UNIVERSITY OF CALIFORNIA

SANTA CRUZ

A-U, G-C; HOW COMPLICATED CAN IT BE? PROTEIN-RNA INTERACTIONS IN TELOMERASE

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

CHEMISTRY

by

Christina Palka

September 2019

The Dissertation of Christina Palka is approved:

_____
Professor Michael Stone, chair

_____
Professor Seth Rubin

_____
Professor Melissa Jurica

_____
Quentin Williams
Acting Vice Provost and Dean of Graduate Studies

Table of Contents

List of figures ..... **v**

Abstract ..... **vi**

Acknowledgements ..... **ix**

Chapter 1: Introduction to Telomerase Biology **1**

References 7

Chapter 2: **11**

A metastable junction in human telomerase RNA is remodeled during enzyme assembly **11**

Abstract: 11

Introduction 12

Results 16

Discussion 29

Materials and Methods: 31

References: 35

Chapter 3: Chemical mapping beyond P4P6 ..... **39**

Introduction 39

Chapter 3.1 - Chemical mapping experimental protocol 43

Chapter 3.2 - Analysis of chemical mapping data using HiTrace 47

Chapter 3.3 - Mutate-and-Map (M2) 68

Chapter 3.4 - Mutate-Map-Rescue (M2R) 79

References 87

Chapter 4: Re-evaluation of the RNA binding properties of the *Tetrahymena thermophila* telomerase reverse transcriptase N-terminal domain 89

Introduction 89

Results 93

Methods 101

References 104

Chapter 5 - Unfinished Projects 109

5.1 - CR4/5 Disease Mutants 109

5.2 - Yeast CR4/5 Chemical Probing 111

5.3 - DNA:RNA Duplex Handling in TERT 113

References: 117

List of figures

Chapter 1:

Figure 1 1 - Telomeres cap the ends of chromosomes. 1

Figure 1 2 - Telomerase is processive. 2

Figure 1 3 - Conserved domain architecture of TR and TERT. 4

Chapter 2:

Figure 2 1 - Conserved protein and RNA domains of the telomerase catalytic core. 13

Figure 2 2 - Chemical mapping of medaka and human CR4/5 domain. 18

Figure 2 3 - Data driven RNA secondary structure prediction of medaka and human CR4/5 domain. 21

Figure 2 4 - Mutation profile of the human CR4/5 domain. 23

Figure 2 5 - Structurally engineered CR4/5 domains exhibit assembly and functional defects. 27

Figure 2 6 - Confocal microscopy reveals structural heterogeneity in the CR4/5 domain and stabilization of a single state when bound to TERT. 29

Chapter 3:

Figure 3 1 - RNA exists in a complex folding landscape. 39

Figure 3 2 - Chemicals are used to modify RNA. 41

Figure 3 3 - RNA construct design and immobilization scheme. 43

Figure 3 4 - High purity of RNA is required for clean RT extension. 45

Figure 3 5 - Visual evaluation of high quality and low quality data using quick_look. 53

Figure 3 6 - Examples of mis-aligned bands viewed during band assignment. 56

Figure 3 7 - Band assignment in HiTrace. 58

Figure 3 8 - Visualization of output of rna_structure and VARNA. 64
Figure 3 9 - Reactivity magnitude affects structure predicted in heterogeneous data sets. 67
Figure 3 10 - Point mutations can cause unexpected structural changes due to RNA's complex folding landscape. 68
Figure 3 11 - Mutation to RNA gives rise to different structural perturbations. 74
Figure 3 12 - Mis-alignment during band assignment step results in missed release event in the rdat. 77
Figure 3 13 - Toy example of the visualization and calculation of a rescue factor from single mutants and double rescues. 81
Figure 3 14 - Only low and medium rescue factors were calculated for single mutant, double mutant rescues for the human CR4/5 domain. 82
Figure 3 15 - Double mutants and quadruple compensatory rescue supports the formation of P6a, P6b, and P5. 84

Chapter 4:

Figure 4 1 - Conserved telomerase subunits from the ciliate Tetrahymena thermophila. 90
Figure 4 2 - Purification and EMSAs testing GST fusion to the TEN domain. 94
Figure 4 3 - Purification and EMSAs testing MBP fusion to the TEN domain. 96
Figure 4 4 - Comparison of RNAs for binding to the TEN domain. 98

Chapter 5:

Figure 5 1 - Structure probing of human CR4/5 disease causing mutants. 110

Figure 5 2 - Structure probing of K. lactis CR4/5. 112

Figure 5 3 - DNA:RNA handling mutants immobilize for FRET analysis. 115

Abstract

Christina Palka

A-U, G-C; How complicated can it be? Protein-RNA interactions in telomerase

The ability for RNA to adopt multiple structures from one sequence presents challenges when predicting RNA secondary structure. Most RNA in the cell is bound to protein, adding additional layers of complexity to the puzzle. Protein-RNA contacts are studied in this body of work within the context of telomerase, a ribonucleoprotein that adds telomeric DNA to the ends of chromosomes allowing cells preserve genomic integrity and prevent the DNA damage. Telomerase is composed of a protein and RNA component, each composed of multiple conserved domains. The contacts formed between these two units are important for understanding telomerase assembly and function. In the first study it is shown that a functionally critical domain within human telomerase RNA is structurally heterogeneous. This heterogeneity is eliminated with telomerase RNA is bound to the telomerase protein. The role of structural heterogeneity in a step-wise assembly pathway of telomerase is discussed. In the second study a previously characterized protein-RNA interaction in *Tetrahymena thermophila* is demonstrated to be due to a protein contaminant rather than a true telomerase protein-telomerase RNA interaction. This allows for the re-evaluation of previously proposed models for which protein-RNA interactions are important for telomerase function. Taken together these two studies shed light on the different protein-RNA interactions important for telomerase assembly and function.

Acknowledgements

I would like to thank my family, friends, and wonderful partner Rafael for their unconditional support and love through graduate school. I'd also like to thank Pacific Edge and Touchstone climbing gyms, where I holed up to write this thesis and Hans Zimmer for his unending soundtracks that powered each and every word in the following manuscript. Finally thank you to the Sierra mountains for reminding me of my own mortality and putting those failed experiments back into proper perspective.

Chapter 1: Introduction to Telomerase Biology

The integrity of genetic material is fundamental to every living organism. Prokaryotes store their DNA in a circular genome, while eukaryotes have linear chromosomes. The development of linear chromosomes within eukaryotes resulted in unique challenges pertaining to the maintenance of genomic stability. Due to the necessity for an RNA primer to anneal and 'prime' DNA polymerase, the ends of the chromosomes cannot be fully replicated and are shortened through each replication cycle [1]. This shortening is compounded by endogenous nuclease activity [2].

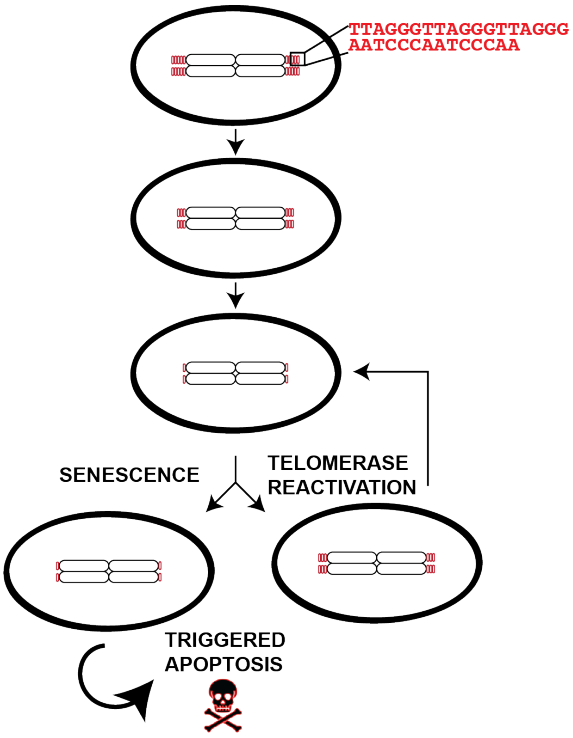


Figure 1 1 - Telomeres cap the ends of chromosomes.
Repetitive telomeric DNA (red) sequences flank the ends of chromosomes and are degraded with cell division.

Further, exposed chromosome ends illicit DNA damage response signals resulting in end-to-end chromosomal fusion [3] [4].

Telomeres, which are composed of a repetitive DNA sequence that is packaged into chromatin with telomere specific binding proteins, protect chromosomes from aberrant fusion [5] (**Figure 1.1A**). After many replications telomeres reach a critical shortness and senescence is triggered [6] [7], to protect the coding region of the genome (**Figure 1.1A**). Telomeres and

their maintenance are particularly crucial in cells that continually divide. The biomolecule responsible for maintaining telomere homeostasis is a uniquely adapted reverse transcriptase named telomerase [8]. Mutations in telomerase genes and promoter regions result in abnormal shortening of telomeres and severe developmental defects [9] [10] [11]. After development, telomerase is expressed in germline cells [12] [13], lymphocytes [14] and some stem cells [15], but its expression is repressed in most somatic cells. Given its importance in continually dividing cells telomerase activity is critical to cancer cells, and indeed inappropriate telomerase re-activation is seen in 90% of cancers [16]. The ability to modulate telomerase activity could serve as a therapeutic for both re-extension of telomeres in developmental diseases and forced senescence in cancers. Efforts towards this goal are hindered by an incomplete understanding of the telomerase assembly pathway and the mechanism by which telomerase maintains telomere length.

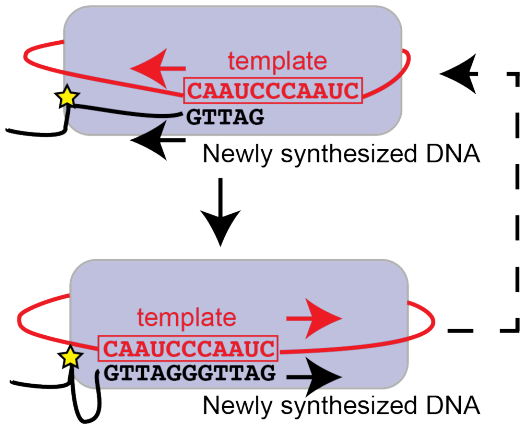


Figure 1 2 - Telomerase is processive.
Coordinated movement between the template (red), telomeric DNA (black) and TERT (purple square) is required for RAP. Contact between telomeric DNA and TERT is maintained through an anchor site distal to the active site (star).

Telomerase is made of an integral protein and RNA component. Together these domains contain the active site and template responsible for the de novo addition of telomeric repeats on the ends of chromosomes. A specialized characteristic of telomerase is its ability to add multiple telomeric repeats in a single telomere binding event in a process called Repeat Addition Processivity (RAP) (**Figure 1.2**). For this

to occur at the end of the template the DNA:RNA hybrid must un-anneal, the new 3' end of the telomere and the template must reposition in the active site, while maintaining contact with telomeric DNA to avoid dissociation. Many open questions remain regarding this mechanism including the source of the anchor point between the telomeric DNA and TERT, the length of the DNA:RNA hybrid in the active site, and the source of mechanistic motion responsible for telomere template repositioning, to name a few.

Broadly, the topic of this thesis is the study of the protein-RNA interactions in telomerase and their role in the coordination of telomerase assembly and the catalysis of telomeric repeats. Telomerase RNA (TR) is composed of the functionally conserved Template-Pseudoknot (t/PK) domain, a distal Stem Terminal Element (STE) and in vertebrates an scRNA H/ACA domain (**Figure 1.3A and B**). The catalytic telomerase reverse transcriptase (TERT) protein is composed of four conserved domains: Telomerase Essential N-terminal (TEN) domain, the telomerase RNA binding domain (TRBD), the Reverse Transcriptase (RT) domain, and the C-terminal Extension (CTE) (**Figure 1.3C**).

Recently, an intermediate resolution cryo electron microscopy (cryo EM) structure of telomerase was published from the model organism *Tetrahymena thermophila* (hereafter referred to as *Tetrahymena*) [17]. The reconstruction allowed atomic modeling of the fully assembled Tetrahymena telomerase holoenzyme with high confidence. An intricate set of protein-RNA interactions were shown, supporting decades of prior biochemical work on important regions of the protein and RNA components. Specifically, the RBD and CTE domain both form a high affinity interaction with the loop of Stem Loop IV [18] [19] (**Figure 1.3** - Blue Star). RBD forms

an additional high affinity interaction with the base of Stem Loop II (**Figure 1.3** - Yellow Star) providing a mechanism for how template boundary definition is achieved during telomere catalysis [20] [21] [22]. A single stranded region of RNA 3' of the template region is sandwiched between an insertion of fingers domain (IFD) in the RT and the CTE (**Figure 1.3** - Pink Star). The TEN domain is positioned above the IFD and is attached to the RBD domain by a flexible, unstructured linker. The published structure does not reveal new insights into proposed TEN-nucleic acid interactions. The PK domain contains the template for telomeric DNA, which must be placed in the RT catalytic active site. Yet the PK fold itself is positioned on the opposite side of the active site and only contains a single point of protein contact with the CTD domain [17] (**Figure 1.3** - Green Star).

While the holoenzyme structure illustrates the end point for telomerase assembly, open questions remain as to how the protein and RNA subunits assemble and how those contacts contribute to telomere catalysis. Telomerase assembly and elucidation of the mechanistic details regarding the telomerase catalytic cycle remains an exciting field of research.

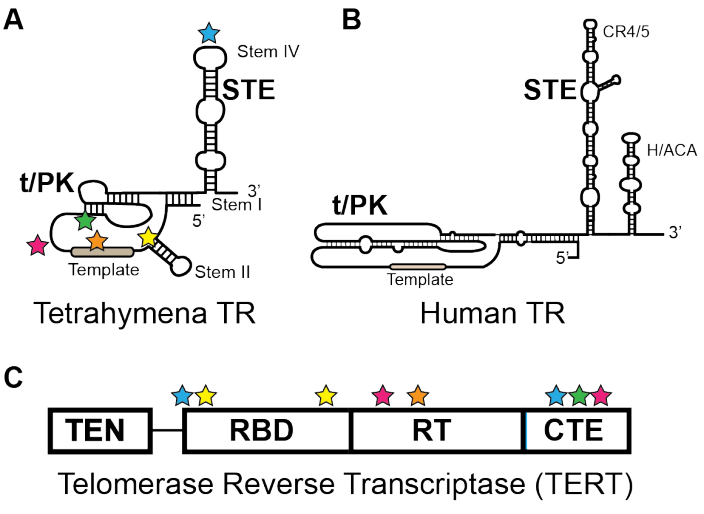


Figure 1 3 - Conserved domain architecture of TR and TERT.

Conserved domains (bold) and specialized names of (A) Tetrahymena TR (B) Human TR (C) TERT. Regions of protein-RNA contacts based on the Tetrahymena cryo EM structural model are approximated on the TERT and TR schematics.

Concurrent with the Tetrahymena structure, a low resolution cryo EM reconstruction of the human telomerase holoenzyme was published [23]. While domain architecture is conserved between *Tetrahymena* and human TERT, a dramatic divergence in TR is observed in both sequence and length making direct comparison between RNAs difficult (**Figure 1.3A and B**). Although the resolution was insufficient to distinguish molecular protein-RNA interactions, the structure revealed the overall human telomerase architecture. Interestingly, the cryo EM structure supports crystallographic studies on fish TRBD-CR4/5 [24] complexes that show extensive contacts between the TR CR4/5 domain and TERT near the core of the enzyme, rationalizing the importance of the distal, yet critical CR4/5 domain. Similar to the Tetrahymena system surprisingly few protein-RNA contacts appear to be formed between the PK and TERT. These findings are supported by data suggesting the PK is folded correctly in the absence of TERT [25] [26].

Both Tetrahymena and human systems are utilized to dissect the role of protein:RNA interactions in telomerase. In Chapter 1 a debated question regarding a potential TR-TEN domain interaction in *Tetrahymena* telomerase is revisited. The TEN domain is known to be important in the telomerase catalytic cycle, but contradictory results have been published regarding the sequence and structure specificity of its interaction with TR [27] [28]. Our work suggests that the TEN domain has very low affinity for TR without noticeable sequence or structure specificity [29]. Previous findings regarding sequence/structure specificity, as published by the Collins lab, were due to a protein contaminant still present after Nickel-affinity purification and resulted in a formal correction. Questions remain as to the identity and functional importance of the TEN-domain binding site within TR and the TEN domains role in RAP.

Chapter 2 is a technical description of chemical mapping, a technique used to probe RNA secondary structure. Chemical mapping is a tool that has been utilized for decades but recent advances have allowed it to become a high throughput tool. After a detailed analysis of a chemical mapping data platform developed by the Das lab I discuss the strengths and weaknesses of the platform as it stands. Applications are discussed in the context of better analysis the potential unwanted structural changes in mutagenesis-function type experiments and the role these large data sets will play in deconvoluting the mysteries of RNA folding.

The technique and analytical pipeline in Chapter 2 are applied in Chapter 3 where I'll present the discovery that structural heterogeneity exists in the CR4/5 domain of human telomerase RNA. This heterogeneity is important for telomerase assembly and function and upon TERT binding to the CR4/5 domain a single RNA structure is stabilized. This may speak to an initial step in TERT-TR assembly, potentially demonstrating a unified requirement for RNA structural change as a step in telomerase assembly across eukaryotes.

Chapter 4 presents incomplete projects. Projects included in this are the chemical mapping of a yeast CR4/5 domain and the potential for FRET experiments to be performed across multiple species. Also discussed are a set of mutations in TERT proposed to affect DNA:RNA primer handling. Preliminary functional analysis and FRET results are shown. Finally, the chemical mapping results for a set of disease mutants in the CR4/5 domain are illustrated and potential FRET experiments are outlined.

References

1. Olovnikov, A.M., *[Principle of marginotomy in template synthesis of polynucleotides].* Dokl Akad Nauk SSSR, 1971. **201**(6): p. 1496-9.
2. Lingner, J. and T.R. Cech, *Purification of telomerase from Euplotes aediculatus: requirement of a primer 3' overhang.* Proc Natl Acad Sci U S A, 1996. **93**(20): p. 10712-7.
3. Muller, H.J., *The remaking of chromosomes.* The Collecting Net, 1938. **13**: p. 181-195.
4. McClintock, B., *The behavior in successive nuclear divisions of a chromosome broken at meiosis.* Proc Natl Acad Sci U S A, 1939. **25**: p. 405-416.
5. Blackburn, E.H. and J.G. Gall, *A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena.* J Mol Biol, 1978. **120**(1): p. 33-53.
6. Hayflick, L., *The Limited in Vitro Lifetime of Human Diploid Cell Strains.* Exp Cell Res, 1965. **37**: p. 614-36.
7. Harley, C.B., A.B. Futcher, and C.W. Greider, *Telomeres shorten during ageing of human fibroblasts.* Nature, 1990. **345**(6274): p. 458-60.
8. Greider, C.W. and E.H. Blackburn, *Identification of a specific telomere terminal transferase activity in Tetrahymena extracts.* Cell, 1985. **43**(2 Pt 1): p. 405-13.

9. Yamaguchi, H., et al., *Mutations of the human telomerase RNA gene (TERC) in aplastic anemia and myelodysplastic syndrome.* Blood, 2003. **102**(3): p. 916-8.

10. Vulliamy, T.J. and I. Dokal, *Dyskeratosis congenita: the diverse clinical presentation of mutations in the telomerase complex.* Biochimie, 2008. **90**(1): p. 122-30.

11. Savage, S.A., *Human telomeres and telomere biology disorders.* Prog Mol Biol Transl Sci, 2014. **125**: p. 41-66.

12. Kolquist, K.A., et al., *Expression of TERT in early premalignant lesions and a subset of cells in normal tissues.* Nat Genet, 1998. **19**(2): p. 182-6.

13. Wright, D.L., et al., *Characterization of telomerase activity in the human oocyte and preimplantation embryo.* Mol Hum Reprod, 2001. **7**(10): p. 947-55.

14. Roth, A., et al., *Telomerase levels control the lifespan of human T lymphocytes.* Blood, 2003. **102**(3): p. 849-57.

15. Blasco, M.A., *Telomeres and human disease: ageing, cancer and beyond.* Nat Rev Genet, 2005. **6**(8): p. 611-22.

16. Kim, N.W., et al., *Specific association of human telomerase activity with immortal cells and cancer.* Science, 1994. **266**(5193): p. 2011-5.

17. Jiang, J., et al., *Structure of Telomerase with Telomeric DNA.* Cell, 2018. **173**(5): p. 1179-1190 e13.

18. Berman, A.J., A.R. Gooding, and T.R. Cech, *Tetrahymena telomerase protein p65 induces conformational changes throughout telomerase RNA (TER) and rescues telomerase reverse transcriptase and TER assembly mutants.* Mol Cell Biol, 2010. **30**(20): p. 4965-76.

19. Robart, A.R., C.M. O'Connor, and K. Collins, *Ciliate telomerase RNA loop IV nucleotides promote hierarchical RNP assembly and holoenzyme stability.* RNA, 2010. **16**(3): p. 563-71.

20. Akiyama, B.M., A. Gomez, and M.D. Stone, *A conserved motif in Tetrahymena thermophila telomerase reverse transcriptase is proximal to the RNA template and is essential for boundary definition.* J Biol Chem, 2013. **288**(30): p. 22141-9.

21. Jansson, L.I., et al., *Structural basis of template-boundary definition in Tetrahymena telomerase.* Nat Struct Mol Biol, 2015. **22**(11): p. 883-8.

22. Bley, C.J., et al., *RNA-protein binding interface in the telomerase ribonucleoprotein.* Proc Natl Acad Sci U S A, 2011. **108**(51): p. 20333-8.

23. Nguyen, T.H.D., et al., *Cryo-EM structure of substrate-bound human telomerase holoenzyme.* Nature, 2018. **557**(7704): p. 190-195.

24. Huang, J., et al., *Structural basis for protein-RNA recognition in telomerase.* Nat Struct Mol Biol, 2014. **21**(6): p. 507-12.

25. Hengesbach, M., et al., *Single-molecule FRET reveals the folding dynamics of the human telomerase RNA pseudoknot domain.* Angew Chem Int Ed Engl, 2012. **51**(24): p. 5876-9.

26. Parks, J.W., et al., *Single-molecule FRET-Rosetta reveals RNA structural rearrangements during human telomerase catalysis.* RNA, 2017. **23**(2): p. 175-188.

27. O'Connor, C.M., C.K. Lai, and K. Collins, *Two purified domains of telomerase reverse transcriptase reconstitute sequence-specific interactions with RNA.* J Biol Chem, 2005. **280**(17): p. 17533-9.

28. Jacobs, S.A., E.R. Podell, and T.R. Cech, *Crystal structure of the essential N-terminal domain of telomerase reverse transcriptase.* Nat Struct Mol Biol, 2006. **13**(3): p. 218-25.

29. Palka, C.P., et al., *Reevaluation of the RNA binding properties of the Tetrahymena thermophila telomerase reverse transcriptase N-terminal domain.* bioRxiv, 2019.

Chapter 2:

A metastable junction in human telomerase RNA is remodeled during enzyme assembly

Abstract:

Telomeres safeguard the genome by suppressing illicit DNA damage responses at chromosome termini. In order to compensate for incomplete DNA replication at telomeres, most continually dividing cells, including many cancers, express the telomerase ribonucleoprotein (RNP) complex. Telomerase maintains telomere length by catalyzing de novo synthesis of short DNA repeats using an internal telomerase RNA (TR) template. TRs from diverse species harbor structurally conserved domains that contribute to RNP biogenesis and function. In vertebrate TRs, the conserved regions 4 and 5 (CR4/5) fold into a three-way junction (3WJ) that binds directly to the telomerase catalytic protein subunit and is required for telomerase function. We have analyzed the structural properties of the hTR CR4/5 domain using a combination of in vitro chemical mapping, endogenous RNP assembly assays, and single-molecule structural analysis. Our data suggest that a functionally essential stem loop within CR4/5 is not stably folded in the absence of the telomerase reverse transcriptase protein subunit in vitro. Rather, the hTR CR4/5 domain adopts a heterogeneous ensemble of conformations. RNA structural engineering intended to bias the folding landscape of the hTR CR4/5 demonstrates that a stably folded 3WJ motif is necessary but not sufficient to promote assembly of a functional RNP complex. Single-molecule measurements on the hTR CR4/5 domain show that RNP assembly

selects for a conformation that is not the major population in the heterogeneous free RNA ensemble. Alternate folds of the hTR CR4/5 domain should serve as new therapeutic targets for small molecules inhibitors of telomerase.

Introduction

The ends of linear chromosomes in eukaryotic cells terminate with highly repetitive DNA sequences that bind to specialized proteins to form telomeres [1]. Telomeres protect coding DNA from degradation and distinguish chromosomal termini from double-stranded breaks in order to evade unwanted recognition by DNA damage response machineries fusion [2][3]. With each round of cell division, the inability of the conventional replication machinery to completely copy the lagging strand template results in gradual telomere attrition. Ultimately the presence of a critically short telomere drives cells into permanent cell growth arrest or apopotosis [4-6]. However, cells that must retain high proliferative capacity maintain telomere length through the action of the telomerase reverse transcriptase [7-10]. Given the importance of maintaining telomere length in dividing cells, germ-line mutations in telomerase genes result in severe developmental defects [11] [12] [13]. In addition, telomerase contributes to the unchecked cell growth that is a hallmark of human cancers [14]. Therefore, efforts to better understand telomerase structure, function, and regulation have direct biomedical significance.

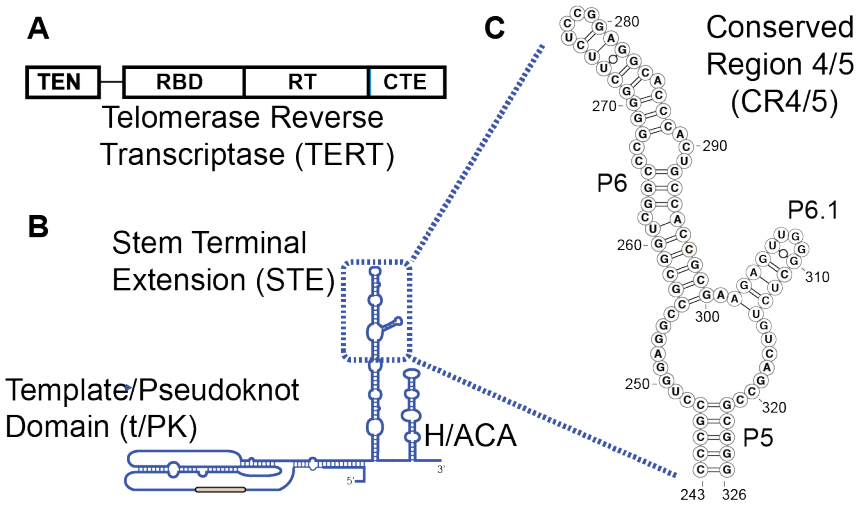


Figure 2 1 - Conserved protein and RNA domains of the telomerase catalytic core.

(A) The conserved domains of telomerase reverse transcriptase (TERT). (B) The conserved domains of telomerase RNA (TR). (C) Conserved helixes P5, P6 and P6.1 in the vertebrate STE, known as Conserved Region 4/5 (CR4/5).

