Title

New strategies for RNA crystallization and applications to the pri-miRNA processing machinery

Author

Shoffner, Grant

Publication Date

2018

UNIVERSITY OF CALIFORNIA

Los Angeles

New strategies for RNA crystallization and applications to the pri-miRNA processing machinery

A dissertation submitted in partial satisfaction of the

Requirements for the degree Doctor of Philosophy

In Biological Chemistry

by

Grant Shoffner

2018

ABSTRACT OF THE DISSERTATION

New strategies for RNA crystallization and applications to the pri-miRNA processing machinery

by

Grant Shoffner
Doctor of Philosophy in Biological Chemistry
University of California, Los Angeles, 2018
Professor Feng Guo, Chair

Although the regulatory, signaling, and catalytic functions of RNA are comparable to their protein counterparts, the fact that RNA represent less than 1% of the atomic structures in the Protein Data Bank highlights the gap in our understanding of the RNA structure-function relationship. Overcoming the challenge of RNA structure determination requires innovation in crystallization techniques and further insight into the principles of RNA crystal packing. We propose a scaffold-directed crystallization method for short RNA fragments and employ this technique to determine the three-dimensional conformation of loop sequences from hairpin-structured microRNA (miRNA) precursors. These data reveal common structured elements of the primary miRNA transcripts that may designate the RNA for processing by the Drosha-DGCR8 complex. We further show that mutation of DGCR8, particularly in cancer, can disrupt this process. To study RNA crystal packing, we devise an *in crystal* selection scheme and show that this selection experiment can generate new lattice contacts in RNA crystals. From these contacts we infer strategies for the rational engineering of RNA lattice contacts. Lastly, we explore the RNA-protein interactions involved in the process of X-chromosome inactivation and uncover a unique Xist RNA-binding region in the regulatory factor SHARP.

The dissertation of Grant Shoffner is approved.

Reid C. Johnson

Pascal Franois Egea

Juli F. Feigon

Todd O. Yeates

Feng Guo, Committee Chair

University of California Los Angeles

2018

TABLE OF CONTENTS

VITA		vii
Chapter 1	Overview	1
	References	6
Chapter 2	Functional sites of the DGCR8 Rhed revealed by phylogenetic analysis, protein footprinting, and in vitro evolution	8
	Introduction	8
	Results	10
	Discussion	18
	Materials and Methods	20
	References	36
Chapter 3	Structures of pri-miRNA terminal loops determined by scaffold-directed crystallography	39
	Introduction	39
	Results	42
	Discussion	48
	Materials and Methods	50
	References	77
Chapter 4	Cancer-associated mutations render the DGCR8 protein defective in pri-miRNA processing	79
	Introduction	81
	Materials and Methods	83
	Results	86
	Discussion	92
	References	107
Chapter 5	In crystal selection to establish new RNA crystal contacts	110
	Introduction	112
	Results	114
	Discussion	119
	Methods	122
	References	137
Chapter 6	An unexpected RNA-binding region of SHARP promotes association with the long non-coding RNA Xist	139
	Introduction	139
	Results	141
	Discussion	145
	Materials and Methods	146
	References	157

LIST OF FIGURES BY CHAPTER		
Chapter 1		
Figure 1	Overview of the canonical miRNA maturation pathway	5
Chapter 2		
Figure 1	Alignment and purification of Rhed homologs	27
Figure 2	Secondary structure predictions for all RNA constructs used in this study	28
Figure 3	Analysis of RNA-Rhed complexes by SEC and absorption spectroscopy	29
Figure 4	*In vitro* evolution scheme for thermostabilization of RNA-binding proteins	32
Figure 5	Melting assays for thermostable variants isolated after three rounds of selection	33
Figure 6	Heme and RNA binding analysis of group mutations of acidic residues in the Rhed	34
Figure 7	Phylogenetic conservation of the C-terminal Rhed sequence and secondary structure prediction	35
Chapter 3		
Figure 1	Analysis of pri-miRNA terminal loops and potential crystallization scaffolds	61
Figure 2	Distance difference matrices for C1' atom positions between YdaO W.T. (PDB ID 4QK8) and pri-miRNA loop fusion structures	63
Figure 3	Atomic structures of pri-miRNA terminal loops determined by scaffold-directed crystallography	64
Figure 4	Simulated annealing composite omit maps calculated for all pri-miRNA loops	
Figure 5	Comparison of loop structures with and without the terminal base pair of the pri-miRNA stem	66
Figure 6	Structures of shorter pri-miRNA loops (4-5nt)	67
Figure 7	Secondary structure predictions for all pri-miRNA fragments used in the Rhed binding assay	68
Figure 8	Gel shift assays for pri-miRNA-Rhed binding	70
Figure 9	Association of the Rhed with pri-miRNA apical junctions	71
Figure 10	Estimating the flexibility of the terminal loop with atomic displacement parameters (ADPs)	72
Figure 11	Cluster analysis of MD trajectories for simulations of eight pri-miRNA loops	74
Figure 12	Comparison of apical junction crystal structures to NMR structures	76
Chapter 4		
Figure 1	Analysis of tumor-associated mutations in DGCR8	96
Figure 2	Pri-miRNA processing activity of tumor-derived mutants	98
Figure 3	Biochemical characterization of DGCR8 Rhed mutants	100
Figure 4	Drosha pull-down assay with N-flag-NC1	102
Figure 5	Tumor-derived point-mutations in the Rhed weaken heme binding	103
Figure 6	K289E and G336E decrease protein structural stability, compared to WT DGCR8	106
Chapter 5		
Figure 1	In crystal selection of P4-P6 variants	127
Figure 2	Generation of a new lattice contact in mutant M1	128
Figure 3	Rearranged lattice interaction in the M2 mutant	129
Figure 4	Lattice contact created by a base-stacking interaction in mutant M3	130
Figure 5	M4 weakens the P6b helix, allowing the L6b loop to establish new lattice contacts	131
Figure S1	Distance difference matrix plots between M1-M3 mutant structures and WT P4-P6 (PDB ID 1GID)	135

Figure S2	Lattice interaction in M1 mutant crystals grown under MPD conditions and without cryoprotection	136
Chapter 6		
Figure 1	Organization of the mSHARP protein	150
Figure 2	SEC analysis of complexes between Xist$^{2-304}$ and SHARP fragments under low salt conditions (80mM NaCl)	151
Figure 3	Complexes between Xist$^{334-725}$ and SHARP fragments	152
Figure 4	Re-analysis of Xist$^{2-181}$-SHARP complexes with intermediate salt concentration (150 mM NaCl) and detergent (0.00145% Triton X-100)	153
Figure 5	Confirmation of RNA-protein complex assembly and estimation of the complex stoichiometry	154
Figure 6	Molecular mass estimates for mSHARP proteins based on calibration of the SEC column	156

LIST OF TABLES

Chapter 2		
Table 1	MS analysis of tryptic peptides from the protein footprinting experiment, covering lysine residues K287, K289, K424, K474, and K488	31
Chapter 3		
Table 1	Data collection and refinement statistics for pri-miRNA loop fusion structures	60
Chapter 4		
Table 1	Dissociation constants of NC1 K289E and G336E for pri-miRNAs	105
Chapter 5		
Table 1	Crystallographic data and refinement statistics	133
Chapter 6		
Table 1	Optimized expression conditions for SHARP fragments	147

VITA

2006-2011	B.A. Mathematics University of California, Berkeley
2006-2011	University of California Regent's and Chancellor's Scholar
2006-2011	Honors standing, University of California, Berkeley
2012-2015	NIH/CMB Ruth L. Kirschstein National Research Service Award (GM007185)
2015-2016	Philip J. Whitcome Fellowship

Publications

First Author

Shoffner, G., Wang, R., Podell, E., Cech, T.R., and Guo, F. *In crystal* selection to establish new RNA crystal contacts. *Under review at Structure*

Shoffner, G., Quick-Cleveland, J., Solorio, K.M., Pickard, C., and Guo, F. Tumor-associated Point Mutations in DGCR8 Disrupts Primary MicroRNA Processing. *In preparation*

Shoffner, G., Shirasaki, D., Loo, R., Loo, J., and Guo, F. Functional sites of the DGCR8 Rhed revealed by phylogenetic analysis, protein footprinting, and *in vitro* evolution. *In preparation*

Shoffner, G., Peng, Z., and Guo, F. Structures of pri-miRNA terminal loops determined by scaffold-directed crystallography. *In preparation*

Contributing Author

Sjölander, K., Datta, R.S., Shen, Y., and Shoffner, G. Ortholog identification in the presence of domain architecture rearrangement. Briefings in Bioinformatics 12(5), 2011

Quick-Cleveland, J., Jacob J.P., Weitz, S.H., Shoffner, G., and Guo, F. The DGCR8 RNA-binding Heme Domain Recognizes Primary-miRNAs by Clamping the Hairpin. Cell Reports 7 (6), 2014

Lal, S., Comer, J.M., Konduri, P.C., Shah, A., Wang, T., Lewis, A., Shoffner, G., Guo, F., and Zhang, L. Heme promotes transcriptional and demethylase activities of Gis1, a member of the histone demethylase JMJD2/KDM4 family. Nucleic Acids Research 46(9), 2018

Presentations

RNA Society Meeting 2015
Madison, WI
Poster: "Missense Mutations from Primary Tumor Tissues cause defects in microRNA processing"

UCLA Molecular Biology Institute Retreat 2016
Lake Arrowhead, CA
Poster: "Missense Mutations from Primary Tumor Tissues cause defects in microRNA processing"

Mentorship

Caitlyn Pickard – Undergraduate (2011-2015)

Chely Tejeda – Undergraduate (2011-2017)

Awarded CARES Fellowship for summer term 2016

Chapter 1: Overview

The driving focus of this dissertation is to further our understanding of the molecular interactions governing recognition of the primary miRNA transcript (pri-miRNA) by the Microprocessor complex (MC). What are the molecular features that define the pri-miRNA? How does this protein complex distinguish these features to differentiate an authentic pri-miRNA from the vast majority of competing RNA molecules in the nucleus? Do deficits in the recognition process produce aberrant phenotypes? These questions have guided the field since the realization that MC largely (though not completely) controls the gateway into the RNA interference pathway.

The primary miRNA genes themselves are a fascinating molecular context in which to find the miRNA progenitors. The initial search into miRNA precursors revealed that the mature ~22nt sequences are sequestered within RNA stem-loop structures found in much longer pol II transcripts (Figure 1A) [1]–[3]. Surprisingly, the hairpins could cohabitate transcripts with other coding and non-coding genes; arose in both introns and exons; and had the habit of clustering together as series of polycistronic hairpins, each bearing a different mature miRNA [1]. Few trends were discernable amongst this diversity of transcripts, and the hairpins themselves were highly variable, with no primary sequence homology between different miRNA families and extensive differences in the base-pairing structure of the hairpin.

Isolation of the pri-miRNA processing activity raised further questions. While the ribonuclease Drosha and RNA-binding protein DGCR8 formed the minimal Microprocessor, numerous other RNA processing factors co-fractioned with these components [4]–[8]. Could Drosha/DGCR8 parse the pri-miRNA hairpins independently, or did the array of accessory proteins function to guide MC to various classes of hairpin? In fact, evidence has arisen for both mechanisms. On one hand, the current “molecular clamp” model posits that Drosha/DGCR8 are sufficient for pri-miRNA recognition, as the combination of the two components can authenticate the pri-miRNA by measuring the length of the hairpin stem (Figure 1B)

[9]–[12] . On the other hand, the literature contains a growing list of accessory factors that function to control the biogenesis of one or more miRNAs, and certain sequence elements in mammalian pri-miRNAs can recruit additional processing determinants [13]–[17].

Structural biology promises to unravel these questions by detailing the exact molecular interactions permitting the MC to identify the pri-miRNA substrate. However, our current understanding is limited to crystal structures of fragments of the protein components, which in the absence of RNA can only hint at the recognition structure [18]–[20]. We attacked this gap in our structural understanding of the processing complex in two ways. First, we pursued the co-crystal structure of the RNA-binding heme domain (Rhed) from DGCR8 in complex with the pri-miRNA. As the Rhed functions to selectively identify the junction regions of the RNA (Figure 1B), this structure will be an important step towards learning the recognition principles. Although the structure escaped us, our extensive screening of crystallization constructs outlined several functional features of the protein (Chapter 2).

Second, we sought to determine the structure of the terminal loop and apical junction of the pri-miRNA (Chapter 3). Absent the structure of the Rhed, we reasoned that atomic insights into the pri-miRNA conformation could point to unique molecular features of the RNA. To achieve this goal, we designed a scaffold-directed crystallization approach that enabled the determination of eight terminal loop and junction structures at near atomic resolution (Chapter 3). The results point toward specific three-dimensional conformations in the terminal loop that are shared across all the pri-miRNA we investigated. This raises the possibility that beyond clamping the pri-miRNA hairpin, the MC may also recognize conserved structural elements in the terminal loop.

In addition to the structural approaches, we investigated the role of MC malfunction in human cancers (Chapter 4). This was a collaborative project between myself, graduate student Jen Quick-Cleveland, technician Jose Jacob, and undergraduate Kristina Solorio. Based on the observation that MC is an important suppressor of tumorigenesis [21], we identified multiple missense mutations in the DGCR8 coding sequence from tumor exome sequencing data. Experiments in cultured cells and with purified mutant proteins proved that the mutations can inhibit DGCR8 function through a variety of mechanisms, including disruption of heme-binding and association with RNA. These data suggest that DGCR8 is highly susceptible to genetic inactivation in cancer. Furthermore, identification of structurally important residues in the Rhed deepens our analysis of the Rhed amino acid sequence (Chapter 2).

Although the above results further our insight into the function of the MC, our experience pointedly highlights the major discrepancy in the wealth of tools available to protein crystallographers versus those studying nucleic acids. I utilized a wide array of time-honored (as well as recently published) crystallization strategies to attack the structure of the Rhed. But producing crystals of the pri-miRNA terminal loop necessitated the development of my own RNA crystallization scaffold, with little guidance available in the literature. Seeing that the RNA field sorely needs fresh approaches, I jumped at the opportunity to test an *in vitro* selection scheme for optimization of RNA crystals (Chapter 5). Using a selection strategy devised by Feng Guo during his post-doctoral training, we show that crystallization of a library of mutant RNA molecules is an effective means for generating new intermolecular contacts in the RNA crystal lattice. The new contacts teach us general strategies for engineering similar interactions into other RNA lattices, pushing forward our abilities to optimize or design RNA crystals.

As the miRNA biogenesis field has hit its stride, exciting new data has also surfaced pointing to the widespread transcription of long non-coding RNA (lncRNA) across the genome [22]. The purpose of much of this RNA is hotly contested, but interaction with RNA-binding proteins are sure to mediate the function of many lncRNAs. This proved true for Xist, a lncRNA responsible for transcriptional silencing of one copy of the X chromosome in female cells [23]. In particular, we investigated the Xist interaction with SHARP, a massive 400 kDa protein that bridges Xist to transcriptional repressors (Chapter 6). Through *in vitro* binding assays, we show that SHARP coordinates Xist using not only its RRM domains, but also a previously uncharacterized RNA-binding region in the middle of the protein. Our analysis of Xist-SHARP assembly points to multiple modes of interaction that may play into the function of this complex in X-chromosome inactivation.

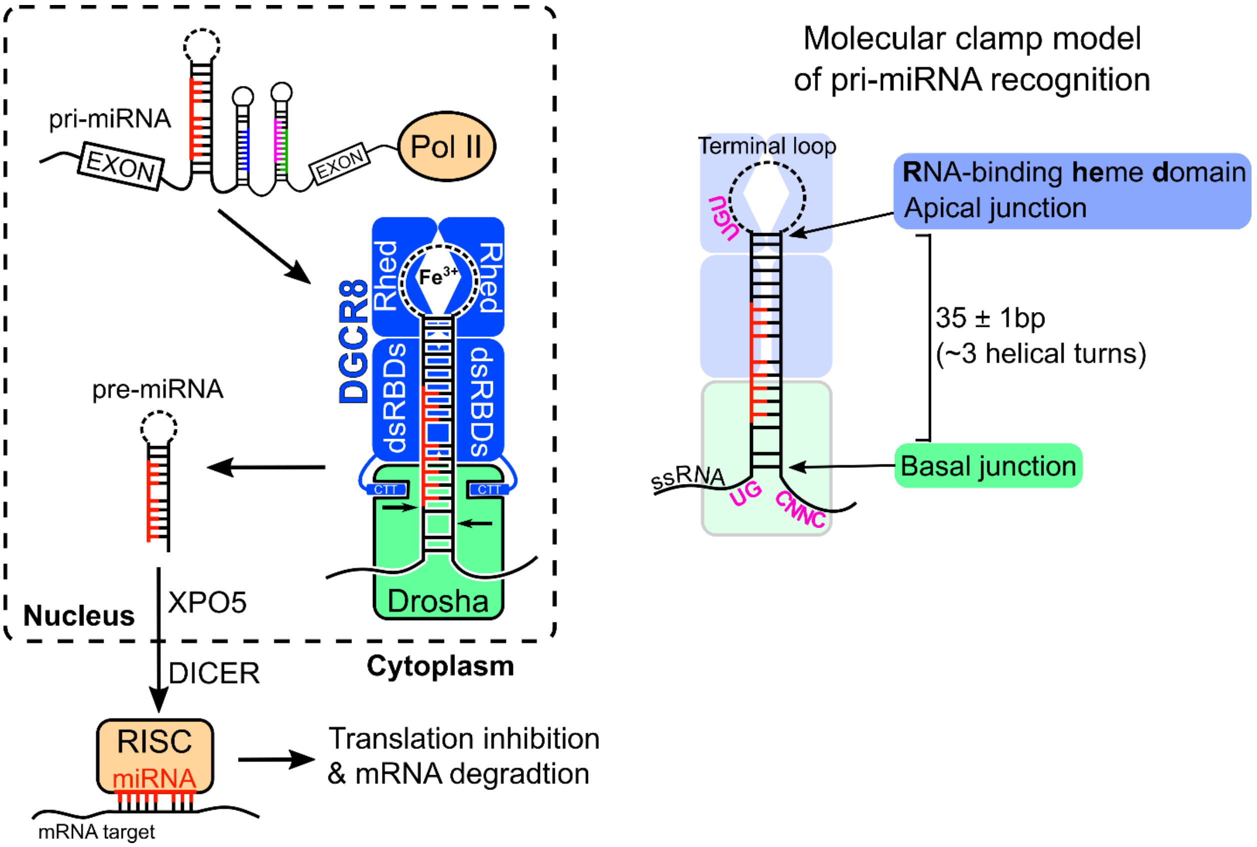


Figure 1. (A) Overview of the canonical miRNA maturation pathway, beginning with transcription of the primary miRNA gene (pri-miRNA). Folding the pri-miRNA into a hairpin structure buries the mature miRNA sequence (red) within the stem. The Microprocessor proteins Drosha (green) and DGCR8 (blue) recognize the hairpin, which requires the Fe^{3+}-heme ligand in the Rhed domain of DGCR8. The Drosha RNase-III domains bind DGCR8 via a C-terminal α-helix (CTT). Cleavage of the hairpin yields the pre-miRNA, which subsequently undergoes nuclear export (XPO5) and further processing (DICER) before loading onto an Argonaut protein in the RNA-induced silencing complex (RISC) to effect repression of cognate mRNAs. (B) The molecular-clamp model predicts that Microprocessor can recognize authentic pri-miRNAs by measuring the length of the hairpin (~35 bp). The two ends of this molecular "ruler" are the DGCR8 Rhed domain and Drosha, which coordinate the apical and basal junctions of the stem-loop, respectively. Additional sequence motifs (pink lettering) at the base and loop can tune the processing efficiency.

References

[1] V. N. Kim, J. Han, and M. C. Siomi, “Biogenesis of small RNAs in animals,” *Nat. Rev. Mol. Cell Biol.*, vol. 10, no. 2, pp. 126–139, Feb. 2009.

[2] Y. Lee, K. Jeon, J.-T. Lee, S. Kim, and V. N. Kim, “MicroRNA maturation: stepwise processing and subcellular localization.,” *EMBO J.*, vol. 21, no. 17, pp. 4663–70, Sep. 2002.

[3] Y. Lee, M. Kim, J. Han, K.-H. Yeom, S. Lee, S. H. Baek, and V. N. Kim, “MicroRNA genes are transcribed by RNA polymerase II,” *EMBO J.*, vol. 23, no. 20, pp. 4051–4060, Oct. 2004.

[4] Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Rådmark, S. Kim, and V. N. Kim, “The nuclear RNase III Drosha initiates microRNA processing,” *Nature*, vol. 425, no. 6956, pp. 415–419, Sep. 2003.

[5] J. Han, Y. Lee, K.-H. Yeom, Y.-K. Kim, H. Jin, and V. N. Kim, “The Drosha-DGCR8 complex in primary microRNA processing.,” *Genes Dev.*, vol. 18, no. 24, pp. 3016–27, Dec. 2004.

[6] A. M. Denli, B. B. J. Tops, R. H. A. Plasterk, R. F. Ketting, and G. J. Hannon, “Processing of primary microRNAs by the Microprocessor complex,” *Nature*, vol. 432, no. 7014, pp. 231–235, Nov. 2004.

[7] R. I. Gregory, K. Yan, G. Amuthan, T. Chendrimada, B. Doratotaj, N. Cooch, and R. Shiekhattar, “The Microprocessor complex mediates the genesis of microRNAs,” *Nature*, vol. 432, no. 7014, pp. 235–240, Nov. 2004.

[8] M. Landthaler, A. Yalcin, and T. Tuschl, “The human DiGeorge syndrome critical region gene 8 and its D. melanogaster homolog are required for miRNA biogenesis,” *Curr. Biol.*, vol. 14, no. 23, pp. 2162–2167, Dec. 2004.

[9] J. Quick-Cleveland, J. Jacob, S. Weitz, G. Shoffner, R. Senturia, and F. Guo, “The DGCR8 RNA-Binding Heme Domain Recognizes Primary MicroRNAs by Clamping the Hairpin,” *Cell Rep.*, vol. 7, no. 6, 2014.

[10] T. A. Nguyen, M. H. Jo, Y.-G. Choi, J. Park, S. C. Kwon, S. Hohng, V. N. Kim, and J.-S. Woo, “Functional Anatomy of the Human Microprocessor,” *Cell*, vol. 161, no. 6, pp. 1374–1387, Jun. 2015.

[11] H. Ma, Y. Wu, J.-G. Choi, and H. Wu, “Lower and upper stem-single-stranded RNA junctions together determine the Drosha cleavage site.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 51, pp. 20687–92, Dec. 2013.

[12] K. M. Herbert, S. K. Sarkar, M. Mills, H. C. Delgado De la Herran, K. C. Neuman, and J. A. Steitz, “A heterotrimer model of the complete Microprocessor complex revealed by single-molecule subunit counting.,” *RNA*, vol. 22, no. 2, pp. 175–83, Feb. 2016.

[13] S. R. Viswanathan, G. Q. Daley, and R. I. Gregory, “Selective blockade of microRNA processing by Lin28,” *Science (80-.).*, vol. 320, no. 5872, pp. 97–100, Apr. 2008.

[14] M. Trabucchi, P. Briata, M. Garcia-Mayoral, A. D. Haase, W. Filipowicz, A. Ramos, R. Gherzi, and M. G. Rosenfeld, “The RNA-binding protein KSRP promotes the biogenesis of a subset of microRNAs,” *Nature*, vol. 459, no. 7249, pp. 1010–1014, Jun. 2009.

[15] M. Ha and V. N. Kim, “Regulation of microRNA biogenesis,” *Nat. Rev. Mol. Cell Biol.*, vol. 15, no. 8,

pp. 509–524, Aug. 2014.

[16] V. C. Auyeung, I. Ulitsky, S. E. McGeary, and D. P. Bartel, “Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing.,” *Cell*, vol. 152, no. 4, pp. 844–58, Feb. 2013.

[17] W. Fang and D. P. Bartel, “The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes,” *Mol. Cell*, vol. 60, pp. 131–145, 2015.

[18] S. Y. Sohn, W. J. Bae, J. J. Kim, K.-H. Yeom, V. N. Kim, and Y. Cho, “Crystal structure of human DGCR8 core,” *Nat. Struct. Mol. Biol.*, vol. 14, no. 9, pp. 847–853, Sep. 2007.

[19] R. Senturia, M. Faller, S. Yin, J. A. Loo, D. Cascio, M. R. Sawaya, D. Hwang, R. T. Clubb, and F. Guo, “Structure of the dimerization domain of DiGeorge Critical Region 8,” *Protein Sci.*, vol. 19, no. 7, pp. 1354–1365, Jul. 2010.

[20] S. C. Kwon, T. A. Nguyen, Y.-G. Choi, M. H. Jo, S. Hohng, V. N. Kim, and J.-S. Woo, “Structure of Human DROSHA.,” *Cell*, vol. 164, no. 1–2, pp. 81–90, Jan. 2016.

[21] M. S. Kumar, J. Lu, K. L. Mercer, T. R. Golub, and T. Jacks, “Impaired microRNA processing enhances cellular transformation and tumorigenesis,” *Nat. Genet.*, vol. 39, no. 5, pp. 673–677, May 2007.

[22] F. Kopp and J. T. Mendell, “Functional Classification and Experimental Dissection of Long Noncoding RNAs.,” *Cell*, vol. 172, no. 3, pp. 393–407, Jan. 2018.

[23] C. A. McHugh, C.-K. Chen, A. Chow, C. F. Surka, C. Tran, P. McDonel, A. Pandya-Jones, M. Blanco, C. Burghard, A. Moradian, M. J. Sweredoski, A. A. Shishkin, J. Su, E. S. Lander, S. Hess, K. Plath, and M. Guttman, “The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3,” *Nature*, vol. 521, no. 7551, pp. 232–236, May 2015.

Chapter 2: Functional sites of the DGCR8 Rhed revealed by phylogenetic analysis, protein footprinting, and *in vitro* evolution

Abstract

In pursuit of the crystal structure of the DGCR8 Rhed domain we generated dozens of protein and RNA variants and performed thousands of crystallization trials. Along the way, we often abandoned mutations or constructs that misbehaved in favor of more promising targets. In hindsight however, the unsuccessful designs themselves help shed light on the function and structure of the protein. This chapter synthesizes results from a variety of structure-oriented experiments to further our insight into the biology of the Rhed. Analyses of four uncharacterized Rhed homologs confirms that the heme- and RNA-binding activities of this domain are conserved across multiple phylogenetic lineages. Using a mass spectrometry-based footprinting approach, we identify three new lysine residues as candidate contacts with the pri-miRNA. Conversely, we show that multiple negatively charged residues in the domain are dispensable for association with heme and RNA. We also detail an *in vitro* evolution scheme for identifying thermostabilizing mutations in the Rhed, and expect this method will be broadly applicable to other RNA-binding domains.

Introduction

The convergence of crystallography and the RNAi pathway proved that structural biology promised deep insights into the function of the RNA processing machinery [1]. The isolation of the Microprocessor complex (MC) ignited the race to determine structures of the factors responsible for cleaving the primary miRNA transcript (pri-miRNA) [2]–[6]. Inspection of the MC components Drosha and DGCR8 already suggested a processing mechanism: the dual RNase-III domains of Drosha would cleave the transcript to yield the pre-miRNA, and DGCR8 would help coordinate the substrate with its pair of double-stranded RNA binding domains (dsRBDs). However, the discovery that DGCR8 bound a heme cofactor [7] immediately complicated the simple model. Not only did the processing machinery associate with a mysterious ligand, but a completely novel heme-binding domain had entered the picture.

In fact, initial biochemical and structural data on the DGCR8 dsRBDs argued this fragment was necessary and sufficient to support processing by Drosha [7], [8]. The structure revealed that the two domains form a tight intramolecular interface, with opposing RNA-binding surfaces that suggested the pri-miRNA hairpin might wrap around the domain [8], [9]. These findings cast doubt on the importance of heme until work in human cells demonstrated a clear requirement for the ligand to activate processing *in vivo* [10]. The crystal structure of an N-terminal fragment of the heme-binding domain, termed the dimerization sub-domain (DSD), revealed the molecular interaction underpinning homo-dimerization of DGCR8 [11]. Dimerization was further connected to assembly of multimeric complexes of DGCR8 on pri-miRNA [12].

The essential role of the heme-binding region was cemented by the finding that this domain also contacts the pri-miRNA, making this the first example of an **R**NA-binding **he**me **d**omain (Rhed) [13]. Multiple studies of the processing complex demonstrated that Rhed functions to recognize the apex of the pri-miRNA stem loop, thus correctly positioning Drosha at the cleavage site near the base of the stem [13], [14]. This model makes additional sense in light of the Drosha crystal structure [15], and is supported by the finding that correct cleavage site selection requires heme [16].

Work from our group has also extensively explored the biochemical basis of the heme-Rhed interaction. While C352 was first proposed as the heme-binding site, additional data proved that a stretch of 4 residues (350-IPCL-353) were essential and constitute a heme-binding motif [7], [10], [17]. One cysteine residue from each copy of the DGCR8 dimer coordinates the Fe^{3+} metal center, likely bonding as the thiolate anion [17], [18]. The iron atom can be swapped for Co^{3+} (as cobalt protoporphyrin IX) while retaining activity of the protein [19]. A broad analysis of heme-binding proteins further suggests that multiple aromatic side chains line the heme-binding cleft for interaction with the porphyrin ring [20].

Determining the atomic structure of the Rhed has been a longstanding goal of our laboratory. To this end, we have dug deep into the crystallographer's toolbox in our search for diffraction-quality Rhed crystals. This included not only optimization of the protein construct, but also screening of a variety of cognate pri-miRNA fragments for co-crystallization. Although not designed to address specific scientific questions, many experiments aimed at generating crystals also revealed interesting details about the function of the protein. Below, we highlight results from several structure-oriented experiments that serendipitously informed our understanding of the Rhed and RNA-Rhed complex.

Results

Conservation of heme-binding across Rhed homologs

Because evolution strongly conserves protein structure—even between sequences with low overall amino acid identity—screening of homologous proteins is an effective tactic for obtaining crystals. Our group has extensively explored DGCR8 sequence space in pursuit of the Rhed crystal structure. Previous work has identified the Rhed sequences from frog (*Xenopus laevis*) and bat star (*Patiria miniata*) as highly soluble, heme-bound proteins amenable to expression in bacteria [21]. To this list we added homologous proteins from Zebrafish (*Danio rerio*), the body louse (*Pediculus humanus*), the tunicate *Oikopleura dioica*, and the tardigrade *Ramazzottius varieornatus* (also known as the "water bear"). In addition, we also tested sequences from worm (*Caenorhabditis elegans*), shrimp (*Litopenaeus vannamei*), and the brown plant hopper (*Nilaparvata lugens*), although we found these were either poorly soluble or heavily proteolyzed (data not shown).

Comparison of the Rhed sequences from these proteins supports our previous analyses of functional residues in this domain (Figure 1A). Within the dimerization sub-domain (residues 276-352), the WW motif (marked with asterisks in Figure 1A) is highly conserved, but the second tryptophan is mutated to leucine in the louse and water bear sequences. Importantly, the heme-binding site (green residues in Figure 1A) is almost perfectly conserved across all homologs. The only exception is from louse, which contains the I350V mutation (human numbering). Residue F448 (marked with a green dot in Figure 1A) is also perfectly conserved, consistent with our observation that this residue is crucial for heme-binding (see Chapter 3). Seeking to identify the RNA binding surface of the domain, we previously identified basic residues which contact the pri-miRNA-substrate (red dots in Figure 1A) [13]. These residues are also conserved between our homologs.

To investigate their biochemical and crystallization properties, we purified these new proteins to homogeneity (Figure 1B). As the human Rhed co-purifies with Fe^{3+}-heme from bacterial cultural, we looked for heme-binding in the purified homologs by electronic absorption spectroscopy (Figure 1C). The diagnostic features of the heme-bound Rhed spectrum are the split-Soret bands at 366 nm and 450 nm, along with a minor peak near 510 nm. Despite variability in the degree of heme loading during over expression, all of the homologous proteins display the characteristic split-Soret bands. The absorbance bands closely match the human spectrum, but for the Zebrafish (drRhed) we see a slight but reproducible shift of the second Soret band to 452 nm. In addition, the water bear spectrum contains a shoulder on the 450 nm peak near 420 nm. We have speculated this spectral feature is associated with a small degree of non-specific heme-binding, so it is uncertain whether this reflects the true absorption spectrum or an artifact from bacterial expression. Nonetheless, the water bear protein clearly binds heme in a similar manner to other Rhed domains.

Our experience with the human Rhed indicated that the central acidic loop is dispensable for heme binding. We also generated a loop deletion construct for the Zebrafish protein (drRhedΔloop) and purified the protein (Figure 1B). The heme-bound absorption spectrum for this construct further confirms that the central loop does not participate in the association with heme (Figure 1C).

The heme-binding results shed new light on our sequence alignment. The crystal structure of the dimerization sub-domain highlighted a central role for the WW motif in protein folding [11]. Therefore, we previously reasoned that dimerization and heme binding would not be conserved in homologs lacking either tryptophan residue. The W329L mutation in both the louse (phRhed) and water bear sequences clearly illustrates that the structure can accommodate a significantly smaller side chain at this position while maintaining heme binding. Furthermore, the I350V substitution in phRhed also proves the IPCL heme-binding motif can accept conservative substitutions in the first position. Lastly, the *Oikopleura* and water bear sequences contain long insertions of highly charged residues in the C-terminus of the domain. As these sequences are likely unstructured, this may highlight a particularly flexible region of the domain that can accept a solvent-exposed loop, similar to the central acidic loop found in all Rhed homologs.

RNA-binding activity of Zebrafish and Louse Rhed homologs

Our current model for pri-miRNA recognition predicts that DGCR8 functions to anchor the Microprocessor complex at the apex of the pri-miRNA stem. The Rhed facilitates anchoring by binding to the double-stranded to single-strand junction between the stem and the terminal loop. To investigate whether recognition of the junction is a conserved function of the Rhed, we tested binding of drRhed to various pri-miRNA fragments. Secondary structure predictions for all RNA used are shown in Figure 2. We used size-exclusion chromatography to analyze the binding reactions, and varied the input ratio of protein to RNA to estimate the stoichiometry of the complexes. For these experiments, we monitored the absorbance at both 260 nm and 450 nm so that co-elution of peaks at both wavelengths reflects assembly of an RNA-Rhed complex.

For a fragment of human pri-miR-23a containing the terminal loop plus 25bp of the stem (pri-miR-23a-A), we observed clear binding with the dsRhed (Figure 3A). Titration of the reaction with dsRhed showed this complex contained two copies the Rhed dimer per RNA. This is surprising given that we expect this RNA to contain a single Rhed binding site at the apical junction. However, when we truncate the stem by only three base pairs (pri-miR-23a-C) and repeat the assay with 1:2 (RNA:protein) ratio we observe two peaks corresponding to bound RNA and free protein (Figure 3B, peaks marked with an asterisk). Reducing the RNA:protein ratio to 1:1 yields a single peak (Figure 3C). Further truncation of the stem by an additional four base pairs (pri-miR-23a-D) does not change the binding ratio (Figure 3E and 3F). Extending the protein construct to cover both the Rhed and dsRBDs (drNC14) does not change the stoichiometry of the complex (Figure 3D). These results suggests that drRhed can not only recognize the apical junction, but can also associate with the double-stranded stem or blunt end at the base of pri-miR-23a-A.

The 1 RNA: 2 protein binding observed for pri-miR-23a-A calls into question our model for the Rhed-RNA interaction, as it may point to unappreciated binding sites on sufficiently long RNA stems. To refute this notion, we designed a RNA construct containing the pri-miR-23a hairpin but capped with a FAB binding site (pri-miR-23a-FAB) [22]. At the base of the stem, we imitated the basal junction by adding short single-stranded segments at the 5' and 3' ends (yellow residues in Figure 2D). We reasoned that tight binding of the FAB at terminal loop would completely block association with the Rhed with any feature at the apex of the stem. Therefore, binding stoichiometry above 1:1 would necessarily reflect binding the protein to the stem region. We first confirmed assembly of the FAB-RNA complex (Figure 3K). We then added human NC1 protein (containing the Rhed, dsRBDs, and C-terminal tail) to the preformed complex. We only observed 1:1 binding for hsNC1 with the RNA (Figure 3L). This strongly argues that the higher stoichiometry complex observed for drRhed with pri-miR-23a-A is an artifact specific to this RNA construct.

Next, we tested whether drRhed and drRhedΔloop could also recognize the basal junction. We employed two constructs as models of the basal junction: 30a-BJ1 consists of two annealed strands covering the pri-miR-30a basal junction with ~20bp of the stem and terminating in a blunt end; 30a-BJ2 is a single strand covering the junction and 8 bp of stem, capped with a GAAA tetraloop. Mixing drRhed with 30a-BJ1 resulted in a single peak for the 1:1 complex (Figure 3G). Likewise, drRhedΔloop bound 30a-BJ2 with the same stoichiometry (Figure 3H). This confirms that drRhed can recognize the basal junction, even when only a short fragment of stem is available for binding. Furthermore, Rhed maintains its junction binding function when we remove the central loop.

Since the drRhed is similar to the human sequence (>70% identity), we also explored whether the more distantly related louse protein (phRhed) could bind pri-miRNA. When mixed with a 186 nt fragment containing the entire pri-miR-30a hairpin, phRhed bound with the expected 1:2 RNA:protein ratio for binding of the domain at both apical and basal junctions (Figure 3I).

Encouraged by our binding results, we performed a preparative SEC run of the pri-miR-30a-drRhed complex (1 RNA: 2 protein), and concentrated the peak fractions for crystallization. Interestingly, the absorption spectrum of the concentrated material displayed some anomalous features (Figure 3J). While the RNA absorption at 260 nm dominates the spectrum, the visible region reveals unexpected peaks. The peak at 369 nm closely matches the first Soret band at 366 nm, but we also observe overlapping peaks at 420 nm and 450 nm. Although difficult to interpret, this may indicate remodeling of the heme-protein interaction upon RNA binding or loss of the heme ligand during purification.

Protein footprinting identifies residues from the Rhed contacting the pri-miRNA

To further dissect the RNA-Rhed interaction we employed a protein footprinting approach analyzed by mass spectrometry. We prepared either the free hsRhed or 1:1 complex with pri-miR-23a-D and briefly exposed the samples to NHS-biotin. This reagent selectively reacts with primary amines at the N-terminus or within lysine side chains. Following SDS-PAGE and trypsin digestion, the mass shift of biotinylated peptides can be detected by mass spectrometry. We expect intermolecular contacts between lysine and the RNA to shield these residues from labeling, relative to solvent exposed lysine. There are 18 lysine residues in hsRhed, and in the free protein we observed biotinylation of all but 5 positions (K356, K357, K408, K431, and K472). As these residues are outside of the dimerization sub-domain, we do not have structural information available on their conformation. Protection of these residues from the labeling reagent may point to salt-bridge interactions in the protein structure. However, the absence of peptides from the spectra can also indicate these are poorly volatized during electro-spray ionization and hence not detected by MS.

For the remaining lysine residues, we limited our analysis only to residues for which we detected labeling in multiple peptides and with strong signal intensity (>10000 counts for at least one ion). Across two replicate MS experiments we observed clear protection of five residues by the RNA (Table 1). These are K287/K289 in the dimerization sub-domain, and K424/K478/K488 in the C-terminal region. We previously demonstrated the role of K424 in RNA binding by mutagenesis of this site, which supports our footprint procedure. Interestingly, the remaining four residues have not been linked to pri-miRNA recognition and point to an unappreciated RNA-binding surface.

In vitro selection of thermostable Rhed variants highlights additional functional sites

Frustrated with recalcitrant RNA-Rhed co-crystallization trials, we looked for a creative solution to optimize our crystallization constructs. The C-terminal region of the Rhed is less well conserved than the dimerization sub-domain, and contains many charged residues. This suggested to us that the C-terminal fragment might be a weakly structured or metastable portion of the protein. If true, the dynamic state of the domain or presence of multiple conformations could poison our crystallization experiments. In an attempt to stabilize the C-terminal region, we devised an *in vitro* evolution strategy (Figure 4A).

Our basic strategy was to engineer the pri-miR-23a hairpin into the 3’-UTR of the bacterial expression construct for human NC1, a truncation of DGCR8 beginning at the Rhed that retains RNA processing activity. This enables linkage of the protein with its own coding sequence via tight binding of NC1 to the pri-miR-23a hairpin in the mRNA. Starting from NC1, we replaced the acidic loop and C-terminal fragment of the Rhed with SpeI/AgeI restriction sites. The in-frame, six base pair restriction sites introduce a non-natural dipeptide sequence at either end of the C-terminal region (TS for SpeI and TG for AgeI). We predicted these modifications would not disrupt the structure since they fall within predicted flexible regions of the protein. We then performed error-prone PCR to mutagenize the C-terminal fragment and ligated this library into the SpeI/AgeI sites (Figure 4B). Sequencing clones from the initial library confirmed that most contained 1-2 mutations (data not shown).

Next, we overexpressed the library in bacteria and purified the RNA-protein complexes. We confirmed the presence of both RNA and NC1 in the purified fraction by absorption spectroscopy, which clearly showed the 450 nm heme peak and a strong signal near 260 nm from the RNA (Figure 4C). SDS-PAGE confirmed the protein was full-length and well purified (Figure 4E). We then heated the RNA-protein library and pelleted denatured material. We recovered the thermostable complexes by affinity purification with anti-FLAG resin. We also monitored the purification by RT-PCR and confirmed RT-dependent amplification of the RNA in our purified complexes (Figure 4D).

Finally, we RT-PCR amplified the C-terminal fragment from the thermostable complexes and used this as input for further selection cycles. Over three rounds of selection we incremented the denaturation temperature from 55°C to 63°C. After sequencing multiple clones from the final library, we selected a few candidate mutations for closer inspection. We measured the thermostability of the clones in two ways, by (1) size exclusion chromatography with fluorescence detection (F-SEC), and (2) monitoring melting of the heme peak by spectroscopy. For F-SEC detection, we tagged the clones with a C-terminal mCherry domain. As mCherry has a high melting temperature (>75°C), we reasoned this group would not interfere with the melting assay. Purified mCherry-tagged proteins were heated at varying temperatures and insoluble protein removed by ultracentrifugation; the soluble fraction was analyzed by F-SEC.

For W.T. NC1Δloop-mCherry we observed complete denaturation of the protein by 60°C (Figure 5A). However, a selected mutant containing a lysine insertion after V455 (V455insK) retained substantial soluble protein even at 60°C (Figure 5B). Comparison of the F-SEC melting curves revealed that V455insK and the triple mutant V423A/V429E/D436E improved survival of the protein at high temperatures (Figure 5C). To corroborate these data by spectroscopy, we analyzed the mutations in the context of RhedΔloop-mCherry, which helps reduce strong background scattering seen for melting of NC1 proteins (data not shown). Tracking the intensity of the 450 nm peak, we found that W.T. RhedΔloop loses heme-binding with a transition temperature ($T_{1/2}$) near 62°C (Figures 5D and 5F). Although we did not see a shift in the transition temperature for V423A/V429E/D436E, we found that V455insK melts approximately 4°C above the W.T. (Figures 5E and 5F).

Counter to our intuition, both clones V423A/V429E/D436E and V455insK create additional charged sites within the already heavily charged Rhed sequence. In the case of V423A/V429E/D436E this does not destabilize the protein, and V455insK actually increases resistance to precipitation (F-SEC assay) and thermostability of the heme-protein interaction. Both V429E and V455insK fall within segments of the Rhed with moderate to high levels of evolutionary conservation, but these substitutions do not appear amongst our Rhed homologs. We note that V455insK is in close proximity to F448, a residue we know to be critical for heme binding. The boost in stability of the heme complex by V455insK suggests that the stretch of residues downstream from F448 are also involved in the interaction with heme, and that V455insK may tweak the structure to enhance binding.

Bulk mutation of acidic residues in the Rhed C-terminal region

We have long suspected that the high frequency of both negatively and positively charge side chains in the C-terminus of Rhed make this domain especially prone to aggregation. This was borne out in our crystallization trials, where the untagged protein formed amorphous precipitates in most conditions, even at concentrations as low as 2 mg/ml. As our data shows many of the positively charged residues in this region are important for RNA binding, we attempted to reduce the negatively charged amino acid content instead. We constructed group mutants that collectively substituted serine for 13 aspartate or glutamate residues: E438S/D432S/E433S (AG1), D436S/E438S/E439S (AG2), E445S/D449S/E451S (AG3), E463S/E470S (AG4), and E477S/E479S (AG5). The targeted acidic positions are highlighted with a blue triangle in Figure 1A.

The group mutants all retain normal heme-binding activity, as confirmed by the absorption spectra of the purified hsRhed proteins (Figure 6A). We do not observe any shift of the Soret bands, which is a strong indication the heme-binding pocket remains intact in all mutants. We next tested if the mutants could recognize their pri-miRNA partner. We mixed each mutant with the 186nt pri-miR-30a construct at a 1 RNA: 2 protein ratio and resolved the complex by SEC (Figure 6B-F). While all mutants bound the pri-miR-30a, we found that several complexes were not stable over the SEC run. In particular, peaks for the free protein near 13 mL are clear for AG2 and AG5 (Figures 6C and 6F, respectively), and to a lesser extent for AG3 and AG4 (Figures 6D and 6E). We speculate that AG2 and AG5 partially disrupt local protein structure, reducing the contact of neighboring lysine or arginine side chains with the RNA.

Discussion

Our exploration of the Rhed family tree has surfaced several insights. As for heme binding, we have revealed a degree of plasticity in sequence features previously thought to be invariable. In particular, the second tryptophan of the WW motif and isoleucine position of the IPCL heme-binding segment can be conservatively mutated without abolishing heme-binding activity. Furthermore, we have demonstrated that diverse Rhed homologs can recognize single-stranded to double-stranded RNA junctions, confirming that this is an evolutionarily conserved function of the domain.

Comparison of our phylogenetic, footprinting, and mutagenesis data supports a new model for the structure of the Rhed C-terminal region (Figure 7A). Predicting the secondary structure of this region with PSIPRED detects several helical and β-strand elements [23]–[25]. We analyzed the helical portions by constructing helical wheel diagrams (Figure 7B and 7C). The projection of Helix 1 (residues 437-446) indicated a possible hydrophobic surface. When we extended the predicted helix by 4 aa (covering 437-450, dashed line in Figure 7A) we observed that residues F448 and F450 also localize on the hydrophobic face of the helix. The positioning of multiple charged side chains on the opposite face strongly indicates this is an amphipathic α-helix. Although R447 sits on the hydrophobic side, we note this residue is poorly conserved and that V, L, and I amino acids are common substitutions at this position. The amphipathic character of this helix also explains why mutagenesis of four of the negatively charged side chains (E436/E439/E445/D449) did not disrupt protein folding, since these occupy the solvent exposed face of the helix. Inclusion of the essential F448 in the hydrophobic region argues this portion of the helix sits in close proximity to the heme ligand.

