SHAPE-SEQ TECHNOLOGIES FOR ANALYZING, UNDERSTANDING, AND DESIGNING RNA STRUCTURES AND FUNCTIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Kyle Edward Watters

August 2016

SHAPE-SEQ TECHNOLOGIES FOR ANALYZING, UNDERSTANDING, AND DESIGNING RNA STRUCTURES AND FUNCTIONS

Kyle Edward Watters, Ph.D.

Cornell University 2016

RNA is one of the most versatile and abundant biological molecules in nature. It maintains roles in a vast array of functions critical to life that include, among others, encoding protein synthesis, catalyzing chemical reactions, storing genetic information, defending the cell, and regulating gene expression.

A critical property of RNA is its ability to fold into intricate three-dimensional structures that determine its function. These three-dimensional structures are directly derived from the sequence of the RNA and are generated by internal interactions between the nucleotides. Determining what these structures are and how they form is key to understanding the many functions RNAs play in cellular life.

One technique used to analyze RNA structure is SHAPE-Seq, which couples local nucleotide flexibility measurements with next-generation sequencing to obtain high-resolution information about RNA structural elements. In this work, I greatly expand the capabilities of the SHAPE-Seq technique both *in vitro* and inside living cells and describe a novel technique for capturing RNA folding during the process of transcription.

Using the capability to measure RNA structures inside living cells, I study the role of RNA structure in regulating both transcription and translation. In studying translational regulation, I link changes in RNA structural state to a functional output. I also discuss the utility of in-cell RNA structural data to guide the design of a transcriptional derepressor using CRISPR interference.

I then study how the process of transcription affects RNA folding using a newly developed cotranscriptional SHAPE-Seq technique. With it, I determine the folding pathways of two regulatory RNAs *in vitro*, and gain new insights into how transcriptional regulators make structural decisions in the timescale of transcription.

I close with investigations into the structural changes that occur during RNA-protein binding in the contexts of two different RNase Ps and the cucumber mosaic virus. In both, I observe protein binding at distinct structural motifs.

The improvements and extensions to SHAPE-Seq described herein represent new frontiers in RNA structural biology and promise to change the way RNA biologists and engineers study and design this fundamental molecule of life.

BIOGRAPHICAL SKETCH

Kyle E. Watters was born in Providence, Rhode Island to Keith and Diane Watters. He attended high school at Greenwich Central in Greenwich, NY where he was deeply interested in chemistry and biology. Kyle got his first taste of biological research in high school, in a program offered through the University at Albany, investigating the effect of magnetic fields on the growth of *Zea mays*.

Kyle obtained a B.S. in chemical engineering at Rensselaer Polytechnic Institute (RPI) in Troy, NY in 2011 with minors in biology and economics. His initial research interests were in the field of materials science and chip manufacturing, prompting him to accept a job in RPI's nanofabrication facility and a summer internship at what is now SUNY Polytechnic in Albany, NY. However, Kyle's interests later shifted from materials science to biochemistry. Diving into the world of biology, Kyle spent much of his free time working in the labs of Johnathan Dordick and Christopher Bystroff engineering GFP and studying psychrophilic enzymes under the guidance of Ivonne Paredes. He spent his summers interning at Wyeth Pharmaceuticals (now Pfizer) in Rouses Point, NY, Genentech in South San Francisco, CA, and Mastermelt America in Sweetwater, TN. He entered into the Ph.D. program at Cornell University in 2011 with long-term entrepreneurial plans and interests in virology.

Kyle then joined the Lucks Lab in the fall, and the rest is history, as documented below. Next, he will be heading to UC Berkeley as a postdoctoral associate in the lab of Jennifer Doudna to study CRISPR systems.

This work is dedicated to Edward Watters, to whom I am forever indebted. May your curiosity, ambition, and pedagogy live on forever.

ACKNOWLEDGEMENTS

To my wife Nicole, thank you for everything, my life only gets better every day we're together. It was you that brought me to Cornell and changed my life. I'll remember grad school not just for science, but also for our milestones: our engangement, our wedding, getting our dog Jack. None of this would have been worth it without you.

To Julius, you have been the supportive and caring mentor that everyone wishes they had. I owe my entire future career to you, and only wish you the absolute best in your new home at Northwestern. You have my utmost gratitude.

To Matt, Susan, and John, thank you for your guidance and support as my committee members. Your insights and suggestions have all made me a better scientist.

To the members of the Lucks Lab, thank you for being the best labmates I could ask for. Great people and excellent scientists; I wish you all the very best.

To Tim, Alex, Jane, and Ray, thank you for all of your hard work as undergrads working with me. This dissertation is as much your work as it is mine. You are all exceptionally talented and will do great things in your futures.

To my friends in Olin, thank you for making the last five years so enjoyable. I hope we all stay in touch and see each other again in the future, some of my best friends were made here.