Telomerase is a multi-subunit ribonucleoprotein (RNP) complex that includes the catalytic telomerase reverse transcriptase (TERT) protein, telomerase RNA (TR), and several additional species-specific holoenzyme proteins that are necessary for proper RNP biogenesis [7]. TERT structure is well-conserved across species and consists of four domains: the telomerase essential N-terminal (TEN) domain, the RNA binding domain (RBD), the reverse transcriptase (RT) domain, and the C-terminal Extension (CTE) (**Fig 2.1A**). In contrast, comparison of TRs across species ranging from yeasts to human reveals an exceedingly high degree of variation in both RNA length and sequence. Interestingly, in spite of this apparent evolutionary divergence, there exist several conserved TR structural elements essential for enzyme assembly and function. These include the universally conserved template/pseudoknot (t/PK) domain and a stem-terminal element (STE) (**Fig. 2.1B**)

[15]. In vertebrate TRs, the STE is thought to fold into an RNA three-way junction (3WJ) often referred to as the conserved regions 4/5 (CR4/5) domain. Furthermore, vertebrate TRs harbor a canonical H/ACA box domain that binds to the general scaRNP biogenesis factors: dyskerin, Nop10, NHP2, and Gar1. With regard to TR primary sequence, the CR4/5 domain is spatially separated from the RNA template that must necessarily reside in the TERT enzyme active site; yet, naturally occurring mutations in hTR CR4/5 can result in human diseases characterized by loss of telomerase function [16] [11] [17].

In hTR, the CR4/5 domain includes three RNA helices (P5, P6, and P6.1) joined together by an expanded RNA junction sequence (**Fig. 2.1C**). Detailed biochemical studies performed on vertebrate TR CR4/5 variants have shown that a stably formed P6.1 helix within the 3WJ is essential for telomerase assembly and function [18, 19]. Protein-RNA crosslinking studies and an atomic-resolution structure of the medaka fish TR 3WJ bound by its cognate TERT-RBD revealed the molecular details of the TERT-RNA interaction [20]. Interestingly, the helical arrangement observed in the medaka protein-RNA complex was substantially altered when compared to the solution structure of the same RNA domain in the absence of protein [21]. More recently, cryoEM structures of the *Tetrahymena* and human telomerase RNPs were reported [22, 23], providing additional details on the arrangement of protein and RNA domains within the fully assembled telomerase RNP complex. Interestingly, both structures, in particular the higher resolution Tetrahymena complex, suggest that an apical stem loop within the STE (P6.1 in hTR) serves as a molecular coupler that lies at the interface of the TERT-CTE and

TERT-RBD domains, providing a plausible explanation for the essential requirement of the P6.1 stem loop.

Here, we set out to characterize the in vitro RNA folding properties of the hTR CR4/5 domain using a combination of chemical mapping, paired together with single-molecule Forster Resonance Energy Transfer (smFRET) experiments. Chemical probing experiments using a variety of RNA modification reagents revealed a substantial degree of reactivity within the region of hTR CR4/5 expected to form the essential P6.1 stem loop structure. Use of chemical reactivity data to guide computational modeling of RNA structure predicts several possible alternative conformations that may be sampled within the hTR CR4/5 structure ensemble. To further validate these structure predictions, we systematically perturbed each nucleotide within the hTR CR4/5 domain, and queried the effects of each mutation on the chemical reactivity profile. The results of these Mutate-and-Map (M^2) experiments [24] [25] reinforce the conclusion that the P6.1 stem loop is not well ordered in vitro. Next, we engineered hTR CR4/5 variants designed to bias the folding energy landscape to favor P6.1 formation. After validating the efficacy of the RNA designs by chemical probing in vitro, selected full-length hTR constructs were transfected into human cells to be endogenously assembled and purified. We find that certain hTR sequence variants intended to stabilize the canonical P6.1 fold displayed marked defects in assembly of functional RNP complexes, demonstrating that a stably folded hTR 3WJ is necessary but not sufficient for telomerase function. Using smFRET to probe the conformational properties of the hTR CR4/5 domain also revealed heterogeneous RNA folding, characterized by at least three distinct FRET states. Interestingly, smFRET measurements made on the same hTR CR4/5

fragment assembled into an active RNP complex show that the conformation of the RNA is substantially altered upon protein binding. Taken together, these results suggest that non-canonical conformations of the P6.1 stem loop, as well as specific junction nucleotides, within the hTR CR4/5 domain are required for faithful RNP assembly. Collectively, our results are consistent with a working model wherein non-canonical TR folds serve as assembly intermediates during telomerase biogenesis, as has been shown for other essential cellular RNPs such as the ribosome and splicesome. Given the central importance of CR4/5 folding in promoting telomerase assembly and function, the hTR folding properties characterized in the present work provide a new framework for developing small molecules that target non-canonical hTR structures with the goal of inhibiting telomerase in cancer cells.

Results

An essential RNA stem loop within the human CR4/5 domain is not stably folded in the absence of TERT

The 3WJ motif is well conserved across many telomerase RNA systems, ranging from yeasts to vertebrates. Many of the RNA structural models that are used to generate hypotheses relating to telomerase function are derived from sequence covariation analysis [15] and/or the use of biochemical mutagenesis [18, 26]. One challenge of methods such as sequence covariation analysis is that the resultant models may not accurately capture the structural properties of the RNA in the absence physiological binding partners. For example, studies of telomerase biogenesis indicate that hTR accumulates in sub-nuclear compartments prior to assembly with the TERT protein subunit [27, 28]. In order to better understand the

structural properties of TRs prior to and during RNP biogenesis, we set out to analyze the secondary structural properties of telomerase 3WJs from two vertebrate systems: medaka fish (*Orzias latipas)* and human. The medaka TR 3WJ serves as an important benchmark in our TR structural analyses because its atomic structure is well characterized in the absence and presence of the TERT-RBD [20, 21].

For each TR system, we used an isolated CR4/5 RNA fragment to facilitate in vitro structure probing. Notably, the isolated hTR CR4/5 domain used in our studies is sufficient to support telomerase function when reconstituted with the hTR t/PK domain and TERT protein (data not shown) [29]. Several sequence elements were added to the TR segment to assist in quantitative data analysis of chemical probing experiments (**Fig. 2.2A**). First, a primer-binding site was appended to the RNA 3'-end for use in the reverse transcriptase reactions required to read out sites of RNA modification. Second, a short RNA hairpin structure flanked by unstructured 'buffer' regions was added to serve as an internal normalization control when calculating chemical relativities (see Methods for details) [30]. De novo structure predictions calculated using the RNAstructure web server [31] yielded lowest free energy conformations with the expected stems that collectively form the 3WJ fold (**Fig. 2.2B**). In the case of the hTR CR4/5 domain, RNAstructure predicts an additional cross-junction clamping helix not typically included in canonical representations of this region of hTR. Furthermore, multiple structures with nearly isoenergetic stability are also predicted, including several conformations lacking the essential P6.1 stem loop (data not shown), highlighting the need for experimental data to validate specific RNA models.

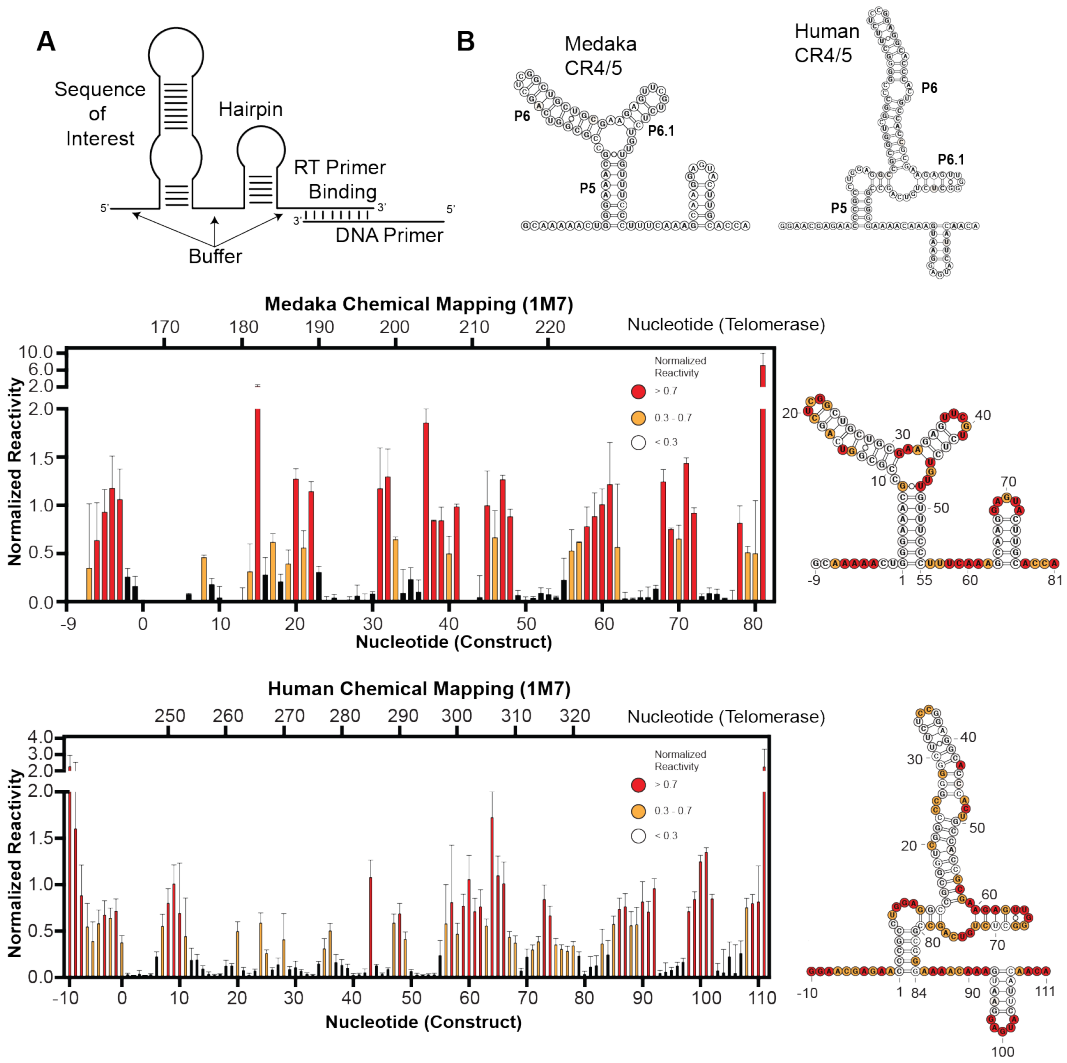


Figure 2 2 - Chemical mapping of medaka and human CR4/5 domain.

(A) Schematic of construct design. (B) Predicted secondary structure of medaka and human CR4/5 domain using RNAstructure. (C and D) Chemical mapping of the medaka and human CR4/5 domain using 1M7. Reactivity was mapped onto predicted secondary structure illustrated in B. Red denotes reactivity values above 0.7, yellow from 0.3 - 0.7, and white below 0.3. Telomerase numerical nomenclature is denoted on the top x-axis and the construct numerical nomenclature is denoted on the bottom x-axis.

To experimentally evaluate each of these structure predictions we performed chemical mapping of each construct. We interrogated the flexibility of the sugar-

phosphate backbone using 1-methyl-7-nitoisatoic anhydride (1M7), a fast acting chemical modifier used in selective hydroxyl acylation analyzed by primer extension (SHAPE) experiments. The 1M7 SHAPE reagent acylates the 2'-hydroxyl of flexible ribose moieties along the RNA backbone, thereby providing a sequence-independent proxy for unstructured regions of an RNA. For each experiment, sites of chemical modification were read out as premature termination sites during RT catalyzed primer extension. The individual reactivity at every position along the RNA was calculated and then normalized to the internal hairpin control signal (see Methods for details). In addition, experiments were also performed using the base-specific reagents dimethyl sulfate (DMS) or 1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-*p*-toluene sulfonate (CMCT), which primarily react with adenine/cytosine or guanine/uracil bases, respectively. In the case of the medaka TR CR4/5 domain, reactivity profiles obtained by all three chemical probing methods (DMS, CMCT, and 1M7) yielded data that support the canonical base pairing arrangement expected for this 3WJ fold, and are highly consistent with the reported solution structure of this same RNA fragment [20] (**Fig. 2.2C**). In contrast, strong reactivity was observed in the region of the hTR CR4/5 domain expected to fold into the P6.1 stem loop (**Fig. 2.2D**). This result is unexpected given the established importance of the P6.1 stem loop structure in promoting telomerase RNP assembly and function [18, 19]. Taken together, these data suggest that the RNAstructure folding algorithm can effectively capture the base pairing configuration of the medaka TR 3WJ, but fails to do so in the more expanded junction/P6.1 region of the hTR CR4/5 domain.

SHAPE-guided modeling of telomerase RNA CR4/5 domain structure

RNAstructure calculates the lowest free energy structures using thermodynamic parameters that are dynamically sampled against databases of structures with well-characterized stabilities [31]. Experimentally derived chemical probing data significantly improves the predictive power of the RNAstructure folding algorithm [32]. For example, SHAPE reactivities are typically used to generate a pseudo-energy change term (deltaG$_{SHAPE}$) for each individual base pair of a predicted structure, which can then be added to the RNAstructure prediction algorithm as a nearest neighbor free energy term [33]. Using this approach, we generated SHAPE-guided models of the medaka TR 3WJ and the hTR CR4/5 domain with the Biers software package that implements the use of the RNAstructure prediction algorithm [24, 25]. In addition to predicting a lowest energy conformation for a particular RNA sequence utilizing SHAPE reactivity data, Biers also includes a nonparametric bootstrapping function to estimate confidence levels in the prediction of each helical element within a particular low energy RNA conformation. Specifically, the bootstrapping function within the Biers software package uses random resampling with replacement of experimental SHAPE data to calculate an ensemble of data-guided RNAstructure outputs. This ensemble is then used to calculate the frequency of each base pair present in each computationally derived replicate. In this way, the resulting bootstrap value for any given helix provides a metric to help evaluate the degree of confidence for each helix given the data that were used to guide the structure calculation. It is important to note that bootstrap values are a statistical tool to analyze computational prediction methods, and should not be interpreted as an indication of the equilibrium conformation(s) present for a particular RNA of interest.

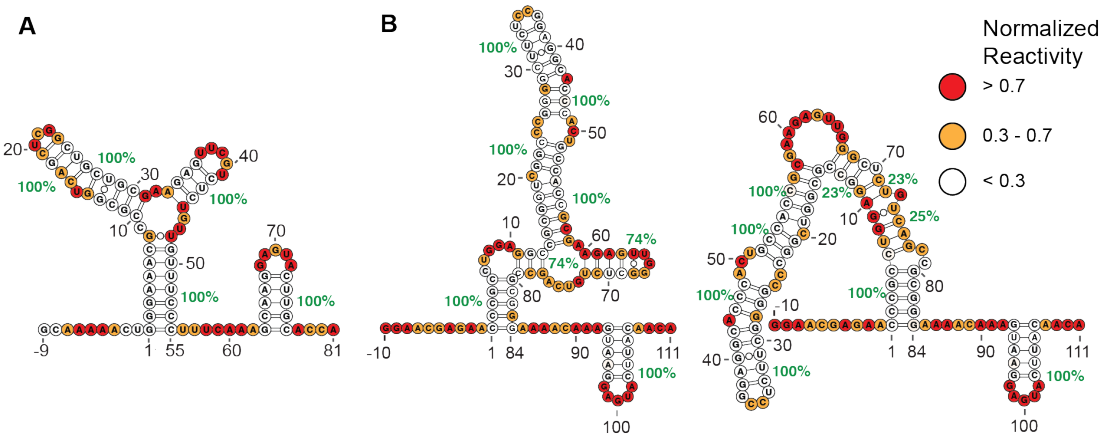


Figure 2 3 - Data driven RNA secondary structure prediction of medaka and human CR4/5 domain.

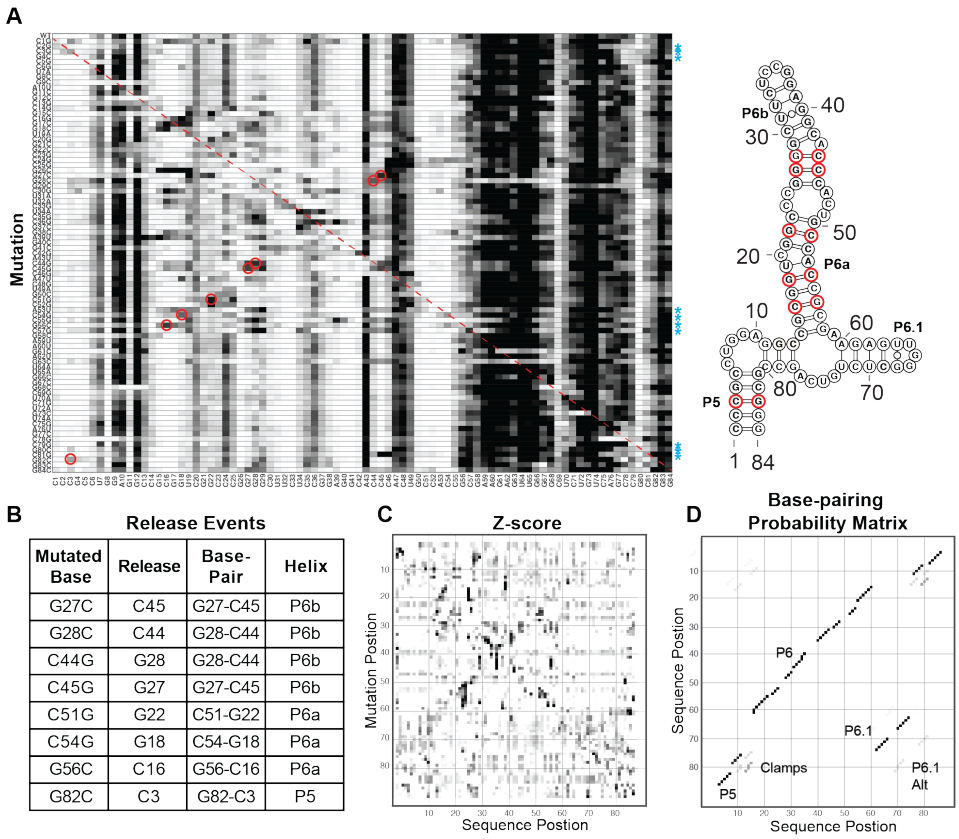
(A - B) 1M7 reactivity data was used to guide RNA structure prediction for medaka and human CR4/5 domains. Bootstrap parametric sampling was preformed to determine confidence intervals for each helix and is displayed in green. Red denotes reactivity values above 0.7, yellow from 0.3 - 0.7, and white below 0.3.

As expected, the addition of the deltaG$_{SHAPE}$ constraints to predictions of the medaka TR CR4/5 yield the canonical 3WJ fold with each of the expect helices being called with very high bootstrap values (**Fig. 2.3A**). We note that structure prediction is performed on the complete RNA construct, including the normalization hairpin which is also predicted with very high confidence. This result indicates that addition of experimentally derived data does not cause the RNAstructure algorithm to deviate in its prediction of the lowest energy conformation for the medaka TR CR4/5. In the case of the hTR CR4/5, the inclusion of deltaG$_{SHAPE}$ constraints in structure calculation recaptures a lowest energy conformation in which the P5, P6, and normalization hairpin are called with very high confidence. In contrast, the bootstrap value calculated for the P6.1 stem is significantly decreased, consistent with the high levels of reactivity in this region (**Fig. 2.3B**). Manual inspection of alternative predicted RNA conformations with nearly equal calculated energies yield a

substantially altered junction region defined by extensive base pairing that is not characteristic of a 3WJ fold (**Fig. 2.3B**). These data-driven structure predictions indicate the hTR CR4/5 domain is structurally heterogeneous within the expanded junction region and the functionally essential P6.1 stem loop is not stably folded in vitro.

Mutli-dimensional Chemical Mapping supports the presence of hTR CR4/5 heterogeneity

Chemical mapping experiments as described above provide important information about which nucleotides are not engaged in base pairing interactions. While these data are useful to help guide computational structure prediction, they lack any information as to the identity of specific base pairing partners within the primary RNA sequence. To address this limitation, a systematic mutagenesis approach was recently reported [24, 25] that permits rapid chemical probing analysis of a panel of RNA mutant constructs designed to explicitly test for the presence of Watson-Crick base pairing in a proposed RNA secondary structural model. If a mutation is made to a base that is engaged in a base pair, then one expects the interacting partner to become accessible to the SHAPE probe. Such release events provide powerful information with which to infer the presence of specific base pairs in an RNA structure. In practice, due to the complexity of RNA folding energetics, individual mutations may also elicit more complex phenotypes that also provide useful information about the folding properties of an RNA.



Mutated Base	Release	Base-Pair	Helix
G27C	C45	G27-C45	P6b
G28C	C44	G28-C44	P6b
C44G	G28	G28-C44	P6b
C45G	G27	G27-C45	P6b
C51G	G22	C51-G22	P6a
C54G	G18	C54-G18	P6a
G56C	C16	G56-C16	P6a
G82C	C3	G82-C3	P5

Figure 2 4 - Mutation profile of the human CR4/5 domain.

(A) Mutations (A to U, U to A, G to C and C to U) were introduced at each base. Each mutant underwent chemical mapping using 1M7 and the resulting reactivity profiles were stacked allowing comparison of at each nucleotide across all mutants. Mutation position is indicated with a red dashed line. Release events observed are boxed in red. (B) Release events identified are outlined in red mapped onto the CR4/5 secondary structure. (C) Mutations and release events are summarized in a table. (C) A Z-score plot was used to calculated statistically significant reactivity events triggered by mutation. (E) A base pairing probability matrix shows the calculated likelihood of each base pair predicted. Base pairs predicted with higher confidence are darker in grey scale and those with lower probability are lighter in grey-scale. The helix that each base pair belongs to are annotated.

To further probe the structure of the hTR CR4/5 domain, we performed this multidimensional chemical mapping (MCM) procedure, generating a set of eighty-

four mutants across the entire RNA construct (**Fig. 2.4A**). The chemical reactivity profiles of all RNA variants are stacked vertically to generate a reactivity tapestry. Signals on the diagonal of the reactivity tapestry represent release events at the engineered site of mutation **(Fig. 2.4A, red dotted line**). Signals that deviate from the wild type reactivity profile indicate changes in reactivity that result from each individual mutation. Many of the single mutant reactivity profiles reveal complex structural rearrangements beyond the simple base pair release event principle. We generally categorize signals from the MCM experiment into two classes: single base pair release events and larger scale structural reorganization. Manual inspection of the data reveals multiple features in the reactivity tapestry that support specific base pairs present within the hTR CR4/5. For example, the C51G mutant displays an increased reactivity at G22, and the G56C mutant results in an increased reactivity at C16, providing direct support for the presence of these two base pairs within the P6 stem (**Fig. 2.4A, red circles**). In addition, the G82C mutation induces increased reactivity at C3, providing evidence for this base pair in the center of the P5 stem. A list of each release event identified is shown in **Fig 2.4B**.

Many mutations introduced did not give rise to clean release events. Interestingly, due to the high G-C content of P5 for many mutations (G to C and C to G), a helix shift is observed rather than a release event (**Fig 2.4A asterix**). We find that mutations introduced at the base of the P6 stem have the unexpected effect of causing substantial structural rearrangement in the CR4/5 domain, evidenced by reduced reactivity in the junction region and increased reactivity within the P6 stem (**Fig. 2.4A, asterix**). While not direct release events these large scale conformational change none-the-less provide interesting qualitative data regarding the folding

landscape of the CR4/5 domain. Unfortunately, the high baseline reactivity and the complexity of the reactivity profile observed in the hTR CR4/5 junction and P6.1 stem loop region precludes unambiguous visual analysis of the MCM data.

To achieve a more quantitative analysis across the reactivity tapestry we generate a Z-score plot, where individual Z-scores report on the statistical significance of deviations in the reactivity level for a given nucleotide compared across all RNA constructs (**Fig. 2.4D**). Using the Z-score, mutations that cause perturbations above the average level of reactivity at that nucleotide across all mutants is scored and was used by RNAstructure to predict the most likely base pairing interactions for each nucleotide. While not all mutations will give rise to a clear release event, even a few known base pairing interactions can strongly guide RNAstructure in the structures predicted. Predicted helices are then mapped onto a base pairing probability matrix with helices predicted with high confidence in dark shades and those with lower confidence in lighter shades. The resulting base-pairing probability matrix (**Fig 2.4E**) predicts P6.1 to be present but also predicts with a series of cross-junction clamps with a high degree of certainty. Additionally, an alternative helix is predicted with lower confidence. These data are consistent with the conclusion that the bulged junction of the CR4/5 is structurally heterogeneous.

P6.1 formation is not sufficient for telomerase assembly

Structural heterogeneity in RNA could be an evolutionary artifact or an important functional property of the RNA. To query these two possibilities mutants that alter the structural heterogeneity at the TJW bulge and in the P6.1 stem were engineered. A Pol-A construct was engineered such that every nucleotide in the junction bulge was

mutated to an A. This construct would eliminate structures other than P6.1 that may exist across the junction, while leaving the P6.1 sequence in tact. Prior studies have established that destabilizing P6.1 causes severe functional defects [18, 19]. To test the effect of altered structural heterogeneity a construct was designed that strengthened the P6.1 construct by mutating an A-U (A60C and U72G) and G-U (U64C) pair to G-C pairs.

Chemical mapping was used to confirm that the structural changes anticipated were achieved (**Fig 2.5A**). Mutations were cloned into full length hTR and transiently transfected with hTERT into HEK 293T cells. Following transfection telomerase was immunoprecipitated using the FLAG tag on TERT and tested using a dot blot hybridization assay to assess for telomerase assembly and a primer extension assay to test for telomerase activity. While FLAG-IP'd WT telomerase was able to extend off of the provided primer and generate telomeric repeats, the Poly-A mutant activity was severely impaired (**Fig 2.5C**). Dot blot analysis confirms that Poly-A assembly is significantly knocked down compared to WT (**Fig 2.5B**). In contrast the P6.1 stabilization mutant shows a slight knockdown in telomerase activity (**Fig 2.5C**) but WT level of assembly (**Fig 2.5B**). Taken together the Poly-A mutants suggests that the formation and sequence identity of the P6.1 helix is not sufficient for telomerase assembly. Mutations stabilizing the P6.1 helix likely alter the structural heterogeneity of the bulged junction but to what extent remains unclear. Because only a slight decrease in telomerase activity was observed in this construct the exact role of structural heterogeneity remains unclear.