The projection for Helix 2 shows that nearly every aspect of the helix presents a charged or polar side chain (Figure 7C). This helix contains K474, which is protected by RNA in our footprinting assay. Furthermore, the helix localizes three negative charges (E463/E470/E477) adjacent to K474, which could explain why mutation of these residues modestly interferes with RNA-binding. Collectively, this suggests that Helix 2 creates a solvent-exposed surface for interaction with RNA. Lastly, we note the presence of a PxxP motif immediately downstream of Helix 2. SH3 domains are known to bind PxxP elements [26]. If this is a protein interaction site, it supports the unstructured conformation of residues between Helix 2 and the terminal β-strand.

Using *in vitro* evolution, we were able to identify a mutation that increases the thermostability of the Rhed (V455insK) and falls within a predicted β-strand of the structure. A new structure prediction for the mutant retains this β-strand (not shown), which is consistent with the insertion maintaining local structure. It will be interesting to learn the mechanism by which this mutation enhances stability.

On a final note, in addition to the above experiments we tried an assortment of other techniques for growing Rhed crystals. One popular strategy we extensively explored is the use of maltose binding protein (MBP) as a fusion partner, connected to the Rhed N-terminus via a fixed α-helical linker [27]. This method drastically improved the solubility of the Rhed in crystallization trials (we routinely screened MBP fusions at 20 mg/ml). Furthermore, addition of mCherry at the C-terminus (MBP-Rhed-mCherry), allowed for crystallization screening up to 50 mg/ml protein. However, given that the structure of the Rhed N- and C-termini are unknown, engineering fusions with stable conformations is difficult guesswork. We also constructed an array of in-line protein fusions based on a method developed by Ray Stevens' group [28]. The idea is to replace internal loops of the target protein with a panel of fusion partners that vary in their N- to C-terminal spacing. Nearly every construct we tested from this series was insoluble or did not bind heme. We also explored multiple RNA engineering ideas [29], although we suspect these could not compensate for the misbehavior of the protein partner.

Materials and Methods

Construction of multiple sequence alignment of Rhed homologs

The human Rhed sequence (residues 276-498) was used in a PSI-BLAST search against the NCBI non-redundant (nr) protein sequence database with default settings [30]. After multiple PSI-BLAST rounds, all hits with significant E values were retained and downloaded. To prepare an initial alignment, we used the LINSI setting ("localpair") in MAFFT with max iterations set to 1000 [31]. As large groups of highly similar sequences can bias the alignment, we next filtered the sequences with Belvu to remove redundant sequences with amino acid identity greater than 80%. Select sequences of interest were manually restored, and the alignment of the filtered sequences was repeated in MAFFT. We show only select sequences from this finalized alignment in Figure 1A.

Cloning and characterization of Rhed homologs

Zebrafish Rhed (drRhed) and NC14 (drNC14, equivalent to human DGCR8$^{276\text{-}726}$) were cloned from a Zebrafish cDNA library (kind gift from Alvaro Sagasti, UCLA) using the forward primer 5'-GCATATGGACGGGGAGGCCGGAGTT-3' and reverse primer 5'-CAGGCGGCCGCCTATGTGGGGGCATCTTGAACAG-3' (drRhed) or 5'-CAGGCGGCCGCTCACTGCTGAAGCTCAATCACAC-3' (drNC14) and cloned into the NdeI/NotI sites of pET17b. Clones containing the insert were verified by Sanger sequencing. We constructed drRhedΔloop (deletion of S368-P402, Zebrafish numbering) using a four-primer PCR strategy and our sequenced clone as template. The forward and reverse primers above were the outside primer set, and internal primers were 5'-GGTCAAGATCAATCTCCGCTGTCGGAG-3' and 5'-GCGGAGATTGATCTTGACCTCATGGAGC-3'.

The *Oikopleura* Rhed sequence (Y270-P489 in the *Oikopleura* numbering) was amplified from a cDNA library (kind gift of Cristian Cañestro, Univ. of Barcelona) with the forward and reverse primer sequences 5'- CAGCCATATGTATGAATACTACACAAAAAGCATTGTGAGG-3' and 5'-ATGCGGCCGCCTATGGTCTGCCTTTTTGCTC-3'. The PCR product was cloned into the NdeI/NotI sites of pET17b and the final clone was sequence verified.

The louse (phRhed) and water bear Rhed sequences were codon optimized for bacterial expression and synthesized as double-stranded gene fragments (Integrated DNA Technologies) with restriction sites added to the termini. These were cloned into pET17b as above and the sequences confirmed by Sanger sequencing.

All proteins were over-expressed and purified as previously described for the human Rhed using a combination of ion exchange and size-exclusion chromatography [13], [32]. Absorption spectra were recorded at room temperature on a Cary 300 Bio UV-Visible spectrophotometer (Varian).

Preparation of RNA and the RNA-binding FAB

All pri-miRNA constructs except 30a-BJ2 and pri-miR-23a-FAB were cloned into pUC19, transcribed, and purified as previously described [13]. Preparation of 30a-BJ2 was identical, except a synthetic oligonucleotide containing the T7 promoter and RNA sequence was used as the transcription template. We generated a transcription template for pri-miR-23a-FAB from our pri-miR-23a-L1 clone (~150 bp fragment containing the full hairpin plus flanking regions) by four-primer PCR. In addition, we added a Hammerhead ribozyme to at the 3’ end of the RNA. The forward primer was 5’-TAATACGACTCACTATAGGAGAGCCACGGCCGGCTGGGGTTCCTGGGGAGGAAACACC-3’, middle primers were 5’- CCGTCGAACTGCTCAGGGTCGGTTGGAAATCCCTGGCAAGGTGTTTCCTCCCCAGGAAC-3’, and 5’-GACCCTGAGCAGTTCGACGGAGTCTAGACTCCGTCCTGATGAGTCCGTGAGG-3’, and the reverse primer was 5’-CCTCTGCAGGCAGTTTTCGTCCTCACGGACTCATCAGGAC-3’. The final PCR product was used as the transcription template, and the remaining transcription and purification steps were identical to the other RNA constructs.

The RNA-binding FAB was over-expressed in BL21(DE3)-RIPL cells (Agilent) following the previously described protocol [22].

SEC analysis of RNA-protein complexes

RNAs were annealed in 100 mM NaCl, 20 mM Tris pH8.0 by heating to 65 °C for 3 min followed by snap cooling on ice. RNA binding reactions (150 µL) contained ~5 µM RNA, 150 mM Nacl, 20 mM Tris pH8.0, and variable concentration of protein (either 5 or 10 µM for 1:1 or 1:2 binding ratios, respectively). The reactions were incubated at room temperature for 20 min. 100 µL of the reaction was injected on a

Superdex 16/30 S200 SEC column (GE Healthcare) at 0.5 mL/min. The running buffer consisted of 80 mM NaCl and 20 mM Tris pH8.0. For the preparative scale experiment, the RNA-protein complex was purified over a HiPrep Sephacryl 26/60 S200 column (GE Healthcare) and the peak fractions concentrated in an Amicon 15 centrifugal filter device (30 kDa MWCO). The concentrated material was analyzed by absorption spectroscopy as above.

Protein footprinting of the RNA-Rhed complex

The 5 µM pri-miR-23a-D-Rhed complex was prepared as above, except the binding reaction contained phosphate buffer instead of Tris. The protein and RNA-protein complexes were exposed to 1 mM NHS-Biotin (ThermoFisher, dissolved in DMSO) for 5 min at room temperature before the reaction was quenched by the addition of 50 mM Tris pH8.0. The samples were mixed with SDS loading dye and run on a 12% polyacrylamide gel. The gel was Coomassie stained and the Rhed band cut out of the gel. The gel slices were destained and an in-gel trypsin digest performed following standard protocols. Peptides were purified by C18 ZipTip (Varian) and dried in a centrifugal evaporator. Two replicate mass spectrometry runs were performed using both MS^{E} and DDA (data-dependent analysis). Results were mined using custom scripts written in Python.

In vitro selection of thermostable Rhed variants and melting temperature determination

The fusion of the human NC1 and pri-miR23a-L1 sequences was accomplished by four-primer PCR. The NC1 fragment was amplified in two steps, first with 5'-GATTACAAGGATGACGACGATAAGGATGGAGAGACAAGTGTGCAG-3'and 5'-CAGGGGTGCCCTACTTTCGAGTCTCCTCCCTTTC-3' (R1). The second round used R1 along with 5'-CCGCATATGGATTACAAGGATGACGACGATAAG-3' (F2). The pri-miR-23a-L1 fragment was amplified with primers 5'-GACTCGAAAGTAGGGCACCCCTGTGCCACGG-3' and 5'-CTGAATTCGCCACCCCGTCCCCGGG-3' (R2). The products were gel purified and then assembled together using the F2 and R2 primers above to amplify from either end. This was cloned into the NdeI/EcoRI sites of a pET17b. Following cloning of this construct, we removed the acidic loop and C-terminus of the Rhed again by four-primer PCR. The first reaction used F2 with 5'-CATACTAGTATGATGCTAAAGCTTATGATGCTAACCGGTAGGGGTGAGGTCACTGCTTTG-3'. The second reaction used 5'-CATACTAGTATGATGCTAAAGCTTATGATGCTAACCGGTAGGGGTGAGGTCACTGCTTTG-3' and R2. These products were reassembled as above using F2 and R2 primers.

Error-prone PCR of the C-terminal region of Rhed used the primers 5'-CAGAccGGTCCACTAGGGGCTGAGGCAG-3' (P1) and 5'-CTCTTTCTTACTAGTGGGTGCATCTTG-3' (P2), and performed essentially as described [33]. Briefly, the 100 µL reaction contained 10 mM Tris pH8.3, 50 mM KCl, 7 mM $MgCl_2$, 10 µM $MnCl_2$, 0.2 mM dGTP, 0.2 mM dATP, 1 mM dCTP, 1 mM dTTP, 0.2 µM each primer, 2 fmol template, and 1 µL Taq polymerase (NEB). The reaction was heated to 95°C for 30 s, followed by 20 cycles of amplification at 95 °C for 15 s and 68 °C for 1 min, with a final extension at 68 °C for 5 min.

The PCR product was gel purified, digested with SpeI/AgeI, and the vector triple digested with SpeI/HindIII/AgeI and phosphatase treated. For ligation, a 3:1 insert:vector ratio was estimated by gel electrophoresis of an aliquot of both components. 20 µL ligation reactions were incubated at 16°C overnight and heat inactivated. The reaction was purified over a G25 spin column (GE Healthcare) and transformed into Electromax DH10B cells (ThermoFisher) by electroporation. The library size was estimated by dilution of the transformation on LB-agar. The transformation was grown overnight in LB media with antibiotic and the plasmid library was isolated by maxi-prep. The library was subsequently transformed into BL21(DE3)-RIPL cells using the same electroporation protocol.

Following overnight culture, the library was diluted into 2L LB media and grown at 37°C to OD_{600nm} = 0.6. Protein expression was induced with IPTG and δALA following our standard protocol [32]. Cells were lysed by sonication in 0.1 M NaCl, 20 mM Tris pH8.0, 3 mM DTT, 1 mM EDTA, and 1X Halt-protease inhibitor cocktail (Pierce). The lysate was clarified by centrifugation and purified over a HiTrap SP-HP column (GE Healthcare), using a gradient to 1M NaCl over 40 min at 3 ml/min. Fractions containing the RNA-protein complex were identified by a combination of absorption spectroscopy, SDS-PAGE, RT-PCR (see below), and SEC analysis (not shown). The purified fraction was heated for 5 min at 55, 60, or 63 °C in the first, second, and third rounds of selection, respectively. Precipitated material was pelleted by centrifugation, and the supernatant mixed with 100 µL anti-FLAG M2 affinity resin (Sigma). The sample was incubated with resin at 4 °C with shaking for 1hr. The resin was then spun down and washed 3 times with 1 mL of buffer (150 mM NaCl, 20 mM Tris pH8.0, 1 mM DTT, 1 mM EDTA). RNA was recovered from the resin by extraction with TRIzol (Invitrogen) and purified using an RNeasy Kit (Qiagen).

The RNA was reverse transcribed using random hexamer primers and Superscript III (ThermoFisher) following the manufacture’s protocol. PCR from the cDNA was performed with Q5 polymerase (NEB) using primers P1 and P2 above. The PCR product was gel purified and used for subsequent rounds of selection.

Clones from the final library were mini-preped and sequenced, and the NC1 region subcloned into a modified pET17b vector containing an N-terminal His6 tag and C-terminal mCherry fusion. These were over-expressed as above and purified by IMAC and SEC columns. The 100 µL aliquots of the purified fraction were heated at 45, 48, 50, 52.5, 55, 57.5, and 60 °C for 5 min, and precipitated material pelleted by ultracentrifugation at 120,000 x g for 10min. The supernatant was analyzed on a custom SEC column packed with Superdex S200 resin (GE Healthcare) coupled to a home-built flow cell in a Jobin/Yvon/Horriba fluorimeter, with excitation set to 587 nm and emission at 610 nm. For the spectroscopic melting assay, the Rhed domain from select clones were subcloned into the same His6/mCherry-pET17b backbone, and purified as above. Using the Cary 300 Bio instrument, proteins were melted from room temperature to 70 °C over ~45min, with scans taken every 20-30 seconds.

Cloning and characterization of acidic group mutants

We used the standard four-primer PCR to mutate acidic positions in the Rhed C-terminus in groups. The outside primers were 5’-CAGCCATATGGATGGAGAGACAAGTGTGC-3’ (forward) and 5’-GCTAGCGGCCGCTCAGGGTGCATCTTGCACTGA-3’ (reverse). For E438S/D432S/E433S (AG1), the middle primers were 5’- GATCAACGGACGAGGATTTGCACACGGAGACTTTGGCCTTCACCTGCC-3’ and 5’-GGCCAAAGTCTCCGTGTGCAAATCCTCGTCCGTTGATCTCGAGGAATTTC-3’. For D436S/E438S/E439S (AG2), the middle primers were 5’- GCTTCGAAAGGAGGAGAGGGAAACGGATTCATCTTTGCACACCTC-3’ and 5’-GAATCCGTTTCCCTCTCCTCCTTTCGAAGCTACCTGGAGAAGCG-3’. For E445S/D449S/E451S (AG3), the middle primers were 5’- CAGTAACTTGCGAAAAGGAAAAACGCTTCGACAGGTAGCTTCGAAATTCCTCG-3’ and 5’- GAAGCTACCTGTCGAAGCGTTTTTCCTTTTCGCAAGTTACTGTGAAAAAATTCAGGAC-3’. For E463S/E470S (AG4), the middle primers were 5’-CCGCTTCATGGACCGATTGAATTGCCGCCGCGAAGCCCAAGTCCTGAATTTTTTCAC-3’ and 5’-GACTTGGGCTTCGCGGCGGCAATTCAATCGGTCCATGAAGCGGAAGCAGGCG-3’. For E477S/E479S (AG5) the middle primers were 5’- GATGGGCCTCGAGGACGACGCCTGCTTCCGCTTCATTTC-3’ and 5’-GAAGCAGGCGTCGTCCTCGAGGCCCATCTTGCCAGCC-3’. These were cloned into pET17b, expressed, and purified as described. Spectroscopy and RNA-binding assays were performed as above.

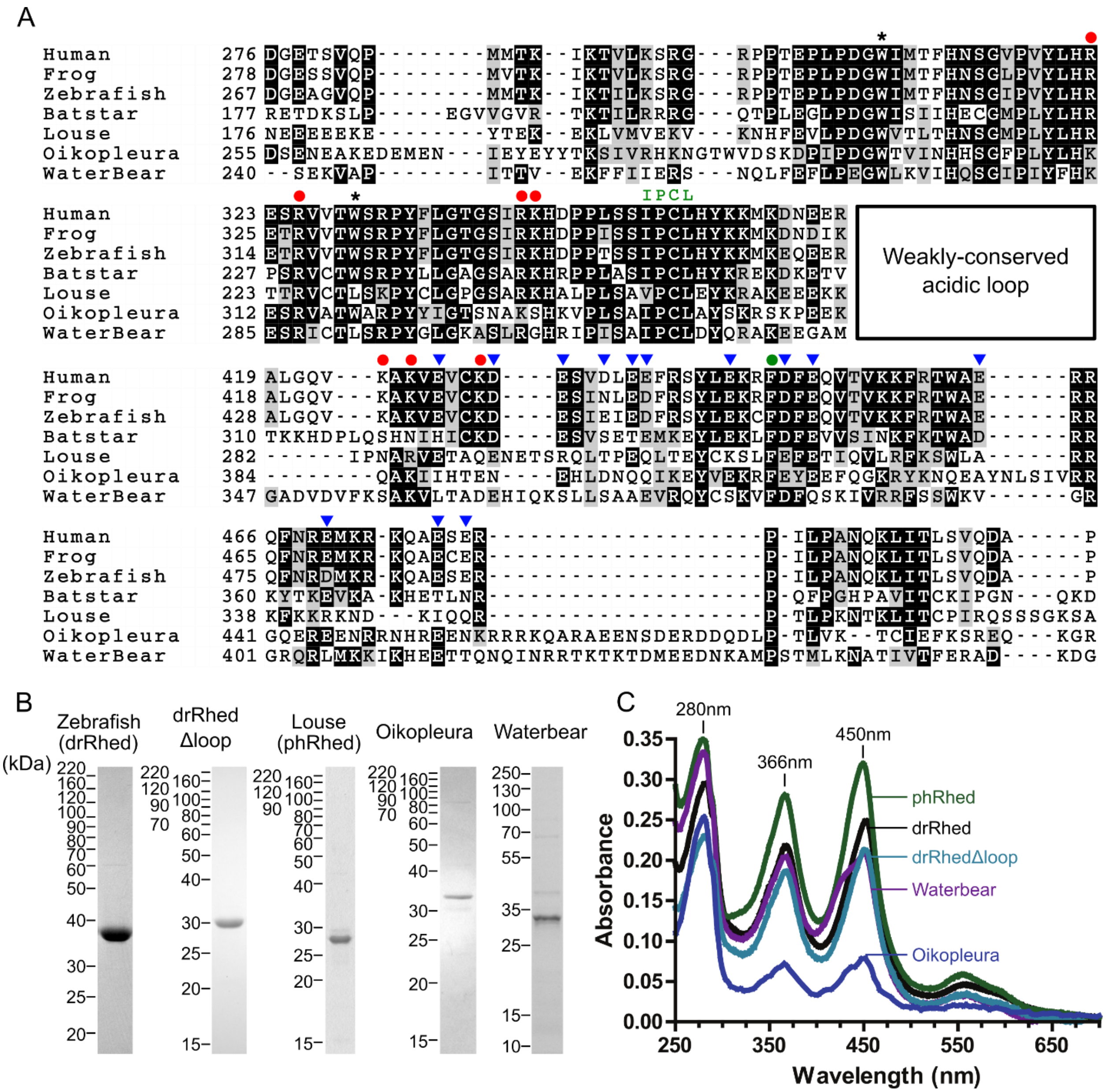


Figure 1. Alignment and purification of Rhed homologs. (A) Multiple sequence alignment of the human Rhed domain (DGCR8$^{276\text{-}498}$) with experimentally characterized homologs. The variable central loop region is omitted for clarity. An asterisk (*) marks the WW motif. The essential heme-binding elements IPCL and F448 are shown as green letters and a dot, respectively. Previously charted RNA-binding residues appear with a red dot. Blue triangles show the positions of acidic residues mutated in the C-terminal region of the domain. (B) Coomassie-stained SDS-PAGE gels of purified Rhed homologs. (C) Electronic absorption spectra of homologs.

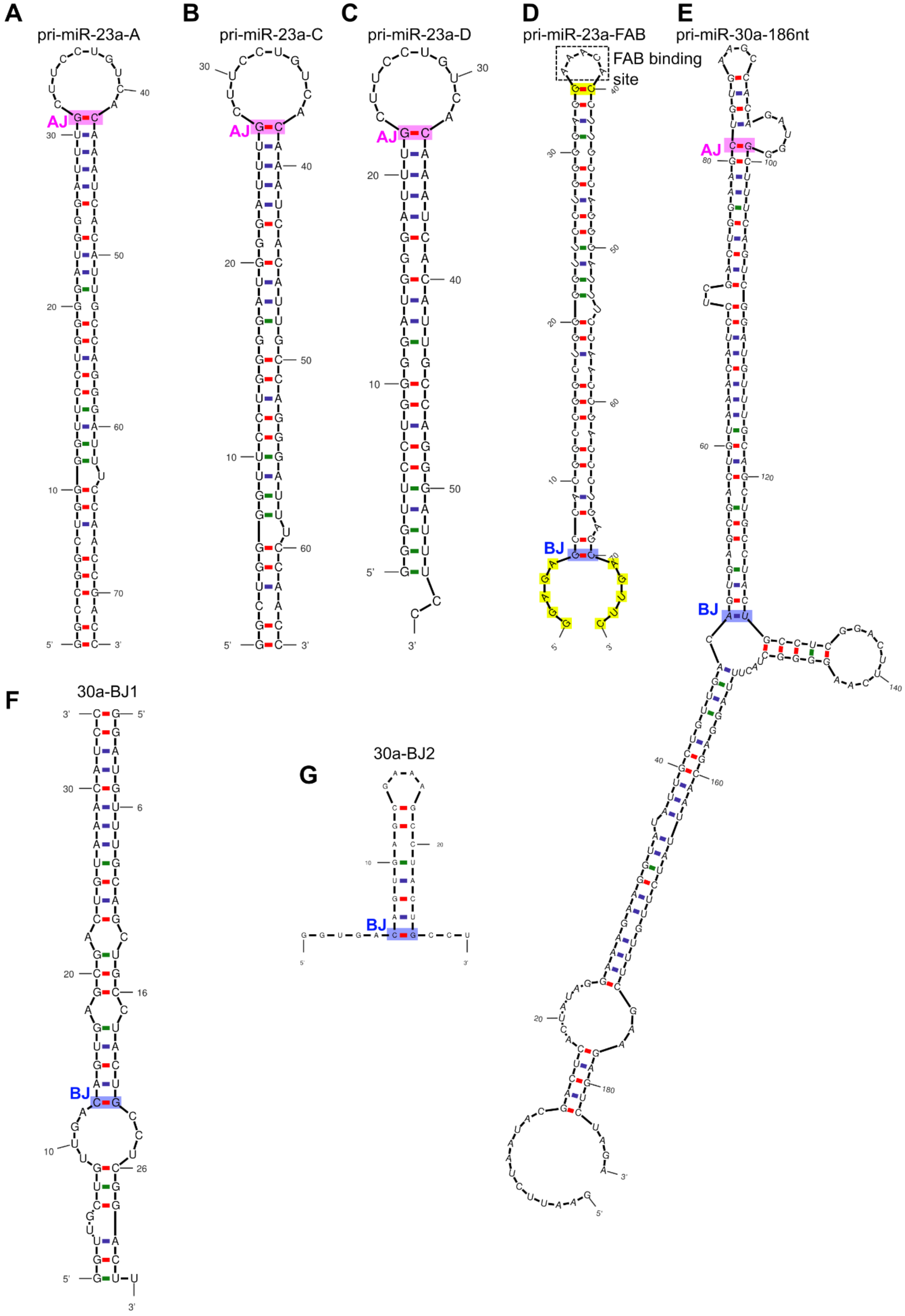


Figure 2. Secondary structure predictions for all RNA constructs used in this study, as generated by mfold. The apical (AJ) and basal junction (BJ) are highlighted in pink and blue, respectively. Non-natural bases engineered in the constructs are shown in yellow.

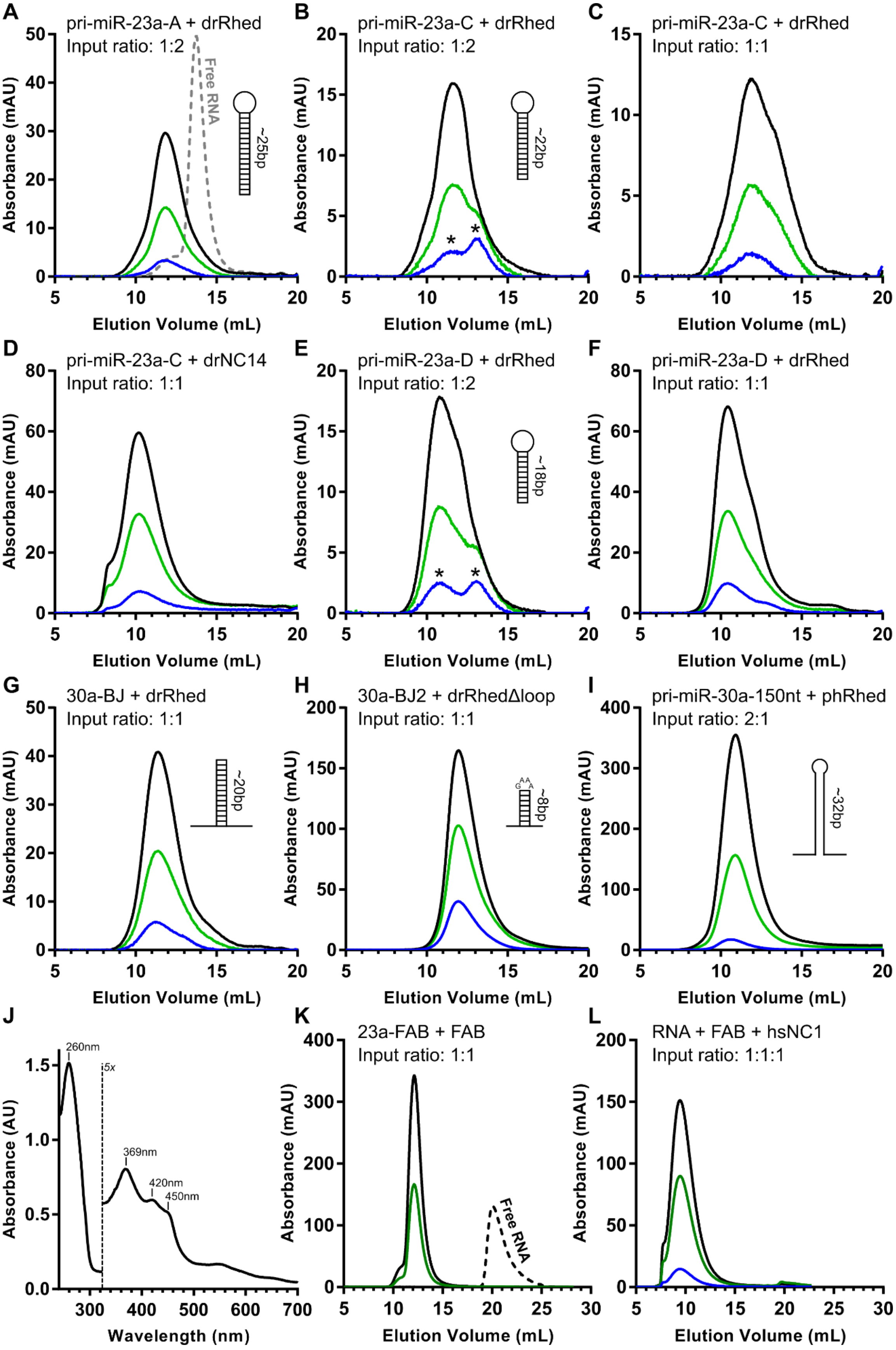

A
pri-miR-23a-A + drRhed
Input ratio: 1:2
Free RNA
~25bp
Absorbance (mAU)
50
40
30
20
10
0
5
10
15
20
Elution Volume (mL)
B
pri-miR-23a-C + drRhed
Input ratio: 1:2
~22bp
*
*
Absorbance (mAU)
20
15
10
5
0
Elution Volume (mL)
C
pri-miR-23a-C + drRhed
Input ratio: 1:1
Absorbance (mAU)
15
10
5
0
Elution Volume (mL)
D
pri-miR-23a-C + drNC14
Input ratio: 1:1
Absorbance (mAU)
80
60
40
20
0
Elution Volume (mL)
E
pri-miR-23a-D + drRhed
Input ratio: 1:2
~18bp
*
*
Absorbance (mAU)
Elution Volume (mL)
F
pri-miR-23a-D + drRhed
Input ratio: 1:1
Absorbance (mAU)
Elution Volume (mL)
G
30a-BJ + drRhed
Input ratio: 1:1
~20bp
Absorbance (mAU)
Elution Volume (mL)
H
30a-BJ2 + drRhedΔloop
Input ratio: 1:1
G A A A
~8bp
Absorbance (mAU)
200
150
100
50
0
Elution Volume (mL)
I
pri-miR-30a-150nt + phRhed
Input ratio: 2:1
~32bp
Absorbance (mAU)
400
300
200
100
0
Elution Volume (mL)
J
260nm
5x
369nm
420nm
450nm
Absorbance (AU)
1.5
1.0
0.5
0.0
300
400
500
600
700
Wavelength (nm)
K
23a-FAB + FAB
Input ratio: 1:1
Free RNA
Absorbance (mAU)
400
300
200
100
0
5
10
15
20
25
30
Elution Volume (mL)
L
RNA + FAB + hsNC1
Input ratio: 1:1:1
Absorbance (mAU)
200
150
100
50
0
Elution Volume (mL)

Figure 3. Analysis of RNA-Rhed complexes by SEC and absorption spectroscopy. (A) Complex of drRhed with ~25bp pri-miR-23a-A stem-loop prepared at a ratio of two Rhed dimer to one RNA. The grey dashed line shows the elution peak of the free RNA. Throughout the figure, the black, green, and blue curves represent the 260 nm, 280 nm, and 450 nm signals, respectively. (B) and (C) show the same experiment with the 18bp stem (pri-miR-23a-C) at 2:1 and 1:1 ratios, and (D) shows binding of drNC14 to the same RNA. (E) and (F) show association of drRhed with the shorter pri-miR-23a-D. (G) Binding of drRhed to the basal junction model 30a-BJ1. (H) Binding of drRhedΔloop to 30a-BJ2. (I) Complex between the louse homolog (phRhed) and pri-miR-30a. (J) Absorption spectrum of SEC-purified and concentrated pri-miR-30a-drRhed complex, with an unexpected peak near 420 nm. (K) Complex of the 23a-FAB RNA with the cognate FAB protein. Elution of the free RNA is shown with a dashed line. (L) Tripartite complex between the RNA-FAB complex shown in (K) and the NC1 protein.

K287 & K289	**m/z**	**z**	**NC3Z-1**	**NC3Z-2**	**23aD-1**	**23aD-2**
275 – MDGETSVQPMMT**K**IK – 289	646.6	3	3231	2063	531	1555
275 – MDGETSVQPMMT**K**IK – 289	641.3	3	12746	8207	7458	6969
275 – MDGETSVQPMMT**K**IK – 289	961.5	2	21419	11087	12159	10891
275 – MDGETSVQPMMT**K**I**K**TVLK – 293	874.4	3	4420	0	0	0
275 – MDGETSVQPMMT**K**I**K**TVLK – 293	874.4	3	1024	614	0	0
275 – MDGETSVQPMMT**K**I**K**TVLK – 293	869.1	3	738	996	817	390
275 – MDGETSVQPMMT**K**I**K**TVLK – 293	874.4	3	1838	2395	0	0
275 – MDGETSVQPMMT**K**I**K**TVLK – 293	869.1	3	5285	419	0	0
275 – MDGETSVQPMMT**K**I**K**TVLK – 293	869.1	3	377	712	0	0
275 – MDGETSVQPMMT**K**I**K**TVLK – 293	863.8	3	556	902	431	449
275 – MDGETSVQPMMTKIKTVL**K** – 293	1295.1	2	1196	0	0	0
276 – DGETSVQPMMT**K**I**K**TVLK – 293	820.1	3	613	0	0	0
288 – I**K**TVLK – 293	464.3	2	6890	1828	492	1452
K424						
409 – DPLGAEAAPGALGQV**K**AK – 426	640.3	3	9244	5348	1999	3778
409 – DPLGAEAAPGALGQV**K**AK – 426	960.0	2	11717	0	0	0
K474 & K488						
473 – R**K**QAESERPILPANQK – 488	697.7	3	2438	3605	0	0
473 – R**K**QAESERPILPANQK – 488	523.5	4	2319	3455	809	461
474 – **K**QAESERPILPANQK – 488	645.7	3	3855	1051	1790	1230
474 – **K**QAESERPILPANQK – 488	968.0	2	31122	27193	8898	1288
474 – **K**QAESERPILPANQKLITLSVQDAP – 498	743.9	4	892	0	0	0
474 – **K**QAESERPILPANQ**K**LITLSVQDAP – 498	1066.9	3	3461	3853	0	0
474 – **K**QAESERPILPANQKLITLSVQDAP – 498	800.4	4	2691	1626	0	0
474 – KQAESERPILPANQ**K**LITLSVQDAP – 498	991.5	3	7077	9966	1053	1651
475 – QAESERPILPANQ**K**LITLSVQDAP – 498	948.8	3	1535	4901	3421	2728

Table 1. MS analysis of tryptic peptides from the protein footprinting experiment, covering lysine residues K287, K289, K424, K474, and K488. The peptide sequence, mass to charge ratio, and charge value are shown. Oxidized methionine residues are colored red, while biotinylated lysines are in green. Columns NC3Z-1 and NC3Z-2 are two replicate analyses of the free protein labeling, where as 23aD-1/23aD-2 are replicate runs of the RNA-protein complex labeling.

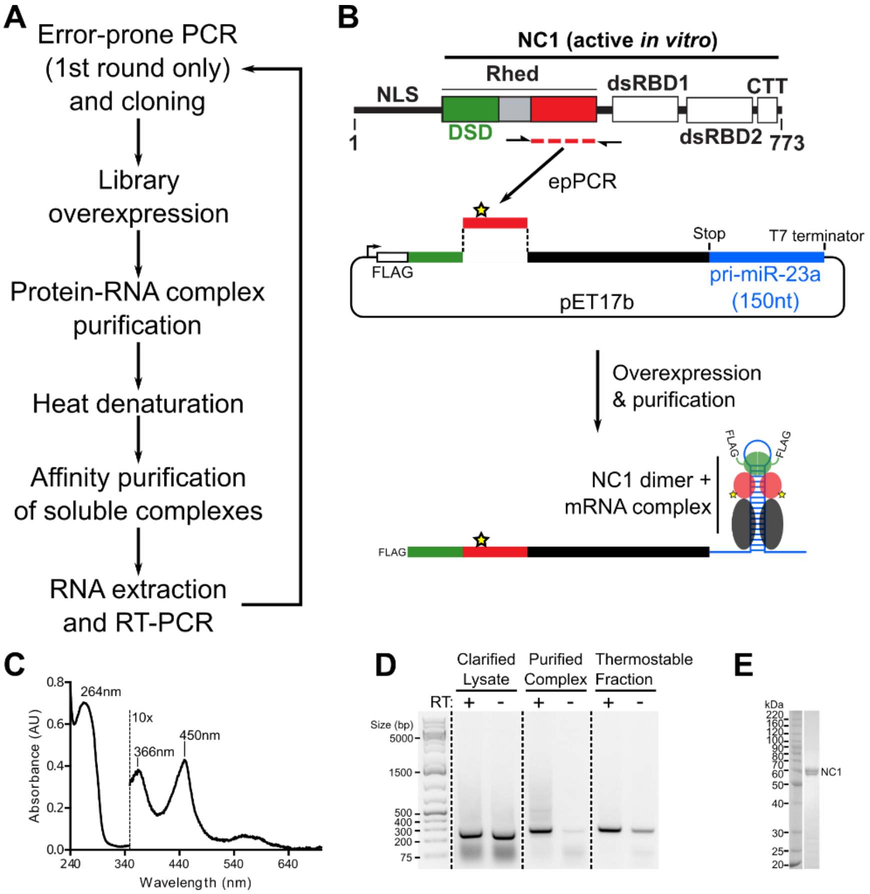


Figure 4. *In vitro* evolution scheme for thermostabilization of RNA-binding proteins. (A) Outline of the selection strategy. (B) Graphic representation of the plasmid construction and mRNA-protein complex used for selection. (C) Absorption spectrum of the purified RNA-protein complexes. (D) RT-PCR analysis of selection steps, using primers specific for the C-terminal region the Rhed (250bp product). (E) Coomassie-stained SDS-PAGE gel of protein present in the RNA-protein complex.

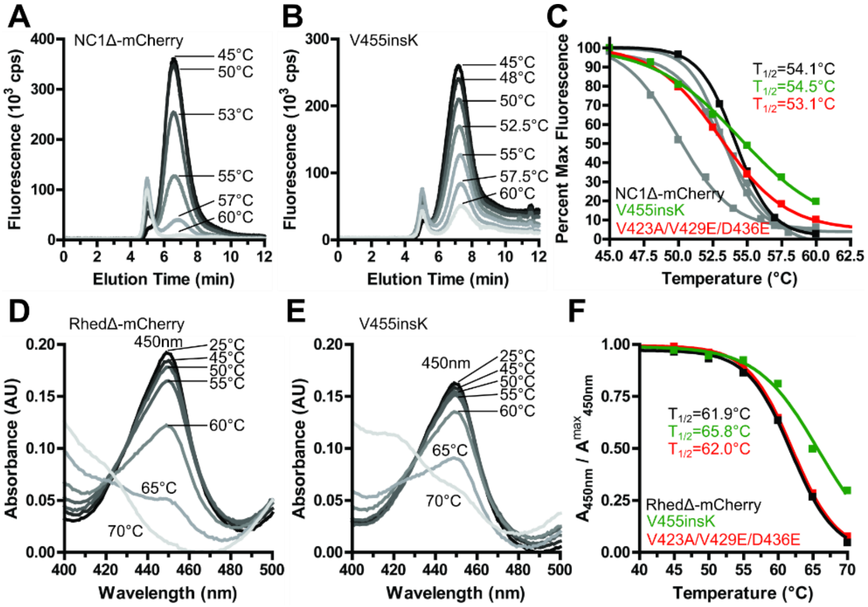


Figure 5. Melting assays for thermostable variants isolated after three rounds of selection. (A) F-SEC melting assay for NC1Δloop-mCherry control. (B) Melting assay the V455insK mutation in the context of NC1Δloop-mCherry. (C) Melting curves for NC1Δloop (black), V455insK (green), the triple mutant V423A/V429E/D436E (red), and additional mutants (grey). (D) Melting monitored by loss of the Soret peak at 450nm, using RhedΔloop-mCherry. (E) Identical to (D) but with V455insK mutation. (F) Melting curves extracted from the 450nm signal in (D) and (E). Curves are colored analogously to (C).

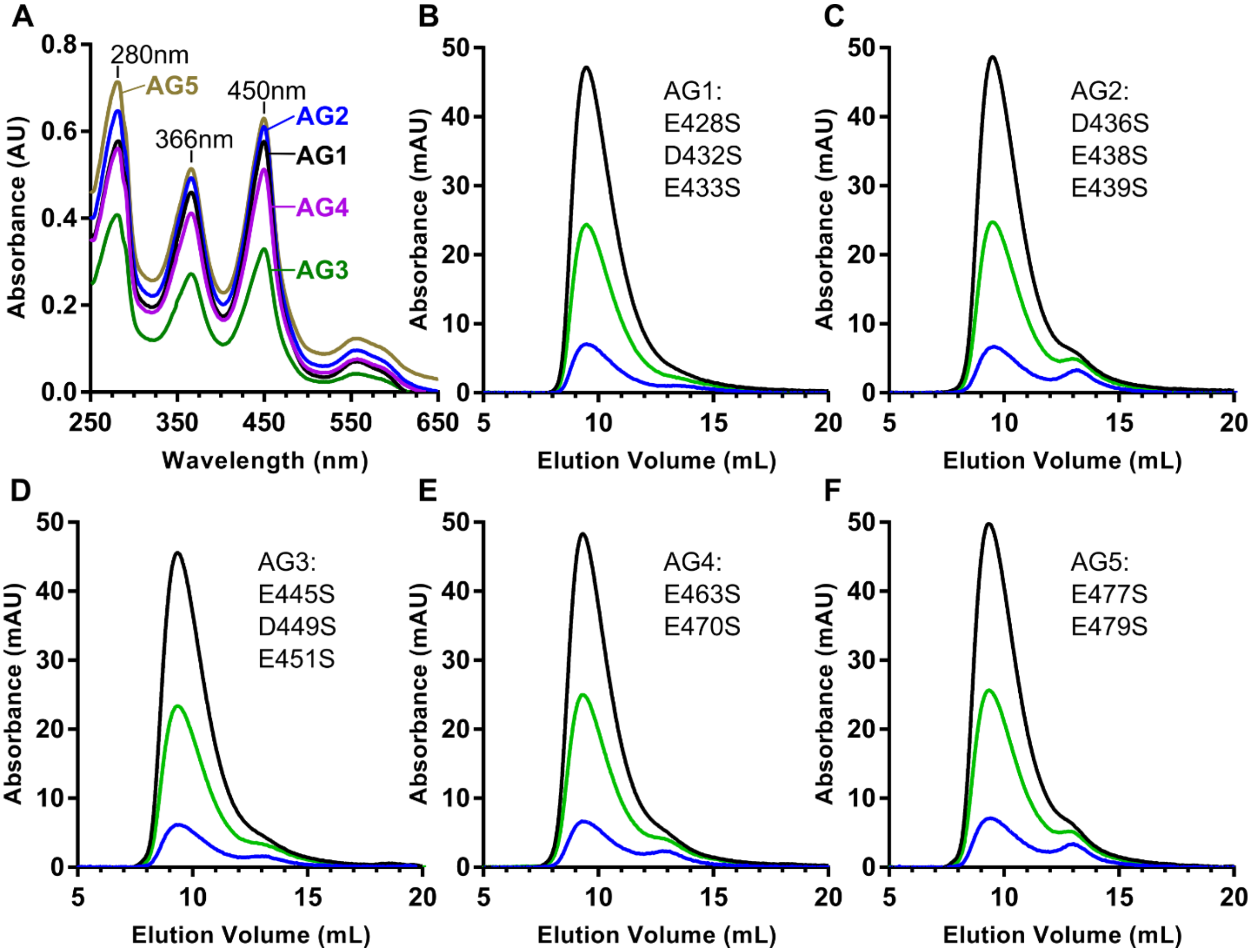


Figure 6. Heme and RNA binding analysis of group mutations of acidic residues in the Rhed. (A) Absorption spectra of all group mutants. The AG5 spectrum was bumped above baseline for clarity. (B)-(F) show SEC analysis of group mutants AG1-AG5 in a 2 dimer to 1 RNA ratio with pri-miR-30a. SEC curves are colored as in Figure 2.

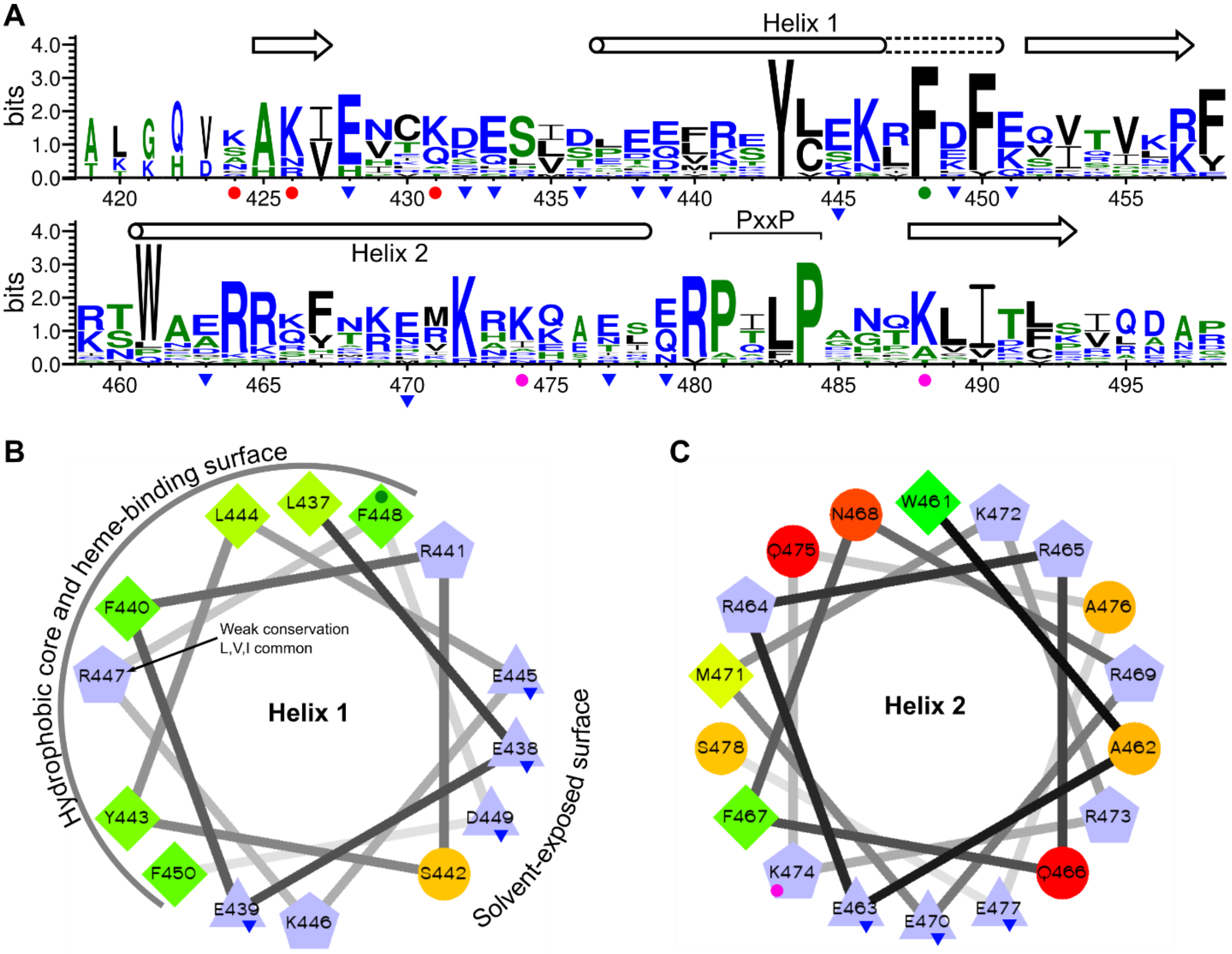


Figure 7. Phylogenetic conservation of the C-terminal Rhed sequence and secondary structure prediction. (A) Logo representation of conservation of residues in the human Rhed C-terminus (positions 419-498). Arrows and tubes show β-strands and α-helix regions predicted by PSIPRED. The extension of Helix 1 to F450 (dashed tube) is based on our analysis. Red dots shown known RNA binding sites. The green dot marks F448, which is important for binding heme. Blue triangles mark sites of mutation of acidic residues. Pink dots indicate new RNA-binding residues reported here. The PxxP motif highlights a potential protein interaction site. Helical wheel projections of Helix 1 and Helix 2 are shown in (B) and (C), respectively.

