic activity

To determine the physiological relevance of EhDbr1 proteins incorporating Zn^{2+} and Mn^{2+}, we completed ICP-AES metal quantitation analyses and in vitro activity assays. Our data revealed that EhDbr1 samples purified and dialyzed in GFB in the absence of added metal contained 1.42 ± 0.0029 molar equivalents of Zn^{2+}, 0.75 ± 0.051 molar equivalents of Mn^{2+} and 0.21 ± 0.11 molar equivalents of Fe^{2+} (Table 2.II), suggesting two primary populations of EhDbr1 – homonucleated with Zn^{2+} and heteronucleated Zn^{2+}/Mn^{2+}. To test the activity of the sample, we completed in vitro debranching assays utilizing a fluorescently-labeled, synthetic substrate. The substrate is a 19-nt RNA, consisting of a 13-nt stem with a 5'-terminal Dylight 547 (Dy547; a Cy3 equivalent) fluorescent dye and a 6 nt branch linked at the 2'-hydroxyl of the branch point adenosine (Figure 2.2A). Debranching reactions were initiated by mixing EhDbr1 with the RNA substrate in reaction conditions (50mM Tris-HCl (pH=7.5) buffer with 25 mM NaCl, 2.5mM DTT and 0.01% BSA) with and without supplemented 1 mM

$Mn^{2+}$and were stopped at indicated time points with formamide before separation via denaturing PAGE.

With equal concentrations of EhDbr1 and substrate (1 μM each), the substrate was cleaved >50% at the first time point (Figure 2.2B), which hindered reliable fit to obtain a cleavage rate constant. We therefore used 1 nM EhDbr1 with 100 nM substrate to obtain an observed rate constant (k_{obs}) of 0.025 s^{-1} for the cleavage reaction that reflects both binding and cleavage for multiple turnovers. At higher concentrations of enzyme, the reactions were too fast for reliable manual aliquots and measurement. The turnover rate (k_{cat}) for EhDbr1 was subsequently determined to be 0.19 ± 0.01 s^{-1} (with K_M = 25.9 ± 3.5 nM) using a dual fluorescently labeled (with Dy547 and Cy5) branched RNA ranging from 20 to 120 nM in a FRET based cleavage assay using a fluorimeter (Appendix Figure C.S1). Details of this assay and additional parameters of related cleavage reactions with EhDbr1 will be reported elsewhere as they are outside of the scope of this Chapter.

Table 2.II. ICP results indicating Zn^{2+} and Mn^{2+} composition of dialyzed samples

Sample	Mols Zn/ Mol Dbr	Mols Mn/ Mol Dbr	Mols Fe/ Mol Dbr
Average	1.42	0.75	0.21
S.D.	0.0029	0.051	0.11

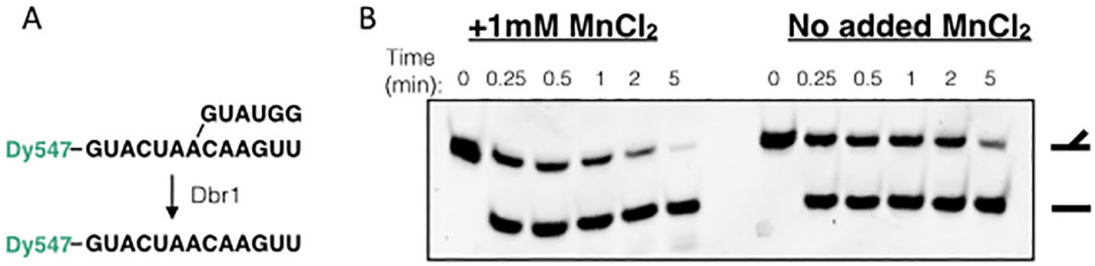


Figure 2.2. Zn^{2+}/Mn^{2+} and Zn^{2+}/Zn^{2+} incorporated Dr1 active sites are active. (A) Schematic of the synthetic 5'-Dy547 labeled, branched RNA substrate and the debranching assay. (B) Debranching assay comparing purified Dbr1 sample activities in the presence and absence of added $MnCl_2$.

Discussion

Here, we present the X-ray crystal structure of the *E. histolytica* Dbr1 (refined to 1.8 Å) that complements existing EhDbr1 structural models. Our model is nearly identical to previously reported structures with the notable exception of the identified active site metal ions – we determined a novel $Zn^{2+/}Mn^{2+}$ heterobinucleation. This metal composition is supported by ICP-AES and the metal configuration within the active site is supported by anomalous scattering experiments with X-rays of different wavelengths. In vitro debranching assays demonstrate that a mixed population of Zn^{2+}/Mn^{2+} heteronucleated and Zn^{2+}/Zn^{2+} homonucleated is comparably active to EhDbr1 samples previously characterized in Clark et al., 2016 as being Fe^{2+}/Zn^{2+} heteronucleated.

The metal configuration in our model incorporates singular aspects of each of the two previous structure models, as the first reported structures identified a β-site Mn^{2+} occupancy (and vacant α-site) and subsequent models identified Zn^{2+} in the α-site (though Fe^{2+} in the β-site) (Clark et al., 2016; Montemayor et al., 2014). The apparent flexible metal binding behavior of EhDbr1 is unsurprising given that the enzyme was determined to be active in the presence of Mn^{2+}, Mg^{2+} and Ni^{2+}(Khalid et al., 2005), and additionally, Fe^{2+} and Zn^{2+} (Clark et al., 2016). Furthermore, several MPE enzymes of the MRE11 family have been shown to be active in the presence of diverse metals or to be heteronucleated or diversely homonucleated in different structure models (reviewed in (Matange et al., 2015)). The incorporation of Zn^{2+} and Mn^{2+} simultaneously, however, has neither been observed nor predicted in EhDbr1.

We propose that the differing methods of metal removal and enrichment during or following purification could be a reason for the difference in active site metals we describe and the other published structures. Thus, experiments probing the metal ion requirements for the catalytic activity of EhDbr1 are ongoing.

Materials And Methods

EhDbr1 construct cloning

The EhDBR1 cDNA, optimized for *E. coli* expression, was purchased from DNA 2.0, Inc. and was cloned from pJ204 into pET22b(+) by PCR amplifying the gene with flanking BamHI and XhoI restriction sites. The pET22b(+)- EhDbr1 construct was generated by completing appropriate restriction enzyme digestions and T4 Ligase ligation reactions. To improve diffraction, the first six residues were removed via Quick change site-directed mutagenesis generating EhDbr1(Δ1-6). All constructs were designed to incorporate a fused, upstream His-12 tag separated from the EhDbr1 cDNA by a TEV protease recognition sequence that facilitates subsequent purification and crystallization.

Expression and purification of recombinant EhDbr1

Bacterial expression cultures were initiated from single colonies of BL21(pRARE) *E. coli*, transformed with pET22b(+) EhDbr1 and plated on LB amp/chloramphenicol. Single colonies were picked and grown in 5 ml of LB with 50 μg/ml ampicillin and 25 μg/ml chloramphenicol, for 16-18 hrs at 37 ^{o}C. Starter cultures were used to inoculate 500 ml of LB (with 50 μg/ml ampicillin and 25

μg/ml chloramphenicol) and cultures were grown at 37 ^{o}C to an OD of 0.8 - 1.0. For induction, cultures were chilled on ice for 15 mins before the addition of the galactose analog, Isopropyl β-D-1-thiogalactopyranoside (IPTG) (1 mM) and ethanol (0.5 %), and subsequently grown at 17 ^{o}C overnight.

Cells were harvested via centrifugation (10 mins at 10,000 x g) approximately 18-20 hours after IPTG induction, lysed in Buffer A (20 mM Tris, pH 8.0, 5 % glycerol, 1 mM β-mercaptoethanol, 750 mM NaCl, 45 mM Imidazole, 1 mM $MnCl_2$) by high-pressure homogenization and cellular debris was removed via ultra-centrifugation (45 mins at 30,000 x g). Clarified cell lysate was batch-bound to Ni-NTA agarose beads then subjected to affinity column chromatography. The resin was washed with Buffer A and His-tagged EhDbr1 was eluted from the Ni-NTA column with Elution buffer (Buffer A with 400 mM Imidazole and 100mM NaCl). Column fractions were analyzed by SDS-Page and Coomassie Brilliant Blue staining (EhDbr1 is ~41 kDa) and elution fractions containing EhDbr1 were pooled and subjected to heparin sepharose chromatography. EhDbr1 was eluted from the heparin column using a gradient of low concentration of NaCl (20mM Tris, pH 8.0, 5% glycerol, 1mM β-mercaptoethanol, 100 mM NaCl) to high concentration of NaCl (20mM Tris, pH 8.0, 5% glycerol, 1mM β-mercaptoethanol, 1M NaCl) buffers. Fractions were analyzed by SDS-PAGE and those containing EhDbr1 were pooled and dialyzed in dialysis buffer (20mM Tris, pH 8.0, 5% glycerol, 1mM β-mercaptoethanol, 200 mM NaCl, 25 mM Imidazole) with TEV protease for 4 hrs at 22 °C and overnight at 4 ^{o}C.

TEV protease treated samples were then bound to equilibrated Ni-NTA agarose beads and subjected to non-affinity column chromatography utilizing Buffer A to wash off TEV-cleaved protein. Flow-through and wash fractions containing EhDbr1 were pooled and concentrated for size exclusion chromatography (SEC). SEC was completed using a HiLoad 16/600 S200 PG column (GE Healthcare Life Sciences) and gel filtration buffer (GFB) (20mM Tris, pH 8.0, 5% glycerol, 1mM β-mercaptoethanol, 200 mM NaCl). Finally, gel filtration fractions containing EhDbr1 were pooled, concentrated and quantified by A_{280} absorbance spectrometry: estimated ε = 64290 $M^{-1}cm^{-1}$. Protein yields ranged from 0.15-0.5mg/L culture.

ICP-AES metal analysis

Two replicate preparations of purified EhDbr1 samples (9.1 x 10^{-9} and 1.1 x 10^{-8} mols) were denatured in 5 ml of 5% HPLC grade Nitric Acid and sent to the Materials Characterization Laboratory at the University of Pennsylvania, University Park. Duplicate samples of each replicate were analyzed for the presence Co^{2+}, Cu^{2+}, Fe^{2+}, Mn^{2+}, Mo^{2+}, Ni^{2+} and Zn^{2+}. Co^{2+} and Mo^{2+} metal ions were undetectable. Fe^{2+}, however, was present at 0.2 ± 0.1 molar equivalents in both replicates and Cu^{2+} and Ni^{2+} were present at 0.1 and 0.2 molar equivalents in one replicate (below detection in the other).

Crystallization of EhDbr1(Δ1-6)

Crystal screens with purified EhDbr1 and EhDbr1(Δ1-6) were completed using the sitting-drop crystal screening method with Hampton Research and Qiagen screening suites. An Art Robbins Crystal Phoenix robot was used for initial screening. Preliminary crystal hits found to support micro-crystal growth were replicated and optimized by manually screening various conditions.

Conditions that supported high resolution diffracting crystals were 24% PEG 8000, 0.1 M HEPES, pH 7.5, and 5 mg/mL EhDbr1(Δ1-6), 4^{o}C. Cryo-protected crystals were grown in the crystallization condition with the addition of 10-20% glycerol. Crystals were initially screened at the University of Pittsburgh X-ray facility using a Saturn Kappa CCD detector and a Cu-K_{α} rotating anode X-ray source. Crystals of sufficient quality were sent to Beamline X-25 at the National Synchrotron Light Source (NSLS) for single wavelength anomalous dispersion data collection using X-rays of 1.1 Å wavelength. Diffraction data were integrated and scaled using HKL2000 and the metal ions were located using SHELXD. Density modification and solvent flattening was performed using RESOLVE as implemented in the PHENIX crystallographic software package. Manual building of the protein model was performed using COOT and the model was refined using REFMAC5 as implemented with the CCP4i suite of crystallographic software.

To differentiate between the anomalous scattering contributions of the Mn^{2+} and Zn^{2+}, diffraction data from a second crystal were collected using X-rays with wavelengths of 1.28 Å (at the absorption peak of Zn) and 1.3 Å (beyond the

Zn^{2+} absorption edge). Phases for each data set were calculated using PHASER, employing our coordinates from the 1.8 Å data set as the input coordinates, with the coordinates for the metal ions removed. $2F_{obs}$-F_{calc} maps were readily interpreted in COOT and anomalous scattering maps were created using FFT in the CCP4i suite using mtz files from PHASER as input. Figures were made using PyMOL. Data collection and refinement statistics are shown in Table 2.I. The coordinate and structure factor files have been submitted to the Protein Data Bank and have been assigned the following PDB: 5UKI.

Synthesis and sequence of branched RNA substrate

Synthesis of the branched RNA substrate was done on the solid-phase in a MerMade 4 synthesizer (Bioautomation Inc, Irving, TX) by including a 2'-*O*-photoprotected adenosine phosphoramidite as the branchpoint residue during otherwise standard RNA coupling. The free 5'-OH of the linear sequence was capped by coupling with Dylight547 (Dy547; a Cy3 equivalent) phosphoramidite (Glen Research corp.) and selective removal of the cyanoethyl group was achieved by treating the CPG beads with a 3:2 mixture of acetonitrile and triethylamine for 90 minutes. The CPG beads were washed extensively with acetonitrile and THF and the synthesis beads were transferred to a 1.8 mL glass vial, 1 mL of acetonitrile was added and photodeprotection of the branchpoint adenosine 2'-hydroxyl group was achieved by irradiating the vial with a 100W long wave UV lamp (UVP-model B 100AP) for 60 mins. Following the photodeprotection, the synthesis beads were washed with acetonitrile and placed

in an empty synthesis column. The synthesis column was placed on the MerMade instrument and washed with 3x1 mL of anhydrous acetonitrile. The 2'-5' branch synthesis was then performed using 'reverse' phosphoramidites (Chemgenes) and 10 min coupling times with ethylthiotetrazole as activator. Deprotection of the branched RNA was conducted by standard methods. The sequence synthesized was 5'-Dy547- GUACUAA-(2'-5'-GUAUGG)-CAAGUU with the underlined A denoting the branchpoint adenosine and 2'-branch sequence in parenthesis. The sequence was purified by polyacrylamide gel electrophoresis, and desalted before further use. MALDI mass spectrometry confirmed the successful synthesis of the branched RNA (mass found 6581.7; mass calculated 6581.1). Full details of the chemical synthesis of the protected branchpoint adenosine as well as solid-phase branched RNA synthesis will be reported elsewhere.

Debranching assays and kinetic characterizations

All debranching assays were completed in 50 mM Tris, pH 7.5, 25 mM NaCl, 2.5 mM DTT, 0.01 % BSA in 10uL reactions with concentrations of purified EhDbr1 and Dy547-labeled branched RNA substrate indicated in text. EhDbr1 reactions were initiated with the addition of appropriate volumes of EhDbr1 sample to reaction mix, incubated at 30 ^{o}C and stopped with 90 % formamide/10 % EDTA, pH 8.0 at appropriate time points. Stopped reactions were electrophoresed at 250 volts in gels containing 20 % (19:1) polyacrylamide:

bisacrylamide and 7 M urea. Substrate and product bands were quantified by a GE Typhoon FLA 9000 imager.

Author Contributions

E.M.R and M.R.M conceptualized and designed the project. E.M.R. cloned, overexpressed and purified Dbr1 proteins; crystallized Dbr1, completed ICP analysis preparations and in vitro debranching assays. E.M.R, A.H. and M.R.M determined the structure of Dbr1. E.P and S.K.D synthesized debranching substrates and completed kinetic characterizations. E.M.R, S.R.D and M.R.M wrote the manuscript.

Acknowledgements

E.P. and S.R.D. gratefully acknowledge financial support from the David Scaife Family Charitable Foundation and by NIH grant R01GM110414. E.R. gratefully acknowledges financial support from the National Science Foundation Graduate Research Fellowship Program [DGE1144152] and the UNCF-Merck Graduate Research Science Initiative. We would like to thank Laura Dassama and Henry Gong of the Materials Characterization Lab at Penn State for completing ICP-AES analyses.

Chapter 3

Comparative analysis of LIN28-RNA binding sites identified at single nucleotide resolution

Contributors: Elizabeth Ransey, Anders Björkbom, Victor S. Lelyveld, Przemyslaw Biecek, Jack W. Szostak, Piotr Sliz

Supplemental Materials for this Chapter are in Appendix D.

At the time of submission, this Chapter is under review for publication at

RNA Biology

Introduction

LIN28 is a highly conserved RNA-binding protein that was first described as the product of the heterochronic gene, *LIN28*, in *C. elegans* (Ambros, 1989; Ambros and Horvitz, 1984; Liu and Ambros, 1989) In mammals, the two LIN28 paralogs, LIN28A and LIN28B, play roles in a wide range of cellular processes, including stem cell self-renewal and pluripotency (Qiu et al., 2010; Viswanathan et al., 2008; Xu et al., 2009) skeletal myogenesis (Polesskaya et al., 2007) glucose metabolism in diabetes (Zhu et al., 2011) and tissue repair (Shyh-Chang et al., 2013). These proteins are upregulated in ~15% of human tumors and cancer cell lines, and elevated expression is associated with a poor prognosis and increased aggression in numerous malignancies, including germ cell tumors, colon cancer, and ovarian cancer (King et al., 2011; Viswanathan et al., 2009). Signalling components in the Wnt pathway have been shown to cooperate with LIN28 to increase the severity and invasiveness of colorectal cancer (Tu et al., 2015) LIN28B also promotes metastasis of colon cancer (King et al., 2011) and tumorigenesis of the intestinal epithelium (Madison et al., 2013).

LIN28 exerts its profound phenotypic effects by acting as a negative regulator of let-7 miRNA biogenesis. Specifically, LIN28 binds let-7 precursors and prevents miRNA maturation, limiting cellular differentiation (Heo et al., 2009; Lehrbach et al., 2009; Newman and Hammond, 2010; Rybak et al., 2008; Viswanathan et al., 2008). High-resolution crystal structures of mouse LIN28A-let-7 complexes and other structural and biochemical data revealed that LIN28 inhibits Dicer processing (Heo et al., 2008; Rybak et al., 2008) through steric hindrance and by locally unwinding the cleavage site (Mayr et

al., 2012; Nam et al., 2011). Furthermore, LIN28 proteins recruit the terminal uridylyltransferase, TUT4, to uridylate bound miRNAs, resulting in degradation by the Dis3l2 exonuclease (Chang et al., 2013; Heo et al., 2012; Heo et al., 2008). Recent developments suggest that LIN28 also functions through let-7 independent mechanisms. Several studies utilizing crosslinking and immunoprecipitation with deep sequencing (CLIP-seq) and photoactivatable ribonucleoside analog CLIP (PAR-CLIP) have identified thousands of potential pre-mRNA or mRNA targets of LIN28 (Cho et al., 2012; Graf et al., 2013; Hafner et al., 2013; Stefani et al., 2015; Wilbert et al., 2012). These methods rely on irradiation by UV light to generate covalent RNA-protein heteroconjugates in live cells, allowing for the isolation of RNA binding proteins (RBPs) by immunoprecipitation and subsequent high-throughput sequencing of crosslinked RNAs. CLIP was originally developed to address limitations of non-covalent RBP immunoprecipitation (RIP) methods (Mili and Steitz, 2004), such as non-specific RNA target capture, loss of lower affinity targets, and a weak signal-to-noise ratio (Ule et al., 2005). Deep sequencing CLIP methods have identified global RNA targets for notable proteins such as Argonaute (Hafner et al., 2010), HuR (Kishore et al., 2011), eIF4AIII (Sauliere et al., 2012), DDX17 (Moy et al., 2014), snoRNA proteins (Granneman et al., 2009), and splicing factors, including PTBP1 and RBFOX (Li et al., 2015). CLIP investigations of the LIN28 paralogs, combined with functional assays, have revealed that both proteins are target-specific post-transcriptional and translational regulators that alter splicing factor abundance and alternative splicing, suppress translation of secretory pathway proteins, and mildly stabilize mRNA targets (Cho et al., 2012; Hafner et al., 2013; Wilbert et al.,

2012). Several CLIP studies have seen enrichment of mRNAs containing the sequence GGAG, the signature LIN28 recognition element present in let-7 miRNAs (Cho et al., 2012; Graf et al., 2013; Wilbert et al., 2012). Furthermore, one report also identified an enrichment of pyrimidine rich binding motifs, consistent with sequences recognized by the LIN28 cold shock domain (CSD) in let-7 targets (Hafner et al., 2013), suggesting that LIN28 interactions with mRNA may have binding determinants that mirror those previously identified in let-7 miRNA targets.

Building on CLIP methods, crosslink induced mutation site (CIMS) analysis has emerged as a powerful bioinformatic tool for the elucidation of single nucleotide resolution crosslink interaction information derived from CLIP-seq datasets. This method aims to identify mutations, primarily deletions and substitutions (Moore et al., 2014; Zhang and Darnell, 2011), that occur during reverse transcription at presumed crosslink sites within CLIP reads (Granneman et al., 2009; Ule et al., 2005). Despite the availability of numerous LIN28 CLIP-seq and PAR-CLIP studies, crosslink profiles from mutational analysis have been reported in only one study (see Cho et al., 2012), which found an increased prevalence of mutations at guanine residues that were apparently localized within a LIN28 GGAG recognition motif. To our knowledge, no other CLIP studies have been examined by CIMS, nor has the crosslinked side chain on LIN28 been concurrently elucidated (Lelyveld et al., 2015).

In this work, we examined the overlap of binding sites identified in five previously published human LIN28 CLIP datasets. For a subset of these, as well as mouse and worm LIN28 datasets, we additionally used CIMS analysis

to identify sequencing mutations presumed to indicate crosslink sites. We found significant discrepancies in determined global RNA binding sites and variability in crosslinked nucleotide identities across diverse data sets, although more agreement was found among mammalian datasets. Additionally, we used mass spectrometry (MS) to characterize products of UV-induced crosslinking within a well-characterized complex of LIN28A and precursor let-7 fragments. This allowed us to identify the discrete position of a crosslinked interaction on both the protein and the RNA components simultaneously. Overall, our work suggests that high-precision analysis methods of RNA-protein crosslinks must be cross-validated to avoid methodology-specific conclusions.

Results

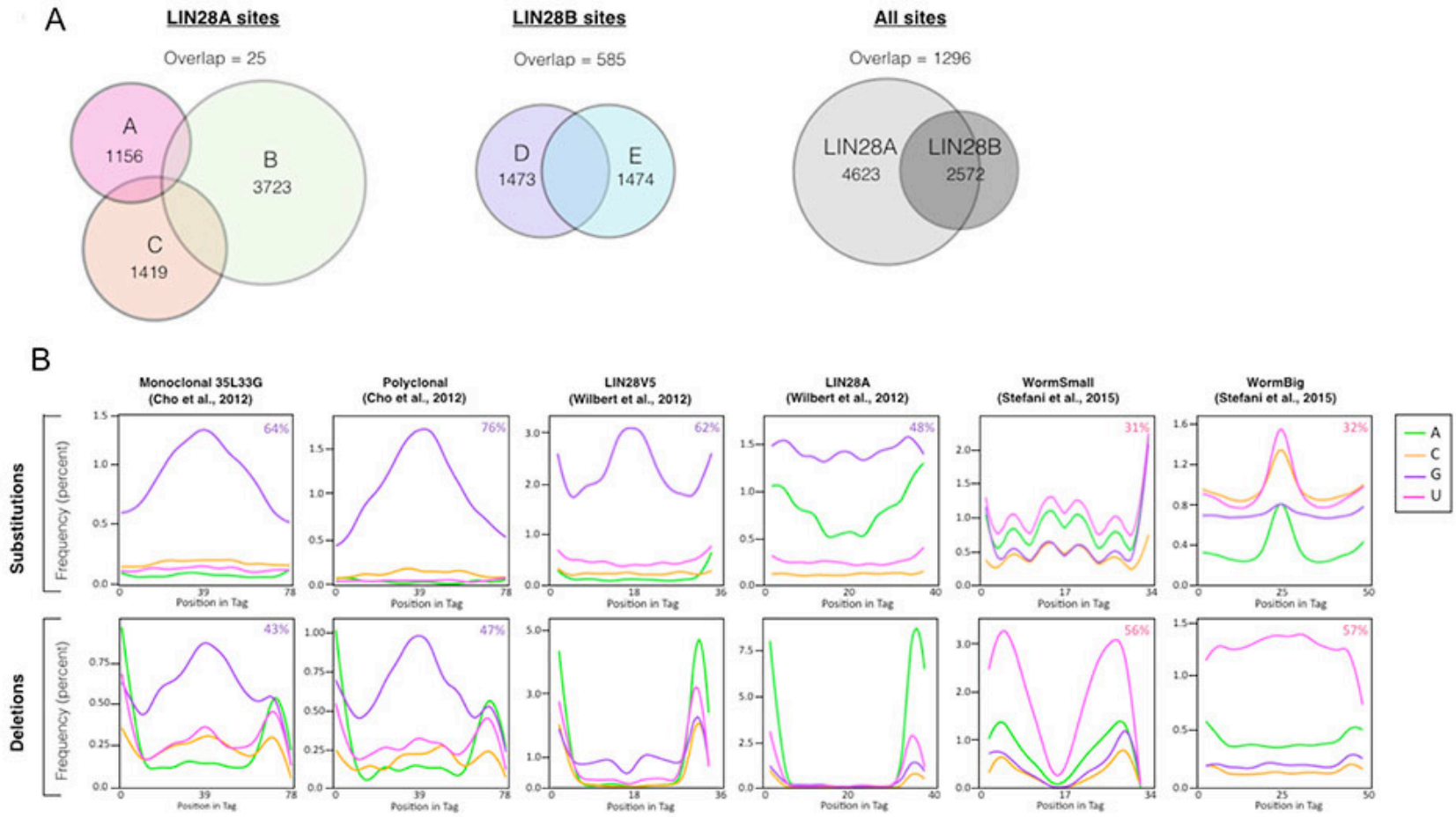
Mammalian CLIP-seq reads are enriched for guanine mutations but largely disagree on LIN28A transcriptomic targets.

If LIN28 can be crosslinked to target transcripts in a robust manner, we expect to find significant concordance between identified binding sites in multiple studies. To evaluate the consistency of LIN28-RNA UV-induced crosslinking and LIN28 genomic binding site determination, we compared identified binding sites within currently available human LIN28A and LIN28B CLIP and PAR-CLIP datasets. We analyzed five datasets from three prominent studies in human cell lines (see Wilbert et al., 2012; Hafner et al., 2013; Graf et al., 2013) (source details in Appendix Table D.SI). Binding sites were identified using Piranha software (Appendix Table D.SII; Appendix

Dataset D.S1) and direct comparisons were achieved by segmenting the hg19 reference genome into bins of 200 nt length and observing CLIP sequence read coverage across all bins for all datasets (Appendix Dataset D.S2). We determined that only a small percentage, 0.7% - 2.1%, of each individual dataset of LIN28A sites were in total agreement across all three datasets. On the other hand, 39.7% of each LIN28B dataset overlapped, though it should be noted that there were only two LIN28B studies (Figure 3.1A, Appendix Table D.SIII; Appendix Dataset D.S2).

Figure 3.1. Bioinformatics analysis shows small overlap between numerous CLIP datasets and identifies mutations within LIN28 CLIP reads. (A) Venn diagrams demonstrating the overlap of five CLIP-seq and PAR-CLIP datasets. Analyzed datasets include only those CLIP experiments performed in human cells, to avoid interspecies transcriptome variation complications. All identified binding sites and overlapping binding sites are listed in Supplemental Datasets 3.1 and 3.2 and an overlap comparison summary is listed in Appendix Table D.S3. (B) Mutation frequency profiles of CLIP reads generated by CIMS analysis. Mono35L33g and Polyclonal datasets are M. musculus; LIN28V5 and LIN28A are H. sapiens datasets; and WormSmall and WormBig are from C. elegans. Selected datasets include only CLIP experiments performed in tissue culture in the absence of photoactivatable ribonucleoside analogs (PAR-CLIP), to avoid well-known mutation bias. Percentage numbers and colors within profile plots indicate dominant nucleotide identity and relative enrichment.

Figure 3.1. (Continued)



RNA-protein UV crosslinking causes observable mutations in CLIP sequencing reads, which are presumed to be indicative of crosslinking sites and can be characterized using CIMS analysis (Moore et al., 2014; Zhang and Darnell, 2011). To validate our bioinformatic workflow, we reanalyzed two mouse LIN28A CLIP datasets for which a CIMS mutational analysis was reported (see Cho et al., 2012). Consistent with that work, our analysis of the monoclonal 35L33G and polyclonal antibody CLIP datasets showed that mutations arose most frequently at guanines (Figure 3.1B). Though we observed similar mutation identities and positions, our frequencies were lower, likely due to differences in filtering parameters. Nonetheless, we determined guanines make up 64% and 76% of substitution sites and 43% and 47% of deletion sites for monoclonal and polyclonal antibody datasets, respectively (Figure 3.1B).

To examine the consistency of this observation across published LIN28 crosslinking studies, we applied our validated analysis to four published datasets for which crosslink-induced mutations have not yet been reported [23, 27]. These include two datasets denoted as LIN28A and LIN28-V5 from Wilbert et al., 2012, which identified mRNA targets of LIN28A in human H9 embryonic stem cells and HEK 293 cells, respectively, as well as two datasets from Stefani et al., 2015, referred to as WormSmall (CLIPseq1) and WormBig (CLIPseq2) which identified mRNA targets of LIN28 in *C. elegans*. Importantly, these four datasets are derived from CLIP-seq, rather than PAR-CLIP, and therefore do not incorporate known mutation biases that occur as a result of employing photoreactive nucleosides (i.e. T to C transitions in 4-thiouridine

and G to A transitions in 6-thioguanosine experiments) (Graf et al., 2013; Hafner et al., 2013).

We observed substantial variability in crosslink-induced mutation enrichment between these four datasets (Figure 3.1B). While the two human LIN28A CLIP datasets showed enrichments for substitutions at guanine residues, no nucleotide dominated deletion mutations. Additionally, our analysis of the *C. elegans* dataset showed that uridine nucleotides were the most frequent points of mutation. Though the enrichment of substitutions occurring at uridines was slight, making up 31% and 32% of mutations for the two datasets, we noted a significant enrichment of uridines as the sites of deletions, which represent 56% and 57% of that type of mutation.

Recombinant model LIN28A and pre-let-7f complexes can be crosslinked *in vitro*.

To evaluate the *in vitro* protein-RNA crosslinking profile of a representative bipartite LIN28A-RNA complex, we generated complexes comprised of a previously reported, truncated mouse LIN28 construct (LIN28A-ΔΔ) and a correspondingly modified pre-element let-7f miRNA substrate, (preE_M-let-7f) (Figure 3.2A, B) (Nam et al., 2011). LIN28A-ΔΔ consists of amino acids D33-K187 of the full-length protein and lacks the random coil N- and C- termini as well as a 9 amino acid internal flexible linker between the CSD and zinc-knuckle domain (ZKD) (Figure 3.2A). PreE_M-let-7f has a 5-nucleotide deletion between the AYYHY (the CSD-binding pyrimidine-rich sequence motif, where Y = C or U and H = A, C, or U) (Hafner et al., 2013;

Nam et al., 2011) and GGAG elements to accommodate the decreased space between the LIN28A binding domains (Figure 3.2B). These truncated components were previously crystallized as a complex, which reflected the interactions between the wild type full-length LIN28A and preE-let-7f, as determined by functional studies (Nam et al., 2011). The binding affinity between LIN28A-ΔΔ and preE_M-let-7f, was comparable to the full-length LIN28A affinity for its corresponding preE-let-7f (47 – 190 nM) (Figure 3.2C), in agreement with previous data (Nam et al., 2011). We also observed that purified LIN28A-ΔΔ:preE_M-let-7f complexes could be crosslinked with comparable efficiency as full-length LIN28A:preE-let-7f (Figure 3.2D). This finding suggests that *in vitro* complexes that incorporate these truncated components are sufficient to mimic *in vivo* binding of the native protein with this miRNA intermediate.

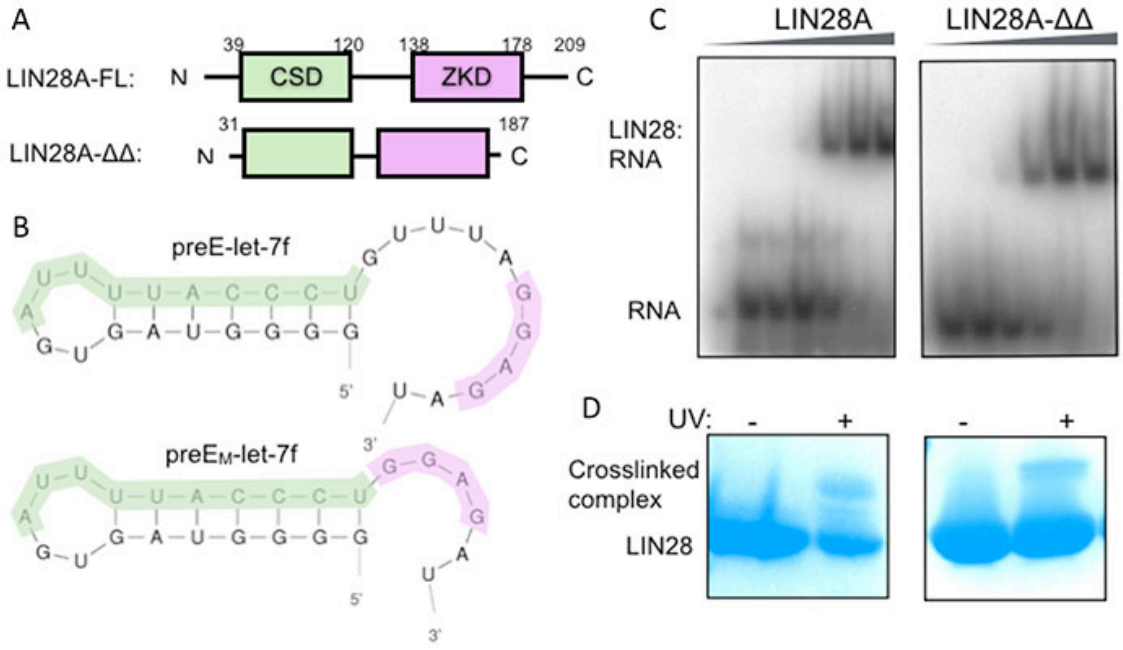


Figure 3.2. LIN28A constructs have high affinity for and crosslink to preE-let-7f targets *in vitro.* (A) LIN28A constructs utilized in biochemistry and UV-crosslinking experiments. LIN28A-FL is full-length and LIN28A-ΔΔ is a truncated version that shortens the flexible linker between the CSD (cold-shock domain) and ZKD (zinc knuckle domain) in addition to shortening the two random coil termini. (B) LIN28A-FL and LIN28A-ΔΔ bind preE-let-7f and $preE_M$-let-7f, respectively. $PreE_M$-let-7f is shortened to accommodate the reduced interdomain linker in LIN28A-ΔΔ. The CSD and ZKD recognize AYYHY (highlighted green) and GGAG sequences (highlighted purple), respectively. (C) Gel shift binding assays with radiolabeled $preE_M$-let-7f probe, mixed with increasing concentrations of LIN28-FL (0, 22, 44, 180, 700 nM, 2.8 μM) and LIN28A-ΔΔ (0, 24, 47, 190, 750 nM, 3 μM). (D) Corresponding SDS-PAGE gels show crosslinked complex bands following UV irradiation.

Mass spectrometry reveals a crosslinking site between the let-7 pre-element terminal loop and LIN28A cold-shock domain.

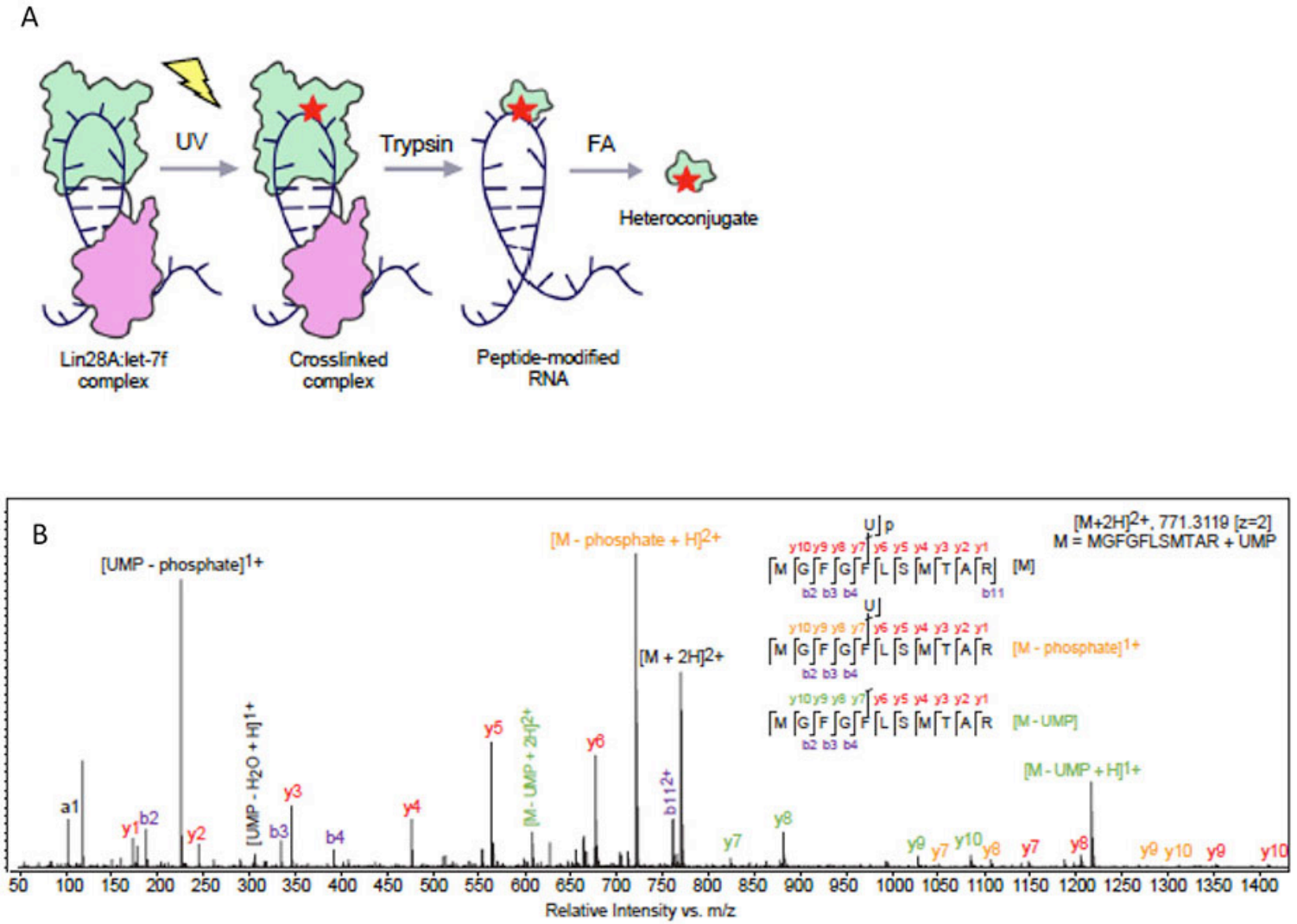
We sought to assign discrete crosslink sites in model recombinant RNA-LIN28 complexes by direct isolation and mass analysis. Complexes of LIN28A-ΔΔ and $preE_M$-let-7f were exposed to 254 nm UV light, and crosslinked complexes were isolated and trypsinized to yield samples of peptide-modified RNA for liquid chromatography-mass spectrometry (LC-MS) analysis (Figure 3.3A). The full-length, unmodified $preE_M$-let-7f RNA has an exact calculated mass of 8052.0600 Da, and crosslinked heteroconjugates were initially sought by scrutinizing species with neutral mass gains that might correspond to tryptic peptide addition. Initially, we observed a discrete species with a low resolution mass of 9268 Da, consistent with a mass gain corresponding to the predicted tryptic fragment MGFGFLSMTAR (residues 51 – 61 in full length LIN28).

To examine this specific peptide-RNA heteroconjugate, trypsinized peptide-RNA samples were subjected to RNA hydrolysis in 50% (v/v) formic acid (FA) at 80 °C for 2 h. Survey spectra were searched for peptide-nucleotide heteroconjugates as either simple mass neutral conjugates or those exhibiting neutral loss of H_2O. A candidate peptide-nucleotide species was observed with an accurate mass of 1540.6120 Da, which is consistent with the heteroconjugate MGFGFLSMTAR-uridine monophosphate (UMP, calculated exact mass 1540.6092 Da, 2.5 ppm mass error). The peptide sequence of the observed heteroconjugate species was confirmed by selecting the $[M + 2H]^{2+}$ ion 771.31 m/z for fragmentation using collision-induced dissociation (CID, Figure 3.3B). The fragmentation spectrum is

complicated by multiple fragmentation pathways, where backbone fragment ions arising from three parent species exist in the tandem MS scans: the selected parent ion (Figure 3.3B, red labels), the ion resulting from loss of phosphate (Figure 3.3B, orange labels), and the ion resulting from complete loss of UMP (Figure 3.3B, green labels). These three species were secondarily fragmented in a manner that generates sequenceable peptide backbone y- and b-ions, giving the expected peptide sequence. Significantly, the y-ions corresponding to UMP and uridine modifications both appear for y7 - y10 but not for smaller products, indicating that the UMP is crosslinked to Phe55 in LIN28A-ΔΔ. To assign the crosslinked position within the pre-element RNA, we sought to observe crosslinked heteroconjugates composed of peptides covalently bound to larger nucleotide species: either di- or tri-nucleotides. Crosslinked heteroconjugates were generated and hydrolyzed as before, with the exception that the acid digest time was reduced (30 min) and temperature lowered (60 °C) to prevent complete RNA hydrolysis. From these samples we were able to identify ions corresponding to the same peptide, MGFGFLSMTAR, neutrally conjugated to UMP, as well as to RNA fragments with the following nucleotide compositions: UU, AU, UUU, AUU and GAU (Table 3.I.). We further confirmed the nucleotide composition and peptide sequence of these species using tandem MS (Appendix Figure D.S1).

Figure 3.3. UV-crosslinking occurs between Phe55 of the LIN28A CSD and Uridine-11 of the preE_M-let-7f terminal loop. (A) Sample preparation workflow for crosslinked peptide-RNA heteroconjugates for LC-MS analysis. (B) Targeted tandem mass spectra identifying product ions that confirm the MGFGFLSMTAR peptide and locate the crosslinked residue as being Phe55; a uridine was identified as the counterpart nucleotide.

Figure 3.3. (Continued)



Analyzing the composition of these species revealed only one consistent overlapping site in the pre-element RNA sequence: uridine-11 (Table 3.I). We concurrently confirmed the identity of U11 using a newly developed RNA site-specific stable isotope labeling technique (Lelyveld et al., 2015). Using pre-element RNA labeled with synthetic isotopes at either the U11 or U12 positions, we observed a mass-shifted isotope distribution exclusively for heteroconjugates arising from complexes formed from RNA labeled at U11. The identified tryptic peptide corresponds to a region within the LIN28A CSD at its binding interface with preE_M-let-7f (Figure 3.4A, B), as observed in a high-resolution crystal structure (PDB ID: 3TS0) (Nam et al., 2011). Within this interface, Phe55 is oriented such that the side chain is (at closest proximity) within 3.5 Å of the uracil moiety of U11 within the pre-element terminal loop, giving a planar angle of ~7.5 degrees between the two aromatic rings. This contact is consistent with a strong π-π interaction between the two residues and suggests that the mass neutral crosslink identified here is physiological.

Table 3.I. Masses and nucleotide composition overlap of crosslinked mono-, di- and tri-nucleotides that identify U11 as the most probable crosslinking counterpart to Phe55.

Species	Mass (Da)	RNA
M + $preE_M$-let-7f	9268	GGGGUAGUGAU$_{11}$UUUACCCUGGAGAU
M + GAU	2214.7149	GAU
M + AUU	2175.6949	AUU
M + UUU	2152.6627	UUU
M + AU	1869.6682	AU
M + UU	1846.6384	UU
M + UMP	1540.612	U

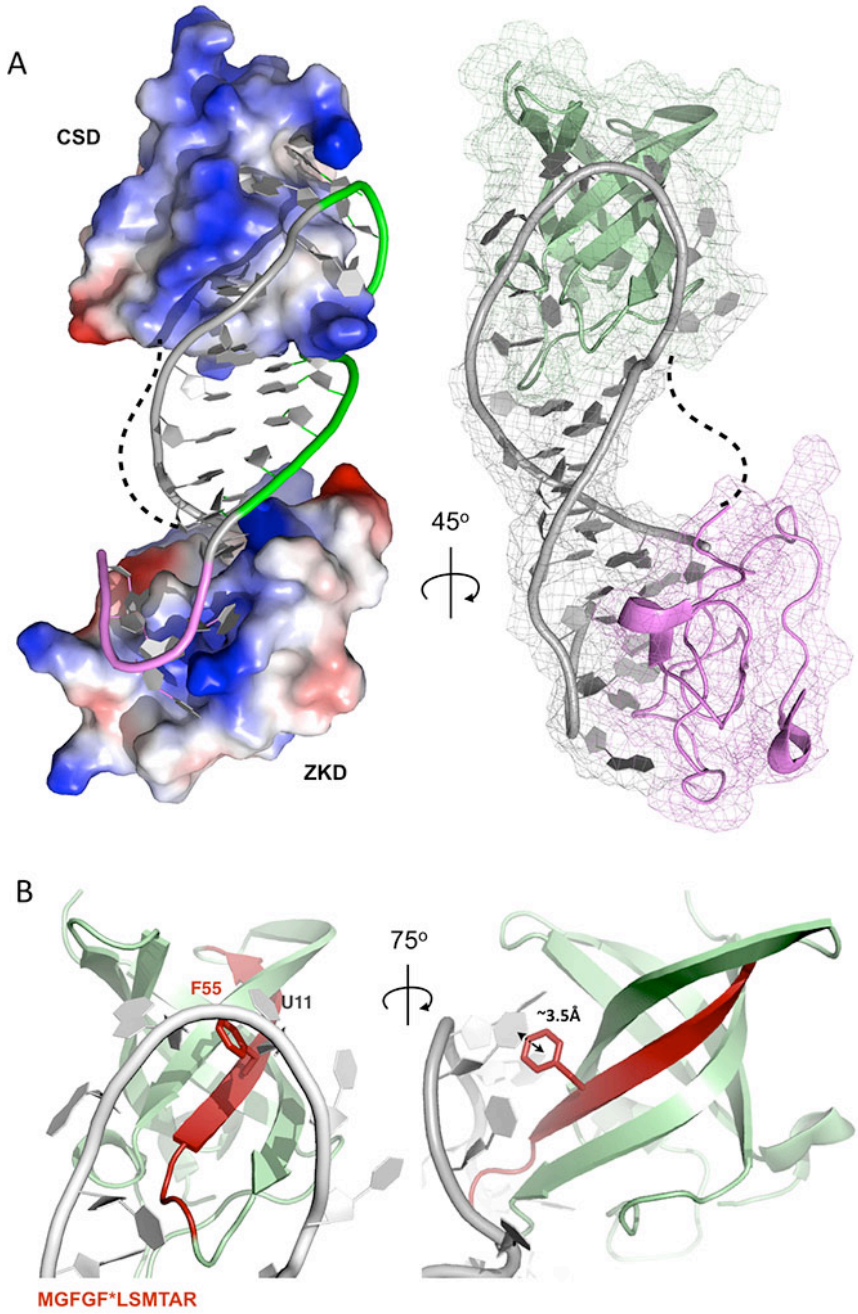


Figure 3.4. UV crosslinked interaction mapped to LIN28A:let-7 crystal structure. (A) Electrostatic and cartoon representations of the known structure of LIN28A-ΔΔ complexed with preE_M-let-7f (PDB ID: 3TS0). In the electrostatic representation, the AYYHY motifs and GGAG motif are highlighted in green and purple, respectively. In the cartoon representation, the CSD and the ZKD are highlighted in green and purple, respectively. The dotted line in each complex represents the unstructured, flexible linker between the two binding domains. (B) Front and side views of the LIN28A-ΔΔ CSD (green) in complex with the preE_M-let-7f terminal loop (grey). Phe55, the identified crosslink site, contacts U11 of the preE_M-let-7f terminal loop. The tryptic peptide MGFGFLSMTAR and Phe55 side chain are highlighted in red

Mutational analyses suggest that *in vitro* crosslinking may occur at sites undetectable by MS.