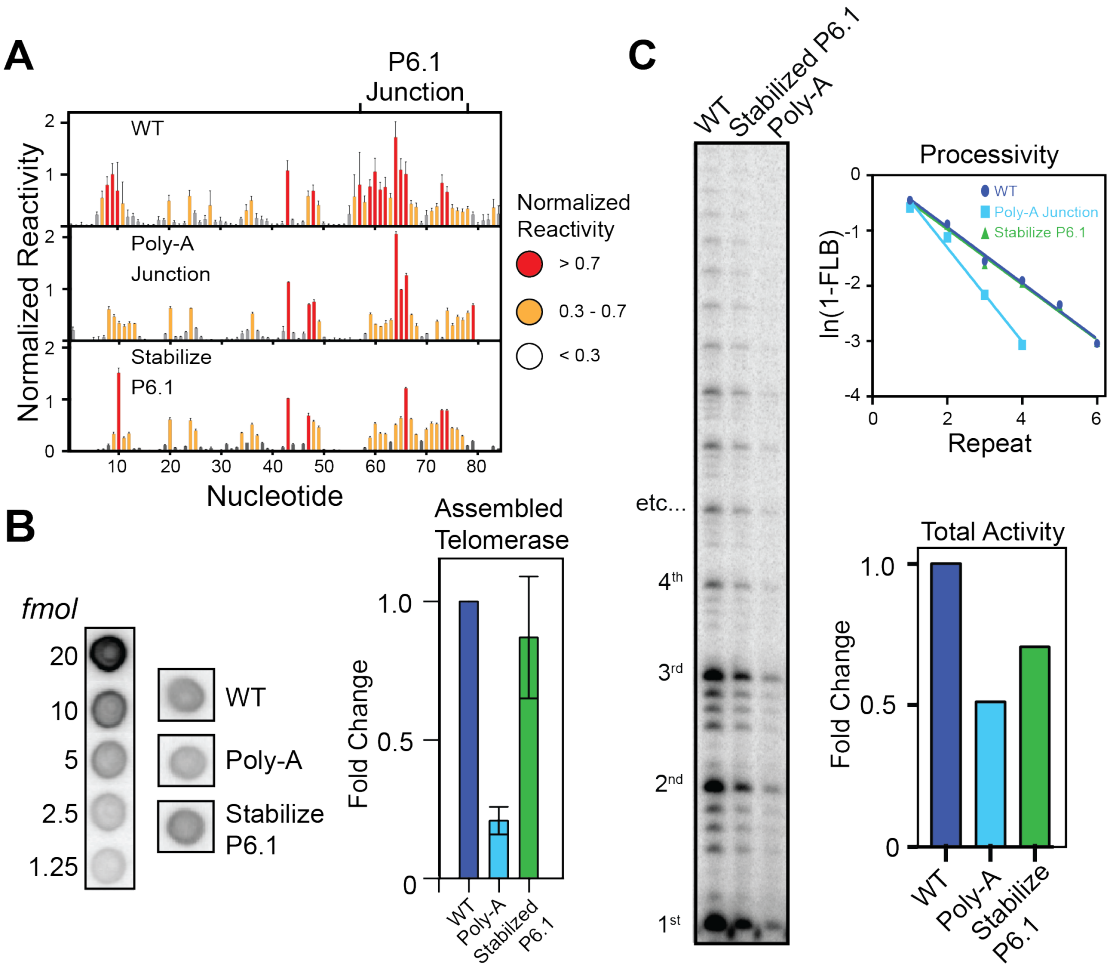


Figure 2 5 - Structurally engineered CR4/5 domains exhibit assembly and functional defects.

(A) 1M7 reactivity profile of WT, Poly-A junction, and Stabilized P6.1 engineering CR4/5 constructs (B) Dot blot was preformed of FLAG IP'd material from cells transiently transfected with WT, Poly-A, or Stabilized P6.1 construct. A standard curve using in vitro transcribed RNA was used to quantify levels of RNA and fold change between WT and mutantswas calculated (C) Primer extension assay of WT, Poly-A junction, and Stabilized P6.1 construct was preformed on IP'd material. Processivity and total activity for each construct was calculated.

Human CR4/5 adopts multiple tertiary structural states

To further query the structural properties of the human CR4/5 domain and the P6.1 stem, we employed a single molecule Förster Resonance Energy Transfer

(FRET) approach. We placed the FRET-coupled dyes Cy3 and Cy5 on P6b (U274) and P6.1 (U312) to create a dye pair that reports on the physical proximity of these two helical elements. In addition, we reconstituted this fluorescently-labeled CR4/5 domain into human telomerase RNPs using rabbit reticulocyte lysates (RRL) by supplying a plasmid encoding FLAG-hTERT and a template/pseudoknot RNA fragment, followed by immunoprecipitation of the RNP. The activity of this labeled-CR4/5 telomerase was comparable to unlabeled telomerase as assessed by a primer extension assay. smFRET measurements of two samples we made using a confocal fluorescence microscope, in which FRET values are extracted from single freely diffusing molecules as they pass through the excitation beam.

The CR4/5 domain in the absence of TERT reveals a strikingly heterogeneous FRET profile with values ranging across the FRET scale from ~0.1 to 1 (**Fig 2.6A**). From this dataset we observe that CR4/5 without TERT forms an ensemble of structural states, with the majority of molecules falling into populations centered at ~0.75 and 0.9 FRET by Gaussian approximation. Molecules reporting these FRET values likely exist in a conformation in which the P6.1 stem in close proximity to P6b. However, human CR4/5 also populates a 0.3 FRET state, suggesting a conformation in which P6.1 is distal to P6b exists. Together, these data lend support to the notion that the human CR4/5 is structurally heterogeneous, capable of adopting tertiary folds that differ in helical arrangement around the 3WJ of CR4/5. Single molecule FRET measurements of CR/45 assembled into telomerase via RRL reports a dramatic restriction of FRET values to about 0.3. This suggests that when assembled with TERT, P6.1 is distal to P6b, suggesting a conformation similar to that seen in the crystallized complex of medaka CR4/5-TERT. Assembly of

human CR/45 with TERT appears to select for a particular architecture of the human 3WJ from among its heterogeneous ensemble.

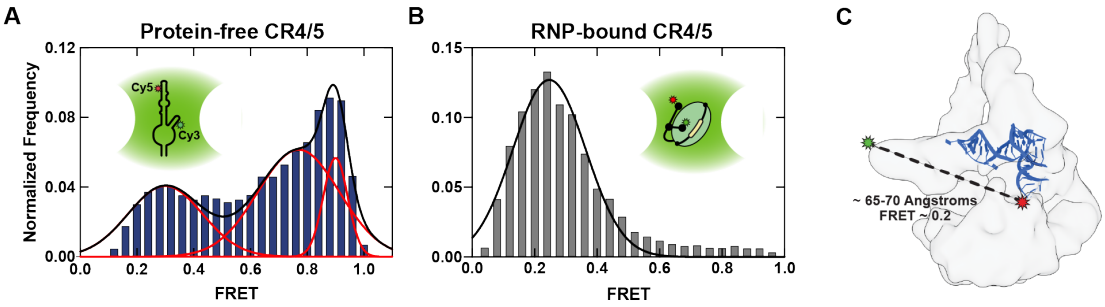


Figure 2 6 - Confocal microscopy reveals structural heterogeneity in the CR4/5 domain and stabilization of a single state when bound to TERT.

(A) Double labeled CR4/5 domain exhibits multiple FRET populations. Protein-free CR4/5 domain is illustrated schematically. (B) CR4/5 is reconstituted into the telomerase RNP and a FRET population centered around 0.3 is observed. The telomerase RNP is illustrated schematically. (C) Cryo EM density of low-resolution human telomerase (REF) with crystal structure of RBD bound medaka CR4/5 (REF) placed in blue. Approximate location of dyes is docked onto density and approximate distance between the dyes based on structure as well as based on FRET was calculated.

Discussion

The importance of RNA heterogeneity and dynamics within the context of RNP assembly and function remains a challenging topic to study due to difficulties in predicting and characterizing RNA structures. These challenges are compounded when attempting to predict how those structural dynamics change upon protein binding.

Understanding the mechanistic details regarding TERT-TR interactions has been hindered by a lack of structural information. Additionally, the assembly pathway of telomerase remains unclear though the transport and biogenesis of TR has been a subject of recent interest and many RNA processing proteins have been proposed to

be responsible for the biogenesis of the mature hTR form. The Poly-A junction CR4/5 construct demonstrates clearly that important protein-RNA contacts involved in telomerase assembly are present in the bulged junction of the CR4/5 domain. However, whether these mutations are involved in TERT-TR interactions or another protein-TR interaction in the assembly pathway remains an open question.

The identification of structural heterogeneity in the CR4/5 domain provides a potential explanation for prior observations in telomerase studies. For example, in high vertebrates an expansion of the CR4/5 domain junction is observed. The nucleotides within this region are conserved despite the presence of hyper-variable sequences flanking the region. Despite this sequence conservation it has been shown that no phylogenetic co-variation exists for P6.1 and the surrounding regions, though co-variation does support the formation of P6 and P5. The presence of multiple structures important for function may account for this observation. The exact identity and whether these alternate structures play a role in the selective recruitment of proteins during the biogenesis pathway remains to be seen.

The role in structural heterogeneity in TERT-TR has been hinted at in studies of other model systems of telomerase. The medaka CR4/5 shows a dramatic tertiary structural change upon TERT binding. Our FRET data is the first to support this model in human telomerase. Interestingly in *Tetrahymena* telomerase a *Tetrahymena* specific protein, P50, binds to Stem IV and induces a strong bend in Stem IV [34] [35]. The current model for Tetrahymena telomerase assembly proposes that after this rearrangement a high affinity interaction between RBD and TR drives holoenzyme assembly. Perhaps in human telomerase structural flexibility is built into the RNA rather than requiring a protein-binding partner.

Overall our data speaks to the importance of structural heterogeneity in the CR4/5 junction domain for telomerase assembly. It remains unclear how human disease associated mutations perturb telomerase function. An increased understanding of the structural intermediates in the CR4/5 domain may allow drugs to be designed to stabilize or rescue an RNA conformation. Additionally, the high-throughput mutagenesis of the CR4/5 domain has identified regions of the RNA that are particularly important in the CR4/5 folding pathway. Selection of an unfavorable assembly state could provide a potential cancer drugs while rescue of the functional assembly state could offer a therapeutic target for telomerase associated developmental diseases.

Materials and Methods:

RNA construct design:

In addition to the RNA sequence of interest, the following RNA sequences were included for data analysis or RNA purification purposes (going from 5' to 3'): T7 transcription binding sequence, buffer, telomerase CR4/5 sequence, buffer, hairpin, buffer, reverse transcription (RT) binding sequence. All sequences used are described in Supplemental Table 1. To check for possible non-desirable interactions between buffers, hairpin, and RT binding sequence RNAstructure was used. Structure prediction analysis was confirmed with subsequent chemical mapping data.

Primer assembly and In vitro transcription:

Primerize was used to generate a series of primers that were used in primer assembly to generate PCR products using recommended protocol (Tian et al 2015). PCR products were purified using Quiagen (#28104) PCR cleanup kit or AMPure XP

Beads (Beckman #A63881) and quality of DNA fragment generated was confirmed using a 2% agarose gel. RNA was generated by in vitro transcription (40mM Tris-HCl pH 7.9, 25mM MgCl2 90mM DTT 2mM spermidine 25mM rNTP) using T7 polymerase and PAGE purified or purified using AMpure XP Beads. RNA quality was confirmed using a 10% PAGE gel.

RNA folding and chemical modification:

Chemical mapping was done as described in detail in Cordero et al 2014. Briefly, 1.2pmols of RNA was unfolded in 50mM Na-HEPES pH 8.0, at 95C for 3 minutes then removed from heat block and allowed to slow cool and refold. After 20 minutes, $MgCl_2$ was added to a final concentration of 10, 5, 1, or 0mM. All experiments except $MgCl_2$ titration were done at 10mM $MgCl_2$. 15uL of folded RNA was aliquot into a 96 well plates.

RNA modifier was added – 0.25% DMS (Sigma #D186309), 25mM CMCT (TCI America #C0793), 5mM 1M7 (generously provided by Dr. Manny Ares) – and allowed to incubate at room temperature for 15 minutes. After chemical modification, reaction was quenched (84mM Na-MES pH 6.0, 500mM NaCl, 2.1nM FAM-A20 primer, Poly-dT Beads washed) at room temperature for 10 minutes. Magnetic beads were pulled down for 7 minutes on 96 well plate magnetic rack. Supernatant was discarded and beads were washed 2x with 100uL 70% ethanol. Beads were allowed to air dry for 10 minutes to remove residual ethanol then resuspended in 2.5uL of nuclease free water.

Reverse transcription and cDNA purification:

Reverse transcription took place under the following conditions - 1x First Strand Buffer, 5mM DTT, 0.8mM dNTP, 20 units of reverse transcriptase Superscript

III. RT reaction was incubated at 48C for 30 minutes. RNA was hydrolyzed with 200mM NaOH at 95C for 3 minutes. Plate was put on ice for 3 minutes and acid quench (1 volume 5M NaCl, 1 volume 2M HCl, 1.5 volume 3M Sodium Acetate) was added to return to pH 7. Magnetic beads were pulled down and washed as described above. The beads were resuspended in Hi-Di Formamide and ROX ladder and cDNA was allowed to elute for 20 minutes. Beads were pulled down and supernatant was kept. Samples were sequenced by capillary electrophoresis by ElimBio.

Capillary Electrophoresis data analysis and secondary structure prediction:

Capillary electrophoresis data was analyzed using the HiTrace software package [36]. Bands were aligned and fit to a Gaussian. For 1D chemical mapping, band intensity was corrected for saturated bands, over-modification, and background subtraction [30]. Reactivity factors were calculated through normalization to an internal hairpin serving as a reference allowing for cross-experimental comparisons [30]. RNAstructure was used to predict RNA secondary structure using the reactivity factors as weights. This process was implemented using Biers as part of the HiTrace workflow. For 2D chemical mapping analysis the band intensity was averaged at that nucleotide across all nucleotides. By calculating the deviation from the average band intensity the Z-score plot was created.

Telomerase transfection and immunoprecipitation:

Constructs were transiently transfected into HEK 293T cells Enzyme was immunoprecipitated (IP) using FLAG-beads (Sigma #A2220) and eluted using FLAG peptide (Sigma #F4799). Telomere extension was measured with 2.5uL of IP'd enzyme at 50mM Tris pH 8.3, 3mM MgCl2, 2mM DTT, 80mM dNTP, 100mM end-labeled telo-primer. Primer extension was allowed to occur at 30C for 1.5 hours.

DNA was phenol chloroform extracted, ethanol precipitated, run on a 12% sequencing gel and imaged on a phosphorus screen

Assembly Assay and Dot Blot:

5uL aliquots of FLAG IP'd telomerase were diluted to 10 uL in formamide loading buffer (90% deionized formamide, 0.1% bromphenol blue, 0.1% xylene cyanole and 1X TBE) and heated at 70◦C for 5 min and placed on ice. The solution was pipetted onto Hybond N+ membrane (GE Lifesciences) and allowed to air dry at room temperature for 1 hour. Sample was cross-linked to the surface using a UV transilluminator. The cross-linked membrane was blocked in 50 ml Church buffer (1% BSA, 1 mM EDTA, 500 mM sodium phosphate pH 7.2, 7% SDS) at 55C for 1 hour. Approximately 3×10^6 cpm of a 5 - 32P labeled DNA oligo was added (sequence: 5 - TATCAGCACTAGATTTTTGGGGTTGAATG-3) and incubated while shaking at 55C for at least 12 hours. The membrane was washed, shaking for 15 min 3x in 0.1X saline-sodium-citrate buffer (15 mM NaCl, 1.5 mM trisodium citrate, pH 7.0) containing 0.1% SDS at room temperature. The membrane was imaged using a phosphor screen (GE Lifesciences) and a typhoon scanner (GE Lifesciences). Quantification of the blot was performed with ImageJ. To determine concentrations, samples were compared against in vitro transcribed telomerase RNA standards dotted onto the same blot.

References:

1. Blackburn, E.H. and J.G. Gall, *A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena.* J Mol Biol, 1978. **120**(1): p. 33-53.
2. Muller, H.J., *The remaking of chromosomes.* The Collecting Net, 1938. **13**: p. 181-195.
3. McClintock, B., *The behavior in successive nuclear divisions of a chromosome broken at meiosis.* Proc Natl Acad Sci U S A, 1939. **25**: p. 405-416.
4. Hayflick, L., *The Limited in Vitro Lifetime of Human Diploid Cell Strains.* Exp Cell Res, 1965. **37**: p. 614-36.
5. Harley, C.B., A.B. Futcher, and C.W. Greider, *Telomeres shorten during ageing of human fibroblasts.* Nature, 1990. **345**(6274): p. 458-60.
6. Allsopp, R.C., et al., *Telomere length predicts replicative capacity of human fibroblasts.* Proc Natl Acad Sci U S A, 1992. **89**(21): p. 10114-8.
7. Greider, C.W. and E.H. Blackburn, *A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis.* Nature, 1989. **337**: p. 331-337.
8. Kolquist, K.A., et al., *Expression of TERT in early premalignant lesions and a subset of cells in normal tissues.* Nat Genet, 1998. **19**(2): p. 182-6.
9. Roth, A., et al., *Telomerase levels control the lifespan of human T lymphocytes.* Blood, 2003. **102**(3): p. 849-57.
10. Blasco, M.A., *Telomeres and human disease: ageing, cancer and beyond.* Nat Rev Genet, 2005. **6**(8): p. 611-22.

11. Yamaguchi, H., et al., *Mutations of the human telomerase RNA gene (TERC) in aplastic anemia and myelodysplastic syndrome.* Blood, 2003. **102**(3): p. 916-8.

12. Vulliamy, T.J. and I. Dokal, *Dyskeratosis congenita: the diverse clinical presentation of mutations in the telomerase complex.* Biochimie, 2008. **90**(1): p. 122-30.

13. Savage, S.A., *Human telomeres and telomere biology disorders.* Prog Mol Biol Transl Sci, 2014. **125**: p. 41-66.

14. Kim, N.W., et al., *Specific association of human telomerase activity with immortal cells and cancer.* Science, 1994. **266**(5193): p. 2011-5.

15. Chen, J.L. and C.W. Greider, *An emerging consensus for telomerase RNA structure.* Proc Natl Acad Sci U S A, 2004. **101**(41): p. 14683-4.

16. Alder, J.K., et al., *Diagnostic utility of telomere length testing in a hospital-based setting.* Proc Natl Acad Sci U S A, 2018. **115**(10): p. E2358-E2365.

17. Boyraz, B., et al., *A novel TERC CR4/CR5 domain mutation causes telomere disease via decreased TERT binding.* Blood, 2016. **128**(16): p. 2089-2092.

18. Chen, J.L., K.K. Opperman, and C.W. Greider, *A critical stem-loop structure in the CR4-CR5 domain of mammalian telomerase RNA.* Nucleic Acids Res, 2002. **30**(2): p. 592-7.

19. Mitchell, J.R. and K. Collins, *Human telomerase activation requires two independent interactions between telomerase RNA and telomerase reverse transcriptase.* Mol Cell, 2000. **6**(2): p. 361-71.

20. Huang, J., et al., *Structural basis for protein-RNA recognition in telomerase.* Nat Struct Mol Biol, 2014. **21**(6): p. 507-12.

21. Kim, N.K., Q. Zhang, and J. Feigon, *Structure and sequence elements of the CR4/5 domain of medaka telomerase RNA important for telomerase function.* Nucleic Acids Res, 2014. **42**(5): p. 3395-408.

22. Jiang, J., et al., *Structure of Telomerase with Telomeric DNA.* Cell, 2018. **173**(5): p. 1179-1190 e13.

23. Nguyen, T.H.D., et al., *Cryo-EM structure of substrate-bound human telomerase holoenzyme.* Nature, 2018. **557**(7704): p. 190-195.

24. Kladwang, W., et al., *A two-dimensional mutate-and-map strategy for non-coding RNA structure.* Nat Chem, 2011. **3**(12): p. 954-62.

25. Tian, S., et al., *High-throughput mutate-map-rescue evaluates SHAPE-directed RNA structure and uncovers excited states.* RNA, 2014. **20**(11): p. 1815-26.

26. Robart, A.R. and K. Collins, *Investigation of human telomerase holoenzyme assembly, activity, and processivity using disease-linked subunit variants.* J Biol Chem, 2010. **285**(7): p. 4375-86.

27. Jady, B.E., et al., *Cell cycle-dependent recruitment of telomerase RNA and Cajal bodies to human telomeres.* Mol Biol Cell, 2006. **17**(2): p. 944-54.

28. Tomlinson, R.L., et al., *Cell cycle-regulated trafficking of human telomerase to telomeres.* Mol Biol Cell, 2006. **17**(2): p. 955-65.

29. Tesmer, V.M., et al., *Two inactive fragments of the integral RNA cooperate to assemble active telomerase with the human protein catalytic subunit (hTERT) in vitro.* Mol Cell Biol, 1999. **19**(9): p. 6207-16.

30. Kladwang, W., et al., *Standardization of RNA chemical mapping experiments.* Biochemistry, 2014. **53**(19): p. 3063-5.

31. Reuter, J.S. and D.H. Mathews, *RNAstructure: software for RNA secondary structure prediction and analysis.* BMC Bioinformatics, 2010. **11**: p. 129.

32. Leonard, C.W., et al., *Principles for understanding the accuracy of SHAPE-directed RNA structure modeling.* Biochemistry, 2013. **52**(4): p. 588-95.

33. Deigan, K.E., et al., *Accurate SHAPE-directed RNA structure determination.* Proc Natl Acad Sci U S A, 2009. **106**(1): p. 97-102.

34. Berman, A.J., A.R. Gooding, and T.R. Cech, *Tetrahymena telomerase protein p65 induces conformational changes throughout telomerase RNA (TER) and rescues telomerase reverse transcriptase and TER assembly mutants.* Mol Cell Biol, 2010. **30**(20): p. 4965-76.

35. Akiyama, B.M., et al., *The C-terminal domain of Tetrahymena thermophila telomerase holoenzyme protein p65 induces multiple structural changes in telomerase RNA.* RNA, 2012. **18**(4): p. 653-60.

36. Yoon, S., et al., *HiTRACE: high-throughput robust analysis for capillary electrophoresis.* Bioinformatics, 2011. **27**(13): p. 1798-805.

Chapter 3: Chemical mapping beyond P4P6

Introduction

RNA secondary structure is dictated by base pairing rules, with A-U and G-C being the canonical set of base pairs. However, RNA base modification, non-canonical base pairing interactions and the ability for RNA to be single stranded allow for multiple possible RNA secondary structures for any given sequence. While there must be one structure with a lowest energy minima, multiple structures frequently exist near this minima, allowing RNAs to exist as a structurally heterogeneous population. (**Fig 3.1**). Additionally, within the cell RNA is frequently bound to protein. Protein binding can bias RNA structure into an energy minima that was is favored by the free RNA. Therefore, when considering RNA structure, the lowest energetic state is not necessarily the only relevant structure. As such, both structural heterogeneity and dynamics must be taken into consideration when examining RNA structure.

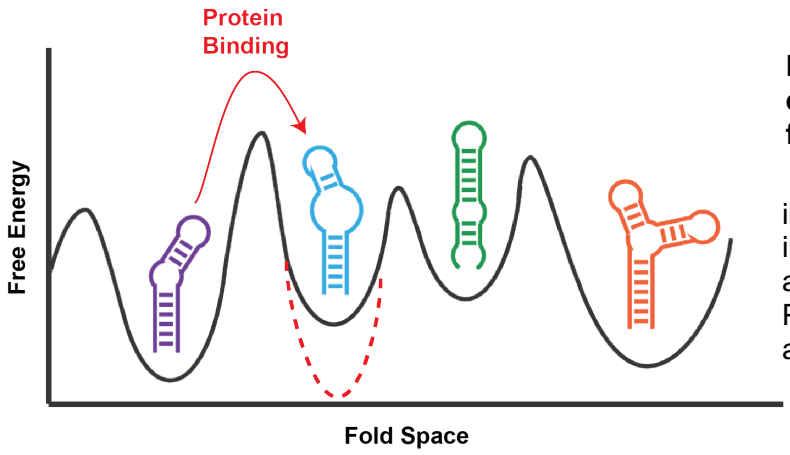


Figure 3 1 - RNA exists in a complex folding landscape. Cartoon schematics illustrate RNA existing in multiple secondary and tertiary structures. Protein binding can alter the landscape.

Structure determination can provide important clues to a bio-molecules Many techniques have been developed to examine nucleic acid structure. X-ray crystallography has provided many atomic resolution structures of nucleic acid and

nucleic acid-protein complexes. However, crystallography is limited in that it captures a static snapshot of systems that are frequently dynamic. Additionally, crystallography may provide incomplete pictures if a system is structurally heterogeneous. Nuclear magnetic resonance (NMR) allows atomic resolution of structure and also captures dynamics but it is severely limited in the size of the molecule that can be studied. Computational tools can query both structure and dynamics but results from these set of tools must always be verified using some experimental method and are additionally limited in the size of molecule and the time scale of potential dynamics. Single molecule techniques such as Forster Resonance Energy Transfer (FRET) provide information on a molecule's dynamics but can only provide low resolution structural information.

Chemical mapping is a structural prediction technique that relies on chemical probes that modify bases when they are *not* involved in a base pairing interaction. These probes are used while the RNA is in solution and can be used in a wide range of conditions or even within a cell, allowing for examination of RNA structure in a semi or fully native state. Identification of what areas of the RNA are single stranded provide valuable restraints when combined with computational secondary structure prediction, allowing for highly accurate secondary structure determination. Different chemical probes modify different bases.

The most commonly used probes are CMCT, DMS, and 1M7. DMS modifies A and C bases on the Watson-Crick face (**Fig 3.2A**) [1]. CMCT modifies G and U bases on the Watson-Crick face (**Fig 3.2B**) [2]. Hypothetically, by combining DMS

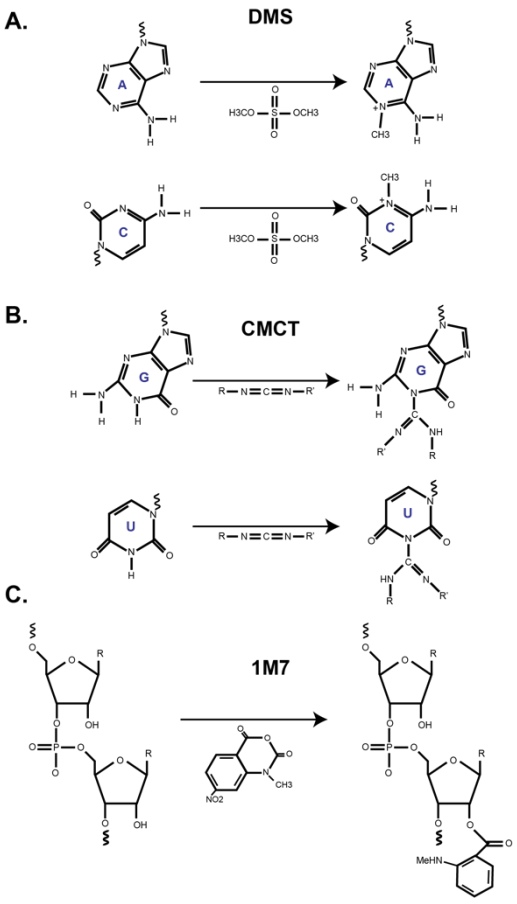


Figure 3 2 - Chemicals are used to modify RNA.

The mechanism of modification for *A.* DMS *B.* CMCT and *C*. 1M7.

and CMCT data base specific modification can be observed for each nucleotide. However, in practice, differences in optimal probing conditions and the fact that CMCT modifies G's with very low efficiency lead to regions of RNA that are predicted to be non-reactive when in fact the probing just does not speak to those bases. Additionally, because two probes are used and they have with different optimal conditions and quenches, high throughput use of these methods is challenging.

