

**Discrimination of ssRNA by Pot1 and Identification of a Novel CypE Aptamer through an  
Optimized RNA SELEX Protocol**

by

NEIL RYAN LLOYD

B.S., Colorado School of Mines, Golden, CO 2012

A thesis submitted to the  
Faculty of the Graduate School of the  
University of Colorado in partial fulfillment  
of the requirement for the degree of  
Doctor of Philosophy  
Department of Biochemistry

2018

This thesis entitled:  
**Discrimination of ssRNA by Pot1 and Identification of a Novel CypE Aptamer through an  
Optimized RNA SELEX Protocol**

written by Neil Ryan Lloyd  
has been approved for the Department of Biochemistry by

---

Deborah S. Wuttke

---

Robert T. Batey

August 22nd, 2018

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

## Abstract

Lloyd, Neil R. (Ph.D., Biochemistry)

### **Discrimination of ssRNA by Pot1 and Identification of a Novel CypE Aptamer through an Optimized RNA SELEX Protocol**

Thesis directed by Professor Deborah S. Wuttke

Protein-ligand specificity forms the fundamental basis for many biological mechanisms with properly tuned binding being required for most biological processes. Aberrant interactions can result in consequences ranging from wasted cellular resources to disease pathologies and death. As such, characterizing interaction specificities is a critical step in understanding biological systems. In this thesis I have characterized the RNA-binding properties of the telomere protection protein Pot1, developed an optimized SELEX protocol to characterize RNA-binding by newly identified RNA-binding proteins, and expanded on the RNA-binding specificity of one of those proteins, the epigenetic regulator CypE.

High fidelity binding to ssDNA, but not ssRNA, is integral to the function of the essential telomere end protection protein Pot1, In *S. pombe*, this presents a unique challenge as the C-terminal domain of the DNA-binding domain, Pot1pC, exhibits non-specific ssDNA recognition, achieved through thermodynamically equivalent alternative binding conformations. Given this malleability, how simultaneous specificity for ssDNA over RNA is achieved was unclear. Examination of the ribose-position specificity of Pot1pC shows that ssDNA specificity is additive but not uniformly distributed across the ligand. High-resolution structures of Pot1pC in complex with RNA-DNA chimeric ligands reveal Pot1pC discriminates against RNA by utilizing conserved non-compensatory binding modes that feature significant rearrangement of the binding interface. These alternative conformations, accessed through both ligand and protein flexibility, recover much, but not all, of the binding energy, leading to the observed reduction in affinities

suggesting that intermolecular interfaces are remarkably sophisticated in their tuning of specificity towards flexible ligands.

Recent discovery of widespread RNA-binding by unexpected RNA binders highlights the need for functional characterization of these non-canonical RNA-binding domains. SELEX, combined with new sequencing technologies, represents an ideal technique to do this. Using CypE, an RNA-binding cyclophilin involved in splicing and chromatin remodeling, I have optimized a selection protocol for other cyclophilins. Selection against CypE, while not identifying an RNA that binds the cyclophilin, reveals an aptamer with 20-fold tighter binding than previously reported with an extended binding interface on the RRM, suggesting RNA as a competitive ligand for CypE and provoking implications for the role of RNA in CypE gene repression.

## **Acknowledgements**

Over the past 6 years, I have found myself in a debt of gratitude to lot of people, without which this work would have been much harder, if not impossible.

First and foremost, I would like to thank Prof. Deborah Wuttke for her excellent mentorship and support during this time. She has pushed me to be a better scientist, provided valuable insights at nearly every turn, and always being there to keep my spirits up. I would also like to thank my committee members Rob Batey, Roy Parker, Rob Kuchta, and Loren Hough for their advice and feedback on my research. In particular, Rob Batey was intimately involved with feedback and discussions for much of this work, and in several cases gave key insights to get me back on track. Day in and out, the Wuttke lab, Marissa, Meagan, Nick, and Leslie provided general moral support, thoughtful scientific discussions, and invaluable editing. Past lab members Karen and Thayne were both key to my initial integration into the lab.

I also need to thank the department and the greater JSCBB community for their support. In particular, Annette was always there to help whenever instruments or experiments were giving me trouble. Amber and Jamie for help with the technical side of preparing sequencing libraries. The people in charge of organizing everything also made doing the science much easier, with many thanks to Kim, Pamela, Lin, and Carla and the others who did their jobs so well I don't even know their names.

The people who kept me sane also need to be acknowledged. My cohort during the first year, my friends and family, and especially Dani who has always been there throughout it all.

## Contents

<b>Chapter 1 – Pot1 Biology</b> .....	<b>1</b>
1.0 – Chapter Overview: .....	1
1.1 – Overview of Telomeres.....	1
1.1.1 – Telomeres and Telomerase.....	1
1.1.2 - Shelterin and the Pot1 protein:.....	3
1.1.3 - Structures of Pot1 Proteins Reveal How Binding Affinity and DNA Specificity are Achieved .....	4
1.1.4 – How the Pot1 subdomains work together .....	12
1.1.5 - How Pot1 Might Regulate Telomerase .....	14
1.2 – Overview of Pot1 and RNA at the telomeres.....	17
1.2.1 - Challenges Facing Telomere End-Protection Proteins .....	17
1.2.2 – Transcribed Telomeres and Cellular RNAs Represent a Pool of Potential Pot1 Substrates.....	18
1.2.3 - RNA Discrimination by mPOT1 .....	20
1.2.4 – RNA Discrimination by full-length <i>S. pombe</i> and Pot1pN .....	21
1.2.5 – RNA Discrimination by Pot1pC and Novel Insight into Protein-Nucleic Acid specificity .....	22
<b>Chapter 2 – Pot1pC Discrimination of RNA Backbones</b> .....	<b>23</b>
2.0 – Chapter Overview: .....	23
2.1 – Introduction.....	23
2.2 – Materials and Methods .....	23
2.2.1 – Protein Expression and Purification.....	23
2.2.2 – Isothermal Titration Calorimetry .....	24
2.2.3 – Crystallization.....	25
2.2.4 – Data Collection and Refinement.....	25
2.3 – Results .....	27
2.3.1 – Pot1pC discriminates against RNA additively by ribose position .....	27
2.3.2 – Discrimination in the 1R Structure .....	30
2.3.3 – Discrimination in the 1-3R Structure .....	33
2.3.4 – Cryptic secondary binding mode is widely used to provide partial thermodynamic compensation .....	37
2.3.5 – Ligand Accommodation in the 7-9R Structure .....	39
2.3.6 - Model for 4-6R and full RNA Pot1pC complex structures .....	40
2.4 - Discussion.....	42

<b>Chapter 3 – Implications of Newly Discovered RNA-Binding By Cyclophilins</b> .....	<b>48</b>
3.0 – Chapter Overview .....	48
3.1 - Identification of Cyclophilins as Non-canonical RNA-Binding Proteins .....	48
3.2 - Cyclophilins are a family of key biological regulatory proteins.....	51
3.3 – Possible Mechanisms of Cyclophilin RNA-binding .....	54
3.3.1 – RNA-binding cyclophilins share features suggestive of an RNA-binding surface ....	54
3.3.2 – Heparin binding Interface of CypA and CypB .....	55
3.3.3 – Surface Residue Variation and Charge Distribution.....	57
3.4 – Cyclophilin Enzymatic Dynamics and the Potential for Allosteric Regulation by RNA....	58
3.4.1 – CypA Conformational Dynamics Reveal Two Allosteric Sites .....	59
3.5 – The Known Functions of the Putative RNA-Binding Cyclophilins .....	61
3.5.1 – The myriad, Wide Ranging Functions of CypA .....	61
3.5.2 - Functions of Cpr1.....	62
3.5.3 – Functions of CypB.....	62
3.5.4 – Functions of CypE.....	63
3.5.5 – Functions of PPIL4/AtCyp59 .....	64
3.5.6 – Functions of CypG .....	65
3.6 – RNA may play mechanistic roles in cyclophilin-mediated regulatory activities .....	65
<b>Chapter 4 – Optimization of SELEX Protocol with MS2 Coat Protein</b> .....	<b>68</b>
4.0 – Chapter Overview: .....	68
4.1 – Introduction.....	68
4.2 – Methods.....	70
4.2.1 – Protein Expression and Purification.....	70
4.2.2 – Library Binding through EMSA .....	71
4.2.4 – SELEX .....	71
4.2.5 – Sanger sequencing .....	72
4.2.6 – High-throughput sequencing .....	72
4.2.7 – Bioinformatics Pipelines .....	73
4.3 – Results .....	76
4.3.1 – Selected RNA pool binds tighter than the round 0 library .....	76
4.3.2 – Sanger sequencing of round 6 RNAs reveal 6 unique sequences containing MS2 binding sites.....	77
4.3.3 – High-Throughput Sequencing of 8 Rounds of SELEX Reveals Library Biases and Enrichment of Sequences .....	78

4.3.4 – QIIME Clustering reveals 19 clusters comprised of >0.5% of all unique sequences .....	82
4.3.5 – FASTAptamer and AptaSUITE Recapitulate the Clustering By QIIME and AptaSUITE Reveals Ubiquitous Presence of Ideal Binding Motif in Selected Pools .....	83
4.4 – Discussion .....	87
4.4.1 – AptaSUITE is the Most Comprehensive HT-SELEX Analysis Pipeline Currently Available .....	87
4.4.2 – MS2 Results Agree with Previous Literature .....	89
4.4.3 – SELEX Protocol Optimization.....	90
<b>Chapter 5 – Identification of a Tight Binding CypE Aptamer through an Optimized SELEX Protocol .....</b>	<b>92</b>
5.0 – Chapter Overview .....	92
5.1 – Introduction.....	92
5.2 – Methods.....	94
5.2.1 – Protein Expression and Purification.....	94
5.2.2 – Expression and Purification of <sup>15</sup> N Labeled Recombinant Protein.....	95
5.2.3 – Electromobility Shift Assays (EMSAs) .....	96
5.2.4 – <i>In vitro</i> peptyl-prolyl isomerase (PPlase) assay .....	96
5.2.5 – SELEX Experiments.....	98
5.2.6 – High-throughput sequencing .....	102
5.2.7 – QIIME and AptaSUITE Analysis .....	105
5.2.8 – NMR-HSQC Titration Experiments.....	105
5.3 – Results .....	107
5.3.1 – Activation of the PPlase Activity of CypA, Full-length CypE and CypE CLD by RNA .....	107
5.3.2 – SELEX with 25N Library for 7 Rounds Resulted in Insufficiently Selected RNA Pools and Reveals Constant-Random Region Pairing .....	108
5.3.3 – Preliminary Evidence of RNA Binding and Benchmark for Selection Libraries.....	111
5.3.4 – SELEX Experiment 2.....	116
5.3.5 – Final optimized SELEX protocol SELEX Experiment 3 Produces Enriched Aptamer Sequences.....	118
5.3.6 – Screening Aptamers for CLD Interactions by PPlase Activity .....	124
5.3.7 – Preliminary Characterization by EMSA Suggests SO-1 Aptamer is the Tightest Binder of Aptamers Tested .....	126
5.3.8 – NMR HSCQ Titration of CypE-RRM, CypE-CLD, and FL-CypE Reveal SO-1 Binding Solely to the CypE-RRM and Independently Behaved Subdomains.....	126
5.4 – Discussion .....	132



5.4.1 – Iteration of the SELEX Protocol Provides Insights into Optimization.....	132
5.4.2 – SELEX Sequencing Results Point to Several Promising Binding Motifs .....	134
5.4.3 – Inconsistent PPlase Assay Warrants Further Controls .....	134
5.4.4 – NMR Characterization of CypE Interfaces Suggests Possible Mechanisms of RNA Regulation.....	135
<b>References .....</b>	<b>136</b>
<b>Appendix A.....</b>	<b>167</b>
Protein Expression and Purification.....	167
Isothermal Titration Calorimetry .....	172
Crystallization .....	188
Data Collection and Refinement.....	189
<b>Appendix B.....</b>	<b>190</b>
Protein Cloning, Expression, and Purification.....	190
Protein Expression and Purification.....	190
Nickel-NTA Affinity Column Purification .....	192
Removing the His-SUMO affinity tag.....	193
Concentration and Size Exclusion Chromatography .....	195
Template PCR and RNA Library Preparation .....	200
PCR amplification of initial SELEX library.....	200
RNA Transcription.....	202
RNA Purification.....	203
Library Binding through EMSA .....	205
SELEX.....	208
Pre-selection against Co-NTA beads .....	208
Binding Equilibrium Reaction, Washing, and Elution .....	209
Reverse Transcription(RT)-PCR .....	209
High-throughput sequencing .....	210
<b>Appendix C.....</b>	<b>224</b>
Overview of the Analysis Pipeline .....	224
Demultiplexing with QIIME 1.9 .....	225
Removing 3' Constant and Illumina Barcode Primer with QIIME 1.9 .....	226
Scripts for Prepping QIIME Processed Data for AptaSUITE.....	230
AptaSUITE Pipeline Scripts .....	231
Example AptaSUITE config file .....	232
QIIME batey_mapping_file/Batey Barcodes (Reverse Complement of primer sequence)	237

filter_otu_mapping_from_otu_table.py .....	243
--	-----

## List of Tables

Table 2.1 Data Collection and Refinement Statistics for Ribose Chimeric Pot1pC Complexes..	26
Table 2.2 Table 2.2 Thermodynamic Impact of Ribose Substitutions. ....	27
Table 3.1 Summary of Human Cyclophilins.....	53
Table 5.1 Summary of the variable selection conditions across the 3 selection experiments.....	98
Table 5.2 List of Primers and Oligos Used in SELEX Experiments.....	103
Table 5.3 Summary of sequencing read statistics and the sequencing method used. ....	110
Table 5.4 Summary of apparent $K_D$ s observed for the initial 50N library through EMSA.....	112
Table 5.5 The Predicted Secondary Structures and Condition Origins of the Most Abundant Cluster Seed Sequences. ....	122
Table 5.6 Enriched 6-mer Seed Sequences Enriched in the First 8 Rounds by Condition.....	124
Table 5.7 Preliminary EMSA Binding Affinities for Condition “Winners.” .....	126

## List of Figures

Figure 1.1 The Shelterin Complex.....	3
Figure 1.2 Domains of Pot1.....	5
Figure 1.3 Disease mutations in the DNA-binding domain of hPOT1 .....	5
Figure 1.4 The structural similarities and differences of hPOT1 and spPot1.....	7
Figure 1.5 The hydrogen bond networks for spPot1 .....	8
Figure 1.6 Plastic accommodation of the DNA ligand for spPot1pC. ....	11
Figure 1.7 RNA Discrimination by mPot1 and Pot1pN.....	21
Figure 2.1 Most Cognate Ligand 2' Hydroxyl Positions and All Thymine Methyl Groups Are Solvent Exposed .....	28
Figure 2.2 An Unfavorable Interaction Between rG1 Hydroxyl, Trp72, and G2 Base is the Strongest Individual Discrimination Determinant .....	31
Figure 2.3 Ligand Electron Density for 1R and 1-3R .....	32
Figure 2.4 1-3R Binds Pot1pC in an Alternative Binding Mode, Recapitulates the Unfavorable Interactions Seen in the 1R Structure with the Additional Loss of the His109-G1 Interaction.....	34
Figure 2.5 Structural Rearrangement of the 1-3R Ligand Driven by Conformational Changes in the Sugar Phosphate Backbone and Alleviation of Steric Clashes at the Bases.....	35
Figure 2.6 Rearrangement of the 1-3R Ligand Results in a Mix of Lost and Compensatory Interactions .....	36
Figure 2.7 1-3R Binds Pot1pC More Like the T4A and G2C DNA Ligands Than the Cognate DNA Ligand .....	38
Figure 2.9 Comparison of Sugar rG1 Sugar Conformations in 1R and 1-3R Complexes and Model for Discrimination at Positions 4-6.. ....	41
Figure 3.1 General Protocol Schematic of the Crosslinking Studies.....	49
Figure 3.2 Structural Features of Cyclophilins Mapped onto CypA.....	55
Figure 3.3 Heparin Binding Residues of CypA and CypB.....	56
Figure 3.4 Electrostatic Surfaces of Select Cyclophilins Implicated as RNA Binding.. ....	57
Figure 3.5 Substrate Conformations in the Active Site. ....	59
Figure 3.6 Independently Coupled Dynamics of Two Sites Reveal Possible Allosteric Sites. ....	60
Figure 3.7 Schematic of CypE Proposed Mechanism of Gene Repression. ....	64
Figure 3.8 Schematic of Possible Mechanisms of RNA Regulation of Cyclophilins .....	66
Figure 4.1 Figure 4.1 MS2 coat protein Consensus Binding Motif. ....	69

Figure 4.2 Figure 4.2 Schematic Diagram of the SELEX Protocol Used.....	72
Figure 4.3 MS2 Binding to the Initial and Selected Libraries.....	77
Figure 4.4 Sanger Sequencing Reveals MS2 consensus binding site in Round 6..	78
Figure 4.5 Length distribution of Round 0 Sequences. ....	79
Figure 4.6 Nucleotide base read distribution of round 0 random region.....	80
Figure 4.7 Nucleotide base read distribution of round 8 random region.....	80
Figure 4.8 Most Reads are Singleton Sequences but Enriched Sequences Become Evident by Rounds 3 and 4.....	82
Figure 4.9 Enriched 9-mer Consensus Reveals Ideal Binding Site Biases Towards the 5' half of the Random Region..	84
Figure 4.11 Second Most Abundant MS2 Aptamer Sequence Utilizes Random Region Pairing with the 5' Constant Region to Form MS2 Binding Site. ....	86
Figure 5.1 Cartoon of SELEX Library Designs. ....	97
Figure 5.2 PPLase Activation of CypA and CypE by RNA. ....	107
Figure 5.3 Quantification of Enzyme Efficiency Activation. ....	108
Figure 5.4 Representative EMSA gels shown for CypE-CLD. ....	111
Figure 5.5 [Ligand] Dependent Binding of FL-CypE by EMSA.....	113
Figure 5.6 Coomassie Staining for Protein Constructs Used in SELEX 1 and 2 and EMSA assays. ....	114
Figure 5.7 NMR HSQC Titration of 15N CypA by 25N Library.....	115
Figure 5.8 Representative Enrichment Statistics for SELEX 2 Reveals No Aptamers. ....	117
Figure 5.9 Sequence enrichment by Round for Selection 3.....	119
Figure 5.10 Abundance Heatmap of Top Cluster Families by Condition and Round..	121
Figure 5.11 Pervasive Activation of the PPLase Activity of CypE-CLD by SELEX Aptamers....	125
Figure 5.12 SO-1 Binding to CypE-RRM by NMR HSQC Titration..	127
Figure 5.13 Comparison of Phe70 (Phe51 native) Chemical Shift Changes During SO-1 Binding. ....	128
Figure 5.14. SO-1 Chemical Shift Changes Map to the Canonical RRM Interface.....	129
Figure 5.15 CypE-RRM Mapping of the Chemical Shift Changes with SO-1 Reveal Overlapping Surface with Previous Interactions. ....	129
Figure 5.17 Comparison of Free HSQCs Reveal Independently Behaved Domains.....	130
Figure 5.16 CypE-CLD HSQC SO-1 Titration Shows Little to No Interaction.....	130
Figure 5.18 SO-1 Binding to FL-CypE by NMR HSQC Titration. ....	131

## **Chapter 1 – Pot1 Biology**

### **1.0 – Chapter Overview:**

The first half of my thesis work focuses on the detailed characterization of RNA discrimination by the **protection of telomeres** protein (Pot1). This chapter is meant to provide the background of Pot1 biology. As such, it gives an overview of what telomeres are, why Pot1 needs to protect them, how it does it and a limited overview of the other players involved, what else Pot1 does in telomere biology, and how RNA and RNA discrimination play into Pot1 biology. Note: Much of the text and figures of this chapter has been previously published in papers I published as first author.<sup>1,2</sup>

### **1.1 – Overview of Telomeres**

#### **1.1.1 – Telomeres and Telomerase**

Telomeres are the nucleoprotein caps at the ends of linear chromosomes<sup>3-7</sup> that buffer against the loss of genomic DNA.<sup>8-11</sup> This specialized heterochromatin comprises a region of repetitive non-coding DNA that terminates in a conserved single-stranded overhang<sup>5,11,12</sup> and protein complexes that tightly bind telomeric DNA. These proteins protect the DNA from degradation, prevent the erroneous recognition of the single-stranded overhang as DNA damage, and regulate the extension of telomeres by the reverse-transcriptase telomerase.<sup>13-18</sup> During DNA replication, daughter strands are shortened because DNA polymerase requires RNA primers in lagging strand synthesis that cannot be replaced by DNA at the extreme 5' ends of the chromosome.<sup>19</sup> Further shortening arises from the replication of the shorter C-rich strand. The loss of telomeric DNA is exacerbated during telomere processing by the action of the Exo1 and Apollo/SNM1B nucleases, which resect the 5' end to create the overhang at mammalian telomeres. The processing pathway to generate these ends in budding yeast is different, involving the Sae2-MRX exonuclease pathway (<sup>20-23</sup> and reviewed in<sup>24</sup>) but in both cases the ends produced by DNA replication are resected. These processing pathways standardize both

the 3' overhang length and the sequence register.<sup>25</sup> Progressive DNA replication and telomere processing leads to the shortening of telomeres until a critically short length triggers cells to cease dividing in a process known as senescence.<sup>26</sup> As the name suggests, this process is thought to be involved in aging, and telomere length correlates with age. As a result, telomere length is a target for age-related therapeutics and health diagnostics.<sup>26-28</sup>

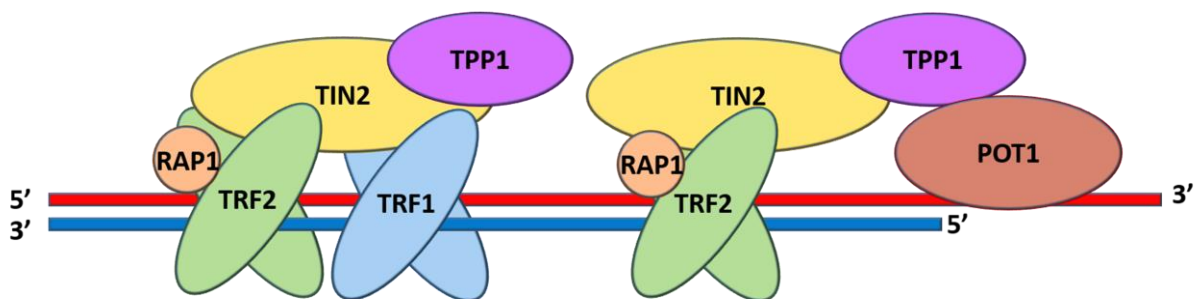
To combat the loss of telomeric DNA, stem cells and unicellular organisms utilize the reverse transcriptase telomerase to replenish telomere length.<sup>29,30</sup> Comprised of a template/scaffolding RNA and a protein subunit related to viral reverse transcriptases,<sup>31-33</sup> telomerase catalyzes the addition of dNTPs to the 3' end of chromosomes by partially aligning the template RNA to 3' overhang.<sup>34,35</sup> This addition of DNA proceeds with high nucleotide processivity and repeat addition processivity whereby a single telomerase molecule can dissociate and realign the template RNA to add multiple DNA repeats.<sup>33,36</sup> Following the addition of repeats to the 3' overhang, standard 5' to 3' DNA synthesis then produces the 5' strand. Together this results in the lengthening on the double-stranded telomeric DNA. As cancer cells must also resolve the end replication problem, it is unsurprising that many cancer cell lines utilize telomerase to do so. Approximately 90% of all human cancer activate telomerase, making it a highly sought out target for cancer therapeutics.<sup>37-39</sup>

Human chromosomal telomeres are typically comprised of 5-15 kb double-stranded DNA and 50-500 nucleotides of a ssDNA 3' strand at the end of eukaryotic chromosomes. Based on the telomerase template, the sequence repeat added to telomeres is GGTTAG. However, high-throughput sequencing has revealed a significant sequence variation in telomeres in both primary and immortalized human cell lines,<sup>11,12,35,36</sup> suggesting a combination of DNA mutations and inconsistent repeat addition have created significant variation in the telomere sequence. An even greater sequence variation has been found the distantly related yeast species, *S. pombe* and *S. cerevisiae*.<sup>20,40-43</sup> The *S. pombe* RNA template should produce a GGTTACA repeat, but appears to do so inconsistently with frequent nucleotide deletions and additions, resulting in a

consensus best represented by the sequence: GGTTAC)(A/AC)<sub>0-1</sub>(G)<sub>0-7</sub>.<sup>44,45</sup> Likewise, the variable *S. cerevisiae* is more accurately described by the sequence (TG)<sub>1-6</sub>TG<sub>2-3</sub> than by the RNA template sequence. The *S. cerevisiae* telomere length also varies from less than 10 to over 70nts.<sup>46-49</sup> All together, these variable sequences and lengths provide substantial challenge for the proteins that interact with them and provides an excellent model system for the study of sequence specificity for DNA-binding proteins.

### 1.1.2 - Shelterin and the Pot1 protein:

The shelterin protein complex is major protein component of the specialized chromatin found at telomeres. The shelterin complex is responsible for capping and protecting the telomere in most eukaryotes. While not conserved the model organism *S. cerevisiae*, the shelterin complex is roughly conserved from fission yeast to humans (reviewed by Palm and de Lange<sup>6</sup>). Comprised of six proteins, the complex contains dsDNA-binding proteins (TRF1 and TRF2 in humans, Taz1 in *S. pombe*), ssDNA-binding proteins (Pot1 in both species), bridging proteins (TIN2 and TPP1 in humans, Rap1, Poz1, and Tpz1 in *S. pombe*) and other associated proteins (RAP1 in humans, Ccq1 in *S. pombe*) (**Figure 1.1**).<sup>50</sup> Telomeres are further protected by the formation of t-loops in humans in which the ssDNA overhang loops back on the double-stranded region via a strand invasion mechanism dependent on topological changes induced in



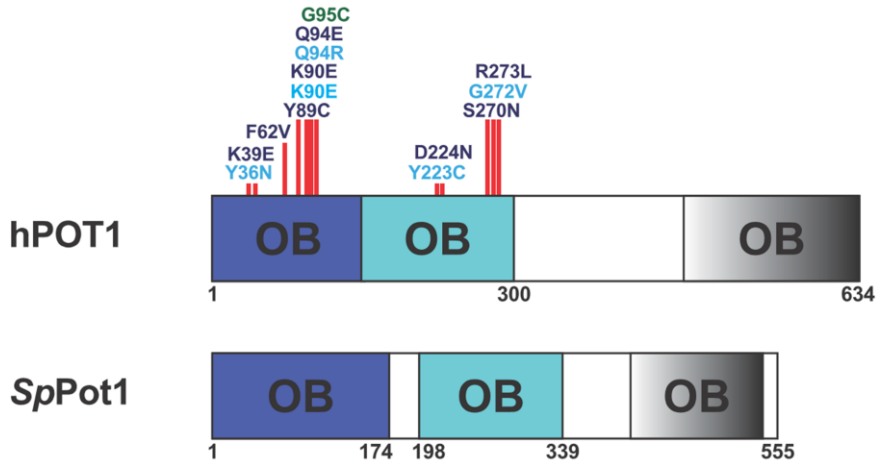
**Figure 1.1 The Shelterin Complex.** Schematic diagram of the human shelterin complex. Recent evidence suggests a core shelterin complex comprising (2)TRF2-(1)TIN2-(1)TPP1-(1)POT1 binds to the ds-ssDNA junction while a shelterin sans POT1 binds within the double-stranded region, though it is unclear if TRF2 and TRF1 can be present together.

telomeric DNA by TRF2.<sup>51-54</sup> Deletion of components of the shelterin complex has been reported to trigger an increase in the volume occupied by telomeric chromatin as well as an increase in DNA-damage response signaling at telomeres. In *S. pombe* and *S. cerevisiae*, both the duplex region and the overhang of the telomere are much shorter and thus do not appear to form t-loops.

Pot1 is the sole protein in shelterin that exhibits autonomous ssDNA-binding activity and is critical for end protection. Disruptions of human POT1 (hPOT1), mouse Pot1a, or chicken Pot1 result in activation of the Rad3-related (ATR) DNA-damage response pathways, chromosomal fusion, and cell death, likely through a failure to exclude the ATR damage sensor RPA.<sup>55-57</sup> Loss of just the ssDNA-binding activity of hPOT1, however, leads to rapid and extensive telomere elongation.<sup>58</sup> Furthermore, knockdown of hPOT1 also disrupts the terminal sequence of the 5' strand, suggesting that hPOT1 sets the register for end resection.<sup>59</sup>

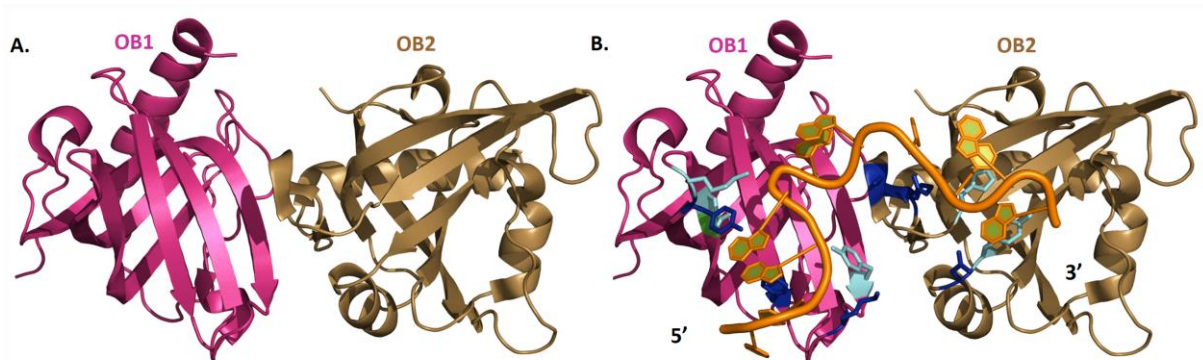
### **1.1.3 - Structures of Pot1 Proteins Reveal How Binding Affinity and DNA Specificity are Achieved**

Pot1, and telomere end-proteins in general, use a common structural topology known as the OB fold to recognize ssDNA. OB-folds are multifunctional domains found throughout biology and are frequently implicated in the recognition of disordered linear polymers, most commonly ssDNA and ssRNA.<sup>60,61</sup> The structural framework is a simple 5-stranded  $\beta$ -barrel elaborated with loops and helical elements to form a virtual platform for polymer recognition whose properties can be tailored to the desired specificity and affinity through a variety of mechanisms. The ligand can bind to a single OB fold, multiple independently binding OB folds, an extended binding interface across several OB folds in tandem, or through homo/hetero-oligomerization.



**Figure 1.2 Domains of Pot1.** Schematic domain map of Pot1 proteins with homologous domains color coded and the predicted C-terminal OB-folds shaded with a gray-black gradient.

The N-terminal portion of Pot1 contains a dual OB-fold that confers full DNA-binding activity while the C-terminal half (predicted to be an OB-fold) interacts with the shelterin component TPP1 in humans/Tpz1 in *S. pombe* (**Figure 1.2**).<sup>62,63,18</sup> Structures of both the complete human DNA-binding domain (DBD) and the 2 OB folds that together comprise the *S. pombe* DBD have been solved.<sup>64–66</sup> hPOT1 adopts an elongated structure comprised of these 2 OB folds that are closely linked together by a short 9 amino acid linker such that the two domains functionally bind



**Figure 1.3 Disease mutations in the DNA-binding domain of hPOT1** **A)** Crystal structure of hPOT1-DNA complex with DNA omitted for clarity (IXJV).<sup>66</sup> OB1 is in magenta and OB2 is light brown. **B)** hPOT1 with DNA ligand shown. The portion of the ligand bound by OB1 (OB1-6mer) is yellow and the portion bound by OB2 (OB2-4mer) is orange. GWAS mutations near the DNA binding interface are shown in cyan for CLL associated mutations, green for glioma associated mutations, and blue for mutations associated with other types of cancer.



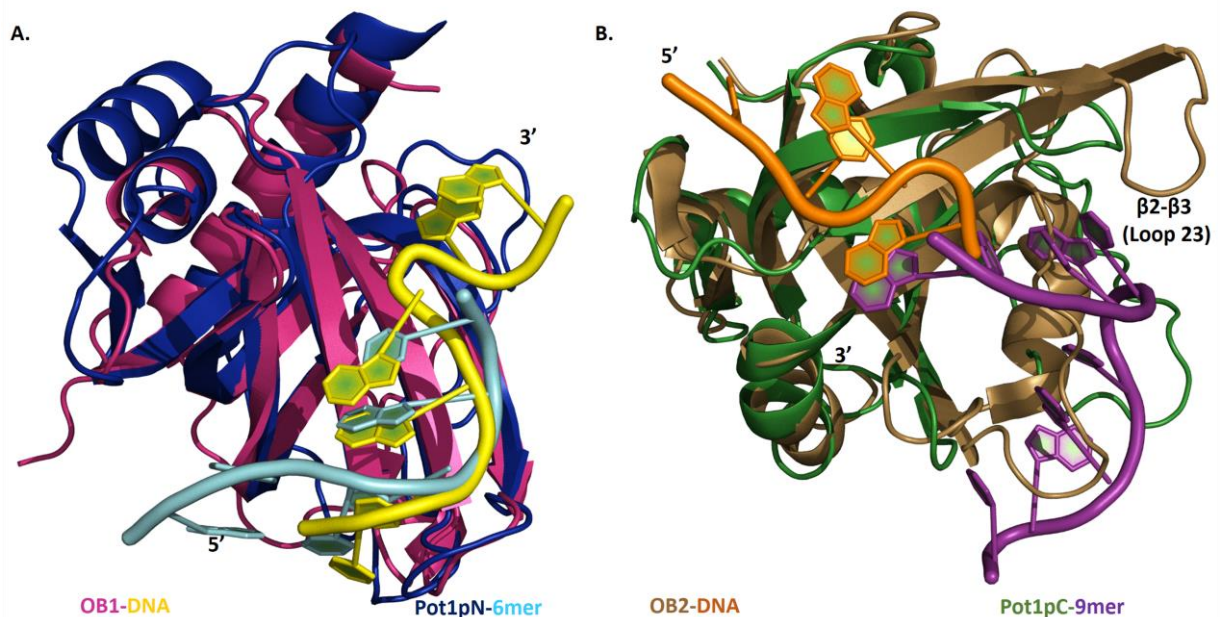
a 10-nt telomeric ssDNA ligand as one contiguous unit with an extensive domain/domain interface (**Figure 1.3A**). hOB1, the N-terminal OB fold, binds the first 6- nt (TTAGGG) with strong specificity, especially for nucleotides 2-5. hOB2, the C-terminal OB fold of the pair, binds to the final 4-nt (TTAG) with less specificity than hOB1 except for the terminal G10. Consistent with the specificity data, hOB1 forms over two-thirds of the hydrogen-bonding interactions between the ligand and the protein (22 out of 31 total). At the interface between the two, the phosphodiester bond of T7 kinks 90° to shift into the binding interface of hOB2.

Recent genome wide association studies (GWAS) have found several mutations in hPOT1, associated with chronic lymphocytic leukemia, familial glioma, and several other cancers types as well as the rare familial disorder Coat's plus.<sup>67-72</sup> In CLL, Pot1 is one of the most frequently mutated genes with 3.5% of CLL cases containing somatic mutations in Pot1. Strikingly, most of the disease-associated point mutations occur at residues contacting DNA in the crystal structure of hPOT1-DBD (**Figure 1.3B**). Some of these mutations appear to disrupt ssDNA-binding, deprotect telomeres, and trigger oncogenic fusions.<sup>67</sup> However, others appear to lack a binding defect, and exercise their influence through other pathways. *In vivo*, deletion of the DNA-binding domain of hPOT1 results in telomere elongation, supporting a role in negative length regulation.<sup>58</sup> Conversely, some of these mutations lead to telomere shortening, currently ascribed to a loss of interaction with another ssDNA binding complex of hCTC1-hSTN1-hTEN1 and suppression of appropriate lagging strand synthesis.<sup>72</sup> This differential impact speaks to the complexity of processing at the telomere and the myriad roles hPOT1 plays.

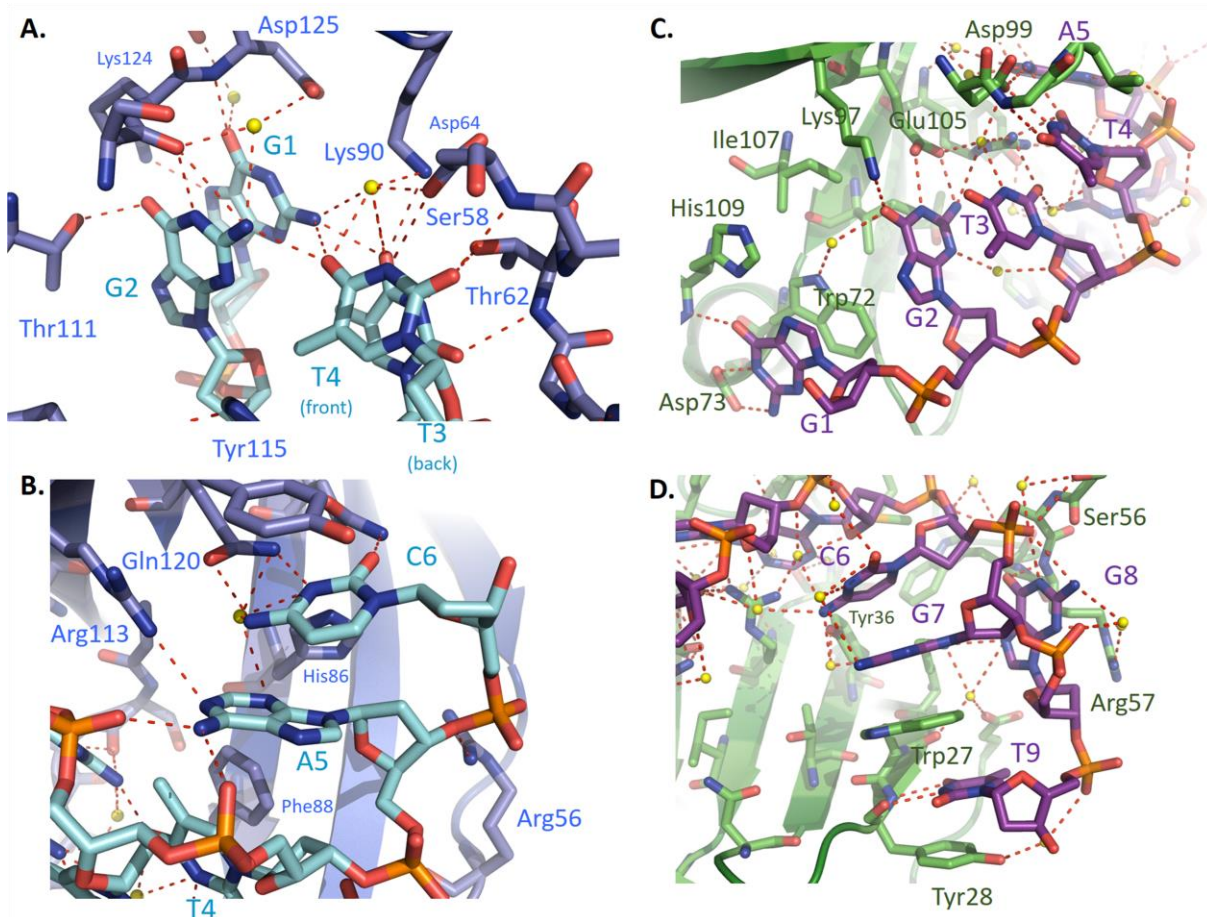
*S. pombe* Pot1 is a functional homologue of hPOT1 and shares a similar domain organization.<sup>18,64-66</sup> This includes an N-terminal DNA-binding domain (Pot1-DBD) composed of two OB-folds (Pot1pN and Pot1pC).<sup>64-66,73</sup> Biochemical experiments already suggest a difference in mechanism of action between the homologs. Pot1pN and Pot1pC can be separated and retain biochemical activity individually, in contrast to hOB1 and hOB2 which appear to function only as a tightly packed unit.<sup>65,66,73-75</sup> This is likely in part due to the expanded

linker between Pot1pN and Pot1pC, composed of 25 proteolytically labile residues as opposed to only 5 disordered residues between hOB1 and hOB2.<sup>66,75</sup> Thus Pot1pN and Pot1pC appear to be flexibly tethered subdomains in contrast to the more tightly packed arrangement between hOB1 and hOB2.<sup>64–66</sup> Curiously, *SpPot1* also binds DNA with an affinity three orders of magnitude stronger than hPOT1 (low pM vs. low nM).<sup>65,73,74,76</sup> An outstanding question is how these differences are related to their respective roles at telomeres.

Pot1pN has significant sequence identity to its human counterpart, hOB1, and, as expected, the protein structures are quite similar.<sup>64,65</sup> This similarity is also evident in the specificity profiles of both domains in which binding is strongly disrupted when the individual nucleotides at positions 2-5 of either ligand are substituted with the complementary base.<sup>74</sup> Notably, four of these nucleotides overlay well in both structures and occupy nearly identical



**Figure 1.4 The structural similarities and differences of hPOT1 and spPot1** **A)** Crystal structures of hPOT1 OB1 (1XJV)<sup>66</sup> overlaid with Pot1pN (1QZH)<sup>65</sup> OB1 is shown in magenta and Pot1pN is shown in blue. The 6mer ligand bound by OB1 (OB1-DNA) is shown in yellow and the 6mer ligand bound by Pot1pN (Pot1pN-6mer) is shown in cyan. **B)** Crystal structures of hPOT1 OB2 (1XJV)<sup>66</sup> overlaid with Pot1pC (4HIK).<sup>67</sup> OB2 is shown in light brown and Pot1pC is shown in green. The 4mer ligand bound by OB2 (OB2-DNA) is shown in orange and the 9mer ligand bound by Pot1pC (Pot1pC-9mer) is shown in purple.



**Figure 1.5** The hydrogen bond networks for spPot1 shown for **A)** Pot1pN (1QZH)<sup>65</sup> nucleotides 1-4, **B)** Pot1pN nucleotides 5-6, **C)** Pot1pC (4HIK)<sup>67</sup> nucleotides 1-3, and **D)** Pot1pC nucleotides 7-9. Pot1pN is in blue and Pot1pN-6mer is in cyan. Pot1pC is in green and Pot1pC-9mer is in purple. Water molecules are shown in yellow and hydrogen bonds are indicated by the dashed red lines.

binding pockets (**Figure 1.4A**). The DNA-binding surfaces of both proteins participate in extensive hydrogen bonding with the Watson-Crick face of the DNA ligand and form several protein DNA stacking interactions (two in Pot1pN and three in hOB1). Additional specificity in Pot1pN appears to be achieved through intramolecular hydrogen bonding and stacking interactions within the DNA ligand itself between the bases of the nucleotides 1-4 (**Figure 1.5A**).

Pot1pC and hOB2 lack sequence identity and exhibit differing biochemical behavior, confounding direct extrapolation between them.<sup>65,66,74,75,77</sup> However, a structural comparison of the two domains reveals the mechanistic basis for their divergent behaviors. Despite their

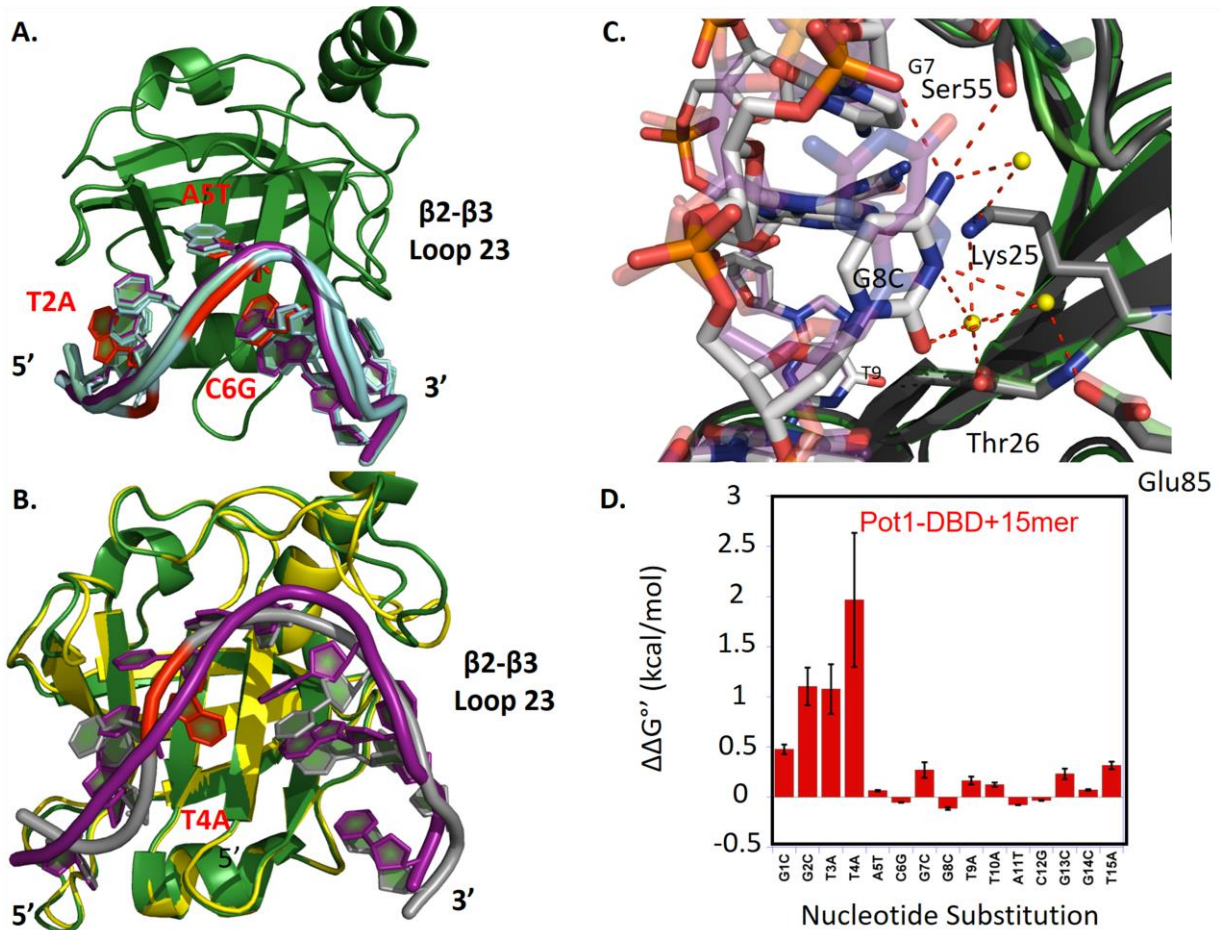
sequence divergence, the overall structures of Pot1pC and hOB2 are strikingly similar and they are easily identified as structural homologues by computational algorithms.<sup>78</sup> However, the clear structural differences between them have had a profound impact on their respective recognition of ssDNA. While hOB2 interacts with only four nucleotides in the structure of hPOT1-DBD bound to DNA, Pot1pC alone binds a minimal 9-nt ligand roughly across the canonical ligand-binding interface of the OB fold)(**Figure 1.4B**).<sup>66</sup> Interestingly, the 9-nt ligand is bent  $\sim 90^\circ$  as it traverses the surface. When compared to the path of human ssDNA along hOB2,<sup>65</sup> it becomes apparent that a substantially different region of the OB-fold barrel is used for ligand binding, which results in a stunning lack of ligand overlap between the two structures. Indeed, the binding pocket for only one nucleotide overlaps between these 2 domains (**Figure 1.4B**). Surprisingly, these dramatic differences in ssDNA-binding activity stem from the reorientation of a single loop connecting strands 2 and 3, which allows the proteins to take advantage of completely different binding surfaces. The path of ssDNA in Pot1pC suggests also that the OB-OB domain interface observed in the human structure is not achievable in *S. pombe*; arrangement of the *S. pombe* N and C OB folds in the human packing geometry leaves a 23 Å gap in the path of ssDNA.<sup>79</sup>

### **A remarkably plastic interface confers non-specificity**

One of the most surprising features of the Pot1pC/9mer structure is the large number (~22) of apparently sequence-specific H-bonds between both the Watson-Crick and Hoogsteen faces and the surface of the protein. This recognition interface is composed of several stacking interactions and a set of base-mediated H-bonds that largely resemble those that confer specificity in the N-terminal domain (**Figure 1.5A, B**).<sup>64</sup> Canonically, sequence-specific recognition is thought to occur through the readout of a pattern of H-bond donor and acceptor atoms characteristic of a nucleotide sequence. Conversely, non-specific nucleic acid recognition is believed to be achieved by stacking/hydrophobic interactions and/or interactions with the

phosphate backbone.<sup>80-83</sup> Thus, the specificity at position 2, for example, would typically be ascribed to the presence of three direct H-bonds between the base and the protein. In the Pot1pC/ssDNA interface, though, the presence of those H-bonding interactions does not predict specificity, for example, examination of the interactions at position 1 reveals 4 direct H-bonds that confer no specificity.<sup>75</sup> Base-mediated H-bonds such as the ones observed here are frequently assigned roles in conferring specificity, and nothing about the chemical nature of the interface suggests a biochemical difference of specificity relative to Pot1pN. Thus, the cognate structure alone cannot be used to predict the biochemical specificity of the interface.

Fortuitously, structures of complexes containing non-telomeric (non-cognate) sequences provided insight to understanding how this seemingly specific interface accommodates other sequences.<sup>75</sup> Despite having similar affinities, study of this series of complexes revealed unanticipated structural changes at the protein/nucleic acid interface. These range from a local reorientation of a base to wholesale reorganization of the interface. For example, only local reorientation at the site of substitution is observed following base alterations at positions 3, 5 and 6. These modest rearrangements do not substantially impact the positioning of the base but neatly compensate for lost H-bonds by forming new ones (**Figure 1.6A**). The overall protein backbone is largely unaffected as well, with only minor changes in protein structure distal to the interface. Some base substitutions lead to more pronounced local changes in both DNA and protein conformation. For example, substitution from guanine to cytosine at position 8 results in a 180° rotation to the base coupled with a rearrangement of the  $\beta$ 2-  $\beta$ 3 loop (L23). As expected, the substitution of cytosine disrupts a suite of H-bonds, but the base's rearrangement creates an equally intricate network of H-bonds with almost completely new intra- and intermolecular partners (**Figure 1.6B**). Although these adjustments are all proximal to the site of base substitution, these structures suggest conformational plasticity within both the protein surface and the DNA that allows the interface to adapt structurally and thermodynamically to the base changes. While some non-cognate ligands can be accommodated by these local (although



**Figure 1.6 Plastic accommodation of the DNA ligand for spPot1pC.**<sup>67</sup> **A)** Non-cognate DNA ligands, T2A (4HIM), A5T (4HJ5), and C6G (4HJ7) in cyan with substituted bases highlighted in red overlay with the cognate Pot1pC-9mer in purple and the cognate Pot1pC protein structure in green (4HIK). Non-cognate protein structures are omitted for clarity. **B)** Structure of Pot1pC bound to T4A (4HIO) non-cognate ligand (non-cognate-Pot1pC protein yellow, T4A gray with A4 highlighted in red) overlay with cognate bound Pot1pC (cognate-Pot1pC protein green, Pot1pC-9mer DNA purple). **C)** Compensatory hydrogen bond network for non-cognate G8C structure (4HJ8) shown. G8C bound Pot1pC in dark gray, G8C ligand in white. Cognate 9mer bound Pot1pC in green and cognate Pot1pC-9mer in purple. **D)** Nucleotide specificity profile for spPot1-DBD binding domain in which single nucleotide positions of the cognate 15mer sequence (GGTTACGGTTACGGT) are individually substituted with the complementary base.<sup>75</sup>

significant) adjustments to the interface, others lead to even larger changes, with a global reorganization of the complex. Substitution of the base at positions 2 and 4 leads to a second binding mode. For example, substitution at position 4 triggers a repositioning of the base, presumably because the large A cannot fit in the pocket previously occupied by a T (**Figure**

**1.6C).** As a result, the base rotates  $\sim 90^\circ$  around the phosphodiester backbone, flipping it out of the original binding pocket into the largely unoccupied space below. This reorientation is stabilized by a stacking interaction with Arg68 and leads to a “chain reaction” of molecular events both 5’ and 3’ of the site of substitution, causing a complete reorganization of the interface marked by an overall 3.05 Å ligand RMSD compared to the cognate ligand structure. For comparison, excluding the flexible L23, hOB2 and Pot1pC have a 1.9 Å RMSD for 125  $\alpha$ -carbons (out of 139). All in all, Pot1pC has at its disposal several structural elements to accommodate sequence heterogeneity, including ligand and protein flexibility (particularly in loop regions), an enlarged binding cleft, and a complex network of H-bonding interactions. These structures in total revealed a sophisticated mechanism of conformational malleability by which Pot1pC is able to accommodate heterogeneous ssDNA ligands with little to no change in the overall thermodynamics of binding. This plasticity is likely shared by other proteins that are either fully non-specific such as RPA<sup>84</sup> or require graded specificity, such as t-RPA (see below). It is an open question as to what biophysical features of the protein and ligand, for example, types of amino acids at the interface or dynamic properties, facilitate this type of malleable recognition. Moreover, this raises the questions of what makes an interface biochemically specific for an inherently flexible ligand and which type of interface is in fact harder to evolve.

#### **1.1.4 – How the Pot1 subdomains work together**

Our structural understanding of the *S. pombe* Pot1-DBD is derived from studies of the individual subdomains, primarily because the full DBD proved intractable to high-resolution structural characterization. This caveat raises the question of how many of its characteristics can be explained through the action of the two subdomains in isolation. A reasonable first measure is to compare the biochemical features of the individual domains to those of the intact DBD (and full-length protein). The full Pot1-DBD shows much of the same specificity trends as

the individual domains but has reduced specificity at A5 and C6 for Pot1pN and G2 for Pot1pC (**Figure 1.6D**).<sup>74,77</sup> However, the absolute specificity for the Pot1pC sequence is dramatically reduced such that complete substitution of the Pot1pC 9mer sequence results in less than a 2-fold change in binding.<sup>74</sup> The full-DBD can also bind a 12mer ligand comprised of two 6mer repeats whereas Pot1pC exhibits no observed binding to a 6mer sequence.<sup>77</sup>

While the structure of the homologous hPOT1 has been solved, the disparate DNA-binding surfaces of Pot1pC and hOB2 make homology modeling unreliable. As noted above, simply docking the *S. pombe* DNA-bound structures in the relative hOB1/hOB2 orientation seen in the crystal structure creates a physically impossible path for the ssDNA to adopt. While it is possible that the DNA completely rearranges in the full DBD relative to the conformation adopted in the individual domains, the similarity of the biochemical features between the two suggests the DBD is more like the individual domain structures than not. The more likely scenario is that the long flexible linker that connects Pot1pN and Pot1pC allows for a domain/domain reorientation that differs considerably from that observed in the human homologue.

Solution NMR strategies provide a complementary tool to x-ray crystallography to probe the overall conformation of the *S. pombe* Pot1-DBD complex. Comparison of the full assigned spectra of the Pot1pN+6mer and the Pot1pC+9mer complexes to that of the Pot1-DBD+15mer allows for high-resolution mapping of regions of difference.<sup>75</sup> Overall, the notion that the whole equals the sum of the parts holds true. The vast majority of assigned residues coincide precisely in chemical shift between the Pot1-DBD and its constituent subdomains, suggesting a large degree of structural similarity. Mapping of the few residues that are shifted pinpoints a potential Pot1pN/Pot1pC interface that is indeed rotated significantly away from the orientation in the human structure.)<sup>75</sup> Interestingly, deletion of the majority of the linker did not lead to any change in affinity for telomeric substrate, indicating that this altered conformation can be accommodated with a relatively short (only 4 amino acid) linker sequence.<sup>75</sup> Furthermore, perturbation of the



putative contact residues within this interface also leads to minimal (less than 2-fold) changes in ssDNA binding affinity.<sup>79</sup> Together, these data support a model where the Pot1pN and Pot1pC subdomains are relatively structurally independent, lacking in a precise, stable protein/protein interface and acting merely as weakly associated partners in binding.

Why this evolutionary divergence and what does it suggest regarding telomere maintenance in general? It is quite common to identify proteins that have relatively similar structures in the absence of identifiable sequence relationship, as structure is generally more conserved than sequence. However, we are unaware of any examples of structurally homologous domains that bind the same ligand via a completely novel interface. The marked differences between hPOT1 and *SpPot1* DBDs may have evolved to accommodate the unusual and specific needs of the telomeres in each species: hPOT1 only needs to recognize a relatively invariant repeat while *SpPot1* must accommodate degenerate sequences and likely does in part via domain-domain rearrangement. Other potential reasons include differences in shelterin, need for t-loop assembly, differences in the length of the overhang, and/or degenerate solutions happened upon by evolution. Conversely, it may be that these two homologues represent the range of conformations needed to be accessed at different points in the process of telomere maintenance.

### **1.1.5 - How Pot1 Might Regulate Telomerase**

*In vitro*, Pot1 inhibits telomerase activity by sequestering the 3' ssDNA overhang that telomerase requires as a substrate, presumably through a simple competition event, suggesting that its intrinsic nature is to restrict access of telomerase to the overhang.<sup>16,85</sup> This is consistent with the observation that deleting a DNA-binding OB fold in hPOT1 leads to significantly longer and more heterogeneous telomeres.<sup>58</sup> *In vitro* addition of hPOT1's direct binding partner within the shelterin complex, TPP1, however, ameliorates this inhibitory effect and significantly increases the repeat addition processivity of telomerase,<sup>17,63,86</sup> by slowing primer dissociation

and aiding translocation, perhaps by increasing the dynamic sliding of Pot1 on DNA.<sup>86,87</sup> TPP1 (or Tpz1 in *S. pombe*) has no significant DNA-binding ability of its own but modestly alters the *in vitro* DNA-binding properties of Pot1.<sup>17</sup> In addition to tethering hPOT1 to the shelterin complex, hTPP1 also recruits telomerase to telomeres *in vivo* through, incidentally, yet another OB fold.<sup>88-</sup>

91

It remains unclear if there is active regulation of hPOT1 binding to ssDNA, or if telomerase simply competes with hPOT1 for access to the 3' end. The bias in end sequence provides some insight- 40% of 3' overhangs in telomerase active human cells terminate in the sequence 5'-GGTTAG-3'.<sup>25</sup> Based on the crystal structure of hPOT1 bound to ssDNA, this sequence should be bound and fully sequestered from telomerase.<sup>65</sup> While different terminal sequences are extendable to some extent, full human telomerase activity requires an unprotected overhang of at least eight nucleotides.<sup>85</sup> Aside from the structural considerations, there are kinetic features to consider as well. As is typical of tight binding interactions, hPOT1/TPP1 dissociates slowly from ssDNA, with a half-life of nearly 30 minutes *in vitro*, pointing to the need for active regulation of POT1 binding.<sup>17</sup> Less is known about *S. pombe* proteins, but a similar mechanism of telomerase recruitment is proposed via a Pot1-Tpz1-Ccq1 complex and *S. pombe* Pot1 has a ssDNA-bound half-life of approximately one hour.<sup>15,18</sup> These common features point to a shared requirement for active regulation to allow telomerase access.<sup>74</sup>

Several lines of data on the ssDNA-binding preferences of *SpPot1* suggest that Pot1 can bind ssDNA in alternative modes, predominantly through malleability in the recognition of ssDNA by the less-specific Pot1pC domain. The first observation is that, in addition to the 15mer binding mode described above (that is closely related to the “sum of the parts” idea), *SpPot1* binds a simple 12mer sequence that comprises 2 repeats of the core telomere sequence – GGTTAC. This clearly must adopt a different conformation than the 6+9 mode described above. At high concentrations of Pot1-DBD, the protein binds the 12mer ligand as a

dimer, suggesting that the Pot1pN of each monomer binds its core specific sequence (GGTTAC, as described above).<sup>92</sup> This suggests that the avidity of Pot1pC for the remaining 3' 6mer is too modest to out compete a second binding event at high concentrations. Indeed, Pot1pC binding of a 6mer in isolation is in the mM range.<sup>79</sup> At lower, more physiologically relevant, concentrations of Pot1-DBD, the dimer is not observed, and gel shift suggests a distinct, as yet structurally uncharacterized, conformation. Preliminary NMR data suggest that the mode of interaction with the Pot1pC part of the DBD is entirely disrupted relative to that present in the 15mer complex.<sup>79</sup> While the precise structural details are elusive, this new conformation has distinct biochemical features relative to the 15mer binding mode, most prominently a 3' end that is more accessible to other end-binding factors.

The ability to observe this second binding mode by gel shift allowed the screening of protein mutants able to induce a similar conformation.<sup>75</sup> In an effort to rationally induce such a conformational change, a panel of mutations was engineered near the binding site of the 3' end of the DNA. In Pot1pC, the 3' end of the oligo forms an interleaved aromatic stack, similar to four teeth of a zipper, with W223 and Y224 (**Figure 1.5B**). Mutation of Y224 in the context of Pot1pC has a drastic effect on binding affinity, however, mutation of Y224 in the context of Pot1-DBD has no effect on affinity. This curious disconnect can be explained through the observation that Pot1-DBD containing this mutation adopts the alternate 12mer binding mode, suggested by the characteristic gel shift. This mode is also induced when alterations in DNA sequence are made at the 3' end at positions 13 or 15, the bases that stack with Y224, or at high salt conditions that disrupt this more electrostatically driven binding mode.

Despite these distinct biochemical and structural features, the 12mer and 15mer binding modes have similar affinities at physiological salt concentrations.<sup>74</sup> This argues that both 1:1 binding modes have to be considered when evaluating biological function. Access to the 3' overhang at the telomere is an essential step in regulating telomerase activity. *In vitro* telomerase extension is inhibited by the presence of Pot1 and can be restored when the Pot1

binding site is moved away from the 3' end.<sup>85</sup> The potential ability of another protein to engage the 3' end in the 12mer, but not 15mer, binding mode suggests that this binding mode does not completely sequester the 3' end and may represent an extendible telomeric state. Does this happen in hPOT1? As noted above, the specificity for the 5' end of the oligonucleotide substrate is shared, and the localization of Pot1 to the telomere via its interaction with TPP1 means it has the flexibility to perhaps shift modes to ones with weaker affinity. The role of this plasticity is an exciting frontier in telomerase regulation.

## **1.2 – Overview of Pot1 and RNA at the telomeres**

### **1.2.1 - Challenges Facing Telomere End-Protection Proteins**

The telomere end-protection proteins, such as Pot1, must overcome several challenges to accomplish their vital functions. First, these proteins must bind telomeres tenaciously to prevent degradation by nucleases as well as occlude telomeric structures from recognition by damage response pathways. Moreover, because these proteins can displace proteins that sense DNA damage, they must have limited binding activity to non-telomeric sequence so as to not interfere with proper recognition of *bona fide* DNA damage and subsequent repair.<sup>93</sup> Furthermore, non-specific binding to other regions of the genome would overshadow the limited binding sites present at telomeres<sup>94</sup> and leave telomeres inadequately protected. These activities are achieved through the combination of unique biochemical properties and association with the shelterin complex. Seemingly counter to this need for specificity, however, sequence variation at telomeres necessitates that telomere binding proteins somehow also accommodate some level of non-specificity.<sup>43,95</sup> TPP1 in humans and Tpz1 in *S. pombe* in part aid to resolve these challenges by bridging Pot1 to the dsDNA binding components of the shelterin complex to increase the avidity of Pot1 to telomeres. Additionally, the evolution of Pot1's recognition features addresses these functional challenges in a remarkable manner as described above by the plasticity evolved at the Pot1pC binding interface. However, beyond the

specificity requirements set up by Pot1 binding to ssDNA is the consideration of the specificity towards ssRNA.

## 1.2.2 – Transcribed Telomeres and Cellular RNAs Represent a Pool of Potential Pot1

### Substrates

Direct transcription of the telomeres, starting from the subtelomeric region and transcribed towards the telomere ends, produces a population of *telomere repeat* containing **RNA** (TERRA) which plays a role in the regulation of telomere length.<sup>96,97</sup> TERRA, containing a complementary sequence to the C-rich strand, can form RNA-DNA hybrid structures known as R-loops.<sup>96,98,99</sup> TERRA is also able to associate with and recruit telomerase to telomeres.<sup>98,100</sup> During the G1/S cell cycle transition, TERRA transcription is upregulated. Then while telomeres are replicated by DNA polymerase and extended by telomerase during S-phase and G2, TERRA is degraded by the exonuclease Rat1 and by RNaseH2.<sup>101</sup> This degradation of TERRA is impaired at critically short telomeres at which both Rat1 and RNaseH2 are inefficiently recruited.<sup>101</sup> Moreover, shortened telomeres also lose the transcriptional silencing marks typical of healthy telomeres, further enriching TERRA at short telomeres. The association between TERRA, the telomeres, and telomerase has subsequently developed into a model where TERRA indicates the presence of critical short telomeres in need to elongation. Consistent with this model, overexpression of TERRA in *S. pombe* results in telomere elongation,<sup>102</sup> though for unclear reasons, the same experiment in *S. cerevisiae* leads to telomere shortening.<sup>103</sup> If TERRA fails to recruit telomerase, the R-loop structures formed between TERRA and telomeric DNA is able to recruit homologous recombination machinery – likely resulting from interaction with the replication machinery.<sup>101</sup> In addition, recent data indicates TERRA plays a role in recruiting the PCR2 transcriptional silencing machinery to telomeres.<sup>104</sup> Together through these functions, TERRA serves as a signaling molecule for shortened telomeres and is able to lead to the

recruitment the various cellular machineries capable of restoring telomere hemostasis through elongation and silencing.<sup>98-100</sup>

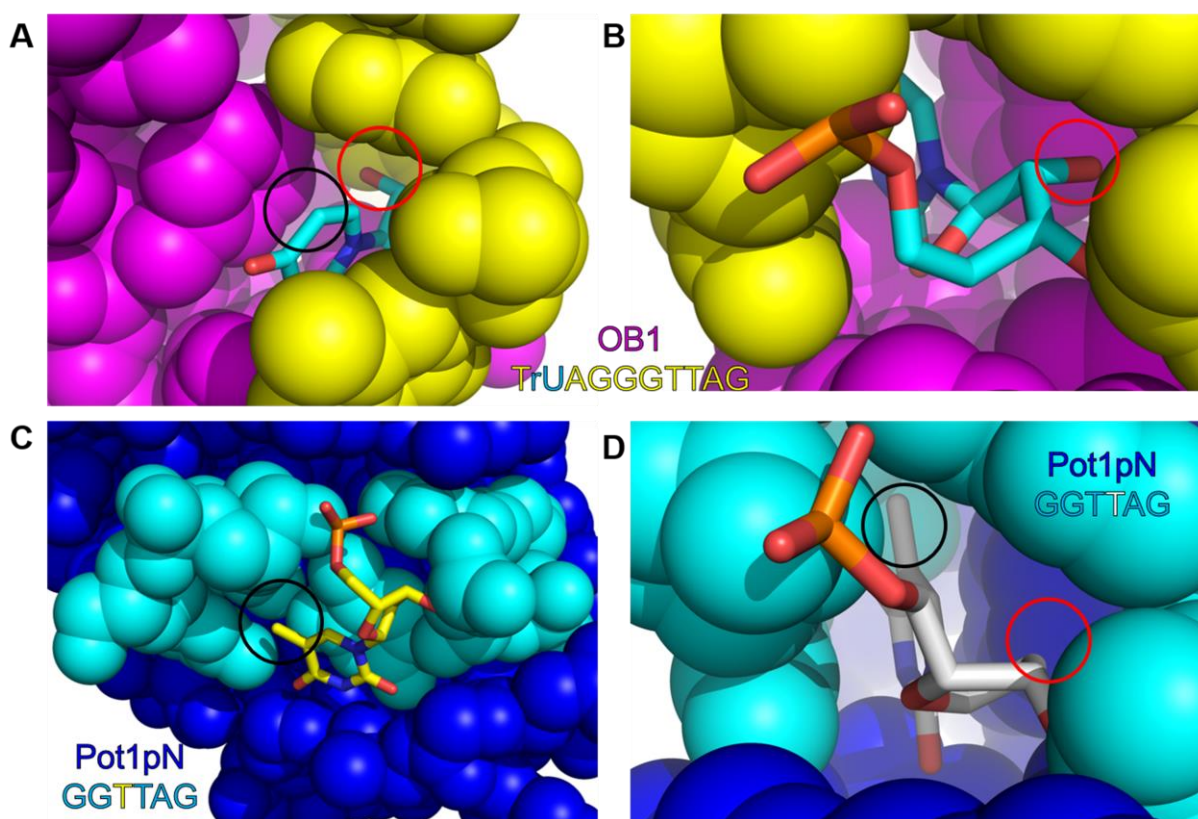
However, this cellular pool of TERRA and other cellular RNAs present a significant specificity challenge for Pot1. Given that its ability to bind ssDNA is needed for proper telomere maintenance,<sup>56</sup> the question arises of how Pot1 discriminates against the vast pool of cellular RNAs. If Pot1 also bound ssRNA, then, at cellular concentrations of each, virtually all Pot1 would be predicted to be sequestered into non-productive Pot1/RNA complexes simply as a result of RNA containing a Pot1 binding sequence by random chance.<sup>64</sup> But the problem is even further exacerbated by TERRA, which on average contains ~30 potential Pot1 binding sites rendered in RNA<sup>97</sup> and is localized in close proximity to Pot1 ssDNA binding sites. Due to the extremely tight binding of Pot1 to single-stranded telomeric sequences and its inhibition of DNA repair pathways, overexpression of Pot1 is likely to cause serious issues through spurious binding throughout the genome. As a result, the necessarily low expression of Pot1 in *S. pombe* presents a stoichiometric problem for Pot1 as it must faithfully and tightly bind the 6 *S. pombe* telomeres without being sequestered by RNA of the same sequence. Thus, strong discrimination against RNA ligands by *S. pombe* Pot1 (at least  $\sim 10^5$  based on the concentration of spurious binding sites in all RNA and the low expression of Pot1) is necessary to prevent Pot1 sequestration by RNA. Failure to discriminate against RNA would likely recapitulate deletion phenotypes and their catastrophic effects on genome stability. Accomplishing the necessary discrimination to prevent this is a challenging prospect as ssDNA and ssRNA are chemically similar, adopt similar conformational structures, and are both highly flexible. Consistent with the biochemical prediction that Pot1 must strongly disfavor RNA binding, experimental results demonstrate that both mammalian and *S. pombe* Pot1 discriminate against RNA of the same cognate sequence.

### 1.2.3 - RNA Discrimination by mPOT1

In the case of mammalian Pot1, the underlying mechanism of discrimination occurs in the region of the protein homologous to *S. pombe* Pot1pN, as the C-terminal domain of the human protein, analogous to Pot1pC, barely engages with the ligand.<sup>65,94</sup> Using a triplet substitution binding strategy, Nandakumar et al. characterized the position ribose specificity of mammalian Pot1.<sup>94</sup> Discrimination in this case appears to result primarily from losing a hydrophobic interaction from the T4 methyl as well as forcing the 2' hydroxyl at that position into a sterically unfavorable interaction in a hydrophobic pocket<sup>94</sup> (structure shown in **Figure 1.7A**). This single nucleotide substitution showed only modestly 4-fold impaired binding for human Pot1 and a 22-fold affinity decrease for mouse Pot1 when in complex with Pot1 alone. However, addition of a subdomain of the shelterin bridging protein TPP1 strongly enhanced the discrimination exhibited by Pot1 by preferentially increasing the affinity of Pot1 for the cognate ssDNA ligand – likely by stabilizing the ssDNA bound conformation. However, while the Pot1-TPP1 complex bound most rNMP substituted ligands tighter than Pot1 alone, the fold-enhancement of binding was much lower than that of the cognate ligand, resulting in an effective  $2-3 \times 10^4$ -fold discrimination of the full RNA ligand and 120 to 470 -fold discrimination at position 4.<sup>94</sup> Interestingly, other positions throughout the ligand also contributed to ssDNA specificity, especially when placed near the discriminating position – suggesting nearby ribose nucleotides reinforce the suboptimal binding geometries at discriminating positions by further limiting the conformational flexibility of the ligand.<sup>94</sup> Together with preferential stabilization of the ssDNA-complex by TPP1, these data suggest subtle additive effects throughout the ligand-Pot1-TPP1 complex contribute to preferential binding to ssDNA in addition to the more apparent mechanisms at position 4.

### 1.2.4 – RNA Discrimination by full-length *S. pombe* and Pot1pN

The full-length DNA-binding domain of *S. pombe* Pot1 discriminates against RNA of the same cognate sequence by at least a factor  $10^6$ .<sup>105</sup> This discrimination is conferred in part by the specificity determining first OB-fold, Pot1pN, which alone disfavors RNA by a factor of >200 using interactions at two positions<sup>64</sup>. While an RNA bound or chimeric complex for Pot1pN has not been solved, predictions for the mechanisms of specificity can be made based on the



**Figure 1.7 RNA Discrimination by mPot1 and Pot1pN.** **A/B)** Structure of hPOT1 (magenta) bound with dTrUd(AGGGTTAG) (yellow) shows discrimination of rU (cyan) at position 4. **(A)** The pocket left by the lost methyl group is highlighted by a black circle and the hydroxyl group highlighted with a red circle. **(B)** A rotated for comparison to **(D)** as both Pot1pN and hPOT1 T4 occupy the same binding pocket **C/D)** Pot1pN (blue) bound to GGTTAG (cyan) shows discrimination at **(C)** T3 (yellow) likely due to the empty space after the loss of the methyl group highlighted with a black circle while at T4 (white) **(D)** both the methyl group, highlighted with black circle, and a 2' hydroxyl, highlighted by a red circle, cause discrimination due to empty space and steric clashes, respectively, in a hydrophobic pocket.



available structures and existing biochemical data. If the RNA adopted the same conformation as the DNA bound complex, there would be readily apparent steric clashes for hydroxyls at two positions and energetically unfavorable hydrophobic pockets left empty by the loss of thymine methyl groups<sup>64</sup>. At both positions (GGTTAC, cognate), a greater than 200-fold discrimination is observed at T3 attributable entirely to the loss of the methyl group (**Figure 1.7B**) while T4 exhibits ~100-fold affinity reduction due to the 2' hydroxyl and ~7-fold reduction due to the loss of the methyl group (**Figure 1.7C**).<sup>64</sup>

### **1.2.5 – RNA Discrimination by Pot1pC and Novel Insight into Protein-Nucleic Acid specificity**

However, to explain the discrimination observed by the full-length protein, Pot1pC must also contribute to the discrimination against RNA. This raises exciting questions regarding the mechanism of how Pot1pC is able to achieve this specificity for ssDNA. In classical systems of nucleic acid specificity, sequence-specificity is widely believed to be achieved through hydrogen bond donor and acceptor patterns that provide shape complementarity not readily satisfied by other species.<sup>80,81,83</sup> In contrast, sequence indiscriminate recognition of nucleic acids, important for the function of proteins such as replication protein A (RPA),<sup>106,107</sup> single-strand break protein (SSB)(Lohman and Ferrari, 1994), DNA polymerase,<sup>108</sup> and others, are thought to be largely driven by nonspecific stacking/hydrophobic interactions and/or electrostatic interactions with the sugar-phosphate backbone.<sup>82,109</sup> However, the structural accommodation of non-cognate sequences by Pot1pC reveals hydrogen bond contacts primarily to the bases in a manner commonly associated with the canonical sequence specific interactions, suggesting this nucleic acid specificity paradigm deserves broadening. Moreover, the additional capability of Pot1pC to also discriminate against ssRNA despite that difference being much more chemically and structurally subtle than base substitutions provides unique and exciting new insights into protein-nucleic acid specificity.

## **Chapter 2 – Pot1pC Discrimination of RNA Backbones**

### **2.0 – Chapter Overview:**

This chapter describes my work on characterizing the RNA specificity of the C-terminal domain of the DNA-binding domain of *S. pombe* Pot1 through a combination of biochemical and structural techniques. Note: Much of the text and figures of this chapter has been previously published in a paper I published as first author.<sup>2</sup>

### **2.1 – Introduction**

To resolve how Pot1pC discriminates against ssRNA despite the remarkable structural plasticity of its interface, we characterized the ribose-position specificity of Pot1pC by measuring binding affinities of RNA and chimeric RNA-DNA ligands containing ribose nucleotide substitutions in the cognate sequence and found that specificity for DNA over RNA is not evenly distributed across the ligand. We also solved 3 high resolution crystal structures of Pot1pC bound to these chimeric RNA-DNA ligands, revealing a widely utilized cryptic binding mode of Pot1pC characterized by the rearrangement of T4 into a new binding pocket and substantial rearrangement of the 3' portion of the interface. These rearrangements allow full thermodynamic accommodation of RNA nucleotide substitutions at positions near the 3' end of the ligand, facilitated by a long and flexible protein loop, but not fully at the 5' end due to suboptimal binding conformations for RNA ligands.

### **2.2 – Materials and Methods**

Detailed protocols for all the methods briefly described here are included in Appendix A

#### **2.2.1 – Protein Expression and Purification**

Pot1pC was expressed and purified using essentially the same method described in Dickey et al. (2013). Briefly, V199D Pot1pC was expressed as an intein-chitin-binding domain fusion in BL21 (DE3) *E. coli* at 18 °C for 20 hours. Following bacterial cell harvesting and lysis, the fusion construct was bound to chitin beads (New England Biolabs) and Pot1pC was cleaved from the intein-chitin binding domain by incubation with 100 mM beta mercaptoethanol ( $\beta$ ME) for 20-40 hrs. at 4 °C. Following elution, Pot1pC was concentrated and injected onto a Superdex 75 column (GE) in 100 mM Tris pH 8.0, 100 mM KCl, 0.1% (w/v) deoxycholate, 3 mM  $\beta$ ME, and 5% (v/v) glycerol. After elution from the size exclusion column, ~99% pure protein was concentrated to 450-600  $\mu$ M, snap frozen in liquid nitrogen, and stored at -70 °C.