References

[1] I. J. MacRae, K. Zhou, F. Li, A. Repic, A. N. Brooks, W. Z. Cande, P. D. Adams, and J. A. Doudna, "Structural Basis for Double-Stranded RNA Processing by Dicer," *Science (80-.).*, vol. 311, no. 5758, pp. 195–198, Jan. 2006.

[2] Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Rådmark, S. Kim, and V. N. Kim, "The nuclear RNase III Drosha initiates microRNA processing," *Nature*, vol. 425, no. 6956, pp. 415–419, Sep. 2003.

[3] J. Han, Y. Lee, K.-H. Yeom, Y.-K. Kim, H. Jin, and V. N. Kim, "The Drosha-DGCR8 complex in primary microRNA processing.," *Genes Dev.*, vol. 18, no. 24, pp. 3016–27, Dec. 2004.

[4] A. M. Denli, B. B. J. Tops, R. H. A. Plasterk, R. F. Ketting, and G. J. Hannon, "Processing of primary microRNAs by the Microprocessor complex," *Nature*, vol. 432, no. 7014, pp. 231–235, Nov. 2004.

[5] R. I. Gregory, K. Yan, G. Amuthan, T. Chendrimada, B. Doratotaj, N. Cooch, and R. Shiekhattar, "The Microprocessor complex mediates the genesis of microRNAs," *Nature*, vol. 432, no. 7014, pp. 235–240, Nov. 2004.

[6] M. Landthaler, A. Yalcin, and T. Tuschl, "The Human DiGeorge Syndrome Critical Region Gene 8 and Its D. melanogaster Homolog Are Required for miRNA Biogenesis," *Curr. Biol.*, vol. 14, no. 23, pp. 2162–2167, Dec. 2004.

[7] M. Faller, M. Matsunaga, S. Yin, J. A. Loo, and F. Guo, "Heme is involved in microRNA processing," *Nat. Struct. Mol. Biol.*, vol. 14, no. 1, pp. 23–29, Jan. 2007.

[8] S. Y. Sohn, W. J. Bae, J. J. Kim, K.-H. Yeom, V. N. Kim, and Y. Cho, "Crystal structure of human DGCR8 core," *Nat. Struct. Mol. Biol.*, vol. 14, no. 9, pp. 847–853, Sep. 2007.

[9] C. Wostenberg, W. G. Noid, and S. A. Showalter, "MD Simulations of the dsRBP DGCR8 Reveal Correlated Motions that May Aid pri-miRNA Binding," *Biophys. J.*, vol. 99, no. 1, pp. 248–256, Jul. 2010.

[10] S. H. Weitz, M. Gong, I. Barr, S. Weiss, and F. Guo, "Processing of microRNA primary transcripts requires heme in mammalian cells.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 5, pp. 1861–6, Feb. 2014.

[11] R. Senturia, M. Faller, S. Yin, J. A. Loo, D. Cascio, M. R. Sawaya, D. Hwang, R. T. Clubb, and F. Guo, "Structure of the dimerization domain of DiGeorge Critical Region 8," *Protein Sci.*, vol. 19, no. 7, pp. 1354–1365, Jul. 2010.

[12] M. Faller, D. Toso, M. Matsunaga, I. Atanasov, R. Senturia, Y. Chen, Z. H. Zhou, and F. Guo, "DGCR8 recognizes primary transcripts of microRNAs through highly cooperative binding and formation of higher-order structures.," *RNA*, vol. 16, no. 8, pp. 1570–83, Aug. 2010.

[13] J. Quick-Cleveland, J. Jacob, S. Weitz, G. Shoffner, R. Senturia, and F. Guo, "The DGCR8 RNA-Binding Heme Domain Recognizes Primary MicroRNAs by Clamping the Hairpin," *Cell Rep.*, vol. 7, no. 6, 2014.

[14] T. A. Nguyen, M. H. Jo, Y.-G. Choi, J. Park, S. C. Kwon, S. Hohng, V. N. Kim, and J.-S. Woo, "Functional Anatomy of the Human Microprocessor," *Cell*, vol. 161, no. 6, pp. 1374–1387, Jun. 2015.

[15] S. C. Kwon, T. A. Nguyen, Y.-G. Choi, M. H. Jo, S. Hohng, V. N. Kim, and J.-S. Woo, “Structure of Human DROSHA.,” *Cell*, vol. 164, no. 1–2, pp. 81–90, Jan. 2016.

[16] A. C. Partin, T. D. Ngo, E. Herrell, B.-C. Jeong, G. Hon, and Y. Nam, “Heme enables proper positioning of Drosha and DGCR8 on primary microRNAs,” *Nat. Commun.*, vol. 8, no. 1, p. 1737, Dec. 2017.

[17] I. Barr, A. T. Smith, Y. Chen, R. Senturia, J. N. Burstyn, and F. Guo, “Ferric, not ferrous, heme activates RNA-binding protein DGCR8 for primary microRNA processing.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 6, pp. 1919–24, Feb. 2012.

[18] I. Barr, A. T. Smith, R. Senturia, Y. Chen, B. D. Scheidemantle, J. N. Burstyn, and F. Guo, “DiGeorge critical region 8 (DGCR8) is a double-cysteine-ligated heme protein.,” *J. Biol. Chem.*, vol. 286, no. 19, pp. 16716–25, May 2011.

[19] I. Barr, S. H. Weitz, T. Atkin, P. Hsu, M. Karayiorgou, J. A. Gogos, S. Weiss, and F. Guo, “Cobalt(III) Protoporphyrin Activates the DGCR8 Protein and Can Compensate microRNA Processing Deficiency,” *Chem. Biol.*, vol. 22, no. 6, pp. 793–802, Jun. 2015.

[20] T. Li, H. L. Bonkovsky, and J. Guo, “Structural analysis of heme proteins: implications for design and prediction,” *BMC Struct. Biol.*, vol. 11, no. 1, p. 13, Mar. 2011.

[21] R. Senturia, A. Laganowsky, I. Barr, B. D. Scheidemantle, and F. Guo, “Dimerization and Heme Binding Are Conserved in Amphibian and Starfish Homologues of the microRNA Processing Protein DGCR8,” *PLoS One*, vol. 7, no. 7, p. e39688, Jul. 2012.

[22] Y. Koldobskaya, E. M. Duguid, D. M. Shechner, N. B. Suslov, J. Ye, S. S. Sidhu, D. P. Bartel, S. Koide, A. A. Kossiakoff, and J. A. Piccirilli, “A portable RNA sequence whose recognition by a synthetic antibody facilitates structural determination,” *Nat. Struct. Mol. Biol.*, vol. 18, no. 1, pp. 100–106, Jan. 2011.

[23] D. T. Jones, “Protein secondary structure prediction based on position-specific scoring matrices,” *J. Mol. Biol.*, vol. 292, no. 2, pp. 195–202, Sep. 1999.

[24] L. J. McGuffin, K. Bryson, and D. T. Jones, “The PSIPRED protein structure prediction server.,” *Bioinformatics*, vol. 16, no. 4, pp. 404–5, Apr. 2000.

[25] D. W. A. Buchan, F. Minneci, T. C. O. Nugent, K. Bryson, and D. T. Jones, “Scalable web services for the PSIPRED Protein Analysis Workbench,” *Nucleic Acids Res.*, vol. 41, no. W1, pp. W349–W357, Jul. 2013.

[26] K. Saksela and P. Permi, “SH3 domain ligand binding: What’s the consensus and where’s the specificity?,” *FEBS Lett.*, vol. 586, no. 17, pp. 2609–2614, Aug. 2012.

[27] A. F. Moon, G. A. Mueller, X. Zhong, and L. C. Pedersen, “A synergistic approach to protein crystallization: Combination of a fixed-arm carrier with surface entropy reduction,” *Protein Sci.*, vol. 19, no. 5, p. NA-NA, May 2010.

[28] E. Chun, A. A. Thompson, W. Liu, C. B. Roth, M. T. Griffith, V. Katritch, J. Kunken, F. Xu, V. Cherezov, M. A. Hanson, and R. C. Stevens, “Fusion Partner Toolchest for the Stabilization and Crystallization of G Protein-Coupled Receptors,” *Structure*, vol. 20, no. 6, pp. 967–976, Jun. 2012.

[29] J. Zhang and A. R. Ferré-D’Amaré, “New molecular engineering approaches for crystallographic

studies of large RNAs," *Curr. Opin. Struct. Biol.*, vol. 26, pp. 9–15, Jun. 2014.

[30] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–402, Sep. 1997.

[31] K. Katoh and D. M. Standley, "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability," *Mol. Biol. Evol.*, vol. 30, no. 4, pp. 772–780, Apr. 2013.

[32] I. Barr and F. Guo, "Primary MicroRNA Processing Assay Reconstituted Using Recombinant Drosha and DGCR8," in *Methods in molecular biology (Clifton, N.J.)*, vol. 1095, 2014, pp. 73–86.

[33] D. S. Wilson and A. D. Keefe, "Random Mutagenesis by PCR," in *Current Protocols in Molecular Biology*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2001, p. 8.3.1-8.3.9.

Chapter 3: Structures of pri-miRNA terminal loops determined by scaffold-directed crystallography

Grant Shoffner, Zhixiang Peng, and Feng Guo

Abstract

Cleavage of pri-miRNA transcripts by the Microprocessor requires the correct determination of the length and orientation of their characteristic hairpin secondary structures, and yet these molecular features do not uniquely specify pri-miRNA substrates. Acting as the apical junction-binding factor, the DGCR8 Rhed domain anchors the Microprocessor at the top of the stem. We wondered if unique structural features of the apical junction or terminal loop might function as specificity determinants, allowing the Rhed to discern true pri-miRNAs from other RNA hairpins. We developed a scaffold-directed crystallization method that enabled structure determination of eight pri-miRNA apical junctions and loops at near atomic resolution. Each structure revealed a surprising degree of tertiary folding in the loop, driven by non-canonical base pairing and base stacking. Despite diversity in the loop length and primary sequence, our structures highlight shared structural elements between all the loops. In particular, nucleotides at the 5' end of the loop are consistently well ordered by base stacking interactions. We suggest that the previously unsuspected pairing interactions and the structured 5' loop residues act as additional determinants for recognition of pri-miRNA.

Introduction

Since the discovery of the Microprocessor complex (MC) over a decade ago, the driving question in the field has been how its core components Drosha and DGCR8 collaborate to specifically recognize their pri-miRNA substrates and identify the appropriate cleavage sites. This proved especially fascinating given that early assessments found no sequence homology between the pri-miRNAs, which arise from both coding and non-coding transcripts and within intronic and exonic regions [1]. One fact was clear: the primary transcript contained the mature miRNA buried within an imperfectly base-paired hairpin structure [2]. This posed a conundrum: how could the Microprocessor possibly distinguish the various pri-miRNA hairpins from all the other RNA hairpins in the nucleus?

What followed was a series of attempts to work out the recognition principles based on mutagenesis of a handful of example pri-miRNAs. Cullen and colleagues first proposed that the MC sought out the single-stranded terminal loop at the apex of hairpin, then measured two helical turns down the stem to find the cut sites; but noted that single-stranded RNA segments flanking the base of the stem were also important [3], [4]. Conversely, Narry Kim's group proposed that the terminal loop was dispensable for processing and that Drosha measured one helical turn up from the base of the stem to reach the cleavage site [5].

Subsequent reports added weight to the importance of the apical junction and terminal loop for directing the MC to the correct hairpins. In particular, several studies showed that mutagenesis of the loop or natural genetic variation in this region can have severe impacts on processing efficiency and cleavage site selection [6], [7]. Furthermore, a growing list of regulatory proteins seemed to modulate processing by association with the terminal loop [8], [9]. The basal versus apical models were partially reconciled with the proposal that certain groups of pri-miRNA are more sensitive to basal stem length whereas others depend on the position of the apical junction [10].

The notion that MC could sense both the upper and lower stem lengths coalesced into the "molecular clamp" model, in which both distances contribute to cut site selection and together enable MC to measure the overall length of the stem [11]. The finding from our group that the Rhed domain from DGCR8 can specifically recognize single-stranded to double-stranded RNA junctions proved that MC could anchor itself at the junction in order to measure the stem [12]. Further analysis of MC assembly indicated that a heterotrimeric complex of Drosha and a DGCR8 dimer bind the stem from both ends, with the Rhed locating the apical junction and Drosha finding the base of the stem [13].

In spite of this detailed description of the processing complex, the field was no closer to elucidating how MC could distinguish the correct substrate purely by measuring the helical length. Surely MC would not cleave any stem with approximately three helical turns? In fact, human MC is blind to pri-miRNA stems from nematodes, despite their correct length [14]. Work from David Bartel's group showed that this confusing result was largely due to short sequence motifs found in some human pri-miRNA but missing from the C. elegans RNAs [14]. Adding the motifs to completely artificial pri-miRNAs was sufficient to activate processing [15]. Only a single sequence element mapped to 5' end of the terminal loop, a UGU motif that enhances processing though possible interaction with the Rhed [13]. In addition to these sequence specificity factors, post-transcriptional modification of certain pri-miRNA can highly upregulate their processing [16]. Finally, more comprehensive studies of pri-miRNA secondary structure revealed that MC tolerates mismatches and bulges only within specific regions of the stem [17].

Collectively these features support a model in which recognition primarily depends on stem length, but an array of additional factors can fine-tune the rate or efficiency of processing for an individual substrate. However, an important part of the recognition mechanism remains largely unexplored: the possibility that MC can identify junctions based on three-dimensional structural elements shared between all or most pri-miRNAs. A unique conformational signature of the loop or apical junction could have escaped previous analyses of sequence or secondary structure because these methods do not completely capture three-dimensional information. Interestingly, the observation that the pri-miR-125a terminal loop can serve as an aptamer domain for binding folic acid supports the idea that the junction and loop can have functional 3D structure [18]. To further explore this possibility we decided to investigate the atomic structure of pri-miRNA apical junctions.

To date the only two structural studies have been performed for the pre-miR-20b and pre-miR-21 stem-loops using NMR spectroscopy [19]–[21]. Although extensive NOEs generated a high-confidence, single conformation for the pre-miR-20b loop, weak signals for pre-miR-21 provided few restraints for the model in the upper stem. We are interested to increase the resolution and throughput of pri-miRNA structure determination using X-ray crystallography. To this end, we developed a scaffold-direct crystallization technique that enables the rapid determination of pri-miRNA apical junction and loop structures. We report eight such structures and biochemical characterization of their interactions with Rhed.

Results

Given that the terminal loop contains the UGU motif, and that the Rhed functions as the junction-binding factor at this end of the stem, we wondered if specific structural features of the terminal loop or junction might define the pri-miRNA substrate. Using MFOLD to predict the secondary structure of all human pri-miRNA hairpins in miRBase, we measured the length of the loop for all annotated pri-miRNA (Figure 1A). Interestingly, we noticed that the majority of hairpins (1,368 out of 1,881, 73%) contained a terminal loop less than 10 nt long. This indicates that for most pri-miRNA recognition events, the Rhed must associate with a relatively short terminal loop in order to access the apical junction.

We employed a scaffold-directed crystallization approach to determine the atomic structures of pri-miRNA terminal loops and junctions. The concept of scaffold-directed crystallography is to fuse the target (unknown) sequence onto a known crystal structure of a possibly unrelated molecule (the scaffold). Ideally, the scaffold permits fusion of the target moiety in such an orientation that it occupies a void in the existing crystal lattice, enabling recrystallization of the fusion construct in the original conditions used to solve the structure of the scaffold. In this case, the scaffold provides a ready molecular replacement solution to determine the structure of the fusion molecule.

To identify a suitable scaffold for pri-miRNA stem-loops, we mined the PDB for RNA crystals containing large lattice cavities. For each RNA structure entry, we tiled the crystal lattice with grid points and then calculated the largest sphere centered at each point that touched no RNA atoms in the structure. We defined R_{max} to be the radius of the largest such sphere for each RNA, which approximates the size of the largest cavity present in the lattice. We plotted this value against the maximum diffraction resolution reported for that structure (Figure 1B). Our previous attempts at designing RNA scaffolds taught us that engineering new moieties into the crystal lattice is simplified when only one molecule occupies the asymmetric unit (green crosses in Figure 1B). Focusing our attention on structures meeting this criterion and with better than 3.5 Å diffraction, we manually reviewed the crystal lattices to find stem-loop portions of the scaffold that terminate inside the lattice cavity. These stem-loops represent potential sites onto which we may fuse a pri-miRNA hairpin. Amongst hundreds of structures surveyed, we identified only one RNA meeting these requirements, the YdaO-type c-di-AMP riboswitch from *Thermoanaerobacter pseudethanolicus* bound to c-di-AMP (PDB ID: 4QK8) [22].

The YdaO riboswitch crystal contains large solvent channels (approx. 30 Å radius, see the grey sphere in Figure 1C). Furthermore, the RNA adopts a pseudo-two-fold symmetric 'cloverleaf' fold (Figure 1D) which positions the short P2 stem inside the solvent channel and away from neighboring molecules in the lattice (green stem-loop highlighted in Figures 1C and 1D). To determine if replacing the GAAA tetraloop with a pri-miRNA stem-loop would disrupt the folding of the riboswitch, we generated fusions between YdaO and pri-miR-9-1. These fusions contained the 14-nt pri-miR-9-1 apical junction plus 0-3 additional base-pairs from the stem. Annealing the RNA in the presence of the c-di-AMP ligand resulted in single bands for all constructs on a native gel (Figure 1E). This result suggests that the engineered pri-mi-RNA sequences do not interfere with the scaffold folding and encouraged us to pursue crystallization of pri-miRNA fusions.

Starting from a set of 48 pri-miRNAs with loops ranging from 4-10 nt, we generated fusions with the YdaO scaffold containing the loop plus a variable number of base pairs from the stem, and screened for crystallization. We succeeded in solving the structures of eight terminal loops (Table 1), all of which crystallized with one base pair from the pri-miRNA stem. In addition, we solved the structures of two of these loops with no base pairs from stem. The refined structures showed very few differences between the W.T. and fusions for residues in the scaffold (C1′ rmsd values ranged from 0.27 to 1.18 Å, Figure 2).

The fusion with the longest loop was 8 nt, from pri-miR-378a (termed 378a+1bp, Figures 3A and 4I). Our structure shows the expected final base pair of the pri-miRNA stem, an A-U pair, stacking atop the final C-G pair of the YdaO P2 stem. As RNA loops can be flexible, they are often not well resolved in the electron density. To our surprise, the $2F_o$-F_c map of 378a+1bp revealed a highly structured conformation with clear density for all residues. The 378a+1bp structure clearly shows that outermost residues of the loop, C1 and A8, form a non-canonical pair, which stacks to the A-U pair of the stem. This extra C-A pair creates a platform onto which bases from the remainder of the loop stack. On the 5' end, C2 and U3 stack above C1, and from the 3' side A4, G5, A6, and A7 stack in four layers above A8. Across the two stacks of bases, a hydrogen bond between O2 oxygen of U3 and N7 position of A7 further stabilizes the loop (Figure 5A).

We also solved the structure of the pri-miR-378a terminal loop with no base pair from the stem (378a + 0bp, Figure 5B), which diffracted to higher resolution (2.79 Å vs 2.95 Å for 378a + 1bp). The models for both loops are in close agreement (1.4 Å rmsd over all non-hydrogen atoms in the loop, Figure 5C). The 378a + 0bp structure confirms the non-canonical C1-A8 pair as well as the C2-A7 H-bond (Figure 5D). The clearer electron density in 378a + 0bp shows the position of U3 rotated slightly into the loop, revealing additional H-bonds between $U3^{O2}$ and $A6^{N7}$ (3.0 Å) as well as $U3^{N3}$ to $A6^{OP2}$ (2.8 Å, Figure 5D). These interactions show that H-bonding coordinates every loop nucleotide except A4 and G5. Interestingly, the fact that 378a + 0bp and 378a + 1bp are nearly identical suggests that the loop conformation is independent from the terminal base pair of the pri-miRNA stem.

We observed a similar set of interactions in the structure of the 7-nt loop of pri-miR-340 (Figure 3B and 4G). The structure confirms the presence of the terminal A-U pair, which is capped by an unexpected

U1-U7 pair. The following G2 and U3 residues from the 5’ end of the loop stack on top of the U-U pair. This leaves just three residues (C4, G5, and U6) in a more flexible conformation at the top of the loop.

The second 7-nt loop structure is of pri-miR-300 (300 + 0bp). In this case, the terminal C-G pair of the scaffold is identical to the last base pair of the pri-miR-300 stem. As in the case of 378a and 340, we observe a non-canonical pairing between U1 and U7 (Figure 3C and Figure 4D). Likewise, a chain of base-stacking interactions between U1, U2, U3 and A4 orders the 5’ end of the loop. Positions C5 and U6 are outside the density and potentially more flexible. When we added in an additional C-G pair to the stem (300 + 1bp), we obtained a broadly similar result (Figure 5E and 5F), albeit at lower resolution (3.96 Å). Although the positions of A4, C5, and U6 are not well defined in the 300 + 1bp model, the U1-U7 pairing and U1-U2 stacking interactions are replicated.

In the structure of pri-miR-202 (6-nt loop), we did not observe non-canonical base pairs. However, similar to other structures, the A1 base at the 5’ end of the loop stacks to the final G-C pair of the pri-miRNA stem (Figure 3D and 4B).

Next, we investigated the structures of shorter pri-miRNA terminal loops (4-5 nt, Figure 6). Strikingly, the structure of pri-miR-208a (5-nt loop) revealed an unpredicted A1-U5 Hoogsteen pair positioned above the final G-C pair from the stem (Figures 6A and 4C). The central 3 nt of the loop (U2, G3, and C4) base-stack together and onto the A1 base in the Hoogsteen pair. Similar to the structure of 202+1bp above, for pri-miR-320b-2 (5-nt loop), the A1 residue of the loop sits atop the terminal A-U pair of the stem (Figure 6B and 4F). In addition, the non-canonical U-U pair from 340+1bp is recapitulated between U1 and U5 in the structure of pri-miR-449c (Figure 6C and 4J). Positions U1, G2, and A3 stack together above the terminal base pair, leaving just U4 in a flexible conformation. Finally, in the tetraloop structure of pri-miR-19b-2 (19b-2 + 1bp) the 5' loop nucleotide U1 stacks above the terminal base pair and a partial stacking interaction of A2 on top of U1 (Figure 6D and 4A). U3 and G4 are mostly outside the electron density, although there may be a contact between G4^{N7} and the 2'-OH of A2 (~2.6 Å). These structures confirm that the non-canonical pairing and base stacking of the 5' loop residues witnessed in longer loop structures also dominate the folding of the shorter loops.

Our pri-miRNA stem-loop structures point toward a common set of structural features defining the terminal loop. To further illustrate these features, we generated a structural alignment of all eight pri-miRNA loops (Figure 6E). We only included the structures with +1bp so that the alignment was achieved simply by superimposing the coordinates of the last G-C pair from the scaffold. First, we always observe the predicted base pair at the apical end of the pri-miRNA stem (5'-1 paired with 3'-1). Because the loops are of different sizes, here we use 5'-1 to represent the first residue from the 5'-end of the pri-miRNA sequence, and 3'-1 to represent the first residue from the 3'-end. Although our structures were determined with the rest of pri-miRNA stem substituted by the YdaO P2 helix, this observation nevertheless indicates that the terminal loops investigated are not longer than their MFOLD-predicted lengths. Second, in all structures the first nucleotide on the 5' end of the loop base-stacks with the terminal base pair (5'-2 stacking with 5'-1/3'-1). In six of the eight loops (378a, 340, 300, 208a, 449c), this base stacking is also accompanied by a non-canonical base pair (5'-2 with 3'-2 pairing), effectively making the terminal loop shorter than predicted by two nucleotides. Third, six of the eight structures also reveal a second level of base-stacking interactions (5'-3 stacked on 5'-2). Other than these common features, other residues of the pri-miRNA loops appear to adopt quite different conformations.

Considering the high degree of structural overlap between the terminal loops, we wondered whether the Rhed could recognize all the apical junctions despite differences in overall loop length. We addressed this question by measuring the affinities of Rhed for pri-miRNA fragments containing the terminal loop plus around 20 bp from the stem (Figure 7). These pri-miRNA fragments do not contain the basal junctions, thereby each having only a single Rhed binding site. We used gel shift binding assays with radiolabeled RNA to determine the Rhed dissociation constant (K_d) for each RNA (Figures 8 and 9).

To ensure we detected specific binding of the pri-miRNA fragment, we included competitor tRNA (0.1 mg/mL) and heparin (5 µg/mL) in the binding reaction (see Methods section).

With the exception of pri-miR-320b-2 (Figure 9E, K_d ~ 9.2 µM), the Rhed bound all pri-miRNA fragments with comparable affinities (K_d = 1.9 – 7.0 µM). We note that pri-miR-340, which contains the UGU motif at the 5' side of the loop, binds the Rhed with similar affinity to other constructs lacking this sequence (Figures 9F). To compare the binding affinity between the loops, we plotted the ΔG of binding versus the MFOLD predicted loop length (Figure 9I). Interestingly, we observed a rough trend toward tighter binding of loops in the 6-8 nt range. However, our structural data indicate that non-canonical base pairing shortens most loops longer than four nucleotides. When we subtracted these non-canonical pairs from the loop length, we found a clearer relationship, where 6 nt loops appear to bind with highest affinity (Figure 9J). The free energy penalty paid by shorter loops indicates the Rhed may need to unwind these structures to access the apical junction.

These results strongly suggest that the Rhed recognizes common structural characteristics of the apical junction and loop. If this is the case, we would expect these features to be stable structural elements. To investigate the dynamics of the junction and terminal loop, we first reviewed the atomic displacement parameters (ADP, also known as the temperature or B factor) refined during structure determination. Plotting the ADP values on the crystal structures revealed a clear trend toward higher stability at the 5’ end of the loop and more flexibility in the 3’ end (Figure 10). For each RNA, the base stacking interactions at the 5’ end are consistently more stable than the 3’ nucleotides. Residues at the top of the loop have very large ADPs, indicating these positions are highly dynamic. To directly compare ADPs between structures, we calculated the average ADP per residue and then plotted these on the same scale (Figure 10I). The peak in ADPs is consistently located near the middle to 3’ end of the loop across all structures.

For a more detailed view into the loop dynamics, we performed molecular dynamics simulations of the junction and loop nucleotides in explicit solvent (Figure 11). For simplicity, the simulation included only the pri-miRNA residues plus two base pairs from the scaffold (grey residues in Figure 11), and we restrained the position of the scaffold nucleotides to prevent unwinding of the strand (see Methods section). The simulations were run at 300 K for 1 µs, and we analyzed the resulting trajectories using two methods. First, we calculated the root-mean-square fluctuation (RMSF) for each residue (Figure 10J). These statistics clearly confirm our previous observation that the loop 5' residues are relatively stable whereas the 3' residues sample wider range of conformations.

Second, we clustered trajectories together to determine the full range of motion of the loop (rainbow colored alignments in Figure 10). During clustering, we adjusted the RMSD cutoff value to obtain a reasonable number of clusters (5-12 clusters per simulation).We found that several RNAs (202+1bp, 300+1bp, and 340+1bp) have very flexible loops which sample a large range of conformations (Figures 11B, 11D, and 11F). Since the clustering is based only on RMSD of individual trajectories, this analysis does not indicate how prevalent a given conformation is during the simulation. To determine which conformations occurred most frequently, we repeated the clustering using the same 1.0 Å cutoff for all RNAs. Figure 10 shows any conformation that represented more than 12% of the total simulation time (green structures). In all cases, the simulation was dominated by 1-3 conformations, and these contain the experimentally observed base stacking features described above, indicating these interactions are stable over the course of the simulation.

Discussion

Our results provide the proof-of-concept that scaffold-directed crystallography can be a powerful tool for RNA structural biology. This method is largely analogous to the popular fixed-arm MBP fusion technique, in which a target protein is linked to MBP in a fixed orientation via a continuous alpha-helical linker [23]. However, our engineering approach specifically positions the target RNA within a lattice void of the scaffold crystal. This results in several additional advantages: (1) the original crystallization conditions can be reused for crystallization of the fusion molecule because the target moiety does not disrupt existing lattice contacts; and (2) the large distances between the target and neighboring molecules minimize lattice distortions in the conformation of the target. Furthermore, since rescreening of a broad array of conditions is unnecessary, a minimal amount of purified fusion RNA is required for crystallization.

Applying this technique to the problem of pri-miRNA recognition provides an atomic-level survey of eight pri-miRNA terminal loop structures. These pri-miRNA loops vary in primary sequence and length, and they display unique three-dimensional folds. This observation in and of itself has important implications. First, terminal loops can form unique and highly-ordered structures which may relate to their individual functions. This agrees with a report that the pri-miR-125a loop can function as an aptamer domain for binding a small metabolite [18]. Second, unpredicted secondary structure in the terminal loop indicates that previous estimates of pri-miRNA helical length are not completely accurate, and calls into question the exact mechanism of how the MC measures the stem length. Do the non-canonical pairs unwind during the recognition event, revealing the true apical junction, or does the MC count the extra base-pair when finding the cut site?

The diversity of terminal loop structures may indicate that the loop conformation is not a specificity determinant for pri-miRNA recognition. However, despite the clear differences in overall conformation, the most striking features of the structures we report are their shared structural elements. Most prominently, we observe base stacking between residues at the 5’ end of the loop and final base pair of the pri-miRNA stem. This stacking interaction minimally involves the first loop nucleotide, but in most loops additional downstream residues join in to form multiple layers of stacked, co-planar bases. This trend is most evident in the structure of the pri-miR-378a loop, in which all of the loop nucleotides are stacked. The fact that base stacking of the 5’ nucleotides arises in all the pri-miRNA investigated, regardless of sequence or loop length, supports the idea that this shared feature may function to uniquely identify pri-miRNA hairpins.

An alternative explanation for this repeated structural element is that our crystallization scaffold somehow induces this conformation in the loop. To investigate this possibility, we compared our structures to the previously reported NMR data for pre-miR-20b and pre-miR-21, two RNA stem-loops that we did not crystallize. Alignment of the pre-miR-20b terminal loop to our crystallographic models confirmed that the shared features described above are reiterated in pre-miR-20b (Figure 12A). Specifically, the stem terminates in a non-canonical G-U pair and the first loop nucleotide (G) stacks on top of the pair. Comparison of the top 20 NMR solutions confirms these are stable features of the molecule (Figure 12B). Although fewer NOE restraints were detected for the pre-miR-21 loop [20], the data generally support the positioning of the first loop nucleotide above the terminal base pair (red U in Figure 12C). In addition, an independent NMR study of pre-miR-21 detected base-stacking in the 5’ loop residues [21]. As NMR represents an orthogonal means of structure determination to our crystallographic analysis, the commonalities in the terminal loop structures strongly argue that these features are not crystallographic artifacts.

As our data represent a small sample of the >1,800 annotated human pri-miRNA hairpins, appreciation of the full terminal loop folding space will require additional crystal structures. To this end, we are endeavoring to enhance the crystallization properties of the YdaO scaffold, following the procedure we develop in Chapter 3. If successful, the improved scaffold may increase the diffraction resolution as well as the range of target molecules amenable to fusion at the P2 stem. We anticipate this will unlock the structures of many more pri-miRNA loops in the near future.

Materials and Methods

Pri-miRNA secondary structure analysis

To gauge the approximate size of the terminal loops we needed to crystallize, we downloaded data from miRBase describing all annotated human “hairpin” sequences and their genomic coordinates. The miRBase hairpin roughly corresponds to the pre-miRNA along with a variable number of additional base pairs from the basal stem. For each hairpin, we used the genomic sequence to extend the RNA an equal number of nucleotides at the 5’ and 3’ ends until the total length equaled 150nt. This 150nt window contained the full pri-miRNA hairpin, plus some single-stranded RNA on either side of the basal junction. We then generated predicted secondary structures for all pri-miRNA hairpins using MFOLD [24], and generally retained the top scoring structures (i.e. with the lowest predicted free energy of folding). We manually reviewed all the predictions to ensure they reflected the expected hairpin structure with mature miRNA sequences derived from either or both strands of the stem; in cases were mfold predicted alternative conformations, we selected the structure with the lowest free energy that contained a stem length of approximately three helical turns. For each structure, we counted the longest stretch of unpaired residues as a proxy for the length of the terminal loop.

PDB mining and identification of YdaO crystallization scaffold

We first filtered the PDB to obtain X-ray structures containing only RNA molecules (no protein or DNA). To identify voids in the crystal lattices, we wrote a custom PyMOL script that implemented a grid search algorithm in the following steps. (1) Generate a “super-cell”, which is a 3 × 3 × 3 block of unit cells (i.e. 21 copies of the unit cell). The cell at the center of this block sees all possible lattice voids, either internally or between unit cells. (2) Using three unit vectors along each of the unit cell axis (i.e. a, b, and c vectors of length 1 Å), iteratively generate grid points of the form 5*i*a + 5*j*b + 5*k*c for integer values of i,j,k less than the respective unit cell edge length divided by 5. This gives grid points with 5 Å spacing. (3) Set $R_{max} = 0$. For each grid point, do the following: Set $R_{local} = 100$; for each C1’ atom in the super-cell, calculate the distance d to the grid point; if d < R_{local}, set R_{local} = d; lastly, if R_{local} > R_{max}, set R_{max} = R_{local}.

We then manually reviewed the structures with large R_{max} and a single molecule in the asymmetric unit to find suitable scaffold molecules. More specifically, we traced the chain looking for any stem-loop that projected into the cavity in the lattice. Amongst several hundred candidates reviewed, only the P2 stem-loop from the YdaO riboswitch (PDB ID: 4QK8) met these conditions [22].

Preparation of YdaO W.T. and pri-miR-9-1 fusion RNA and native gel electrophoresis

We initially designed the W.T. YdaO construct to contain a T7 promoter sequence at the 5’ end and HDV ribozyme on the 3’ side, along with flanking EcoRI and BamHI restriction sites. This fragment was synthesized as a gene block (IDT), double digested and cloned into the pUC19 plasmid. The clone was verified by Sanger sequencing. To replace the P2 loop nucleotides with the pri-miRNA stem-loop, we used a two-step PCR protocol. All reactions were performed with Q5 high-fidelity DNA polymerase (New England Biolab) following the manufacture’s recommended reaction setup and cycling conditions. All reactions contained the same reverse primer which annealed to the 3’ end of HDV and contained the BamHI site (5’-CGTGGATCCGGTCCCATTC-3’). For the first step of PCR, the forward primer contained the pri-miRNA sequence plus around 20 nt upstream and downstream on the scaffold. The forward primers for pri-miR-9-1 fusions were 5’-CTATAGGTTGCCGAATCCGTGGTGTGGAGTCTGGTACGGAGGAACCGCTTTTTG-3’ (pri-miR-9-1 + 0bp); 5’-CTATAGGTTGCCGAATCCAGTGGTGTGGAGTCTTGGTACGGAGGAACCGCTTTTTG-3’ (pri-miR-9-1 + 1bp); 5’-CTATAGGTTGCCGAATCCGAGTGGTGTGGAGTCTTCGGTACGGAGGAACCGCTTTTTG -3’ (pri-miR-9-1 + 2bp); 5’- CTATAGGTTGCCGAATCCAGAGTGGTGTGGAGTCTTCUGGTACGGAGGAACCGCTTTTTG-3’ (pri-miR-9-1 + 3bp). This PCR product was gel purified and 1 µL was used as template for a second round of PCR. The second round contained the same reverse primer, and all reactions used a forward primer which annealed to the common scaffold residues and added the T7-promoter and EcoRI site (5'-GCAGAATTCTAATACGACTCACTATAGGTTGCCGAATCC-3'). The second round product was gel purified, digested, and ligated into pUC19. Clones containing the desired insert were sequence verified.

For WT and pri-miR-9-1 fusion constructs we prepared maxi-prep DNA and linearized the plasmid by overnight digestion with BamHI (NEB). Transcription reactions contained ~400 µg linearized template, 40 mM Tris pH7.5, 25 mM $MgCl_2$, 4 mM DTT, 2 mM spermidine, 40 µg inorganic pyrophosphatase (Sigma), 0.7 mg T7 RNA polymerase, and 3 mM each rNTP in a total volume of 5 mL. After a 4.5 hr incubation at 37°C, the final $MgCl_2$ concentration was adjusted to 40 mM with a 2M $MgCl_2$ stock solution, and the reactions were incubated for an additional 45 min. Despite the elevated Mg^{2+} concentration, we observed only partial cleavage by the HDV ribozyme. Reactions were ethanol precipitated and purified over denaturing 10% polyacrylamide slab gels. The desired product was visualized by UV shadowing and excised from the gel. Gel pieces were crushed and extracted overnight in 30 mL TEN buffer (150 mM NaCl, 20 mM Tris pH7.5, 1 mM EDTA) at 4°C. We then spun down the gel pieces and concentrated the RNA in an Amicon 15 centrifugal filter device with 10-kDa molecular weight cutoff (MWCO). RNA was buffer exchanged three times into 10 mM HEPES pH7.5 and concentrated to ~50 µL final volume.

For analysis on a native gel, 5 µM RNA stock solutions were prepared by dilution of the purified RNA into 5 mM Tris pH 7.0. Next, 2.5 µL RNA was mixed with an equal volume of 2X annealing buffer containing 35 mM Tris pH7.0, 100 mM KCl, 10 mM $MgCl_2$, and 20 µM c-di-AMP. Mixtures were heated at 90°C for 1 min followed by snap cooling on ice and then a 15-min incubation at 37°C. The annealed RNA was mixed with a 2X loading dye containing 40 mM Tris pH 7.0, 50 mM KCl, 5 mM $MgCl_2$, 20% (v/v) glycerol, and xylene cyanol. 5 µL of each reaction was run on a 10% polyacrylamide gel with Tris-borate (TB) running buffer. The gel was stained in Sybr Green II and scanned on a Typhoon 9410 Variable Mode Imager (GE Healthcare).

Preparation of pri-miRNA-YdaO fusions for crystallization

Given the poor HDV self-cleavage efficiency we observed for the pri-miR-9-1 fusions, we elected to change strategy. Instead of employing a ribozyme to create homogeneous 3’ ends, we used PCR to generate transcription templates in which the final two bases of the YdaO scaffold were 2’-O-methylated on the anti-sense DNA strand. The modifications have been shown to reduce un-templated nucleotide addition by T7 RNA polymerase [25]. We utilized a three-stage PCR approach to create the transcription templates. All reactions below contained the same reverse primer, 5’-mCmUCCTTCCTTTATTGCCTCC-3’, where ‘m’ indicates the 2’-O-methylated sugar. For the first stage of PCR, we set up a 50 µL reaction with Q5 polymerase to amplify the 3’ fragment of YdaO with the forward primer 5’-GGTACGGAGGAACCGCTTTTTG-3’ and performed 30 cycles of amplification. The product was gel purified and 1 µL was used as template for the next round. In the second stage, we used a unique forward primer for each construct containing the pri-miRNA loop and stem sequence which annealed to the 3’ YdaO fragment from the first stage. The primer sequences were

5’-CTATAGGTTGCCGAATCCATATGTGGTACGGAGGAACCGCTTTTTG-3’ (19b-2 + 1bp);

5’-CTATAGGTTGCCGAATCCGATCTGGCGGTACGGAGGAACCGCTTTTTG –3’ (202 + 1bp);

5’-CTATAGGTTGCCGAATCCGATGCTCGGTACGGAGGAACCGCTTTTTG –3’ (208a + 1bp);

5’-CTATAGGTTGCCGAATCCCTTTACTTGGGTACGGAGGAACCGCTTTTTG –3’ (300 + 1bp);

5’-CTATAGGTTGCCGAATCCAAAGTTGGTACGGAGGAACCGCTTTTTG–3’ (320b-2 + 1bp);

5’-CTATAGGTTGCCGAATCCATGTCGTTTGGTACGGAGGAACCGCTTTTTG–3’ (340 + 1bp);

5’-CTATAGGTTGCCGAATCCACCTAGAAATGGTACGGAGGAACCGCTTTTTG–3’ (378a + 1bp); and

5’-CTATAGGTTGCCGAATCCATGATTTGGTACGGAGGAACCGCTTTTTG–3’ (449c + 1bp).

This reaction was also 50 µL and used Q5 polymerase for 30 cycles. The product from the second stage was analyzed by agarose gel electrophoresis to confirm amplification, and 40 µL of the reaction was used as template for the next stage without further purification. The third stage consisted of a 2-mL PCR reaction using the Phusion high-fidelity DNA polymerase (Thermo-Fisher) and set up according to the manufacturer's directions. The forward primer was 5'-GCAGAATTCTAATACGACTCACTATAGGTTGCCGAATCC-3'. The 2-mL volume was divided into 100 µL aliquots in a 96-well PCR plate for 35 cycles of PCR.

The third stage PCR product was purified over a HiTrap Q HP column (GE Healthcare). Buffer A contained 10 mM NaCl and 10 mM HEPES pH 7.5; Buffer B was identical but with 2 M NaCl. The column was equilibrated with 20% Buffer B and the desired DNA product was eluted with a linear gradient to 50% B over 10 min at 2 ml/min. We analyzed the peak fractions on an agarose gel to confirm they contained a single band of the correct size. The peak fractions were then pooled and concentrated in an Amicon device (10 kDa MWCO), and then washed with water to remove excess salt. The concentration of the DNA template (~200 µL final volume) was determined by UV absorbance.

Transcription reactions were set up as described above for pri-miR-9-1 fusions, but in a 10-mL volume and containing 2.8 fmol DNA template. Reactions were run for 4 hr at 37°C followed by phenol-chloroform extraction. The transcription was concentrated in an Amicon device (10 kDa MWCO) and washed with 0.1 M trimethylamine-acetic acid solution (TEAA, pH 7.0). The RNA (~2 mL) was injected onto a Waters XTerra MS C18 reverse phase HPLC column (3.5 µm particle size, 4.6x150 mm) thermostated at 54°C. TEAA and 100% acetonitrile were used as mobile phases. The column was washed with 6% acetonitrile and the RNA eluted with a gradient to 17% acetonitrile over 80 min at 0.4 ml/min. Peak fractions were analyzed on denaturing 10% polyacrylamide gels. Pure fractions were pooled and buffer exchanged into 10 mM HEPES pH 7.0 using an Amicon device. The RNA was concentrated to <50 µL final volume and the concentration determined by UV absorbance.

RNA crystallization and structure determination

All RNA-c-diAMP complexes were prepared as described [22]. Briefly, a solution containing 0.5 mM RNA, 1 mM c-di-AMP, 100 mM KCl, 10 mM $MgCl_2$, and 20 mM HEPES pH7.0 was heated to 90°C for 1 min, snap cooled on ice, and equilibrated for 15 min at 37°C immediately prior to crystallization. Screening was performed in 24-well plates containing 0.5 mL well solution; the hanging drops consisted of 1 µL RNA plus 1 µL well solution. Plates were incubated at room temperature, and crystals generally grew to full size (100 µm to over 200 µm) within one week. For 19b-2 + 1bp, the well solution contained 1.7 M $(NH_4)_2SO_4$, 0.2 M Li_2SO_4, and 0.1 M HEPES pH 7.1. For 202 + 1bp, 208a + 1bp, and 320b-2 + 1bp the well contained 1.9 M $(NH_4)_2SO_4$, 0.2 M Li_2SO_4, and 0.1 M HEPES pH 7.4. The well solution for 378a + 0bp contained 1.7 M $(NH_4)_2SO_4$, 0.2 M Li_2SO_4, and 0.1 M HEPES pH 7.4. For the remaining constructs crystallization was performed in 96 well plates with hanging drops consisting of 0.4 µL RNA plus 0.4 µL well solution. For 300 + 1bp, the well solution contained 1.88 M $(NH_4)_2SO_4$, 0.248 M Li_2SO_4, and 0.1 M HEPES pH 7.4, and for 300 + 0bp it held 1.90 M $(NH_4)_2SO_4$, 0.158 M Li_2SO_4, and 0.1 M HEPES pH 7.4 Construct 340 + 1bp crystallized from a well solution containing 1.89 M $(NH_4)_2SO_4$, 0.214 M Li_2SO_4, and 0.1 M HEPES pH 7.4. Construct 378a + 1bp crystallized from 1.63 M $(NH_4)_2SO_4$, 0.272 M Li_2SO_4, and 0.1 M HEPES pH 7.4. For construct 449c + 1bp, the well contained 1.89 M $(NH_4)_2SO_4$, 0.128 M Li_2SO_4, and 0.1 M HEPES pH 7.4

All crystals were briefly soaked in a cyroprotectant solution containing 20% (w/v) PEG 3350, 20% (v/v) glycerol, 0.2 M $(NH_4)_2SO_4$, 0.2 M Li_2SO_4, and 0.1 M HEPES pH7.3, and then flash frozen in liquid nitrogen. Data were collected at 100 K at the Advanced Photon Source Beamline 24-ID-C or Advanced Light Source Beamline 8.3.1. Data were indexed, integrated, and scaled using XDS [26].

To solve the structures, we started with the YdaO c-di-AMP riboswitch structure from *Thermoanaerobacter pseudethanolicus* (PDB ID: 4QK8) and removed residues 14-17 from the model, corresponding to the GAAA tetraloop on the P2 stem of the riboswitch. We then performed a rigid body fit of the model to data using Phenix [27]. This produced an excellent initial model with $R_{work} < 30\%$. We then inspected the electron density map in region of the P2 stem. For all RNAs, additional density for the missing base-pair and loop could clearly be seen in the $2F_o$-F_c and difference maps. We then modeled in the missing residues in Coot [28]. In cases where the density was unclear, we stopped modeling with an incomplete loop and performed an additional round of coordinate, ADP, and TLS parameter refinement with Phenix. This typically revealed additional density for the missing residues. Once the loop was completely modeled, we performed subsequent rounds of refinement and manual adjustment as above until reasonable R factors and model geometry were obtained.