To validate the experimentally determined crosslinking site and characterize its contribution to substrate binding, we measured *in vitro* binding affinities of single and double mutants at the experimentally determined crosslinking site, Phe55, and an adjacent contact point residue, Trp46 within the LIN28A-ΔΔ construct (Figure 3.5A). Using gel shift binding assays, we found that a single conservative mutation of Phe55 to tyrosine (F55Y) had a minimal effect on binding (K_D: 100 – 200 nM), whereas an alanine mutation (F55A) at the same position resulted in a significant decrease in affinity (K_D: ~700 nM) (Figure 3.5B). Predictably, a double mutation further decreased affinity for preE_M-let-7f. The F55Y/W46A mutant had an estimated K_D of ~1.1 μM and binding affinity for preE_M-let-7f was comparatively abolished in the F55A/W46A mutant (K_D: >38 μM) (Figure 3.5B). These results demonstrate significant contribution of Phe55 and the nearby Trp46 to LIN28-RNA binding affinity, even though their contributions to binding specificity are small (Nam et al., 2011). Despite the range of observed affinities, SDS-PAGE experiments confirmed that all mutant constructs were able to crosslink preE_M-let-7f, though to varying extents (Figure 3.5C).

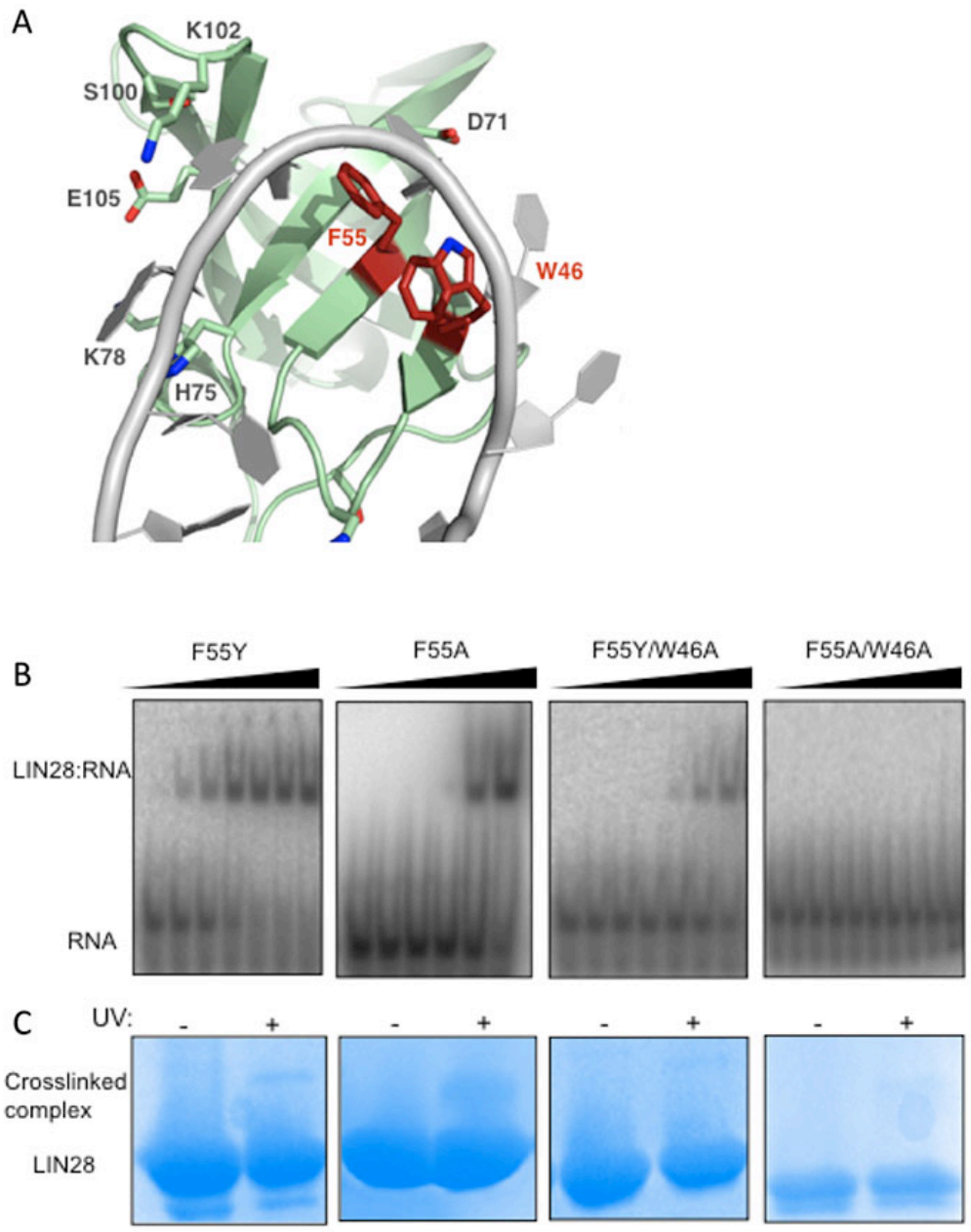


Figure 3.5. LIN28A-ΔΔ CSD mutations at the LIN28A-ΔΔ:RNA interface affect binding affinity but cannot completely abolish crosslinking. (A) Front view of the LIN28A-ΔΔ CSD, highlighting in red the identified crosslinking site, Phe55, and the additional site of mutation, Trp46. Other side chains of the CSD at the binding interface are labeled. (B) Gel shift binding assays with radiolabeled preE_M-let-7f probe, mixed with increasing concentrations of LIN28A-ΔΔ mutant constructs: F55Y (0, 100, 200, 400, 800 nM, 1.6, 3.2 µM), F55A (0, 22, 44, 180, 700 nM, 2.8 µM), F55Y/W46A (0, 70, 140, 280, 560 nM, 1.1, 2.2 µM) and F55A/W46A (75, 150, 300, 600 nM, 1.2, 2.4, 4.8, 9.6, 19, 39 µM). (C) Corresponding SDS-PAGE gels show crosslinked complex bands following UV irradiation.

Discussion

CLIP methods have identified thousands of putative mRNA targets of LIN28, with a significant number of binding sites in both translated and untranslated regions of the transcriptome. While LIN28 unambiguously inhibits processing of the let-7 precursors it targets, reports indicate that LIN28 can both enhance and suppress translation of subsets of RNAs with numerous cellular consequences (Cho et al., 2012; Polesskaya et al., 2007; Qiu et al., 2010; Wilbert et al., 2012; Xu et al., 2009), prompting questions about the specificity and pervasiveness of interactions, as well as concerns over reproducibility and experimental variation. The nature of LIN28 recognition of RNA likely further complicates target identification efforts as structural and biochemical studies have revealed that LIN28 recognizes let-7 precursors through a bi-partite interaction mediated by two binding domains with distinct recognition characteristics (Mayr et al., 2012; Nam et al., 2011; Piskounova et al., 2008).

Our bioinformatics analysis of several publicly available LIN28 CLIP datasets revealed major inconsistencies in identified transcript binding sites, and slight variation in the specific nucleotide crosslink sites. In a bulk comparison, we found only 0.6% - 1.9% of each LIN28A dataset completely overlapped, whereas 38% of each of the two LIN28B datasets overlapped. Our CIMS analysis of selected human and mouse LIN28 datasets confirmed the previously reported enrichment of mutations at guanine residues, although crosslinking-induced mutations were enriched at uridines in the *C.elegans* datasets. Whether the observed differences in CIMS identities result from

experimental variations or species differences is unclear, since the exact mechanism of mutation following UV crosslinking remains uncharacterized.

To identify the precise crosslink site in a model recombinant LIN28A protein-RNA complex, we used LC-MS to analyze covalent RNA-peptide heteroconjugates containing a series of short overlapping RNA sequence fragments. We have separately confirmed this assignment using a novel, single-nucleotide isotope labeling technique (Lelyveld et al., 2015). Out of 11 ssRNA bases that maintain the LIN28A-ΔΔ:preE_M-let-7f complex, our MS analysis identified only a single crosslink site. The covalent crosslink between Phe55 of the LIN28A CSD and a terminal loop uridine of preE_M-let-7f, corresponds to a previously observed tight interfacial contact seen in the crystal structure of the complex (Nam et al., 2011). Aside from the single crosslink of uridine to Phe55, we were unable to identify other sites of modification, including between the ZKD and the GGAG recognition element, which is considered to be specific and critical for high affinity binding to miRNA targets (Desjardins et al., 2012; Loughlin et al., 2012; Mayr et al., 2012; Nam et al., 2011). Interestingly, mutation of Phe55 did not completely inhibit crosslinking of the *in vitro* complexes suggesting the presence other crosslink sites that have not yet been detected by MS.

Consistent with our observations, a recent report (see Lelyveld et al., 2015) demonstrated that in a systematic MS/MS analysis of 124 RBPs crosslinked *in vivo*, MS-detectable crosslinking nearly exclusively occurred at uridines. In one dataset, 89% of crosslinking events are confirmed at uridines and at least one uridine is present in crosslinked di- and tri-nucleotides in the remaining 11% of cases, although LIN28 was not included in that report.

Furthermore, they noted that protein-RNA crosslinks are typically few in number (~1-3 crosslinks per complex) and are frequently mediated by phenylalanine residues (24% of identified amino acid crosslink sites). On the other hand, the only previous report to feature CIMS profiles of LIN28 CLIP datasets (see Cho et al., 2012) identified primarily guanine points of mutation and, specifically, guanines within the critical GGAG recognition element of the let-7 miRNA family. Consistent with their global CIMS analysis, the crosslinking site mutation profile determined for pre-let-7f, (the target of our MS analyses, though ours is truncated), showed crosslink induced mutations primarily at guanines within the GGAG motif and no mutations at the uridine position that we identified, or indeed at any other uridine within the terminal loop.

Collectively, the results of our case study of existing LIN28 binding site data and our biophysical investigation of LIN28A crosslinking to a let-7 precursor suggest that the two highest resolution crosslink site detection methods may benefit from orthogonal, complementary analyses to gain complete crosslink site identification data. Given that precise identification of protein-RNA interactions on a global scale provides an invaluable tool for mechanistic and functional studies of various cellular processes, further development of comprehensive analysis methods is needed to identify, reproducibly and with high confidence, the precise binding sites between diverse RBPs and their RNA targets. One potential opportunity might be found in the advancement of CLIP methods that examine crosslinking between individual domains of proteins and their RNA targets, such as iDo-PAR-CLIP (individual domain PAR-CLIP), with the addition of a higher precision MS

analysis step. Furthermore, unambiguous and comprehensive structural data generated from key techniques such as X-ray crystallography and nuclear magnetic resonance (NMR) may be used to further scrutinize or validate CLIP determined binding sites.

Material And Methods

Comparison of global LIN28 targets

Human LIN28 CLIP datasets were obtained from the Gene Expression Omnibus (GEO): GSE44615 (Hafner et al., 2013), GSM980594 and GSM980593 (Wilbert et al., 2012) and GSM1140829 (Graf et al., 2013). Poor quality reads were removed and adapters were trimmed using the *cutadapt* tool from *FASTQC* with quality test (-q 20). Fastq files were mapped to the hg19 genome assembly (genome.ucsc.edu) using *bowtie2* with options –p 18 and -N 1 allowing for one mismatch. *Picard* software was used to collapse PCR duplicates. Detailed information concerning statistics of filtered reads and mappings are available in Appendix Table D.SI. Enriched binding sites were determined using *Piranha* software with bin size set to 200 nt genomic intervals and sequences of identified binding sites are listed in Appendix Dataset D.S1. Overlapping binding sites were determined using a proprietary Python script and sequences of determined overlapping sites are listed in Appendix Dataset D.S2. Total overlaps are summarized in Appendix Table D.SIII. The Piranha processing summary is listed in Appendix Table D.SII.

Workflow for CIMS analysis

CLIP datasets for mouse (Cho et al., 2012) and human (Wilbert et al., 2012) LIN28 were obtained through Gene Expression Omnibus (GEO), with accessions SRR458758 (monoclonal 353L33G), SRR458760 (polyclonal), SRR531464 (LIN28A) and SRR531465 (LIN28V5). The WormBig and WormSmall datasets from Stefani et al., 2015 were transferred via personal communication. The SRA files were downloaded and converted to FASTQ files using SRA Toolkit 2.5.2. FastQC determined the sequence quality and presence of adapters. Bowtie2 indexes were created for hg19, mm10 and ce10 for human (*H. sapiens*), mouse (*M. musculus*), and worm (*C. elegans*), respectively. Bowtie2 was run with the parameters -p 18 -N 1. Samtools was used to sort and convert mappings into SAM files. The SAM files were then read into R and the following filters applied: low quality reads were removed (if for any nucleotide the quality is below D then the whole read is removed) and all exact sequence duplicates were collapsed to eliminate the potential for PCR duplicates. For each mutation type (substitutions/deletions), distributions were derived separately. For substitutions, only reads with a single substitution were considered. Substitution positions were identified based on the SAM files, fields for tag MD type Z. For deletions, only reads with a single deletion were considered. The positions of deletions were identified based on CIGAR string based on the SAM files. Stringent conditions for the number of substitutions and deletions (1 per read) lead to filtering out reads with adapter sequences. Further analyses of obtained distributions were strand specific, were performed with dplyr package, and were visualized with ggplot2 package for R.

Expression and purification of LIN28A proteins

All recombinant LIN28A protein constructs were overexpressed from pET21a or pETDuet expression vectors. BL21 Rosetta cell colonies transformed with construct plasmids were used to inoculate 100ml LB starter cultures and incubated overnight (~18-20 hrs) in a shaker incubator at 37 ^{o}C. The next day, 10 ml of starter culture was used as inoculation for every 1 L of expression culture (typically 2-4 L). Expression cultures were incubated at 37 ^{o}C until reaching O.D. of 0.6-0.8, at which point cultures were induced with 0.5 mM IPTG. Following induction, cultures were incubated overnight at 18 ^{o}C with shaking and harvested the next day via centrifugation. Proteins were purified via Ni^{2+} affinity, cationic exchange and size exclusion chromatography, as previously described (Nam et al., 2011).

Electrophoretic mobility shift assays (EMSAs) and SDS-PAGE

For binding analyses, LIN28A was serially diluted into a low-salt binding buffer (20 mM Bis-Tris, pH 7, 100 mM NaCl, 50 μM $ZnCl_2$, 5% glycerol and 5 mM DTT) supplemented with yeast tRNA (to a final concentration of 1 mg/ml) and the RNAse inhibitor Ribolock (Thermofisher Scientific, CAT: EO0381). Precursor let-7 RNA probes were synthesized from IDT and labeled with ^{32}P via the T4 Polynucleotide Kinase (New England BioLabs, CAT: M0201S). ^{32}P-labeled RNA substrates (<1 nM) were incubated with LIN28A protein dilutions of increasing concentration for ~30 minutes at RT before

samples were run on a 10% native gel. Gels were vacuum dried and pressed to radiolabel sensitive film overnight. Films were imaged the following day using scanning phosphorimager. For SDS-PAGE analyses, complexes of LIN28A proteins with pre-let-7f RNAs were exposed to UV irradiation and pre- and post-crosslinked samples were compared on SDS-PAGE gels and stained with Coomassie Blue.

UV crosslinking and hydrolysis of LIN28 and let-7 complexes

Purified complexes of LIN28 proteins and preE-let-7f targets were buffer exchanged into crosslinking buffer (20 mM Bis-Tris, pH 7, 100 mM NaCl, 50 μM $ZnCl_2$, and 1 mM DTT), and UV crosslinking was completed by irradiating samples three times at 300 mJ/cm^2 at 254 nm in a Stratalinker 1800. Crosslinked complexes were separated from non-crosslinked complexes via denaturing urea gel. Gels were stained with ethidium bromide and imaged. Crosslinked complex RNA bands (indicated by the increase in size over the free RNA bands) were excised and RNA was eluted from gel via electro-elution into a dialysis bag, 3 x 30 mins at 100V. Eluted, crosslinked complexes were concentrated and buffer exchanged into 8 M urea, 50 mM bis-Tris, pH 7. Trypsin/Lys-C (Promega, CAT: V5072) was added to a final ratio of 1:25 (w/w). Digestions were incubated at 37 ^{o}C for 3 hrs with shaking. Samples were diluted to lower the urea concentration to <1 M and digestions continued overnight at 37 ^{o}C.

Modified RNA was purified from digested samples via anion exchange chromatography. Specifically, samples were loaded onto a DEAE

column (GE Healthcare) with low-salt buffer (20 mM Bis-Tris, pH 6, 10% glycerol and 5 mM DTT) and eluted with a high-salt buffer fast gradient (20 mM Bis-Tris, pH 6, 10% glycerol and 5 mM DTT, 2 M NaCl). Chromatography fractions containing RNA were pooled, flash frozen, and lyophilized. Dried samples were resuspended in LC-MS grade water and quantified by their optical density at 260 nm.

LC-MS

All samples were separated on an Agilent 1200 HPLC coupled to a solvent degasser, auto sampler, diode array detector, and column oven (Agilent Technologies). Separations were performed using two solvent systems. Native and tryptic peptide-modified, full-length RNA samples were separated on a 100 mm x 1 mm i.d. Xbridge C18 column with a particle size of 3.5 µm (Waters). The solvent system was based on a previously published reverse phase ion pairing LC-MS method, and we used 200 mM HFIP with 1.25 mM trimethylamine at pH 7.0 in buffer A and methanol in buffer B (Apffel et al., 1997). The column was heated to 60.0 °C and the flow rate was 100 µL/min. Injection volumes were 10 - 25 µL. Mobile phase B was increased from 5% to 15% from 0 - 20 minutes and then from 15% to 60% over an additional 20 minutes of run time. Absorbance was monitored at 260 nm with a reference wavelength at 380 nm and a 2 s response time. Tryptic, crosslinked peptides digested in formic acid were separated on a 150 mm x 1.0 mm i.d., micro bore rapid resolution SB-C18 column with a particle size of 3.5 um (Agilent Technologies). Buffer A was water with 0.1% FA and buffer B was acetonitrile with 0.1% FA. The column was heated to 40.0 °C and the flow rate was 150 µL/min. Injection volumes were 5 - 20 µL. Buffer B was

increased from 5% to 45% over 0 - 20 minutes.

All samples were analyzed on an Agilent G6520A accurate-mass QTOF coupled to the LC system described above, operating in extended dynamic range mode. The system was calibrated on the same day of analysis and reference masses were continuously infused for online mass correction. Full length RNA crosslinked to tryptic peptides (resultant of tryptic digest of crosslinked, full-length LIN28) were separated using the HFIP/TEA solvent system and analyzed in negative ion mode from 239 - 3200 m/z with a scan rate of 1 spectrum/s using the following settings: drying gas flow, 8 L/min; drying gas temperature, 325 °C; nebulizer pressure, 30 psig; capillary voltage, 3500 V; fragmentor, 200 V; and skimmer, 65 V. Tryptic, crosslinked peptides digested in formic acid were separated using the water + 0.1% FA and acetonitrile + 0.1% FA system and analyzed in positive ion mode from 104 - 3000 m/z with a scan rate of 1 spectrum/s in MS1 mode and 1.67 spectrum/s for MS acquisition and 1.2 spectrum/s for MS/MS acquisition in targeted MS/MS mode using the following settings: drying gas flow, 8 L/min: drying gas temperature, 325 °C; nebulizer pressure, 35 psig; capillary voltage, 4500 V; fragmentor, 175 V; and skimmer: 65 V. In targeted MS/MS mode, ions were fragmented using collision induced dissociation with nitrogen gas, collision potentials of 10 – 30 V and an isolation width of ~4 amu.

LC-MS data analysis and crosslink site identification

Tryptic peptides crosslinked to mono- or dinucleotides were first identified based on MS1 data using a database composed of predicted tryptic peptides with up to two missed cleavages crosslinked to all possible mono-

and dinucleotides using Agilent's Find by Formula algorithm in the MassHunter software package. The analysis allowed for the loss of water so we could identify both mass neutral crosslinks and those leading to loss of water. The peptides identified in survey spectra with covalent UMP, di-, and trinucleotide modifications were then analyzed in targeted MS/MS experiments to validate the peptide sequence and characterize the species. The targeted MS/MS spectra for the heteroconjugate MGFGFLSMTAR-UMP were manually compared to theoretical product ion spectra in which the UMP was attached to different amino acids in the peptide as to determine the position of the crosslink.

Author Contributions

E.M.R, A.B and V.S.L. conceptualized and designed the project. E.M.R cloned, overexpressed and purified LIN28 proteins, completed binding assays and generated LIN28-let-7 crosslinked complexes and digested heteroconjugates. A.B. and V.S.L. completed LC-MS experiments and analysis. P.B. completed bioinformatics analysis of CIMS sites. E.M.R, V.S.L, J.W.S and P.S. wrote the manuscript.

Acknowledgments

We would like to express our gratitude to Areum Han and Kristina Holton for preliminary analyses and troubleshooting of the bioinformatics.

Funding

This work was supported by the National Cancer Institute [R01CA163647 to P.S.]; the National Science Foundation Graduate Research

Fellowship Program [DGE1144152 to E.R.]; the UNCF-Merck Graduate Research Science Initiative [E.R.]; and from the Academy of Finland [A.B.]. Funding for open access charge: National Institutes of Health. J.W.S. is an Investigator of the Howard Hughes Medical Institute.

Chapter 4

Pinpointing RNA-Protein Cross-Links with Site-Specific Stable Isotope-Labeled Oligonucleotides

Contributors: Victor S. Lelyveld, Anders Björkbom, Elizabeth M. Ransey, Piotr Sliz and Jack W. Szostak

Supplemental Materials for this chapter are in Appendix E.

This Chapter originally appeared in *The Journal of the American Chemical Society*, Vol. 137 (2015). http://pubs.acs.org/doi/abs/10.1021/jacs.5b10596

Background

High affinity RNA-protein interactions are critical to cellular function, but directly identifying the determinants of binding within these complexes is often difficult. Here, we introduce a stable isotope mass labeling technique to assign specific interacting nucleotides in an oligonucleotide-protein complex by photo-cross-linking. The method relies on generating site-specific oxygen-18-labeled phosphodiester linkages in oligonucleotides, such that covalent peptide-oligonucleotide cross-link sites arising from ultraviolet irradiation can be assigned to specific sequence positions in both RNA and protein simultaneously by mass spectrometry. Using Lin28A and a let-7 pre-element RNA, we demonstrate that mass labeling permits unambiguous identification of the cross-linked sequence positions in the RNA-protein complex.

Results and Discussion

Proteins recognize structured RNAs to form regulatory and catalytic complexes. UV and chemical cross-linking have been widely used to interrogate these biochemical interactions under native solution-phase conditions, but it is still challenging to pinpoint the specific interacting residues within the linear RNA and protein sequences (Kramer et al., 2014). Photo-cross-linking under ultra-violet illumination followed by purification and sequencing (Hafner et al., 2010; Ule et al., 2003) or mass analysis (Kramer et al., 2014) may yield significant structural insights, but neither method is, by itself, sufficient to unambiguously assign cross-link sites to specific side chains on both proteins and nucleic acids

simultaneously. The incorporation of artificial photoreactive bases (e.g., 4-thiouridine or 6-thioguanosine) is useful to resolve ambiguities and increase the yield of covalent photoproducts (Hafner et al., 2010). It is, however, possible that the resultant cross-links do not precisely represent biologically relevant interactions since the modified nucleotides themselves may alter the interaction of interest.

Here we present an alternative approach based on site-specific stable isotope labeling that can be used for efficient identification of interacting sites by liquid chromatography-mass spectrometry (LC-MS). The technique relies on a facile, inexpensive, and fully automated method to generate individual ^{18}O-labeled phosphodiester linkages during oligonucleotide synthesis. These site-specifically mass labeled oligonucleotides can be prepared in a straightforward manner by standard solid-phase synthesis, such that no cumbersome organic synthesis is necessary to generate specific mass labeled DNA or RNA probes. Site-specifically labeled oligonucleotides can be cross-linked to interacting proteins (Figure 4.1A), and the resulting covalent product can be subjected to hydrolysis and analyzed by MS. Covalent nucleotide-labeled peptides prepared in this manner exhibit a unique isotopic distribution exclusively when an ^{18}O mass label is retained adjacent to the cross-link site.

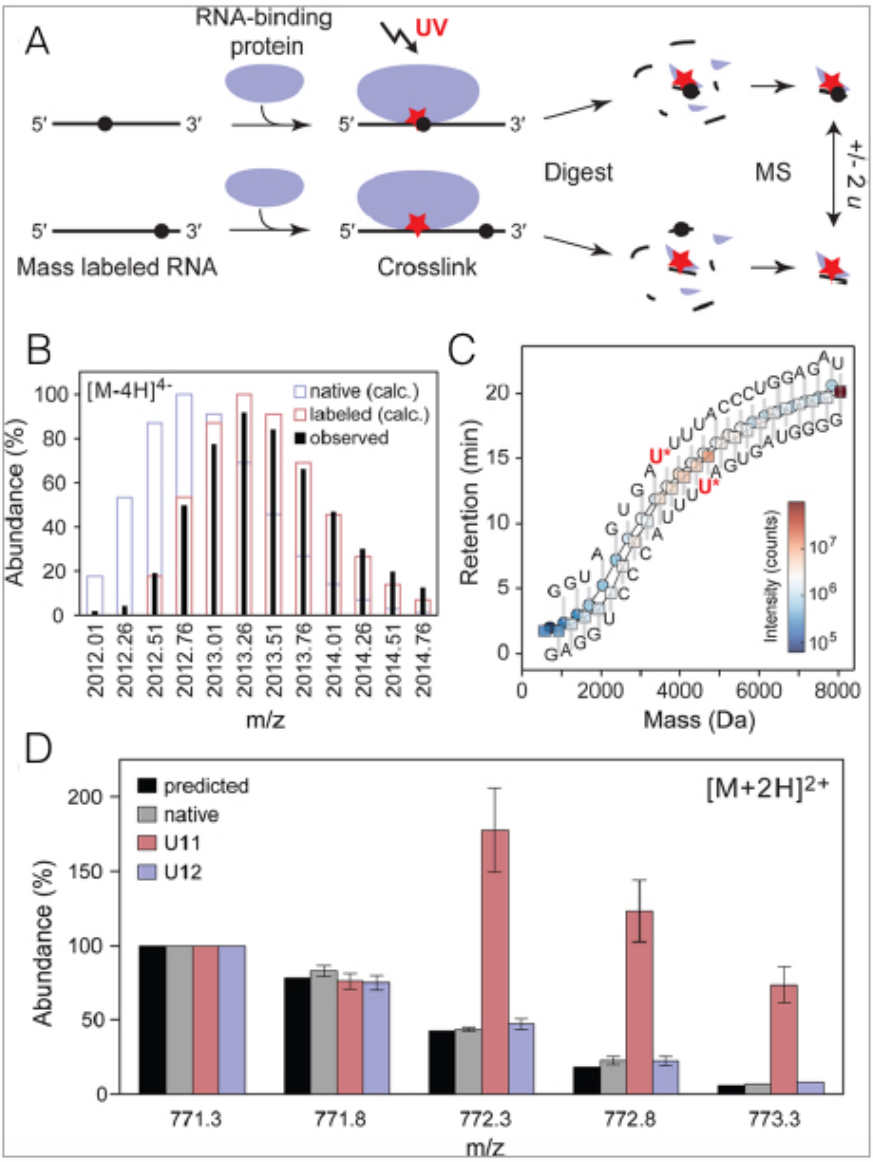
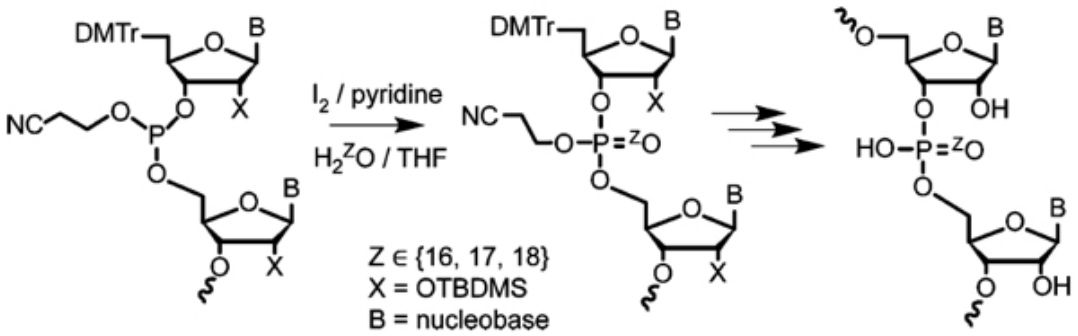


Figure 4.1. Identifying RNA-protein contacts with site-specific RNA mass labeling. (A) Sample preparation workflow. Oligonucleotides are prepared with selectively enriched 18O phosphates at distinct positions (black circle), separately incubated with protein, UV cross-linked, hydrolyzed to peptide-nucleotide fragments, and analyzed by MS. (B) Calculated and observed isotopic distributions for [M − 4H]4 − ion of 25-nt preEM -let-7f RNA (GGGGUAGU-GAU11UUUACCCUGGAGAU) labeled with 18O at the phosphodiest-er following U11. (C) Direct LC-MS sequencing of U11-labeled RNA, with the mass labeled position indicated as U* (red). (D) Isotope distribution of [MGFGFLSMTAR + UMP + 2H]2+, a tryptic peptide ion derived from cross-linking Lin28A in the presence of preEM-let-7f with natural abundance (native) or 18O mass labeled at the 3′phosphodiester following U11 or U12. All spectra are normalized to the abundance of the monoisotopic ion 771.3 m/z.

To demonstrate the utility of site-specific RNA mass labeling, we examined the solution-phase interaction of a precursor fragment of the let-7 miRNA and Lin28A, a high affinity RNA-binding protein that is one of four critical regulators whose overexpression yields an induced pluripotent stem cell (iPS) state (Yu et al., 2007b). Mature let-7 plays a regulatory role in inflammation (Iliopoulos et al., 2009) and has been implicated in tumorigenesis (Yu et al., 2007a). Dicer activity on the let-7 precursor, prelet-7, releases a stem-loop fragment "pre-element," preE-let-7, and the mature let-7 miRNA. Binding of Lin28A to the preE region of prelet-7 inhibits let-7 microRNA processing (Piskounova et al., 2011). Based on structural evidence, the interaction between let-7 precursor RNAs and Lin28A protein is mediated through contacts in both loop and stem regions of preE-let-7 (PDB 3TS0) (Nam et al., 2011). Consistent with crystallographic contacts between a truncated preE-let-7 RNA (dubbed preEM-let-7f) and the cold shock domain of a loop-minimized Lin28A (Lin28ΔΔ), we recently observed (Ransey et al., 2017, in preparation) a native solution-phase cross-link between these domains by tandem MS.

To pinpoint the precise ribonucleotides in the let-7 pre-element that are photo-cross-linked to Lin28A upon UV exposure, we now report the use of stable isotope labeling of a pre-element RNA and high-resolution mass spectrometry. Our recent study implicated a uridine at the cross-linking site, most likely in the loop region of the pre-element hairpin (Ransey et al., 2017, in preparation). We therefore synthesized the previously co-crystallized 25-nt stem-loop preEM-let-7f RNA (Nam et al., 2011) carrying a single ^{18}Omass label on one of two possible

uridine positions in the pre-element hairpin: U11 or U12 in the RNA sequence GGGGUAGUGAU11U12UUACCCUGGAGAU, where the mass label is 3' to the indicated nucleoside in each case. Solid-phase oligonucleotide synthesis by the phosphoramidite method proceeds by iterative cycles of protected nucleoside phosphite addition to a deprotected terminal hydroxyl on the growing oligonucleotide chain, followed by rapid iodine-mediated oxidation to generate an O-protected phosphate triester (Scheme 4.1) (Scaringe et al., 1990; Wincott et al., 1995). The source of oxygen equivalents is H2O, typically applied to the solid support in a solution of I_2, pyridine, and tetrahydrofuran (THF). This oxidation method has been used to generate regiospecifically isotope-labeled DNA dinucleo-tides and short oligonucleotides for NMR studies using isotope-enriched water (Connolly and Eckstein, 1984; Potter et al., 1983; Shah, 1984). It has also recently been shown that the oxidation mix can be completely substituted with one formulated with enriched water ($H_2{}^{18}O$) to generate RNA that is uniformly labeled at every backbone position in the resulting oligonucleotide (Hamasaki et al., 2013). Rather than labeling all positions, we used two alternative oxidation mixes in an automated synthesizer, such that isotopically enriched phosphodiester nonbridging oxygens were incorporated in a highly site-specific manner. Here, the heavy oxidation mix was formulated with 20 mM I2 and 97% enriched $H_2{}^{18}O$ water in the volumetric ratio 2:78:20 $H_2{}^{18}O$:THF:pyridine.



Scheme 4.1. Site-Specific Stable Isotope Labeling by Iodine Oxidation during Solid-Phase Oligonucleotide Synthesis

Mass labeled oligonucleotides maintained the expected +2 Da shift following a typical deprotection protocol (Figure 4.1B). Crude oligonucleotides were purified by the DMT-on method using C18 cartridges, followed by HPLC purification. Oligonucleotides mass labeled at a single position were typically enriched by 90% starting from 97% enriched $H_2{}^{18}O$ (Figure 4.1B).

We confirmed the position of the mass label in the oligonucleotide sequence by direct LC-MS sequencing (Figure 4.1C) by a variation of the method recently described (Bjorkbom et al., 2015). A database of chemical formulas of all possible single-cut hydrolytic fragments with or without ^{18}O enrichment was generated, such that misincorporation of the label would be observable. For the U11-labeled RNA, the label appeared exclusively as a mass difference corresponding to [UMP + 2 Da − H2O] between 11-nt and 12-nt hydrolytic fragments on both 5' and 3' sequence ladders. Fragments with higher mass carried the +2 Da label on both 5' and 3' ladders.

While the labeled nonbridging phosphodiester oxygens are stable under biological conditions (Hamasaki et al., 2013), we noted that phosphodiester hydrolysis of mass labeled RNA can result in label exchange with bulk solvent (Castleberry et al., 2009) when the newly generated nucleotide carries a 3' (2')-monophosphate, which would result in label loss. Acid hydrolysis of RNA generates a terminal 3' (2')-monophosphate product (Oivanen et al., 1998), and we therefore sought to minimize the impact of exchange when generating digests by this method. We compared 50% (v/v) FA:water (~12 N) digestion at 60 and 80

°C over time (Appendix Figure E.S1). The ratio of ^{16}O:^{18}O UMP increased steadily over digest time at 80 °C, indicating significant phosphate oxygen exchange, such that the +2 Da species was nearly at natural abundance levels after 2 h. Under these conditions, the first-order rate constant for exchange was 1.1 h−1. When digested at 60 °C, the ^{18}O mass label was well retained even after 2.5 h of digestion, with a significantly slower exchange rate of 0.28 h−1. Alternatively, enzymatic RNA digestion with nuclease P1, which leaves a 5'-phosphorylated product after cleavage, resulted in negligible loss of the mass label (Appendix Figure E.S3). For the cross-linked RNA-peptide species analyzed in Figures 4.1D, 4.2, and Appendix Figure E.S2, we chose a 2 h RNA digest in 50% (v/v) FA at 60 °C.

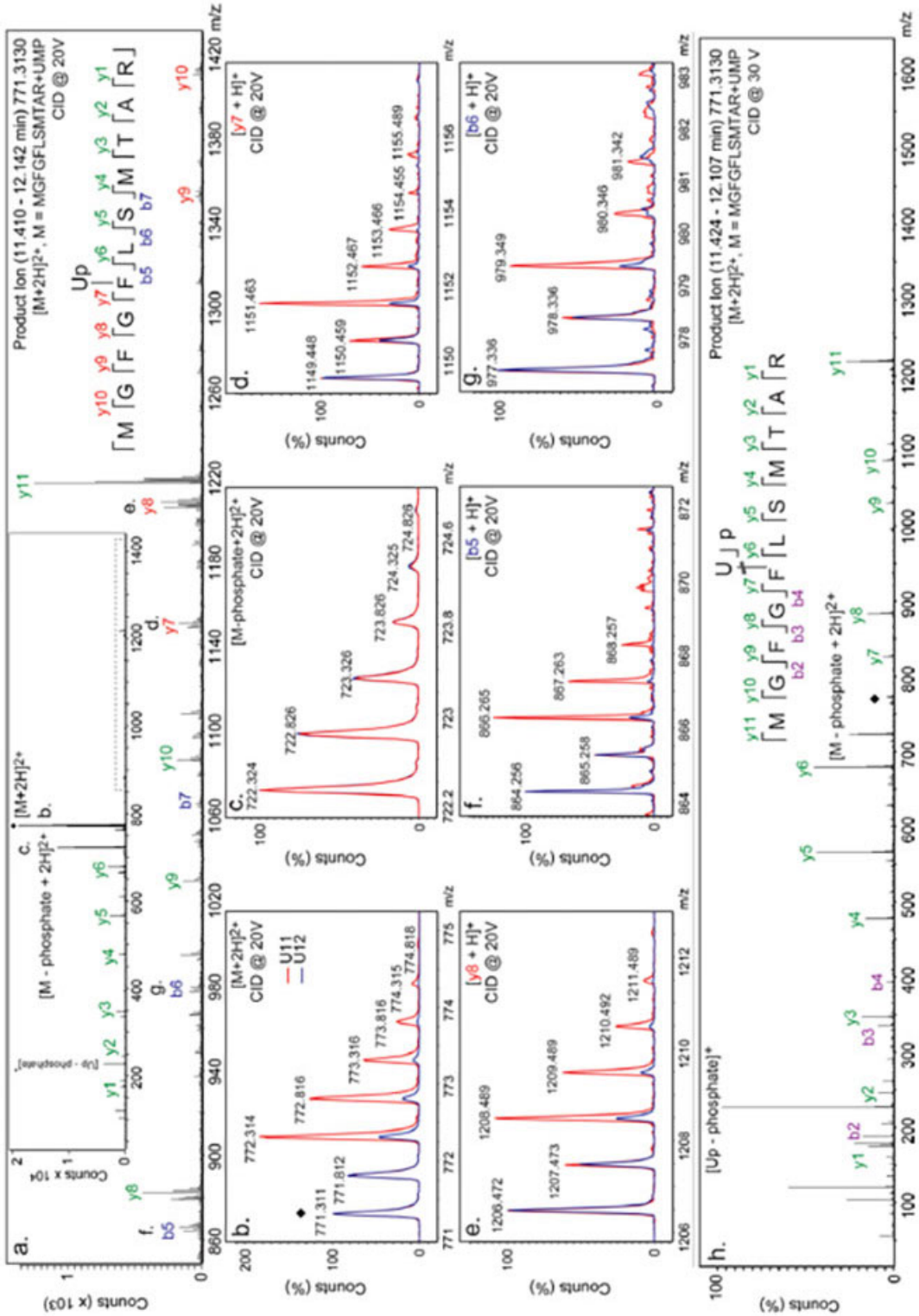
Recombinant Lin28ΔΔ protein was complexed with un-labeled, U11-labeled, or U12-labeled preEM-let-7f RNA, photo-cross-linked with 254 nm UV light, and hydrolyzed with trypsin and FA. Using this procedure, we recently identified a tryptic peptide, MGFGFLSMTAR, cross-linked to uridine mono-phosphate (UMP) by acid digestion and tandem mass spectrometry (Ransey et al., 2017, in preparation). Using mass labeled preE_M-let-7f RNA oligonucleotides, we observed no significant deviation from the expected isotopic distribution for the peptide-UMP cross-linked species arising from unlabeled RNA or RNA mass labeled in the U12 position (Figure 4.1D). However, highly significant ^{18}O enrichment ([M + 2H + 2 Da]2+ = 772.3 m/z) is apparent for cross- linked peptide generated from U11-labeled RNA, thereby assigning the uridine in the cross-linked species to U11 in the preE_M-let-7f sequence.

Using tandem MS/MS with CID, we confirmed the position and identity of the observed peptide-nucleotide species (Figure 4.2). The cross-linked peptide ion corresponding to MGFGFLSMTAR-UMP (771.3 m/z) was isolated for fragmentation with a 4 m/z isolation width (771−775 m/z), such that the isotope distributions for both unlabeled and labeled peptides would be simultaneously monitored (Figure 4.2A, windowed on the high mass range for clarity). At lower CID energy (20 V), the characteristic product ions y7 (Figure 4.2D) and b5 (Figure 4.2F) carrying uridine monophosphate are observed, as well as larger ions (Figure 4.2E,G) consistent with UMP modification at phenylalanine-55 (F55). Higher CID energy confirms the underlying peptide sequence (Figure 4.2h).

Product ions carrying UMP showed a characteristic mass-shifted isotope distribution when carrying the mass label. When the RNA used for complexation was isotope labeled following U12, the resulting fragmentation spectrum for the cross-linked peptide shows a native isotopic distribution. However, when the same complexes were formed from U11 ^{18}O-labeled RNA, an enriched isotope distribution is observed for all peptide fragments carrying UMP. Loss of phosphate in the CID fragmentation spectra of U11-labeled peptide (Figure 4.2C) showed recovery of a natural abundance isotope distribution in the remaining uridine-peptide fragment ion.

Figure 4.2. Simultaneous assignment of cross-linked nucleotide and amino acid sequence position by MS/MS. (A) A subset of product ions (840−1420 m/z) generated by CID at 20 V from the selected ion 771.3130 m/z (z = 2, isolation width 4 m/z), showing a fragmentation pattern consistent with uridine monophosphate (Up) attached by a mass-neutral linkage to phenylalanine in the fifth sequence position of the peptide MGFGFLSMTAR from Lin28A. Green y-ions are those derived from peptide fragmentation following loss of Up, shown fully in panel (h). Inset: Full range of product ions showing region (dotted box) magnified in the main panel. (B-G) Selected product ions arising from samples prepared with 18O labeling after position U11 (red) or U12 (blue) in preEM-let-7f, with apparent +2 Da enrichment derived only from U11 products. (B) Isotope distribution of remaining unfragmented selected ion (black diamond) magnified from the spectra in (A). (C-G) Magnified views of the product ion isotope distributions in panel (A). (H) Full scan of product ions from CID at 30 V with the same selected ion (black diamond), showing peptide sequence fragments for y1−11 and b2−4 arising from Up loss, with [Up − PO4H2]1+ shown as a prominent product ion. Intensities in panels (B-G) are normalized to the first isotope.

Figure 4.2. (Continued)



We also observed the same peptide cross-linked to di- and trinucleotides, as a result of incomplete RNA digestion (Appendix Figure E.S2). From U11-labeled peptide-RNA complexes, we observe the signature mass-shifted isotope distribution for the cross-linked nucleotides U + 2 Da, AU + 2 Da, and GAU + 2 Da, AUU + 2 Da, strongly implying that U11 is the cross-link site. For the U12-labeled species, the peptide cross-linked to U, AU, and GAU all show natural abundance isotope levels, but we also observe some enrichment in the isotope distribution on the cross-linked trinucleotide AUU + 2 Da (Figure 4.2B). For this latter fragment, the enriched cross-linked trinucleotide must compositionally include the U12 mass label, but its absence in smaller cross-linked species strongly implies that U12 is adjacent to the cross-link site but not directly linked in the UV-catalyzed reaction.

Given the known sequence of preE_M-let-7, these data taken together unambiguously assign the UV-induced photo-cross-link position to uridine U11 in the RNA and phenylalanine F55 in Lin28A in this complex. Furthermore, mass labeling also allows for non-native isotopic enrichment to be used as an analytical signature for identification of interesting ions. Peptide preparations for tandem mass analysis are typically highly complex, and UV cross-linking further increases complexity in the absence of additional purification steps. The ability to differentiate cross-linked RNA-peptide species from a complex background could significantly improve precursor ion selection and analysis, possibly alleviating the need for rigorous sample enrichment steps. RNA with partially enriched phosphodiester positions could be used to uniquely identify cross-linked peptide-

nucleotide products by examining high-resolution survey spectra for bimodal isotope distributions exhibiting +2 Da mass shifts. (Back et al., 2002; Castleberry et al., 2009; Meng and Limbach, 2005; Wallis et al., 2001).

Crude maps of the interaction surfaces in RNA-protein complexes could in principle be elucidated by an extension of our method that makes use of an efficient search strategy based on oligonucleotides labeled at multiple phosphate positions. Uridine cross-links are predominantly observed with 254 nm UV irradiation and ESI-LC-MS (Kramer et al., 2014). Therefore, the number of mass labeled positions in a small RNA need only be, in the worst case, a function of the uridine content, and not all adjacent uridines must be labeled since trinucleotide-carrying peptides are often observed. The expected number of uridines is ca. six for single-stranded small RNA of mean base composition and length 20−25. All singly labeled oligonucleotides could be synthesized and analyzed individually in a cross-linking MS experiment, or multiple labels could be used for larger RNAs to search the space with higher efficiency, ruling out candidate positions by deduction.

Because our method relies only on ^{18}O-enriched water, introducing stable site-specific phosphodiester mass labels into synthetic oligonucleotides is significantly less expensive and more generally accessible than other conservative RNA labeling schemes that seek to minimize chemical perturbation, such as those that rely on isotopically enriched ^{13}Cor ^{15}N nucleosides. Photoreactive thiolated nucleoside analogs such as 4-thiouridine or 6-thioguanosine have the potential to affect RNA structure and protein interactions,

since their affinity and specificity within RNA duplexes is measurably different from their native counterparts (Heuberger et al., 2015; Sheng et al., 2014). As such, isotopic phosphodiester labeling is a prudent strategy for identification and confirmation of RNA-protein cross-linking sites under physiological conditions.

Additional Methods

Synthesis of phosphodiester isotope labeled oligonucleotides.

Solid phase oligonucleotide synthesis was performed using the phosphoramidite method and typical procedures on an automated synthesizer (Expedite 8909). Positions were mass labeled using an oxidation step that draws from the AUX line, typically used for phosphorothioate synthesis. Oxidation for unlabeled positions used typical I2/THF/pyridine/water oxidation mix (Glen Research), whereas labeled positions were oxidized using a fresh preparation of "heavy oxidation mix" containing 20 mM I2 and 2:78:20 (v:v:v) H2 ^{18}O:THF:pyridine, starting from 97% enriched H2 18O water (Cambridge Isotopes). The AUX line was thoroughly washed with anhydrous acetonitrile and dried with nitrogen prior to priming with the heavy oxidation mix. This mix was used for oxidation only to generate isotopically enriched nucleotide positions carrying a 3' phosphodiester labeled with ^{18}O, whereas all other positions were oxidized with manufacturer-supplied oxidation mix (Glen Research) that is formulated with natural abundance water and 20 mM I2. RNA phosphoramidites (Chemgenes or Glen Research) were used for synthesis and deprotected based on the manufacturer's guidelines. Briefly, columns were treated in 1:1 (v:v)

aqueous ammonium hydroxide:methylamine (AMA, mixed in equal parts from 30% aqueous ammonium hydroxide, 40% aqueous methylamine) at 65 °C for 10 min, followed by drying under a stream of nitrogen. Desilylation was performed in DMSO with 1:3 HF:TEA for 2.5 h at 65°C and quenched in Tris buffer (Glen Research). Crude oligonucleotides were purified by the DMT-on method using C18 cartridges with on-column detritylation in aqueous 2% TFA, as per the cartridge manufacturer's protocol (Glen Research), followed by HPLC purification.

Oligonucleotide UV cross-linking to LIN28A.