### **2.2.2 – Isothermal Titration Calorimetry**

Pot1pC stored at -70 °C was thawed and dialyzed overnight at 4 °C in buffer containing 20 mM potassium phosphate pH 8.0, 150 mM NaCl, and 3 mM  $\beta$ ME. Oligonucleotides obtained from Integrated DNA Technologies were resuspended in the same dialysis buffer. Heats of dilution experiments showed no detectable heat evolved and thus were not subtracted from binding experiments. All experiments were performed in triplicate on a MicroCal ITC200 (GE Healthcare) at 25 °C. The sample cell was loaded with 230  $\mu$ L of 5-100  $\mu$ M Pot1pC into which buffer matched nucleic acid at approximately 10-fold higher concentration was titrated as follows: one 0.2  $\mu$ L dummy injection, followed by nineteen 2  $\mu$ L injections, and a final 1.3  $\mu$ L injection. Data were integrated and fit by nonlinear least-squares fitting to a single binding site model using Origin ITC Software (OriginLab, Northampton, MA).

Ultraviolet (280 and 260 nm) absorbance measurements were used to calculate protein and nucleic acid concentrations, using extinction coefficients provided by ExPASy ProtParam and Integrated DNA Technologies, respectively.

### 2.2.3 – Crystallization

Crystals were grown using the hanging drop vapor diffusion method at 4 °C. Drops contained 1 µL of mother liquor and 1 µL of a solution of 1:1 protein:ssRNA/DNA (5-15 mg/mL). Crystallization conditions for each complex are listed in Table 2. The 1-3R and 7-9R crystals were obtained by two-step seeding with the cognate ssDNA complex crystals providing the initial seeds and then the resulting low-quality 1-3R and 7-9R crystals as the seeds for a second round of seeding. Seeds were generated by vortexing seed crystals in mother liquor (Seed Bead crystal kit, Hampton Research) and resulting microcrystals were transferred to the hanging drop by dipping a cat whisker into the seed solution and swiping it through the drop. Crystals were cryoprotected by sequentially transferring the crystal in mother liquor solutions supplemented with 5%, 10%, 15%, and 20% (v/v) ethylene glycol and flash frozen in liquid nitrogen.

### 2.2.4 – Data Collection and Refinement

X-ray diffraction data for 1R was collected at the Advanced Light Source (ALS) Beamline 8.2.1 and the data sets for 1-3R and 7-9R were collected at the ALS Beamline 8.2.2. Reflections were indexed using iMOSFLM<sup>110</sup> and scaled using Scala within the CCP4 program suite<sup>111</sup>. The phases were solved through molecular replacement using the coordinates of cognate Pot1pC without ssDNA (4HIK)<sup>66</sup> as a starting model in PHENIX<sup>112,113</sup> followed by rigid body refinement using PHENIX Refine<sup>114,115</sup>. The non-cognate RNA ligands were built into the electron density manually in Coot<sup>116</sup> and subsequent refinement was performed in the PHENIX program suite with manual adjustment in Coot. The final models were validated using PHENIX.validate and MolProbity<sup>117</sup> to assess quality (statistics for final models can be found in Table 2.1).

	1R 9mer rGGTTACGGT	1-3R 9mer rGrGrUTACGGT	7-9R 9mer GGTTACrGrGrU
RCSB PDB ID	5USB	5USN	5USO
<b>Data Collection</b>			
Space group	P 21 21 21	P 21 21 21	P 21 21 21
<b>Cell dimensions</b>			
a, b, c	41.29, 58.01, 65.91	41.64, 59.8, 66.08	44.56, 57.61, 66.76
$\alpha$ , $\beta$ , $\gamma$	90 90 90	90 90 90	90 90 90
Resolution (Å)	33.64 - 1.615 (1.673 - 1.615)	44.339 - 1.9 (1.968 - 1.9)	43.62 - 2.0 (2.072 - 2.0)
Rmerge	0.064	0.131	0.163
$I/\sigma$	119.88 (6.00)	101.80 (10.01)	131.41 (6.77)
Completeness (%)	96.58 (90.87)	98.97 (96.98)	94.39 (92.36)
Redundancy	7.4 (5.5)	12.7 (12.9)	22.1 (17.8)
<b>Refinement</b>			
Resolution	33.64 - 1.615 (1.673 - 1.615)	44.339 - 1.9 (1.968 - 1.9)	43.62 - 2.0 (2.072 - 2.0)
No. Reflections	20220 (1852)	13388 (1284)	12121 (1177)
$R_{work}/R_{free}$	0.1759 / 0.2027	0.1829 / 0.2231	0.2102 / 0.2457
No. atoms	1697	1557	1495
Protein	1300	1194	1188
Ligand/ion	186	187/1	187
Water	211	175	120
<b>B-Factors</b>			
Protein	21.61	21.31	31.86
Ligand/ion	29.09	31.69/35.12	40.58
Water	34.28	30.17	36.5
<b>Rmsds</b>			
Bond Lengths (Å)	0.016	0.005	0.009
Bond Angles (°)	1.49	0.66	0.99
Crystallization Conditions	50 mM Tris, 0.2 mM sodium formate, 20% PEG 8K	100 mM Tris pH 8.4, 0.2 mM sodium formate, 15% PEG 4K	100 mM Tris pH 7.5, 0.2 mM sodium formate, 15% PEG 4K

**Table 2.1 Data Collection and Refinement Statistics for Ribose Chimeric Pot1pC Complexes** RNA nucleotides in red. Each structure determined by one crystal. Highest resolution shell in parentheses.

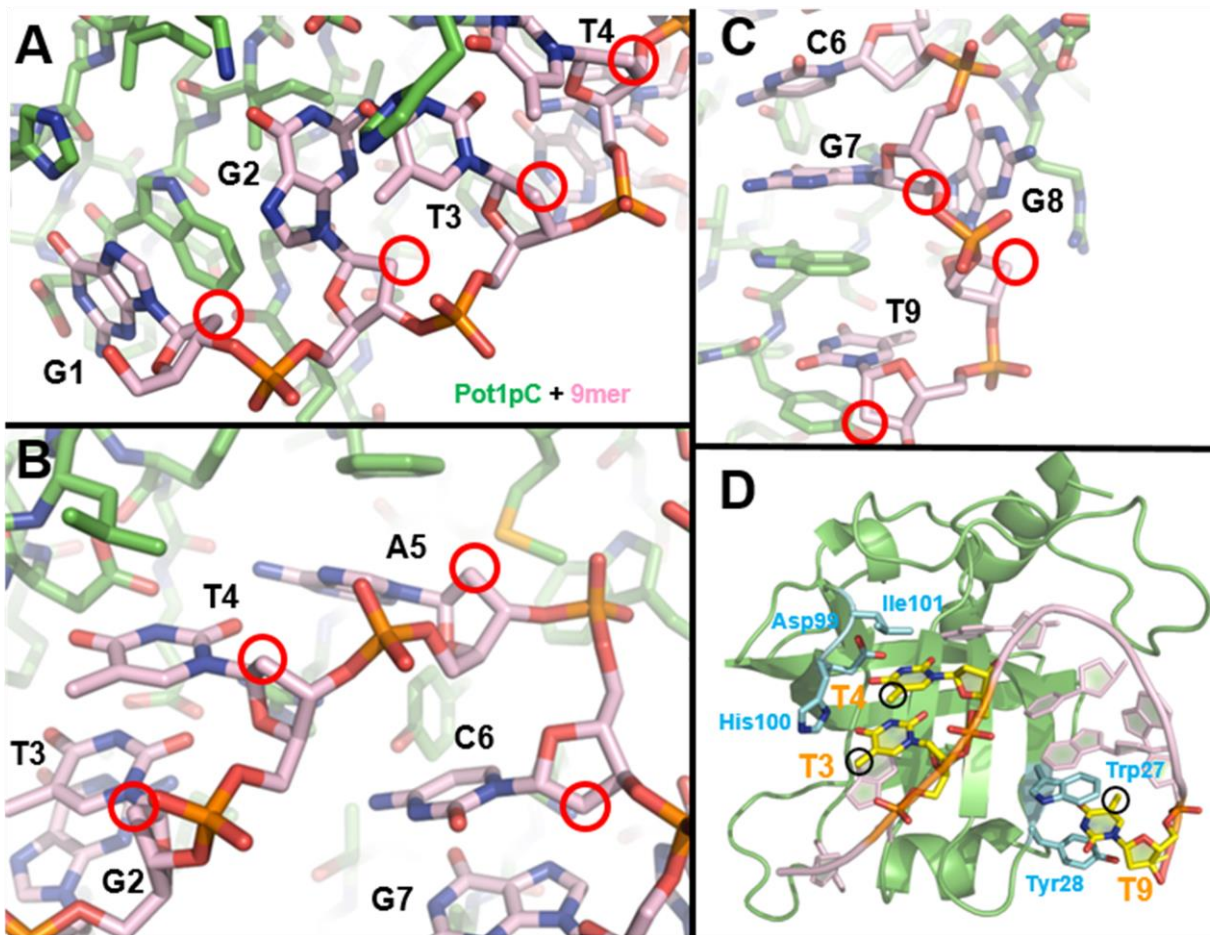
## 2.3 – Results

### 2.3.1 – Pot1pC discriminates against RNA additively by ribose position

Pot1pC minimally recognizes a 9 nucleotide-long sequence (9mer) of ssDNA of the sequence GGTACGGT with an apparent binding dissociation constant ( $K_D$ ) of 24 nM<sup>66</sup>. Full substitution of this cognate ssDNA sequence with ribose nucleotides (1-9R) results in a substantial loss of affinity by ~80-fold as measured by isothermal titration calorimetry (ITC, **Table 2.2**, with raw data and fitted curves shown in Appendix A), demonstrating that both the Pot1pN and Pot1pC subdomains of Pot1-DBD contribute to RNA discrimination. One possibility is that this discrimination is achieved through recognition of the specific chemical differences

dNTP/rNTP <sup>a</sup>	$K_D$ (nM) <sup>b</sup>	Fold Change <sup>c</sup>	$\Delta H$ (kcal/mol) <sup>b</sup>	$T\Delta S$ (kcal/mol) <sup>b</sup>
Cognate <sup>d</sup> GGTACGGT	24 <sup>d</sup>	-	-29 <sup>d</sup>	-18 <sup>d</sup>
(349)dU GGUUACGGU	60 ± 16	2.5	-30 ± 1.7	-20 ± 2
1-3R GGU <sup>1</sup> TACGGT	228 ± 24	9.5	-34 ± 1	-24 ± 1
1R GGU <sup>1</sup> TACGGT	92 ± 2	3.8	-32 ± 0.3	-22 ± 0.3
2R GGT <sup>2</sup> TACGGT	56 ± 10	2.3	-30 ± 1	-20 ± 1
3R GGU <sup>3</sup> TACGGT	48 ± 6	2.0	-28 ± 0.7	-18 ± 0.8
4-6R GGT <sup>4</sup> UACGGT	154 ± 2	6.4	-29 ± 0.7	-20 ± 0.7
4R GGT <sup>4</sup> UACGGT	57 ± 9	2.4	-29 ± 2	-19 ± 2
5R GGT <sup>5</sup> TACGGT	53 ± 14	2.2	-30 ± 1	-20 ± 1
6R GGT <sup>6</sup> TACGGT	60 ± 4	2.5	-26 ± 0.7	-16 ± 0.7
7-9R GGT <sup>7</sup> TACGGU	44 ± 7	1.8	-30 ± 2	-20 ± 2
1-9R GGUUACGGU	1930 ± 110	80	-16 ± 2	-8.1 ± 2

**Table 2.2 Thermodynamic Impact of Ribose Substitutions.** <sup>a</sup> Substituted nucleotides in red <sup>b</sup> Apparent  $K_D$ ,  $\Delta H$ , and  $T\Delta S$  are averaged from triplicate ITC experiments with standard error of the mean. Data shown in Appendix A. <sup>c</sup> Fold change is relative to cognate DNA  $K_D$  (24 nM) <sup>d</sup> Values from Dickey *et al.* (2013)



**Figure 2.1 Most Cognate Ligand 2' Hydroxyl Positions and All Thymine Methyl Groups Are Solvent Exposed** **A)** Expanded view of the cognate ligand (pink) and Pot1pC (green) for nucleotides 1-4 with the would-be position of a 2' hydroxyl highlighted with red circles. A 2' hydroxyl at G1 would likely cause mild steric clashes with neighboring nucleotides in the cognate position, but G2, T3, and T4 hydroxyls are solvent exposed. **B)** Expanded view of the cognate ligand (pink) and Pot1pC (green) for nucleotides 2-6 with the would-be position of a 2' hydroxyl highlighted with red circles. A 2' hydroxyl at A5 may be unfavorably forced into a hydrophobic pocket at this position. **C)** Expanded view of the cognate ligand (pink) and Pot1pC (green) for nucleotides 6-9 with the would-be position of a 2' hydroxyl highlighted with red circles. 2' hydroxyls at C6 and G7 would likely cause mild steric clashes with neighboring bases, but a G8 2' hydroxyl may actually form a favorable hydrogen bond with neighboring residues and a T9 2' hydroxyl is solvent exposed. **D)** The thymine groups of the cognate ligand (pink) bound to Pot1pC (green) are highlighted in yellow with black circles around the methyl groups. As revealed by the neighboring protein residue side chains in cyan, the protein makes little to no hydrophobic contacts to the methyl groups of the ligand.

between DNA and RNA of the same sequence. However, in the ssDNA complex, the methyl groups of the three thymine bases of the cognate sequence, as well as most of the ribose 2' hydroxyl positions, are primarily solvent exposed (**Figure 2.1**), with little to no interaction with the protein. In this context and given the multiple binding modes observed for non-cognate ssDNA ligands, predicting how Pot1pC discriminates against RNA cannot be confidently discerned from the ssDNA complex structure.

To resolve how the differences between ssRNA and ssDNA contribute to Pot1pC RNA discrimination, we probed affinity changes as a function of nucleotide position by using chimeric DNA/RNA ligands containing combinations of deoxyribose and ribose nucleotides, first by nucleotide triplets and then by individual nucleotides for the triplets that exhibited significant discrimination. In addition, the protein specificity for thymine methyl groups was examined through deoxyuridine substitutions. Following a previous strategy employed by Nandakumar et al.<sup>94</sup> for characterizing mammalian Pot1 RNA discrimination, we tested binding with ligands that group ribose substitutions in triplets at the first three nucleotides (1-3R), fourth through sixth (4-6R) nucleotides, and the last three nucleotides (7-9R) of the cognate ssDNA sequence. These experiments reveal that Pot1pC discriminates against ssRNA primarily at the first 6 nucleotides, with modest binding reductions observed in the 1-3R and 4-6R ligands and little to no affinity change for the 7-9R ligand (**Table 2.2**). Together, the sum of the energetic differences between the cognate ligand and the triplets is consistent with the energetic loss for the full RNA 9mer, suggesting that the binding perturbations are additive and not cooperative.

Individually, the nucleotide position that shows the most significant impact on binding affinity with the addition of the ribose 2' hydroxyl is G1 (1R). Substitution of this position with rG leads to a ~4-fold reduction in affinity (**Table 2.2**). In contrast to the methyl specificity exhibited by Pot1pN<sup>64</sup>, neither full substitution of the cognate sequence with uracil (T3-4-9dU) nor the ligands containing dT to rU substitutions (positions 3, 4, and 9; tested by ligands 3R, 4R, and 7-9R, respectively) significantly impact affinity, suggesting the addition of hydroxyl groups on the

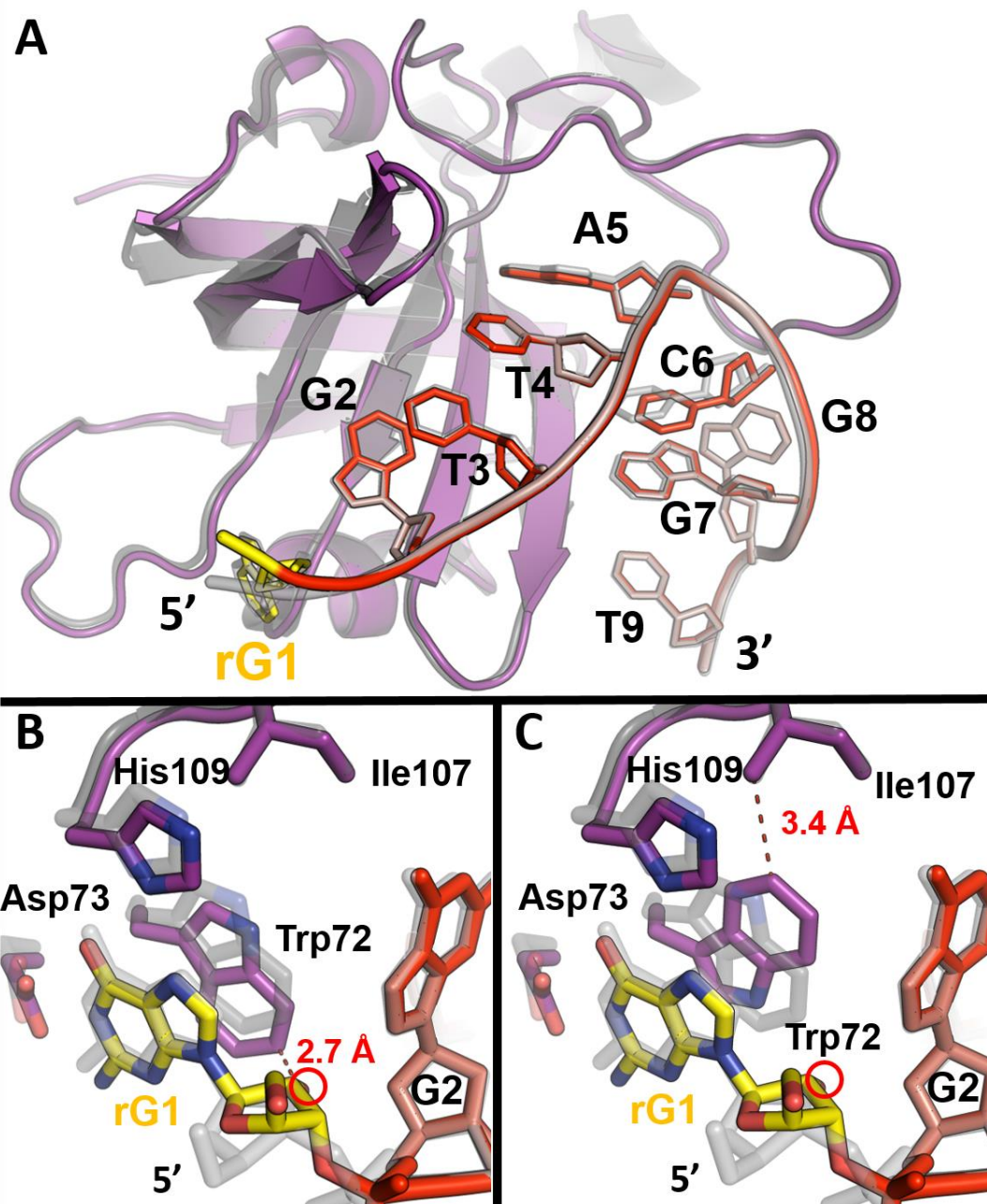


ribose moiety are primarily responsible for the loss of Pot1pC affinity. In the structural context of the cognate ssDNA Pot1pC complex, this is not unexpected as these methyl groups are primarily solvent exposed except for intra-molecular base stacking between T3 and T4 and a limited contribution to the aromatic stack between T9 and Trp27 and Tyr28 (**Figure 2.1**). Because dU and rU substitutions did not exhibit significant affinity loss at these positions, binding with rT substitutions was not tested.

### **2.3.2 – Discrimination in the 1R Structure**

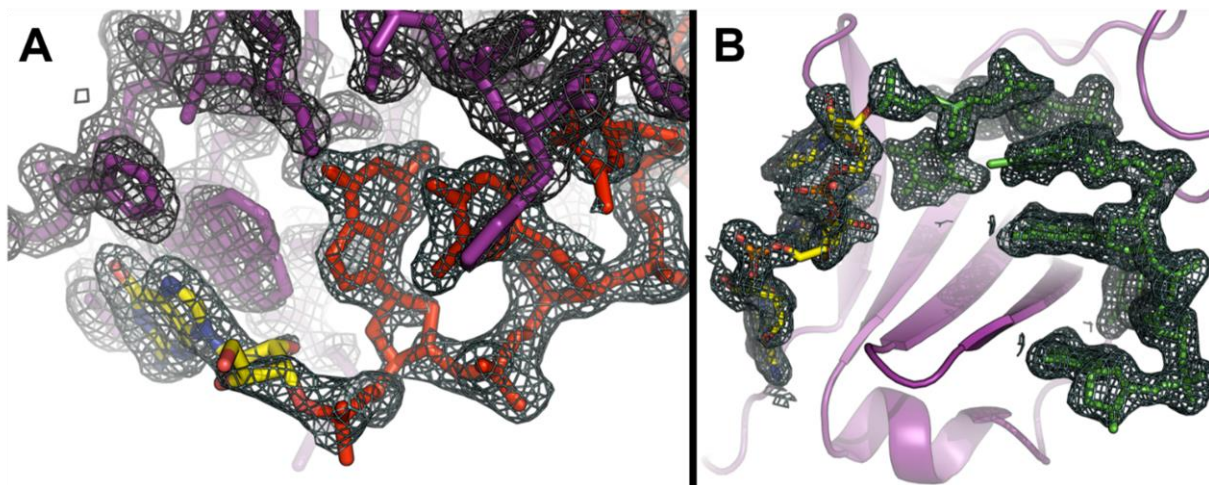
#### ***A suboptimal binding geometry between rG1 hydroxyl, Trp72, and the G2 base is the strongest individual discrimination determinant***

Following the lessons learned from the ssDNA-Pot1pC structures in which unexpected interactions were formed upon base substitution, we sought to resolve the underlying structural mechanisms responsible for Pot1pC 2' hydroxyl discrimination by solving high resolution structures of Pot1pC bound to chimeric ligand species. Diffraction quality crystals of the full RNA bound complex remained elusive, likely due to its weak affinity. Instead, noting the additivity observed in the thermodynamics of the chimeric ligands, we solved chimeric complexes where individual or groups of sites in the ssDNA were replaced with their ribose equivalent. For individual site replacement, we targeted the position that displays the most discrimination, (1R; which contains a dG to rG substitution at position 1). Using conditions established previously, we were able to obtain diffraction quality crystals and solved this structure (PDB ID: 5USB) to 1.62 Å resolution,  $R_{\text{work}}/R_{\text{free}}$  (0.1759 / 0.2027) using molecular replacement with the cognate bound Pot1pC structure (4HIK)<sup>66</sup>. The structure of the 1R complex overlays closely to the cognate structure aligning with a root-mean-square deviation (rmsd) of 1.16 Å (protein) and 0.75 Å (ligand) (**Figure 2.2A**).



**Figure 2.2 An Unfavorable Interaction Between rG1 Hydroxyl, Trp72, and G2 Base is the Strongest Individual Discrimination Determinant** **A)** 1R bound Pot1pC (5USB) shows high similarity to cognate Pot1pC (4HIK). Overlay shown for cognate (DNA; white) Pot1pC (gray) complex and 1R (red, rG1 substitution yellow) Pot1pC (purple) complex. **B/C)** Enlarged view of the rG1 binding site reveals most cognate binding features are maintained in the 1R complex. However, the rG1 2' hydroxyl forces Trp72 into two alternative conformations (Shown separately in panels B and C for clarity) with unfavorably close contact with either the 2' hydroxyl of rG1 (**B**) or Ile107 (**C**). The 2' hydroxyl is highlighted by red circle in panels B and C.

Curiously, despite binding with a ~4-fold weaker affinity, the 1R complex maintains all of the hydrogen bond and stacking interactions observed in the cognate structure (**Figure 2.2B**). Closer examination of the 2' hydroxyl of position 1 suggests the loss of affinity is the result of an unfavorably close contact between the ribose moiety of rG1 and the cognate positioning of Trp72 (**Figure 2.2B/C**). While the conformation of the ribose is somewhat ambiguous due to weaker electron density and higher B-factor than the neighboring groups, its relative position is constrained by the strong electron density for the base and phosphate moieties (**Figure 2.3A**). Thus, favorable ribose conformations place the 2' hydroxyl (as modeled) or alternatively the ring O4' oxygen (not shown) into an unfavorably close contact to Trp72 (2.7 Å or 2.3 Å between heavy atoms, respectively; **Figure 2.2B**) and close enough for the 2' hydroxyl to form a hydrogen bond with the ribose ring oxygen of G2. Careful examination of the electron density of Trp72 (**Figure 2.3A**) suggests this residue partially alleviates this steric clash by adopting another conformation (**Figure 2.2C**). However, the alternative conformation also has an unfavorably close contact to Ile107 (3.4 Å between heavy atoms) which is likely why Trp72



**Figure 2.3 Ligand Electron Density for 1R and 1-3R** **A)** Electron density map (2mFo-DFc at 1  $\sigma$ ; Pot1pC purple, 1R red with rG1 highlighted in yellow) of the G1 binding pocket suggests Trp72 adopts an alternative conformation to partially alleviate a close contact between the 2' OH and Trp72 by flipping into close contact with Ile107. **B)** Electron density map (2mFo-DFc at 1  $\sigma$ ; Pot1pC purple, 1-3R green with 1-3R highlighted in yellow) of the 1-3R ligand reveals a well-defined electron density for most regions of the ligand.

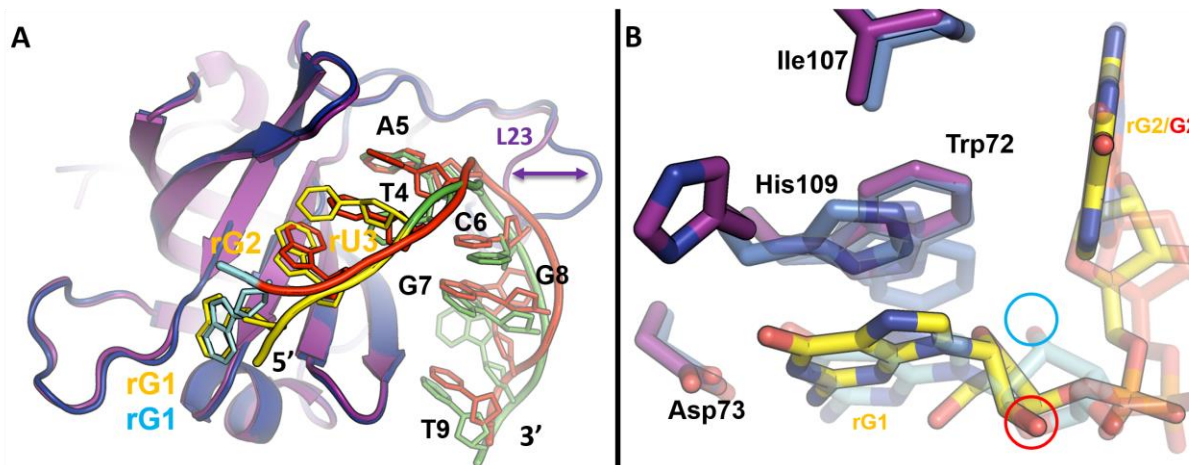
adopts a conformation that is positioned to clash with a 2' hydroxyl at position 1 (**Figure 2.2B**) in the cognate ligand bound structure. Together, these suboptimal conformations enforced by the positions of the rG1 guanine and phosphate likely comprise the mechanism of discrimination for this ligand, as it can only be bound in a non-ideal configuration resulting in the observed loss in affinity. While the rG1 substitution can force accommodation, it does not recapitulate the full binding energy of the cognate DNA due to unfavorable interactions between Trp72 and the ribose moiety or the neighboring Ile107.

### **2.3.3 – Discrimination in the 1-3R Structure**

#### ***Pot1pC plasticity partially, but not fully, compensates for lost interactions in presence of the 2' hydroxyl at positions 1-3***

To better understand the above structure and expand our understanding to additional positions, we solved the structure of the first triplet complex (1-3R; d(GGT) to r(GGU); PDB ID: 5USN, 1.9 Å,  $R_{\text{work}}/R_{\text{free}}$ : 0.1829/0.2231). This complex shows dramatic rearrangement of both the protein and ligand relative to the 1R (1.53 Å protein, 2.34 Å ligand rmsd; **Figure 2.2A**) and cognate structures (1.55 Å protein, 2.39 Å ligand rmsd; **Figure 2.4A**). The conformational changes are especially significant in the 3' portion of the ligand and  $\beta 2$ - $\beta 3$  loop (L23), a region distant from the site of modification, suggesting that despite the propagation of conformational changes in the ligand backbone, the protein interface is able to make similarly large adjustments to compensate.

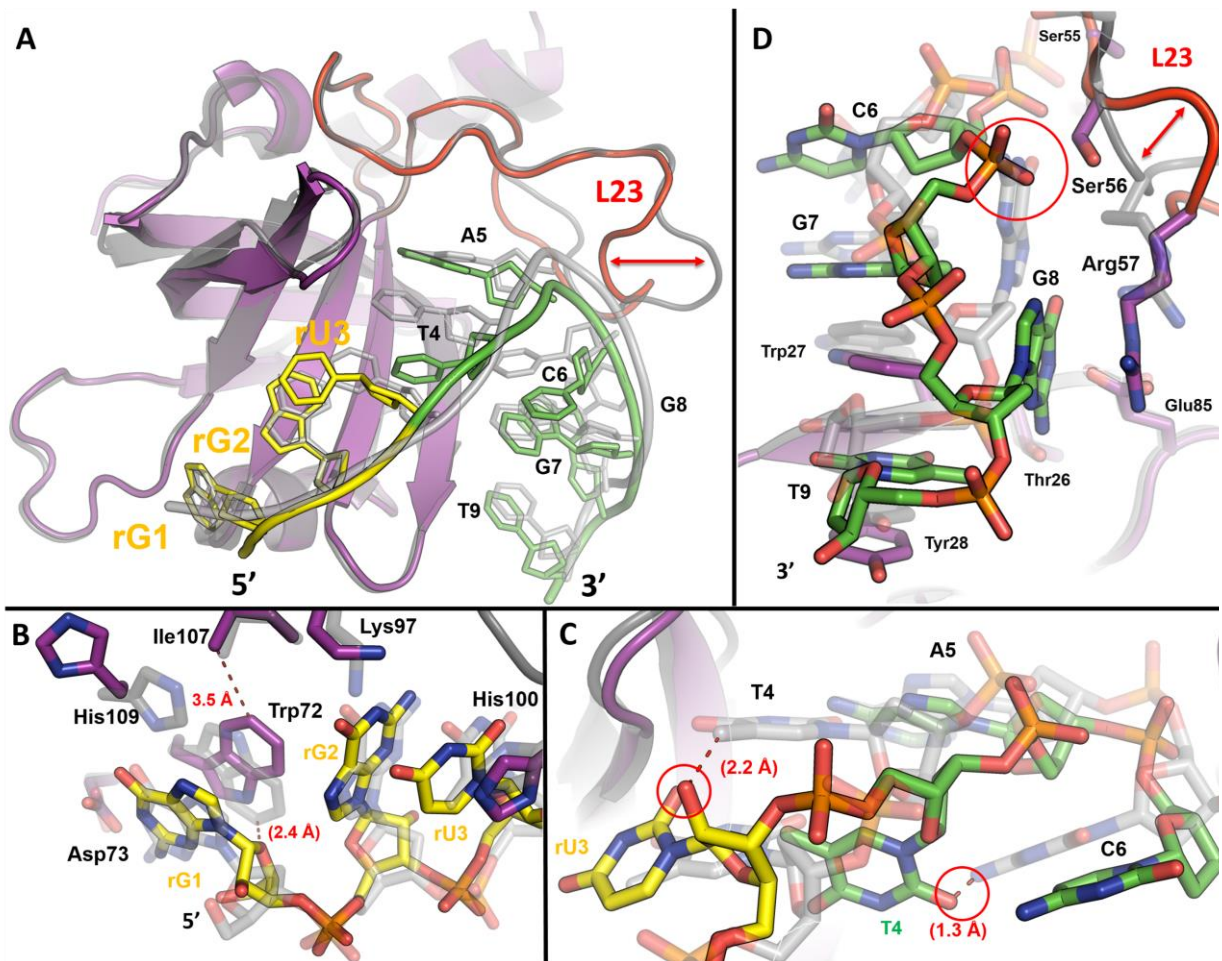
In this structure, relative to both the cognate and 1R structures, the rG1 ribose (**Figure 2.4B**) is flipped, roughly swapping the positions of the 2' hydroxyl and 5' carbons. Without



**Figure 2.4** 1-3R Binds Pot1pC in an Alternative Binding Mode, Recapitulates the Unfavorable Interactions Seen in the 1R Structure with the Additional Loss of the His109-G1 Interaction **A)** 1-3R bound Pot1pC (5USN) shows significant conformational changes compared to 1R bound Pot1pC (5USB). Overlay shown for the 1R (red, 1R substitution cyan) bound Pot1pC (blue) compared to the 1-3R (green, 1-3R substitutions yellow) bound Pot1pC (purple). The conformational change of L23 is highlighted by a purple arrow. **B)** Comparison of the 1R and 1-3R G1 binding sites reveals a shared steric clash between Trp72 and Ile107 to accommodate the rG1 2' hydroxyl despite different ribose conformation models (2' hydroxyls highlighted with circles; 1R cyan, 1-3R red). In addition, the 1-3R complex loses the interaction between His109 and G1.

further reorganization, this ribose conformation would have put the ring O4' oxygen within a predicted 2.4 Å of the cognate conformation of Trp72 (**Figure 2.4B**) based on structural alignments. Presumably to avoid this unfavorable contact, Trp72 rotates roughly 180° into a new conformation, placing it into close proximity to Ile107 and forming a new hydrogen bond between the rG1 ring O4' oxygen and the indole nitrogen of Trp72.

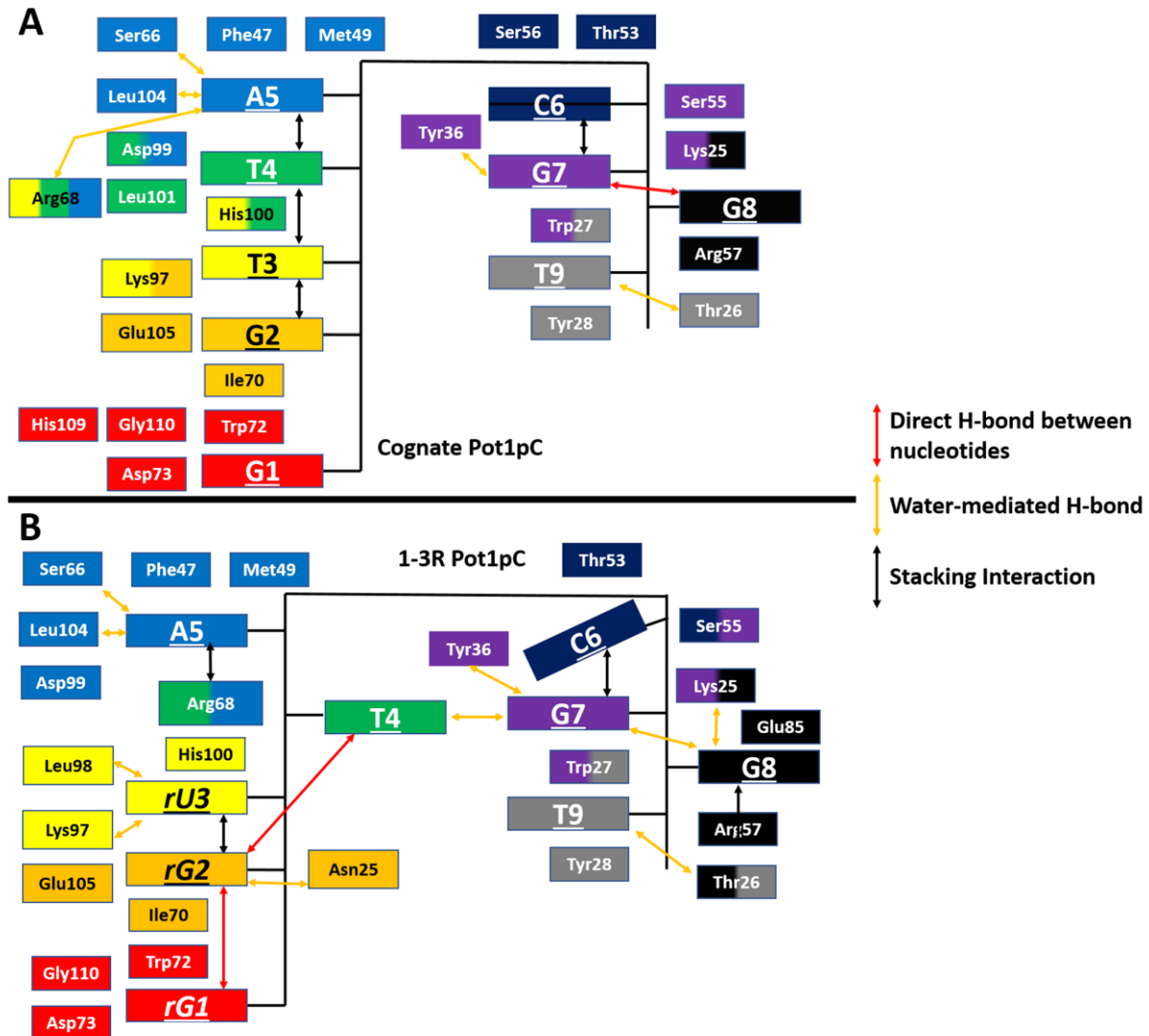
The 1-3R structure contains additional changes near position 1 (**Figure 2.4B**) relative to those observed in the 1R structure. The 1-3R rG1 maintains the hydrogen bonds to the amide backbone of Gly110 and the side chain of Asp73, but, relative to the 1R and cognate DNA complexes (**Figures 2.2B and 2.4B**), His109 rotates ~120 degrees out the binding pocket and no longer forms the hydrogen bond with N7 of G1. Thus, the 1-3R complex supports the steric



**Figure 2.5 Structural Rearrangement of the 1-3R Ligand Driven by Conformational Changes in the Sugar Phosphate Backbone and Alleviation of Steric Clashes at the Bases** **A)** 1-3R bound Pot1pC (5USN) reveals significant conformational changes compared to the cognate bound Pot1pC (4HIK). Overlay shown for the cognate DNA (white) bound Pot1pC (gray) compared to the 1-3R (green, 1-3R substitutions yellow) bound Pot1pC (purple). The shift of the protein backbone near the 3' portion of the ligand (red arrow) illustrates the conformational plasticity of L23 (red). **B)** Comparison of the 5' portion of the 1-3R complex (1-3R substitutions yellow; Pot1pC, purple) with the cognate Pot1pC (DNA, white; Pot1pC, gray) reveals several lost interactions and an unfavorable steric clash. Predicted distances between cognate and 1-3R complexes are in red parentheses, observed distance in red without parentheses. **C)** Comparison of nucleotides 3-6 of the 1-3R ligand (ligand green, substitutions yellow) to cognate reveals substantial rearrangement of non-substituted nucleotides. A would-be steric clash between rU3 2' hydroxyl (red circle) and T4 and rearrangement of the ligand backbone shifts T4 into a new binding pocket. This new position clashes with the cognate position of C6 (red circle) and results in another nucleotide shift. Predicted distances between cognate and 1-3R complexes are in red in parentheses. **D)** Comparison of the 3' portion of the 1-3R ligand reveals additional changes compared to the cognate complex. The shift of C6 positions the phosphate of G7 into the cognate position of G8 (red circle). G8 flips and L23 residues shift to accommodate the new positions of the phosphate groups and the G8 base while Glu85 and Thr26 are already poised to form new hydrogen bonds with the G8 base. The shift downward of the G7 and T9 bases are matched by corresponding shifts in Trp27 and Tyr28.

the lost interaction with His109.

The 1-3R ligand dramatically diverges in conformation far beyond the sites of substitution, with conformational changes encompassing the entire length of the nucleic acid (**Figure 2.5A**). The disruptions begin near the site of substitution and appear to be due to a chain of conformational changes that occur to alleviate close contacts. The shift in the positions



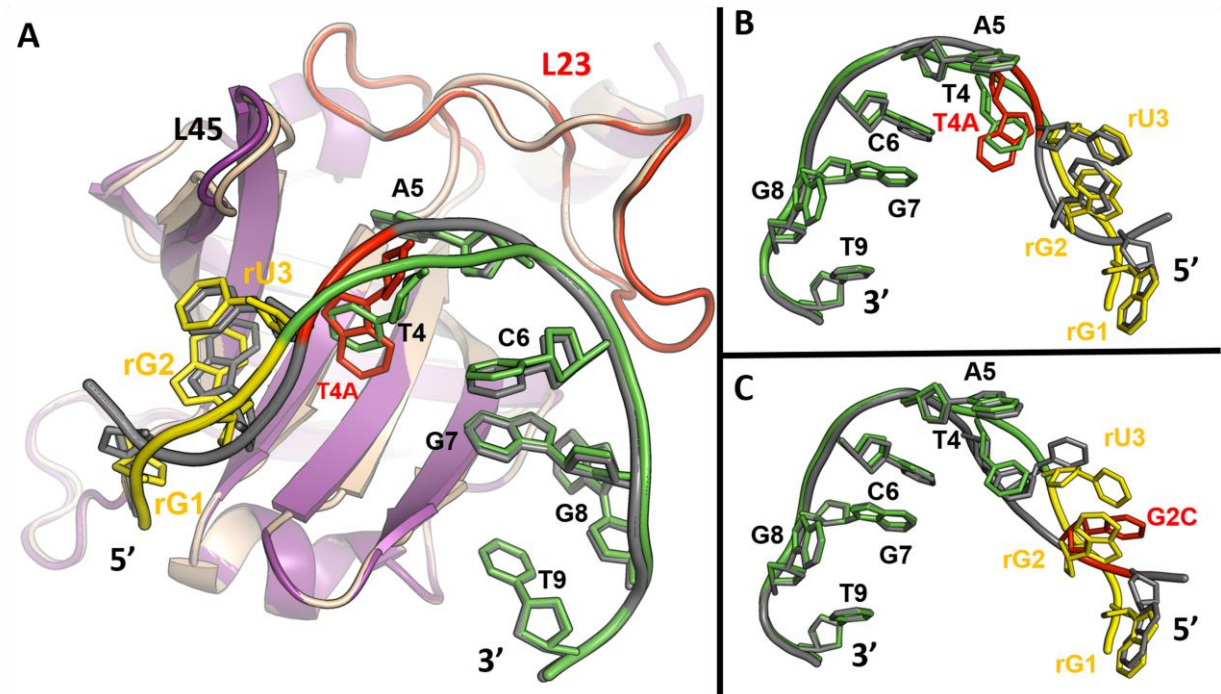
**Figure 2.6 Rearrangement of the 1-3R Ligand Results in a Mix of Lost and Compensatory Interactions** **A)** Schematic summarizing the cognate and **B)** the 1-3R ligand-protein interactions. Nucleotide interacting residues are color coded by interaction partner. Multicolored residues indicate interactions with more than one nucleotide. Red arrows indicate hydrogen bonds between nucleotides; orange arrows, a water mediated interaction; and black arrows, stacking interactions.

of rG2 and rU3 disrupts a cognate G2-Lys97 interaction and changes the T3-His100 hydrogen bond interaction to a  $\pi$ -stacking interaction (**Figure 2.5B**). The combination of a predicted close contact between the rU3 2' hydroxyl and the cognate position T4 and the conformational change in the sugar-phosphate backbone results in T4 swinging into a new binding pocket (**Figure 2.5C**) and disrupting most T4-protein interactions. In this new pocket, T4 clashes with the cognate position of C6, forcing it to adopt a new position (**Figure 2.5C**) while A5 maintains its cognate interactions and binding pocket. Rearrangement of the C6 base positions the G7 phosphate into the cognate position of G8 (**Figure 2.5D**). The rearrangement of G7 is accommodated by shifts in L23 residues Ser56 and Arg57 while disrupted G8-protein interactions are compensated by new G8 interactions with Thr26 and Glu85. The changes in C6 and G8 also shift G7 and T9 downwards (**Figure 2.5D**), but these changes are accommodated by a corresponding movement of Trp27 and Tyr28 which stack with G7 and T9. A summary of the ligand protein interactions for cognate and the 1-3R complexes are summarized schematically in **Figures 2.6A and 2.6B**, respectively. Notably, similar structural changes for positions 6-9 and L23 are seen in several other Pot1pC complexes that exhibit no binding defect, suggesting the unfavorable interactions at rG1 and lost interactions for rG1 and T4 are the driving mechanisms behind the 1-3R ligand discrimination.

### **2.3.4 – Cryptic secondary binding mode is widely used to provide partial thermodynamic compensation**

Even though the interface is quite different than in the cognate structure, the binding mode of the 1-3R structure shows striking similarity to the structural changes observed in two completely different non-cognate complexes, the T4A Pot1pC complex and G2C (1-3R vs. T4A: 0.862 Å protein, 1.30 Å ligand; 1-3R vs. G2C: 0.883 Å protein, 1.83 Å ligand; 1-3R vs. cognate: 1.55 Å protein, 2.39 Å ligand rmsd) <sup>66</sup>. These structures were characterized by the base at position 4 flipping 55° away from the  $\beta$ 4- $\beta$ 5 loop (L45) and towards the  $\beta$ -barrel. In the 1-3R





**Figure 2.7 1-3R Binds Pot1pC More Like the T4A and G2C DNA Ligands Than the Cognate DNA Ligand** **A)** Comparison of the 1-3R (5USN) and T4A (4HIO) complexes reveals high similarity. Overlay shown for 1-3R bound Pot1pC (1-3R green, substitutions yellow; Pot1pC, purple) and T4A bound Pot1pC (ligand gray, T4A substitution red). T4 of 1-3R occupies the same binding pocket as A4 of T4A, L23 (red). **B)** Overlay of 1-3R and T4A ligands shown rotated  $\sim 180^\circ$  relative to (A) shows high similarity between the T4A and 1-3R binding modes, with greater agreement in the 3' portion of the ligands. **C)** Overlay of the 1-3R and G2C (4HID, gray, G2C substitution red) ligands shows high similarity in the G2C and 1-3R binding modes.

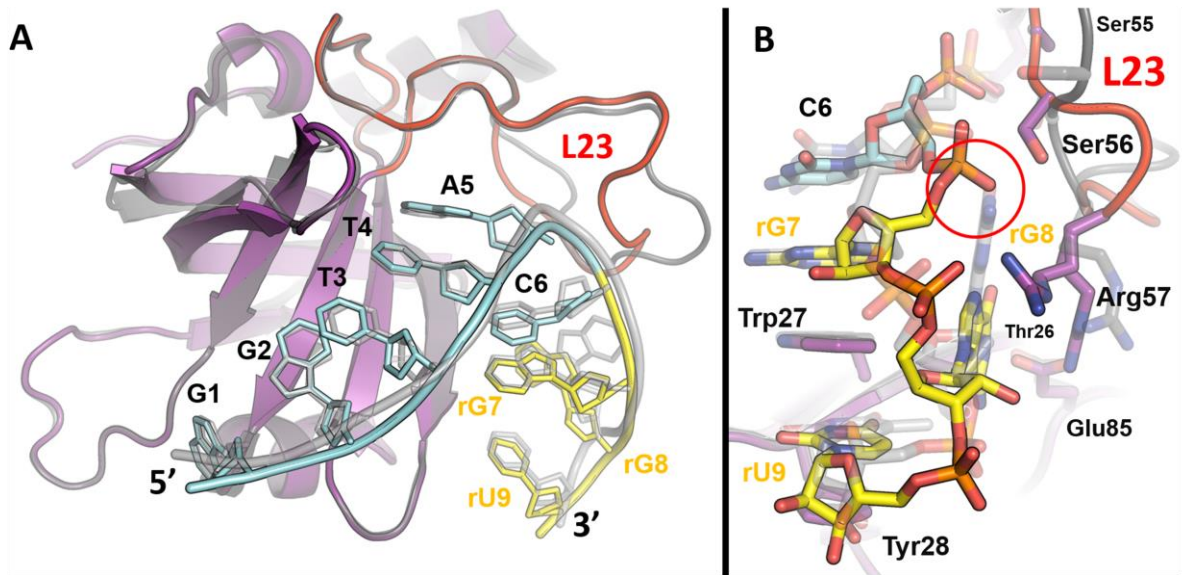
complex T4 rotates into the same binding pocket observed for the adenine of the T4A DNA substitution (**Figure 2.7A/B**) as well as T4 in the G2C complex (**Figure 2.7C**). Like 1-3R, both T4A and G2C diverge in ligand conformation relative to the cognate ligand well beyond the sites of substitution. The large rearrangement of the 3' portion of the ligand likely results from an overlap in the cognate position of C6 and the new position of nucleotide 4. While the T4A DNA binding mode has an equivalent affinity to the cognate DNA ligand, both the 1-3R and G2C complexes have reduced affinity despite ostensibly binding in the same binding mode (1-3R, 9-fold; G2C, 36-fold). The common difference is that these ligand complexes flip T4 into the adenine binding pocket of T4A. The use of the T4A adenine binding pocket appears to not fully compensate for the interactions lost upon T4 flipping into this pocket as thymine is not large

enough to reach the  $\beta$ -barrel residues in this pocket like T4A, resulting in the observed net loss of binding affinity due to suboptimal complex geometries.

### 2.3.5 – Ligand Accommodation in the 7-9R Structure

#### ***Backbone alterations at the 3' end of the ligand are readily accommodated by ligand and L23 structural rearrangement***

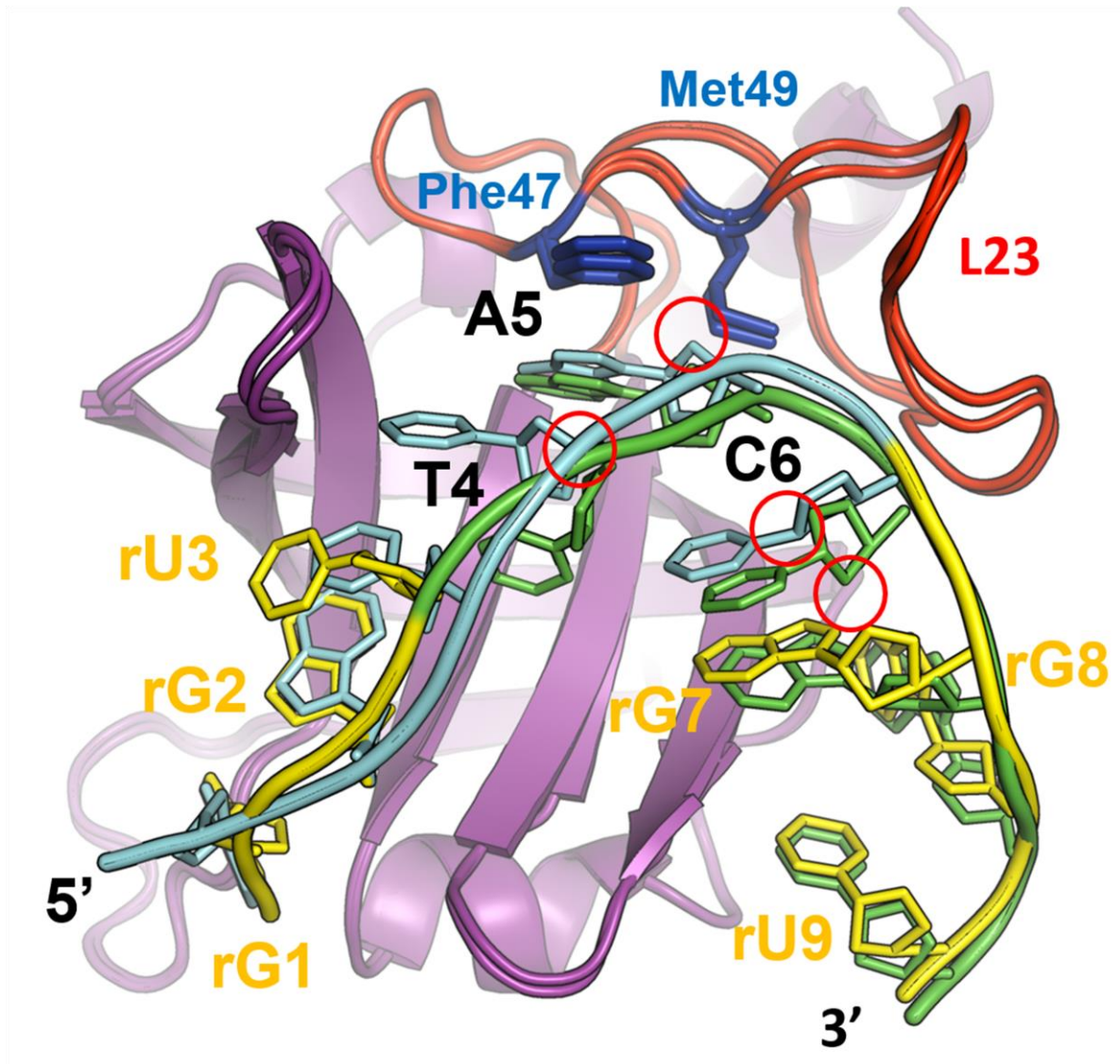
In contrast to the thermodynamic consequences of introducing riboses at positions 1-6, ribose incorporation at positions 7-9 has no impact on binding affinity. To address how this full accommodation is achieved, we solved the crystal structure (7-9R; d(GGT) to r(GGU); PDB ID: 5USO, 2.0 Å,  $R_{\text{work}}/R_{\text{free}}$ : 0.2102/0.2457). Instead of simple non-specific recognition of the backbone at these positions, the crystal structure of 7-9R reveals rearrangement of the ligand for nucleotides 6-9 compared to the cognate structure (1.31 Å protein, 1.63 Å ligand rmsd; **Figure 2.8A**). As expected from the cognate structure in which the simple addition of the 2' hydroxyl at G7 would clash with the base position of G8, the sugar orientation of rG7 shifts such that the 2' hydroxyl is instead solvent exposed (**Figure 2.8B**). This rearrangement of the sugar-phosphate backbone has consequences elsewhere in the ligand, altering the orientation of C6 as well as the rG7 phosphate resulting in the phosphate clashing with the cognate position of the G8 base. In turn, the rG8 base rotates 180° around the glycosidic bond, removing the intramolecular hydrogen bond between G8 and the G7 phosphate. As seen in the 1-3R complex, the L23 residues Ser56 and Arg57 rearrange to accommodate these shifts while the side chains of Thr26 and Glu85 are statically poised to interact with the new position of rG8. In addition, the rG8/rU9 backbone shifts downward relative to the cognate backbone. Despite these changes, the rU9 base is readily accommodated in essentially the same relative positioning as the cognate T9 hydrophobic stack between Trp27 and Tyr28 (**Figure 2.8B**). In large part, the flexibility of the sugar phosphate backbone and L23 facilitates this accommodation and contributes to the plasticity of the interface.



**Figure 2.8 7-9R Binding Causes Rearrangement of L23 to Form Thermodynamically Compensatory Interactions in Response to Ligand Backbone Conformational Changes** **A)** 7-9R bound Pot1pC (5USO) reveals rearrangement of L23 and G8 in response to substitutions relative to cognate (4HIK). Overlay shown for the cognate DNA (white) bound Pot1pC (gray) compared to the 7-9R (cyan, 7-9R substitutions yellow) Pot1pC (purple). **B)** Comparison of the binding pocket of the 7-9R substitutions (7-9R ligand cyan, substitutions yellow; Pot1pC, purple) with the cognate DNA ligand (white; Pot1pC, gray). Despite all 2' hydroxyls being solvent exposed, rearrangement of the sugar phosphate backbone forces a shift of rG8 due to rG7 phosphate moving into the cognate position of G8 (red circle). Rearrangement of the L23 (red) residues accommodate the shift in ligand backbone while Glu85 and Thr25 are poised to interact with the flipped G8.

### 2.3.6 - Model for 4-6R and full RNA Pot1pC complex structures

The structures we have solved allows us to develop a model for how the full RNA 9mer binds. Comparison of the ligand conformations for the 1-3R and 7-9R complexes reveals that the binding mode of nucleotides 5-9 is highly similar (**Figure 2.9**). This apparent compatibility of these binding modes, which also share many features with the binding mode of the T4A and G2C DNA ligands, suggests that the full RNA 9mer may also bind in this major binding mode. Based on the overlap of positions 5 and 6 in all Pot1pC structures, the speculation that they



**Figure 2.9 Comparison of Sugar rG1 Sugar Conformations in 1R and 1-3R Complexes and Model for Discrimination at Positions 4-6.** Comparison of the 1-3R (5USN) and 7-9R (5USO) complexes (all substitutions yellow; Pot1pC purple, 1-3R ligand green, and 7-9R ligand cyan) suggest compatibility between the binding modes of these complexes. Both L23 (red) and the 3' portion of the ligands agree reasonably well, suggesting the full 1-9R complex may bind in a similar binding mode. The predicted positions of the 2' hydroxyl of positions 4-6 are highlighted with red circles for this binding mode. The U4 and A5 2' hydroxyls are both poised to force a steric clash in the 1-3R binding mode. U4 with the A5 base while A5 is positioned in a clash with Phe47 and Met49 (blue). The C6 2' hydroxyl shows more ambiguity of position with a close contact with G7 in the 1-3R complex but is positioned to form a potential hydrogen bond with Lys25 in the 7-9R complex.

apparent energetic additivity for the triplets compared to the full RNA 9mer. If this were the case, we can speculate on the mechanisms of discrimination for the 4-6R ligand based on the ligand overlap of these complexes. At nucleotide rU4, the 2' hydroxyl would clash with the rA5 position in the 1-3R binding position while being solvent exposed in the cognate/7-9R T4 position. For position 5, the rA5 2' hydroxyl would unfavorably occupy a hydrophobic pocket created by Phe47 and Met49 (**Figure 2.8**). In addition, rC6 presents an apparent steric clash between the would-be 2' hydroxyl and the base of G7 in the 1-3R binding conformation (**Figure 2.8**) but could form a new hydrogen bond with Lys25 in the 7-9R conformation. While this speculation for the 4-6R mechanisms of discrimination should be taken with caution in the absence of definitive structural information for these substitutions, features of L23 suggest the plasticity used to bind the 3' portion of the ligand does not extend to Phe47 and Met49. The neighboring residue Phe46 is buried in the hydrophobic core of the protein and nearby prolines, Pro48 and Pro51, severely limit the backbone conformations favorable for Phe47 and Met49, suggesting accommodation is likely to resemble the suboptimal geometry of Trp72 rather than wholesale rearrangement of Arg57 and Ser56.

## 2.4 - Discussion

*S. pombe* Pot1 has the challenge of recognizing an inherently degenerate telomere sequence, G<sub>2-8</sub>TACGGT(A)<sup>43,95</sup>, with both high specificity and affinity. In doing so, it must discriminate against ssDNA with similar sequence as well as the much more abundant RNA ligands containing the identical sequence. *Sp*Pot1 accomplishes this DNA specificity by having evolved a modular DNA-binding domain comprising a sequence-specific binding domain, Pot1pN<sup>64</sup>, and a non-specific binding domain, Pot1pC<sup>44,74,66</sup>, which contribute equally to the full-length binding affinity. Surprisingly though, both the sequence-specific and sequence non-specific domains discriminate against the telomeric sequence rendered in RNA<sup>64,105</sup>. An *a priori* rationale to explain the full extent of how Pot1pC achieves this is difficult to formulate as the

bulk of the interaction between Pot1pC and its ligands are primarily base-mediated hydrogen bond interactions and comparatively few interactions with the sugar-phosphate backbone<sup>66</sup>. Moreover, this protein/nucleic acid interface exhibits remarkable plasticity capable of both subtle and dramatic structural rearrangements of the protein and ligand to form thermodynamically equivalent complexes. At first glance, the three thymine bases in the cognate ligand could suggest that uridine substitution may explain ssRNA discrimination, but the methyl groups only interact in limited aromatic stacking interactions and are partially solvent exposed (**Figure 2.1D**). Likewise, the 2' hydroxyl groups are mostly solvent exposed in the cognate and non-cognate structures (**Figure 2.1A-C**). Therefore, neither the substitution of thymines to uridine nor the addition of 2' hydroxyl groups into the cognate structural conformation provides a satisfactory explanation for the 80-fold reduction in affinity observed for ssRNA.

Our set of chimeric ssRNA-ssDNA Pot1pC complex structures allow us to identify the features of the Pot1pC binding interface that provide specificity for ssDNA over ssRNA while binding ssDNA with a surprising level of non-specificity. Overall, the underlying Pot1pC specificity for DNA ligands appears to result from forcing RNA ligands into suboptimal binding geometries in regions of the protein with less conformational flexibility. In these regions of the protein, which are responsible for interacting with the first 5 nucleotides, base substitutions are accommodated by residues poised for new interactions with alternative bases, but unlike L23 are limited primarily to rotameric rearrangements of side chains. The 1R and 1-3R structures reveal that the first nucleotide of the 9mer ligand is responsible for the largest individual discrimination observed for Pot1pC through an unfavorable interaction between the ribose moiety and residues on the protein surface (**Figure 2.2B/C**). Speculatively, 4-6R also places the 2' hydroxyl of rA5 in a similar clash based on the ligand conformations seen in all solved Pot1pC complexes (**Figure 2.9**). In addition, differential ligand flexibility appears to underlie the mechanism of discrimination at other positions. While ssDNA and ssRNA are both highly flexible ligands and ssDNA can generally adopt the same conformations as ssRNA, the reverse is not

strictly true - the presence of the 2' hydroxyls in RNA change the energy landscape favorable for different backbone conformations. In the case of the 1-3R structure, the sugar orientation of nucleotide 3 switches from the cognate binding mode to that seen in the T4A and G2C binding modes with the accompanying rearrangement of T4 into a non-compensatory binding pocket in which the side chains that interact with the T4A substitution are unable to reach the smaller thymine. The 7-9R structure shows a similar structural impact of differential flexibility but exemplifies the difficulty of predicting the biochemical impact of ribose nucleotides at specific positions based on the structure of the Pot1pC cognate complex. The cognate orientation of the G7 ribose suggests a steric clash with the G8 base would result in the loss of binding affinity. However, this clash is alleviated through rearrangement of the sugar-phosphate backbone, rotation of the G8 base about the glycosidic bond, and concomitant changes in loop conformation which underscore the ability of the protein to exhibit a flexibility as dramatic as the more obviously flexible ligand. The features of the rearrangement in the 3' portion of the 7-9R ligand and L23 are largely recapitulated in the 1-3R, G2C, and T4A complexes. Our ability to solve a range of structures has revealed that rather than adopting a mashup of many conformations, Pot1pC has a widely utilized alternative binding mode that is employed partially or in full in response to myriad chemical modifications.

Structures of related ssDNA-binding proteins reveal other mechanisms utilized for RNA discrimination. In the case of mammalian Pot1, the underlying mechanism of discrimination occurs in the region of the protein homologous to *S. pombe* Pot1pN, as the C-terminal domain of the human protein, analogous to Pot1pC, barely engages with the ligand<sup>65</sup>. Discrimination in this case appears to result primarily from losing a hydrophobic interaction from a thymine methyl as well as forcing the 2' hydroxyl at that position into a sterically unfavorable interaction in that same hydrophobic pocket<sup>94</sup> in a manner similar to our speculated mechanism of discrimination at A5. However, other positions throughout the ligand also contributed to ssDNA specificity, especially when placed near the discriminating position – suggesting nearby ribose nucleotides

reinforce the suboptimal binding geometries at discriminating positions by further limiting the conformational flexibility of the ligand<sup>64</sup>. While an RNA bound or chimeric complex for Pot1pN has not been solved, predictions for the mechanisms of specificity can be made based on the available structures and existing biochemical data. If the RNA adopted the same conformation as the DNA bound complex, there would be readily apparent steric clashes for hydroxyls at two positions and an energetically unfavorable hydrophobic pocket left empty by the loss of a thymine methyl group<sup>64</sup>. At both positions (GGTTAC, cognate), a greater than 200-fold discrimination is observed at T3 attributable entirely to the 2' hydroxyl while T4 exhibits ~100-fold affinity reduction due to the methyl and ~8-fold reduction due to the 2' hydroxyl<sup>64</sup>. The presumed 2' hydroxyl steric clash at Pot1pN's T4 results in a binding defect in line with our observations for the 1R and 4-6R substitutions whereas the loss of methyl interactions has a much greater effect for *S. pombe* Pot1pN but not mammalian Pot1. *S. pombe* Pot1pN more strongly engages with the methyl group of T4, with empty space left behind in hydrophobic pockets in their absence whereas Pot1pC has far fewer close contacts to the methyl groups of its ligand. Together, these suggest that both mechanisms of discrimination can strongly favor ssDNA over ssRNA but depend on the context of the ligand-protein contacts. Thymines buried in hydrophobic pockets such as Pot1pN T4 can strongly discriminate against ssRNA or have only modest effects when not buried such as the analogous position in mammalian Pot1 which does not engage the T4 methyl to the same extent<sup>64</sup>. Discrimination against 2' hydroxyls appears to primarily result from steric clashes with the ribose moiety or forcing suboptimal binding geometries resulting from rearrangement of the sugar-phosphate backbone in response to the differential conformational flexibility of ribose and deoxyribose moieties. However, the impact of suboptimal binding geometries largely depends on the strength and specificity of the interactions being disrupted and can also range from zero to modest as seen for Pot1pC or strong in the case of Pot1pN.



Radical conformational adjustment to confer non-specific binding has been observed in other systems. The atypical Puf domain of Puf5 allows for the specific binding of RNA sequences of variable length (8 to 12 nt) through rearrangement of the ligands, though without the concomitant rearrangement of the protein<sup>118</sup>. The change in RNA ligand conformation places spacer nucleotides into non-specific pockets at the protein interface or arranges them in stacking interactions with other nucleotides. A similar mechanism of binding is seen for the *Oxytricha nova* telomere end-binding protein in which non-cognate bases are flipped out of the binding interface and neighboring nucleotides are shuffled into 'cognate-like' conformational register<sup>119</sup>. In other systems, the ability of plastic interfaces to accommodate cryptic ligand specificities through similar mechanisms may play important, but currently unappreciated, biological roles. For example, the plasticity exhibited by SH2 domains, PLC $\gamma$ 1 and SH2B1, for phospho-tyrosine ligands divergent from cognate specificities reveal the potential for these proteins to play roles in additional signaling pathways.<sup>120</sup>

A key result of these studies of altered complexes is that the degree of specificity exhibited for flexible ligands is carefully tuned through the use of alternative conformations. In particular, the extended L23 of Pot1pC, which provides key structural rearrangement for ligand accommodation, may provide hints for the mechanisms of specificity at other protein interfaces. Long loops, while often showing poor electron density, may well play general roles in ligand accommodation, especially for chemically diverse ligands. These sophisticated mechanisms of conformational malleability observed for Pot1pC are likely shared by other single-stranded nucleic acid binding proteins that are either fully non-specific, such as RPA,<sup>84</sup> or address the same non-degenerate specificity requirements as *S. pombe* Pot1, such as the *S. cerevisiae* Cdc13-Stn1-Ten1 complex<sup>121</sup> and human CST.<sup>122,123</sup> Moreover, recognition of intrinsically disordered peptides shares many of the same fundamental features and challenges of single-stranded nucleic acid recognition in that disordered proteins are highly flexible and are often comprised of similarly low complexity sequences. As a result, some proteins may recognize

some degenerate low complexity peptide sequence by providing an interface rich with potential favorably interacting residues and discriminate against other similar disordered peptides through differential flexibility preventing full interface utilization by those ligands.

## **Chapter 3 – Implications of Newly Discovered RNA-Binding By Cyclophilins**

### **3.0 – Chapter Overview**

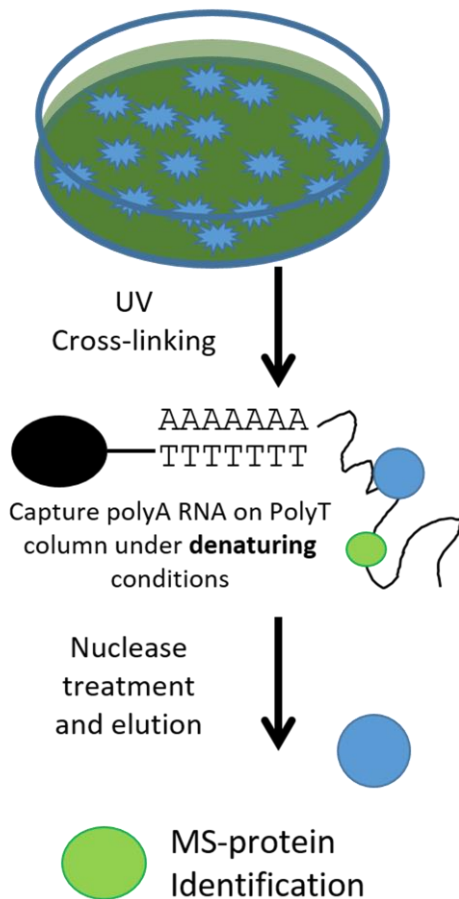
The second half of my thesis is motivated by the exciting discovery that over 40% of the proteins that bind mRNA in cells do not contain identifiable RNA-binding domains.<sup>124–127</sup> Because of our general interest in protein-nucleic acid specificity and the new avenues of regulation of biology by RNA, we began to study the cyclophilin family of proline isomerases as a model system for these novel RNA-binding domains. This chapter provides a brief description of these global studies and what was discovered, why we chose the cyclophilin family as a model system, and a general overview of cyclophilin biology with a focus on how it relates to RNA.

### **3.1 - Identification of Cyclophilins as Non-canonical RNA-Binding Proteins**

RNA-proteins interactions govern the behavior of many facets of biology, such as differential splicing patterns of mRNA,<sup>128–130</sup> repression and activation of translation of specific mRNAs,<sup>130–132</sup> as well as stability and degradation of RNAs,<sup>133–135</sup> and even regulation of chromatin states.<sup>136–139</sup> As such, characterizing the protein domains involved in RNA binding has been a core feature in understanding these biologies. Emerging from this large body of work, numerous classical RNA binding domains have been identified<sup>140,141</sup> such as Oligonucleotide-Oligosaccharide-Binding fold (OB),<sup>60,61,142</sup> RNA-recognition motifs (RRM),<sup>81,143,144</sup> Zn-finger domains,<sup>145–147</sup> K-homology domains (KH),<sup>148,149</sup> double-stranded RNA binding motifs (dsRBM),<sup>150–152</sup> among several others.<sup>118,136,137,140,141,153</sup>

Following in this tradition, revolutionary new technologies allow us to understand protein-nucleic acid interactions in unprecedented breadth as global transcriptomics and proteomic experiments become increasingly feasible, opening up exciting new avenues of investigation.

### Cross-linking to find full mRNA proteome



**Figure 3.1 General Protocol Schematic of the Crosslinking Studies** (Baltz et al. 2012, Castello et al. 2012, Mitchel et al. 2013, and Kwon et al. 2013)

Unexpectedly, unbiased cross-linking and mass spectroscopy studies performed by Baltz *et al.*<sup>124</sup> and Castello *et al.*<sup>125</sup> strongly implicated a surprising number of proteins as direct, RNA-binders, even though they contain no known nucleic acid binding motifs. These cross-linking studies incorporated photoreactive nucleosides into all RNAs, cross-linked to bound proteins with UV light, and stringently purified PolyA+ RNAs for subsequent mass spectrometric identification of the *directly* bound proteins (a schematic of the general protocol is shown in **Figure 3.1**). Analysis the proteins identified reveals that nearly ~40% of the mRNA-bound proteome do not contain any canonical RNA-binding domains with many of the proteins identified being metabolic enzymes.<sup>124–127</sup> RNA binding to metabolic enzymes has long been known for several classical systems. For example, the iron-responsive element-binding protein 1 (IRP1)/aconitase 1 (ACO1) is an iron-sulfur

cluster binding enzyme involved in the citric acid cycle that also binds an RNA stem-loop motif in the apo-cofactor state.<sup>154,155</sup> The RNA motif is enriched in the mRNAs of many iron metabolism genes, providing a feedback mechanism between iron metabolism gene expression and an enzyme dependent on that cofactor for activity.<sup>156,157</sup> As the multiple naming conventions for IRP1/ACO1 suggest, the connection between many of these functions have historically been discovered by research groups in unconnected fields connecting the same protein in different biological contexts through sequencing. In large part, the significance of these global

crosslinking studies is in revealing pervasive, direct, interactions between proteins and RNAs that were wholly unsuspected for many of these proteins until now. Identification of these new interactions represents an opportune landscape to find additional mechanisms of cellular regulation and signaling. Further characterization of several of these identified metabolic and housekeeping genes have borne out RNA-regulatory relationships. For example, recent characterization of RNA binding by the housekeeping gene glyceraldehyde-3-phosphate dehydrogenase suggests competitive binding between RNA and its NAD<sup>+</sup> cofactor can result in a metabolic-state dependent translation repression of mRNAs bound by GAPDH.<sup>156,158</sup> In another enzyme system, protein kinase R, double-stranded RNA leads to dimerization and kinase mediated inhibition of translation in response to viral infection.<sup>159</sup> One particularly interesting and promising domain family identified is the cyclophilin-like domain (CLD).

These seminal studies identified CypA, CypB, CypG, and PPIL4 (see Table 1) as putative RNA binders and Castello *et al.*<sup>125</sup> additionally identified CypE (also known as Cyp33). A comparable study from the lab of Dr. Roy Parker likewise identified Cpr1, the *Saccharomyces cerevisiae* homologue of CypA, as an RNA-binding protein.<sup>127</sup> Subsequent work by Castello *et al.*<sup>160</sup> identified RNA crosslinked peptides for CypA, CypB, CypD, CypF, and PPIL4. Altogether, these works have implicated 7 of the 17 human cyclophilins as direct RNA-binding proteins. The identification of CypE and PPIL4 as RNA-binding proteins is perhaps unsurprising as both proteins contain an RRM or predicted RRM.<sup>161</sup> CypE is a two-domain protein with an N-terminal RNA-recognition motif domain (RRM) that binds an AU-rich sequence, though the precise consensus sequence remains unknown. However, direct interaction between CypA, CypB, and CypF with RNA is remarkable as CypA and CypB have been extensively studied and all three proteins consist solely of a single CLD, which is a domain that had not been previously observed to bind RNA (Castello 2012/2016, Baltz). Notably, the structurally unrelated FKBP family of peptidyl prolyl isomerases were identified as direct RNA-binding proteins by these

same studies, suggesting the evolution of widespread RNA-dependent functions of peptidyl prolyl isomerases.

Recent RNase dependent sucrose gradient migration experiments corroborate the connection of cyclophilins to RNA biology. Currently unpublished work by the Diederichs Lab (data available at [r-deep.dkfz.de](http://r-deep.dkfz.de))<sup>162</sup> has additionally implicated CypC, PPWD1, and CWC27 as cyclophilins bound in RNA-dependent complexes. As core components of the spliceosome, the identification of PPWD1 and CWC27 as forming RNA-dependent complexes is expected, but the connection of CypC (comprised solely of a single CLD) to complexes involving RNA has not been previously reported.

Moreover, other work has shown direct interactions between CLDs and RNA. Both human CypA and Cpr1 inhibit the viral replication of Tomato bushy stunt tombuvirus through direct interaction with the viral RNA,<sup>163</sup> and *Piriformospora indica* CypA (PiCypA) also directly binds RNA<sup>164</sup> but native RNA ligands for these proteins remain unknown. In addition, CypB is the fortuitous target of an RNA-SELEX aptamer developed to identify biomarkers of pancreatic cancer, with an estimated low nanomolar affinity.<sup>165</sup> The sum of these data, alongside the biological importance cyclophilins and the wealth of structural data available for the cyclophilin family (14 of the 17 human proteins have high-resolution structures of the cyclophilin domain,<sup>166</sup> including 9 of the 10 implicated as RNA binding), points towards the CLD domain representing a unique and important target for further characterization of its non-canonical RNA-binding activity. Moreover, their involvement in RNA processing provides a rich space in which binding native RNAs may play mechanistic regulatory roles.