Simulated annealing composite omit maps were calculated in Phenix. In the case of 19b-2 + 1bp, 202 + 1bp, 320b-2 + 1bp, 340 + 1bp, and 378a + 1bp the standard annealing temperature (5000 °C) and other default parameters produced reasonable maps (Figure 4). However, for 300 + 0bp, 300 + 1bp and 378a + 0bp the default settings generated noisy maps with regions of broken density. To improve the quality of the maps, we reduced the annealing temperature to 1000 °C and excluded the bulk solvent mask from the omitted regions. This type of composite omit map is known as Polder map and prevents the solvent mask from obscuring weaker density [29]. Figure 4 shows the Polder maps for these three structures.

RNA-Rhed gel shift binding assays

Human heme-bound Rhed protein was over-expressed and purified from bacteria by ion exchange and size exclusion chromatography as previously described [12]. Radiolabeled pri-miRNA stem-loops (Figure 6) were prepared by *in vitro* transcription. DNA templates consisted of anti-sense oligonucleotides covering the desired sequence plus the T7 promoter, annealed with a second oligo complementary to the T7 promoter sequence [30]. and 50 fmol were added to each reaction. The 20 µL transcription reaction contained 40 mM Tris pH 7.5, 25 mM $MgCl_2$, 4 mM DTT, 2 mM spermidine, 2 µg T7 RNA polymerase, 0.5 mM rATP, and 3 mM each of rUTP, rCTP, and rGTP; along with 3 nmol α-^{32}P-ATP (10 µCi). Transcriptions were run at 37°C for 2 hr and the RNA purified over a denaturing 15%

polyacrylamide gel. Gel slices containing the labeled RNA were extracted overnight at 4°C in TEN buffer, isopropanol precipitated, and resuspended in 40 µL water.

RNAs diluted in 100 mM NaCl, 20 mM Tris pH 8.0 and heated at 90°C for 1 min followed by snap cooling on ice. The annealed RNA was added to binding reactions containing 10% (v/v) glycerol, 0.1 mg/ml yeast tRNA, 0.1 mg/ml BSA, 5 µg/ml heparin, 0.01% (v/v) octylphenoxypolyethoxyethanol (IGEPAL CA-630), 0.25 unit RNase-OUT ribonuclease inhibitor, xylene cyanol, 20 mM Tris pH 8.0, and 0-20 µM Rhed protein. The final salt concentration of the solution was 150 mM NaCl. Binding reactions were incubated at room temperature for 30 min prior to loading on a 10% polyacrylamide gel. The gel and gel running buffer contained 80 mM NaCl, 89.2 mM Tris base, and 89.0 mM boric acid (pH 8.2 final). Gels were run at 110 V for 45 min at 4°C, and then dried and exposed to a storage phosphor screen. Screens were subsequently scanned on a Typhoon instrument (GE Healthcare). The free and bound RNA bands were quantified with Quantity One software (BioRad) and fit with the Hill equation in GraphPad Prism. As a control, we included pri-miR-23a-57nt (Figure 7I); we previously investigated the binding of the Rhed to this RNA using a filter binding assay [12]. For this construct we measured Kd = 4.2 ± 0.043 (mean ± SE, 3 replicate experiments, data not shown). This confirms that our previously reported apical junction models bind with comparable affinity to the pri-miRNA investigated here with gel shift assays.

Molecular Dynamics Simulations

Coordinates corresponding to the pri-miRNA residues plus two G-C pairs from the P2 stem of the scaffold were extracted from each crystal structure. Hydrogens were added to the model in GROMACS, and the RNA was dissolved in a truncated dodecahedral box with TIP3P water molecules. The box was sufficiently large to space the RNA at least 1 nm from any periodic copy of itself. Next, K^+ and Cl^- ions were added to the system to neutralize the net charge and give a final KCl concentration of 0.1 M. The CHARMM27 force field, Verlet cutoff scheme, and particle-mesh Ewald electrostatics were employed for all calculations. The system was energy minimized until the maximum force acting on any atom was less than 900 kJ/mol/nm. The final potential energy of the system was in the range of -1.3 x 10^5 kJ/mol.

Next the system was equilibrated in two steps, first in the NVT ensemble and then in the NPT ensemble. Both simulations ran at 300 K over 2 ns using a 2 fs time step. During NVT temperature is controlled by velocity rescaling. For NPT, the Parrinello-Rahman barostat is used to maintain pressure at 1 bar. For production MD runs, position restraints were applied to the G-C pairs from the scaffold, and all pri-miRNA nucleotides were unrestrained. All simulations were run in NPT with 2 fs time stepping for a total of 1 µs. Trajectories were analyzed using the rmsf and clustering functions in GROMACS. For clustering analysis, we performed simple linkage clustering with varying cutoffs, and retained the centroid structure from each cluster. Clusters with a reasonable number of members (<12) are visualized in Figure 10.

Acknowledgements

We thank Duilio Cascio and Michael Sawaya from the UCLA-DOE Crystallization and X-ray Diffraction facilities for guidance in data collection and processing. Malcolm Capel and Kay Perry at the Advanced Photon Source sector 24 (NE-CAT) also assisted with data collection; NE-CAT is supported by funding from the NIH (P41 GM103403 and S10 RR029205), and DOE (DE-AC02-06CH11357). We also thank George Meigs and Jane Tanamachi at ALS beamline 8.3.1 for assistance; BL 8.3.1 is supported by the UC Multicampus Research Programs and Initiatives grant (MR-15-328599), and NIH (R01 GM124149 and P30 GM124169).

RNA	19b-2 + 1bp	202 + 1bp	208a + 1bp	300 + 0bp	300 + 1bp	320b-2 + 1bp	340 + 1bp	378a + 0bp	378a + 1bp	449c + 1bp
Data collection										
Space group	P3$_1$21	P3$_1$21	P3$_1$21	P3$_1$21	P3$_1$21	P3$_1$21	P3$_1$21	P3$_1$21	P3$_1$21	P3$_1$21
Cell dimensions										
a,b,c (Å)	115.3, 115.3, 115.3	114.9, 114.9, 115.3	114.7, 114.7, 114.8	113.1, 113.1, 114.1	112.9, 112.9, 114.2	114.6, 114.6, 115.1	114.8, 114.8, 115.6	113.8, 113.8, 115.0	114.9, 114.9, 114.7	114.6, 114.6, 115.3
α, β, γ (°)	90, 90, 120	90, 90, 120	90, 90, 120	90, 90, 120	90, 90, 120	90, 90, 120	90, 90, 120	90, 90, 120	90, 90, 120	90, 90, 120
Resolution (Å)	75.5 – 2.85 (2.92 – 2.85)	75.4 – 2.71 (2.81 – 2.71)	75.1 – 2.95 (3.06 – 2.95)	74.3 – 3.08 (3.19 – 3.08)	74.3 – 3.96 (4.10 – 3.96)	75.2 – 2.80 (2.90 – 2.80)	75.4 – 2.99 (3.10 – 2.99)	74.8 – 2.79 (2.89 – 2.79)	75.2 – 2.95 (3.05 – 2.95)	75.2 – 3.12 (3.23 – 3.12)
R$_{meas}$(%)[1]	6.1 (162)	8.3 (217)	9.0 (220)	9.2 (143)	9.3 (123)	6.2 (158)	10.1 (202)	5.9 (140)	6.7 (154)	12.2 (169)
R$_{p.i.m.}$(%)[1]	1.4 (35.8)	1.4 (33.6)	2.0 (47.6)	2.1 (34.4)	2.2 (29.8)	2.0 (49.9)	1.6 (32.4)	1.4 (32.7)	1.5 (34.6)	2.8 (37.9)
I/σ	32.8 (2.02)	29.5 (2.1)	23.0 (2.0)	18.9 (1.7)	14.0 (2.0)	21.1 (1.5)	27.1 (2.4)	27.7 (2.1)	25.5 (1.9)	14.9 (1.7)
CC$_{1/2}$	99.9 (81.6)	99.8 (81.4)	100 (83.2)	100 (90.3)	100 (91.8)	99.8 (76.6)	100 (84.8)	100 (84.8)	100 (84.0)	99.8 (83.2)
Completeness (%)	100.0 (99.8)	99.8 (98.6)	99.9 (99.0)	99.2 (92.4)	99.1 (90.8)	99.9 (99.7)	99.9 (100)	99.1 (88.5)	99.8 (97.5)	99.6 (94.4)
Redundancy	19.8 (17.9)	39.7 (39.1)	19.9 (20.7)	18.8 (15.3)	18.5 (14.3)	9.95 (9.92)	38.2 (35.7)	20.0 (17.0)	19.4 (18.9)	19.7 (18.9)
Refinement										
Resolution (Å)	75.5 – 2.85	75.4 – 2.71	75.1 – 2.95	74.3 – 3.08	74.3 – 3.96	75.2 – 2.80	75.4 – 2.99	74.8 – 2.79	75.2 – 2.95	75.2 – 3.12
Unique reflections	21124	24330	18726	15788	7536	17866	18192	21687	18836	15858
R$_{work}$ / R$_{free}$	0.184 / 0.212	0.180 / 0.195	0.160 / 0.178	0.163 / 0.191	0.142 / 0.177	0.167 / 0.191	0.151 / 0.181	0.179 / 0.204	0.174 / 0.172	0.157 / 0.179
No. atoms										
RNA	2750	2803	2792	2771	2811	2752	2814	2800	2839	2772
Mg^{2+}	4	3	3	4	4	4	4	4	4	5
K$^+$	1	1	1	1	1	1	1	1	1	1
Na$^+$		2								
SO$_4$	5	5	10	5	5	5	5	5	5	5

Table 1. Data collection and refinement statistics for pri-miRNA loop fusion structures. [1]R-factors are defined as:

$$\mathrm{R}_{meas} = \sum_{hkl} \sqrt{N(hkl)/(N(hkl)-1)} \times \sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl} \sum_i I_i(hkl)$$

$$\mathrm{R}_{p.i.m.} = \sum_{hkl} \sqrt{1/(N(hkl)-1)} \times \sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl} \sum_i I_i(hkl)$$

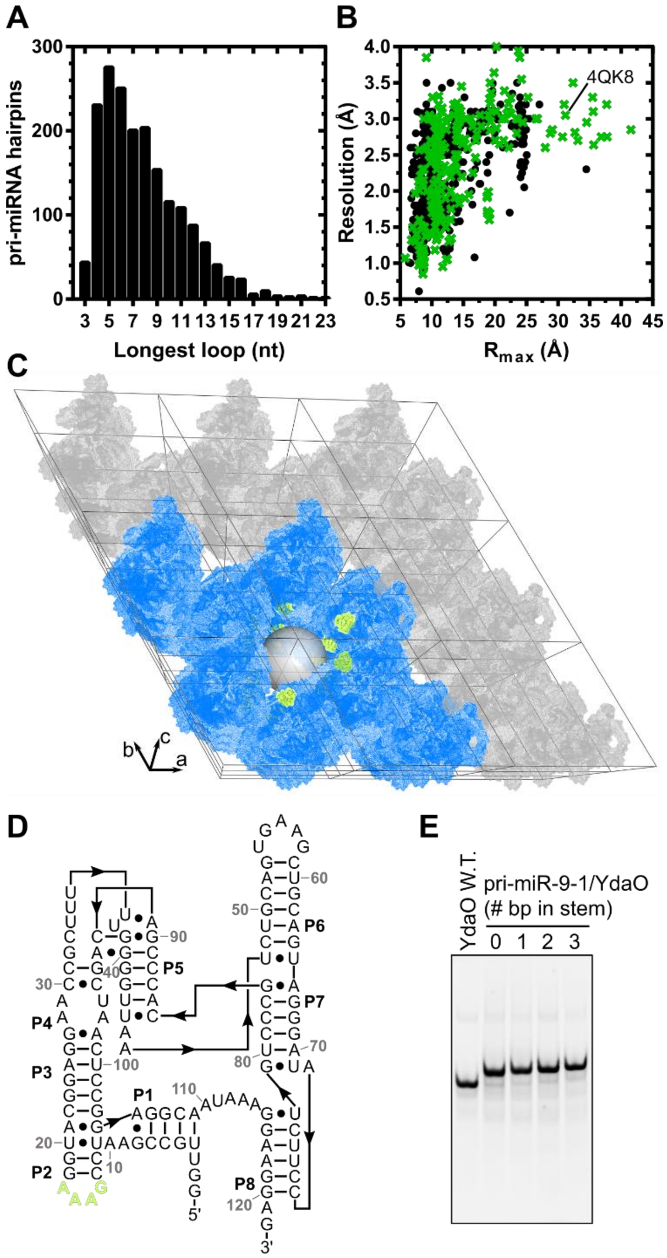


Figure 1. Analysis of pri-miRNA terminal loops and potential crystallization scaffolds. (A) Distribution of pri-miRNA terminal loop lengths calculated from their predicted mfold secondary structures. (B) Comparison of the largest spherical cavity (with radius R_{max}) present in each RNA crystal structure against the diffraction resolution of the structure. Crystal forms with a single molecule in the asymmetric unit are shown in green, and all others in black. (C) Crystal packing in the YdaO-type ci-di-AMP riboswitch (PDB ID 4QK8). Molecules surrounding a large central channel are colored blue, and a grey sphere (radius 31 Å) is positioned in the channel to show its size. Stem loops terminating inside the

channel are green. (D) Secondary structure of the YdaO riboswitch. (D) Native gel of W.T. YdaO and fusions between the YdaO scaffold and the pri-miR-9-1 terminal loop with 0-3 base pair from the stem.

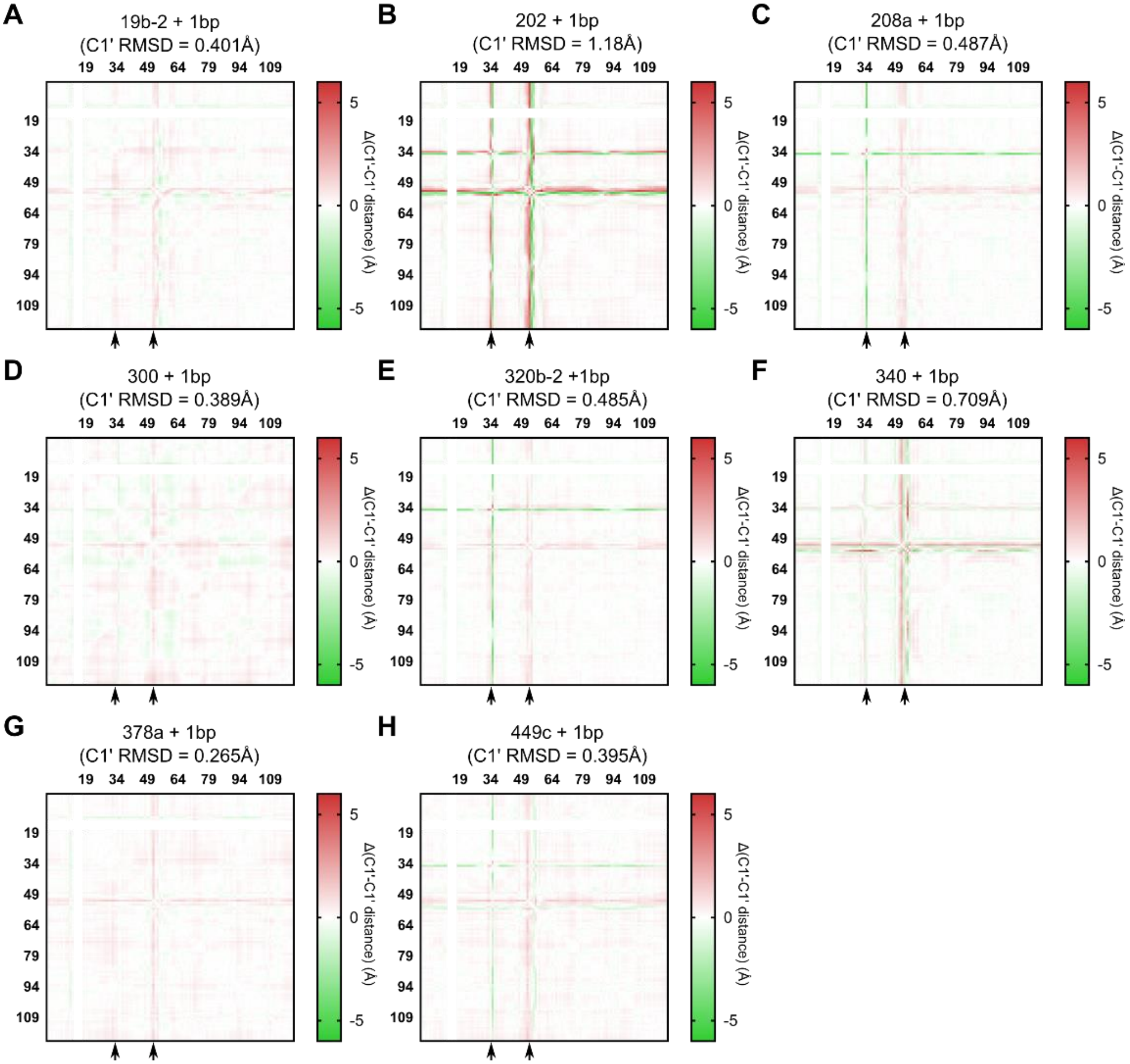


Figure 2. Distance difference matrices for C1′ atom positions shared between YdaO W.T. (PDB ID 4QK8) and pri-miRNA loop fusion structures. All panels are scaled to the same distance range (-5 to 5 Å). The RMSD over all C1′ atoms is shown. Arrows indicate a flexible internal loop (residues 34-35) and the terminal loop of stem P6 (vicinity of residue 55) which are not well resolved in the electron density and so have variable conformations in the different structures.

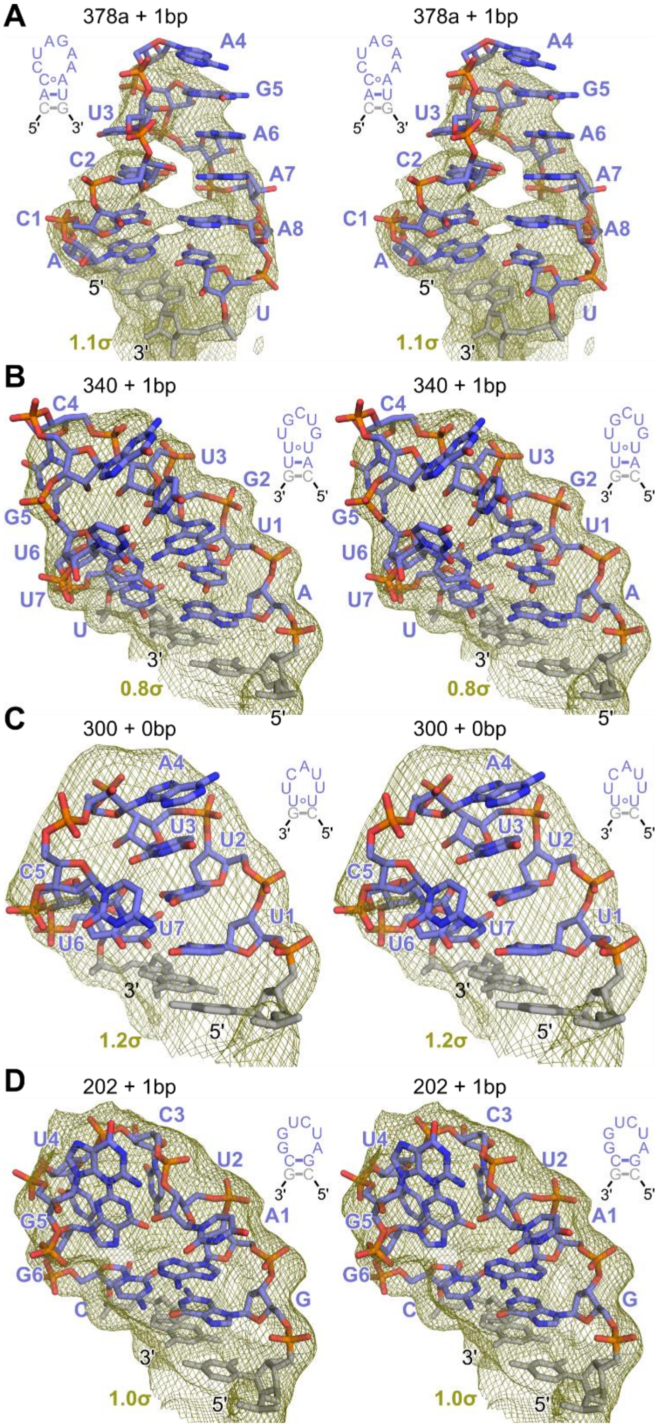


Figure 3. Atomic structures of pri-miRNA terminal loops determined by scaffold-directed crystallography. Each structure appears in stereographic view, where the last base pair from the scaffold P2 stem is colored grey and all pri-miRNA residues are blue. Inset shows the secondary structure of the

loop. The 2Fo-Fc electron density map is contoured at the level shown in each panel. (A) pri-miR-378a (378a + 1bp). (B) pri-miR-340 (340 + 1bp). (C) pri-miR-300 (300 + 0bp). (D) pri-miR-202 (202 + 1bp).

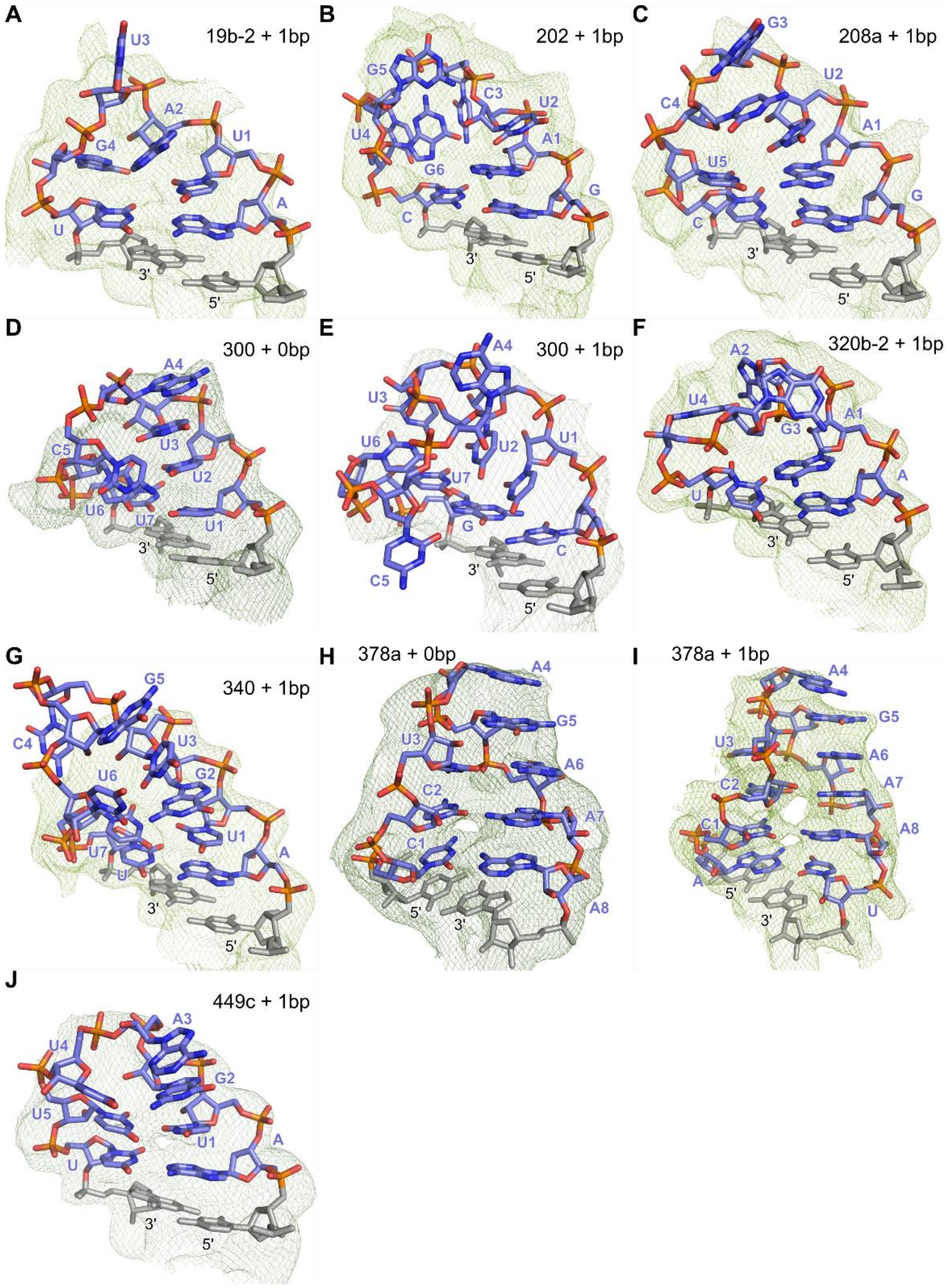


Figure 4. Simulated annealing composite omit maps calculated for all pri-miRNA loops. Color scheme is the same as Figure 3. All maps are contoured to 1.1σ. See the Material and Methods section for details on calculation of individual maps.

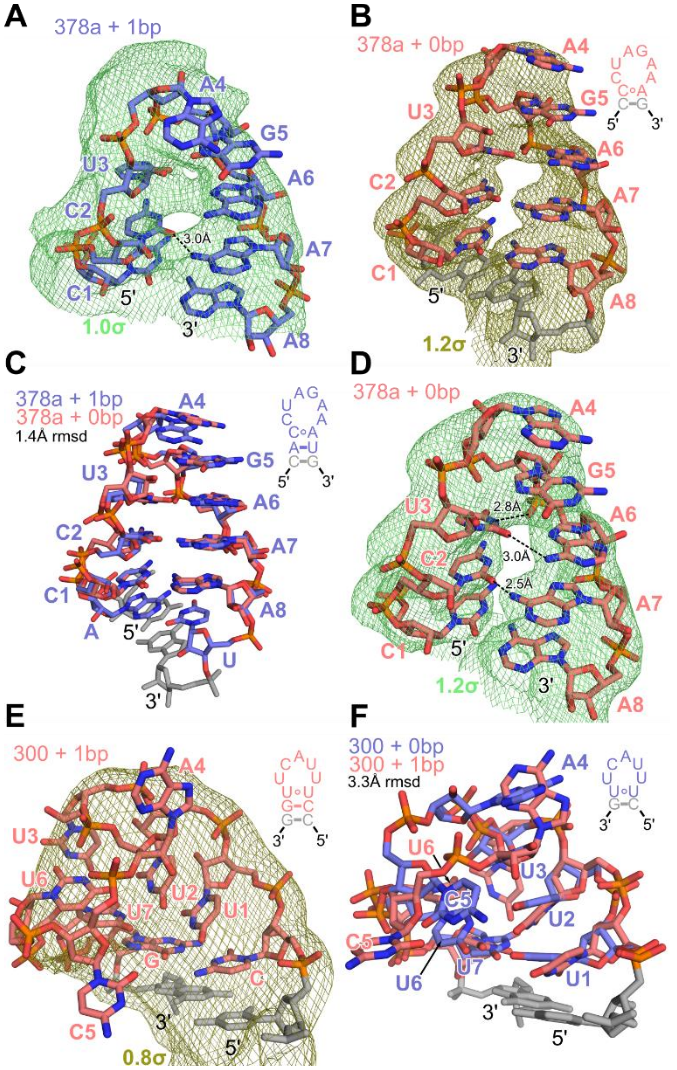


Figure 5. Comparison of loop structures with and without the terminal base pair of the pri-miRNA stem. (A) 378a + 1bp showing C2-A7 H-bond. Throughout the figure, density represents the $2F_o$-F_c electron density map contoured at the indicated σ level. (B) Structure of pri-miR-378a terminal loop with no base pair from the stem (378a + 0bp). (C) Alignment between 378a + 1bp structure (Figure 3A) and 378a + 0bp. The RMSD of all heavy atoms in the loop is shown. (D) Alternative view of 378a + 0bp showing H-bond interactions. (E) Structure of the pri-miR-300 terminal loop with one base pair from the stem (300 + 1bp). (D) Alignment of 300 + 0bp (Figure 3C) and 300 + 1bp loops.

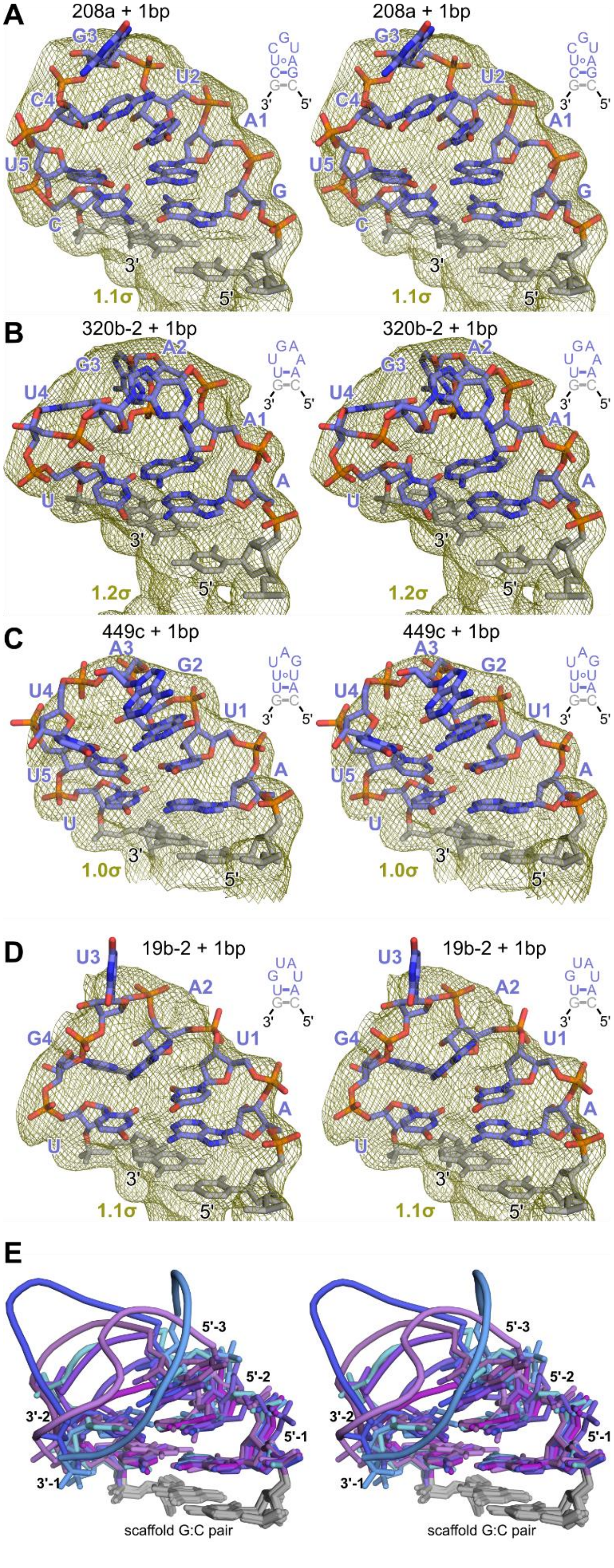

A
208a + 1bp
G3
U2
C4
A1
U5
G
C
3'
5'
1.1σ
C G U
U∘A
C–G
G=C
3' 5'
B
320b-2 + 1bp
G3
A2
U4
A1
A
U
3'
5'
1.2σ
GA
U A
U–A
G=C
3' 5'
C
449c + 1bp
A3
G2
U4
U1
A
U5
U
3'
5'
1.0σ
U A G
U∘U
U–A
G=C
3' 5'
D
19b-2 + 1bp
U3
A2
G4
U1
A
U
3'
5'
1.1σ
UA
G U
U–A
G=C
3' 5'
E
5'-3
5'-2
5'-1
3'-2
3'-1
scaffold G:C pair

Figure 6. Structures of shorter pri-miRNA loops (4-5nt). For (A)-(D) color scheme is identical to Figure 3. (A) pri-miR-208a (208a + 1bp). (B) pri-miR-320b-2 (320b-2 + 1bp). (C) pri-miR-449c (449c + 1bp). (D) pri-miR-19b-2 (19b-2 + 1bp). (E) Structural alignment of all eight loops shown in Figures 3 and 4. Positions that align between most or all of the structures are labeled.

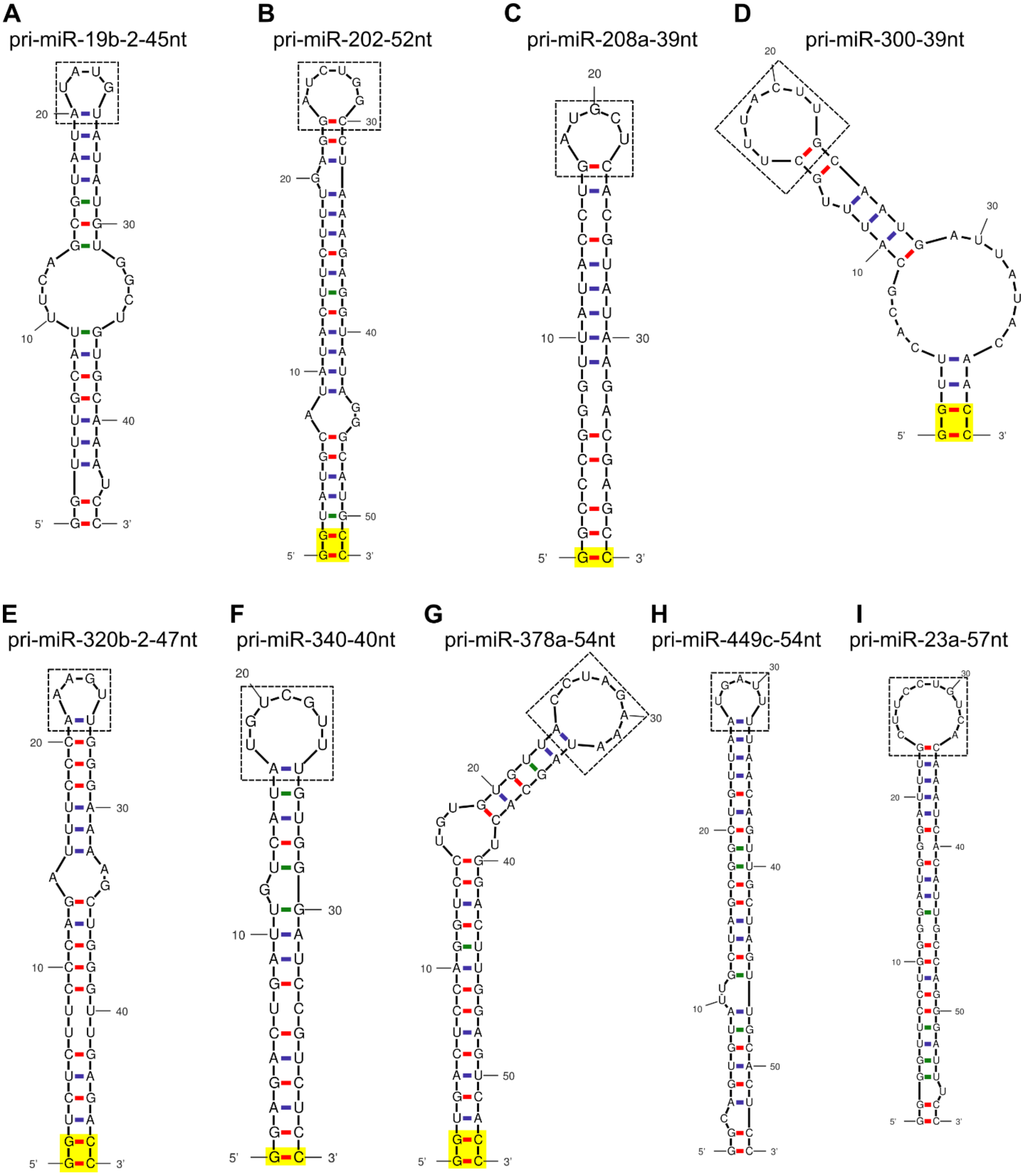


Figure 7. Secondary structure predictions for all pri-miRNA fragments used in the Rhed binding assay. Additional G-C pairs added to the base of the stem to enhance transcription are highlighted in yellow. The box shows the sequence of loop and terminal base pair of the stem used to determine crystal structures.

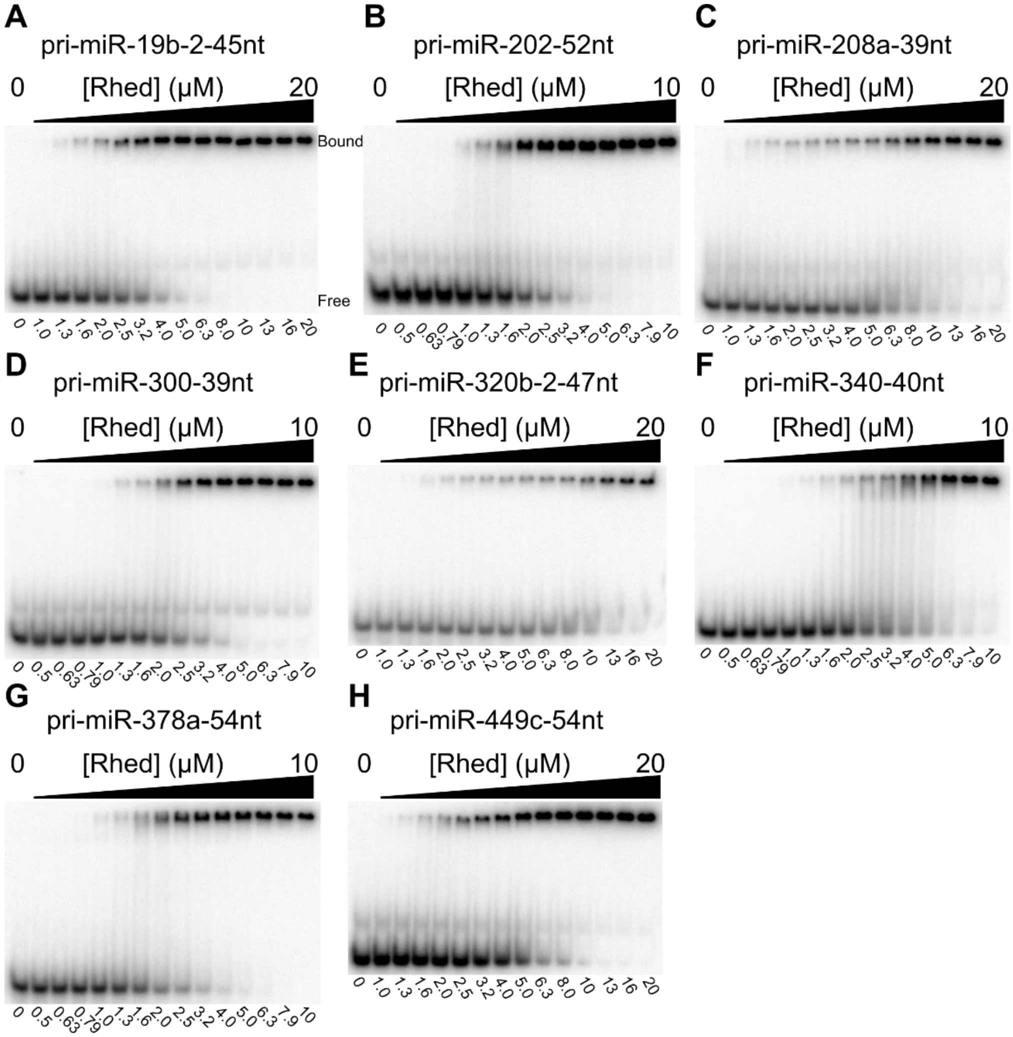


Figure 8. Example gel shift assays for each pri-miRNA fragment binding to the Rhed. The free RNA and protein-bound species are labeled in (A). Rhed dimer concentrations (µM) used in the binding reactions are shown below the gel.

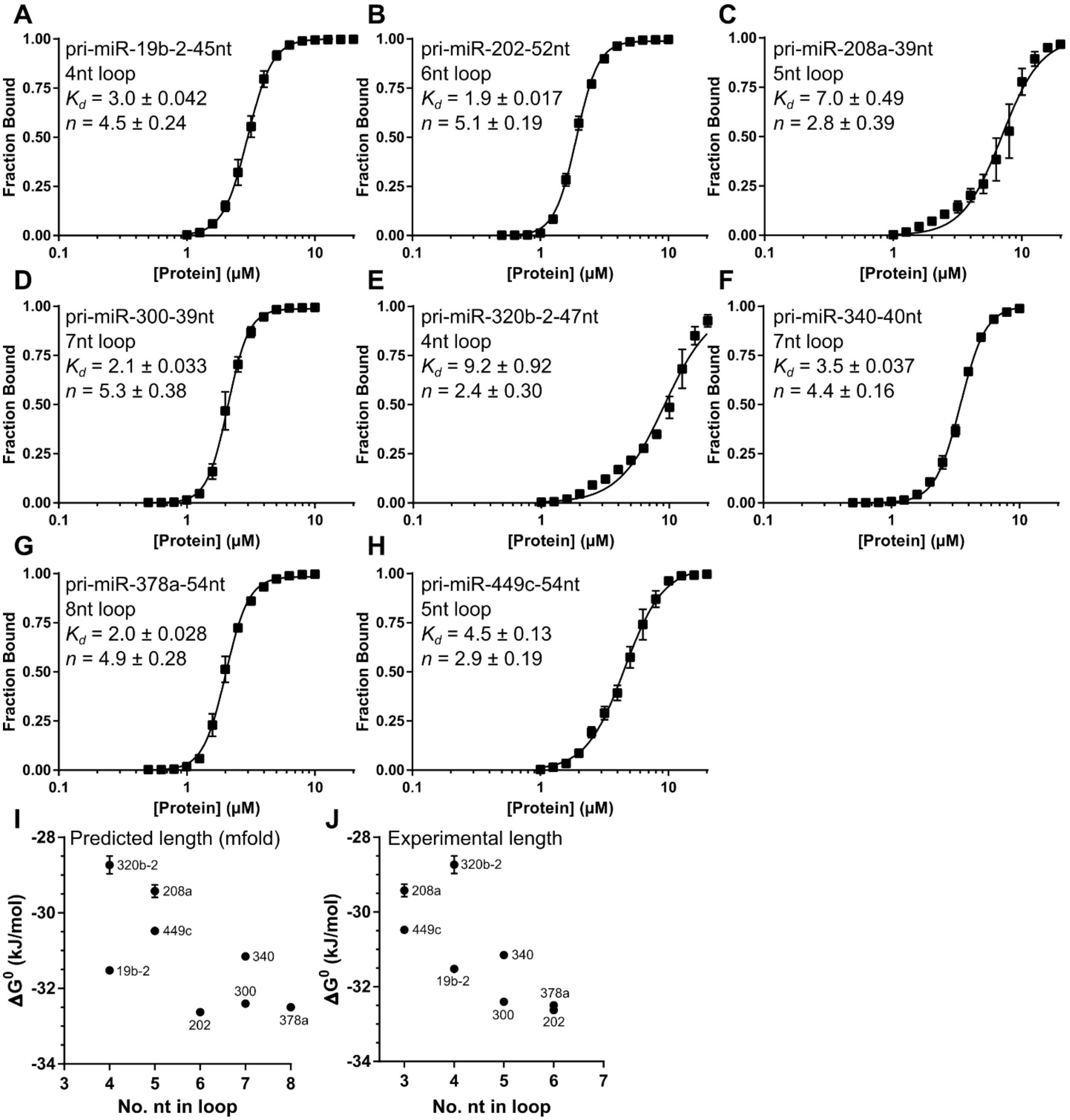


Figure 9. Association of the Rhed with pri-miRNA apical junctions. (A) – (H) Quantification of gel shift assays shown in Figure 8. Data points represents the mean fraction bound ± standard error from three replicate experiments. Data were fit with the Hill equation and the dissociation constant (K_d) and Hill coefficient (n) are shown (± SE). (I) Comparison of the free energy of Rhed binding (RTln(K_d)) to the length of the terminal loop, as measured by MFOLD. (J) Same as (I) but with loop length corrected based on our crystal structures.

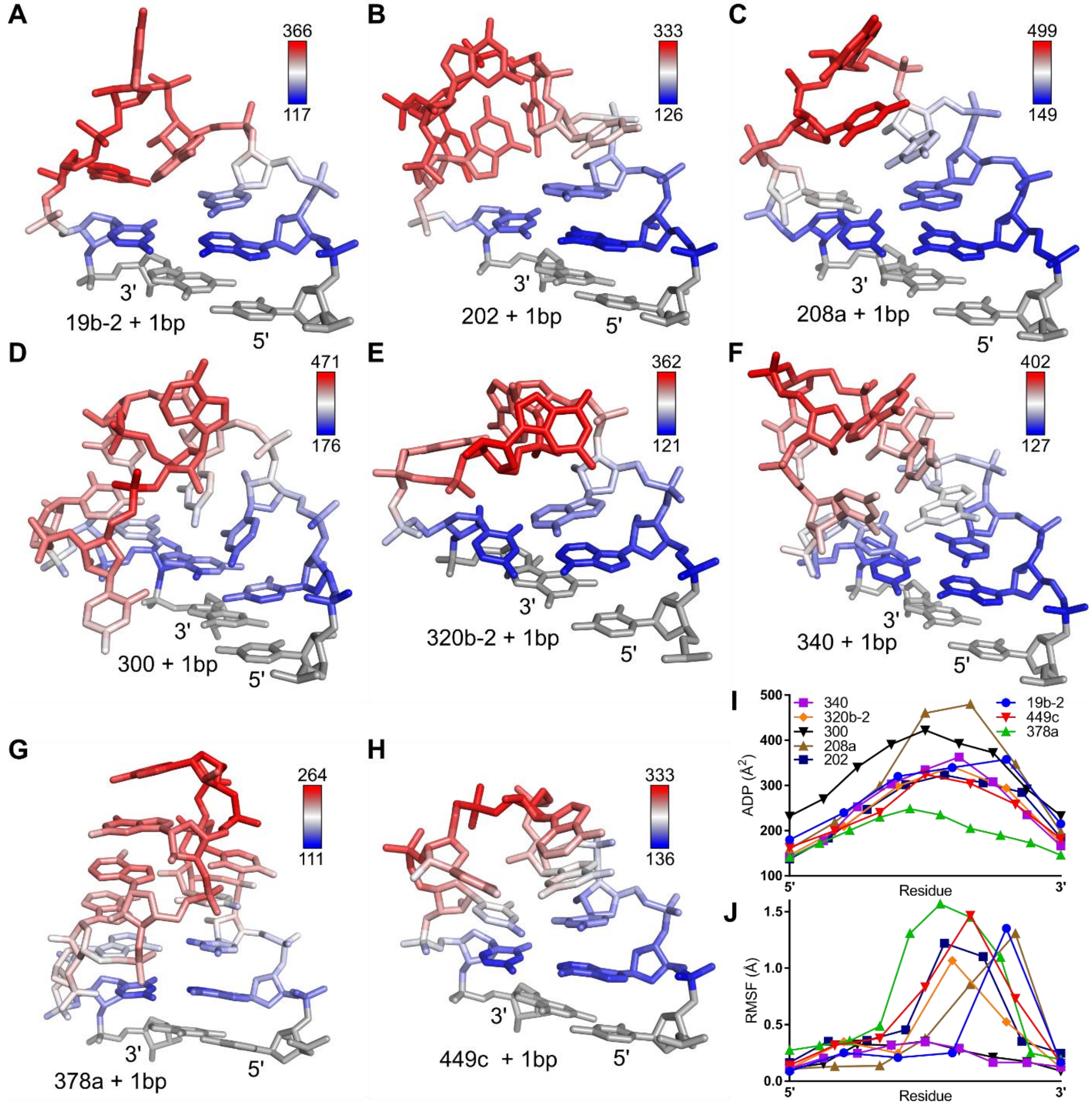


Figure 10. Estimating the flexibility of the terminal loop with atomic displacement parameters (ADPs). Each structure in (A) – (H) is colored with the lowest ADPs in blue to the highest in red. The inset shows the range of ADP plotted. (I) Average ADP per residue, with all loops plotted on the same scale. The 5' and 3' end represent the terminal base pair of the pri-miRNA stem loop. (J) Root-mean-square fluctuations (RMSF, Å) determined for each residue by molecular dynamics. Symbols and coloring are identical to (I).