Recombinant protein preparation of the loop minimized Lin28ΔΔ variant of Lin28A and purification from E. coli has been previously described (Nam et al., 2011). Complexes of Lin28ΔΔ and preE-let-7f were crosslinked in 20 mM Bis-Tris, pH 7,100 mM NaCl, 50 μM ZnCl2, and 1 mM DTT by 254 nm irradiation in three pulses of 300 mJ/cm2 in a fluorescent bulb crosslinker (Stratagene 1800). Crosslinked complexes were enriched by denaturing PAGE and electroelution, followed by denaturation in 8 M urea, 50 mM Bis-Tris, pH 7. Protein was digested enzymatically with Trypsin/Lys-C (Promega) added to a final ratio of 1:25 (w/w) at 37 °C for 3 h, followed by dilution to <1 M urea and overnight incubation at 37 °C. Modified RNAs were enriched by anion exchange chromatography on diethylaminoethyl resin in S6 20 mM Bis-Tris, pH 6, 10% glycerol and 5 mM DTT using a linear gradient from 0 – 2 M NaCl. Eluted fractions containing RNA were pooled, flash frozen, and lyophilized. Crosslinked tryptic peptide-RNA conjugates

were prepared for MS analysis by hydrolysis in 50% (v/v) formic acid (FA) at 60 °C for 2 h, flash frozen, and lyophilized to dryness. Shorter digests tend to enrich for dimer and trimer species, and it may often be preferable to use a 30 min digest at 60 °C to observe these species in greater abundance. Dried samples were resuspended in LC-MS grade water for analysis.

LC-MS analysis

Samples were separated and analyzed on an Agilent 1200 HPLC coupled to an Agilent 6520 accurate-mass Q-TOF equipped with a dual ESI source, solvent degasser, auto sampler, diode array detector and column oven. RNA oligonucleotide characterization was performed in negative mode with ion pairing reverse phase chromatography (Bjorkbom et al., 2015). RNA sequencing was performed by digesting 45 pmol of purified oligonucleotides for 5 min at 40 °C in 50% (v/v) FA/water, followed by flash freezing, lyophilization, resuspension in water, and LC-MS analysis. On-line separations were performed with aqueous mobile phase (A) as 200 mM 1,1,1,3,3,3-hexafluoro-2-propanol (HFIP) with 1.25 mM triethylamine (TEA) at pH 7.0 and organic mobile phase (B) as methanol across a 100 mm Å~ 1 mm i.d. Xbridge C18 column with a particle size of 3.5 µm (Waters). Cross-linked nucleotide-peptide conjugates were mass analyzed as recently described3. Briefly, separations were performed using an Agilent ZORBAX SB-C18 column 1.0 mm i.d. Å~ 150 mm length with 3.5 µm particle size with a solvent elution gradient from (A) water with 0.1 % formic acid to (B) acetonitrile with 0.1 % formic acid. The flow rate was 0.1 mL/min, and all

separations were performed with the column maintained at 40 °C. Mass analysis was performed in positive mode, and tandem mass analysis was performed by collision induced dissociation (CID) using nitrogen as the collision gas and a 4 m/z isolation width.

Characterization of phosphodiester mass label exchange during hydrolysis.

Total RNA digests for isotope exchange studies were generated either by enzymatic or acid hydrolysis. To examine 18O isotope exchange when the hydrolytic product carries a 3′ or 2′monophosphate, U11-labeled RNA was hydrolyzed in 50% (v/v) formic acid at either 60 °C or 80 °C in a thermocycler with a heated lid, and samples were taken at the indicated time points, frozen on dry ice, and lyophilized. In both cases, the isotope distribution of product nucleotides were analyzed by ESI-LC-MS using the ion pairing separation method described above. The exchange rate, k_x, was estimated by fitting the observed ratio of 16O-UMP to 18O-UMP, *l*/*h*, over time, where the observable *l* and *h* species (light and heavy, respectively) were estimated by considering hydrolysis and exchange of the UMP content of the starting material, *L* and *H* respectively, to be first-order irreversible processes under these conditions. We considered an approximated reaction scheme,

$$(1) \qquad H \overset{k}{\rightarrow} h \overset{k_x}{\rightarrow} l \overset{k}{\leftarrow} L \quad ,$$

where *k* is the hydrolysis rate of both *L* and *H* starting material (neglecting any

kinetic isotope effect) and h_0 and l_0 are the maximum observable counts of heavy and light monomer, respectively, within the starting material, which were constrained based on their known initial stoichiometry. We solved the system of first-order differential equations to obtain the time-dependent observable counts of *l* and *h*, where we assumed no significant difference in ionization efficiency between them:

$$h = \frac{k}{k_x - k} h_0 \left(e^{-kt} - e^{-k_x t}\right) \quad (2)$$

$$l = \frac{k}{k_x - k} h_0 e^{-k_x t} - \left(\frac{k_x}{k_x - k} h_0 + l_0\right) e^{-kt} + h_0 + l_0 \ . \quad (3)$$

Alternatively, to examine ^{18}O isotope exchange with bulk solvent when the product nucleotide carries the labeled monophosphate in the ribose 5′position, the preE-let-7 RNA isotope labeled at U11 (90 pmol) was digested enzymatically with 1 U nuclease P1 from *P. citrinum* (USBio) for 2 h at 42 °C in the enzyme's diluted storage buffer alone (3 mM NaOAc, pH 5.3, 0.5 mM $ZnCl^2$, and 5 mM NaCl).

Author contributions

V.S.L and A.B. conceptualized and designed the project. E.M.R generated LIN28-let-7 crosslinked complexes and digested heteroconjugates. V.S.L. completed the stable isotope labeling of let-7 substrates and LC-MS experiments and analyses. V.S.L, A.B. and J.S.W wrote the manuscript.

Acknowledgements

J.W.S. is an Investigator of the Howard Hughes Medical Institute. This work was supported by grants from the Simons Foundation to J.W.S. (290363) and from the National Cancer Institute to P.S. (NIH R01CA163647). A.B. was supported by a fellowship from the Academy of Finland, and E.M.R. is supported by the National Science Foundation Graduate Research Fellowship (DGE1144152) and by the UNCF-Merck Graduate Research Science Initiative. We further thank Drs. L. Li, A. Pal, and A. Fahrenbach for helpful discussions.

Chapter 5

DEAD-Box Helicase P72 phase separates in vitro and associates with liquid-like ActD induced Nucleolar caps in vivo.

Contributors: Elizabeth M. Ransey, Piotr Sliz

Supplementary Materials for this Chapter are in Appendix F.

At the time of submission, this Chapter is in preparation for publication.

Introduction

Compartmentalization of diverse biochemical, cellular processes and pathways is a vital feature of cellular organization. Though spatial separation has typically relied on lipid bilayers functioning as membrane barriers, an increasing number of cellular bodies and organelles are being characterized as existing independently of a membrane, particularly those within the nucleus (nuclear bodies) (Phair and Misteli, 2000) . Liquid-liquid phase separation (LLPS), the condensation of concentrated protein components (especially RNA binding proteins, RBPs) into dense, yet dynamic bodies has emerged as critical to the assembly and maintenance of both nuclear RNP bodies (such as nucleoli, histone locus bodies, cajal bodies and pml bodies) and cytoplasmic (such as stress granules, P bodies and germ cell granules) granules and membraneless organelles (reviewed in(Mitrea and Kriwacki, 2016)).

While new sequence features of proteins that undergo LLPS are constantly being characterized, generally, these proteins contain intrinsically disordered regions (IDRs) with embedded low complexity (LC) motifs, as well as globular, RNA-binding domains (Decker et al., 2007; Kato et al., 2012; Lee et al., 2016; Sun et al., 2011) (Lin et al., 2015; Weber and Brangwynne, 2012). Sequences within the IDRs contribute to weak, multivalent interactions (electrostatic, dipole-dipole, cation-pi, etc.) homo- and heterotypically with other proteins and RNAs inside phase separated droplets. The weak character of these interactions causes intermolecular shuffling, allowing for even very dense droplets to maintain fluidity. These interactions additionally make droplets and

membraneless bodies environmentally responsive both in vitro and in the cell and show sensitivity to protein concentration, ionic strength and temperature (Lin et al., 2015; Nott et al., 2015). Importantly, RNA binding is increasingly shown to drive RNP droplet assembly and affect droplet stability (Berry et al., 2015). In at least one case, different RNAs have been shown to confer different viscoelastic properties on the phase separation of a single protein (Zhang et al., 2015).

DEAD-Box RNA Helicase, P72 is a major nucleoplasmic protein with numerous roles in RNA metabolism. P72 has been shown to interact with RNA Polymerase II and other transcription factors such as CBP, P300 and PCAF (Janknecht, 2010). Additionally, P72 has been identified as an accessory factor to the Microprocessor, the ~670 kDa complex that cleaves miRNA precursors (Gregory et al., 2004), and has been deemed important for efficient miRNA processing (Fukuda et al., 2007). Though P72 has neither been characterized as a component of a membraneless organelle nor observed to phase separate in vitro, a mass spectrometry (MS) proteomic analysis determined that P72 is enriched at HeLa cell nucleoli (the largest membraneless organelle in the cell) after transcriptional arrest by the RNA Pol I inhibitor and cancer drug, Actinomycin D (ActD) (Andersen et al., 2005). Furthermore, it has been reported (though direct evidence has not been shown) that P72 localizes to a particular substructure of the ActD treated HeLa cell nucleoli, called dark nucleolar caps (DNCs) (Shav-Tal et al., 2005). The basis of ActD cap formation is unclear.

Here, we made the first observation and characterizations of the in vitro liquid-liquid phase separation behavior of the DEAD-Box RNA helicase, P72. We

demonstrated that the behavior is consistent with numerous other proteins recently characterized as underlying the formation of well-known membraneless organelles (such as germ granules, stress granules and the nucleolus (Elbaum-Garfinkle et al., 2015; Lin et al., 2015; Nott et al., 2015)). Namely, we showed phase separated droplet responsiveness/sensitivity to protein concentration, ionic strength and, in a limited instance, temperature. Our data suggest that both terminal IDRs contribute to phase separation as removal of both IDRs appears to have a compounded deleterious effect on LLPS behavior. In vivo, we confirmed the ActD induced cap localization of P72 in HEK 293 cells. Lastly, our imaging data suggests that the DNCs may form via a liquid-like condensation mechanism, whereby small droplets grow and fuse over time on the periphery of ActD treated nucleoli. Combined, the data suggest that IDR sequences of P72 (and the associated LLPS), may contribute to DNC structural attributes, structures which may be more liquid-like than previously appreciated.

Results

P72 phase separates in vitro.

Due to numerous determinations that the multivalent interactions that contribute to separation often center on weak electrostatic interactions, a general trend in the assessment of in vitro liquid-liquid phase separation of proteins has been to examine the salt and protein concentration dependence. Under the model of polymer phase separation (Flory, 1942 and Huggins, 1942), the expectation is that increasing salt concentrations and lower protein concentrations will prevent or diminish phase separation, while lowering the ionic

strength and increasing in the protein will have the opposite effect. To assess the ability of P72 to phase separate in vitro, we chose to complete a systematic turbidity assay using recombinant full-length P72. Purified, concentrated P72 was dialyzed or diluted into buffers at pH 7 with varying concentrations of NaCl to reach final concentrations of 25-200 mM NaCl and 0.3-10 μM P72. Sample turbidity was determined by measuring absorbance at 340nm on a Nanodrop spectrometer (b=1cm) and A_{340}=0.1 was established as the threshold above which, liquid-like droplets could be observed via light microscopy (Figure 5.1A).

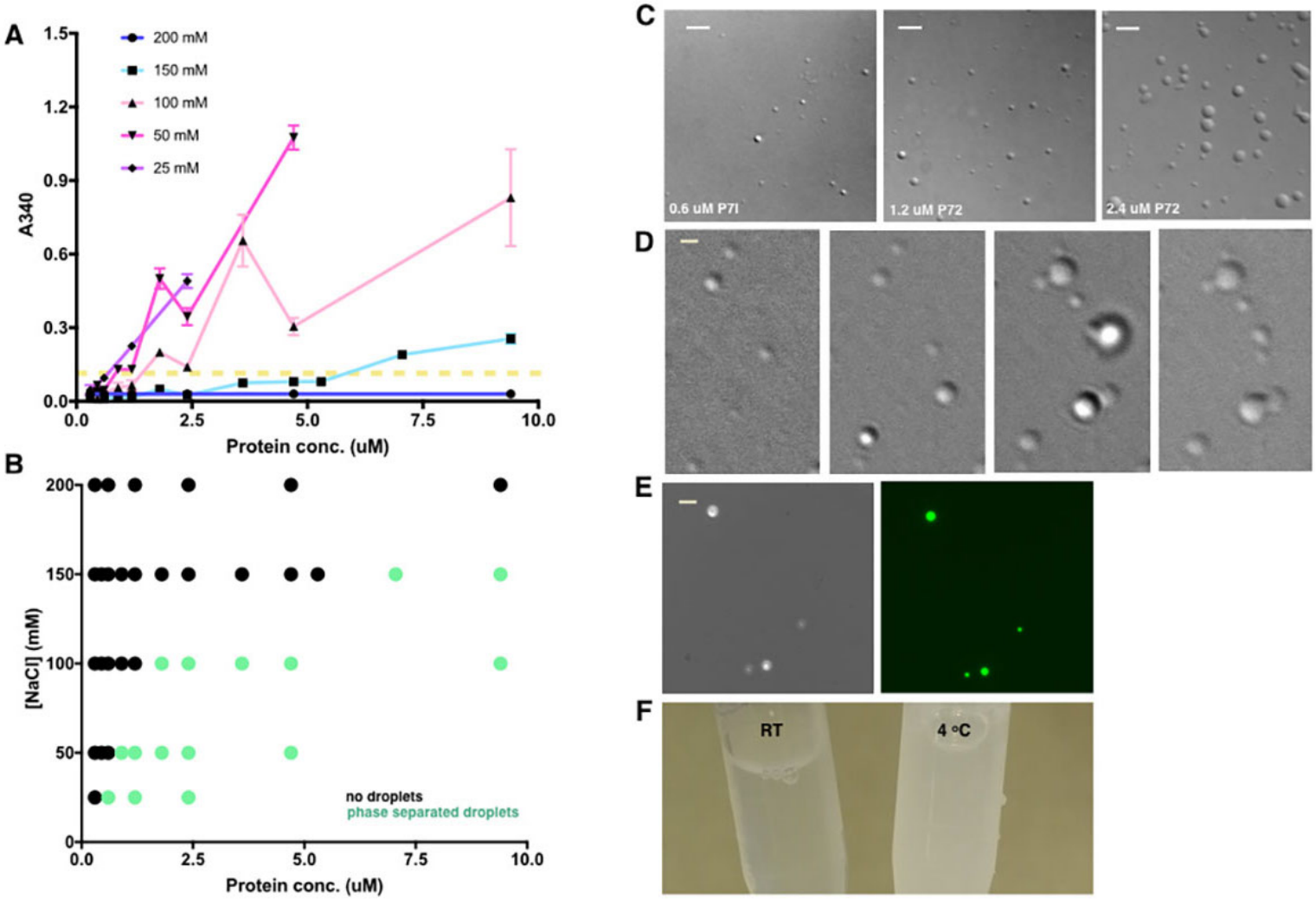
Applying the A_{340}=0.1 threshold, we were able to systematically measure turbidity of combinations of different component concentrations to generate a phase diagram (Figure 5.1B). Numerous observations were consistent with the expectations of a phase separating protein. Namely, we observed that P72 was turbid at low micromolar concentrations of protein at near physiological ionic strength (1.8 μM P72, 100 mM NaCl). We further observed that lower salt concentrations increased turbidity at lower protein concentrations, with the lowest component concentrations being 0.6 μM protein and 25 mM NaCl (Figure 5.1A).

We observed that increasing concentrations of protein at a single salt concentration resulted in the generation of liquid-like droplets of increasing size (Figure 5.1C). Additionally, we observed that droplets of P72 fuse over time (Figure 5.1D). We confirmed that the droplets were proteinaceous by combining purified GFP-labeled P72 in a 1:30 ratio with unlabeled P72 in a condition previously shown to support phase separation (Figure 5.1E). Finally, we observed that at certain concentrations of protein (>20 μM) and physiological

salt, lowering the temperature of the sample by placing on ice resulted in a reversible visible increase in turbidity, though we could not accurately record an A_{340} from chilled samples (Figure 5.1F).

Figure 5.1. P72 phase separates in vitro. (A) Turbidity measurements of increasing P72 concentrations at various ionic strengths. (B) Phase separation diagram of P72 as assessed by A_{340} turbidity measurement, green dots indicate phase separation and black dots indicate no separation. $A_{340} \geq .01$ was determined to indicate phase separation and separation was confirmed by light microscopy. (C) Representative light microscopy images of P72 phase separation occurring at the weakest ionic strength (25mM NaCl), scale bar = 4um. (D) Light microscopy images demonstrating droplet fusion/growth over time (2.4 uM P72, 25 mM NaCl), scale bar = 2 um. (E) DIC and FITC images of P72:GFP-P72 (50:1) demonstrating that droplets are proteinaceous. (F) Eppendorf tubes of a single concentration of P72 (25 uM) at 100 mM NaCl, showing increased turbidity at 4°C.

Figure 5.1. (Continued)



P72 in vitro phase separation is dependent upon IDRs.

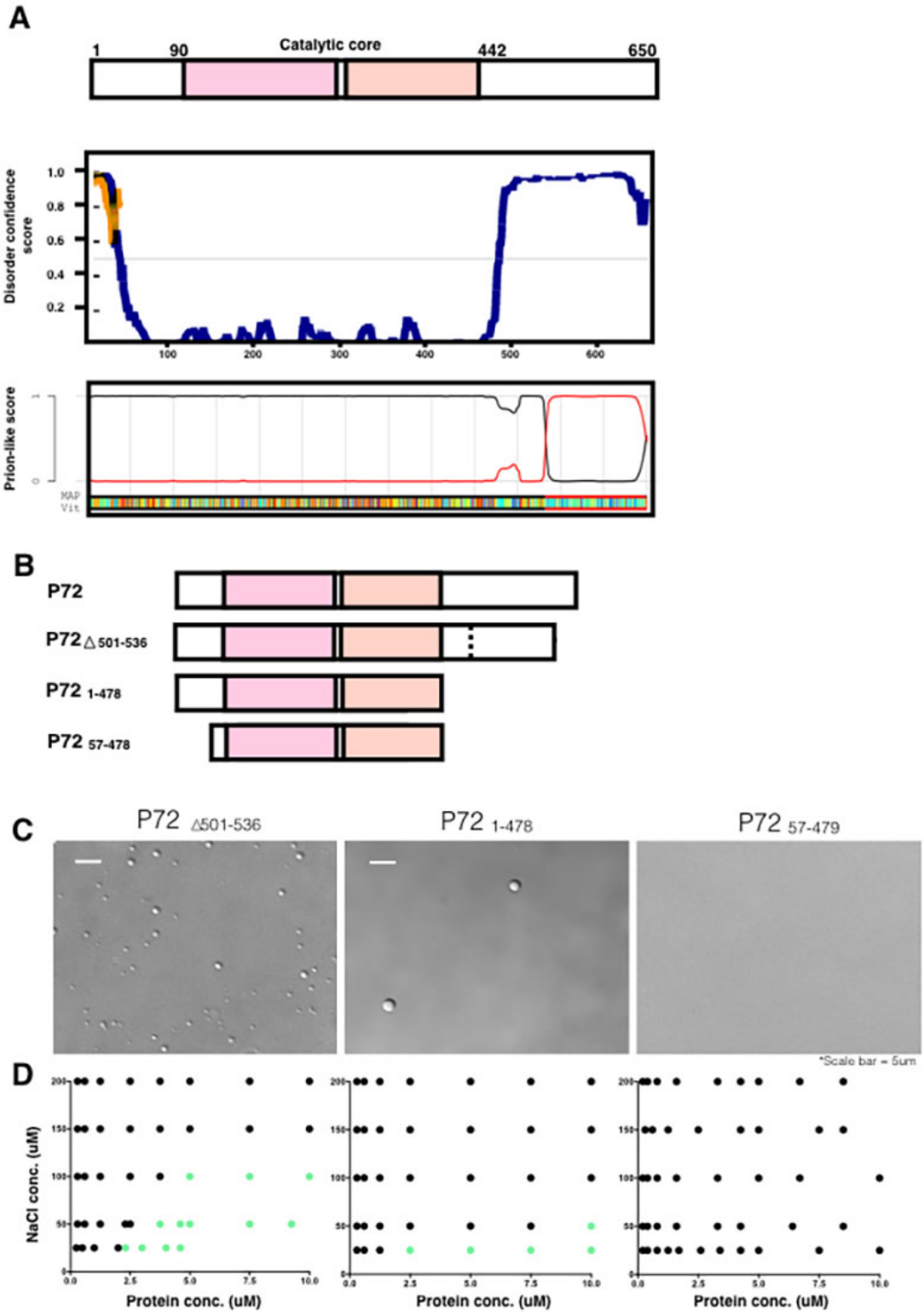
P72 has two predicted significantly disordered N- and C- terminal extensions (referred to as IDRs) that flank a globular, conserved DEAD-box catalytic core and are presumed to be the sites that confer the diverse and specialized activities of this family of enzymes (reviewed in (Cordin et al., 2006)). Combined the IDRs make up approximately half of the protein. The N-terminal IDR (residues 1-90) has three RGG motifs associated with nucleic acid binding (Thandapani et al., 2013), in addition to two nuclear localization sequences (Li et al., 2017). The C-terminal IDR (residues 442-650) contains a predicted prion-like domain (plaac.mit.edu) (PLD) (Figure 5.2A).

To test whether the P72 terminal IDRs contribute to in vitro phase separation, we generated constructs of P72 IDR truncations and overexpressed and purified the proteins to complete phase separation trials as described previously for the full-length protein. Truncation constructs included P72 $_{\Delta 501\text{-}536}$, which removed a 36 residue charged block just before the C-terminal PLD; $P72_{1\text{-}478}$, which removed the entire C-terminal IDR and $P72_{57\text{-}478}$, which removed both IDRs (Figure 5.2B). Using the same assay that was described for the full-length protein, we observed that removal of the charged block of the C-terminal IDR had relatively little effect on the ability of P72 to phase separate (phase separating at 5 μM protein in 100 mM NaCl). Alternatively, removal of the entire C-term IDR significantly shifted the phase separation diagram of P72, requiring greater concentrations of protein to phase separate (Figure 5.2C,D and Appendix Figure F.S1). Finally, removal of both IDRs predictably had the most dramatic effect and

prevented phase separation within the parameters of our assay (Figure 2C,D and Appendix Figure F.S1). We additionally observed that this construct was prone to irreversible precipitation at concentrations higher than 20 μM in physiological osmolarity (data not shown).

Figure 5.2. P72 IDRs support liquid-liquid phase separation. (A) Schematic of the full-length P72 sequence features. The catalytic core (residues 90-442) is made up of two RecA domains, the N-terminal RecA domain contains the identifying DEAD motif that is the namesake for this family of RNA helicases. Schematic of the full-length protein is centered over structural disorder prediction (generated by DISOPRED3, http://bioinf.cs.ucl.ac.uk/psipred/) and prion-like domain prediction (PLAAC, http://plaac.wi.mit.edu/).(B) Schematics of IDR truncation constructs analyzed for phase separation. (D) Representative light microscopy images of phase separation at lowest separating protein and salt concentrations, scale bar = 5um. (E) Phase separation diagram of P72 as assessed by A_{340} turbidity measurement, green dots indicate phase separation and black dots indicate no separation. $A_{340} \geq .01$ was determined to indicate phase separation.

Figure 5.2. (Continued)



P72 localizes to dark nucleolar caps in response to Actinomycin D (ActD) treatment in HEK 293 cells.

Upon transcription inhibition by ActD, the nucleolar compartments segregate and rearrange via a mechanism that is not well understood. In addition to the original three compartments (granular component (GC), dense fibrillar component (DFC), and fibrillar center (FC), 'dark nucleolar caps' (DNCs) (named so due to visible characteristics in phase contrast microscopy experiments), form around the segregating nucleoli (Shav-Tal et al., 2005). DNCs are ostensibly composed primarily of nucleoplasmic proteins and pre-rRNA and are associated with several RBPs including RBM14 (PSP2), an essential component of nuclear paraspeckles (Hennig et al., 2015). To date visualization experiments concerning both ActD nucleolar caps and P72 endogenous localization have primarily utilized HeLa cells as a model system. We, however, use HEK 293 cells, thus, we first demonstrated that P72 localizes to the nucleoplasm in both HeLa and HEK 293 untreated cells using fixed cell immunofluorescence. In these experiments, we use Fibrillarin (FIB), as a nucleolar marker (DFC component) we can clearly observe that P72 is excluded from these regions containing FIB (Figure 5.3A).

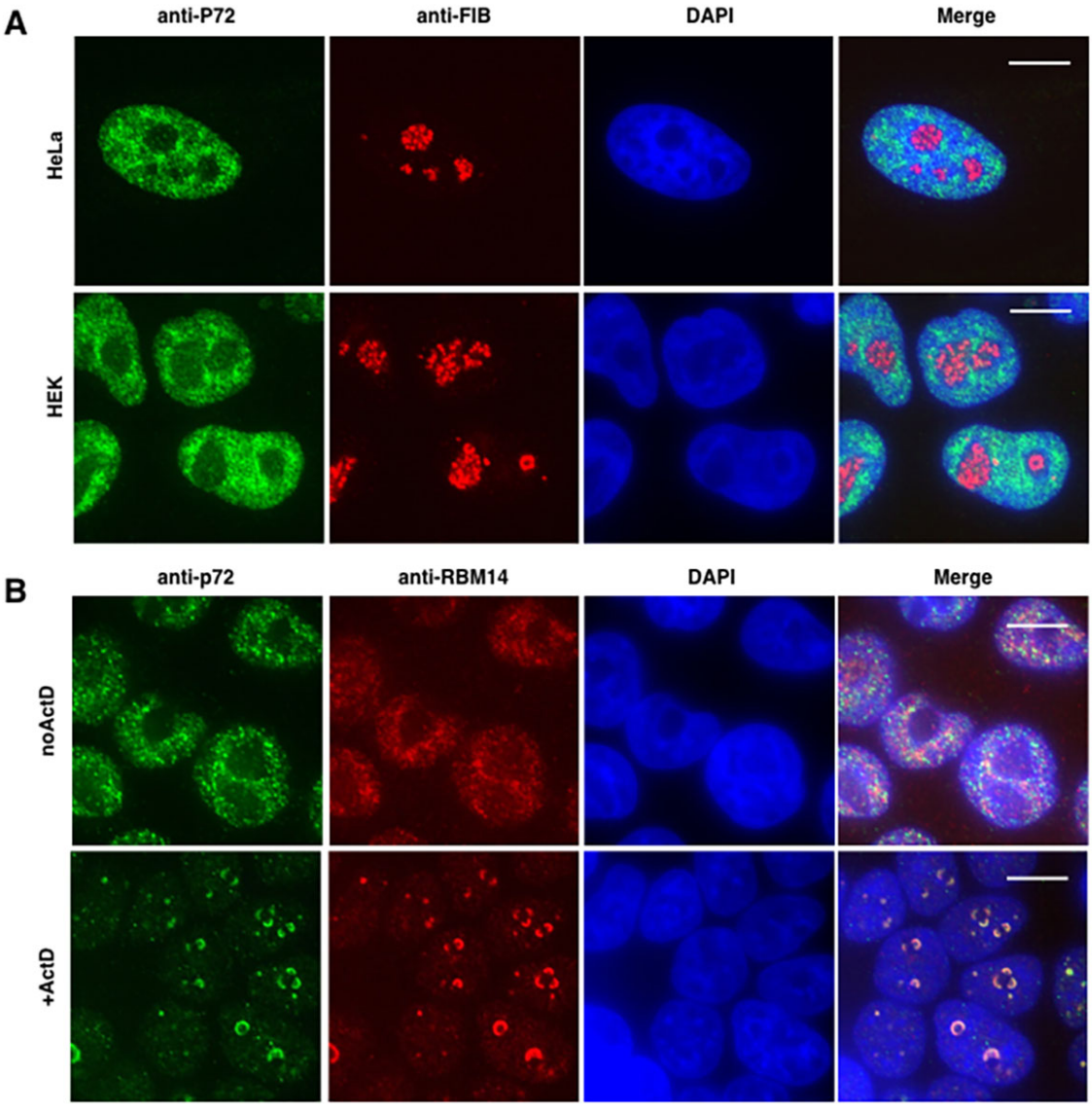


Figure 5.3. P72 endogenous localization. (A). Confocal images show P72 maintains a nucleoplasmic localization in both HeLa and HEK cells. Fibrillarin (FIB) is a nucleolar marker. (B) Confocal immunofluorescent images demonstrating that P72 localizes to dark nucleolar caps associated with the RNA binding protein RBM14. Scale bar = 10 um.

Though P72 was reported (in text) to localize to ActD induced nucleolar caps (Hennig et al., 2015), and data from a previous MS proteomic analysis of HeLa cell nucleoli determined significant enrichment of P72 at the nucleolus after ActD treatment (Andersen et al., 2005) – no direct evidence has been shown of P72 at DNCs. To observe P72 localization in response to ActD treatment, we treated live HEK cells with 2.5 µg /ml ActD and incubated at 37 ^{o}C for 4 hrs. At the end of the treatment period, cells were fixed with 4% paraformaldehyde, and immunostained for RBM14 and P72. Confocal imaging confirmed that P72 localizes to DNCs (Figure 5.3B).

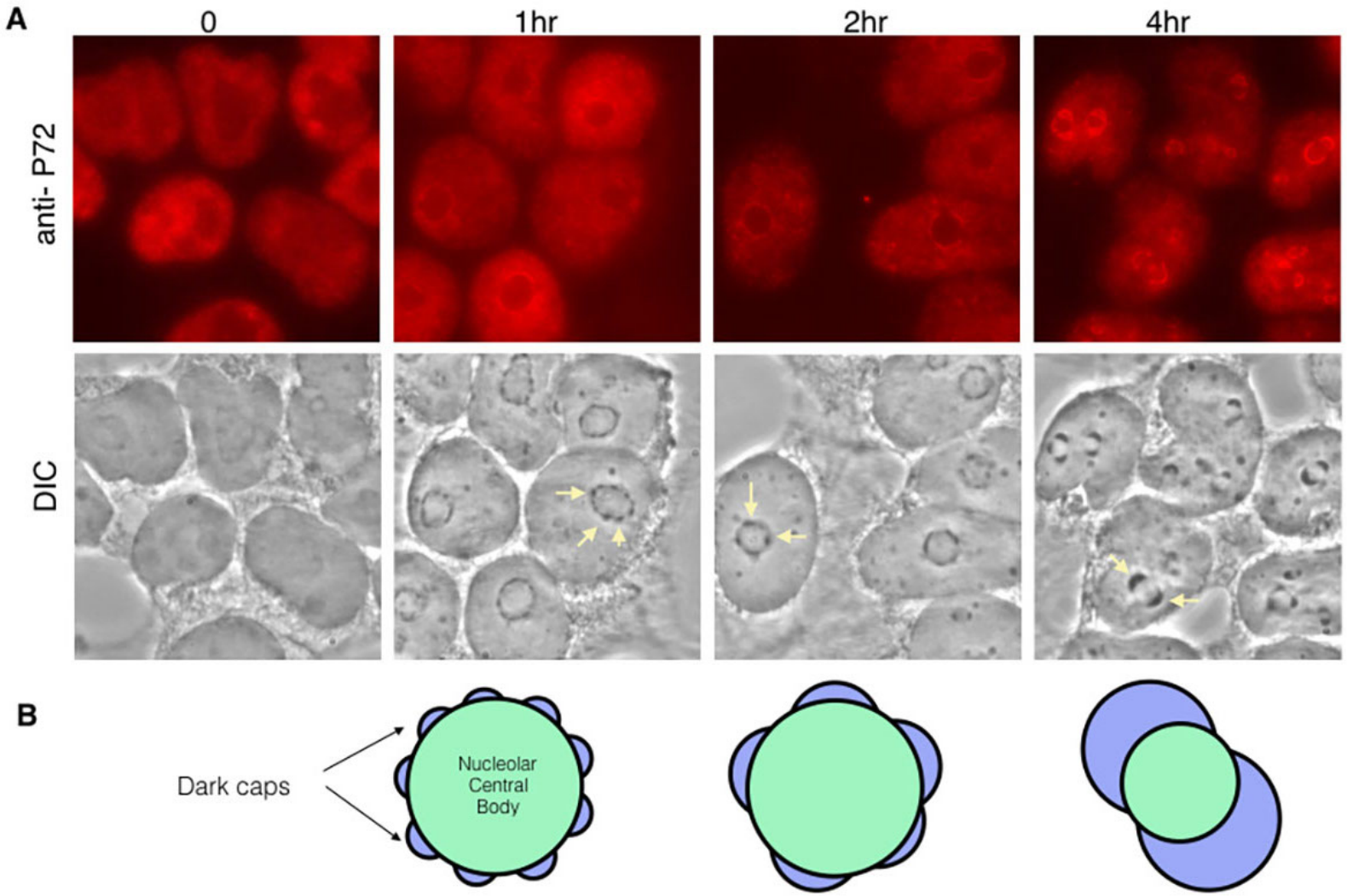
ActD Induced DNCs appear to form through a liquid-like condensation mechanism.

Though numerous DNC components have been annotated and FRAP experiments have determined that these bodies exchange with the nucleoplasm and require energy to form (Shav-Tal et al., 2005), the mechanism of formation is unclear, as fine details of DNCs are only observed several hours after treatment. To observe and characterize cap formation we completed a series of staggered ActD treatments (2.5 µg /ml) over a several hour period to allow for variable treatment lengths. At each time point (0 hr, 1 hr, 2 hrs and 4 hrs), treatment progression was halted by fixing cells with 4% paraformaldehyde, at which point cells were immunostained for P72 (Figure 5.4A). At the 4 hr time point, we were able to see dark caps in transmitted light (Brightfield) that correspond to P72 immunostained DNCs (Figures 5.3B and 5.4A), thus, throughout the brightfield

images, we associated the phase contrasted dark puncta as being DNCs or proto-DNC structures. Interestingly, at 1 hr post ActD treatment, numerous tiny proto-DNC structures or puncta are visible dispersed on the periphery of the nucleoli. At 2 hours, the puncta become noticeably less numerous and larger and this trend of growing in size and decreasing in number continues throughout the time course (Figure 5.4A). At 4 hrs, on average 2-3 caps are clearly visible for each nucleoli. This same pattern was observed for time course treatments with 1 μg/ml and 5 μg/ml concentrations of ActD, which cover the typical concentration spectrum within studies utilizing ActD as a cellular stress reagent (Appendix Figure F.S2). From the initial data, we posited that these observations show liquid-like fusion or condensation mechanism formation and growth (Figure 5.4B).

Figure 5.4. Dark Nucleolar caps appear to form via a condensation mechanism. (A) Timelapse images showing changes in dark nucleolar cap formation over time. At early time points, small dark puncta appear and encircle the segregating nucleolar central body, over time puncta appear to become less numerous and grow larger (yellow arrows). (B) Schematic model of liquid-like fusion behavior of dark nucleolar caps as observed in images.

Figure 5.4. (Continued)



While observations from fixed cells indicated a clear pattern, the fixation and staining of cell plates would not allow us to follow cap formation in individual cells to potentially observe fusion events. Thus, to allow us to observe potential cap growth or fusion in individual cells, we generated a mammalian expression construct where the DEAD-Box catalytic core of the protein is replaced with GFP, leaving the IDR termini on either side (P72-swap) to act as a DNC probe (Figure 5.5A). We confirmed that GFP alone does not localize to ActD caps (Figure 5.5B). We found that whereas exogenous expression of GFP-labeled full-length P72 mislocalized the protein to the nucleolus (in the absence of ActD) (Appendix Figure F.S3), the exogenously expressed P72-swap construct localized to the P72 endogenous compartment – the nucleoplasm (Figure 5.5C). Upon ActD treatment (2.5 μg/ml, 4 hrs) we found that P72-swap localized to DNCs (as indicated by RBM14 staining), which are adjacent to FIB associated caps (light nucleolar caps (LNCs) (Shav-Tal et al., 2005) (Figure 5.5C), making P72-swap a suitable probe to monitor DNC formation in live cells.

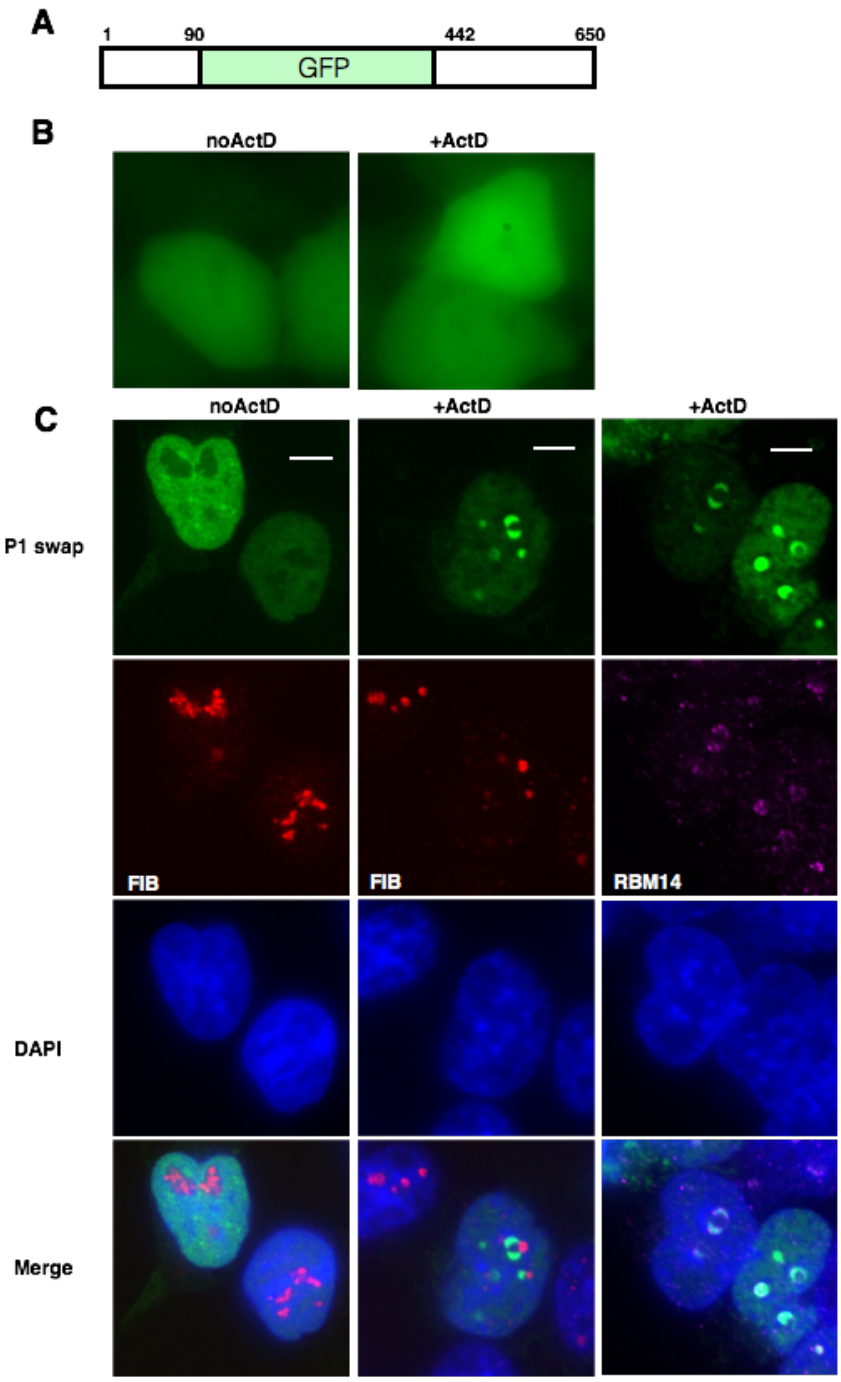


Figure 5.5. P72 IDR construct is a good probe to monitor live cell ActD cap formation. (A) Schematic of the P1swap construct. (B) Widefield immunofluorescent images showing that GFP does not form discernible cap structures in response to ActD treatment. (C) Confocal immunofluorescent images demonstrating that P72 localizes to dark nucleolar caps associated with the RNA binding protein RBM14, scale = 5um

To monitor DNC formation in live cells, we transfected the P72-swap construct into plated HEK 293 cells. ActD treatment (2.5 μg/ml) was initiated 20 hrs post transfection and successive widefield images at 15 minute intervals were taken. Cells could only be imaged for ~1hr due to photobleaching and phototoxicity, thus, each 1 hr range of micrographs observes a different cell or group of cells (Figure 5.6). Within the first hour, we observed that the P72-swap construct concentrated relatively uniformly at the apparent nucleolar surface, beginning at 30 minutes (Figure 5.6). Beginning in the second hour, at time = 0 mins numerous fluorescent puncta can be seen at a larger and a smaller nucleoli in one cell nucleus and at time = 45mins, they are clearly larger and fewer. In the hour 3 and 4 ranges, there is much less observable DNC growth or movement, though it is unclear whether this is due to increased photo sensitivity due to prolonged drug exposure, though caps are larger still, with some appearing to be elongated around the spherical nucleolar bodies.

Figure 5.6. Dark Nucleolar caps appear to form via a condensation mechanism. Widefield fluorescent images of the time dependent condensation of GFP-P72-IDR construct which colocalizes with RBM14, a DNC indicator; scale bar - 5um.

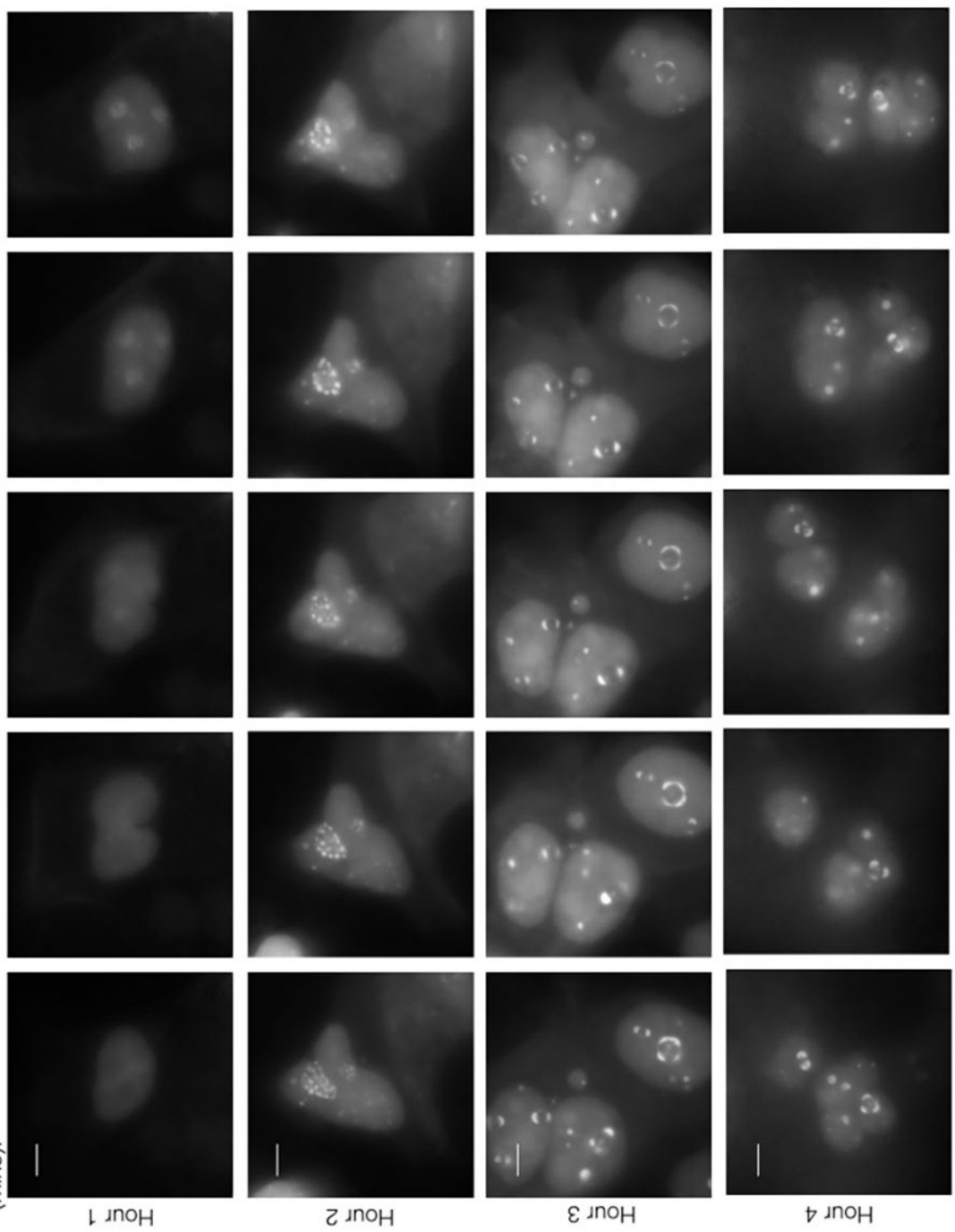


Figure 5.6. (Continued)

Finally, to quantify changes in DNC puncta number and size, we repeated this experiment on another plate of P72-swap transfected cells, but used confocal imaging instead of widefield. At 45 minutes, 1, 2, 3 and 4 hrs post ActD treatment, Z stacks of images were acquired to facilitate 3-D reconstruction of the cells. As the intense and repeated illumination of confocal acquisition ultimately killed the cells very quickly, we could not acquire time lapses but instead imaged different plate positions at different time points. At each time point, at least 2 plate positions were recorded, capturing several observable cell nuclei (8-11, Appendix Table F.SI) from which proto-DNC puncta and DNCs were counted and statistics generated.

Though we generally observe 2-3 nucleoli per cell (before ActD treatment), we note that some nucleoli are noticeably larger than others, furthermore, following ActD treatment, boundaries of individual nucleoli are not consistently, clearly discernable throughout all samples. Therefore, we chose to represent our data as puncta counts per nuclei. Our count data confirm that the number of puncta per nuclei is decreasing (Figure 5.7A, Appendix Table F.SI.), from an average of 114 ± 30 puncta at 45 mins to 7 ± 5 puncta at 4 hrs. We also showed that the counts at each time point are significantly different from each other with the exception of the counts at 2 and 3 hrs, though this is likely due to the high variance in sample counts at 2 hrs (Appendix Table F.SI.).