### **3.2 - Cyclophilins are a family of key biological regulatory proteins**

Found in all domains of life, cyclophilins are a family of proteins sharing a common domain of approximately 109 amino acids.<sup>166</sup> So named due to their binding activity towards the immunosuppressant drug cyclosporin A, most CLD family members exhibit peptidyl-proline

isomerase activity,<sup>166</sup> which catalyzes the interconversion of proline between the *cis* and *trans* conformations. For many proteins, proline isomerization is thought to be the rate-limiting step in folding, as the uncatalyzed reaction occurs on the order of seconds, resulting in the CLD family isomerases playing important roles as protein chaperones.<sup>163,166–168</sup> Furthermore, as modulators of protein structure, these proteins also serve regulatory roles in a number of cell signaling pathways by altering conformational states of specific targets.<sup>169</sup> Thus far, 17 cyclophilins have been identified in the human genome,<sup>166</sup> with evidence that eight are essential in at least one cell type,<sup>170,171</sup> but the functions of most members of the family and their targets remain unknown.<sup>166</sup>

Despite their namesake, the discovery of cyclophilins as the targets of the immunosuppressant cyclosporin A (CspA) was initially misleading towards elucidating cyclophilin native functions. CspA forms a complex with CypA and this CspA-CypA complex binds to calcineurin, inhibiting its phosphatase activity.<sup>172</sup> In humans this serendipitous complex is responsible for the subsequent inhibition of T-cell activation that results in the useful pharmacological immunosuppressive properties of CspA. The fungal scarab beetle pathogen that produces CspA likely benefits from its insecticidal and antifungal properties.<sup>173–175</sup> The mechanisms underlying the insecticidal properties are unclear, but the ability to form the CspA-CypA-calcineurin complex is conserved in *S. cerevisiae* and other fungi is likely related its antifungal activity.<sup>176,177,175</sup> However, in terms of the native functions of cyclophilins, CypA does *not* regulate the activity of calcineurin in the absence of CspA, limiting the implications of this gain of function complex.<sup>178</sup> However, CspA broadly inhibit the peptyl prolyl isomerase activity of most cyclophilins and has been a useful tool for characterizing the isomerase activity of cyclophilins.<sup>166,178,179</sup> Moreover, several cyclophilins do play major roles in inflammation and infection independently of calcineurin, providing an impetus to develop non-immunosuppressive inhibitors of cyclophilins. .<sup>169,180–184</sup> Though the precise nature of their functions remains elusive,

many cyclophilins have been implicated at many levels of RNA biology ranging from gene expression, chromatin remodeling, and RNA processing as summarized in **Table 3.1**.

Cyclophilin	Localization <sup>183</sup>	Size <sup>183</sup>	Cyclosporin A Binding, K <sub>D</sub> <sup>166</sup>	RNA-Binding?	Essential Gene <sup>170,171</sup>	Cellular Roles <sup>183</sup>
CypA (PPIA)	Cytoplasm; Nucleus, Secreted	18 kDa	Y, 6.8 nM	B, C, P	N	Inflammation; tumor progression <sup>180,185,186</sup>
CypB (PPIB)	ER; Secreted; Cell surface	20 kDa	Y, 8.4 nM	B, C, P, D	Y	Secretory pathway; inflammation <sup>187-189</sup>
CypC (PPIC)	Cytoplasm; ER; Secreted	33 kDa	Y, 7.7 nM	D	N	Circulating tumor cell survival <sup>190,191</sup>
CypD [Cyp40] (PPID)	Cytoplasm	41 kDa	Y, 61 nM	P	N	Hsp90 chaperone complex <sup>192</sup>
CypE [Cyp33] (PPIE)	Nucleus	33 kDa	Y, 6.9 nM	C, D, K	Y	mRNA processing; Chromatin remodeling <sup>193-196</sup>
CypF [CypP3/CypD] (PPIF)	Mitochondria	22 kDa	Y, 6.7 nM	P	N	Mitochondrial permeability <sup>197,198</sup>
CypG (PPIG)	Nucleus	88 kDa	Y, 51 nM	B, C, D, K	N	Splicing; interaction with RNA pol II <sup>199,200</sup>
CypH (PPIH)	Nucleus; Cytoplasm	19 kDa	Y, 160 nM		Y	mRNA processing; splicing <sup>201-203</sup>
CypL1 (PPIL1)	Nucleus	18 kDa	Y, 9.8 nM		Y	mRNA processing <sup>204-206</sup>
CypL60 (PPIL2)	Nucleus; Golgi	59 kDa	n.d.		Y	Cell surface expression of CD147 <sup>207,208</sup>
CypJ (PPIL3)	Nucleus	18 kDa	n.d.		N	mRNA processing <sup>209</sup>
PPIL4 [Cyp57] (PPIL4)	Nucleus	57 kDa	n.d.	B, C, P, D, K	Y	
PPIL6 (PPIL6)	Unknown	35 kDa	n.d.		N	
CypNK (NKTR)	Nucleus	150 kDa	Y, 488 nM		N	Tumor recognition in NK cells <sup>210,211</sup>
RanBP2 [Nup358] (RANBP2)	Nucleus	358 kDa	n.d.		Y	Nuclear pore complex <sup>212,213</sup>
PPWD1 (PPWD1)	Nucleus	73 kDa	Y, 168 nM	D	Y	mRNA processing <sup>193</sup>
SDCCAG-10 (CWC27)	Nucleus	54 kDa	n.d.	D	N	

**Table 3.1 Summary of Human Cyclophilins.** Y – indicate yes; N – indicates No; n.d – indicates not determined. The RNA-binding column indicates in which studies the cyclophilin was implicated as RNA binding. B refers to Batiz et al. 2012, C refers to Castello et al. 2012, P refers to Castello et al. 2016, K refers to Kwon et al. 2013, D refers to the currently unpublished Diederichs Lab database (r-deep.dkfz.de)

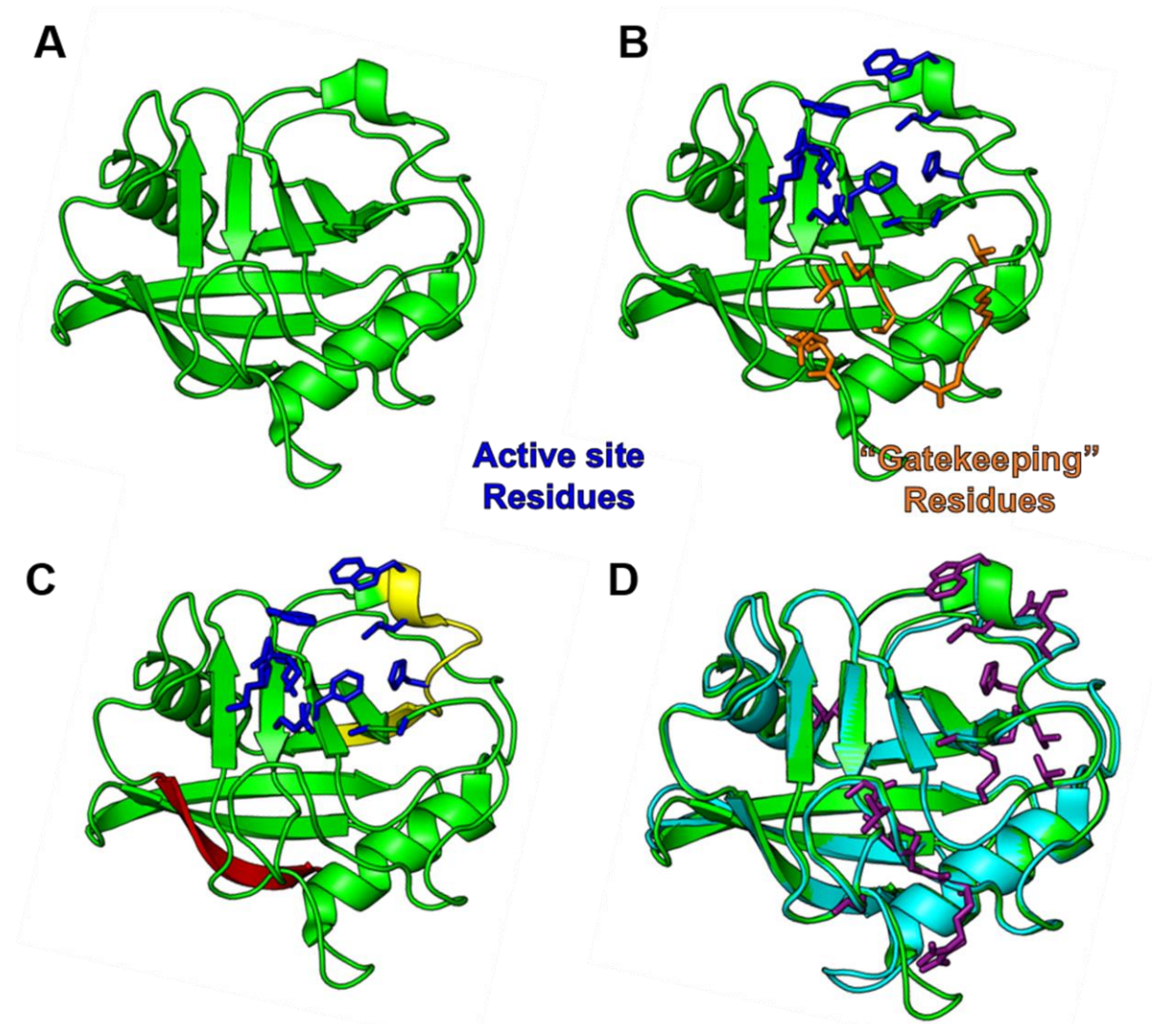


### 3.3 – Possible Mechanisms of Cyclophilin RNA-binding

#### 3.3.1 – RNA-binding cyclophilins share features suggestive of an RNA-binding surface

The most well characterized cyclophilin is CypA. As the smallest human cyclophilin, this protein exemplifies the typical cyclophilin domain architecture. The cyclophilin structural fold is comprised of 8 anti-parallel beta-sheets and 2 alpha helices organized into a beta-barrel structure anchored together with a hydrophobic core (structure of CypA shown in **Figure 3.2A**).<sup>214</sup> The active site residues form a binding pocket comprised of residues from  $\beta$ -strands 3, 4, 6 and the long loops nearby (**Figure 3.2B**).<sup>178,214</sup> Flanking the active site are the so called “gatekeeping” residues which show the greatest sequence variation between CLDs and have proposed to be involved in the substrate specificity of the different paralogs (**Figure 3.2B**).<sup>166</sup>

Several biochemical and structural studies have suggested features of the RNA binding of CLDs. Castello et al. 2016<sup>160</sup> identified several CLD peptides crosslinked to RNA. The CypA crosslinked peptide maps onto the structure directly adjacent to the active site, even containing several active site residues (**Figure 3.2C**). More distant from the active site, the conserved peptide sequence found crosslinked in CypF maps to the C-terminus which is on the surface of the protein distal to the active site. Likewise, the peptide crosslink for CypB also maps to the C-terminus – although this sequence is extended relative to other CLDs, and is involved in heparin binding by CypB, described in a section below. The surface residues involved in RNA binding by *Pi*CypA were structurally mapped via NMR chemical shifts changes upon addition of RNA (human CypA conserved residues shown in **Figure 3.2D**).<sup>164</sup> Intriguingly, many of the residues showing significant changes map to the “gatekeeping” residues and several active site residues.

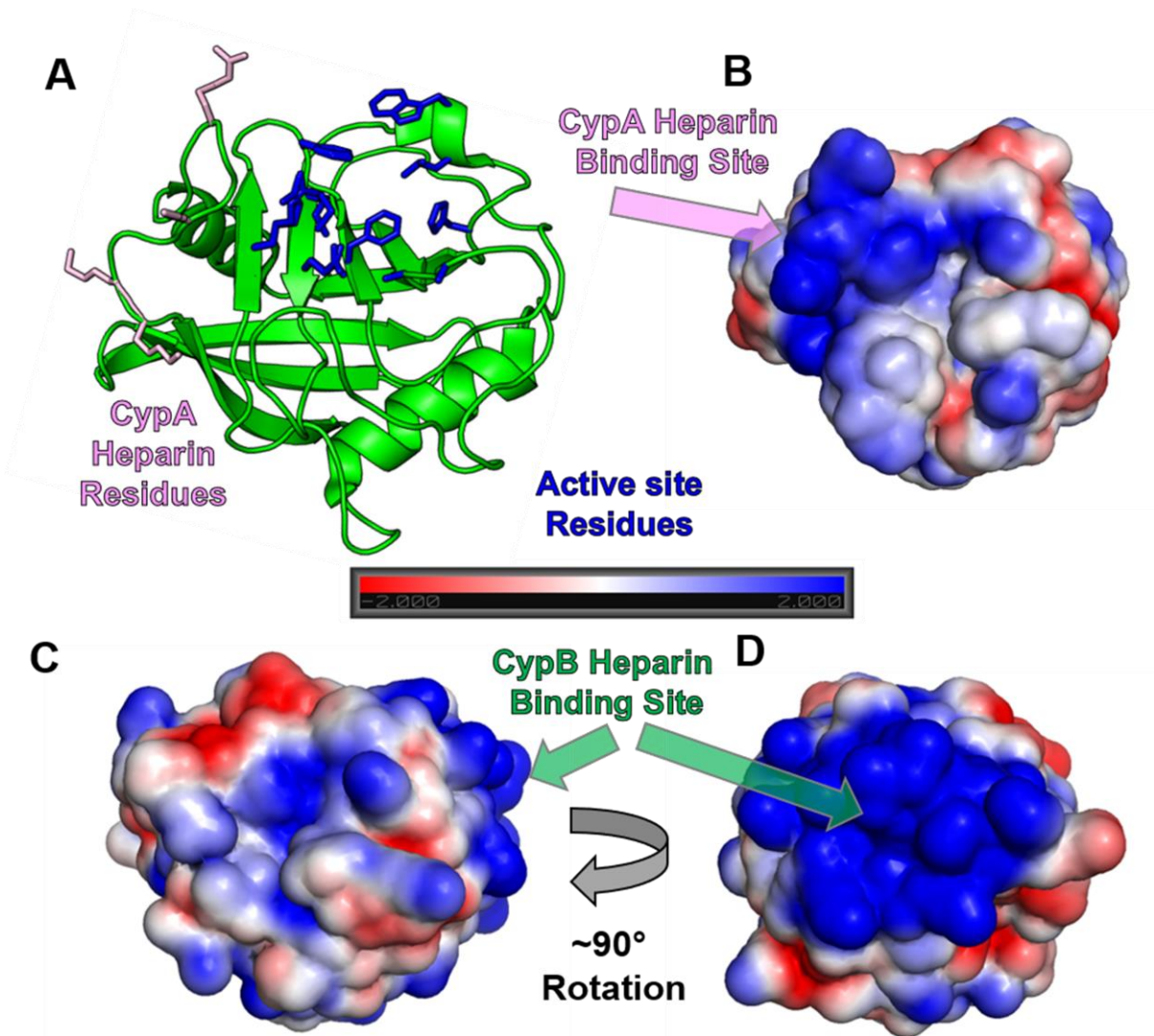


**Figure 3.2 Structural Features of Cyclophilins Mapped onto CypA (3K0M).** **A)** Cartoon structure of CypA is shown. **B)** CypA with active site residues in blue and “gatekeeping” residues in orange. **C)** Crosslinked peptide sequence mapped onto the structure of CypA (yellow CypA-peptide, red CypF-peptide on conserved CypA residues). **D)** Overlay of the crystal structures of PiCypA (3K0N) and CypA with conserved residues implicated in PiCypA RNA binding shown in purple.

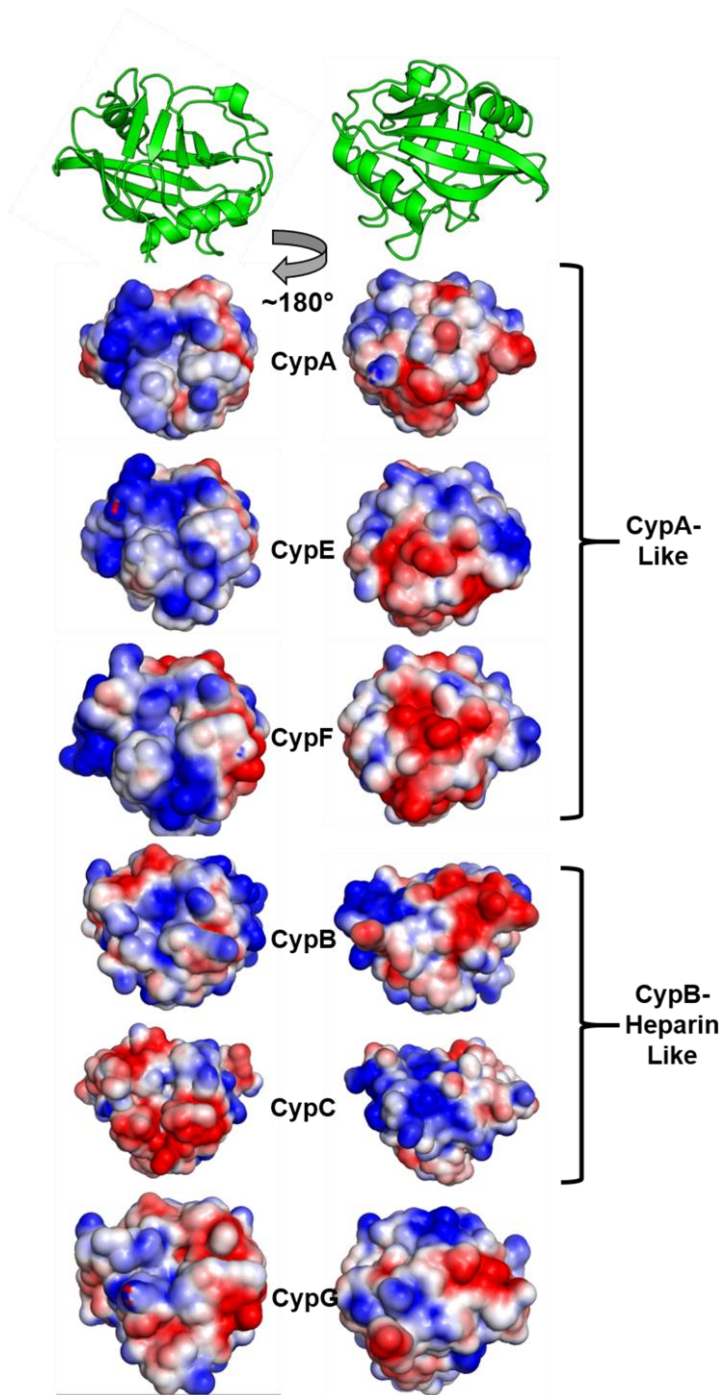
### 3.3.2 – Heparin binding Interface of CypA and CypB

Insights into nucleic acid binding may be obtained from examination of the binding of CLDs to another negatively charged oligosaccharide. Both CypA and CypB bind to heparin and this interaction is vital to the infectivity of HIV particles as the interaction serves to anchor the viral particle to the cell surface.<sup>215,216</sup> The interaction with CypA is dependent on four positively charged residues R148, K151, K154, and K155 (**Figure 3.3A/B**)<sup>215</sup> which are in a loop adjacent

to the active site residues K55. The characterized heparin binding interface for CypB is much more extensive and comprises both the N and C-terminal  $\beta$ -strands as well as other structurally close positive amino acids (**Figure 3.3C/D**).<sup>216,217</sup> This interface has been proposed to mediate several unique activities of CypB such as being a more potent agonist of chemotaxis as well as triggering the adhesion of T-lymphocytes to fibronectin.<sup>188,189,218</sup> While the binding site for CypB is unique to it, CypB, CypE, and CypF conserve the charge seen in that loop for CypA. As



**Figure 3.3 Heparin Binding Residues of CypA and CypB.** **A)** Heparin binding residues highlighted on the CypA structure (3K0M) in pink. **B)** Electrostatic surface of CypA in same orientation as (A) **C)** Electrostatic surface of CypB (3ICH) shown in the same domain orientation as CypA in panel (A). **D)** Rotated electrostatic surface of CypB highlight the basic surface patch involved in heparin binding. Electrostatic surface potentials were calculated using the Adaptive Poisson-Boltzmann Solver plugin for PyMOL.



**Figure 3.4 Electrostatic Surfaces of Select Cyclophilins Implicated as RNA Binding.** Domain orientation of each protein is the same as the orientation of the cartoon structure of CypA shown at the top of the column. Electrostatic surface potentials were calculated using the Adaptive Poisson-Boltzmann Solver plugin for PyMOL.

heparin and RNA are both negatively charged oligosaccharides, either of these sites implicated in heparin binding site presents a site for potential interaction for RNA.

### 3.3.3 – Surface Residue Variation and Charge Distribution

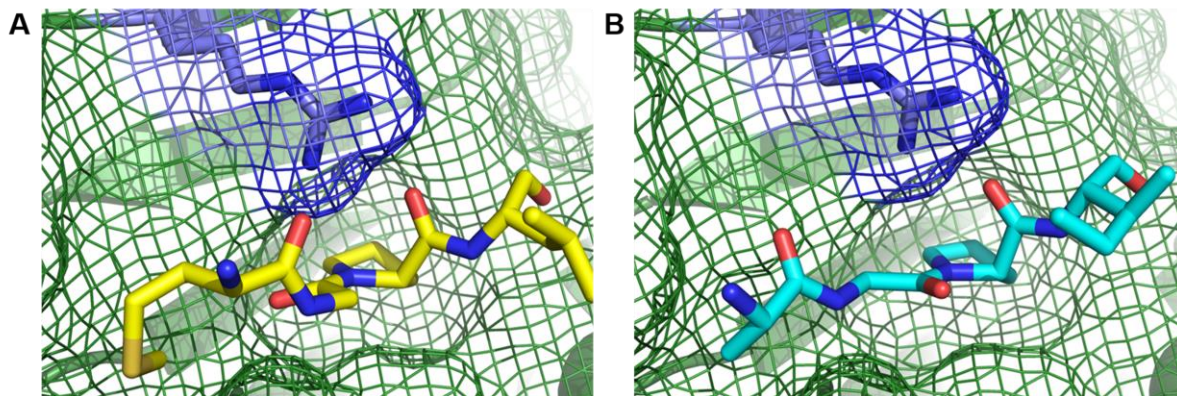
Despite the 14 solved structures of human CLDs aligning with root mean square distances of 0.4 Å to 1.0 Å and sequence identity ranging from 61% to 86%,<sup>166</sup> the sequence divergence of surface residues and charge localization suggests if RNA-binding activity is conserved among other CLDs, then the features of binding may differ. For example, the large positively charged surface in CypB involved in heparin binding that crosslinks to RNA is distinct from the heparin binding residues and crosslinking peptide for CypA.<sup>160</sup> However, this

extends to other CLDs as the “gatekeeping” residues show significant variation between CLDs (**Figure 3.4**).<sup>166</sup> For instance, the “gatekeeping” residues implicated in *Pi*CypA RNA binding are primarily acidic for CypG, CypC, PPWD1, and SDCCAG-10 with basic surface patches elsewhere on the proteins. The latter three proteins are implicated as being present in RNA-dependent complexes by sucrose-gradient experiments,<sup>162</sup> suggesting they may not directly bind RNA. Moreover, the CLD of CypG is one domain in an 88 kDa protein which also includes an Arg/Ser-rich domain,<sup>199,200</sup> suggesting CypG might not bind RNA through the CLD. However, these differences raise the possibility that RNA may bind at alternative surfaces of the CLD and may manifest in alternative RNA specificities. There appears to be three sites where these basic patches are common – the CypB heparin binding site, near the CypA heparin binding residues, and among the “gatekeeping” residues.

### **3.4 – Cyclophilin Enzymatic Dynamics and the Potential for Allosteric Regulation by RNA**

Due to its clinical importance, ease of structural characterization, and the advantage that the substrate of CypA is not consumed during enzymatic activity, CypA has served as a model system for enzyme dynamics during catalysis.<sup>181,214,219</sup> Using the powerful combination of protein mutants, millisecond dynamics extracted from NMR experiments, and shorter timescales dynamics from molecular simulations– the CypA catalyzed interconversion of *cis-trans* prolyl conformations in several peptides has been characterized extensively. As is typical of enzymes, the active site of CypA stabilizes the transition state – thereby lowering the activation energy necessary for the chemical step of enzyme catalysis.<sup>219</sup> This is accomplished through a hydrophobic pocket surrounding the proline residue with some hydrogen bonds to the peptide backbone. The double-bond character of the prolyl-peptide bond is reduced by the arginine residue that hydrogen bonds with the proline carbonyl oxygen (R55 in CypA). A number of other residues at the active site are involved in hydrogen bonding with *cis*, *trans*, or transition state within the +1-residue backbone. Mutations that disrupt these hydrogen bonds or decrease the

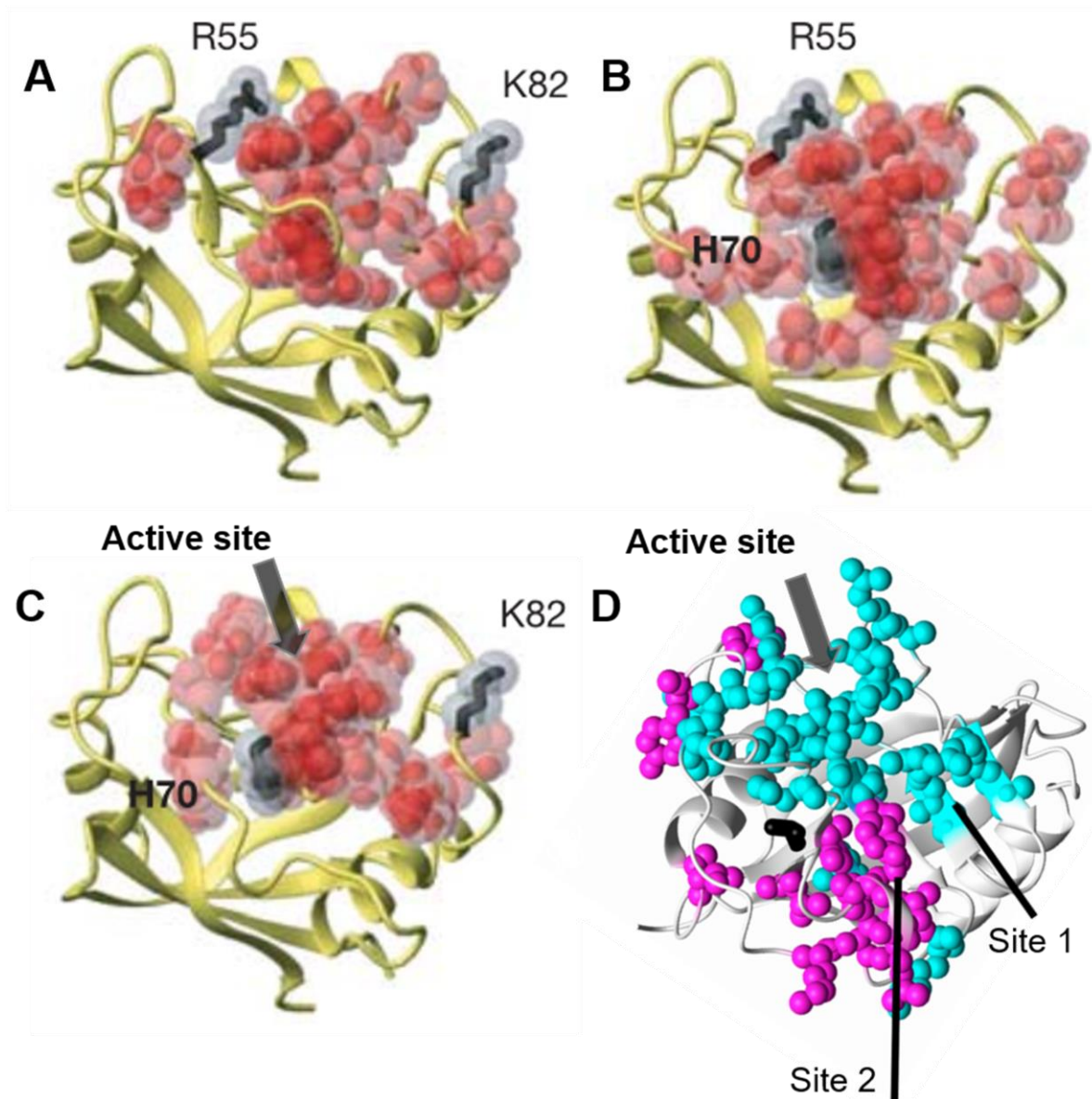
hydrophobicity of the active site result in catalytic defects (shown in blue in **Figure 3.2B**).<sup>178,219,220</sup> The *cis* and *trans* bound conformations of active site have been captured with various peptides (with the substrate conformations shown in **Figure 3.5A/B**)<sup>220</sup>



**Figure 3.5 Substrate Conformations in the Active Site. A)** *cis*-substrate (yellow) bound to the CypA active site (1M9D) Arg55 highlighted in blue with the CypA surface shown as a green mesh. **B)** *trans*-substrate (cyan) bound to CypA active site (1M9C)

### 3.4.1 – CypA Conformational Dynamics Reveal Two Allosteric Sites

Careful study of the native protein in bound and free states reveal the dynamics are largely similar, demonstrating that the enzyme intrinsically samples the conformations needed by catalysis with the substrate harvesting the energy of this movement to bind, transition through the chemical step, and to release.<sup>221</sup> Consistent with this, mutations distant to the active site of CypA can have large impacts on enzymatic activity without affecting the chemical step of the isomerization due to disruptions in conformational dynamics that affect other enzymatic steps.<sup>214,222</sup> Analysis of these mutations in CypA reveal two dynamically distinct allosteric sites within the enzyme.<sup>214,222</sup> The first site, near the active site, was initially revealed by a network of common chemical shift changes caused by several mutations (**Figure 3.6A-C**).<sup>214</sup> Subsequent NMR relaxation experiments with other mutants in the presence and absence of substrate binding support the coupled dynamics of this allosteric site (**Figure 3.6D**).<sup>222</sup> In addition, these experiments reveal another allosteric site dynamically distinct from the first.



**Figure 3.6 Independently Coupled Dynamics of Two Sites Reveal Possible Allosteric Sites.** **A-C)** Common differential chemical shifts upon residue mutation (i.e. mutation of R55 and K82 result in similar chemical shift changes for the residues highlighted in red) with the mutated residues shaded in black. Adapted from Fraser et al. **D)** NMR relaxation experiments studying the effects of mutants and substrate binding recapitulate the allosteric site observed in A-C (shown in cyan) and find another dynamically coupled site (shown in magenta) Adapted from Dochi et al.

The close proximity and overlap of these allosteric sites with the potential RNA binding surfaces implicated by current data suggest that binding of RNA or heparin is likely to alter enzyme activity. In fact, RNA interaction with several multi-domain cyclophilins has been

observed to modulate activity both ways. CypE enzyme activity *increases* when RNA is bound to the full-length protein.<sup>195</sup> In contrast, another RRM- and CLD-containing protein from *Arabidopsis thaliana*, AtCyp59, is *inhibited* upon RNA binding to the full-length protein.<sup>223</sup> At present, it is unclear how RNA does this – as it could influence conformational dynamics or perhaps occlude substrate binding or release. Characterizing how RNA influences activity in either direction will be important for understanding the function of RNA in CLD biology.

### **3.5 – The Known Functions of the Putative RNA-Binding Cyclophilins**

#### **3.5.1 – The myriad, Wide Ranging Functions of CypA**

CypA has been characterized to play roles in a myriad of cellular and extracellular processes such as but not limited to cellular trafficking, cell signaling, differentiation, gene expression, and protein folding.<sup>167,168,183</sup> Moreover, CypA is widely localized to the nucleus, cytoplasm, and is even secreted.<sup>183</sup> In the nucleus, CypA regulates and interacts with various transcription factors such as YY1 and Zpr1.<sup>224,225</sup> One notable regulatory interaction of CypA and several other nuclear cyclophilins is through their interaction with the circadian rhythm protein BMAL1 which acts as a transcription regulatory hub.<sup>226</sup> A key Trp-Pro bond in BMAL1 contributes to two conformational states that interact with a different set of transcriptional activators and repressors. This interaction and the inhibition of the PPIase activity of cyclophilins by cyclosporin A may explain why patients taking cyclosporin A experience lengthened circadian rhythms.<sup>226</sup> This ability to modulate other intracellular processes through conformational changes extends to kinase signaling molecules such as the tyrosine kinase Itk which is downregulated by CypA.<sup>169,227,228</sup> Extracellular CypA mediates potent pro-inflammatory responses by stimulating the pro-inflammatory signals MMP-2, MMP-9, and Interleukin-8 and exerts chemotactic activity for neutrophils, monocytes, and T-cells.<sup>183,185</sup> These roles contribute to the joint inflammation and cartilage degradation seen in rheumatoid arthritis and inflammatory lung diseases.<sup>207,229–231</sup> CypA also plays vital roles in maturation of many different viruses



including HIV, HPV, Hep B, Hep C, measles, and many others<sup>183,184,207,215,232,233</sup> Likely as a result of its roles in cell-to-cell signaling, CypA also has enhanced expression in malignancy.<sup>234,235</sup> Targeting these PPlase mediated activities of CypA has substantial therapeutic value, with several non-immunosuppressive derivatives of cyclosporin A and other unrelated fungal toxins currently in development. The most promising, Alisporivir, is being pursued as a therapeutic in the treatment of Hep C.<sup>183,236,237</sup>

### **3.5.2 - Functions of Cpr1**

Cpr1 is the budding yeast homologue of CypA and shares 65% sequence identity with its human counterpart. Like CypA, Cpr1 also regulates a broad range of biological processes and is present in both the cytoplasm and nucleus.<sup>238</sup> Several known Cpr1 interactions include modulation of the histone-deacetylase complexes Sin3-Rpd3 and Set3.<sup>238–240</sup> Like human CypA, Cpr1 bound to CsA also forms a non-native complex with the yeast calcineurin homolog causing a recovery defect following growth arrest. Deletion of the cyclophilin homologue in yeast (Cpr1) leads to viable yeast with a range of phenotypes consistent with its multiple roles as a cellular regulator, including sensitivity to stressful growth conditions, susceptibility to a range of chemicals, and, interestingly, increased accumulation of RNAs.<sup>239,241,242</sup>

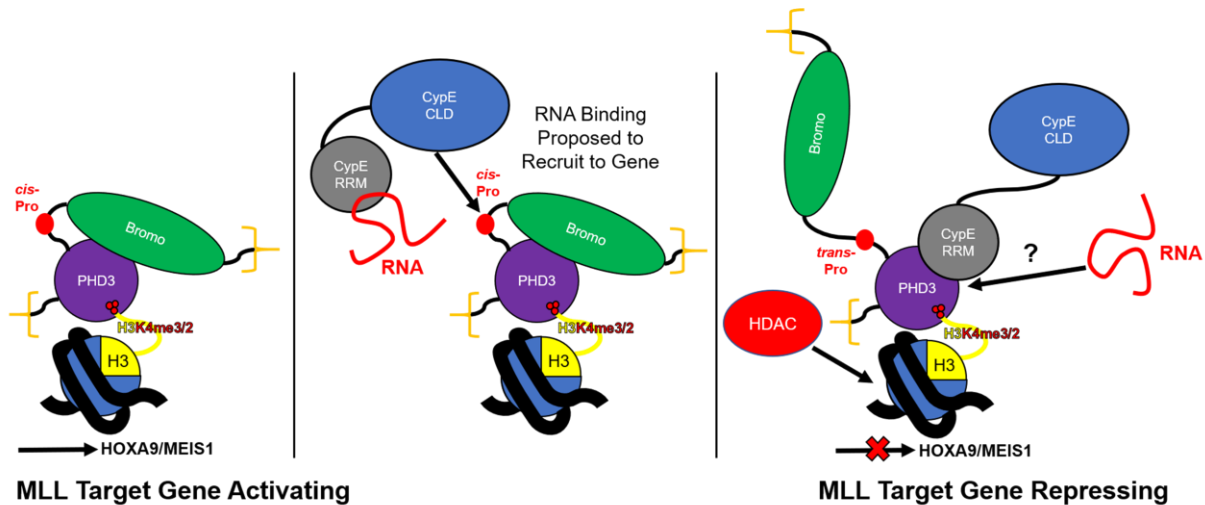
### **3.5.3 – Functions of CypB**

Like most cyclophilins, CypB has strong structural resemblance to CypA. The major differences between the two proteins are the addition of residues to the spatially close N- and C-termini and in two loop regions.<sup>214,217</sup> The additional residues on the N-terminus of CypB encode an ER signal peptide sequence that results in ER localization and CypB secretion.<sup>243</sup> In the ER, CypB plays important roles protein folding. In particular, deletions or mutations in CypB causes defects in the export of procollagen and can result in severe osteogenesis imperfecta, a connective tissue disorder.<sup>244</sup> Related to its extracellular functions, the additional residues on

the C-terminus also encode a heparin binding consensus sequence that results in several unique extracellular CypB activities compared to CypA. For instance, CypB is a more potent agonist of chemotaxis and only CypB can trigger T lymphocytes adhesion to fibronectin.<sup>188,218</sup> However, CypA and CypB functions appear to highly overlap with respect to many of their extracellular roles in inflammation and their roles in viral lifecycles. For example, CypB can substitute for CypA in the mature HIV viral particles.<sup>215,216</sup>

### 3.5.4 – Functions of CypE

CypE demonstrates an example in which RNA binding to a CLD-containing protein is proposed to play an important mechanistic role. CypE isomerase activity is enhanced in the presence of mRNA.<sup>195</sup> Through the RRM domain, CypE binds an AU-rich RNA sequence.<sup>195,245</sup> This RRM domain also specifically interacts with the third PHD (PHD3) finger domain of the mixed lineage leukemia (MLL) proto-oncoprotein with a 2  $\mu$ M affinity in a manner suggesting it is mutually exclusive to RNA binding.<sup>195,245,246</sup> As CypE PPIase activity is required to alter the conformation of MLL to reveal the occluded RRM-PHD3 binding,<sup>246</sup> the increase in PPIase activity upon RNA binding may be one potential mechanism of regulation by RNA. The *cis-trans* isomerization of MLL also allows binding of histone deacetylase 1 to MLL, and is required *in vivo* for MLL-mediated epigenetic repression of the MEIS1 and HOXA9 target genes.<sup>195,245,246</sup> Notably, CypE is recruited to these loci independently of MLL. Extensive precedent for lncRNA regulation of epigenetic chromatin state through recruitment of chromatin modifying complexes,<sup>247</sup> suggests non-coding RNAs may be involved in recruiting CypE to these repressed genes. Moreover, RNAs may play additional mechanistic roles by mediating MLL binding through a mutually exclusive interaction with the RRM domain. Importantly, the precise consensus sequences bound by the CypE RRM has not been rigorously defined, limiting the identification of native RNAs that interact with CypE. Knowledge of this consensus sequence



**Figure 3.7 Schematic of CypE Proposed Mechanism of Gene Repression.**

and the extent to which the CLD participates in this binding interaction will help elucidate the possible *in vivo* RNA binding partners potentially involved in this pathway.

Based on its abundance in spliceosome complexes, CypE is also a core component of the spliceosome, forming part of the stable ribonucleoprotein core of catalytically active C-complex.<sup>248–250</sup> However, its role in splicing is poorly characterized.

### 3.5.5 – Functions of PPIL4/AtCyp59

Our functional knowledge of PPIL4 is scant beyond that it is found in the B-complex of the spliceosome.<sup>250</sup> The protein is 57 kDa and comprised of a cyclophilin-like domain, an RRM, a bipartite nuclear localization sequence, and a lysine rich domain.<sup>204</sup> The cyclophilin-like domain has relatively low sequence identity for other cyclophilins (36% identity with CypA) and mutations in nearly half of the 13 residues essential for interaction with cyclosporin A.<sup>166,204</sup> However, homologs of PPIL4 are present in a wide variety of other organisms including *Arabidopsis thaliana* (44% identity) in which it is better characterized and known as AtCyp59.

The RRM of AtCyp59 binds to a consensus motif best described by the sequence GYNRCCR, which was determined by genomic SELEX.<sup>223</sup> Like CypE, AtCyp59 exhibits RNA-

dependent isomerase activity, however, unlike CypE, addition of RNA containing the AtCyp59 consensus motif results in inhibition of PPIase activity.<sup>223</sup> Remarkably, the consensus sequence is found in 70% of all annotated transcripts in Arabidopsis with preferential enrichment in exons. In addition, AtCyp59 interacts with the proline-rich C-terminal domain of RNA polymerase II and splicing proteins.<sup>251</sup> Altogether, these suggest AtCyp59 may play a general role connecting RNA transcription with RNA processing. Moreover, this function appears to be at least partially conserved as the *S. pombe* Rct1 homolog also interacts with the CTD of Pol II and overexpression levels follow the same pattern of reduced phosphorylation of the CTD.<sup>252</sup>

### **3.5.6 – Functions of CypG**

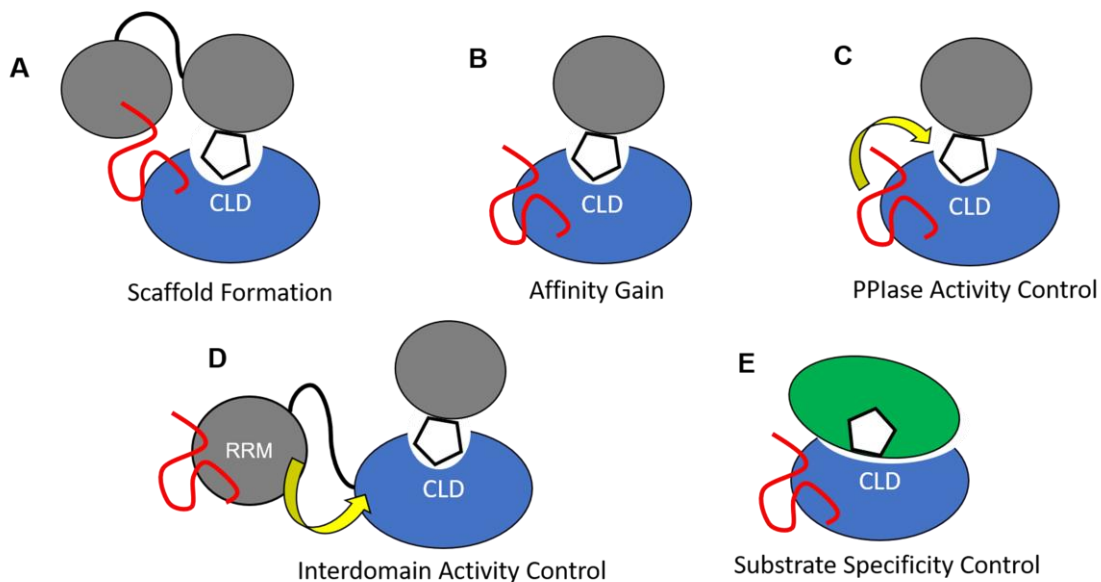
As an 88 kDa protein, CypG contains a cyclophilin domain, a Nopp140-like domain, and a serine/arginine-rich domain.<sup>199,253</sup> The serine/arginine-rich domain is necessary for CypG to interact with the CTD of Pol II.<sup>200</sup> As CypG associates with many splicing factors and is found in the catalytically active C-complex of the spliceosome<sup>250</sup> in addition to its enrichment in nuclear speckles,<sup>254</sup> it has been proposed as possible mediator between transcription and splicing.<sup>253</sup> Interestingly, the absence of CypG in natural killer cells and the similar domain architecture of NKTR has led to the proposal that NKTR is a natural killer cell specific paralog of CypG.<sup>211,253</sup>

### **3.6 – RNA may play mechanistic roles in cyclophilin-mediated regulatory activities**

As noted above, CLDs are involved in regulating a number of key biological pathways, including several associated with RNA processing. Anecdotal observations support the RNA proteome results by linking RNA-binding to CLD activity in several systems. For example, CypE binds an AU-rich RNA through its RNA-binding domain that also interacts with the PHD3 domain from the MLL proto-oncogene, the isomerase activity of CypE is modulated as a result of these interactions, which enhances the binding of histone deacetylase 1 to MLL.<sup>195,245,246</sup> A direct CLD-RNA interaction between CypA and Cpr1 and viral RNA has been shown to inhibit

the viral replication of Tomato bushy stunt tombuvirus.<sup>163</sup> Finally, a fungal cyclophilin, *Piriformospora indica* CypA (PiCypA), is known to directly bind RNA.<sup>164</sup> CypE enzyme activity *increases* when RNA is bound to the full-length protein.<sup>195</sup> In contrast, another RRM- and CLD-containing protein from *Arabidopsis thaliana*, AtCyp59, is *inhibited* upon RNA binding to the full-length protein.<sup>223</sup> The response to RNA binding could explain the lack of congruence between the *in vitro* peptide specificity and the known *in vivo* protein targets.<sup>166</sup>

The role of an RNA-binding function for CLDs is open to speculation. However, several proteins containing the FKBP domain, another major proline isomerase domain, have also been identified as RNA-binding,<sup>125,160,255</sup> suggesting RNA may play fundamental roles in the biological function of many peptidyl-prolyl isomerases. RNA could impact cyclophilin activities in several ways (**Figure 3.8**). RNA could serve as a scaffolding molecule, increasing the avidity of cyclophilins for substrates with weak affinity or regulating sub-cellular localization. Likewise, RNA binding could stabilize or destabilize substrate binding. The presence of allosteric sites on CypA also suggests RNA binding could control PPIase activity by altering conformational dynamics. For CypE and other multidomain cyclophilins, PPIase activity could be modulated by



**Figure 3.8 Schematic of Possible Mechanisms of RNA Regulation of Cyclophilins**

RNA binding in another protein domain. RNA binding to “gatekeeping” residues could alter substrate specificity and explain the disconnect between *in vitro* and *in vivo* peptide specificity. Alternatively, cyclophilins may regulate RNA through their interactions by activating or repression translation or altering the trafficking or stability of RNAs. However, cyclophilins interact with RNA, characterizing the RNA-specificity of these proteins is critical to understanding how RNA plays a role in cyclophilin biology and may provide novel insights into the biology of the less characterized cyclophilins.

## **Chapter 4 – Optimization of SELEX Protocol with MS2 Coat Protein**

### **4.0 – Chapter Overview:**

This chapter discusses my validation of our selection protocol and bioinformatics pipeline for general use in our lab and more importantly as a validation for our use of this technique to characterize RNA interactions with cyclophilins. As an overview, I discuss the strengths and weakness of the SELEX approach, and how recent technological advances in sequencing have made this technique more powerful. By using the previously characterized MS2-coat protein as target, I compare the results of our selection protocol and high-throughput sequencing to the literature on MS2 RNA binding and discuss the bioinformatics pipelines we have tested with this dataset.

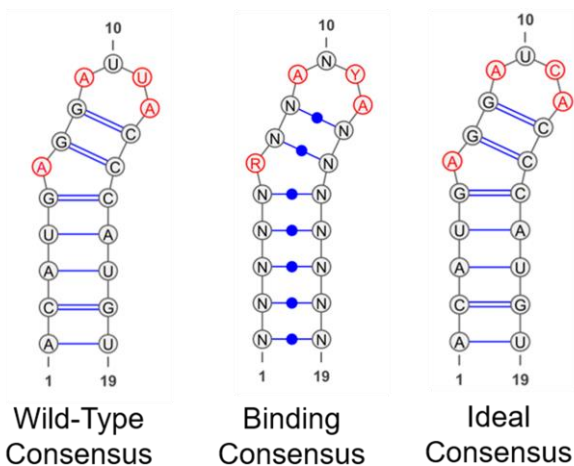
### **4.1 – Introduction**

Historically, SELEX has been a very powerful technique to determine the consensus binding sequence for RNA and DNA binding proteins and has successfully determined the consensus binding motifs for many nucleic acid binding proteins.<sup>256–261</sup> Adaptation of the principles of the technique to phage-display has further allowed for the characterization of many other types of interactions<sup>262</sup> and use of non-native nucleotides and other chemical modifications<sup>263</sup> have improved upon the ability of SELEX to produce highly specificity, high-affinity ligands for diagnostic purposes<sup>264–266</sup> as well as potentially therapeutic molecules.<sup>267,268</sup>

However, early experiments were hampered by limitations of old technologies – suffering from poor sequencing depth and having to over-select to compensate for this, resulting in the accumulation of selection biases as a result of PCR and reverse transcription artifacts. Now, with the power of recent developments in high-throughput sequencing, many of these weaknesses can be addressed through new strategies. Sequencing of every (and fewer rounds) of selection can now provide extremely deep insights into the emergence of aptamer sequences round by round, how motifs are enriched, and to characterize the biases inherent in the protocol.

For example, the enrichment of near consensus sequences in early rounds gives broader insights into a range of biologically relevant interactions beyond the tightest binding sequence. This timing works out well as concomitant advances in mass-spectroscopy have allowed for unprecedented insight into the proteome (see chapter 3 for details), and global studies have revealed a large number of novel RNA-binding proteins<sup>124–127</sup> for which SELEX is an ideal technique to characterize their RNA motif specificity.

With the goal of using RNA SELEX to characterize non-canonical RNA binding proteins, such as the cyclophilins described in the preceding chapter, as well as the possibility of identifying functional regions of non-coding RNA through genomic SELEX, we needed to develop and validate a selection protocol, high-throughput sequencing library design, and a bioinformatics pipeline for general use in our lab. We chose MS2 for the following reasons – it binds a small motif very tightly (**Figure 4.1**),<sup>260</sup> binding has been extensively characterized structurally and biochemically,<sup>269</sup> and the protein is easy expressed and purified. The sum of these features allows a clear and unambiguous metric by which to judge whether our selection protocol successfully enriches for tightly binding aptamers and test bioinformatics pipelines for



**Figure 4.1** MS2 coat protein Consensus Binding Motif. The consensus binding motif is shown for the MS2 coat protein, as derived from a genomic SELEX experiment and recapitulated by recently by ref

ease of implementation and capability.

To this end, we performed RNA-SELEX on MS2 coat protein over eight rounds of selection, deep sequenced the resulting libraries on an Illumina NextSeq instrument, and evaluated several published bioinformatic pipelines with our MS2 aptamer dataset. We found that at our sequencing depth and using the AptaSUITE pipeline,<sup>270</sup> we were able to find enrichment of the MS2 binding sites



as early as round 2, and unambiguously identify several clusters of unrelated sequences all containing the consensus sequence in round 4. Moreover, from all sequences we were able to extract a 9-nt motif comprising of the sequence specific portion the consensus binding sequence consistent with the previously characterized energetics of binding. As a result, our we have validated that our selection protocol can successfully enrich binding motifs for MS2 and have found that the AptaSUITE pipeline is both the easiest to implement and provides the most in depth analysis options of the bioinformatics pipelines we have tested.

## **4.2 – Methods**

Detailed protocols for all the methods briefly described here are included in Appendix B

### **4.2.1 – Protein Expression and Purification**

6xHis-maltose binding protein-MS2 fusion protein containing the MS2 V29/dIFG mutation reported to prevent oligomerization (generously gifted by Prof. Robert Batey; Addgene #67717) was transformed into BL21 (DE3) *E. coli* and selected on LB plates supplemented with kanamycin. Single colonies were then picked for a 40 mL 37 °C overnight growth with the same antibiotic selection. Using 10 mL of the overnight growths, 1L growths were inoculated and grown in 2L baffled flasks containing antibiotic at 37 °C and shaken at 180 rpm for 2-3 hrs to an O.D.<sub>600</sub> of 0.6-0.8 before being induced with 1 mM IPTG. After induction, the growth temperature was decreased to 18-20 °C and the cultures were harvested, pelleted by spinning at 15K RPMs in a Fiberlite F21-8 rotor (ThermoFisher), and frozen at -20 °C after 18-20 hrs of growth.

Frozen pellets were thawed in lysis buffer (40-50 mL final volume) supplemented with a Roche EDTA-free protease inhibitor tablet before being sonicated. Lysed cells were then spun at 15K RPM and the supernatant fraction was incubated with Ni-NTA beads equilibrated with lysis buffer for 0.5-1 hr. After 3 washes with lysis buffer, the captured protein fraction was eluted with lysis buffer supplemented with 350 mM imidazole in two 15-20 mL fractions. Eluted protein

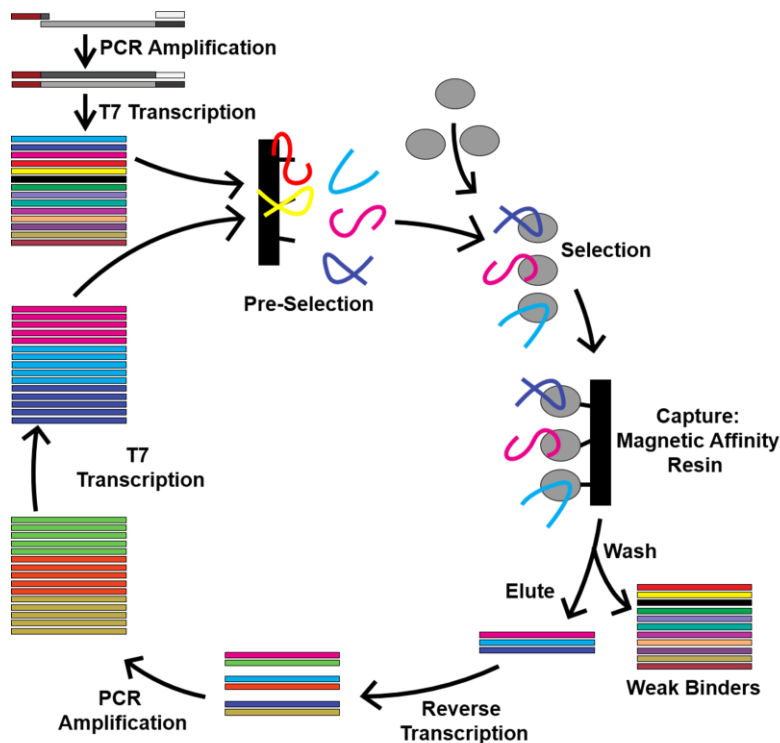
was then concentrated to ~1.5-2 mL in 10K MWCO Sartorius concentrators The Superdex G200 column (GE) and further purified with size-exclusion chromatography on a Akta FPLC. After elution fractions were combined and concentrated to ~800  $\mu$ M aliquoted, and flash frozen in liquid nitrogen for later use.

#### **4.2.2 – Library Binding through EMSA**

To quantify the binding affinity of the target proteins for RNA ligands, EMSAs were performed using radiolabeled RNA ligands produced by T7 *in vitro* transcription and purified protein. The 5' phosphate of transcribed RNA ligands were removed using calf intestinal phosphatase (CIP, NEB) and then 5' labeled with  $^{35}$ P using T4 polynucleotide kinase (PNK, NEB) and  $^{35}$ P- $\gamma$  ATP. Labeled ligand with a final concentration of 5 nM was added to 2-fold serial dilutions of the purified protein ranging from 200  $\mu$ M to 0 nM final concentration in SELEX buffer (defined in Chapter 5 and Appendix B) supplemented with 10% glycerol. Samples were loaded onto a 0.5X TBE 8% polyacrylamide gel and run at 200V at room temperature for 15-20 minutes. The gels were then dried and exposed on a phosphor screen and imaged on an Amersham Typhoon Imaging System. The resulting images were quantified in ImageQuant 5.0 and fit to the quadratic binding equation in Excel using Solver by minimizing the sum of the least squares difference between the data and fit (details in Appendix B)

#### **4.2.4 – SELEX**

A schematic diagram of the SELEX protocol used here is shown in **Figure 4.2**. Detailed descriptions of the protocols used for selection against MS2 are described in Chapter 5.2.5 under SELEX Experiment 2 as well as Appendix B.



**Figure 4.2 Schematic Diagram of the SELEX Protocol Used.**

#### 4.2.5 – Sanger sequencing

PCR primers identical to the SELEX primers with restriction enzyme sites appended to the ends BamHI for forward and Xho1 for reverse (sequence detailed in Appendix B), respectively, were used to amplify and clone round 6 sequences into a pET21a vector. Single colonies were then picked, grown overnight, and miniprepmed plasmid was then Sanger sequence at Genewiz, resulting in 6 unique aptamer sequences.

#### 4.2.6 – High-throughput sequencing

A description of the library preparation and sequencing protocol are described in Chapter 5 and in more detail in Appendix B.

## 4.2.7 – Bioinformatics Pipelines

Detailed descriptions of the scripts and analysis using the bioinformatics pipelines are available in Appendix C

### 4.2.7.1 – QIIME 1.9

QIIME is an open-source bioinformatics pipeline developed by the Robin Knight lab and others for microbiome analysis.<sup>271</sup> This software is typically used for characterizing the microbial populations found in various samples through genus-level variations found in the 16S ribosomal RNA. However, the pipeline parallels most of the necessary analysis for SELEX aptamer identification; it takes raw high-throughput sequencing data, demultiplexes and quality filters it, clusters based on sequence similarity, and provides a suite of phylogenetic and diversity analyses and visualizations. Thus, I used it for the analysis of the selected sequences. The biggest limitation is the normal QIIME 1.9 default distribution does not contain a script to separate sequences based on the clusters identified, but that script is available on the QIIME google user forums (and reproduced in Appendix C).

The scripts and all of the options used with this pipeline are fully described in Appendix C. Briefly summarized here, the FASTQ file outputs from the Illumina NextSeq run were demultiplexed based on the associated library barcodes, filtered for quality reads, and had the 3' constant regions/adaptor sequence subsequently trimmed to leave only the random region. Sequences for all rounds of SELEX at all conditions were then rank-sorted into a FASTA file, filtered for read abundance >5 reads, and then clustered against each other at 45% similarity (~23/50 matches) with the preference of using more abundant sequences as cluster seed sequences. Clusters comprising sequences less than 0.5% of the total fraction of sequences were then filtered out leaving 19 aptamer clusters. The seed sequence of each of these 19 clusters was then used to do clustering alignments of all sequences over all rounds of selection

to query the abundance of each cluster in each sample to monitor the emergence of the winning families.

#### **4.2.7.2 – FASTAptamer**

FASTAptamer is a bioinformatics pipeline designed for analysis of high-throughput sequencing data from SELEX experiments, going from quality processed sequencing files to identification of enriched aptamer sequences.<sup>272</sup> To test the suitability of FASTAptamer for a general SELEX pipeline, I ran our sequences from through the recommended pipeline. In the words of the Donald H. Burke lab at the University of Missouri, group who maintains and distributes it, FASTAptamer is a suite of perl scripts run through a command line terminal developed to “perform the simple tasks of counting, normalizing, ranking, and sorting the abundance of each unique sequence in a population, comparing sequence distributions for two populations, clustering sequences into sequence families based on Levenshtein edit distance, calculating fold-enrichment for all of the sequences across populations, and search degenerately for nucleotide sequence motifs.”<sup>272</sup> The scripts included in the distribution are FASTAptamer-Count, FASTAptamer-Compare, FASTAptamer-Cluster, FASTAptamer-Enrich, and FASTAptamer-Search. FASTAptamer-Count normalizes an input FASTQ file into an abundance-sorted FASTA file in which each sequence is given an identifier based on abundance-rank, number of reads, and reads normalized to all reads in units of reads per million. FASTAptamer-Compare generates a comparison of the abundance of shared sequences between two FASTAptamer-Count output FASTA files, useful for comparing rounds of SELEX to each other, for example. FASTAptamer-Cluster takes the FASTA output from FASTAptamer-Count to group similar sequences together to identify families of aptamers. FASTAptamer-Enrich calculates the enrichment for sequences in up to three input files which can be output files from Count or Cluster to show fold-enrichment for each individual sequence along with the associated sequence length, rank, read, normalized reads, the cluster information

if available, and which samples/files the sequence was found in. Lastly, FASTAptamer-Search reports the sequences containing up to two IUPAC-IUBMB formatted motifs and can output those sequences into a separate FASTA file.

The scripts and all of the options used with this pipeline are fully described in Appendix C. Briefly, using QIIME demultiplexed FASTQ files for each round of SELEX sequences were rank-sorted and normalized to total reads for each round. Sequences were then clustered against each other with a Levenshtein edit distance of 7 with higher abundance sequences forming the initial seeds of each new cluster and a minimum abundance of 25 reads per million. Alternatively, all sequences from all rounds were clustered against each other with the same criteria to identify any sequence clusters present in more than one condition.

#### **4.2.7.3 – AptaSUITE**

Like FASTAptamer, I also tested AptaSUITE for general use as our bioinformatics pipeline following SELEX sequencing. AptaSUITE is an open-source collection of software for the comprehensive analysis of HT-SELEX experiments developed in java, allowing for platform-independent usage.<sup>270</sup> Still under development by the National Center of Biotechnology Information of the NIH, AptaSUITE currently contains a number of previously developed software packages including AptaPLEX,<sup>273</sup> AptaSIM,<sup>270</sup> AptaCLUSTER,<sup>274</sup> AptaTRACE,<sup>275,276</sup> and AptaMUT.<sup>277</sup> Together, this pipeline allows for demultiplexing of barcoded sequencing data, clustering of selected sequences, prediction of aptamer secondary structure and motif identification, and analysis of “mutations” among related sequences within clusters. Moreover, the package contains a graphical user interface that provides several useful visualizations such as sequence motif logos, sequence abundance per round, and predicted secondary structure – among others.

The scripts and all of the options used with this pipeline are fully described in Appendix C. Briefly, FASTQ files demultiplexed and quality filtered from QIIME were appended with the 5'

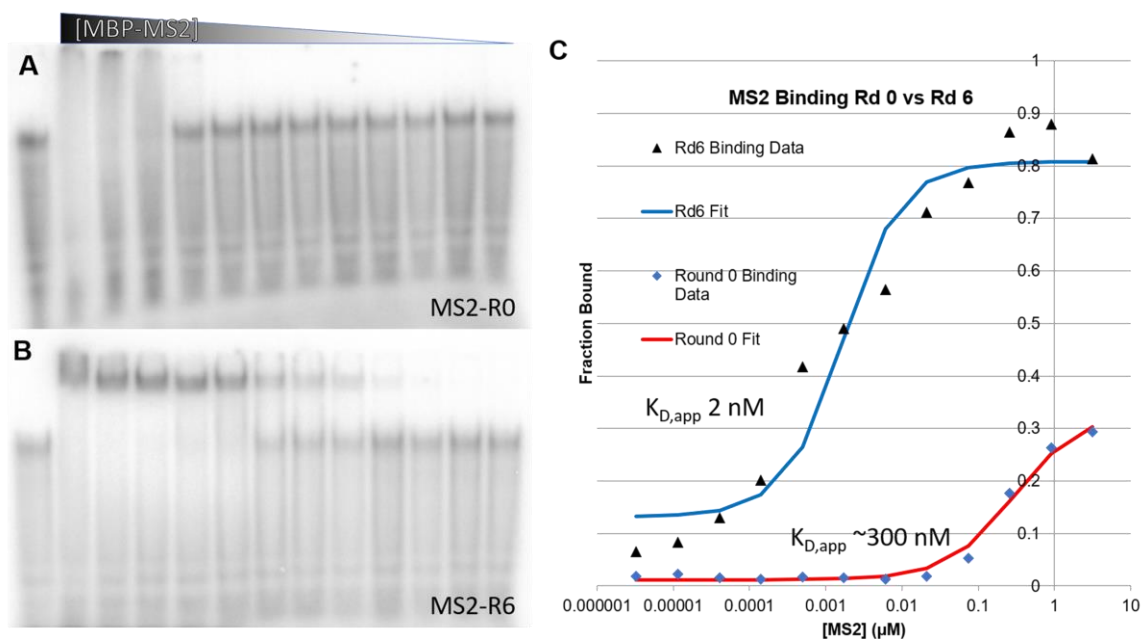
constant region sequence with dummy quality data. The full RNA sequence was included to facilitate more accurate secondary structure prediction despite the 5' constant region not being directly sequenced as the read primer was complementary to the 5' constant region and reads began at the first nucleotide of the random region. The AptaPLEX<sup>273</sup> feature of AptaSUITE is capable of demultiplexing as well, but the process in either case is computationally expensive and was not duplicated. Sequences were then clustered with AptaCLUSTER.<sup>274</sup> This program approximates the Levenshtein edit distance alignments used in the FASTAptamer-Cluster<sup>272</sup> through first filtering out sequences from each cluster beyond an upper edit distance prior to performing the more computationally expensive pairwise comparison of the remaining sequences. Like FASTAptamer-Cluster, the seed sequence for each cluster is preferentially the most abundant sequence not previously clustered. AptaTRACE predicts the secondary structure of each aptamer sequence and performs k-mer analysis on the sequence pool.<sup>270,275</sup> For example, for a k=6, each possible 6mer nucleotide sequence is counted and compared to the abundance of all 6-mers to identify particular 6-mer sequences that are enriched within the selected pools. The motifs are then put into the context of the secondary structure predictions to provide both a sequence and structural logo motif. AptaMUT<sup>277</sup> can characterize the “mutations” among clustered sequences and provides enrichment/depletion data between rounds of selection to give insight into history of the sequence pool.

## **4.3 – Results**

### **4.3.1 – Selected RNA pool binds tighter than the round 0 library**

To determine whether the selection protocol indeed selected for RNAs that bound more tightly to MS2, we performed electrophoretic mobility shift assays to assess the affinity of initial pool of RNA of RNA as well as various RNA pools after several rounds of selection. The initial library pool was found to bind RNA with a  $K_D$  of approximately 300 nM (**Figure 4.3**). MS2 binding to the RNA pool after 6 rounds of selection was at a  $K_D$  of approximately 2 nM (**Figure**

4.3B/C), indicating that the RNA pool was successfully enriched for RNA with a greater affinity for MS2 than the initial population and near the reported affinity for previous the tightest MS2 RNA aptamer which is 2-3 nM.<sup>269</sup>

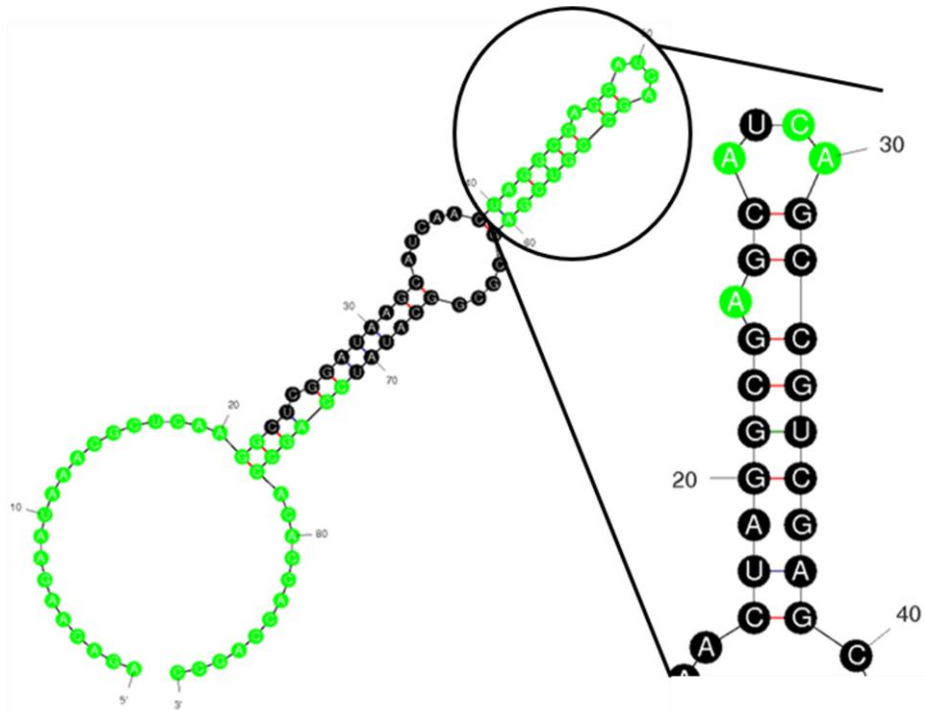


**Figure 4.3 MS2 Binding to the Initial and Selected Libraries. A)** EMSA showing MS2 binding to the Round 0 library. **B)** EMSA showing MS2 binding to the Round 6 library. **C)** Binding fit shown for MS2 binding to the Round 0 and 6 libraries.

#### 4.3.2 – Sanger sequencing of round 6 RNAs reveal 6 unique sequences containing MS2 binding sites

Prior to committing the cost of a high-throughput sequencing run, we wanted to assess whether our pool of RNAs were able to reveal any MS2 binding sites using the conventional method of cloning sequences and then submitting them for Sanger sequencing. Thus, we sequenced 6 of the “winning” sequences. The resulting 6 clones revealed unique sequences that contain the predicted MS2 consensus binding motif of a stem loop of with a loop sequence



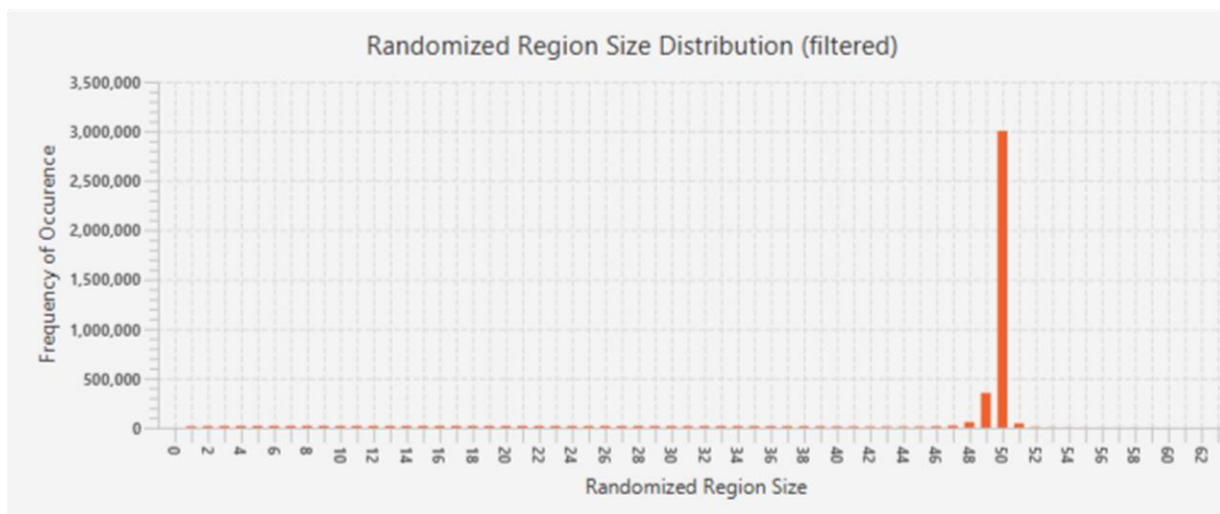


**Figure 4.4 Sanger Sequencing Reveals MS2 consensus binding site in Round 6.** Predicted secondary structure shown for one of the sequenced clones (mfold webserver). Constant regions are highlighted in green as well as the MS2 consensus binding site. On zoom-in, the bases with the greatest specificity for MS2 are highlighted.

of ANYA and a bulged purine within the helix (**Figure 4.1**). The predicted secondary structures for a representative sequence and the associated MS2 binding site is shown in **Figure 4.4**<sup>278</sup>

### 4.3.3 – High-Throughput Sequencing of 8 Rounds of SELEX Reveals Library Biases and Enrichment of Sequences

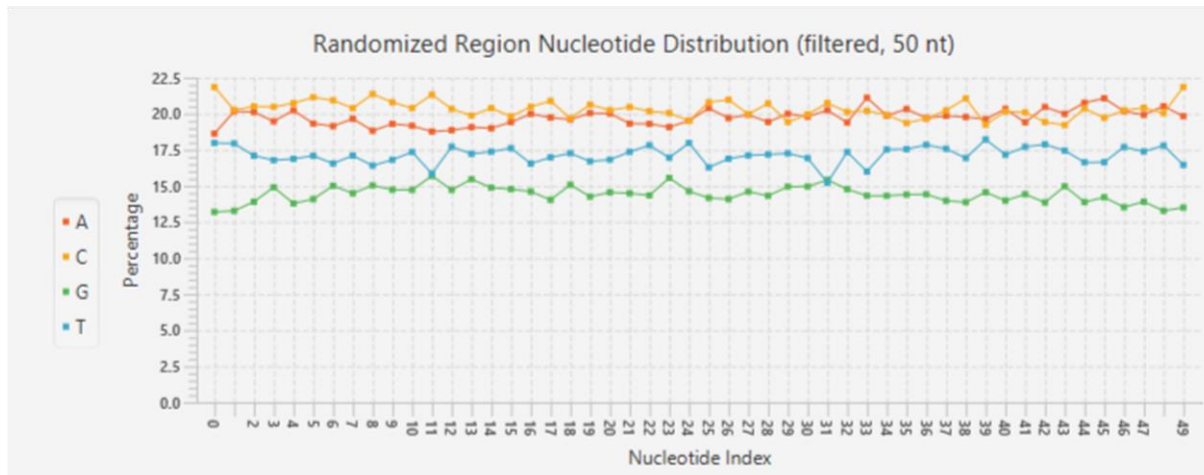
Based on the success of the pilot Sanger sequences, we sequenced both the final winning pool as well as each round of the selection resulting in 9 RNA pools including the initial library to access the behavior of the libraries throughout the selection in terms of overall sequence length, decline in sequence diversity, emergence of any MS2 consensus binding sites, as well as any artifacts produced from the selection protocol itself. Sequencing of the MS2



**Figure 4.5 Length distribution of Round 0 Sequences.** Screenshot from AptaSuite

RNA pools results in over 32 M reads with each sample (1 per round) containing 1.3 M to 9.1 M reads (average 4 M).

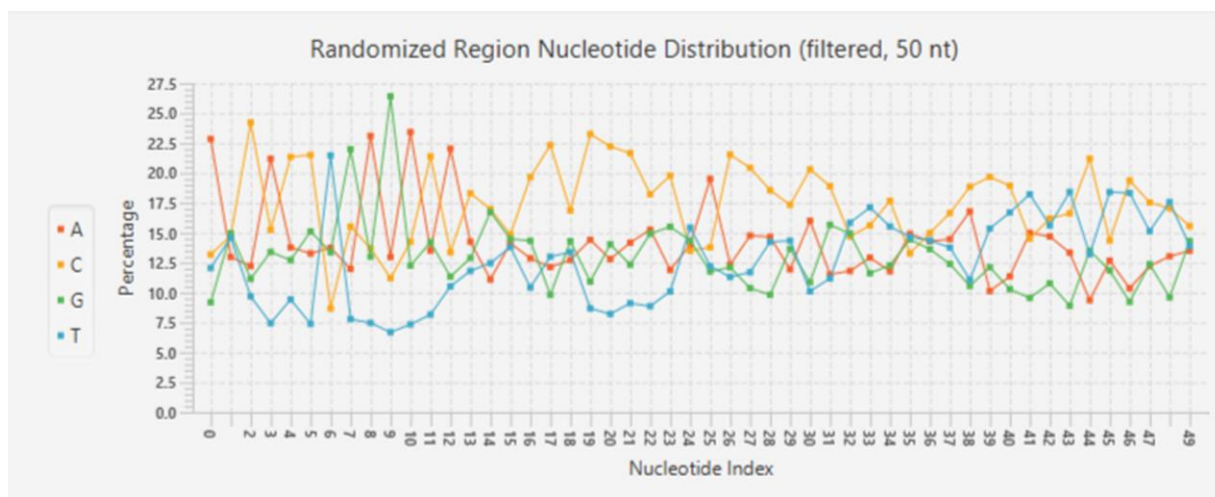
Overall, 64% of the random region reads were 50-nt with ~7.5% 49-nt (distribution shown in **Figure 4.5**), with consistently low levels of other random region lengths, suggesting the initial library synthesized by IDT and then PCR amplified was primarily made of the DNA products we expected. IDT reports their chemical synthesis process as 99.549% efficient,<sup>279</sup> which for the 90-nt template would correspond to 66.58% of the final product being the correct size – providing the bulk of the explanation for our size distribution. Because these reads were sequenced, the constant region primers must have been present in the sequences that produced each read or at least present in the initial library with a sequence close enough for primer annealing. For the extremely short reads, such as those below 10-nt, these molecules are likely the result of primer concatemers as IDT synthesis would be very unlikely to produce a significant amount of truncated products with the constant regions correct for PCR amplification but missing the intervening 50 random nucleotides. The distribution of intermediate shorter random regions in the library is likely the combination of diminishing efficiency of chemical synthesis as well as serendipitous internal priming in the random region, producing truncated



**Figure 4.6 Nucleotide base read distribution of round 0 random region.** Nucleotide index refers to position of the nucleotide in the random region which is defined as the sequence between the constant regions. Screenshot from AptaSuite

products. Likewise, insertions or deletions during PCR amplification could contribute to broadening of the distribution.