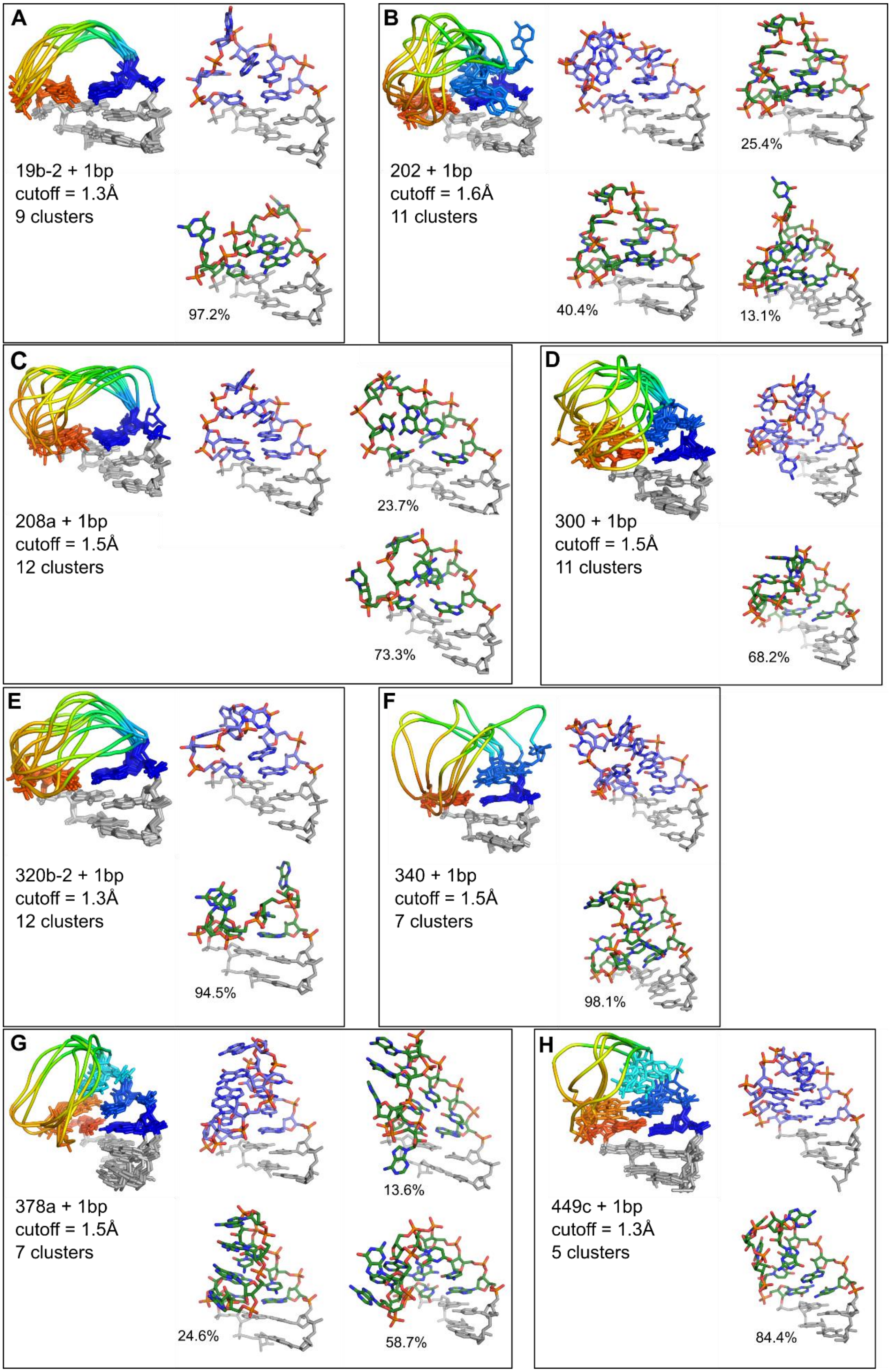

A
19b-2 + 1bp
cutoff = 1.3Å
9 clusters
97.2%
B
202 + 1bp
cutoff = 1.6Å
11 clusters
25.4%
40.4%
13.1%
C
208a + 1bp
cutoff = 1.5Å
12 clusters
23.7%
73.3%
D
300 + 1bp
cutoff = 1.5Å
11 clusters
68.2%
E
320b-2 + 1bp
cutoff = 1.3Å
12 clusters
94.5%
F
340 + 1bp
cutoff = 1.5Å
7 clusters
98.1%
G
378a + 1bp
cutoff = 1.5Å
7 clusters
13.6%
24.6%
58.7%
H
449c + 1bp
cutoff = 1.3Å
5 clusters
84.4%

Figure 11. Cluster analysis of MD trajectories for simulations of eight pri-miRNA loops. In all panels, grey residues represent base pairs from the scaffold that were restrained during the simulation. Rainbow colored structures are alignments of multiple clusters (5’ blue to 3’ red). For comparison, the original crystal structure is shown in light blue. Individual clusters are shown in green, and the fraction of simulation time spent in this conformation is given as a percentage next to each structure.

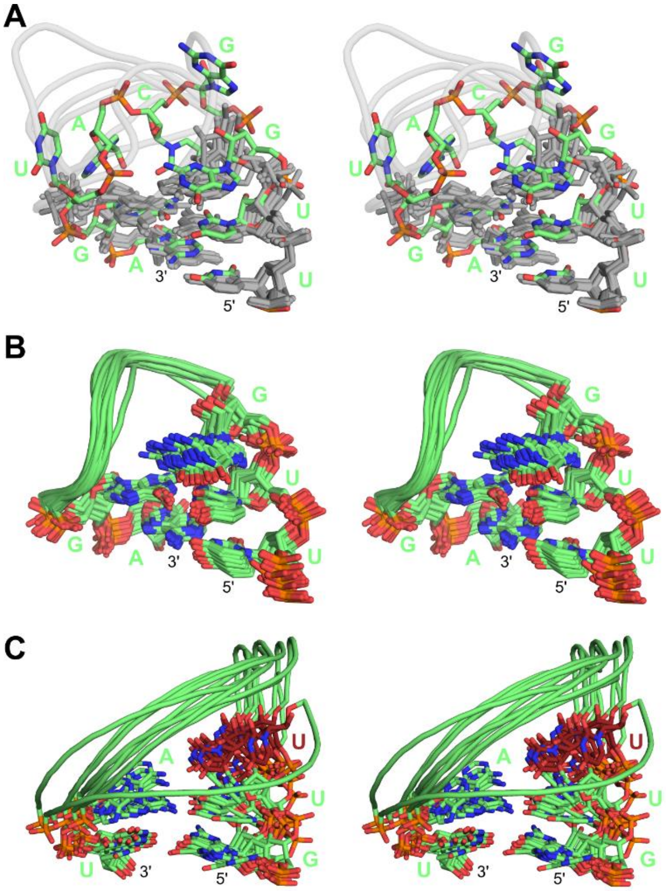


Figure 12. Comparison of apical junction crystal structures to NMR structures. (A) Stereographic view of the alignment between pri-miRNA structures from Figure 4E (grey) and the NMR structure of pre-miR-20b (PDB ID: 2N7X). Only one representative of the NMR ensemble is shown for clarity. The pre-miR-20b junction and loop sequence 5'-UUGGCAUGA-3' is labeled and colored green. (B) NMR ensemble for pri-miR-20b, showing the stable positioning of the G residue at the 5' end of the loop above the apical junction (U-G pair). (C) NMR ensemble structure of the apical junction and loop of pre-miR-21 (PDB ID: 5UZT). The semi-ordered U-A pair (green) forms the apical junction, and a U at the 5' end of the loop stacks on top (red residue).

References

[1] V. N. Kim, J. Han, and M. C. Siomi, “Biogenesis of small RNAs in animals,” *Nat. Rev. Mol. Cell Biol.*, vol. 10, no. 2, pp. 126–139, Feb. 2009.

[2] Y. Lee, K. Jeon, J.-T. Lee, S. Kim, and V. N. Kim, “MicroRNA maturation: stepwise processing and subcellular localization.,” *EMBO J.*, vol. 21, no. 17, pp. 4663–70, Sep. 2002.

[3] Y. Zeng, R. Yi, and B. R. Cullen, “Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha.,” *EMBO J.*, vol. 24, no. 1, pp. 138–48, Jan. 2005.

[4] Y. Zeng and B. R. Cullen, “Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences.,” *J. Biol. Chem.*, vol. 280, no. 30, pp. 27595–603, Jul. 2005.

[5] J. Han *et al.*, “Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex.,” *Cell*, vol. 125, no. 5, pp. 887–901, Jun. 2006.

[6] X. Zhang and Y. Zeng, “The terminal loop region controls microRNA processing by Drosha and Dicer,” *Nucleic Acids Res.*, vol. 38, no. 21, pp. 7689–7697, Nov. 2010.

[7] X. Xiong, X. Kang, Y. Zheng, S. Yue, and S. Zhu, “Identification of loop nucleotide polymorphisms affecting MicroRNA processing and function,” *Mol. Cells*, vol. 36, no. 6, pp. 518–526, Dec. 2013.

[8] M. Trabucchi *et al.*, “The RNA-binding protein KSRP promotes the biogenesis of a subset of microRNAs,” *Nature*, vol. 459, no. 7249, pp. 1010–1014, Jun. 2009.

[9] M. Ha and V. N. Kim, “Regulation of microRNA biogenesis,” *Nat. Rev. Mol. Cell Biol.*, vol. 15, no. 8, pp. 509–524, Aug. 2014.

[10] J. M. Burke, D. P. Kelenis, R. P. Kincaid, and C. S. Sullivan, “A central role for the primary microRNA stem in guiding the position and efficiency of Drosha processing of a viral pri-miRNA.,” *RNA*, vol. 20, no. 7, pp. 1068–77, Jul. 2014.

[11] H. Ma, Y. Wu, J.-G. Choi, and H. Wu, “Lower and upper stem-single-stranded RNA junctions together determine the Drosha cleavage site.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 51, pp. 20687–92, Dec. 2013.

[12] J. Quick-Cleveland, J. Jacob, S. Weitz, G. Shoffner, R. Senturia, and F. Guo, “The DGCR8 RNA-Binding Heme Domain Recognizes Primary MicroRNAs by Clamping the Hairpin,” *Cell Rep.*, vol. 7, no. 6, 2014.

[13] T. A. Nguyen *et al.*, “Functional Anatomy of the Human Microprocessor,” *Cell*, vol. 161, no. 6, pp. 1374–1387, Jun. 2015.

[14] V. C. Auyeung, I. Ulitsky, S. E. McGeary, and D. P. Bartel, “Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing.,” *Cell*, vol. 152, no. 4, pp. 844–58, Feb. 2013.

[15] W. Fang and D. P. Bartel, “The Menu of Features that Define Primary MicroRNAs and Enable De Novo Design of MicroRNA Genes,” *Mol. Cell*, vol. 60, pp. 131–145, 2015.

[16] C. R. Alarcón, H. Lee, H. Goodarzi, N. Halberg, and S. F. Tavazoie, “N6-methyladenosine marks primary microRNAs for processing,” *Nature*, vol. 519, no. 7544, pp. 482–485, Mar. 2015.

[17] C. Roden *et al.*, “Novel determinants of mammalian primary microRNA processing revealed by systematic evaluation of hairpin-containing transcripts and human genetic variation.,” *Genome Res.*, vol. 27, no. 3, pp. 374–384, Mar. 2017.

[18] N. Terasaka, K. Futai, T. Katoh, and H. Suga, “A human microRNA precursor binding to folic acid discovered by small RNA transcriptomic SELEX.,” *RNA*, vol. 22, no. 12, pp. 1918–1928, Dec. 2016.

[19] Y. Chen *et al.*, “Rbfox proteins regulate microRNA biogenesis by sequence-specific binding to their precursors and target downstream Dicer,” *Nucleic Acids Res.*, vol. 44, no. 9, pp. 4381–4395, May 2016.

[20] M. D. Shortridge, M. J. Walker, T. Pavelitz, Y. Chen, W. Yang, and G. Varani, “A Macrocyclic Peptide Ligand Binds the Oncogenic MicroRNA-21 Precursor and Suppresses Dicer Processing,” *ACS Chem. Biol.*, vol. 12, no. 6, pp. 1611–1620, Jun. 2017.

[21] S. Chirayil, Q. Wu, C. Amezcua, and K. J. Luebke, “NMR Characterization of an Oligonucleotide Model of the MiR-21 Pre-Element,” *PLoS One*, vol. 9, no. 9, p. e108231, Sep. 2014.

[22] A. Gao and A. Serganov, “Structural insights into recognition of c-di-AMP by the ydaO riboswitch,” *Nat. Chem. Biol.*, vol. 10, no. 9, pp. 787–792, Sep. 2014.

[23] A. F. Moon, G. A. Mueller, X. Zhong, and L. C. Pedersen, “A synergistic approach to protein crystallization: Combination of a fixed-arm carrier with surface entropy reduction,” *Protein Sci.*, vol. 19, no. 5, p. NA-NA, May 2010.

[24] M. Zuker, “Mfold web server for nucleic acid folding and hybridization prediction.,” *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3406–15, Jul. 2003.

[25] C. Kao, M. Zheng, and S. Rüdisser, “A simple and efficient method to reduce nontemplated nucleotide addition at the 3 terminus of RNAs transcribed by T7 RNA polymerase.,” *RNA*, vol. 5, no. 9, pp. 1268–72, Sep. 1999.

[26] W. Kabsch, “XDS,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 66, no. 2, pp. 125–132, Feb. 2010.

[27] P. D. Adams *et al.*, “PHENIX: A comprehensive Python-based system for macromolecular structure solution,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 66, no. 2, pp. 213–221, Feb. 2010.

[28] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan, “Features and development of Coot,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 66, no. 4, pp. 486–501, Apr. 2010.

[29] D. Liebschner *et al.*, “Polder maps: improving OMIT maps by excluding bulk solvent,” *Acta Crystallogr. Sect. D Struct. Biol.*, vol. 73, no. 2, pp. 148–157, Feb. 2017.

[30] J. F. Milligan, D. R. Groebe, G. W. Witherell, and O. C. Uhlenbeck, “Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates.,” *Nucleic Acids Res.*, vol. 15, no. 21, pp. 8783–98, Nov. 1987.

Chapter 4: Cancer-associated mutations render the DGCR8 protein defective in pri-miRNA processing

Grant Shoffner[1], Jen Quick-Cleveland[1], Kristina Solorio[1], Jose P. Jacob[1], and Feng Guo[1,2]

[1] Department of Biological Chemistry, David Geffen School of Medicine,

[2] Molecular Biology Institute,

University of California, Los Angeles, California

G. Shoffner and J. Quick-Cleveland contributed equally to this article.

Corresponding author: Feng Guo, University of California, Los Angeles, 611 Charles E. Young Drive East, 202 Boyer Hall, Los Angeles, CA 90095. Phone: 310-206-4576; Fax: 310-206-1929; Email: fguo@mbi.ucla.edu.

Running title: Cancer-associated DGCR8 mutations cause microRNA defects

Abstract (238/250 words)

Tumors can harbor many mutations, but distinguishing disease-driving from passenger mutations remains a challenge in cancer etiology. Mutations that drive tumorigenesis provide crucial targets for therapeutic development. The microRNA (miRNA) biogenesis pathway is directly linked to tumor formation and disease progression, and miRNAs are often globally repressed in cancerous cells. Drosha and DGCR8, which form the core of the Microprocessor complex (MC), are frequently mutated in tumors. The mechanistic details that underlie global repression of miRNAs, and how susceptible the MC is to genetic inactivation in cancer cells remains unknown. In this work, we selected seven tumor-derived mutations in DGCR8 for detailed characterization using cellular and biochemical assays. We found that four of these mutations cause substantial loss of pri-miRNA processing activity. We discovered that E518K, a mutation found in ~3% of Wilms tumors, is inactive in pri-miRNA processing and this defect is due to misfolding of the mutant protein. The DGCR8 protein carrying F448L, a mutation found in colon cancer, does not only lose its pri-miRNA processing activity, but also cause a moderate dominant negative effect when expressed in cells expressing wild-type DGCR8. The DGCR8-F448L mutant does not bind the essential Fe(III) heme cofactor. This mutation also disrupts the association between DGCR8 and its ribonuclease partner Drosha. The mutations K289E and G336E both render DGCR8 partially active, by decreasing the DGCR8 affinities for Fe(III) heme and pri-miRNAs, as well as reducing the protein structural stability. Altogether, we have identified four DGCR8 missense mutations that severely disrupt the miRNA biogenesis pathway and thereby are likely to promote neoplastic processes. Further, our study link defective mutants to molecular mechanism underlying DGCR8 structure and function relationship. In particular, we reveal a new function for the Rhed in promoting MC assembly in cells.

Introduction

The genesis and development of cancer remains an engrossing biological problem. The intricacy of the molecular pathways involved has made progress in research and successful treatment challenging. Data collected on cancer tissues has led to the insight that there are many mutations in cancerous cells (1–3). Untangling disease-driving versus spurious mutations presents a daunting challenge. This is further complicated by that fact that collective mutations often differ between patients, even with the same kind of cancer (4,5). To overcome the challenge in treating heterogeneous tumors, it is important to identify potential driver mutations for pathogenesis. Estimates for mutations that drive cancer development range from 10,000 (5) to, more recently, as few as 3 (6), and vary depending on cancer type.

miRNA dysregulation is a hallmark of malignant cells (7–9). Coukos *et al.* discovered that there is a high frequency of genomic alterations at miRNA loci (1). About 50% of miRNA genes were found in cancer-associated genomic regions linking miRNAs with cancer pathogenesis (11). Specific miRNAs are often overexpressed or suppressed in cancer cells. Additionally, there is a global down-regulation of miRNA expression in many cancers (2). Some miRNAs, such as let-7, miR-143/145, miR-200 are tumor suppressors (11–13). Some others, such as the miRNA cluster miR-17~92 and miR-155 are oncogenic (14). In non-cancerous cells, global reduction of miRNA expression has an overall effect of favoring cellular transformation and tumorigenesis (3,4).

miRNA biogenesis proteins undergo changes that are strongly implicated in cancer development. In the canonical miRNA maturation pathway, primary transcripts (pri-miRNAs) are cleaved in the nucleus by the Microprocessor Complex (MC), which contains the RNase Drosha and its obligate RNA-binding partner DGCR8. The resulting precursor miRNAs (pre-miRNA) leave the nucleus via Exportin-5. Once in the cytoplasm, the RNase Dicer performs another cleavage step. The single-stranded ~22-nt mature miRNAs are incorporated into effector Argonaute proteins forming the *R*NA- *i*nduced *s*ilencing *c*omplex (RISC.). RISC uses the mature miRNA as a guide to bind cognate mRNAs ultimately leading to repression of their expression (16–18). Expression of Drosha and Dicer is often decreased in ovarian cancer specimens and such decreases correlate with poor prognosis (5). Reduction of DGCR8 or Dicer expression enhances cellular transformation and tumorigenesis in mice. This demonstrates a direct contribution of abnormal miRNA biogenesis to tumor formation (3,4). Hypoxia, another hallmark of cancerous cells, causes coordinated repression of miRNA biogenesis proteins Dicer, Drosha, TARPB2, XPO5 and DGCR8 in breast cancer cell lines (6). Drosha and DGCR8 mutations and deletions appear in 15% of Wilms tumors, a cancer that originates in the kidney and is the most common pediatric renal malignancy in the United States (7,8). Conversely, DGCR8 specifically plays a role in prostate cancer progression. In *PTEN* knock-out mouse model for prostate cancer, loss of DGCR8 expression inhibits tumor progression (9). Expression of the miRNA biogenesis proteins varies depending on cancer type; in prostate cancer these proteins are upregulated, while in most lung and ovarian cancers they are downregulated (9). This tissue-specific regulation and role of miRNA in cancer cells makes dissecting general molecular mechanisms even more challenging.

DGCR8 is a unique RNA-binding protein. It contains a RNA-binding heme domain (Rhed) in the central region that is required for recognition and processing of pri-miRNA substrates (Fig. 1A) (10,11). Rhed is flanked by an N-terminal region required for nuclear localization and tandem double-stranded RNA-binding domains (dsRBD1 and dsRBD2) on the C-terminal side. Using the Rhed, DGCR8 forms a dimer to bind a ferric heme cofactor (12,13). DGCR8 binds Fe(III) heme with >10^7-fold higher affinity than that for the Fe(II) form and, importantly, only Fe(III) heme can activate pri-miRNA processing activity (14). DGCR8 ligates the heme iron using two cysteine side chains, one (C352) contributed by each subunit (15). This heme ligation configuration generates an electron-rich environment for the heme iron that confers extraordinary redox specificity to DGCR8. In addition, DGCR8 is regulated by phosphorylation (16), sumoylation (17), acetylation (18), caspase-mediated proteolytic cleavage (19), feedback cleavage of pri-miRNA-like hairpins in its mRNA (20,21) and competition from non-pri-miRNAs (22). Therefore, the DGCR8 protein is a point of convergence between many regulatory pathways that control miRNA biogenesis.

The motivation of the current study is two-fold. First, since DGCR8 is frequently mutated in tumors, we wanted to investigate whether these changes in DGCR8 could cause pri-miRNA processing defects. Second, biochemical characterization of cancer-associated DGCR8 mutations would shed light on the structure-and-function relationship of this protein. Using a combination of cellular and biochemical assays, we show that several cancer-associated point mutations disrupt pri-miRNA processing activity, thereby are likely to contribute to pathogenesis. Some of the mutations highlight the importance of the Rhed in DGCR8 function.

Materials and Methods

Live-cell pri-miRNA processing reporter assay

Tet-on Hela and U2-OS cells were grown at 37°C with 5% CO_2 in DMEM medium containing 4.5 g/L glucose, L-glutamine and 110 mg/L sodium pyruvate supplemented with 10% Tet-System-Approved FBS (Clonetech Laboratories). For imaging, cells were first passaged into 3.5-cm dishes with a glass coverslip bottom. Cells were transfected at ~80% confluency with Effectene (Qiagen), and the expression of reporters was immediately induced by addition of 1 μg/mL doxycycline. About 20 hr posttransfection, cells were washed and exchanged into DMEM without phenol red. Fluorescence imaging was performed using a Nikon Eclispse Ti microscope with an EMCCD camera (Andor iXon). We used a 10x objection lens and an exposure time of 0.3 ms with gain at 0. Fluorescence was detected with filter cubes for eYFP (excitation filter at 510 ± 10 nm band pass and emission filter at 535 ± 15 nm band pass) and mCherry (excitation filter at 535 ± 25 nm band pass and emission filter at 610 ± 10 nm band pass). Images were exported as 16-bit tiff files.

Images were analyzed mostly as described (23), except that a new Matlab (The MathWorks) script was written for integrating fluorescence intensities of individual cells in a streamlined fashion. Using the image analysis toolbox in Matlab, our program first performs segmentation to determine the locations of individual nuclei from pairs of eYFP and mCherry images acquired from the same field of view. A mask is created such that pixels with intensity above a user-defined threshold are considered to be in nuclei (both eYFP and mCherry expressed contain a nuclear localization sequence) and are assigned a value of 1. The values of all other pixels are set to 0. We choose threshold values that maximize the number of cells detected. The perimeter of a nucleus is determined for each image of a pair separately, and then a common mask is chosen for the two images based on the following criteria. If the same cell is found in both images, the mask with smaller perimeter is chosen to avoid including cytosols, which sometimes have low fluorescent signals. If a cell is detected in only one image, the perimeter is kept, allowing for cells that are dim in the other channel to be included. The center of the final perimeter for each cell is determined. A background fluorescence value is calculated for each image by averaging the intensities of the pixels outside the identified nuclei. The background is then subtracted from the intensities of all pixels. The total intensity of an individual cell is calculated as the sum of the pixel intensities within the perimeter. Dead cells often have very strong fluorescent signals, therefore cells with top 10% fluorescent intensities are excluded in the analyses. Overlapping cells that are detected as one object and can be removed manually. A list of total eYFP and mCherry intensities and their center coordinates for individual cells is exported to Microsoft Excel, scatter plots are generated, and linear fit of eYFP versus mCherry signals, forced to go through the origin, gives the fluorescence slopes.

To examine the expression level of N-flag-DGCR8, we prepared nuclear extracts from transfected cells and performed immunoblotting using an anti-DGCR8 antibody as described previously (19).

DGCR8 bacterial expression and purification

NC1, Rhed, and DSD, including WT and tumor-derived mutants, were expressed and purified as previously reported (24). Briefly, DGCR8 proteins were expressed in *E. coli* strain BL21(DE3) CodonPlus (Agilent Technologies). Cells were grown to late-log phase. Expression was induced with 1 mM IPTG for 4 hr at 37°C. For NC1 and Rhed, δ-aminolevulinic acid was added to a final concentration of 1 mM to facilitate heme loading to the proteins. Cells were harvested by centrifugation at 5000xg, for 5 min at 4°C. Protein was initially purified using a SP cation-exchange column (GE Healthcare). Column equilibration and loading were performed using Buffer *A*, which contained 100 mM NaCl and 20 mM Tris pH 8.0. Linear gradient elution was performed with Buffer B containing 2 M NaCl and 20 mM Tris pH 8.0. Fresh dithiothreitol (DTT) was added to buffers at a final concentration of 1 mM. Fractions containing the DGCR8 protein were concentrated in a centrifugal concentrator with appropriate MWCO (Millipore), and loaded onto a Superdex-200 size exclusion chromatography column (GE Healthcare) equilibrated with a SEC buffer containing 400 mM NaCl and 20 mM Tris pH 8.0. Peak fractions were again concentrated and analyzed using a SDS 12%-polyacrylamide gel to determine purity.

Pri-miRNA binding and processing assays

The pri-miRNA binding assays were performed as described previously (12). Briefly, serial dilutions of purified DGCR8 protein were incubated with a trace amount of pri-miRNA uniformly labeled with ^{32}P in 100 mM NaCl and 20 mM Tris pH 8.0 at room temperature for 30 min. The mixtures were filtered through stacked nitrocellulose (EMD Millipore) and Hybond-N+ positively changed nylon (GE Healthcare) membranes. The autoradiography images of the membranes were analyzed using Quantity One (Bio-Rad, version 4.4.1). The data were fit using PRISM (GraphPad, version 6).

Reconstituted pri-miRNA processing assays were performed as described (24).

Heme dissociation assay

Purified DGCR8 protein at 3-20 μM concentration was incubated in SEC buffer at room temperature with 4x or 5x molar excess of horse skeletal apomyoglobin (SIGMA). Electronic absorption spectra were monitored over time using a double-beam Varian CARY 300 Bio spectrophotometer with spectral bandwidth set to 1.0 nm, and averaging time at 0.1 sec. The scanning kinetics program was used to automatically collect scans at set intervals. The absorption values over time were fit with a one-phase exponential decay equation using PRISM.

Results

Tumor-derived point-mutations in DGCR8 cause pri-miRNA processing defects

To investigate the susceptibility of DGCR8 to genetic inactivation in tumors, we searched the Catalogue of Somatic Mutations in Cancer (COSMIC) for mutations in the DGCR8 locus (25). From the 138 mutations annotated in the database, we observed that the majority (70%) arose from carcinomas, with cancers of intestinal origin occurring at the highest frequency (22%) (Fig. 1B). When classified by their effects on protein coding, 11% resulted in frameshift or nonsense mutations (Fig. 1B, right panel), clearly deleterious to protein function. We noticed that 34% of the mutations produced missense substitutions. It is not clear whether these mutations affect DGCR8 function and contribute to oncogenesis.

We elected to analyze seven missense DGCR8 mutations for their functional implications (Fig. 1A). Three mutations are located in Rhed, including K289E (change in charge), G336E (change in flexibility and charge), and F448L (change in aromaticity). Four mutations are in dsRBD1 (E518K, R527C, R527H and R570Q). E518K is a hot-spot mutation that occurs in about 3% of Wilms Tumors (7,8). Although E518K has been shown to alter miRNA expression profile, the underlying mechanism is unknown.

Comparing protein sequences across multiple homologs revealed that all sites of mutation are highly conserved, with the exception of R527 (Fig. 1C). For those mutations falling within crystal structures of DGCR8 (13,26,27), we determined their spatial proximity to residues with known functional significance (Fig. 1D). G336 co-localizes with R322 and R325 on the surface of the DSD (Fig. 1D, left panel) (13,27). We have previously shown that these basic residues are involved in binding pri-miRNA (10). Interestingly, E518, R527, and R570 on dsRBD1 are not spatially close to the previously reported RNA-binding patches on helix 2 (Fig. 1D, right panel, labeled as D1H2 and D2H2) (26).

We first tested the effects of tumor-derived DGCR8 mutations using a previously-developed live-cell pri-miRNA processing reporter assay (Fig. 2A) (23). The reporter uses a bi-directional promoter that drives the expression of two fluorescent proteins, eYFP and mCherry. We inserted a pri-miRNA hairpin into the 3'UTR of the mCherry expression cassette while eYFP serves as normalization. Cleavage of the pri-miRNA moiety by the MC results in a decrease in mCherry protein expression as the fusion mRNA is destabilized. Slopes from linear fits to the mCherry vs eYFP intensities of individual cells indicate pri-miRNA processing efficiency. Overexpression of wild-type (WT) DGCR8 causes a significant increase in processing efficiency, providing a convenient way to test mutants (Fig. 2B). To assure generality of our conclusions, we employed two pri-miRNA reporters (pri-miR-9-1 and pri-miR-30a) in two human cell lines (HeLa and U2OS Tet-on), respectively.

The reporter assays showed that four of the seven DGCR8 mutations cause severe defects in pri-miRNA processing in both HeLa and U2-OS cell lines. In HeLa, overexpression of WT DGCR8 nearly doubles the mCherry vs eYFP slope (Fig. 2B). K289E and G336E are significantly less active than overexpression of WT DGCR8, while R527C, R527H and R570Q have higher activity). Overexpression of F448L is significantly lower than the 'reporter only' control, suggesting a potential dominant negative effect. E518K is also defective in pri-miRNA processing with no statistically significant differences from the 'reporter only' control. In U2-OS cells using a pri-miR-30a reporter, we observed a similar trend. Overexpression of WT DGCR8 results in a 30% increase in mCherry vs eYFP slope. K289E, G336E, F448L and E518K are defective in pri-miRNA processing, with F448L and E518K showing no statistically significant differences in pri-miRNA processing over the 'reporter only' control. Immunoblot analysis of transfected cells showed that the DGCR8 mutants were expressed at levels similar to the wild type (Fig. 2B), ruling out the possibility that the reduced pri-miRNA processing efficiency was due to uneven protein expression levels.

E518K disrupts the folding of DGCR8

To investigate why the cancer-associated mutations are defective in pri-miRNA processing, we engineered these mutations into a bacterial expression plasmid for a truncated DGCR8 construct called NC1, which contains all domains important for biochemical pri-miRNA processing activity (Fig. 1A). Using a previously established protocol, we expressed and purified these mutants, except for E518K. This mutant was expressed as an insoluble form under our standard condition. We screened a variety of pH, salt, and additive conditions, but the protein remained insoluble. Therefore, we conclude that E518K disrupts the folding of DGCR8. This is consistent with crystal structure of DGCR8 core, which shows that E518 is involved in an extensive hydrogen-bonding network (26).

F448L abolishes DGCR8 binding to the essential heme cofactor

In contrast to purified WT NC1 and the K289E and G336E mutants, F448L did not display the yellow-brown color characteristic of the DGCR8-bound Fe(III) heme cofactor. The WT NC1 protein has intense absorption peaks arising from the ligation of Fe(III) heme by the two cysteine side chains (15). However, the electronic absorption spectrum of purified NC1 F448L does not have the signature heme bands at 366, 450 and 556 nm (Fig. 3B). Furthermore, we attempted to reconstituted potential NC1 F448L-heme complex by incubating the apo form of the protein with Fe(III) heme, but did not observe any absorption peaks corresponding to specific DGCR8-heme interaction. Therefore, we conclude the F448L mutation abolishes heme binding to DGCR8.

Next, we tested the activity of the purified NC1 mutants using pri-miRNA processing assays reconstituted with purified recombinant Drosha and ^{32}P-labeled substrates (Fig. 3C). Consistent with the results from cellular assays, we observed greatly attenuated activity for F448L, with very little pre-miR-9-1 and pre-miR-21 produced. Substantial amount of pre-miR-30a was generated by this mutant but still much less that that from the WT. This is consistent with pri-miR-30a being less sensitive to DGCR8 defects in biochemical activity assays as previously reported (10). In contrast, NC1 G336E is only slightly less active than the WT and NC1 K289E is just as active as the WT. These observations are consistent with our previous report that the reconstituted pri-miRNA processing assay is much more forgiving than the cellular assay (10), probably due to the absence of other cellular RNAs and other nuclear proteins as competitors. Overall, the biochemical assays confirm that F448L mutant has severe pri-miRNA processing defects. This is likely to be caused by the disruption of heme binding.

F448L also disrupts the interaction between DGCR8 and Drosha

To determine if the cancer-associated mutations might decrease DGCR8 activity by interfering with MC assembly, we developed a pull-down assay. This assay examines the interaction between purified recombinant N-terminally FLAG-tagged NC1 (N-flag-NC1) and His_6-Drosha$^{390-1374}$. Drosha is selective precipitated by N-flag-NC1 (compare lanes 11 and 5 in Fig. 4A). Their co-precipitation does not depend on pri-miRNA (compare lane 11 with lanes 14 and 8), indicative of a direct interaction. These observations are consistent with previous biochemical and structural analyses supporting a direct interaction between DGCR8 and Drosha (28-30). We then tested the DGCR8 mutants using the pull-down assay. We observed that whereas the N-flag-NC1 K289E and G336E mutants associate with Drosha to the same extent as the WT, F448L fails to do so (Fig. 4B). The disruption of DGCR8-Drosha interaction and interference with the DGCR8-heme interaction provide clear rationales for the inactivity of F448L mutant. Biochemical assays and crystal structure indicate that Drosha directly binds DGCR8 in the CTT region (28,30). Our results suggest that heme-binding to DGCR8 may be required for strong association with Drosha. It is possible that the Rhed provides additional binding sites for Drosha.

K289E and G363E weaken heme binding, reduce pri-miRNA affinity, and destabilize the structure

We used quantitative assays to measure how the K289E and G336E mutations may affect DGCR8's affinities for heme and pri-miRNAs, as well as the protein's structural stability. To examine changes in heme-binding affinity, we performed a kinetic heme-dissociation assay in which purified

Rhed mutants were incubated with a large molar excess of a heme scavenger protein, apomyoglobin. Heme molecules that dissociate from DGCR8 are quickly sequestered by apomyglobin, which has a K_d of 3 x 10^{-15} M for heme (31), and become unavailable to bind back to DGCR8. The heme transfer results in a decrease in the absorbance at 450 nm (A_{450}), and a concurrent increase in absorbance at 409 nm (A_{409}) indicating formation of metmyoglobin (Fe(III) heme-bound myoglobin).

We previously showed that DGCR8 binds Fe(III) heme extremely tightly, with no visible changes in A_{450} observed for human NC1 over 4-5 days (15). Here we measured changes in heme dissociation rate in the context of the Rhed. In almost 3 days, A_{450} had only decreased by very little (Fig. 5A). In contrast, both K289E and G336E Rhed mutants lost heme much faster. By fitting A_{450} over time with a one-phase exponential decay equation, we deduced the half-lives ($t_{1/2}$) and dissociation rate (k_{off}) to be 16.3 hr and 1.70 x10^{-5} s^{-1} for K289E, and 16.1 hr and 1.73 x10^{-5} s^{-1} for G336E. By comparison, the previously characterized P351A mutant has a shorter $t_{1/2}$ of 0.39 hr and higher k_{off} of 5 x 10^{-4} s^{-1} and this severe heme-binding defect renders DGCR8 inactive in HeLa cells (15,23). Thus, we conclude that both K289E and G336E weaken DGCR8's association with heme, but the effects are moderate comparing to all known inactive heme-binding-deficient mutants.

To measure mutational effects on pri-miRNA-binding affinity, we used a filter binding assay to measure the dissociation constant (K_d) for each mutant in the context of Rhed. Three pri-miRNAs were used. As we have shown in the past, wild-type NC1 binds all three pri-miRNAs with K_d values 10-16 nM (10). The affinities of NC1-K289E for pri-miR-21 and pri-miR-30a are similar to the WT, but the affinity for pri-miR-9-1 is decreased, with K_d of 36 nM (Table 1). NC1-G336E binds pri-miR-30a with a WT-level affinity, but had lower affinity for both pri-miR-9-1 and pri-miR-21, with average K_d values of 52 and 50 nM, respectively. Importantly, the K289E and G336E mutations render the Rhed unable to bind the pri-miRNA substrate without the help of the dsRBDs (Table 1).

We also investigated whether the K289E and G336E mutations reduce DGCR8 structural stability. We performed thermo-melting assays to compare the mutants with WT in the context of NC1. In these assays, A_{450} is monitored to assess association with heme and thereby serves as a sensitive indicator of proper protein folding. We observed that the thermostability of both K289E and G336E mutants is decreased relative to WT (Fig. 6A). Although we were not able to determine their exact melting temperatures due to substantial background scattering from aggregated proteins, visual inspection of the melting curves indicates that K289E and G336E destabilize the NC1 proteins by 2-3˚C. The effects of G336E are more pronounced, as below 50˚C NC1 G336E loses A_{450} more quickly than the WT and K289E.

K289E and G336E are located within the independently folded dimerization sub-domain (DSD) of the Rhed. The DSD expresses very well in *E. coli* and has a crystal structure available (13,27). We analyzed DSD mutants by size-exclusion chromatography (Fig. 6B). Whereas WT DSD elutes at 59.2 ml, we observed a decreased elution volume (~56.3 ml) for both K289E and G336E that was still greater than the void volume of the column. We believe that the apparent increase in elution volume is caused by the dimeric DSD mutants partially unfolding, effectively increasing the hydrodynamic radius of the molecule.

Taken together, our biochemical characterization of the K289E and G336E mutants revealed moderate reductions in heme binding, pri-miRNA binding and structural stability. Any one or combination of these changes could result in the pri-miRNA processing deficiency observed in cells. The biochemical effects of G336E are more pronounced, consistent with this mutant being more defective than K289E.

Discussion

Identifying driver mutations that favor uncontrolled growth and prevent apoptosis is critical for understanding cancer development and for formulating novel treatment strategies. A great example concerns B-Raf, a serine/threonine kinase involved in cell-growth signaling pathways. A point mutation, V600E, constitutively activates B-Raf and occurs in ~50% of melanomas (43)(44,45). The drug Trametinib has been developed to treat patients carrying this mutation (32). Because it has been demonstrated that disruption of the miRNA maturation pathway leads to a pro-growth phenotype and the development of tumors (3), we analyzed tumor-associated DGCR8 mutations. We found that, out of 138 mutations in the COSMIC database, 11 cause frameshifts that most likely result in loss of function. Four out of the seven point-mutations we selected for experimental testing, are severely defective in pri-miRNA processing. Thus, a substantial fraction of the DGCR8 tumor-associated mutations render the protein defective and are likely to directly contribute to pathogenesis.

These tumor-associated mutations are presumably somatic, and arise heterozygously during tumor development. They may cause pri-miRNA processing deficiency by the following mechanisms. First, losing 50% of functional DGCR8 proteins directly results in a relatively mild reduction in pri-miRNA processing efficiency, which has been shown to reduce the maturation of a small fraction of miRNAs, as observed in DiGeorge syndrome patients and *Dgcr8*$^{+/-}$ mouse model (5,33,34). However, such reduction in pri-miRNA processing has not been linked to tumor development. A second mechanism is that one mutated DGCR8 can form a heterodimer with a WT polypeptide and in turn cause dominant negative effects. Consistent with such an expectation, expression of F448L in HeLa cells results in a pri-miRNA processing efficiency lower than the vector control (Fig. 2B). Third, these DGCR8 mutations could result in a very severe defect, even a complete loss in pri-miRNA processing when combined with another event that reduces or abolishes the expression of the second copy of *DGCR8* gene. Such changes have been observed in the Wilms tumor patients carrying the E518K mutation (22,23). Through these mechanisms, pre-cancerous cells have many paths to dysregulate miRNA maturation in favor of proliferation and tumorigenesis.

There has been a great interest to understand the mechanistic underpinnings responsible for the global repression of miRNA expression in cancer. The reported molecular explanations included high tendency of miRNA genes to be located in fragile genomic loci (35), loss of expression of proteins involved in miRNA biogenesis (5,6), and abnormal signaling that regulates miRNA processing (36). We and others showed that DGCR8 mutations found in tumors can cause functional and structural defects that lead to miRNA processing deficits (7,8). Therefore, tumors have multiple ways of achieving global miRNA repression. We recently identified a heme analog compound, cobalt (III) protoporphyrin IX, that can bind and hyper-activate the DGCR8 protein, and compensate heterozygous deletion of the *DGCR8* gene (37). Agonist compounds like this offer a potential way of correcting the pri-miRNA processing defects in tumors carrying *DGCR8* mutations. We imagine that such therapeutic strategies will become a common practice in the era of personalized medicine.

The DGCR8 mutations characterized here lend insight into how DGCR8 functions in pri-miRNA processing. Three of the defective mutants, K298E, G336E, and F448L, are located in the Rhed, which our group previously demonstrated is required for processing activity in cells (23). We showed that this domain participates in pri-miRNA recognition by binding pri-miRNA junctions (10). We also demonstrated that this domain contains a conserved caspase cleavage site and that caspase-mediated cleavage of Rhed inactivates DGCR8 (19). Still others have provided evidence that this domain is critical for pri-miRNA processing accuracy (11). Our work here highlights that the Rhed of DGCR8 is particularly sensitive to perturbation in tumors. F448L, K289E, and G336E weaken the heme-binding affinity to various degrees, reinforcing the physiological importance of DGCR8-heme interaction, though its exact function remains to be defined.

Our data show that both K289E and G336E differentially reduce the affinities for different pri-miRNA hairpins. This observation suggests that these mutations are likely to cause differential reduction of miRNA expression, as opposed to a uniform repression. Humans have close to 2,000 annotated pri-miRNA hairpins (38). The predicted secondary structures of human pri-miRNA can be quite different from each other, providing an explanation of why the DGCR8 mutations affect some pri-miRNAs more than others.

The heme-binding deficient F448L failed to associate with Drosha. This observation suggests that the Rhed and its association with heme are important for MC formation. The C-terminal tail of DGCR8 has been shown to mediate a low-affinity interaction with Drosha (11,30). However, the CTT-Drosha interaction appears to be insufficient to maintain the MC in our pull-down assays. Therefore, our study reveals a new function of the Rhed in mediating interaction between Drosha and DGCR8, in addition to binding pri-miRNAs.