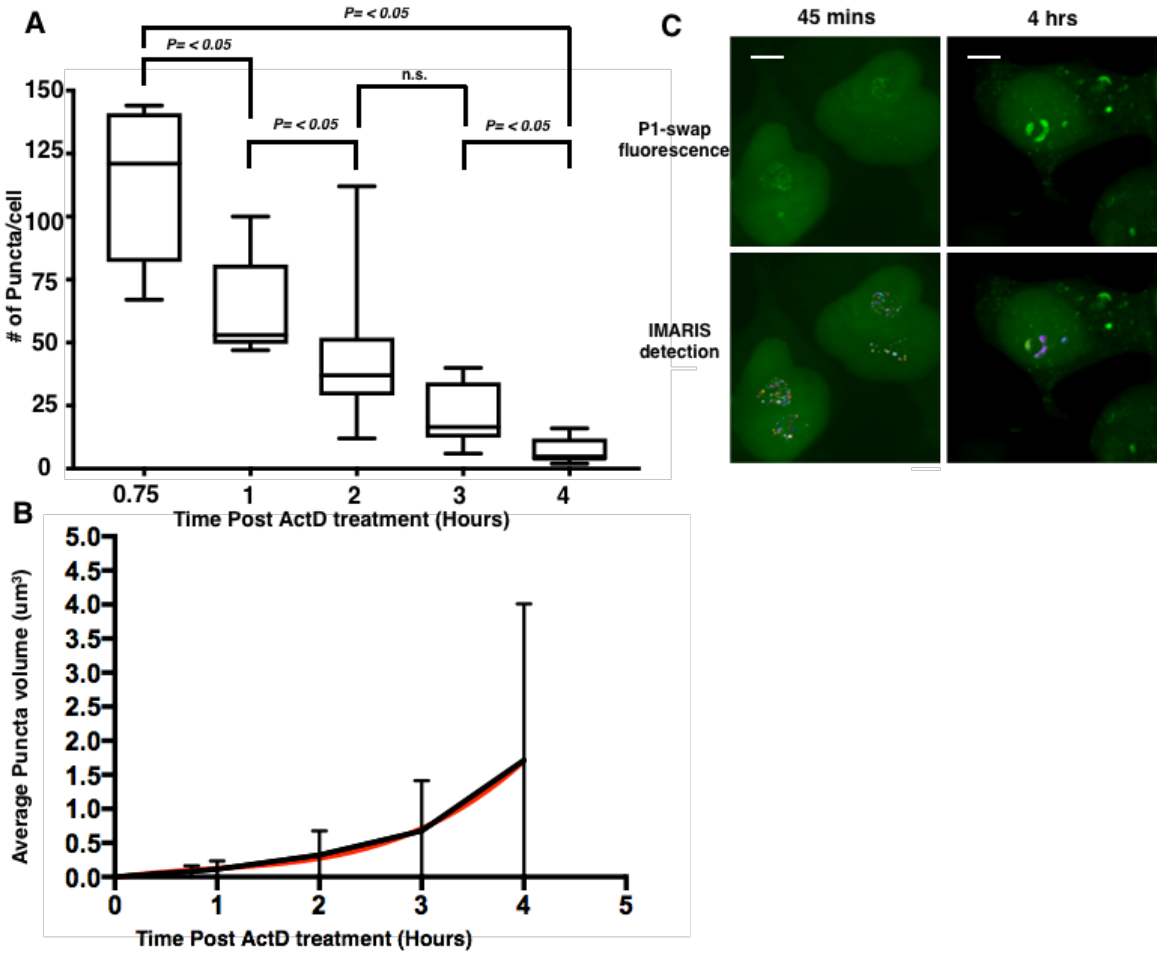


Figure 5.7. Count and volume quantitation of DNC puncta growth. (A) Box plot showing distribution of counts of puncta per nuclei, n= 8-11 cells at each time point (statistical details listed in Appendix Table. F.SI.). Statistical P values were calculated using unranked Kolmogorov-Smirnov test.(B) Graph of the mean and standard deviation of DNC puncta volume at different time points fit with a third order polynomial regression (red line, $Y=0.0454x^3 -0.1297x^2+ 0.2212x-0.01098$) indicating steady volume increase over time (statistical details listed in Appendix Table F.SII). (C) Example 3D reconstructions from confocal Z-stacks of P1-swap expressing, ActD treated HEK 293 cells. Puncta are highlighted (multi-colored) by the IMARIS program and indicate the counts and volume of puncta measured. Times indicated are following 2.5 ug/ml ActD treatment; scale bar = 5 um.

Using computational image analysis software (IMARIS), we were also able to determine the volume of all identified proto-DNC and DNC puncta at each time point. In agreement with imaging data, we calculated that the average size of puncta increased from 0.1 ± 0.1 μm^3 to 1.7 ± 2.6 μm^3 (Figure 5.7B,C, Appendix Table F.SII.).

Discussion

Herein, we have made the first characterizations of the DEAD-Box Helicase, P72 as a protein that undergoes LLPS in vitro. We showed that low micromolar concentrations of protein separate in vitro in low salt conditions and we have attributed this behavior to sequences within the disordered N- and C-terminal IDRs of the protein. Qualitatively, our phase diagrams (generated from turbidity measurements) appear to be consistent with the Flory-Huggins polymer phase model, though more data would be required (including with temperature variables) to be sure. We also made observations that P72 droplets fuse in vitro upon making contact, we confirmed that droplets were proteinaceous and we demonstrated increased turbidity at lower temperature.

We showed that P72 does relocalize to DNCs in ActD treated HEK 293 cells and we found that the two terminal IDRs (in the absence of the catalytic core) were sufficient to relocalize GFP to DNCs. Thus we used a GFP labeled, IDR only construct (P1-swap) as a probe to monitor cap formation in real time.

We demonstrated that the caps appear to form via a condensation mechanism. Apparently, small 'proto'-cap puncta form around segregating nucleoli at early time points and then can be observed to grow larger and less numerous over time. Fusion of proto-cap puncta is suggestive of liquid-like behavior, the spherical nature of droplets in time course widefield images additionally suggests a liquid-like behavior. Furthermore, we were able to estimate both the number and size of the growing DNCs and show that the values of each are changing significantly.

Though we were able to observe in vitro, that the P72 IDRs contribute to phase separation, we were unable to identify the precise sequences required for P72 DNC localization in ActD treated HEK 293 cells.

However, following ActD treatment, we observed no significant overexpression of P72 (Appendix Figure F.S4), suggesting post-translational modifications, such as stress induced phosphorylation or changes in the availability of binding partners, RNA substrate or a combination thereof, likely underlie its ActD induced localization in vivo (as opposed to a concentration dependent mechanism).

Though, a previous study reported that DNCs are ostensibly comprised of only nucleoplasmic proteins (and pre-rRNA), imaging and colocalization studies have determined that recruitment of these proteins appears to be specific to a subset of RBPs, and not a general phenomenon (Shav-Tal et. al., 2005). The observation that P72 phase separates in vitro and is found to localize at DNCs, complements the idea that DNCs may be functional liquid bodies. This characterization, would be a first for the DNCs and would open numerous

avenues as to exploring the role of subgroups of nucleoplasmic RBPs in cellular stress responses.

Methods

Cloning

For cloning of in vitro P72 constructs, the human DDx17 cDNA, optimized for E. coli expression, was purchased as a gene fragment from IDT and truncation constructs were cloned via PCR amplification with appropriate ligation overhangs. All recombinant P72 constructs were cloned into LIC (Ligase Independent Cloning) vectors, either plasmid 2BT (His-6x-TEV-orf, amp resistance) or plasmid 2GT (His-6x-GST-TEV-orf, amp resistance), gifted from the Berkeley MacroLab, using In-Fusion HD Cloning Kit (Clontech). Mammalian expression P72 constructs were cloned from synthetic gene fragment from IDT and truncation constructs were generated via PCR amplification with appropriate ligation overhangs. Mammalian expression constructs were cloned into the C3-eGFP vector, gifted from the Blacklow Lab, as previously described.

Overexpression and purification of recombinant P72 proteins

Bacterial expression cultures were started from single colonies of BL21 Rosetta E. coli, transformed with P72 constructs and plated on LB amp/chloramphenicol media. Single colonies were picked and used to inoculate 100ml LB starter cultures and incubated overnight (~18-24 hrs) in a shaker incubator set to 37oC. The following day, 10 ml of starter culture was used to

inoculate every 1L of expression culture (typically 4-6 L). Expression cultures were incubated at 37oC until reaching O.D. of 0.4-0.5, at which point temperature was dropped to 20oC. Once cultures reached O.D. of 0.7-0.9, cultures were induced with 0.5 mM IPTG. Cultures were incubated overnight at 18oC and harvested the next day with centrifugation (4,000 x g, 25 mins). For lysis, cell pellets were resuspended in lysis buffer (100mM Bis-Tris, pH 7, 2M NaCl and 20% glycerol), sonicated, and centrifuged (18,000 x g, 30 mins). Clarified lysate was batch bound to lysis buffer equilibrated Ni-NTA resin for 1 hr prior to completion of Ni2+ affinity column chromatography. Following batch bind, supernatant was allowed to flow through and resin was washed with wash buffer (50 mM Bis-Tris, pH 7, 1M NaCl, 10% glycerol, 20 mM Immidazole, pH 7 and 1 mM DTT). His-tagged proteins were eluted with elution buffer (20 mM Bis-Tris, pH 6, 1M NaCl, 10% glycerol, 200 mM Immidazole, pH 6 and 1 mM DTT). Fractions were analyzed by SDS-Page with Coomassie staining and elution fractions containing P72 proteins were pooled. TEV protease was added to pooled fractions and samples were dialyzed in dialysis buffer (20 mM Bis-Tris, pH 6, 300 mM NaCl, 10% glycerol and 5 mM DTT) overnight. The following day, TEV-treated samples were subjected to either Ni2+ non affinity chromatography to further remove contaminant proteins (sample was batch bound to Ni_NTA resin and only flow-through was collected) followed by SPHP cationic exchange chromatography, or were immediately subjected to SPHP cationic exchange chromatography. P72 proteins were eluted from SPHP column with a gradient of low salt ion exchange buffer (20 mM Bis-Tris, pH 6, 100 mM NaCl, 10% glycerol,

and 5 mM DTT) to high salt ion exchange buffer (20 mM Bis-Tris, pH 6, 2 M NaCl, 10% glycerol, and 5 mM DTT). Ion exchange elution fractions were analyzed via SDS-Page and fractions containing P72 were pooled and concentrated for size exclusion chromatography (SEC) via spin centrifugation filter units. SEC was completed using a HiLoad 16/600 S75 PG column (GE Healthcare Life Sciences) and gel filtration buffer (20 mM Bis-Tris, pH 7, 1 M NaCl, 5% glycerol, and 5 mM DTT). SEC fractions containing P72 were pooled concentrated and quantified using a nanodrop.

Phase Transition Assay

Samples of purified proteins were dialyzed or diluted with buffer composed of 20 mM Bis-Tris, pH 7, 5% glycerol and 5 mM DTT and various concentrations of NaCl to achieve desired protein and salt concentrations. Samples were incubated at room temperature for 10 minutes and absorbance at 340nm (A_{340}) was measured in duplicate for two experimental replicates on a NanoDrop 2000 spectrophotomoeter. The threshold for positive phase separation was determined to be A_{340}=0.1, as visible turbidity and microscopic droplets could be detected at this point.

Microscopy of in vitro droplets

Samples of phase separated purified proteins were incubated at room temperature 10 minutes before 5 uL aliquots were placed on the coverslip of a 70% isopropanol washed Mattek dish. Sample drops were covered with

siliconized glass coverslip. Phase separated droplets were imaged (widefield) on an inverted Nikon Ti Fluorescence microscope to achieve either DIC or FITC images.

Cell culture

Both HeLa (ATCC: CCL-2) and HEK 293 cells (gifted from the Blacklow lab, HMS) are adherent and were maintained in DMEM supplemented with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin and incubated at 37^{o}C with 5% CO_2. Cells were passaged every 2-3 days (or upon reaching 60-90% confluency) by treating with Trypsin/EDTA and allocating cells into new plates.For live and fixed cell imaging experiments, HEK cells were transfected with GFP-P72 constructs using Lipofectamine 3000 kit with the included protocol (Thermo Fisher Scientific) for 20 hrs prior to either visualization or experimental drug treatment.

Antibodies

Primary antibodies included Rabbit Anti-DDx17 (Abcam, ab24601); Mouse Anti-DDx17 (P72) (Santa Cruz, (C-9) sc-271112); Rabbit Anti-DDx5 (P68) (Abcam, ab126730 [EPR7239]); Rabbit Anti-Fibrillarin (Abcam, ab5821); Mouse Anti-Fibrillarin (Abcam, ab4566 [38F3]) and Rabbit Anti-RBM14 (Abcam, ab70636). Secondary antibodies included Goat Anti-Rabbit-FITC (Abcam, ab6798); Donkey Anti-Mouse-AlexaFluor 594 (Abcam, ab150112); Donkey Anti-

Rabbit-AlexaFluor 647 (Abcam, ab150075); Goat Anti-Rabbit-HRP (Abcam, ab205718) and Goat Anti-Mouse-HRP (Abcam, ab205719).

Immunofluorescence and Fixed Cell Imaging

HEK cells were grown onto collagen coated Mattek dishes (No. 1.5 coverslips) (with and without drug treatment) were washed twice with PBS and fixed for 10 minutes with 4% paraformaldehyde. Plates were washed twice with PBS and then cells were permeabilized and blocked in 2% fetal bovine serum (FBS) and 0.2% triton x-100. Cells were stained with appropriate primary antibodies (1:200) in 2%FBS/0.2% triton overnight at 4^{o}C. Following primary stain, cell plates were washed three times with 2FBS/0.2% triton for 5 minutes and stained with appropriate secondary antibodies for 1 hr followed by three more washes with 2FBS/0.2% triton for 5 minutes. Coverslips (inside the dishes) were covered with Vectashield mounting media (with DAPI). Prior to double labelling experiments, each antibody was imaged individually to check for channel cross talk. Fixed cells were imaged using a Yokogawa spinning disk confocal on an inverted Nikon Ti fluorescence microscope. In some instances, widefield images were also taken using an inverted Nikon Ti fluorescence microscope.

ActD Treatment

Actinomycin D (ENZO Therapeutics, Inc.) was reconstituted in 100% ethanol to a stock concentration of 1 mg/ml. Working concentrations of ActD

were 1-5 µg/ml in DMEM/10%FBS and without penicillin-streptomycin. Treatments were initiated by replacing cell control media with ActD supplemented media, cells were incubated at 37 ^{o}C until desired time point. Treatments were halted by cell fixation with 4% paraformaldehyde and cells were stained as previously described.

Colocalization Correlation Analysis

Prior to confocal imaging, dual- fluorescent samples, single-fluorescent (either overexpressed GFP or conjugated-antibody stained were imaged and evaluated for cross-channel bleed through. A Mander's coefficient of 0.834 was calculated for the overlap of endogenous P72 and endogenous RBM14 after ActD treatment.

Live Cell Imaging

HEK 293 cells were grown onto collagen coated Mattek dishes (No. 1.5 coverslips) and mounted in either a Tokai Hit INU microscope stage heated chamber or an Okolab Stage Top Incubator warmed to 37 °C, 5% CO_2. Widefield images were taken using an inverted Nikon Ti fluorescence microscope. Images were acquired with a Hamamatsu ORCA R2 cooled-CCD camera controlled with MetaMorph 7.2 software. For time-lapse experiments post ActD treatment, images were collected every 15 min, using an exposure time of 500 ms and 2×2 camera binning. Nikon Ti-E inverted microscope equipped with Plan Apo 100x (oil) NA 1.4 objective lens and the Perfect Focus System for maintenance of focus over time. Multiple stage positions were collected using a Prior ProScan

motorized stage. For the confocal ActD induced DNC puncta quantitation, images were collected with a Yokogawa spinning disk confocal on an inverted Nikon Ti fluorescence microscope with Plan Apo 100x (oil) NA 1.4 objective lens. At each time-point, 28-40 z-series optical sections were collected with a step-size of 0.25 microns, using the Nikon Ti-E internal focus motor . Z-series are displayed as maximum z-projections, brightness, and contrast were adjusted using Image J software.

Cell Lysate preparation and Western Blotting

To extract HEK cell lysates, plated cells were washed twice with room temperature PBS, trypsinized, collected and centrifuged at 2,000 x g for 5 minutes. Pellets were washed twice with PBS followed by centrifugation and aspiration. Pellets were lysed by the addition of 50 mM Tris, pH 7.5, 150 mM NaCl, 1 mM EDTA, 0.5% Nonidet P-40 (NP-40), 10% glycerol and 1X protease and phosphatase inhibitor cocktail (Thermo), and incubation at 4^{o}C with rotation for 30 minutes. Following incubation, lysate was centrifuged for 25 minutes at 13,000 x g and supernatant was removed and quantified via Bradford assay. For analysis, samples were boiled in SDS-Page sample buffer (5% glycerol, 2% SDS, 62.5 mM Tris, pH 6.8, 2% 2-mercaptoethanol, 0.01% bromophenol blue) and separated on 4-20% SDS-Page gels. Separated proteins were transferred to PVDF membrane in transfer buffer (25 mM Tris, 190 mM glycine, 20% methanol, final pH 8.3), for 1 hr at 110V. Subsequently, membranes were blocked in 5% milk for 1 hr, followed by 1hr incubation with primary antibody (1:1000) in 5% milk

with shaking. Blots were then washed three times with TBST(20 mM Tris, pH 7.5, 150 mM NaCl and 0.1% Tween 20) for 5 minutes and incubated with Horse radish peroxidase conjugated secondary antibody (1:5000), followed by three washes with TBST for 5 minutes. Blots were developed using Optiblot ECL kit (Abcam) and visualized via Biorad phosphorimager. To restain individual blots for different proteins, blots were stripped with mild stripping buffer (200 mM glycine, 0.1% w/v SDS, 1% Tween 200), 2 x 10 minute washes with shaking, followed by 2 x 10 minute PBS washes with shaking and 2 x 5 minute TBST washes with shaking. After stripping, blots were blocked in 5% milk for 1 hr in preparation for antibody staining as previously described.

Acknowledgments

The authors acknowledge equipment used in the Nikon Imaging Center at Harvard Medical School and generous support from the NIC staff – particularly Dr. Anna Jost and Dr. Jennifer Waters – the microscopy experimental design and training. The authors would like to acknowledge two undergraduate students – Cassandra Sunga (Boston University) and Rubye Peyser (Wesleyan College) for initial contributions and efforts in optimizing protein purification as well as working out phase assay conditions. We acknowledge Dr. Brandon Zimmerman for guidance and training in the completion of cell based assays.

Chapter 6

Discussion

In this dissertation, I address questions persisting in the RNA biology field concerning the comprehensive understanding of numerous, complex processes in RNA metabolism by investigating and characterizing proteins that affect RNA processing pathways.

In Chapter 2, I determined the highest resolution crystal structure of the RNA Lariat debranching enzyme (Dbr1) and unearthed a novel $Zn^{2+/}Mn^{2+}$ heterobinucleation that supports enzymatic activity. Despite significant conservation of active sites throughout the MPE family enzymes, Dbr1 is the only protein in eukaryotes to recognize 2'-5' bonds, thus, this finding will contribute to ongoing and future investigations of the basis for the vital substrate recognition and cleavage.

CLIP methods are popular and powerful techniques used to rapidly gather massive amounts of RBP target data and mechanistic insight into protein-RNA interactions. The methods, however, are largely used as stand-alone techniques without structural or biochemical validation. In Chapter 3, I present a comparative analysis of LIN28-pre-let-7 UV-induced crosslinking using a bioinformatic survey of existing CLIP datasets, as well as a tandem mass spectrometry (MS/MS) interrogation of in vitro crosslinked complexes. My data found very little consistency in results amongst different investigative groups and further found potential methodological biases in the highest resolution CLIP analysis methods, revealing the need for comprehensive analysis and validation of crosslinking techniques. Furthermore, in Chapter 4, I describe the development of a new

stable-isotopic labeling technique to resolve ambiguity for the high precision determination of cross-link sites as a complement to CLIP-based studies.

Finally, in Chapter 5, I made the first observations of the DEAD-Box protein, P72 phase separating in vitro. I biochemically characterized the phenomenon and determined that the behavior was typical of other RBPs that have been shown to undergo LLPS in vitro or in vivo. Additionally, I localized P72 to particular nucleolar structures (DNCs) in transcriptionally inactive HEK 293 cells and showed that the stress induced DNCs behave ostensibly like liquids in their formation. Compounded, the findings in this chapter potentially allude to an uncharacterized function of P72 and also, shed light on underlying characteristics of the bodies.

Though each section of the work presented here represents a complete, concise study, each unit would benefit from further investigations. For example, though I discovered a novel feature of the Dbr1 active site, a co-crystal structure with a non-hydrolyzable branched substrate would be more informative to the catalytic mechanism, as well as, substrate specificity. At the point of completion of my involvement with the Dbr1 project, I had characterized such a substrate as a competitive inhibitor of Dbr1. Co-crystallization experiments are currently underway with collaborators. Likewise, in the UV crosslinking project, sequencing of the in vitro crosslinked RNA (as opposed to solely an MS analysis) would close the gap between the two highest resolution identification techniques and allow us to determine whether there are crosslink sites at guanines being missed by MS and uridine crosslink sites being missed by sequencing. This

experiment is currently underway. Finally, the novel characterization of the phase separation of P72 and the related characterization of the liquid-like behavior of ActinomycinD induced nucleolar caps open up numerous, exciting avenues for exploration that may move beyond the scope of the original P72 project. Specifically, the determination that ActD caps may be liquid bodies suggests some functional and potentially active role to the bodies in response to cell stress. Attempts to disrupt or prevent the caps (by targeting RBM14) and discern the cellular consequences are ongoing.

Appendix A

Data publication with the structural biology data grid supports live analysis

Contributors: Peter A. Meyer, Stephanie Socias, Jason Key, Elizabeth Ransey, Emily C. Tjon, Alejandro Buschiazzo, Ming Lei, Chris Botka, James Withrow, David Neau, Kanagalaghatta Rajashankar, Karen S. Anderson, Richard H. Baxter, Stephen C. Blacklow, Titus J. Boggon, Alexandre M.J.J. Bonvin, Dominika Borek,Tom J. Brett, Amedeo Caflisch, Chung-I Chang, Walter J. Chazin, Kevin D. Corbett,Michael S. Cosgrove, Sean Crosson, Sirano Dhe-Paganon, Enrico Di Cera, Catherine L. Drennan, Michael J. Eck, Brandt F. Eichman, Qing R. Fan, Adrian R. Ferr´e-D'Amar´e, J. Christopher Fromme, K. Christopher Garcia, Rachelle Gaudet, Peng Gong, Stephen C. Harrison, Ekaterina E. Heldwein, Zongchao Jia, Robert J. Keenan, Andrew C. Kruse, Marc Kvansakul,Jason S. McLellan, Yorgo Modis, Yunsun Nam, Zbyszek Otwinowski, Emil F. Pai, Pedro Jose´ Barbosa Pereira, Carlo Petosa, C.S. Raman, Tom A. Rapoport, Antonina Roll-Mecak, Michael K. Rosen, Gabby Rudenko, Joseph Schlessinger, Thomas U. Schwartz, Yousif Shamoo, Holger Sondermann, Yizhi J. Tao, Niraj H. Tolia, Oleg V. Tsodikov, Kenneth D. Westover, Hao Wu, Ian Foster, James S. Fraser, Filipe R.N.C. Maia, Tamir Gonen, Tom Kirchhausen, Kay Diederichs, Merce` Crosas & Piotr Sliz

This Appendix originally appeared in *Nature Communications*, Vol. 7 (2016).

doi: 10.1038/ncomms10882

Abstract

Access to experimental X-ray diffraction image data is fundamental for validation and reproduction of macromolecular models and indispensable for development of structural biology processing methods. Here, we established a diffraction data publication and dissemination system, Structural Biology Data Grid (SBDG, url: data.sbgrid.org), to preserve primary experimental datasets that support scientific publications. Datasets are accessible to researchers through a community driven data grid, which facilitates global data access. Our analysis of a pilot collection of crystallographic datasets demonstrates that the information archived by SBDG is sufficient to reprocess data to statistics that meet or exceed the quality of the original published structures. SBDG has extended its services to the entire community and is used to develop support for other types of biomedical datasets. It is anticipated that access to the experimental datasets will enhance the paradigm shift in the community towards a much more dynamic body of continuously improving data analysis.

Introduction

Access to one of the most powerful tools in structural biology, X-ray crystallography allows determination of the structure (atomic coordinates) of proteins, nucleic acids, small molecule compounds and macromolecular complexes to atomic-level resolution. Crystallographic data continue to be a primary source of mechanistic understanding of macromolecules, the implications of which extend from basic research to translational studies and the rational design of therapeutics. Reflecting the significance of the technique, the number of published macromolecular crystal structures has rapidly grown to more than 100,000 and numerous investigators within

structural biology have been awarded the Nobel Prize, including Drs. Kendrew, Perutz, Watson, Crick, Wilkins, Hodgkin, Klug, Deisenhofer, Michel, Huber, Walker, MacKinnon, Kornberg, Ramakrishnan, Steitz, Yonath, and Kobilka.

To support the needs of a growing structural biology community, a global network of synchrotron beamlines (Bilderback et al., 2005) has been established and made available to researchers. These facilities remain the predominant source for crystallographic data collection. While the data collection process has become increasingly streamlined, deployment of a data management infrastructure to archive original diffraction images has been slow and uncertain (Guss and McMahon, 2014). With the exception of a modest number of data storage systems dedicated to the support of individual synchrotron beamlines (Meyer et al., 2014), or specific structural genomics projects (Elsliger et al., 2010), storage of diffraction image datasets is typically the responsibility of primary investigators. Access to these original experimental datasets is therefore dependent on the policies of individual laboratories, which vary in storage organization, institutional resources, and researcher turnover. There is no universal archiving system to store X-ray diffraction datasets, and raw datasets are rarely made publicly available. In the cases where datasets are available, their distribution format can vary significantly. A typical data set of 360 images collected on modern detectors is 5 GB, and structure determination can involve one to tens of data sets, making the logistics of storing diffraction data for many protein structures a daunting task.

The benefits of easy and public access to experimental data are numerous (Kroon-Batenburg and Helliwell, 2014). Access to primary data would support

community efforts to continuously improve existing models and identify new features through complete reprocessing of experimental data (Joosten et al., 2009; Terwilliger and Bricogne, 2014; Wall et al., 2014a) with modern software tools and improved criteria (Karplus and Diederichs, 2012). Further, original data may provide a basis for validating questionable existing structures while mistakes in structure determination may be identified earlier (Janssen et al., 2007; Matthews, 2014; Tanley et al., 2013). Additionally, access to a diverse volume of raw data can be used to develop improved software to address limitations of existing programs. Finally, access to a collection of varied experimental data will undoubtedly benefit the training and education of practitioners. The Worldwide Protein Data Bank (Berman et al., 2003; Berman et al., 2014) (wwPDB) has illustrated how these achievements can be realized with the collection of reduced experimental data, in the form of structure factor amplitudes. Complementing this resource by preserving raw experimental data and making it available to a broad community promises a profound scientific impact in structural biology and other biomedical disciplines that face the challenges of preserving large datasets.

While the primary role of the SBGrid Consortium (URL: sbgrid.org) has been to curate and support a collection of data processing software applications and to organize community-wide computing support (Morin et al., 2013), SBGrid has also been active in the management of raw, experimental datasets. In 2012, SBGrid prototyped a system based on Globus technology (Chard et al., 2015; Foster, 2005, 2011; Stokes-Rees et al., 2012) to move diffraction data between Harvard, The Advanced Photon Source, and the Stanford Synchrotron Radiation Lightsource (Stokes-Rees et al., 2012).

To support the outstanding needs of the global structural community, we have established a publication system for experimental diffraction datasets that supports published structural coordinates: the Structural Biology Data Grid (SBDG). The SBDG project was initiated with a collection of X-ray diffraction image datasets as well as a few additional dataset types contributed by many SBGrid Consortium laboratories. The collection supports a diverse subset of over 68 peer-reviewed publications and represents a sampling of numerous structure determination approaches. To evaluate the utility of such a Data Grid, we reprocessed all published diffraction datasets in this initial collection with modern software and compared the derived statistics against those reported in the original publications. We also demonstrate that by integrating the storage resources of multiple research groups and institutions, the Data Grid is poised to deliver a novel community driven data-preservation system to support various types of structural biology and biomedical datasets.

Results

Structural Biology Data Grid

The SBDG is a centralized data publication service — a repository for discovering, downloading, and depositing large structural biology datasets. We developed the SBDG to support the need of the SBGrid community to archive and disseminate X-ray diffraction image datasets, i.e. images recorded on X-ray detectors, which support published structures. More than 90% of SBGrid laboratories use X-ray crystallography in their research, and SBGrid investigators have contributed over eleven thousand X-ray structures to the PDB. The SBDG complements the PDB, which archives derived data –

merged and post-refined data from diffraction images and the resulting refined coordinates of macromolecular structural models. The Data Grid has been developed in collaboration with the Data Science team at Harvard's Institute for Quantitative Social Science, and it conforms to progressive data science standards (Appendix Table A.I). The SBDG limits its collection to datasets that support journal publications, referred to as "primary data". For X-ray diffraction data, this primary data consists of experimental diffraction images supporting a derived structural model and journal publication. Release of this primary data by the SBDG coincides with publication of the resulting manuscript and for the structural biology datasets of related PDB files. As of September 1st, 2015, the SBDG stores a diverse collection of 117 datasets, including 111 X-ray diffraction datasets and a handful of other data types including computational decoys and datasets from MicroED, lattice light-sheet microscopy, and molecular dynamics (Appendix Table A.II). These published datasets, contributed by 50 laboratories with diffraction datasets collected at 11 synchrotron facilities (Appendix Figure A.1) and several home sources, originated 94 structures and 68 journal publications. The X-ray diffraction datasets range in size from 126 MB (Lee and Raman, 2015) to 20 GB (Rudenko, 2015) with a mean of 4.9 GB and a total of 573 GB of storage. Extrapolating from this initial collection, which is quite diverse and registers at just over 0.5 TB, our current 100TB file system could immediately support roughly twenty thousand X-ray diffraction datasets (Appendix Figure A.2).

Appendix Table A.I. Data Science Standards

Disclosure	Software tools developed under this program will be incorporated into open source software and released to the community. Manuscripts and white papers describing various phases of the project will be released on a regular basis.
Adoption	All biomedical image data will be converted to the master formats, such as OME-TIFF or HDF5. Community tools to create, analyze, and manipulate diffraction images will be extended to include support for these formats. All biomedical data are assigned Digital Object Identifiers through the CDL EZID system, and follow modified DataCite and Dataverse metadata schemas. Associated metadata are registered with the International DOI Foundation, making it virtually permanent and independent of SBGrid and Harvard computing infrastructure. All datasets published through the SBDG will be citable using Force11 recommendations.
Transparency	Files within individual datasets will be deposited in their original format (no archives or encryption allowed). Self-documentation: The majority of diffraction datasets are self-documented and include the basic information required for reprocessing in the images themselves. Additional information will be collected during deposition and will include dataset representation (the ability to use the data to be processed), reference (relation to PDB files, publications, and other datasets), context (e.g. a native dataset or a derivative used for phasing), fixity (checksums), and provenance (typically the data collection facility and the project member who deposits the original dataset). With conversion to master formats, all secondary information will be appended to the image metadata along with all original headers.
External Dependencies	The ability to reprocess some older datasets and verify master format conversions could depend on access to a specific version of data processing software. As datasets enter our repository, they will be reprocessed with our Data Reprocessing Pipeline (one of several we will develop as part of our Data Mining Pipelines). Data Reprocessing Pipelines will be archived within our system, issued DOIs, and interlinked with the datasets. It is worth noting that, since 2002, SBGrid has been archiving structural biology applications and, therefore, has access to previous software versions that might be required to reprocess older datasets.
Licensing	Biomedical datasets will be deposited under the Creative Commons Zero license, supporting future development of data validation services and database replications and migrations.
Technical Protection Mechanism	The security of the deposited data will be maintained by the DAA. The DAA will join with the Library of Congress sponsored National Digital Stewardship Alliance (NDSA) and the data architect working on the project will ensure that NDSA recommendations are being followed.

Appendix Table A.II. Reference Subset. 12 X-ray diffraction datasets from the SBDG pilot collection were identified as particularly suitable for software testing and teaching activities. In addition datasets from molecular dynamics, lattice light-sheet microscopy and MicroED represent an invaluable subset.

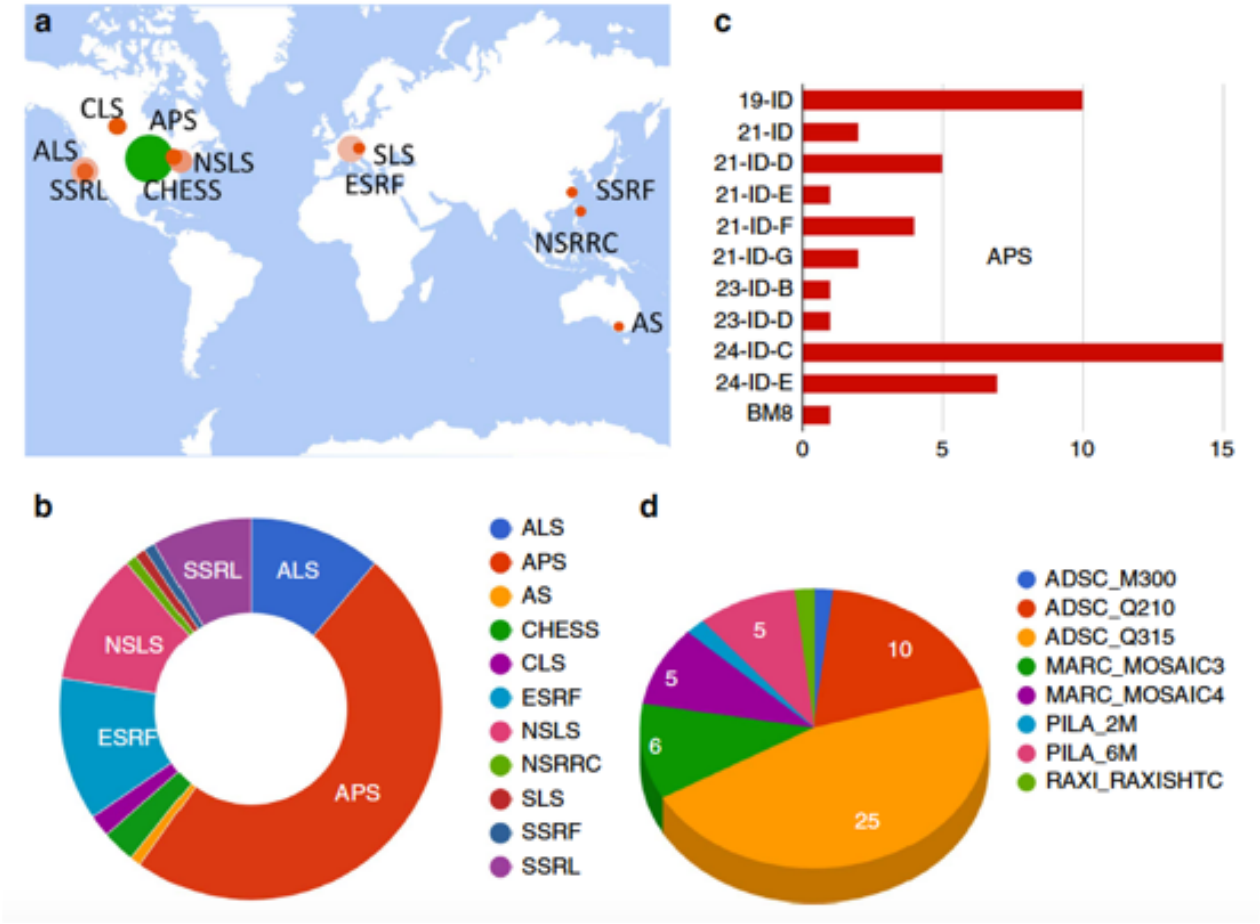
Dataset	Description
10.15785/SBGRID/5 Boggon Laboratory **Reference Case 1:** MR/Multi-crystal averaging.	Datasets from 5 crystals of SNX17 FERM domain in complex with a peptide corresponding to KRIT1's NPxY2 motif. Separate integration of the datasets and scaling together allows a complete 3.0 Å dataset for molecular replacement solution (original paper used 4GXB as a search model) and structure refinement.
10.15785/SBGRID/117 Baxter Laboratory **Reference Case 2:** MR/Low-resolution, twinned with rotational pseudosymmetry.	3.70 Å dataset collected on a crystal of thioester-containing protein 1 *S1 allele (TEP1*S1). Initial data processing suggested $P4_32_12$, but one of the two molecules (~1300 aa. each) in the ASU overlapped with its symmetry-mate. Comparison of alternative scenarios in refinement identified the true space group as $P4_3$ with twinning and rotational pseudosymmetry. Refinement was completed with TLS, NCS (local) and external restraints derived by *ProSMART* [65] using TEP1*R1 (PDB 4D94) as reference.
10.15785/SBGRID/62 Modis Laboratory **Reference case 3:** U SAD/Low-resolution.	4.5 Å dataset of a uranyl acetate derivative used for a challenging structure determination by SAD. Certain images had streaky features and were excluded from data reprocessing. The height and definition of peaks in anomalous difference Patterson maps was improved by omitting certain images near the end of the data collection run.
10.15785/SBGRID/111 Ferré-D'Amaré Laboratory **Reference Case 4:** Ba/K SAD; 91 nt RNA-chromophore complex	2.5 Å dataset collected at ALS BL 5.0.2 using 6.0 keV X-rays from a crystal of 'Spinach' a fluorescent RNA analog of GFP. Although anomalous signal was very weak, a heavy atom substructure comprised of one barium and six potassium ions resulted in good quality SAD electron density maps.
10.15785/SBGRID/3 Sliz Laboratory **Reference Case 5:** Zn SAD; 4 Zn/ASU protein/RNA complex	2.9 Å Zn SAD dataset was sufficient to determine a crystal structure of Lin28/let-7d protein-microRNA complex. X-ray beam size was adjusted to maximize flux and minimize radiation damage. One swapped-dimer is located in each asymmetric unit. Two native zinc atoms are located in each tandem CCHC zinc knuckles domain.

Appendix Table A.II. Reference Subset (Continued).

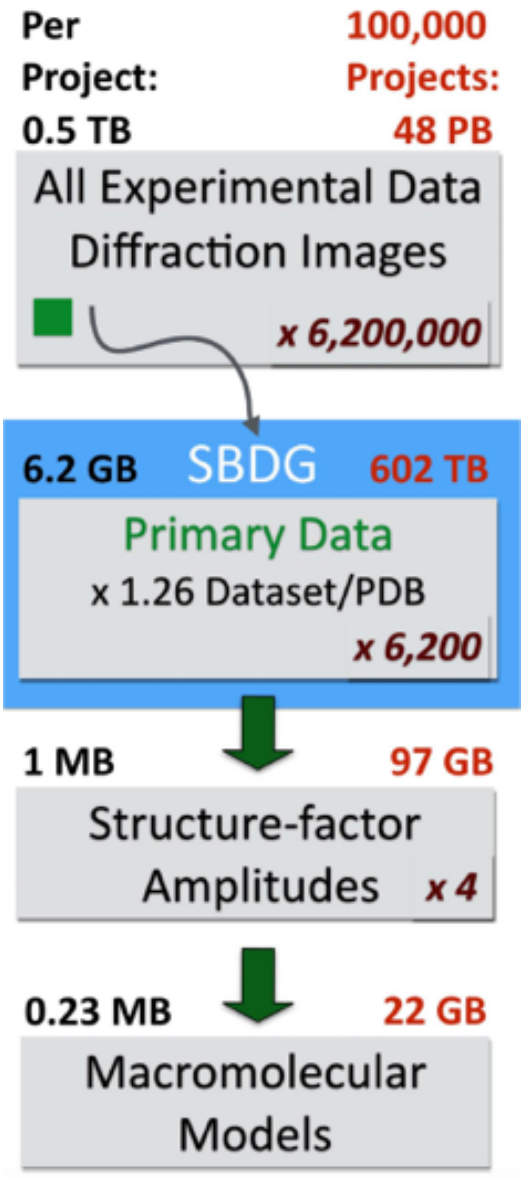
10.15785/SBGRID/123 Heldwein Laboratory **Reference Case 6:** 3.29-Å SeMet SAD 9 Se/ASU	This 3.29-Å selenomethionine SAD data set, collected at 0.9789 Å wavelength at BNL X25 beamline, was sufficient to determine the phases and to trace the structure of HSV-2 gH/gL complex[66]. There are 9 Se sites in the ASU. During integration in HKL2000, χ^2 appeared very large for some sectors of the data set. These correlated with crystal orientation and likely resulted from a large difference in cell edges (a=b=88 Å vs c=333 Å).
10.15785.SBGRID/179 Schwartz Laboratory **Reference Case 7:** MR-SAD at 7.0 Å	Contaminating *E.coli* protein 4FCC_A, acting as a crystallization chaperone, was found readily by MR. Using these MR phases seven $(Ta_6Br_{12})^{2+}$-positions could be found in the 8.8 Å derivative dataset 180. The combined MR-SAD phases were sufficient to position two copies of Nup37 (4FHL) and two copies of Nup120 in the asymmetric unit.
10.15785/SBGRID/218 10.15785/SBGRID/78 Rudenko Laboratory **Reference Case 8:** MR-SAD at 2.65 Å (44 Se atoms/ASU)	3.25 Å dataset (#218) from a crystal of the selenomethionyl neurexin 1alpha ectodomain and 2.65 Å higher resolution native dataset (#78), both collected at APS using multiple settings. The structure has 2 molecules/ASU with a total of 14 ordered domains and ~2000 residues. Molecular replacement successfully placed 8 LNS domains (using a single LNS domain as a search model, i.e. ~9% of the scattering mass) generating phases which could be used to reveal 37 out of 44 Se atoms/ASU in the 3.25 Å SeMet SAD data set. Refinement was completed using dataset #78.
10.15785/SBGRID/9 Tao Laboratory **Reference case 9:** 3.25 Å dataset used for MR with a 9-Å cryo-EM envelope	A 3.25-Å resolution dataset was collected at APS LS-CAT. The structure was determined by molecular replacement using a 9-Å resolution cryo-EM reconstruction as a phasing model. Solvent flattening and 15-fold noncrystallographic symmetry averaging were applied during phase extension.
10.15785/SBGRID/83 Drennan Laboratory **Reference Case 10:** MR/large unit cell, anisotropic.	Diffraction data from different regions of a crystal of Isobutyryl-coenzyme A mutase fused, a 250 kDa dimeric enzyme. This crystal had a large unit cell (a,b = 319 Å, c = 344 Å) and the data were anisotropic. Separate integration of the 6 wedges with individually adjusted resolution limits and scaling together yields a complete 3.35 Å dataset that can be used for molecular replacement.
10.15785/SBGRID/125 Kruse Laboratory (data collected in Kobilka Laboratory) **Reference Case 11:** MR, lipidic cubic phase	Diffraction data for lipidic cubic phase crystals of human M_2 muscarinic acetylcholine receptor bound to the agonist iperoxo, the allosteric modulator LY2119620, and the conformationally-selective nanobody Nb9-8.

Appendix Table A.II. Reference Subset (Continued).

DOI: 10.15785/SBGRID/68 Fraser Laboratory **Reference case 12:** X-ray diffuse scattering	1.2 Å dataset collected at SSRL provides a high-resolution standard dataset of the enzyme Cyclophilin to examine the influence of data collection temperature to compare to XFEL data, and to measure X-ray diffuse scattering.

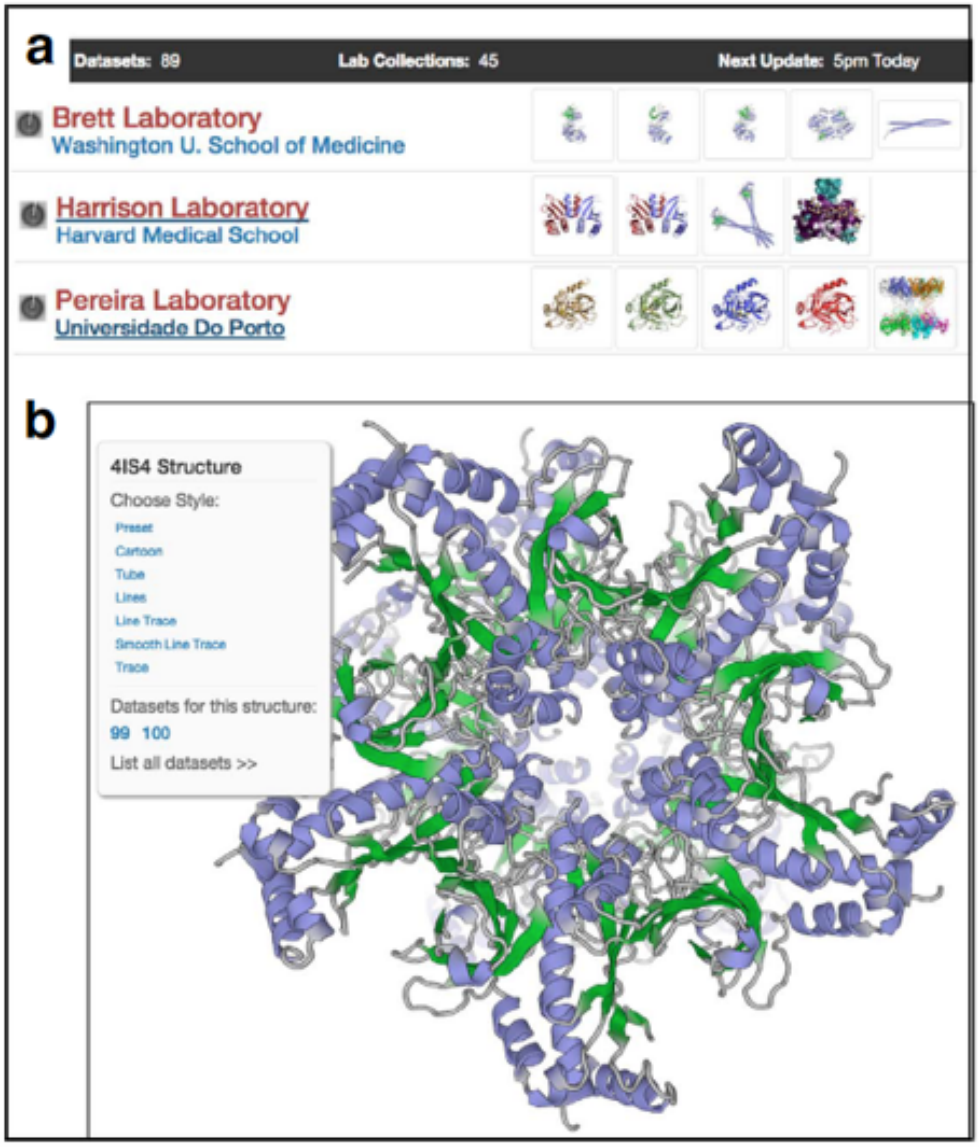


Appendix Figure A.1. Data collection statistics for the pilot subset of 112 data sets. (a,b) Data sets were collected from synchrotrons on four continents (in addition to laboratory sources, which are not broken down geographically) and originate from eleven synchrotron facilities: Advanced Light Source, Advanced Photon Source, Australian Synchrotron, Cornell High Energy Synchrotron Source, Canadian Light Source, European Synchrotron Radiation Facility, National Synchrotron Light Source, National Synchrotron Radiation Research Center, Swiss Light Source, Shanghai Synchrotron Radiation Facility, and Stanford Synchrotron Radiation Lightsource. World map image courtesy of the U.S. Geological Survey. (c) Breakdown of data sets collected at the Advanced Photon Source beamlines. (d) Data sets cover a range of detector types, including Area Detector Systems Corporation M300, Q210 and Q315, Rayonix MarMosaic, Dectris Pilatus 2M and 6M, R-AXIS HTC, and MAR345.



Appendix Figure A.2. Estimation of storage requirements for different stages of the structural biology pipeline, based on the SBDG pilot collection. For structure factor amplitudes and PDB models file sizes were obtained from a subset of 96 PDB depositions derived from the pilot data sets. On average, SBDG stores 1.26 data sets per PDB file. Numbers in red indicated the estimated storage requirements to accommodate data sets for 100,000 structures. We estimate that for each primary data set, additional 100 datasets are collected at national facilities. Primary data refers to original experimental diffraction images supporting the derived structural model, as distinguished from all experimental data (screening images, inferior quality data sets, and so on). For crystallographic experiments, reduced data refers to the integrated intensities (or amplitudes, which do not materially affect storage requirements).

The SBDG's collection of datasets can be accessed from the data.sbgrid.org website. On the home page, deposited datasets are organized into laboratory and institutional collections (Appendix Figure A.3A). Hyperlinked collection pages provide a list of selected datasets along with the dataset's corresponding data Digital Object Identifier (DOI), a link to the journal publication, the PDB ID, a link to the PDB entry, and a link to the depositors' laboratory website. The website molecular viewer, PV (doi:10.5281/zenodo.12620), offers visitors an option to view structures in a manipulatable cartoon representation (Appendix Figure A.3B). With multiple high-quality viewing options and flexible search functionality, users of the SBDG website can easily identify a small subset of relevant datasets.



Appendix Figure A.3. Organized display of data collections at SBDG. (a) Graphical view of Laboratory and Institutional Collections within the SBDG; (b) PV structure viewer, displaying a published model with links to its two primary deposited data sets.

Persistent dataset pages are an important element for any research data repository because they typically provide a landing URL, which resolves from a given DOI (Starr et al., 2015). The SBDG does not advertise unique codes, but instead distinguishes datasets by fully qualified DOIs. From each SBDG collection page or viewer page a user can access those unique Dataset Pages (Appendix Figure A.4), which offer additional information for each dataset including download instructions and the fully formatted dataset citation for inclusion in manuscripts, following best practices set by the Joint Declaration of Data Citation Principles (Martone, 2014). A Dataset Page can also be located by searching the SBDG for a PDB code, although often several related datasets are used to determine a single set of macromolecular coordinates. As the Data Grid is developed, the Dataset Pages will include additional functionality, with more information on how to reprocess datasets, extended data statistics, and discussion forums allowing users to annotate datasets after publication. Taken together, the uniquely defined Dataset Pages provide a comprehensive and persistent location for individual datasets.