High-throughput sequencing of the initial library also reveals a nucleotide composition bias in the random region, favoring A and C (~20%), a small depletion of T/U (~17.5%), and disfavoring G (~15%). This nucleotide distribution appears to be consistently distributed throughout the random region (**Figure 4.6**). This distribution reveals a compositional bias is

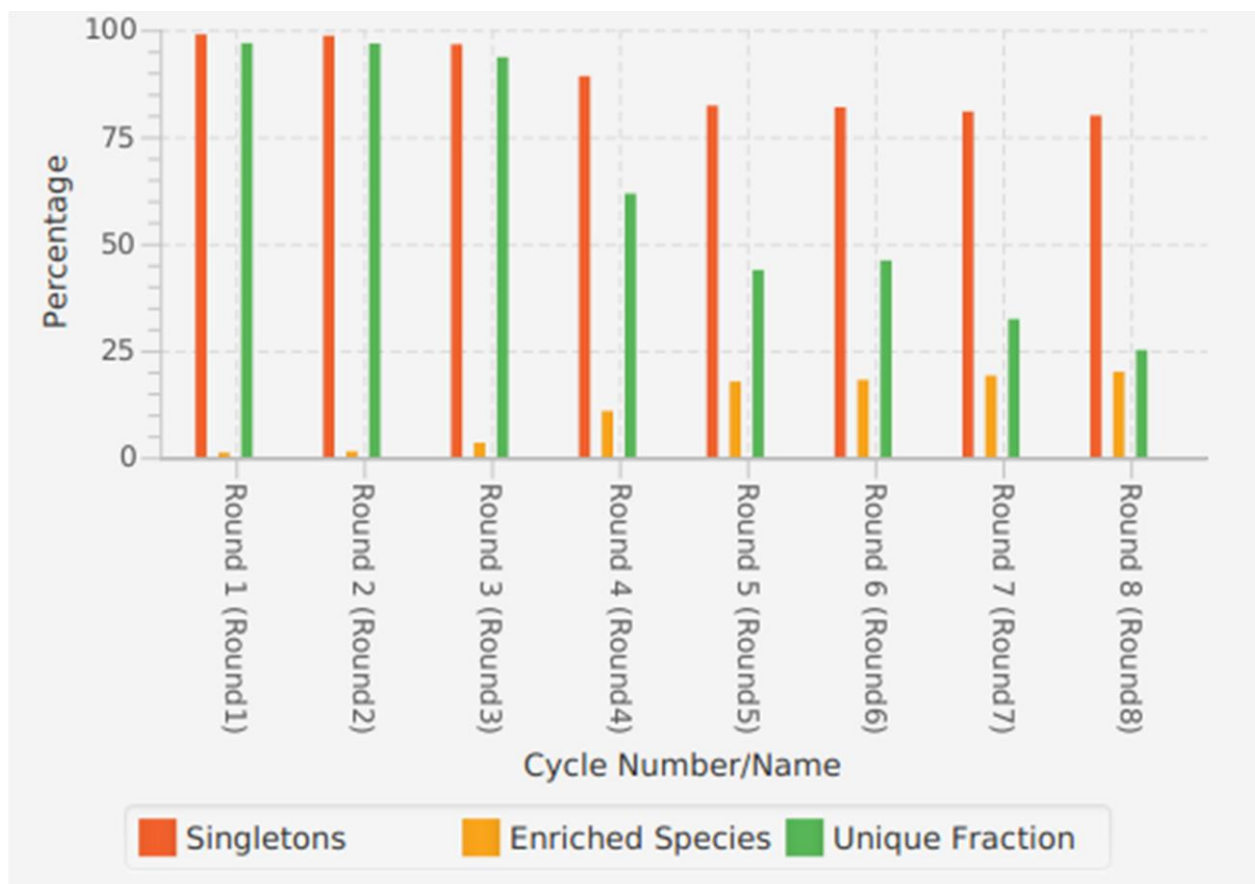


**Figure 4.7 Nucleotide base read distribution of round 8 random region.** Nucleotide index refers to position of the nucleotide in the random region which is defined as the sequence between the constant regions. Screenshot from AptaSuite

present in the initial library. Of the two possible sources for this bias, chemical synthesis or PCR amplification, the chemical synthesis bias is the likeliest explanation for the initial library bias. This is the result of two limitations in the chemical synthesis, differential reactive efficiencies of nucleotide precursors and the order in which the synthesizer injects the nucleotides during synthesis. For an additional cost, hand mixing the phosphoramidites during the synthesis can eliminate this bias or to purposely bias the nucleotide content.<sup>280,281</sup> However, the uniformity of the distribution shows high sequence diversity, consistent with the observation of all sequences other than primer concatemers producing only a single read each. This pattern is contrasted by the distribution of the final round (**Figure 4.7**) in which there are strong positional biases indicative of lower sequence diversity due to the process of the selection.

Throughout the selection, even in round 8, singleton reads (sequences comprising a single read count) dominate the pools (**Figure 4.8**) – however, this should not be interpreted as a failure of the selection to produce enrichment of aptamer sequences – simply that most of the sequences differ in at least one nucleotide over the 50N random region or differ by an insertion or deletion. This is reflected in the difference between the singleton metric and the unique fraction metric. While many of the singleton sequence only appear once within that round, the decrease in the unique fraction indicates many of those sequences are present in at least one previous round. Many of the singleton sequences present within the pool are increasingly related to other sequences present in the library – with many of them likely emerging due to PCR derived mutations. Moreover, even unrelated sequences that appear only once within the last round are still likely to contain a binding consensus but not cluster with other sequences due to the high sequence diversity allowed within the rest of the random region. However, the enriched species metric indicates that the library pool trends towards a subset of sequences (**Figure 4.8**). The enrichment of these sequences could be the result of a variety of factors – high affinity binding that allows the sequence to survive the binding and wash steps, preferential efficiency during reverse transcription, or PCR amplification, or even a stochastic early

enrichment that cascades throughout the selection due to a higher early population. Notably, significantly abundant clusters are extractable as early as round 4.



**Figure 4.8 Most Reads are Singleton Sequences but Enriched Sequences Become Evident by Rounds 3 and 4.** Screenshot from AptaSuite. Singleton reads are sequences that appear once within a round. Enriched Species correspond to sequences that can be tracked between rounds and grow in abundance. The Unique Fraction are sequences unrelated to sequences found in other rounds.

#### 4.3.4 – QIIME Clustering reveals 19 clusters comprised of >0.5% of all unique sequences

Using QIIME to cluster sequences, we found 19 clusters containing more than 0.5% of all unique sequences to find the most highly represented families of sequences. Notably, despite the 19 seed sequences having variable sequences that results in different clusters, all of the sequences contain the MS2 binding consensus sequence, indicating the selection enriched the MS2 consensus motif within the context of many different random regions. Moreover, on closer inspection, all the clusters also contain the tight binding ANCA loop sequence flanked by

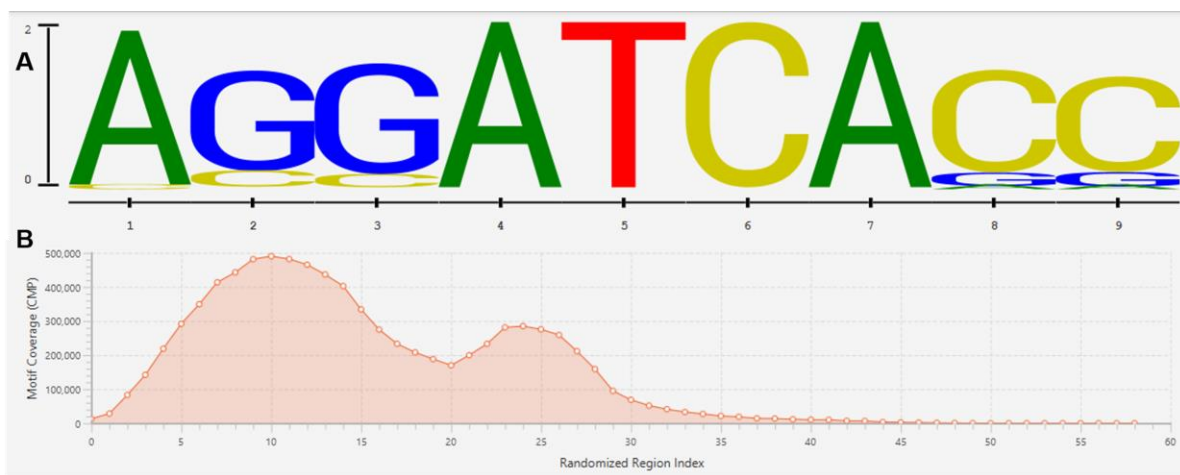
two G-C pairs and a bulged adenine. All but one of the cluster seeds contain the optimal AUCA loop sequence. Searching for the combinations of AUCA string with flanking GC pairs and the adenine bulge reveals almost all sequences appearing more than 10 times contain the ideal binding sequence, with pervasive representation throughout sequences at even lower read counts. As this motif is the tightest binding MS2 aptamer sequence, it is unsurprising that despite the overall sequence diversity of the pool, the round 6 library bound with an affinity essentially equivalent to the tightest aptamer.

#### **4.3.5 – FASTAptamer and AptaSUITE Recapitulate the Clustering By QIIME and AptaSUITE Reveals Ubiquitous Presence of Ideal Binding Motif in Selected Pools**

As FASTAptamer and AptaSUITE use essentially the same method of clustering,<sup>270,272</sup> it should come as no surprise that both pipelines produced nearly identical clustering results. However, due to the parallel implementation of AptaSUITE and its utilization of localized sensitivity hashing (in which several subsets of nucleotides in the sequence are algorithmically converted to strings and compared to each other as a quick approximation of alignments) to filter sequences pre-alignment, clustering with all sequences was feasible rather than sequences based on a threshold abundance. This resulted in 98K clusters, although most clusters are comprised of a single unique sequence of low read abundance. Consistent with the clustering results seen in QIIME, many of the seed sequences used for clusters are identical between the two pipelines, although the lower similarity threshold used for QIIME resulted in the combination of clusters seen in AptaSUITE and FASTAptamer. Likewise, predicated secondary structures of the seed sequences for the most abundant clusters contain a clear MS2 binding site.

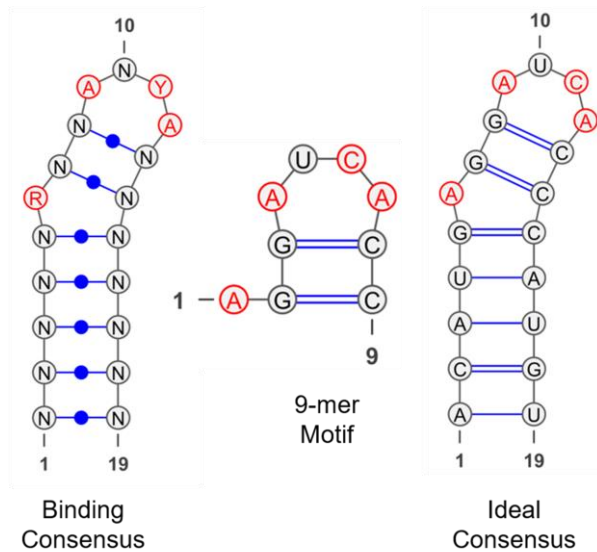
Tracking the emergence of aptamers in this selection may provide clues as to how early solutions may emerge for selection against other targets. The first aptamer containing sequences with more than one read appears in round 2. However, the abundance of this

“winning” sequence is below short reads resulting from primer concatemers, suggesting it is unlikely that this observation would produce an experimental lead where the solution was not already known. By clustering, it appears that by round 4 an aptamer solution with enough abundance has appeared to safely consider the enrichment as significant. This aptamer does not maintain its rank as most abundant throughout subsequent rounds, but it does contain an ideal MS2 binding site. However, as the tight binding for the pool by EMSA demonstrated prior to sequencing, MS2 binding sites are enriched throughout the pool without any single exact 50-nt nucleotide sequence comprising the majority of the pool. Due to the length of the random region, clustering does not immediately reveal this fact as the presence of the perfect 19-nt consensus site still leaves 31 other nucleotides that are unlikely to meet similarity thresholds without the overall pool being over selected for a few dominant sequences. As a result, for longer random regions such as the one we have used here, the enrichment of shorter k-mers (where  $k=6$  refers to all  $4^6$  6-mers sequence combinations) is likely to be more generally informative for the earlier rounds.



**Figure 4.9 Enriched 9-mer Consensus Reveals Ideal Binding Site Biases Towards the 5' half of the Random Region. A)** Sequence logo motif shows sequence preference of the identified motif **B)** Abundance of the motif as a function of nucleotide position within the random region.

Using the AptaTRACE in the AptaSUITE pipeline, enrichment of 6-mer sequences reveal the presence of an ideal MS2 consensus motif in 42% of all sequences. Because the positional enrichment of several related 6-mers overlapping, alignment of the 6-mers produces a 9-mer sequence motif comprising the bulged A, the two-base pair



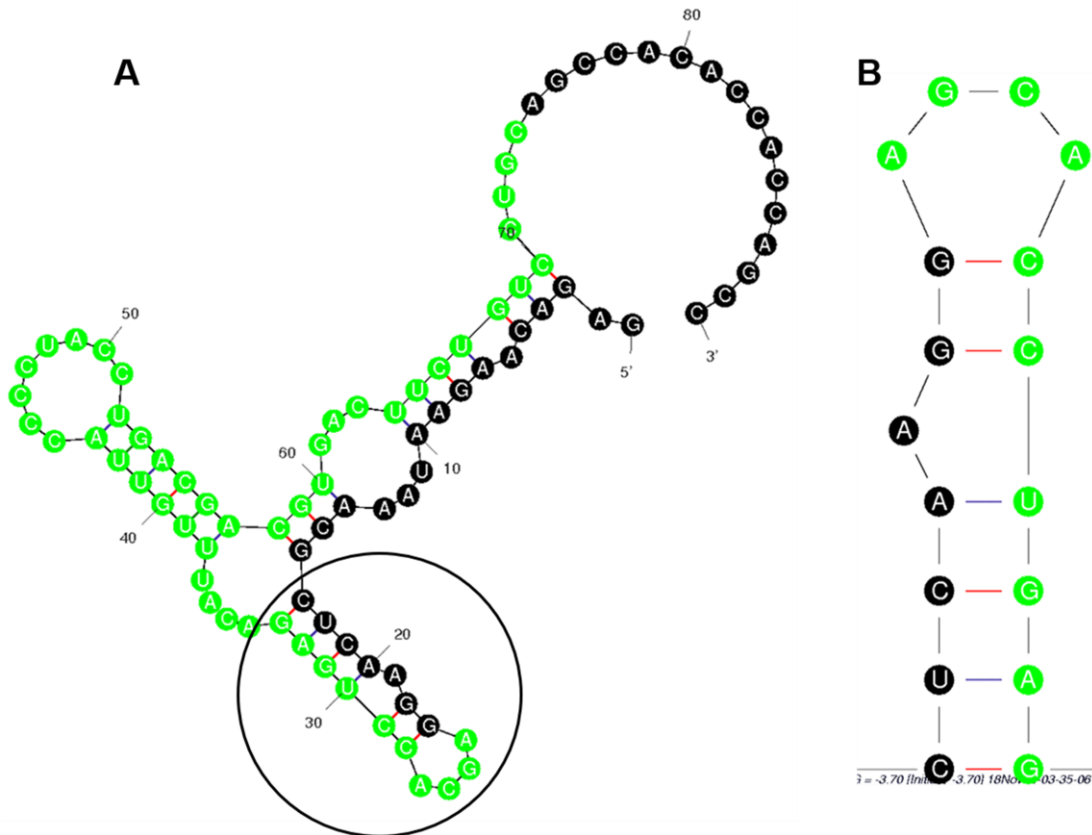
**Figure 4.10 Predicted Secondary Structure of the 9-mer Motif Compared to Consensus and Ideal Binding Sequence**

stem, and the four-nucleotide loop of the ideal MS2 consensus (the logo of which is shown in **Figure 4.9A**). Remarkably, the AptaTRACE k-mer enrichment analysis was able to identify significant enrichment of the AUCA loop sequence with as little as two rounds of sequencing data, suggesting the k-mer enrichment analysis is a powerful tool for revealing sequence motifs which when combined with the abundance and mutational information available in clustering data can provide further insights into the sequence and structural contexts that motif enriched alongside.

As demonstrated previously, this sequence comprises the stem loop and bulge portion of the tightest binding MS2 aptamer, which in the context of an extended helix binds with 2-3 nM affinity (**Figure 4.10**). Interestingly, this motif is enriched in the first half of the random region, with a bimodal distribution centered at 10 and 25-nts into the random region and almost entirely absent in the 3' portion of the random region (**Figure 4.9B**). The enrichment in the 5' region is partially explained by an artifact of the selection, as revealed by the 2<sup>nd</sup> most abundant sequence.

The second most abundant sequence contains an MS2 binding site but does so through pairing between the random region and the 5' constant region. In this constant-random region





**Figure 4.11 Second Most Abundant MS2 Aptamer Sequence Utilizes Random Region Pairing with the 5' Constant Region to Form MS2 Binding Site.** The 50N random region is highlighted in green and the constant regions are shown in black. **A)** The predicted secondary structure (mfold webserver)<sup>233</sup> is shown with the MS2 binding site within the black circle and zoomed in in **(B)**

pairing, the first nucleotide of the random region starts the ANYA loop and completes the motif with the sequence 5'-CCUGA-3' to pair with last nucleotides of the 5' constant region, 5'-UCA(A)GG-3' (**Figure 4.11**).<sup>278</sup>

High-throughput sequencing and affinity binding have both revealed that our selection protocol is capable of producing sequences previously characterized as high affinity aptamers for MS2.<sup>260,269</sup> By EMSA, this affinity manifests as a  $K_D$  as tight as the tightest MS2 aptamer as early as round 6. Analysis of the enriched sequence motifs in the selection bores this out, as 42% of the sequences contain the identified 9-mer sequence motif necessary to form the secondary structure of the tightest MS2 aptamer. Together, these data validate that our

selection protocol enriches for more tightly binding sequences within an initial RNA pool, and that analysis of high-throughput sequencing can reveal those more tightly binding sequences.

Moreover, high-throughput sequencing has provided massive depth to our insights into the selected populations, e.g., revealing biases in the initial pool likely due to the chemical synthesis of the DNA template. The AptaSUITE pipeline also reveals some of the challenges of interpreting the clustering of the long 50N random region used in our protocol as the 19-nt consensus is not sufficient to form a large enough clustering seed to capture all the sequences containing the motif with the recommended similarity thresholds. Complimentary to clustering, sequence motif enrichment should facilitate more detailed examination of enriched clusters of sequences by suggesting which part of the sequence is likely to be the most important, while still providing the advantages of sampling an increased sequence space and providing informational on compatible structural and sequence contexts that motif can be found in.

#### **4.4 – Discussion**

High-throughput sequencing has provided an unprecedented insight into the selected populations of SELEX experiments, allowing for far fewer rounds of selection to reveal consensus binding sites without having to discard insightful sequence variation by over-selecting the RNA pool as necessitated by Sanger sequencing. However, the magnitude of data produced by high-throughput sequences also necessitates a dedicated bioinformatic pipeline to avoid drowning in the overwhelming riches of data, as well as validated strategies in place to determine the relevant features selected.

##### **4.4.1 – AptaSUITE is the Most Comprehensive HT-SELEX Analysis Pipeline Currently Available**

While QIIME has a pipeline that can be adapted to HT-SELEX analysis and a high level of customization and visualization options, it requires a lot of improvisation from the typical

pipeline described in the documentation which makes it challenging to use for this purpose. In addition, the latest release does not currently contain all of the necessary scripts to perform the analysis described here and the 1.9 release is no longer actively supported.

FASTAptamer, despite being designed to process HT-SELEX data for aptamer analysis is computationally slow for large datasets since it is limited to a single core implementation. As a result, the increasingly large datasets produced by advancing technologies such as NEXTSeq means that a large portion of the dataset must be discarded prior to analysis. Moreover, the lack of a user-interface beyond the command line, and the necessity of exporting outputs round by round into spreadsheet programs such as Excel limits both the speed of analysis and the user-base capable of effectively utilizing it. As AptaSUITE utilizes essentially the same core pipeline, it supplants FASTAptamer with its additional advantages.

The AptaSUITE pipeline provides many tools to analyze the huge amount of data produced by high-throughput sequencing in addition to several useful visualization tools to express trends within the rounds of selection. The multicore implementation of AptaSUITE, in addition to local hash filtering for clustering, dramatically reduces the amount of time necessary to run AptaSUITE compared to FASTAptamer. Moreover, it allows for scaling to a supercomputer for even faster analysis. Like FASTAptamer, AptaSUITE is built in a platform-independent package, allowing usage on Windows, Mac, and Linux if Perl (for FASTAptamer) or java (for AptaSUITE) is installed. AptaSUITE also has several other advantages over FASTAptamer including the incorporation of AptaTRACE which allows for built-in secondary structure prediction (though computationally slow) and k-mer motif analysis. In addition, AptaMUT provides a way to track enrichment and depletion trends within clusters, giving incredible insight into the sequence selection pressure among highly related sequences. It also allows for a number of features not described here such as selection simulations and the ability to include sequencing data from counter-selections.

#### 4.4.2 – MS2 Results Agree with Previous Literature

Our MS2 aptamer results agree with the sequences previously identified as MS2 aptamers.<sup>260,269</sup> In particular, the most abundant sequence motif we have identified comprises the nucleotides that confer the greatest specificity to MS2 binding, and this is reflected in the high affinity for MS2 observed for the overall pool. One observation that has emerged from the clustering data is the presence of low-level reads of highly abundant sequences with point mutations. Given the initial diversity of the sequencing pool, the presence of a large population of sequences within 1 or 2-point mutations from the highly abundant sequences strongly suggests PCR or RT induced mutations rather than those sequences representing species from the original library due to the  $4^{50}$  sequences in the initial sequence space. The ability to analyze these mutations for enrichment and depletion is an attractive tool for analyzing which portions of the aptamer sequence are important for binding based on enrichment and depletion profiles or different mutations. Unfortunately, in this dataset the read depth for those mutant species is not adequate for significant statistical analysis, even for the most abundant sequence. This is likely due the high salt wash during the selection protocol but would also be further complicated by the multiple PCR amplification steps between the selection and our sequencing output. The high salt wash caused our selection to be highly stringent, such that mutations that disrupted the specific contacts in the aptamer sequence were likely strongly depleted during each round as they would be unlikely to be able to rely on non-specific interactions to survive within the pool. In this stringency regime, it would be more difficult to accumulate enough weaker binding sequences for statistically relevant read counts – though with enough subsequent rounds with greater sequence convergence towards one or two highly abundant sequence we could probably infer positional importance based on which mutations we never see. This stringency is also likely the reason we do not see substantial emergence of less tightly binding, non-optimal MS2 sites that have appeared in other selections.<sup>260,282</sup> Redesign of the selection protocol may allow for this type of mutational analysis. Less stringent selection conditions would facilitate less

ideal binding motifs to appear in significant read counts and having fewer PCR steps between the selection and the sequencing would improve the sensitivity of the analysis as mutations would be more directly connected to the selection itself rather than introduced in intervening steps. One possibility would be to use a lower fidelity polymerase or error prone PCR during the RT-PCR step to introduce additional mutations for high population enriched species followed by use of high fidelity polymerases during the preparation steps for high-throughput sequencing.

One exciting prospect is that the level of depth by high-throughput sequencing could provide insight into the energetic binding difference among close to optimal sequence/structural motifs similarly the information provided by RNA Bind N' Seq<sup>283,284</sup> experiments or the massive parallel binding performed by Buenrostro et al.<sup>269</sup> However, the SELEX protocol complicates these relationships as sequences are exponentially enriched during the PCR step through multiple rounds. If the sequence library is too diverse to see decent statistical representation in the first round, it would be difficult to deconvolute the energetic relationships later on. Further complicating that potential analysis with this dataset is the differential impact of the high-salt wash on each of the point mutations within the ideal binding motif, as weaker binders are more likely to depend on the charge-charge interactions disrupted by the high salt wash – changing the binding landscape as part of the selection protocol in a manner difficult to systematically correct for in the calculation of relative binding energies.

#### **4.4.3 – SELEX Protocol Optimization**

MS2 is likely an ideal SELEX candidate that might not be representative of success for other proteins that bind more weakly or have faster dissociation rates. However, the trends in the bioinformatics strategy and trends with our library construction are likely to hold true for other systems. With the high sequence diversity of our library pools and the long 50N random region, clustering of sequences has produced hard to interpret data. The long random region means that even if relatively large (such as 25-nt) and invariant consensus motifs emerge, the

sequence similarity between aptamers containing identical binding sites are still likely to be below 50%. The level of diversity is almost certainly invaluable in identifying near-consensus binding sites with biological affinity and providing information about covariation, but it also makes the initial identification of the ideal consensus challenging. This could likely be alleviated in two ways – adding additional rounds of selection to the point of over-selection and/or complimenting analysis of abundant sequences/clusters by looking at the enriched k-mers to narrow down a start point for which regions of an 50N aptamer are important for binding.

Our selection here has also revealed a number of artifacts arising from our protocol, such as the initial distribution of nucleotide base composition, constant-random region pairing, and a positional bias for the MS2 binding site. The biggest disadvantage to the compositional bias in the initial library is that it means we are under-sampling certain combinations of sequences such as G-rich sequences. Constant-random region pairing has a similar effect of skewing the effective sampling of particular secondary structures if a disproportionate subset of the population forms these interactions. However, in terms of identifying binding motifs, this is likely to present a greater problem for genomic SELEX due to the selection of non-genomic motifs that will not map to the genome. Based on the observation of a positional bias of the MS2 binding site in our sequences, there may be an additional advantage to using a longer random region if this is an artifact of reverse transcription in the context of strong secondary structure. More in depth characterization of the secondary structural propensity in the 3' half of the random region for these sequences will be necessary to test this.

Excitingly, this work with MS2 has validated that our selection protocol is capable producing aptamer “winners” and has provided an excellent model system to implement a robust bioinformatics pipeline to handle the magnitude of sequences emerging from high-throughput sequencing. The protocol and these optimizations are likely to prove extremely valuable in characterizing non-canonical RNA-binding by other systems such as the cyclophilins by RNA SELEX.

## **Chapter 5 – Identification of a Tight Binding CypE Aptamer through an Optimized SELEX Protocol**

### **5.0 – Chapter Overview**

This chapter describes the optimization of my SELEX protocol for identifying RNA sequences that bind to cyclophilins. In optimizing this protocol, I discovered a number of enriched aptamers and sequence motifs similar, but distinct from the published consensus RNA binding sequence for CypE. Characterization of the interaction between CypE and the most abundant aptamer revealed a tight binding sequence with an affinity and an extended binding interface, suggesting it could compete *in vivo* with other CypE binding partners.

### **5.1 – Introduction**

Among the remarkable ~40% of proteins identified as RNA-binding proteins without known RNA-binding domains in global studies of the RNA interactome,<sup>124–127,160</sup> the cyclophilin-like domain (CLD) stands out. Repeatedly, consistently, and across kingdoms of life, the CLD has been implicated as a non-canonical RNA-binding domain. Several additional lines of evidence in more focused studies support these global studies. CypA and several *S. cerevisiae* homologues including Cpr1 have been shown to interact directly with viral RNA and inhibit viral packaging.<sup>163</sup> Another yeast CypA from *P. indica* has also been shown to interact with RNA and NMR chemical shift mapping of the interface places the binding surface in close proximity of the isomerase active site.<sup>164</sup> Moreover, a large number of cyclophilins contain canonical RNA binding domains such as RRM for (human) CypE and PPIL4,<sup>161,245</sup> (*A. thaliana*) AtCyp59 which also contains a Zinc-finger and a Arg/Ser-rich domain,<sup>251</sup> and (*S. pombe*) Rct1.<sup>252</sup> In addition, several human cyclophilins also have Arg/Ser-rich domains such as CypG and NKTR.<sup>199,200</sup> As such, it appears that interaction with RNA plays an important and conserved role in cyclophilin function. In that context, it is perhaps not surprising that cyclophilins are widely involved in RNA biological processes such as transcription and RNA processing.

Moreover, the CLD represents an excellent model of non-canonical RNA-protein interactions. The cyclophilin family is a biologically prominent domain targeted by several clinical drugs in practice or currently in development.<sup>190,237</sup> In addition, the available structural and biochemical data strongly facilitates further characterization of RNA interactions with several model cyclophilins. CypA is a well-characterized model system for enzymatic activity and dynamics,<sup>214,219,222</sup> and most cyclophilins are easily expressed and purified. Moreover, a majority of human cyclophilin domains have high-resolution structures available portending well for the feasibility of crystallizing RNA-cyclophilin complexes.<sup>166</sup> However, in the absence of high-quality complex crystals, the structural data current available facilitates the mapping of RNA-proteins interfaces by NMR chemical shift mapping.<sup>214,245</sup> Here, the literature on cyclophilins provides another boon with chemical shift assignments for CypA and the RRM domain of CypE already available. With these advantages in mind, we have chosen CypA and its yeast homologue as ideal candidates for further characterization of the non-canonical RNA-binding activities of the CLD alone as well as CypE to further characterize RNA binding in the context of an RRM domain alongside the CLD.

While the possible functions of CypA and Cpr1 RNA binding is largely conjectural at this stage, CypE presents a cyclophilin family member known to interact with RNA (both described in more detail in Chapter 3). This interaction has been demonstrated unequivocally through *in vitro* binding of purified components and has been proposed to play a regulatory role in CypE gene repression.<sup>195,245,246</sup> However, the consensus sequence described in the literature was originally defined by a SELEX experiment from 44 sequences obtained by cloning and Sanger sequencing<sup>285</sup> and is quite weak at ~200  $\mu\text{M}$ , or about 100-fold weaker than the affinity measured for the MLL1-peptide interaction.<sup>245</sup> Moreover, the sequence was defined by enrichment over a selection against the CLD, which, in light of the recent evidence that the CLD may be a RNA-binding domain itself, may have occluded the discovery of an extended and tighter binding consensus motif. This suggests further characterization of the RNA sequence



specificity of CypE may result in extended binding motifs capable of more tightly interacting with one or both domains.

The first step towards understanding the potential role of RNA binding by cyclophilins is identifying the RNAs and the motifs contained within them responsible for interaction with the CLD. In an effort to better understand this unexplored area of biology, we have pursued the RNA sequence and structural preferences of our model cyclophilins (CypA, Cpr1, and CypE) using *in vitro* RNA selection strategies (SELEX)<sup>256,259,261</sup> for the advantages described in the preceding chapter. We have identified a large number of aptamer families enriched in the selections against CypE and several enriched sequence motifs that differ from the published consensus. Validation of binding by one of these aptamers has revealed an interaction 20-fold tighter than the published consensus sequence that interacts solely with the RRM domain. Additionally, the CypE selection has provided an optimized selection protocol for the affinity regime in which other cyclophilins likely bind RNA through a systematic sampling of buffer salt and protein concentrations, as well as insights on library design such as the use of unstructured constant regions annealed with DNA primers to mitigate constant-random region pairing. The optimized protocol should provide greater insight into cyclophilin-RNA interactions going forward while additional validation of the CypE aptamers and minimization of the sequence and structural motifs involved in binding will help elucidate the RNA binding partners of CypE *in vivo*.

## **5.2 – Methods**

Detailed methods and protocols for the experiments briefly described here are available in Appendix B

### **5.2.1 – Protein Expression and Purification**

To allow for protein binding to the Co-NTA affinity column during selection, 6xHis-tagged proteins were cloned into pET15b, pET21b, and pET28b plasmids (company) using NdeI and

XhoI restriction sites. This resulted in the following constructs; N-terminally 6xHis-tagged CypA (pET15b), Cpr1 (pET15b), full-length CypE (FL-CypE) (pET28b), and CypE-RRM domain (pET28b); and C-terminally 6xHis-tagged -CypE-CLD (pET21b). Plasmids (~50 ng) were transformed into BL21 (DE3) *E. coli* and selected on LB plates supplemented with kanamycin (for full-length and RRM) or ampicillin (for CypA, Cpr1 and CLD). Single colonies were then picked for a 40 mL 37 °C overnight growth with the same antibiotic selection. Using 10 mL of the overnight growths, 1L growths were inoculated and grown in 2L baffled flasks containing the respective antibiotic at 37 °C and shaken at 180 rpm for 2-3 hrs to an O.D.<sub>600</sub> of 0.6-0.8 before being induced with 1 mM IPTG. After induction, the growth temperature was decreased to 18-20 °C and the cultures were harvested, pelleted by spinning at 15K RPMs in a Fiberlite F21-8 rotor (ThermoFisher), and frozen at -20 °C after 18-20 hrs of growth.

Frozen pellets were thawed in lysis buffer (100 mM Tris pH 8 at 4 °C, 1000 mM NaCl, 10% glycerol, 0.1% Triton-X100, 10 mM imidazole; 40-50 mL final volume) supplemented with a Roche EDTA-free protease inhibitor tablet before being sonicated. Lysed cells were then spun at 15K RPM and the supernatant fraction was incubated with Ni-NTA beads equilibrated with lysis buffer for 0.5-1 hr. After 3 washes with lysis buffer, the captured protein fraction was eluted with lysis buffer supplemented with 350 mM imidazole in two 15-20 mL fractions. Eluted protein was then concentration down to ~1.5-2 mL in Sartorius concentrators (10K MWCO for CypA, Cpr1, full-length, and CypE CLD, 5K MWCO for CypE RRM). The concentrated protein was then injected onto a Superdex G75 (CypA, Cpr1, CypE RRM and CLD) or Superdex G200 (full-length CypE) column (both GE) and further purified with size-exclusion chromatography on a Akta FPLC. After elution fractions were combined and concentrated to ~400 µM to 2 mM (with yields ranging from 3 mg/L growth for CypE-CLD to 32 mg/L growth for FL-CypE yield), aliquoted, and flash frozen in liquid nitrogen for later use.

## 5.2.2 – Expression and Purification of <sup>15</sup>N Labeled Recombinant Protein

<sup>15</sup>N-labeled recombinant protein for NMR experiments was generated through the same protocol as described above with the following exceptions. The 2L growth was performed using minimal media supplemented with <sup>15</sup>N ammonium sulfate or ammonium chloride (recipe details in Appendix B) and the slower growth rate using this media required 4-6 hours to reach O.D.<sub>600</sub> of 0.6-0.8 prior to induction.

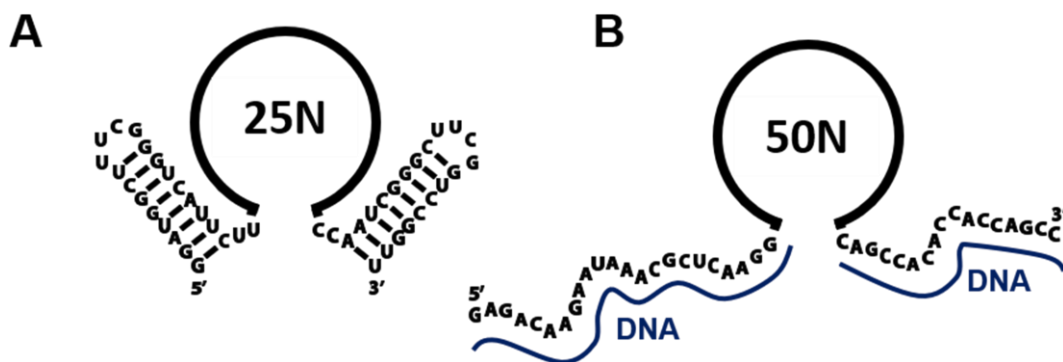
### **5.2.3 – Electromobility Shift Assays (EMSAs)**

To quantify the binding affinity of the target proteins for RNA ligands, EMSAs were performed using radiolabeled RNA ligands produced by T7 *in vitro* transcription and purified protein. The 5' phosphate of transcribed RNA ligands were removed using calf intestinal phosphatase (CIP, NEB) and then 5' labeled with <sup>35</sup>P using T4 polynucleotide kinase (PNK, NEB) and <sup>35</sup>P-γ ATP. Labeled ligand with a final concentration of 5 nM was added to 2-fold serial dilutions of the purified protein ranging from 200 μM to 0 nM final concentration in SELEX buffer (defined below) supplemented with 10% glycerol. Samples were loaded onto a 0.5X TBE 8% polyacrylamide gel and run at 200V at room temperature for 15-20 minutes. The gels were then dried and exposed on a phosphor screen and imaged on an Amersham Typhoon Imaging System. The resulting images were quantified in ImageQuant 5.0 and fit to the quadratic binding equation in Excel using Solver by minimizing the sum of the least squares difference between the data and fit (details in Appendix B)

### **5.2.4 – *In vitro* peptyl-prolyl isomerase (PPIase) assay**

The isomerase activity of recombinant proteins was tested using a previously described assay.<sup>179</sup> Tetrapeptide substrate (N-succinyl-Ala-Ala-Pro-Phe p-nitroanilide; Sigma-Aldrich) was resuspended in a 0.5M LiCl trifluoroethanol solution, which has been previously reported to shift the *cis-trans* population from 12% *cis* in aqueous solution to ~70% *cis*, to a 40 mM concentration. 1-2 uL of 40 mM of this substrate was then added to a reaction with a final

volume of 200  $\mu\text{L}$  of 50 mM Tris pH 7.0, 135 mM KCl, 15 mM NaCl, 2 mM  $\text{MgCl}_2$ , 2.5-5 mM LiCl, and 2.5  $\mu\text{M}$   $\alpha$ -chymotrypsin, 0.5-2 nM recombinant cyclophilin protein, and 200-400  $\mu\text{M}$  substrate with the reaction kept at 4  $^\circ\text{C}$  through a temperature-controlled Peltier. After mixing the reaction volume with a pipette, the UV-vis absorbance of the cleavage product, *p*-nitroaniline, of the *trans*-conformation by  $\alpha$ -chymotrypsin was monitored at 410 nm starting  $\sim$ 10s after addition of substrate. The background thermal isomerization was monitored in the same manner without the addition of recombinant cyclophilin. The effect of RNA on this reaction was monitored with the same conditions, except recombinant cyclophilin was incubated with 1-20  $\mu\text{M}$  RNA for 1 hour prior to addition to the reaction with RNA concentrations in the final reaction volume ranging from 100 nM to 2  $\mu\text{M}$ . For the effect of heparin on CypA activity, CypA was incubated with heparin for 1 hour prior to addition to the reaction with a final concentration of  $\sim$ 8 mg/mL heparin. Total substrate concentration by calculating the total concentration of product at saturation with a  $\epsilon$  of 8800  $\text{M}^{-1}\text{cm}^{-1}$  at 410 nm. Substrate concentration at each point was calculated by subtract the product concentration from total concentration at the that point. The rate of reaction was calculated as the change in substrate concentration at time points 10s apart and divided by 10. Relative catalytic efficiency was calculated by linear fitting of Rate vs. substrate concentration and comparison of the slope between conditions.



**Figure 5.1 Cartoon of SELEX Library Designs** **A)** Library design for 25N library used in SELEX experiment 1 with predicted secondary structure of the constant regions shown. **B)** Library design for 50N library used in SELEX experiments 2 and 3 with the DNA primers annealed to the constant regions.

Exp	Rounds	[Protein]	Binding Buffer	Wash Buffer	Library	Tags	Target
1	7	500 nM (1) 100 nM (2) 25 nM (3-7)	50 mM Tris pH 7 150 mM NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole	50 mM Tris pH 7 1M NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole	25N SHAPE	6xHis	CypA, Cpr1, CypE, RRM, CLD
2	8	100 nM	<b>“Physiological”</b> 50 mM Tris pH 7 135 mM KCl 15 mM NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole	50 mM Tris pH 7 135 mM KCl 1M NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole	50N Anneal	6xHis; 6xHis-MBP	CypA, Cpr1, CypE, RRM, CLD, MBP- MS2
3	15	100 nM 500 nM 1000 nM	<b>“Physiological”</b> 50 mM Tris pH 7 135 mM KCl 15 mM NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole; <b>“Low Salt”</b> 50 mM Tris pH 7 45 mM KCl 5 mM NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole	<b>“Physiological”</b> 50 mM Tris pH 7 135 mM KCl 15 mM NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole; <b>“Low Salt”</b> 50 mM Tris pH 7 45 mM KCl 5 mM NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole	50N Anneal	<b>Alternating</b> 6xHis-MBP; 10xHis- SUMO; <b>Last 7 R</b> 6xHis	CypE

**Table 5.1 Summary of the variable selection conditions across the 3 selection experiments.** Exp is the experiment number; Rounds indicates that number of rounds selected; [Protein] is the protein concentration used through the selection; Binding Buffer and Wash Buffers are the buffer conditions used in the binding equilibrium step and wash steps, respectively; Library indicates the length of the random region and whether the constant regions were the structure shape construct or unstructured and annealed with DNA primers; Tags indicate the protein tags on the targets (MBP, Maltose-Binding Protein; SUMO; Small Ubiquitin-like Modifier); and Target indicates the proteins selected against.

### 5.2.5 – SELEX Experiments

As we performed our SELEX experiments against our cyclophilin constructs, we realized that our initial selection conditions were not ideal for enrichment of CLD binding aptamers. As a result, we performed several iterations of our protocol to optimize these conditions. The different library constructs and protocol differences between SELEX trials are highlighted in **Table 5.1** In addition cartoons of the two library designed we used here are shown in **Figure 5.1**

### 5.2.5.1 – SELEX Experiment 1 - 7 rounds, All CLD Constructs

The following steps (except amplification of the initial library) were repeated 7 times for this SELEX experiment using the following protein constructs, 6xHis-CypA, 6xHis-Cpr1, 6xHis-CypE, 6xHis-CypE-RRM, and 6xHis-Cype-CLD.

#### PCR amplification of initial library

The initial DNA template of the library was produced by PCR amplification of the complementary DNA sequence chemically synthesized by IDT. To obtain an idealized 1X coverage of the  $1 \times 10^{15}$  possible sequences in the 25N library, 2 nmols of the DNA template was used to generate the initial library. This template was split into ten 100  $\mu$ L aliquots of the following reaction conditions: 2  $\mu$ M DNA template, 5  $\mu$ M primers, 1 mM dNTPs, 1X Taq Buffer (10 mM Tris pH 8.3, 50 mM KCl, 1.5 mM  $MgCl_2$ ), and 1U/50  $\mu$ L Taq. Following a 5 min 95 °C hot start, PCR amplification was performed for 10 cycles of 95 °C for 45s, 55 °C for 45s, 68 °C for 45s.

#### RNA Transcription

RNA was *in vitro* transcribed using the T7 RNA polymerase system. PCR template (10% of final volume) was added to a reaction volume with the final buffer of 40 mM Tris pH 7.9, 24 mM  $MgCl_2$ , 1 mM DTT, 2 mM spermidine. 1U/100  $\mu$ L of T7 RNA polymerase and inorganic pyrophosphatase were then added and incubated at 37 °C for 4-16 hours.

#### RNA Purification

RNA was purified by gel purification. RNA was mixed with 2X loading buffer (95% formamide, 0.5M EDTA, 0.1% bromophenol blue) and then loaded onto an 8M urea 8% polyacrylamide denaturing slab gel and run at 20-30W for 2-4 hours. RNA bands were visualized by UV shadowing on Fluor-Coated TLC Plate (Fisher Scientific), cut out, and then

crushed and soaked between 2 hours to overnight in 0.5X TE pH 7.5 buffer. The gel particles were filtered using 0.22  $\mu$ M cellulose-acetate filters (ThermoScientific), before being concentrated on a 5K MWCO centrifuge concentrator (Sartorius). Once the RNA volume reached  $\sim$ 0.5 mL, the 1 mL IDT nuclease free water was added and spun again, with the process repeated three times to remove residual urea. RNA concentration and purity was assessed using a NanoDrop Spectrometer using extinction coefficients predicted by IDT Oligo Analyzer.

### **Pre-selection against Co-NTA beads**

RNA was first refolded by incubation at 80 °C for 5 minutes followed by snap cooling on ice. This RNA at a concentration of 7.7  $\mu$ M for the first round of selection and 1.1  $\mu$ M for all subsequent rounds was then pre-incubated with Co-NTA beads in 1.1X selection buffer for 15 minutes. The Co-NTA beads were then separated to the side of the tube with a magnetic stand while the supernatant was added to the binding equilibrium reaction.

### **The Selection - Binding Equilibrium, Washing, and Elution**

Protein (500 nM Rd 1, 100 nM Rd 2, and 25 nM Rd3-7) was incubated in 1X SELEX buffer (recipe in Table 5.1) for 1 hour with  $\sim$ 7  $\mu$ M pre-selected RNA for the first two rounds and  $\sim$ 1  $\mu$ M pre-selected RNA for subsequent rounds. Co-NTA beads were then added and incubated for 15 minutes prior to separation with a magnetic stand. Supernatant was removed and the Co-NTA resin was wash 3X with wash buffer. After the final wash, 20  $\mu$ L of 1X SELEX buffer supplemented with 350 mM imidazole was then used to resuspend the Co-NTA resin. After 15 minutes, the resin was again separated with a magnetic stand and the supernatant used as the input for a reverse transcriptase reaction.

## **Reverse Transcription(RT)-PCR**

First 1  $\mu$ M RT primer complimentary to the 3' region of the RNA was added to 14  $\mu$ L of eluted RNA. In a thermocycler, the protein was denatured at 80 °C for 10 minutes and then cooled to 4 °C over ~15 minutes. After annealing, 4  $\mu$ L of 5X RT buffer (100 mM Tris pH 7.5, 50 mM NaCl, 50 mM MgCl<sub>2</sub>, 5 mM DTT) was added along with 1  $\mu$ L of 10 mM dNTPs and 1U of reverse transcriptase. The RT reaction was performed at 60 °C for 20 minutes followed by 80 °C for 10 minutes utilizing a thermostable group II intron reverse transcriptase.<sup>286</sup> The full RT reaction was then used as the template for a 500  $\mu$ L PCR reaction aliquoted into 100  $\mu$ L with the following reaction conditions: 20 $\mu$ L/500 $\mu$ L RT-PCR template, 1  $\mu$ M primers, 0.5 mM dNTPs, 1X Taq Buffer (10 mM Tris pH 8.3, 50 mM KCl, 1.5 mM MgCl<sub>2</sub>), and 1U/50  $\mu$ L Taq. PCR amplification was performed for 10 cycles of 95 °C for 45s, 55 °C for 45s, 68 °C for 45s.

### **5.2.5.2 – SELEX Experiment 2 - 8 rounds, all CLD Constructs + MS2**

Following the sequencing results of SELEX experiment 1 producing no identifiable enrichment in sequences, we iteratively changed several of the selection conditions. First, we added 6xHis-MBP-MS2 as a positive control, as described in Chapter 4. We also altered the library design to use constant regions with low propensity to form secondary structure as well an extended 50N random region. The extension of the random region was done with the idea that the additional sequence would still cover all of the sequence space of the 25N library as well as sample more as well. We then added a DNA primer annealing step in an attempt to sequester the constant regions and prevent interaction with the random region. We also slightly modified the binding buffer to 135 mM KCl and 15 mM NaCl which is more physiologically representative than 150 mM NaCl. Moreover, the selection was performed for an additional round for a total of eight rounds of selection.

As a result, the protocol used here was largely identical to the protocol in the SELEX experiment 1. While the amount of DNA template used in the generation of the initial library



does not provide an ideal 1X coverage, the amount of material necessary to cover  $1 \times 10^{30}$  sequences is not feasible. The other major difference was the addition of 2X molar ratio of DNA primers complimentary to the RNA constant regions during the re-folding step prior to pre-selection. The snap-cooling was not changed to a slower annealing due to the concern that intermolecular random region pairing would lead to a larger issue than intramolecular pairing, so DNA-RNA annealing efficiency was sacrificed to mitigate intermolecular interactions.

### **5.2.5.3 – SELEX Experiment 3 - 15 rounds FL-CypE**

Our third iteration of the SELEX experiment tried to systematically sample conditions for CypE as it was the most likely cyclophilin to produce a positive result because of its RRM domain. Comparison of the success of MS2 to enrich for binding aptamers with the CLDs suggested a stringency issue in the selection as the most obvious difference between the systems was their RNA affinity regime. To address this, we tested three protein concentrations with the lowest equal to the concentration used in the previous selections alongside 5 and 10-fold higher protein concentrations. We also tested a lower salt condition of 1/3 the concentration of the previous experiments. Combining these two-variable series led to six parallel CypE selections. To further address our concerns about selection stringency, we eliminated the high 1M salt wash used in the previous experiment and instead change the wash buffer to the same conditions as the binding buffer.

### **5.2.6 – High-throughput sequencing**

Details of the sequences submitted to sequencing are highlighted in **Table 5.2**

Primer or Oligo Name	Sequence
25N Library (RNA Seq)	GGATGGCTTTTCGGGTCATTCTT(N) <sub>25</sub> CCAATCGGGCTTCGGTCCGGT T
25N Library DNA Template	CTCTGTTCTTATTTGCGAGTTCC(N) <sub>25</sub> GTCGGTGTGGTGGTCCG
25N T7 Fwd. PCR Primer	ATATATATGGGTAATACGACTCACTATAGGGAGACAAGAATAAACGC TCAAGG
25N Rev. PCR/RT-Primer	GGCTGGTGGTGTGGCTG
25N P5 Illumina Adapter	AATGATACGGCGACCACCGAGATCTACACATATATATGGGTAATACG ACTCACTATAGG
25N 3' seq adapter	CCGAACCGGACCGAAGCCCGGGCTGGTGGTGTGGCTG
P3-Barcode Index Primer	CAAGCAGAAGACGGCATAACGAGAT(N) <sub>12</sub> AGTCAGTCAGCCGAACCGG ACCGAAGCCCG
25N Sequencing Read Primer	GGGTAATACGACTCACTATAGGGAGACAAGAATAAACGCTCAAGG
Indexing Read Primer	CGGGCTTCGGTCCGGTTCGGCTGACTGACT
50 Library (RNA Seq)	GAGACAAGAATAAACGCTCAAGG(N) <sub>50</sub> CAGCCACACCACCAGCC
50N Library DNA Template	GGCTGGTGGTGTGGCTG(N) <sub>50</sub> CCTTGAGCGTTTATTCTTGTCTC
50N T7 Fwd. PCR Primer	ATATATATGGGTAATACGACTCACTATAGGGAGACAAGAATAAACGC TCAAGG
50N Rev. PCR Primer/RT/3' Annealing Primer	GGCTGGTGGTGTGGCTG
5' Annealing Primer	CTCTGTTCTTATTTGCGAGTTCC
50N P5 Illumina Adapter	AATGATACGGCGACCACCGAGATCTACACATATATATGGGTAATACG ACTCACTATAGG
50N 3' seq adapter	CCGAACCGGACCGAAGCCCGGGCTGGTGGTGTGGCTG
50N Sequencing Read Primer	GGGTAATACGACTCACTATAGG GAGACAAGAATAAACGCTCAAGG

**Table 5.2 List of Primers and Oligos Used in SELEX Experiments**

### 5.2.6.1 – Preparing SELEX Libraries for Sequencing

To submit our SELEX libraries to high-throughput sequencing, Illumina adapter sequences had to first be appended onto the sequences for proper adherence to the Illumina cell. We did this by PCR amplification of our libraries with primers containing the Illumina adapter sequences and sequences complimentary to the constant regions. In both cases, the 5'

P5 Illumina adapter with a T7-5' constant region sequence required only 1 step of PCR for addition while the 3' P3 Illumina adapters also containing 12mer indexing barcodes required 2 PCR-steps for addition.

### **PCR Step 1**

Using the P5'-T7-5' constant primer and our 3' adapter primer, we amplified our libraries for 8 cycles of 95 °C for 45s, 55 °C for 45s, 68 °C for 45s with 100 µL reaction volumes of the following concentrations: 1 µM input library, 5 µM primers, 1 mM dNTPs, 1X Taq Buffer (10 mM Tris pH 8.3, 50 mM KCl, 1.5 mM MgCl<sub>2</sub>), and 1U/50 µL Taq. The PCR products were then cleaned up using a E.Z.N.A. Cycle Pure Kit (Omega).

### **PCR Step 2**

We used the product from PCR Step 1 as the template for the second step to add the P3-barcode indexing primer. Using the P5'-T7-5' constant primer and our P3-barcoding primer, we amplified our libraries for 8 cycles of 95 °C for 45s, 55 °C for 45s, 68 °C for 45s with 100 µL reaction volumes of the following concentrations: 1 µM PCR Step 1 product, 5 µM primers, 1 mM dNTPs, 1X Taq Buffer (10 mM Tris pH 8.3, 50 mM KCl, 1.5 mM MgCl<sub>2</sub>), and 1U/50 µL Taq. The PCR products were then cleaned up using a E.Z.N.A. Cycle Pure Kit (Omega).

### **Pooling and Quality Control**

The resulting libraries were then quantified using a Nanodrop spectrometer and pooled together at rough equimolar concentrations. The combined pool was then gel purified to select the correctly sized products on a native 1X TBE 8% polyacrylamide gel. Following a crush and soak in 1X TE pH 7, the pooled sample was filtered using a 0.22 µM cellulose-acetate filter (ThermoScientific) and submitted to the CU Boulder BioFrontiers Sequencing Facility for quality control and sequencing. The size distribution of the pool was quantified using a High Sensitivity

D1000 ScreenTape system and the concentration was determined using Qubit Fluorometric Quantitation.

#### **5.2.6.2 – Illumina MiSeq Sequencing of SELEX 1**

The first SELEX experiment was sequenced on an Illumina MiSeq instrument using a V2 MiSeq 50 cycle kit for 50 base single-end reads through the CU Boulder BioFrontiers Sequencing Facility. The PhiX concentration used was 30%. The custom read and indexing primers used are shown in **Table 5.2**.

#### **5.2.6.3 – Illumina NEXTSeq Sequencing of SELEX 2 and 3**

The second and third SELEX experiments were sequenced on an Illumina NEXTSeq instrument using a V2 High output 75 cycle kit for 75 base single-end reads through the CU Boulder BioFrontiers Sequencing Facility. The PhiX concentration used in the SELEX experiment 2 sequencing was 30% and the PhiX concentration used in the SELEX experiment 3 sequencing was 50%. The custom read and indexing primers used are shown in **Table 5.2**.

#### **5.2.7 – QIIME and AptaSUITE Analysis**

The analysis pipelines used here are described in Chapter 4 and Appendix C.

#### **5.2.8 – NMR-HSQC Titration Experiments**

To observe gain insight into the binding interface between RNA and cyclophilins, we performed NMR-HSQC titration experiments to map residues with significant changes in chemical shift. All NMR experiments were  $^1\text{H}$ - $^{15}\text{N}$  HSQC experiments performed at 25 °C on a Varian Inova-600 MHz spectrometer using a z-axis gradient HCN room-temperature probe. The pulse sequences used were Varian BioPack pulse sequences. Between data collections, the following variables were optimized:  $^1\text{H}$  and  $^{15}\text{N}$  pulse widths,  $\text{tpwrsf}_n$  and  $\text{tpwrsf}_d$  water

suppression pulses, and tof carrier frequency. Spectra were processed with NMRPipe and analyzed in CcpNmr Analysis.

#### **5.2.8.1 – 25N Titration of CypA**

The initial  $^1\text{H}$ - $^{15}\text{N}$  HSQC of 6xHis-tag CypA at 300  $\mu\text{M}$  was collected for 1 hour 15 min in (50 mM Sodium phosphate pH 6.5, 3 mM DTT, 10%  $\text{D}_2\text{O}$ ). Concentrated 25N RNA library in was then titrated into CypA in a stepwise fashion of 0.25 molar ratio per step from 0.25 to 1.5 molar ratio, each with 1 hour 15 min HSQC data collection and the concentration of CypA ranging from 300  $\mu\text{M}$  for free CypA to 162  $\mu\text{M}$  for the 1.5 molar ratio HSQC. Chemical peak assignments were transferred for residues with peaks overlapping with the assignments available for BMRB Entry 17218.

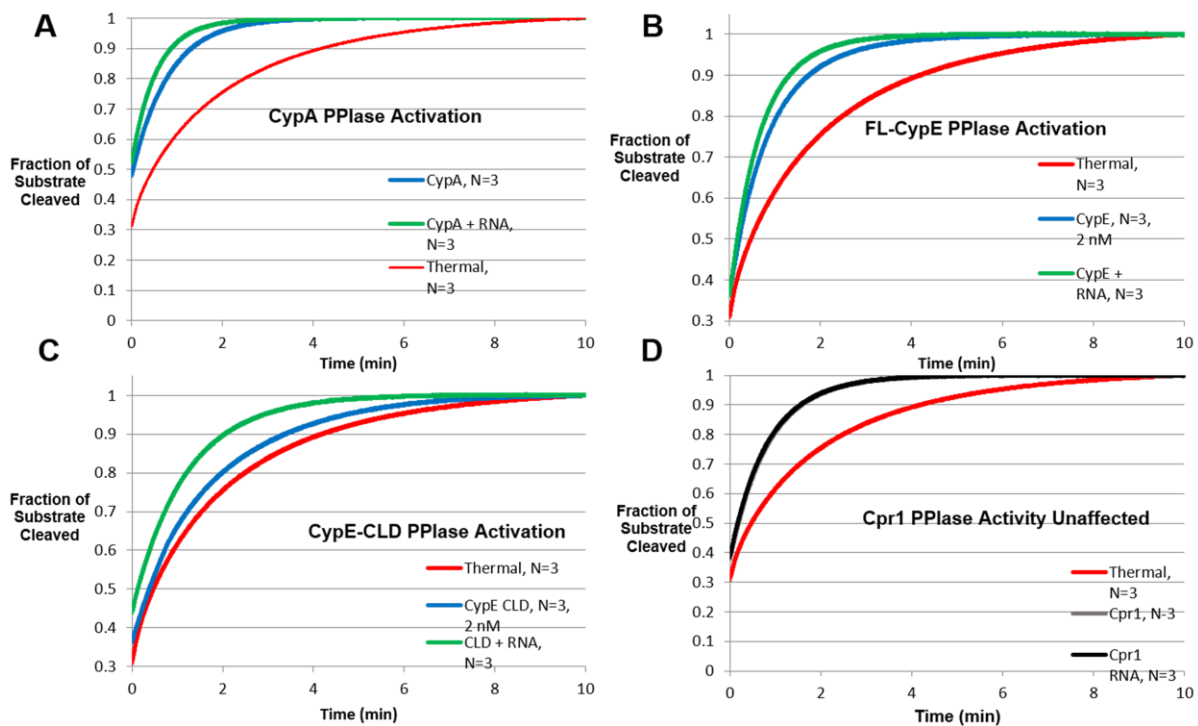
#### **5.2.8.2 – SO-1 Titration of Full-length CypE, RRM, and CLD**

The HSQC titrations of FL-CypE, CypE-RRM, and CypE-CLD followed the same basic procedure as the CypA titration with 25N with the following differences. The buffer conditions used in these experiments was 50 mM Tris pH 7.5, 135 mM KCl, 15 mM NaCl, 2 mM  $\text{MgCl}_2$ , 10%  $\text{D}_2\text{O}$ . To avoid a buffer mismatch likely seen in the CypA titration, SO-1 RNA was precipitated in 70% ethanol, washed, air dried, and resuspended in the NMR buffer. Initial concentrations of protein were 200  $\mu\text{M}$  for the free protein and went down to 150  $\mu\text{M}$  in the final titration point. Chemical peak assignments for CypE-RRM were transferred for residues with peaks overlapping with the assignments available for BMRB Entry 16989.

## 5.3 – Results

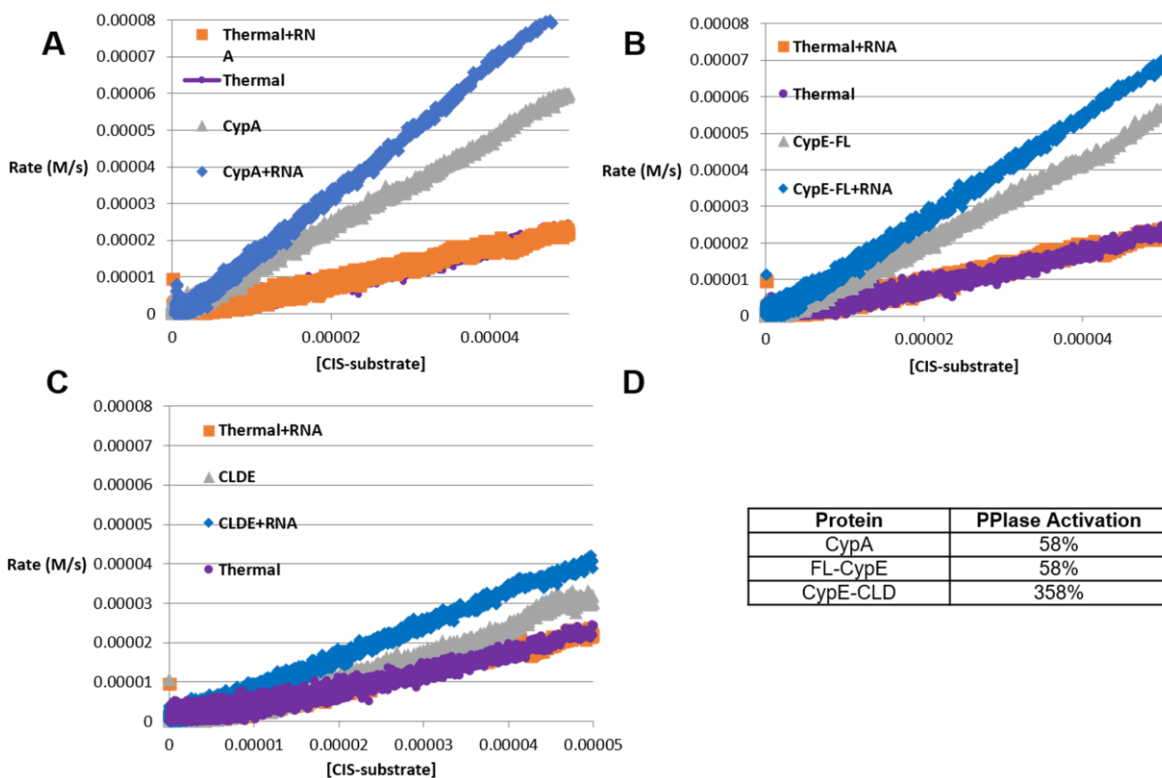
### 5.3.1 – Activation of the PPlase Activity of CypA, Full-length CypE and CypE CLD by RNA

In the context that the PPlase of full-length CypE has been previously reported to be activated in the presence of mRNA, and AtCyp59 has been reported to be inhibited by its consensus RNA sequence, we wanted to test whether the addition of our initial RNA library had any effect on the PPlase activity of CypA, Cpr1, and CypE (full-length and the CLD alone). Excitingly, CypA and CypE (both full-length and the CLD) show increased tetrapeptide isomerase activity in this *in vitro* assay while the activity of Cpr1 is unaffected. Because CypA has also been reported to bind heparin, we tested that as well and it showed inhibition of PPlase



**Figure 5.2 PPlase Activation of CypA and CypE by RNA.** A) Reaction progress curve shown for CypA in the presence and absence of RNA. Thermal background control shown in red, CypA shown in blue, CypA+25N RNA library shown in green B) FL-CypE in the presence and absence of RNA. Thermal background control shown in red, FL-CypE shown in blue, and FL-CypE+25N RNA library shown in green. C) CypE-CLD in the presence and absence of RNA. Thermal background control shown in red, CypE-CLD shown in blue, and CypA+25N RNA library shown in green. D) Cpr1 in the presence and absence of RNA. Thermal background control shown in red, Cpr1 shown in gray, Cpr1 +25N RNA shown in black.

activity. These data are summarized in **Figure 5.2**. Michaelis-Menton experiments for this system are technically challenging due to the fast rate of the thermal interconversion of proline conformations. However, by using the extinction coefficient of the product at 410 nM to calculate substrate concentration and by calculating the rate between different time points, we can estimate the relative activation of the catalytic efficiency. For CypA relative catalytic efficiency is activated by 58%, full-length CypE by 58%, and CypE CLD by 358% (shown in **Figure 5.3**).



**Figure 5.3 Quantification of Enzyme Efficiency Activation.** In these plots of rate vs. [substrate]  $K_{cat}/K_M$  is proportional to the slope of the linear fit **A)** Shown for CypA **B)** for FL-CypE, and **C)** for CypE-CLD.

### 5.3.2 – SELEX with 25N Library for 7 Rounds Resulted in Insufficiently Selected RNA

#### Pools and Reveals Constant-Random Region Pairing

Based on the data suggesting that cyclophilins bind RNA along with the activation of the PPIase activity of CypA and CypE with our random 25N SELEX library, we performed seven

Rounds of RNA SELEX to reveal the subpopulation of tightly binding RNAs within our initial library and identify enriched motifs through high-throughput sequencing.

In this first selection experiment, we designed a 25N RNA library flanked by structured sequences previously used as SHAPE cassettes (cartoon representation shown in **Figure 5.1A**). The rationale behind using these constant regions was the thought that the stem loops formed by the constant regions would prevent or at least mitigate interaction between the constant regions with the random regions as has been previously reported<sup>260</sup> while still being efficiently reverse transcribed and amplified. Several other solutions to avoid random-region pairing with the constant region have been used in the literature. One, switching out the constant region using restriction enzyme sites and careful library design, and two, annealing complimentary primers to the RNA prior to selection.<sup>260</sup> However, the appeal of simplifying our selection protocol through the initial library design of independently folding constant regions and the compatibility of the design with an existing sequence barcoding library led us to test the effectiveness of these SHAPE cassette as constant regions in our first selection.

We performed RNA SELEX with this 25N library against 6XHis-tagged CypA, Cpr1, CypE, CypE-RRM, and CypE-CLD for seven rounds of selection utilizing a Co-NTA resin to capture RNA-protein complexes before reverse-transcription and PCR amplification of the selected RNA sequences. Due to the recent successful utilization of high-throughput sequencing to reveal binding motifs in “RNA-bind N’ Seq” experiments<sup>283,284</sup> analogous to a single round of selection done in parallel at several protein concentrations, we performed a limited selection of only seven rounds. With these experiments, we wanted to probe whether the enhanced sequencing depth of high-throughput sequencing was adequate to reveal aptamer sequences at much earlier rounds than SELEX has traditionally been able to accomplish. The resulting libraries for rounds 4-7 were barcoded and pooled together at roughly equimolar concentrations prior to being sequenced on an Illumina MiSeq instrument (described in more detail in Appendix B). This sequencing run produced ~40M sequences, ~32M of which passed



quality filtering (Phred Score >20), resulting in an average of ~400K reads per sample with a range of 150K-400K reads and summarized in **Table 5.3**.

<b>SELEX Experiment</b>	<b>Total Reads</b>	<b>Reads that Passed Quality Filtering</b>	<b>Sequencing Method</b>
1	40M	32M	MiSeq
2	406M	346M	NEXTSeq
3	184M	180M	NEXTSeq

**Table 5.3. Summary of sequencing read statistics and the sequencing method used.**

At the sequencing depth we obtained, we did not observe significant sequence enrichment for any particular aptamer in round seven. In fact, the most abundant sequences reach only 7 reads per sample. This level of read depth precludes accurate calculation of the loss of sequence diversity from initial library but provides an upper limit of a  $10^9$ -fold loss of diversity. Moreover, k-mer analysis failed to reveal any significantly enriched sequence motifs. This raised several questions regarding the selection – are more rounds necessary? Was something else about the selection causing poor enrichment? The low 25 nM protein concentration in the last 5 rounds raised the possibility that protein-RNA binding may have been outnumbered by background binding, suggesting a higher protein concentration may help subsequent selections.

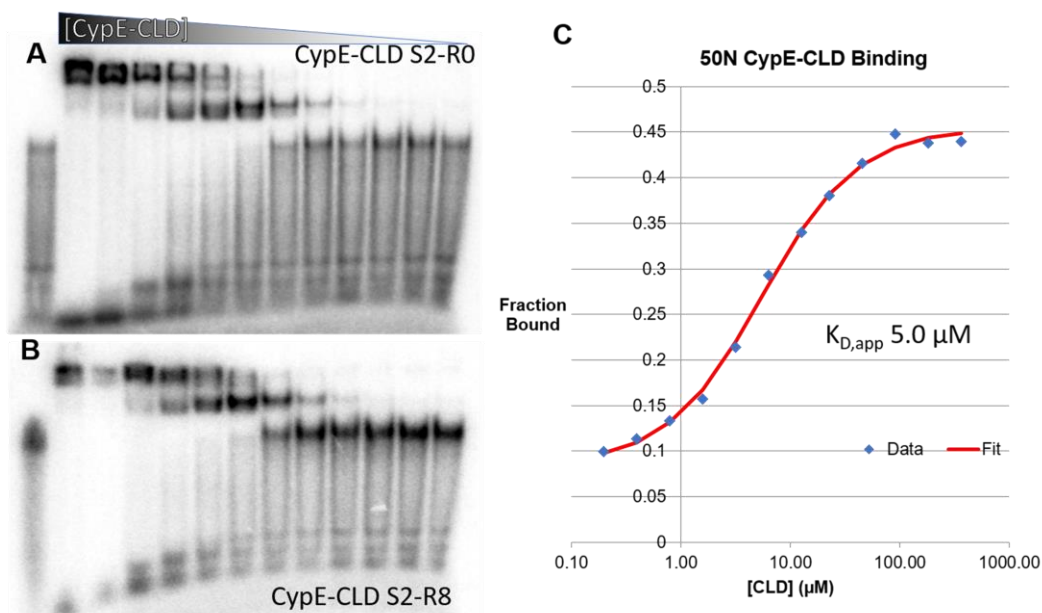
In examining the 6mer counts to gain insight into selection artifacts that may manifest in our sample, it became clear that sequences complementary to the constant regions and A-rich sequences were generally enriched as evidenced by the nucleotide composition of the random region (shown in Appendix C). The enrichment of sequences complimentary to the random region suggested a selection pressure biased towards disrupting the secondary structure, perhaps because sequences that interrupted those structures were more efficiently reverse transcribed and/or PCR amplified. Sampling of the predicted secondary structure of the (albeit poorly enriched) most abundant sequences bore this out, as many of these sequences had

predicted secondary structures involving interactions between the random region and the constant region.

Based on these data, we conclude that this selection did not selected sequences adequately to reveal individual aptamer sequences at our read depth – possibility as a result of background binding or insufficient selection rounds – and additionally that the flanking SHAPE sequences did not successfully prevent constant-random region pairing. Together, these results suggested optimization of conditions was required and an alternative method of mitigating constant-random region pairing should be used.

### 5.3.3 – Preliminary Evidence of RNA Binding and Benchmark for Selection Libraries

To understand whether our selection experiments were enriching for tighter binding RNAs, we needed to establish a benchmark for binding to the initial pool of RNAs to which to compare our selected aptamers. Preliminary binding of the library by EMSAs showed biologically relevant  $K_D$ s (representative gel shown in **Figure 5.4** with others in Appendix B) for



**Figure 5.4 Representative EMSA gels shown for CypE-CLD binding to A) the 50N round 0 library and B) the 50N SELEX 2 Round 8 library. C) Quantification shown and fit shown.**

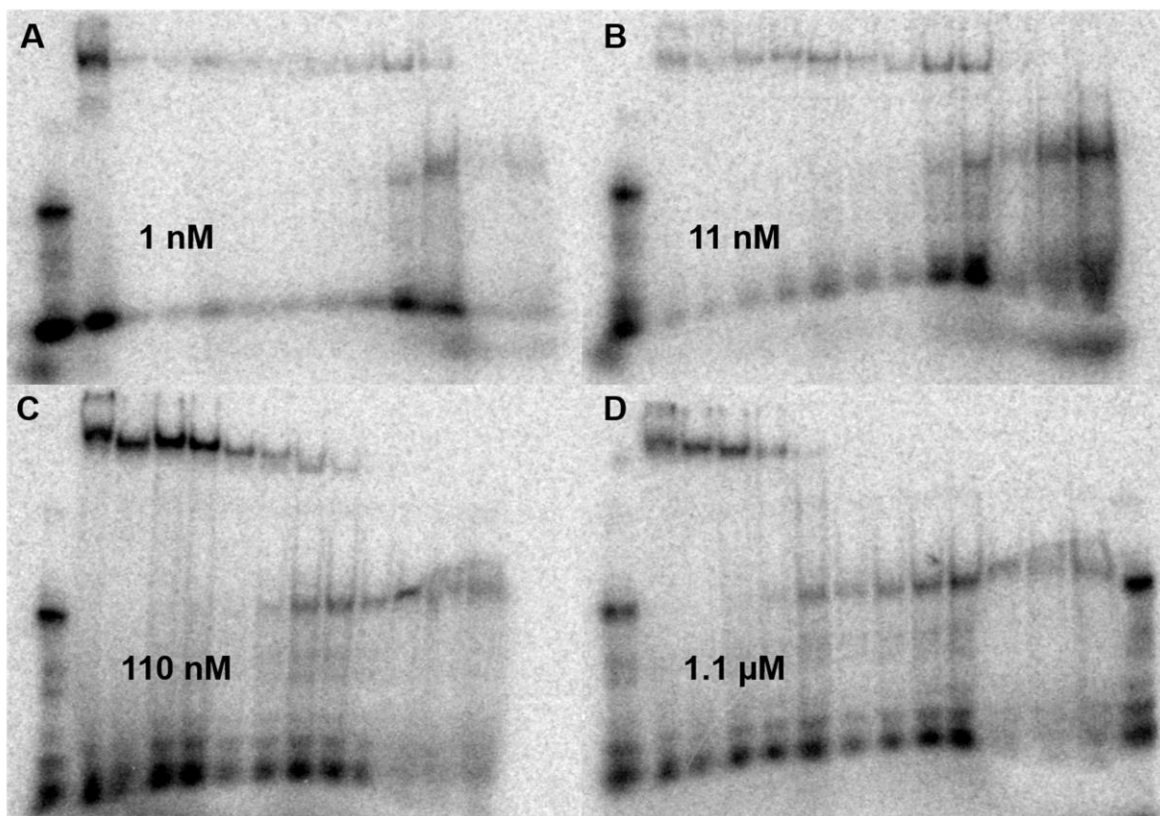
all of the protein constructs tested, with  $K_D$ s ranging from 1-50  $\mu$ M and summarized in **Table 5.4**. It is important to note that these affinities later proved to be artifactual, i.e., they could not be replicated using alternative binding strategies such as ITC or NMR titrations. I hypothesize that they are due to a tight binding, trace contaminant from *E. coli* carried through the purification. I have evidence, described below, suggesting this is the case for CypA, FL-CypE, CypE-RRM, and CypE-CLD. Because all of the protein constructs were purified through nearly identical protocols, this issue is likely present in all the protein preparations.

<b>Protein</b>	<b>50N, <math>K_{D, app}</math> (<math>\mu</math>M)</b>
CypA	51
Cpr1	20
CypE	1.7
CypE-RRM	36
CypE-CLD	5

**Table 5.4 Summary of apparent  $K_D$ s observed for the initial 50N library through EMSA.**

***Ligand-Concentration Dependent Affinity for FL-CypE and CypE-CLD by EMSA***

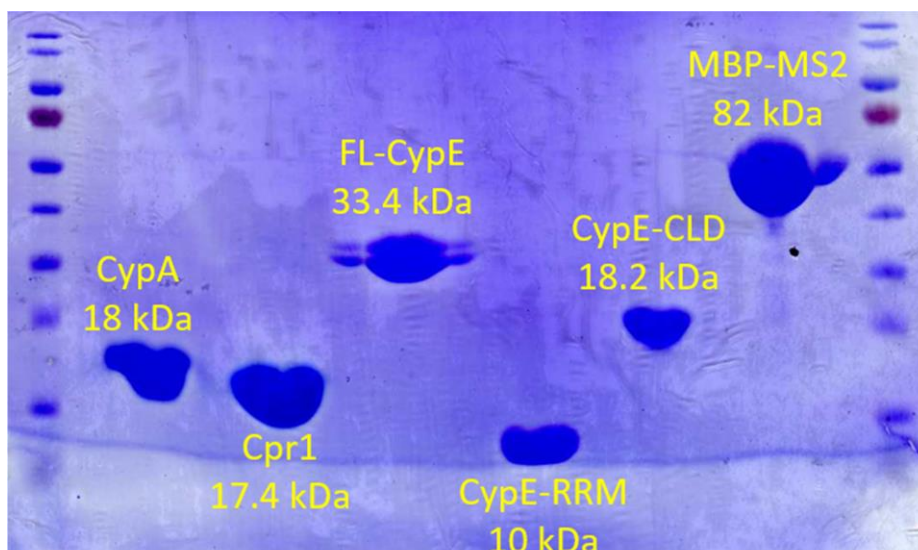
While validating the interaction between the RNA, CypE, and its subdomains, we observed discrepancies between EMSA experiments. Initially thought to be an activity difference between protein preps, EMSA experiments testing RNA binding by FL-CypE, CypE-RRM, and CypE-CLD all showed much reduced binding compared to initial tests. Due to decreasing signal from  $^{32}$ P-labeled ligands, the experiments with the reduced apparent affinities had higher ligand concentrations. This motivated us to test binding at four ligand concentrations 1 nM, 11 nM, 110 nM, and 1.1  $\mu$ M for FL-CypE (**Figure 5.5**) and CypE-CLD which revealed the measured affinities had a strong dependence of the concentration of ligand present. Formally this could be attributed to several phenomena: a very low active protein concentration at odds with the isomerase activity assay, a very strong transition in the effective activity of the ligand through a



**Figure 5.5 [Ligand] Dependent Binding of FL-CypE by EMSA** A) EMSA gel with SO-1 ligand, B) 11 nM SO-1 ligand, C) 110 nM SO-1 ligand, D) 1.1  $\mu$ M SO-1 ligand. [Protein] is directly comparable between gels for identical lanes.

concentration dependent structural rearrangement or oligomerization that does not manifest in a gel mobility change, or an extremely tight binding trace contaminant within our purified protein stocks. The most likely explanation is the present of a contaminant from *E. coli*. Due to the design of the EMSA experiments in which the ligand was at trace concentrations  $200-10^5$  below the apparent  $K_D$ , even an extremely small, but tight binding, contaminant could lead to shifts of the labeled ligand at high protein concentrations used for the apparent  $\mu$ M binding observed. Using the stoichiometry of the binding reactions containing higher concentrations of ligand when the ligand is fully shifted, we can estimate the potential active contaminant fraction at about  $\sim 0.1\%$  which is consistent with its absence on Coomassie and silver-stained PAGE-SDS gels (Figure 5.6).

As these EMSA experiments were used as a benchmark for SELEX enrichment of tighter binding aptamers, it should come as no surprise that this assay did not reveal an increase in binding affinity as any of the selections progressed for any of the cyclophilins as we are unlikely to have selected a significant population of species for tighter binding to a trace contaminant. Even so, the potential presence of an RNA binding contaminant raises concerns for the selections. However, the stoichiometry of the selection experiments should mitigate that undesired bias as the amount of RNA bound to the cyclophilins should greatly exceed RNA bound to the trace contaminant.