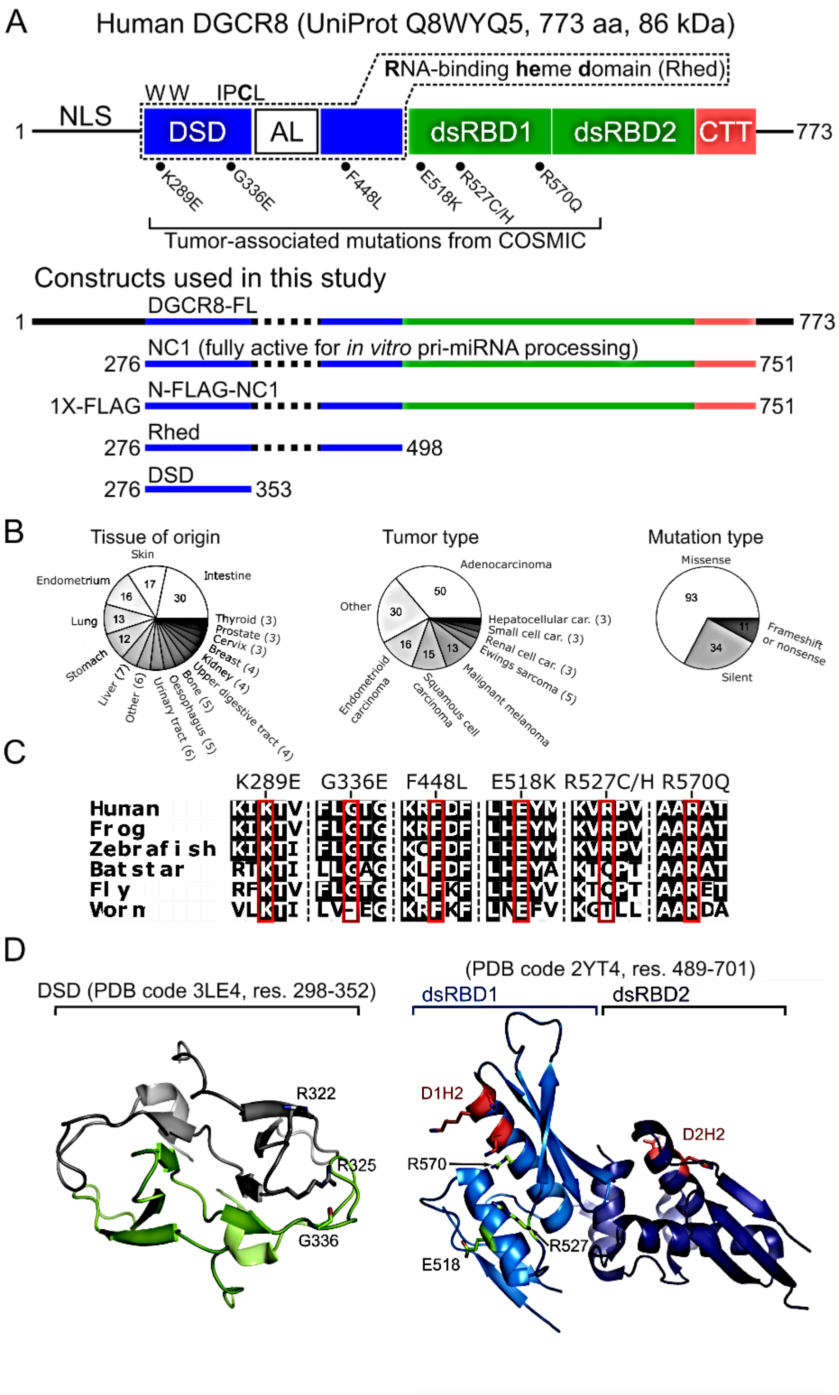

A
Human DGCR8 (UniProt Q8WYQ5, 773 aa, 86 kDa)
RNA-binding heme domain (Rhed)
WW IPCL
1 NLS DSD AL dsRBD1 dsRBD2 CTT 773
K289E G336E F448L E518K R527C/H R570Q
Tumor-associated mutations from COSMIC
Constructs used in this study
1 DGCR8-FL 773
276 NC1 (fully active for in vitro pri-miRNA processing) 751
1X-FLAG N-FLAG-NC1 751
276 Rhed 498
276 DSD 353
B
Tissue of origin
Skin 17
Intestine 30
Endometrium 16
Lung 13
12
Stomach
Thyroid (3)
Prostate (3)
Cervix (3)
Breast (4)
Kidney (4)
Upper digestive tract (4)
Bone (5)
Oesophagus (5)
Urinary tract (6)
Other (6)
Liver (7)
Tumor type
Adenocarcinoma 50
Other 30
Hepatocellular car. (3)
Small cell car. (3)
Renal cell car. (3)
Ewings sarcoma (5)
Malignant melanoma 13
Squamous cell carcinoma 15
Endometrioid carcinoma 16
Mutation type
Missense 93
Frameshift or nonsense 11
Silent 34
C
K289E G336E F448L E518K R527C/H R570Q
Hunan KIKTV FLGTG KRFDF LHEYM KVRPV AARAT
Frog KIKTV FLGTG KRFDF LHEYM KVRPV AARAT
Zebrafish KIKTI FLGTG KCFDF LHEYM KVRPV AARAT
Batstar RTKTI LLGAG KLFDF LHEYA KTQPT AARAT
Fly RFKTV FLGTG KLFKF LHEYV KTQPT AARET
Vorm VLKTI LV-EG KRFKF LNEFV KGTLL AARDA
D
DSD (PDB code 3LE4, res. 298-352)
R322
R325
G336
(PDB code 2YT4, res. 489-701)
dsRBD1 dsRBD2
D1H2
D2H2
R570
R527
E518

Figure 1. **Analysis of tumor-associated mutations in DGCR8**. (A) Domain structure of DGCR8 and expression constructs. The RNA-binding heme domain (Rhed) is shown as an N-terminal dimerization subdomain (DSD), a central acidic loop (AL), and a C-terminal subdomain. A WW motif and the IPCL heme-binding site are labeled. (B) Classification of 138 somatic DGCR8 mutations observed in the COSMIC database. (C) Multiple sequence alignment of DGCR8 homologs highlighting mutations considered in this study. (D) Crystal structures of the DGCR8 DSD (left panel) and dsRBDs (right panel). Basic residues previously shown to interact with RNA are highlighted in red (D1H2 and D2H2) (26).

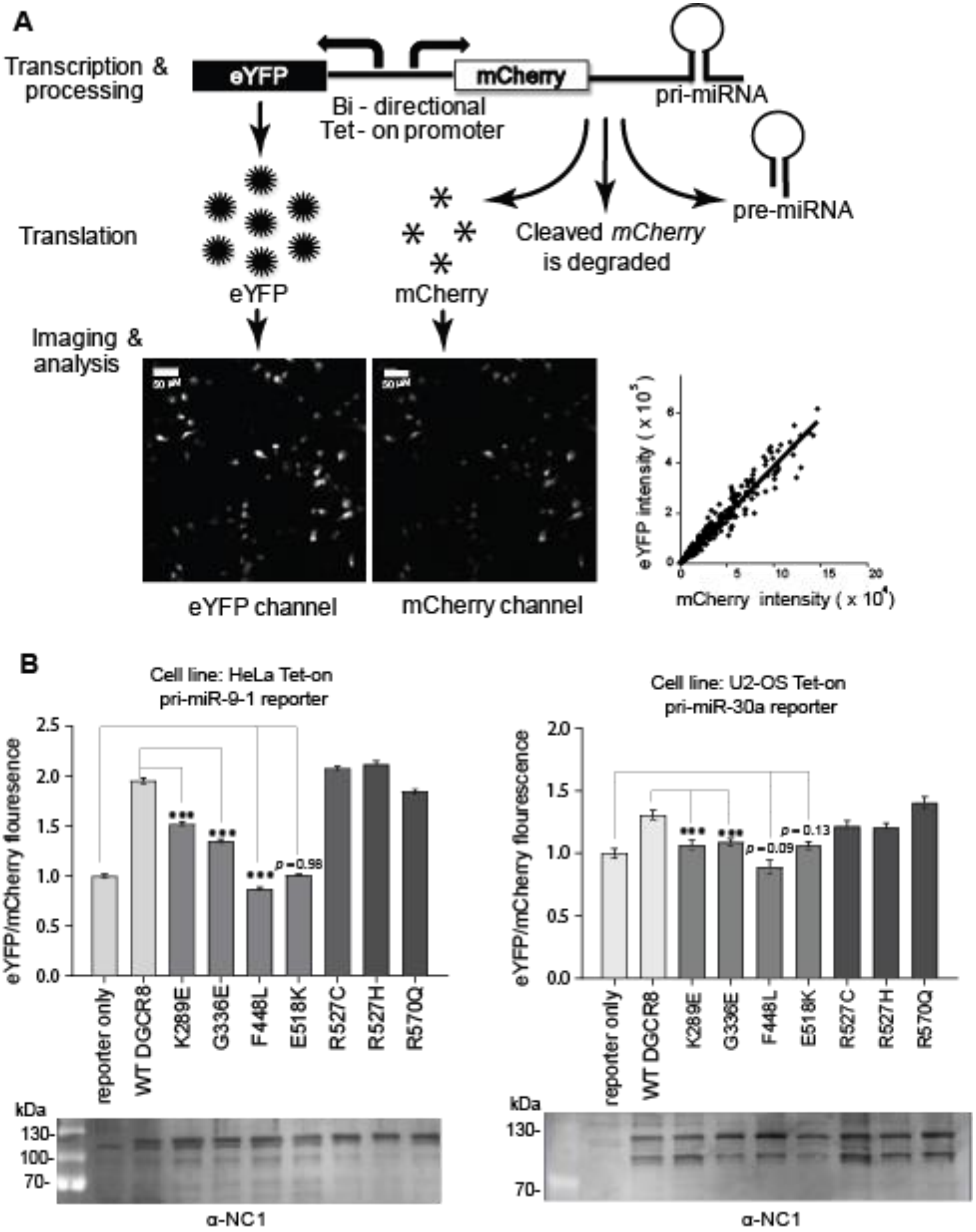

A
Transcription & processing
eYFP
Bi - directional
Tet - on promoter
mCherry
pri-miRNA
pre-miRNA
Cleaved *mCherry* is degraded
Translation
eYFP
mCherry
Imaging & analysis
50 μM
50 μM
eYFP channel
mCherry channel
eYFP intensity (x 10^5)
0
2
4
6
mCherry intensity (x 10^4)
0
5
10
15
20
B
Cell line: HeLa Tet-on
pri-miR-9-1 reporter
eYFP/mCherry flouresence
0.0
0.5
1.0
1.5
2.0
2.5
***
***
***
p = 0.98
reporter only
WT DGCR8
K289E
G336E
F448L
E518K
R527C
R527H
R570Q
kDa
130-
100-
70-
α-NC1
Cell line: U2-OS Tet-on
pri-miR-30a reporter
eYFP/mCherry flourescence
0.0
0.5
1.0
1.5
2.0
***
***
p = 0.09
p = 0.13
reporter only
WT DGCR8
K289E
G336E
F448L
E518K
R527C
R527H
R570Q
kDa
130-
70-
α-NC1

Figure 2. Pri-miRNA processing activity of tumor-derived mutants. (A) Live-cell processing assay reporter and experimental schematic. MC processing activity is measured as a decrease in mCherry fluorescence intensity with eYFP serving as a normalization control. Scale bar on cell images is 50 nm. (B) Pri-miRNA processing activity of DGCR8 mutants in both HeLa and U2-OS cells with immunoblots indicating expression of each mutant is near or above DGCR8 overexpression.

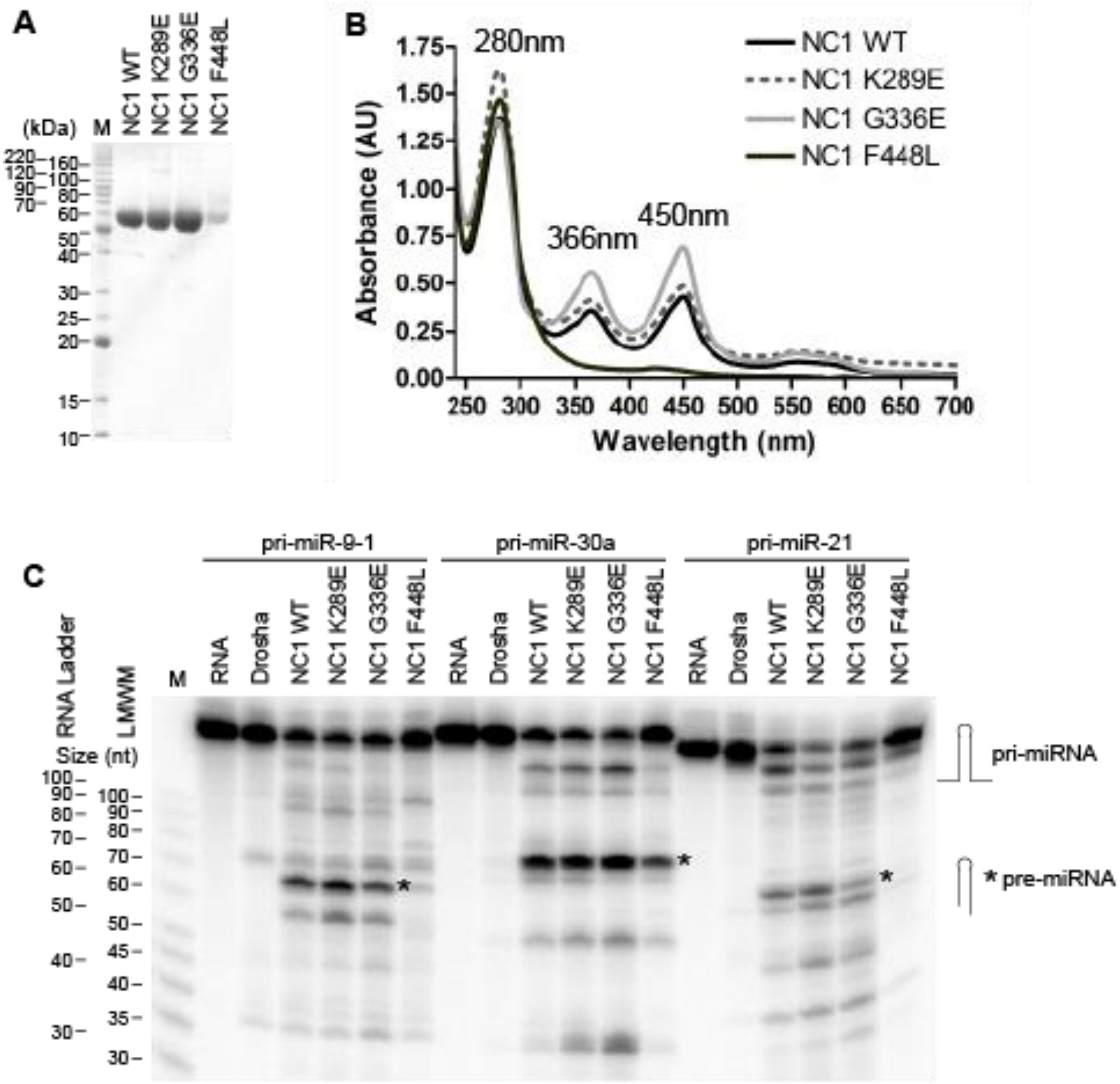


Figure 3.

Biochemical characterization of DGCR8 Rhed mutants. (**A**) Coomassie -stained SDS gel of purified NC1 proteins. (**B**) Electronic absorption spectra of NC1 WT and mutant proteins. (**C**) In vitro Microprocessor assay comparing the three Rhed mutants across three pri-miRNA substrates. We previously determined the relative electrophoretic migration of the Low Molecular Weight Marker (LMWM, Affymetrix) and the

RNA Decade ladder (ThermoFisher). Pri-miRNA substrates are 150 nt and pre-miRNA products are between 50-60 nt.

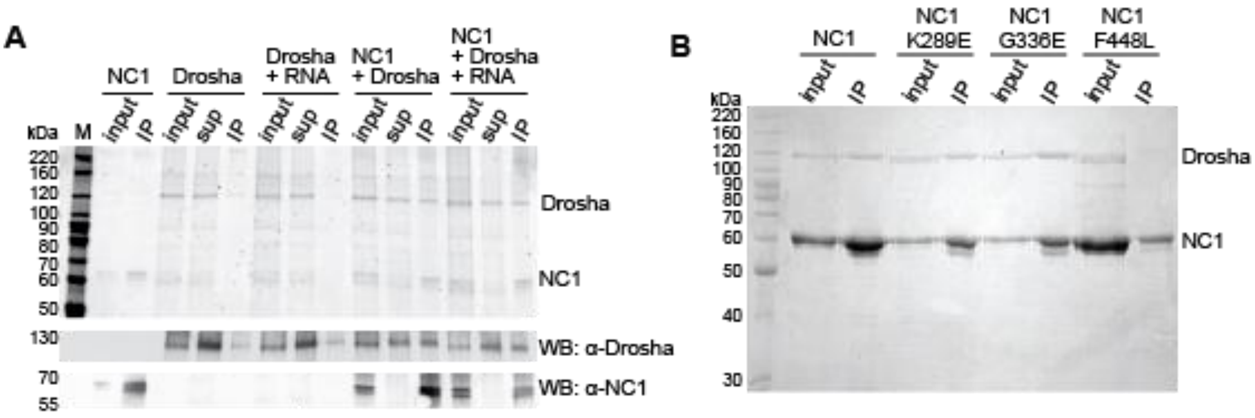


Figure 4.

Drosha pull-down assay with N-flag-NC1. (**A**) Validation of the pull down experiment showing that Drosha selectively precipitates in the presence of N-FLAG-NC1. Top panel shows a Sypro Red-stained SDS gel and lower panels are immunoblots for Drosha and NC1. (**B**) Coomassie-stained SDS gel of the pull-down assay for mutations in the Rhed.

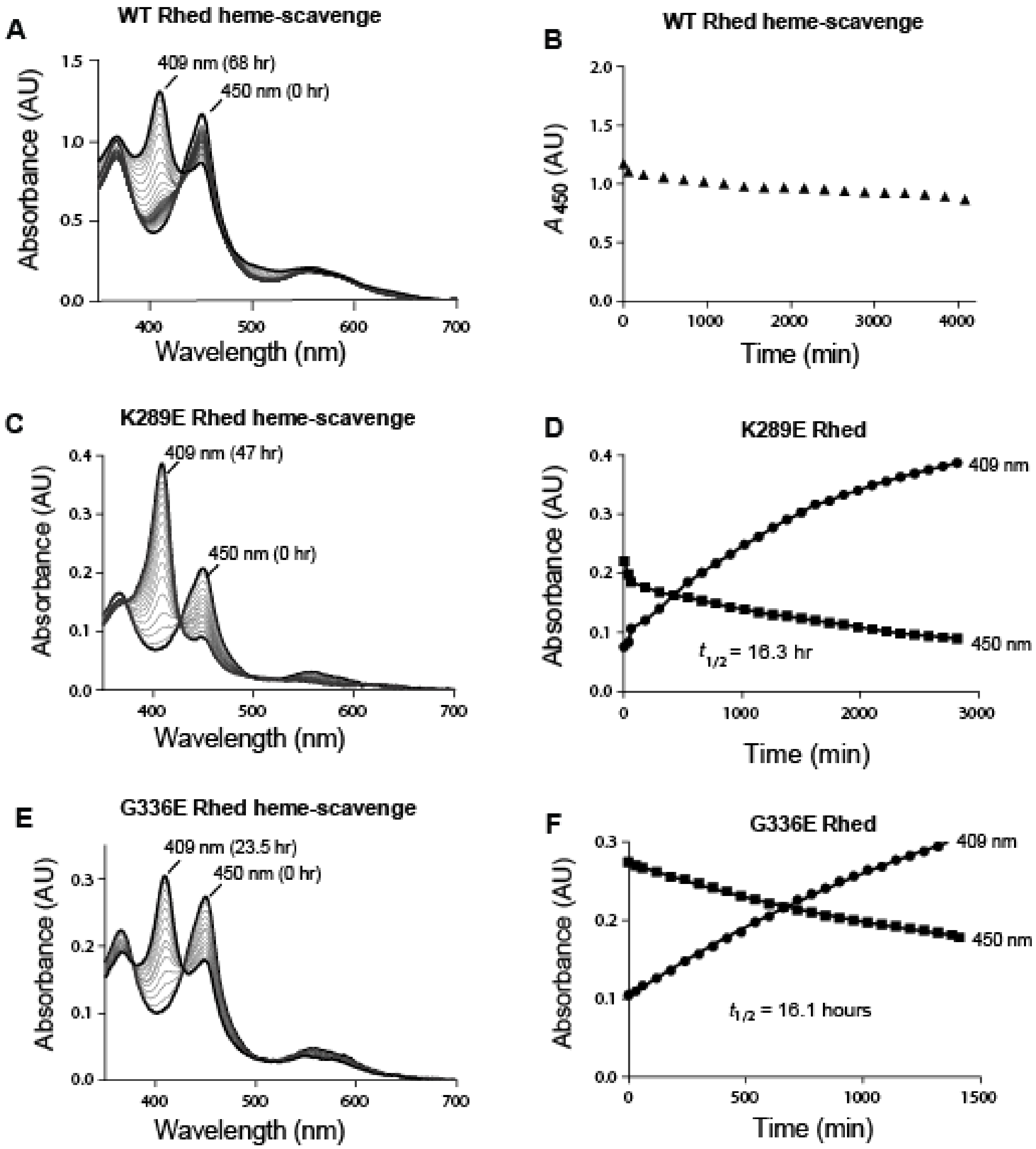


Figure 5.

Tumor-derived point-mutations in the Rhed weaken heme binding. (**A**) Measurement of the Rheds affinity for Fe(III) heme by heme-scavenging with a molar excess of apomyoglobin. Decrease in the 450 nm söret peak is monitored over time by spectrophotometry. The displayed time points are every hour till 28 hours, then every 4 hours to 68 hours. (**B**) Plot of the Rheds 450 nm absorbance over time shows virtually no loss of heme binding over ~3 days. (**C-D**) Measurment of K289E Rhed and G336E Rhed affinity for Fe (III) heme by heme scavenging with molar excess of apomyoglobin. The displayed time points for K289E Rhed are at 1 hour, then every 3 hours to 47 hr. The displayed time points for the G336E Rhed are at 1 hour and every 2 hr till 23.5 hr. (**E-F**) Plots of 409 nm absorbance from ferric myoglobin and concomitant decrease in 450 nm absorbance from DGCR8 Söret peak Data is fit with one-phase decay to calculated the half-life of heme association with each mutant.

Table 1. Dissociation constants of NC1 K289E and G336E for pri-miRNAs.

Protein	***K*d in nM**		
	pri-miR-9-1	pri-miR-21	pri-miR-30a
WT NC1	10 ± 1	12 ± 4	16 ± 1
K289E NC1	36 ± 6	11 ± 5	11 ± 1
G336E NC1	52 ± 14	49 ± 14	11 ± 2

NOTE: The K_d values are means ± SD from three repeats. The data of WT NC1, highlighted in grey, are from a previous report (10).

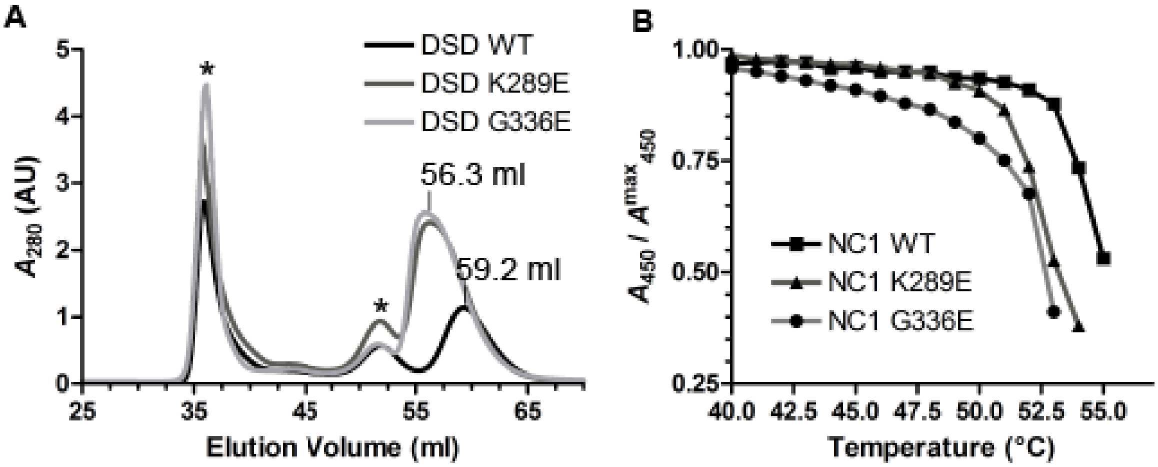


Figure 6.

K289E and G336E decrease protein structural stability, compared to WT DGCR8. (**A**) Size-exclusion chromatography analysis of DSD mutant proteins. Aggregated and RNA-bound protein peaks are marked with an asterisk. (**B**) Melting curves for NC1 mutant proteins analyzed at the 450 nm absorption peak of the heme-protein complex.

References

1. Zhang L, Huang J, Yang N, Greshock J, Megraw MS, Giannakakis A*, et al.* microRNAs exhibit high frequency genomic alterations in human cancer. Proc Natl Acad Sci U S A 2006;103:9136-41.

2. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D*, et al.* MicroRNA expression profiles classify human cancers. Nature 2005;435:834-8.

3. Kumar MS, Lu J, Mercer KL, Golub TR, Jacks T. Impaired microRNA processing enhances cellular transformation and tumorigenesis. Nat Genet 2007;39:673-7.

4. Kumar MS, Pester RE, Chen CY, Lane K, Chin C, Lu J*, et al.* Dicer1 functions as a haploinsufficient tumor suppressor. Genes Dev 2009;23:2700-4.

5. Merritt WM, Lin YG, Han LY, Kamat AA, Spannuth WA, Schmandt R*, et al.* Dicer, Drosha, and outcomes in patients with ovarian cancer. N Engl J Med 2008;359:2641-50.

6. Bandara V, Michael MZ, Gleadle JM. Hypoxia represses microRNA biogenesis proteins in breast cancer cells. BMC Cancer 2014;14:533.

7. Walz AL, Ooms A, Gadd S, Gerhard DS, Smith MA, Guidry Auvil JM*, et al.* Recurrent DGCR8, DROSHA, and SIX homeodomain mutations in favorable histology Wilms tumors. Cancer Cell 2015;27:286-97.

8. Wegert J, Ishaque N, Vardapour R, Georg C, Gu Z, Bieg M*, et al.* Mutations in the SIX1/2 pathway and the DROSHA/DGCR8 miRNA microprocessor complex underlie high-risk blastemal type Wilms tumors. Cancer Cell 2015;27:298-311.

9. Belair CD, Paikari A, Moltzahn F, Shenoy A, Yau C, Dall'Era M*, et al.* DGCR8 is essential for tumor progression following PTEN loss in the prostate. EMBO reports 2015;16:1219-32.

10. Quick-Cleveland J, Jacob JP, Weitz SH, Shoffner G, Senturia R, Guo F. The DGCR8 RNA-binding heme domain recognizes primary microRNAs by clamping the Hairpin. Cell Rep 2014;7:1994-2005.

11. Nguyen TA, Jo MH, Choi YG, Park J, Kwon SC, Hohng S*, et al.* Functional Anatomy of the Human Microprocessor. Cell 2015;161:1374-87.

12. Faller M, Matsunaga M, Yin S, Loo JA, Guo F. Heme is involved in microRNA processing. Nat Struct Mol Biol 2007;14:23-9.

13. Senturia R, Faller M, Yin S, Loo JA, Cascio D, Sawaya MR*, et al.* Structure of the dimerization domain of DiGeorge Critical Region 8. Protein Sci 2010;19:1354-65.

14. Barr I, Smith AT, Chen Y, Senturia R, Burstyn JN, Guo F. Ferric, not ferrous, heme activates RNA-binding protein DGCR8 for primary microRNA processing. Proc Natl Acad Sci U S A 2012;109:1919-24.

15. Barr I, Smith AT, Senturia R, Chen Y, Scheidemantle BD, Burstyn JN*, et al.* DiGeorge Critical Region 8 (DGCR8) is a double-cysteine-ligated heme protein. J Biol Chem 2011;286:16716-25.

16. Herbert KM, Pimienta G, DeGregorio SJ, Alexandrov A, Steitz JA. Phosphorylation of DGCR8 increases its intracellular stability and induces a progrowth miRNA profile. Cell Rep 2013;5:1070-81.

17. Zhu C, Chen C, Huang J, Zhang H, Zhao X, Deng R*, et al.* SUMOylation at K707 of DGCR8 controls direct function of primary microRNA. Nucleic acids research 2015;43:7945-60.

18. Wada T, Kikuchi J, Furukawa Y. Histone deacetylase 1 enhances microRNA processing via deacetylation of DGCR8. EMBO Rep 2012;13:142-9.

19. Gong M, Chen Y, Senturia R, Ulgherait M, Faller M, Guo F. Caspases cleave and inhibit the microRNA processing protein DiGeorge Critical Region 8. Protein Sci 2012;21:797-808.

20. Han J, Pedersen JS, Kwon SC, Belair CD, Kim YK, Yeom KH*, et al.* Posttranscriptional crossregulation between Drosha and DGCR8. Cell 2009;136:75-84.

21. Triboulet R, Chang HM, Lapierre RJ, Gregory RI. Post-transcriptional control of DGCR8 expression by the Microprocessor. RNA 2009;15:1005-11.

22. Sellier C, Freyermuth F, Tabet R, Tran T, He F, Ruffenach F*, et al.* Sequestration of DROSHA and DGCR8 by expanded CGG RNA repeats alters microRNA processing in fragile X-associated tremor/ataxia syndrome. Cell Rep 2013;3:869-80.

23. Weitz SH, Gong M, Barr I, Weiss S, Guo F. Processing of microRNA primary transcripts requires heme in mammalian cells. Proc Natl Acad Sci U S A 2014;111:1861-6.

24. Barr I, Guo F. Primary microRNA processing assay reconstituted using recombinant Drosha and DGCR8. Methods Mol Biol 2014;1095:73-86.

25. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J*, et al.* COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res 2017;45:D777-D83.

26. Sohn SY, Bae WJ, Kim JJ, Yeom KH, Kim VN, Cho Y. Crystal structure of human DGCR8 core. Nat Struct Mol Biol 2007;14:847-53.

27. Senturia R, Laganowsky A, Barr I, Scheidemantle BD, Guo F. Dimerization and heme binding are conserved in amphibian and starfish homologues of the microRNA processing protein DGCR8. PloS one 2012;7:e39688.

28. Kwon SC, Nguyen TA, Choi YG, Jo MH, Hohng S, Kim VN*, et al.* Structure of Human DROSHA. Cell 2016;164:81-90.

29. Herbert KM, Sarkar SK, Mills M, Delgado De la Herran HC, Neuman KC, Steitz JA. A heterotrimer model of the complete Microprocessor complex revealed by single-molecule subunit counting. RNA 2016;22:175-83.

30. Han J, Lee Y, Yeom KH, Kim YK, Jin H, Kim VN. The Drosha-DGCR8 complex in primary microRNA processing. Genes Dev 2004;18:3016-27.

31. Hargrove MS, Krzywda S, Wilkinson AJ, Dou Y, Ikeda-Saito M, Olson JS. Stability of myoglobin: a model for the folding of heme proteins. Biochemistry 1994;33:11767-75.

32. Holderfield M, Deuker MM, McCormick F, McMahon M. Targeting RAF kinases for cancer therapy: BRAF-mutated melanoma and beyond. Nat Rev Cancer 2014;14:455-67.

33. de la Morena MT, Eitson JL, Dozmorov IM, Belkaya S, Hoover AR, Anguiano E*, et al.* Signature MicroRNA expression patterns identified in humans with 22q11.2 deletion/DiGeorge syndrome. Clin Immunol 2013;147:11-22.

34. Stark KL, Xu B, Bagchi A, Lai WS, Liu H, Hsu R*, et al.* Altered brain microRNA biogenesis contributes to phenotypic deficits in a 22q11-deletion mouse model. Nat Genet 2008;40:751-60.

35. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E*, et al.* Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A 2002;99:15524-9.

36. Mori M, Triboulet R, Mohseni M, Schlegelmilch K, Shrestha K, Camargo FD*, et al.* Hippo signaling regulates microprocessor and links cell-density-dependent miRNA biogenesis to cancer. Cell 2014;156:893-906.

37. Barr I, Weitz SH, Atkin T, Hsu P, Karayiorgou M, Gogos JA*, et al.* Cobalt(III) Protoporphyrin Activates the DGCR8 Protein and Can Compensate microRNA Processing Deficiency. Chemistry & biology 2015;22:793-802.

38. Chang TC, Pertea M, Lee S, Salzberg SL, Mendell JT. Genome-wide annotation of microRNA primary transcript structures reveals novel regulatory mechanisms. Genome Res 2015;25:1401-9.

Chapter 5: In crystal selection to establish new RNA crystal contacts

Grant M. Shoffner[1], Ruixuan Wang[1], Elaine Podell[2], Thomas R. Cech[2], and Feng Guo[1,*]

[1]Department of Biological Chemistry, David Geffen School of Medicine, University of California, Los Angeles, CA 90095.

[2]Howard Hughes Medical Institute, Department of Chemistry and Biochemistry, University of Colorado BioFrontiers Institute, Boulder, Colorado 80309-0596.

[*] Correspondence: fguo@mbi.ucla.edu

SUMMARY

Crystallography is a major technique for determining large RNA structures. Obtaining diffraction-quality crystals has been the bottle neck. Although several RNA crystallization methods have been developed, the field strongly needs new approaches. Here we invented an *in crystal* selection strategy for identifying mutations of a target RNA that enhance its crystallizability. The strategy includes constructing an RNA pool containing random mutations, obtaining crystals, and amplifying the sequences enriched by crystallization. We demonstrated a proof-of-principle application to the P4-P6 domain from the *Tetrahymena* ribozyme. We further determined the structures of four selected mutants. All four establish new crystal lattice contacts while maintaining the native structure. Three mutants achieve this by relocating bulges and one by making a helix more flexible. *In crystal* selection provides opportunities to improve crystals of RNAs or RNA-ligand complexes. Our results also suggest that mutants may be rationally designed for crystallization by "walking" a budge along the RNA chain.

Keywords: X-ray crystallography, structural biology, in vitro selection, group I intron, crystal lattice contacts.

INTRODUCTION

RNA molecules often fold into defined three-dimensional structures to carry out their functions. While crystallography has had considerable success in determining atomic structures of RNAs, it is challenging to obtain crystals for large RNAs. Even when an RNA does crystallize, the crystals often do not diffract or diffract poorly. This challenge may be attributed to several features of RNAs. They have polyanionic backbones. Therefore, RNAs have to overcome charge-charge repulsion, often via counter ions, in order to fold properly and pack orderly into crystals. Indeed, structural stability has been shown to be an important factor in obtaining diffraction-quality crystals (Juneau and Cech, 1999; Juneau et al., 2001). Furthermore, unlike proteins, RNAs often fold with most of their side chains (bases) facing inwards, in double-stranded helices for example, making it harder to establish unique lattice contacts.

Great efforts in the field of RNA structural biology have been devoted to develop methods for facilitating crystallization (Ke and Doudna, 2004; Zhang and Ferre-D'Amare, 2014). Homologues are routinely explored in crystallization trials. Crystallization screens tailored for RNAs and RNA-protein complexes have been designed (Doudna et al., 1993; Scott et al., 1995). *In vitro* selection has been employed to identify mutations that improve structural stability and aid crystallization (Guo and Cech, 2002; Guo et al., 2004; Juneau and Cech, 1999; Juneau et al., 2001). Knowledge on RNA tertiary interactions, such as the GAAA tetraloop/tetraloop receptor interactions, has been put to use in engineering potential lattice contacts (Ferre-D'Amare et al., 1998; Golden et al., 1997; Reiter et al., 2010). Protein-binding sites may be engineered into RNAs so that proteins such as U1A and antibodies are included in crystallization and may bridge lattice interactions (Ferre-D'Amare and Doudna, 2000; Koldobskaya et al., 2011; Shechner et al., 2009; Ye et al., 2008). Many RNAs, such as the T-box riboswitch, naturally contain the Kink-turn structural motif and thereby can be cocrystallized with Kink-turn binding proteins (Zhang and Ferre-D'Amare, 2013). These methods have allowed successful structure determination for a substantial number of examples. Nevertheless, crystallization remains the major bottleneck. There is a great need to expand the RNA crystallography toolbox.

Unlike proteins, RNAs can carry functions while fully encoding their “genetic” information. This property led to the RNA World hypothesis (Gesteland et al., 2005) and the invention of *in vitro* selection methods (Tuerk and Gold, 1990; Wilson and Szostak, 1999). *In vitro* selection is also called *S*ystematic *E*volution of *L*igands by *EX*ponential amplification (SELEX). The method starts with a diverse pool of sequences that is subjected to iterative rounds of selection based on biophysical or functional properties, RT-PCR amplification of the “winner” sequences, and transcription to produce the next-generation library. Random mutagenesis may be included between rounds of selection to make “*in vitro* evolution.” *In vitro* selection is a powerful tool for both biological discoveries and biotechnological applications, which range from identifying binding sites of RNA-binding proteins to developing nucleic acid aptamers with high affinities for specific target molecules (Nimjee et al., 2017).

Here we invented an *in crystal* selection strategy to identify RNA variants with improved abilities to incorporate into a crystal lattice (Figure 1A). The strategy is *in vitro* selection using crystallization as the selection criterion. The method includes introduction of random mutations to a target RNA molecule and crystallization of the library mixture. The latter step is somewhat counterintuitive as it is generally believed in crystallography that crystallization requires the subject macromolecules to be chemically and conformationally homogeneous (Doudna, 1997; Ferre-d'Amare and Doudna, 1997). As a proof-of-concept, we applied this schema to the wild-type (WT) 160-nt P4-P6 domain from the *Tetrahymena* self-splicing group I intron, which contains extensive secondary and tertiary structures but is non-trivial to crystallize (Cate et al., 1996). Crystal structures of selected mutants show how they engage in new lattice interactions.

RESULTS

Design and implementation of the *in crystal* selection strategy

We worked out an *in crystal* selection procedure enabling direct selection for crystallizability (Figure 1A). We generated an initial mutant library of P4-P6 through 10 cycles of mutagenic PCR, yielding a DNA sequence pool with most constructs expected to contain 0-2 mutations. We randomly picked 12 clones for sequencing and observed 7 mutant and 5 wild-type P4-P6 sequences. Following *in vitro* transcription and denaturing gel purification, the RNA library was annealed and subjected to crystallization by vapor diffusion. As the WT P4-P6 was known to crystallize in essentially the same crystal form under conditions containing either 2-Methyl-2,4-pentanediol (MPD) or polyethylene glycol (PEG) 1000, we compared selection conditions with either as the precipitant. Large (around 200 μm) crystals appeared in both MPD and PEG1000 conditions within 4 days. Harvesting crystals from the drops yielded the initial pool of selection "winners." We subjected the winner RNA libraries to a further round of selection, but only obtained large numbers of smaller crystals that were difficult to manipulate. Therefore, we opted to investigate the winner pools after a single round of selection.

We randomly picked a total of 24 clones from the MPD and PEG1000 pools for sequencing. The results revealed 11 mutants and 13 WT sequences (Figure 1B). The mutations distribute throughout the P4-P6 sequence and secondary structures (Figure 1C), except in the invariable RT-PCR primer annealing regions at the 5'- and 3'-ends. Even with the P4-P6 three-dimensional structures taken into account, it was not obvious how these mutations affect the ability of the RNA to join a crystal lattice.

Structure determination of P4-P6 mutants

To determine how these mutations might alter the RNA structure and crystal-lattice contacts, we crystallized individual mutants and collected diffraction data for four of them (Figure 1C and Table 1).

For each construct we re-screened crystallization in both MPD and PEG1000 conditions, with or without spermine or cobalt hexamine as additives. M1-M4 and the WT P4-P6 control crystallized within 1-5 days. Under the conditions in our laboratory, the time required for crystallizing WT P4-P6 is much shorter than the 1-2 months previously reported (Cate et al., 1996). As the WT structure was solved using crystals treated with isopropanol, we examined the effects of soaking crystals in mother liquor supplemented with 5%, 10%, or 15% (v/v) isopropanol.

The space group and unit cell dimensions of the mutant P4-P6 crystals (Table 1) closely matched those of the WT (PDB ID 1GID; $P2_12_12_1$; *a* = 74.8 Å, *b* = 128.7 Å, *c* = 145.9 Å). The crystals diffracted to 2.80-3.95 Å resolution, either the same as the WT (M4) or moderately worse (M1-M3). We solved the structures by refining the WT coordinates against the new data (see the Methods section for details). The overall structures of the M1-M4 RNAs aligned well with that of the WT, all with backbone C1' r.m.s.d. under 1 Å (both molecules A and B in the asymmetric unit are included in the calculation, Figure S1). There are no large changes in intermolecular positioning and orientation, as indicated by distance difference matrix plots (Figures S1 and 5B).

Swapping base-pairs in M1 and M2 produces new lattice interactions

The M1 (U131A) mutation causes a local structural rearrangement that establishes new lattice contacts. We first solved the M1 structure using crystals grown from a solution containing 22% (v/v) PEG1000 and soaked in mother liquor supplemented with 10% (v/v) isopropanol. The electron density map indicated a substantial local conformational rearrangement relative to the WT (Figure 2A and 2B). In the WT crystal, U131 pairs with A192 in the middle of P5a helix and the adjacent U130 forms a single-nucleotide bulge (Figure 2C). In the lattice, molecules A and B (as defined by the deposited coordinates) are related by a pseudo-twofold symmetry. U130 and U131 are about 7 and 10 Å from the closest atoms in the other molecule, respectively (Figure 2C), too far to mediate direct lattice contacts. Mutation of U131 to adenine disrupts the original base pair with A192, allowing U130 to rotate into the helix to restore the pairing, while pushing A131 out toward the neighboring molecule (Figure 2D). Although the electron density map does not unambiguously define the position of the whole A131 base, we estimate that the N1 imine of A131 base in molecule A is within hydrogen-bonding distance to the 2' hydroxyl of U244 in molecule B (Figure 2D), and that the N6 amino and N7 nitrogen of A131 in molecule B contact the 2' hydroxyl of U245 in molecule A. Therefore, we conclude that U131A establishes new lattice contacts by shifting the bulge by one residue to a closer location and better orientation toward the neighboring molecule and by replacing the pyrimidine with a larger purine base.

To investigate whether the conformational and lattice rearrangements between WT and M1 structures might result from differences in crystal preparation, we solved the structure of M1 using crystals grown with MPD as the precipitant and were frozen without soaking. We observed base-pairing between U130 and A192 as well as relocation of A131 to contact the neighboring molecule (Figure S2), in agreement with the M1 structure under PEG1000 condition and soaked with isopropanol. This result supports the notion that the new lattice interactions are stable under different conditions and thereby are caused by the mutation identified using *in crystal* selection.

We observed a similar theme in the structure of M2, which contains A125U and G126U. The electron density in this region of the M2 structure showed repositioning of residues124 and 125 (compare Figures 3A and 3B). In the WT structure, G126 forms a non-canonical base pair with A196, stacking on top of the P5a stem (Figure 3C). C124 stacks on top of G126 and is further stabilized by hydrogen bonds with C197 and G201. This conformation forces the intervening A125 to the exterior of the molecule. In molecule B, the A125 C8 carbon is 3.0 Å from the U205 O3' and 3.5 Å from the A206 OP1 oxygen on a symmetry-related molecule A (Figure 3C). These contacts are closer than C-O van der Waals distance of 3.7 Å and therefore are potential steric clashes or (mostly weak) C-H...O hydrogen bonds. In molecule B of the M2 structure, G126U replaces the G-A pair with a Watson-Crick U-A pair (Figure 3D). Importantly, substitution of A125 with uracil allows this smaller base to assume the position previously occupied by C124. The net result is that C124 is flipped out of the structure and points towards the 5' and 3' tails of a symmetry-related molecule A, creating an opportunity for the C124 N4 amino group to hydrogen-bond with the U106 O4 or C260 2'-OH of the symmetry-related chain. Similar structural rearrangements appear to occur in molecule A, with analysis limited by the lower quality of electron density in that region. Residues 124-126 in Molecule A are not involved in lattice contacts in either WT or M2 crystals. In sum, the M2 structure demonstrates that repositioning of the bulge from position A125 to C124 creates new intermolecular hydrogen bonds and ameliorates a potential clash, while maintaining the core structure of the RNA.

An insertion mutation creates a base-stacking lattice contact

M3, which contains G134A and U185AA, involves a region critical for tertiary interactions of P4-P6. U185AA replaces U185 in the A-rich bulge (in sequence AAUAA) with two adenines (annotated in M3 as A185 and A185*), creating a stretch of six adenines. U185AA may also be viewed as U185A substitution plus A185* insertion. In the WT P4-P6 structure, this region of the RNA adopts a unique “corkscrew turn” conformation around two magnesium ions and is buttressed by a Hoogsteen base pair between A187 and U135 (Figure 4A). This base pair is then capped by a Watson-Crick pairing of C189 and G134. In this conformation, the A-rich bulge does not make substantial lattice interactions with a neighboring molecule.

In the M3 structure, the A185* insertion lets A187 bulge out and establish a new lattice contact. The other two mutations, U185A and G134A, do not cause a large shift of these residues. A185* occupies the position of A186 in WT, pushing A186 to the position of A187 in WT. A186 forms a Watson-

Crick pair with U135 in M3, replacing the A187-U135 Hoogsteen pair in WT. The M3 electron density maps of molecule B revealed that A187 bulges away from the structure to stack with the base of A125 that is flipped out of a symmetry-related molecule A (Figure 4B). Hence, the M3 structure demonstrates that a single-nucleotide insertion can generate a lattice interaction by complementing a dangling residue on an opposing molecule through a base-stacking interaction.

M1, M2 and M3 show a common theme that a bulge can migrate along the RNA chain by one or more residues to bridge a new crystal lattice contact. In M1 and M2, the bulge migration occurs along a helical strand, whereas in M3, the bulge walks along a strand involved in tertiary folding.

A mutation weakens the P6b stem, allowing the L6b loop to make new lattice contacts

We additionally solved the structure of M4, containing an A230U mutation in the P6b stem. While this mutation breaks the Watson-Crick pair with U244, we did not observe disruption of the stem. U230 assumes the position of A230 as in the WT structure (Figure 5A). However, with the smaller base, U230 stacks less extensively with neighboring bases and forms only one hydrogen bond with U244. We used Distance Difference Matrix Plot (DDMP) to compare the M4 and WT structures and identified a conformational change in the lower P6b stem loop (Figure 5B). The weakened base stacking and pairing at U230 allow the lower P6b stem of molecule B to shift by about 1 Å (Figure 5C). The L6b loops in both molecules A and B adopt conformations different from those of the WT and M1-M3 structures, so that the U236 base from molecule A stacks with U236 from a symmetry-related molecule B and also interacts with U199 from another molecule B’ (Figure 5D). In the WT structure, the occupancy for the L6b residues was set to 0, indicative of poor electron density and low confidence in their actual conformation. In the M1, M2 and M3 structures, there is continuous electron density for the backbone of L6b loops, but little support for direct crystal contacts. In M4, on the immediate upper side of A230U is a region of P4-P6 that makes extensive lattice contacts (Figure 1C). In a pseudo-twofold symmetric interaction between molecule A and a symmetry-related molecule B, 2'-OH of U228 forms a hydrogen bond with 2'-OH of U247, 2'-OH of G227 hydrogen-bonds to U249 phosphate, and U249 base-stacks with another U249. These lattice contacts, common to the WT and all mutant P4-P6, hold the molecules in the crystal lattice while letting the lower P6b stem explore new contacts. Thus, we conclude that M4 establishes new crystal contacts by weakening a helical stem and giving the adjacent stem loop flexibility to shift.

DISCUSSION

Our study provides the first proof-of-concept that *in crystal* selection is a viable strategy for enhancing the crystallizability of biological macromolecules. Our exercise revealed several features. First, all four winner mutants we structurally characterized establish new lattice contacts, suggesting that their mutations make it more favorable for them to be incorporated into the crystal lattice by lowering the free-energy change during this process. Second, all four structures are essentially identical to WT, indicating that in crystal selection requires the mutations in the winners not to disrupt the folding or cause large conformational changes. By inspecting the secondary and three-dimensional structures, we expect most other selected P4-P6 mutants also to fold the same as WT. Thus, when a mixed population of molecular variants is subjected to crystallization, only those variants with sufficiently similar conformation are admitted to the growing crystal lattice. This is a welcoming benefit for studying structure-and-function relationship of target RNA molecules. Furthermore, this feature allows *in crystal* selection to be used for identifying and characterizing variants that do not change the overall structure but improve other desired properties.