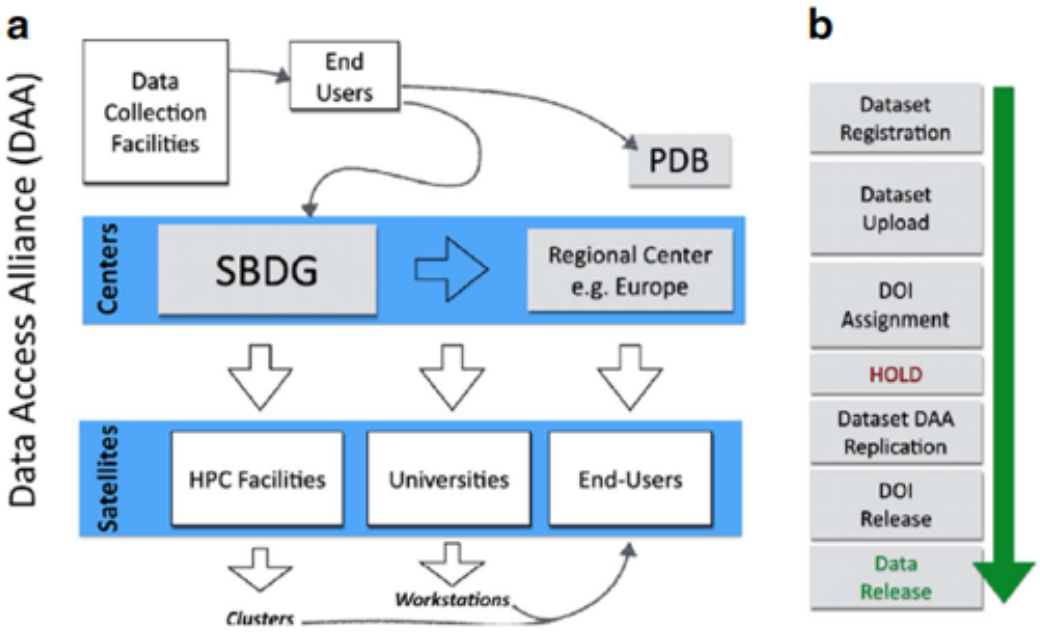
Appendix Figure A.4. SBDG persistent data set landing page (the target of a DOI resolver for a published data set). Data set metadata are displayed, as are instructions for downloading and verifying the data set.

Dataset Access

All datasets in the SBDG are readily and freely accessible to the community. Access rights were formalized with adoption of the Creative Commons Zero license (CC0), which supports dedication of research results to the public domain and is used by many open-data projects. This license allows use and redistribution of data for both commercial and non-commercial purposes without requiring additional agreements. The CC0 license does not affect patents or trademark rights of contributors, and is similar to the licensing terms that are used for macromolecular models released by the wwPDB.

Although datasets can be downloaded individually, their size can make this cumbersome. Physical access to SBDG datasets is facilitated through a data grid infrastructure that is supported by members of the Data Access Alliance (DA, Appendix Figure A.5A). The DA is a voluntary and open organization of research-data-storage providers and is being developed in collaboration with the Globus Project. The DA has two aims: 1) to minimize the chance of data loss by replicating SBDG datasets, and 2) to facilitate global data access through its members. Although it is expected that DA membership and architecture will evolve rapidly, in its current state the DA framework already provides a global solution for data dissemination. DA centers in Europe, Asia, North America, and South America replicate the entire SBDG collection and provide local access to members of regional communities. There are four DA centers: Harvard Medical School in the USA, Uppsala University in Sweden, Shanghai Institutes for Biological Sciences in China, and Institut Pasteur de Montevideo in Uruguay. As a secondary service, DA centers can provide local, direct access to datasets for their institutional research groups. For example, Harvard Medical School hosts the entire

collection and provides direct access to all data from its computing center. The DA infrastructure is further extended by the DA satellites, which replicate fractions of SBDG datasets in their local storage for direct access by members of individual institutions. This mode of participation provides an attractive option for research institutions to develop local archives of all primary data generated by the local community. For example, the NE-CAT (sector 24-ID) synchrotron beamline at the Advanced Photon Source, in Argonne, IL, replicates all SBDG datasets that originate from NE-CAT beamlines and makes them available to beamline staff and users. Another SBGrid member and DA Satellite, Yale University, replicates all datasets from Yale laboratories on its institutional storage and makes them accessible to structural biology workstations through the Network File System (NFS). We expect that, as research storage infrastructure catches up with the capacities required to archive larger collections of diffraction datasets, some DA satellites will elect to replicate a larger fraction of SBDG archives and make them available to the general community.



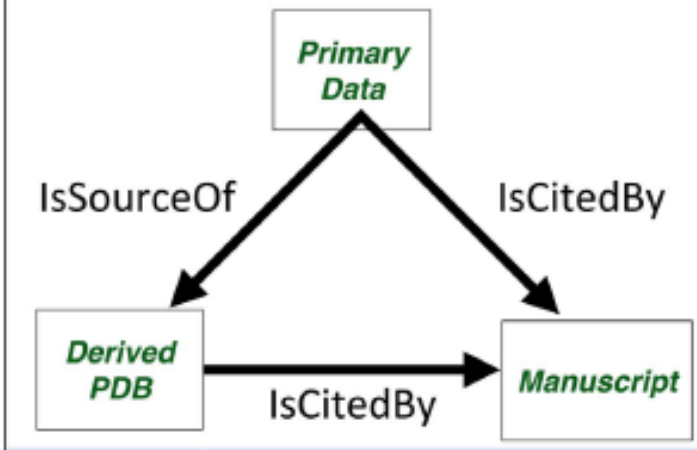
Appendix Figure A.5. Experimental data flow and publication. (a) Flow of Primary Experimental Data. Data sets collected at synchrotrons are moved to end-users' computers for processing and structure determination. Subsequently refined macromolecular models are deposited at PDB and primary data is uploaded to SBDG. From SBDG, data sets are replicated to DAA centres and eventually copied to DAA Satellites. End-users can access data sets by downloading from DAA centres and by direct access from Satellites. (b) Flowchart for data publication.

While the DA offers a variety of data access options that will support growth of the repository, members of the community can also download individual datasets directly from SBGrid servers at Harvard using an rsync protocol. Instructions for downloading individual datasets are provided on the Dataset View Pages, and effectively all datasets can be downloaded using the following command: “rsync -av rsync://data.sbgrid.org/DOI .”, where DOI is the digital object identifier for a particular dataset. The rsync utility, which is native to Linux and OS X systems, is particularly suitable for downloading large data files and can be restarted in case of interruption. After download, the data integrity of individual datasets can be verified by following instructions on the Data Grid website. With a well-defined and permissive CC0 access license and multiple channels for accessing data (four DA sites and the rsync download mechanism) our initial infrastructure is well suited to support expansion of the data collection.

Data Publication Cycle

For many SBGrid laboratories, interest in data deposition is driven by a desire to better organize research data and comply with institutional, federal, and project-specific data preservation requirements. During the pilot phase, data deposition privileges were limited to SBGrid-member laboratories. With recent funding to further support the project, the Data Grid is now open to the entire structural biology community. Non-SBGrid groups would first need to register with the SBDG to obtain proper deposition credentials.

Wide adoption of data-preservation systems is often hindered by the complexities involved in the data-deposition process itself. To mitigate this problem, SBDG deposition involves two simple steps: registration and uploading (Appendix Figure A.5B). To register a dataset, the depositor completes a web form with basic information about the sample, data-collection facility, related objects (e.g. publication, PDB code), and authorship; this information is mapped to the DataCite schema (Appendix Figure A.6). Many details necessary for dataset reprocessing – beam center, distance, wavelength, etc. – are automatically included with most datasets in the form of an image header generated by the data-collection software at the time of collection, simplifying the registration process. A principal investigator is authorized to sponsor depositions as a recognized member of the community and must approve each deposit. This system allows maximum flexibility when accepting data for deposition, facilitating the upload of complex datasets that otherwise could be challenging to validate. Following registration, a DOI is reserved for the dataset and the user is provided with data transfer instructions. Data deposition is handled by an automated script provided by SBDG and run on the depositor's computer, which uploads the data and checks for data integrity after upload. Upon verification, the primary data are either released in the bi-weekly SBDG release or placed on hold. As with the PDB, release of data placed on hold will coincide with publication.



Location	URL for the dataset web page
Creator	Group Name, e.g. Harrison Laboratory
Title	Sample name
Publisher	Structural Biology Data Grid
Publication Year	Year of dataset release
Resource Type	Diffraction Dataset
Abstract	short description, e.g. Native Dataset
Methods	Any additional information required to process.
Subject	Protein, RNA, DNA and/or ligand.
Contributor	Co-author of the data publication
Contributor's attributes	Additional roles, such as data depositor, project leader or data collector.
Date/Collected	Date of data collection (usually in the header)
Date/Submitted	Date of submission
Date/Available	Data of release
PDB	ID of the derived PDB file
Publication	DOI of the manuscript
Related Datasets	DOIs of other datasets supporting the PDB
Rights	CC0 License
Geolocation	Data collection facility (synchrotron beamline)

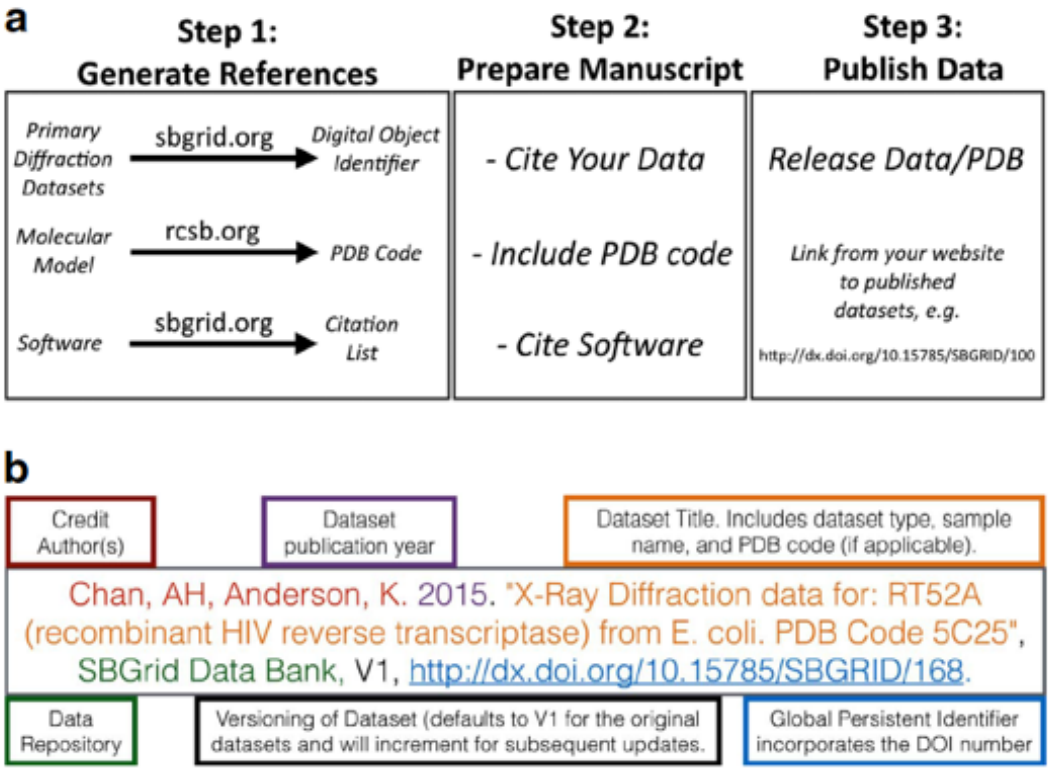
Appendix Figure A.6. DataCite metadata schema used for primary data sets within the SBDG. Information associated with the DOI record for a primary data set through the EZID system.

The two-step publication process is complemented by behind-the-scenes data replication, DOI registrations, and data analysis. All X-ray diffraction images are currently post-processed using data processing pipelines that provide a post-publication data review that will be shared with depositors and the community in the next phase of the SBDG project. We are building additional tools to help increase data-deposition rates, including automatic reminders sent to consortium members to encourage them to deposit data for previously published work.

Data Citation

Research data are the legitimate and citable product of research (Bourne et al., 2012; Martone, 2014) and, therefore, the SBDG recommends that depositors and data users cite all data deposited with the SBDG in the standard reference section of their manuscripts following well established community standards (Altman and Crosas, 2013; Altman and King, 2007; Martone, 2014). Data citation examples are provided on individual dataset pages (Appendix Figure A.4). The SBDG complements our AppCiter application (Socias et al., 2015), which facilitates citation of research software. Both services are now presented to users in a unified publication support workflow (Appendix Figure A.7A). In step 1, the user deposits research-related data that are put on hold until publication. A set of DOIs and corresponding data citations are then generated and provided to the end-user. Users can also use AppCiter to generate a list of software citations for all scientific software used in the project. In step 2, all research data and scientific software citations are included in the References section of the manuscript. In step 3 the user, anticipating manuscript publication, contacts relevant databases to

request release of the primary and supporting data. This process should, ideally, take place prior to manuscript publication and be timed to coincide with the publication date, allowing the community to access the data when the manuscript is released. When preparing future publications that refer to completed structures, scientists should reference the relevant publications and macromolecular models, unless they are referring to a specific dataset. For specific datasets, authors should explicitly reference experimental data using the corresponding data citation (Appendix Figure A.7B). Citation metrics for published datasets will be comparable to those obtained for journal publications.



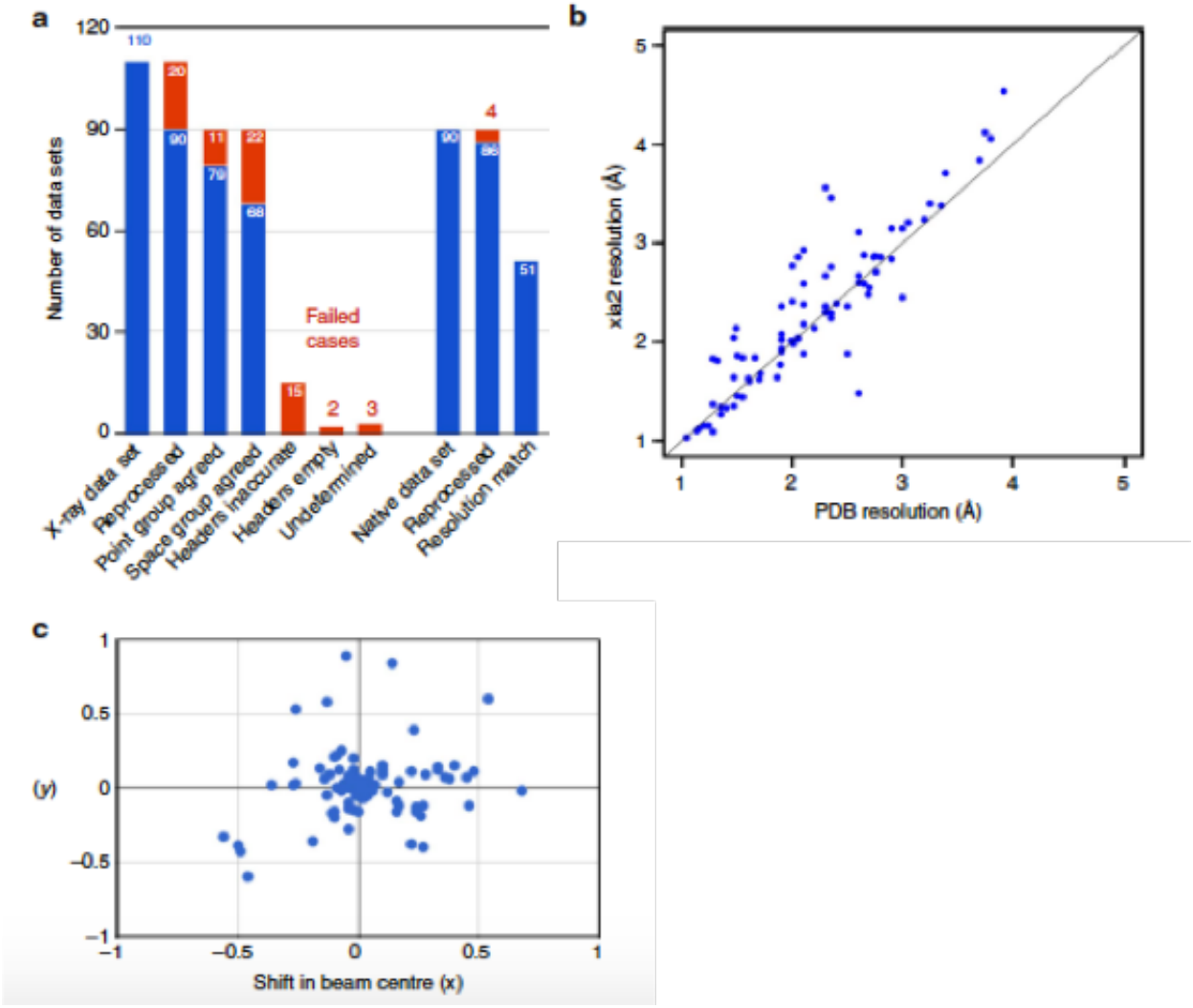
Appendix Figure A.7. Data publication guidelines. (a) Flowchart illustrating publication guidelines incorporating software and data citations. (b) Data Citation guidelines, adapted from Dataverse Best Practices Guidelines that were developed based on Force 11 Joint Declaration of Data Citation Principles.

Data Grid Content

Ease of data deposition and community-wide interest facilitated growth of the initial collection of X-ray diffraction datasets when it opened to the SBGrid community in May 2015. The datasets deposited during the pilot collection phase represent a wide cross-section of structures and a diverse subset of journal articles and structure-determination methods. For example, 68 structures derived from data deposited in the SBDG have been determined by molecular replacement (MR), while 4 have been solved by Multiple-wavelength Anomalous Diffraction (MAD), 4 by Single Isomorphous Replacement with Anomalous Scattering (SIRAS) and 15 by Single-wavelength Anomalous Diffraction (SAD) (Hunter and Westover, 2015b). The highest resolution dataset extended to 1.04 Å, and the lowest resolution dataset (Gilman and McLellan, 2015) to 5.5 Å. The structures ranged in molecular weight from 8.1 kDa (Feldkamp and Chazin, 2015) to 426 kDa (Tolia, 2015). The solvent content of these structures ranged from 32% (Hunter and Westover, 2015a) to 85% (Corbett and Harrison, 2015) and the longest unit cell edge was reported to be 525.29 Å (Gajadeera and Tsodikov, 2015).

For a proof of concept, released datasets in the SBGridDB were reprocessed with *xia2* (Evans, 2006; Evans and Murshudov, 2013; Kabsch, 2010; Leslie, 1999; Waterman et al., 2013; Winn et al., 2011; Winter et al., 2013) in a fully automated manner (Appendix Figure A.8A). 90 of the 110 released datasets with a corresponding PDB ID were successfully reprocessed. 86 of those 90 datasets represented high-resolution, native data and for 51 of those *xia2* decision making determined a high resolution limit within 0.1 Å of the published structure (Appendix Figure A.8B). The point group determined by reprocessing agreed with that of the published structure in 79

cases; for 65 of these the space groups agreed. The lower degree of recovery of space groups, in comparison to point groups, is attributed to ambiguity in screw axis determination at this stage of data processing. To provide insight into the most common failure modes, datasets for which *xia2* did not produce a set of integrated intensities were investigated using iMOSFLM (Battye et al., 2011). Twelve of the failure cases could be attributed to absent or inaccurate information in the image headers: while accuracy of the beam center annotation varied within the pilot collection (**Appendix** Figure A.8C), ten datasets had visually incorrect beam center information, two had missing header information. The cause of failure for the eight remaining datasets was not definitively determined from the datasets alone; however, consulting the reprocessing instructions provided by depositors clarified this for five of these datasets. The reprocessing instructions also suggested that many of the datasets for which *xia2* was able to produce integrated intensities, but with resolution or symmetry disagreeing with the deposited structure, could be attributed to incorrect header information. One outlying reprocessing case for which a significantly higher resolution was determined than originally reported was also investigated. For this case, one of four reprocessing attempts for the dataset reported a resolution higher than that supported by merging statistics. This discrepancy was resolved by a software update.



Appendix Figure A.8. Reprocessing of X-ray diffraction data sets. (a) Analysis of 110 X-ray diffraction data sets that supported previously published PDB coordinates. Most of the failures (represented in red) were due to inaccurate or incomplete image-header information. In several of these cases, depositors provided annotations correcting this information; (b) Comparison of resolution determined by automated xia2 reprocessing with published resolution. Includes data sets not used for final refinement of published structures; (c) Shift in direct beam position from image headers and refined value following successful reprocessing with xia2.

In addition to estimates of the Bragg intensities, diffraction images can also be analyzed for additional features (Helliwell and Mitchell, 2015). A well-known example is the isotropic solvent ring that generally appears around ~3-4 Å resolution (Welberry, 2004). However, diffraction images also contain anisotropic diffuse scattering signals under and between the Bragg peaks that derive from two-point correlations of electron density fluctuations (Wall et al., 2014a). Analysis of this diffuse scattering could therefore provide information about protein, nucleic acid, and lipid structural dynamics and correlated motions, potentially leading to new mechanistic insights (Wall et al., 1997) or to validating sampling schemes and energy functions for molecular dynamics simulations (Wall et al., 2014b). One dataset on the model enzyme Cyclophilin A is currently deposited (Appendix Table A.III) to be used as "gold-standard" to compare the influence of temperature on data collection (Wall, 2009) and to assess consistency between XFEL and synchrotron data (Fraser, 2015). This dataset can now also be analyzed for diffuse-scattering features, which could distinguish between models of correlated motion suggested by NMR experiments.

X-ray Diffraction Reference Subset and Other Collections

To take advantage of Data Grid diversity, we have selected a small subset of cases that could be used to support software development and teaching of data processing and diverse structure determination techniques (Appendix Table A.II). This subset includes high-resolution (1.2 Å), low-resolution (4.5 and 7.0 Å), anisotropic and twinned datasets. Additionally datasets that supported a variety of experimental phasing approaches (e.g. phasing with selenium, zinc, uranium, barium/potassium) and

molecular replacement cases (eg. with a 9 Å EM envelope) are included. The subset also incorporates diffraction data for crystals grown in lipidic cubic phase and an example of multi-crystal averaging.

Additionally SBDG is suited to support various other primary data types that are being generated by members of the consortium, and those pilot collections will seed development of community-wide data analysis systems. MicroED is a promising new technique (Nannenga et al., 2014; Shi et al., 2013) and inclusions of the early microcrystal datasets might stimulate the community to explore this technique and to fine-tune data processing software. Examples of MicroED datasets that are included in the pilot collection include three MicroED datasets that were used to determine structures of the toxic core of α-synuclein (Reyes et al., 2015), catalyse (de la Cruz et al., 2015) and lysozyme (Shi and Gonen, 2015). Other types of datasets in our pilot collection include a 55 GB computational decoy dataset for 55 complexes with associated HADDOCK scores (Vangone and Bonvin, 2015), a 2 μs Desmond (Bowers et al., 2006) MD trajectory (Sliz, 2015), and a recently collected Lattice Light Sheet Microscopy (Chen et al., 2014; Kural et al., 2015) dataset with in-vivo imaging of zebrafish embryos (Upadhyayula and Kirchhausen, 2015). Here the engagement with domain experts and respective communities will be also required to establish data validation pipelines and effective DA distribution models.

Discussion

We have developed a flexible data publication system, the Structural Biology Data Grid, to support deposition of a variety of large primary datasets. The data repository complements the wwPDB efforts by preserving the raw data that supports

PDB-deposited structure models. The pilot phase of the project, which was limited to SBGrid laboratories, demonstrated both feasibility and strong participation, with the deposition and publication of 117 datasets (as of September 1st, 2015, collected over 3 months). To support annotated data-collection, we have established data processing pipelines that will evolve the post-deposition data-analysis process. For example, the pipeline presented in the results section allows depositors and SBDG curators to quickly identify image-header problems, and parameters that are refined or corrected will be included in the expanded Dataverse schema (Crosas, 2011, 2013; Crosas et al., 2015; King, 2007). The outliers and failures of the current reprocessing pipeline illustrate areas of potential improvement to metadata accuracy and the pipeline itself. Data depositors and other community members will be able to provide data annotations to assist with the convergence of this process. Access to this growing collection of X-ray diffraction datasets will support the proposed paradigm shift in the community (Terwilliger and Bricogne, 2014) from the static archive towards a much more dynamic body of continuously improving refined models.

Despite being in the age of “big data science”, universal storage of large, biomedical datasets is an issue that has not yet been resolved, as infrastructure and support responsibilities have not been well defined. Shifting the burdens of data management from individual research groups and institutions to global infrastructures is an effective and economical strategy to address this issue that has previously been proven successful by the wwPDB and would now be demonstrated by the SBDG. By virtue of the consortium’s global presence, SBDG is well positioned to stimulate community-wide participation. SBGrid may facilitate integration of the Data Grid with

regional projects and facility-related efforts to preserve primary diffraction datasets. This data distribution model is similar to those established in other fields. For example, the Data Preservation Alliance (www.data-pass.org) replicates and indexes quantitative data for the social sciences. Data collected at the Large Hadron Collider are made available under a multi-tier processing and storage framework. As a large international consortium backed by diverse funding mechanisms and DA storage contributions of its members, SBGrid is uniquely capable of bypassing grant limitations that would otherwise deter such a long-term global infrastructure effort. Given recently secured funding to support data curation and technology integration under the Dataverse research data management, and with gradual community investment, SBDG is poised to scale up to support the entire community.

While access to experimental data is critical to ensuring research reproducibility, metadata quality is also crucial. Datasets that are poorly annotated have limited use to the research community. With a focus on deployment of a sustainable and flexible data management infrastructure, the SBDG takes a unique approach on metadata preservation. The repository employs an accommodating DataCite schema, which preserves basic information about experiments. The depositions are self-moderated by contributing laboratories, with data publication subject to approval of the principal investigators. As our results demonstrated, this approach worked well for the vast majority of datasets deposited in the SBDG, 82% of which were automatically reprocessed with current data processing software and the majority of the remaining datasets could be easily reprocessed manually. This success rate for reprocessing diffraction datasets was achieved without any explicit quality control to ensure that the

datasets contained sufficient information for reprocessing – in other words, using image headers as the only source of experimental (geometry and detector) parameters. Two possibilities under consideration for maintaining and improving this success rate are allowing depositors to annotate updated experimental parameters (for example, beam center) and explicit checks for metadata required for reprocessing prior to data publication. To facilitate interoperability with other projects and further stimulate uniform data evaluations, we will work in parallel to develop tools that will support download of archived datasets in community accepted master formats supporting intrinsic metadata, such as OME-TIFF or HDF5. This process will allow annotation of downloadable datasets with additional information from analysis pipelines, and will be guided by feedback from projects that interface with SBDG. Ideally, publication of datasets will encourage the communities to adopt standardized formats and ensure complete population of experimental metadata with adequate accuracy to support reprocessing.

While the SBDG immediately serves the well-defined area of X-ray crystallography, our pilot project has demonstrated that our infrastructure can preserve additional data types, such as decoy datasets for NMR computations or MicroED datasets. SBDG will duplicate XFEL datasets that are currently accessible through the Coherent X-ray Imaging Data Bank (http://www.cxidb.org/) and support their distribution by DA. In addition, SBDG will collaborate with MicroED and XFEL collection curators who will moderate development of community driven efforts to automate data analysis pipelines to parallel automatic processing of X-ray diffraction datasets with packages like DIALS or *xia2*. We envision that the tools and technologies that arise from this project will ultimately lead to the development of a fully featured, primary data

publication system. Features of such a system would include the capability of supporting a variety of experimental data types and automatic incorporation of pertinent dataset information during data collection at local, regional and national facilities. The integration of primary data management with a base set of scientific software enables repositories to progress towards dynamically improving sources of knowledge, as well as providing an integrated computing environment for ongoing research.

In summary, we have presented the Structural Biology Data Grid, a new system for the preservation and publication of large experimental datasets. The system is the latest product of SBGrid's mission to maintain a community-wide research-software infrastructure. Through disclosure, adoption, transparency, management of external dependencies, permissible licensing, and technical protection mechanisms, the SBDG is committed to compliance with evolving community standards of data preservation. We expect that the widespread sharing of experimental data will support methods development and will ultimately lead to better quality of structural models that are subject to continuous methods improvement.

Experimental Procedures

Current Implementation

The databank deposition process involves five stages: 1) recording associated metadata, 2) local checksum calculation, 3) data transfer, 4) post-transfer verification, and 5) public identifier registration.

A publicly accessible web frontend is used for handling user interactions with the databank. Built using the Python-based Django web framework, this frontend runs on an

Ubuntu 14 LTS Server with a PostgreSQL 9.3 database. It collects the necessary metadata during deposition and informs the backend systems about deposition requests. A cryptographic checksum (SHA1 SUM, FIPS 180-4) is calculated prior to data transfer. This ensures that the dataset is unchanged. Data transfer is handled by rsync over ssh. Once data transfer is complete, the databank verifies that the dataset has been transferred uncorrupted, or reports a problem with the dataset. If necessary, extraneous files (intermediate data files, processing or transfer scripts) are removed, data files are uncompressed and checksums re-computed. In the event of any modifications to the dataset, an unmodified copy is stored in an offline file system. Upon dataset release, the DOI reserved during dataset registration is registered using the recorded metadata, and the dataset (including checksum information) is made available for download over anonymous rsync.

Metadata Schema

DOIs are issued through EZID, through the Harvard University Library and the California Digital Library. Metadata are organized following the DataCite schema (Appendix Figure A. 6).

Reprocessing Details

Datasets that had been publicly released by September 1st, 2015 were reprocessed by *xia2* in a fully automated manner. For each dataset, four attempts were made to reprocess using options "-2d", "-3d", "-3dii" and "-dials", using MOSFLM, XDS, XDS (indexing with peaks from all images) and DIALS, respectively. AIMLESS (Evans

and Murshudov, 2013) and POINTLESS (Evans, 2006) were used by xia2 for spacegroup determination. A dataset was considered successfully reprocessed if any of these attempts succeeded, and comparisons to the originally published structure were done with the best matching result. Investigation of unsuccessfully reprocessed datasets was performed using iMOSFLM (Leslie, 1999). This investigation was performed “blinded” to the reprocessing instructions provided by depositors, in order to better investigate the limits of relying solely on diffraction images.

Data Alliance

Released datasets are distributed to Data Alliance mirror sites using the same mechanism as individual dataset distribution. Dataset checksums enable accurate data transfer. Users can select a mirror site by picking an appropriate rsync URL for data download.

Author Contributions:

All authors contributed to the current study, including intellectual input and editing of the manuscript. P.A.M., S.S. and P.S. developed the data grid system and A.B., M.L., C.B., J.W., D.N., K.R., J.K., F.R.N.C.M., I.F., MC and P.S. implemented the Data Access Alliance infrastructure. P.A.M, K.D. and P.S. analysed the data. K.S.A, R.H.B., S.C.B., T.J.B., D.B., T.J.B., A.C., C.I.C., W.J.C., K.D.C., M.S.C., S.C., S.D.P., E.D.C., C.L.D., M.J.E., B.F.E., Q.R.F., A.R.F., J.S.F., J.C.F., K.C.G., R.G., P.G., S.C.H., E.E.H., Z.J., R.J.K., A.C.K., M.K., J.S.M., Y.M., Y.N., Z.O., E.F.P., P.J.B.P., C.P., C.S.R., T.A.R., A.R., M.K.R., G.R., J.S., T.U.S., Y.S., H.S., Y.J.T., N.H.T., O.V.T., K.D.W., H.W.

and P.S. contributed X-ray diffraction data sets. A.M.J.J.B contributed the HADDOCK docking decoys data set. T.G., T.K., and P.S. contributed MicroED, Lattice Light-Sheet Microscopy and Molecular Dynamics data sets, respectively. P.A.M., E.R and P.S. Analysed the data and wrote the paper.

Acknowledgements:

Development of the Structural Biology Data Grid is funded by The Leona M. and Harry B. Helmsley Charitable Trust 2016PG-BRI002 to PS and MC. Development of citation workflows is supported NSF 1448069 (to PS). DA is being developed as a pilot project of the National Data Service, with additional funds to support storage and technology development, including NIH P41 GM103403 (NE-CAT) and 1S10RR028832 (HMS) and DOE DE-AC02-06CH11357; NIH 1U54EB020406-01, Big Data for Discovery Science Center; and NIST 60NANB15D077 (Globus Project). AB acknowledges Ariel Chaparro for assistance with the DA setup (Inst Pasteur Montevideo). Collections of pilot datasets were supported by various grants (see Appendix Table A.II).

Appendix B

LIN28 zinc knuckles domain is required and sufficient to induce let-7 oligouridylation.

Contributors: Longfei Wang*, Yunsun Nam*, Anna K. Lee, Chunxiao Yu, Kira Roth, Cassandra Chen, Elizabeth M. Ransey, Piotr Sliz

*Denotes equal contribution

Supplemental materials for this Appendix are at the end of this Appendix.

This Appendix originally appeared in *Cell Reports* (18), 2017.

https://doi.org/10.1016/j.celrep.2017.02.044

Abstract

LIN28 is an RNA binding protein that plays crucial roles in pluripotency, glucose metabolism, tissue regeneration, and tumorigenesis. LIN28 binds to the let-7 primary and precursor microRNAs through bipartite recognition and induces degradation of let-7 precursors (pre-let-7) by promoting oligouridylation by terminal uridylyltransferases (TUTases). Here, we report that the zinc knuckle domain (ZKD) of mouse LIN28 recruits TUT4 to initiate the oligouridylation of let-7 precursors. Our crystal structure of human LIN28 in complex with a fragment of pre-let-7f-1 determined to 2.0 A˚ resolution shows that the interaction between ZKD and RNA is con-strained to a small cavity with a high druggability score. We demonstrate that the specific interaction between ZKD and pre-let-7 is necessary and sufficient to induce oligouridylation by recruiting the N-terminal fragment of TUT4 (NTUT4) and the formation of a stable ZKD:NTUT4:pre-let-7 ternary complex is crucial for the acquired processivity of TUT4.

Introduction

LIN28 was first discovered as a heterochronic gene in C. elegans (Moss et al., 1997) and was later shown to encode an RNA binding protein that suppresses the let-7 family of microRNAs in mice and humans (Heo et al., 2008; Newman et al., 2008; Rybak et al., 2008; Viswanathan et al., 2008; Lehrbach et al., 2009). Because the let-7 family negatively regulates oncogenes such as MYC, KRAS, and HMGA2 (Johnson et al., 2005; Mayr et al., 2007; Sampson et

al., 2007), LIN28 expression is closely associated with numerous cancers, including ovarian and colon cancers (Peng et al., 2010; Permuth-Wey et al., 2011; King et al., 2011). Human LIN28A and LIN28B are upregulated in a spectrum of tumors (~15%), and the elevated expression of either protein often results in increased cancer aggression and poor prognoses (Viswanathan et al., 2009). Transgenic overexpression of LIN28 in mice has been shown to lead to T cell lymphoma, neuroblastoma, and intestinal adenocarcinoma (Tu et al., 2015; Beachy et al., 2012; Molenaar et al., 2012). Studies have revealed that in addition to driving cell transformation, LIN28 is required for the maintenance of Wilms' tumor and hepatocellular carcinoma in mice (Urbach et al., 2014; Nguyen et al., 2014). This wealth of in vitro and in vivo data make LIN28A and LIN28B attractive therapeutic targets in LIN28-dependent tumors.

LIN28 functions as a let-7 suppressor by binding to the terminal loop region (let-7 pre-element, also known as the preE-let-7) of both primary let-7 (pri-let-7) and intermediate let-7 precursor (pre-let-7) transcripts; thus, it blocks cleavage by Microprocessor and Dicer, respectively (Heo et al., 2008; Viswanathan et al., 2008; Rybak et al., 2008), and effectively prevents let-7 maturation. In our previously reported high-resolution crystal structures (Nam et al., 2011), the RNA binding domains of LIN28, the cold shock domain (CSD) and the zinc knuckles domain (ZKD, or the Cys-Cys-His-Cys [CCHC]-type CCHCx2 domain), engage two distinct regions of precursor RNAs of various preE-let-7. The CSD binds to a stem-loop structure in the pre-element (preE), while the ZKD interacts with a highly conserved GGAG motif near the 30 end of the preE. The

linker connecting the two folded domains is flexible, thereby accommodating LIN28 binding to diverse let-7 family members (Nam et al., 2011). Evaluation of LIN28 truncation constructs demonstrated that the CSD has a higher affinity for let-7 than the ZKD and that binding of the CSD induces a conformational change in the terminal loop of pre-let-7 (Mayr et al., 2012). In the presence of RNA, the ZKD folds into a unique conformation that has been captured in both crystal and nuclear magnetic resonance spectroscopy (NMR) structures (Nam et al., 2011; Loughlin et al., 2011) and selectively recognizes the let-7 30 GGAG motif. The only member of the let-7 family that escapes this pathway is the human let-7-a3 (Triboulet et al., 2015).

In addition to binding pre-let-7 to inhibit processing events, LIN28 promotes degradation of pre-let-7 in the cytoplasm through the recruitment of terminal uridylyltransferases (TUTases) (Heo et al., 2008, 2009; Hagan et al., 2009). As the first step of the degradation pathway, TUTases processively extend the 30 ends of pre-let-7 with uridines. Subsequently, the 30-50 exonuclease, Dis3l2, recognizes and degrades oligouridylated pre-let-7 (Chang et al., 2013). Seven TUTases have been identified in humans, but only TUT4 (ZCCHC11) and TUT7 (ZCCHC6) have been shown to actively oligouridylate pre-let-7 (Thornton et al., 2012). TUT4 and TUT7 are the largest TUTases (~180 kDa compared to the typical ~50 kDa) and contain an extended N-terminal region with additional domains, including a C2H2 zinc finger, an inactive nucleotidyltransferase domain, and a poly(A) polymerase (PAP)-associated domain. The N-terminal region of TUT4 or TUT7 has no catalytic activity, but it is

required for LIN28-mediated oligouridylation (Thornton et al., 2012). Although it has been shown by single-molecule immunoprecipitation that a 3% fraction of TUT4 extracted from human cells associates with the LIN28B:pre-let-7a-1 complex (Heo et al., 2009; Yeom et al., 2011), how LIN28 drives complex formation and the mechanism of TUTase recruitment remains unknown. In addition to oligouridylation, TUT4, TUT7, and TUT2 monouridylate the 30 ends of a group of micro-RNA precursors, including most let-7 microRNAs, known as group II microRNAs. This monouridylation, in contrast to the oligouridylation process, promotes the biogenesis of group II microRNAs by extending the 1 nt 30 overhang and enhancing Dicer activity (Heo et al., 2012).

LIN28 and TUTases have been implicated in mRNA processing, although this activity seems to occur independent of LIN28:TUTase interactions. Using cross-linking immunoprecipitation (CLIP) and next-generation sequencing methods, it has been demonstrated that LIN28A and LIN28B bind to a large fraction of the transcriptome (Hafner et al., 2013; Cho et al., 2012; Wilbert et al., 2012). Others have reported that TUT4 and TUT7 oligouridylate poly (A)-lacking mRNAs and enhance their decay in a LIN28-independent pathway (Lim et al., 2014). The exact interplay among LIN28, TUTase, and LIN28:TUTase-driven mechanisms is still under investigation.

Here we describe distinct roles for the LIN28 CSD and ZKD in the oligouridylation of pre-let-7 microRNAs. We show that LIN28 ZKD binding is sufficient to promote oligouridylation of pre-let-7 by TUT4 and mediates the complex formation between LIN28:pre-let-7 and TUT4 through direct interactions.

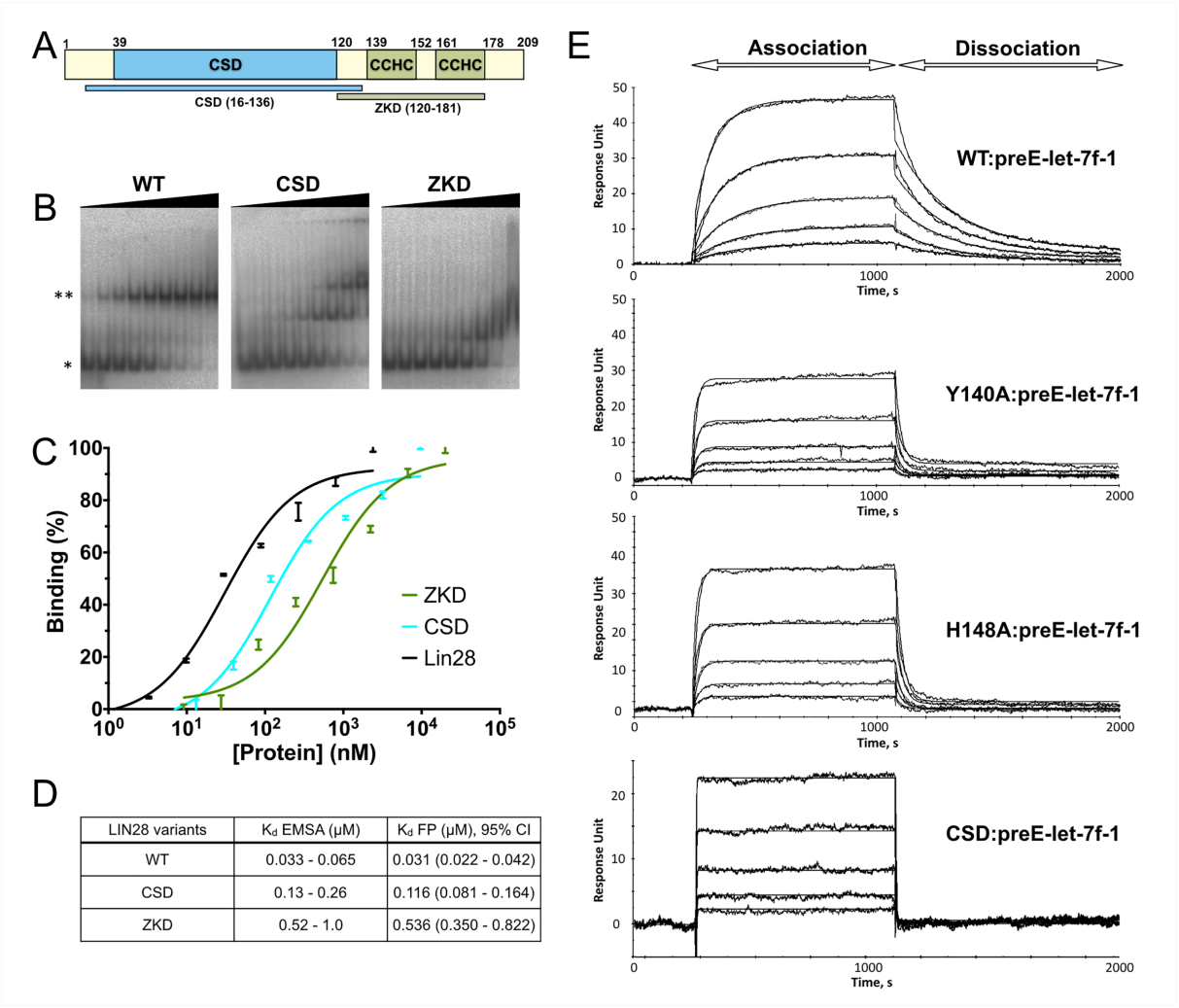
Our data highlight the crucial role of ZKD in LIN28-mediated inhibition of let-7 microRNAs. LIN28 ZKD stabilizes the LIN28:let-7 complex by preventing complex dissociation. The ZKD also directly recruits TUT4 and serves as a processivity factor for the oligouridylation activity of TUTase. By analyzing the crystal structure of human LIN28 in complex with preE-let-7f-1, we evaluated the feasibility of targeting ZKD with a small molecule.

RESULTS

The ZKD of LIN28 Is Critical to Reduce the Dissociation Rate of the LIN28:pre-let-7 Complex

We previously reported that mouse LIN28A binds to preE-let-7d (preE sequences of let-7d) with a dissociation constant (K_D) of 33–130 nM, while the individual CSD (residues 16–136) and the ZKD (residues 120–181) bind to the corresponding RNA fragment, with binding affinities of 130–520 and 520–2,100 nM, respectively (Nam et al., 2011). Now we extended this data to demonstrate that mouse LIN28A binds to a pre-let-7, pre-let-7g, with a K_D of 33–65 nM, while the individual CSD (residues 16–136) and the ZKD (residues 120–181) bind to the corresponding RNA fragment, with binding affinities of 130–260 and 520–1,004 nM, respectively (Appendix Figures B.1A, 1B, and 1D). To validate these results, we completed fluorescence polarization (FP) as-says using synthesized 30 fluorescein amidite (FAM)-labeled preE-let-7f-1 RNA oligonucleotides. We observe similar binding affinity patterns, with a KD of 31 nM for the wild-type

protein and 116 and 536 nM for the CSD and ZKD, respectively (Appendix Figures B.1C and 1D).



LIN28 variants	K_d EMSA (μM)	K_d FP (μM), 95% CI
WT	0.033 - 0.065	0.031 (0.022 - 0.042)
CSD	0.13 - 0.26	0.116 (0.081 - 0.164)
ZKD	0.52 - 1.0	0.536 (0.350 - 0.822)

Appendix Figure B.1. The ZKD of LIN28 Is Critical to Reduce the Dissociation Rate of the LIN28:pre-let-7 Complex. (A) Schematic representation of mouse LIN28A and truncations used for EMSA.(B) EMSAs with mouse pre-let-7g as probe, mixed with increasing concentrations (16, 33, 65, 130, 260, 520, 1,004, 2,100, and 4,200 nM) of LIN28, CSD, and ZKD. *Free probe; **complex. A second CSD molecule can bind to pre-let-7g at high CSD concentrations, which results in a 2:1 CSD:pre-let-7g complex. (C) FP assays with FAM-labeled mouse preE-let-7f-1 as probe, mixed with increasing concentrations of mouse LIN28A, CSD, and ZKD. (D) Comparison of the dissociation constants (KD) of mouse LIN28A, CSD, and ZKD for let-7 RNA from EMSA and FP assay (Figure S1). (E) Comparison of the dissociation rates between mouse LIN28A mutants using SPR assay, with preE-let-7f-1 immobilized to the sensor surface, followed by the injection of LIN28 mutants.

Because ZKD mutations can abrogate the ability of LIN28 to suppress let-7 maturation (Nam et al., 2011), we decided to investigate why the ZKD is critical for LIN28's function, despite its low affinity to let-7 sequences. To monitor the kinetics of LIN28-let-7 binding in real time, we developed a surface plasmon resonance (SPR) assay. In this assay, the mouse let-7f-1 pre-element (preE-let-7f-1) was immobilized on CM5 sensor chips and LIN28A was injected to the sensor surface. Association and dissociation were measured as response units (RUs) upon binding of LIN28 and its release in subsequent wash steps. For wild-type LIN28A, we calculated an association rate of 4.76×10^{4} ms^{-1} and a dissociation rate of 4.38×10^{-3} s^{-1} with a corresponding K_D of 92 nM (Appendix Figures B.1E; Appendix Table B.I), in agreement with electrophoretic mobility shift assay (EMSA) and FP results. To determine the contribution of the ZKD to the binding kinetics between LIN28A and let-7f-1, we first individually mutated two ZKD residues, Y140 and H148, which are key residues that directly interact with the GGAG motif of let-7 (Nam et al., 2011). Alanine substitutions of either Y140 or H148 in-crease the dissociation rate of LIN28A from let-7f-1 ~10-fold, indicating that the ZKD strongly contributes to complex stability (Appendix Figures B. 1E; Appendix Table B.I). The ZKD Y140A and H148A mutations also increase the association rate of LIN28A, which results in relatively moderate changes in KD (Appendix Figures B.1E; Appendix Table B.I). To complement these experiments, we evaluated the kinetics of LIN28A using a mutated preE-let-7f-1. We mutated the critical GGAG motif to a UGAG sequence of this preE and confirmed that changes in the dissociation and association rates have the

same effect as mutations of the ZKD RNA binding residues (Appendix Figure B.S1).