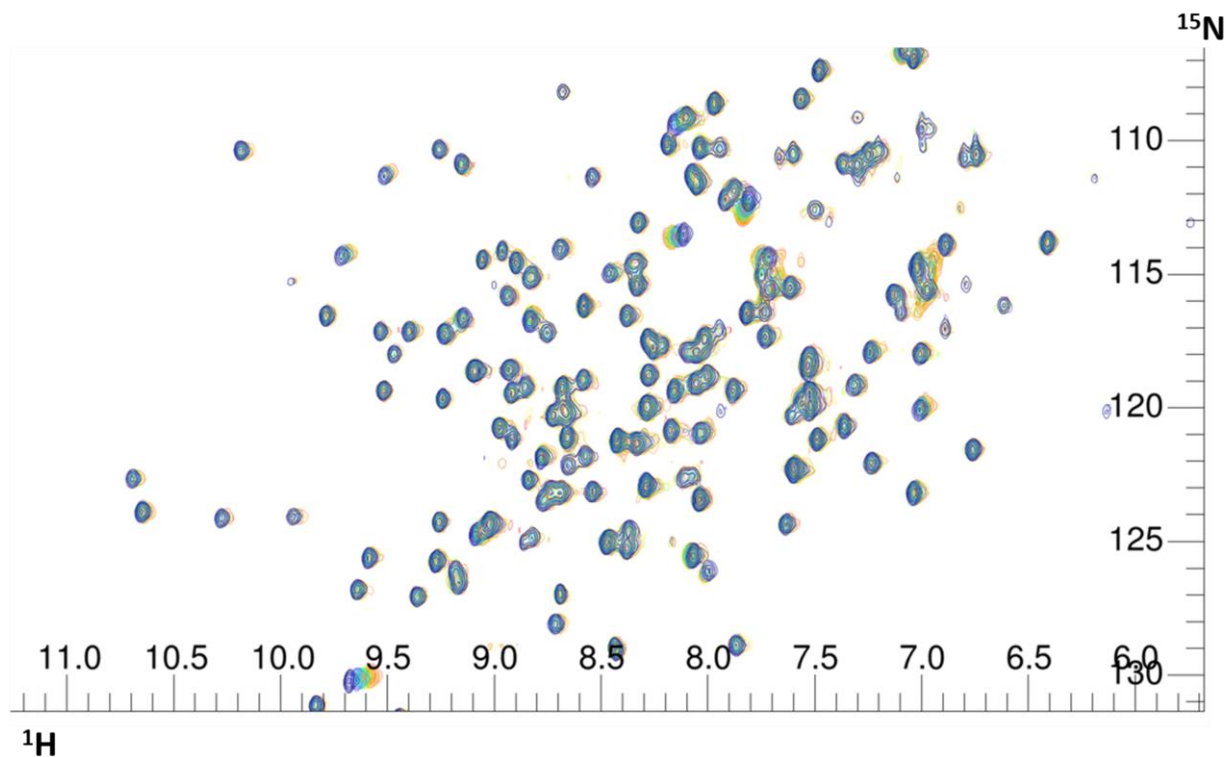


**Figure 5.6 Coomassie Staining for Protein Constructs Used in SELEX 1 and 2 and EMSA assays.**

#### ***NMR HSCQ Titration of CypA by 25N Shows Little to No Interaction with Native Protein***

An NMR HSCQ titration of  $^{15}\text{N}$ -CypA with the 25N SELEX 1 library revealed results consistent with the hypothesis is present in most of the protein stocks. Despite an apparent  $\sim 50$   $\mu\text{M}$  by EMSA, comparison of HSQC of CypA alone to the HSQC of 1:1.5 molar ratio of protein to RNA at concentrations  $>162$   $\mu\text{M}$  revealed significant chemical shift change primarily in a tag residue (**Figure 5.7**) – likely a result of a pH mismatch between the protein and RNA buffer. It's worth noting that some of the residues showing modest chemical shift differences are residues

involved in heparin binding (described in more detail in Chapter 3) – though it is unclear if the change in pH is responsible for these shifts. However, based on the EMSA affinity and concentrations present, at least half of the CypA should have been in complex with RNA, but the magnitude of the chemical shift changes for native protein residues, especially compared to the relative size of the two molecules, is inconsistent with the full ligand shift by EMSA at comparable protein concentrations. Moreover, no precipitation, or decrease in signal-to-noise inconsistent with dilution of the protein was observed, suggesting that if CypA does interact with the RNA, it does so to a much lower extent than suggested by our EMSA data. As this is the 25N random library, the possibility remains that a small population of RNA could interact with CypA.

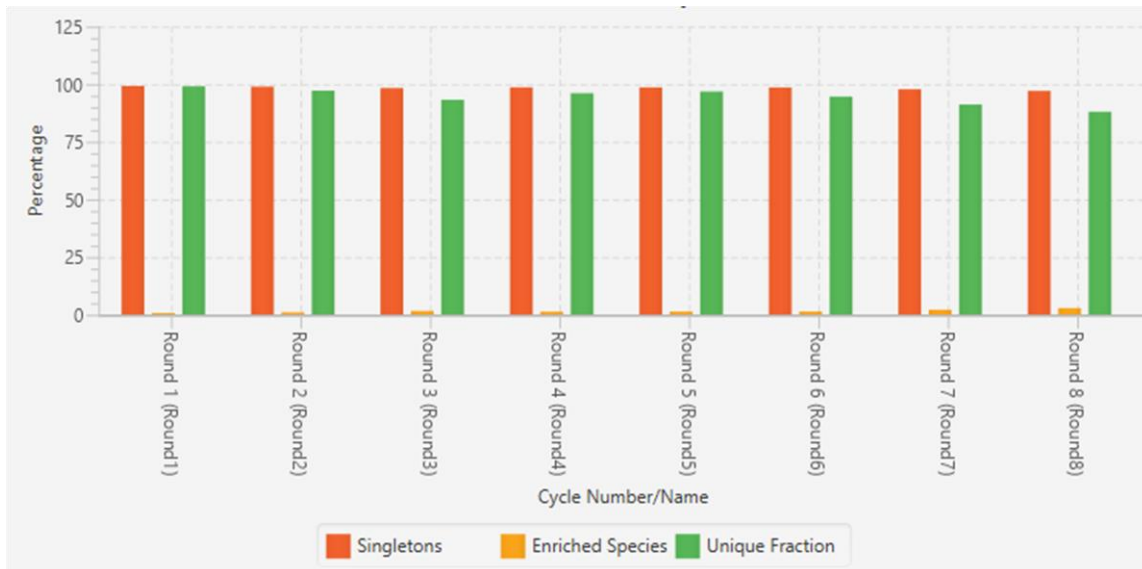


**Figure 5.7 NMR HSQC Titration of  $^{15}\text{N}$  CypA by  $^{25}\text{N}$  Library** Addition of  $^{25}\text{N}$  RNA Round 0 library to  $^{15}\text{N}$ -labeled 6xHis-CypA at stoichiometric ratios of 0 (blue contours) to 1:1.5 protein:RNA (red contours; final protein concentration  $\sim 162 \mu\text{M}$ )

### 5.3.4 – SELEX Experiment 2

In our second SELEX experiment, we took an alternative strategy with the design of our constant regions as we did we our first SELEX experiment. Choosing a sequence with very low propensity of forming secondary structure, we decided to add a primer annealing step to mitigate constant-random region pairing. Despite being less effective than changing the constant region during the selection, adding a primer annealing step is much less technically challenging and does not require specialized single-stranded restriction enzymes or ligation steps. In addition, with the idea that a larger random region samples a greater sequence space, we decided to use a 50N library. To validate our modified selection protocol, we performed this selection on the MS2-coat protein as a positive control, described in depth in Chapter 3. We also used EMSAs to benchmark whether the selection enriched for more a more tightly binding RNA pool by performing EMSAs every other round, although this strategy was misleading for the reasons described above.

Using this RNA library, we performed RNA SELEX against CypA, Cpr1, CypE, CypE-RRM, CypE-CLD, and MS2 coat protein for a total of 8 rounds. As alluded to earlier, our benchmarking EMSA assay did not reveal an increase in binding affinity as the selection progressed for any of the cyclophilins, so we stopped the experiment at round 8. However, presumably because MS2 coat protein binds RNA in a much tighter affinity regime, the EMSA did reveal an increase in the affinity of the MS2-RNA pool for MBP-MS2 (as shown in Chapter 4). This differential observation suggested the selection protocol in principle worked, but that perhaps some feature of the RNA-protein interaction differed between the cyclophilins and MS2, with subsequent sequencing coincidentally revealing this to be the case.



**Figure 5.8 Representative Enrichment Statistics for SELEX 2 Reveals No Aptamers.** Enrichment graph shown for FL-CypE, with graphs for other targets shown in Appendix C

Deep-sequencing with an Illumina NextSeq instrument alongside the MS2 selection results, described in Chapter 4, reveal similar results for the cyclophilin targets as the first SELEX trial, despite ~10-fold greater read depth per sample. This sequencing run produced a total of 406 M sequences, 346 M of which passed quality filtering (Phred score >20). However, none of the targeted protein selections revealed any significantly detectable enrichment of aptamers or sequence motifs (with FL-CypE shown in **Figure 5.8**) other than the MS2 positive control.

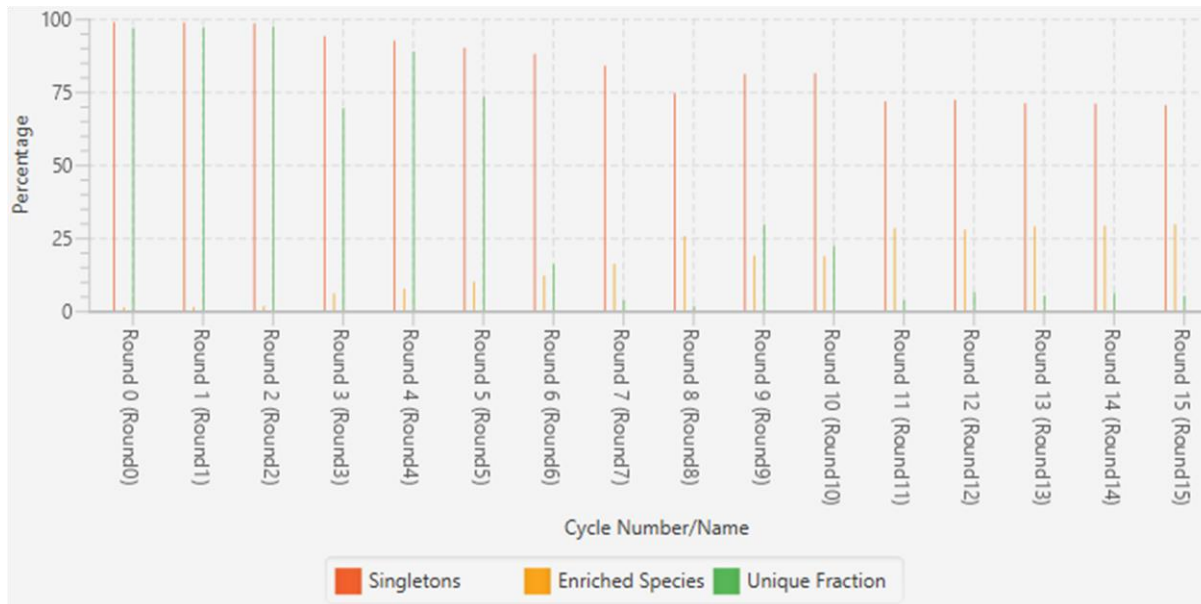
An autopsy of the experiment suggested several possibilities to explain why the protocol was selective against MS2 but not the cyclophilin targets. The most likely difference was believed to be the 1 M salt washes during the selection which may have been less deleterious to the tightly binding MS2-RNA interaction. Other concerns were a difference arising from the relative affinities and concentrations of RNA and cyclophilins and the possibility that interaction between the cyclophilin targets and RNA could occlude the affinity tag. As a result, we decided to modify the selection protocol in several ways. First, we replaced the high-salt washes with washes using the same binding buffer, and systematically sampled various protein and salt



concentrations to establish a selection regime more amenable for the cyclophilins. Finally, we also added a much larger solubility tag to the 6xHis-tag during the first half of the selection in an effort to eliminate the possibility that RNA binding 6xHis-CypE would prevent retention on the Co-NTA column due to the weaker efficiency of Co-NTA compared to Nickel-NTA, the close relative size of the target proteins and RNA ligand, and to mirror the 6xHis-MBP tag on the MS2 positive control used here (though demonstratedly not necessary from the literature).<sup>260</sup> This added the complication that RNA aptamers could emerge for the solubility tag, so the tag was alternated between Maltose-Binding Protein (6xHis-MBP) and Small Ubiquitin-like Modifier (10xHis-SUMO) each round to provide a negative selection pressure against aptamers that bound either tag.

### **5.3.5 – Final optimized SELEX protocol SELEX Experiment 3 Produces Enriched Aptamer Sequences**

For the third and final selection, we decided to address all of these concerns and focus on systematically optimizing the conditions for a single target. With the prior success of SELEX with CypE, it was the promising test candidate. Using six selection conditions of physiological salt at 100, 500, and 1000 nM protein and 1/3 physiological salt at 100, 500, and 1000 nM protein, I attempted to systematically sample the selection space of FL-CypE. In this selection, I used the same library design with primer annealing as used in the second SELEX experiment and added alternating selections against different SUMO/MBP fusion constructs of CypE for 8 rounds. Because the benchmarking assay did not show increased affinity for the pool, I continued for another 7 rounds under the assumption that the pool had not yet been over-selected due to unchanged affinity and to gain further insight through sequencing of every round. Beginning in round 9, I switched the selection back to 6x-His tagged -FL-CypE to further avoid aptamers for either solubility tag as the pool decreased in diversity. We submitted all 91 samples for sequencing on an Illumina NextSeq instrument to try and gain insight into the



**Figure 5.9 Sequence enrichment by Round for Selection 3 (L1000).** A representative enrichment graph by round is shown for L1000.

progression of the selection. From this sequencing run we produced 184 M reads, 157 M of which passed quality filtering (Phred score >20).

For this optimized SELEX protocol, in contrast to our previous strategies, every one of our selection conditions produced highly enriched aptamer sequence clusters (**Figure 5.9**) and revealed enriched sequence motifs through k-mer analysis. Due to the sheer number of sequences, our initial analysis pipeline focused on identifying sequences or motifs warranting further validation and more focused characterization. To do this, we used QIIME to answer questions about how the six conditions compared to each other with the hypothesis that if sequences clustered into the same aptamer family, then that “convergence” of independent selection conditions to the same solution would be compelling evidence for those sequences containing a high affinity binding motif. To do this, we combined all quality filtered sequences from all conditions and rounds (including round 0) and rank-sorted them with FASTX Collapser. Then, filtering for all sequences that appear 5 or more times combined among samples, we clustered the sequences against each other using the QIIME implementation of uclust at a similarity of 80%, preferentially picking cluster seeds based on the most abundant, currently

unclustered sequence based on the rank-sorted order of the input file. Further filtering of the resulting clusters to sequence families comprising greater than 0.5% of the total sequences revealed 16 clusters. We designated the name of each unique sequence by the designation of SELEX Oligo (SO-#) with the number corresponding to its rank abundance. The name of the resulting cluster families follows this with each cluster designated by its seed sequence (the cluster\_1 cluster has the most abundant sequence as its seed). Notably, the cluster families are not just SO-1 to SO-16 as several of the top 16 sequences cluster with other families, such as SO-5 clustering with SO-1.

### ***Cluster Family Abundance Per Condition and Round Reveals Unique Solutions for Each Condition Followed by Cross-Contamination***

Using the 16 seed sequences as a reference sequence file, we then clustered all sequencing reads against those 16 sequences to reveal cluster abundance and enrichment as a function of round and condition. A heatmap of cluster abundance in each condition and round is shown in **Figure 5.10**. Prior to round 7 and 8, each condition has a unique cluster family that emerges as the most abundant sequences within that condition. However, stark changes in the abundance profiles of these clusters as well as the sudden highly abundant emergence of clusters dominant in other conditions strongly suggests cross-contamination. Closer examination of the sequences clustering together between selection conditions reveal the exact same sequences being present in both conditions – a very unlikely event considering the severe under sampling of the possible sequence space of the initial library, especially in combination with the massive enrichment (low to undetectable in one round to greater than 40 K reads in the next). The cross contamination between selection conditions suggest comparison of relative motif and sequence abundance due to protein and salt concentrations is problematic as the issue is likely present throughout the protocol but only clearly evident once a large enough contamination source occurred. It is also unclear at what step this contamination

		Round 5	Round 6	Round 7	Round 8	Round 9	Round 10	Round 11	Round 12	Round 13	Round 14	Round 15
Low Salt 100 nM Protein	cluster_1	0%	0%	0%	1%	75%	37%	6%	35%	42%	46%	56%
	cluster_4	0%	10%	25%	42%	0%	0%	23%	24%	14%	1%	0%
	cluster_8	0%	0%	0%	0%	17%	2%	1%	0%	0%	0%	0%
	cluster_196	0%	0%	0%	0%	0%	0%	0%	27%	40%	50%	41%
	unclustered	100%	90%	74%	57%	6%	26%	70%	14%	4%	2%	2%
Low Salt 500 nM Protein	cluster_1	0%	0%	14%	66%	76%	65%	58%	21%	24%	22%	20%
	cluster_8	0%	5%	21%	20%	0%	0%	18%	0%	0%	0%	0%
	cluster_21	0%	0%	0%	0%	0%	0%	2%	12%	14%	15%	11%
	cluster_27	0%	0%	0%	0%	0%	0%	5%	21%	23%	27%	34%
	cluster_111930	0%	0%	0%	0%	0%	0%	2%	5%	9%	13%	19%
	unclustered	99%	94%	65%	13%	24%	32%	16%	40%	29%	22%	15%
Low Salt 1000 nM Protein	cluster_1	0%	77%	94%	85%	1%	30%	68%	43%	58%	59%	67%
	cluster_9	0%	0%	0%	0%	16%	0%	0%	0%	0%	0%	0%
	cluster_16	0%	0%	0%	0%	0%	0%	0%	6%	20%	23%	22%
	cluster_27	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	1%
	cluster_196	0%	0%	0%	0%	0%	0%	0%	1%	1%	2%	2%
	unclustered	99%	23%	6%	15%	80%	51%	32%	51%	21%	15%	7%
Phys. Salt 100 nM Protein	cluster_1	0%	0%	0%	0%	9%	48%	2%	40%	59%	43%	44%
	cluster_2	0%	0%	1%	3%	75%	40%	3%	2%	0%	0%	0%
	cluster_9	0%	1%	3%	11%	0%	0%	14%	13%	6%	6%	3%
	cluster_26	0%	0%	0%	0%	0%	0%	0%	1%	9%	13%	17%
	unclustered	99%	99%	86%	86%	16%	11%	81%	44%	26%	38%	35%
Phys. Salt 500 nM Protein	cluster_1	0%	0%	0%	3%	0%	1%	15%	65%	47%	52%	83%
	cluster_2	3%	16%	40%	68%	0%	0%	42%	26%	9%	4%	2%
	cluster_4	0%	0%	0%	0%	0%	19%	0%	0%	0%	0%	0%
	cluster_6	0%	0%	0%	0%	30%	0%	0%	0%	0%	0%	0%
	unclustered	97%	84%	59%	30%	67%	79%	42%	9%	44%	44%	15%
Phys. Salt 1000 nM Protein	cluster_1	0%	0%	0%	0%	29%	0%	0%	1%	1%	1%	0%
	cluster_3	0%	1%	1%	4%	0%	5%	5%	6%	8%	10%	17%
	cluster_6	0%	3%	3%	17%	0%	27%	20%	17%	13%	7%	5%
	cluster_22	0%	0%	0%	0%	0%	5%	0%	13%	13%	16%	18%
	cluster_196	0%	0%	0%	0%	13%	0%	0%	0%	0%	0%	0%
	unclustered	100%	96%	96%	79%	16%	63%	75%	64%	66%	66%	59%

**Figure 5.10 Abundance Heatmap of Top Cluster Families by Condition and Round.** The value of each cell is the percentage of total reads clustered into that family within that sample (round and condition). The magnitude and shading of the cell from blue to red is a visual representation of the percentage. The heatmap is truncated prior to round 5 because no clusters comprise greater than 1% of the total sequences in earlier rounds.

occurred – whether during the selection proper or in the subsequent library preparation in which all of the selected libraries were barcoded in parallel. However, the prospect of cross-contamination during the selection does presents the possibility that we unintentionally performed competition experiments in which “winning” sequences between conditions were pitted against each other during the selections.

In this respect, the dominance of SO-1 in several of the conditions strongly suggests this sequence warrants further characterization. Moreover, a sequence variant of SO-1 in which a truncation occurred (SO-3), likely due to the RT/3' PCR primer annealing within the SO-1 sequence, also became dominant within the library pools in which it appeared, becoming abundant enough to reach the third most abundant rank by read count. Notably, the SO-3 truncation was observed during the RNA purification step in multiple conditions during selections, indicating that this sequence was not introduced during the barcoding steps.

SELEX Oligo	Sequence – Lowest Energy Structure (kcal/mol) Dot-bracket Structure – Energy of prediction	Origin
SO-1	UGGUAGACCAGUGAUAAUUAACCUGAUGCGUGGAGGAUAUCGAGUUGUCC -6.70 (((.....)))..(((((((.....(((((((.....).)))))))))))). -6.70 ..(((.....(((((((.....(((((((.....).)))))))))))).)).. -6.40 (((.....))).....((.....)).((((.....))) -6.30	L1000 or L500
SO-3	UGGUAGACCAGUGAUAAUUAACCUGAUG -1.20 (((.....)))..... -1.20 .....(((.....)))... -0.80	P1000 (After X- contam.)
SO-2	CAGGUGUGUGACUACGAAAGAACAUAUAACACAAAAGAGUCCCGUGCC -5.40 ..(((((((.....(((((((.....).))))))))))..))))) -5.40 ..(((.....(((((((.....(((((((.....).))))))))))..))..)) -5.10	P500
SO-4	UGGCCGGCCCAUCCCGACUGCCGGGUGAUAGACUCUUUAGCGAUUUUUGG -9.00 ..(((((((.....).))))))(((((.....))))..... -9.00 ..(((.....(((((((.....).))))..)).((((.....)))))).. -9.00	L100
SO-6	AGGUGCCUCAAAUCCGCAUAAGAUAACAACAUGGAGUGAAGCGCUCCCC -10.50 ..(((((((.....(((((((.....).))))))))))..)).. -10.50 ..(((((((.....(((((((.....).))))))))))..)).. -10.40	P1000
SO-8	AGAACAUAUUACAAAGACUGAGCGUUUUAAAGUCUCCUCAUGUGCCCC -3.90 .....((((.....)))..... -3.90 .....(((.....((((.....))))))..... -3.40	L500
SO-9	AAAGUGAGAUAGGUAACAACAAGAAUAUAUACCUAUCUUGCC -8.50 ..(((((((.....(((((((.....).))))))))))..)).. -8.50	P100







**Table 5.5 The Predicted Secondary Structures and Condition Origins of the Most Abundant Cluster Seed Sequences.** The SO-# indicates the rank-sort abundance of each sequence. Representative secondary structures are shown for the lowest energy structures along with the kcal/mol energy of folding. The origin indicates which selection condition (prior to round 8, expect for SO-3) that the SO first appeared.

***Secondary Structure Calculation and Enrichment K-mer Analysis Suggests Enrichment of Single-Stranded AAY-rich Motifs***

Canonically, RRM capable of binding RNA do so at single-stranded regions of RNA and interact with 4 to 6 nucleotides through each RRM. To assess likely regions of interaction between the selected sequences and CypE through the RRM domain, we predicted the

secondary structures of our top cluster sequences and searched for enriched 6-mer sequences. A summary of the sequences and lowest energy secondary structure predictions, in dot-bracket notation (Vienna)<sup>287</sup> are shown for the 50N random sequence for each of the “winning” sequence families in **Table 5.5**. While the selection included constant regions during the bind and retention of the RNAs, DNA primers ideally were annealed to the constant regions during this step. However, in actuality, the folding space was quite complicated during the selection with 4 possible RNA-primer annealing states (free, 5' annealed, 3' annealed, both annealed). The 6-mer enrichment was analyzed using AptaTRACE, a subscript in the AptaSUITE program.<sup>270,275</sup> Using a comparison of all sequences with reads less than 10 counts as the background, the enriched 6-mer sequences for each condition was calculated for the first 8 rounds (**Table 5.6**). Together, the secondary structure prediction and the enriched k-mer analysis suggest the RNA region involved in at least binding to the CypE-RRM is single-stranded AAY-rich sequences, consistent with the known preference polyA and polyU RNA.<sup>195</sup> Including the motif for L100 suggests the consensus might be broader and accommodate RYRAYA.

While the knowledge about RRM s gives us insight into the features likely involved in RNA binding to half of CypE and what to look for in our bioinformatic analysis, the potential RNA interactions with the CLD are less clear. Several lower abundance sequence motifs emerged from the k-mer analysis – with pyrimidine-rich sequences appearing in several conditions at low ~1% frequency above background and preferentially appearing in the 3' region (as present in SO-6 and SO-8). However, similar motifs also emerge from the MS2 selection, suggesting this motif may actually result from a bias from in our RT or PCR steps – not from CLD binding. To address these, we decided to take an experimental screening approach to quickly assay for interaction with the CLD.

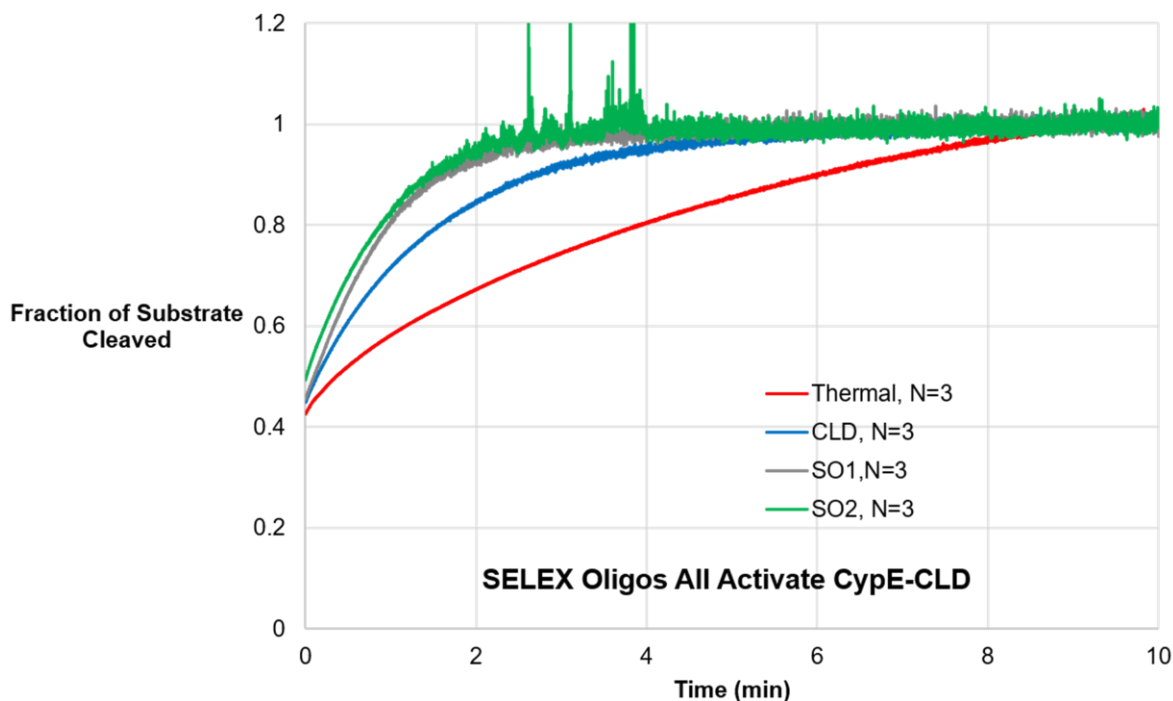
Selection Condition	Top Enriched 6-mer	6-mer Freq.	Motif Logo
L100	GTGATA	42%	
L500	ATAATT	84%	
L1000	AATTAA	85%	
P100	ATAATA	49%	
P500	ATAATA	81%	
P1000	AATAAC	41%	

**Table 5.6 Enriched 6-mer Seed Sequences Enriched in the First 8 Rounds by Condition.** The selection condition indicates the condition in which the 6-mer sequence is enriched with the Freq. indicates how frequently that 6-mer sequence is found in reads with counts greater than 10 compared to reads with total counts lower than 10. The Motif Logo corresponds to the relative abundance of each nucleotide at the each position of the 6-mer motif and surrounding overlapping enriched 6mers (e.g. GATAAT + ATAATT + TAATTA for L500).

### 5.3.6 – Screening Aptamers for CLD Interactions by PPlase Activity

Because we already have an PPlase activity assay in place for CypE, it seemed like an excellent starting point to screen some of our RNAs, and if we saw differential effects such as activation, inhibition, or magnitudinal differences in either effect, then that would point towards the most biologically interesting motifs.

To do so, we *in vitro* transcribed our 16 seed clusters (i.e. the most abundant sequence from each of our 16 QIIME clusters) and tested the PPlase activity of CypE-CLD with these RNAs. All the RNAs tested activated the PPlase activity and did so with similar magnitudes (**Figure 5.11**). However, this was strange as not all the RNAs were at the same concentration in the preliminary assay and some of the seed sequences were intentionally chosen due to their relatively low read count. To rule out a systematic salt effect being carried through from the RNA purification, SO-1 was dialyzed against the protein in the reaction buffer, but the assay result was unchanged, with RNA still having the same activation behavior. Subsequent characterization of SO-1, as described below, reveals that the RNA does *not* bind the CypE-CLD when the domain is in isolation, and so this effect is likely to be mediated through some other interaction in the assay. Together these point towards more controls being necessary to explain the physical meaning of the PPlase activation.



**Figure 5.11 Pervasive Activation of the PPlase Activity of CypE-CLD by SELEX Aptamers.** Activation of SO-1 and SO2 shows representative activation by all SELEX Oligos.



### 5.3.7 – Preliminary Characterization by EMSA Suggests SO-1 Aptamer is the Tightest Binder of Aptamers Tested

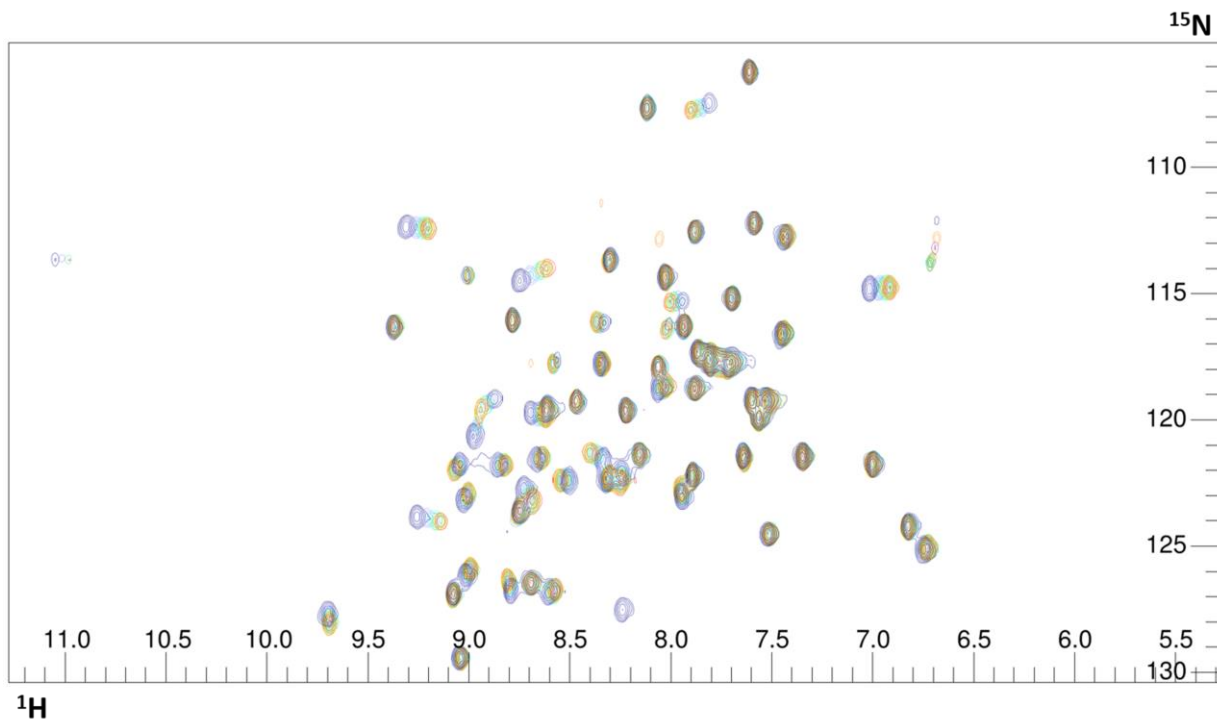
With this in mind, we instead focused on characterizing the affinities by EMSA of the most abundant “winners” from each of the selection conditions (with preliminary data shown in **Table 5.7**) rather than a larger set of sequences. To address the tight-binding RNA contaminant issue that plagued the benchmarking assay, we re-purified FL-CypE by purifying the His-SUMO construct already on hand with the addition of Ulp1 cleavage and second nickel affinity column clean-up steps prior to SEC. The alternative purification scheme produced a protein stock that does not exhibit ligand concentration dependent affinities in our EMSA assay, indicating the shifts are due to FL-CypE and not the previously observed contaminant.

SELEX Oligo	$K_{D, \text{Apparent}}$ ( $\mu\text{M}$ ) by EMSA	Replicates
SO-1	6.6	3
SO-2	21	2
SO-6	~38	2
SO-9	~87	2

**Table 5.7 Preliminary EMSA Binding Affinities for Condition “Winners.”**

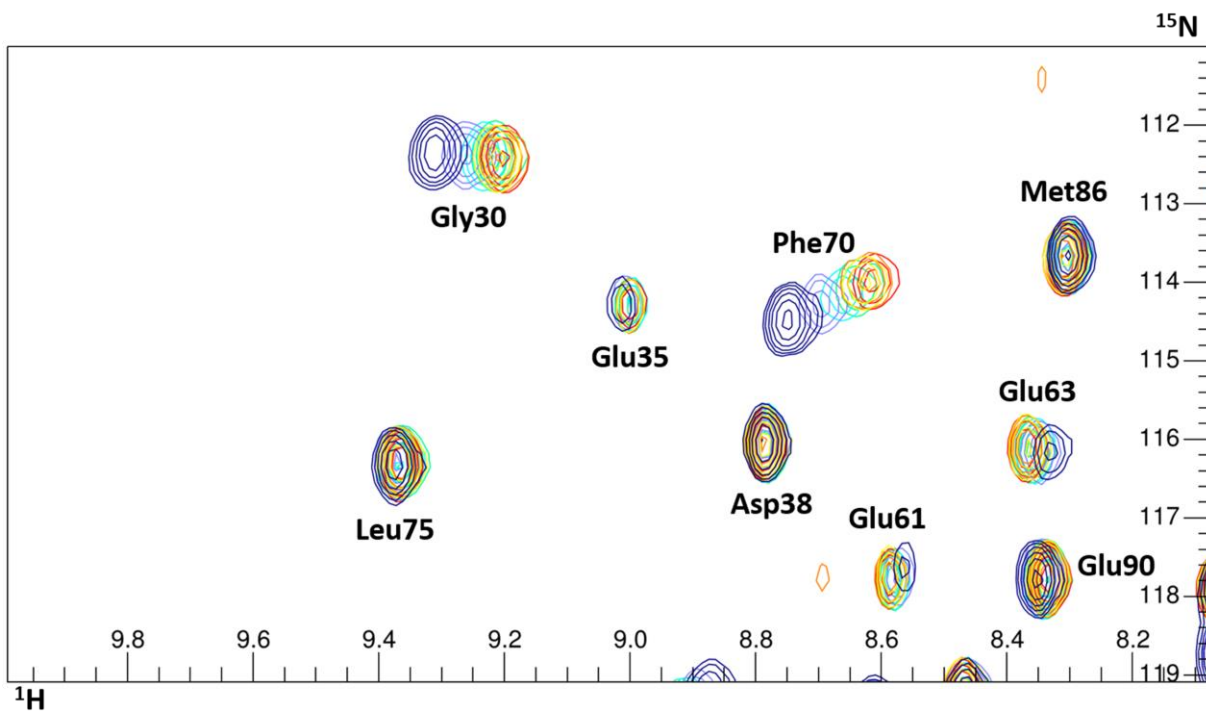
### 5.3.8 – NMR HSCQ Titration of CypE-RRM, CypE-CLD, and FL-CypE Reveal SO-1 Binding Solely to the CypE-RRM and Independently Behaved Subdomains

NMR HSQC titration experiments allow for the characterizing of the structural interface of our RNA aptamers and CypE. By adding RNA at a fixed 0.25 molar ratio between 1:0 and 1:1.5 protein:RNA, we can obtain information about the chemical environment of most residues in our protein in the free and complexed state.



**Figure 5.12 SO-1 Binding to CypE-RRM by NMR HSQC Titration.** Free RRM HSQC color blue, with each 0.25 molar ratio step going cyan, green, yellow, orange, and red as the final 1.5 molar ratio point.

The complex between CypE-RRM and SO-1 occurs in a fast-exchange regime which results in characteristic peak walking behavior as the overall equilibrium of the free and complexed protein changes as RNA is added (**Figure 5.12**). This allows for the changing chemical shift data for individual residues to be tracked as a function of complex formation without having to reassign the chemical shifts of those residues. Moreover, it allows us to observe saturation of the protein by RNA as the peaks stop walking once 1:1 stoichiometry is reached whereas an interaction in which the concentration of either the protein or RNA is below the  $K_D$  the chemical shifts of residue peaks will continue to change as RNA is added above a 1:1 molar ratio. While the  $K_D$  predicted by EMSA is likely too low to accurately measure by NMR chemical shift changes due to poor signal to noise issues at protein concentrations below 7  $\mu\text{M}$ , the stoichiometric titration behavior of CypE-RRM with 0.25 to 1.5 molar ratios of SO-1 are consistent with the estimated  $K_D$ .



**Figure 5.13 Comparison of Phe70 (51 native) Chemical Shift Changes During SO-1 Binding.**

Fortunately, the HSQC chemical shift assignments are available for CypE-RRM, allowing the transfer of 70% of assignments for the peaks we have observed (selected residues are shown in **Figure 5.13**). In mapping the significantly shifted residues onto the solved RRM structure,<sup>245</sup> we can observe that SO-1 interacts with the canonical RRM binding residues with additional chemical shifts in the loop between  $\beta$ -sheets 2 and 3 (**Figure 5.14**). In addition, previous titration experiments for CypE-RRM have revealed the surface responsible for RRM binding to the literature consensus sequence AAYAAA and the PHD3 peptide, allowing for comparison between our chemical shift changes and theirs. Remarkably, the interaction surface of the RRM-SO-1 complex shows overlaps with binding surfaces of both the AAYAAA ligand and the PHD3 peptide (**Figure 5.15**).

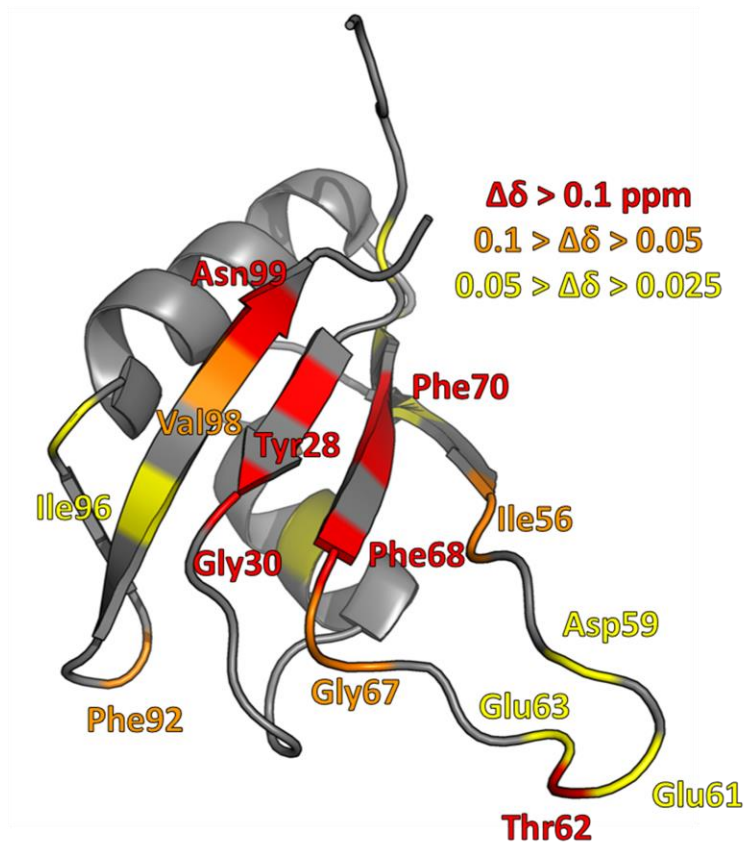


Figure 5.14. SO-1 Chemical Shift Changes Map to the Canonical RRM Interface.

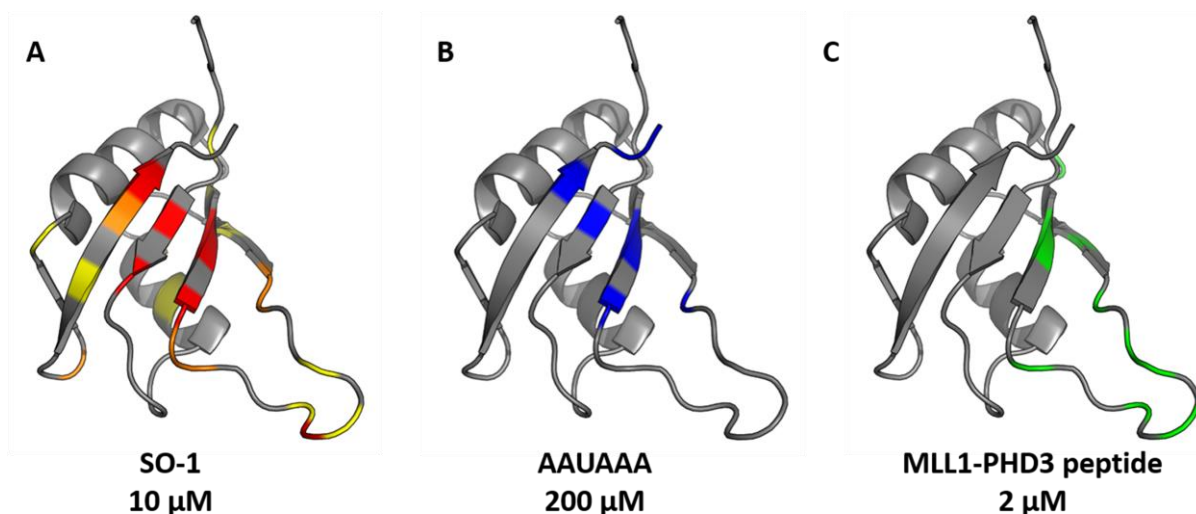
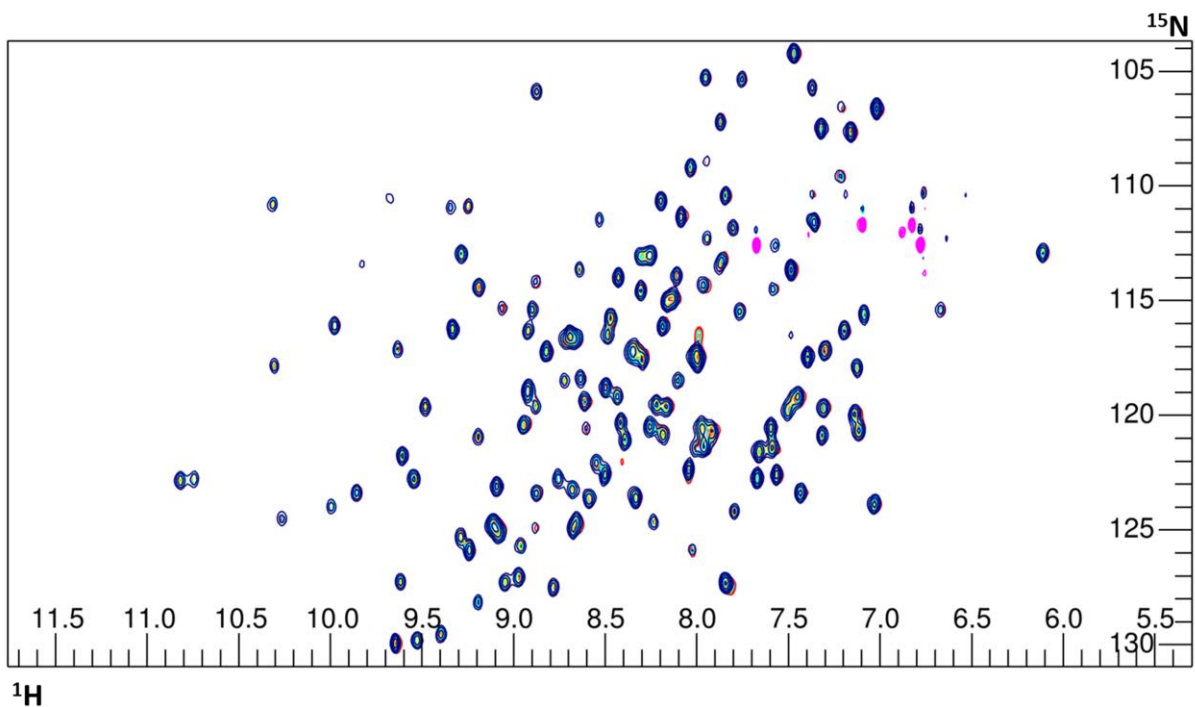
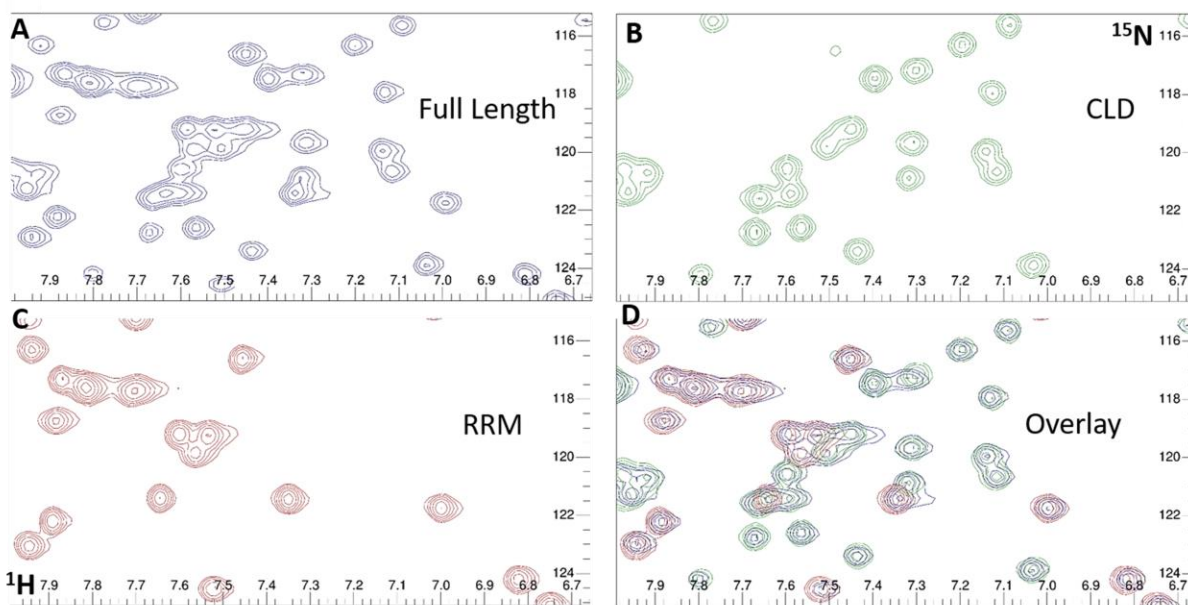


Figure 5.15 CypE-RRM Mapping of the Chemical Shift Changes with SO-1 Reveal Overlapping Surface with Previous Interactions. A) Chemical shift changes upon SO-1 binding mapped onto CypE-RRM structure B) AAUAAA interactions mapped C) PHD3 peptide interactions mapped (3MDF)



**Figure 5.17 CypE-CLD HSQC SO-1 Titration Shows Little to No Interaction.** Free CLD HSQC color blue, with each 0.25 molar ratio step going cyan, green, yellow, orange, and red as the final 1.5 molar ratio point.

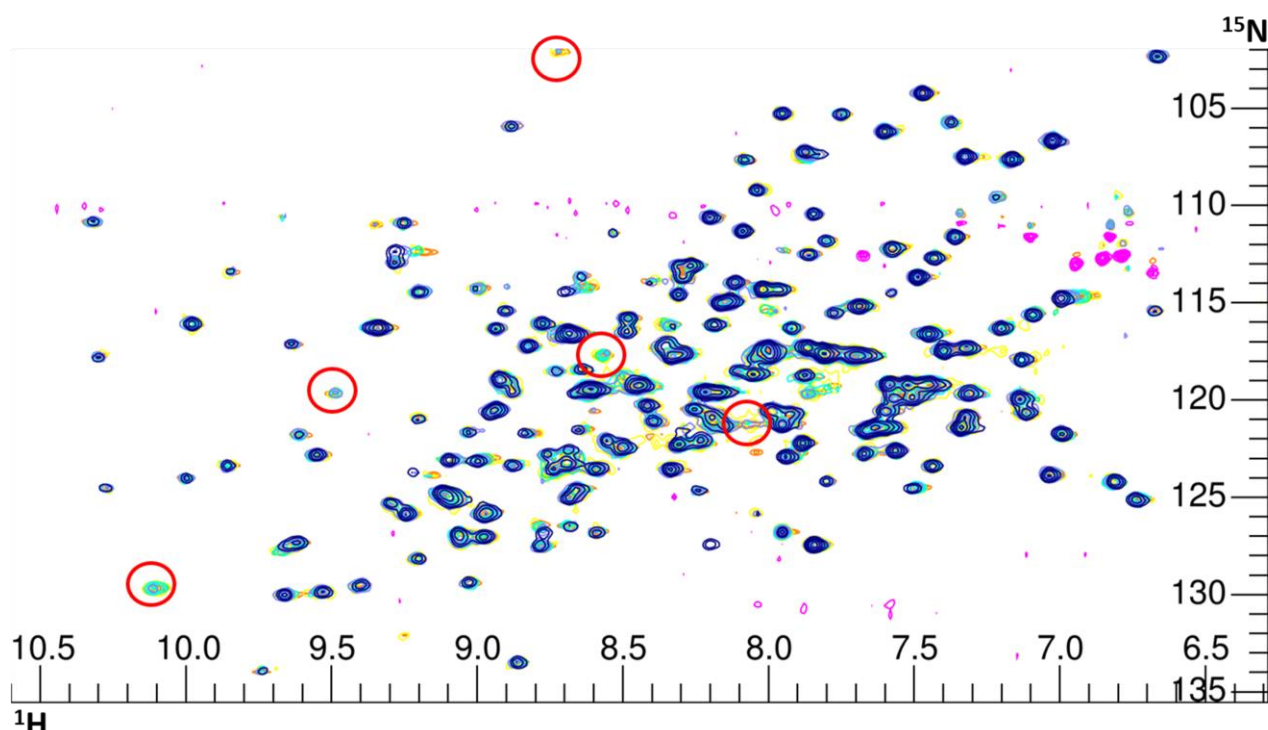


**Figure 5.16 Comparison of Free HSQCs Reveal Independently Behaved Domains** A) FL-CypE B) CypE-CLD C) CypE-RRM D) Overlay of all three.

Notably, the titration of CypE-CLD with SO-1 shows no significant changes in the HSQC spectrum upon the addition of RNA (**Figure 5.16**), strongly suggesting that the two molecules

do not interact with each at a biologically relevant affinity. Beyond the absence of any peak walking or appearance/disappearance of any peaks, there was no observed precipitation in the tube, and the decreased signal-to-noise for the observed peaks is consistent with the dilution of the protein – eliminating the possibility of slow tumbling, invisible complex.

Despite being 36 kDa, full-length protein shows excellent signal-to-noise (**Figure 5.17A**). Overlay of the HSCQ of the full-length with the subdomains reveals that the addition of the subdomains HSQC spectra largely recapitulates full-length spectra, with additional peaks from the CLD likely arising from the C-terminal His-tag not shared by the FL-CypE. This suggests that the two subdomains largely act independently of each other. Remarkably, addition of the 16 kDa SO-1 ligand results in a resolvable peak walking similar to the behavior exhibited by the



**Figure 5.18 SO-1 Binding to FL-CypE by NMR HSQC Titration.** Free FL-CypE HSQC color blue, with each 0.25 molar ratio step going cyan, green, yellow, orange, and red as the final 1.5 molar ratio point. The red circles highlight slow-exchange peaks of unknown origin that appear with the addition of RNA

RRM domain alone. Some notable differences, however, are the appearance of several new peaks for which the residues responsible are unknown (**Figure 5.18**).

## **5.4 – Discussion**

### **5.4.1 – Iteration of the SELEX Protocol Provides Insights into Optimization**

Through our iteration of the SELEX protocol in these three selection trials, we have gained several insights into our library design and selection conditions – resulting in an optimized protocol for CypE. Because CLDs frequently bind peptide sequences in the same  $\mu\text{M}$   $K_D$  range as SO-1,<sup>181,226,245</sup> the conditions used to enrich for it are likely to be amenable for binding of biologically relevant RNAs to other CLDs. Comparison of the enrichment results for CypE in selections 2 and 3 reveals high levels of detectable enrichment of sequences in selection 3 but not 2. As the binding conditions used for selection 2 was replicated in one of the conditions for selection 3, the most likely inappropriate condition in the first two selections is the high 1M salt wash, which likely causes the protocol to be too stringent. In the results from our first iteration, ubiquitous constant-random region pairing revealed that the hope that structured constant regions would fold independently did not prove to be the case in practice – however, it is still unclear if that would be an issue in a selective protocol where a high affinity aptamer was produced. Additional testing with a lower salt wash regime would provide a more appropriate comparison to the annealed primer strategy that also had its failures to prevent constant region annealing in selection 3 as well as in the high salt wash selection for MS2 in which the second most abundant MS2 aptamer utilized the 5' constant region to from the MS2 consensus site.

While tempting to compare the relative low, medium, and high protein concentrations in selection 3, the evidence of cross contamination between those conditions make it challenging to conclude any general trends in the effect of stringency on the relative enrichment of aptamer sequences. However, as all six conditions produced enriched aptamer sequences, it seems like an appropriate selection regime has been determined. Further lowering of the protein

concentrations for future cyclophilin selections may result in additional noise from selection artifacts and background binding of the resin beads in a manner like the effect of a high salt wash.

In future selections, benchmarking the progress of the experiment would provide a lot of valuable feedback during the selection and subsequent validation. Adding a SUMO cleavage and nickel-column clean-up step for FL-CypE appeared to resolve the contamination issue. With SUMO-CypA and SUMO-Cpr1 constructs on hand, a similar purification protocol could work well for those – though SEC may prove less useful due to the similar size of SUMO and the CypA and Cpr1 proteins. Alternatively, the typical purification scheme for recombinant CypA utilizes two ion-exchange column steps for purification, which may also eliminate the contaminant that appears to carry through the nickel affinity step.<sup>181</sup> Replacement of the size-exclusion step with an ion-exchange step could also be sufficient to remove the contaminant carried through in our current protocol.

Speculatively, the most likely contaminant present in our protein stocks is the *E. coli* protein Hfq, for several reasons. Hfq has previously been reported as a common contaminant in nickel-affinity chromatography.<sup>288</sup> In addition, Hfq binds RNA extremely tightly,<sup>258</sup> binding its consensus with a 50 pM  $K_D$  which is consistent with the extremely low contaminant levels indicated by our PAGE-SDS staining and EMSA results. If this is the case, it raises the question of whether the presence of Hfq could have affected our selections. One concerning result is that the Hfq binding motif (AAYAAAYAA)<sup>258</sup> is essentially the motif enriched in all of the selections. However, that might just be a coincidence given the similarity of the CypE published ligand of AAYAAA and its preference for polyU and polyA RNA.<sup>245,285</sup> One feature of Hfq RNA binding does suggest it would probably have a limited impact in the selection as Hfq binds RNA on two faces of a toroidal hexamer and the six histidines that bind the Ni/Co-NTA column are involved in one of those binding interactions. Moreover, the affinity of the SO ligands, SO-1 in particular,



as tighter than the published RNA ligand for CypE strongly suggests CypE-RNA binding was the dominant selective pressure during Selection 3.

#### **5.4.2 – SELEX Sequencing Results Point to Several Promising Binding Motifs**

Excitingly, analysis of the enriched sequence motifs from experiment 3 reveal several strikingly similar, but distinct, motifs compared to the literature consensus sequence. One, AAYAAYAA as well as other AY-rich sequences appear to be an extension of the AAYAAA literature consensus while the motif contained by SO-1 appear to have a GATA core. Counter to the expectation that repeats of the an AAYAA consensus sequence would bind tighter, SO-6 and SO-9, which contain 3-4 potential AAYAA binding sites depending on the register, bind weaker than SO-1 which contains only one similar site. While the aptamer we have validated thus far, SO-1, interacts with the CypE-RRM domain but not the CypE-CLD domain in isolation, the slow exchange peaks that appear in the full-length CypE titration suggests the RRM binding may not be the full picture. Alternatively, the structural context of the RRM binding site in SO-1 may form additional contacts to the protein not present in the weaker binding SO ligands.

#### **5.4.3 – Inconsistent PPlase Assay Warrants Further Controls**

Our observation of CypE activation by RNA agrees with the previous observation that FL-CypE is activated by mRNA.<sup>195</sup> However, we also have evidence this activation appears to occur independently of RNA binding as SO-1 does not interact with the CypE-CLD by NMR titration but does activate its PPlase activity in this assay. While it is possible that RNA interacts very weakly with the CypE-CLD, especially in the context that the substrate itself has a mM  $K_D$ , the population of bound CLD would be extremely small and likely short-lived. Strangely, several genuine binding interactions inhibit PPlase activity in this assay as is the case for CypA and heparin described here, as well as *AtCyp59* when bound to its consensus RNA.<sup>223</sup> These inconsistent results highlight the need for ZZ-exchange experiments with physiological

ligands.<sup>181</sup> Because CypE binds to two MLL PHD3-bromo peptides,<sup>246</sup> studying RNA binding and PPIase activity in that context is an important next step in characterizing CypE regulatory functions.

#### **5.4.4 – NMR Characterization of CypE Interfaces Suggests Possible Mechanisms of RNA Regulation**

The NMR experiments we have performed here have given significant insight into possible mechanisms regarding this function. Comparison of the HSQCs from the two subdomains with full-length protein suggest the two domains behave independently of each other in solution. This has interesting implications for PPIase regulation as it is unclear how the binding of RNA to the RRM could have an allosteric effect on the CLD when they do not appear to share an interface and the linker between them is long and flexible. One possibility is that the RNA itself mediates an effect on the CLD. Definitive assignment of the CLD and the slow-exchange peaks that appear as a result of titration of the full-length protein will be important in testing this hypothesis.

The binding of SO-1 to the RRM also raise other questions. Notably, the increased affinity of SO-1 compared to the published motif puts it within an order of magnitude of the RRM-PHD3 interaction.<sup>245</sup> Moreover, NMR titration with this sequence reveals an extended interface that overlaps with the interface of the RRM-PHD3 interaction. Together, these suggest RNA may be a biologically relevant competitor to the CypE-MLL interaction and could regulate CypE through that mechanism. To that end, minimizing the SO-1 sequence involved in this interaction may allow for identification of RNAs that bind *in vivo* through genomic alignments.

## References

- (1) Lloyd, N. R., Dickey, T. H., Hom, R. A., and Wuttke, D. S. (2016) Tying up the Ends: Plasticity in the Recognition of Single-Stranded DNA at Telomeres. *Biochemistry* 55, 5326–5340.
- (2) Lloyd, N. R., and Wuttke, D. S. (2018) Discrimination against RNA Backbones by a ssDNA Binding Protein. *Structure* 26, 722-733.e2.
- (3) McClintock, B. (1939) The Behavior in Successive Nuclear Divisions of a Chromosome Broken at Meiosis. *Proc. Natl. Acad. Sci.* 25, 405–416.
- (4) Gottschling, D. E., and Zakian, V. A. (1986) Telomere Proteins: Specific Recognition and Protection of the Natural Termini of *Oxytricha* Macronuclear DNA. *Cell* 47, 195–205.
- (5) McElligott, R., and Wellinger, R. J. (1997) The terminal DNA structure of mammalian chromosomes. *EMBO J.* 16, 3705–3714.
- (6) Palm, W., and de Lange, T. (2008) How Shelterin Protects Mammalian Telomeres. *Annu. Rev. Genet.* 42, 301–334.
- (7) Blackburn, E. H., Epel, E. S., and Lin, J. (2015) Human telomere biology: A contributory and interactive factor in aging, disease risks, and protection. *Science* 350, 1193–1198.
- (8) Watson, J. D. (1972) Origin of Concatemeric T7 DNA. *Nature. New Biol.* 239, 197–201.
- (9) Hayflick, L., and Moorhead, P. (1973) The Serial Cultivation of Human Diploid. *Read. Mamm. Cell Cult.* 25, 66.
- (10) Olovnikov, A. M. (1973) A theory of marginotomy: the incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon. *J. Theor. Biol.* 41, 181–190.

- (11) Makarov, V. L., Hirose, Y., and Langmore, J. P. (1997) Long G tails at both ends of human chromosomes suggest a C strand degradation mechanism for telomere shortening. *Cell* 88, 657–666.
- (12) Moyzis, R. K., Buckingham, J. M., Cram, L. S., Dani, M., Deaven, L. L., Jones, M. D., Meyne, J., Ratliff, R. L., and Wu, J.-R. (1988) A highly conserved repetitive DNA sequence, (TTAGGG)<sub>n</sub>, present at the telomeres of human chromosomes. *Proc. Natl. Acad. Sci.* 85, 6622–6626.
- (13) Van Steensel, B., Smogorzewska, A., and de Lange, T. (1998) TRF2 Protects Human Telomeres from End-to-End Fusions. *Cell* 92, 401–413.
- (14) Smogorzewska, A., Karlseder, J., Holtgreve-Grez, H., Jauch, A., and de Lange, T. (2002) DNA Ligase IV-Dependent NHEJ of Deprotected Mammalian Telomeres in G1 and G2. *Curr. Biol.* 12, 1635–1644.
- (15) Tomita, K., and Cooper, J. P. (2008) Fission yeast Ccq1 is telomerase recruiter and local checkpoint controller. *Genes Dev.* 22, 3461–3474.
- (16) Kelleher, C., Kurth, I., and Lingner, J. (2005) Human Protection of Telomeres 1 (POT1) Is a Negative Regulator of Telomerase Activity *In Vitro*. *Mol. Cell. Biol.* 25, 808–818.
- (17) Wang, F., Podell, E. R., Zaug, A. J., Yang, Y., Baciú, P., Cech, T. R., and Lei, M. (2007) The POT1–TPP1 telomere complex is a telomerase processivity factor. *Nature* 445, 506–510.
- (18) Miyoshi, T., Kanoh, J., Saito, M., and Ishikawa, F. (2008) Fission Yeast Pot1-Tpp1 Protects Telomeres and Regulates Telomere Length. *Science* 320, 1341–1344.
- (19) Brutlag, D., Schekman, R., and Kornberg, A. (1971) A possible role for RNA polymerase in the initiation of M13 DNA synthesis. *Proc. Natl. Acad. Sci.* 68, 2826–2829.
- (20) Larrivé, M., LeBel, C., and Wellinger, R. J. (2004) The generation of proper constitutive G-tails on yeast telomeres is dependent on the MRX complex. *Genes Dev.* 18, 1391–1396.

- (21) Wu, P., van Overbeek, M., Rooney, S., and de Lange, T. (2010) Apollo Contributes to G Overhang Maintenance and Protects Leading-End Telomeres. *Mol. Cell* 39, 606–617.
- (22) Lam, Y. C., Akhter, S., Gu, P., Ye, J., Poulet, A., Jose`phe, M., Panis, G., Bailey, S. M., Gilson, E., Legerski, R. J., and Chang, S. (2010) SNMIB/Apollo protects leading-strand telomeres against NHEJ-mediated repair. *EMBO J.* 29, 2230–2241.
- (23) Wu, P., Takai, H., and de Lange, T. (2012) Telomeric 3' Overhangs Derive from Resection by Exo1 and Apollo and Fill-In by POT1b-Associated CST. *Cell* 150, 39–52.
- (24) Bonetti, D., Martina, M., Falcettoni, M., and Longhese, M. P. (2014) Telomere-end processing: mechanisms and regulation. *Chromosoma* 123, 57–66.
- (25) Sfeir, A. J., Chai, W., Shay, J. W., and Wright, W. E. (2005) Telomere-End Processing: the Terminal Nucleotides of Human Chromosomes. *Mol. Cell* 18, 131–138.
- (26) Bodnar, A. G., Ouellette, M., Frolkis, M., Holt, S. E., Chiu, C.-P., Morin, G. B., Harley, C. B., Shay, J. W., Lichtsteiner, S., and Wright, W. E. (1998) Extension of life-span by introduction of telomerase into normal human cells. *Science* 279, 349–352.
- (27) Lindsey, J., McGill, N. I., Lindsey, L. A., Green, D. K., and Cooke, H. J. (1991) *In vivo* loss of telomeric repeats with age in humans. *Mutat. Res.* 256, 45–48.
- (28) Baker, D. J., Wijshake, T., Tchkonja, T., LeBrasseur, N. K., Childs, B. G., van de Sluis, B., Kirkland, J. L., and van Deursen, J. M. (2011) Clearance of p16Ink4a-positive senescent cells delays ageing-associated disorders. *Nature* 479, 232–236.
- (29) Greider, C. W., and Blackburn, E. H. (1985) Identification of a Specific Telomere Terminal Transferase Activity in *Tetrahymena* Extracts. *Cell* 43, 405–413.
- (30) Greider, C. W., and Blackburn, E. H. (1989) A telomeric sequence in the RNA of *tetrahymena* telomerase required for telomere repeat synthesis. *Nature* 337, 331–337.
- (31) Yu, G.-L., Bradley, J. D., Attardi, L. D., and Blackburn, E. H. (1990) *In vivo* alteration of telomere sequences and senescence caused by mutated *tetrahymena* telomere RNAs. *Nature* 344, 126–132.

- (32) Lingner, J., Hughes, T. R., Shevchenko, A., Mann, M., Lundblad, V., and Cech, T. R. (1997) Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science* 276, 561–567.
- (33) Martin, A. A., Dionne, I., Wellinger, R. J., and Holm, C. (2000) The function of DNA polymerase  $\alpha$  at telomeric G tails is important for telomere homeostasis. *Mol. Cell. Biol.* 20, 786–796.
- (34) Greider, C. W. (1991) Telomerase is processive. *Mol. Cell. Biol.* 11, 4572–4580.
- (35) De Lange, T., Shiue, L., Myers, R. M., Cox, D. R., Naylor, S. L., Killery, A. M., and Varmus, H. E. (1990) Structure and variability of human chromosome ends. *Mol. Cell. Biol.* 10, 518–527.
- (36) Feng, J., Funk, W. D., Wang, S.-S., Weinrich, S. L., Avilion, A. A., Chiu, C.-P., Adams, R. R., Chang, E., Allsopp, R. C., Yu, J., and others. (1995) The RNA component of human telomerase. *Science* 269, 1236–1241.
- (37) Kim, N. W., Piatyszek, M. A., Prowse, K. R., Harley, C. B., West, M. D., Ho, P. d L., Coviello, G. M., Wright, W. E., Weinrich, S. L., and Shay, J. W. (1994) Specific association of human telomerase activity with immortal cells and cancer. *Science* 266, 2011–2015.
- (38) Ruden, M., and Puri, N. (2013) Novel anticancer therapeutics targeting telomerase. *Cancer Treat. Rev.* 39, 444–456.
- (39) Rousseau, P., and Autexier, C. (2015) Telomere biology: Rationale for diagnostics and therapeutics in cancer. *RNA Biol.* 12, 1078–1082.
- (40) Shampay, J., Szostak, J. W., and Blackburn, E. H. (1984) DNA sequences of telomeres maintained in yeast. *Nature* 310, 154–157.
- (41) Singer, M. S., and Gottschling, D. E. (1994) TLC1: Template RNA component of *Saccharomyces cerevisiae* Telomerase. *Science* 266, 404–409.

- (42) Cooper, J. P., Nimmo, E. R., Allshire, R. C., and Cech, T. R. (1997) Regulation of telomere length and function by a Myb-domain protein in fission yeast. *Nature* 385, 744–747.
- (43) Leonardi, J., Box, J. A., Bunch, J. T., and Baumann, P. (2008) TER1, the RNA subunit of fission yeast telomerase. *Nat. Struct. Mol. Biol.* 15, 26–33.
- (44) Trujillo, K. M., Bunch, J. T., and Baumann, P. (2005) Extended DNA Binding Site in Pot1 Broadens Sequence Specificity to Allow Recognition of Heterogeneous Fission Yeast Telomeres. *J. Biol. Chem.* 280, 9119–9128.
- (45) Webb, C. J., and Zakian, V. A. (2008) Identification and characterization of the *Schizosaccharomyces pombe* TER1 telomerase RNA. *Nat. Struct. Mol. Biol.* 15, 34–42.
- (46) Wellinger, R. J., Wolf, A. J., and Zakian, V. A. (1993) Origin Activation and Formation of Single-Strand TG<sub>1-3</sub> Tails Occur Sequentially in Late S Phase on a Yeast Linear Plasmid. *Mol. Cell. Biol.* 13, 4057–4065.
- (47) Wellinger, R. J., Wolf, A. J., and Zakian, V. A. (1993) *Saccharomyces* Telomeres Acquire Single-Strand TG<sub>1-3</sub> Tails in Late S Phase. *Cell* 72, 51–60.
- (48) Forstemann, K., and Lingner, J. (2001) Molecular Basis for Telomere Repeat Divergence in Budding Yeast. *Mol. Cell. Biol.* 21, 7277–7286.
- (49) Wellinger, R. J., and Zakian, V. A. (2012) Everything You Ever Wanted to Know About *Saccharomyces cerevisiae* Telomeres: Beginning to End. *Genetics* 191, 1073–1105.
- (50) Lim, C. J., Zaug, A. J., Kim, H. J., and Cech, T. R. (2017) Reconstitution of human shelterin complexes reveals unexpected stoichiometry and dual pathways to enhance telomerase processivity. *Nat. Commun.* 8.
- (51) Stansel, Rachel M., de Lange, T., and Griffith, Jack D. (2001) T-loop assembly *in vitro* involves the binding of TRF2 near the 3' telomeric end. *EMBO J.* 20, 5532–5540.
- (52) Denchi, E. L., and de Lange, T. (2007) Protection of telomeres through independent control of ATM and ATR by TRF2 and POT1. *Nature* 448, 1068–1071.

(53) Amiard, S., Doudeau, M., Pinte, S., Poulet, A., Lenain, C., Faivre-Moskalenko, C., Angelov, D., Hug, N., Vindigni, A., Bouvet, P., Paoletti, J., Gilson, E., and Giraud-Panis, M.-J. (2007) A topological mechanism for TRF2-enhanced strand invasion. *Nat. Struct. Mol. Biol.* 14, 147–154.

(54) Benarroch-Popivker, D., Pisano, S., Mendez-Bermudez, A., Lototska, L., Kaur, P., Bauwens, S., Djerbi, N., Latrick, C. M., Fraasier, V., Pei, B., Gay, A., Jaune, E., Foucher, K., Cherfils-Vicini, J., Aeby, E., Miron, S., Londoño-Vallejo, A., Ye, J., Le Du, M.-H., Wang, H., Gilson, E., and Giraud-Panis, M.-J. (2016) TRF2-Mediated Control of Telomere DNA Topology as a Mechanism for Chromosome-End Protection. *Mol. Cell* 61, 274–286.

(55) Churikov, D., Wei, C., and Price, C. M. (2006) Vertebrate POT1 Restricts G-Overhang Length and Prevents Activation of a Telomeric DNA Damage Checkpoint but Is Dispensable for Overhang Protection. *Mol. Cell. Biol.* 26, 6971–6982.

(56) Denchi, E. L., and de Lange, T. (2007) Protection of telomeres through independent control of ATM and ATR by TRF2 and POT1. *Nature* 448, 1068–1071.

(57) Gong, Y., and de Lange, T. (2010) A Shld1-Controlled POT1a Provides Support for Repression of ATR Signaling at Telomeres through RPA Exclusion. *Mol. Cell* 40, 377–387.

(58) Loayza, D., and De Lange, T. (2003) POT1 as a terminal transducer of TRF1 telomere length control. *Nature* 423, 1013–1018.

(59) Hockemeyer, D., Sfeir, A. J., Shay, J. W., Wright, W. E., and de Lange, T. (2005) POT1 protects telomeres from a transient DNA damage response and determines how human chromosomes end. *EMBO J.* 24, 2667–2678.

(60) Murzin, A. G. (1993) OB (oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *EMBO J.* 12, 861.

(61) Theobald, D. L., Mitton-Fry, R. M., and Wuttke, D. S. (2003) Nucleic Acid Recognition by OB-Fold Proteins. *Annu. Rev. Biophys. Biomol. Struct.* 32, 115–133.

(62) Theobald, D. L., and Wuttke, D. S. (2004) Prediction of Multiple Tandem OB-Fold Domains in Telomere End-Binding Proteins Pot1 and Cdc13. *Structure* 12, 1877–1879.



- (63) Xin, H., Liu, D., Wan, M., Safari, A., Kim, H., Sun, W., O'Connor, M. S., and Songyang, Z. (2007) TPP1 is a homologue of ciliate TEBP- $\beta$  and interacts with POT1 to recruit telomerase. *Nature* 445, 559–562.
- (64) Lei, M., Podell, E. R., Baumann, P., and Cech, T. R. (2003) DNA self-recognition in the structure of Pot1 bound to telomeric single-stranded DNA. *Nature* 426, 198–203.
- (65) Lei, M., Podell, E. R., and Cech, T. R. (2004) Structure of human POT1 bound to telomeric single-stranded DNA provides a model for chromosome end-protection. *Nat. Struct. Mol. Biol.* 11, 1223–1229.
- (66) Dickey, T. H., McKercher, M. A., and Wuttke, D. S. (2013) Nonspecific Recognition Is Achieved in Pot1pC through the Use of Multiple Binding Modes. *Structure* 21, 121–132.
- (67) Ramsay, A. J., Quesada, V., Foronda, M., Conde, L., Martínez-Trillos, A., Villamor, N., Rodríguez, D., Kwarciak, A., Garabaya, C., Gallardo, M., López-Guerra, M., López-Guillermo, A., Puente, X. S., Blasco, M. A., Campo, E., and López-Otín, C. (2013) POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. *Nat. Genet.* 45, 526–530.
- (68) Speedy, H. E., Di Bernardo, M. C., Sava, G. P., Dyer, M. J. S., Holroyd, A., Wang, Y., Sunter, N. J., Mansouri, L., Juliusson, G., Smedby, K. E., Roos, G., Jayne, S., Majid, A., Dearden, C., Hall, A. G., Mainou-Fowler, T., Jackson, G. H., Summerfield, G., Harris, R. J., Pettitt, A. R., Allsup, D. J., Bailey, J. R., Pratt, G., Pepper, C., Fegan, C., Rosenquist, R., Catovsky, D., Allan, J. M., and Houlston, R. S. (2013) A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. *Nat. Genet.* 46, 56–60.
- (69) Bainbridge, M. N., Armstrong, G. N., Gramatges, M. M., Bertuch, A. A., Jhangiani, S. N., Doddapaneni, H., Lewis, L., Tombrello, J., Tsavachidis, S., Liu, Y., Jalali, A., Plon, S. E., Lau, C. C., Parsons, D. W., Claus, E. B., Barnholtz-Sloan, J., Il'yasova, D., Schildkraut, J., Ali-Osman, F., Sadetzki, S., Johansen, C., Houlston, R. S., Jenkins, R. B., Lachance, D., Olson, S. H., Bernstein, J. L., Merrell, R. T., Wrensch, M. R., Walsh, K. M., Davis, F. G., Lai, R., Shete, S., Aldape, K., Amos, C. I., Thompson, P. A., Muzny, D. M., Gibbs, R. A., Melin, B. S., Bondy, M. L., and The Gliogene Consortium. (2014) Germline Mutations in Shelterin Complex Genes Are Associated With Familial Glioma. *JNCI J. Natl. Cancer Inst.* 107, dju384–dju384.

(70) Shi, J., Yang, X. R., Ballew, B., Rotunno, M., Calista, D., Fagnoli, M. C., Ghiorzo, P., Bressac-de Paillerets, B., Nagore, E., Avril, M. F., Caporaso, N. E., McMaster, M. L., Cullen, M., Wang, Z., Zhang, X., Bruno, W., Pastorino, L., Queirolo, P., Banuls-Roca, J., Garcia-Casado, Z., Vaysse, A., Mohamdi, H., Riazalhosseini, Y., Foglio, M., Jouenne, F., Hua, X., Hyland, P. L., Yin, J., Vallabhaneni, H., Chai, W., Minghetti, P., Pellegrini, C., Ravichandran, S., Eggermont, A., Lathrop, M., Peris, K., Scarra, G. B., Landi, G., Savage, S. A., Sampson, J. N., He, J., Yeager, M., Goldin, L. R., Demenais, F., Chanock, S. J., Tucker, M. A., Goldstein, A. M., Liu, Y., and Landi, M. T. (2014) Rare missense variants in POT1 predispose to familial cutaneous malignant melanoma. *Nat. Genet.* 46, 482–486.