A major advancement in chemical mapping occurred when a non-sequence specific probe was developed [3]. Instead of modifying the RNA on the Watson-Crick face like CMCT and DMS, 1-methyl-7-nitroisatoic anhydride (1M7) modifies the RNA on the phosphate backbone (**Fig 3.2C**). It is only able to modify the backbones of bases that are flexible, and the flexibility of the backbone is related to whether that base is forming a base pairing interaction [3]. A number of 1M7-like molecules that interact at different time scales have been developed allowing time-resolved studies of RNA dynamics [4].

After a base has been modified by a chemical probe, a read-out of where that modification has occurred must occur. Reverse transcription has been commonly

used, as reverse transcriptases (RTs) will terminate the production of the cDNA product when it reaches a modified base. In this way a full array of cDNA products is produced that corresponds to where the RNA is single stranded. Initially detection of cDNA fragments was done using sequencing gels. Early work examining RNA structure, including seminal work done on the ribosome [5] [6], yeast tRNA [7] used multiple chemical probes, multiple reverse transcriptase start sites, and what I can only imagine was tens of thousands of hours in the lab and innumerable shattered sequencing gels. These studies laid the ground-work for the rules we currently use when thinking about RNA structure today.

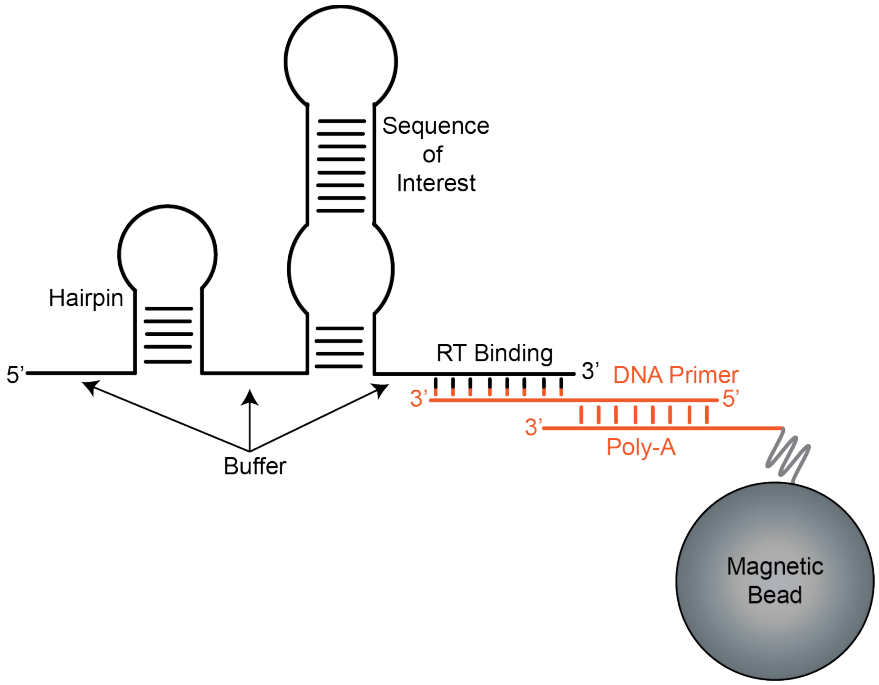
However, sequencing gels are limited in their size resolution, not to mention the technical expertise and time required to run them. The rise of the sequencing era has resulted in a number of new techniques that can be used to examine the cDNA population including capillary electrophoresis [8] and direct sequencing [9]. In this chapter, I describe a series of chemical mapping experimental protocols that implements chemical mapping using DMS, CMCT, and 1M7 with capillary electrophoresis used as the modification readout. Development of this experimental pipeline and the platforms used to analyze the data was done in Rhiju Das's lab at Stanford University. This technical and data analysis platform development is documented in a long string of papers out of the Das lab [10] [11] [12] [13] [14]. However, a singular, comprehensive document outlining the technical details required to implement the protocol, data analysis and troubleshooting does not exist. In this chapter, I will attempt to put together such a document allowing for the easy adaptation of this powerful technique into other lab's RNA structural characterization tool kit.

Chapter 3.1 - Chemical mapping experimental protocol

Construct design

Additional sequences must be added to the RNA of interest to preform chemical mapping. Given the potential for unwanted interaction between the sequence of interest and the appended sequences, a detailed description of why the sequences are added and how to minimize unwanted interactions is provided below.

Figure 3 3 - RNA construct design and immobilization scheme.
Additional RNA sequences necessary for structure probing are illustrated and labeled in black. DNA immobilization scheme is illustrated in orange.



Some of the RNA sequences added relates to the RNA immobilization scheme required for the exchange of buffers throughout the experimental protocol. RNA immobilization is achieved by linking the RNA to magnetic beads through a DNA primer (**Fig 3.3**). In addition to immobilization the region of DNA-RNA complementarity serves as the docking site for the RT. The docking of the RT will

melt any structure present immediately downstream of the binding region, therefore, a buffer of at least 4 nucleotides is required to separate the RT binding site and the sequence of interest. A buffer of at least 6 nucleotides at the 5' end of the RNA is also preferred, as the data generated at the very beginning and end of the run suffers in quality due to bleed over in signal from the longest and shortest bands. The RT binding site and the two buffer sequences are the minimal components that must be included in the RNA construct design.

Additional components that aid in data analysis are the inclusion of a hairpin, separated from the sequence of interest with a buffer (**Fig 3.3**). The hairpin serves as an internal positive control and allows for intensity normalization across experiments [10]. Placing the hairpin on the 5' end of the RNA also provides a nice 'my RT step worked well' sanity check.

The hairpin sequence should form a stable helix, and the loop section of the hairpin should include at least one A, C, G and U bases for accurate normalization if CMCT and DMS probing is going to be used. This will be explained further in the upcoming data analysis section. The buffer regions can be anything that does not interact with the RNA of interest. Before making RNA, run structure prediction on the full construct in RNAstructure to check for unwanted interactions. A final check to ensure a lack of interaction between the sequence of interest and add-ons is performed in the analysis of chemical mapping data. All buffers and the loop of the hairpin should be reactive in chemical probing and the helix of the hairpin should be non-reactive.

After finalizing a construct design enter the full sequence, plus a T7 promoter sequence on the 5' end, into Primerize (https://primerize.stanford.edu/) [15]. Primerize will automatically generate a series of overlapping primers that can be used in a PCR primer assembly reaction. One consideration to be aware of is that there is a certain chance of a base being skipped during primer synthesis. Because the primers are used as the template in the PCR reaction, a missed base will be incorporated into the template generated and into the final RNA. In a template made with only 4 primers no mistakes were made as read out by sequencing, however, in a template made from 12 primers many skipped bases were observed (data not shown). For longer RNA consider the use of ultra-mers which have a lower mutation rate. The PCR reaction, purification, and in vitro transcription are described in detail on the Primerize website - https://primerize.stanford.edu/.

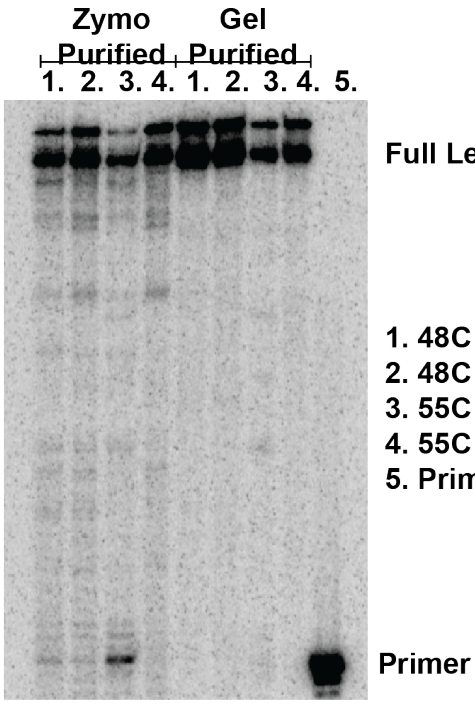


Figure 3 4 - High purity of RNA is required for clean RT extension. RT extension was tested using either Zymo purified RNA or gel extracted RNA in the presence of absence of RNasin. The difference in background noise of RT extension is shown on a sequencing gel.

One comment worth mentioning is the stringency required for RNA purity in order to obtain clean CE data. In complete transcript products found at a low level in a population of RNA will be amplified in the RT step (**Fig 3.4**). This

contamination will result in a low signal to noise ratio in the CE data. Use gel purification or AMPure Beads to purify in vitro transcribed RNA for the cleanest RT results.

Running the experiment:

Chemical mapping requires a series of highly timed steps. The protocol takes from 3.5 - 8 hours to run depending on the expertise of the user, some of the reagents are expensive, and the results are not received until 48 hours later. Therefore, troubleshooting is time intensive and frustrating. To minimize time spent agonizing over whether poor results are due to buffer contamination or preparation errors, prepare stocks of buffers, filter solution when possible, aliquot reagents into useful quantities, and store appropriately. An outline of every step during chemical mapping is described in Cordero et al. 2014 and therefore will not be reiterated in great detail here.

Chapter 3.2 - Analysis of chemical mapping data using HiTrace

The following is meant to be a 'I'm-a-wet-lab-scientist-not-a-programmer' description of the data analysis platform**.** An excellent and extensive tutorial put together by the Das lab is available in RiboKit, and described here [11]. While the Das Lab tutorial provides an excellent start for a data set that runs perfectly, outlined here are some thoughts on what to do when things run slightly less than smoothly.

The lines of code that are required to run this data processing pipeline are outlined below. The code presented in this chapter show extra input parameters that may be necessary during troubleshooting. Unlike the descriptive scripts presented in this chapter, an example script is also provided which can be run in MatLab without any adjustments. This dataset and script is located in the jazz data_share folder under Christina **JAZZ FILENAME**. A good way to learn this pipeline is to run through the example script using the example data. While executing the script read through this outline and the Das lab tutorial.

A: Before starting:

It is important for the user to provide a clear view of the data processing steps and save the analyzed data. Save a .m file outlining the scripts used to generate the data. Saving the .mat file which will save all the input parameters and data generated. Additionally, saving screenshots of steps of the analysis process and general conclusions in a powerpoint or other electronic document provides a reference file to point to for a lab notebook.

Download RDATKit, Biers, VARNA, and RNAstructure:

The programs below only work on Mac systems. To my knowledge they have never been used on Windows. Be sure each of these are downloaded into the same folder on your computer. Text written in `Courier` should be excecuted in the command line.

1 - Download RDATKit (https://ribokit.github.io/RDATKit/install/#MATLAB) by copying the text below into terminal:

```
git clone https://github.com/ribokit/RDATKit.git
```

2 - Download Biers (https://ribokit.github.io/Biers/install/) by copying the text below into terminal:

```
git clone https://github.com/ribokit/Biers.git
```

3 - Download VARNA (http://varna.lri.fr/index.php?page=downloads)

- Download the Applet one

4 - Download RNAstructure:

http://rna.urmc.rochester.edu/RNAstructureDownload.html

- Install the Text (Command-Line) Interface version
- **IMPORTANT**! There is also an RNAstructure associated with the RDATKit installed above. You also need this version and in the following steps you must point terminal to this folder, **not** the RDATKit RNAstructure folder.

5 - Next edit .bash_profile .

To find your .bash_profile use *one of the two* following options:

Finding .bash_profile (command line)
- Open the terminal and type: `ls -a`
  - This reveals all the hidden folders
- Open .bash_profile with the text editor (vim, etc) of your choice
- Add the below lines of code

Finding .bash_profile (GUI)
- Navigate to home folder (ie – usually your name)
- Type Shift+Command+. (dot)

- Edit .bash_profile to add the below lines of code

Edit .bash_profile to include the following lines of text:

```
export DATAPATH=/path/RNAstructure/data_tables/
export VARNA=/path/VARNA.jar
alias matlab=/Applications/MATLAB_R20xxa.app/bin/matlab
```

Replace text in red with:

1) 'path' in the code with the appropriate path to the target directory
2) VARNA.jar with the exact name of the VARNA app you downloaded
3) Replacing the x's in MATLAB_R20xxa.app with the correct numbers from your MATLAB version

Quit (don't just close) Terminal and then re-open .bash to confirm that you've added and saved correct code to your .bash_profile or .bashrc text file. Run this code in the terminal:

```
echo $DATAPATH
# Should return the path to path/RNAstructure/data_tables
echo $matlab
# Should return the path to your MATLAB application
```

A: Quick_look

Download the .abi files from capillary electrophoresis (CE) run into a folder. A good organizational practice is to save each experiment in a separate folder with the date the data was collected. The first script used to visualize CE data is quick_look. This script generates a series of five panels each portraying the data in different ways. See Step 1 of Das Lab Tutorial for images of the five panels or execute the first line in the example script to generate them yourself. The first panel provides a rough overlay of some of the data files with the ROX ladder in red and the FAM signal in blue. The second panel shows the unprocessed data in greyscale with the lanes in the order defined in 'trace_subset'. The third panel shows a rough approximation of the linear alignment of the bands. The final aligned data and

reference ladders are illustrated in the fourth and fifth panel respectively. All panels are saved in a 'Figures' folder.

To run quick_look execute the following line of code in the MatLab command window, replacing the variables in red with values appropriate to the experiment being run. Each variable in **bold and underlined** is described in detail below. Do not change the presence of absence of ' ' or the kind of brackets. They are important in MatLab. The example script provided does not include many of the inputs or outputs in the script outlined in the chapter. They are outlined here to aid with potential troubleshooting and may not be necessary for every analysis.

```
[d_align, d_ref, ylimit, labels] = QUICK_LOOK( {'dirnames'},
[ylimit], [trace_subset], signals_and_ref, dye_names,
lane_names, ['moreOptions'] )
```

dirnames (directory names) points the script to the folder containing the sequencing data. Be sure the 'Current Folder' window in Matlab is in the folder that contains the folder holding the data or the script will throw an error.

ylimit defines how far into the capillary electrophoresis run to search for signal. Initially allow quick_look to auto-determine an appropriate y-limit by leaving the bracket empty. The numerical range for the y-limit value should be determined by examining the y-axis of the second panel (Panel 2: All data) that Matlab generates, not the final Panel 4 that is most commonly viewed. Changing the y-limit may be necessary to obtain a better alignment or if full length product signal is low. This will be described in greater detail in the 'Examining data quality' and 'Further Alignment' sections.

trace_subset defines which lanes will be included in the analysis. Only one RNA sequence can be analyzed at a time, though multiple conditions in the same

RNA can be analyzed together. Select the sequencing lanes to be analyzed by inputting numbers separated by commas. The script parses through the file names and defines the order by the name of the well that the sample was in. For example, well A01 corresponds to 1, B01 to 2, C01 to 3 … A02 to 9, B02 to 10, etc. The simplest way to change the order of the lanes is to change the order of the numbers input in the brackets (ie - [4,3,2,1]). This will order the lanes with D01 first, C01 second, B01 third, and A01 first. However, the order can also be changed by adjusting the actual file names. After moving beyond the quick_look step, the lanes will have an associated number that corresponds to the new order, not the original data (ie- first number in the brackets is now 1, second number in the brackets is now 2, etc). It is best to place identical conditions, both saturated and dilute, next to each other. (ie - NoMod Saturated, NoMod Diluted, 1M7 Saturated, 1M7 Dilute, A, C, G, T). This format ensures that identical samples/conditions are next to each other in the subsequent alignment steps.

The **'moreOptions'** command can be used to look at the data without mean signal intensity normalization, without alignment or without leakage correction. The mean signal normalization that takes place can sometimes hide aspects of the data that may contribute to difficulties in processing. For this reason, it can be instructive to look at the data without this processing step. A non-normalized data set will appear as black square in panel 4, so visualization of data must be done by graphing the output d_align values as in **Fig 3.2B**. An example line of code including this option is commented out, but present, in the example script.

Quick_look processing and data manipulation:

Two main processing steps occur in quick_look: alignment and mean signal intensity normalization. Initial alignment is achieved through the signal generated by the ROX reference ladders. Careful alignment of the bands generated due to chemical modification is one of the most important aspects of the experiment and plays a critical role in the quality and reproducibility of the data between experiments. Normalizing to the mean signal intensity in each lane is necessary to iron out variability between overall intensity between lanes. Variability between lanes can be due to low RT efficiency, loss of beads during the chemical probing experiment, a faulty capillary electrophoresis run, elution efficiency, or poor primer-RNA binding to name a few.

It is important to be aware of how the data is being visualized in quick_look. Panels 1 and 2 show the actual intensity of the signal. Visualization in the rest of the panels is achieved using a sliding greyscale based on the highest and lowest intensity in a particular window. This is done lane by lane. Therefore the bands shown in panels 3 - 5 are not an accurate representation of the actual signal intensity in each lane. This is purely a visual adjustment and does not change the actual numerical values of the data. Analysis of the intensity of lanes should be done by looking at the actual values in the final d_align output as discussed below.

Examining the quality of the data after quick_look:

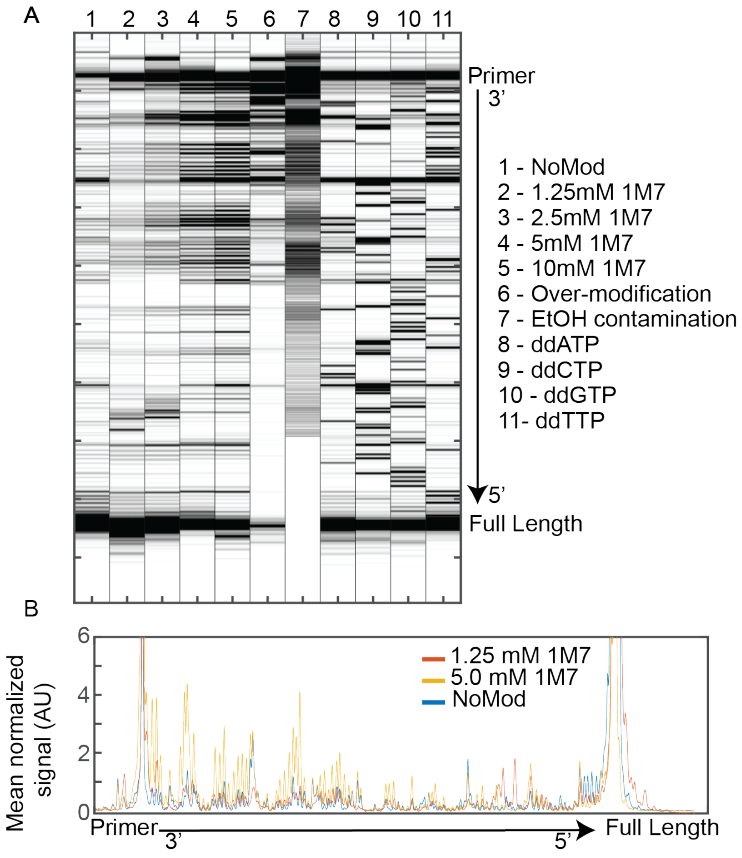


Figure 3 5 - Visual evaluation of high quality and low quality data using quick_look.

(A). Panel 4 from quick_look illustrating lanes examples of over-modification, low full length product, EtOH contamination as well as examples of high quality data. *(B)* Graphing d_align data reveals signal difference between NoMod, 1.25mM 1M7 and 5mM 1M7. These differences are not well portrayed in the quick_look panel due to sliding grey-scale visual adjustments.

The ladder lanes and chemically modified lanes should have high signal where there is a stop and be white in most other places (**Fig 3.5A - Lanes 4,5 and 8 - 11**). Failure to achieve this can be due to over modification, low full length product, high background to noise ratio, or ethanol (EtOH) contamination. Lanes that exhibit these qualities should be excluded from further analysis.

Over modification can occur if the concentration of chemical modifier is too high and many modifiers are added to a single RNA. This results in the RT being unable to reverse transcribe through to the end of the RNA. A signature of over modification is high signal at the 3' end of the RNA and a rapid trailing off of signal,

with very little signal in full length product (**Fig 3.5A - Lane 6**). Over-modification can be avoided by doing a chemical modifier titration (**Fig 3.5A - Lanes 2 - 5**). The lowest amount of modifier that still provides sufficient signal to noise ratio should be used.

Low full-length product can occur even in the absence of over-modification if the RT step is not occurring efficiently. Another sign of poor RT efficiency is having high background noise, even in the ladder and NoMod lanes. Optimization of the RT step may be necessary if this is observed.

Signal to noise ratio can be examined visually in the panel 4 output by Matlab. However, the sliding grey-scale intensity adjustment makes values in regions of very low intensity seem brighter than they actually are (**Fig 3.5A - Compare Lane 1,2, and 4**). Therefore, to visualize actual signal intensity, graph the d_align values (**Fig 3.5B**). In an experiment where the chemical modification agent is at sufficiently low enough to achieve approximately single hit conditions, all lanes should show the highest signal in the primer and full-length band. However, the signal generated by the modifiers should be well above that of the NoMod lane.

EtOH contamination is easy to spot as the lane will look smeary and grey (**Fig 3.5A - Lane 7**). There will also be no signal in the panel 5, the reference ladder channel.

B: Parameter Input

After determining which lanes to use for analysis a series of parameters are input that will be utilized throughput the remainder of the pipeline.

```
sequence = 'your_RNA_sequence_here';
structure = 'your_predicted_RNA_structure_here';
offset = length of 5' buffer
first RT nucleotide = length(sequence) - 20 + offset;
```

```
data_types = {'chem_mod_condition_saturated'},
'chem_mod_condition_dilute','ddATP', 'ddCTP', 'ddGTP', 'ddTTP'};
```

sequence is the RNA's sequence, including all buffers, hairpin and RT binding sequence. It is not possible to analyze the data without including all aspects of the RNA as this information is necessary in the band assignment step.

structure is the dot-bracket notation of your full RNA construct. This structure input will not be used in any aspect of the data processing or structure prediction and is simply present to guide the eye in subsequent band-assignment steps.

offset and **first_RT_nucleotide** are present for the numbering of the RNA. Offset should equal the number of nucleotides in the 5' buffer. First_RT_nucleotide takes the offset and the length of the RNA minus 20 (the length of the RT binding site) to give the value of where the first nucleotide of your sequence of interest should begin.

data_types is used in the band assignment step to place circles over where the ladder lanes should be. In the chemically modified lanes, circles are also placed using predicted structural information provided in the 'structure' parameter. The script recognizes certain strings so straying from the format in the example script will result in lanes and conditions not being recognized in the band assignment step.

C: Further Alignment:

An additional alignment step is often necessary. Unlike quick_look which uses the reference ladders, fine alignment takes the uses the intensity of the bands of interest for alignment.

```
align_blocks = {first_lane_number:last_lane_number};
d_align_before_more_alignment = d_align;
d_align_dp_fine = align_by_DP_fine(d_align_before_more_alignment,
align_blocks);
d_align_= d_align_dp_fine;
```

align_blocks defines which lane should be aligned to each other. Alignment is done based on the intensity of the bands so 1M7 should be aligned to only 1M7 lanes, CMCT to CMCT lanes, etc. If multiple conditions are being analyzed at once they should be annotated with commas separating each condition (ie - `align_blocks = {3:8, 9:14};`). Include both dilute and saturated samples. Ladders should not be included in any of the alignments. Visual assessment of this step and in band assignment is helped significantly if all conditions were grouped together in quick_look. If a different alignment technique is necessary don't forget to re-run quick_look to reset the original d_align variable.

Troubleshooting alignment:

A good alignment is critical to ensure accurate quantification of band intensity across lanes. If the alignment is poor, the line on which the Gaussian will be fit will cover a band in one lane but not the other in the band assignment step (**Fig 3.6**). This will be interpreted incorrectly as variability between lanes or conditions.

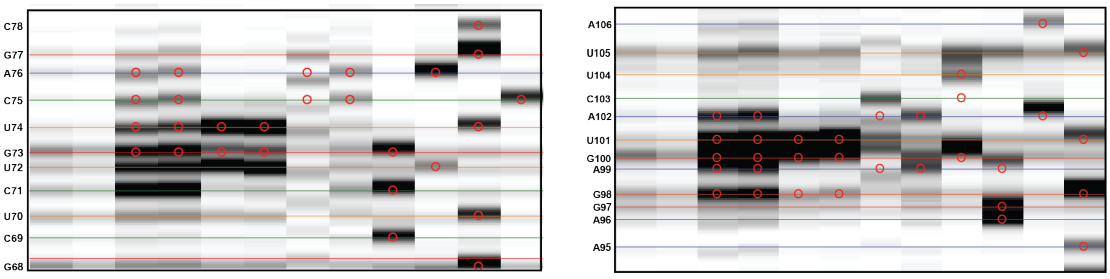


Figure 3 6 - Examples of mis-aligned bands viewed during band assignment. Poor alignment results in the inability to place a single line over the middle of the band resulting in poor Guassian fit in the next steps.

The best place to check for a good alignment is in the band assignment step. Frequently alignment within a single conditions will look good, but between modifiers or between dilute and saturated samples alignment will be poor. If the user is only

compromising the integrity of a few bands in the whole RNA it is probably fine to process the whole data set, but if the user is consistently compromising the placement of a band for one modifier over another (**Fig 3.6**) try some of the following troubleshooting techniques.

Steps that **do** achieve a better alignment:

1) Expand the y_limit in the quick_look step to include more reference ladders. This will often make the alignment better especially at the top and bottom of the lane.
2) Remove lanes bad lanes. The script is searching for the most intense bands across all lanes. If one lane has a very different signal profile, or higher background than the others, the bad lane will throw off a good alignment for the rest of the lanes.
3) Process CMCT, 1M7, and DMS data separately.

Steps that **do not** (in my experience) achieve a better alignment:

1) Two other alignment scripts are described in the HiTrace tutorial align_by_DP() and align_by_DP_using_ref(). Align_by_DP_using_ref() is already used in the quick_look alignment. These two scripts have not resulted in better alignment results in my hands.
2) Running the alignment multiple times.

Data the exhibits poor alignment is likely of poor quality and even if is not of poor quality the resulting Gaussian fits will be of poor quality. If necessary quick_look and subsequent data analysis can be run on a single condition (saturated, dilute, and ladders). Bands can be assigned for each condition individually and datasets merged

after guassians have been fit to the bands. In this way the user can be confident that each band is in the optimal place across all lanes.

D: Band Assignment

After a proper alignment and parameter input band assignment should be straight forward, if slightly labor intensive.

```
xsel = []; clf;
[xsel, seqpos, area_pred] = annotate_sequence(d_align, xsel,
sequence, offset, data_types, first_RT_nucleotide, structure);
```

xsel defines a new variable and clears whatever panel is currently selected in Matlab.

annotate_sequence generates an interactive panel on which the user will assign bands to a location on the data set. The output of this step is the coordinates of the assigned bands.