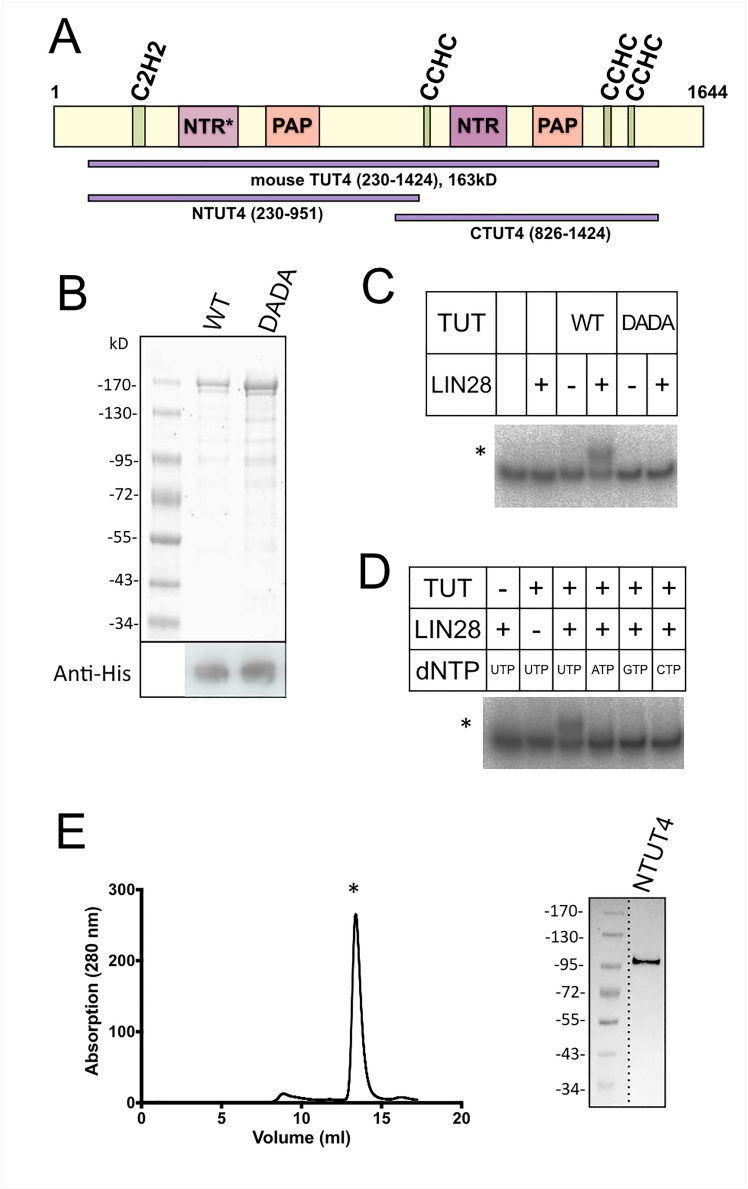
Appendix Table B.I. SPR Results of Mouse LIN28A Variants

Protein	k_{on} (1/ms)	k_{off} (1/s)	K_D (nM)	X^2
WT	4.76E+04	4.38E-03	92	0.145
H148A	1.31E+05	0.0363	276	0.194
H162A	9.23E+04	0.0353	382	0.218
Y140A	1.04E+05	0.0427	410	0.271
CSD	4.59E+05	0.37	808	0.147
WT/UGAG	6.05E+04	0.0234	387	0.131
E151K	4.89E+04	5.54E-03	113	0.086
I167A	5.28E+04	5.25E-03	99.5	0.148

We subsequently completed SPR experiments with an isolated CSD (LIN28A residues 16–136). In comparison to the full-length LIN28A containing either Y140A or H148A mutation, the dissociation rate for the isolated CSD is 10-fold higher. This result indicates that the ZKD contributes to complex stability even in the presence of mutations. When compared to wild-type LIN28A, the dissociation rate of the CSD alone is 100-fold higher, confirming that the ZKD plays a critical role in stabilizing the LIN28:microRNA complex.

Recombinant LIN28 and TUT4 Are Sufficient for Oligouridylation of pre-let-7 by TUT4

To address the role of the ZKD in recruiting TUT4 in the let-7 degradation pathway, we purified E. coli expressed recombinant wild-type and mutant mouse TUT4 (residues 230–1,424, 163 kDa) and established an oligouridylation assay using recombinant LIN28A, TUT4, and chemically synthesized pre-let-7g (Appendix Figures B.2A and 2B). Incubation of 50 end-labeled pre-let-7g with recombinant TUT4 and uridine-50-triphosphate (UTP) yields oligouridylated RNA (Appendix Figure B.2C). When the catalytic aspartate residues are mutated (D1026A and D1028A mutant, DADA), oligouridylation activity is no longer detectable under the same conditions. Purified TUT4 can only yield slower migrating species in the presence of UTP, but not with ATP, cytidine-50-triphosphate (CTP), or guanosine triphosphate (GTP) (Appendix Figure B.2D). Thus, recombinant TUT4 manifests LIN28-dependent and uridine-specific polymerase activity on pre-let-7g.



Appendix Figure B.2. Recombinant LIN28 and TUT4 Are Sufficient for Oligour-idylation of pre-let-7 by TUT4. (A) Schematic representation of mouse TUT4 and N-terminal TUT4 (NTUT4) constructs purified (Figures S3 and S4). (B) SDS-PAGE and western blot of purified mouse TUT4 and its catalytic dead construct DADA. Each lane of the SDS-PAGE and western blot shows the same sample but in different gels. (C) Oligouridylation assays with pre-let-7g as probe, carried out using re-combinant TUT4 and UTP (Figure S5). *Oligouridylated pre-let-7. (D) Oligouridylation assay with pre-let-7g as probe, carried out using recom-binant TUT4 and ribonucleotide triphosphate (ATP, GTP, CTP, and UTP). (E) Super elongation complex (SEC) chromatogram and SDS-PAGE of re-combinant N-terminal TUT4.

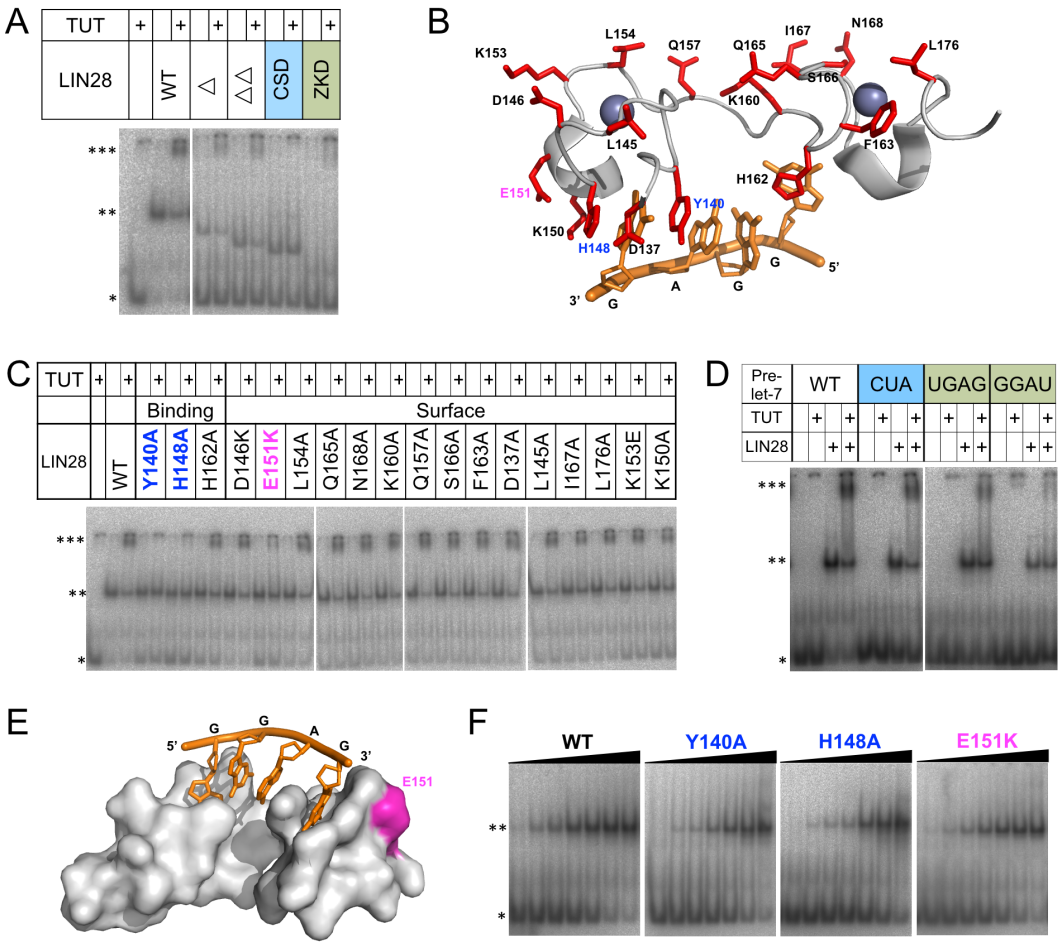
We generated truncation constructs of mouse TUT4, containing mostly the N-terminal (NTUT4) or the C-terminal region (amino acids 230–951 or 826–1,424, respectively). As shown in Appendix Figure B.2E, we found that recombinant NTUT4 is mono-disperse and pure, while the C-terminal construct is not fully homogeneous and barely yields enough isolated protein for binding experiments. These results demonstrate successful in vitro purification of recombinant mouse TUT4 and TUT4 truncation constructs with activity and high purity.

The LIN28 ZKD Mediates the Formation of the LIN28:pre-let-7:TUT4 Ternary Complex

The interaction between LIN28 and TUT4 has been reported to be transient (Yeom et al., 2011). To investigate the ability of TUT4 to interact with LIN28:pre-let-7 complexes, we performed EMSAs using recombinant mouse LIN28A and catalytically dead recombinant mouse TUT4, with radiolabeled pre-let-7g. As shown in Appendix Figure B.3A, a band corresponding to the ternary complex is observed as a super-shifted band when both full-length polypeptides are present. When flexible regions of LIN28A are removed, such as in LIN28D (residues 16–184, removed N and C termini) or LIN28DD (LIN28D with additional internal deletion of the interdomain linker, residues 116–135), LIN28 retains the ability to recruit TUTase. However, isolated CSD without ZKD is unable to form ternary complexes while isolated ZKD retains some ability to recruit TUT4. Thus,

in vitro binding studies suggest that the small ZKD (~7 kDa) is the region of LIN28 responsible for recruiting TUTase.

To further dissect the LIN28:pre-let-7:TUT4 interaction, we evaluated a series of ZKD mutations of LIN28A, including three known let-7 binding pocket residues and 15 surface-exposed residues located outside of the RNA binding interface of the ZKD (Appendix Figure B.3B). As shown in Appendix Figure B.3C, mutation of key residues that bind RNA, Y140A and H148A, abolishes complex formation, suggesting that forming a proper ZKD:GGAG complex is critical for TUT4 recruitment. Similarly, changing the sequence of the ZKD binding site from GGAG to GGAU or UGAG leads to a substantial loss of TUTase recruitment. In contrast, mutation of the CSD binding let-7 sequence (GAU to CUA) has no noticeable effect on complex formation (Appendix Figure B.3D). The other residue that affected complex formation, E151, is a surface-exposed residue (Appendix Figure B.3E). To confirm that E151K does not affect RNA binding, we completed EMSAs with pre-let-7g and SPR assays with preE-let-7g and observed no significant difference in affinity or kinetics (Appendix Figures B.3F and 3.S1; Appendix Table B.I). These results suggest that E151 residue in LIN28A is likely to contact TUT4 directly.

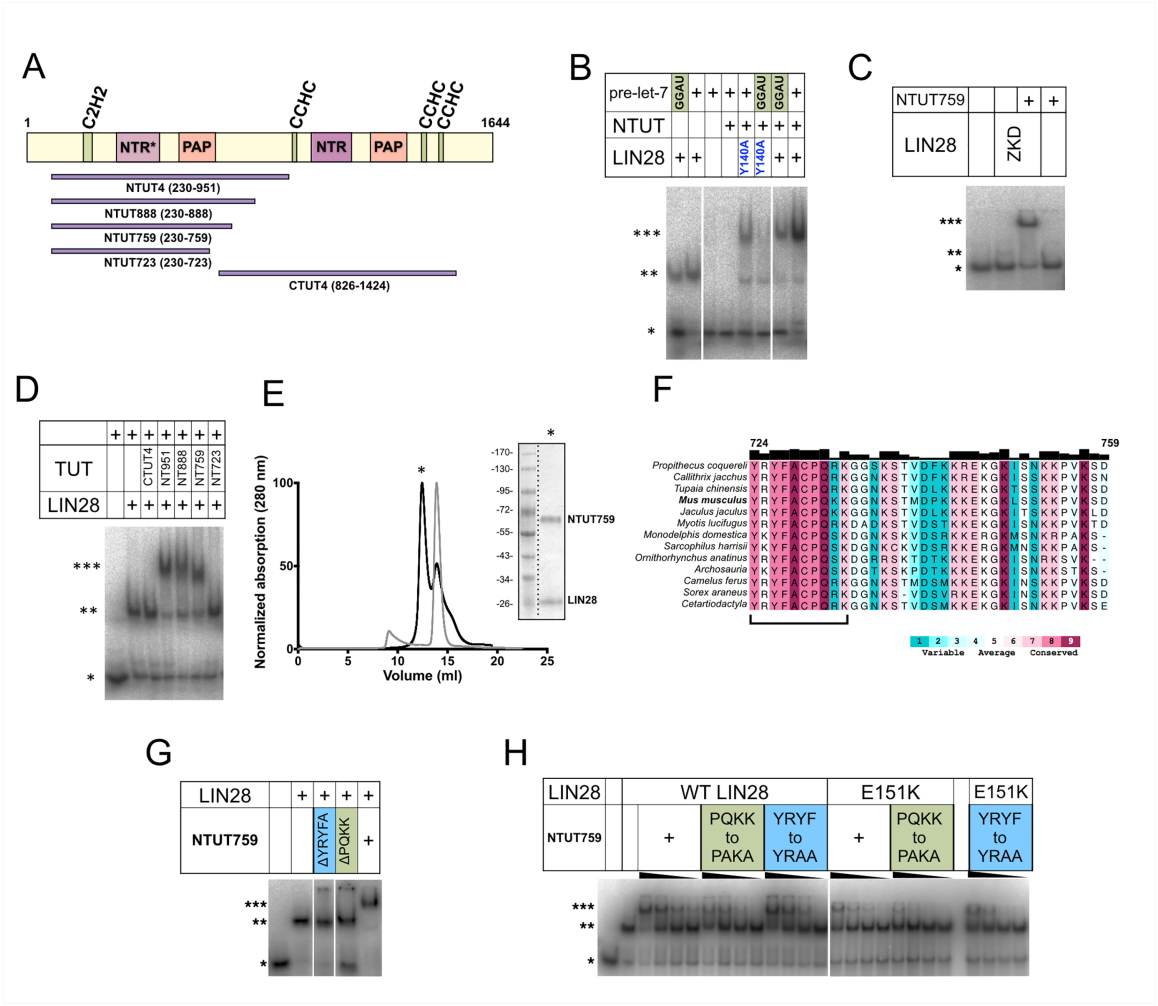


Appendix Figure B.3. The LIN28 ZKD Mediates the Formation of the LIN28:pre-let-7:TUT4 Ternary Complex. (A) EMSAs with pre-let-7g as probe, mixed with mouse LIN28A truncation constructs and TUT4. (B) View of ZKD interacts with GGAG RNA. GGAG RNA is in orange. The ZKD is represented by gray. The tested RNA binding and surface residues are in red. The model used is mouse LIN28A (PDB: 3TS2) (Nam et al., 2011). (C) EMSAs with pre-let-7g as probe, mixed with mouse LIN28A mutants and TUT4. (D) EMSAs with pre-let-7g mutants as probe, mixed with mouse LIN28A and TUT4. (E) Surface view of the LIN28A ZKD. The surface residues E151 is in magenta. The model used is mouse LIN28A (PDB: 3TS2) (Nam et al., 2011). (F) EMSAs with pre-let-7g as probe, mixed with increasing concentrations (16, 33, 65, 130, 260, 520, and 1,004 nM) of mouse LIN28A variants. *Free probe; **LIN28A:pre-let-7g complex; ***LIN28A:pre-let-7g:TUT4 complex.

The N-terminal Region of TUT4 Binds LIN28:pre-let-7 Complexes

To identify the TUT4 region that interacts with LIN28, we performed EMSAs with several mouse TUT4 constructs (Appendix Figure B.4A). The N-terminal TUT4 construct, NTUT4230–951, is sufficient to form a ternary complex and, similar to TUT4, is sensitive to GGAG to GGAU, as well as LIN28A Y140A mutations (Appendix Figure B.4B). The N-terminal NTUT4230–759 construct also forms a complex with pre-let-7g and the ZKD of LIN28A (Appendix Figure B.4C). The C-terminal TUT4, in contrast, does not interact with LIN28A:pre-let-7g (Appendix Figure B.4D).

We next sought to map the minimal region of mouse TUT4 that binds to the LIN28A:pre-let-7g complex. For this purpose, we designed three additional constructs with C-terminal truncations of NTUT4230–951 (NTUT951), NTUT723, NTUT759, and NTUT888 (Appendix Figure 4A)—and purified them using the NTUT4230–951 purification protocol. The shortest construct, NTUT723, does not form a stable ternary complex (Appendix Figure 4D). The LIN28A:pre-let-7g: NTUT759 complex was isolated using size exclusion chromatography (Appendix Figure B.4E). We reasoned that deleting the 36 residues that differentiate NTUT723 from NTUT759 caused the loss of NTUT723 binding to LIN28A:pre-let-7g.



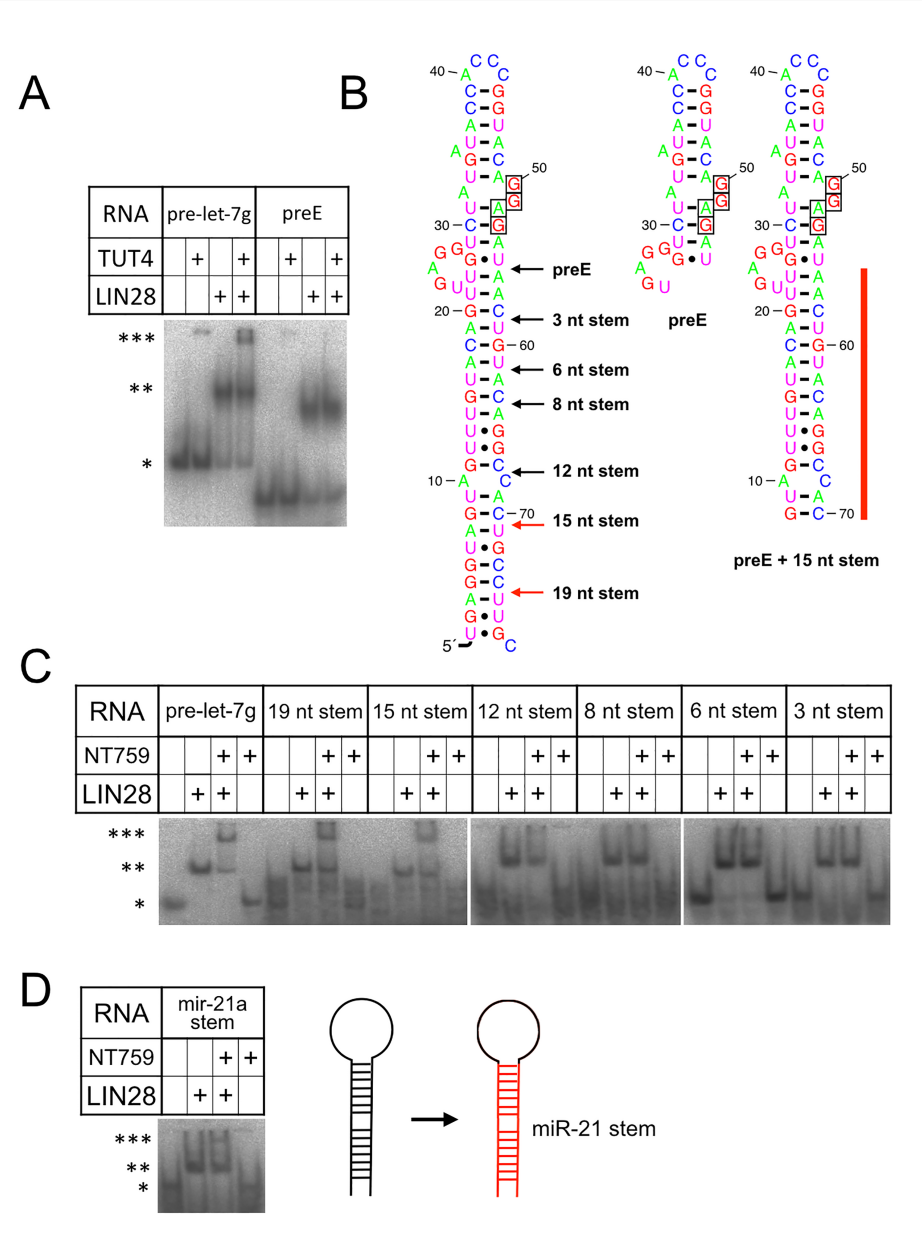
Appendix Figure B.4. The N-terminal Region of TUT4 Binds LIN28:pre-let-7 Complexes. (A) Schematic representation of N-terminal TUT4 and C-terminal TUT4 constructs used in EMSAs. (B) EMSAs with pre-let-7g as probe, mixed with mouse LIN28A and NTUT4. (C) EMSAs with pre-let-7g as probe, mixed with ZKD and NTUT759 truncations. (D) EMSAs with pre-let-7g as probe, mixed with mouse LIN28A, NTUT4 and CTUT4 truncations. (E) Size exclusion chromatograms of NTUT759 (gray) and LIN28A:NTUT759:pre-let-7g ternary complex (black). (F) Conservation analysis of the 36 residues of the C-termini of NTUT759. The analysis was performed using ConSurf. All residues are color coded based on conservation score. (G) EMSAs with pre-let-7g as probe, mixed with LIN28A and NTUT759 truncations. (H) EMSAs with pre-let-7g as probe, mixed with LIN28A surface mutants and NTUT759 mutants. Concentrations for each NTUT759 mutants are 3, 1, 0.33, and 0.11 mM. *Free probe; **LIN28A:pre-let-7g complex (B, D, G, and H) and ZKD:pre-let-7g complex (C); ***LIN28A:NTUT4:pre-let-7g ternary complex (B, D, G, and H) and ZKD:NTUT759:pre-let-7g ternary complex (C).

To further investigate the role of this region, we carried out a sequence conservation search using the ConSurf server (Glaser et al., 2003) and identified a highly conserved motif containing ten residues YRYFACPQKK (724–733) (Appendix Figure B.4F). Homology modeling of NTUT4 against the C. elegans ortholog of TUT4, Cid1, revealed that the conserved region corresponds to the loop region following the C-terminal a helix (hereafter referenced as motif M). To interrogate the functional relevance of the motif M, we introduced two NTUT759 deletion constructs, DYRYFA and DPQKK. In EMSAs, we observed that both deletions abolished the association between NTUT759 and LIN28A:pre-let-7g, indicating that this motif is required for ternary complex formation (Appendix Figure B.4G). More discrete NTUT759 mutations, M1 (YRYF to YRAA) and M2 (PQKK to PAKA) also resulted in reduced complex formation (Appendix Figure B.4H). Combining M1 or M2 mutations with the LIN28A E151K mutation demonstrated a synergistic effect and led to even further reduction in complex formation. These data reveal that the N-terminal fragment of TUTase is sufficient to drive formation of the ternary complex and that the ten-residue YRYFACPQKK motif plays a role in this interaction.

Assembly of the LIN28:pre-let-7:TUT4 Ternary Complex Requires the Stem Region of RNA

To investigate the role of pre-let-7 RNA in recruitment of TUTase, we tested the effects of truncated let-7g precursors. LIN28A, including ZKD, specifically binds the terminal loop region and does not require any of the stem

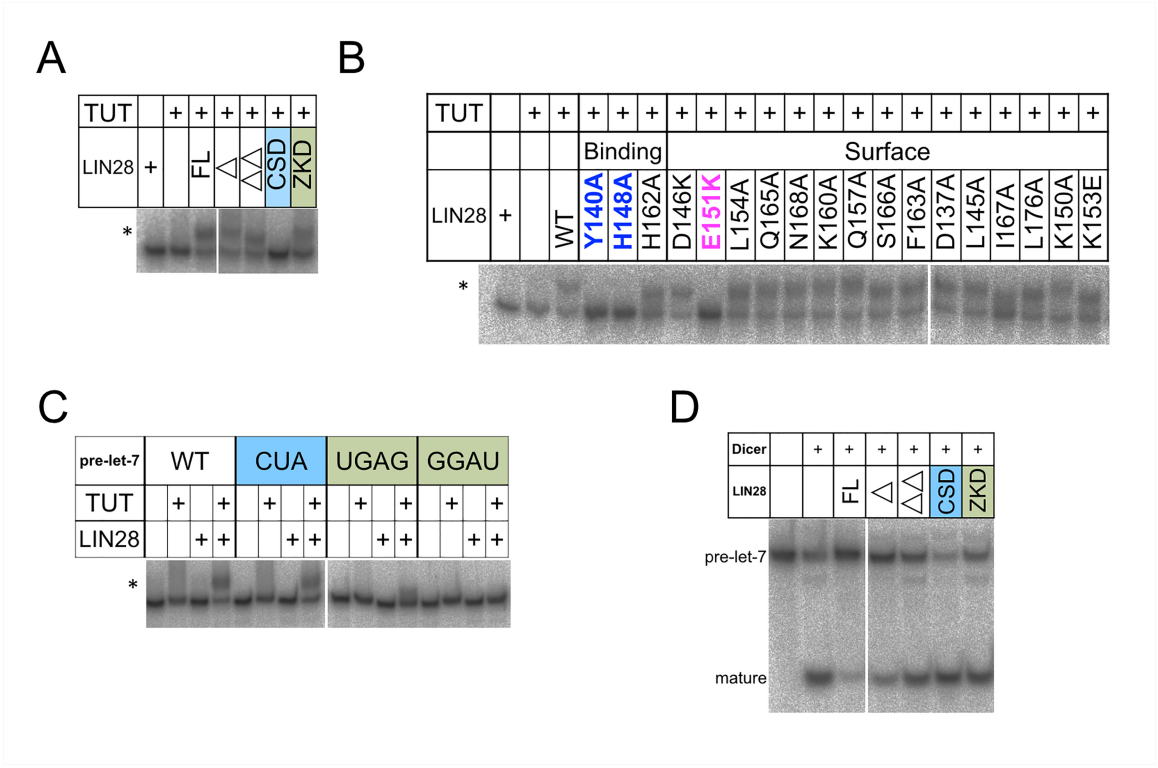
region to form stable complexes with pre-let-7g (Nam et al., 2011). However, isolated preE does not support recruitment of TUT4 (Appendix Figure B.5A). We performed a systematic truncation of pre-let-7g to shorten the stem length distal to the LIN28A binding site and found that double-stranded RNA (dsRNA) of about 15 nucleotides on each side is necessary to form a stable ternary complex (Figures B.5B and B.5C). When the stem region is swapped with that of another microRNA (miR-21), ternary complex can still assemble, though less efficiently (Appendix Figure B.5D). These data suggest that more than a full turn of dsRNA stem, in addition to LIN28 bound to preE, is required to recruit TUT4. A sufficient length of the dsRNA is critical, while the sequence content may vary.



Appendix Figure B.5. Assembly of the LIN28:pre-let-7: TUT4 Ternary Complex Requires the Stem Region of RNA. (A) EMSAs using pre-let-7g and preE-let-7g, mixed with mouse LIN28A and TUT4. (B) Schematic view of all pre-let-7 truncation constructs used in the EMSAs. GGAG sequences are boxed (Figure S6). (C) EMSAs with pre-let-7g truncations as probe, mixed with LIN28A and NTUT759. (D) EMSAs with pre-let-7g that has stem region of miR-21, mixed with LIN28A and NTUT759. *Free probe; **LIN28A:pre-let-7g and LIN28A: preE-let-7g complexes (A) and LIN28:pre-let-7g complex (C and D); ***LIN28A:pre-let-7g:TUT4 ternary complex (A) and LIN28:pre-let-7g:NTUT759 ternary complex (C and D).

The ZKD Plays Distinct Roles in Dicer and TUT4 Regulation

Our domain mapping results for assembly of the ternary complexes indicate that the ZKD of LIN28 may be sufficient to recruit TUT4 to pre-let-7. We used in vitro oligouridylation assay to test whether complex formation is sufficient to activate TUT4 activity. LIN28A missing flexible linker or terminal regions is capable of promoting oligouridylation of pre-let-7g (Appendix Figure B.6A). Moreover, isolated ZKD can activate TUT4 activity, while the CSD does not, consistent with their differential abilities to recruit TUT4. Mutations of the ZKD surface residues can also have a detrimental effect on oligouridylation of pre-let-7g (Appendix Figure B.6B). Mutations of ZKD that affect RNA binding (Y140A and H148A) lead to no detectable TUT4 activity, consistent with how they affect ternary complex formation. Another assembly mutation of the ZKD, E151K, also leads to loss of oligouridylation, reinforcing the correlation between the ternary complex assembly and the TUT4 activation. We identified an additional surface residue of ZKD, I167, that does not show significant defects in TUT4 recruitment but can decrease its polymerase processivity. Our results suggest that proper interaction between ZKD and preE of let-7 is required to recruit TUT4. In addition to bringing the three components together, oligouridylation may be sensitive to conformational changes such as those caused by certain surface residues of ZKD (I167 and K153).



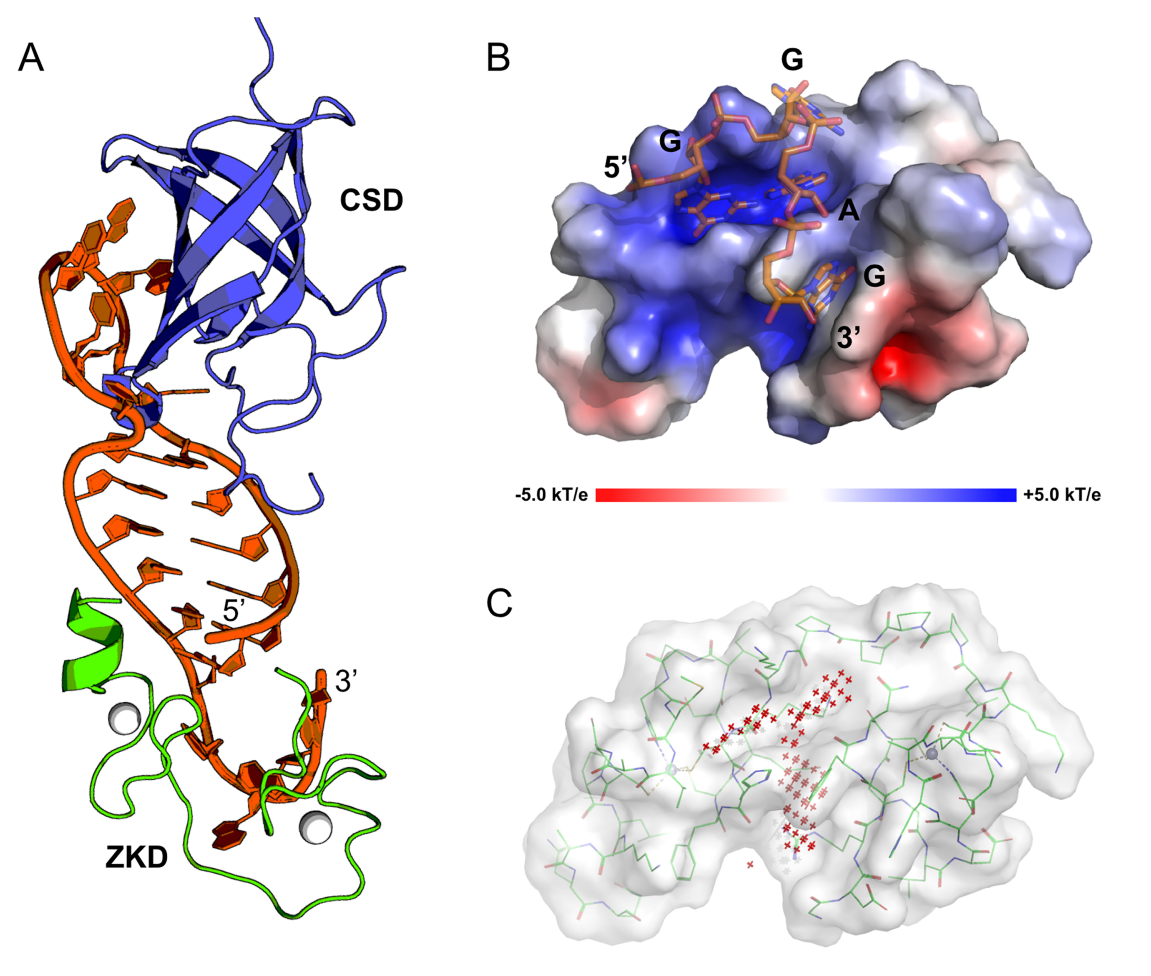
Appendix Figure B.6. The ZKD Plays Distinct Roles in Dicer and TUT4 Regulation. (A) Oligouridylation assays with pre-let-7g as probe, carried out using full-length mouse TUT4 and LIN28A truncation constructs. (B) Oligouridylation assays with pre-let-7g as probe, carried out using full-length mouse TUT4 and LIN28A mutants. (C) Oligouridylation assays with pre-let-7g mutants as probe, carried out using mouse TUT4 and LIN28A. (D) Dicer processing assays with pre-let-7g and LIN28A full-length and truncation constructs. *Oligouridylated pre-let-7g.

We also tested RNA mutations that affect ternary complex assembly for their effects on oligouridylation. When the ZKD binding site in pre-let-7g is altered (from GGAG to GGAU), we observe a dramatic loss of TUT4 activity (Appendix Figure B.6C). GGAG to UGAG mutation is less effective and only causes reduction in TUT4's processivity. In contrast, mutating the CSD binding site (GAU to CUA) does not significantly change the level of oligouridylation. All RNA mutations were tested at high-enough concentrations of LIN28 to support binary complex formation. The detrimental effects of ZKD binding site mutations on oligouridylation suggest that it is important for ZKD to interact with the cognate sequence GGAG to recruit TUT4 and activate its function on pre-let-7g.

The ZKD binds to the 30 end of the preE, which is adjacent to the Dicer cleavage site. Given the critical role that the ZKD plays in recruitment of TUT4, we tested how it may affect Dicer activity. Full-length mouse LIN28A can effectively block Dicer cleavage of pre-let-7g, as previously reported (Appendix Figure B.6D). However, the CSD is unable to inhibit Dicer, and the ZKD can only partially inhibit Dicer processing. At the same concentrations, ZKD can elicit oligouridylation almost as effectively as the full-length LIN28A (Appendix Figure B.6A). These results suggest that both domains of LIN28 are required for efficient Dicer inhibition and that the ZKD does not play a dominant role, in contrast to how it affects TUT4.

ZKD Is a Potential Therapeutic Target

Given the crucial role of ZKD in recruiting TUT4 and subsequent degradation of pre-let-7 RNAs, modulation of LIN28A activity in humans by targeting the ZKD using small molecules could pro-vide a general mechanism to regulate the LIN28 pathway. Our results suggest that a small change in the ZKD region may be enough to block TUT4 recruitment or activation, making the ZKD a focused target for an inhibitor against LIN28. To provide an atomic model for the human ZKD bound to let-7, we deter-mined a crystal structure of the human LIN28DD:preE-let-7f-1 complex at 2.0 A˚ resolution (Appendix Table B.SI). The refined structure is overall similar to the mouse model (PDB: 3TS0), with a root-mean-square deviation of 0.28 A˚ (Appendix Figure B.7A). The ZKD has a deep and positively charged pocket that engages the G1 and A3 of the GGAG let-7 motif (Appendix Figure B.7B). This ZKD binding pocket scored favorably (SiteScore 0.897) as a potential druggable site, as determined by the predictor Sitemap (Schrodinger) (Appendix Figure B.7C) (Halgren, 2009). Sitemap also identified additional pockets in the CSD, but in contrast to the ZKD pocket, the CSD sites had only moderate scores of 0.63 and 0.58, respectively. Our data indicate that the functionally critical ZKD has interesting surface characteristics that could be exploited by small molecules.



Appendix Figure B.7. ZKD Is a Potential Therapeutic Target. (A) Overview of human LIN28DD:preE-let-7f-1. The CSD is in blue, the ZKD is in green, zinc is represented as gray spheres, and preE-let-7f-1 is in orange. (B) Surface potential map of the ZKD RNA binding site. Blue represents a positively charged surface, and red indicates a negatively charged surface. The GGAG RNA is in orange. (C) Identification of the druggable pocket within the ZKD. Pocket identification was performed using Sitemap (Schrodinger). Red crosses represent the pocket identified as suitable for targeting. The surface of ZKD is in white, and all residues are in green.

Discussion

Here, we report that in addition to inhibiting the maturation of let-7, LIN28 initiates pre-let-7 degradation by recruiting TUTase through a direct interaction mediated by its ZKD. We show that the formation of the LIN28:pre-let-7:TUT4 ternary complex is a critical step for LIN28-mediated oligouridylation of pre-let-7. We also identify the minimal regions of each component that are required to assemble stable ternary complexes. Our results shed light on the molecular basis of TUTase recruitment by LIN28 and highlight the role of LIN28 ZKD as a crucial factor in regulating degradation of pre-let-7.

In high-resolution structures of mouse LIN28A in complexes with preE-let-7, the LIN28A CSD and ZKD manifest fundamentally different modes of RNA binding (Nam et al., 2011; Mayr et al., 2012; Loughlin et al., 2011). Our SPR data reveal that the CSD alone has fast-on and fast-off binding kinetics, sup-porting the relative promiscuity of the CSD described in previous studies (Mayr et al., 2012). Fast association and dissociation of the CSD effectively samples RNA sequences for GGAG motif recognition by the ZKD. Given that the CSD motif is a common feature of nucleic acid binding proteins, we believe our findings may provide important insights into diverse instances of RNA and DNA regulation by related factors.

The ZKD interacts with a short, four-base GGAG fragment that interdigitates with LIN28 side chains. Our SPR kinetic data show that despite its low binding affinity, the interaction between the GGAG sequence and the ZKD

increases the half-life of the full-length complex by two orders of magnitude and therefore dramatically stabilizes the LIN28A:pre-let-7f-1 complex. The sequence of CSD and ZKD binding events remains speculative, but it is likely that the CSD binds and remodels the let-7 sequence in the first step of the interaction, as demonstrated by Mayr et al., 2012. Once CSD forms the first contact and brings the ZKD to the vicinity, CCHC motifs and the GGAG sequence can bind to arrive at a unique conformation as a stable complex. This model is supported by evidence that isolated ZKD undergoes a conformational change upon binding RNA (Nam et al., 2011; Loughlin et al., 2011).

The role of the ZKD in stabilizing the LIN28:let-7 complex is crucial for TUT4 recruitment. Our data show that ZKD-RNA binding interface mutations—Y140A and H148A of LIN28A, as well as GGAG to GGAU and GGAG to UGAG pre-let-7 mutations—abolish the ternary complex assembly, indicating that conformational change in the ZKD upon binding to the GGAG RNA of pre-let-7 is a required step for TUT4 recruitment. Enzymatic activity of TUT4 is even more sensitive to these mutations. Our data are consistent with a model in which the core of LIN28:TUT4 interaction is via the ZKD and the N-terminal fragment of TUT4. Additional surface residues (e.g., E151 and I167) of ZKD that do not affect RNA binding are likely to be important for interacting with TUT4 directly. Therefore, we propose a specific model for LIN28-dependent oligouridylation of pre-let-7 RNAs. LIN28 ZKD interacts with the specific sequence of pre-let-7 near the 30 end of its preE, which causes ZKD:pre-let-7 to form a stable complex with a unique conformation. The composite surface of ZKD and the dsRNA region of

pre-let-7 then allows TUT4 to recognize and bind pre-let-7 for sufficient time to complete many consecutive uridylation cycles at the 30 end of pre-let-7.

It was previously suggested by single-molecule methods that LIN28 serves as a processivity factor by localizing TUT4 in the pre-let-7 vicinity; however, the molecular basis of TUT4 recruitment has remained unclear (Yeom et al., 2011). In particular, stable complexes of pre-let-7, LIN28, and TUT4 have been difficult to observe. Here we demonstrate, using purified components, that we can detect stable ternary complexes, which can generate oligouridylated pre-let-7 in vitro upon addition of UTP. In addition, our mapping studies identify the necessary and sufficient portions of each component to form a ternary complex: the ZKD of LIN28, N-terminal region of TUT4, and preE adjacent to an about 15-nt-long dsRNA stem. In the full-length complex, we expect a local interaction between ZKD and N-terminal TUT4 to bring the catalytic domain in the C-terminal region of TUT4 close to the 30 tail of pre-let-7. It is possible that the interaction between ZKD and N-terminal TUT4 serves as the switch between monouridylation and oligouridylation (Heo et al., 2009, 2012; Hagan et al., 2009). This proposed model of oligouridylation is similar to the model of polyadenylation: processivity factors like cleavage and polyadenylation specificity factor (CPSF) and poly(A) binding protein II (PABII) bind to RNA and bring PAP close to its substrate (Viphakone et al., 2008). Studies have proposed that mRNAs can also be oligouridylated (Lim et al., 2014) and it will be interesting to investigate whether other LIN28 target RNAs (for example, mRNA) are substrates for TUT4 and whether processivity factors other than LIN28 exist.

One important question is how our model, developed with mouse LIN28A and TUT4 constructs, applies to other homologs. Sequence analysis and structural evaluation suggest that mouse LIN28B and human LIN28 paralogs are functionally equivalent in the recruitment of TUTase. The major differences between sequences of LIN28A and LIN28B paralogs are in the C-terminal region, which in LIN28B contains a nuclear localization sequence outside of the previously crystallized functional fragment (Nam et al., 2011). Sequence identity within the critical ZKD is high (96.8% between mouse and human LIN28A and 82.3% between human LIN28A and LIN28B), and the residues involved in GGAG recognition (Y140, H148, H162, M170, and K159), as well as the two surface residues involved in TUT4 recruitment (E151 and I167), are conserved (Appendix Figure B.S2). Previous studies have pro-posed that TUT4 and TUT7 are functionally redundant (Thornton et al., 2012). We identified a motif in TUT4 (YRYFACPQKK) to be important for interacting with LIN28, and it is conserved between TUT4 in mice and humans and between TUT4 and TUT7 paralogs (Appendix Figures B.S3 and B.S4). Therefore, the specific regions of LIN28 and TUT4 that we identify to be critical for complex formation and oligouridylation are conserved through evolution.

There is an increasing body of evidence linking overexpression of LIN28 proteins in mature cells to oncogenesis (Peng et al., 2010; Permuth-Wey et al., 2011; King et al., 2011; Viswanathan et al., 2009; Tu et al., 2015; Beachy et al., 2012; Molenaar et al., 2012; Urbach et al., 2014; Nguyen et al., 2014). Therapeutic agents to target this pathway have not yet been identified, partly

because of the difficulty in disrupting the extensive LIN28-RNA binding interface. Our data demonstrate that the ZKD is the required component of the let-7 degradation pathway, which points to the ZKD as a possible target for suppressing LIN28-dependent cancers. Although most RNA binding proteins reuse a small subset of RNA binding scaffolds, the tandem CCHC ZKD motif in LIN28 appears to be rare and may be blocked by a specific inhibitor. To facilitate development of potential therapeutics, we have determined a high-resolution crystal structure of human LIN28DD:preE-let-7f-1. The structure of the complex reveals a potentially druggable pocket located within the ZKD between the two zinc knuckles. Although the presence of this pocket is partially attributed to the RNA-bound conformation of the ZKD, small molecules that mimic the GGAG RNA, or bind to parts of the pocket and allosterically block the ZKD:RNA interactions, could effectively inhibit the LIN28/let-7 pathway.

Our studies highlight the role of the ZKD of LIN28 in repressing let-7 by promoting its degradation: first stabilizing the LIN28: let-7 complex and then recruiting TUT4 through direct interactions with N-terminal TUT4, which leads to the oligouridylation of pre-let-7 followed by poly(U)-specific degradation by Dis3L2. The LIN28/let-7 pathway is critical for driving and maintaining factors in several human cancers. Here we present a cavity of LIN28 ZKD suitable for targeting by small molecules that can serve as a potential therapeutic target for cancer.

Experimental Procedures

Constructs

All mouse LIN28A constructs were previously described (Nam et al., 2011), including LIN28A (1–209), LIN28D (16–184), LIN28DD (16–184, with an internal deletion of residues 126–135), CSD (16–126), and ZKD (135–184). The mouse TUT4 constructs were derived from mouse TUT4 (NP_780681.2). TUT4 expression constructs are in the pGEX expression vector with an N-terminal glutathione S-transferase tag and a C-terminal hexahistidine tag. N-terminal TUT4 and C-terminal TUT4 expression constructs are in the Macrolab vector (Addgene plasmid 29707) with an N-terminal hexahistidine tag and a glutathione S-transferase tag.

Protein Purification

Mouse LIN28A and human LIN28DD constructs were purified as previously described (Nam et al., 2011). The mouse TUT4 constructs were overexpressed in E. coli strain BL21(DE3) Rosetta pLysS cells and purified by affinity chromatography using Nickel-nitrilotriacetic acid (Ni-NTA) resin (QIAGEN) affinity purification, followed by glutathione Sepharose resin (glutathione S-transferase [GST], GE Healthcare) affinity purification. The buffer used in Ni-NTA elution is 50 mM Tris (pH 8.0), 300 mM NaCl, 10% glycerol, 1 mM DTT, 100 mM ZnCl2, and 300 mM imidazole (pH 8.0). The buffer used in GST elution is 50 mM Tris (pH 8.0), 300 mM NaCl, 30% glycerol, 5 mM DTT, 100 mM ZnCl2, and 20 mM glutathione. The N-terminal and C-terminal TUT4 were purified by affinity

chromatography using Ni-NTA resin (QIAGEN) affinity purification, followed by ion exchange chromatography (HiTrap SP, GE Healthcare). The buffers used in ion exchange chromatography were 20 mM Tris (pH 8.0), 10% glycerol, and 5 mM b-mercaptoethanol (BME), with a 100 mM to 1 M gradient of NaCl. The size exclusion chromatography was performed on Superdex 200 (GE Healthcare) using buffer containing 20 mM Tris (pH 8.0), 300 mM NaCl, 1 mM MgCl2,50 mM ZnCl2, 5% glycerol, and 5 mM BME.

EMSAs

Mouse pre-let-7g RNA was purified by PAGE after in vitro transcription; fol-lowed by double-ribozyme cleavage, as described in Walker et al. (2003); and radiolabeled with ATP [g-32P] using a T4 polynucleotide kinase. Other RNA oligonucleotides (oligos) were chemically synthesized and purchased from Integrated DNA Technologies (IDT). Reactions were performed with labeled pre-let-7g probes incubated with protein in a buffer containing 20 mM Tris (pH 7.5), 75 mM NaCl, 10 mM DTT, 3 mM MgCl2,50mM ZnCl2 0.21 mg/mL yeast tRNA, 0.13 U/mL RNase inhibitor (RNaseOUT, Thermo Fisher Scientific), and 10% glycerol. Reactions were incubated for 20 min and resolved on 8%–10% native polyacrylamide gels.

TUTase In Vitro Uridylation Assay

Recombinant mouse TUT4 and LIN28A proteins were incubated with radiolabeled pre-let-7g at 37 ^{o}C in a reaction containing 20 mM Tris (pH 7.5), 5%

glycerol, 6 mM MgCl2, 6 mM DTT, 0.14 U/mL RNase inhibitor, 50 mM ZnCl2,40mM KCl, and 167 mM UTP for 30 min. The reactions were then stopped by adding 1% SDS and 25 mM EDTA, followed by protease K treatment at 50□C for 30 min. The reactions were resolved on 15% denaturing polyacrylamide gels.