(71) Robles-Espinoza, C. D., Harland, M., Ramsay, A. J., Aoude, L. G., Quesada, V., Ding, Z., Pooley, K. A., Pritchard, A. L., Tiffen, J. C., Petljak, M., Palmer, J. M., Symmons, J., Johansson, P., Stark, M. S., Gartside, M. G., Snowden, H., Montgomery, G. W., Martin, N. G., Liu, J. Z., Choi, J., Makowski, M., Brown, K. M., Dunning, A. M., Keane, T. M., López-Otín, C., Gruis, N. A., Hayward, N. K., Bishop, D. T., Newton-Bishop, J. A., and Adams, D. J. (2014) POT1 loss-of-function variants predispose to familial melanoma. *Nat. Genet.* 46, 478–481.

(72) Takai, H., Jenkinson, E., Kabir, S., Babul-Hirji, R., Najm-Tehrani, N., Chitayat, D. A., Crow, Y. J., and de Lange, T. (2016) A POT1 mutation implicates defective telomere end fill-in and telomere truncations in Coats plus. *Genes Dev.* 30, 812–826.

(73) Croy, J. E., Podell, E. R., and Wuttke, D. S. (2006) A New Model for *Schizosaccharomyces pombe* Telomere Recognition: The Telomeric Single-stranded DNA-Binding Activity of Pot11-389. *J. Mol. Biol.* 361, 80–93.

(74) Altschuler, S. E., Dickey, T. H., and Wuttke, D. S. (2011) *Schizosaccharomyces pombe* Protection of Telomeres 1 Utilizes Alternate Binding Modes To Accommodate Different Telomeric Sequences. *Biochemistry* 50, 7503–7513.

(75) Dickey, T. H., and Wuttke, D. S. (2014) The telomeric protein Pot1 from *Schizosaccharomyces pombe* binds ssDNA in two modes with differing 3' end availability. *Nucleic Acids Res.* 42, 9656–9665.

(76) Loayza, D., Parsons, H., Donigian, J., Hoke, K., and de Lange, T. (2004) DNA Binding Features of Human POT1: A nonamer 5'-TAGGGTTAG-3' Minimal Binding Site, Sequence Specificity, and Internal Binding Site to Multimeric Sites. *J. Biol. Chem.* 279, 13241–13248.

(77) Croy, J. E., Altschuler, S. E., Grimm, N. E., and Wuttke, D. S. (2009) Nonadditivity in the Recognition of Single-Stranded DNA by the *Schizosaccharomyces pombe* Protection of Telomeres 1 DNA-Binding Domain, Pot1-DBD. *Biochemistry* 48, 6864–6875.

(78) Holm, L., and Rosenstrom, P. (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res.* 38, W545–W549.

(79) Dickey, Thanye H. (2014) Structural Plasticity in the Recognition of ssDNA by the Telomeric Protein Pot1. Ph.D. Dissertation, University of Colorado Boulder, Boulder, CO.

(80) Messias, A. C., and Sattler, M. (2004) Structural Basis of Single-Stranded RNA Recognition. *Acc. Chem. Res.* 37, 279–287.

(81) Cléry, A., Blatter, M., and Allain, F. H.-T. (2008) RNA recognition motifs: boring? Not quite. *Curr. Opin. Struct. Biol.* 18, 290–298.

(82) Record, T. M., Lohman, T. M., and De Haseth, P. (1976) Ion Effects on Ligand-Nucleic Acid Interactions. *J. Mol. Biol.* 107, 145–158.

(83) Croy, J. E., and Wuttke, D. S. (2006) Themes in ssDNA recognition by telomere-end protection proteins. *Trends Biochem. Sci.* 31, 516–525.

(84) Chen, R., and Wold, M. S. (2014) Replication protein A: Single-stranded DNA's first responder: Dynamic DNA-interactions allow replication protein A to direct single-strand DNA intermediates into different pathways for synthesis or repair. *BioEssays* 36, 1156–1161.

(85) Lei, M., Zaug, A. J., Podell, E. R., and Cech, T. R. (2005) Switching Human Telomerase On and Off with hPOT1 Protein *in Vitro*. *J. Biol. Chem.* 280, 20449–20456.

(86) Latrick, C. M., and Cech, T. R. (2010) POT1–TPP1 enhances telomerase processivity by slowing primer dissociation and aiding translocation. *EMBO J.* 29, 924–933.

- (87) Hwang, H., Buncher, N., Opresko, P. L., and Myong, S. (2012) POT1-TPP1 Regulates Telomeric Overhang Structural Dynamics. *Structure* 20, 1872–1880.
- (88) Nandakumar, J., Bell, C. F., Weidenfeld, I., Zaug, A. J., Leinwand, L. A., and Cech, T. R. (2012) The TEL patch of telomere protein TPP1 mediates telomerase recruitment and processivity. *Nature* 492, 285–289.
- (89) Zhong, F. L., Batista, L. F. Z., Freund, A., Pech, M. F., Venteicher, A. S., and Artandi, S. E. (2012) TPP1 OB-Fold Domain Controls Telomere Maintenance by Recruiting Telomerase to Chromosome Ends. *Cell* 150, 481–494.
- (90) Sexton, A. N., Youmans, D. T., and Collins, K. (2012) Specificity Requirements for Human Telomere Protein Interaction with Telomerase Holoenzyme. *J. Biol. Chem.* 287, 34455–34464.
- (91) Zhang, Y., Chen, L.-Y., Han, X., Xie, W., Kim, H., Yang, D., Liu, D., and Songyang, Z. (2013) Phosphorylation of TPP1 regulates cell cycle-dependent telomerase recruitment. *Proc. Natl. Acad. Sci.* 110, 5457–5462.
- (92) Nandakumar, J., and Cech, T. R. (2012) DNA-induced dimerization of the single-stranded DNA binding telomeric protein Pot1 from *Schizosaccharomyces pombe*. *Nucleic Acids Res.* 40, 235–244.
- (93) Flynn, R. L., Centore, R. C., O’Sullivan, R. J., Rai, R., Tse, A., Songyang, Z., Chang, S., Karlseder, J., and Zou, L. (2011) TERRA and hnRNPA1 orchestrate an RPA-to-POT1 switch on telomeric single-stranded DNA. *Nature* 471, 532–536.
- (94) Nandakumar, J., Podell, E. R., and Cech, T. R. (2010) How telomeric protein POT1 avoids RNA to achieve specificity for single-stranded DNA. *Proc. Natl. Acad. Sci.* 107, 651–656.
- (95) Lee, M., Hills, M., Conomos, D., Stutz, M. D., Dagg, R. A., Lau, L. M. S., Reddel, R. R., and Pickett, H. A. (2014) Telomere extension by telomerase and ALT generates variant repeats by mechanistically distinct processes. *Nucleic Acids Res.* 42, 1733–1746.
- (96) Luke, B., and Lingner, J. (2009) TERRA: telomeric repeat-containing RNA. *EMBO J.* 28, 2503.

- (97) Bah, A., Wischnewski, H., Shchepachev, V., and Azzalin, C. M. (2012) The telomeric transcriptome of *Schizosaccharomyces pombe*. *Nucleic Acids Res.* *40*, 2995–3005.
- (98) Cusanelli, E., and Chartrand, P. (2015) Telomeric repeat-containing RNA TERRA: a noncoding RNA connecting telomere biology to genome integrity. *Front. Genet.* *6*.
- (99) Michelini, F., Jalihal, A. P., Francia, S., Meers, C., Neeb, Z. T., Rossiello, F., Gioia, U., Aguado, J., Jones-Weinert, C., Luke, B., Biamonti, G., Nowacki, M., Storici, F., Carninci, P., Walter, N. G., and d’Adda di Fagagna, F. (2018) From “Cellular” RNA to “Smart” RNA: Multiple Roles of RNA in Genome Stability and Beyond. *Chem. Rev.* *118*, 4365–4403.
- (100) Smekalova, E., and Baumann, P. (2013) TERRA –A Calling Card for Telomerase. *Mol. Cell* *51*, 703–704.
- (101) Graf, M., Bonetti, D., Lockhart, A., Serhal, K., Kellner, V., Maicher, A., Jolivet, P., Teixeira, M. T., and Luke, B. (2017) Telomere Length Determines TERRA and R-Loop Regulation through the Cell Cycle. *Cell* *170*, 72-85.e14.
- (102) Moravec, M., Wischnewski, H., Bah, A., Hu, Y., Liu, N., Lafranchi, L., King, M. C., and Azzalin, C. M. (2016) TERRA promotes telomerase-mediated telomere elongation in *Schizosaccharomyces pombe*. *EMBO Rep.* *17*, 999.
- (103) Pfeiffer, V., and Lingner, J. (2012) TERRA Promotes Telomere Shortening through Exonuclease 1–Mediated Resection of Chromosome Ends. *PLOS Genet.* *8*, e1002747.
- (104) Montero, J. J., López-Silanes, I., Megías, D., F. Fraga, M., Castells-García, Á., and Blasco, M. A. (2018) TERRA recruitment of polycomb to telomeres is essential for histone trimethylation marks at telomeric heterochromatin. *Nat. Commun.* *9*.
- (105) Altschuler, S. E. (2011) Characterization of single-stranded DNA binding and small molecule inhibition of *S. pombe* Pot1. University of Colorado Boulder.
- (106) Wold, M. S. (1997) Replication protein A: a heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. *Annu. Rev. Biochem.* *66*, 61–92.

- (107) Bochkarev, A., and Bochkareva, E. (2004) From RPA to BRCA2: lessons from single-stranded DNA binding by the OB-fold. *Curr. Opin. Struct. Biol.* 14, 36–42.
- (108) Beese, L. S., Derbyshire, V., and Steitz, T. A. (1993) Structure of DNA polymerase I Klenow fragment bound to duplex DNA. *Sci.-N. Y. THEN Wash.* 260, 352–352.
- (109) Wilson, K. A., Kellie, J. L., and Wetmore, S. D. (2014) DNA–protein  $\pi$ -interactions in nature: abundance, structure, composition and strength of contacts between aromatic amino acids and DNA nucleobases or deoxyribose sugar. *Nucleic Acids Res.* 42, 6726–6741.
- (110) Batty, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R., and Leslie, A. G. W. (2011) *iMOSFLM*: a new graphical interface for diffraction-image processing with *MOSFLM*. *Acta Crystallogr. D Biol. Crystallogr.* 67, 271–281.
- (111) Evans, P. R. (2011) An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr. D Biol. Crystallogr.* 67, 282–292.
- (112) McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) *Phaser* crystallographic software. *J. Appl. Crystallogr.* 40, 658–674.
- (113) Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) *PHENIX*: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* 66, 213–221.
- (114) Terwilliger, T. (2004) SOLVE and RESOLVE: automated structure solution, density modification and model building. *J. Synchrotron Radiat.* 11, 49–52.
- (115) Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H., and Adams, P. D. (2012) Towards automated crystallographic structure refinement with *phenix.refine*. *Acta Crystallogr. D Biol. Crystallogr.* 68, 352–367.

- (116) Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010) Features and development of *Coot*. *Acta Crystallogr. D Biol. Crystallogr.* 66, 486–501.
- (117) Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) *MolProbity*: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* 66, 12–21.
- (118) Wilinski, D., Qiu, C., Lapointe, C. P., Nevil, M., Campbell, Z. T., Tanaka Hall, T. M., and Wickens, M. (2015) RNA regulatory networks diversified through curvature of the PUF protein scaffold. *Nat. Commun.* 6, 8213.
- (119) Theobald, D. L., and Schultz, S. C. (2003) Nucleotide shuffling and ssDNA recognition in *Oxytricha nova* telomere end-binding protein complexes. *EMBO J.* 22, 4314–4324.
- (120) McKercher, M. A., Guan, X., Tan, Z., and Wuttke, D. S. (2017) Diversity in peptide recognition by the SH2 domain of SH2B1. *Proteins Struct. Funct. Bioinforma.*
- (121) Anderson, E. M., Halsey, W. A., and Wuttke, D. S. (2003) Site-Directed Mutagenesis Reveals the Thermodynamic Requirements for Single-Stranded DNA Recognition by the Telomere-Binding Protein Cdc13. *Biochemistry* 42, 3751–3758.
- (122) Wan, B., Tang, T., Upton, H., Shuai, J., Zhou, Y., Li, S., Chen, J., Brunzelle, J. S., Zeng, Z., Collins, K., Wu, J., and Lei, M. (2015) The Tetrahymena telomerase p75–p45–p19 subcomplex is a unique CST complex. *Nat. Struct. Mol. Biol.* 22, 1023–1026.
- (123) Hom, R. A., and Wuttke, D. S. (2017) Human CST Prefers G-Rich but Not Necessarily Telomeric Sequences. *Biochemistry* 56, 4210–4218.
- (124) Baltz, A. G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., Wyler, E., Bonneau, R., Selbach, M., Dieterich, C., and Landthaler, M. (2012) The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Mol. Cell* 46, 674–690.
- (125) Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B. M., Strein, C., Davey, N. E., Humphreys, D. T., Preiss, T., Steinmetz, L. M., Krijgsveld, J., and Hentze,

M. W. (2012) Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell* 149, 1393–1406.

(126) Kwon, S. C., Yi, H., Eichelbaum, K., Föhr, S., Fischer, B., You, K. T., Castello, A., Krijgsveld, J., Hentze, M. W., and Kim, V. N. (2013) The RNA-binding protein repertoire of embryonic stem cells. *Nat. Struct. Mol. Biol.* 20, 1122–1130.

(127) Mitchell, S. F., Jain, S., She, M., and Parker, R. (2013) Global analysis of yeast mRNPs. *Nat. Struct. Mol. Biol.* 20, 127–133.

(128) Bell, L. R., Maine, E. M., Schedl, P., and Cline, T. W. (1988) Sex-lethal, a *Drosophila* sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins. *Cell* 55, 1037–1046.

(129) Fu, X.-D., and Ares, M. (2014) Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.* 15, 689–701.

(130) Glisovic, T., Bachorik, J. L., Yong, J., and Dreyfuss, G. (2008) RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* 582, 1977–1986.

(131) Babitzke, P., Baker, C. S., and Romeo, T. (2009) Regulation of Translation Initiation by RNA Binding Proteins. *Annu. Rev. Microbiol.* 63, 27–44.

(132) Harvey, R. F., Smith, T. S., Mulrone, T., Queiroz, R. M. L., Pizzinga, M., Dezi, V., Villeneuve, E., Ramakrishna, M., Lilley, K. S., and Willis, A. E. (2018) *Trans*-acting translational regulatory RNA binding proteins. *Wiley Interdiscip. Rev. RNA* 9, e1465.

(133) Colgan, D. F., and Manley, J. L. (1997) Mechanism and regulation of mRNA polyadenylation. *Genes Dev.* 11, 2755–2766.

(134) Minvielle-Sebastia, L., and Keller, W. (1999) mRNA polyadenylation and its coupling to other RNA processing reactions and to transcription. *Curr. Opin. Cell Biol.* 11, 352–357.

(135) Hasan, A., Cotobal, C., Duncan, C. D. S., and Mata, J. (2014) Systematic Analysis of the Role of RNA-Binding Proteins in the Regulation of RNA Stability. *PLoS Genet.* (Copenhaver, G. P., Ed.) 10, e1004684.



- (136) Pal-Bhadra, M., Bhadra, U., and Birchler, J. A. (2002) RNAi Related Mechanisms Affect Both Transcriptional and Posttranscriptional Transgene Silencing in *Drosophila*. *Mol. Cell* 9, 315–327.
- (137) Zilberman, D. (2003) ARGONAUTE4 Control of Locus-Specific siRNA Accumulation and DNA and Histone Methylation. *Science* 299, 716–719.
- (138) Bernstein, E. (2005) RNA meets chromatin. *Genes Dev.* 19, 1635–1655.
- (139) Lee, J. T. (2012) Epigenetic Regulation by Long Noncoding RNAs 338, 6.
- (140) Chen, Y., and Varani, G. (2005) Protein families and RNA recognition: Protein families and RNA recognition. *FEBS J.* 272, 2088–2097.
- (141) Lunde, B. M., Moore, C., and Varani, G. (2007) RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* 8, 479–490.
- (142) Flynn, R. L., and Zou, L. (2010) Oligonucleotide/oligosaccharide-binding fold proteins: a growing family of genome guardians. *Crit. Rev. Biochem. Mol. Biol.* 45, 266–275.
- (143) Maris, C., Dominguez, C., and Allain, F. H.-T. (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression: The RRM domain, a plastic RNA-binding platform. *FEBS J.* 272, 2118–2131.
- (144) Daubner, G. M., Cléry, A., and Allain, F. H.-T. (2013) RRM–RNA recognition: NMR or crystallography...and new findings. *Curr. Opin. Struct. Biol.* 23, 100–108.
- (145) Brown, R. S. (2005) Zinc finger proteins: getting a grip on RNA. *Curr. Opin. Struct. Biol.* 15, 94–98.
- (146) Hall, T. M. T. (2005) Multiple modes of RNA recognition by zinc finger proteins. *Curr. Opin. Struct. Biol.* 15, 367–373.

- (147) Font, J., and Mackay, J. P. (2010) Beyond DNA: Zinc Finger Domains as RNA-Binding Modules, in *Engineered Zinc Finger Proteins* (Mackay, J. P., and Segal, D. J., Eds.), pp 479–491. Humana Press, Totowa, NJ.
- (148) Thisted, T., Lyakhov, D. L., and Liebhaber, S. A. (2001) Optimized RNA Targets of Two Closely Related Triple KH Domain Proteins, Heterogeneous Nuclear Ribonucleoprotein K and  $\alpha$ CP-2KL, Suggest Distinct Modes of RNA Recognition. *J. Biol. Chem.* 276, 17484–17496.
- (149) Valverde, R., Edwards, L., and Regan, L. (2008) Structure and function of KH domains: Structure and function of KH domains. *FEBS J.* 275, 2712–2726.
- (150) Rytter, J. M. (1998) Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. *EMBO J.* 17, 7505–7513.
- (151) Masliah, G., Barraud, P., and Allain, F. H.-T. (2012) RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cell. Mol. Life Sci.*
- (152) Banerjee, S., and Barraud, P. (2014) Functions of double-stranded RNA-binding domains in nucleocytoplasmic transport. *RNA Biol.* 11, 1226–1232.
- (153) Weiss, M. A., and Narayana, N. (1998) RNA recognition by arginine-rich peptide motifs. *Biopolymers* 48, 167–180.
- (154) Koeller, D. M., Casey, J. L., Hentze, M. W., Gerhardt, E. M., Chan, L. N., Klausner, R. D., and Harford, J. B. (1989) A cytosolic protein binds to structural elements within the iron regulatory region of the transferrin receptor mRNA. *Proc. Natl. Acad. Sci.* 86, 3574–3578.
- (155) Walden, W. E., Selezneva, A. I., Dupuy, J., Volbeda, A., Fontecilla-Camps, J. C., Theil, E. C., and Volz, K. (2006) Structure of Dual Function Iron Regulatory Protein 1 Complexed with Ferritin IRE-RNA. *Science* 314, 1903–1908.
- (156) Castello, A., Hentze, M. W., and Preiss, T. (2015) Metabolic Enzymes Enjoying New Partnerships as RNA-Binding Proteins. *Trends Endocrinol. Metab.* 26, 746–757.

- (157) Ciesla, J. (2006) Metabolic enzymes that bind RNA: yet another level of ceullar regulatory network? *Acta Biochim. Pol.* 53, 11–32.
- (158) Mukhopadhyay, R., Jia, J., Arif, A., Ray, P. S., and Fox, P. L. (2009) The GAIT system: a gatekeeper of inflammatory gene expression. *Trends Biochem. Sci.* 34, 324–331.
- (159) Dabo, S., and Meurs, E. (2012) dsRNA-Dependent Protein Kinase PKR and its Role in Stress, Signaling and HCV Infection. *Viruses* 4, 2598–2635.
- (160) Castello, A., Fischer, B., Frese, C. K., Horos, R., Alleaume, A.-M., Foehr, S., Curk, T., Krijgsveld, J., and Hentze, M. W. (2016) Comprehensive Identification of RNA-Binding Domains in Human Cells. *Mol. Cell* 63, 696–710.
- (161) Zeng, L., Zhou, Z., Zhao, W., Wang, W., Huang, Y., Cheng, C., Xu, M., Xie, Y., and Mao, Y. (2001) Molecular cloning, structure and expression of a novel nuclear RNA-binding cyclophiln-like gene (PPIL4) from human fetal brain. *Cytogenet. Cell Genet.* 95, 43–47.
- (162) German Cancer Research Center. Database for RNA-Dependent Proteins.
- (163) Kovalev, N., and Nagy, P. D. (2013) Cyclophilin A Binds to the Viral RNA and Replication Proteins, Resulting in Inhibition of Tombusviral Replicase Assembly. *J. Virol.* 87, 13330–13342.
- (164) Trivedi, D. K., Bhatt, H., Pal, R. K., Tuteja, R., Garg, B., Johri, A. K., Bhavesh, N. S., and Tuteja, N. (2013) Structure of RNA-interacting Cyclophilin A-like protein from *Piriformospora indica* that provides salinity-stress tolerance in plants. *Sci. Rep.* 3.
- (165) Ray, P., Rialon-Guevara, K. L., Veras, E., Sullenger, B. A., and White, R. R. (2012) Comparing human pancreatic cell secretomes by in vitro aptamer selection identifies cyclophilin B as a candidate pancreatic cancer biomarker. *J. Clin. Invest.* 122, 1734–1741.
- (166) Davis, T. L., Walker, J. R., Campagna-Slater, V., Finerty, P. J., Paramanathan, R., Bernstein, G., MacKenzie, F., Tempel, W., Ouyang, H., Lee, W. H., Eisenmesser, E. Z., and Dhe-Paganon, S. (2010) Structural and Biochemical Characterization of the Human

Cyclophilin Family of Peptidyl-Prolyl Isomerases. *PLoS Biol.* (Petsko, G. A., Ed.) 8, e1000439.

(167) Nigro, P., Pompilio, G., and Capogrossi, M. C. (2013) Cyclophilin A: a key player for human disease. *Cell Death Dis.* 4, e888–e888.

(168) Wang, P., and Heitman, J. (2005) The cyclophilins. *Genome Biol.* 6.

(169) Colgan, J., Asmal, M., Neagu, M., Yu, B., Schneidkraut, J., Lee, Y., Sokolskaja, E., Andreotti, A., and Luban, J. (2004) Cyclophilin A Regulates TCR Signal Strength in CD4+ T Cells via a Proline-Directed Conformational Switch in Itk. *Immunity* 21, 189–201.

(170) Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., and Sabatini, D. M. (2015) Identification and characterization of essential genes in the human genome. *Science* 350, 1096.

(171) Blomen, V. A., Májek, P., Jae, L. T., Bigenzahn, J. W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F. R., Olk, N., Stukalov, A., Marceau, C., Janssen, H., Carette, J. E., Bennett, K. L., Colinge, J., Superti-Furga, G., and Brummelkamp, T. R. (2015) Gene essentiality and synthetic lethality in haploid human cells. *Science* 350, 1092.

(172) Huai, Q., Kim, H.-Y., Liu, Y., Zhao, Y., Mondragon, A., Liu, J. O., and Ke, H. (2002) Crystal structure of calcineurin–cyclophilin– cyclosporin shows common but distinct recognition of immunophilin–drug complexes 6.

(173) Dreyfuss, M., H $\heartsuit$ rrri, E., Hofmann, H., Kobel, H., Pache, W., and Tschertter, H. (1976) Cyclosporin A and C: New metabolites from *Trichoderma polysporum* (Link ex Pers.) Rifai. *Eur. J. Appl. Microbiol.* 3, 125–133.

(174) Hodge, K. T., Krasnoff, S. B., and Humber, R. A. (1996) *Tolyocladium inflatum* Is the Anamorph of *Cordyceps subsessilis*. *Mycologia* 88, 715.

(175) Odom, A. (1997) Calcineurin is required for virulence of *Cryptococcus neoformans*. *EMBO J.* 16, 2576–2589.

(176) Poor, F., Parent, S. A., Morin, N., Dahl, A. M., Ramadan, N., Chrebet, G., Bostian, K. A., and Nielsen, J. B. (1992) Calcineurin mediates inhibition by FK506 and cyclosporin of recovery from  $\alpha$ -factor arrest in yeast. *Nature* 360, 682.

(177) Breuder, T., Hemenway, C. S., Movva, N. R., Cardenas, M. E., and Heitman, J. (1994) Calcineurin is essential in cyclosporin A- and FK506-sensitive yeast strains. *Proc. Natl. Acad. Sci.* 91, 5372–5376.

(178) Zydowsky, L. D., Etzkorn, F. A., Chang, H. Y., Ferguson, S. B., Stolz, L. A., Ho, S. I., and Walsh, C. T. (1992) Active site mutants of human cyclophilin A separate peptidyl-prolyl isomerase activity from cyclosporin A binding and calcineurin inhibition. *Protein Sci.* 1, 1092–1099.

(179) Kofron, J. L., Kuzmic, P., Kishore, V., Colon-Bonilla, E., and Rich, D. H. (1991) Determination of Kinetic Constants for Peptidyl Prolyl Cis-Trans Isomerases by an Improved Spectrophotometric Assay. *Biochemistry* 30, 6127–6134.

(180) Arora, K., Gwinn, W. M., Bower, M. A., Watson, A., Okwumabua, I., MacDonald, H. R., Bukrinsky, M. I., and Constant, S. L. (2005) Extracellular Cyclophilins Contribute to the Regulation of Inflammatory Responses. *J. Immunol.* 175, 517–522.

(181) Bosco, D. A., and Kern, D. (2004) Catalysis and Binding of Cyclophilin A with Different HIV-1 Capsid Constructs <sup>†</sup>. *Biochemistry* 43, 6110–6119.

(182) de Wilde, A. H., Pham, U., Posthuma, C. C., and Snijder, E. J. (2018) Cyclophilins and cyclophilin inhibitors in nidovirus replication. *Virology* 522, 46–55.

(183) Frausto, S., Lee, E., and Tang, H. (2013) Cyclophilins as Modulators of Viral Replication. *Viruses* 5, 1684–1701.

(184) Peel, M., and Scribner, A. (2013) Cyclophilin inhibitors as antiviral agents. *Bioorg. Med. Chem. Lett.* 23, 4485–4492.

(185) Sherry, B., Yarlett, N., Strupp, A., and Cerami, A. (1992) Identification of cyclophilin as a proinflammatory secretory product of lipopolysaccharide-activated macrophages. *Proc. Natl. Acad. Sci.* 89, 3511–3515.

- (186) Lee, J., and Kim, S. S. (2010) Current implications of cyclophilins in human cancers. *J. Exp. Clin. Cancer Res.* 29, 97.
- (187) Pakula, R., Melchior, A., Denys, A., Vanpouille, C., Mazurier, J., and Allain, F. (2007) Syndecan-1/CD147 association is essential for cyclophilin B-induced activation of p44/42 mitogen-activated protein kinases and promotion of cell adhesion and chemotaxis. *Glycobiology* 17, 492–503.
- (188) Marcant, A., Denys, A., Melchior, A., Martinez, P., Deligny, A., Carpentier, M., and Allain, F. (2012) Cyclophilin B Attenuates the Expression of TNF- in Lipopolysaccharide-Stimulated Macrophages through the Induction of B Cell Lymphoma-3. *J. Immunol.* 189, 2023–2032.
- (189) Allain, F., Vanpouille, C., Carpentier, M., Slomianny, M.-C., Durieux, S., and Spik, G. (2002) Interaction with glycosaminoglycans is required for cyclophilin B to trigger integrin-mediated adhesion of peripheral blood T lymphocytes to extracellular matrix. *Proc. Natl. Acad. Sci.* 99, 2714–2719.
- (190) Friedman, J., and Weissman, I. (1991) Two Cytoplasmic Candidates for Immunophilin Action Are Revealed by Affinity for A new Cyclophilin: OOne in the Presence and One in the Absence of CsA. *Cell* 66, 799–806.
- (191) Obermayr, E., Castillo-Tong, D. C., Pils, D., Speiser, P., Braicu, I., Van Gorp, T., Mahner, S., Sehouli, J., Vergote, I., and Zeillinger, R. (2013) Molecular characterization of circulating tumor cells in patients with ovarian cancer improves their prognostic significance — A study of the OVCAD consortium. *Gynecol. Oncol.* 128, 15–21.
- (192) Pirkl, F., and Buchner, J. (2001) Functional analysis of the hsp90-associated human peptidyl prolyl Cis/Trans isomerases FKBP51, FKBP52 and cyp40 1 Edited by R. Huber. *J. Mol. Biol.* 308, 795–806.
- (193) Jurica, M. S., Licklider, L. J., Gygi, S. P., Grigorieff, N., and Moore, M. J. (2002) Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis. *RNA* 8, 426–439.
- (194) Kim, J.-O., Nau, M. M., Allikian, K. A., Makela, T. P., Alitalo, K., Johnson, B. E., and Kelley, M. J. (1998) Co-amplification of a novel cyclophilin-like gene (PPIE) with L-myc in small cell lung cancer cell lines. *Oncogene* 17, 1019–1026.

- (195) Wang, Y., Han, R., Zhang, W., Yuan, Y., Zhang, X., Long, Y., and Mi, H. (2008) Human CyP33 binds specifically to mRNA and binding stimulates PPIase activity of hCyP33. *FEBS Lett.* 582, 835–839.
- (196) Mi, H., Kops, O., Zimmermann, E., Jaschke, A., and Tropschug, M. (1996) A nuclear RNA-binding cyclophilin in human T-cells. *FEBS Lett.* 398, 201–205.
- (197) Schinzel, A. C., Takeuchi, O., Huang, Z., Fisher, J. K., Zhou, Z., Rubens, J., Hetz, C., Danial, N. N., Moskowitz, M. A., and Korsmeyer, S. J. (2005) Cyclophilin D is a component of mitochondrial permeability transition and mediates neuronal cell death after focal cerebral ischemia. *Proc. Natl. Acad. Sci.* 102, 12005–12010.
- (198) Nakagawa, T., Shimizu, S., Watanabe, T., Yamaguchi, O., Otsu, K., Yamagata, H., Inohara, H., Kubo, T., and Tsujimoto, Y. (2005) Cyclophilin D-dependent mitochondrial permeability transition regulates some necrotic but not apoptotic cell death. *Nature* 434, 652–658.
- (199) Nestel, F., Karen, C., Harper, S., Pawson, T., and Anderson, S. K. (1996) RS cyclophilins: Identification of an NK-TR1-related cyclophilin. *Genes Genomes* 180, 151–155.
- (200) Bourquin, J.-P., Stagljar, I., Meier, P., Moosmann, P., Silke, J., Baechi, T., Georgiev, O., and Schaffner, W. (1997) A serine/arginine-rich nuclear matrix cyclophilin interacts with the C-terminal domain of RNA polymerase II. *Nucleic Acids Res.* 25, 2055–2061.
- (201) Horowitz, D. S., Kobayashi, R., and Krainer, A. R. (1997) A new cyclophilin and the human homologues of yeast Prp3 and Prp4 form a complex associated with U4/U6 snRNPs\*. *RNA* 3, 1374–1387.
- (202) Teigelkamp, S., Achsel, T., Mundt, C., Gothel, S.-F., Cronshagen, U., Lane, W. S., Marahiel, M., and Lührmann, R. (1998) The 20kD protein of human [U4/ U6.U5] tri-snRNPs is a novel cyclophilin that forms a complex with the U4/ U6-specific 60kD and 90kD proteins. *RNA* 4, 127–141.
- (203) Horowitz, D. S., Lee, E. J., Mabon, S. A., and Misteli, T. (2002) A cyclophilin functions pre-mRNA splicing. *EMBO J.* 21, 470–480.

(204) Ozaki, K., Fujiwara, T., Kawai, A., Shimizu, F., Takami, S., Okuno, S., Takeda, S., Shimada, Y., Nagata, M., Watanabe, T., Takaichi, A., Takahashi, E., Nakamura, Y., and Shin, S. (1996) Cloning, expression and chromosomal mapping of a novel cyclophilin-related genes (PPIL1) from human fetal brain. *Cytogenet. Cell Genet.* 72, 242–245.

(205) Jurica, M. S., and Moore, M. J. (2003) Pre-mRNA Splicing: Awash in a Sea of Proteins. *Mol. Cell* 12, 5–14.

(206) Xu, C., Zhang, J., Huang, X., Sun, J., Xu, Y., Tang, Y., Wu, J., Shi, Y., Huang, Q., and Zhang, Q. (2006) Solution Structure of Human Peptidyl Prolyl Isomerase-like Protein 1 and Insights into Its Interaction with SKIP. *J. Biol. Chem.* 281, 15900–15908.

(207) Pushkarsky, T., Yurchenko, V., Vanpouille, C., Brichacek, B., Vaisman, I., Hatakeyama, S., Nakayama, K. I., Sherry, B., and Bukrinsky, M. I. (2005) Cell Surface Expression of CD147/EMMPRIN Is Regulated by Cyclophilin 60. *J. Biol. Chem.* 280, 27866–27871.

(208) Wang, B. B., Hayenga, K. J., Payan, D. G., and Fisher, J. M. (1996) Identification of a nuclear-specific cyclophilin which interacts with the proteinase inhibitor eglin c. *Biochem. J.* 314, 313–319.

(209) Mesa, A., Somarelli, J. A., and Herrera, R. J. (2008) Spliceosomal immunophilins. *FEBS Lett.* 582, 2345–2351.

(210) Rinfret, A., Collins, C., Menard, R., and Anderson, Stephen K. (1994) The N-Terminal Cyclophilin-Homologous Domain of a 150-Kilodalton Tumor Recognition Molecule Exhibits Both Peptidylprolyl cis-trans-Isomerase and Chaperone Activities. *Biochemistry* 33, 1668–1673.

(211) Anderson, S. K., Gallinger, S., Roder, J., Frey, J., Young, H. A., and Ortaldo, J. R. (1993) A cyclophilin-related protein involved in the function of natural killer cells. *Proc. Natl. Acad. Sci.* 90, 542–546.

(212) Beddow, A. L., Richards, S. A., Orem, N. R., and Macara, I. G. (1995) The Ran/TC4 GTPase-binding domain: Identification by expression cloning and characterization of a conserved sequence motif. *Proc. Natl. Acad. Sci.* 92, 3328–3332.



- (213) Wu, J., Matunis, M. J., Kraemer, D., Blobel, G., and Coutavas, E. (1995) Nup358, a Cytoplasmically Exposed Nucleoprotein with Peptide Repeats, Ran\_GTP Binding Sites, Zinc Fingers, a Cyclophilin A Homologous Domain, and a Leucine-rich Region. *J. Biol. Chem.* 270, 14209–14213.
- (214) Fraser, J. S., Clarkson, M. W., Degnan, S. C., Erion, R., Kern, D., and Alber, T. (2009) Hidden alternative structures of proline isomerase essential for catalysis. *Nature* 462, 669–673.
- (215) Saphire, A. C., Bobardt, M. D., and Gallay, P. A. (1999) Host cyclophilin A mediates HIV-1 attachment to target cells via heparans. *Eur. Mol. Biol. Organ.* 18, 6771–6785.
- (216) Hanouille, X., Melchior, A., Sibille, N., Parent, B., Denys, A., Wieruszeski, J.-M., Horvath, D., Allain, F., Lippens, G., and Landrieu, I. (2007) Structural and Functional Characterization of the Interaction between Cyclophilin B and a Heparin-derived Oligosaccharide. *J. Biol. Chem.* 282, 34148–34158.
- (217) Kozlov, G., Bastos-Aristizabal, S., Määttänen, P., Rosenauer, A., Zheng, F., Killikelly, A., Trempe, J.-F., Thomas, D. Y., and Gehring, K. (2010) Structural Basis of Cyclophilin B Binding by the Calnexin/Calreticulin P-domain. *J. Biol. Chem.* 285, 35551–35557.
- (218) Denys, A., Allain, F., Carpentier, M., and Spik, G. (1998) Involvement of two classes of binding sites in the interactions of cyclophilin B with peripheral blood T-lymphocytes. *Biochem. J.* 336, 689–697.
- (219) Eisenmesser, E. Z., Millet, O., Labeikovsky, W., Korzhnev, D. M., Wolf-Watz, M., Bosco, D. A., Skalicky, J. J., Kay, L. E., and Kern, D. (2005) Intrinsic dynamics of an enzyme underlies catalysis. *Nature* 438, 117–121.
- (220) Howard, B. R., Vajdos, F. F., Li, S., Sundquist, W. I., and Hill, C. P. (2003) Structural insights into the catalytic mechanism of cyclophilin A. *Nat. Struct. Biol.* 10, 475–481.
- (221) Villali, J., and Kern, D. (2010) Choreographing an enzyme's dance. *Curr. Opin. Chem. Biol.* 14, 636–643.

- (222) Doshi, U., Holliday, M. J., Eisenmesser, E. Z., and Hamelberg, D. (2016) Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. *Proc. Natl. Acad. Sci.* 113, 4735–4740.
- (223) Bannikova, O., Zywicki, M., Marquez, Y., Skrahina, T., Kalyna, M., and Barta, A. (2013) Identification of RNA targets for the nuclear multidomain cyclophilin atCyp59 and their effect on PPLase activity. *Nucleic Acids Res.* 41, 1783–1796.
- (224) Yang, M.-W., Inouye, C. J., and Seto, E. (1996) Cyclophilin A and FKBP12 Interact with YY1 and Alter its Transcriptional Activity. *J. Biol. Chem.* 270, 15187–15193.
- (225) Ansari, H., Greco, G., and Luban, J. (2002) Cyclophilin A Peptidyl-Prolyl Isomerase Activity Promotes Zpr1 Nuclear Export. *Mol. Cell. Biol.* 22, 6993–7003.
- (226) Gustafson, C. L., Parsley, N. C., Asimgil, H., Lee, H.-W., Ahlback, C., Michael, A. K., Xu, H., Williams, O. L., Davis, T. L., Liu, A. C., and Partch, C. L. (2017) A Slow Conformational Switch in the BMAL1 Transactivation Domain Modulates Circadian Rhythms. *Mol. Cell* 66, 447-457.e7.
- (227) Min, L., Fulton, D. B., and Andreotti, A. (2005) A CASE STUDY OF PROLINE ISOMERIZATION IN CELL SIGNALING. *Front. Biosci.* 10, 385–397.
- (228) Brazin, K. N., Mallis, R. J., Fulton, D. B., and Andreotti, A. (2002) Regulation of the tyrosine kinase Itk by the peptidyl-prolyl isomerase cyclophilin A. *Proc. Natl. Acad. Sci.* 99, 1899–1904.
- (229) Yurchenko, V., Pushkarsky, T., Li, J.-H., Dai, W. W., Sherry, B., and Bukrinsky, M. (2005) Regulation of CD147 Cell Surface Expression: INVOLVEMENT OF THE PROLINE RESIDUE IN THE CD147 TRANSMEMBRANE DOMAIN. *J. Biol. Chem.* 280, 17013–17019.
- (230) Yang, Y., Lu, N., Zhou, J., Chen, Z. -n., and Zhu, P. (2008) Cyclophilin A up-regulates MMP-9 expression and adhesion of monocytes/macrophages via CD147 signalling pathway in rheumatoid arthritis. *Rheumatology* 47, 1299–1310.
- (231) Kim, H., Kim, W.-J., Jeon, S.-T., Koh, E.-M., Cha, H.-S., Ahn, K.-S., and Lee, W.-H. (2005) Cyclophilin A may contribute to the inflammatory processes in rheumatoid

arthritis through induction of matrix degrading enzymes and inflammatory cytokines from macrophages. *Clin. Immunol.* 116, 217–224.

(232) Lin, K., and Gallay, P. (2013) Curing a viral infection by targeting the host: The example of cyclophilin inhibitors. *Antiviral Res.* 99, 68–77.

(233) Madan, V., Paul, D., Lohmann, V., and Bartenschlager, R. (2014) Inhibition of HCV Replication by Cyclophilin Antagonists Is Linked to Replication Fitness and Occurs by Inhibition of Membranous Web Formation. *Gastroenterology* 146, 1361-1372.e9.

(234) Yang, H., Chen, J., Yang, J., Qiao, S., Zhao, S., and Yu, L. (2007) Cyclophilin A is upregulated in small cell lung cancer and activates ERK1/2 signal. *Biochem. Biophys. Res. Commun.* 361, 763–767.

(235) Howard, B. A., Furumai, R., Campa, M. J., Rabbani, Z. N., Vujaskovic, Z., Wang, X.-F., and Patz, Jr., E. F. (2005) Stable RNA Interference–Mediated Suppression of Cyclophilin A Diminishes Non–Small-Cell Lung Tumor Growth In vivo. *Cancer Res.* 65.

(236) Garcia-Rivera, J. A., Bobardt, M., Chatterji, U., Hopkins, S., Gregory, M. A., Wilkinson, B., Lin, K., and Gallay, P. A. (2012) Multiple Mutations in Hepatitis C Virus NS5A Domain II Are Required To Confer a Significant Level of Resistance to Alisporivir. *Antimicrob. Agents Chemother.* 56, 5113–5121.

(237) Phillips, S., Chokshi, S., Chatterji, U., Riva, A., Bobardt, M., Williams, R., Gallay, P., and Naoumov, N. V. (2015) Alisporivir Inhibition of Hepatocyte Cyclophilins Reduces HBV Replication and Hepatitis B Surface Antigen Production. *Gastroenterology* 148, 403-414.e7.

(238) Arévalo-Rodríguez, M., and Heitman, J. (2005) Cyclophilin A Is Localized to the Nucleus and Controls Meiosis in *Saccharomyces cerevisiae*. *Eukaryot. Cell* 4, 17–29.

(239) Kim, I. S., Yun, H. S., Kwak, S. H., and Jin, I. N. (2007) The Physiological Role of CPR1 in *Saccharomyces cerevisiae* KNU5377 against Menadione Stress by Proteomics. *J. Microbiol.* 45, 326–332.

(240) Kim, I.-S., Yun, H., Jin, I., and Yoon, H.-S. (2011) Cyclophilin A Cpr1 Protein Modulates the Response of Antioxidant Molecules to Menadione-induced Oxidative

Stress in *Saccharomyces cerevisiae* KNU5377Y. *Osong Public Health Res. Perspect.* 2, 171–177.

(241) Andersen, K. S., Bojsen, R., Sorensen, L. G. R., Nielsen, M. W., Lisby, M., Folkesson, A., and Regenber, B. (2014) Genetic Basis for *Saccharomyces cerevisiae* Biofilm in Liquid Medium. *G3amp58 GenesGenomesGenetics* 4, 1671–1680.

(242) Yoon, H.-S. (2010) Expression of Yeast Cyclophilin A (Cpr1) Provides Improved Stress Tolerance in *Escherichia coli*. *J. Microbiol. Biotechnol.* 20, 974–977.

(243) Price, E. R., Zydowsky, L. D., Jin, M., Baker, H., McKeon, F. D., and Walsh, C. T. (1991) Human cyclophilin B: A second cyclophilin gene encodes a peptidyl-prolyl isomerase with a signal sequence. *Proc. Natl. Acad. Sci.* 88, 1903–1907.

(244) van Dijk, F. S., Nesbitt, I. M., Zwikstra, E. H., Nikkels, P. G. J., Piersma, S. R., Fratantoni, S. A., Jimenez, C. R., Huizer, M., Morsman, A. C., Cobben, J. M., van Roij, M. H. H., Elting, M. W., Verbeke, J. I. M. L., Wijnaendts, L. C. D., Shaw, N. J., Högler, W., McKeown, C., Sistermans, E. A., Dalton, A., Meijers-Heijboer, H., and Pals, G. (2009) PPIB Mutations Cause Severe Osteogenesis Imperfecta. *Am. J. Hum. Genet.* 85, 521–527.

(245) Hom, R. A., Chang, P.-Y., Roy, S., Musselman, C. A., Glass, K. C., Selezneva, A. I., Gozani, O., Ismagilov, R. F., Cleary, M. L., and Kutateladze, T. G. (2010) Molecular Mechanism of MLL PHD3 and RNA Recognition by the Cyp33 RRM Domain. *J. Mol. Biol.* 400, 145–154.

(246) Park, S., Osmers, U., Raman, G., Schwantes, R. H., Diaz, M. O., and Bushweller, J. H. (2010) The PHD3 Domain of MLL Acts as a CYP33-Regulated Switch between MLL-Mediated Activation and Repression. *Biochemistry* 49, 6576–6586.

(247) Cao, J. (2014) The functional role of long non-coding RNAs and epigenetics. *Biol. Proced. Online* 16.

(248) Bessonov, S., Anokhina, M., Will, C. L., Urlaub, H., and Lührmann, R. (2008) Isolation of an active step I spliceosome and composition of its RNP core. *Nature* 452, 846–850.

- (249) Bessonov, S., Anokhina, M., Krasauskas, A., Golas, M. M., Sander, B., Will, C. L., Urlaub, H., Stark, H., and Luhrmann, R. (2010) Characterization of purified human Bact spliceosomal complexes reveals compositional and morphological changes during spliceosome activation and first step catalysis. *RNA* 16, 2384–2403.
- (250) Hegele, A., Kamburov, A., Grossmann, A., Sourlis, C., Wowro, S., Weimann, M., Will, C. L., Pena, V., Lührmann, R., and Stelzl, U. (2012) Dynamic Protein-Protein Interaction Wiring of the Human Spliceosome. *Mol. Cell* 45, 567–580.
- (251) Gullerova, M. (2006) AtCyp59 is a multidomain cyclophilin from *Arabidopsis thaliana* that interacts with SR proteins and the C-terminal domain of the RNA polymerase II. *RNA* 12, 631–643.
- (252) Chang, A.-Y., Castel, S. E., Ernst, E., Kim, H. S., and Martienssen, R. A. (2017) The Conserved RNA Binding Cyclophilin, Rct1, Regulates Small RNA Biogenesis and Splicing Independent of Heterochromatin Assembly. *Cell Rep.* 19, 2477–2489.
- (253) Schiene-Fischer, C. (2015) Multidomain Peptidyl Prolyl cis/trans Isomerases. *Biochim. Biophys. Acta BBA - Gen. Subj.* 1850, 2005–2016.
- (254) Lin, C. L., Leu, S., Lu, M. C., and Ouyang, P. (2004) Over-expression of SR-cyclophilin, an interaction partner of nuclear pinin, releases SR family splicing factors from nuclear speckles. *Biochem. Biophys. Res. Commun.* 321, 638–647.
- (255) Dilworth, D., Upadhyay, S. K., Bonnafous, P., Edo, A. B., Bourbigot, S., Pesek-Jardim, F., Gudavicius, G., Serpa, J. J., Petrotchenko, E. V., Borchers, C. H., Nelson, C. J., and Mackereth, C. D. (2017) The basic tilted helix bundle domain of the prolyl isomerase FKBP25 is a novel double-stranded RNA binding module. *Nucleic Acids Res.* 45, 11989–12004.
- (256) Tuerk, C., and Gold, L. (1990) Systematic Evolution of Ligands by Exponential Enrichment: RNA ligands to Bacteriophage T4 DNA Polymerase. *Science* 249, 505–510.
- (257) Chen, F., Hu, Y., Li, D., Chen, H., and Zhang, X.-L. (2009) CS-SELEX Generates High-Affinity ssDNA Aptamers as Molecular Probes for Hepatitis C Virus Envelope Glycoprotein E2. *PLoS ONE* (Bereswill, S., Ed.) 4, e8142.

(258) Lorenz, C., Gesell, T., Zimmermann, B., Schoeberl, U., Bilusic, I., Rajkowitsch, L., Waldsich, C., von Haeseler, A., and Schroeder, R. (2010) Genomic SELEX for Hfq-binding RNAs identifies genomic aptamers predominantly in antisense transcripts. *Nucleic Acids Res.* 38, 3794–3808.

(259) Manley, J. L. (2013) SELEX to Identify Protein-Binding Sites on RNA. *Cold Spring Harb. Protoc.* 2013, pdb.prot072934-pdb.prot072934.

(260) Shtatland, T. (2000) Interactions of Escherichia coli RNA with bacteriophage MS2 coat protein: genomic SELEX. *Nucleic Acids Res.* 28, 93e – 93.

(261) Reid, D. C., Chang, B. L., Gunderson, S. I., Alpert, L., Thompson, W. A., and Fairbrother, W. G. (2009) Next-generation SELEX identifies sequence and structural determinants of splicing factor binding in human pre-mRNA sequence. *RNA* 15, 2385–2397.

(262) Smith, G. P., and Petrenko, V. A. (1997) Phage Display. *Chem. Rev.* 97, 391–410.

(263) Vaught, J. D., Bock, C., Carter, J., Fitzwater, T., Otis, M., Schneider, D., Rolando, J., Waugh, S., Wilcox, S. K., and Eaton, B. E. (2010) Expanding the Chemistry of DNA for in Vitro Selection. *J. Am. Chem. Soc.* 132, 4141–4151.

(264) Gold, L., Ayers, D., Bertino, J., Bock, C., Bock, A., Brody, E. N., Carter, J., Dalby, A. B., Eaton, B. E., Fitzwater, T., Flather, D., Forbes, A., Foreman, T., Fowler, C., Gawande, B., Goss, M., Gunn, M., Gupta, S., Halladay, D., Heil, J., Heilig, J., Hicke, B., Husar, G., Janjic, N., Jarvis, T., Jennings, S., Katilius, E., Keeney, T. R., Kim, N., Koch, T. H., Kraemer, S., Kroiss, L., Le, N., Levine, D., Lindsey, W., Lollo, B., Mayfield, W., Mehan, M., Mehler, R., Nelson, S. K., Nelson, M., Nieuwlandt, D., Nikrad, M., Ochsner, U., Ostroff, R. M., Otis, M., Parker, T., Pietrasiewicz, S., Resnicow, D. I., Rohloff, J., Sanders, G., Sattin, S., Schneider, D., Singer, B., Stanton, M., Sterkel, A., Stewart, A., Stratford, S., Vaught, J. D., Vrkljan, M., Walker, J. J., Watrobka, M., Waugh, S., Weiss, A., Wilcox, S. K., Wolfson, A., Wolk, S. K., Zhang, C., and Zichi, D. (2010) Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *PLoS ONE* (Gelain, F., Ed.) 5, e15004.

(265) Brody, E., Gold, L., Mehan, M., Ostroff, R., Rohloff, J., Walker, J., and Zichi, D. (2012) Life's Simple Measures: Unlocking the Proteome. *J. Mol. Biol.* 422, 595–606.

- (266) Ganz, P., Heidecker, B., Hveem, K., Jonasson, C., Kato, S., Segal, M. R., Sterling, D. G., and Williams, S. A. (2016) Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. *JAMA* 315, 2532.
- (267) Keefe, A. D., and Cload, S. T. (2008) SELEX with modified nucleotides. *Curr. Opin. Chem. Biol.* 12, 448–456.
- (268) Viores, S. A. (2006) Pegaptanib in the treatment of wet, age-related macular degeneration. *Int. J. Nanomedicine* 6.
- (269) Buenrostro, J. D., Araya, C. L., Chircus, L. M., Layton, C. J., Chang, H. Y., Snyder, M. P., and Greenleaf, W. J. (2014) Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat. Biotechnol.* 32, 562–568.
- (270) Hoinka, J., Backofen, R., and Przytycka, T. M. (2018) AptaSUITE: A Full-Featured Bioinformatics Framework for the Comprehensive Analysis of Aptamers from HT-SELEX Experiments. *Mol. Ther. Nucleic Acids* 11, 515–517.
- (271) Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336.
- (272) Alam, K. K., Chang, J. L., and Burke, D. H. (2015) FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections. *Mol. Ther. - Nucleic Acids* 4, e230.
- (273) Hoinka, J., and Przytycka, T. (2016) AptaPLEX – A dedicated, multithreaded demultiplexer for HT-SELEX data. *Methods* 106, 82–85.
- (274) Hoinka, J., Berezhnoy, A., Sauna, Z. E., Gilboa, E., and Przytycka, T. M. (2014) AptaCluster – A Method to Cluster HT-SELEX Aptamer Pools and Lessons from Its Application, in *Research in Computational Molecular Biology* (Sharan, R., Ed.), pp 115–128. Springer International Publishing, Cham.

(275) Dao, P., Hoinka, J., Takahashi, M., Zhou, J., Ho, M., Wang, Y., Costa, F., Rossi, J. J., Backofen, R., Burnett, J., and Przytycka, T. M. (2016) AptaTRACE Elucidates RNA Sequence-Structure Motifs from Selection Trends in HT-SELEX Experiments. *Cell Syst.* 3, 62–70.

(276) Dao, P., Hoinka, J., Wang, Y., Takahashi, M., Zhou, J., Costa, F., Rossi, J., Burnett, J., Backofen, R., and Przytycka, T. M. (2016) AptaTRACE: Elucidating Sequence-Structure Binding Motifs by Uncovering Selection Trends in HT-SELEX Experiments. *ArXiv160403081 Cs Q-Bio*.

(277) Hoinka, J., Berezhnoy, A., Dao, P., Sauna, Z. E., Gilboa, E., and Przytycka, T. M. (2015) Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. *Nucleic Acids Res.* 43, 5699–5707.

(278) Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406–3415.

(279) Integrated DNA Technologies. (2011) Chemical Synthesis of Oligonucleotides.

(280) Pollard, J., Bell, S. D., and Ellington, A. D. (2000) Design, Synthesis, and Amplification of DNA Pools for In Vitro Selection. *Curr. Protoc. Nucleic Acid Chem.* 00, 9.2.1-9.2.23.

(281) Takahashi, M., Wu, X., Ho, M., Chomchan, P., Rossi, J. J., Burnett, J. C., and Zhou, J. (2016) High throughput sequencing analysis of RNA libraries reveals the influences of initial library and PCR methods on SELEX efficiency. *Sci. Rep.* 6.

(282) Stockley, P. G., Stonehouse, N. J., Murray, J. B., Goodman, S. T. S., Talbot, S. J., Adams, C. J., Liljas, L., and Valegard, K. (1995) Probing sequence-specific RNA recognition by the bacteriophage MS2 coat protein. *Nucleic Acids Res.* 23, 2512–2518.

(283) Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P. A., and Burge, C. B. (2014) RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Mol. Cell* 54, 887–900.

(284) Lambert, N. J., Robertson, A. D., and Burge, C. B. (2015) RNA Bind-n-Seq, in *Methods in Enzymology*, pp 465–493. Elsevier.



(285) Solanki, J. A. (2011) Role of Non-Coding RNA NC4 in MLL and CYP33 Mediated Regulation of HOXC8. Loyola University, Chicago.

(286) Mohr, S., Ghanem, E., Smith, W., Sheeter, D., Qin, Y., King, O., Polioudakis, D., Iyer, V. R., Hunicke-Smith, S., Swamy, S., Kuersten, S., and Lambowitz, A. M. (2013) Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* 19, 958–970.

(287) Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.* 6, 26.

(288) Bolanos-Garcia, V. M., and Davies, O. R. (2006) Structural analysis and classification of native proteins from *E. coli* commonly co-purified by immobilised metal affinity chromatography. *Biochim. Biophys. Acta BBA - Gen. Subj.* 1760, 1304–1313.

## Appendix A

This Appendix contains the detailed protocols used in chapter 2

### Protein Expression and Purification

Pot1pC was expressed and purified using the same construct and essentially the same method described in Dickey et al. (2013) with the rationale behind the construct described below.

The C-terminal domain of *s. pombe* Pot1 (residues 198-339) was cloned into pTXB1, between the NdeI and SpeI with a C-terminal chitin-binding domain linked to Pot1pC through a self-cleaving intein linker (Pot1pC-Intein-ChitinBindingDomain). The chitin-binding domain is used as an affinity purification tag and to increase the expression, stability, and solubility of the protein construct. The intein linker, in the presence of a reducing agent such as  $\beta$ -mercaptoethanol or dithiothreitol, will self-cleave the affinity tag, allowing elution of Pot1pC containing the native protein sequence. The domain boundaries of Pot1pC (198-339) were defined by limited proteolysis experiments alongside  $^1\text{H}$ - $^{15}\text{N}$  HSQC comparison and ITC for the 178-389 construct, all of which indicate residues 198-339 form the structural core of Pot1pC. The V199D was introduced for protein solubility in the initial screening for small soluble protein constructs, so it is unknown if the core native sequence is also stably expressed and soluble.

Pot1 (Pombe)

MGEDVIDSLQLNELLNAGEYKIGELTFQSIRSSQELQKNTIVNLFQIVKDFTPSRQSLH  
GTKDWVTTVYLWDPTCDTSSIGLQIHLFSKQGNDLPVIKQVGPPLLHQITLRSYRDRTQ  
GLSKDQFRYALWPDFSSNSKDTLCPQMPRLMKTGDKEEQFALLLNKIWDEQTNKHKNGE  
LLSTSSARQNQTGLSYP

SVSFSLLSQITPHQRCSFYAQVIKTWYSDKNFTLYVTDYTENE  
LFFPMSPYTSSSRWRGPFGRFSIRCILWDEHDFYCRNYIKEGDYVVMKNVRTKIDHLGYL  
ECILHGDSAKRYNMSIEKVDSEEPNELNEIKSRKRLYVQ

NCQNGIEAVIEKLSQSQQSENP

FIAHELKQTSVNEITAHVINEPASLKLTTISTILHAPLQNLLKPRKHRLRVQVDFWPKS  
LTQFAVLSQPPSSYVWMFALLVRDVSNTLPVIFFDSDAAELINSSKIQPCNLADHPQMT  
LQLKERLFLIWGNLEERIQHHISKGESPTLAAEDVETPWFDIYVKEYIPVIGNTKDHQSL  
TFLQKRWRGFGTKIV

Pot1pC 198-339 (V199D)

SDFSLLSQITPHQRCSFYAQVIKTWYSDKNFTLYVTDYTENE  
LFFPMSPYTSSSRWRGPFGRFSIRCILWDEHDFYCRNYIKEGDYVVMKNVRTKIDHLGYL  
ECILHGDSAKRYNMSIEKVDSEEPPELNEIKSRKRLYVQ

### **Pot1pC Protein Purification**

$\epsilon_{280} = 32890 \text{ M}^{-1} \text{ cm}^{-1}$

Number of amino acids: 141

Molecular weight: 16914.1

Theoretical pI: 6.87 (cleaved)

### **Day 1 – Transform Pot1pC pTBX1 in Intein-CBD vector into DE3 *E. coli* such as BL21 (DE3).**

- 1 - Thaw competent cells on ice
- 2 – Add ~50-100 ng of plasmid to autoclaved sterile 1.5-1.7 Eppendorf tube (typically 1-2  $\mu\text{L}$  of stock plasmid).
- 3 – Gently pipet thawed cells to plasmid Eppendorf tube and incubate on ice for 15-30 min
- 4 – Heat shock cells at 42 °C for 45s or 37 °C for 90s
- 5 – Incubate heat shocked cells on ice for 5 minutes
- 6 – Add 700  $\mu\text{L}$  of LB to cells and gently mix

- 7 – Incubate transformed cells at 37 °C with shaking (~180 rpm) for 45-60 minutes.
- 8 – Spin cells at 5000g in a microcentrifuge, decant supernatant with ~100 µL of LB remaining.
- 9 – Resuspend cell pellet and spread on LB agar plates supplemented with 100 mg/mL ampicillin.
- 10 – Allow plate to dry before incubating at 37 °C overnight.

### **Day 2 – Inoculate starter culture.**

- 1 – Using a sterile loop, pipet tip, or toothpick, pick a single colony of transformed cells or transformed glycerol stock and inoculate 40 mL of sterile autoclaved LB in a 125 mL baffled flask supplemented with 100 mg/mL ampicillin. You will use 10 mL of starter culture per liter of growth you plan to do, so upscale if you plan to do more than 4L of growth. Be sure to use a baffled flask with a capacity at least twice as much as the volume of culture you are using.
- 2 – Incubate starter culture at 37 °C overnight with ~180 rpm shaking.

### **Day 3 – Growth and Expression**

- 1 – Pitch 10 mL of starter culture into each 1L growth (1L LB in 2L baffled flasks) supplemented with 100 mg/mL ampicillin. Optional – pellet a mL of cells and resuspend in sterile 50% glycerol for a glycerol stock.
- 2 – Grow 1L growths at 37 °C with shaking ~180 rpm until an OD<sub>600</sub> of 0.5-0.8. Typically, the cultures will reach this OD<sub>600</sub> after 2-3 hours and have a doubling time around 20-30 minutes.
- 3 – Incubate cultures on ice for 40 min
- 4 – Induce with 500 µl of 1M IPTG per liter.
- 5 – Grow at 18 °C overnight (~18-22 hrs.)

### **Day 4 – Pellet cells and freeze**

- 1 – Spin the cells down in 500 mL centrifuge bottles with the F10-6 Fiberlite or J10 rotor in the floor centrifuges at ~5000 rpm (roughly 9000g). Use 1 bottle per 1L growth, keeping the cultures

separate, spinning up to ~400 mL at a time. Do not fill the centrifuge bottles to the top or the bottle will likely leak – fill up to the lip of the bottle and check that the rubber seal on the lid is flush with the plastic.

2 – Decant as much liquid as possible and scrap pellets into 50 mL tubes and freeze at -20 °C or continue with the next steps.

#### **Day 4/5 – Protein Purification**

1 – Resuspend pellet in 50 mL of lysis buffer (20 mM Tris pH 8.5 500 mM NaCl) with a Roche EDTA-free inhibitor tablet.

\*Note that this prep is done mostly at 4 °C and the pH of Tris buffers has a significant temperature dependence so either pH the buffer at 4 °C or use consult a chart for the equivalent buffer pHed at room temperature (~7.9).

2 – Pre-equilibrate 10 mL chitin beads (20 mL of slurry) per liter of culture with 100 mL of lysis buffer per 10 mL beads using a peristaltic pump and a Kontes column. Set the flow rate to 2 mL/min.

3 – Using the Misonix Sonicator 3000 with a ½” tip, sonicate 10-12 times with 10s pulses and 45s rests at power=8. If you observe any foaming, pause the sonication and lower the power. You may need to adjust the tip so that it remains completely submerged during each pulse.

4 – Spin 30 minutes at least 15,000g in the 30 mL Oakridge tubes in the F21 rotor in the floor centrifuge. Do not fill the tubes above the lip and check that the seal is flush with the cap, or they may leak.

5 – Pour the supernatant over pre-equilibrated chitin beads and set the flow rate to 0.5 ml/min.

6 – Wash with 20 column volumes of lysis buffer at 2 ml/min (200 mL for 1L purification).

7 – Allow the supernatant to flow just above the top of the beads and then blow the beads into a 50 mL conical vial. Resuspend in an equivalent volume of lysis buffer at add 135 µL of β-mercaptoethanol for each 10 mL of beads. Gently mix beads and incubate 20-48 hours.

#### **Day 5/6 – Concentration and SEC.**

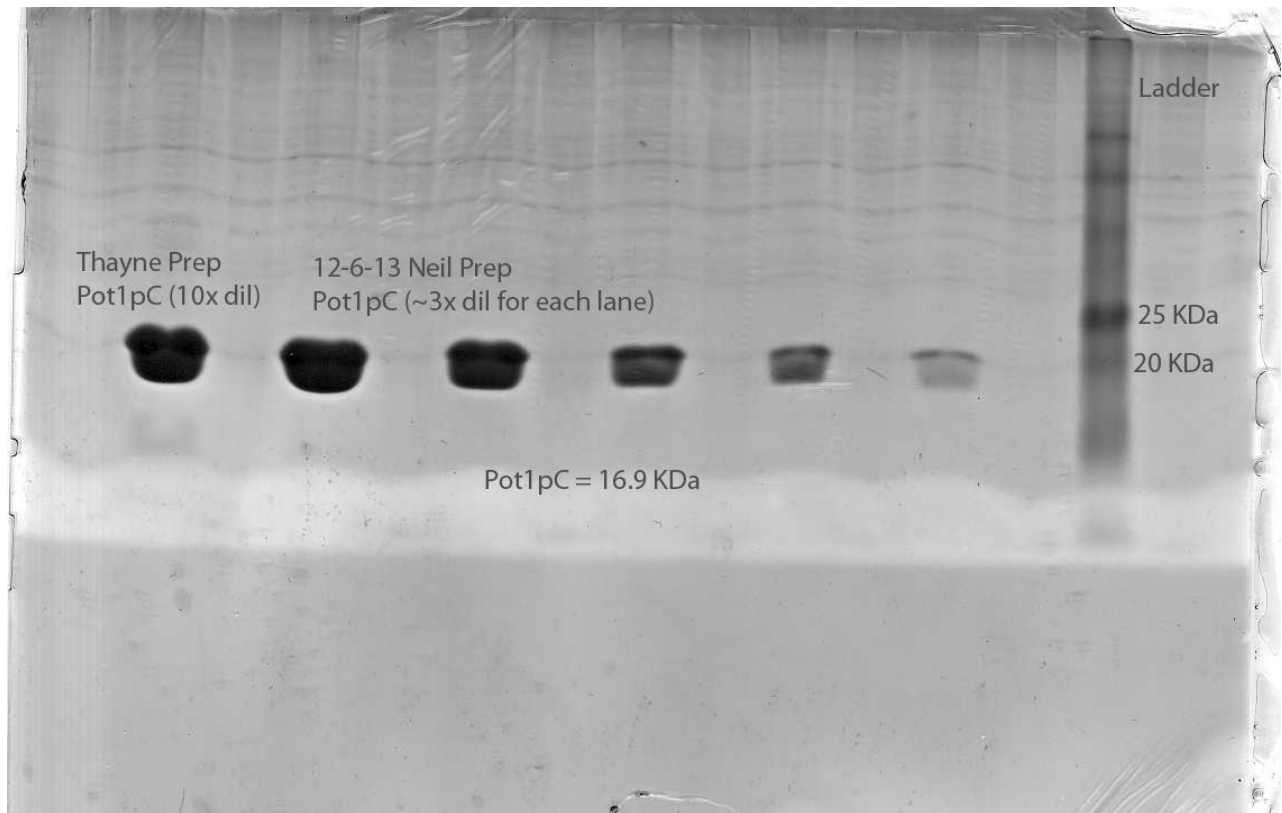
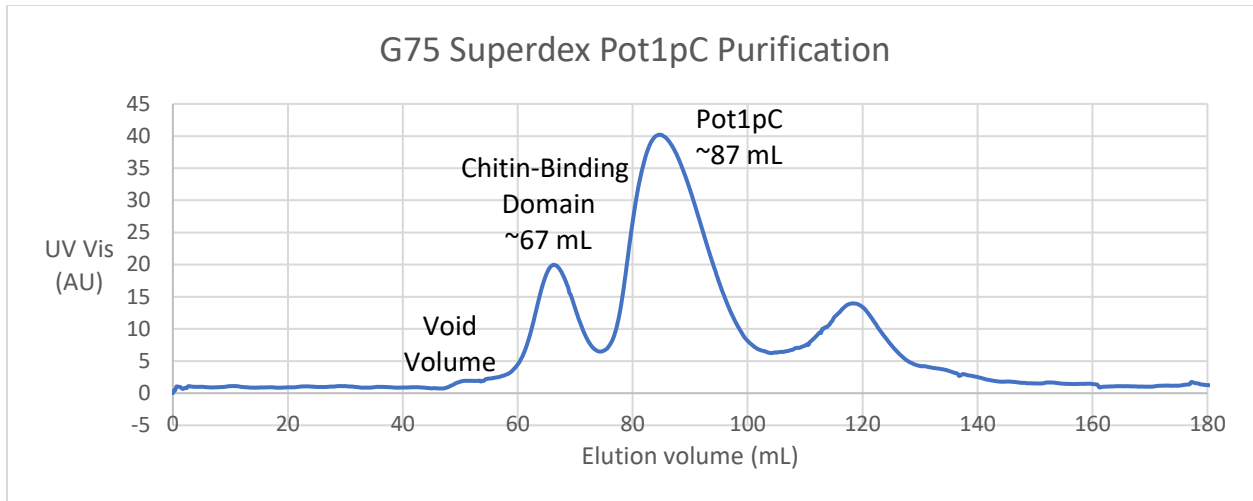
1 – Pre-equilibrate the G75 Superdex column with filtered and degassed storage buffer (50 mM KPhos pH 8.0, 150 mM NaCl, 3 mM  $\beta$ ME, and 0.1% (w/v) deoxycholate. You will need at least 180 mL for the equilibration wash and another 180 mL for the fraction collection, so having at least 500 mL is recommended.

2 - Pour beads back into Kontes column, collect flow through, and rinse with an additional ~25 mL of lysis buffer.

3 - Concentrate flow through and wash on a 5/10K MWCO concentrator (you lose less with the 5K MWCO, but it takes a lot of longer to concentrate) down to ~2 mL. To avoid high-concentration and precipitation at the membrane interface, you should gently mix the solution every 15-20 minutes during concentration. If you observe significant precipitation, you should filter the concentrate immediately. Depending on the protein concentration, if you observe precipitation you may want to proceed to SEC with multiple injections if the eluate is still above 2 mL.

4 – When the eluate is concentrated to ~ 2 mL, filter the solution and inject onto the G75 column with a flow rate of ~ 1 mL. I observe free Pot1pC eluting around ~87 mL, though Thayne has previously observed ~95 mL. Some chitin-binding domain typically elutes as well with a peak around ~67 mL. If you have issues with overlapping Pot1pC and chitin-binding domain, running the flow through over additional chitin beads prior to concentration should help with that issue.

5 – Concentrate the Pot1pC fractions centered around 80-100 mL, flash freeze in liquid nitrogen, and store at -70 °C. My typical yields were about ~500  $\mu$ L of about 400-600  $\mu$ M protein. The protein can be stored at higher concentrations but tends to produce more precipitate upon thawing.



### Isothermal Titration Calorimetry

ITC was used to characterize the binding affinity for RNA-substituted 9mer oligos. All experiments were performed in at least triplicate on a MicroCal iTC200 (GE Healthcare) at 25 °C. The sample cell was loaded with ~200  $\mu$ L of 5-100  $\mu$ M Pot1pC into which buffer matched

nucleic acid at approximately 10-fold higher concentration was titrated as follows: one 0.2  $\mu\text{L}$  dummy injection, followed by nineteen 2  $\mu\text{L}$  injections, and a final 1.3  $\mu\text{L}$  injection. Data were integrated and fit by nonlinear least-squares fitting to a single binding site model using Origin ITC Software (OriginLab, Northampton, MA).

### **Day 1 – Setup overnight dialysis**

- 1 – Thaw -70 °C stored Pot1pC on ice
- 2 – Filter 1L of 20 mM KPhos pH 8.0, 150 mM NaCl, and 3 mM  $\beta\text{ME}$  and place in a ~2L beaker or another wide opening container.
- 3 – Dilute some thawed Pot1pC with the dialysis buffer and load Pot1pC solution into a 10K MWCO Thermofisher Slide-A-Lyzer Mini Dialysis tube. When loading, be careful not to touch the membrane with the pipette tip.
- 4 – Place the Dialysis tube into a foam float device, careful to position the top of the protein solution so that it is level with the buffer solution – if the protein solution meniscus is too far below the buffer meniscus, the pressure differential will cause some unwanted dilution of the protein.
- 5 – Cover the beaker or container with foil/lid and allow to dialyze overnight.

### **Day 2 – Step up ITC**

- 1 – Carefully pipette the dialyzed protein solution into an Eppendorf tube and centrifuge at 15000g for 10 minutes at 4 °C to pellet any protein precipitate that appeared overnight.
- 2 – Pipette the protein supernatant to another Eppendorf tube and keep on ice.
- 3 – Filter ~20 mL of dialyzed buffer solution to be used for protein and ligand dilution as well as washing the sample cell. Store on ice in a 50 mL conical vial.
- 4 – Nanodrop Pot1pC to quantify protein concentration (using ExPasy predicted  $\epsilon_{280}$  of 32890  $\text{M}^{-1} \text{cm}^{-1}$ )
- 5 – Using the MicroCal software, you can predict the concentrations of protein and ligand that will give you the best quality data based on the  $K_D$  and the enthalpy of binding. For most of the



ligands I tested that bound in with a low nanomolar  $K_D$ , 5-10  $\mu\text{M}$  Potp1C with a 10-fold higher concentration of ligand gave the best results. The weakest ligand, the 1-9R oligo required ~10-fold more material.

6 – Based on the concentrations predicted to give the best data, dilute your protein concentration with the filtered dialysis buffer. You will need at least 230  $\mu\text{L}$  of protein to load the ~200  $\mu\text{L}$  cell for each experiment, but for viscous solutions like protein, 300  $\mu\text{L}$  should be used for loading. For triplicate experiments, you will need 900  $\mu\text{L}$  of protein diluted to 5-10  $\mu\text{M}$  or the concentration suggested for your affinity/enthalpy range.

7 – Using the same dialysis buffer, dilute the IDT oligo ligand to ~1 mM. You will then want to nanodrop a 1/10-fold dilution of the 1 mM stock solution and calculate the concentration using the predicted IDT extinction coefficient at 260 nm. I recommend doing this higher concentration stock solution because in my experience, the initial concentration is usually quite a bit off the 1 mM it should be.

8 – Using the ~1 mM stock, corrected with the measured nanodrop concentration, dilute an aliquot of the ligand to the concentration needed for ITC. Each ITC experiment will use ~40  $\mu\text{L}$  of ligand, but loading the syringe requires at least 60-80  $\mu\text{L}$  suggested.

9 – If the instrument has not been used recently, replace the solution in the reference cell with MilliQ water.

10 – Do a combined cell and syringe wash on the instrument. Prior to the first run and in between runs, a detergent wash is recommended, followed by a regular combined cell and syringe wash. To do a detergent wash, load the sample cell with 10% Contrad70 prior to running the wash.

11 – Prior to loading your sample, wash the sample cell manually with a syringe with your dialysis buffer. I recommend taking an aliquot of the filtered buffer so that you can wash out the sample loading syringe between samples/washing without contaminating the rest of the buffer.

12 – Ensure that as little liquid as possible is in the sample cell and then load your sample. To load your sample, pull up 300  $\mu\text{L}$  of your protein solution. Then insert the syringe into the sample cell and inject your protein sample slowly to start. After about 50  $\mu\text{L}$ , pause, and then stepwise inject your sample in sharp  $\sim 50$   $\mu\text{L}$  steps\*. Once you have injected about 280  $\mu\text{L}$ , remove the syringe slowly and then pull up any sample above the silver ring at the top of the sample cell.

\*This is to help dislodge any bubbles that are on the side of sample cell which can cause artifacts while establishing the baseline and during the experiment if dislodged.

13 – Load the titrant by first putting the syringe in the rest position as indicated by the software screen. If any liquid comes out of the titrant syringe while the plunger is dropping, stop the loading process. Place the syringe back into the cleaning apparatus, tighten the cord, and perform a new syringe wash – the liquid in the syringe is methanol that will mess up your experiment and indicative of a loose connection during the drying step of the wash.

14 – After the plunger has reached the bottom, place the syringe in the loading position with your titrant sample in a PCR tube, then load the syringe. In my experience, the syringe is almost never fills without a bubble the first time. Reload the syringe if a bubble remains, keeping the titrant syringe in the titrant sample during the entire process (instead of putting it in the rest position).

15 – When the titrant is properly loaded, disconnect the tubing and place the titrant syringe in the sample cell.

16 – Step up the program to 19-21 injections based on the calibrated loading volume of the titrant syringe. The first injection should be a 0.2-5  $\mu\text{L}$  dummy injection, followed by 19-20 2  $\mu\text{L}$  injections and a final injection of the remaining volume  $>1$   $\mu\text{L}$ . I used a reference power of  $\sim 8$  and an initial delay of 10 minutes and 1000 rpm spinning. If the baseline reference power is not within 0.5 of your target, there is likely to be bubble within your sample cell. Larger differences  $>1$  from the reference power suggests that the cell is dirty and requires additional cleaning. If

the reference power is > 0.5 away from your target power, I would recommend reloading your sample cell.

### **Data Analysis**

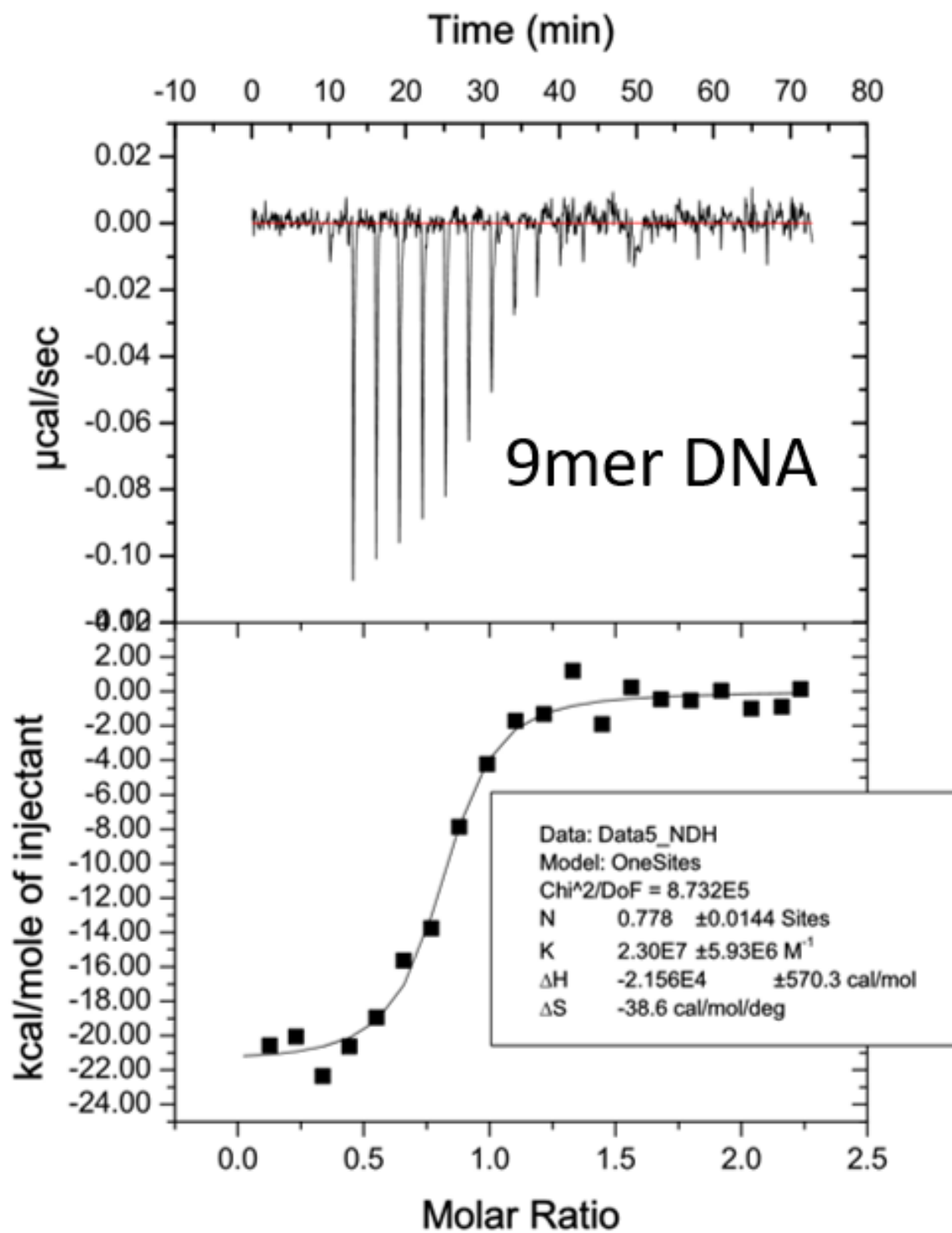
Data were integrated and fit by nonlinear least-squares fitting to a single binding site model using Origin ITC Software (OriginLab, Northampton, MA).