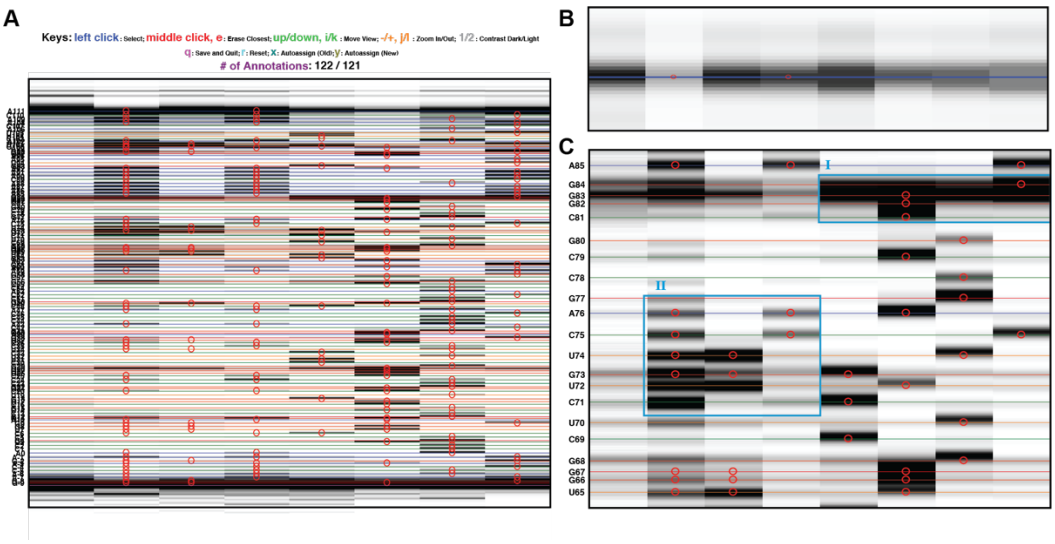


Figure 3 7 - Band assignment in HiTrace.

(A) Full panel view of the initial band assignment panel. Circles are placed over ladder in a sequence dependent manner. Navigation panel is shown above. *(B)* When assigning bands, the line should be placed in the exact middle of the signal profile. *(C)* I: Guanine repeats run in a compressed manner in CE. II - Circles are place in a structure parameter dependent way for 1M7 and a sequence AND structure parameter dependent way for CMCT and DMS. Circles in chemically modified lanes are NOT placed in an intensity dependent manner

The navigation commands are at the top of the panel (**Fig 3.7A**): left and right keyboard arrows zoom in and out. Up and down arrows are used to move up and down the panel. 1 and 2 are used to change the intensity of the bands displayed. 'e' will erase one band and 'r' will erase all bands. 'q' will save your progress. At the very top of the panel is '# of Annotations:'. Annotations include both the primer and the full length product so the number of annotations should always be one more than the number of nucleotides in the sequence (ie - '# of Annotations: 122/121). If the number is not sequence length + 1 a band somewhere has been deleted or added. It is can be difficult to track where the deletion or addition occurred so when this happens it's best to start over with the band assignment.

Assign the first and last band by zooming in so only two or three bands are in view. Decrease or increase the band intensity so that the middle of the band is nearly saturated (**Fig 3.7B**). Click on the exact center of the band. A colored line will appear. Navigate to the bottom of the gel and repeat with the bottom band. Click 'x' and all the bands in the between the top and bottom band will be assigned.

Red circles are placed over each location where intensity should be present in the ladder based on the sequence parameter input (**Fig 3.7A,B, and C**). If a DMS or CMCT conditions is provided it will only place circles on areas that are predicted to be unstructured and also contain an A or C for DMS or G or U for CMCT (**Fig 3.7C**) If 1M7 was used all regions predicted to be unstructured will have a circle. Ladder assignment and the sequence annotated on the side already takes into consideration the fact that cDNA is the reverse compliment of the actual sequence and also that the RT stops one nucleotide before the modification.

If the red circles are not appearing accurately over the ladder either the sequence input is incorrect, the wrong RNA is being probed, or the bands are too weak to be assigned properly. In regions where there is high G content the separation between the bands may be poor (**Fig 3.7C I**), however, directly before and after such areas the ladder should match up well, giving the user confidence in the placement of the bands. The accuracy of the circle placed over the over the chemically modified lanes may or may not be correct depending on the accuracy of the structure prediction input (**Fig 3.7C II**).

Autoassign places lines on the band that is most intense. This often results in the band being placed on the ladder. Where the band is placed is where the guassian will be fit and where data will be collected, therefore, proper band assignment technique is ***critical***. To ensure proper band assignment go through and check/adjust where each band has been placed by hovering the cross hair over the band, pressing 'e' to erase and left clicking to replace in the correct location. Place each band in the same way as the top and bottom band were placed with a zoom in such that only five to eight bands are visible, adjusting the intensity such that the exact middle of the band is slightly undersaturated. In regions with very low reactivity it may be difficult to decide where to place the line, even after increasing the intensity of the band. In this case use the ladder to guide the placement. In these cases the calculated intensity of the band will be very low regardless of where the line is placed. Move through the entire tapestry in this way. Get into some state of zen (or crack open a beer) and enjoy the process. When you have placed each band press 'q' to save the assignments.

At this time running `xsel = []; clf;` will clear all the band assignments the user just so carefully made. The tapestry can be saved at any point during the assignment process by pressing `q` and can be re-opened by running the *only the second line of code* again.

F: Fit to Guassian:

The next line of code fits a guassian to the region where the line was placed (xsel) and quantifies the band intensity. **area_peak** is the output that contains the fit intensity values for all bands.

```
[area_peak, darea_peak] = fit_to_gaussians(d_align, xsel);
```

G: More Parameter Input:

The following are parameters that are used for data normalization.

```
saturated_idx = [1,3,5,7]
diluted_idx = saturated_idx + 1;
saturated_array = area_peak(:, saturated_idx);
diluted_array = area_peak(:, diluted_idx);
saturated_error = darea_peak(:,saturated_idx);
diluted_error = darea_peak(:,diluted_idx);
bkg_col = [1,1,1,1];
ref_segment = 'GAGUA';
ref_peak = get_ref_peak(sequence, ref_segment, offset);
sd_cutoff = 1.5;
```

saturated_idx defines what lanes are the saturated. In this instance there is one saturated NoMod lane and three chemical modification lanes. Change the numbers in brackets depending on the number of lanes that you have. This script setup relies on the dilute samples being positioned directly after their saturated partner in the quick_look step. If this is not the case the lane number can be defined manually with brackets like saturated_idx is.

bkg_col defines the background lanes. Use ‘1’ if your NoMod lane was in lane 1, ‘2’ if your NoMod lane was in lane 2, etc. There should be equal number of

bkg_col to number of lanes in saturated_idx. If multiple background conditions were used, change the numbers accordingly.

ref_segment defines the sequence of the hairpin. Be sure ref_peak only returns 5 numbers. If the hairpin sequence is present in multiple areas of the sequences the normalization will not work. If the user changed the hairpin sequence in the RNA construct design, change ref_segment to the appropriate sequence.

H: Calculate Normalized Reactivity:

A series of panels is generated as you click through each of the normalization steps allowing for visualization of how the data is being manipulated. A detailed description of how normalization is calculated is provided by the Das lab [10]. The only parameter to adjust in the below script is the number of chemical modification lanes that are present. Include a lane for both the saturated and the dilute lanes.

```
[normalized_reactivity, normalized_error, seqpos_out] =
get_reactivities(saturated_array, diluted_array, saturated_error,
diluted_error, bkg_col, ref_peak, seqpos, [],
{'nomod_saturated','nomod_dilute','1M7_saturated','1M7_dilute','repe
at_with_conditions_as_necessary'}, sequence, offset, sd_cutoff);
```

Three normalization steps:

Data normalization is described in detail [10] and therefore will not be reiterated here. Briefly three normalization steps take place: saturation correction, attenuation correction, and hairpin normalization.

I: Average Across Replicates:

Change the name of the output (**d_1M7/da_1M7**) depending on the reactivity being calculated. If multiple chemical modifications were analyzed this is the step where they are split and placed in different arrays. Change the lane numbers to only encompass the lanes that contain the conditions to be analyzed.

```
[d_1M7, da_1M7, flags] =
average_data_filter_outliers(normalized_reactivity(:,
[lane#:lane#]), normalized_error(:, [lane#:lane#]), [], seqpos_out,
sequence, offset);
```

If CMCT and DMS data are being analyzed this script merges the DMS and CMCT reactivities such that if an A or C is present the DMS reactivity at that sequence position is used and if a G or U is present the CMCT reactivity is used. The new combined normalized reactivity array is output at **d_DMS_CMCT**

```
for i = [1:length(sequence)-20] ;
    if sequence(i) == 'A' || sequence(i) == 'C';
        d_DMS_CMCT(i) = d_DMS(i);
    else
        d_DMS_CMCT(i) = d_CMCT(i);
    end;
end
d_DMS_CMCT = transpose(d_DMS_CMCT)
```

J: Predict Secondary Structure:

rna_structure uses the calculated normalized reactivity profile (**d_1M7** or **d_DMS_CMCT**) and the sequence to predict an RNA secondary structure. If a nucleotide is highly reactive a penalty is placed on structures that place that nucleotide in a base pairing interaction. Change **structure_WT** and **bpp_WT** to an appropriate name describing the RNA construct and the chemical modification used. These steps will only work if MatLab is started via terminal. If bash is not set up to summon Matlab, an alternate method to start MatLab from the terminal is to type the path to it in your terminal (ie -

```
/Applications/MATLAB_R2017a.app/bin/matlab)
```

```
[structure_WT, bpp_WT] = rna_structure(sequence, d_1M7, offset,
seqpos_out, [], 100, 0);

output_varna('WT', sequence, structure_WT, structure, structure_WT,
offset, [], [], [d_1M7; nan(20, 1)], bpp_WT);
```

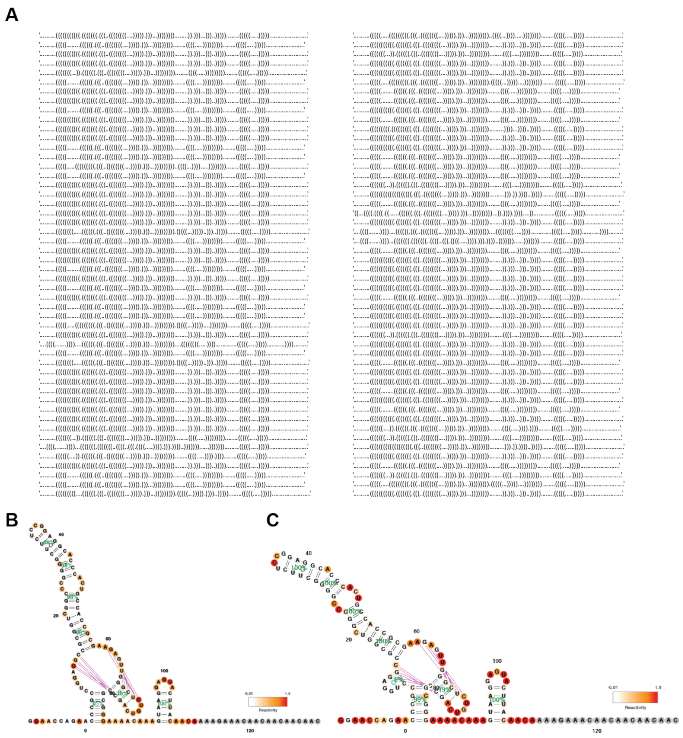


Figure 3 8 - Visualization of output of rna_structure and VARNA.

(A) Bootstrap sampling generates 100 dot-bracket predicted structure. *(B)* 1M7 and *(C)* merged CMCT-DMS data visualization using VARNA. Reactivities are mapped in from white to red. Difference in predicted structure vs input structure are visualized with pink lines.

During this step the data is being randomly sampled and bootstrapping is used to determine a confidence parameter for each helix predicted. This sampling is repeated **100** times. After this step is complete a series of structures, annotated in dot bracket form, is output in the MatLab command window (**Fig 3.8A**). In addition, the top predicted structure is saved as a **structure_WT** variable and the base pairing probability for each nucleotide is saved as **bpp_WT**.

The next line of code uses the structure_WT and bpp_WT variable along with many of the other parameters entered previously to visualize the top RNA structure with VARNA (**Fig 3.8B and C**). The output image is the best predicted secondary structure using the mapping data as weights! Congratulations! Take a moment to admire your work (or stare in horror/confusion at the monstrosity you've created). Check to ensure all buffers and hairpins are displaying appropriate reactivity. Do not ignore unexplained reactivity in hairpin helixes, or lack or reactivity in the buffer

regions. They may alternate structures. Keep in mind that the output structure is only a model and does not reflect the only structural model generated.

VARNA is a nice tool for RNA visualization but a little bit finicky. As bootstrapping may have generated structures other than the top hit output in structure_WT the user may wish to visualize reactivities mapped onto other structures. Unfortunately manipulation of the structure after VARNA is called will make the reactivities disappear. An easy way to change the structure used by VARNA is to change the structure_WT variable. However, often times other visual modifications may be desired. Easy-ish custom visualization in VARNA is possible with the addition of the line of code - `disp(command_without_output)-` into line 181 of the varna_fig script. This generates a line of text in the MatLab command window. This text contains everything necessary to generate the VARNA output and the sequence, structure, and color parameters can be modified in the text file. After the residual /n's have been removed from the text the script can be pasted into command line to call VARNA and generate an image with the desired modifications.

To remove pink lines that show differences in the structure prediction you provide in the 'sequence' variable, set sequence = [ ]. More interesting visualization options are available in VARNA using the right mouse button.

Conclusions and Discussion:

1D chemical mapping has been used for decades to examine RNA structure. It is a powerful tool and with recent technological developments is amenable, to some extent, to high-throughput techniques. However, RNA construct design is one aspect of the experimental pipeline that prevents this technique from being truly high-throughput. Of the three RNA constructs that were initially probed in this experiment,

two of them had some sort of interaction between the buffer and the sequence of interest. It is not responsible to use the same buffer sequence and RT sequence for every RNA. Careful thought with regard to construct design is required. A script that optimizes the design of the buffer region would increase the throughput of this step.

Chemical mapping works well on structured RNAs. But one aspect of RNA structure prediction that is lacking in the field is structure prediction of heterogeneous RNA populations. Chemical mapping produces an average reactivity values at each nucleotide. If multiple secondary structures are present the reactivity at each nucleotide will not fit any of the actual structures well and structure prediction suffers. One aspect of the data analysis that could be potentially improved is the weighting system by which reactivities are implemented into the structure prediction.

While prior work by the Das lab has reported no change in structures predicted upon changing the reactivity values in magnitude, this may not be the case for heterogeneous RNAs. RNA structure without any weights predicts the CR4/5 with each helix being called with 100% confidence (**Fig 3.9A**). When secondary structure is predicted with penalties applied based on reactivity, P6.1 is predicted with a fairly high confidence, despite having significant reactivity on one side of the helix (**Fig 3.9B**). When the reactivity is multiplied by 1.5 an entirely new structure is predicted at the junction (**Fig 3.9C**). This new structure, termed P6.1 appears to fit the reactivity profile better than the P6.1, but only RNA structure needs to apply a severe penalty on the structure predicted to move away from the ‘no weights’ structural prediction. It is simple to say that likely neither of the structures is accurate and the uncertainty is a property of a heterogeneous system. But for chemical mapping to move beyond the realm of qualitative structural prediction a better way of

deconvoluting RNA structural heterogeneity through computational approaches is necessary.

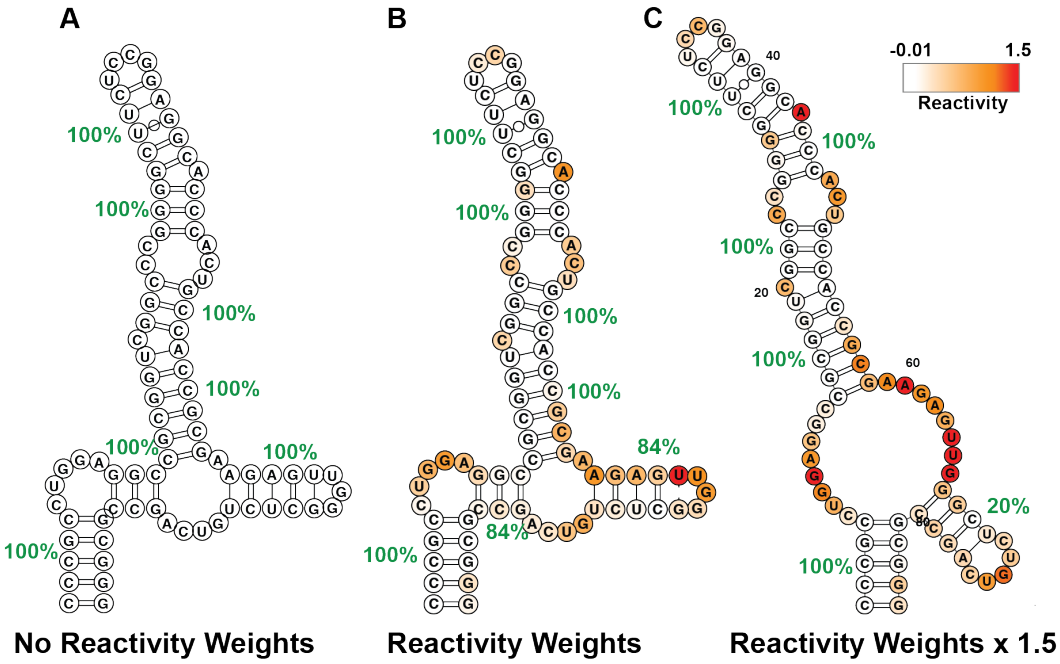


Figure 3 9 - Reactivity magnitude affects structure predicted in heterogeneous data sets.

RNAstructure prediction with (A) no 1M7 weights (B) 1M7 weights (C) 1.5x 1M7 weight. Bootstrapped derived confidence is shown by each helix in green.

Some potential solutions to this problem have been presented in the literature. In particular the Illumina sequencing version of this protocol multiple probes can be added to the RNA as the chemical modification is read out as a sequencing error rather than a stop in the RT. In this way if multiple structures are present, classes of probes should fall together as a population, allowing for better secondary structure classification. Whether this will work in practice for the CR4/5 domain remains to be explored.

Chapter 3.3 - Mutate-and-Map (M2)

Introduction:

The premise behind Mutate-and-Map is simple. 1D analysis provides data on regions that are unstructured. By mutating through the RNA, base pairs that were previously in a helix should be released when their partner is mutated. In this way a region of the RNA that previously reported negative data is now providing exact information on it's base pairing partner. However, RNA folding landscapes can be complicated and unexpected folding events can occur even with single base changes (**Fig 3.10**)

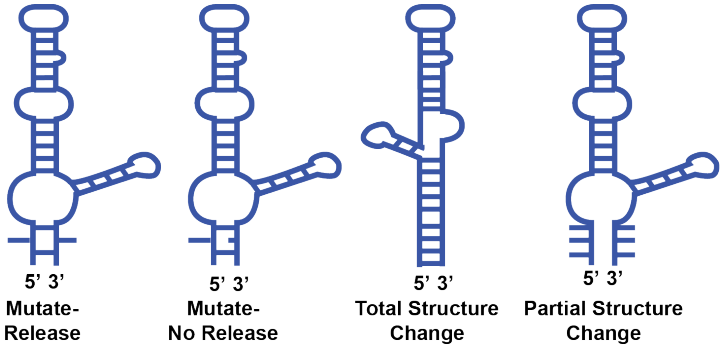


Figure 3 10 - Point mutations can cause unexpected structural changes due to RNA's complex folding landscape.

M2 making the RNA/executing the experiment:

The technical challenge of generating a large number of point mutants is circumvented using primer assembly. Follow the excellent protocol put together by the Das lab on the Primerize website when ordering plates of primers for an M2 experiment. (https://primerize.stanford.edu/design_2d/). A series of .xls sheets is output that can be directly input into IDT's website.

A few notes:

1. Adjust the sequence offset and mutation start and end position in primerize (advanced options) so the first mutated nucleotide listed as 1, or whatever

numbering system is used for the RNA of interest. The 5' buffer that is present will mess up the numbering if this is not corrected.

2. Don't forget to select wet as the shipping option. Shipping oligos dry significantly decreases the cost but at the expense of having to resuspend (likely hundreds) of oligos by hand. There is a high potential for error and/or contamination associated with this and therefore the monetary cost is offset.
3. Order extra of the WT primers (present in A01 in all plates). The user must fill the regions of the plate that utilize the WT primer. Alternately the user may fill in the WT sequences for the rest of the plate. This will result in the easiest assembly but will significantly increase the cost of the plates.

After receiving the plates of primers assemble the primers using the same protocol as a single primer assembly reaction. If on all plates, assembling all the A01 primers will give a WT RNA, B02 assembly will make the first mutant, C02 the second, etc. Purify PCR product purification using AMPure Beads (Fisher Scientific #NC9933872) according to manufacturer's instructions. Check the concentration of each PCR purified product using Nanodrop. Failure rate tends to be low and very few reactions should need to be repeated.

The purified PCR products are used as the template for an in vitro transcription reaction as outlined in the 1D analysis portion. Run a 50uL reaction volume. After in vitro transcription purify RNA with AMPure beads. Check RNA concentration with Nanodrop and repeat reactions that failed. Again failure rate should be low. Check purity of RNA on a denaturing PAGE. Apply the same level of stringency for RNA purity as discussed in the 1D section of this chapter (**Chapter 2.1 - Figure 2.4**).

M2 Chemical Mapping

Prepare a plate of RNA at 1.2uM. 2uL of this reaction is used in a 15uL reaction. The protocol is described in detail [16]. The same workflow is used as 1D analysis. Ladders made from the WT sequence may or may not be utilized, but are not necessary for band assignment or analysis, and must be removed if used before data analysis. In addition to the plate of mutants it is best to include one or two additional WT samples and a NoMod condition on the plate as well. It is not necessary to submit plates of diluted samples as a true normalized reactivity will not be calculated for analysis.

M2 Analysis

M2 analysis is executed in a similar manner to the 1D analysis described in the prior section with the following differences:

A: Before starting:

No additional programs or downloads are required.

B: Initial visual inspection of data quality

Run quick_look as described before. Instead of removing poor quality lanes replace them with WT. Removing the lanes will mess up the sequence annotation in the coming steps. Not replacing them with WT will significantly hinder alignment.

C: Parameter Input

Define parameters as before with the addition of running the line to replace T's with U's. If this line is not run an error regarding sequence identity will be displayed in later steps. Instead of data_types written in strings the data_types are a number/lane reflecting where the mutations begin. If more than one WT lanes are used at the beginning of the experiment change the ***2*** in the line - for i = 2:85; - to start the numbering where the mutation count begins.

```
sequence = 'your_RNA_sequence_here';
sequence = strrep(sequence, 'T', 'U');
structure = 'your_predicted_RNA_structure_here';
offset = length of 5' buffer
first RT nucleotide = length(sequence) - 20 + offset;


data_types{1} = 'NaN';
for i = 2:85;
    data_types{i} = [num2str(i - 1)];
end;
```

D: Further Alignment:

Further alignment is the same as 1D analysis.

E: Band Assignment

The interactive panel is generated as before. This line of circles moving from the bottom left to the top right indicate the mutations made and highlight where reactivities should be found. The other set of circles indicate where release events should occur based on the input structure.

There are no ladders present in this experiment though they could be used. If ladders were used they must be removed after the fit_to_guassian step to avoid generating errors in the data annotation step. As before bands must be placed carefully to encompass as many of the reactivities as possible. In regions of structural heterogeneity it can be difficult (impossible) to choose line line that encompasses all reactivities. This will be discussed further in Discussion and Conclusions.

In an M2 experiment a normalized reactivity is not calculated. Instead the absolute reactivity value at each nucleotide is compared across all mutants. Because no normalization occurs the intensity of the full length and primer band is not needed. Therefore, at the end of band assignment, delete the full length band assignment. Failure to do so will result in very different calculated signal intensity and will mess

up comparison between experiments. The error can be caught by noticing an error message regarding band annotation vs data annotation generated when running data_annotation in the 'Output to RDAT' section. Unfortunately this error will not terminate the data_annotation process and therefore is easy to miss.

F: Fit to Guassian:

This is the same as before.

G: Output to RDAT

Instead of inputting more parameters and calculating a normalized reactivity the raw area_peak is used for analysis. This is possible because each nucleotide will be compared to itself across all constructs. It is not possible to calculate a true normalized value unless the user also submitted a dilute sample of the plate, enabling accurate quantification of the signal at the full length and primer. However, scripts were written that execute attenuation correction (atten_corr2.m) and hairpin normalization (norm_hairpin.m). This allows for the calculation of an approximate normalized intensity. These scripts are in the **jazz data_share** folder. No large change in data quality or data interpretation was observed when these corrections were made (data not shown).

Instead of using normalized reactivities to analyze data a mutation profile tapestry will be generated and submitted to both quantitative and qualitative analysis. To make this tapestry generate an rdat file. Input rdat parameters by revise the comments and filenames to suit the experiment.

```
filename = 'name_of_rdat.rdat';
name = 'area_peak';
comments = {'write_about_the-experiment_done'};
annotations = { 'experimentType:MutateAndMap',
'chemical:buffer_and_salt_condition',
'chemical:MgCl2_concentration',
```

```
'chemical:additional_chem_modifications', 'temperature:RT',
'modifier:SHAPE'};
```

Data annotation can be done using one of the two the loops below. In the first loop no additional input is required. In the second loop a .txt file is required. This .txt file is output by primerize when the primer plates are generated. Make sure the .txt file is in the folder shown in the 'Current Folder' panel on Matlab.

```
% loop generates a cell of mutation labels, e.g. U55A
data_annotations{1} = 'mutation:WT';
for j = 2:size(area_peak, 2);
    data_annotations{j} = {['mutation:', sequence(j - 1 - offset),
num2str(j - 1), DNA2RNA(complement(sequence(j - 1 - offset)))]};
end;

% loop generates a cell of mutation labels based on .txt file
construct_names = read_constructs('180817_primer_keys.txt');
for j = 1:size(area_peak, 2);
    data_annotations{j} = {['mutation:',
strrep(strrep(strrep(construct_names{j}, 'T', 'U'), 'WU', 'WT'),
'Lib1-', '')]};
end;
```

Finally, an rdat file is generated using the script below.

```
output_workspace_to_rdat_file(filename, name, sequence, offset,
seqpos, area_peak, structure, annotations, data_annotations,
darea_peak, [], [], [], comments);
d_rdat = show_rdat(filename);
```

A panel is generated that shows your rdat. To visualize only your sequence of interest use show_rdat_ROI.m **(JAZZ**). The user can change the region of interest by changing the offset_fivePrime and offset_threePrime variable at the beginning of the script. The greyscale intensity range can be changed by adjusting the colormap found on line 81 of the script.

H. Analysis of M2 Tapestry

Observing where known mutations induce changes elsewhere in the tapestry can validate potential helices. There are two important aspects of M2 analysis: visual and quantitative.

The strongest evidence for a base pairing event is the mutate-release event. An ideal release event occurs only when the base pairing mutation is made, and is not reactive when other mutations are made (**Fig 3.11**). This event is a powerful indication that a particular base pair is formed. Even a few clean events provide strong evidence for a helix only one particular structure is likely to contain a particular set of base pairing interactions.