SPR

The immobilization of neutravidin to the CM5 sensor chip (GE Healthcare) was performed as previously described (Wang et al., 2011). 50 biotinylated preE-let-7f-1 was purchased from IDT with the sequence of biotin-50-GGGGUAGU GAUUUUACCCUGUUUAGGAGAU-30. The biotin-labeled preE-let-7 was applied to the sensor chip and captured by immobilized neutravidin. Sensor-grams were measured by injection of mouse LIN28A at various concentrations at the flow rate of 50 mL/min in running buffer containing 20 mM Bis-Tris (pH 6.0), 10% glycerol, 300 mM NaCl, 5 mM MgSO4, 0.05% NP-40, 5 mM DTT, and 50 mg/mL tRNA. Regeneration was performed using running buffer plus 2 M NaCl. A channel containing neutravidin, but without preE-let-7f-1, was used as a reference channel. All sensorgrams were repeated at least once. All experiments were performed on a Biacore 3000 (GE Healthcare). Fitting of sensorgrams was carried out using the BIAevaluation software suit.

Dicer In Vitro Processing Assay

Dicer expression construct (Addgene plasmid 19873) and purification were carried out as described previously (Landthaler et al., 2008). Radiolabeled

pre-let-7g was prepared in the same manner as EMSA probes. Dicer assays were performed as described (De and Macrae, 2011) in a reaction containing 20 mM Tris (pH 7.5), 5% glycerol, 3.2 mM MgCl2, 5 mM DTT, 50 mM NaCl, and 150 mM ZnCl2.

FP Assay

The FP assay was carried out by titrating the LIN28 into the mixture of FAM-labeled preE-let-7f-1 (2 nM final concentration) and buffer M (100 mM sodium chloride, 20 mM Tris-HCl [pH 7.0], 5 mM magnesium chloride, 10% v/v glycerol, 5 mM dithiothreitol, and 0.1% v/v NP-40). The assay plate was briefly vortexed, and air bubbles were removed by spinning at 1,000 rpm. An EnVision plate reader (PerkinElmer) was used to measure the FP.

Crystallography

Crystals of the human LIN28DD:preE-let-7f-1 complex were produced by vapor diffusion using the hanging drop method. Concentrated complexes (10 mg/mL) were mixed 1:1 mL with reservoir solution. Reservoir solution contained 1.1 M citrate (pH 8.0). Crystals were harvested with mother liquor supplemented with 20% glycerol and frozen in liquid nitrogen.

Structure Determination

Datasets were collected on beamline 24 at the Advanced Photon Source. Diffraction data were indexed and scaled using XDS (Kabsch, 2010) and SCALA

(Evans, 2006) in a work flow provided by autoPROC (Vonrhein et al., 2011). Data resolution was established based on CC1/2 criteria. The structure of LIN28DD:preE-let-7f-1 was determined by molecular replacement with the CSD and CCHC domains from crystal structure of mouse LIN28 (PDB: 3TS2) as search models using Phaser (McCoy et al., 2007). Several rounds of model building and refinement were completed in COOT (Emsley and Cowtan, 2004) and PHENIX, respectively. The RNA structure was remodeled in RCRANE. Final data collection and refinement statistics are presented in Appendix Table B.SI. Coordinates and structure factors have been deposited as PDB: 5UDZ.

Accession Numbers

The accession number for the coordinates and structure factors reported in this paper is PDB: 5UDZ.

Author contributions

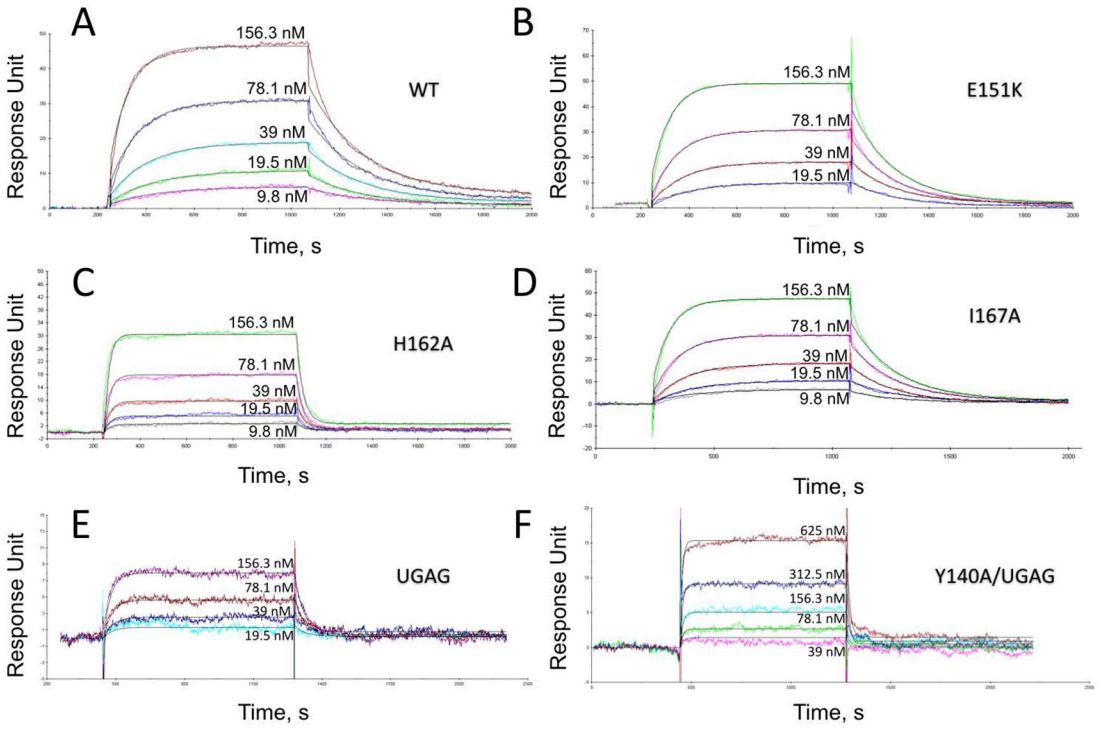
L.W., Y.N. designed the experiments; L.W., Y.N., A.K.L., C.Y. K.R. and C.C. conducted the experiments. L.W., Y.N., and P.S. organized and interpreted the data, and L.W. and P.S. wrote the manuscript with the help of Y.N., E.M.R. and C.Y.

Acknowledgements

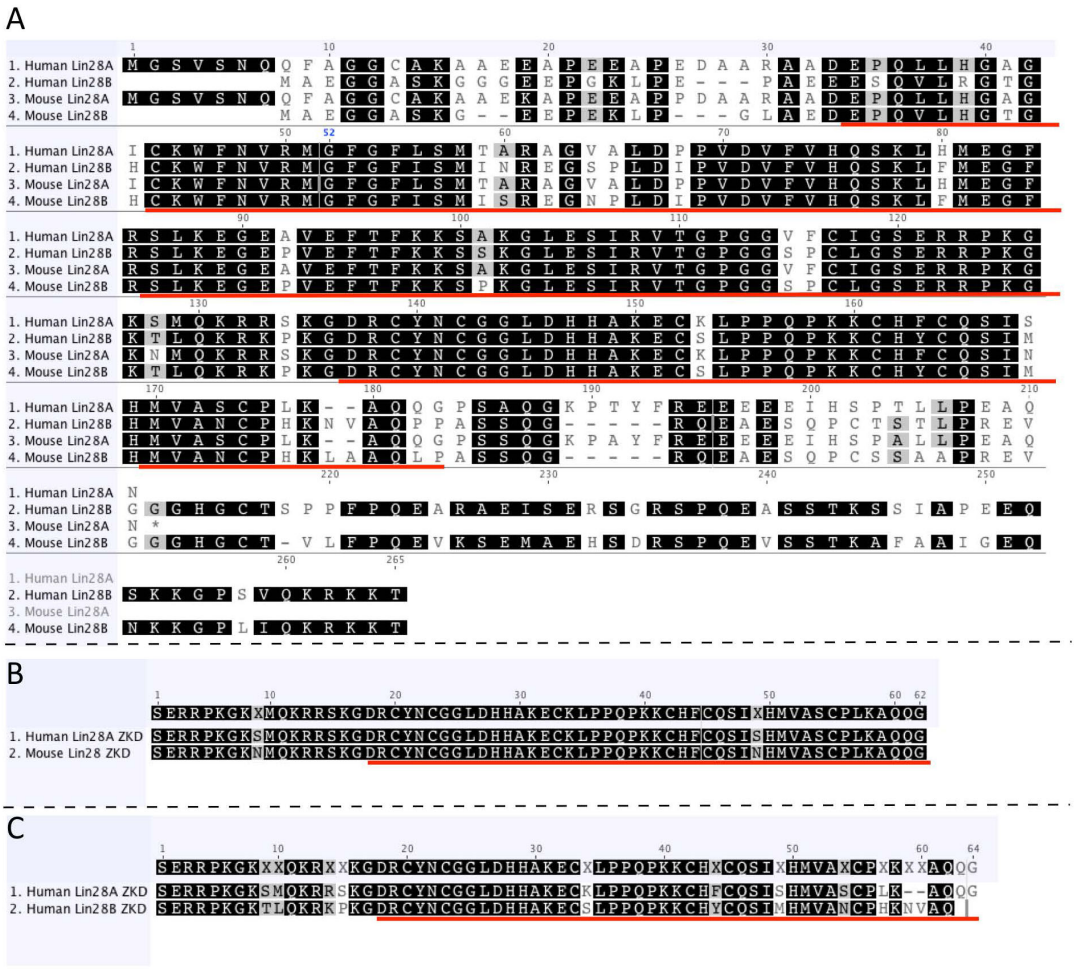
This work was supported by a grant to P.S. from the National Cancer Institute (R01CA163647).

Appendix B – Supplemental Materials

LIN28 zinc knuckles domain is required and sufficient to induce let-7 oligouridylation.



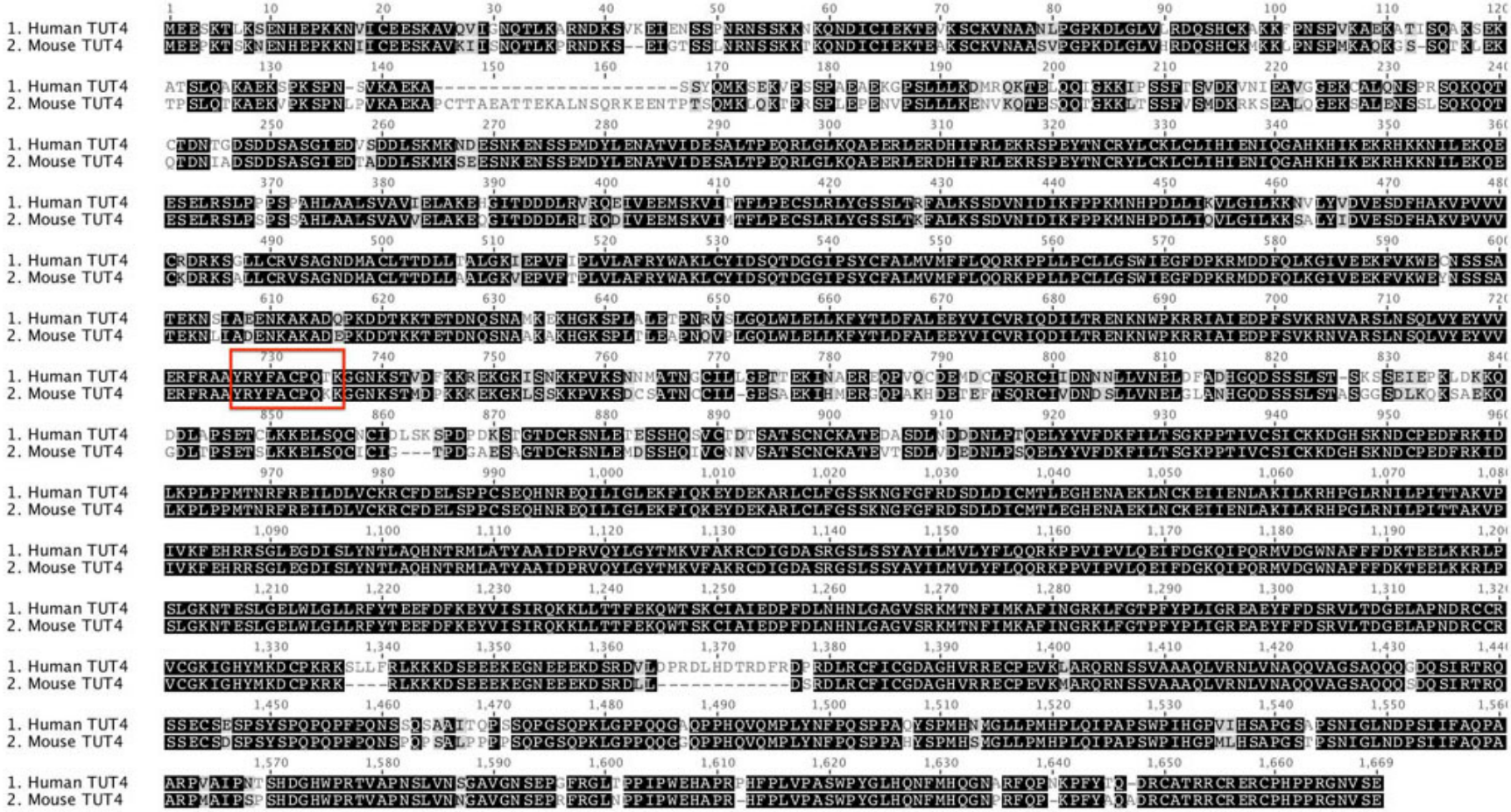
Appendix Figure B.S1. SPR studies of LIN28 and preE-let-7f variants, related to Appendix Figure B.1. (A) Sensorgrams of wild-type mouse LIN28 with immobilized mouse preE-let-7f. (Band D) Sensorgrams of LIN28 ZKD surface mutants with immobilized mouse preE-let-7f. (C) Sensorgrams of LIN28 ZKD RNA-binding mutant with immobilized mouse preE-let-7f. (E) Sensorgrams of LIN28 with mutated mouse preE-let-7f (GGAG to UGAG). (F) Sensorgrams of LIN28 ZKD RNA-binding mutant Y140A with mutated mouse preE-let-7f (GGAG to UGAG).



Appendix Figure B.S2. Sequence alignment of human and mouse LIN28 paralogs, related to Appendix Figure B.1. (A) Sequence alignment of the CSDs from human and mouse LIN28 paralogues. (B) Sequence alignment of the ZKDs from human and mouse LIN28A. (C) Sequence alignment of the ZKDs from human LIN28A and LIN28B. Red lines indicate LIN28 crystal construct. The alignment is performed using Geneious.

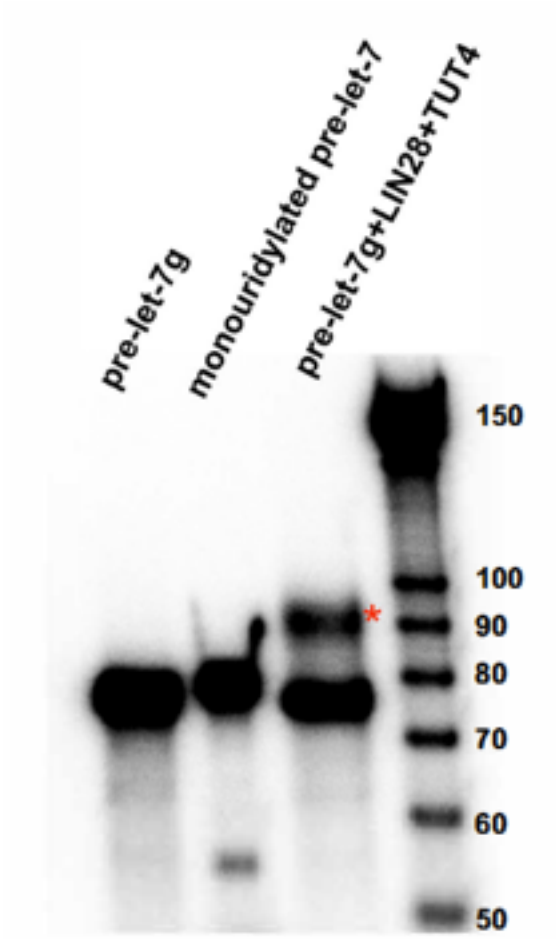
Appendix Figure B.S3. Sequence alignment of human and mouse TUT4, related to Appendix Figure B.2. The sequence identity between human and mouse TUT4 is 86.4% . Red box indicate motif M. The alignment is performed using Geneious.

Appendix B.S3. (Continued)

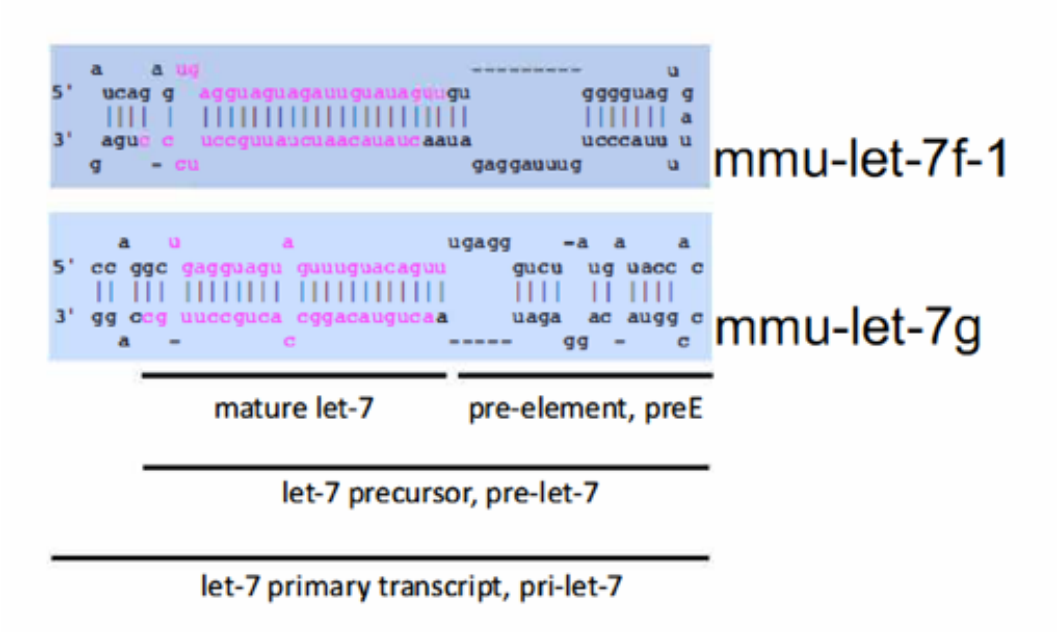


Appendix Figure B.S4. Sequence alignment of human TUT4 and TUT7, related to Appendix Figure B.2. The sequence identity between human TUT4 and TUT7 is 40.2%. Red box indicate motif M. The alignment is performed using Geneious.

Appendix B.S4. (Continued)



Appendix Figure B.S5. Comparison between monouridylation and oligouridylation, related to Appendix Figure B.2. Urea-PAGE of pre-let-7g, monuridylated pre-let-7g, oligouridylated pre-let-7g and Decade marker (Thermo, AM7778). *, oligouridylated pre-let-7g.



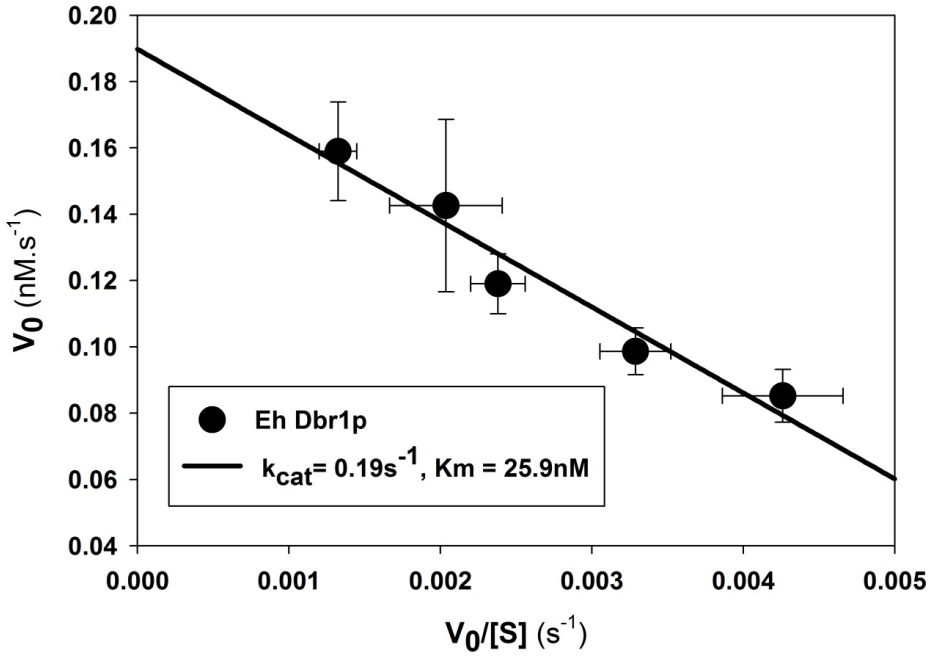
Appendix Figure B.S6. Sequence and secondary structure predictions of mouse let-7f-1 and let-7g, related to Appendix Figure B.5. Sequences and secondary structure predictions are acquired from mirBase.

Appendix Table B.SI. Data collection and refinement statistics, related to Appendix Figure B.7

	Human LIN28:preE-let-7f-1 Complex
Wavelength (A)	0.97
Resolution range (A)	32.52 - 2.0 (2.071 -2.0)
Space group	P 41
Unit cell	76.429 76.429 105.615 90 90 90
Total reflections	109233 (9194)
Unique reflections	40547 (3981)
Multiplicity	2.7 (2.3)
Completeness (%)	99.00 (96.96)
Mean I/sigma (I)	10.79 (1.73)
Wilson B-factor	23.52
R-merge	0.09052 (0.6189)
R-meas	0.1127
R-work	0.1662 (0.2623)
R-free	0.1975 (0.2993)
Number of atoms	3653
macromolecules	3110
ligands	4
water	539
Protein residues	328
RMS (bonds)	0.013
RMS (angles)	1.40
Ramachandran favored (%)	97
Ramachandra outliers (%)	0.74
Clashscore	5.43
Average B-factor	23.10
macromolecules	21.70
ligands	14.30
solvent	31.30

Appendix C

Supplemental Materials to Chapter 2



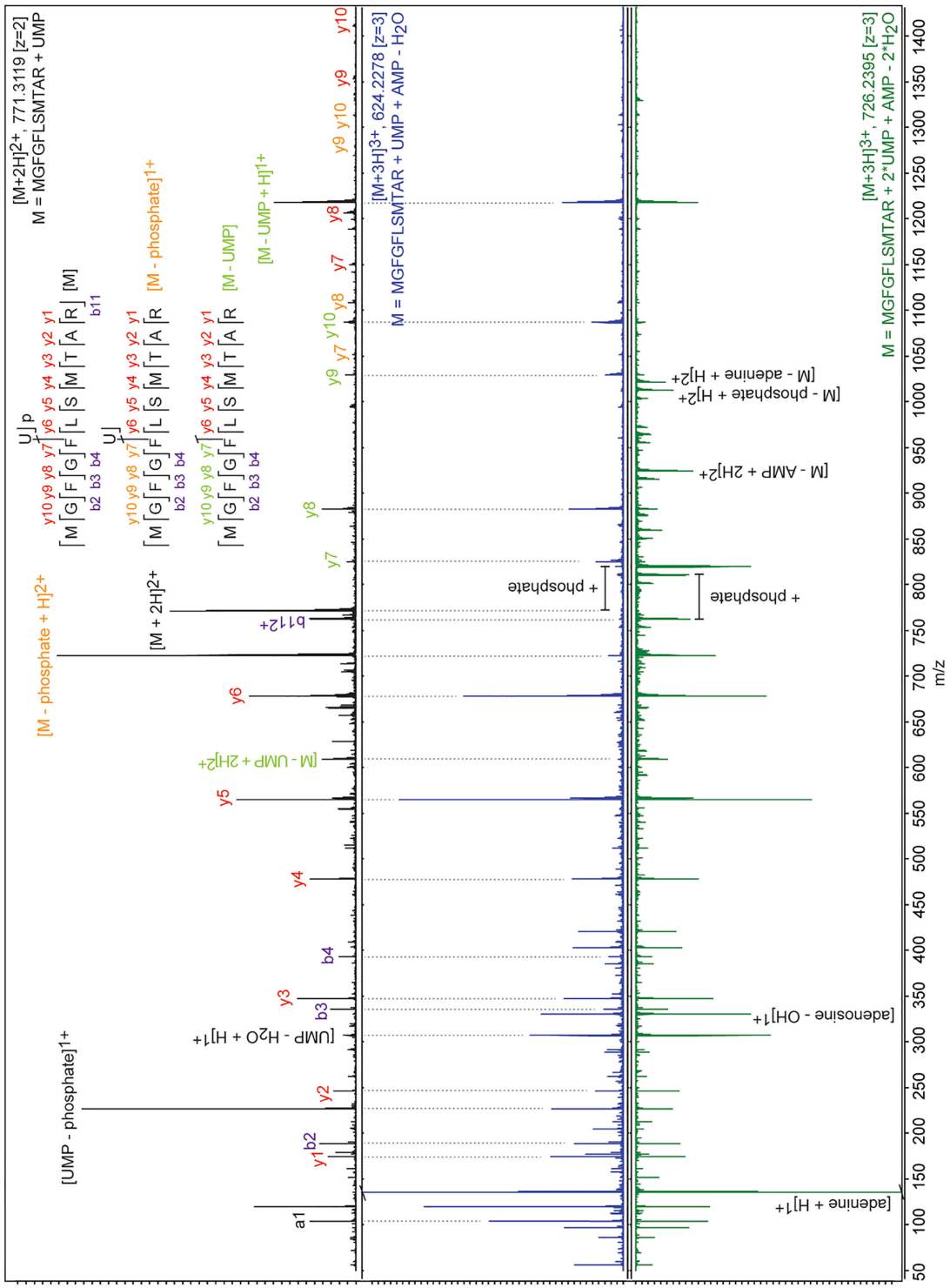
Appendix Figure C.S1. Eadie-Hofstee diagram for analysis of Dbr1 debranching kinetics

Appendix D

Supplemental Materials to Chapter 3

Appendix Figure D.S1. Targeted tandem mass spectra confirming the identity of di- and tri- nucleotide heteroconjugates. Tandem MS fragmentation patterns from three parent species of MGFGFLSMTAR-uridine heteroconjugates that identify Phe55 as the site of uridine crosslinking (top panel). Characteristic ions indicating the presence of adenosine and uridine phosphates were observable following CID of the selected ions 624.2278 m/z (z = +3) (middle panel) and 726.2395 m/z (z = +3) (bottom panel), corresponding to the peptide conjugates with the compositional addition of nucleotides AU and AUU, respectively. The y- and b-ions arising from peptide backbone fragmentation following neutral loss of the nucleotide species were again consistent with the peptide sequence MGFGFLSMTAR.

Appendix Figure D.S1. (Continued)



Appendix Table D.SI. Binding site overlap analysis source information

Human LIN28 CLIP Datasets for Binding Site Overlap Analysis					
Dataset	**Reference**	**CLIP method**	**Protein**	**Cells**	**No. of enriched sites (hg19)**
A	Hafner et al., 2013	PAR-CLIP (4SU)	LIN28A	HEK 293	1156
B	Wilbert et al., 2012	CLIP-seq	LIN28A	H9	3723
C	Wilbert et al., 2012	CLIP-seq	LIN28A	HEK 293	1419
D	Hafner et al., 2013	PAR-CLIP (4SU)	LIN28B	HEK 293	1473
E	Graf et. al., 2013	iDo-PAR-CLIP (4SU and 6SG)	LIN28B	HEK 293	1474

Appendix Table D.SII. Processing Output Summary

		GSE44615 (Hafner et al., 2013)			
Datasets		**GSM1087848**	**GSM1087849**	**GSM1087850**	**GSM1087851**
		SRR764666	**SRR764667**	**SRR764668**	**SRR764669**
Download Size		6.7G	39M	232M	7.1G
FASTQC:average_quality_first_base/low_quality_first_base /length_of_read. If only two values – no low quality bases Red – low quality dataset, yellow – medium quality dataset, green – good quality dataset		25/31/51	34/35/36	22/29/36	22/51 red, 5-14 green
Quality after low-quality filtering					
Quality after low-quality filtering and adapters removal.		32/(25-51)	(25-36)	29/(25-36)	31/33/(25-51)
hg19 mapping					
only quality-filtered [fraction, absolute value]		2.79% 7M	3.57% 90K	0.4% 30K	0.26% 300K
quality-filtered and removed adapter [fraction, absolute value]		22.3% 17M	20% 160K	13% 100K	8.12% 4M
Improvement of mapping		2.428571429	1.777777778	3.333333333	13.33333333
Quality check after mapping strategy					
Removal of all reads with score below D					
hg19		0	96K	0	0
picard					
After duplicate removal		160K	82K	57K	130K
fraction of survivors		0.009411765	0.5125	0.57	0.0325
Piranha hits	**binsize**				
	10	1029	774	361	992
	20	1020	509	350	867
	30	1045	518	345	854
	50	1080	500	326	839
as suggested by creators	**200**	1156	541	335	908

Appendix Table D.SII. Processing Output Summary (Continued).

Datasets		GSM980593 (Wilbert et al., 2012)			GSM980594 (Wilbert et al., 2012)
		SRR531463	SRR531464		SRR531465
Download Size		52M	357M		288M
FASTQC:average_quality_first_base/low_quality_first_base /length_of_read. If only two values – no low quality bases Red – low quality dataset, yellow – medium quality dataset, green – good quality dataset		10/25/40	21/32/40		34/36
Quality after low-quality filtering					
Quality after low-quality filtering and adapters removal.		13/(25-40)	23/(25-40)		(25-36)
hg19 mapping					
only quality-filtered [fraction, absolute value]		16.2% 200K	9.1% 1.1M		8.4% 900K
quality-filtered and removed adapter [fraction, absolute value]		77.4% 400k	90% 2.4M		92% 1.6M
Improvement of mapping		2	2.181818182		1.777777778
Quality check after mapping strategy					
Removal of all reads with score below D					
hg19		0	0		54K
picard					
After duplicate removal		230K	630K		270K
fraction of survivors		0.575	0.2625		0.16875
Piranha hits	**binsize**				
	10	2234	4767		2508
	20	1973	5457		1637
	30	1514	4695		2076
	50	1346	3347		1843
as suggested by creators	**200**	1153	2241		1419

Appendix Table D.SII. Processing Output Summary (Continued).

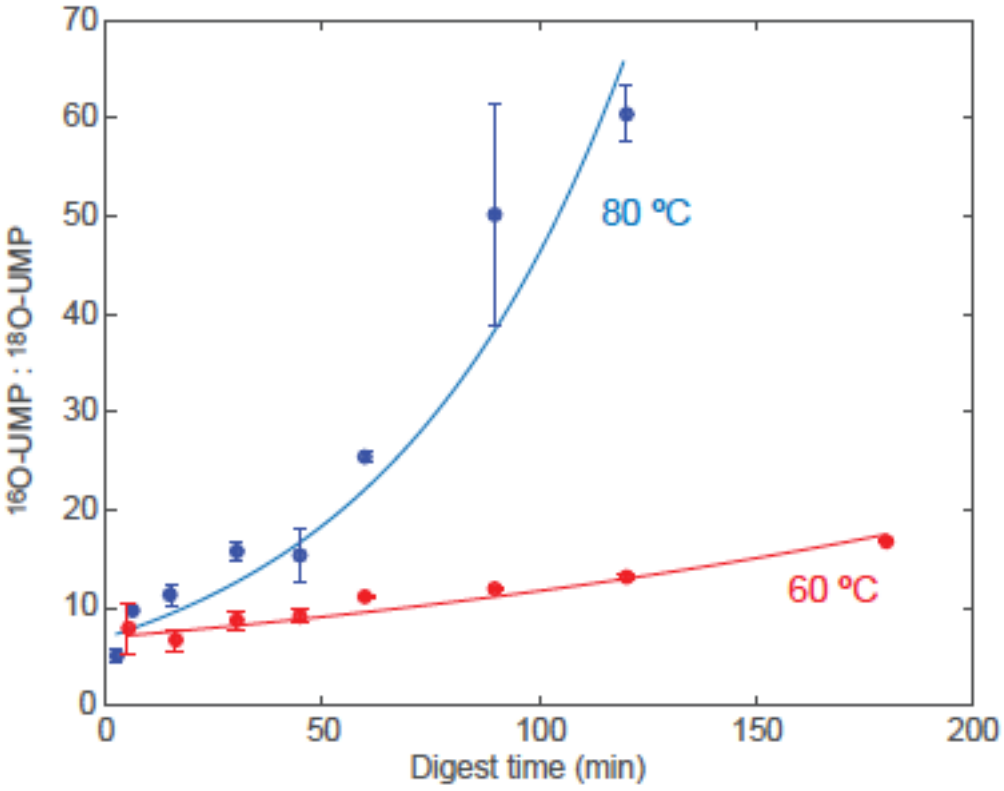
Datasets		GSM1140829 (Graf et al., 2013)					
		SRR850551	SRR850552	SRR850553	SRR850554	SRR850555	SRR850556
Download Size		903M	269M	42M	66M	217M	288M
FASTQC:average_quality_first_base/low_quality_first_base /length_of_read. If only two values – no low quality bases Red – low quality dataset, yellow – medium quality dataset, green – good quality dataset		42/44/51	41/42/51	40/41/51 49-	42-46/51 ~ y	42/43/51	.
Quality after low-quality filtering							
Quality after low-quality filtering and adapters removal.		(25-51)	41/44/(25-51	40/42/(25-51	(25-51)	44/45/(25-51)	42/44/(25-51
hg19 mapping							
only quality-filtered [fraction, absolute value]		0.11% 28K	0.1% 7K	0.4% 4.5K	0.28% 5K	0.06% 2.5K	0.26% 21K
quality-filtered and removed adapter [fraction, absolute value]		1.72% 400K	1.74% 130K	11.2% 70K	10.6% 110K	0.59% 35K	0.29% 21K
Improvement of mapping		14.285714	18.571429	15.555556	22	14	1
Quality check after mapping strategy							
Removal of all reads with score below D							
hg19		0	0	0	0	0	0
picard							
After duplicate removal		75K	67K	47K	65K	7.6K	2.5K
fraction of survivors		0.1875	0.5153846	0.6714286	0.5909091	0.217142857	0.1190476
Piranha hits	binsize						
	10	829	1038	304	434	43	25
	20	570	523	264	358	38	16
	30	558	550	251	339	35	13
	50	544	571	253	349	56	11
as suggested by creators	200	619	673	277	433	40	10

Appendix Table D.SIII. Overlapping binding site summary

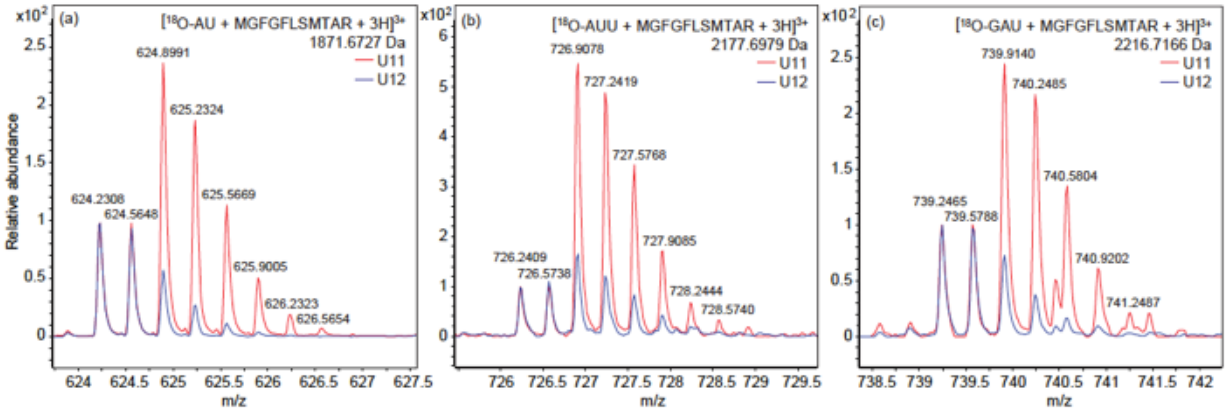
		Individual dataset overlaps				
		Hafner LIN28A	Hafner LIN28B	Wilbert LIN28A	Wilbert LIN28A	Graf LIN28B
		GSM '48	GSM '495051	GSM '93	GSM '94	GSM '29
Hafner LIN28A	GSM '48	-	877	88	268	398
Hafner LIN28B	GSM '495051	877	-	144	300	566
Wilbert LIN28A	GSM '93	88	144	-	480	177
Wilbert LIN28A	GSM '94	268	300	480	-	342
Graf LIN28B	GSM '29	398	566	177	342	-

Appendix E

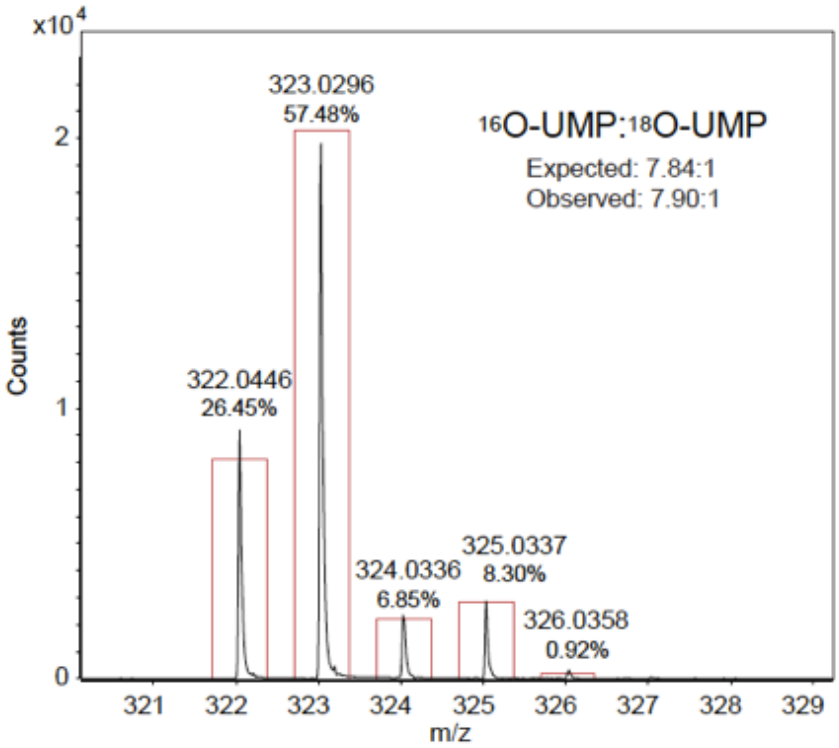
Supplementary Material to Chapter 4



Appendix Figure E.S1. Isotopic exchange of 18O label on uridine 3′(2′)-monophosphate with bulk solvent under acidic conditions. The oligonucleotide preE-let-7f was isotope labeled at U11 and hydrolyzed in 50% (v/v) formic acid, observed by ESI-LC-MS. Isotope ratio of 16O:18O vs. digest time at 80 °C (blue) and 60 °C (red) temperatures are shown. The fit lines correspond to a simple irreversible exchange model, where 18O-UMP is converted to 16O-UMP at a rate kex during hydrolysis of the starting material. In this case, kex = 1.1 h-1 at 80 °C and 0.28 h-1 at 60 °C.



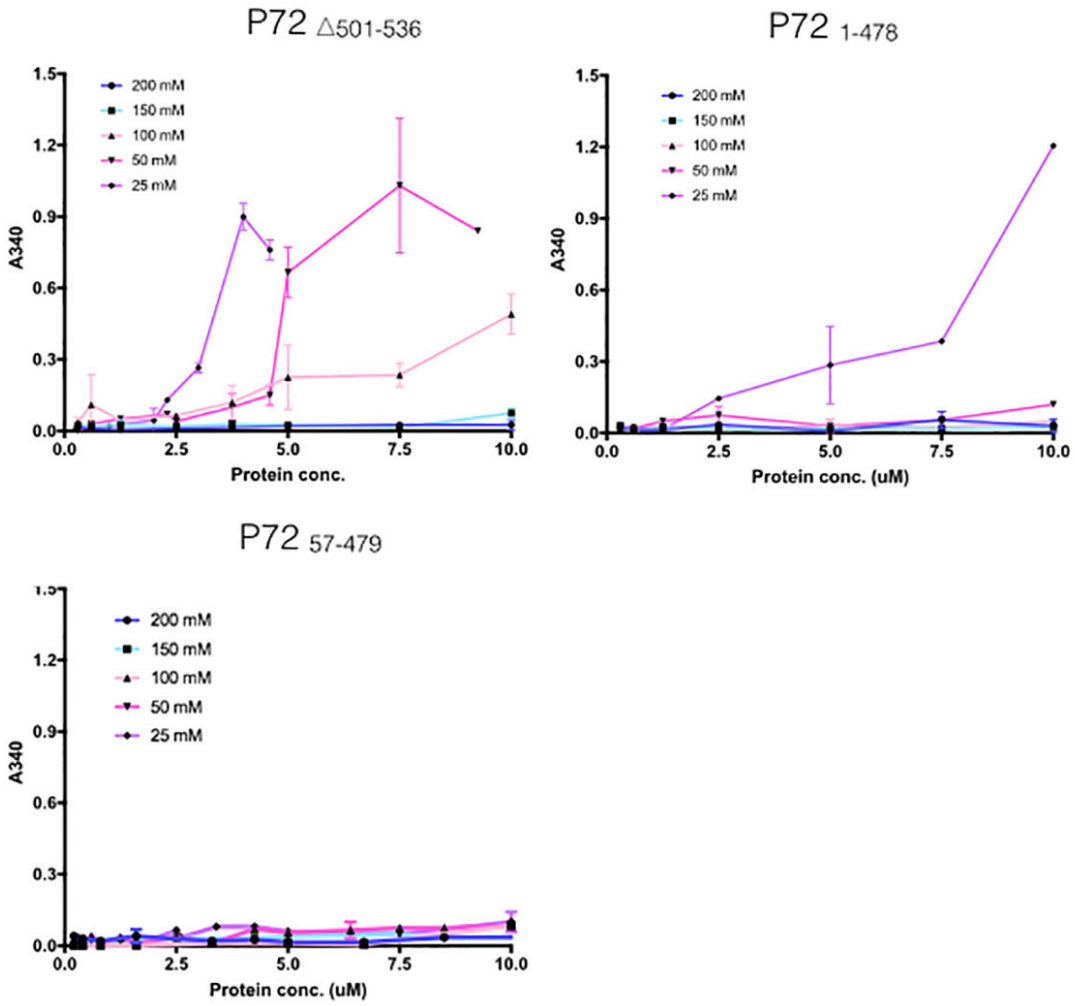
Appendix Figure E.S2. Isotopic distributions of peptide cross-linked di- and trinucleotides. Tryptic peptide ions arising from U11 (red) and U12 (blue) 18O labeled RNA-protein complexes, consistent with the peptide MGFGFLSMTAR cross-linked to the compositionally-defined ribonucleotides (a) AU, (b) AUU, and (c) GAU.



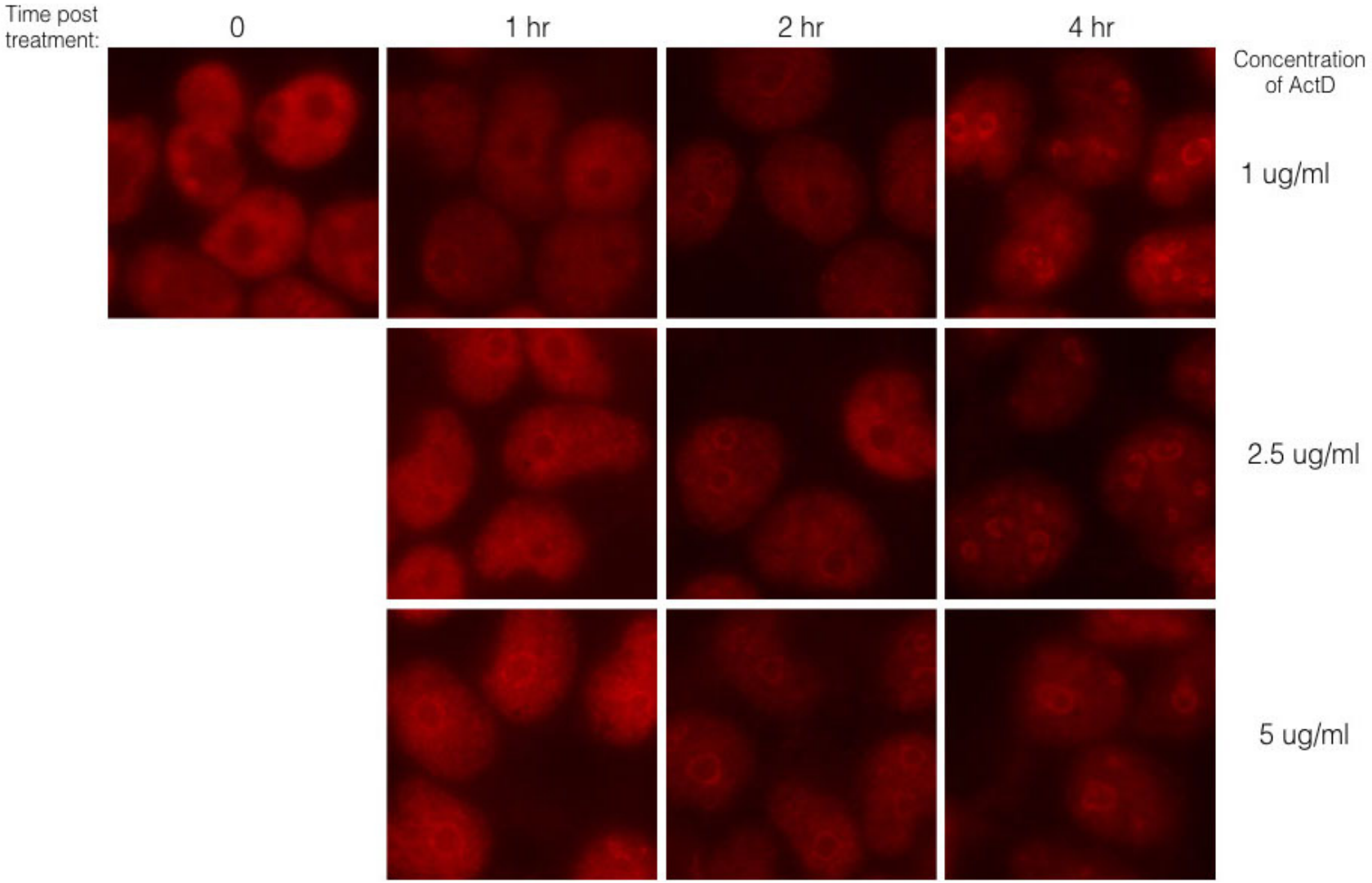
Appendix Figure E.S3. Enzymatic digest of preE-let-7 RNA isotope labeled at U11. The isotope distribution of mononucleotides in the mass region of uridine monophosphate (UMP) were examined by negative mode ESI-MS following digestion of 90 pmol with 1U nuclease P1 for 2 h at 42 ^{o}C. The survey spectrum shows the overlapping isotopic distributions of cytidine monophosphate (CMP, measured [M-H]1- = 322.0446 m/z) and UMP (measured [M-H]1- = 323.0296 m/z), demonstrating an enrichment of the UMP + 2 Da isotope (measured [M-H]1- = 325.0337 m/z). Relative isotope ratios (normalized to 323 m/z) are shown before correction for natural abundances, and the measured stoichiometry is shown after correction. The calculated isotope distribution of CMP and UMP from digesting the U11-labeled sequence GGGGUAGUGAU11UUUACCCUGGAGAU is shown (red boxes). The expected stoichiometry of 16O-UMP:18O-UMP is 7.84:1, given that the sequence has 8 uridines, one of which is labeled with ~90% 18O enrichment. The observed 16O-UMP:18O-UMP stoichiometry after correcting for natural abundances is 7.90:1, indicating negligible exchange of the mass label with bulk solvent. The natural abundance correction was performed by calculating integrated intensities, I, for each species as: ICMP = I322m/z, I16O-UMP = I323m/z - 0.1128 x ICMP, and I18O-UMP = I325m/z - 0.0239 x I16O-UMP – 0.002 x ICMP.