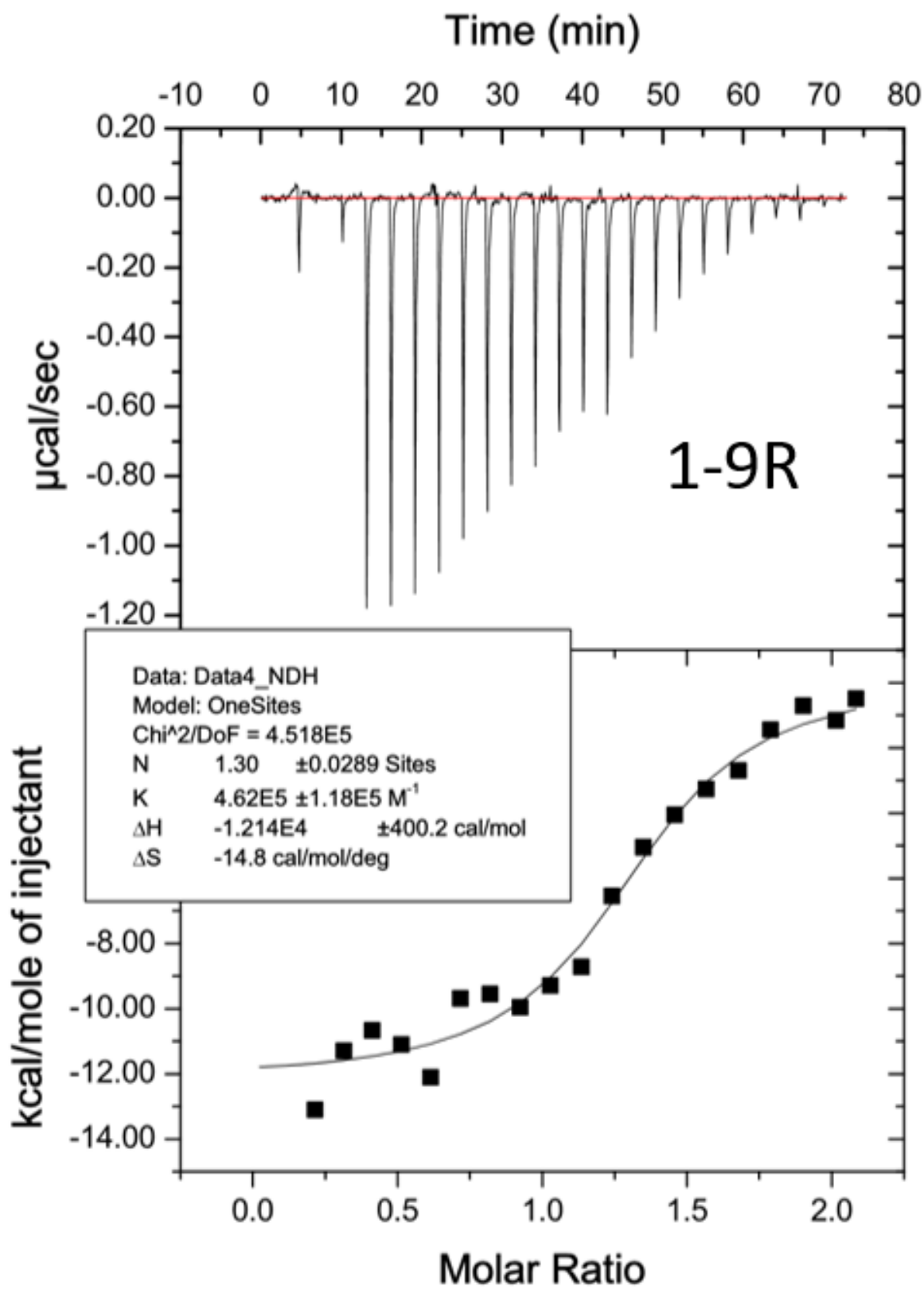
1 – Load the ITC data into the Origin ITC Software by clicking read data and then selecting your experiment.

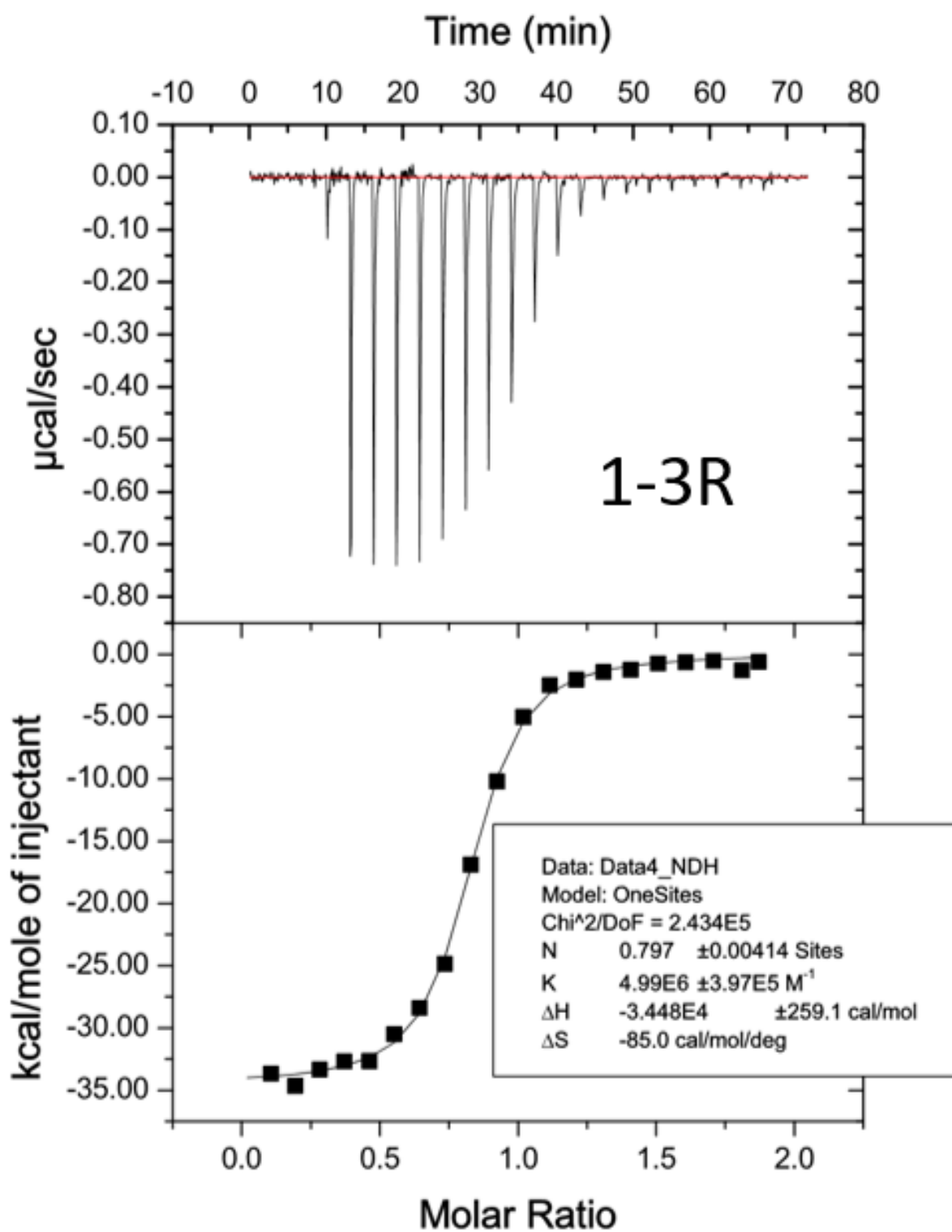
2 – Clicking on integrate peaks will convert your raw ITC data into a  $\Delta H$ /stoichiometry curve. You can adjust the baseline and the window of integration for each of your peaks if the auto-baseline is significantly off from the measured baseline.

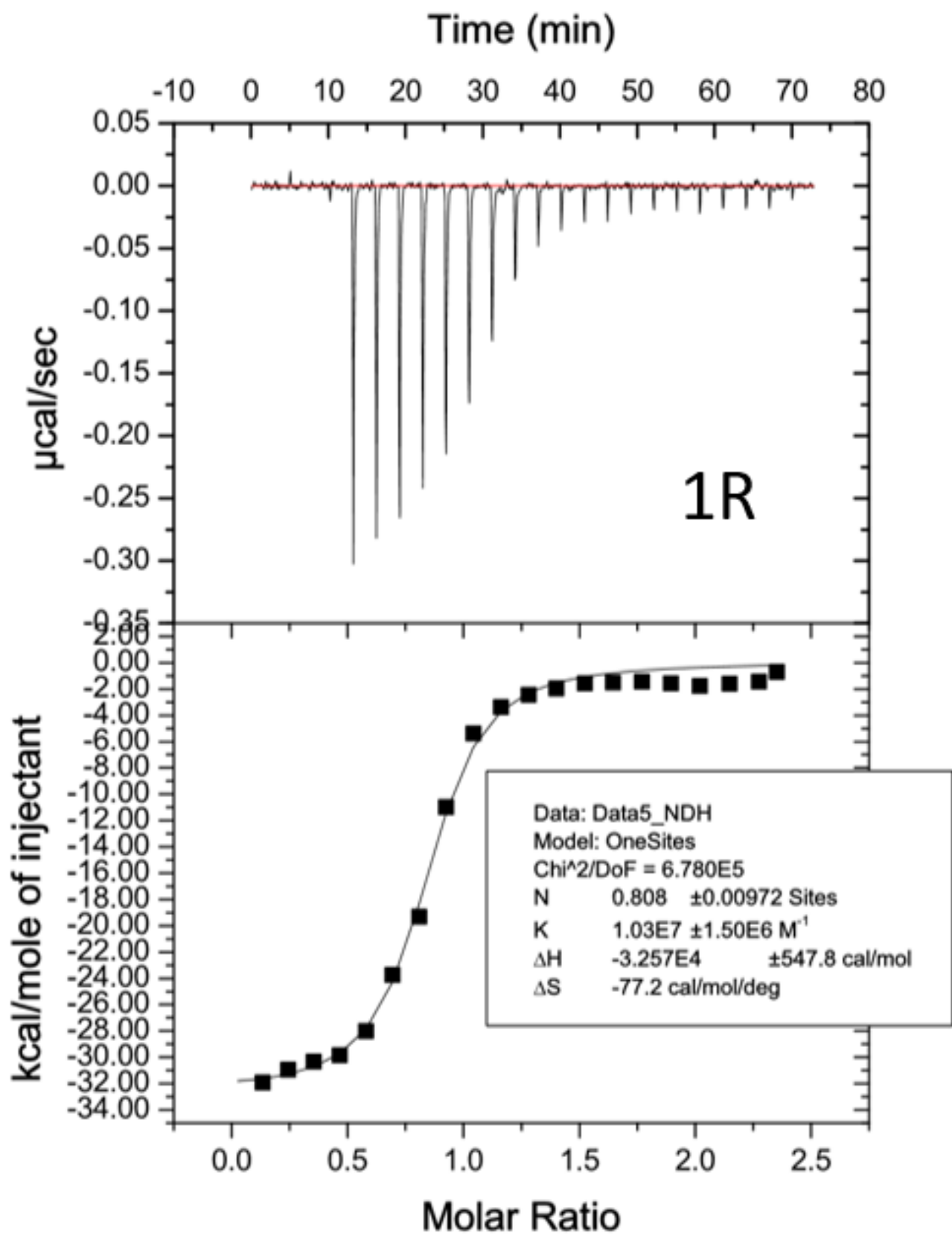
3 – Using remove bad data, the first and last data point should be removed. The first data point is the dummy injection and usually inaccurate while the last injection can sometimes have residual bubbles or incomplete injections that can artifacts.

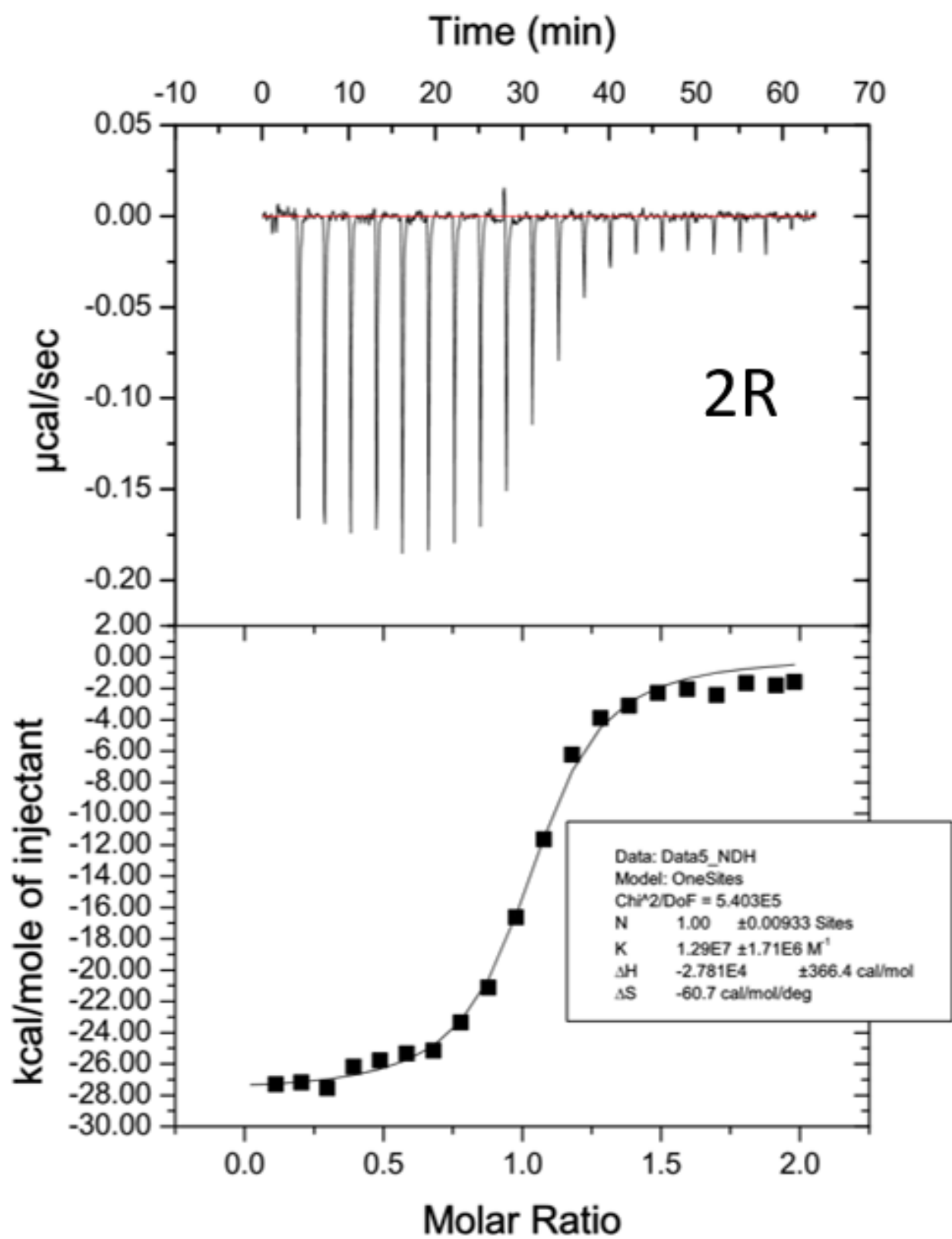
4 – The data can then be fit based on several models. I used the one site binding model to fit my data. Representative fits for my ligands are shown below.



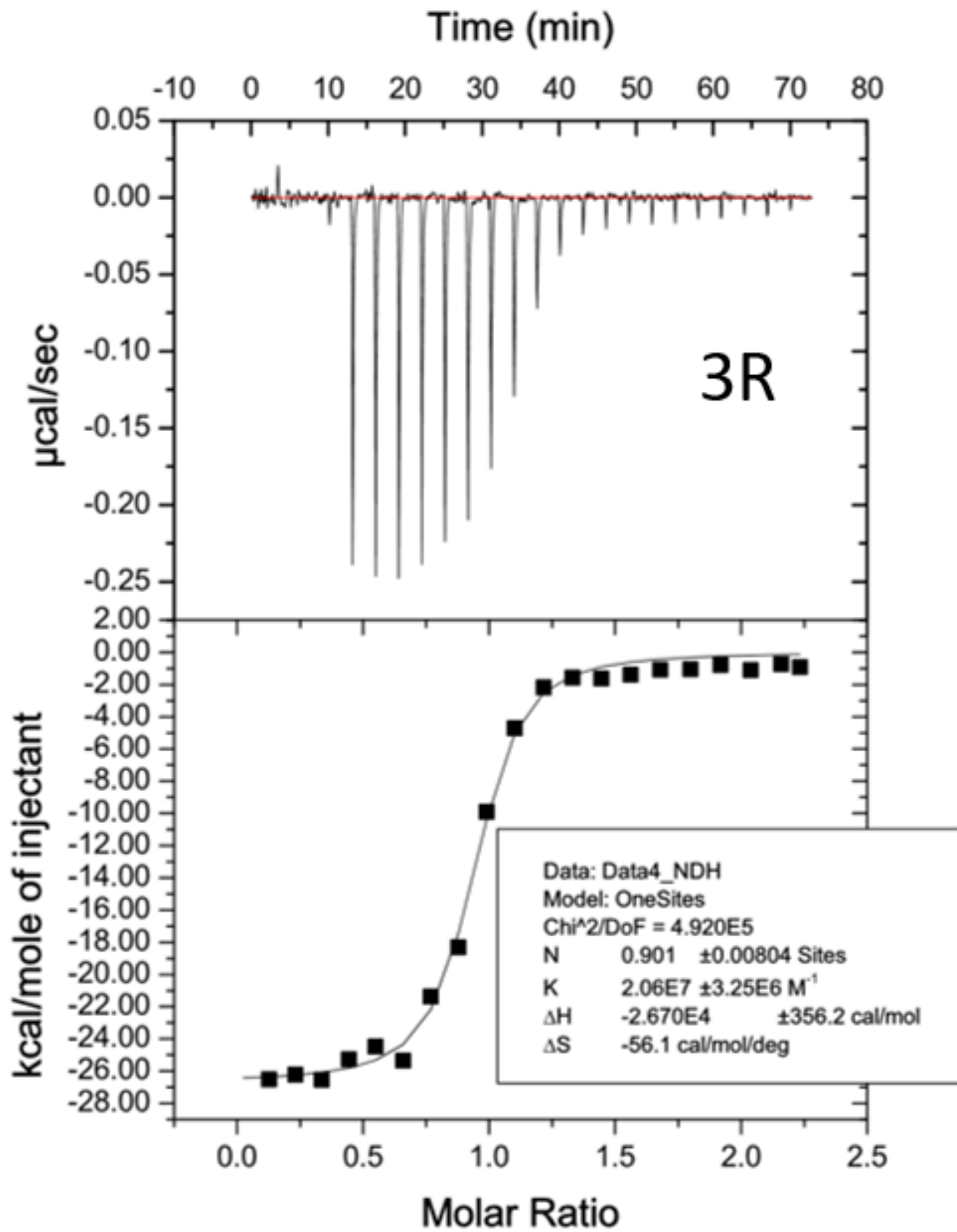


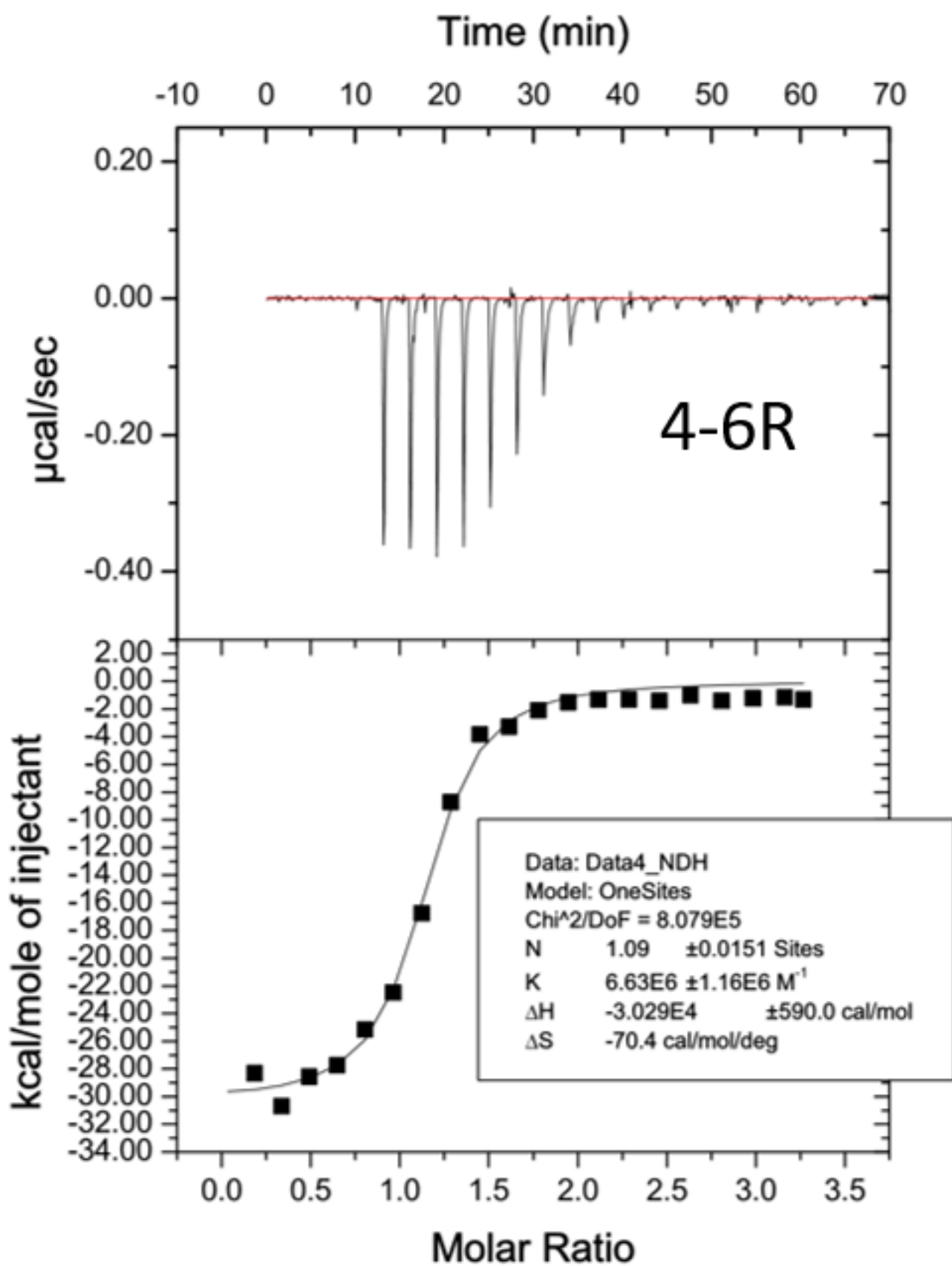


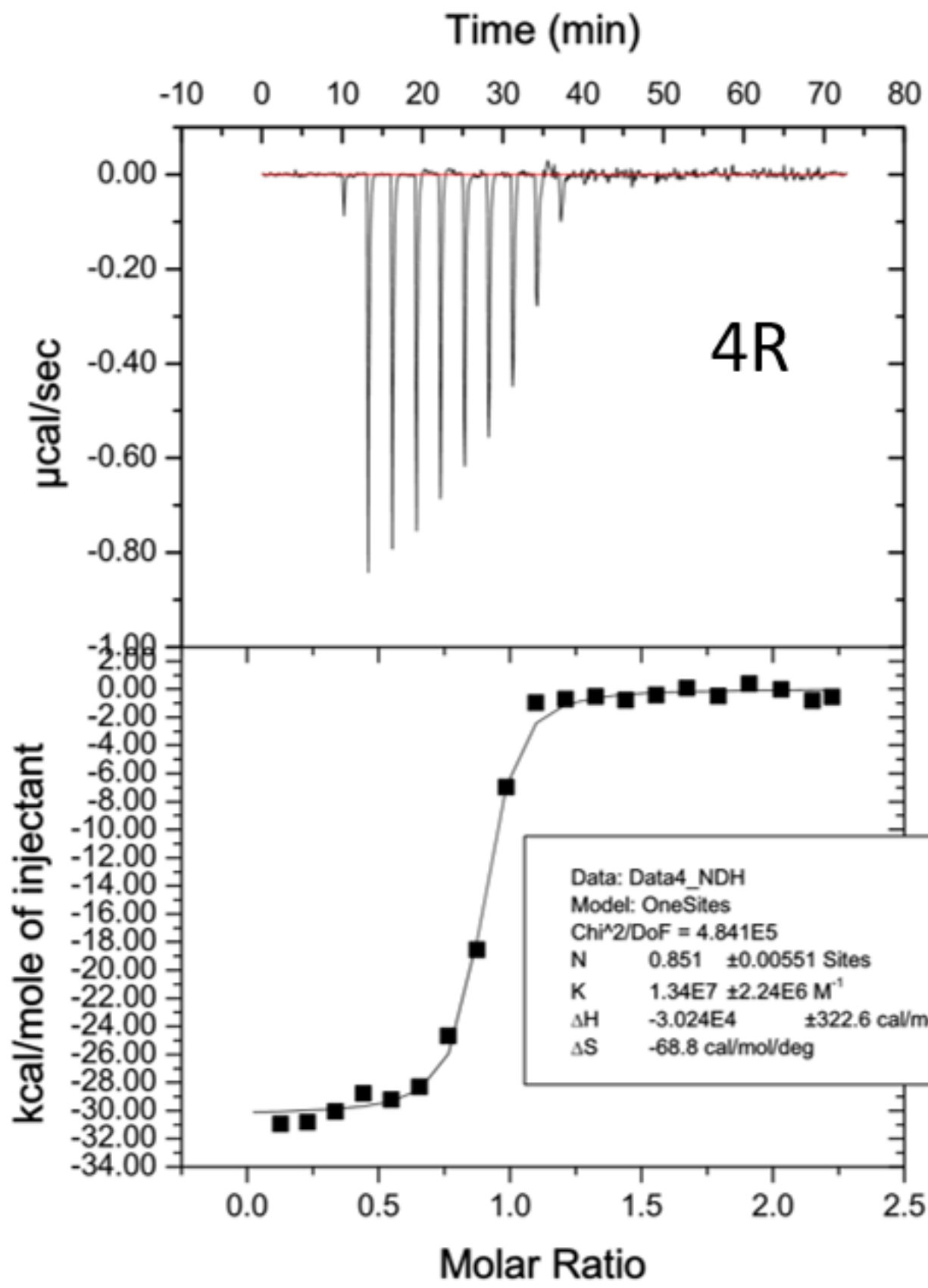


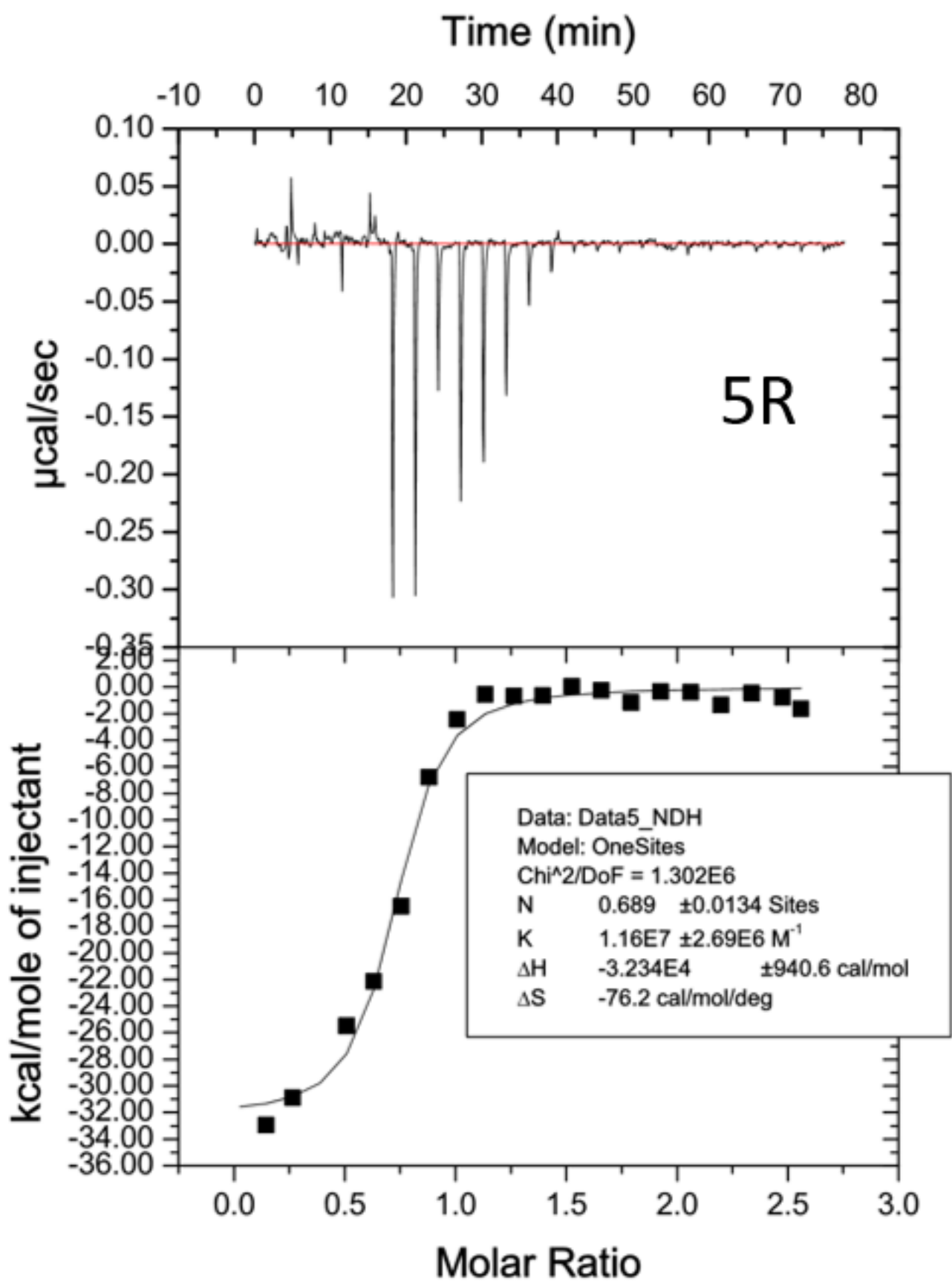


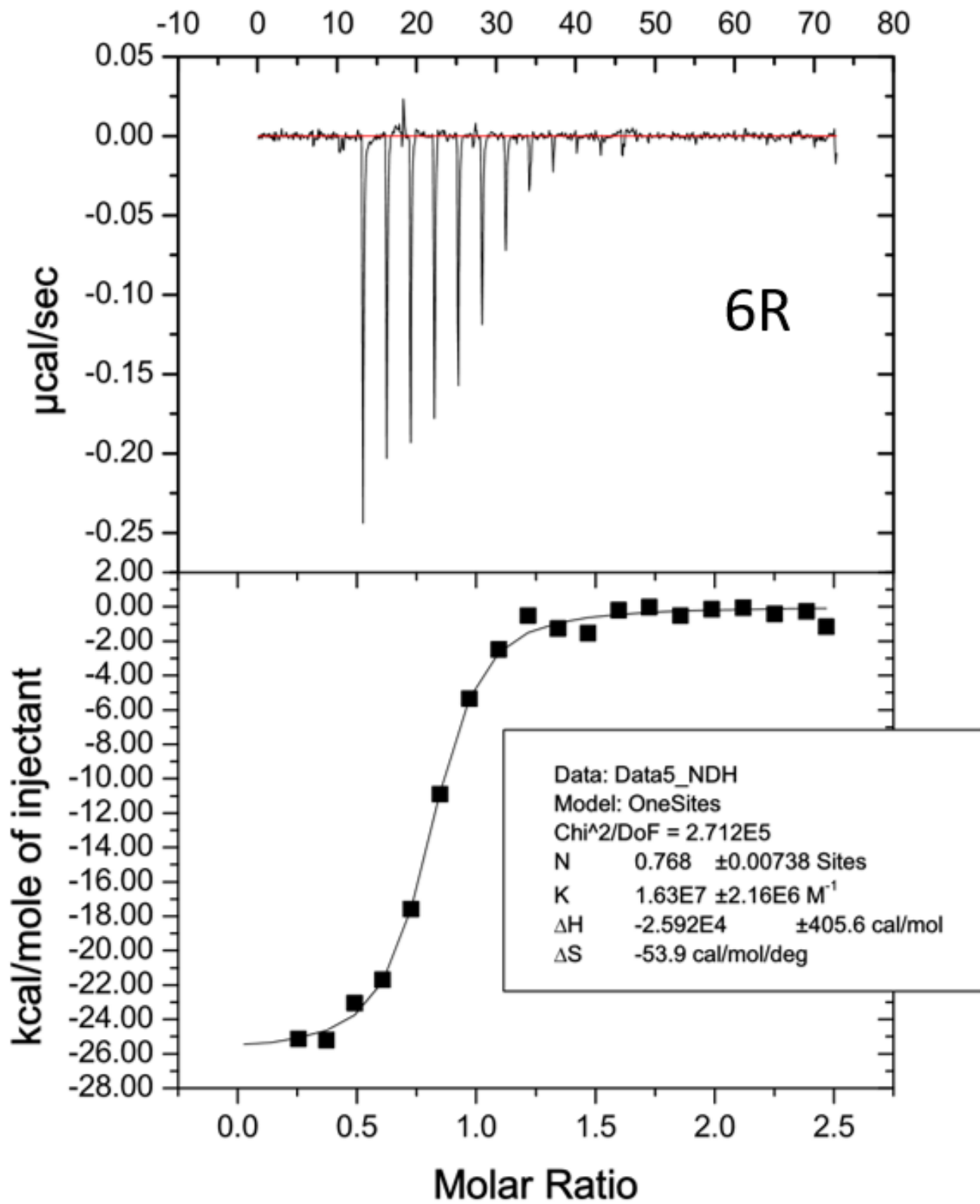


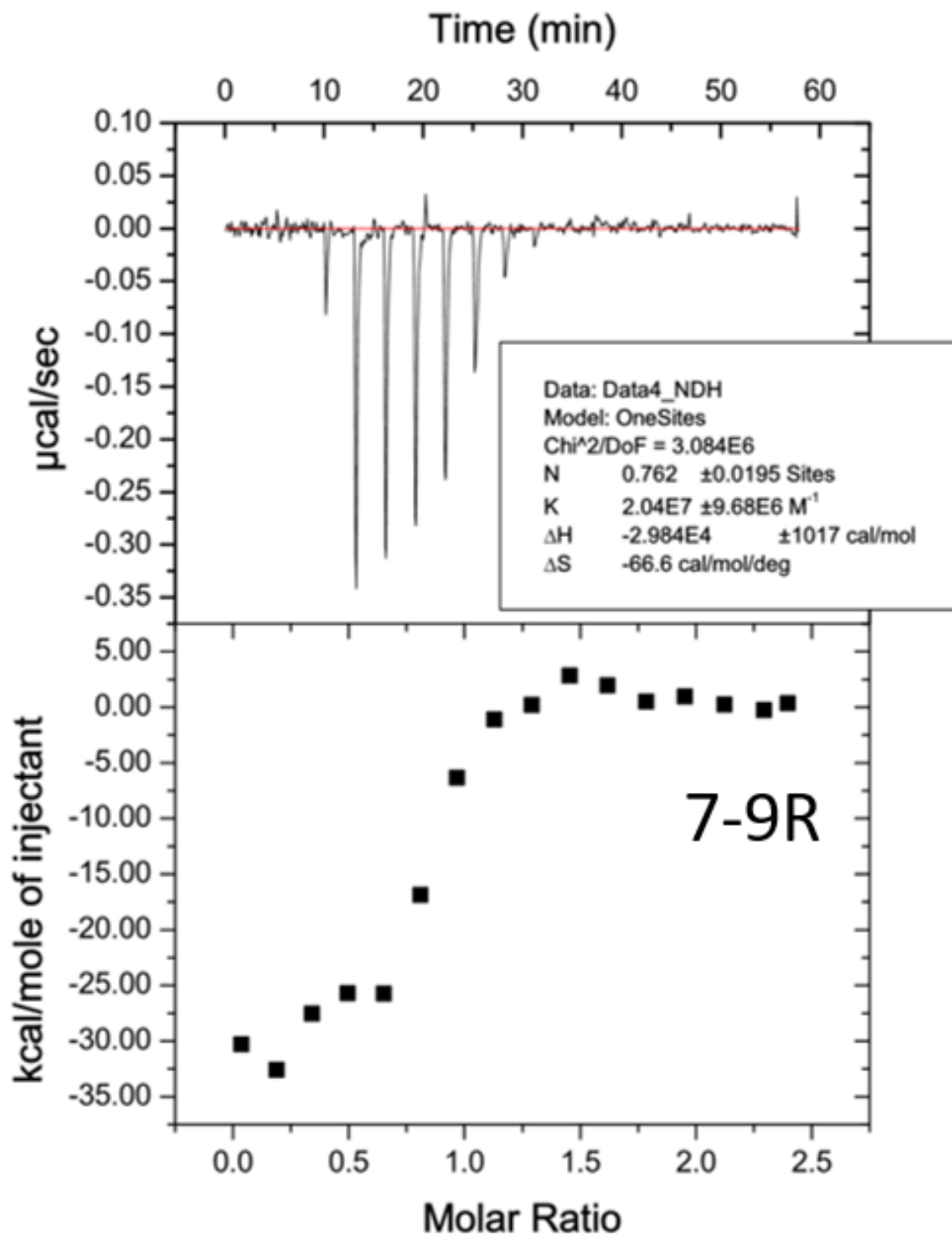












## Crystallization

I initially screened for crystal complexes for Pot1pC bound to 9mer DNA, the 1R, triplets, and full RNA ligands by microscale droplets (0.2  $\mu$ L total volume) via the Phoenix Dropsetter Robot with the Rigaku Wizard I/II, Hampton Research Natrix 1/2, and Hampton Research PEG/Ion I/II screens. The majority of my candidate crystal hits were at conditions similar to the conditions used to crystallize the alternative complexes in Dickey 2013. Select hit are shown below.

With these initial hits in mind, my strategy to get crystals was to focus my screens around the conditions that produced the Pot1pC complexes previously, varying the pH, PEG molecular weight, PEG w/v% and salt additives. I did this in a 24 well hanging drop format. A summary of conditions used are listed below. In each well, 500  $\mu$ L mother liquor was mixed up in the wells, and then chilled at 4  $^{\circ}$ C overnight. This overnight chilling step was to help prevent the precipitation of the complexes upon addition to the mother liquor as higher temperatures tended to cause immediate precipitation. In each well, I screened 6 droplets for each oligo comprising 1  $\mu$ L mother liquor and 1  $\mu$ L Pot1pC-ssRNA/DNA at 5/10/15 mg/mL and 1:1 and 1:1.5 protein:nucleic acid. In this condition space, the 9mer DNA and the 1R produced high-quality diffraction crystals while the triplets produced small spindly sea-urchin like crystals. The 1-9R complexes also produced crystals but they were a different morphology altogether, resembling small spheres.

To facilitate improved crystal morphology of the triplet and 1-9R complexes, the cognate DNA complex crystals were used to seed these droplets. Cognate crystal droplets were resuspended in mother liquor, and then vortexed in the presence of a plastic bead according to the Hampton Research PTFE Seed Bead Kit. A cat whisker was then dipped into the resulting seed solution and streaked through droplets. This resulted in a significant increase in the number of crystals produced for each complex. These triplet crystals were then used to seed new droplets and produced diffraction quality crystals for both the 1-3R and 7-9R complexes.

Crystals were cryoprotected by sequentially transferring the crystal in the mother liquor solutions supplemented with 5%, 10%, 15%, and 20% (v/v) ethylene glycol and flash frozen in liquid nitrogen. Most of the crystals were small enough to be transferred with the 0.2-0.5  $\mu\text{m}$  loops and were kept in the transfer solutions ~5 minutes each. Many of the crystals transferred during the cryoprotection stage were lost, damaged, or destroyed, so 4-5 crystals for each complex were collected for synchrotron screening.

### **Data Collection and Refinement**

X-ray diffraction data for 1R was collected at the Advanced Light Source (ALS) Beamline 8.2.1 and the data sets for 1-3R and 7-9R were collected at the ALS Beamline 8.2.2. Reflections were indexed using iMOSFLM<sup>110</sup> and scaled using Scala within the CCP4 program suite<sup>111</sup>. The phases were solved through molecular replacement using the coordinates of cognate Pot1pC without ssDNA (4HIK)<sup>66</sup> as a starting model in PHENIX<sup>112,113</sup> followed by rigid body refinement using PHENIX Refine<sup>114,115</sup>. The non-cognate RNA ligands were built into the electron density manually in Coot<sup>116</sup> and subsequent refinement was performed in the PHENIX program suite with manual adjustment in Coot. The final models were validated using PHENIX.validate and MolProbity<sup>117</sup> to assess quality (statistics for final models can be found in Table 2.1).



## Appendix B

This appendix contains the detailed experimental protocols and data for Chapter 4 and 5

### Protein Cloning, Expression, and Purification

Construct	Base Plasmid	Restriction Sites	Tag (N/C-term)	Resistance	Superdex Column	$\epsilon_{280}$ ( $M^{-1}cm^{-1}$ )
MBP-MS2 (V29/dIFG)	pET30b	NdeI/HindIII	6xHis-MPB (N)/Thrombin	Kan	G200	83310
CypA	pET15b	NdeI/XhoI	6xHis (N)	Amp	G75	8730
Cpr1	pET15b	NdeI/XhoI	6xHis (N)	Amp	G75	13075
FL-CypE	pET28b	NdeI/XhoI	6xHis (N)/Thrombin	Kan	G200/75	24200
CypE-RRM	pET28b	NdeI/XhoI	6xHis (N)/Thrombin	Kan	G75	2980
CypE-CLD	pET21b	NdeI/XhoI	6xHis (C)	Amp	G75	9970
MBP-CypE	pET30b	NdeI/XhoI	6xHis-MPB (N)/Thrombin	Kan	G200	90550
SUMO-CypE	pET28b	BamHI/XhoI	10xHis-SUMO (N)	Kan	G200	25690

### Protein Expression and Purification

For both protein purification and column binding during selection, the cyclophilin and MS2 constructs all have His-tags. During the protein purification, Nickel-NTA beads are used due to their higher affinity for the His-tag while the SELEX protocol uses Co-NTA beads to physically separate the resin from the solution to facilitate washing despite reduced affinity for the His-tag.

#### Day 1 – Transform vector into DE3 *E. coli* such as BL21 (DE3).

- 1 - Thaw competent cells on ice
- 2 – Add ~50-100 ng of plasmid to autoclaved sterile 1.5-1.7 Eppendorf tube (typically 1-2  $\mu$ L of stock plasmid).
- 3 – Gently pipet thawed cells to plasmid Eppendorf tube and incubate on ice for 15-30 min
- 4 – Heat shock cells at 42 °C for 45s or 37 °C for 90s

- 5 – Incubate heat shocked cells on ice for 5 minutes
- 6 – Add 700  $\mu$ L of LB to cells and gently mix
- 7 – Incubate transformed cells at 37 °C with shaking (~180 rpm) for 45-60 minutes.
- 8 – Spin cells at 5000g in a microcentrifuge, decant supernatant with ~100  $\mu$ L of LB remaining.
- 9 – Resuspend cell pellet and spread on LB agar plates supplemented with 100 mg/mL ampicillin or 50 mg/mL Kanamycin.
- 10 – Allow plate to dry before incubating at 37 °C overnight.

### **Day 2 – Inoculate starter culture.**

- 1 – Using a sterile loop, pipet tip, or toothpick, pick a single colony of transformed cells or transformed glycerol stock and inoculate 40 mL of sterile autoclaved LB in a 125 mL baffled flask supplemented with 100 mg/mL ampicillin or 50 mg/mL Kanamycin. You will use 10 mL of starter culture per liter of growth you plan to do, so upscale if you plan to do more than 4L of growth. Be sure to use a baffled flask with a capacity at least twice as much as the volume of culture you are using.
- 2 – Incubate starter culture at 37 °C overnight with ~180 rpm shaking.

### **Day 3 – Growth and Expression**

- 1 – Pitch 10 mL of starter culture into each 1L growth (1L LB in 2L baffled flasks) supplemented with 100 mg/mL ampicillin or 50 mg/mL Kanamycin. Optional – pellet a mL of cells and resuspend in sterile 50% glycerol for a glycerol stock.
- 2 – Grow 1L growths at 37 °C with shaking ~180 rpm until an OD<sub>600</sub> of 0.5-0.8. Typically, the cultures will reach this OD<sub>600</sub> after 2-3 hours and have a doubling time around 20-30 minutes.
- 3 – Incubate cultures on ice for 40 min
- 4 – Induce with 0.5-1 mL of 1M IPTG per liter.
- 5 – Grow at 18 °C overnight (~18-22 hrs.)

#### **Day 4 – Pellet cells and freeze**

1 – Spin the cells down in 500 mL centrifuge bottles with the F10-6 Fiberlite or J10 rotor in the floor centrifuges at ~5000 rpm (roughly 9000g). Use 1 bottle per 1L growth, keeping the cultures separate, spinning up to ~400 mL at a time. Do not fill the centrifuge bottles to the top or the bottle will likely leak – fill up to the lip of the bottle and check that the rubber seal on the lid is flush with the plastic.

2 – Decant as much liquid as possible and scrap pellets into 50 mL tubes and freeze at -20 °C or continue with the next steps.

#### **Nickel-NTA Affinity Column Purification**

1 – Resuspend pellet in 50 mL of lysis buffer (50 mM Tris pH 8.5, 1000 mM NaCl, 10 mM imidazole pH 8.3, 10% glycerol, 0.1% v/v Triton 100X) with a Roche EDTA-free inhibitor tablet.

\*Note that this prep is done mostly at 4 °C and the pH of Tris buffers has a significant temperature dependence so either pH the buffer at 4 °C or use consult a chart for the equivalent buffer pH from room temperature. The high salt, glycerol, imidazole, and triton are used to reduce non-specific interactions with the column resin due to a co-purifying nuclease.

2 – Pre-equilibrate 2-5 mL Nickel NTA beads (4-10 mL of slurry; use larger volumes for beads that have been regenerated >5 times) per liter of culture with 100 mL of lysis buffer in a Kontes column.

3 – Using the Misonix Sonicator 3000 with a ½” tip, sonicate 10-12 times with 15s pulses and 15s rests at power=7-8. If you observe any foaming, pause the sonication and lower the power. You may need to adjust the tip so that it remains completely submerged during each pulse.

- 4 – Spin 30 minutes at least 15,000g in the 30 mL Oakridge tubes in the F21 rotor in the floor centrifuge. Do not fill the tubes above the lip and check that the seal is flush with the cap, or they may leak.
- 5 – Pour the supernatant over pre-equilibrated nickel-NTA beads and rock on a shaker at 4°C for 0.5-1 hr.
- 6 – Wash with 50 mL of lysis buffer in 3 steps of ~16 mL each.
- 7 – Allow the supernatant to flow just above the top of the beads and then add 15 mL of lysis buffer supplemented with 350 mM imidazole. Rock on a shaker at 4°C for 15 minutes and then allow the resin to reform the column by gravity. Collect elution.
- 8 – Add another 15 mL of 350 mM imidazole lysis buffer and collect the flow through to combine with the first elution.

### **Removing the His-SUMO affinity tag**

While the SELEX protocol requires His-tagged constructs for the capture of the RNA-bound protein complex, many of the constructs used in this work have removable affinity tags. The MBP-MS2, MBP-CypE, FL-CypE, and CypE-RRM constructs all have thrombin cleavage sites between the solubility/affinity tag and the N-terminus of the protein, allowing removal of the affinity tag for to produce native or near native sequences of those protein constructs. In addition, the His-SUMO for SUMO-CypE can also be removed by Ulp1 cleavage leaving an N-terminal Ser residue to the native protein sequence (I also made His-SUMO-Cpr1 and His-SUMO-CypA constructs but did not use those constructs in any of the experiments described in this thesis). Removal of the tag followed by a subsequent nickel-column clean-up step is likely to be sufficient to remove much of or eliminate the RNA-binding contaminant described in Chapter 5.

Dialyze the eluent from the nickel-affinity step overnight in the protein storage buffer you plan to use (If using thrombin to cleave the tag, use a Tris buffer with calcium supplemented – calcium is recommended for the cleavage reaction but will precipitate with phosphate buffers). This is most easily done by using large volume dialysis tubing such as the Spectrum Spectra/Por Dialysis tubing from Thermofisher.

1 – Cut the desired volume of tubing for the eluent dialysis, clamping each end with a plastic clamp (such as chip bag clamps)

2 – Place the dialysis tubing in a large volume lidded container with at least 1L of dialysis buffer and a stir bar and allow to equilibrate for ~10-30 minutes at 4 °C to remove residual glycerol in the membrane. The tubing can be suspended above the stir bar by tying the clamps with floss overhanging the lid and tightening the lid over the floss.

3 – Unclamp one side of the dialysis tubing, being sure that the other side is securely clamped. Then load the dialysis tubing with your sample using a large volume pipet, clamp, and resuspend the tubing in the dialysis buffer.

4 – Dialyze overnight at 4 °C and filter the sample (it is not uncommon to see a substantial precipitate appear; it does not appear to impact protein yield, so it is likely something other than protein).

5 – Add protease to the sample for cleavage during the concentration step (if you observe poor cleavage efficiency, add protease to the concentrated sample and allow overnight cleavage at 4 °C).

6 – Prior to SEC, run the cleaved sample over nickel beads to remove SUMO and Ulp1/uncleaved protein. The thrombin enzyme does not have a His-tag so if it is not important to separate uncleaved protein from cleaved or you have good cleavage efficiency, then this step is not necessary, but may remove the RNA binding contaminant.

## Concentration and Size Exclusion Chromatography

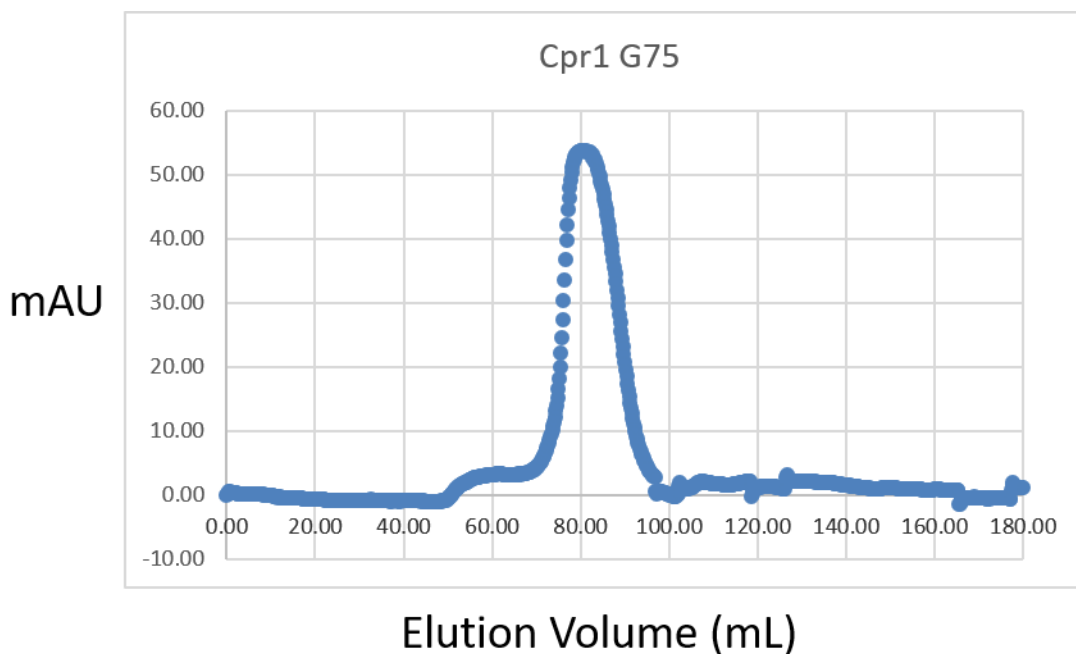
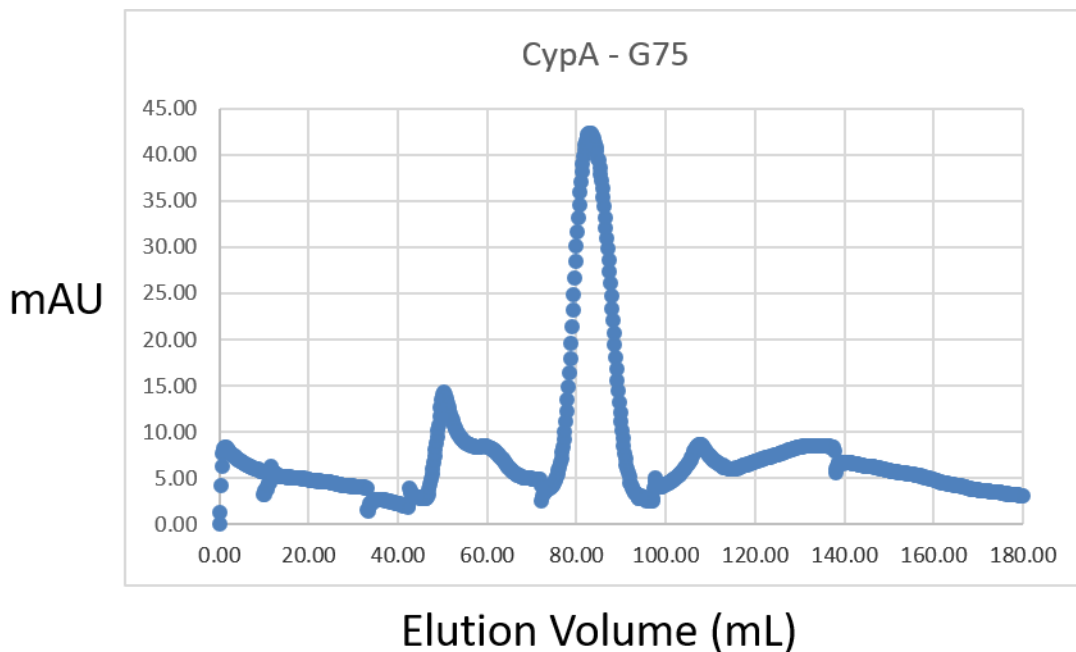
1 – Pre-equilibrate the G75 or G200 Superdex column with filtered and degassed storage buffer (50 mM Tris pH 8.0, 135 mM KCl, 15 mM NaCl, 10% glycerol. You will need at least 180 mL for the equilibration wash and another 180 mL for the fraction collection, so having at least 500 mL is recommended.

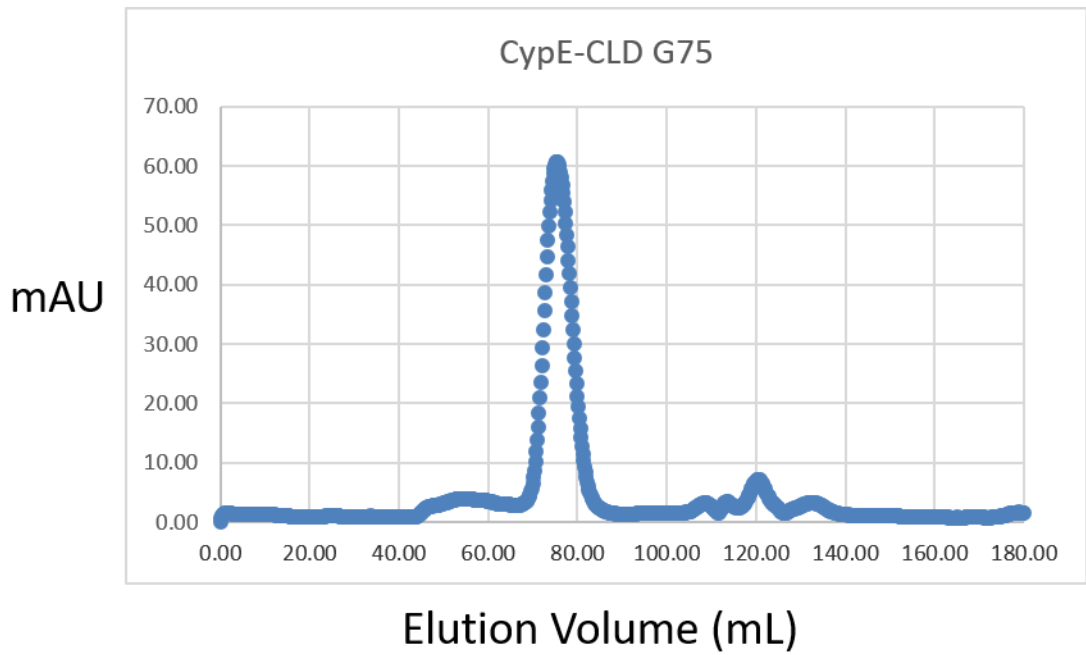
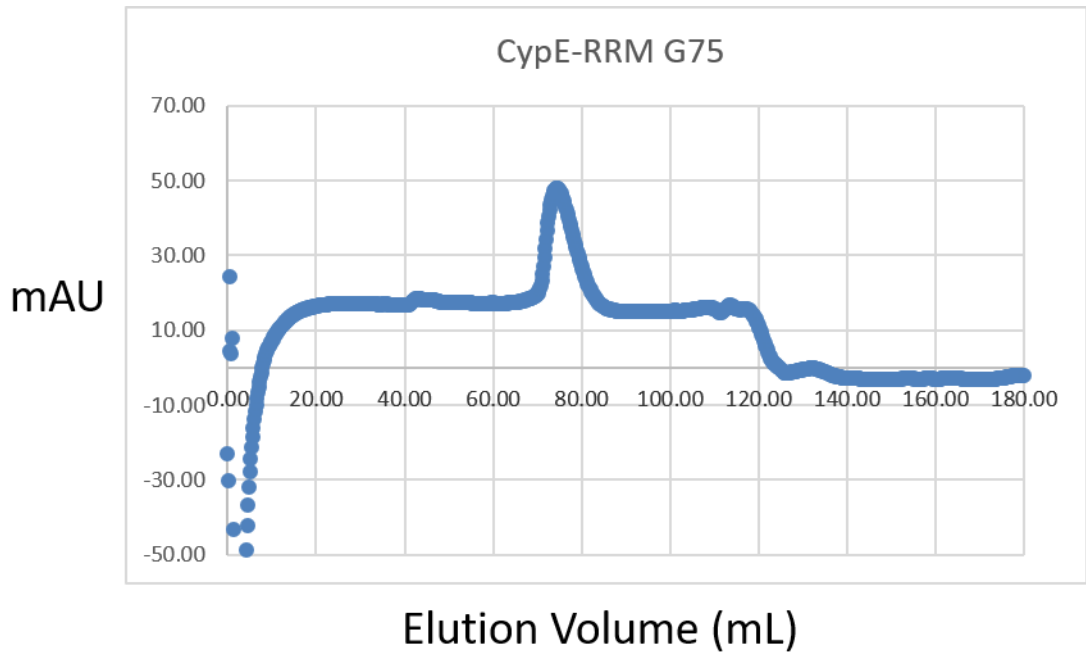
2 - Concentrate the affinity column flow through or dialyzed sample with protease on a 5/10K MWCO concentrator (5K for CypE-RRM, others can use 10K MWCO) at 4 °C down to ~2 mL. To avoid high-concentration and precipitation at the membrane interface, you should gently mix the solution every 15-20 minutes during concentration. If you observe significant precipitation, you should filter the concentrate immediately. Depending on the protein concentration, if you observe precipitation you may want to proceed to SEC with multiple injections if the eluate is still above 2 mL. Alternatively, lowering the salt concentration by mixing the eluent with the storage buffer can help prevent precipitation.

4 – When the eluent is concentrated to ~ 2 mL, filter the solution and inject onto the G75 or G200 column with a flow rate of ~ 1 mL. The expected peaks for each of the constructs are summarized below:

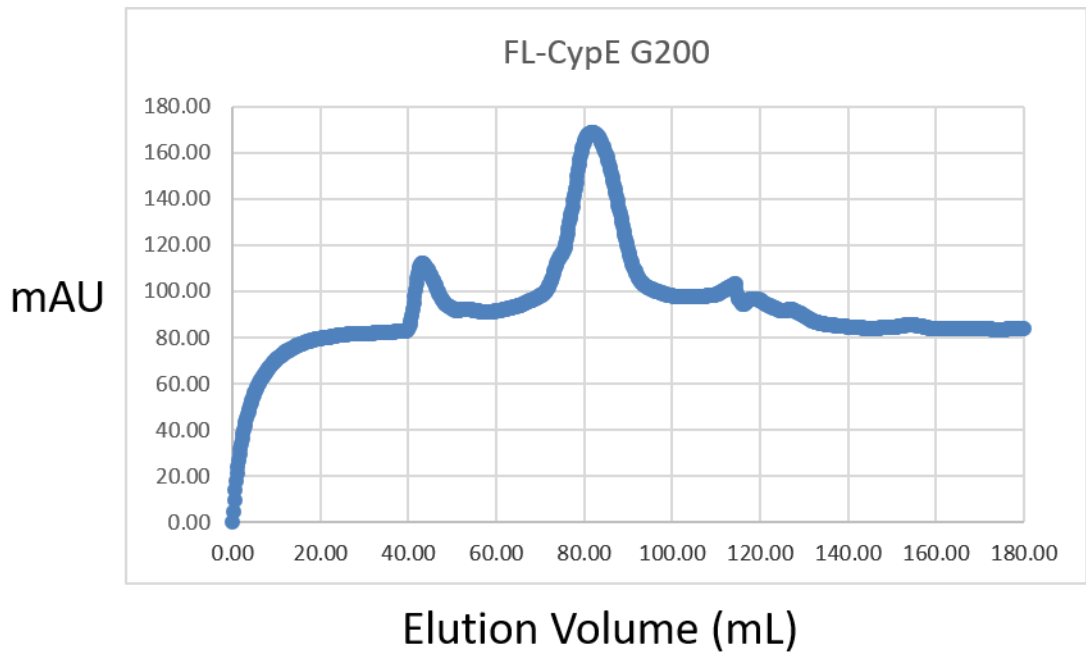
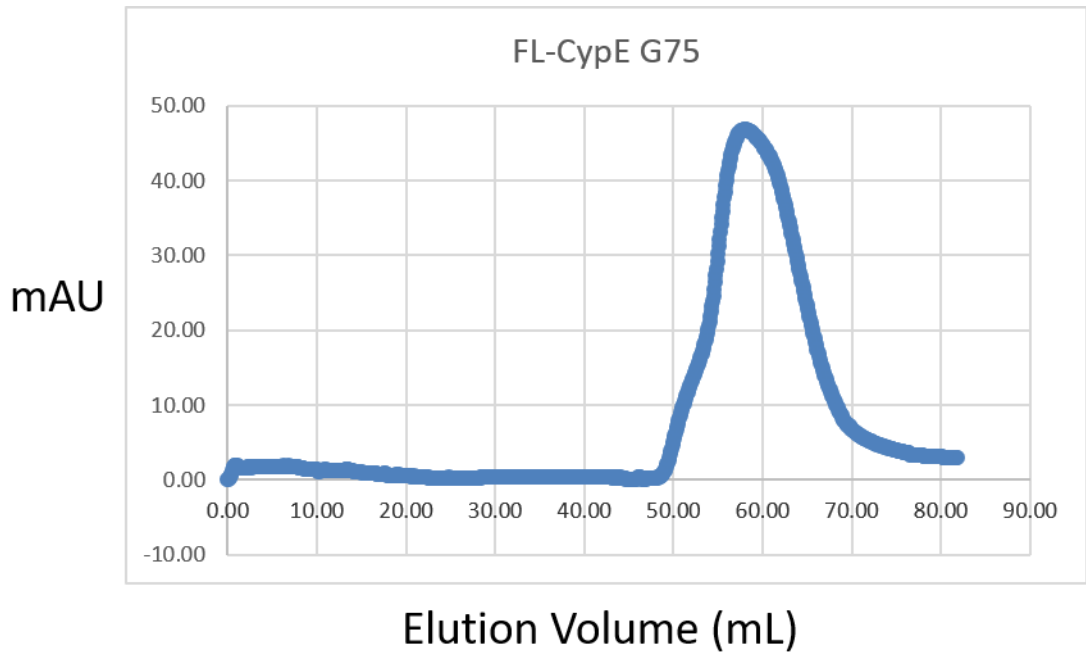
<b>Construct</b>	<b>SEC Column</b>	<b>Observed Elution Peak</b>
MBP-MS2 (V29/dIFG)	G200	~90 mL
CypA	G75	~83 mL
Cpr1	G75	~83 mL
FL-CypE	G75/G200	~59 mL/~83 mL
CypE-RRM	G75	~76 mL
CypE-CLD	G75	~76 mL
MBP-CypE	G200	~85 mL
SUMO-CypE	G200	~79 mL

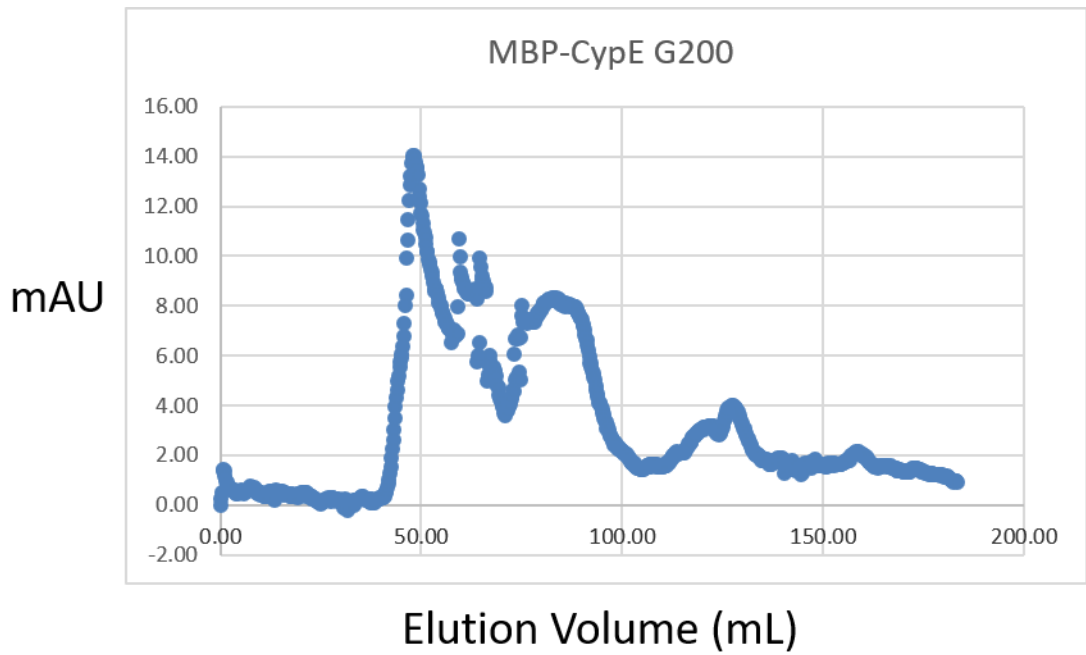
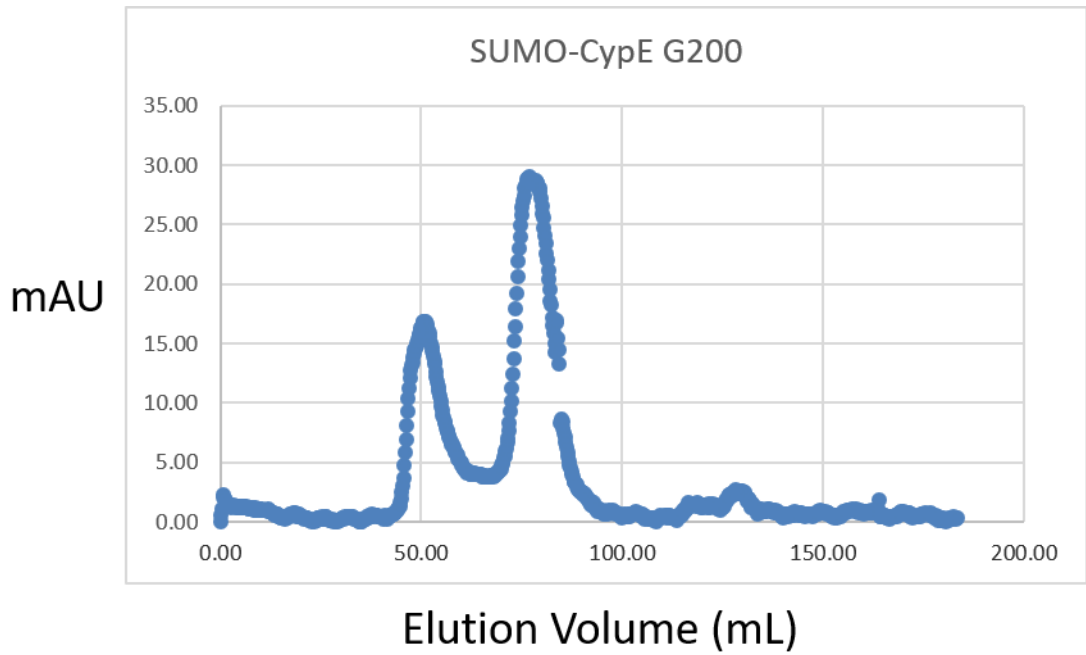
5 – Concentrate the protein fractions centered around the elution peak using a fresh concentrator, flash freeze in liquid nitrogen, and store at -70 °C. My typical yields were about ~500  $\mu$ L of about 400  $\mu$ M to 2 mM protein (highest for CypE-RRM and FL-CypE, lowest for CypE-CLD).











## Template PCR and RNA Library Preparation

Primer or Oligo Name	Sequence
25N Library (RNA Seq)	GGATGGCTTTTCGGGTCATTCTT(N) <sub>25</sub> CCAATCGGGCTTCGGTCCGGT T
25N Library DNA Template	CTCTGTTCTTATTTGCGAGTTCC(N) <sub>25</sub> GTCGGTGTGGTGGTCCG
25N T7 Fwd. PCR Primer	ATATATATGGGTAATACGACTCACTATAGGGAGACAAGAATAAACGC TCAAGG
25N Rev. PCR/RT-Primer	GGCTGGTGGTGTGGCTG
25N P5 Illumina Adapter	AATGATACGGCGACCACCGAGATCTACACATATATATGGGTAATACG ACTCACTATAGG
25N 3' seq adapter	CCGAACCGGACCGAAGCCCGGGCTGGTGGTGTGGCTG
P3-Barcode Index Primer	CAAGCAGAAGACGGCATAACGAGAT(N) <sub>12</sub> AGTCAGTCAGCCGAACCGG ACCGAAGCCCG
25N Sequencing Read Primer	GGGTAATACGACTCACTATAGGGAGACAAGAATAAACGCTCAAGG
Indexing Read Primer	CGGGCTTCGGTCCGGTTCGGCTGACTGACT
50 Library (RNA Seq)	GAGACAAGAATAAACGCTCAAGG(N) <sub>50</sub> CAGCCACACCACCAGCC
50N Library DNA Template	GGCTGGTGGTGTGGCTG(N) <sub>50</sub> CCTTGAGCGTTTATTCTTGTCTC
50N T7 Fwd. PCR Primer	ATATATATGGGTAATACGACTCACTATAGGGAGACAAGAATAAACGC TCAAGG
50N Rev. PCR Primer/RT/3' Annealing Primer	GGCTGGTGGTGTGGCTG
5' Annealing Primer	CTCTGTTCTTATTTGCGAGTTCC
50N P5 Illumina Adapter	AATGATACGGCGACCACCGAGATCTACACATATATATGGGTAATACG ACTCACTATAGG
50N 3' seq adapter	CCGAACCGGACCGAAGCCCGGGCTGGTGGTGTGGCTG
50N Sequencing Read Primer	GGGTAATACGACTCACTATAGG GAGACAAGAATAAACGCTCAAGG

## PCR amplification of initial SELEX library

The amount of material used to generate the initial RNA library was determined based on the number of sequences theoretically in a 25N library ( $4^{25}$  sequences, or  $1.13 \times 10^{15}$ ; 2

nmols X Avogadro's number is  $\sim 1.2 \times 10^{15}$ ). Statistically, this does not actually provide full sequence coverage due to stochastic sequence duplicates and chemical synthesis bias. For the larger 50N libraries, reaching even this pseudo-coverage requires more material than the mass of an average car, so for those libraries, I used the same amount as the 25N library.

The DNA template of the library was chemically synthesized by IDT as the reverse complement of the desired RNA sequences of the 5' and 3' constant regions as well as a complementary T7 promoter sequence attached to the 5' constant sequence (5' to 3' is reverse complementary 3' sequence, 25 or 50N, reverse complementary 5' sequence and reverse complementary T7 promoter sequence). Because of the amount of material necessary to reach pseudo-single coverage of the 25N library, the PCR amplification was separated into multiple aliquots – which also has the advantage of mitigating the possibility of exceedingly well amplifying sequences from overwhelmingly dominating the pool. In addition, an excess of primers and dNTPs were also used with fewer rounds of PCR to accommodate the high concentration of DNA template in the reaction. Notably, running the resulting sample on a native polyacrylamide gel sometimes produces doublet bands which collapse to a single band on a denaturing polyacrylamide gel, suggesting heat cycles beyond the point where primers or dNTPs are exhausted results in products in which annealed constant regions are mismatched in the random region with slower migration speeds due to single-stranded regions of DNA. As T7 RNA polymerase is more efficient with double-stranded template, reducing the number of PCR cycles and/or increasing primers and dNTP concentrations to eliminate or reduce mismatched templates is desirable.

2 nmol of DNA template was split into ten 100  $\mu$ L aliquots of the following reaction conditions: 2  $\mu$ M DNA template, 5  $\mu$ M primers, 1 mM dNTPs, 1X Taq Buffer (10 mM Tris pH 8.3, 50 mM KCl, 1.5 mM MgCl<sub>2</sub>), and 1U/50  $\mu$ L Taq.

#### DNA Template PCR Reaction

100  $\mu$ L 10X Taq buffer (100 mM Tris pH 8.3, 500 mM KCl, 15 mM MgCl<sub>2</sub>)  
2 nmol IDT DNA template

5 nmols T7-Fwd Primer (5')  
5 nmols Rev/RT/Annealing Primer (3')  
100  $\mu$ L 10 mM dNTPs  
20  $\mu$ L Taq  
IDT nuclease free water to 1 mL

5 min 95 °C hot start,  
10 cycles  
95 °C for 45s,  
55 °C for 45s,  
68 °C for 45s.

Combine PCR reactions for RNA transcription

### **RNA Transcription**

RNA was *in vitro* transcribed using the T7 RNA polymerase system. PCR template (10% of final volume) was added to a reaction volume with the final buffer of 40 mM Tris pH 7.9, 24 mM MgCl<sub>2</sub>, 1 mM DTT, 2 mM spermidine. 1U/100  $\mu$ L of T7 RNA polymerase and inorganic pyrophosphatase were then added and incubated at 37 °C for 4-16 hours.