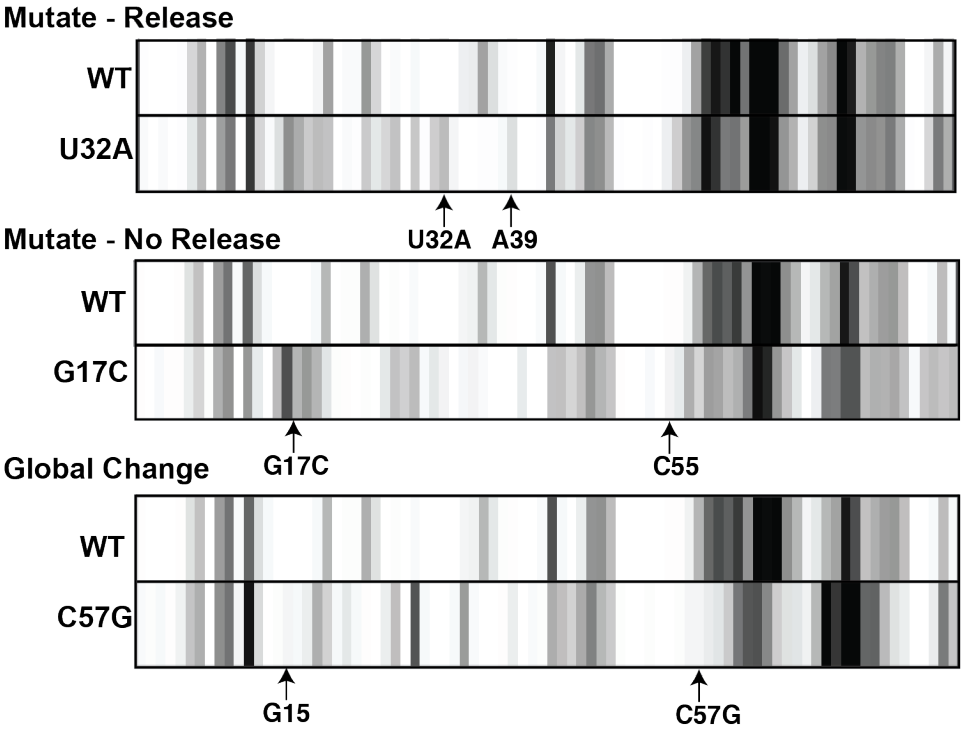


Figure 3 11 - Mutation to RNA gives rise to different structural perturbations.

Three kinds of structural perturbations are illustrated: Mutate and clean release event. Mutate and no release event. Mutate and global changes in RNA structure.

Other classes of perturbations can be seen in the M2 tapestry including an increase in reactivity in the mutation, but not the base pairing partner (**Fig 3.11**). This could be evidence against the predicted helix. Alternately, because 1M7 modifies the

phosphate backbone based on the level of flexibility, if the helix is still intact the backbone may not be flexible enough to be modified.

Another form of perturbation that is observed is a global structural change of the RNA. While not particularly informative with regards to the presence or absence of a predicted helix, these mutations may hold interesting information on the folding landscape of RNA.

Many of the mutations made will fall somewhere within these observed perturbations. Many things must be taken into consideration in a very complex landscape of reactivities. This complexity can make confident assignment of perturbation events challenging. Therefore, it is important to pair visual analysis with quantitative analysis. To do this a Z-score is calculated [17] by taking the mean intensity of a band across all nucleotides and looking at which nucleotides exhibit reactivity that fall outside this mean.

Calculating Z-score

To calculate a Z-score run the following. Make sure **filename** is pointing to the correct rdat. The .txt file generated contains the values for calculated Z-score. Filtering of the Z-score can be done at this point. No large difference in analysis or outcome has been obsevered using the current regiment of Z-score filtering scripts.

```
Z = output_Zscore_from_rdat('hCR45.txt', {filename});

% filtering of Z score
Z_cutoff_mean = Z;
for i = 1:size(Z_cutoff_mean, 1);
    Z_cutoff_mean(i, Z_cutoff_mean(i, :) >= mean(Z_cutoff_mean(i,
:))) = 0;
end;

Z_cutoff_1std = Z;
for i = 1:size(Z_cutoff_1std, 1);
    Z_cutoff_1std(i, Z_cutoff_1std(i, :) >= -std(Z_cutoff_1std(i,
:)) + mean(Z_cutoff_1std(i, :))) = 0;
```

```
end;
```

Predicting RNA structure based on Z-score

The Z-score is implemented into RNAstructure to predict a most-likely secondary structure model in the same way structure was predicted in 1D analysis. Perhaps more informative than the VARNA model, a panel showing the base pairing probability matrix and a visualization of the Z-score is generated.

```
[structure name, bpp name] = rna_structure(sequence, [], offset,
seqpos, Z, 100, 0);

output_varna('2D_hCR45', sequence, structure name, structure,
structure name, offset, [], [], [], bpp name);

print_bpp_Z(bpp name, Z, -15, 'title_of_graph');
```

Conclusions/Strengths/Limitations of M2

The power of the M2 lies in the ability for the user to create many mutations and find a few pieces of evidence per helix that rises above the intrinsic noise of the experiment. Mutate-and-Map works for well structured helices, but largely fails in regions of the RNA that are structurally heterogeneous. These regions of the tapestry exhibit high reactivity and therefore, upon mutation, still exhibit high reactivity. Z-score filtering actually excludes regions that contain a high degree of reactivity. The script can be modified to include these regions but the resulting analysis is often muddled and inconclusive.

Alignment is again one of the most important aspect of the M2 analysis. In an M2 experiment each RNA contains a different sequence and will hypothetically exhibit different reactivity profiles. This results in a degradation of alignment quality in the region where changes are seen. Unfortunately, the regions that change often

contain the information the user is looking for and the areas that need the most accurate alignment/quantification. Many potential release events are missed in the base assignment step (**Fig 3.12**). Because of this, in its current form M2 analysis dramatically under-estimates the number of release events that can be quantified. Currently there is no work around this lack of alignment and as a result accurate quantification of changes in heterogeneous regions suffers. Some potential strategies to improve alignment include having a base-by-base ladder labeled with a different fluorophore present within the lane that contains the reactivities. This would potentially allow for a base-by-base alignment across all lanes in the initial quick_look step. An alternative approach involves implementing an approach used by the SAFA software in which the user can manually trace where the bands lie.

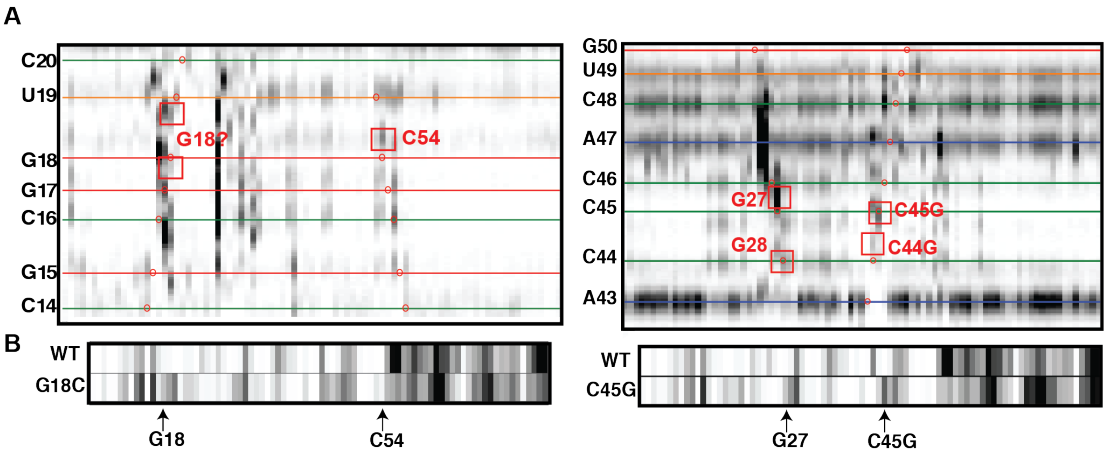


Figure 3 12 - Mis-alignment during band assignment step results in missed release event in the rdat.

(A) Two examples of regions of heterogeneous structure in the M2 band assignment step. Red boxes illustrate likely missed release events due to mis-alignment. *(B)* The processed RDAT output confirms that release events will not be calculated or observed.

Testing RNA structure using mutation profiling generates a lot of data. Only specific kinds of disruptions will lend evidence for the presence of a helix. Another

interesting aspect of this data set are point mutations that induce a large conformational change. While not helpful in determining the structure of the RNA of interest, point mutations that dramatically alter the folding landscape provides interesting information on RNA folding. These regions of the RNA structure that serve as a corner stone for the nucleation of a particular RNA structure may provide interesting targets for therapeutics. An interesting study would be to look at these regions across multiple RNAs and see whether they correspond to known disease causing mutations.

The increasing number of M2 data sets available through the Das lab and other provides a wealth of information not previously accessible for scientists studying computational RNA structure prediction. This aspect of RNA folding was not analyzed in the scope of my work but provides exciting opportunities for scientists interesting in de novo in silico prediction of RNA folding.

Chapter 3.4 - Mutate-Map-Rescue (M2R)

Introduction:

Scientists often discover how biological circuits and machines are put together by breaking them. They demonstrate mastery in understanding by fixing what they broke. In M2, the RNA was broke and a series of hypotheses based on predicted helices was tested. To conclusively demonstrate that a helix is present, the broken helix is rescued by a series of compensatory mutations in a series of experiments called Mutate-Map-Rescue (M2R). This single mutant, double-double rescue mutant strategy did not yield useable data for the CR4/5 domain. Instead, double mutants and quadruple rescues were required to yield sufficiently significant helical perturbation. This chapter summarizes the kinds of analysis done for traditional M2R and the reasoning behind transitioning to a more disruptive system.

Primerize M2R

M2R RNA is generated using Primerize like 1D and M2 experiments. Generate primers to test different sets of base pairs by inputting structures in dot-bracket notation into Primerize. Two different strategies exist for running the experiment. Users can select ‘include single mutants’ in the advanced option panel on the right, and the plate that is generated will contain both single mutants in the base pair and the rescue double mutant next to them (ie - A01: WT; B01:Mut1; C01:Mut2; D01:Rescue_of_mut1_and2). This pattern will continue for each base pair in each helix entered. Alternately, if the user has already run experiments on an M2 plate just the M2R plate can be generated. This is the default setting.

Make M2R RNA

Make RNA as previously described.

Probe M2R RNA

M2R chemical mapping is performed in the same way as M2 mapping.

Analyze M2R RNA

An important aspect of the visual analysis is having all the RNAs that are going to be analyzed in a single data processing step. Therefore, if data is being drawn from different experiments, copy all the data that will be included into a single folder. Run the full data processing pipeline as outlined in the M2 analysis such that a single rdat is generated. The only variation on the data analysis is that to generate data annotation, use the .txt file output by Primerize.

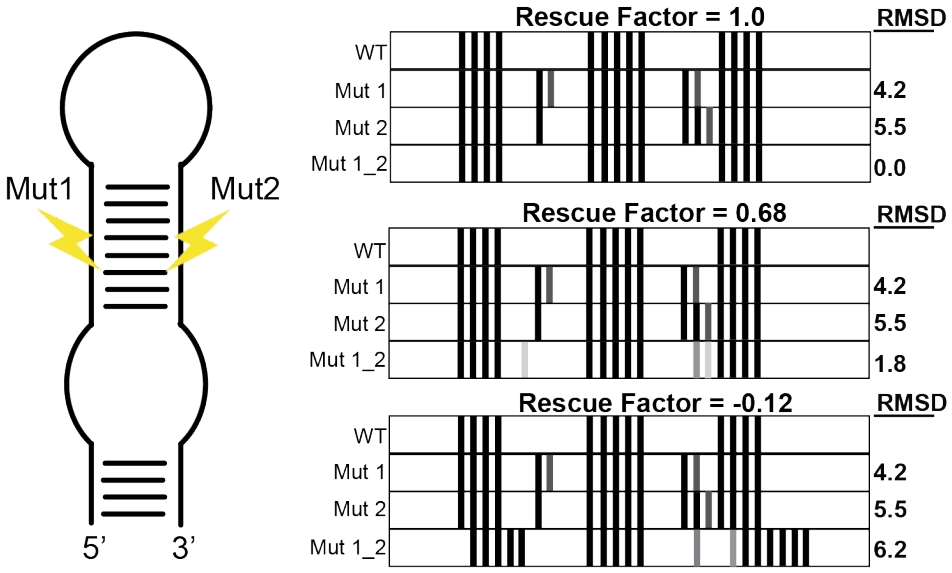
```
construct_names = read_constructs('primer_keys.txt');
for j = 1:size(area_peak, 2);
    data_annotations{j} = {['mutation:',
strrep(strrep(strrep(construct_names{j}, 'T', 'U'), 'WU', 'WT'),
'Lib1-', '')]};
end;
```

Generate the rdat as before:

```
output_workspace_to_rdat_file(filename, name, sequence, offset,
seqpos, area_peak_norm, structure, annotations, data_annotations,
darea_peak, [], [], [], comments);
d_rdat = show_rdat(filename);
```

Unlike M2 analysis, examination of the full rdat is not particularly useful. Instead, the user examines a quartet of RNA constructs: WT, Mutant 1, Mutant 2, and Double Mutant Rescue. If the single mutants disrupt a base pairing interaction, a visual analysis of the reactivity profiles comparing the single mutants and the WT will reveal differences. If the double mutant successful restored the base pairing interaction, the reactivity should look similar to the WT. It is likely possible to draw information from two rdats to generate quartets, but I did not put the time into

generating a script that could do so. Thus, as mentioned before, all mutants and rescues must be in one rdat to generate quartets.



Rescue factor = 1 - (RMSD Mut1_2/(max)RMSD of Mut1 or Mut2)

Figure 3 13 - Toy example of the visualization and calculation of a rescue factor from single mutants and double rescues.

A toy helix with two compensatory mutations is illustrated. Three potential reactivity profiles are shown. An RMSD is calculated comparing the mutants to the WT and a rescue factor is calculated using the equation on the bottom.

Similar to M2 analysis, M2R analysis has a quantitative analysis that accompanies the visual analysis. Differences in reactivity profiles are quantified by calculating an RMSD, comparing each mutant to the WT. Then a rescue factor is calculated (**Equation Fig 3.13**). A high rescue factor was defined as 0.7 - 1.0, a medium rescue factor as 0.3 - 0.69, and a low rescue factor as 0.0 - 0.29.

To calculate an RMSD that is representative of the reactivity profile across multiple experiments, five replicates of the M2 and M2R plates were generated. Data processing was preformed and an RMSD was calculated for each mutant and double

mutant (**Fig 3.14A**). RMSDs were calculated only for the region of interest, excluding the reactivity of the buffers and hairpins. The RMSDs were plot and outliers were excluded. At least three data points for each mutant was present in the final dataset. The average RMSD of each mutant was used to calculate a rescue factor. Only low to mid-rescue factors were obtained for all helices (**Fig 3.14 C/D**).

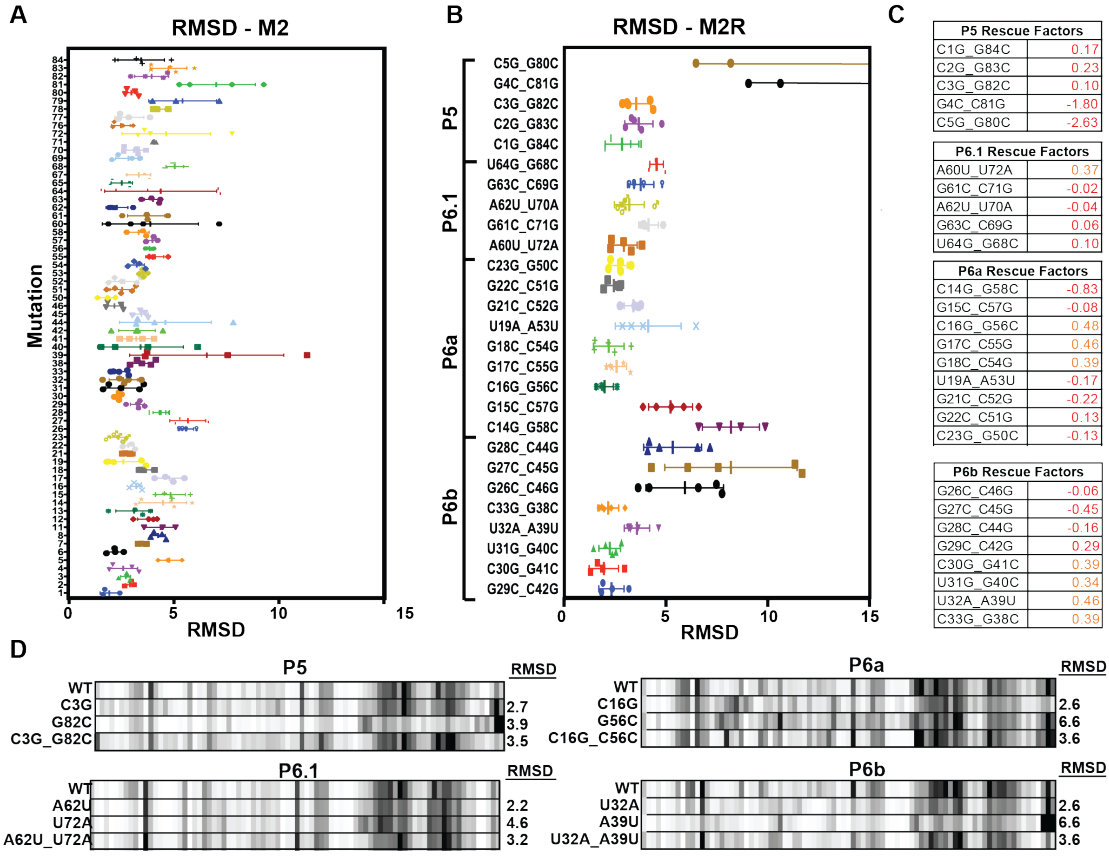


Figure 3 14 - Only low and medium rescue factors were calculated for single mutant, double mutant rescues for the human CR4/5 domain.

(*A and B*) Five replicates of the M2 and M2R chemical mapping was preformed and their RMSD's compared were calculated, outliers, excluded and averages calculated. (*C*) Rescue factors for P5, P6a, P6b, and P6.1 were calculated. (*D*) Quartets of select mutants and rescues.

Discussion and Conclusions:

A lack of rescue factors can be interpreted in a few ways. One is that none of the helices in the CR4/5 domain are stably formed. This interpretation is not supported by the 1D data which shows a lack of reactivity in the vast majority of the P6 stem (**Chapter 2: Fig 2.2 and 2.3**) and perturbation consistent with the formation of P6 is observed in the M2 data. These experiments support the idea that P6 and P5 are stably folded. Another interpretation is that the rescue factors are not accurately reflecting a rescue. Low rescue factors can be obtained when a double mutant displays a high RMSD. Another way to obtain low rescue factors is to have a low level of perturbation calculated for the single mutants. If the single mutants are not causing a large disruption then even if the double mutant rescue aspects of the reactivity profile, a low rescue factor will still be calculated.

Data analysis was performed a number of different ways in an attempt to tease out significant rescue factors from the data sets. Analysis was performed on the data sets after implementing hairpin normalization and attenuation correction. No improvement of rescue factors was observed (data not shown). To determine whether alignment was causing variability between experiments, data were processed in smaller sets, only including mutations in one helix. No improvement of rescue factors was observed (data not shown). If low level of perturbation was the problem with the data analysis pipeline, perhaps looking just at the region that was mutated would yield a higher RMSD and rescue factor. Therefore a script was written that calculated the RMSD in a rolling window, dependent on the point of mutation. The script was written to allow the user to choose the window size for anlaysis. For example, at nucleotide 20, if a window of 5 was chosen, nucleotides between 15 - 25

would be used for the RMSD calculation. Rescue factors obtained were still low (data not shown).

In an effort to obtain an increased level of perturbation, double mutants and quadruple rescue mutants were generated for P6a, P6b, and P5. An exhaustive scan of the base pairs was not done. Instead two mutations and their corresponding rescue mutations were introduced into the P5, P6a, P6b, and P6.1 stem (**Fig 3.15**). RNAs were probed using the 1D chemical mapping experimental protocol. In the double mutant and double mutant rescue RNAs, a significant change in reactivity profile and significant rescue factors were obtained. These data suggest that P6a and P6b and P5 are formed in solution and that the issue with the M2R experiment was a lack of overall perturbation in the single mutant constructs.

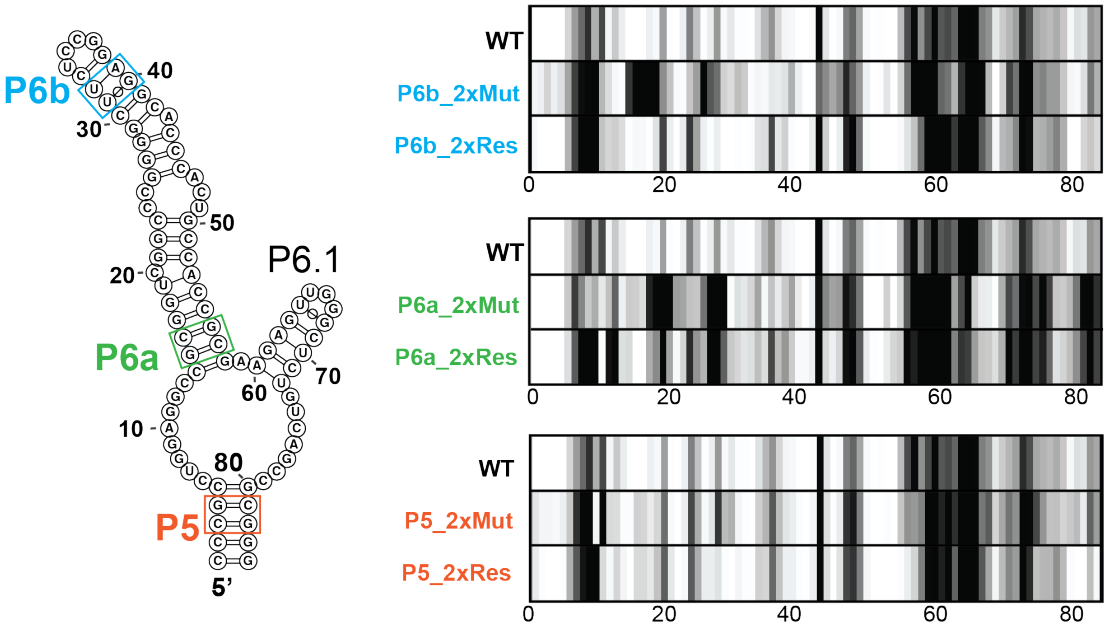


Figure 3 15 - Double mutants and quadruple compensatory rescue supports the formation of P6a, P6b, and P5.

Double mutants and the compensatory rescues are boxed and color coordinated. The corresponding reactivity profiles are also visualized

The Das lab first published the pipeline for M2 experiments in 2010 [14] [17]. Analysis of whether a double mutant was rescued or not was performed by visual analysis by 'experts'. In my hands, examination of the CR4/5 M2 and M2R datasets was not conclusive. It was difficult to determine whether changes and then rescues were significant. If aspects of the rescue mutant were still different, was the mutant considered rescued? How much of a change should be observed to consider a mutant perturbed? Unable to reconcile these considerations in the visual analysis, a quantitative approach was deemed necessary.

To achieve this quantification an approach from a more recent Das lab publication was utilized that introduced rescue factors [13]. However, careful examination of the text revealed that double mutants and double rescues were used in this set of experiments. It is unclear if single mutant and double rescue will achieve a high enough level of perturbation to obtain significant rescue factors. For future users trying to verify helices using rescue factors, skipping the single-mutant-rescue and moving directly to double-mutant-rescue constructs seems to be the recommended route. A helpful addition to the Primerize website would be an addition of the double mutant-double rescue pipeline as well as a potential clarification that significant rescue factors and RMSDs are unlikely to be calculated from normal M2 and M2R experiments.

Another aspect that may be useful for users to consider is the A, C, G, and T content of the RNA. The CR4/5 domain has a very high G and C content. Therefore it seems likely that when a G was mutated to a C, rather than the release of the nucleotide, helical sliding occurred, changing the identity of the base pairing interactions, but not releasing the mutated base. This is especially noticeable in the

P5 double mutant, double rescue (**Fig 3.15**). Perhaps instead of mutating to the opposing base pair the introduction of A's and U's for G's and C' may have generated more noticeable release events and rescues.

References

1. Brookes, P. and P.D. Lawley, *The reaction of mono- and di-functional alkylating agents with nucleic acids.* Biochem J, 1961. **80**(3): p. 496-503.
2. Gilham, P.T., *An Addition Reaction Specific for Uridine and Guanosine Nucleotides and its Application to the Modification of Ribonuclease Action.* J. Am. Chem. Soc., 1962. **84**(4): p. 687-688.
3. Wilkinson, K.A., E.J. Merino, and K.M. Weeks, *Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution.* Nat Protoc, 2006. **1**(3): p. 1610-6.
4. Mortimer, S.A. and K.M. Weeks, *Time-resolved RNA SHAPE chemistry.* J Am Chem Soc, 2008. **130**(48): p. 16178-80.
5. Noller, H.F., et al., *Secondary structure model for 23S ribosomal RNA.* Nucleic Acids Res, 1981. **9**(22): p. 6167-89.
6. Inoue, T. and T.R. Cech, *Secondary structure of the circular form of the Tetrahymena rRNA intervening sequence: a technique for RNA structure analysis using chemical probes and reverse transcriptase.* Proc Natl Acad Sci U S A, 1985. **82**(3): p. 648-52.
7. Romby, P., et al., *Comparison of the tertiary structure of yeast tRNA(Asp) and tRNA(Phe) in solution. Chemical modification study of the bases.* J Mol Biol, 1987. **195**(1): p. 193-204.
8. Mitra, S., et al., *High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis.* Nucleic Acids Res, 2008. **36**(11): p. e63.

9. Mortimer, S.A., et al., *SHAPE-Seq: High-Throughput RNA Structure Analysis.* Curr Protoc Chem Biol, 2012. **4**(4): p. 275-97.

10. Kladwang, W., et al., *Standardization of RNA chemical mapping experiments.* Biochemistry, 2014. **53**(19): p. 3063-5.

11. Kim, H., et al., *HiTRACE-Web: an online tool for robust analysis of high-throughput capillary electrophoresis.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W492-8.

12. Cordero, P. and R. Das, *Rich RNA Structure Landscapes Revealed by Mutate-and-Map Analysis.* PLoS Comput Biol, 2015. **11**(11): p. e1004473.

13. Tian, S., W. Kladwang, and R. Das, *Allosteric mechanism of the V. vulnificus adenine riboswitch resolved by four-dimensional chemical mapping.* Elife, 2018. **7**.

14. Kladwang, W. and R. Das, *A mutate-and-map strategy for inferring base pairs in structured nucleic acids: proof of concept on a DNA/RNA helix.* Biochemistry, 2010. **49**(35): p. 7414-6.

15. Tian, S., et al., *Primerize: automated primer assembly for transcribing non-coding RNA domains.* Nucleic Acids Res, 2015. **43**(W1): p. W522-6.

16. Cordero, P., et al., *The mutate-and-map protocol for inferring base pairs in structured RNA.* Methods Mol Biol, 2014. **1086**: p. 53-77.

17. Tian, S., et al., *High-throughput mutate-map-rescue evaluates SHAPE-directed RNA structure and uncovers excited states.* RNA, 2014. **20**(11): p. 1815-26.