Appendix F

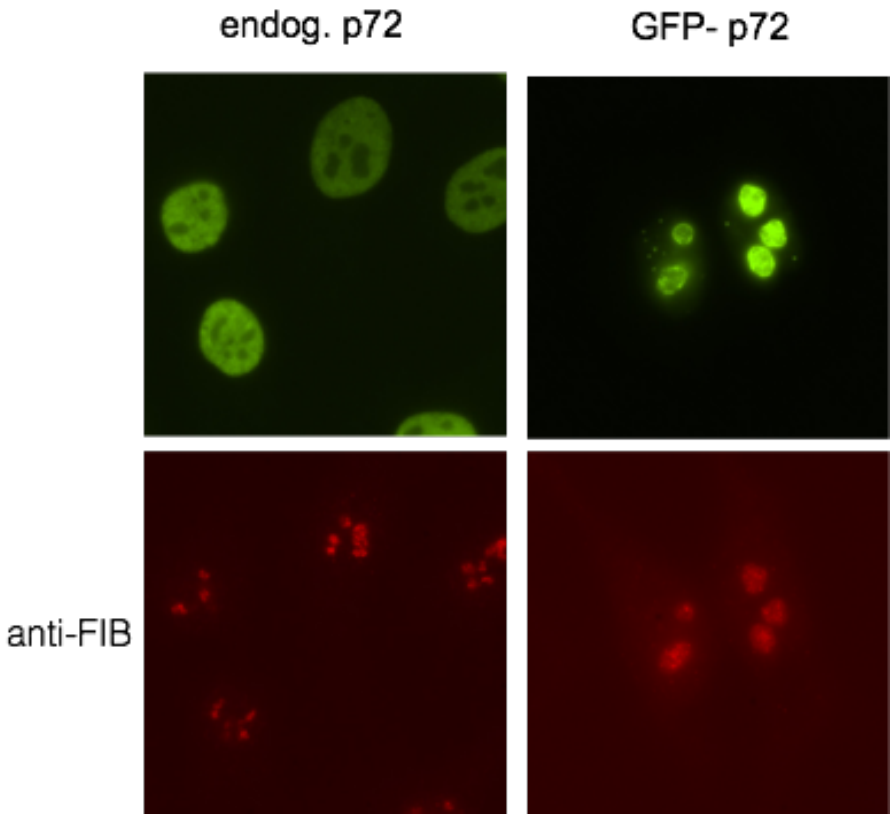
Supplemental Material to Chapter 5



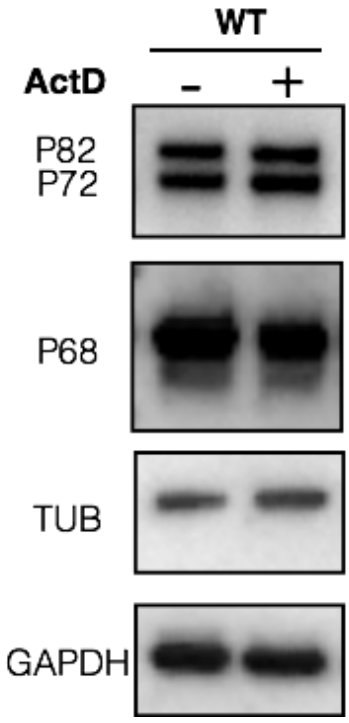
Appendix Figure F.S1. Turbidity measurements of increasing P72 concentrations at various ionic strengths, related to Figure 5.2.



Appendix Figure F.S2 Systematic ActD treatment of HEK cells, show same pattern of cap formation, stained with anti-P72, related to Figure 5.4.



Appendix Figure F.S3. Overexpression of full-length P72 results in nucleolar mislocalization



Appendix Figure F.S4. Western blot reveals expression levels of P72 appear unchanged following ActD treatment (2.5ug/ml, 4hrs)

Appendix Table F.SI. Statistics on puncta counts per cell (nuclei)

Time Point	45 mins	1hr	2hr	3hr	4hr
Number of values	8	9	11	10	11
Minimum	67	47	12	6	2
25% Percentile	82	49.5	29	12.3	4
Median	121	53	37	16.5	4
75% Percentile	141	81	52	34.3	12
Maximum	144	100	112	40.0	16
Mean	113.6	65.4	43.6	21.5	7.4
Std. Deviation	29.8	19.4	25.9	11.9	5.0
Std. Error of Mean	10.5	6.5	7.8	3.8	1.5
Lower 95% CI of mean	88.7	50.5	26.2	13.0	4.0
Upper 95% CI of mean	138.6	80.4	61.1	30.0	10.8
Sum	909	589	480	215	81

Appendix Table F.SII. Statistics on puncta volume (um^3)

Time Point	45 mins	1hr	2hr	3hr	4hr
Total number of values	909	589	483	215	79
Number of excluded values	0	0	0	0	0
Number of binned values	909	589	483	215	79
Minimum	1.1E-06	3.6E-03	2.0E-06	2.9E-02	6.3E-02
25% Percentile	1.1E-02	4.1E-02	8.0E-02	3.4E-01	3.7E-01
Median	4.3E-02	8.0E-02	1.8E-01	5.9E-01	7.3E-01
75% Percentile	1.0E-01	1.6E-01	3.9E-01	8.9E-01	2.0E+00
Maximum	1.1E+00	6.9E-01	3.3E+00	3.0E+00	1.7E+01
Mean	8.0E-02	1.1E-01	3.2E-01	6.8E-01	1.7E+00
Std. Deviation	1.1E-01	1.1E-01	4.0E-01	4.9E-01	2.6E+00
Std. Error of Mean	3.8E-03	4.4E-03	1.8E-02	3.3E-02	2.9E-01
Lower 95% CI of mean	7.3E-02	1.1E-01	2.9E-01	6.1E-01	1.1E+00
Upper 95% CI of mean	8.8E-02	1.2E-01	3.6E-01	7.4E-01	2.3E+00

Bibliography

Ambros, V. (1989). A hierarchy of regulatory genes controls a larva-to-adult developmental switch in C. elegans. Cell *57*, 49-57.

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.

Altman, M. and King, G. (2007). A proposed standard for the scholarly citation of quantitative data. D-lib Magazine 13(3/4).

Altman, M. and Crosas, M. (2013). The evolution of data citation: From principles to implementation. IASSIST Quarterly 37, 62.

Ambros, V., and Horvitz, H.R. (1984). Heterochronic mutants of the nematode Caenorhabditis elegans. Science *226*, 409-416.

Apffel, A., Chakel, J.A., Fischer, S., Lichtenwalter, K., and Hancock, W.S. (1997). Analysis of Oligonucleotides by HPLC-Electrospray Ionization Mass Spectrometry. Analytical chemistry *69*, 1320-1325.

Armakola, M., Higgins, M.J., Figley, M.D., Barmada, S.J., Scarborough, E.A., Diaz, Z., Fang, X., Shorter, J., Krogan, N.J., Finkbeiner, S.*, et al.* (2012). Inhibition of RNA lariat debranching enzyme suppresses TDP-43 toxicity in ALS disease models. Nature genetics *44*, 1302-1309.

Back, J.W., Notenboom, V., de Koning, L.J., Muijsers, A.O., Sixma, T.K., de Koster, C.G., and de Jong, L. (2002). Identification of cross-linked peptides for protein interaction studies using mass spectrometry and 18O labeling. Analytical chemistry *74*, 4417-4422.

Battye, G. T., Kontogiannis, L., Johnson, O., Powell, H. R., and Leslie, A. G. (2011). iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. Acta crystallographica. Section D, Biological crystallography 67(Pt 4), 271–281 (0).

Beachy, S.H., Onozawa, M., Chung, Y.J., Slape, C., Bilke, S., Francis, P., Pineda, M., Walker, R.L., Meltzer, P., and Aplan, P.D. (2012). Enforced expression of Lin28b leads to impaired T-cell development, release of inflammatory cytokines, and peripheral T-cell lymphoma. Blood 120, 1048–1059.

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide protein data bank. Nature Structural & Molecular Biology 10(12), 980–980.

Berman, H., Kleywegt, G., Nakamura, H., and Markley, J. (2014). The protein data bank archive as an open data resource. Journal of Computer-Aided Molecular Design 28(10), 1009–1014.
Bilderback, D. H., Elleaume, P., and Weckert, E. (2005). Review of third and next generation synchrotron light sources. Journal of Physics B: Atomic, Molecular and Optical Physics 38(9).

Berry, J., Weber, S.C., Vaidya, N., Haataja, M., and Brangwynne, C.P. (2015). RNA transcription modulates phase transition-driven nuclear body assembly. Proceedings of the National Academy of Sciences of the United States of America *112*, E5237-5245.

Bjorkbom, A., Lelyveld, V.S., Zhang, S., Zhang, W., Tam, C.P., Blain, J.C., and Szostak, J.W. (2015). Bidirectional Direct Sequencing of Noncanonical RNA by Two-Dimensional Analysis of Mass Chromatograms. J Am Chem Soc *137*, 14430-14438.

Bourne, P. E., Clark, T. W., Dale, R., de WArd, A., Herman, I., Hovy, E. H., Shotton, D., Bourne, P. E., Clark, T. W., Dale, R., et al. (2011). Improving the future of research communications and e-Scholarship (Dagstuhl Perspectives Workshop 11331). Dagstuhl Manifestos 1(1), 41–60.

Bowers, K., Chow, E., Xu, H., Dror, R., Eastwood, M., Gregersen, B., Klepeis, J., Kolossvary, I., Moraes, M., Sacerdoti, F., Salmon, J., Shan, Y., and Shaw, D. (2006) Scalable algorithms for molecular dynamics simulations on commodity clusters. In SC 2006 Conference, Proceedings of the ACM/IEEE, 43–43.

Castleberry, C.M., Lilleness, K., Baldauff, R., and Limbach, P.A. (2009). Minimizing 18O/16O back-exchange in the relative quantification of ribonucleic acids. Journal of mass spectrometry : JMS *44*, 1195-1202.

Chang, H.M., Triboulet, R., Thornton, J.E., and Gregory, R.I. (2013). A role for the Perlman syndrome exonuclease Dis3l2 in the Lin28-let-7 pathway. Nature 497, 244–248.

Chapman, K.B., and Boeke, J.D. (1991). Isolation and characterization of the gene encoding yeast debranching enzyme. Cell *65*, 483-492.

Chard, K. et al. Globus Data Publication as a Service: Lowering Barriers to Reproducible Science. In e-Science (e-Science), 2015 IEEE 11th International Conference on, 401–410 (IEEE, 2015).

Chen, B.-C., Legant, W. R., Wang, K., Shao, L., Milkie, D. E., Davidson, M. W., Janetopoulos, C., Wu, X. S., Hammer, J. A., Liu, Z., English, B. P., Mimori-Kiyosue, Y., Romero, D. P., Ritter, A. T., Lippincott-Schwartz, J., Fritz-Laylin, L., Mullins, R. D., Mitchell, D. M., Bembenek, J. N., Rey- mann, A.-C., Bhme, R.,

Grill, S. W., Wang, J. T., Seydoux, G., Tulu, U. S., Kiehart, D. P., and Betzig, E. (2014). Lattice light-sheet microscopy: Imaging molecules to embryos at high spatiotemporal resolution. Science 346(6208).

Cho, J., Chang, H., Kwon, S.C., Kim, B., Kim, Y., Choe, J., Ha, M., Kim, Y.K., and Kim, V.N. (2012). LIN28A is a suppressor of ER-associated translation in embryonic stem cells. Cell 151, 765–777.

Clark, N.E., Katolik, A., Roberts, K.M., Taylor, A.B., Holloway, S.P., Schuermann, J.P., Montemayor, E.J., Stevens, S.W., Fitzpatrick, P.F., Damha, M.J.*, et al.* (2016). Metal dependence and branched RNA cocrystal structures of the RNA lariat debranching enzyme Dbr1. Proceedings of the National Academy of Sciences of the United States of America *113*, 14727-14732.

Conklin, J.F., Goldman, A., and Lopez, A.J. (2005). Stabilization and analysis of intron lariats in vivo. Methods *37*, 368-375.

Connolly, B.A., and Eckstein, F. (1984). Assignment of resonances in the 31P NMR spectrum of d(GGAATTCC) by regiospecific labeling with oxygen-17. Biochemistry *23*, 5523-5527.

Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. Cell *136*, 777-793.

Corbett, K. D. and Harrison, S. (2015). X-Ray diffraction data for: S. cerevisiae Csm1-Mam1 complex. PDB code 4EMC. V1, http://dx.doi.org/10.15785/SBGRID/24.

Cordin, O., Banroques, J., Tanner, N.K., and Linder, P. (2006). The DEAD-box protein family of RNA helicases. Gene *367*, 17-37.

Crosas, M. A data sharing story. Journal of eScience Librarianship 1, 7 (2013).

Crosas, M. (2011). The dataverse network® : an open-source application for sharing, discovering and preserving data. D-lib Magazine 17, 2.

Crosas, M., Honaker, J., King, G., and Sweeney, L. (2015). Automating open science for big data. ANNALS of the American Academy of Political and Social Science 659, 260–273.

De, N., and Macrae, I.J. (2011). Purification and assembly of human Argo-naute, Dicer, and TRBP complexes. Methods Mol. Biol. 725, 107–119.

Decker, C.J., Teixeira, D., and Parker, R. (2007). Edc3p and a glutamine/asparagine-rich domain of Lsm4p function in processing body

assembly in Saccharomyces cerevisiae. The Journal of cell biology *179*, 437-449.

Desjardins, A., Yang, A., Bouvette, J., Omichinski, J.G., and Legault, P. (2012). Importance of the NCp7-like domain in the recognition of pre-let-7g by the pluripotency factor Lin28. Nucleic acids research *40*, 1767-1777.

Domdey, H., Apostol, B., Lin, R.J., Newman, A., Brody, E., and Abelson, J. (1984). Lariat structures are in vivo intermediates in yeast pre-mRNA splicing. Cell *39*, 611-621.

Elbaum-Garfinkle, S., Kim, Y., Szczepaniak, K., Chen, C.C., Eckmann, C.R., Myong, S., and Brangwynne, C.P. (2015). The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. Proceedings of the National Academy of Sciences of the United States of America *112*, 7189-7194.

Elsliger, M.-A., Deacon, A. M., Godzik, A., Lesley, S. A., Wooley, J., Wu ̈thrich, K., and Wilson, I. A. (2010). The JCSG high-throughput structural biology pipeline. Acta Crystallographica Section F 66(10), 1137–1142.

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

Evans, P. (2006). Scaling and assessment of data quality. Acta Crystallogr. D Biol. Crystallogr. 62, 72–82.
Evans, P. R. and Murshudov, G. N. (2013). How good are my data and what is the resolution? Acta Crystallographica Section D 69(7), 1204–1214.

Feldkamp, M. D. and Chazin, W. J. (2015). X-Ray diffraction data for: Human RPA32C. PDB code 4OU0. V1. http://dx.doi.org/10.15785/SBGRID/92.

Flory, P.J. (1942). Thermodynamics of high polymer solutions. J. Chem. Phys. 10, 51.

Foster, I. (2005). Globus toolkit version 4: Software for service-oriented systems. In Network and Parallel Computing, Jin, H., Reed, D., and Jiang, W., editors, volume 3779 of Lecture Notes in Computer Science, 2–13. Springer Berlin Heidelberg.

Foster, I. (2011). Globus Online: Accelerating and democratizing science through cloud-based services. IEEE Internet Computing 70–73.

Fraser, J. S. (2015). X-Ray diffraction data for: Cyclophilin a. PDB code 4YUO. V1. http://dx.doi.org/10.15785/SBGRID/68.

Fukuda, T., Yamagata, K., Fujiyama, S., Matsumoto, T., Koshida, I., Yoshimura, K., Mihara, M., Naitou, M., Endoh, H., Nakamura, T.*, et al.* (2007). DEAD-box RNA helicase subunits of the Drosha complex are required for processing of rRNA and a subset of microRNAs. Nature cell biology *9*, 604-611.

Gajadeera, C. S. and Tsodikov, O. V. (2015). X-Ray diffraction data for: Inorganic pyrophosphatase from staphylococcus aureus in complex with mn2+. PDB code 4RPA. V1. http://dx.doi.org/10.15785/SBGRID/22.

Galvis, A.E., Fisher, H.E., Nitta, T., Fan, H., and Camerini, D. (2014). Impairment of HIV-1 cDNA synthesis by DBR1 knockdown. Journal of virology *88*, 7054-7069.

Gilman, A. and JS, M. (2015). X-Ray diffraction data for: Motavizumab and AM14 in complex with prefusion RSV f. PDB code 4ZYP. V1. http://dx.doi.org/10.15785/SBGRID/155.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by sur-face-mapping of phylogenetic information. Bioinformatics 19, 163–164.

Graf, R., Munschauer, M., Mastrobuoni, G., Mayr, F., Heinemann, U., Kempa, S., Rajewsky, N., and Landthaler, M. (2013). Identification of LIN28B-bound mRNAs reveals features of target recognition and regulation. RNA biology *10*, 1146-1159.

Granneman, S., Kudla, G., Petfalski, E., and Tollervey, D. (2009). Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. Proceedings of the National Academy of Sciences of the United States of America *106*, 9613-9618.

Gregory, R.I., Yan, K.P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. Nature *432*, 235-240.

Guss, J. and B, M. (2014). How to make deposition of images a reality. Acta Crystallographica Section D 70(Pt 10).

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M.*, et al.* (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129-141.

Hafner, M., Max, K.E., Bandaru, P., Morozov, P., Gerstberger, S., Brown, M., Molina, H., and Tuschl, T. (2013). Identification of mRNAs bound and regulated by human LIN28 proteins and molecular requirements for RNA recognition. RNA 19, 613–626.

Hagan, J.P., Piskounova, E., and Gregory, R.I. (2009). Lin28 recruits the TUTase Zcchc11 to inhibit let-7 maturation in mouse embryonic stem cells. Nat. Struct. Mol. Biol. 16, 1021–1025.

Halgren, T.A. (2009). Identifying and characterizing binding sites and assess-ing druggability. J. Chem. Inf. Model. 49, 377–389.
Heo, I., Joo, C., Cho, J., Ha, M., Han, J., and Kim, V.N. (2008). Lin28 mediates the terminal uridylation of let-7 precursor microRNA. Mol. Cell 32, 276–284.

Hamasaki, T., Matsumoto, T., Sakamoto, N., Shimahara, A., Kato, S., Yoshitake, A., Utsunomiya, A., Yurimoto, H., Gabazza, E.C., and Ohgi, T. (2013). Synthesis of (1)(8)O-labeled RNA for application to kinetic studies and imaging. Nucleic acids research *41*, e126.

Helliwell, J. R. and Mitchell, E. P. (2015). Synchrotron radiation macromolecular crystallography: science and spin-offs. IUCrJ 2(2), 283–291.

Hennig, S., Kong, G., Mannen, T., Sadowska, A., Kobelke, S., Blythe, A., Knott, G.J., Iyer, K.S., Ho, D., Newcombe, E.A*., et al.* (2015). Prion-like domains in RNA binding proteins are essential for building subnuclear paraspeckles. The Journal of cell biology *210*, 529-539.

Heo, I., Ha, M., Lim, J., Yoon, M.J., Park, J.E., Kwon, S.C., Chang, H., and Kim, V.N. (2012). Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs. Cell 151, 521–532.

Heo, I., Joo, C., Cho, J., Ha, M., Han, J., and Kim, V.N. (2008). Lin28 mediates the terminal uridylation of let-7 precursor MicroRNA. Molecular cell *32*, 276-284.

Heo, I., Joo, C., Kim, Y.K., Ha, M., Yoon, M.J., Cho, J., Yeom, K.H., Han, J., and Kim, V.N. (2009). TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. Cell 138, 696–708.

Heuberger, B.D., Pal, A., Del Frate, F., Topkar, V.V., and Szostak, J.W. (2015). Replacing uridine with 2-thiouridine enhances the rate and fidelity of nonenzymatic RNA primer extension. J Am Chem Soc *137*, 2769-2775.

Hopfner, K.P., Karcher, A., Shin, D., Fairley, C., Tainer, J.A., and Carney, J.P. (2000). Mre11 and Rad50 from Pyrococcus furiosus: cloning and biochemical characterization reveal an evolutionarily conserved multiprotein machine. Journal of bacteriology *182*, 6036-6041.
Huggins, M.L. (1942). Some Properties of Solutions of Long-chain Compounds. J. Phys. Chem. 46, 151–158.

Hunter, J. C. and Westover, K. D. (2015). X-Ray diffraction data for: Human GTPase KRAS G12C bound to GDP. PDB code 4LDJ. V1. http://dx.doi.org/10.15785/SBGRID/158.

Hunter, J. C. and Westover, K. D. (2015). X-Ray Diffraction data for: Human GT-Pase KRAS G12R bound to GDP. PDB Code 4QL3, volume V1. Structural Biology Data Grid. http://dx.doi.org/10.15785/SBGRID/160.

Iliopoulos, D., Hirsch, H.A., and Struhl, K. (2009). An epigenetic switch involving NF-kappaB, Lin28, Let-7 MicroRNA, and IL6 links inflammation to cell transformation. Cell *139*, 693-706.

Janknecht, R. (2010). Multi-talented DEAD-box proteins and potential tumor promoters: p68 RNA helicase (DDX5) and its paralog, p72 RNA helicase (DDX17). American journal of translational research *2*, 223-234.

Janssen, B. J., Read, R. J., Brunger, A. T., and Gros, P. (2007). Crystallography: crystallographic evidence for deviating C3b structure. Nature 448(7154).

Johnson, S.M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., Labourier, E., Reinert, K.L., Brown, D., and Slack, F.J. (2005). RAS is regulated by the let-7 microRNA family. Cell 120, 635–647.

Joosten, R. P., Salzemann, J., Bloch, V., Stockinger, H., Berglund, A.-C., Blanchet, C., Bongcam-Rudloff, E., Combet, C., Da Costa, A. L., Deleage, G., Diarena, M., Fabbretti, R., Fettahi, G., Flegel, V., Gisel, A., Kasam, V., Kervinen, T., Korpelainen, E., Mattila, K., Pagni, M., Reichstadt, M., Breton, V., Tickle, I. J., and Vriend, G. (2009). PDB REDO: automated re-refinement of x-ray structure models in the PDB. Journal of Applied Crystallography 42(3), 376–384.

Kabsch, W. (2010). Integration, scaling, space-group assignment and post-refinement. Acta Crystallogr. D Biol. Crystallogr. 66, 133–144.

Kabsch, W. (2010). XDS. Acta Crystallographica Section D 66(2), 125–132.

Karplus, P. A. and Diederichs, K. (2012). Linking crystallographic model and data quality. Science 336(6084), 1030–1033.

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.

Khalid, M.F., Damha, M.J., Shuman, S., and Schwer, B. (2005). Structure-function analysis of yeast RNA debranching enzyme (Dbr1), a manganese-dependent phosphodiesterase. Nucleic acids research *33*, 6349-6360.

King, C.E., Cuatrecasas, M., Castells, A., Sepulveda, A.R., Lee, J.S., and Rustgi, A.K. (2011). LIN28B promotes colon cancer progression and metas-tasis. Cancer Res. 71, 4260–4268.

King, G. (2007). An introduction to the Dataverse Network as an infrastructure for data sharing. Sociological Methods & Research 36, 173–199.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nature methods *8*, 559-564.

Koonin, E.V. (1994). Conserved sequence pattern in a wide variety of phosphoesterases. Protein science : a publication of the Protein Society *3*, 356-358.

Kramer, K., Sachsenberg, T., Beckmann, B.M., Qamar, S., Boon, K.L., Hentze, M.W., Kohlbacher, O., and Urlaub, H. (2014). Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. Nature methods *11*, 1064-1070.

Kroon-Batenburg, L. M. J. and Helliwell, J. R. (2014). Experiences with making diffraction image data available: what metadata do we need to archive? Acta Crystallographica Section D 70(10), 2502–2509.

Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E., and Cech, T.R. (1982). Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. Cell *31*, 147-157.

Kural, C., Akatay, A. A., Gaudin, R., Chen, B.-C., Legant, W. R., Betzig, E., and Kirchhausen, T. (2015). Asymmetric formation of coated pits on dorsal and ventral surfaces at the leading edges of motile cells and on protrusions of immobile cells. Molecular Biology of the Cell 26(11), 2044–2053.

Landthaler, M., Gaidatzis, D., Rothballer, A., Chen, P.Y., Soll, S.J., Dinic, L., Ojo, T., Hafner, M., Zavolan, M., and Tuschl, T. (2008). Molecular characteriza-tion of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. RNA 14, 2580–2596.

Lee, D. and Raman, C. (2015). X-Ray Diffraction data for: Escherichia coli DOS Br complex. PDB Code 1V9Z, volume V1. Structural Biology Data Grid. http://dx.doi.org/10.15785/SBGRID/137.

Lee, K.H., Zhang, P., Kim, H.J., Mitrea, D.M., Sarkar, M., Freibaum, B.D., Cika, J., Coughlin, M., Messing, J., Molliex, A.*, et al.* (2016). C9orf72 Dipeptide Repeats Impair the Assembly, Dynamics, and Function of Membrane-Less Organelles. Cell *167*, 774-788 e717.

Lehrbach, N.J., Armisen, J., Lightfoot, H.L., Murfitt, K.J., Bugaut, A., Balasu-bramanian, S., and Miska, E.A. (2009). LIN-28 and the poly(U) polymerase PUP-2 regulate let-7 microRNA processing in Caenorhabditis elegans. Nat. Struct. Mol. Biol. 16, 1016–1020.

Lelyveld, V.S., Bjorkbom, A., Ransey, E.M., Sliz, P., and Szostak, J.W. (2015). Pinpointing RNA-Protein Cross-Links with Site-Specific Stable Isotope-Labeled Oligonucleotides. Journal of the American Chemical Society.

Leslie, A. G. W. (1999). Integration of macromolecular diffraction data. Acta Crystallographica Section D 55(10), 1696–1702.

Li, K., Mo, C., Gong, D., Chen, Y., Huang, Z., Li, Y., Zhang, J., Huang, L., Li, Y., Fuller-Pace, F.V.*, et al.* (2017). DDX17 nucleocytoplasmic shuttling promotes acquired gefitinib resistance in non-small cell lung cancer cells via activation of beta-catenin. Cancer letters.

Li, Y.I., Sanchez-Pulido, L., Haerty, W., and Ponting, C.P. (2015). RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. Genome research *25*, 1-13.

Lim, J., Ha, M., Chang, H., Kwon, S.C., Simanshu, D.K., Patel, D.J., and Kim, V.N. (2014). Uridylation by TUT4 and TUT7 marks mRNA for degradation. Cell 159, 1365–1376.

Lin, Y., Protter, D.S., Rosen, M.K., and Parker, R. (2015). Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. Molecular cell *60*, 208-219.

Liu, Z.C., and Ambros, V. (1989). Heterochronic genes control the stage-specific initiation and expression of the dauer larva developmental program in Caenorhabditis elegans. Genes & development *3*, 2039-2049.

Lopez, P.J., and Seraphin, B. (2000). YIDB: the Yeast Intron DataBase. Nucleic acids research *28*, 85-86.

Loughlin, F.E., Gebert, L.F., Towbin, H., Brunschweiger, A., Hall, J., and Allain, F.H. (2012). Structural basis of pre-let-7 miRNA recognition by the zinc knuckles of pluripotency factor Lin28. Nat. Struct. Mol. Biol. 19, 84–89.

Lukong, K.E., Chang, K.W., Khandjian, E.W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. Trends in genetics : TIG *24*, 416-425.

Madison, B.B., Liu, Q., Zhong, X., Hahn, C.M., Lin, N., Emmett, M.J., Stanger, B.Z., Lee, J.S., and Rustgi, A.K. (2013). LIN28B promotes growth and tumorigenesis of the intestinal epithelium via Let-7. Genes & development *27*, 2233-2245.

Martone, M. (2014). Data citation synthesis group: Joint declaration of data citation principles. FORCE11.https://www.force11.org/datacitation

Matange, N., Podobnik, M., and Visweswariah, S.S. (2015). Metallophosphoesterases: structural fidelity with functional promiscuity. The Biochemical journal *467*, 201-216.

Matthews, B. W. (2007). Five retracted structure reports: Inverted or incorrect? Protein Science 16(6), 1013–1016.

Mayr, C., Hemann, M.T., and Bartel, D.P. (2007). Disrupting the pairing be-tween let-7 and Hmga2 enhances oncogenic transformation. Science 315, 1576–1579.

Mayr, F., Schutz, A., Doge, N., and Heinemann, U. (2012). The Lin28 cold-shock domain remodels pre-let-7 microRNA. Nucleic acids research *40*, 7492-7506.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. J. Appl. Cryst. 40, 658–674.

Meng, Z., and Limbach, P.A. (2005). Quantitation of ribonucleic acids using 18O labeling and mass spectrometry. Analytical chemistry *77*, 1891-1895.
Meyer, G. R., Aragão, D., Mudie, N. J., Caradoc-Davies, T. T., McGowan, S., Bertling, P. J., Groenewegen, D., Quenette, S. M., Bond, C. S., Buckle, A. M., and Androulakis, S. (2014). Operation of the Australian Store.Synchrotron for macromolecular crystallography. Acta Crystallographica Section D 70(10), 2510–2519.

Mili, S., and Steitz, J.A. (2004). Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. Rna *10*, 1692-1694.

Mitrea, D.M., and Kriwacki, R.W. (2016). Phase separation in biology; functional organization of a higher order. Cell communication and signaling : CCS *14*, 1.

Molenaar, J.J., Domingo-Ferna´ ndez, R., Ebus, M.E., Lindner, S., Koster, J., Drabek, K., Mestdagh, P., van Sluis, P., Valentijn, L.J., van Nes, J., et al.(2012).

LIN28B induces neuroblastoma and enhances MYCN levels via let-7 suppression. Nat. Genet. 44, 1199–1206.

Montemayor, E.J., Katolik, A., Clark, N.E., Taylor, A.B., Schuermann, J.P., Combs, D.J., Johnsson, R., Holloway, S.P., Stevens, S.W., Damha, M.J.*, et al.* (2014). Structural basis of lariat RNA recognition by the intron debranching enzyme Dbr1. Nucleic acids research *42*, 10845-10855.

Moore, M.J., Zhang, C., Gantman, E.C., Mele, A., Darnell, J.C., and Darnell, R.B. (2014). Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. Nature protocols *9*, 263-293.

Morin, A., Eisenbraun, B., Key, J., Sanschagrin, P. C., Timony, M. A., Ottaviano, M., and Sliz, P. (2013). Collaboration gets the most out of software. eLife 2.

Moss, E.G., Lee, R.C., and Ambros, V. (1997). The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. Cell 88, 637–646.

Moy, R.H., Cole, B.S., Yasunaga, A., Gold, B., Shankarling, G., Varble, A., Molleston, J.M., tenOever, B.R., Lynch, K.W., and Cherry, S. (2014). Stem-loop recognition by DDX17 facilitates miRNA processing and antiviral defense. Cell *158*, 764-777.

Nam, K., Lee, G., Trambley, J., Devine, S.E., and Boeke, J.D. (1997). Severe growth defect in a Schizosaccharomyces pombe mutant defective in intron lariat degradation. Molecular and cellular biology *17*, 809-818.

Nam, Y., Chen, C., Gregory, R.I., Chou, J.J., and Sliz, P. (2011). Molecular basis for interaction of let-7 microRNAs with Lin28. Cell *147*, 1080-1091.

Nannenga, B. L., Shi, D., Leslie, A. G., and Gonen, T. (2014). High-resolution structure determination by continuous-rotation data collection in MicroED. Nature Methods 11(9), 927–930.

Newman, M.A., Thomson, J.M., and Hammond, S.M. (2008). Lin-28 interaction with the Let-7 precursor loop mediates regulated microRNA processing. RNA 14, 1539–1549.

Nguyen, L.H., Robinton, D.A., Seligson, M.T., Wu, L., Li, L., Rakheja, D., Comerford, S.A., Ramezani, S., Sun, X., Parikh, M.S., et al. (2014). Lin28b is sufficient to drive liver cancer and necessary for its maintenance in murine models. Cancer Cell 26, 248–261.

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. Molecular cell *57*, 936-947.

Oivanen, M., Kuusela, S., and Lonnberg, H. (1998). Kinetics and Mechanisms for the Cleavage and Isomerization of the Phosphodiester Bonds of RNA by Bronsted Acids and Bases. Chemical reviews *98*, 961-990.

Okamura, K., Hagen, J.W., Duan, H., Tyler, D.M., and Lai, E.C. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell *130*, 89-100.

Ooi, S.L., Samarsky, D.A., Fournier, M.J., and Boeke, J.D. (1998). Intronic snoRNA biosynthesis in Saccharomyces cerevisiae depends on the lariat-debranching enzyme: intron length effects and activity of a precursor snoRNA. Rna *4*, 1096-1110.

Padgett, R.A., Konarska, M.M., Grabowski, P.J., Hardy, S.F., and Sharp, P.A. (1984). Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. Science *225*, 898-903.

Peng, S., Maihle, N.J., and Huang, Y. (2010). Pluripotency factors Lin28 and Oct4 identify a sub-population of stem cell-like cells in ovarian cancer. Onco-gene 29, 2153–2159.

Permuth-Wey, J., Kim, D., Tsai, Y.Y., Lin, H.Y., Chen, Y.A., Barnholtz-Sloan, J., Birrer, M.J., Bloom, G., Chanock, S.J., Chen, Z., et al.; Ovarian Cancer Asso-ciation Consortium (2011). LIN28B polymorphisms influence susceptibility to epithelial ovarian cancer. Cancer Res. 71, 3896–3903.

Phair, R.D., and Misteli, T. (2000). High mobility of proteins in the mammalian cell nucleus. Nature *404*, 604-609.

Piskounova, E., Viswanathan, S.R., Janas, M., LaPierre, R.J., Daley, G.Q., Sliz, P., and Gregory, R.I. (2008). Determinants of microRNA processing inhibition by the developmentally regulated RNA-binding protein Lin28. The Journal of biological chemistry *283*, 21310-21314.

Piskounova, E., Polytarchou, C., Thornton, J.E., LaPierre, R.J., Pothoulakis, C., Hagan, J.P., Iliopoulos, D., and Gregory, R.I. (2011). Lin28A and Lin28B inhibit let-7 microRNA biogenesis by distinct mechanisms. Cell *147*, 1066-1079.

Polesskaya, A., Cuvellier, S., Naguibneva, I., Duquet, A., Moss, E.G., and Harel-Bellan, A. (2007). Lin-28 binds IGF-2 mRNA and participates in skeletal

myogenesis by increasing translation efficiency. Genes & development *21*, 1125-1138.

Potter, B.V., Eckstein, F., and Uznanski, B. (1983). A stereospecifically 18O-labelled deoxydinucleoside phosphate block for incorporation into an oligonucleotide. Nucleic acids research *11*, 7087-7103.

Qiu, C., Ma, Y., Wang, J., Peng, S., and Huang, Y. (2010). Lin28-mediated post-transcriptional regulation of Oct4 expression in human embryonic stem cells. Nucleic acids research *38*, 1240-1248.

Reyes, F., Rodriguez, J., and Gonen, T. (2015). Micro-Electron diffraction data for: alpha-synuclein. PDB code 4RIL. V1. http://dx.doi.org/10.15785/SBGRID/193.

Ruby, J.G., Jan, C.H., and Bartel, D.P. (2007). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83-86.

Rudenko, G. (2015). X-Ray Diffraction data for: neurexin 1alpha extracellular domain. PDB Code 3QCW, volume V1. Structural Biology Data Grid. http://dx.doi.org/10.15785/SBGRID/78

Ruskin, B., and Green, M.R. (1985). An RNA processing activity that debranches RNA lariats. Science *229*, 135-140.

Rybak, A., Fuchs, H., Smirnova, L., Brandt, C., Pohl, E.E., Nitsch, R., and Wulczyn, F.G. (2008). A feedback loop comprising lin-28 and let-7 controls pre-let-7 maturation during neural stem-cell commitment. Nat. Cell Biol. 10, 987–993.

Sampson, V.B., Rong, N.H., Han, J., Yang, Q., Aris, V., Soteropoulos, P., Petrelli, N.J., Dunn, S.P., and Krueger, L.J. (2007). MicroRNA let-7a down-regulates MYC and reverts MYC-induced growth in Burkitt lymphoma cells. Cancer Res. 67, 9762–9770.

Sauliere, J., Murigneux, V., Wang, Z., Marquenet, E., Barbosa, I., Le Tonqueze, O., Audic, Y., Paillard, L., Roest Crollius, H., and Le Hir, H. (2012). CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. Nature structural & molecular biology *19*, 1124-1131.

Scaringe, S.A., Francklyn, C., and Usman, N. (1990). Chemical synthesis of biologically active oligoribonucleotides using beta-cyanoethyl protected ribonucleoside phosphoramidites. Nucleic acids research *18*, 5433-5441.

Shah, D.O.L.K.G.D.G. (1984). Facile synthesis and phosphorus-31 NMR spectra of a double-labeled oligonucleotide d(Ap(17O)Gp(18O)Cp(16O)T). Journal of the American Chemical Society *106*, 4302-4303.

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. Molecular biology of the cell *16*, 2395-2413.

Sheng, J., Larsen, A., Heuberger, B.D., Blain, J.C., and Szostak, J.W. (2014). Crystal structure studies of RNA duplexes containing s(2)U:A and s(2)U:U base pairs. J Am Chem Soc *136*, 13916-13924.

Shi, D. and Gonen, T. (2015). Micro-Electron diffraction data for: Hen egg white lysozyme . PDB code 3J6K. V1. http://dx.doi.org/10.15785/SBGRID/185.

Shi, D., Nannenga, B. L., Iadanza, M. G., and Gonen, T. (2013). Three-dimensional electron crystallography of protein microcrystals. eLife 2.

Shyh-Chang, N., Zhu, H., Yvanka de Soysa, T., Shinoda, G., Seligson, M.T., Tsanov, K.M., Nguyen, L., Asara, J.M., Cantley, L.C., and Daley, G.Q. (2013). Lin28 enhances tissue repair by reprogramming cellular metabolism. Cell *155*, 778-792.

Sliz, P. (2015). Molecular dynamics trajectory of human O-GlcNAc transferase. PDB code 3PE4. http://dx.doi.org/10.15785/SBGRID/190.

Socias, S., Morin, A., Timony, M., and Sliz, P. (2015). Appciter: A web application for increasing rates and accuracy of scientific software citation. Structure 23(5), 807 – 808.

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R. R., Duerr, R., HAk, L. L., Haendel, M., Herman, I., Hodson, S., Hourcl, J., Kratz, J. E., Lin, J., Nielsen, L. H., Nurnberger, A., Proell, S., Rauber, A., Sacchi, S., Smith, A., Taylor, M., and Clark, T. (2015). Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science 1, e1, 5.

Stefani, G., Chen, X., Zhao, H., and Slack, F.J. (2015). A novel mechanism of LIN-28 regulation of let-7 microRNA expression revealed by in vivo HITS-CLIP in C. elegans. Rna *21*, 985-996.

Stokes-Rees, I., Levesque, I., Murphy, IV, F. V., Yang, W., Deacon, A., and Sliz, P. (2012). Adapting federated cyberinfrastructure for shared data collection facilities in structural biology. Journal of Synchrotron Radiation 19(3), 462–467.

Sun, Z., Diaz, Z., Fang, X., Hart, M.P., Chesi, A., Shorter, J., and Gitler, A.D. (2011). Molecular determinants and genetic modifiers of aggregation and toxicity for the ALS disease protein FUS/TLS. PLoS biology *9*, e1000614.

Tanley, S. W. M., Diederichs, K., Kroon-Batenburg, L. M. J., Schreurs, A. M. M., and Helliwell, J. R. (2013). Experiences with archived raw diffraction images data: capturing cisplatin after chemical conversion of carboplatin in high salt conditions for a protein crystal. Journal of Synchrotron Radiation 20(6), 880–883.

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

Terwilliger, T. C. and Bricogne, G. (2014). Continuous mutual improvement of macromolecular structure models in the PDB and of X-ray crystallographic software: the dual role of deposited experimental data. Acta Crystallograph- ica Section D 70(10), 2533–2543.

Thornton, J.E., Chang, H.M., Piskounova, E., and Gregory, R.I. (2012). Lin28-mediated control of let-7 microRNA expression by alternative TUTases Zcchc11 (TUT4) and Zcchc6 (TUT7). RNA 18, 1875–1885.

Tolia, N. H. (2015). X-Ray diffraction data for: Erythrocyte binding antigen 140. PDB code 4GF2. V1. http://dx.doi.org/10.15785/SBGRID/115.

Triboulet, R., Pirouz, M., and Gregory, R.I. (2015). A single let-7 microRNA bypasses LIN28-mediated repression. Cell Rep. 13, 260–266.

Tu, H.C., Schwitalla, S., Qian, Z., LaPier, G.S., Yermalovich, A., Ku, Y.C., Chen, S.C., Viswanathan, S.R., Zhu, H., Nishihara, R., et al. (2015). LIN28 co-operates with WNT signaling to drive invasive intestinal and colorectal adeno-carcinoma in mice and humans. Genes Dev. 29, 1074–1086.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212-1215.

Upadhyayula, S. and Kirchhausen, T. (2015). Lattice Light-Sheet microscopy data for: Developing zebrafish embryo. http://dx.doi.org/10.15785/SBGRID/187.

Urbach, A., Yermalovich, A., Zhang, J., Spina, C.S., Zhu, H., Perez-Atayde, A.R., Shukrun, R., Charlton, J., Sebire, N., Mifsud, W., et al. (2014). Lin28 sus-tains early renal progenitors and induces Wilms tumor. Genes Dev. 28, 971–982.

Vangone, A. and Bonvin, A. M. (2015). HADDOCK docking models. V1. http://dx.doi.org/10.15785/SBGRID/131.

Viphakone, N., Voisinet-Hakil, F., and Minvielle-Sebastia, L. (2008). Molecular dissection of mRNA poly(A) tail length control in yeast. Nucleic Acids Res. 36, 2418–2433.

Viswanathan, S.R., Daley, G.Q., and Gregory, R.I. (2008). Selective blockade of microRNA processing by Lin28. Science 320, 97–100.

Viswanathan, S.R., Powers, J.T., Einhorn, W., Hoshida, Y., Ng, T.L., Toffanin, S., O'Sullivan, M., Lu, J., Phillips, L.A., Lockhart, V.L., et al. (2009). Lin28 pro-motes transformation and is associated with advanced human malignancies. Nat. Genet. 41, 843–848.

Vonrhein, C., Flensburg, C., Keller, P., Sharff, A., Smart, O., Paciorek, W., Womack, T., and Bricogne, G. (2011). Data processing and analysis with the autoPROC toolbox. Acta Crystallogr. D Biol. Crystallogr. 67, 293–302.

Walker, S.C., Avis, J.M., and Conn, G.L. (2003). General plasmids for pro-ducing RNA in vitro transcripts with homogeneous ends. Nucleic Acids Res. 31, e82.

Wall, M. (2009). Methods and software for diffuse x-ray scattering from protein crystals. In Micro and Nano Technologies in Bioanalysis, Foote, R. S. and Lee, J. W., editors, volume 544 of Methods in Molecular Biology, 269–279. Humana Press.

Wall, M. E., Adams, P. D., Fraser, J. S., and Sauter, N. K. (2014). Diffuse x-ray scattering to model protein motions. Structure 22(2), 182–184.

Wall, M. E., Clarage, J. B., and Jr, G. N. P. (1997). Motions of calmodulin characterized using both bragg and diffuse x-ray scattering. Structure 5(12), 1599 – 1612.

Wall, M. E., Van Benschoten, A. H., Sauter, N. K., Adams, P. D., Fraser, J. S., and Terwilliger, T. C. (2014). Conformational dynamics of a crystalline protein from microsecond-scale molecular dynamics simulations and diffuse x-ray scattering. Proceedings of the National Academy of Sciences 111(50), 17887–17892.

Wallis, T.P., Pitt, J.J., and Gorman, J.J. (2001). Identification of disulfide-linked peptides by isotope profiles produced by peptic digestion of proteins in 50% (18)O water. Protein science : a publication of the Protein Society *10*, 2251-2271.

Wang, L., Zhao, F., Li, M., Zhang, H., Gao, Y., Cao, P., Pan, X., Wang, Z., and Chang, W. (2011). Conformational changes of rBTI from buckwheat upon bind-ing to trypsin: implications for the role of the P(8)0 residue in the potato inhibitor I family. PloS One 6, e20950.

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. Journal of molecular biology *337*, 635-645.

Waterman, D. G., Winter, G., Parkhurst, J. M., Fuentes-Montero, L., Hat- tne, J., Brewster, A., Sauter, N. K., and Evans, G. (2013).The DIALS framework for integration software. CCP4 Newsletter on Protein Crystallography 49, 13–15.

Weber, S.C., and Brangwynne, C.P. (2012). Getting RNA and protein in phase. Cell *149*, 1188-1191.

Wilbert, M.L., Huelga, S.C., Kapeli, K., Stark, T.J., Liang, T.Y., Chen, S.X., Yan, B.Y., Nathanson, J.L., Hutt, K.R., Lovci, M.T., et al. (2012). LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. Mol. Cell 48, 195–206.

Wincott, F., DiRenzo, A., Shaffer, C., Grimm, S., Tracz, D., Workman, C., Sweedler, D., Gonzalez, C., Scaringe, S., and Usman, N. (1995). Synthesis, deprotection, analysis and purification of RNA and ribozymes. Nucleic acids research *23*, 2677-2684.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A., and Wilson, K. S. (2011). Overview of the CCP4 suite and current developments. Acta Crystallographica Section D 67(4), 235–242.

Winter, G., Lobley, C. M. C., and Prince, S. M. (2013). Decision making in xia2. Acta Crystallographica Section D 69(7), 1260–1273.

Welberry, T. (2004). Diffuse X-Ray Scattering and Models of Disorder. International Union of Crystallography Monographs on Crystallography. OUP Oxford.
Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S.*, et al.* (2002). The genome sequence of Schizosaccharomyces pombe. Nature *415*, 871-880.

Xu, B., Zhang, K., and Huang, Y. (2009). Lin28 modulates cell growth and associates with a subset of cell cycle regulator mRNAs in mouse embryonic stem cells. Rna *15*, 357-361.

Ye, Y., De Leon, J., Yokoyama, N., Naidu, Y., and Camerini, D. (2005). DBR1 siRNA inhibition of HIV-1 replication. Retrovirology *2*, 63.

Yeom, K.-H., Heo, I., Lee, J., Hohng, S., Kim, V.N., and Joo, C. (2011). Single-molecule approach to immunoprecipitated protein complexes: insights into miRNA uridylation. EMBO Rep. 12, 690–696.

Yu, F., Yao, H., Zhu, P., Zhang, X., Pan, Q., Gong, C., Huang, Y., Hu, X., Su, F., Lieberman, J.*, et al.* (2007a). let-7 regulates self renewal and tumorigenicity of breast cancer cells. Cell *131*, 1109-1123.

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R.*, et al.* (2007b). Induced pluripotent stem cell lines derived from human somatic cells. Science *318*, 1917-1920.

Zhang, C., and Darnell, R.B. (2011). Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nature biotechnology *29*, 607-614.

Zhang, H., Elbaum-Garfinkle, S., Langdon, E.M., Taylor, N., Occhipinti, P., Bridges, A.A., Brangwynne, C.P., and Gladfelter, A.S. (2015). RNA Controls PolyQ Protein Phase Transitions. Molecular cell *60*, 220-230.

Zhu, H., Shyh-Chang, N., Segre, A.V., Shinoda, G., Shah, S.P., Einhorn, W.S., Takeuchi, A., Engreitz, J.M., Hagan, J.P., Kharas, M.G.*, et al.* (2011). The Lin28/let-7 axis regulates glucose metabolism. Cell *147*, 81-94.