#### RNA Transcription Reaction (Round 0 Library)

1 mL 10X Transcription buffer (400 mM Tris pH 7.9, 240 mM MgCl<sub>2</sub>, 10 mM DTT, 20 mM spermidine)  
1 mL PCR reaction  
100  $\mu$ L T7 RNA polymerase  
100  $\mu$ L Inorganic pyrophosphatase  
400  $\mu$ L 100 mM rATP  
400  $\mu$ L 100 mM rUTP  
400  $\mu$ L 100 mM rCTP  
400  $\mu$ L 100 mM rGTP  
Nuclease free water to 10 mL

Incubate at 37 °C for 16 hours, add 2X loading buffer (95% formamide, 0.5M EDTA, 0.1% bromophenol blue) and run on denaturing acrylamide gel for purification

#### RNA Transcription Reaction (Other Rounds)

25  $\mu$ L 10X Transcription buffer (400 mM Tris pH 7.9, 240 mM MgCl<sub>2</sub>, 10 mM DTT, 20 mM spermidine)  
50  $\mu$ L PCR reaction  
2.5  $\mu$ L T7 RNA polymerase  
2.5  $\mu$ L Inorganic pyrophosphatase  
10  $\mu$ L 100 mM rATP  
10  $\mu$ L 100 mM rUTP  
10  $\mu$ L 100 mM rCTP  
10  $\mu$ L 100 mM rGTP  
Nuclease free water to 250  $\mu$ L

Incubate at 37 °C for 2-4 hours, add 250 µL 2X loading buffer (95% formamide, 0.5M EDTA, 0.1% bromophenol blue) and run on denaturing acrylamide gel for purification

### **RNA Purification**

RNA was purified by gel purification. RNA was mixed with 2X loading buffer (95% formamide, 0.5M EDTA, 0.1% bromophenol blue) and then loaded onto an 8M urea 8% polyacrylamide denaturing slab gel and run at 20-30W for 2-4 hours. RNA bands were visualized by UV shadowing on Fluor-Coated TLC Plate (Fisher Scientific), cut out, and then crushed and soaked between 2 hours to overnight in 0.5X TE pH 7.5 buffer. The gel particles were filtered using 0.22 µM cellulose-acetate filters (ThermoScientific), before being concentrated on a 5K MWCO centrifuge concentrator (Sartorius). Once the RNA volume reached ~0.5 mL, the 1 mL IDT nuclease free water was added and spun again, with the process repeated three times to remove residual urea. RNA concentration and purity were assessed using a NanoDrop Spectrometer using extinction coefficients predicted by IDT Oligo Analyzer.

Our slab gels are ~50 mL volume, make ~56 mL for pouring. Use tape to make ~500 µL loading wells (I usually do three 500 µL samples per gel). Put plastic spacers between the gel plates on the edges, clamp in place with large binder clips (3 clips on the bottom and 2 clips on each of the two sides near the bottom and middle of the plates), ensuring that the spacers on the edges of the gel plates do not have any gaps that would leak. If you have tape on the well comb, wet the comb with a little water so it is easier to slide between the plates while the gel is being cast.

#### Gel Recipe

23 mL of 20% acrylamide, 1X TBE, 8M urea  
33 mL of 1X TBE 8M Urea  
560 µL 10% APS  
50 µL TEMED

Cast the gel in one smooth motion if possible by pipetting or pouring the gel recipe into the plates. Gently shake the gel if any bubbles appear to dislodge them. Once the bubbles are gone, insert the well comb and then clamp the top of both sides of the gel and allow the gel to set ~15 minutes.

Setup the gel apparatus by removing the bottom gel spacer, the well comb, and clamping the gel with binder clips to create a top and bottom reservoir of 1X TBE running buffer (remember to clamp the top reservoir drainage tube). Using a syringe with a bent needle, dislodge any air bubbles at the bottom of the plate so that the bottom of the gel makes full contact with the lower reservoir. Set the power source to a constant wattage of 20-30W. Immediately prior to loading your sample, flush out the well with running buffer via the bent needle syringe to remove any urea that has diffused into the well and may interfere with sample loading. Run the gel for 3-4 hours or until the bromophenol blue dye front is near the bottom of the gel.

Disassemble the gel apparatus and use a gel wedge to pull apart the gel plates. Transfer the gel to plastic wrap, being careful to note the sample orientation if running multiple samples. Put the wrapped gel onto an UV-fluorescent TLC plate, turn off the lights, and expose the gel to UV light using a black light. Quickly (to prevent UV-damage) mark the presence of the RNA bands using a sharpie (larger tips tend to work better). Remove the gel from the TLC plate and cut the bands out with razor blades. Using the tips of the blades, transfer the cut bands to fresh tubes (1.7 Eppendorf for 250  $\mu$ L transcriptions, 15/50 mL tubes for larger transcriptions or as the size of the band requires), and then using a pipette tip, crush the gel into fine pieces by rubbing the gel pieces against the side of the tube. Add cold 0.5X TBE and shake the tube 1 hr. to overnight at 4 °C. Filter the gel bits – for large volumes, use a syringe tip filter; for small volumes, load the crush and soak mixture on a centrifuge filter tube and spin.

At this point, you can either do an ethanol precipitation by adding 70% ethanol and chilling the sample in the ultralow or -20 °C freezer followed by ethanol washes of the pellet or

concentrate the filtered RNA on a large 5K MWCO centrifuge concentrator or mini 3K MWCO centrifuge concentrators. If using the concentrators, once the volume is reduced, wash the RNA sample ~3 times with nuclease free water to remove residual urea from the sample.

### **Library Binding through EMSA**

To quantify the binding affinity of the target proteins for RNA ligands, EMSAs were performed using radiolabeled RNA ligands produced by T7 *in vitro* transcription and purified protein. The 5' phosphate of transcribed RNA ligands were removed using calf intestinal phosphatase (CIP, NEB) and then 5' labeled with <sup>35</sup>P using T4 polynucleotide kinase (PNK, NEB) and <sup>35</sup>P-γ ATP. Labeled ligand with a final concentration of 5 nM was added to 2-fold serial dilutions of the purified protein ranging from 200 μM to 0 nM final concentration in SELEX buffer supplemented with 10% glycerol. Samples were loaded onto a 0.25X TBE 8% polyacrylamide gel and run at 200V at room temperature for 15-20 minutes. The gels were then dried and exposed on a phosphor screen and imaged on an Amersham Typhoon Imaging System. The resulting images were quantified in ImageQuant 5.0 and fit to the quadratic binding equation in Excel using Solver by minimizing the sum of the least squares difference between the data and fit.

#### CIP Reaction

50 pmol RNA  
5 μL NEB Buffer 2 or CutSmart Buffer  
1 μL CIP  
Nuclease free water to 50 μL

37 C, 60 min

#### Phenol/Chloroform/isoamyl alcohol extraction:

Equal volume PCI (25:24:1) pH 6.7 (50 μL)  
Vortex 15 sec  
Centrifuge 15 sec, max speed  
Transfer aqueous phase (top) to new tube  
Add 1 μL glycogen (20 mg/mL – 10 μL stocks in door of -20 – aliquoted from store bought stock)  
Add 125 μL 100 % ethanol



Incubate on ice or in freezer, 30 min/overnight  
Spin max speed 10 min, RT  
Wash with 100  $\mu$ L 70% ethanol (make sure the water used to prep is nuclease free)  
Air dry 3 min  
Add 7  $\mu$ L IDT water

Label RNA via T4 PNK:

7  $\mu$ L CIP RNA  
1  $\mu$ L 10X T4 PNK Buffer  
1  $\mu$ L PNK  
0.5-1  $\mu$ L  $\gamma$ -<sup>32</sup>P-ATP

37 C, 30 min  
Heat to denature enzyme 15 minutes at 65 C  
Ice, 5 min  
Vortex G25 to resuspend the resin. Snap off bottom and crack the lid; insert into collection tube and briefly centrifuge (7 s).  
Add 40  $\mu$ L pre-chilled 0.5X TE to dilute to 1  $\mu$ M in a volume of 50  $\mu$ L.  
Transfer G25 into low-binding 1.5 mL eppy, apply diluted labeling reaction, and briefly centrifuge (7 sec). Hold on ice.  
Assuming 100% recovery, stock is at 1000 nM.

Gel Shift

Thaw radiolabeled RNA on ice

Dilute RNA in 0.5X TE to a concentration  $\sim$ 10X your final desired ligand concentration (1  $\mu$ L per sample well)

Heat shock 90 °C for 5 min

Cool on ice for at least 5 min

Load the top row of a 96 well plate with 1X SELEX buffer supplemented with 10% glycerol (50 mM Tris pH 7, 135 mM KCl, 15 mM NaCl, 2 mM MgCl<sub>2</sub>, 10 mM Imidazole, 10% glycerol).  
Pipette 10  $\mu$ L per gel you intend to run multiplied by 1.25

Add equal volume of 2X protein stock to the first column of the first row, mix well, and then serially dilute the protein by 2 over the columns.

Aliquot the protein dilutions (10  $\mu$ L) to lower rows for each gel. I typically only do 4 gels per 96-well plate to help keep track of which wells are which while loading, skipping every other row.

Make a protein free well with volume 10  $\mu$ L buffer per gel.

Add 1  $\mu$ L of refolded RNA ligand to each well containing protein aliquots and 1  $\mu$ L per gel for the protein free well.

Incubate at RT for 30 min to 1 hr.

Cast 8% native 0.25X TBE acrylamide minigels, outline and number the wells on the plates with a sharpie, and step up the power supply to run at 200V in 0.25 TBE running buffer.

While running, load each well with  $\sim$ 6-8  $\mu$ L of sample, including a protein free lane as well as a bromophenol blue loading dye in an empty lane to track dye migration.

Run 15-20 min

Dry gels and expose a phosphor screen 2 hr. to O/N depending on level of radioactivity and image the screen on a Typhoon.

### Quantification

Crop each image for each gel, rotate so the lanes are roughly level and save as a separate tif file. Open the file in ImageQuant, drawing a grid of 13-14 boxes over the unbound species. Copy the same grid and resize to cover the bound species. Generate a volume report for each grid, with the lanes corresponding to a protein concentration/free RNA. Add an additional column corresponding to the sum of the bound and unbound counts, and then plot bound/total as a function of protein concentration. In another column, plot the fraction bound equation with dummy values for the  $K_D$ , saturation offset (S), background offset (O), and protein activity (N)

$$Fraction\ Bound = S * \left( \frac{N * [Protein]_t}{K_D + N * [Protein]_t} \right) + O$$

Take the difference between the equation fit and the observed experimental value for each protein concentration, square that value, and then sum the square differences. Using the Excel Solver addon, minimize the sum of the square differences by optimizing the values of the  $K_D$ , S, O, and N.

## SELEX

### Pre-selection against Co-NTA beads

1 – Refold RN by incubating at 80 °C for 5 minutes followed by snap cooling on ice. If annealing DNA primers to constant region, add 2X molar ratio of primers to RNA during this step.

Exp	Rounds	[Protein]	Binding Buffer	Wash Buffer	Library	Tags	Target
1	7	500 nM (1) 100 nM (2) 25 nM (3-7)	50 mM Tris pH 7 150 mM NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole	50 mM Tris pH 7 1M NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole	25N SHAPE	6xHis	CypA, Cpr1, CypE, RRM, CLD
2	8	100 nM	<b>“Physiological”</b> 50 mM Tris pH 7 135 mM KCl 15 mM NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole	50 mM Tris pH 7 135 mM KCl 1M NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole	50N Anneal	6xHis; 6xHis-MBP	CypA, Cpr1, CypE, RRM, CLD, MBP- MS2
3	15	100 nM 500 nM 1000 nM	<b>“Physiological”</b> 50 mM Tris pH 7 135 mM KCl 15 mM NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole; <b>“Low Salt”</b> 50 mM Tris pH 7 45 mM KCl 5 mM NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole	<b>“Physiological”</b> 50 mM Tris pH 7 135 mM KCl 15 mM NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole; <b>“Low Salt”</b> 50 mM Tris pH 7 45 mM KCl 5 mM NaCl 2 mM MgCl <sub>2</sub> 10 mM Imidazole	50N Anneal	<b>Alternating</b> 6xHis-MBP; 10xHis-SUMO; <b>Last 7 R</b> 6xHis	CypE

2 – Add 1.1X selection buffer and incubate RNA with Co-NTA beads for 15 minutes

3 – Separate Co-NTA beads using a magnetic stand, transfer supernatant to binding equilibrium reaction.

### **Binding Equilibrium Reaction, Washing, and Elution**

4 – Incubate protein with RNA in 1X SELEX buffer for 1 hr.

5 – Add 1  $\mu$ L Co-NTA beads, gently mix, and incubate 15 min

6 – Place the sample tube on a magnetic stand to separate the Co-NTA resin, ~5 min.

7 – Remove supernatant and wash 3 times with wash buffer, separating the resin ~30s to 1 min between each wash.

8 – Add 20  $\mu$ L of 1X SELEX buffer supplemented with 350 mM imidazole and incubate 15 min.

9 – Separate Co-NTA beads and carefully remove supernatant for input in RT reaction

### **Reverse Transcription(RT)-PCR**

First 1  $\mu$ M RT primer complimentary to the 3' region of the RNA was added to 14  $\mu$ L of eluted RNA. In a thermocycler, the protein was denatured at 80 °C for 10 minutes and then cooled to 4 °C over ~15 minutes. After annealing, 4  $\mu$ L of 5X RT buffer (100 mM Tris pH 7.5, 50 mM NaCl, 50 mM MgCl<sub>2</sub>, 5 mM DTT) was added along with 1  $\mu$ L of 10 mM dNTPs and 1U of reverse transcriptase. The RT reaction was performed at 60 °C for 20 minutes followed by 80 °C for 10 minutes utilizing a thermostable group II intron reverse transcriptase.<sup>286</sup> The full RT reaction was then used as the template for a 500  $\mu$ L PCR reaction aliquoted into 100  $\mu$ L with the following reaction conditions: 20 $\mu$ L/500 $\mu$ L RT-PCR template, 1  $\mu$ M primers, 0.5 mM dNTPs, 1X Taq Buffer (10 mM Tris pH 8.3, 50 mM KCl, 1.5 mM MgCl<sub>2</sub>), and 1U/50  $\mu$ L Taq. PCR amplification was performed for 10 cycles of 95 °C for 45s, 55 °C for 45s, 68 °C for 45s.

#### RT-Annealing

14  $\mu$ L eluted RNA

1  $\mu$ L 20  $\mu$ M RT Primer

In thermocycler: 80 °C for 10 minutes then cool to 4 °C.

#### RT-Reaction

Add 4  $\mu\text{L}$  RT Buffer (100 mM Tris pH 7.5, 50 mM NaCl, 50 mM  $\text{MgCl}_2$ , 5 mM DTT) and 1  $\mu\text{L}$  RT

In thermocycler: 60  $^\circ\text{C}$  for 20 minutes followed by 80  $^\circ\text{C}$  for 10 minutes.

#### RT-PCR Reaction

50  $\mu\text{L}$  10X Taq buffer (100 mM Tris pH 8.3, 500 mM KCl, 15 mM  $\text{MgCl}_2$ )

20  $\mu\text{L}$  RT Reaction

5  $\mu\text{L}$  100  $\mu\text{M}$  T7-Fwd Primer (5')

5  $\mu\text{L}$  100  $\mu\text{M}$  Rev/RT/Annealing Primer (3')

25  $\mu\text{L}$  10 mM dNTPs

10  $\mu\text{L}$  Taq

IDT nuclease free water to 0.5 mL

5 min 95  $^\circ\text{C}$  hot start,

10 cycles

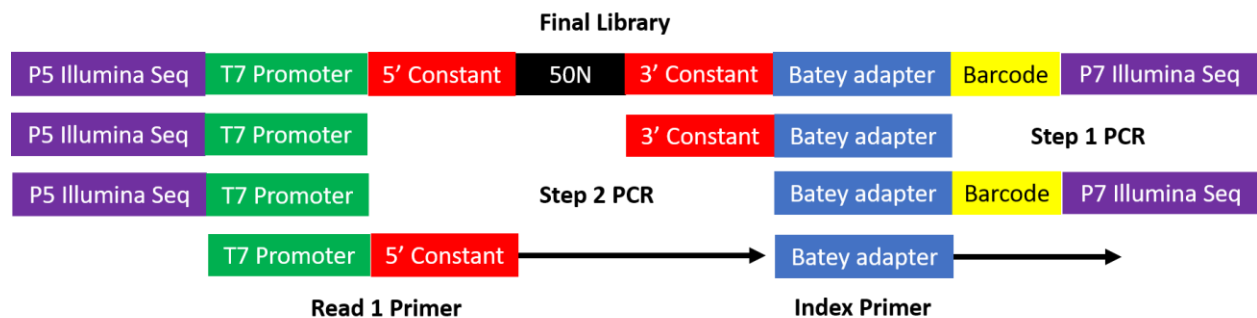
95  $^\circ\text{C}$  for 45s,

55  $^\circ\text{C}$  for 45s,

68  $^\circ\text{C}$  for 45s.

### High-throughput sequencing

The libraries for high-throughput sequencing required the Illumina Adapter sequences appended to the 5' and 3' ends of the DNA libraries produced by each round of selection. In addition, sequencing multiple libraries at the same time required a unique barcode sequence corresponding to each library as part of one of those appended sequences. To convert the selection libraries to this sequence, I performed two steps of PCR using adapter sequence primers as well as the Illumina primers. After the correct library size was generated, the libraries were pooled at a roughly equimolar ratio (as determined by nanodrop) and gel purified on a native 8% polyacrylamide gel, crush and soaked, and then concentrated prior to submission to the CU Boulder Biofrontiers Sequencing Core for MiSeq or NextSeq sequencing.



PCR step 1

10  $\mu$ L 10X Taq buffer (100 mM Tris pH 8.3, 500 mM KCl, 15 mM MgCl<sub>2</sub>)  
10  $\mu$ L PCR library  
2  $\mu$ L 100  $\mu$ M P5-T7-primer  
2  $\mu$ L 100  $\mu$ M Batey adapter primer  
2  $\mu$ L 10 mM dNTPs  
2  $\mu$ L Taq  
IDT nuclease free water to 100  $\mu$ L

5 min 95 °C hot start,

10 cycles

95 °C for 45s,

55 °C for 45s,

68 °C for 45s.

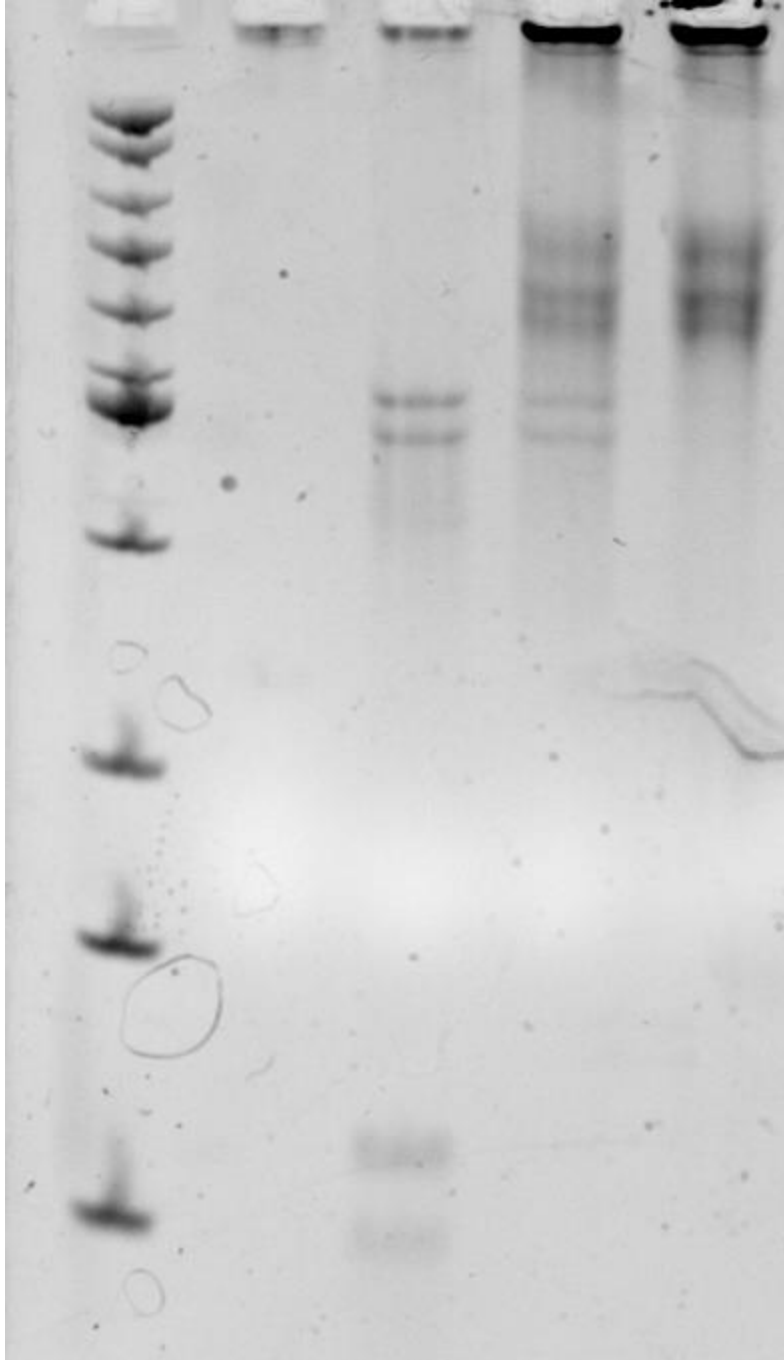
Perform a PCR-clean up step with the Omega Cycle Pure Kit – be sure to follow the steps for adding isopropanol to increase the retention efficiency for short PCR products

PCR step 2

10  $\mu$ L 10X Taq buffer (100 mM Tris pH 8.3, 500 mM KCl, 15 mM MgCl<sub>2</sub>)  
10  $\mu$ L PCR library from step 1 (cleaned up)  
2  $\mu$ L 100  $\mu$ M P5-T7-primer  
2  $\mu$ L 100  $\mu$ M Batey barcode primer  
2  $\mu$ L 10 mM dNTPs  
2  $\mu$ L Taq  
IDT nuclease free water to 100  $\mu$ L

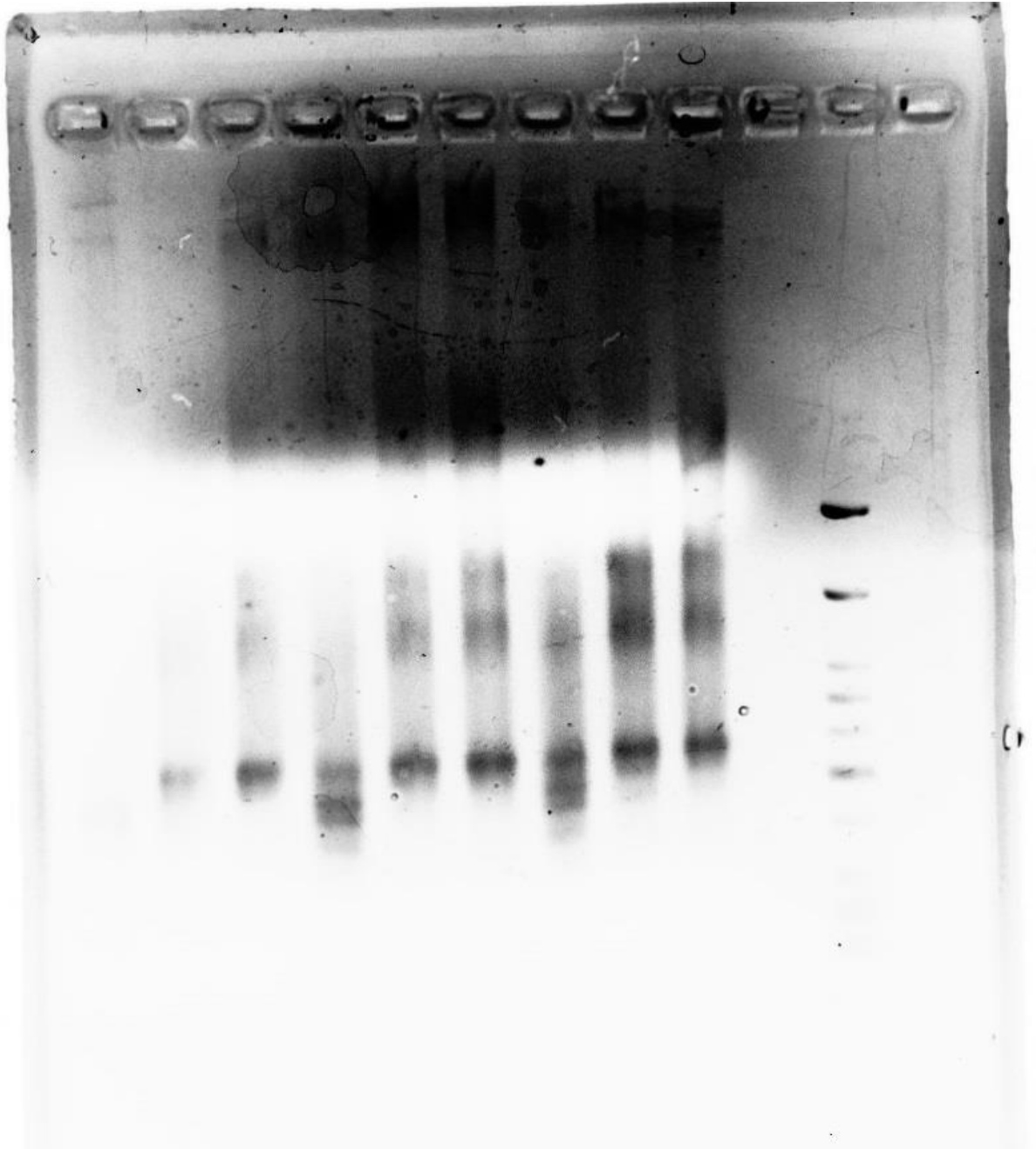
Perform a PCR-clean up step with the Omega Cycle Pure Kit – be sure to follow the steps for adding isopropanol to increase the retention efficiency for short PCR products

Nanodrop each library to get a rough idea of the concentration of each and pool each sample at a roughly equimolar ratio. Load the combined sample on a native 8% acrylamide gel with one large well made by taping the comb with a lane for a standard DNA ladder. Cut out the ladder lane with a razor blade as well as a small portion of the sample lane, stain the cut ladder piece with ethidium bromide and image with UV – marking the regions containing the sample. Re-align the cut ladder piece to the original gel and cut out the unstained sample using the stain as a reference. Crush and soak the cut region as described for the RNA gel purification above.



The gel above is shown for the library preparation for Selection 1. The doublet is the result of an alternative register for the annealing of the barcoding primer but does not appear matter for the sequencing or data analysis, especially as the sequence 3' of the constant region is discarded. The first lane is a 50-nt ladder, the second lane is empty, the third lane is the first PCR product, the fourth lane and fifth lanes are the second PCR product with 5 and 10 rounds of

amplification. The apparently higher MW species above the expected doublet is likely improperly annealed dsDNA from the library. The high MW species in stuck in the well are the motivation for size purifying the library as shown on the agarose gel below. The ladder is identical and the lanes are the same as the above gel but for 3 different samples.





## CypE-RRM HSQC Assignments

RRM assignments were transferred from (Hom et al. 2010) based on similar chemical shifts. Note that while the plasmid construct was identical, I did not cleave the His-tag via the thrombin site as Hom et al did so my spectra have several peaks not seen in the Hom et al. assignments. The assignment file (tab delimited) is reproduced exactly for both the no-RNA and 1.5 RNA titration. Assignments denoted with a ? are tentative assignments.

### No RNA Peak List

Number	#	Position F1	Position F2	Assign F1	Assign F2	Height
	Volume	Line Width F1 (Hz)	Line Width F2 (Hz)			
1	1688	8.97844	120.65697	99Asn[15]	99Asn[16]	4.30E+04
	2.88E+05	24.36528	36.76119			
2	1695	8.8746	119.1767	99Asn[13]	99Asn[14]	1.69E+04
	34.37619	34.3586				1.18E+05
3	1670	9.02785	123.17937	98Val[107]	98Val[108]	3.85E+04
	2.68E+05	26.55271	36.51102			
4	1680	8.8563	121.79454	96Ile[33]	96Ile[34]	3.42E+04
	28.50111	37.04299				2.42E+05
5	1696	8.46437	119.27222	95Thr[90]	95Thr[91]	2.55E+04
	1.78E+05	26.48519	39.71953			
6	1697	7.60023	119.20985	94Arg[100]	94Arg[101]	5.52E+04
	3.80E+05	25.04747	37.80611			
7	1656	9.69823	127.68102	92Phe[5]	92Phe[6]	3.74E+04
	2.63E+05	27.6963	36.66981			
8	1658	9.0761	126.84555	91Leu[68]	91Leu[69]	4.55E+04
	24.03349	37.41274				3.12E+05

9	1702	8.35202	117.7828	90Glu[92]	90Glu[93]	4.37E+04	
		2.89E+05	22.19142	35.87248			
10	1715	7.69946	115.19874	89Ser[58]	89Ser[59]	4.63E+04	
		3.18E+05	24.51817	37.2153			
11	1683	8.66634	121.55792	88Glu[27]	88Glu[28]	3.47E+04	
		2.41E+05	26.37693	37.96949			
12	1692	8.22567	119.61017	87Asn[64]	87Asn[65]	4.80E+04	
		3.18E+05	23.69434	36.45694			
13	1721	8.30274	113.65651	86Met[46]	86Met[47]	4.09E+04	
		2.77E+05	24.26749	36.06207			
14	1722	7.44285	112.77444	85Asn[52]	85Asn[53]	4.32E+04	
		2.96E+05	24.61557	37.36517			
15	1708	7.45508	116.62186	84Asp[56]	84Asp[57]	5.76E+04	
		3.78E+05	21.07694	37.98518			
16	1700	7.88359	118.74976	83Ile[96]	83Ile[97]	4.16E+04	
		2.85E+05	27.21395	36.46479			
17	1676	7.8906	122.19549	82Ala[86]	82Ala[87]	5.66E+04	3.71E+05
		20.87438	38.37888				
18	1687	7.3495	121.43781	81Ala[82]	81Ala[83]	7.68E+04	5.00E+05
		20.98497	36.98449				
19	1681	6.99911	121.75255	79Ala[76]	79Ala[77]	5.75E+04	
		3.79E+05	22.4875	36.42465			
20	1667	6.82126	124.21317	78Asp[74]	78Asp[75]	4.93E+04	
		3.42E+05	24.81309	38.48269			
21	1709	9.37574	116.32495	75Leu[37]	75Leu[38]	3.35E+04	
		2.33E+05	25.84685	37.22715			

22	1660	8.60622	126.84902	73?Phe[111]	73?Phe[112]	3.55E+04	
		2.48E+05	27.00751	39.43573			
23	1655	9.04889	129.42425	72Glu[70]	72Glu[71]	3.80E+04	
		2.64E+05	26.64313	36.39402			
24	1672	8.72931	122.69732	71Val[31]	71Val[32]	3.77E+04	
		2.66E+05	29.41148	37.79209			
25	1717	8.74903	114.4813	70Phe[11]	70Phe[12]	4.19E+04	
		2.92E+05	27.0824	36.30006			
26	1689	8.69308	119.72074	69Ala[25]	69Ala[26]	3.66E+04	
		2.51E+05	26.16234	36.62367			
27	1716	7.01474	114.8037	68Phe[7]	68Phe[8]	5.43E+04	
		3.74E+05	24.96016	37.06226			
28	1720	11.04849	113.65469	67Gly[88]	67Gly[89]	8073.12695	
		5.60E+04	30.84804	38.75439			
29	1701	8.06355	117.90206	65His[60]	65His[61]	2.21E+04	
		1.63E+05	27.8651	97.1324			
30	1711	8.33028	116.15887	63Glu[43]	63Glu[44]	1.29E+04	
		9.13E+04	29.68389	36.75618			
31	1728	7.81337	107.46439	62Thr[9]	62Thr[10]	1.61E+04	
		1.14E+05	30.41148	38.68964			
32	1704	8.56377	117.62781	61Glu[23]	61Glu[24]	8636.5293	
		6.01E+04	23.10528	41.25011			
33	1673	8.5056	122.40477	59Asp[29]	59Asp[30]	3.60E+04	2.55E+05
		26.56397	40.14828				
34	1669	8.75106	123.59527	56?/46?Ile([117]/[119])			
				56?/46?Ile([120]/[118])	4.47E+04	3.07E+05	24.4196
							38.59409

35	1663	9.01703	126.17211	55Gln[19]	55Gln[20]	4.32E+04
		3.05E+05	25.93921	39.10086		
36	1671	7.94476	123.04974	54Ile[35]	54Ile[36]	4.94E+04
		3.41E+05	25.65011	36.5847		
37	1690	7.56248	119.85135	53Asp[102]	53Asp[103]	4.74E+04
		3.50E+05	27.05175	51.72749		
38	1677	9.04269	121.75978	52Thr[17]	52Thr[18]	3.16E+04
		2.23E+05	28.80302	35.88275		
39	1666	7.52156	124.53926	51Ile[78]	51Ile[79]	3.81E+04
		2.70E+05	27.70406	37.94721		
40	1718	8.03191	114.34196	50Asp[48]	50Asp[49]	5.54E+04
		3.67E+05	22.90078	37.03506		
41	1729	7.61375	106.22219	49Gly[113]	49Gly[114]	4.78E+04
		3.18E+05	22.37109	37.59716		
42	1725	7.59117	112.19873	48Phe[54]	48Phe[55]	3.94E+04
		2.64E+05	24.77118	35.21109		
43	1723	7.87942	112.55568	45Phe[50]	45Phe[51]	3.97E+04
		2.74E+05	26.53254	35.47928		
44	1691	8.61468	119.54645	42His[104]	42His[105]	5.37E+04
		3.71E+05	25.11576	37.67409		
45	1699	8.06175	118.73969	41Leu[62]	41Leu[63]	5.17E+04
		3.54E+05	23.8938	39.16562		
46	1674	8.32124	122.32884	40Val[66]	40Val[67]	6.10E+04
		4.15E+05	23.50575	40.96297		
47	1686	7.64425	121.40993	39Lys[84]	39Lys[85]	5.84E+04
		3.80E+05	21.10488	36.58445		

48	1713	8.78671	116.04964	38Asp[41]	38Asp[42]	3.83E+04	
		2.53E+05	23.09635	37.57496			
49	1664	6.74845	125.13847	37Asp[72]	37Asp[73]	4.94E+04	
		3.47E+05	27.12509	37.29199			
50	1698	7.5332	119.25661	36Val[98]	36Val[99]	5.92E+04	4.16E+05
		26.39012	40.87477				
51	1719	9.01411	114.27585	35Glu[39]	35Glu[40]	1.11E+04	
		7.70E+04	24.46966	39.85188			
52	1727	8.11798	107.64594	31Gly[115]	31Gly[116]	3.16E+04	
		2.18E+05	25.47782	40.12308			
53	1724	9.31107	112.35301	30Gly[1]	30Gly[2]	3.78E+04	
		2.60E+05	26.98811	36.16138			
54	1657	8.23891	127.52157	29Val[21]	29Val[22]	3.89E+04	
		2.70E+05	27.576	36.03053			
55	1668	9.25675	123.83894	28Tyr[3]	28Tyr[4]	3.44E+04	
		2.39E+05	26.39439	38.10534			
56	1703	7.70392	117.73348	26?Val[109]	26?Val[110]	5.91E+04	
		4.32E+05	34.85699	40.17749			
57	1712	8.00966	116.19635	25Arg[94]	25Arg[95]	7187.47021	
		5.23E+04	84.70551	42.07556			
58	1730	8.73067	102.13704	None	None	7053.93262	4.76E+04
		31.23725	34.9751				
59	1714	7.94245	115.30474	None	None	1.45E+04	1.02E+05
		30.88219	35.47127				
60	1710	7.93901	116.31013	None	None	4.33E+04	2.94E+05
		24.61673	37.18626				

61	1707	7.87044	117.32996	None	None	5.83E+04	3.92E+05	
		24.18056	37.61985					
62	1705	7.81435	117.6136	None	None	1.08E+05	7.56E+05	
		25.86625	38.92128					
63	1685	8.1608	121.39172	None	None	3.46E+04	2.41E+05	27.24577
			37.67067					
64	1684	8.33601	121.57299	None	None	3.47E+04	2.50E+05	
		27.07464	87.37425					
65	1675	8.24282	122.07331	None	None	4.07E+04	2.94E+05	
		30.66294	39.32074					
66	1662	8.69253	126.44745	None	None	3.48E+04	2.46E+05	
		31.31486	37.01487					
67	1659	8.79417	126.79778	None	None	3.76E+04	2.58E+05	
		24.47471	36.97398					
68	1616	-5.22435	102.73522	None	None	3956.98828	2.87E+04	
		22.79717	46.06664					
69	1614	-5.58954	103.04905	None	None	3801.71582	2.06E+04	
		15.86504	33.20053					
70	1603	-5.68098	112.41639	None	None	3905.61255	2.60E+04	
		25.46695	37.23333					
71	1602	-4.77339	112.398	None	None	3849.34814	2.44E+04	
		22.54338	33.68535					
72	1499	-4.76674	126.77367	None	None	4023.70068	2.70E+04	
		18.89183	55.43092					

### SO-1 1.5:1 RRM Peak List

Number	#	Position F1	Position F2	Assign F1	Assign F2	Height
	Volume	Line Width F1 (Hz)	Line Width F2 (Hz)			
1	32	8.94206	119.58905	99Asn[13]	99Asn[14]	1.86E+04
	1.41E+05	31.85891	48.16215			
2	14	9.00319	122.8924	98Val[107]	98Val[108]	3.37E+04
	2.32E+05	24.80183	37.59524			
3	23	8.82009	121.79354	96Ile[33]	96Ile[34]	3.09E+04
	2.15E+05	27.64779	36.128			
4	34	8.46967	119.24708	95Thr[90]	95Thr[91]	3.07E+04
	2.11E+05	23.76419	39.36605			
5	35	7.59375	119.27031	94Arg[100]	94Arg[101]	6.06E+04
	4.14E+05	23.62915	40.33237			
6	2	9.68907	128.0833	92Phe[5]	92Phe[6]	2.99E+04
	2.13E+05	28.89848	38.34934			
7	3	9.07784	126.94578	91Leu[68]	91Leu[69]	4.66E+04
	3.17E+05	23.88953	36.95503			
8	41	8.33473	117.78079	90Glu[92]	90Glu[93]	4.88E+04
	3.25E+05	22.6241	36.22479			
9	53	7.69566	115.17871	89Ser[121]	89Ser[122]	5.16E+04
	3.53E+05	24.07307	38.4403			
10	25	8.63636	121.50515	88Glu[27]	88Glu[28]	3.15E+04
	2.20E+05	25.7219	39.35604			
11	33	8.22147	119.63761	87Asn[64]	87Asn[65]	4.62E+04
	3.11E+05	23.26748	37.06994			
12	59	8.30965	113.6995	86Met[46]	86Met[47]	3.84E+04
	2.64E+05	25.38042	37.44411			
13	61	7.42418	112.72812	85Asn[52]	85Asn[53]	4.38E+04
	2.99E+05	24.621	37.32462			
14	46	7.44146	116.57119	84Asp[56]	84Asp[57]	6.02E+04
	3.94E+05	21.5166	36.8468			
15	39	7.88079	118.78137	83Ile[96]	83Ile[97]	4.00E+04
	2.71E+05	23.6241	38.28692			
16	21	7.89628	122.23133	82Ala[86]	82Ala[87]	5.54E+04
	3.73E+05	21.84916	39.92865			

17	28	7.34575 4.65E+05	121.46252 21.1755	81Ala[82] 37.80661	81Ala[83]	7.10E+04	
18	24	7.00408 3.53E+05	121.78559 23.28999	79Ala[76] 36.84338	79Ala[77]	5.28E+04	
19	10	6.82801 3.58E+05	124.24174 23.59383	78Asp[74] 38.15299	78Asp[75]	5.27E+04	
20	47	9.36553 2.36E+05	116.30819 25.69512	75Leu[37] 39.34619	75Leu[38]	3.36E+04	
21	4	8.57231 2.11E+05	126.73539 28.33309	73?Phe[111] 39.8185	73?Phe[112]	2.96E+04	
22	1	9.03593 2.52E+05	129.46083 26.21278	72Glu[70] 36.04814	72Glu[71]	3.65E+04	
23	12	8.74484 3.18E+05	123.60215 28.12626	71Val[31] 39.20417	71Val[32]	4.50E+04	
24	57	8.61519 1.41E+05	114.00245 35.99785	70Phe[11] 35.6657	70Phe[12]	1.97E+04	
25	54	6.91482 2.78E+05	114.76289 25.99198	68Phe[7] 37.49669	68Phe[8]	4.01E+04	
26	58	10.97776 5.86E+04	113.66497 45.14419	67Gly[88] 41.9282	67Gly[89]	7658.49512	
27	42	8.06541 2.28E+05	117.88393 26.49024	65His[60] 39.446	65His[61]	3.26E+04	
28	50	8.37045 1.76E+05	116.11999 27.00867	63Glu[43] 40.00008	63Glu[44]	2.48E+04	
29	66	7.90236 1.46E+05	107.73747 31.04362	62Thr[9] 34.52065	62Thr[10]	2.08E+04	
30	40	8.58629 1.57E+05	117.78097 22.26981	61Glu[23] 36.61633	61Glu[24]	2.37E+04	
31	16	8.5501 27.09366	122.39933 40.2218	59Asp[29]	59Asp[30]	3.14E+04	2.24E+05
32	13	8.68431 56?/46?Ile([120]/[118])	123.19229 3.16E+04	56?/46?Ile([117]/[119]) 2.32E+05	32.4274	42.04085	
33	7	8.9944 26.25663	125.89521 37.05993	55Gln[19]	55Gln[20]	3.57E+04	2.49E+05
34	15	7.95032 3.51E+05	122.75607 23.59771	54Ile[35] 39.53745	54Ile[36]	5.14E+04	



35	30	7.56818	119.93757	53Asp[102]	53Asp[103]	5.09E+04	
		3.50E+05	23.87245	49.8683			
36	22	9.07198	121.98823	52Thr[17]	52Thr[18]	2.97E+04	
		2.10E+05	28.08823	37.49785			
37	9	7.51138	124.51643	51Ile[78]	51Ile[79]	3.80E+04	
		2.68E+05	27.21589	37.84099			
38	55	8.02569	114.3518	50Asp[48]	50Asp[49]	6.48E+04	
		4.15E+05	20.22245	36.84981			
39	67	7.60922	106.22351	49Gly[113]	49Gly[114]	4.58E+04	
		3.04E+05	22.33034	36.98583			
40	64	7.5846	112.19665	48Phe[123]	48Phe[124]	3.71E+04	2.53E+05
		24.24304	37.03406				
41	62	7.88332	112.53379	45Phe[50]	45Phe[51]	3.49E+04	
		2.41E+05	25.1569	37.51621			
42	31	8.61781	119.70334	42His[104]	42His[105]	6.94E+04	
		4.87E+05	25.00478	42.31238			
43	38	8.02448	118.65054	41Leu[62]	41Leu[63]	4.05E+04	
		2.75E+05	25.19532	38.15291			
44	19	8.3033	122.32102	40Val[66]	40Val[67]	5.79E+04	3.95E+05
		27.93728	37.77982				
45	26	7.63526	121.51995	39Lys[84]	39Lys[85]	5.53E+04	
		3.65E+05	21.42114	37.7386			
46	51	8.78594	116.06228	38Asp[41]	38Asp[42]	3.69E+04	
		2.48E+05	22.90116	38.57431			
47	8	6.72913	125.0732	37Asp[72]	37Asp[73]	4.50E+04	
		3.19E+05	28.37616	38.57698			
48	36	7.51613	119.26007	36Val[98]	36Val[99]	5.81E+04	
		3.93E+05	24.68813	37.93812			
49	56	8.99918	114.28369	35Glu[39]	35Glu[40]	1.44E+04	
		1.02E+05	30.09018	38.86421			
50	65	8.11915	107.66383	31Gly[115]	31Gly[116]	3.36E+04	
		2.35E+05	26.77623	37.10198			
51	63	9.20098	112.42617	30Gly[1]	30Gly[2]	2.83E+04	
		2.00E+05	28.64236	37.47766			
52	11	9.13751	124.01657	28Tyr[3]	28Tyr[4]	1.70E+04	
		1.22E+05	32.14296	38.67103			

53	43	7.71797 4.20E+05	117.84872 83.19212	26?Val[109] 38.85119	26?Val[110]	5.71E+04
54	48	8.02455 1.15E+05	116.42861 33.67188	25Arg[94] 40.55517	25Arg[95]	1.58E+04
55	68	8.72162 34.62416	102.12758 41.92686	None	None	8513.59863 6.39E+04
56	52	8.00699 33.33855	115.33078 37.60533	None	None	1.74E+04 1.25E+05
57	49	7.93048 23.34664	116.2741 38.55062	None	None	3.95E+04 2.68E+05
58	45	7.86135 24.19919	117.33725 39.84771	None	None	5.42E+04 3.73E+05
59	44	7.80549 22.86663	117.60078 41.18853	None	None	1.15E+05 7.87E+05
60	29	8.40195 34.71962	121.27351 36.86391	None	None	2.13E+04 1.54E+05
61	27	8.15335 24.15767	121.35808 37.18768	None	None	3.45E+04 2.36E+05
62	17	8.25029 59.83965	122.48995 40.32218	None	None	4.13E+04 3.01E+05
63	6	8.69223 31.59348	126.47873 37.55209	None	None	2.96E+04 2.12E+05
64	5	8.80998 22.62487	126.35253 40.48107	None	None	3.45E+04 2.33E+05

## Appendix C

This appendix contains a description of the bioinformatic analysis pipeline and scripts used in Chapters 4 and 5.

### Overview of the Analysis Pipeline

All of my sequencing experiments (both MiSeq and NextSeq) utilized multiplexed samples in which each sample was associated with a specific 12-nt barcode sequence read as part of the indexing read and correlated to a sequencing read based on the physical coordinates of the sequenced cluster on the instrument chip. Making sense of the sequencing data requires separating all of the sequences based on which sample originated the sequence, especially as many of the samples sequences were from unrelated experiments combined for the sequencing run. Demultiplexing of sequences is the process by which this is accomplished and can be performed in AptaSUITE, QIIME, or through the Illumina Software. I did it in QIIME 1.9 alongside quality-control filtering, though the process appears to be more straight-forward and user-friendly for AptaSUITE.

After demultiplexing and quality filtering, I removed the 3' constant region sequences from each of the reads (the read primer contains the 5' constant region so that sequencing reads began with the first nucleotide of the random region and so it was not read and did not need to be removed). For ease of import into AptaSUITE with the QIIME processed reads, I used a script to add in perfect 5' and 3' constant regions and another script to add perfect dummy quality data to trick the AptaSUITE pipeline into accepting the same demultiplexed and quality filtered data from QIIME.

Next, I did clustering with the sequences to try to identify common families of aptamer sequences, using both AptaSUITE and QIIME. Because the AptaSUITE clustering algorithm approximately produces the same results as the one used by FASTAptamer with less computing power, I did not do clustering with FASTAptamer. This clustering was done as a function of each SELEX condition over all rounds of selection. The clustering done in QIIME

was done with all of the sequences from each condition and round to assess whether there were any common “winning” sequences between conditions. To reduce the computational resources necessary for this, the combined sequences were collapsed into an abundance sorted fasta file using Fastx\_collapser to reduce the number of pairwise calculations and file size by eliminating redundant sequences and sorting sequences by sequence count. Then, using all sequences that appeared >5 times, sequences were pairwise aligned with uclust, preferentially starting new clusters based on the most abundant, unclustered sequence.

In addition, AptaSUITE provides a motif finder in which each k-mer (for a 6mer, k=6) in reads appearing more than a certain threshold (I chose 10) is compared to each k-mer found in all sequences appearing less than the threshold to identify sequence motifs enriched during the selections under the assumption that reads <10 are representative of the unselected pool.

### **Demultiplexing with QIIME 1.9**

I used the split\_libraries\_fastq.py script of QIIME 1.9 to demultiplex the sequenced libraries with the following script command.

```
split_libraries_fastq.py -i Undetermined_S0_R1_001.fastq.gz -b  
Undetermined_S0_I1_001.fastq.gz -m batey_barcode_map -q 19 -o demultiplexed/
```

The -i is the fastq file produced for sequencing reads, -b is the fastq file produced for indexing reads, -m is the mapping file which contains the barcode sequences and sample IDs (and possibly other metadata if you so wish). The -q value indicates reads with a Phred quality score below 20 are filtered out (20 = 99% accuracy, with each step of 10 being an additional order of magnitude accuracy, i.e. 30 = 99.9%). The -o refers to the output directory in which each sample is split into an individual fasta file based on the same ID (in this case Barcode###, e.g. all sequencing reads for sample 1 with barcode 1 were all separated into

Barcode001.fasta). The contents of the “batey\_barcode\_map” are pasted at the end of this appendix with the sample IDs corresponding to the Batey lab barcode library, the barcodes are the reverse complement of the IDT primer sequences, the linker sequence is the sequence of the 3’ constant region and adapter sequence from the barcoding primer, and the reverse primer the reverse complement of the 3’ constant region for a subsequent truncation script.

### **Removing 3’ Constant and Illumina Barcode Primer with QIIME 1.9**

The 3’ constant region and subsequent Illumina/Barcode adapter sequences were not present in the RNA libraries used in the selection but present in the sequencing reads due to the library preparation for sequencing. For subsequent analysis, these sequences were removed using the `truncate_reverse_primer.py` script as shown below.

```
truncate_reverse_primer.py -f Barcode###.fasta -m batey_barcode_map -o  
Trim_Barcode###/ -z truncate_remove
```

This script was used for each Barcode ID, corresponding to each round/sample. `-f` is the fasta input file (Barcode001.fasta for sample 1, Barcode002.fasta for sample 2, etc.), `-m` is the mapping file, and `-o` is the output directory for that corresponding library. The `-z truncate_remove` indicates that the script removes the 3’ constant region sequence and all subsequent sequence as well as removing the reads without the 3’ constant region sequence. This script allows for a default of 2 mismatches in the read sequence compared to the primer sequence (fewer allowed mismatches throws out more sequences, but runs much faster)

Following this, I used several Fastx tools from the FASTX Toolkit by the Hannon Lab for further read processing. Using `fastx_collapser`, duplicate sequences were removed and then rank-sorted based on read count. Then, using `fastx_clipper`, reads shorter than 20 nucleotides were removed, and then `fastx_collapser` was used again to re-rank the sequences minus the filtered sequences.

```
fastx_collapser -f Trim_Barcode###/Barcode###sta_rev_primer_truncated.fna -o Trim_Barcode###/coll.fna
fastx_clipper -l20 -i Trim_Barcode###/coll.fna -o Trim_Barcode###/coll_120.fna
fastx_collapser -f Trim_Barcode###/coll_120.fna -o Trim_Barcode###/sorted.fna
```

The -f options are the .fasta (or equivalent .fna) file input from the previous step. The -o option is the output directory and the name of the output file (X.fna)

From here, I clustered the top 500 sequences for each sample library for Selections 1 and 2 (MS2 and each cyclophilin target construct, 50N and 25N), and all sequences that appeared 5 or more times for Selection 3. For selection 3, I wanted to be able to compare like clusters between the sample conditions to identify any “winning” clusters that were common between the parallel selections, so the Barcode###sta\_rev\_primer\_truncated.fna files for all of those rounds were concatenated using the Unix cat command prior to the fastx scripts described above. For both analyses, the top 500 or >5 metric was chosen to identify the most abundant clusters while reducing the computational resources involved in the clustering. The filtered .fasta files were generated using the Unix head command as follows:

```
head -n 1000 Trim_Barcode###/sorted.fna > Trim_Barcode###/sorted_top500.fna
head -n 1734112 Selection3_sorted.fna > Selection3_count5_sorted.fna
```

Head -n 1000 takes the top 1000 lines of a document which are then piped (>) into the sorted\_top500.fna file. An -n of 1000 is used because each sequence takes up 2 lines in the fasta file format. The 1734112 value for selection 3 was determined by using the grep command combined with the tail command to find the line value of the last sequence read appearing more than 5 times.

```
grep -e '-5' Selection3_sorted.fna > grep5.fna
tail -n 2 grep5.fna
```

The rank sorted value of the last sequence with 5 or more reads (rank 867,056) and is the last 2 lines of the grep5.fna file was then multiplied by 2 to get the line count for the count 5 fasta file.

Using these files, the sequences were then clustered using the QIIME 1.9 uclust clustering implementation in the pick\_otus.py script.

```
#Selection 1 and 2 (MS2)
pick_otus.py -s 0.45 -BD --word_length 4 --stepwords 4 -i top500s_sort.fna -o
otu0.45/ --threads 8

#Selection 3 (CypE, 6 conditions)
pick_otus.py -s 0.8 -BD --word_length 4 --stepwords 4 -i
Selection3_count5_sorted.fna -o count5_otus/ --threads 8
```

The -s indicates the similarity (45%/80% respectively), -i is the input fasta file, -o is the output directory, --word\_length is the size of the string compared in each pairwise alignment, and --stepwords is approximately how many words are expected to be in the target sequence. Higher values of for word\_length and stepwords increase the speed of the computation but reduce accuracy. I did half the default values of both, but I am uncertain what the actual effect of those are on the ultimate accuracy of the clustering (i.e. which of the new clusters should actually be part of a previous cluster). The clustering results were then formatted into a .biom format table and filtered for cluster abundance.

```
#MS2
make_otu_table.py -i otu0.45/top500s_sort_otus.txt -o otu0.45/top500.biom
#Selection 3 CypE
make_otu_table.py -i count5_otus/count5_otus.txt -o count5_otus/count5.biom

#MS2
filter_otus_from_otu_table.py -i otu0.45/top500.biom -o
otu0.45/top500_filter005.biom --min_count_fraction 0.05
#Selection 3 CypE
filter_otus_from_otu_table.py -i count5_otus/count5.biom -o
count5_otus/count5_filter005.biom --min_count_fraction 0.05
```

The reformatting makes subsequent analysis more straightforward and allows for export into a .tsv format that is more interpretable than the list of otus (clusters) and their associated sequences. The filter\_otus\_from\_otu\_table.py filters the biom table for a specified value, in this

case for a minimum count fraction of 5%. From there, the following script was used to pull out the clusters remaining in the filtered .biom table and the associated sequence ids. This script is not part of the default QIIME 1.9 distribution, so I have provided the full contents of the script at the end of the appendix in case the original download is no longer available. It also needs to be executed with the python command followed by the path to the script file location.

```
python ~/filter_otu_mapping_from_otu_table.py -i
otu0.45/top500_filter005.biom -o otu0.45/filtered_otu.txt -m
otu0.45/top500s_sort_otus.txt
```

The -i value is the filtered biom table, the -o is the output file, and -m is the original clustering file. Using this new clustering file, I then used pick\_rep\_set.py to get a representative sequence for each cluster (in this case, the most abundant/seed sequence for each cluster).

```
pick_rep_set.py -i top500s_sort_otus.txt -f top500s_sorted.fna -o
top500_otu_rep_set.fna
```

This produces a reference fasta file that can then be used to cluster all sequences of a particular selection (using a concatenated trim file for all samples in a selection prior to the fastx\_collapse command), which would then allow for tracking of the cluster abundance as a function of round.

```
pick_otus.py -m uclust_ref -r top500_otu_rep_set001.fna -s 0.45 -C --
word_length 4 --stepwords 4 -i combined_sta_rev_primer_truncated.fna -o
rep_set/ --threads 8
```

```
make_otu_table.py -i rep_set_otus.txt -m barcode_map_complete -o
output_directory/rep_otu.biom
```

```
make_otu_heatmap.py -i otu.biom -o heatmap.pdf -m barcode_map_complete -
absolute_abundance
```

```
biom convert -i table.biom -o table.from_biom.txt --to-tsv
```



## Scripts for Prepping QIIME Processed Data for AptaSUITE

To add in the 5' and 3' constant regions to the demultiplexed and trimmed fasta files, I used the unix sed command. The first script below adds the 5' constant regions (GAGA...) to the beginning of every other line starting with the second line (every other to skip over the >identifying line for each sequence). This script then puts this output into a new file. The second script then edits that file, adding the 3' constant region (CAGC...) to the end of every other line starting with the second line.

```
sed '2~2s/^/GAGACAAGAATAAACGCTCAAGG/' demultiplexed.fasta >
demultiplexed_constant.fasta
sed -i '2~2s/$/CAGCCACACCACCAGCC/' demultiplexed_constant.fasta
```

The following is the text of the fasta\_to\_fastq.pl script used to convert the demultiplexed fasta files with the constant regions added in to Fastq files as required by AptaSUITE (this will be unnecessary once support for fasta is added to AptaSUITE). The script puts in perfect dummy quality data so the sequences so that AptaSUITE does not reject any of the sequences for quality scores.

```
#Copyright (c) 2010 LUQMAN HAKIM BIN ABDUL HADI (csilhah@nus.edu.sg)
#
#Permission is hereby granted, free of charge, to any person obtaining
a copy of this software and associated documentation files
#(the "Software"), to deal in the Software without restriction,
including without limitation the rights to use, copy, modify,
#merge, publish, distribute, sublicense, and/or sell copies of the
Software, and to permit persons to whom the Software is
#furnished to do so, subject to the following conditions:
#
#The above copyright notice and this permission notice shall be
included in all copies or substantial portions of the Software.
#
#THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND,
EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES
#OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND
NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE
#LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN
ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR
```

```
#IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE
SOFTWARE.
```

```
#!/usr/bin/perl
use strict;

my $file = $ARGV[0];
open FILE, $file;

my ($header, $sequence, $sequence_length, $sequence_quality);
while(<FILE>) {
    chomp $_;
    if ($_ =~ /^>(.)+/) {
        if($header ne "") {
            print "\@".$header."\n";
            print $sequence."\n";
            print "+".$sequence_length."\n";
            print $sequence_quality."\n";
        }
        $header = $_;
        $sequence = "";
        $sequence_length = "";
        $sequence_quality = "";
    }
    else {
        $sequence .= $_;
        $sequence_length = length($sequence);
        for(my $i=0; $i<$sequence_length; $i++) {$sequence_quality
.= "I"}
    }
}
close FILE;
print "\@".$header."\n";
print $sequence."\n";
print "+".$sequence_length."\n";
print $sequence_quality."\n";
```

## AptaSUITE Pipeline Scripts

```
java -Xmx30G -jar AptaSUITE-0.8.8/AptaSUITE-0.8.8.jar -parse -cluster -
predict structure -trace -config config_files/config_l100
```

The Graphical Interface of AptaSUITE can be opened in terminal using the following command. The -Xmx30G refers to the maximum amount of ram allocated to the java program, in this case 30 Gb. Windows users may need to add the -d64 option to run the script using more than 2 Gb of ram.

```
java -Xmx30G -jar AptasSUITE-0.8.8/AptasSUITE-0.8.8.jar
```

### Example AptasSUITE config file

```
# Experiment configuration
Experiment.name = "SELEX against target l100 Cyclophilin E"
Experiment.description = "16 rounds of selection including the initial pool"

Experiment.primer5 = GAGACAAGAATAAACGCTCAAGG
# OPTIONAL, only specify if the 3' primer was part of the sequenced data.
# If not specified, we need to specify the randomized region size
Experiment.primer3 = CAGCCACACCACCAGCC
# Experiment.randomizedRegionSize = 50

### Selection Cycle Information ###
SelectionCycle.name = Round0
SelectionCycle.name = Round1
SelectionCycle.name = Round2
SelectionCycle.name = Round3
SelectionCycle.name = Round4
SelectionCycle.name = Round5
SelectionCycle.name = Round6
SelectionCycle.name = Round7
SelectionCycle.name = Round8
SelectionCycle.name = Round9
SelectionCycle.name = Round10
SelectionCycle.name = Round11
SelectionCycle.name = Round12
SelectionCycle.name = Round13
SelectionCycle.name = Round14
SelectionCycle.name = Round15
SelectionCycle.round = 0
SelectionCycle.round = 1
SelectionCycle.round = 2
SelectionCycle.round = 3
SelectionCycle.round = 4
SelectionCycle.round = 5
SelectionCycle.round = 6
SelectionCycle.round = 7
SelectionCycle.round = 8
SelectionCycle.round = 9
SelectionCycle.round = 10
SelectionCycle.round = 11
SelectionCycle.round = 12
SelectionCycle.round = 13
SelectionCycle.round = 14
SelectionCycle.round = 15
SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
```

```

SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
SelectionCycle.isControlSelection = False
SelectionCycle.isCounterSelection = False
SelectionCycle.isCounterSelection = False
SelectionCycle.isCounterSelection = False
SelectionCycle.isCounterSelection = False
SelectionCycle.isCounterSelection = False
SelectionCycle.isCounterSelection = False
SelectionCycle.isCounterSelection = False
SelectionCycle.isCounterSelection = False
SelectionCycle.isCounterSelection = False
SelectionCycle.isCounterSelection = False
SelectionCycle.isCounterSelection = False
SelectionCycle.isCounterSelection = False
SelectionCycle.isCounterSelection = False
SelectionCycle.isCounterSelection = False
SelectionCycle.isCounterSelection = False

# If the data has previously been de-multiplexed using a third party tool and
# is
# present as one file per selection cycle, set this value to true. The
# default is false.
AptaplexParser.isPerFile = True

# An equal number of files as there are selection cycles must be specified
# and
# in the same order
AptaplexParser.forwardFiles = /home/neil/apta/seqs/Rd0_test.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.1_5p.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.2_5p.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.3_5p.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.4_5p.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.5_5p.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.6_5p.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.7_5p.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.8_5p.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.9_5p.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.10_5p.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.11_5p.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.12_5p.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.13_5p.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.14_5p.fastq
AptaplexParser.forwardFiles = /home/neil/apta/seqs/l100.15_5p.fastq

# One or more input files for the forward reads. If the data was is not
# paired-end,
# specify the single-end data here.
#AptaplexParser.forwardFiles = path/to/forward/reads1.fastq
#AptaplexParser.forwardFiles = path/to/forward/reads2.fastq
#AptaplexParser.forwardFiles = path/to/forward/readsN.fastq

```

```

# One or more input files for the reverse reads. The number and order of the
files
# must coincide with the forwardFiles.
#AptaplexParser.reverseFiles= path/to/reverse/reads1.fastq
#AptaplexParser.reverseFiles= path/to/reverse/reads2.fastq
#AptaplexParser.reverseFiles = path/to/reverse/readsN.fastq

# The five prime barcodes. Must be comma separated and in the same order
# as SelectionCycles
#AptaplexParser.barcodes5Prime = ATGCGT, GACGAC, GGTACC, TCGTAG, CCATGG

# OPTIONAL (specify only if present in the sequencing data), the three
# prime barcodes. Must be in order of SelectionCycles and in 5' to 3' of
# the Forward Read.
#AptaplexParser.barcodes3Prime = TAGCCA, ATCGAT, AATCAA, ATCGTA, GGTTAA

# For paired-end data only. The smallest overlap required between the forward
and
# reverse read when creating a single contig out of the two.
#AptaplexParser.PairedEndMinOverlap = 15

# Maximal number of mutations in the overlapping region for a sequence to be
accepted
#AptaplexParser.PairedEndMaxMutations = 5

# Highest score of the current quality. 55 for phred model.
#AptaplexParser.PairedEndMaxScoreValue = 55

# Maximal number of mutations allowed in the barcodes
#AptaplexParser.BarcodeTolerance = 1

# Maximal number of mutations allowed in the primer regions
AptaplexParser.PrimerTolerance = 3

# If DNA aptamers were used during the selection, it is likely that they were
sequenced in reverse complement order
# By setting this option to true, Aptaplex will automatically convert the
cDNA back into DNA
AptaplexParser.StoreReverseComplement = False

# Specifies the reader for the sequences depending on the input format (case
sensitive).
# Current options are: FastqReader, RawReader
AptaplexParser.reader = FastqReader

# The default back-end for storing aptamer sequence information
AptamerPool.backend = MapDBAptamerPool

#AptapLEX processes the reads in parallel using a producer-consumer model.
The size of the queue containing the items to be processed can be controlled
with
#AptaplexParser.BlockingQueueSize = 500

### APTACLUSTER OPTIONS ###
# Length of the randomized region in the aptamers
Aptacluster.RandomizedRegionSize = 50

```

```

# The number of LSH iterations to be performed
Aptacluster.LSHIterations = 5

# The kmer size used for the distance calculations
Aptacluster.KmerSize = 3

### APTASIM OPTIONS ###
# Fastq file containing training sequences
# Aptasim.HmmFile = /path/to/training/data.fastq.gz

# Degree of the Markov model
Aptasim.HmmDegree = 2

# Length of the randomized region in the aptamers
Aptasim.RandomizedRegionSize = 50

# Number of (unique) sequences in the initial pool
Aptasim.NumberOfSequences = 500000

# Number of high affinity sequences in the initial pool
Aptasim.NumberOfSeeds = 100

#The minimal affinity for seed sequences (INT range: 0-100)
Aptasim.MinSeedAffinity = 80

# Maximal count of remaining sequences
Aptasim.MaxSequenceCount = 10

# The maximal sequence affinity for non-seeds (INT range: 0-100)
Aptasim.MaxSequenceAffinity = 25

# If no training data is specified, create pool based on this distribution
(order A,C,G,T)
Aptasim.NucleotideDistribution = 0.25,0.25,0.25,0.25

# The percentage of sequences that remain after selection (DOUBLE range: 0-1)
Aptasim.SelectionPercentage = 0.20

# Mutation rates for individual nucleotides (order A,C,G,T)
Aptasim.BaseMutationRates = 0.25,0.25,0.25,0.25

# Mutation probability during PCR (DOUBLE range: 0-1)
Aptasim.MutationProbability = 0.05

# PCR amplification efficiency (DOUBLE range: 0-1)
Aptasim.AmplificationEfficiency = 0.995

### APTATRACE OTIONS ###
# Defines the size of the k-mers that will be used during the motif
# extraction procedure of AptaTRACE. In other words, it defines the initial
motif
# size
AptaTRACE.KmerLength = 6

```

```

# Occasionally, motifs might co-occur within the same aptamer or aptamer
family.
# In order to better understand this relationship, we have developed a post-
processing
# add-on that uncovers these relationships. To activate this option, the this
parameter
# has to be set to True.
AptaTRACE.FilterClusters = True

# If, in addition to the motifs, a list of all aptamers that contain the
motif are to
# be saved in a separate file, set this parameter to true.
AptaTRACE.OutputClusters = True

# AptaTRACE uses a background model to identify statistically significant
changes
# in secondary structure contexts. This model is generated from aptamers
which do
# not undergo selection and are therefore present in small numbers in the
pools.
# The parameter alpha specifies which sequences should be included in the
background
# model, i.e. all sequences whose number of occurrences is smaller than, or
equal
# to this value are taken into account.
AptaTRACE.Alpha = 10

# The default back-end for storing the counts of each aptamer in a
# particular selection cycle
SelectionCycle.backend = MapDBSelectionCycle

# The default back-end storing the secondary structure information
StructurePool.backend = MapDBStructurePool

# In order to avoid time-consuming disk I/O, a bloom filter is used to store
the information
# whether an aptamer is present in the pool or not. This value should be at
least a large as
# the total number of reads that were sequenced.
MapDBAptamerPool.bloomFilterCapacity = 500000000

# The corresponding collision probability of the bloom filter. The smaller
this value the
# more memory this data structure will consume.
MapDBAptamerPool.bloomFilterCollisionProbability = 0.001

# To prevent the creation of large files on disk which would yield slower
lookup times,
# the pool is partitioned into smaller units with the capacity as specified
below.
MapDBAptamerPool.maxTreeMapCapacity = 1000000

# The corresponding collision probability of the bloom filter used for the
selection
# cycle implementation. The capacity has the same value as
MapDBAptamerPool.bloomFilterCapacity
MapDBSelectionCycle.bloomFilterCollisionProbability = 0.001

```

```

# The corresponding collision probability of the bloom filter used for the
structure
# information. The capacity has the same value as
MapDBAptamerPool.bloomFilterCapacity
MapDBStructurePool.bloomFilterCollisionProbability = 0.001

# To prevent the creation of large files on disk which would yield slower
lookup times,
# the structure information is partitioned into smaller units with the
capacity as specified below.
MapDBStructurePool.maxTreeMapCapacity = 500000
Aptacluster.LSHDimension = 37
Experiment.projectPath = /home/neil/apta/0.8/l100
Performance.maxNumberOfCores = 8

```

### QIIME batey\_mapping\_file/Batey Barcodes (Reverse Complement of primer sequence)

#SampleID	BarcodeSequence	LinkerPrimerSequence	ReversePrimer	Treatment
DOB Description				
Barcode001	GGAGACAAGGGA	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode002	AATCAGTCTCGT	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode003	AATCCGTACAGC	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode004	ACACCTGGTGAT	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode005	TATCGTTGACCA	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode006	TTACTGTGCGAT	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode007	AGGCTACACGAC	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode008	CTAACCTCCGCT	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode009	GAACCAAAGGAT	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode010	GTATGCGCTGTA	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode011	GTACATACCGGT	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode012	TCCGACACAATT	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode013	CCAGTGTATGCA	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode014	CCTCGTTCGACT	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode015	TGAGTCACTGGT	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode016	GACTTGGTATTC	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			
Barcode017	TACACGATCTAC	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT	
	GCTGGTGGTGTGGCTG			



Barcode018	GCACACACGTTA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode019	CACGCCATAATG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode020	CAGGCGTATTGG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode021	GGATCGCAGATC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode022	GCTGATGAGCTG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode023	AGCTGTTGTTTG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode024	GGATGGTGTTGC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode025	GCGATATATCGC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode026	TAGGATTGCTCG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode027	ATGTGCACGACT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode028	ACGCGCAGATAC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode029	GACTTTCCTCG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode030	ATCCCGAATTTG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode031	GTTGGTCAATCT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode032	TAGCTGTAACT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode033	CAGTGCATATGC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode034	TCACGGGAGTTG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode035	CTGCTAACGCAA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode036	TTAGGGCTCGTA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode037	TCTAGCGTAGTG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode038	TCGAGGACTGCA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode039	CGGAGCTATGGT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode040	AAGAGATGTCGA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode041	TCCAAGTGTTT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode042	TACAGATGGCTC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode043	ACGTGTACCCAA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode044	AAGGAGCGCCTT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode045	CGATCCGTATTA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT

Barcode046	GTCTAATTCCGA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode047	TCCGAATTCACA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode048	ACGCCACGAATG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode049	GGCCACGTAGTA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode050	TAGGAACTGGCC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode051	CTAGCGAACATC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode052	GACAGGAGATAG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode053	ATTCTGTGAGT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode054	GAGGCTCATCAT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode055	TCCTCTGTCGAC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode056	CTATTTGCGACA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode057	AGTAGAGGGATG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode058	CGCAGCGGTATA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode059	AATGCCTCAACT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode060	GGTGTCTATTGT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode061	GTCAATTGACCG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode062	ATGAGACTCCAC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode063	GAATCTTCGAGC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode064	ACACGTAAGCCT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode065	GAGTGGTAGAGA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode066	GAAGTTGGAAGT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode067	TTCCTAGGTGAG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode068	GCACGACAACAC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode069	ATCGATCTGTGG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode070	CTTGTGTCGATA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode071	TGAGCCGGAATC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode072	GCGGCAATTACG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode073	GAACTAGTCACC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT

Barcode074	GACGGAACCCAT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode075	CAAGCATGCCTA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode076	CCTGAACTAGTT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode077	CTTCGGCAGAAT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode078	ACGGGACATGCT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode079	GTCATATCGTAC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode080	GGAAACCACCAC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode081	TTGCGCATACTA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode082	ACATTCAGCGCA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode083	ACTGACAGCCAT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode084	CGAGAAGAGAAC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode085	AGGCATCTTACG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode086	CAGCTAGAACGC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode087	TCCCAGAACAAC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode088	AGCTGGAAGTCC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode089	CACGGTTGTGAG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode090	GAGGAATAGCAG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode091	CAGCGGTGACAT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode092	ATCGGCGTTACA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode093	AGATGTTCTGCT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode094	CCACCTACTCCA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode095	GAATAGAGCCAA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode096	GTACGTGGGATC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode097	GAAGAAGCGGTA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode098	TGTTATCGCACA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode099	TCGTCGATAATC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode100	ATTGGGCTAGGC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode101	ACCACATACATC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT

Barcode102 AACACAAGGAGT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode103 AATGTCCGTGAC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode104 TACTTCGCTCGC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode105 GCTTCGGTAGAT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode106 CTTACACCAAGT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode107 TGACCTCCAAGA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode108 ACAAGGAGGTGA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode109 TATCAGGTGTGC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode110 TGTAATTGTGCG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode111 AATGGAGCATGA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode112 AGCTTGACAGCT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode113 TCTGTTGCTCTC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode114 AGTTCCCGAGTA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode115 AGCCTAAGCACG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode116 ATACCTTCGGTA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode117 GAATGATGAGTG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode118 CGTCCGAAATAC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode119 GCAGGATAGATA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode120 GACTCTTGCAA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode121 TCTTCCGCTACT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode122 GTACCTAATTGC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode123 ACTCACGGTATG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode124 GTCTACACACAT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode125 ATACTTCGCAGG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode126 ATGTCGAGAGAA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode127 TCTACGGAGAGC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode128 GGTCAGCTTAAC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode129 ACGGCATGGCAT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCCGGGCTTCGGTCCGGTTCGGCTGACTGACT

Barcode130	CGTGACAATGTC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode131	ATGGTTGTTGGC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode132	CCTAGTACTGAT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode133	ATCGCTCGAGGA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode134	TAACGCTTGGGT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode135	AATCTTGCTGCA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode136	TGCAATGTTGCT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode137	TAACACCACATC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode138	GACACATTTCTG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode139	CTCTACCTCTAC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode140	TAGCGGATCACG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode141	CGCCAAATAACC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode142	GTATTACGATCC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode143	TTGATGCTATGC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode144	CACATCTAACAC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode145	GCATGGCTCTAA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode146	CCATAGGGTTCA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode147	TGGCAAGACTCT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode148	TCGGAGTGTGTTG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode149	TCAACAGCATCG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode150	TTATGCAGTCGT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode151	ATTAGTTCGCGT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode152	CCATACATAGCT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode153	ATGATGACCCGT GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode154	GTGGGATGTTTC GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode155	CTCGAGAGTACG GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode156	AACGAGAACTGA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT
Barcode157	CAACACGCACGA GCTGGTGGTGTGGCTG	CAGCCACACCACCAGCCC GGGCTTCGGTCCGGTTCGGCTGACTGACT

Barcode158	CCATGCGATAAC	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode159	CCTCTCGTGATC	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode160	GCCTGAATTTAC	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode161	GTCCGAAACT	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode162	TAAACCGCGTGT	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode163	CTAGATTTGCCA	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode164	TAAGGTAAGGTG	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode165	CAGGAAGGTTAA	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode166	TGGCATAACGCA	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode167	ACTATTGTACG	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode168	CGAGTTGTAGCG	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode169	CGACTGTCTTAA	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode170	GCTCAGTGCAGA	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode171	TACTAATCTGCG	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode172	ATGTGGGACCCA	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode173	TATGCACCAAGTG	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode174	AGAGCCTACGTT	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode175	CGGACTACAAC	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode176	CGGGTTTGACGA	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode177	TGGCACCGATTA	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode178	CTACCGGATCAA	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode179	AGCAAACACCCG	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode180	AACCGCGGTCAA	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		
Barcode181	GATTGGTTGCAC	CAGCCACACCACCAGCCC	GGGCTTCGGTCCGGTTCGGCTGACTGACT
	GCTGGTGGTGTGGCTG		

### filter\_otu\_mapping\_from\_otu\_table.py

```
#!/usr/bin/env python

__author__ = "William Walters"
```

```

__copyright__ = "Copyright 2011"
__credits__ = ["William Walters"]
__license__ = "GPL"
__version__ = "1.0"
__maintainer__ = "William Walters"
__email__ = "William.A.Walters@colorado.edu"

from biom import load_table

from qiime.util import parse_command_line_parameters, get_options_lookup,\
    make_option, create_dir, qiime_open

options_lookup = get_options_lookup()
script_info={}
script_info['brief_description']="""Finds the OTU IDs in a supplied OTU
table, filters all IDs not matching these
in the supplied OTU mapping file to create a filtered OTU mapping file as
output. The purpose of this would be to
backtrack to unclustered read data but have all reads removed that were
filtered along the way."""
script_info['script_description']=""""""

script_info['output_description']="""A filtered OTU mapping file (can be used
with -m input with filter_fasta.py)"""
script_info['required_options']= [\
    make_option('-i', '--otu_table',type='existing_filepath',
                help='OTU table (biom) filepath'),
    make_option('-m', '--otu_mapping',type='existing_filepath',
                help='OTU mapping file, tab-separated lines of OTU
ID<tab>seq1<tab>seq2...'),
    make_option('-o', '--output_mapping',
                help='output filtered OTU mapping file. WILL OVERWRITE IF
FILE ALREADY EXISTS.')
]
script_info['optional_options']= []

script_info['version'] = __version__

def main():
    option_parser, opts, args =\
        parse_command_line_parameters(suppress_verbose=True, **script_info)

    output_mapping_f = open(opts.output_mapping, "w")

    # Get OTU table OTU IDs (should be taxa strings in this case)
    otu_table_data = load_table(opts.otu_table)

    obs_ids = set(otu_table_data._observation_ids)

    for line in open(opts.otu_mapping, "U"):
        curr_id = line.split('\t')[0]
        if curr_id in obs_ids:
            output_mapping_f.write("%s" % line)

```

```
if __name__ == "__main__":  
    main()
```