Chapter 4: Re-evaluation of the RNA binding properties of the *Tetrahymena thermophila* telomerase reverse transcriptase N-terminal domain

Authors: Christina Palka[1], Aishwarya P. Deshpande[2], Michael D. Stone[1*], Kathleen Collins[2*]

[1] Chemistry and Biochemistry, University of California at Santa Cruz, Santa Cruz, CA 95064

[2] Molecular and Cell Biology, University of California at Berkeley, Berkeley, CA 94720

Introduction

Telomerase and other reverse transcriptases (RTs) discriminate their appropriate templates using sequence- and structure-specific RNA recognition. Retroelement RTs use their entire bound RNA as template [1]. In contrast, telomerase copies only a short region within the TR subunit of an active ribonucleoprotein (RNP) complex [2]. Telomerase biogenesis stably co-folds TERT and TR in a hierarchical series of induced conformational changes that ultimately determine the region of TR accessible to the active site [3]. Template 5'-boundary fidelity is essential for precise repeat synthesis and is strictly enforced in most but not all telomerase holoenzymes [4]. This boundary is set by hindrance from 5' template-flanking RNA secondary structure or RNA-protein interaction [5, 6], and in vertebrate enzymes it depends in large part also on sequence-specific recognition of the template-product duplex [7, 8].

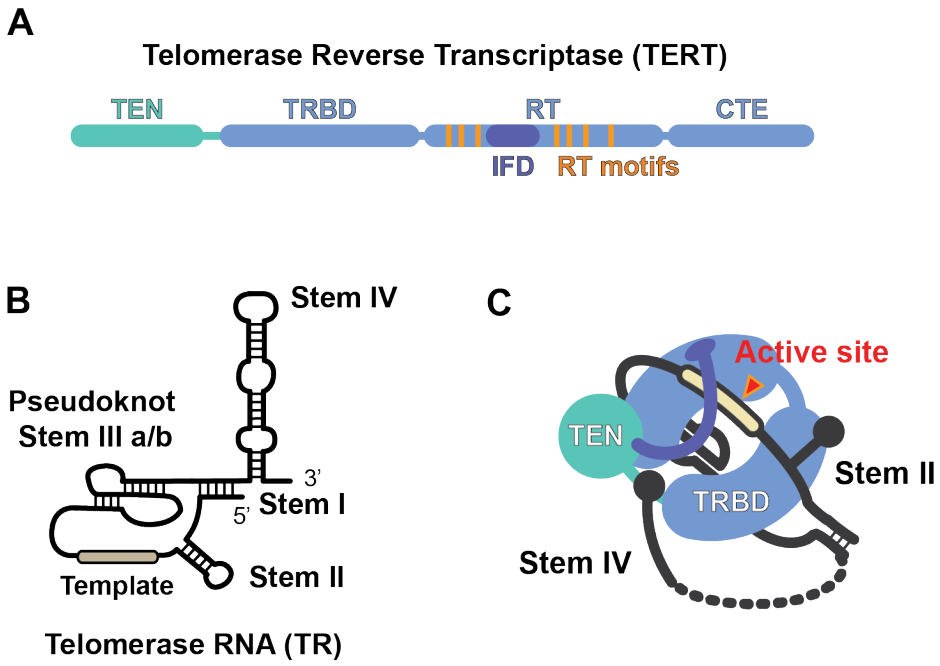


Figure 4 1 - Conserved telomerase subunits from the ciliate Tetrahymena thermophila.

(A) A schematic illustration of TERT domain organization, including: the Telomerase Essential N-terminal (TEN) domain, telomerase RNA binding domain (TRBD), the reverse transcriptase (RT) domain, and the telomerase C-terminal extension (CTE). The conserved insertion of fingers (IFD, purple) motif within the RT domain and canonical RT motifs (orange) are depicted. (B) Secondary structure of *Tetrahymena* TR with conserved stem elements and RNA pseudoknot. The region of TR that serves as the template during telomere repeat synthesis is demarcated in beige. (C) Cartoon model of the three-dimensional architecture of *Tetrahymena* telomerase RNP enzyme based on cryoEM structure [12, 15, 30]. The model highlights the complex topological arrangement of the protein and RNA subunits within the assembled RNP complex that is in part established by a protein-protein contact between the TEN domain and the IFD motif. Colors in model are as described in panel (A).

Critical RNA interactions of TERT and retroelement RTs are mediated by a protein domain immediately preceding the ubiquitously conserved RT motifs (**Fig 4.1A**). Within TERT, this high-affinity telomerase RNA binding domain (TRBD) binds and positions both a template 5'-flanking region (Stem II of ciliate TR, **Fig 4.1B**) and a distant, independently folded stem-loop motif (Stem IV of ciliate TR, **Fig 4.1B**) [6, 9-14]. TRBD interfaces with TR establish the tertiary structure of activity-essential TR motifs in the catalytic core of ciliate and human telomerase holoenzymes [14, 15].

Together the TERT TRBD, the RT domain with active-site motifs (RT), and the following TERT C-terminal extension (CTE) form the “TERT ring” encircling the active site cavity (Figure 1A)[16, 17]. Placement of the template in the vicinity of the active site cavity requires the template 3’-flanking region to traverse to the opposite side of the TERT ring from the TRBD-bound template 5’-flanking region, and for the TR path 3’ of the template to ultimately encircle the entire circumference of TERT ring (**Fig 4.1C**) [12, 14, 15]. The structural determinants of most of the TR path, including that of the template 3’ flanking region, are not yet established.

In telomerase holoenzymes of the model ciliate *Tetrahymena thermophila* (henceforth *Tetrahymena)* and human cells, the TERT TEN domain is perched atop the TERT ring off to the RT-CTE side, instead of above the physically connected TRBD (**Fig 4.1A and 4.1C**). The TR template 3’-flanking region threads past one side of the TEN domain as TR wraps around the CTE to the opposite face of TERT (**Figure 4.1C**). *Tetrahymena* and human TR take a generally similar path despite complete divergence of the template 3’-flanking region sequence and structure, which in *Tetrahymena* TR is entirely single-stranded but in human TR has only a short single-stranded stretch followed by a paired region. Unfortunately, because the structural snapshots captured to date are not at atomic resolution, the position(s) of single-stranded RNA interaction with TERT are not yet possible to decisively infer. The vertebrate-specific paired stem flanking the template 3’ end does approach closely to the human TERT TEN domain [14], but the holoenzyme cryo-EM density in this region is not well fit by the *Tetrahymena* TEN domain structure (the only TEN domain structure at atomic resolution)[18], leaving ambiguous the potential for double-stranded RNA contact.

Ciliate, yeast, and human TEN domains can be expressed autonomously from the rest of TERT, and when purified they have been reported to interact with single-stranded DNA and/or TR as assayed by native gel electrophoresis, filter binding, cross-linking, NMR, and single-molecule assays [18-26]. Functionality of the recombinant, isolated TEN domain is supported by its co-assembly with TEN-less TERT and TR to reconstitute full enzyme activity [27, 28]. However, this complementation requires co-expression or co-assembly in a cell extract [28, 29], suggesting that conformational changes of the autonomously folded TEN domain may be necessary for productive protein-protein or protein-RNA domain interactions. Indeed, structures of the TERT TEN domains from *Tetrahymena* and the thermophilic yeast *Hansenula polymorpha* reveal disordered regions likely to be constrained within a fully assembled holoenzyme [18, 26].

The isolated *Tetrahymena* TEN domain has been reported by two groups to bind *Tetrahymena* TR, but with at least 100-fold lower affinity than the nanomolar binding of the *Tetrahymena* TERT TRBD [18, 21]. Initial studies from the Collins lab suggested that TEN domain interaction with TR had some dependence on TR sequence in two regions, but no single TR region was sufficient for interaction [21]. With the benefit of recent insights about telomerase RNP domain architecture [14, 30], we sought to better characterize *Tetrahymena* TEN domain interaction with TR. We conclude that the isolated TEN domain has lower affinity for TR than previously reported and that it lacks obvious sequence specificity of interaction. Previous studies' findings about nucleic acid interaction specificity have technical concerns discussed in detail below. Overall, TEN domain structure/function relationships remain an elusive goal for future understanding.

Results

A bacterial contaminant with RNA binding activity co-purifies with 6xHis-tagged TEN domain

Previous studies bacterially expressed and purified an N-terminally six-histidine (6xHis) tagged *Tetrahymena* TERT TEN domain to investigate its TR and DNA binding activities, atomic resolution structure, and requirements for functional complementation with *Tetrahymena* TERT ring and other telomerase holoenzyme subunits [18, 21, 22, 28]. Here we purified this polypeptide using a similar process of affinity chromatography with nickel resin. TEN domain was assessed for purification from bacterial proteins by SDS-PAGE and Coomassie staining (**Fig 4.2A, lane 1**) and for RNA binding by electrophoretic gel mobility shift assays (EMSAs). Consistent with previously reported TR binding affinities [18, 21], the nickel-affinity (Ni-affinity) purified TEN domain appeared to robustly bind TR in assays with limiting radiolabeled TR (~1 nM) and a large excess of 6xHis-TEN protein (micromolar range, **Fig 4.2B**). In a separate experiment, we assessed the ability of unlabeled (cold) TR to compete for binding of radiolabeled TR. Unexpectedly, we observed a strong reduction in the amount of TEN-TR complex with sub-stoichiometric levels of cold competitor TR (**Fig 4.2C**), for example with only 75 nM cold TR added to 750 nM TEN domain.

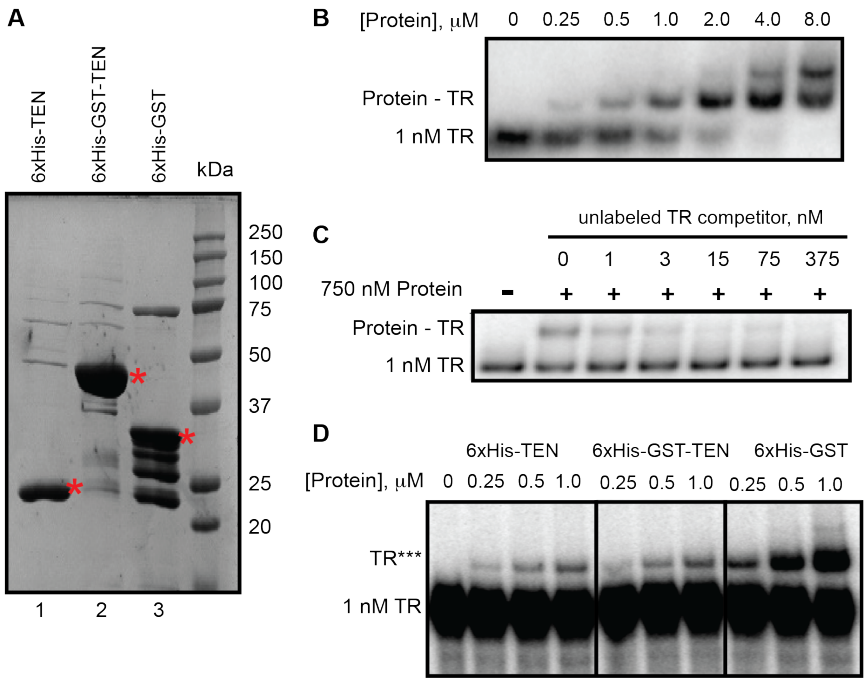


Figure 4 2 - Purification and EMSAs testing GST fusion to the TEN domain.

(A) SDS-PAGE analysis of the indicated proteins. (B) EMSA analysis of 1 nM radiolabeled TR incubated with the indicated amounts of 6xHis-TEN protein. (C) Competition EMSA experiment with 750 nM protein and 1 nM radiolabeled TR (~ 50% of TR bound). Unlabeled TR competitor was added to the binding reaction at the indicated concentrations. (D) EMSA of Ni purified 6xHis-TEN, 6xHis-GST-TEN, and 6xHis-GST (no TEN) with 1 nM radiolabeled TR. TR*** indicates a protein-dependent mobility shift of the radiolabeled TR.

This result is unexpected because a higher concentration of cold TR should be necessary to saturate RNA binding to the TEN domain and thus exclude radiolabeled TR. One possible explanation for this observation was that the fraction of 6xHis-TEN competent to bind TR is exceedingly low, vperhaps due to alternative protein conformation and/or aggregation. To address the state of protein aggregation, we further purified 6xHis-TEN by gel filtration chromatography. Contrary to our expectations, the well-defined peak of 6xHis-TEN at the monomer retention time of a Superdex 200 column exhibited less TR binding than the input when normalized to

TEN domain concentration (data not shown). This raised the possibility that a contaminant protein other than 6xHis-TEN might be responsible for the observed mobility shift.

To explore this possibility, we more than doubled the molecular weight of the TEN-domain polypeptide by inserting a glutathione S-transferase (GST) tag immediately following the 6xHis tag at the TEN domain N-terminus. We purified 6xHis-GST-TEN, as well as a 6xHis-GST negative control, using the same Ni-affinity method employed for 6xHis-TEN. SDS-PAGE analysis showed the expected sizes for the GST fusion proteins (**Fig 4.2A**, **lanes 2-3**). Protein samples were normalized to the same molar fusion protein concentration and titrated into binding reactions with limiting TR (~1 nM). The purified protein samples all gave the same mobility shift (indicated as TR***), which was maximal in amount with the negative control 6xHis-GST sample (**Fig 4.2D**). This result implicates a bacterial contaminant rather than the *Tetrahymena* TEN domain as the source of TR mobility shift here, and by extension likely in previous RNA binding assays as well, even though no candidate for such a contaminant protein is evident by SDS-PAGE in proportion to the mobility shift activity (compare **Fig 4.2A and 2D**).

Elevated TR concentration enables detection of RNA binding by the TEN domain

In additional experiments, we used different TEN domain fusion proteins and different assay conditions to investigate TR interaction with the *Tetrahymena* TEN domain. To this end, in parallel, we expressed and Ni-affinity purified 6xHis-TEN and TEN domain N-terminally tagged with 6xHis also bearing a maltose binding protein (MBP) tag at its C-terminus (6xHis-TEN-MBP) to nearly triple the mass of the fusion

protein relative to 6xHis-TEN alone. Also in parallel we purified two negative control samples: TEN domain with C-terminal MBP tag but no 6xHis tag (TEN-MBP) and N-terminally 6xHis-tagged MBP (6xHis-MBP). All of the fusion polypeptides were soluble and all but the TEN-MBP fusion protein were enriched by Ni-affinity chromatography, as determined by SDS-PAGE and colloidal Coommassie staining (**Fig 4.3A**).

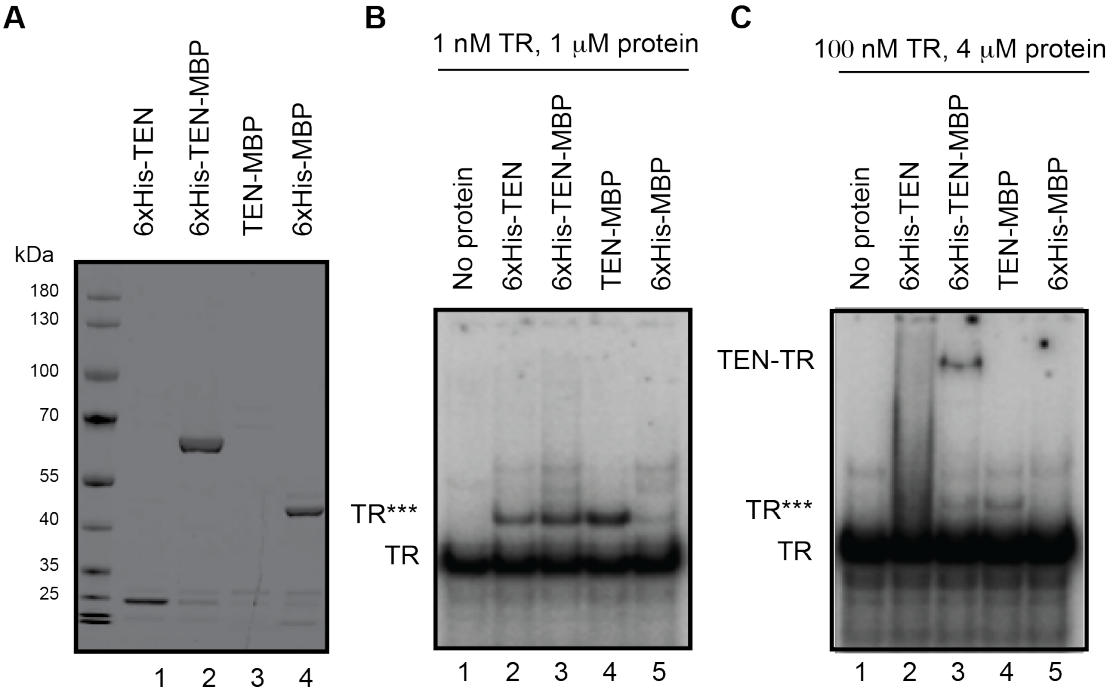


Figure 4 3 - Purification and EMSAs testing MBP fusion to the TEN domain.

(A) SDS-PAGE analysis of 6xHis-TEN (lane 1), 6xHis-TEN-MBP (lane 2), TEN-MBP (lane 3), and 6xHis-MBP (lane 4). (B) EMSA analysis of 1 nM radiolabeled TR and 1 μM of the indicated protein sample or control equivalent. TR*** indicates a protein-dependent mobility shift of the radiolabeled TR. (C) EMSA analysis of low specific activity 100 nM TR (with 99 nM cold TR plus 1 nM radiolabeled TR) and 4 μM of the indicated protein sample or control equivalent.

To use the negative control TEN-MBP purification, the sample eluted from nickel resin was normalized to the 6xHis-MBP sample by equivalent eluted volume rather than protein amount. First, as for the experiments described above, we used

limiting radiolabeled TR (1 nM) and excess protein (1 μM) (**Fig 4.3B**). As described above for 6xHis-GST-TEN, the 6xHis-TEN-MBP protein did not change the position of mobility shift observed with 6xHis-TEN sample (**Fig 4.3B**, **lanes 1-3**). Although TEN-MBP lacking the 6xHis tag was not enriched by Ni-affinity purification, this negative control sample gave the maximal mobility shift intensity (**Fig 4.3B**, lane 4). On the other hand, the mobility shift was only marginally detectable for the negative control 6xHis-MBP sample (**Fig 4.3B**, **lane 5**), consistent with the EMSA signal arising from an RNA-binding contaminant enriched by Ni-affinity resin in competition with Ni-affinity resin binding to a 6xHis-tagged protein. These results parallel results described above in suggesting that the predominant TR mobility shift does not correspond to TR binding by the TEN domain. Also, again, no candidate for the contaminant bacterial protein that mediates the TR shift is evident among the proteins detected by SDS-PAGE and gel staining.

If the contaminant protein is of very low abundance, its TR mobility shift would disappear if limiting radiolabeled TR was mixed with unlabeled TR to generate lower specific activity RNA (as was indeed observed, **Fig 4.2C**). We therefore assayed the panel of purified proteins and controls for RNA binding using radiolabeled TR diluted with cold TR to a final TR concentration of 100 nM, mixed with 4 uM protein. In this binding condition, a different position of mobility shift was produced by 6xHis-TEN than by 6xHis-TEN-MBP, and neither of these shifts occurred with the negative control samples of 6xHis-MBP or TEN-MBP lacking a 6xHis tag (**Fig 4.3C**). The TR shift by TEN domain was more discrete for the protein with both 6xHis and MBP tags compared to the protein with 6xHis tag alone (**Fig 4.3C**, compare lanes 2 and 3), suggesting that MBP fusion improved TEN domain folding or its retention of RNA

during gel electrophoresis. Taken together, these results demonstrate that lower specific activity TR at high concentration can be used to detect RNA binding by the TEN domain.

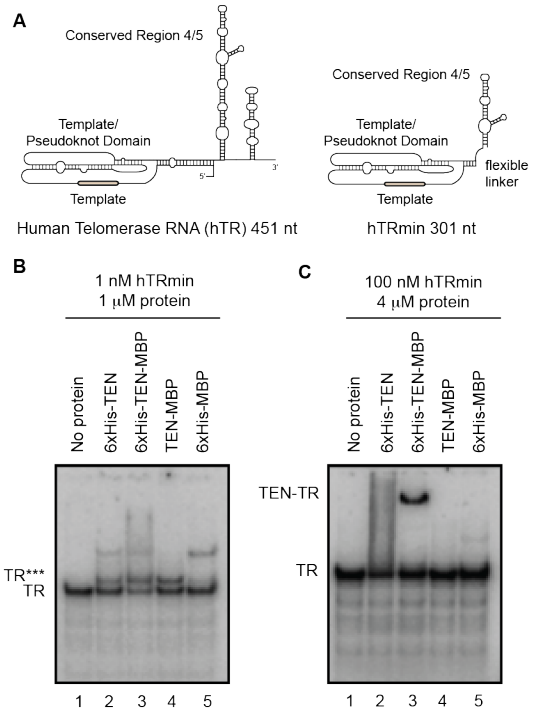


Figure 4 4 - Comparison of RNAs for binding to the TEN domain.

(A) Secondary structure schematics for full-length human TR (hTR) and a minimal activity-reconstituting human TR (hTRmin). *(B)* EMSA analysis of 1 nM radiolabeled hTRmin and 1 μM of the indicated protein sample or control equivalent. TR*** indicates a protein-dependent mobility shift of the radiolabeled hTRmin. *(C)* EMSA analysis of low specific activity 100 nM hTRmin and 4 uM of the indicated protein sample or control equivalent

Having identified appropriate assay conditions to detect TR interaction with the TEN domain, we sought to characterize the interaction specificity. For this purpose we compared TEN domain binding to *Tetrahymena* TR with its binding to a streamlined version of human TR containing the activity-essential TR regions joined by a short linker (hTRmin, **Fig 4.4A**), which has a length about twice that of *Tetrahymena* TR (**Fig 4.1B**) [29]. Like *Tetrahymena* TR, hTRmin reconstitutes telomerase catalytic activity with its respective TERT translated in cell extract [29]. *Tetrahymena* and human TR primary sequences are largely divergent, and even shared secondary structure elements such as the pseudoknot vary in size (**Figu 4.1B** and **4.4A**)[8]. In EMSAs, hTRmin was variably shifted in mobility by the panel of tagged *Tetrahymena*

TEN domain proteins and negative controls. In assays with limiting 1 nM hTRmin and 1 uM TEN domain or control purification, the prominent mobility shift was the same in samples with 6xHis-TEN and 6xHis-TEN-MBP (**Fig 4.4B**, lanes 1-3). The negative control sample TEN-MBP gave maximal intensity of this mobility shift, and 6xHis-MBP gave the minimum, corresponding with results observed for mobility shift of *Tetrahymena* TR under conditions of limiting RNA (compare **Figu 4.3B** and **4.4B**). In comparison, EMSAs using 100 nM lower specific activity hTRmin and 4 uM protein demonstrated mobility shift dependence on the TEN domain: a different position of mobility shift was produced by 6xHis-TEN than by 6xHis-TEN-MBP, and neither of these shifts occurred with the control samples (**Figure 4.4C**). These results parallel those obtained under the same binding conditions for *Tetrahymena* TR (**Figure 4.3C**), indicating that the *Tetrahymena* TEN domain RNA-binding activity is not specific for *Tetrahymena* TR. Finally, we did not find any individual motif within *Tetrahymena* or human TR that bound to the isolated TEN domain, whether using a 1 nM or 100 nM RNA mobility shift condition (data not shown).

Discussion

We show here that a trace amount of a bacterial protein with high affinity for TR co-purifies with 6xHis-tagged *Tetrahymena* TEN domain. Comparing across the purification samples described above and other sets, we did not find a protein evident by staining after SDS-PAGE that correlated in abundance with the amount of 1 nM TR mobility shift. A very low level of the contaminant is consistent with detection of a mobility shift only with high specific activity TR, for which we estimate ~40 pg of a 20 kDa protein contaminant would be sufficient. Mobility shift assays with lower specific

activity TR were useful in detecting an RNA interaction authentic to the TEN domain. Although the trace-level contaminant still binds TR, its RNP becomes less abundant than the TEN domain RNP of interest.

Likely contamination of *Tetrahymena* TEN domain in previously published assays [21] prompts a revision of prior conclusions about TEN domain affinity and specificity of RNA binding. Rather than ~0.5 uM [21] or ~0.1 uM [18], we suggest that TEN domain affinity for TR is substantially greater than 1 uM. Largely similar to the RNA binding specificity of the contaminant, TEN domain interaction with TR does not require a unique region of *Tetrahymena* TR but is sensitive to RNA length (data not shown), perhaps due to RNA length-dependent formation of secondary structure. It is important to note that both the *Tetrahymena* and yeast TEN domain with determined structure do not have a homogeneous well-folded conformation; instead, large portions of the small domain are disordered [18, 26]. This likely reflects the absence of interactions made in holoenzyme context. Recent studies in several organisms have led to increasing appreciation of the TEN domain as a nexus of interactions between the catalytically active TERT ring RNP and critical telomerase holoenzyme and telomere proteins required for telomerase function and coordination at chromosome ends [31]. Because the active RNP TEN domain conformation may depend on the interacting proteins, properties of the autonomous TEN domain have uncertain significance for its roles in holoenzyme context.

Cryo-EM studies of *Tetrahymena* and human telomerase holoenzyme subunit architecture place the TEN domain above the TERT CTE, close to where the template 3' flanking region begins its circumnavigation to the back side of TERT [14, 30]. The TR template 3'-flanking region changes dramatically in structure during TR folding with

TERT [32-35] and likely during the catalytic cycle as well [31, 36]. For *Tetrahymena* TR, the mature RNA fold is not adopted without TERT: the template 3' end, template 3'-flanking region, and some of the future pseudoknot form a long, snap-back hairpin [32-34]. Whether or not the TEN domain plays a role in refolding TR remains to be addressed.

Methods

Expression and purification of RNAs

Tetrahymena TR and human hTRmin were transcribed from linearized plasmids largely as previously described [29, 37], using T7 RNA polymerase, and then RNAs were purified by denaturing PAGE. RNA purity was verified by denaturing PAGE with SYBR Gold or ethidium bromide staining. RNA concentrations were determined by Nanodrop spectrometer (ThermoFisher).

Expression and purification of proteins

Tetrahymena TERT TEN domain (amino acids 1-195) was expressed in fusion with a polypeptide tag or tag combination as indicated in the text, in parallel with expression of the large tags alone. All polypeptides were expressed using pET28 vectors in *E. coli* BL21(DE3) cells. Transformed cells were grown at 37°C until an optical density of approximately 0.6 was reached, at which point cultures were shifted to lower temperature for induction of protein expression by addition of approximately 1.0 mM isopropyl 1-thio-β-D-galactopyranoside. Aliquots of purified protein were stored at -80°C after flash freezing in liquid nitrogen.

For the experiments in Figure 2, protein was expressed by overnight induction at 18°C. Harvested cells were resuspended in TENA (20 mM Tris-HCl, 250

mM NaCl, 10 mM imidazole, 10% glycerol, 2 mM 1,4-dithiothreitol (DTT); pH 8.0). Cells were lysed via cell disruptor, after which slurry was clarified by centrifugation. Supernatant was mixed with Ni Sepharose Excel resin (GE Healthcare) pre-equilibrated with the lysis buffer and allowed to rotate end-over-end at 4°C for 3 hours. Resin was collected and washed at 4°C with approximately 10 column volumes of lysis buffer until no protein came off the column as determined by Bradford assay. Bound protein was eluted in 1 mL fractions with TENA adjusted to 250 mM imidazole. Protein was then dialyzed back into TENA lacking imidazole and concentration was determined by Nanodrop.