We observe local improvements of the electron density maps in the vicinity of the selected mutations relative to the WT structure. For example, in WT P4-P6 the solvent-exposed, bulged residue U130 is disordered and hence reported with occupancy set to zero (Cate et al., 1996). In the M1 structure we cannot only identify U130 as it forms a base pair with A192, but also resolve the position of the bulged A131, which is stabilized by the new lattice interaction. This suggests that *in crystal* selection may be an effective strategy for stabilizing loops or other disordered regions within the context of the crystal lattice.

While we characterize M1-M4 individually, we notice that both M2 and M3 induce changes involving A125. M2 mutates A125 to a uracil that takes over the position of C124. In molecule B, such a local structural rearrangement allows P4-P6 to avoid a potentially unfavorable interaction with a neighboring molecule and instead positions C124 for new lattice contacts (Figure 3). A125 in molecule A is stacked on the new lattice contact induced by the M3 mutations (Figure 4). Furthermore, in the previous structure of P4-P6 ΔC209 mutant at 2.25 Å resolution (Juneau et al., 2001), the deletion results in A125 forming favorable lattice contacts not seen in the WT crystals—the base of A125 in molecule A forms a hydrogen bond with the U185 phosphate in a symmetry-related molecule B and the 2'-OH of A125 in molecule B hydrogen-bonds to the U205 phosphate of a symmetry-related molecule A. These new contacts are reflected by comparative SHAPE analysis of P4-P6 ΔC209 in solution and in crystals (Vicens et al., 2007). Thus, A125 may be considered a hotspot in the P4-P6 RNA for engineering new lattice contacts. In the larger, catalytically active *Tetrahymena* ribozyme, A125 mediates an inter-domain interaction by stacking with the A324 base in the L9 loop (Guo et al., 2004).

None of the mutants characterized here diffracted to higher resolution than WT P4-P6 when crystallized individually. We speculate that the lack of improvement in resolution may result from the complexity of P4-P6 crystal form, in which two molecules occupy the asymmetric unit. When selecting for mutant molecules that can successfully join the growing crystal lattice, each mutant needs only occupy one position in the asymmetric unit in order to become a winner. However when crystallized individually, a mutant has to fill both positions, perhaps negatively affecting crystal quality. Therefore we anticipate that the *in crystal* selection approach will garner superior results with crystal forms containing a single molecule per asymmetric unit.

In crystal selection is most applicable to situations in which crystals of a RNA, RNA-protein, or RNA-ligand complex have been obtained but do not diffract well. Indeed, in our experience working with the *Tetrahymena* group I intron and other large RNAs, most crystal forms do not diffract well (Golden et al., 1997). Therefore, we expect *in crystal* selection to be widely useful. Most other crystallization methods, as summarized in the Introduction, focus on obtaining crystals *ab initio*. Thus, *in crystal selection* and other methods are very complementary to each other and can be used in combination. Further, in protein crystallography, “cross-seeding” using microcrystals of a distantly homologous protein or even non-macromolecular solid substances can produce crystals for a new protein (Abuhammad et al., 2017). We imagine that *in crystal* selection may be used to obtain novel crystals by cross-seeding.

Our P4-P6 mutant structures highlight principles of RNA crystal packing that may be used to rationally design crystallization constructs. In the M1-M3 structures, bulged residues are repositioned or introduced to form hydrogen bonds or base-stacking across different molecules, while the local base pairing and the overall structure are preserved. As such, bulged RNA bases can be “walked” along the chain to a desired position without destabilizing the core packing of the molecule. Based on the M1 structure, we propose that existing bulged residues in a secondary structure may be repositioned in either direction along a helix, so that some of these variants might fortuitously create new lattice contacts. Similarly, new bulged residues may be inserted. Such bulge residues add or alter features to the “boring” surface of an RNA helix. The lesson from the current study is that both A and smaller C/U residues should be tried. Our M2 and M3 structures suggest that the bulge engineering strategy may be applied to unpaired surface regions of the target RNA. Admittedly, in the absence of three-dimensional structures, these constructs are hard to design. However, chemical or nuclease footprinting can identify solvent-accessible regions. Additionally, sequence variation from different species may provide guidance. For instant, the U185 mutated in M3 is the only residue in the A-rich bulge that is not completely conserved. The "bulge walking" approach may be used both for crystallization *ab initio* and for improving existing RNA crystals.

METHODS

Construction of P4-P6 mutant library

The library was constructed by mutagenic PCR using forward primer

5'-CAGTGAATTCAAT*TAATACGACTCACTATA***GGAAT**-3' (underlined, EcoRI site; italic, T7 promoter; bold, transcribed P4-P6 sequence) and reverse primer

5'-GCAGGTCGACTCTAGA*CTCTTC*-3' (underlined, XbaI site for cloning; italic, EarI site for linearization)**.** Error-prone PCR reactions contained 10 mM Tris-HCl pH 8.8, 50 mM KCl, 4.76 mM MgCl$_2$, 0.56 mM dATP, 0.9 mM dCTP, 0.20 mM dGTP, 1.40 mM dTTP, 0.5 MnCl$_2$, 0.5 µM each primer, 3 ng/µl WT P4-P6 template, and 0.08 U/µl Taq DNA polymerase (Roche). Reactions were initially heated at 94°C for 5 min, followed by 10 cycles of 94°C for 1 min, 52°C for 1 min, 72°C for 3 min, and a final 10-min incubation at 72°C. The PCR product was purified on a 2% agarose gel, digested with EcoRI and XbaI, and ligated into pUC19. The ligation reaction was transformed into XL-1 Blue MRF' electrocompetent *E. coli* cells (Stratagene). We estimated the library contains 10 million independently transformed cells.

RNA transcription and purification

RNA samples for in crystal selection and for individual crystallization were generated using *in vitro* transcription. We first isolated mega/maxiprep plasmid DNA in the pUC19 vector containing the RNA sequence preceded by a T7 promoter and followed by an EarI restriction site. Linear templates for run-off transcription were prepared by overnight digestion with EarI followed by ethanol precipitation and resuspension in water. 130 µg of template DNA was then added to a 5 ml transcription reaction containing 40 mM Tris pH7.5, 25 mM $MgCl_2$, 4 mM DTT, 20 mM spermidine, 3 mM each NTP, 350 µl purified T7 RNA polymerase, and 2 µg inorganic pyrophosphatase (Sigma). After 4 hr of incubation at 37°C, reactions were ethanol-precipitated and the RNA purified over a 6% polyacrylamide denaturing gel using a BIO-RAD tube gel apparatus. To elute RNA from the tube gel we used 0.5X TBE buffer supplemented with 3 M urea. Fractions were analyzed by UV absorbance and denaturing PAGE, and pure fractions were pooled, buffer exchanged into 10 mM Na HEPES pH 7.5, and concentrated using an Amicon centrifugal filter device.

In crystal selection

Purified P4-P6 mutant library was annealed at 6 mg/ml by first preparing a solution of RNA in 10 mM NaCl, 5 mM $MgCl_2$, and 5 mM Na HEPES pH 7.5. The RNA was heated to 55°C for 10 min followed by slow cooling of the heat block to around 40°C. The solution was then supplemented with 20 mM $MgCl_2$ and allowed to cool further to room temperature. The annealed RNA library was screened for crystallization at room temperature by hanging drop vapor diffusion by mixing 2 µl RNA with 1 µl of well solution containing 24-30% (v/v) MPD or 22% PEG1000 with 50 mM sodium cacodylate pH 6.0 and 0.5 mM spermine. Crystals were observed in MPD and PEG1000 conditions within 4 days and grew to at least 200 µm in size. A large crystal from each condition was extensively washed in mother liquor and dissolved in 5 µl water. Reverse transcription was performed using 4 µl RNA, primer with sequence 5'-

TGAACTGCATCCATATCAACAG-3', and SuperScript II (Invitrogen) at 42°C for 50 min. The cDNA product was amplified with Pfu Turbo DNA polymerase and primers 5'-CAGTGAATTCAAT*TAATACGACTCACTATA* **GGAATTGCGGGAAAGGGGTC**-3' (underlined, EcoRI site; italic, T7 promoter; bold, part of transcribed P4-P6 sequence), and 5'-CGACTCTAGA*CTCTTC*G**TGAACTGCATCCATATCAACAG**-3' (underlined, XbaI site for cloning; italic, EarI site; bold, P4-P6 sequence). The PCR product was restriction digested, ligated into pUC19, and transformed as described in library construction.

Crystallization and structure determination

Mutant RNA was transcribed, purified, and annealed as described in in crystal selection. M1 crystals prepared from PEG1000 solutions were produced at room temperature via hanging drop vapor diffusion by mixing 1 µl annealed RNA solution at 6 mg/ml with 2 µl of a well solution containing 22% (v/v) PEG1000, 50 mM sodium cacodylate pH 6.1, and 0.5 mM spermine. Single crystals were harvested and briefly soaked in a fresh well solution containing 10% isopropanol before flash frozen in liquid nitrogen. For M1 crystals obtained by MPD, 1 µl of a 6 mg/ml annealed RNA solution was mixed with 2 µl of a well solution containing 26% (v/v) MPD, 50 mM sodium cacodylate pH 6.0, and 1 mM cobalt hexamine. Crystals were flash-frozen without isopropanol treatment.

Crystals for M2 were grown by mixing 1 µl of 6 mg/ml RNA with 2 µl of well solution composed of 28% (v/v) MPD, 50 mM sodium cacodylate pH 6.0, and 0.5 mM spermine. M3 was crystallized under the same condition as M2 except that the MPD concentration in well solution was 31% (v/v). M2 and M3 crystals were frozen without isopropanol treatment.

M4 was crystallized from a drop containing 1 µl annealed RNA at 6 mg/ml and 2 µl of well solution composed of 22% (v/v) PEG1000, 50 mM sodium cacodylate pH 6.1, and 0.5 mM spermine. Prior to freezing, crystals were sequentially soaked in mother liquor solutions supplemented with 5%, 10%, and finally 15% (v/v) isopropanol.

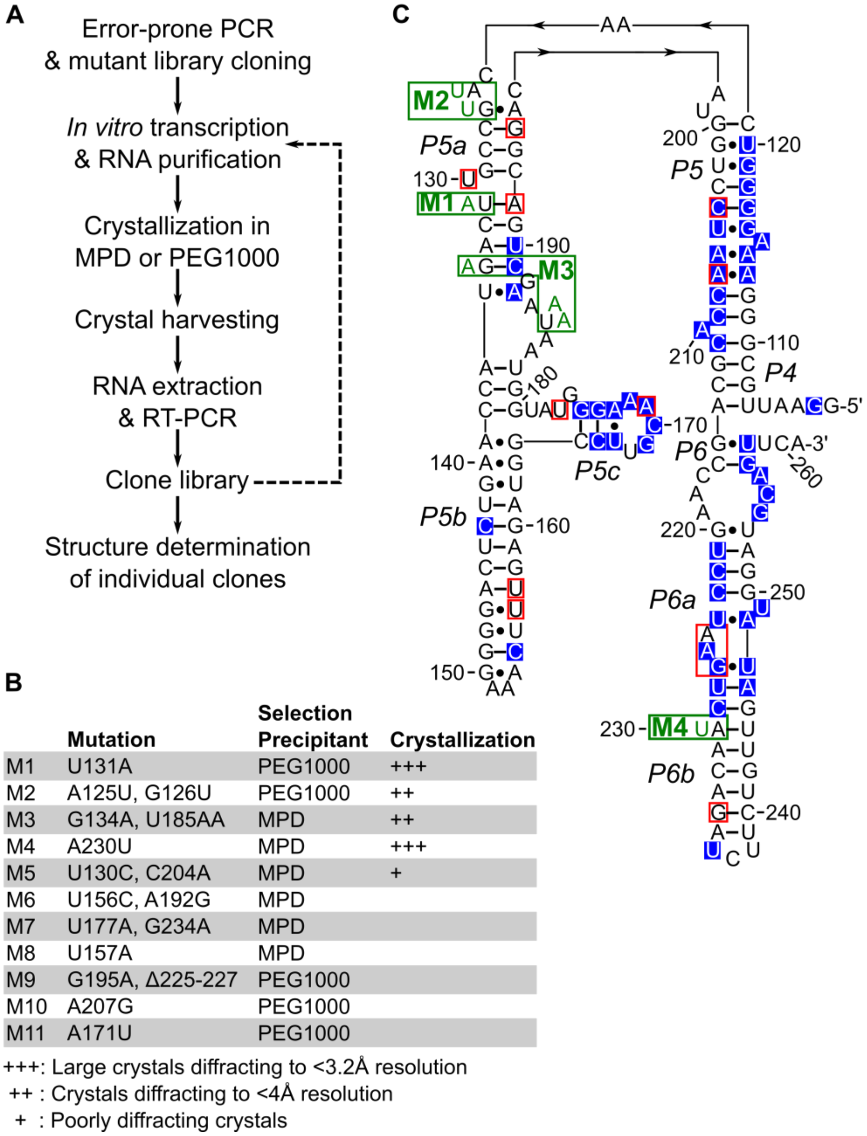
Diffraction data were collected at 100K at the Advanced Photon Source beamline 24-ID-C using an energy of ~12.6 keV. Data were indexed, integrated, and scaled using XDS (Kabsch, 2010). To solve a structure, we started with the WT P4-P6 coordinates (PDB code 1GID) and deleted the mutated residues and an additional residue on the 5' and 3' sides of the mutation(s). We then generated an initial solution by performing rigid-body refinement of the deleted model to the data in PHENIX (Adams et al., 2010) using the default settings, which generally resulted in working and free R factors in the 30-35% range. Next we inspected the difference election density map; if additional residues near the site of mutation appeared to adopt a new conformation, we further reduced the phase bias of the search model by deleting those residues and repeating the rigid-body refinement. Magnesium ions that did not fit the new electron density were removed. Once an acceptable starting model was obtained, we reset all atomic occupancies to 1 and manually modeled in the missing and mutated residues using Coot (Emsley et al., 2010) . The model was further refined with autoBUSTER (Global Phasing Limited) (Smart et al., 2012), using the TLSbasic macro with target restraints from the initial deleted 1GID model and without B-factor refinement. Group B factors (one parameter per residue) were subsequently refined in PHENIX and the model was manually adjusted in Coot. These final refinement steps were iterated until an acceptable solution was obtained.

AUTHOR CONTRIBUTIONS

F.G. and T.R.C. conceived the project over 15 years ago. F.G. and E.P. performed the *in crystal* selection. G.M.S. rekindled the project. G.M.S., R.W. and F.G. prepared and crystallized the P4-P6 mutants and determined their structures. G.M.S. and F.G. wrote the manuscript with input from T.R.C.

ACKNOWLEDGMENTS

We would like to thank Duilio Cascio and Michael Collazo in the UCLA-DOE crystallization and X-ray diffraction facilities for technical supports. We thank Igor Kourinov at the Advanced Photon Source sector 24 (NE-CAT) for assistance with data collection; NE-CAT is supported by funding from the NIH (P41 GM103403 and S10 RR029205), and DOE (DE-AC02-06CH11357). The authors declare that they have no conflict of interest.



	Mutation	Selection Precipitant	Crystallization
M1	U131A	PEG1000	+++
M2	A125U, G126U	PEG1000	++
M3	G134A, U185AA	MPD	++
M4	A230U	MPD	+++
M5	U130C, C204A	MPD	+
M6	U156C, A192G	MPD	
M7	U177A, G234A	MPD	
M8	U157A	MPD	
M9	G195A, Δ225-227	PEG1000	
M10	A207G	PEG1000	
M11	A171U	PEG1000	

+++: Large crystals diffracting to <3.2Å resolution
++ : Crystals diffracting to <4Å resolution
+ : Poorly diffracting crystals

Figure 1. *In crystal* selection of P4-P6 variants. (A) Schematic of the *in crystal* selection strategy. Dashed line represents iterative selections that were performed but not used in structure determination. (B) Summary of selected mutants. (C) Mutations projected on the P4-P6 secondary structure. Green boxes show mutants that are structurally characterized. Red boxes show all other mutation sites. Bases highlighted with blue background fall within 5Å of a neighboring molecule in the crystal lattice. Note that unpaired residues at the bend between P5 and P5a are drawn in a way reflecting their direct stacking on top of the P5 and P5a stems. Such drawing makes it easier to understand the structural changes induced by the M2 mutations.

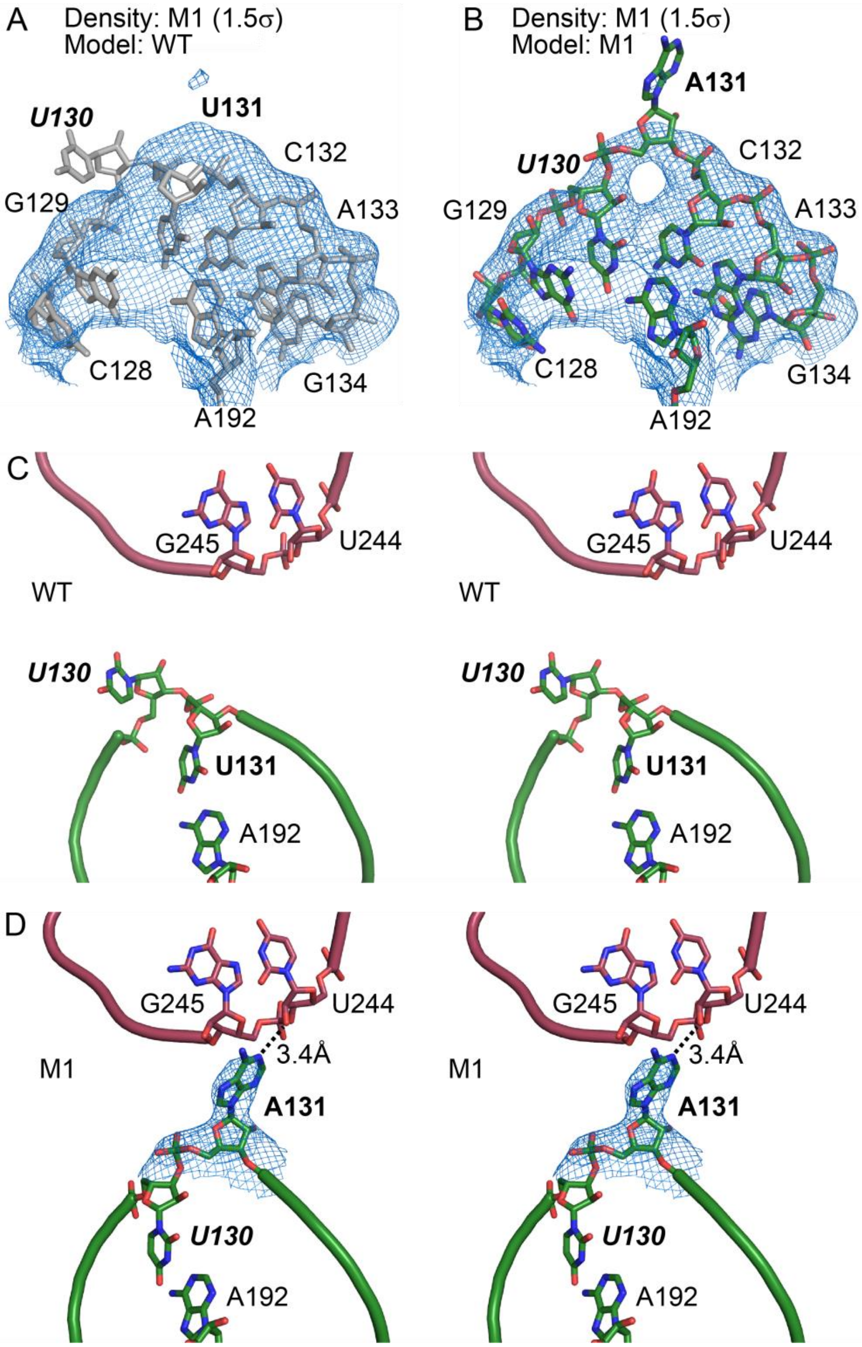


Figure 2. Generation of a new lattice contact in mutant M1. (A) Model of WT P4-P6 (grey) fit to the 2Fo-Fc map (blue) of the M1 mutant contoured at 1.5σ, showing residues at positions 130 and 131 do not fit the electron density. (B) Model of M1 fits into the 2Fo-Fc map as in (A). (C) Stereoscopic view of WT P4-P6 crystal, showing P5 region of molecule A in green and P6 region of molecule B in red. (D) The 3.14 Å resolution M1 structure shows that swapping of base-pairs positions A131 in contact with neighboring molecule. Blue mesh shows σA-weighted 2Fo-Fc map contoured at 0.8σ level. Same coloring as (C).

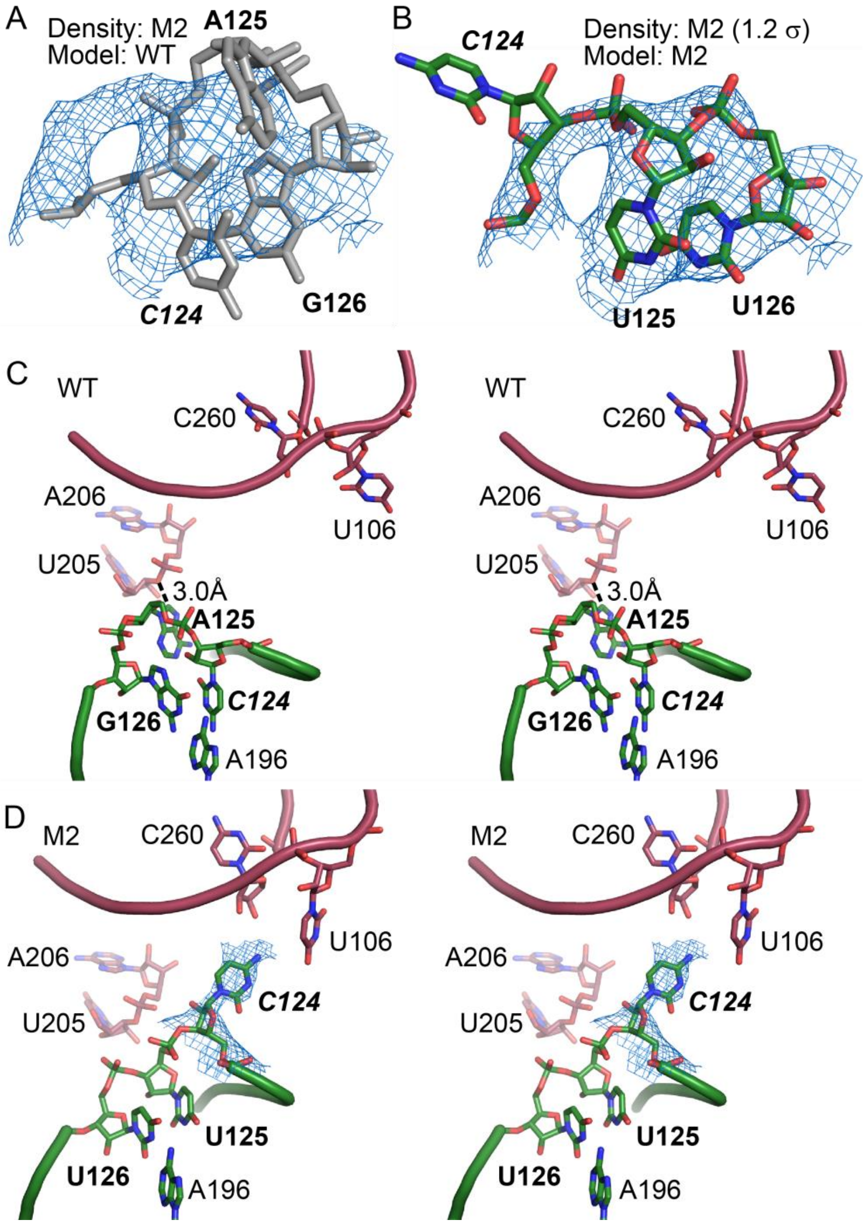


Figure 3. Rearranged lattice interaction in the M2 mutant. (A) Structure of WT P4-P6 (grey) compared to the 2Fo-Fc map from the M2 mutant at 1.2σ. (B) Fitting of the M2 mutaions to the electron density in (A) reveals the conformational rearrangment in the mutation structure. (C) Stereo view of the J5/5a hinge of WT P4-P6 (molecule B) in green and symmetry-related molecule A in red. (D) The same region in the 3.70 Å resolution M2 structure. Electron density map (2Fo-Fc, 0.8σ) in blue showing the new position of C124.

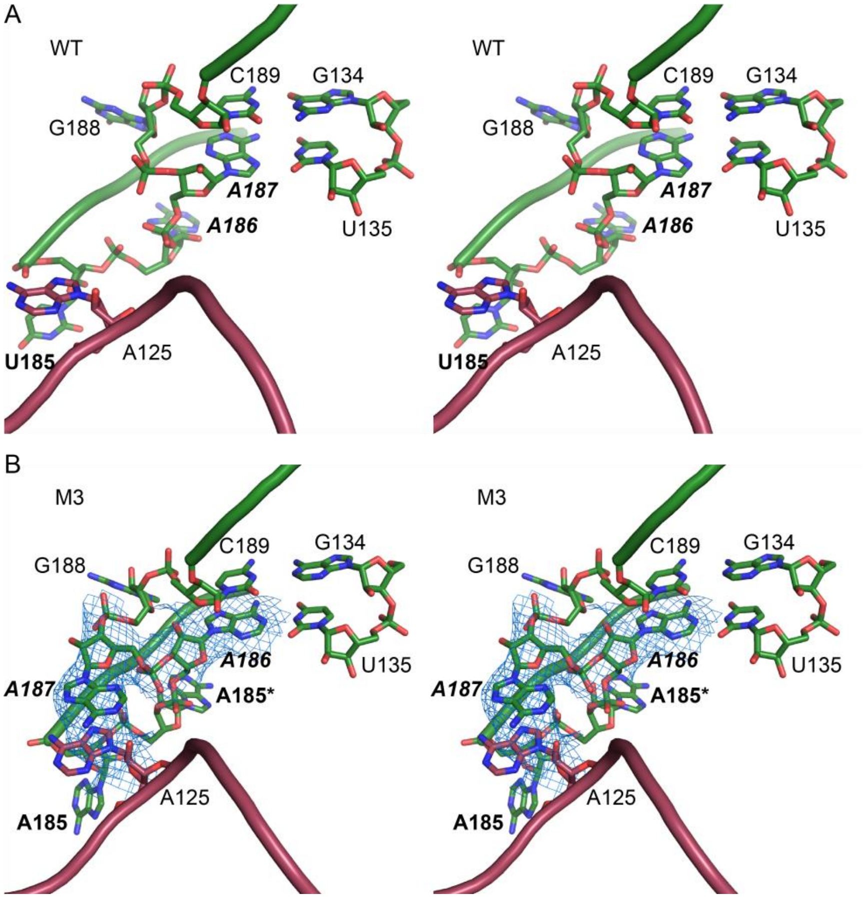


Figure 4. Lattice contact created by a base-stacking interaction in mutant M3. (A) Stereo view of the A-rich bulge (green) and J5/5a region of a symmetry-related molecule in the WT structure (red). (B) The M3 structure at 3.95 Å resolution shows that inserted A185* and A186 take the approximate positions of A186 and A187 in the WT structure. Such a shift of register leads A187 to bulge out and produce new base stacking with the bulged A125 of molecule A. Blue mesh shows 2Fo-Fc map contoured at 1.0 σ level.

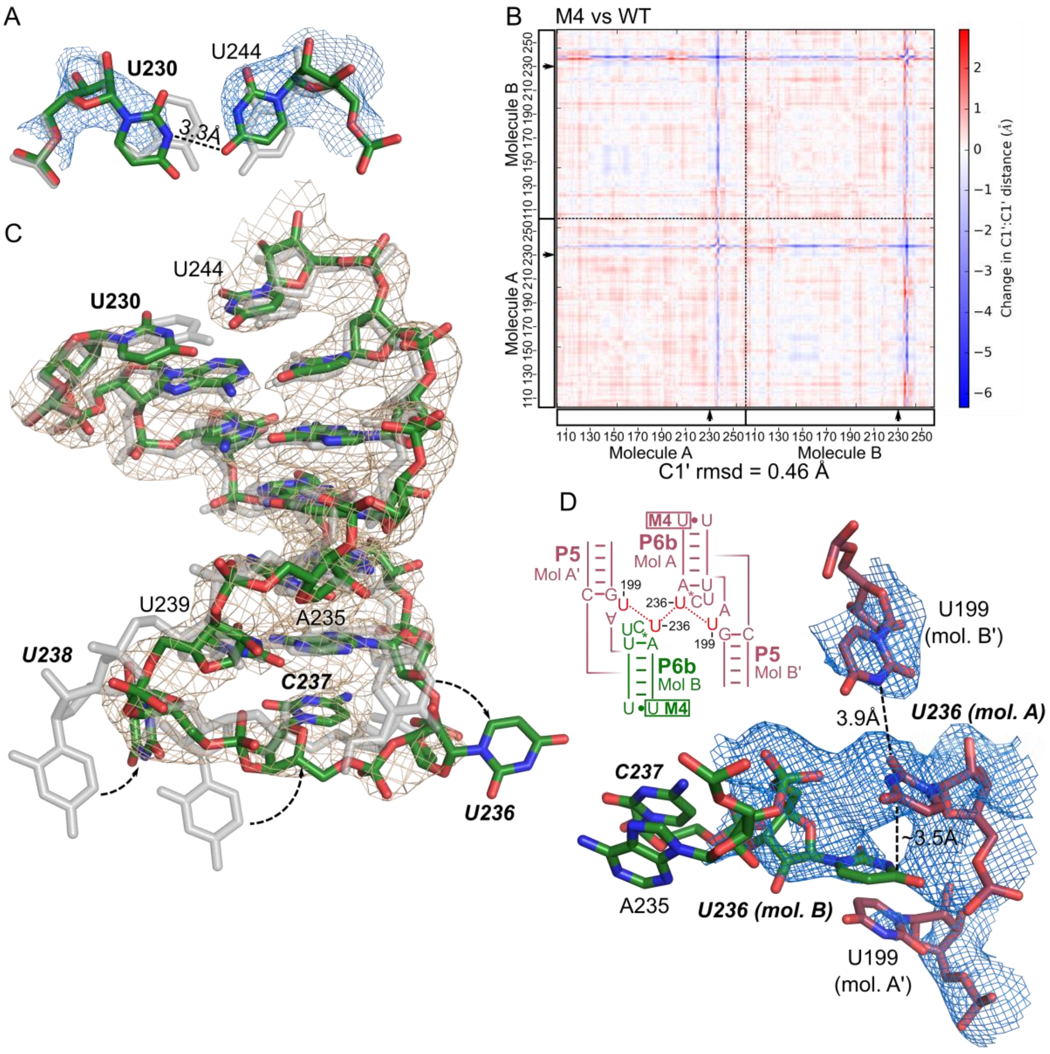


Figure 5. M4 weakens the P6b helix, allowing the L6b loop to establish new lattice contacts.

(A) The A230U mutation site in the 2.8-Å-resolution M4 structure (with atoms colored red, green and blue), compared to the WT structure shown in silver. The structures were superimposed by aligning all molecule B atoms outside of residues 230-244. Blue mesh shows the immediately surrounding 2Fo-Fc map contoured at 1.2 σ level. (B) Distance difference matrix plot between M4 and WT structures. Both molecules A and B in the asymmetric unit are shown, separated by a dash line. Mutated sites are marked with arrows. Calculated r.m.s. deviation for all C1' atoms in the asymmetric unit is shown below the plot. (C) The P6b stem loops in M4 and WT structures superimposed and colored as in (A). 2Fo-Fc map contoured at 1.2 σ level is shown as gray mesh. (D) New crystal contacts established by M4. Blue mesh shows 2Fo-Fc map contoured at 0.6 σ. A schematic of the new lattice contacts involving U236 in L6b and U199 in J5a/5 from four different molecules.

Table 1. Crystallographic data and refinement statistics.

Crystals	M1	M1	M2	M3	M4
Growth conditions	PEG1000	MPD	MPD	MPD	PEG1000
Soaking	Soaked with isopropanol	No	No	No	Soaked with isopropanol
Data collection					
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Cell dimensions					
a, b, c (Å)	74.7, 130.0, 146.4	75.5, 133.0, 147.5	75.6, 132.9, 146.8	76.7, 130.9, 146.4	74.3, 129.5, 146.1
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 90
Resolution (Å)[a]	66.5 – 3.14 (3.22 – 3.14)	67.2 – 3.14 (3.22 – 3.14)	64.2 – 3.70 (3.8 – 3.70)	66.1 – 3.95 (4.06 – 3.95)	63.6 – 2.80 (2.87 – 2.80)
R_{meas} (%)[b]	9.5 (90.6)	11.9 (60.0)	7.4 (104.4)	9.8 (133.3)	9.9 (175.7)
$R_{p.i.m.}$(%)[b]	2.7 (24.2)	3.6 (19.9)	3.3 (44.2)	2.9 (36.0)	2.9 (45.4)
I/σ	16.3 (2.98)	12.9 (3.62)	13.0 (1.92)	15.6 (2.09)	15.0 (1.46)
$CC_{1/2}$	99.6 (89.9)	99.1 (93.9)	99.7 (78.7)	99.9 (86.2)	99.5 (76.0)
Completeness (%)	98.9 (87.9)	95.9 (82.6)	99.5 (96.9)	99.5 (94.4)	99.3 (90.9)
Redundancy	12.9 (11.8)	10.3 (8.1)	5.4 (5.0)	12.9 (12.4)	12.9 (11.3)
Refinement					
Resolution (Å)	63.8 – 3.14	67.2 – 3.14	64.2 – 3.70	66.1 – 3.95	63.6 – 2.80
No. of unique reflections	25,212	25,498	16,335	13,323	35,173

% of reflections in test set	10%	10%	10%	10%	5%
R_{work} / R_{free}	0.216 / 0.227	0.247 / 0.265	0.217 / .224	0.247 / 0.255	0.217 / 0.231
No. of atoms					
RNA	6,770	6,776	6,778	6,812	6,746
Mg^{2+}	24	24	24	7	24
K^{+}	0	0	0	0	1

[a] Values in parentheses are statistics for the highest resolution bin.

[b]

$$\mathrm{R}_{meas} = \sum_{hkl} \sqrt{N(hkl)/(N(hkl)-1)} \times \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$$

$$\mathrm{R}_{p.i.m.} = \sum_{hkl} \sqrt{1/(N(hkl)-1)} \times \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$$

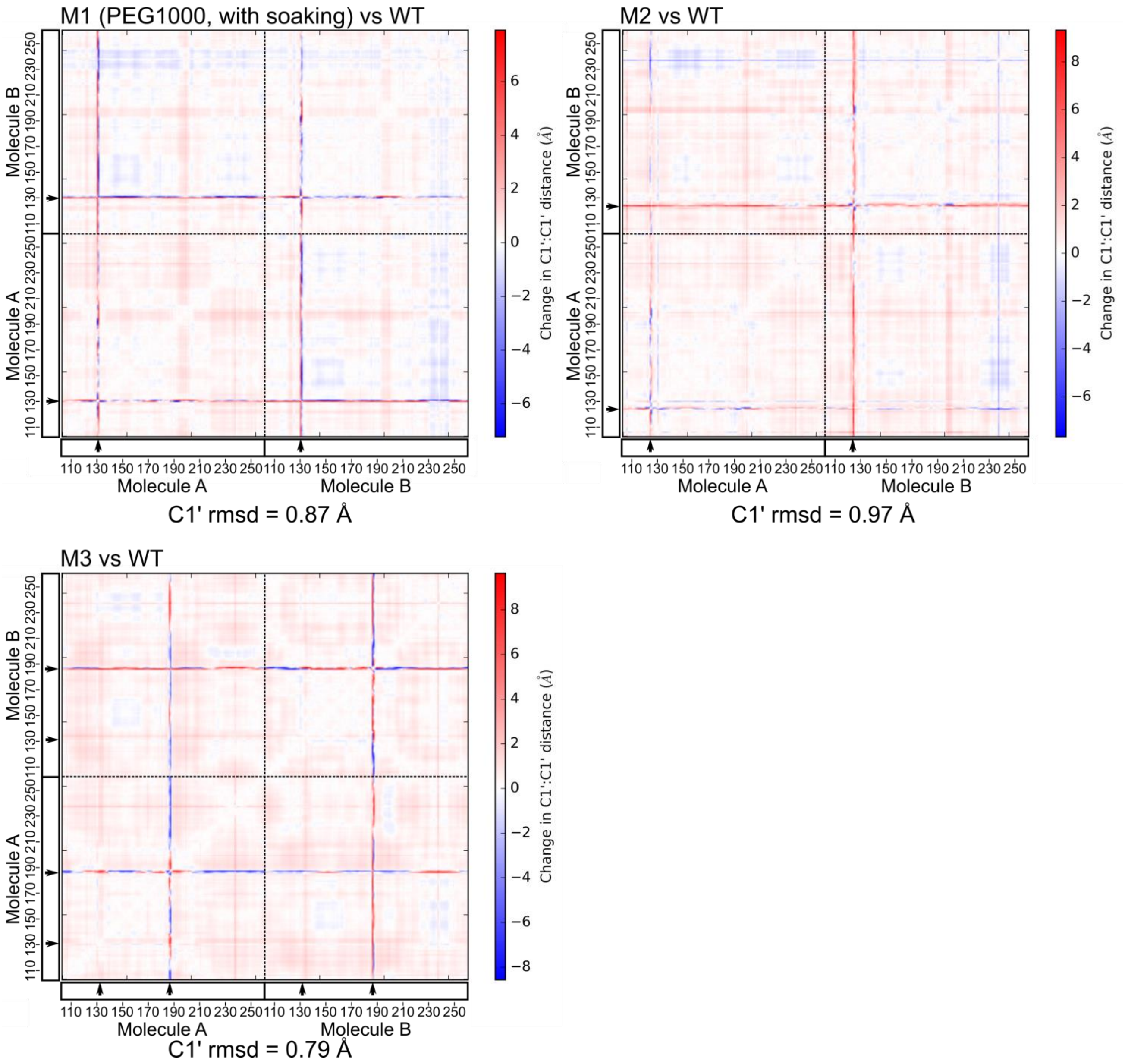


Figure S1. Distance difference matrix plots between M1-M3 mutant structures and WT P4-P6 (PDB ID 1GID). Both molecules A and B in the asymmetric unit are shown, separated by a dash line. Mutated sites are marked with arrows. Calculated r.m.s. deviations for all C1' atoms in the asymmetric unit are shown below the plots.

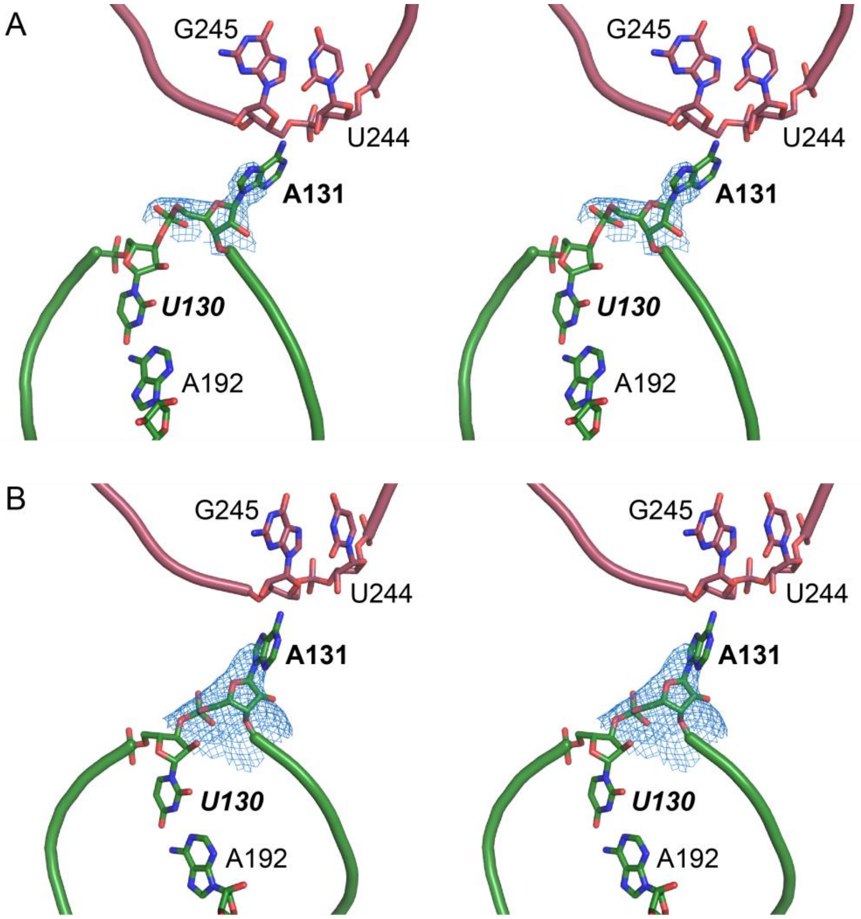


Figure S2. Lattice interaction in M1 mutant crystals grown under MPD conditions and without cryoprotection. The structure was determined at 3.14 Å resolution. Stereo view of the P5a region of molecule A (green) showing U130-A192 base pair and proximity of A131 to neighboring molecule (red). Blue mesh shows 2Fo-Fc maps contoured at 0.6σ.

References:

Abuhammad, A., McDonough, M.A., Brem, J., Makena, A., Johnson, S., Schofield, C.J., and Garman, E.F. (2017). “To cross-seed or not to cross-seed”: a pilot study using metallo-β-lactamases. Cryst. Growth Des. *17*, 913-924.

Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W.*, et al.* (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr *66*, 213-221.

Cate, J.H., Gooding, A.R., Podell, E., Zhou, K., Golden, B.L., Kundrot, C.E., Cech, T.R., and Doudna, J.A. (1996). Crystal structure of a group I ribozyme domain: principles of RNA packing. Science *273*, 1678-1685.

Doudna, J.A. (1997). Preparation of homogeneous ribozyme RNA for crystallization. Methods Mol. Biol. *74*, 365-370.

Doudna, J.A., Grosshans, C., Gooding, A., and Kundrot, C.E. (1993). Crystallization of ribozymes and small RNA motifs by a sparse matrix approach. Proc. Natl. Acad. Sci. U. S. A. *90*, 7829-7833.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr. D Biol. Crystallogr. *66*, 486-501.

Ferre-d'Amare, A.R., and Doudna, J.A. (1997). Establishing suitability of RNA preparations for crystallization. Determination of polydispersity. Methods Mol. Biol. *74*, 371-377.

Ferre-D'Amare, A.R., and Doudna, J.A. (2000). Crystallization and structure determination of a hepatitis delta virus ribozyme: use of the RNA-binding protein U1A as a crystallization module. J. Mol. Biol. *295*, 541-556.

Ferre-D'Amare, A.R., Zhou, K., and Doudna, J.A. (1998). A general module for RNA crystallization. J. Mol. Biol. *279*, 621-631.

Gesteland, R.F., Cech, T.R., and Atkin, J.T. (2005). The RNA World, Third Edition edn (Cold Spring Harbor Laboratory Press).

Golden, B.L., Podell, E.R., Gooding, A.R., and Cech, T.R. (1997). Crystals by design: a strategy for crystallization of a ribozyme derived from the Tetrahymena group I intron. J. Mol. Biol. *270*, 711-723.

Guo, F., and Cech, T.R. (2002). Evolution of Tetrahymena ribozyme mutants with increased structural stability. Nat. Struct. Biol. *9*, 855-861.

Guo, F., Gooding, A.R., and Cech, T.R. (2004). Structure of the Tetrahymena ribozyme: base triple sandwich and metal ion at the active site. Mol. Cell *16*, 351-362.

Juneau, K., and Cech, T.R. (1999). In vitro selection of RNAs with increased tertiary structure stability. RNA *5*, 1119-1129.

Juneau, K., Podell, E., Harrington, D.J., and Cech, T.R. (2001). Structural basis of the enhanced stability of a mutant ribozyme domain and a detailed view of RNA--solvent interactions. Structure *9*, 221-231.

Kabsch, W. (2010). XDS. Acta Crystallogr. D Biol. Crystallogr. *66*, 125-132.

Ke, A., and Doudna, J.A. (2004). Crystallization of RNA and RNA-protein complexes. Methods *34*, 408-414.

Koldobskaya, Y., Duguid, E.M., Shechner, D.M., Suslov, N.B., Ye, J., Sidhu, S.S., Bartel, D.P., Koide, S., Kossiakoff, A.A., and Piccirilli, J.A. (2011). A portable RNA sequence whose recognition by a synthetic antibody facilitates structural determination. Nat. Struct. Mol. Biol. *18*, 100-106.

Nimjee, S.M., White, R.R., Becker, R.C., and Sullenger, B.A. (2017). Aptamers as Therapeutics. Annu Rev Pharmacol Toxicol *57*, 61-79.

Reiter, N.J., Osterman, A., Torres-Larios, A., Swinger, K.K., Pan, T., and Mondragon, A. (2010). Structure of a bacterial ribonuclease P holoenzyme in complex with tRNA. Nature *468*, 784-789.

Scott, W.G., Finch, J.T., Grenfell, R., Fogg, J., Smith, T., Gait, M.J., and Klug, A. (1995). Rapid crystallization of chemically synthesized hammerhead RNAs using a double screening procedure. J Mol Biol *250*, 327-332.

Shechner, D.M., Grant, R.A., Bagby, S.C., Koldobskaya, Y., Piccirilli, J.A., and Bartel, D.P. (2009). Crystal structure of the catalytic core of an RNA-polymerase ribozyme. Science *326*, 1271-1275.

Smart, O.S., Womack, T.O., Flensburg, C., Keller, P., Paciorek, W., Sharff, A., Vonrhein, C., and Bricogne, G. (2012). Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. Acta Crystallogr D Biol Crystallogr *68*, 368-380.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science *249*, 505-510.

Vicens, Q., Gooding, A.R., Laederach, A., and Cech, T.R. (2007). Local RNA structural changes induced by crystallization are revealed by SHAPE. RNA *13*, 536-548.

Wilson, D.S., and Szostak, J.W. (1999). In vitro selection of functional nucleic acids. Annu. Rev. Biochem. *68*, 611-647.