For the experiments in Figures 3 and 4, protein was expressed by 4 hour induction at room temperature. Harvested cells were washed with 1x PBS containing 200 μM phenylmethylsulfonyl-fluoride (PMSF) before freezing at -80°C. Thawed cells were resuspended in TENB (20 mM Tris-HCl, 2 mM $MgCl_2$, 50 mM NaCl, 20 mM imidazole, 10% glycerol, 0.05% NP-40, 1 mM DTT; pH 8.0) with 200 μM PMSF, 1:200 of protease inhibitor cocktail (Sigma) and 1 mg/ml lysozyme. The slurry was gently stirred at 4°C for 45 min, followed by sonication for 3 min with 10-sec on/off pulses. Lysate was clarified by centrifugation, mixed with nickel-nitrilotriacetic acid-agarose resin (NiNTA, Qiagen) and allowed to rotate end-over-end at 4°C for 4 hours. Resin was collected and washed at 4°C with 3 changes of a large bead-volume excess of TENB for 15 min each. Bound protein was eluted in TENB adjusted to 300 mM imidazole. Protein concentration was determined by Bradford assay.

EMSAs

For radiolabeling, purified RNAs were treated with shrimp alkaline phosphatase at 37°C for 1 h then end-labeled using T4 polynucleotide kinase and γ-^{32}P-ATP at 37°C for 1 h. Complexes were resolved by electrophoresis at 4°C on native 5% acrylamide gels (37.5:1 acrylamide:bis acrylamide, 0.5x Tris borate-EDTA, and 4% glycerol added for gels in Figures 3 and 4). Gels were dried and exposed to phosphorimager screens. Products were visualized by scanning on a Typhoon (GE healthcare).

For the experiments in Figure 2, the binding reaction used binding buffer A unless otherwise specified (20 mM Tris-Base, 1 mM $MgCl_2$, 50 mM NaCl, 10% glycerol, 1 mM DTT; pH 8.0) with 0.1 mg/mL yeast tRNA (Sigma) and 0.1 mg/mL BSA (NEB). Reactions were incubated on ice for 10 min. For the experiments in Figures 3 and 4, radiolabeled RNA was spiked into unlabeled RNA, heated to 70°C for 3 min, and slow-cooled to room temperature. RNA and protein were diluted in binding buffer B (20 mM Tris-HCl, 1 mM $MgCl_2$, 100 mM NaCl, 10% glycerol, 5 mM DTT, 0.25 µl RNasin (Promega), trace bromophenol blue; pH 8.0) and incubated at room temperature for 20 min. Similar results were obtained with or without the presence of 0.1 mg/mL yeast tRNA (Sigma) and 0.1 mg/mL BSA (acetylated BSA from New England Biolabs).

References

1. Han JS. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. *Mob DNA* **1**: 15 (2010).
2. Greider CW and Blackburn EH. A telomeric sequence in the RNA of *Tetrahymena* telomerase required for telomere repeat synthesis. *Nature* **337**: 331-7 (1989).
3. Egan ED and Collins K. Biogenesis of telomerase ribonucleoproteins. *RNA* **18**: 1747-59 (2012).
4. Blackburn EH and Collins K. Telomerase: an RNP enzyme synthesizes DNA. *Cold Spring Harb Perspect Biol* **3** (2011).
5. Lai CK, Miller MC, and Collins K. Template boundary definition in *Tetrahymena* telomerase. *Genes Dev* **16**: 415-20 (2002).
6. Jansson LI, Akiyama BM, Ooms A, Lu C, Rubin SM, and Stone MD. Structural basis of template-boundary definition in *Tetrahymena* telomerase. *Nat Struct Mol Biol* **22**: 883-8 (2015).
7. Wu RA and Collins K. Sequence specificity of human telomerase. *Proc Natl Acad Sci USA* **111**: 11234-5 (2014).
8. Podlevsky JD and Chen JJ. Evolutionary perspectives of telomerase RNA structure and function. *RNA Biol*: 1-13 (2016).
9. Lai CK, Mitchell JR, and Collins K. RNA binding domain of telomerase reverse transcriptase. *Mol Cell Biol* **21**: 990-1000 (2001).
10. O'Connor CM and Collins K. A novel RNA binding domain in *Tetrahymena* telomerase p65 initiates hierarchical assembly of telomerase holoenzyme. *Mol Cell Biol* **26**: 2029-36 (2006).

11. Bley CJ, Qi X, Rand DP, Borges CR, Nelson RW, and Chen JJ. RNA-protein binding interface in the telomerase ribonucleoprotein. *Proc Natl Acad Sci USA* **108**: 20333-8 (2011).

12. Jiang J*, et al.* The architecture of *Tetrahymena* telomerase holoenzyme. *Nature* **496**: 187-92 (2013).

13. Huang J*, et al.* Structural basis for protein-RNA recognition in telomerase. *Nat Struct Mol Biol* **21**: 507-12 (2014).

14. Nguyen THD, Tam J, Wu RA, Greber BJ, Toso D, Nogales E, and Collins K. Cryo-EM structure of substrate-bound human telomerase holoenzyme. *Nature* **557**: 190-5 (2018).

15. Jiang J*, et al.* Structure of *Tetrahymena t*elomerase reveals previously unknown subunits, functions, and interactions. *Science* **350**: aab4070 (2015).

16. Gillis AJ, Schuller AP, and Skordalakes E. Structure of the *Tribolium castaneum* telomerase catalytic subunit TERT. *Nature* **455**: 633-7 (2008).

17. Mitchell M, Gillis A, Futahashi M, Fujiwara H, and Skordalakes E. Structural basis for telomerase catalytic subunit TERT binding to RNA template and telomeric DNA. *Nat Struct Mol Biol* **17**: 513-8 (2010).

18. Jacobs SA, Podell ER, and Cech TR. Crystal structure of the essential N-terminal domain of telomerase reverse transcriptase. *Nat Struct Mol Biol* **13**: 218-25 (2006).

19. Xia J, Peng Y, Mian IS, and Lue NF. Identification of functionally important domains in the N-terminal region of telomerase reverse transcriptase. *Mol Cell Biol* **20**: 5196-207 (2000).

20. Moriarty TJ, Marie-Egyptienne DT, and Autexier C. Functional organization of repeat addition processivity and DNA synthesis determinants in the human telomerase multimer. *Mol Cell Biol* **24**: 3720-33 (2004).

21. O'Connor CM, Lai CK, and Collins K. Two purified domains of telomerase reverse transcriptase reconstitute sequence-specific interactions with RNA. *J Biol Chem* **280**: 17533-9 (2005).

22. Jacobs SA, Podell ER, Wuttke DS, and Cech TR. Soluble domains of telomerase reverse transcriptase identified by high-throughput screening. *Protein Sci* **14**: 2051-8 (2005).

23. Finger SN and Bryan TM. Multiple DNA-binding sites in *Tetrahymena* telomerase. *Nucleic Acids Res* **36**: 1260-72 (2008).

24. Sealey DC, Zheng L, Taboski MA, Cruickshank J, Ikura M, and Harrington LA. The N-terminus of hTERT contains a DNA-binding domain and is required for telomerase activity and cellular immortalization. *Nucleic Acids Res* **38**: 2019-35 (2010).

25. Shastry S, Steinberg-Neifach O, Lue N, and Stone MD. Direct observation of nucleic acid binding dynamics by the telomerase essential N-terminal domain. *Nucleic Acids Res* **46**: 3088-102 (2018).

26. Petrova OA*, et al.* Structure and function of the N-terminal domain of the yeast telomerase reverse transcriptase. *Nucleic Acids Res* **46**: 1525-40 (2018).

27. Robart AR and Collins K. Human telomerase domain interactions capture DNA for TEN domain-dependent processive elongation. *Mol Cell* **42**: 308-18 (2011).

28. Eckert B and Collins K. Roles of telomerase reverse transcriptase N-terminal domain in assembly and activity of *Tetrahymena* telomerase holoenzyme. *J Biol Chem* **287**: 12805-14 (2012).

29. Wu RA and Collins K. Human telomerase specialization for repeat synthesis by unique handling of primer-template duplex. *EMBO J* **33**: 921-35 (2014).

30. Jiang J, Wang Y, Susac L, Chan H, Basu R, Zhou ZH, and Feigon J. Structure of Telomerase with Telomeric DNA. *Cell* **173**: 1179-90 e13 (2018).

31. Wu RA, Upton HE, Vogan JM, and Collins K. Telomerase Mechanism of Telomere Synthesis. *Annu Rev Biochem* **86**: 439-60 (2017).

32. Mihalusova M, Wu JY, and Zhuang X. Functional importance of telomerase pseudoknot revealed by single-molecule analysis. *Proc Natl Acad Sci USA* **108**: 20339-44 (2011).

33. Cole DI, Legassie JD, Bonifacio LN, Sekaran VG, Ding F, Dokholyan NV, and Jarstfer MB. New models of *Tetrahymena* telomerase RNA from experimentally derived constraints and modeling. *J Am Chem Soc* **134**: 20070-80 (2012).

34. Cash DD and Feigon J. Structure and folding of the *Tetrahymena* telomerase RNA pseudoknot. *Nucleic Acids Res* **45**: 482-95 (2017).

35. Deshpande AP and Collins K. Mechanisms of template handling and pseudoknot folding in human telomerase and their manipulation to expand the sequence repertoire of processive repeat synthesis. *Nucleic Acids Res* **46**: 7886-901 (2018).

36. Berman AJ, Akiyama BM, Stone MD, and Cech TR. The RNA accordion model for template positioning by telomerase RNA during telomeric DNA synthesis. *Nat Struct Mol Biol* **18**: 1371-5 (2011).

37. Autexier C and Greider CW. Functional reconstitution of wild-type and mutant *Tetrahymena* telomerase. *Genes Dev* **8**: 563-75 (1994).

Chapter 5 - Unfinished Projects

5.1 - CR4/5 Disease Mutants

Introduction:

Germ-line mutations in the CR4/5 domain result in severe developmental defects including bone marrow failure, aplastic anemia, and liver fibrosis, among others [1] [2] [3]. Design of therapeutics targeting telomerase is hindered by an incomplete understanding of the role of CR4/5 in normal telomerase assembly and function. It has been proposed that the formation of the P6.1 stem is important for assembly [1] [4] and the loop of P6.1 forms protein contacts necessary for telomerase function [5]. An examination of the location of the disease mutants reveals that one of the mutants falls in the helix of P6.1 and likely disrupts P6.1 stability. Another is in the P6.1 loop and likely disrupts protein-RNA contacts. However, all other disease mutations fall outside the P6.1 helix and their role in the dis-function of telomerase remains unclear (**Fig 5.1A**). In light of previously discussed studies regarding structural heterogeneity in the bulged junction of the CR4/5 domain, chemical mapping was performed to determine whether disease mutants alter CR4/5 structural equilibrium.

Results:

RNA mutants were generated using Primerize and chemical mapping and data analysis was performed as previously described. With the exception of G309U, which is in P6.1 loop, all predicted structures depart from the canonically predicted P6.1 structure (**Fig 5.1B**). G325U, a mutation in P5, showed the most dramatic divergence in structure, though this data must be repeated as low RT efficiency was observed in this sample.

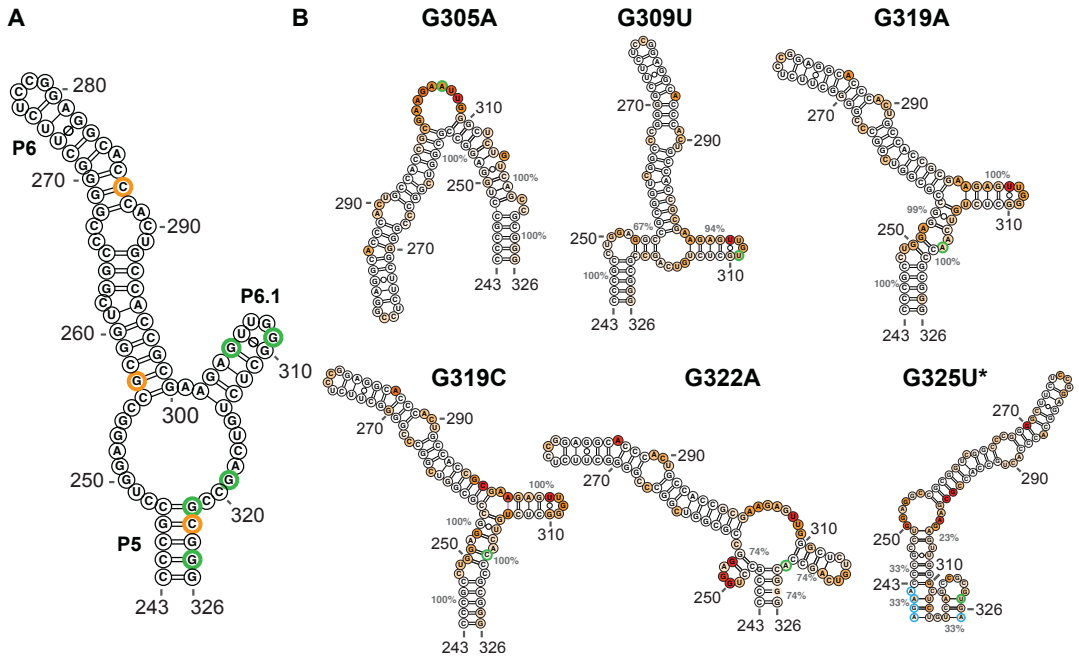


Figure 5 1 - Structure probing of human CR4/5 disease causing mutants.

(A) Canonical CR4/5 structure with disease mutants that were probed (green) and mutants that were not probed (yellow). (B) Chemical mapping results of disease mutants with reactivities mapped onto predicted structures. Disease mutants shown in green. Base pairing predictions with buffer nucleotides shown in blue.

Discussion:

The most frequently studied aspect of the CR4/5 domain is role of the formation of P6.1 and the P6.1 loop sequence requirement for telomerase activity. However, all but two disease associated mutations in the CR4/5 domain fall outside the P6.1 helix and loop. Three mutations are associated with the junction bulge and two with the P5 stem. These mutations, outside the 'important helix' of the CR4/5 domain, again suggest that dynamics at the junction play an important role within telomerase biology in human disease states.

In all constructs P6 is stably formed. At the junction, only G309U, which mutates the loop of P6.1, forms a canonical P6.1 structure and all mutants exhibit a variety of alternate helices and cross-junction clamps. Because the junction region of

the CR4/5 domain is structurally heterogeneous, there is only so much chemical mapping can reveal about the individual structures within the ensemble. However, these mutants are ideal candidates for FRET experiments. Experiment characterizing the structural states of these mutant RNAs using FRET may not only reveal the nature of disease mutants in telomerase but may also shine light on the different structure present in the native RNA. An excellent follow up to FRET-chemical mapping experiment might be design of an oligo based drug or small molecule therapeutic to restore the native structural equilibrium.

5.2 - Yeast CR4/5 Chemical Probing

Introduction

Studying TR has been made difficult by the lack of sequence conservation between species. This challenge is most obvious in yeast which show a massive expansion in TR length, making even identification of conserved domains difficult [6]. Through phylogenetic co-variation studies and structural prediction alignment it was proposed that in yeast CS5a was equivalent to P5, CS5b to P6 and CS6 to P6.1 [7]. Mutation and functional screens in the model system Kluyveromyces lactis (K. lactis) generated a minimal construct, demonstrating a functional role for this region of the RNA, supporting the in silico modeling.

Prior work by graduate student Cherie Musgrove utilized the minimized K. lactis system for a series of FRET experiments and data generated hinted at some interesting dynamics in a subset of the RNA population were observed (see Cherie thesis for details). As a follow-up to this work chemical probing was used to analyze the proposed minimal K. lactis three-way junction structure.

Results:

The minimal yeast CR4/5-like construct was probed using the same protocol described previously. Data generated was analyzed as described in Chapter 2.2. Similar to all other RNA constructs used in that series of experiments a series of buffers, hairpins and RT binding sequences were added to the sequence of interest. Normalized were used as weight and a proposed secondary structure was generated (**Fig 5.2**).

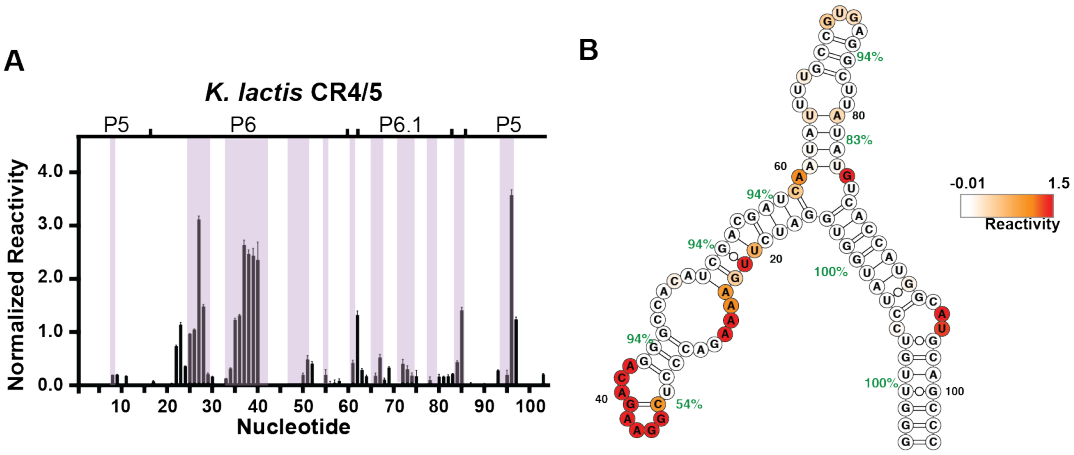


Figure 5 2 - Structure probing of K. lactis CR4/5.

(A) Normalized reactivity of 1M7 on the *K. lactis* CR4/5. Purple bars highlight regions that are expected to be reactive based on the predicted secondary structure. (B) Normalized 1M7 reactivities were used at weights and mapped onto the predicted secondary structure. Bootstrap confidence values for each helix are in green percentages.

Discussion:

The predicted secondary structure output by RNAstructure is the same as the published predicted structure [7]. However, a careful analysis of the data reveals several bulges and loops in the RNA that are not reactive, despite being predicted as single stranded. This lack of reactivity is not due to poor RT efficiency as the distal buffer region shows robust signal (data not shown). This suggests that this is may

not be the correct secondary structure or alternately that a series of tertiary interactions are shielding these regions of the RNA.

Cherie's FRET data also suggests that a large-scale conformational rearrangement occurs between P5 (CS5a) and P6 (CS5b). To date human and yeast CR4/5 RNA constructs have been studied in the lab using FRET and both reveal structurally heterogeneous and dynamic systems. The crystal structure and NMR structure of the medaka CR4/5 domain also show two dramatically different tertiary structures. It would be exciting to perform additional experiments using FRET on these three systems, potentially revealing a set of conserved structures and dynamics for this domain of TR.

To obtain a more detailed view on RNA tertiary structure an M2 experiment in which single (or double) mutants are made across the full RNA could be performed. Mutations that result in an increase in reactivity in the distal loops and bulges would provide evidence for a tertiary interaction.

5.3 - DNA:RNA Duplex Handling in TERT

Introduction:

One of the most studied aspects of telomerase is its ability to add multiple telomeric repeats in a single binding event, or Repeat Addition Processivity (RAP). For this to occur, after the template has been fully copied and the new telomere has been made - 1) the template:DNA hybrid must be melt 2) the template must be repositioned such that 3' end is again positioned in the active site 3) the new end of the telomere must be repositioned and annealed to the template (**Introduction - Figure 1.2**). The order, conformational changes, and nucleic handling requirements to achieve RAP has been a source of intense scrutiny in the telomerase field for

decades and a wide range of models have been proposed [8] [9] [10] [11]. A series of mutations in human TERT were identified that were proposed to play a role in RNA:DNA duplex handling during the RAP [12]. Using gel based nuclease experiments it was proposed that these different mutants had different primer handling kinetics [12]. The model proposed in the Wu paper states that mutations made in the Thumb-helix, Thumb-loop, and T-motif are involved in single stranded DNA handling while mutations in motif 3 are involved in template positioning (**Fig 5.3A**). In this study, to better resolve potential nucleic handling defects select mutants were reconstituted with dye labeled PK select mutations were cloned and reconstituted with dye labeled PK and assayed using FRET.

Results:

Y667E (motif 3), L681E (motif 3), L958E (thumb loop), and K981E (thumb helix) were cloned into hTERT plasmid and expressed in the presence of in vitro transcribed hTR PK and hTR CR4/5 in rabbit reticulocyte lysate. To obtain enough material to detect activity using primer extension assays in vitro transcribed PK and CR4/5 were present at 10uM concentration during the reconstitution as opposed to 1uM. Using primer extension activity assays, functional defects in the mutants were recapitulated (**Fig 5.3B**).

TERT mutants were reconstituted with U42 dye labeled PK and assayed using smFRET. Each mutant is capable of forming a complex with dye labeled primer annealed to the surface of the slide (**Fig 5.3C**). When activity buffer was flown into the cells not all of WT telomerase moved to a low FRET state (**Fig 5.3C**). Whether this lack in activity was due to poor reconstitution, problems with the slide

surface, or something else was never resolved because I moved to working on the CR4/5 project.

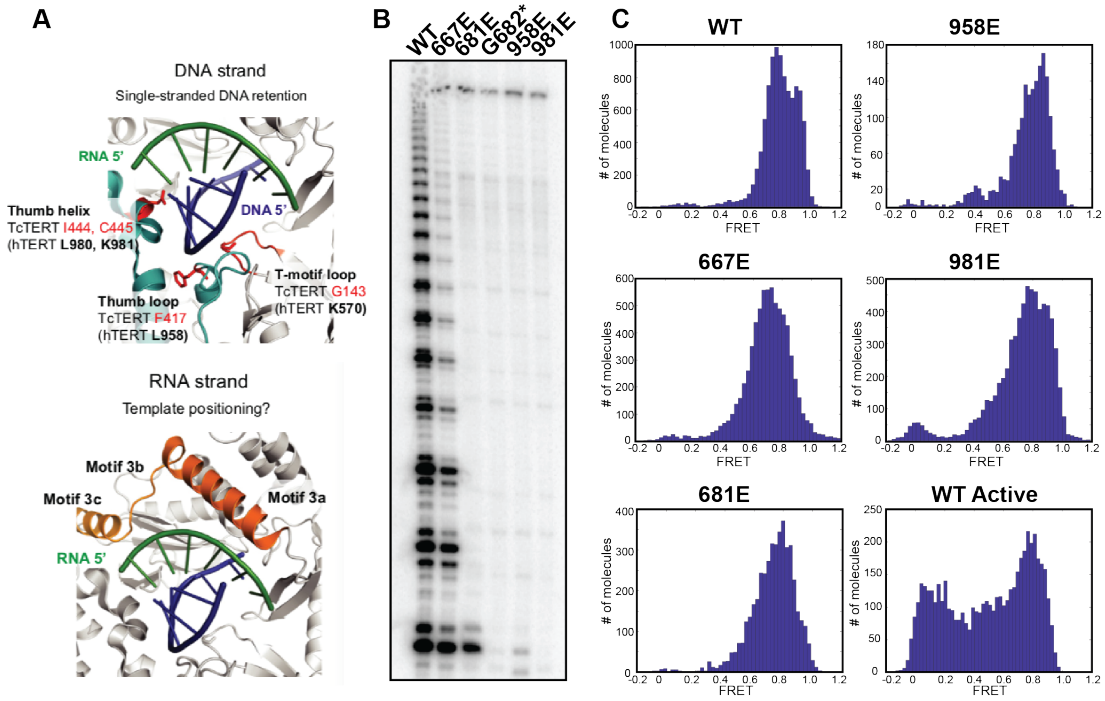


Figure 5 3 - DNA:RNA handling mutants immobilize for FRET analysis. (A) Adaptation of figure from Wu 2017 on model for template:primer handling. (B) Mutants display functional defects as previously described. *Indicates a mutant that was made but did not undergo FRET analysis. (C) FRET distribution of each mutant. The final panel exhibits FRET distribution after activity buffer has been added.

Discussion:

If you don't want to do any of the massively painful cloning and want to reap the glorious rewards of revealing more clues behind the mystery of RAP, then this project is for you. FRET experiments rely on the complex being immobilized on the surface using a primer:template hybrid, therefore it was not clear from the outset that these mutants, specifically chose to disrupt the DNA:RNA interaction, could be able to be immobilized. However, all mutants appears to be competent for the experiment,

though some mutants displayed low molecule counts overall. After flowing in activity buffer into the immobilized telomerase complexes an analysis of either enzyme dissociation, PK dynamics as a read out for template movement [13], or template:primer dynamics as observed in individual traces may reveal more information regarding the effect of these mutations on duplex handling.

References:

1. Alder, J.K., et al., *Diagnostic utility of telomere length testing in a hospital-based setting.* Proc Natl Acad Sci U S A, 2018. **115**(10): p. E2358-E2365.
2. Yamaguchi, H., et al., *Mutations of the human telomerase RNA gene (TERC) in aplastic anemia and myelodysplastic syndrome.* Blood, 2003. **102**(3): p. 916-8.
3. Boyraz, B., et al., *A novel TERC CR4/CR5 domain mutation causes telomere disease via decreased TERT binding.* Blood, 2016. **128**(16): p. 2089-2092.
4. Chen, J.L., K.K. Opperman, and C.W. Greider, *A critical stem-loop structure in the CR4-CR5 domain of mammalian telomerase RNA.* Nucleic Acids Res, 2002. **30**(2): p. 592-7.
5. Mitchell, J.R. and K. Collins, *Human telomerase activation requires two independent interactions between telomerase RNA and telomerase reverse transcriptase.* Mol Cell, 2000. **6**(2): p. 361-71.
6. Chen, J.L. and C.W. Greider, *An emerging consensus for telomerase RNA structure.* Proc Natl Acad Sci U S A, 2004. **101**(41): p. 14683-4.
7. Brown, Y., et al., *A critical three-way junction is conserved in budding yeast and vertebrate telomerase RNAs.* Nucleic Acids Res, 2007. **35**(18): p. 6280-9.
8. Yang, X., et al., *Knockdown of telomeric repeat binding factor 2 enhances tumor radiosensitivity regardless of telomerase status.* J Cancer Res Clin Oncol, 2015. **141**(9): p. 1545-52.

9. Berman, A.J., et al., *The RNA accordion model for template positioning by telomerase RNA during telomeric DNA synthesis.* Nat Struct Mol Biol, 2011. **18**(12): p. 1371-5.

10. Qi, X., et al., *RNA/DNA hybrid binding affinity determines telomerase template-translocation efficiency.* EMBO J, 2012. **31**(1): p. 150-61.

11. Wu, R.A. and K. Collins, *Human telomerase specialization for repeat synthesis by unique handling of primer-template duplex.* EMBO J, 2014. **33**(8): p. 921-35.

12. Wu, R.A., J. Tam, and K. Collins, *DNA-binding determinants and cellular thresholds for human telomerase repeat addition processivity.* EMBO J, 2017. **36**(13): p. 1908-1927.

13. Parks, J.W., et al., *Single-molecule FRET-Rosetta reveals RNA structural rearrangements during human telomerase catalysis.* RNA, 2017. **23**(2): p. 175-188.