Ye, J.D., Tereshko, V., Frederiksen, J.K., Koide, A., Fellouse, F.A., Sidhu, S.S., Koide, S., Kossiakoff, A.A., and Piccirilli, J.A. (2008). Synthetic antibodies for specific recognition and crystallization of structured RNA. Proc. Natl. Acad. Sci. U. S. A. *105*, 82-87.

Zhang, J., and Ferre-D'Amare, A.R. (2013). Co-crystal structure of a T-box riboswitch stem I domain in complex with its cognate tRNA. Nature *500*, 363-366.

Zhang, J., and Ferre-D'Amare, A.R. (2014). New molecular engineering approaches for crystallographic studies of large RNAs. Curr Opin Struct Biol *26*, 9-15.

Chapter 6: An unexpected RNA-binding region of SHARP promotes association with the long non-coding RNA Xist

Abstract

During X chromosome inactivation, the lncRNA Xist recruits the transcriptional co-repressor SHARP to chromatin through direct association with the Xist 5' region. The molecular interface between these two components remains undefined. To dissect this interaction, we performed *in vitro* binding assays with reconstituted Xist-SHARP complexes. Surprisingly, we determined that a fragment from the central region of SHARP strongly binds pieces of the Xist RNA, despite the absence of annotated RNA binding domains in this region. In addition, the N-terminal RRMs of SHARP can also bind Xist, but these may form a multimeric complex and do not appear to be specific for any portion of the Xist 5' segment. These data support a model in which assembly of the Xist-SHARP RNP employs multiple RNA-binding activities from both folded and unstructured portions of the SHARP protein.

Introduction

Despite the widespread transcription of long non-coding RNAs (lncRNAs) across the human genome, relatively few of these diverse transcripts have a proven molecular function or biological significance [1]. An outstanding exception is Xist, the lncRNA responsible for driving the transcriptional silencing of one copy of the pair of X chromosomes in female cells, a process termed X-chromosome inactivation (XCI). As an essential mediator between the inactive X chromosome (Xi) and the protein factors suppressing gene expression, Xist represents a model system for the study of lncRNA-protein interactions. In this light, insights into Xist-protein complexes would not only deepen our understanding of XCI, but also provide a framework for investigating the wide landscape of unexplored lncRNAs. However, the door to structural investigations of Xist silencing complexes would remain shut until multiple reports detailed the complete Xist-protein interactome [2]–[4]. Given the opportunity, we initiated a collaboration with the group of Kathrin Plath at UCLA to push forward our structural understanding of the XCI silencing machinery.

In the current model for XCI, silencing commences with localization of the Xist RNA at the Xi chromosome by Scaffold attachment factor A (Saf-A, also known as hnRNP-U) [5]. Saf-A acts as the physical link between chromatin and Xist, generating a cloud of RNA molecules coating the Xi that is visible by fluorescence microscopy. In turn, Xist functions as a platform for recruiting additional protein effectors to the Xi chromatin [6]. These effectors generally produce silencing through two distinct mechanisms: (1) physical relocation of Xi to the nuclear lamina (mediated by the Xist-Lbr interaction [7]); and (2) chromatin modifications which inhibit Pol II binding and transcription.

Effectors within this second category first surfaced when McHugh et al. employed RNA affinity purification coupled to mass spectrometry (RAP-MS) to systematically identify proteins associating with Xist in mouse embryonic stem cells [2]. By covalently crosslinking RNA-protein interactions and co-purifying the complexes under denaturing conditions, the authors succeeded in isolating specific Xist binding partners within cells. By far the most highly enriched partner was the SMRT/HDAC1-associated repressor protein (SHARP), a member of the SPEN family of transcriptional regulatory factors conserved from plants to animals. This interaction was independently confirmed by analogous crosslinking-based protocols ([3] and [4]), as well as through genetic [8] and shRNA screening [9]. As these results hinted at an essential biological interaction between Xist and SHARP, we elected to further investigate the structural basis of this complex.

Although the first SPEN family members were uncovered as developmental regulators in fruit flies [10] and rats [11], the human homolog SHARP was initially identified through a two-hybrid screen for SMRT interacting partners [12]. A known transcriptional co-repressor, SMRT also interacts with the histone deacetylase HDAC3 in silencing heterochromatin [13]. Given the Xist-SHARP interaction detected by

McHugh et al., this clearly pointed to a role for SHARP in recruiting transcriptional repressors to the Xi chromosome [2]. Accordingly, the SHARP domain structure is well tailored for bridging RNA and protein partners (Figure 1A). Spanning some 3600 residues (~400 kDa), SHARP includes four N-terminal RNA recognition motifs (RRMs) and a C-terminal SPOC domain that associates with SMRT [14]. Large regions of the protein are predicted to be disordered (Figure 1B). The structure of RRM2-4 revealed that RRM3 and 4 form an intramolecular interface, whereas RRM2 is connected by a flexible linker [15]. Furthermore, a helical extension on the C-terminus of RRM4 likely tunes the specificity of this domain toward double-stranded RNA, enabling RRM3-4 to bind a combination of single-stranded and helical RNA elements *in vitro*.

However, the long length of the Xist RNA (approx. 18kb) as well as its complex secondary structure make anticipating the SHARP binding site challenging. Xist is principally composed of six sets of tandem repeats (named A-F), which are conserved across placental mammals [6]. The A repeats, consisting of a 26nt block copied ~8.5 times and connected by U-rich sequences, is essential for Xist silencing activity [16]. Chemical probing *in vitro* as well as within cells has produced a variety of models for the structure of the A repeats, and ultimately Xist may adopt an ensemble of alternative configurations ([6], [17], [18]). Some light was shed by CLIP-seq analysis of SHARP, which indicated that the protein largely crosslinks to the U-rich linkers between the A repeats, but also associates with other regions of the RNA [19]. Given the multiple primary and secondary structural features available for SHARP binding, we elected to take an unbiased approach to defining the Xist-SHARP complex and consider all possible modes of interaction.

Results

Our experience working with the Rhed domain from DGCR8 reminded us that not all functional protein domains are discernable by secondary structure predictions, and that in the case of RNA-associated factors, these domains can possess unanticipated RNA-binding activity [20]. We suspected that by focusing our attention on the SHARP RRM2-4 region, we might miss important RNA-binding sites elsewhere in the protein. To further explore the mouse SHARP sequence (mSHARP), we tiled the protein with overlapping constructs and generated codon optimized expression cassettes for production in bacteria (Figure 1A).

With the expectation that over-expression of the mammalian genes in E. coli would prove difficult, we fused a hexa-histidine and maltose-binding protein tag (His6-MBP) to the N-terminus of each construct to enhance yield and solubility. We extensively screened expression parameters, including growth temperature, IPTG concentration, and expression time, as well as co-expression with protein folding chaperones (see Materials & Methods section). This resulted in high-levels of soluble protein for 7/10 constructs, which collectively cover ~70% of the protein sequence (Figure 1A). Proteolysis of unstructured segments made purification challenging, but we were able to achieve reasonable purity (Figure 1C) for simple RNA-binding studies. In addition to constructs from the mouse homolog, we also purified RRM2-4 from the human sequence by an identical procedure (HsSHARP RRM2-4).

With the knowledge that SHARP primarily associates with the 5’ region of Xist, we transcribed and purified fragments of Xist and tested for binding to our SHARP proteins. Using size-exclusion chromatography (SEC) we first analyzed Xist$^{2\text{-}304}$ in isolation (Figure 2A) and then looked for a shift of the RNA upon mixing with each protein construct (Figure 2B-I). Xist$^{2\text{-}304}$ represents the 5’ most fragment of the RNA immediately preceding the A repeats. Comparing the absorbance at 260nm (solid curve, Figure 2) and 280nm (dashed curve), we can clearly identify complexes with mSHARP$^{1650\text{-}2150}$ and mSHARP$^{2050\text{-}2550}$ (Figure 2F and 2G). We potentially also observe minor complexes with mSHARP RRM2-4 and mSHARP$^{440\text{-}846}$ (Figures 2C and 2D, marked with an asterisk). It is possible these represent weaker interactions that are not stable over the SEC run (approx. 20 min).

We next tested for binding to the A repeats (Xist$^{334-725}$). As above, we only observed a substantial shift of the RNA alone peak (Figure 3A) after mixing with mSHARP$^{1650-2150}$ and mSHARP$^{2050-2550}$ (Figures 3F and 3G). The weak interaction between Xist$^{334-725}$ and RRM2-4 or mSHARP$^{440-846}$ is even less pronounced than for Xist$^{2-304}$ (Figure 3C and 3D, note small shift toward high-molecular weight species eluting near 8 mL). Interestingly, we did not detect any binding between the human RNA-binding domains RRM2-4 and either mouse Xist$^{2-304}$ (Figure 2B) or Xist$^{334-725}$ (Figure 3B), despite conservation of the A repeat structure between these species.

An alternative explanation for the minor shift of RNA observed for mouse RRM2-4 and mSHARP$^{440-846}$ could be aggregation of the complex under the low salt conditions (80mM NaCl) used for SEC analysis. This could especially effect mSHARP$^{440-846}$, which is expected to contain a stretch of unstructured residues at the C-terminus. To account for this possibility, we retested binding of mSHARP$^{440-846}$ to all of our Xist 5' fragments and modified the SEC running buffer. The revised buffer contained an intermediate salt concentration (150mM NaCl) along with 5% (v/v) glycerol and 0.00145% (v/v) Triton X-100 to stabilize the protein. Despite the increased ionic strength, we clearly observed association of the protein with all of the Xist fragments (Figure 4). However, in no case did we obtain a complete shift of the RNA peak, despite mixing both components with approximate 1:1 stoichiometry. Furthermore, in the case of Xist$^{2-181}$ (Figure 4B) and Xist$^{727-993}$ (Figure 4H), the absorbance at 280nm is only slightly below the 260nm reading. This potentially indicates these species contain significantly more protein than RNA.

To confirm the presence of RNA in these peaks we analyzed fractions from the SEC runs by denaturing PAGE and detected the RNA by Sybr-Green II staining (Figures 5A and 5B). For mSHARP RRM2-4 (abbreviated mRRM2-4 in Figure 5) and mSHARP$^{440-846}$ we found small amounts of RNA present in the high-molecular weight fractions 4-6 when mixed with either Xist$^{2-181}$ or Xist$^{727-993}$. This is consistent with a complex containing more copies of the protein than RNA. As noted before, we saw a near complete shift of the RNA on the gel when mixed with mSHARP$^{1650-2150}$ or mSHARP$^{2050-2550}$.

In the case of mSHARP$^{440-846}$ binding with Xist$^{727-993}$, we estimated the RNA to protein ratio by titrating the reaction with additional protein and separating the components by SEC (Figure 5C – 5G). Near complete binding occurred at an approximate molar ratio of one RNA to eight proteins (Figure 5G). Because this peak elutes near the void volume of the column, we cannot confidently determine if this represents true assembly of an RNA-protein complex or merely co-aggregation of the two components. To further understand the constitution of the material in this peak, we analyzed a fraction of the SEC eluate by negative-stain electron microscopy (Figure 5H; we thank Jen Quick-Cleveland for collecting these images). We observed a partially homogeneous population of small particles with a distinctive two-lobe structure (see arrows and close-up views in Figure 5H). These particles measured 46 ± 3.9 nm (mean ± SD, n = 10) along their longest axis.

In sum, we identified four fragments of mSHARP capable of associating with the 5' region of Xist *in vitro* (RRM2-4, mSHARP$^{440-846}$, mSHARP$^{1650-2150}$, and mSHARP$^{2050-2550}$). Surprisingly, our disorder prediction did not detect any appreciable structure with either mSHARP$^{1650-2150}$ or mSHARP$^{2050-2550}$, and RNA-binding activity has not been linked to this region before. One means to gauge the extent of protein folding is to estimate the molecular mass based on the elution volume from the SEC column. Monomeric, globular proteins will elute near their expected size, whereas proteins in an extended or soluble unfolded conformation will have anomalous elution volumes at higher than expected molecular weights.

We applied this strategy to our four Xist-binding hits (Figure 6). For RRM2-4, the protein eluted in two peaks, with the minor peak at 15.5 mL agreeing with the expected molecular mass of 75 kDa for the MBP fusion protein (Figure 6A). The major peak at 11.9 mL corresponds to a mass of 440 kDa, or roughly 6 times the size of the monomeric protein. We also observed anomalous elution for mSHARP$^{440-846}$, but the proximity of the peak to the column void prevents estimation of the size (Figure 6B). Both constructs mSHARP$^{1650-2150}$ and mSHARP$^{2050-2550}$ eluted in the 470 kDa range, which is much larger than their monomeric size of ~95 kDa (Figure 6C and 6D). These data support the prediction that mSHARP$^{1650-2150}$

and mSHARP$^{2050-2550}$ are largely unstructured, and suggest that RRM2-4 is either meta-stable or weakly self-associates.

Discussion

In this study we employed simple biochemical analyses to roughly define the interactions governing the association of the Xist lncRNA with the regulatory effector SHARP. Importantly, we report comprehensive bacterial expression conditions and a purification scheme for constructs covering most of the mSHARP sequence. As recombinant expression of eukaryotic proteins in bacteria is often a major hurdle to *in vitro* characterization, we hope these insights will prove useful for future investigations of SHARP.

Based on our SEC analysis of reconstituted Xist-SHARP complexes, we arrive at two major conclusions. First, the predominant Xist-SHARP interaction arise through protein fragments covering residues 1650 – 2550, a region with no predicted structure or function. In our assay, these SHARP fragments strongly shift the Xist RNA to smaller elution volumes, indicating a stoichiometric complex between the two components. These interactions were reproducible even in the presence of increased ionic strength and detergent. The relatively promiscuous binding of mSHARP$^{1650-2150}$ and mSHARP$^{2050-2550}$ to all of the Xist 5' fragments we tested offer another clue to assembly of Xist-SHARP complexes. As these RNA constructs differ significantly in sequence and secondary structure, this argues against a specific interaction between this portion of SHARP and Xist. Alternatively, these “sticky” unstructured polypeptide segments may act to non-specifically bind SHARP in the Xist cloud surrounding the Xi chromosome. Our SEC data support the notion that these regions are in an extended or coil conformation, but cannot totally rule out the presence of folded domains.

Second, we observed constructs containing the N-terminal RRMs generated unusual complexes with Xist fragments. These RRMs are the only predicted RNA-binding activity within the SHARP domain structure, and the human RRM2-4 have been previously reported to bind other lncRNA sequences *in vitro* [15]. However, our SEC data pointed to large molecular mass assemblies formed between the RRMs and Xist, and our negative stain electron micrographs suggested these are at least partially ordered complexes.

The combination of these two observations supports a model in which multiple copies of SHARP can associate with a single Xist RNA. Binding of the RRMs could initiate the interaction, which is then glued together by unstructured segments of the protein. We speculate that this mode of association may be important for stably localizing SHARP at the Xi chromosome to permanent maintain transcriptional silencing in female somatic cells. Considering that unstructured or low-complexity regions are frequently found in other transcription factor, we anticipate that this type of interaction influence other lncRNA-related processes.

Material and Methods

Cloning, expression, and purification of SHARP proteins

We obtained amino acid sequences for the human (ID: Q96T58) and mouse (ID: Q62504) SHARP proteins from UniProtKB. For each tiled construct, we generated a codon optimized DNA sequence for expression in bacteria, and manually removed common restriction sites. The sequences were synthesized; cloned into the NdeI/NotI sites of pET17b, and sequence verified (GenScript). We then subcloned these genes into a modified version of pET17b which contains a N-terminal hexahistidine tag and maltose binding protein (MBP) moiety followed by a PreScission Protease site and an in-frame KpnI restriction site. The untagged SHARP constructs were PCR amplified and ligated into the KpnI/NotI sites. To determine suitable expression conditions, we screened a variety of factors, including cell line, temperature, and co-expression with chaperones (ArcticExpress system, Agilent). Table 1 summarizes the optimal conditions we identified. Protein identity was confirmed by SDS-PAGE and Western blotting (data not shown).

Construct	Cell Line	Temperature	Co-expression
hsSHARP RRM2-4	Rosetta/pLysS	37 °C	None
mSHARP RRM2-4	Rosetta/pLysS	37 °C	None
mSHARP$^{440-846}$	BL21(DE3)-RIPL	37 °C	None
mSHARP$^{1250-1750}$	Rosetta/pLysS	37 °C	None
mSHARP$^{1650-2150}$	BL21(DE3)-RIPL	18 °C	pArcticExpress
mSHARP$^{2050-2550}$	Rosetta/pLysS	37 °C	None
mSHARP$^{2850-3350}$	BL21(DE3)-RIPL	18 °C	pArcticExpress
mSHARP$^{3250-3644}$	BL21(DE3)-RIPL	37 °C	None

Tabel 1. Optimized expression conditions for SHARP fragments.

All constructs were transformed into the indicated cell type, and co-transformed with the chaperone plasmid as required. Single colonies were grown at 37 °C overnight and diluted 1:100 into 2L day cultures in LB media. Growth was continued at 37 °C until cultures reached OD_{600nm} = 0.6, at which point the cultures were transitioned to the expression temperature shown in Table 1 and induced with 1 mM IPTG. Cultures at 37 °C were expressed for 4 hrs and harvested. Low temperature expressions were performed overnight (14 - 18 hrs).

For purification, pellets were resuspended in 40 mL Buffer A (0.5 M NaCl, 20 mM EPPS pH8.0, 1 mM EDTA, 1 mM DTT) supplemented with 1 mM PMSF and sonicated. The lysates were clarified by centrifugation and the supernatants loaded on 5 mL Amylose resin (NEB) packed in an XK 16 column (GE Healthcare). The column was washed extensively with Buffer A, and eluted with Buffer A supplemented with 10 mM D-(+)-maltose. The elution was concentrated in an Amicon 15 centrifugal filter device (30 kDa MWCO) to a final volume of 0.5 mL. The sample was further purified over a Superdex 10/300 GL S200 SEC column (GE Healthcare), using a running buffer of 0.4 M NaCl, 20 mM EPPS pH8.0, 1 mM EDTA, 1 mM DTT. The peak fractions were analyzed by SDS-PAGE, pooled, and concentrated as above. The absence of nucleic acid contamination and the protein concentration was determined by absorption spectroscopy.

RNA transcription and purification

Transcription templates for each RNA construct were generated by PCR using primers bracketing the indicated regions. The forward primer was designed to contain the T7 promoter sequence followed by a GG dinucleotide and then the first Xist base-pair. The cloned mouse Xist sequence was used as template. Large scale PCR reactions (3.2 mL total volume, divided into 100 µL aliquots across 96-well plates) were performed with Vent polymerase (NEB) according to the manufacture's recommended protocol. The PCR product was ethanol precipitated and resuspended in water. For the Xist$^{334-725}$, the DNA template was further purified over a 6% polyacrylamide native gel to obtain a single band.). Transcription reactions contained the PCR-amplified template, 40 mM Tris pH7.5, 25 mM $MgCl_2$, 4 mM DTT, 2 mM spermidine, 40 µg inorganic pyrophosphatase (Sigma), 0.7 mg T7 RNA polymerase, and 3 mM each rNTP in a total volume of 5 mL. For Xist$^{334-725}$ the reaction volume was doubled to 10 mL. Reactions were incubated at 37 °C for 4 hrs, except the Xist$^{334-725}$ reaction was performed at 42 °C for 2 hrs. The reactions were ethanol precipitated and purified over denaturing 10% polyacrylamide gels. The desired band was excised under UV shadowing and the RNA eluted from the gel slice. The RNA was concentrated and washed with 10 mM HEPES pH7.0 in an Amicon device.

Preparation of Xist-SHARP complexes, EM, and protein mass estimates

For binding reactions, the RNA was annealing in a solution containing 150 mM NaCl, 5 mM HEPES pH7.0 by heating to 65 °C for 3 min and snap cooling on ice. The 100 µL binding reactions contained 5 µM RNA, 150 mM NaCl, 20 mM Tris pH8.0, and a variable amount of protein. Following a 20 min incubation at room temperature, the reactions were analyzed on the S200 column using either of the running buffers described in the Result section. For denaturing gel analysis, 0.5 mL fractions were collected from each run. For Xist$^{727\text{-}993}$ binding with mSHARP$^{440\text{-}846}$, the peak fraction from the 1:8 run was spotted on a holey carbon grid, blotted, stained with uranyl acetate, and imagine on a TF20 electron microscope (FEI). For calibration of the SEC column, we used the same running buffer as for purification of the proteins, and analyzed the Bio-Rad gel filtration standard. Proteins were re-run on the column to accurately determine their elution volumes.

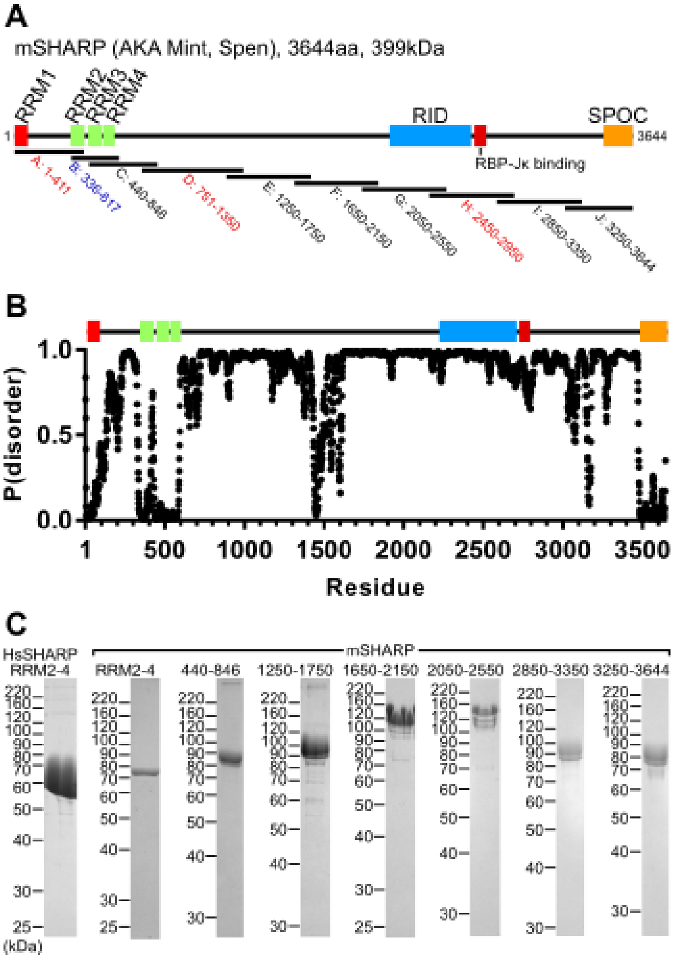


Figure 1. Organization of the mSHARP protein. (A) Domain structure of mSHARP, including the N-terminal RRMs, central nuclear receptor interaction site (RID), and C-terminal SPOC domain. Expression constructs used for recombinant production in bacteria appear below. We were unable to achieve bacterial expression for three constructs (red). The fragment covering the known crystal structure of RRM2-4 is highlighted in blue. (B) Structural disorder prediction for mSHARP from DISOPRED2 webserver [21]. Vertical axis reports the likelihood of each residue to occupy an unstructured state. (C) SDS-PAGE analysis of purified SHARP constructs.

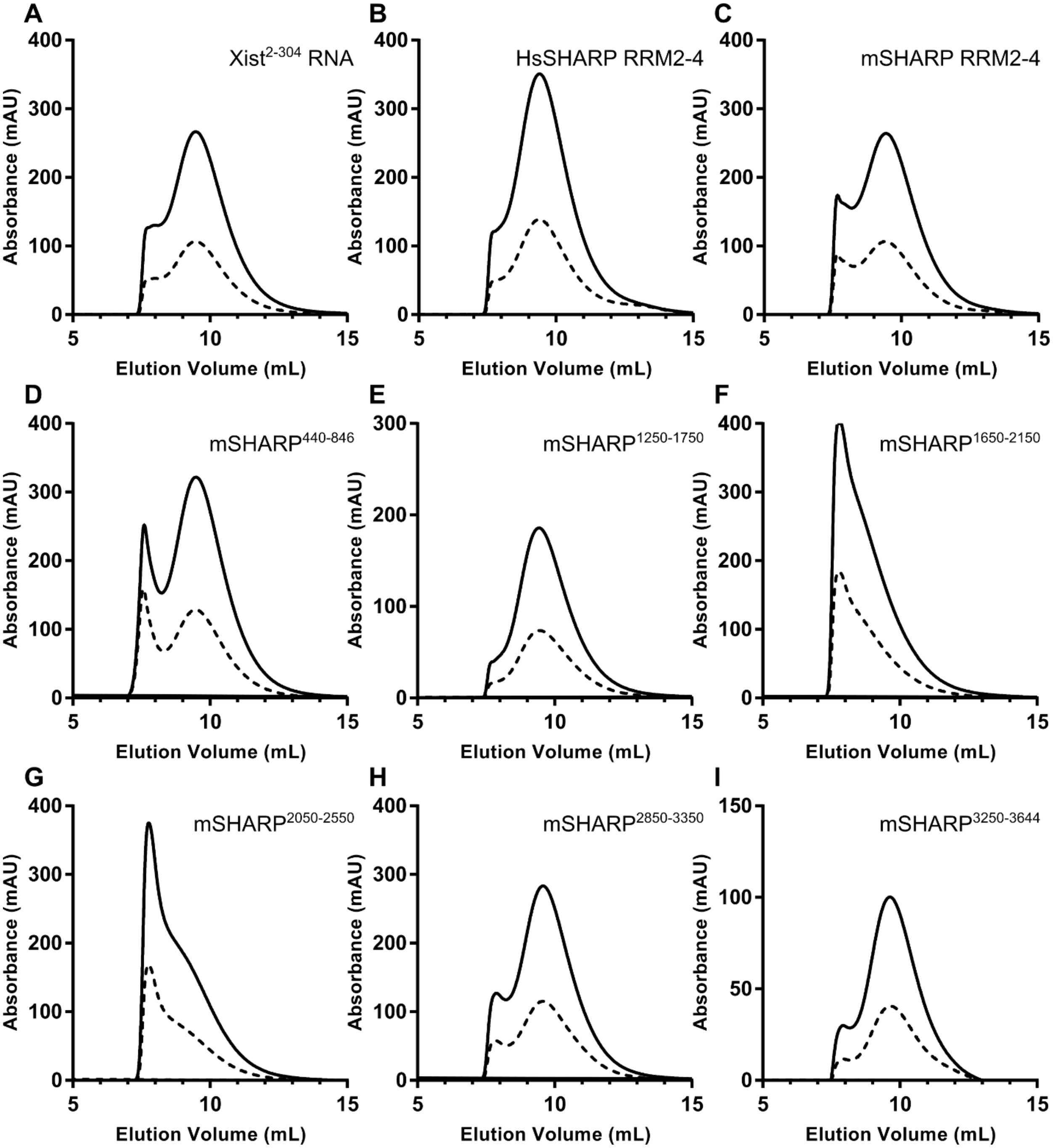


Figure 2. SEC analysis of complexes between Xist$^{2-304}$ and SHARP fragments under low salt conditions (80mM NaCl). Throughout the chapter, the solid SEC curve represents the 260 nm signal and the dashed curve is the 280 nm channel.

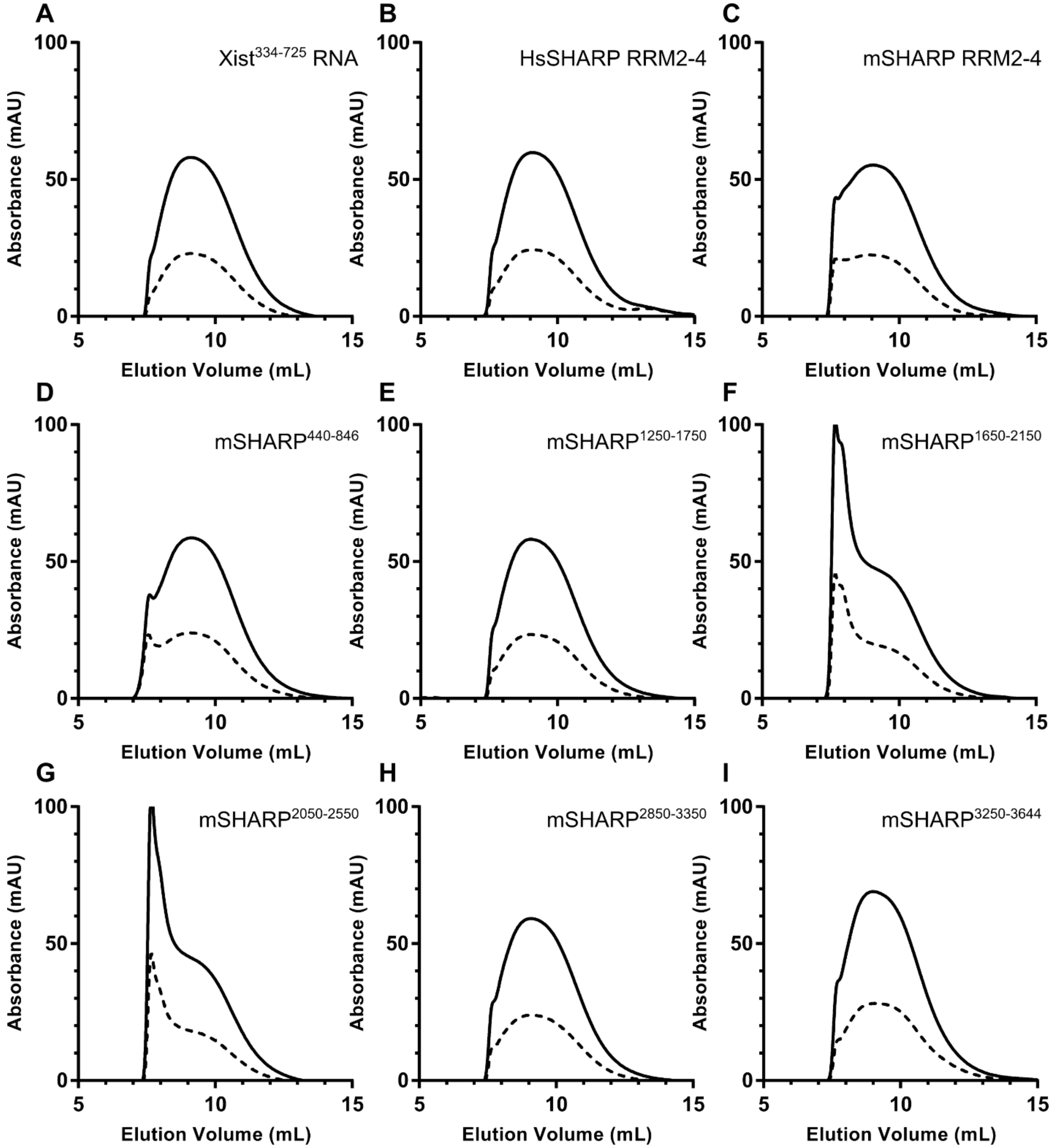


Figure 3. Complexes between Xist$^{334-725}$ and SHARP fragments. These experiments utilize the same low salt (80 mM NaCl) running buffer used in Figure 2.

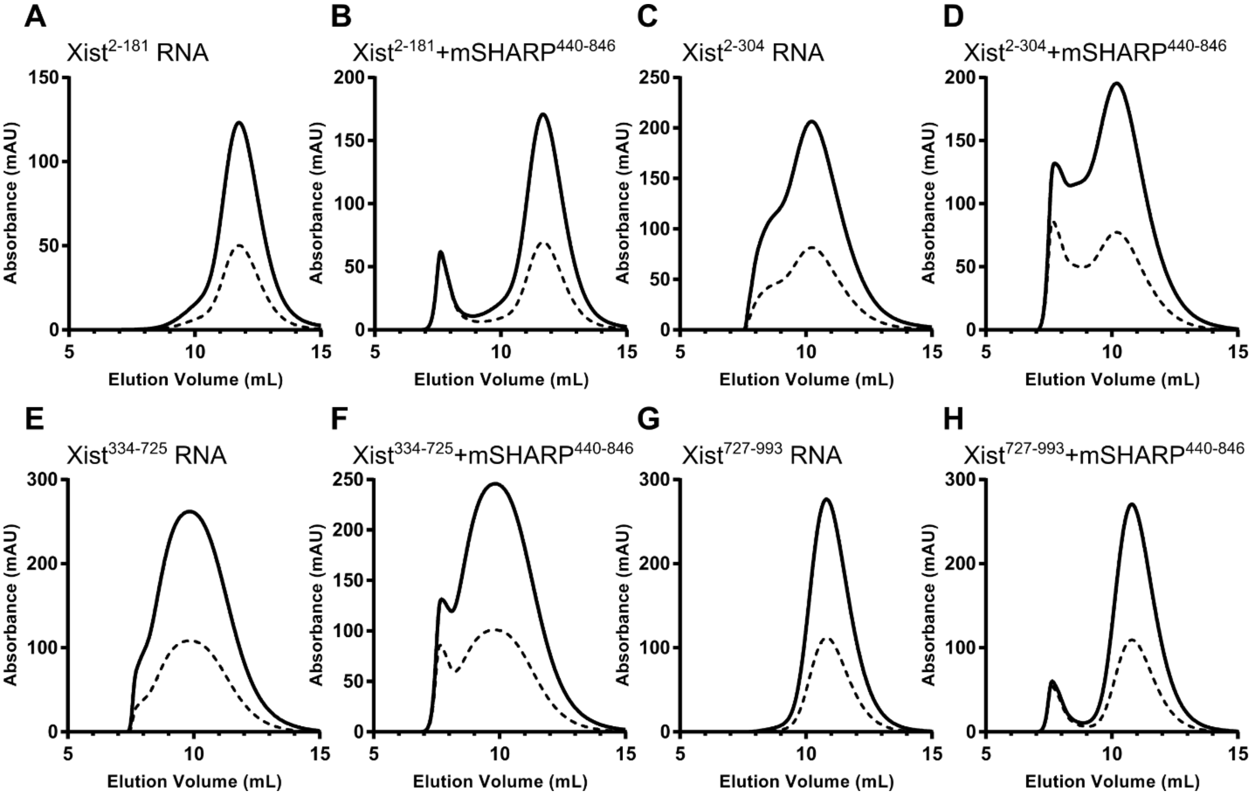


Figure 4. Re-analysis of Xist-SHARP complexes with intermediate salt concentration (150 mM NaCl) and detergent (0.00145% Triton X-100).

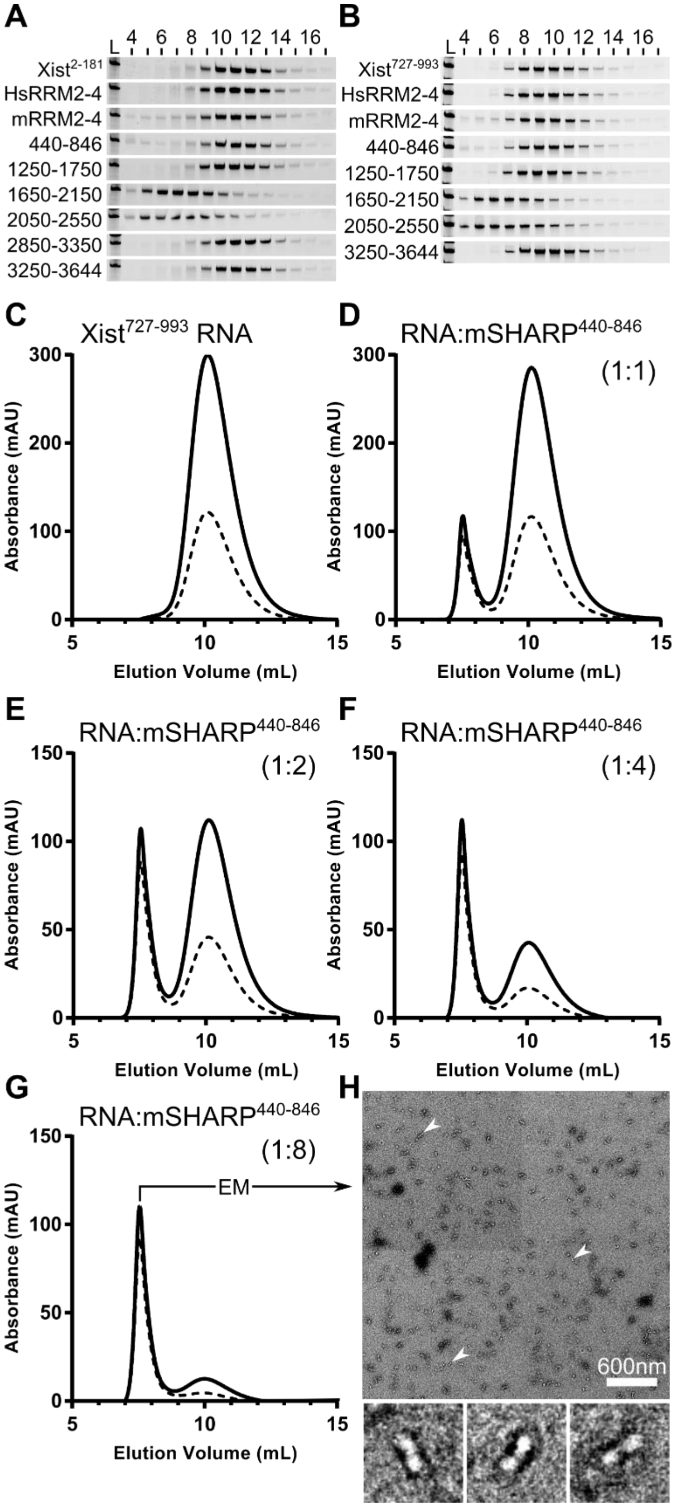


Figure 5. Confirmation of RNA-protein complex assembly and estimation of the complex stoichiometry. (A) Denaturing polyacrylamide gel analysis of SEC fractions for complexes with Xist$^{2-181}$. Fraction 4 captures the void volume of the column and subsequent fractions are later elution times. Each gel slice

shows the 200 nt band of the ladder (L). (B) Denaturing gels of fractions for Xist$^{727-993}$ complexes. The ladder bands are the same 200 nt fragment as in (A). (C) Elution profile of isolated Xist$^{727-993}$. (D) – (G): Titration of Xist727-993 with mSHARP$^{440-846}$, with the approximate molar ratio shown in parenthesis. (H) Negative-stain electron micrograph (11500X magnification) of material from peak fraction in (G). Close-up views show particles labeled with arrows.

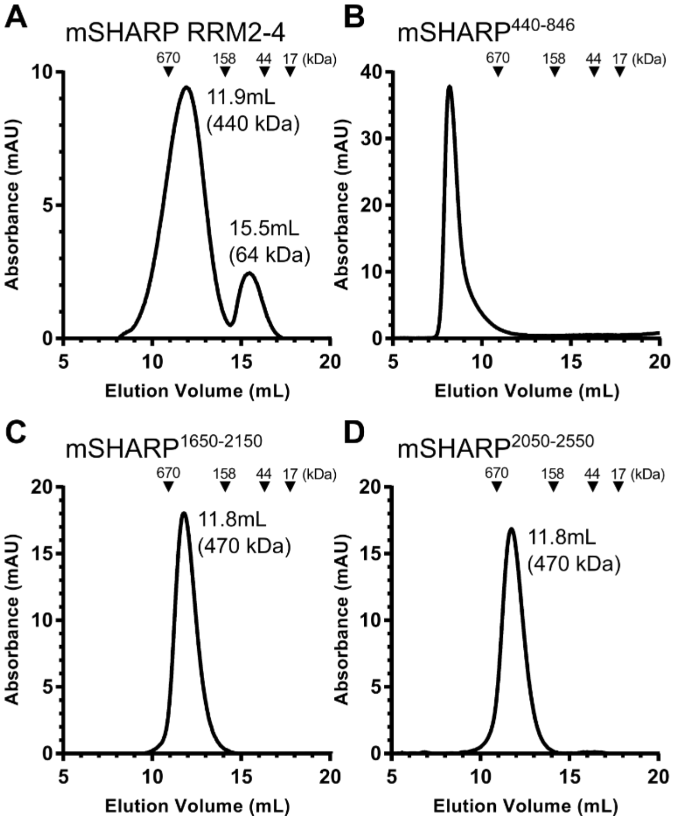


Figure 6. Molecular mass estimates for mSHARP proteins based on calibration of the SEC column. Elution volumes for proteins of known mass are shown with a triangle. The approximate mass of peak is labeled, except for (B) where the protein elutes too near the void volume.

References

[1] F. Kopp and J. T. Mendell, “Functional Classification and Experimental Dissection of Long Noncoding RNAs.,” *Cell*, vol. 172, no. 3, pp. 393–407, Jan. 2018.

[2] C. A. McHugh, C.-K. Chen, A. Chow, C. F. Surka, C. Tran, P. McDonel, A. Pandya-Jones, M. Blanco, C. Burghard, A. Moradian, M. J. Sweredoski, A. A. Shishkin, J. Su, E. S. Lander, S. Hess, K. Plath, and M. Guttman, “The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3,” *Nature*, vol. 521, no. 7551, pp. 232–236, May 2015.

[3] C. Chu, Q. C. Zhang, S. T. da Rocha, R. A. Flynn, M. Bharadwaj, J. M. Calabrese, T. Magnuson, E. Heard, and H. Y. Chang, “Systematic discovery of Xist RNA binding proteins.,” *Cell*, vol. 161, no. 2, pp. 404–16, Apr. 2015.

[4] A. Minajigi, J. Froberg, C. Wei, H. Sunwoo, B. Kesner, D. Colognori, D. Lessing, B. Payer, M. Boukhali, W. Haas, and J. T. Lee, “Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation.,” *Science*, vol. 349, no. 6245, p. aab2276, Jul. 2015.

[5] Y. Hasegawa, N. Brockdorff, S. Kawano, K. Tsutui, K. Tsutui, and S. Nakagawa, “The matrix protein hnRNP U is required for chromosomal localization of Xist RNA.,” *Dev. Cell*, vol. 19, no. 3, pp. 469–76, Sep. 2010.

[6] G. Pintacuda, A. N. Young, and A. Cerase, “Function by Structure: Spotlights on Xist Long Non-coding RNA,” *Front. Mol. Biosci.*, vol. 4, p. 90, Dec. 2017.

[7] C.-K. Chen, M. Blanco, C. Jackson, E. Aznauryan, N. Ollikainen, C. Surka, A. Chow, A. Cerase, P. McDonel, and M. Guttman, “Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing.,” *Science*, vol. 354, no. 6311, pp. 468–472, Oct. 2016.

[8] A. Monfort, G. Di Minin, A. Postlmayr, R. Freimann, F. Arieti, S. Thore, and A. Wutz, “Identification of Spen as a Crucial Factor for Xist Function through Forward Genetic Screening in Haploid Embryonic Stem Cells,” *Cell Rep.*, vol. 12, no. 4, pp. 554–561, Jul. 2015.

[9] B. Moindrot, A. Cerase, H. Coker, O. Masui, A. Grijzenhout, G. Pintacuda, L. Schermelleh, T. B. Nesterova, and N. Brockdorff, “A Pooled shRNA Screen Identifies Rbm15, Spen, and Wtap as Factors Required for Xist RNA-Mediated Silencing,” *Cell Rep.*, vol. 12, no. 4, pp. 562–572, Jul. 2015.

[10] B. J. Dickson, A. van der Straten, M. Domínguez, and E. Hafen, “Mutations Modulating Raf Signaling in Drosophila Eye Development,” *Genetics*, vol. 142, no. 1, 1996.

[11] Elizabeth P. Newberry, and Tammy Latifi, and D. A. Towler*, “The RRM Domain of MINT, a Novel Msx2 Binding Protein, Recognizes and Regulates the Rat Osteocalcin Promoter†,” 1999.

[12] Y. Shi, M. Downes, W. Xie, H. Y. Kao, P. Ordentlich, C. C. Tsai, M. Hon, and R. M. Evans, “Sharp, an inducible cofactor that integrates nuclear receptor repression and activation.,” *Genes Dev.*, vol. 15, no. 9, pp. 1140–51, May 2001.

[13] S.-H. You, H.-W. Lim, Z. Sun, M. Broache, K.-J. Won, and M. A. Lazar, “Nuclear receptor co-repressors are required for the histone-deacetylase activity of HDAC3 in vivo,” *Nat. Struct. Mol. Biol.*, vol. 20, no. 2, pp. 182–187, Feb. 2013.

[14] M. Ariyoshi and J. W. R. Schwabe, “A conserved structural motif reveals the essential transcriptional repression function of Spen proteins and their role in developmental signaling.,” *Genes Dev.*, vol. 17, no. 15, pp. 1909–20, Aug. 2003.

[15] F. Arieti, C. Gabus, M. Tambalo, T. Huet, A. Round, and S. Thore, “The crystal structure of the Split End protein SHARP adds a new layer of complexity to proteins containing RNA recognition motifs,” *Nucleic Acids Res.*, vol. 42, no. 10, pp. 6742–6752, Jun. 2014.

[16] A. Wutz, T. P. Rasmussen, and R. Jaenisch, “Chromosomal silencing and localization are mediated by different domains of Xist RNA,” *Nat. Genet.*, vol. 30, no. 2, pp. 167–174, Feb. 2002.

[17] R. Fang, W. N. Moss, M. Rutenberg-Schoenberg, and M. D. Simon, “Probing Xist RNA Structure in Cells Using Targeted Structure-Seq,” *PLOS Genet.*, vol. 11, no. 12, p. e1005668, Dec. 2015.

[18] N. Liu, K. I. Zhou, M. Parisien, Q. Dai, L. Diatchenko, and T. Pan, “N 6-methyladenosine alters RNA structure to regulate binding of a low-complexity protein,” *Nucleic Acids Res.*, vol. 45, no. 10, pp. 6051–6063, Jun. 2017.

[19] Z. Lu, Q. C. Zhang, B. Lee, R. A. Flynn, M. A. Smith, J. T. Robinson, C. Davidovich, A. R. Gooding, K. J. Goodrich, J. S. Mattick, J. P. Mesirov, T. R. Cech, and H. Y. Chang, “RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure,” *Cell*, vol. 165, no. 5, pp. 1267–1279, May 2016.

[20] J. Quick-Cleveland, J. Jacob, S. Weitz, G. Shoffner, R. Senturia, and F. Guo, “The DGCR8 RNA-Binding Heme Domain Recognizes Primary MicroRNAs by Clamping the Hairpin,” *Cell Rep.*, vol. 7, no. 6, 2014.

[21] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, “Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life,” *J. Mol. Biol.*, vol. 337, no. 3, pp. 635–645, Mar. 2004.