To my collaborators, Venkat, Jeremy, and Keith, thank you all for working with me over these last few years. Working with all of you, I have learned so much more than I

ever would have expected.

Finally, to my family, thank you for all of your kindness and support. Mom and Dad, all of my life's accomplishments are due to the two of you. Your constant encouragement got me to where I am today and I am eternally grateful. Babcia and Dzaidziu, I will never forget all that you did for me. I would have never made it through undergrad without you.

Funding: This work was supported by the National Science Foundation Graduate Research Fellowship Program (Grant No. DGE-1144153). Kyle is a Fleming Scholar in the Robert Frederick Smith School of Chemical and Biomolecular Engineering at Cornell University.

TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Table of Contents . . . vii
List of Tables . . . xv
List of Figures . . . xvi

1 Introduction **1**
1.1 The role of RNA structure in biology . . . 1
1.2 RNA structures that regulate translation in prokaryotes . . . 4
1.3 RNA structures that regulate transcription in prokaryotes . . . 8
1.4 Cotranscriptional RNA folding and its role in structure formation . . . 10
1.5 Understanding RNA function by measuring RNA structure . . . 11
1.6 Developing new SHAPE-Seq technologies for new contexts . . . 19
1.6.1 Chapter 2: Developing SHAPE-Seq v2.0 . . . 19
1.6.2 Chapter 3: Characterizing Cellular RNA Structure and Function with in-cell SHAPE-Seq . . . 20
1.6.3 Chapter 4: Further improvement of *in vitro* SHAPE-Seq . . . 23
1.6.4 Chapter 5: Cotranscriptional folding of a riboswitch at nucleotide resolution . . . 24
1.6.5 Chapter 6: Measuring the cotranscriptional folding pathway of the pT181 transcriptional attenuator . . . 25
1.6.6 Chapter 7: Design of a CRISPR sgRNA Derepressor Using Insight from SHAPE-Seq . . . 27
1.6.7 Chapter 8: Structural Analysis of Cucumber Mosaic Virus RNA3 29
1.6.8 Chapter 9: Structural Features of Protein Binding on RNase P and its Substrates . . . 30

2 SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing **31**
2.1 Abstract . . . 31
2.2 Introduction . . . 32
2.3 Materials and Methods . . . 37
2.3.1 RNA Preparation . . . 37
2.3.2 RNA Modification . . . 38
2.3.3 QuSHAPE . . . 39
2.3.4 SHAPE-Seq Reverse Transcription . . . 40
2.3.5 SHAPE-Seq Second Adapter Ligation . . . 41
2.3.6 SHAPE-Seq PCR, Library QC, and Sequencing . . . 42
2.3.7 SHAPE-Seq Data Analysis . . . 43
2.3.8 Computational Modeling . . . 44

2.4 Results . . . 45
2.4.1 Investigation and Optimization of SHAPE-Seq Library Preparation 45
2.4.2 Adapter Ligation . . . 45
2.4.3 PCR Amplification . . . 49
2.4.4 Assessing the Reproducibility of SHAPE-Seq . . . 51
2.4.5 SHAPE-Seq v2.0: Removing RNA Sequence Requirements with Universal RT Priming . . . 52
2.4.6 3′ Ligation Methods and Other Protocol Adjustments . . . 53
2.4.7 Comparison of SHAPE-Seq v2.0 and v1.0 . . . 54
2.4.8 Using SHAPE-Seq Reactivities as Constraints in Thermodynamic RNA Folding Algorithms . . . 56
2.5 Discussion . . . 58
2.6 Acknowledgements . . . 61
2.7 Funding . . . 61

3 Simultaneous Characterization of Cellular RNA Structure and Function with in-cell SHAPE-Seq 62
3.1 Abstract . . . 62
3.2 Introduction . . . 63
3.3 Materials and Methods . . . 66
3.3.1 Platform (plasmid) construction . . . 67
3.3.2 Strains, growth media, and RNA expression . . . 67
3.3.3 *in vitro* RNA purification . . . 68
3.3.4 RNA modification and fluorescence assay . . . 68
3.3.5 RNA extraction . . . 70
3.3.6 Reverse transcription . . . 70
3.3.7 Adapter ligation . . . 71
3.3.8 Quality control . . . 71
3.3.9 dsDNA sequencing library construction . . . 72
3.3.10 Next-generation sequencing . . . 72
3.3.11 Data analysis . . . 73
3.3.12 Structure folding predictions . . . 73
3.4 Results . . . 74
3.4.1 A standardized platform for characterizing RNA structures, interactions, and regulatory function in cells . . . 74
3.4.2 Characterizing cellular RNA structures of synthetic riboregulators that activate translation . . . 76
3.4.3 Characterizing the cellular RNA interactions and function of synthetic riboregulators that activate translation . . . 80
3.4.4 Quantitatively linking ribosome binding site reactivity with gene expression . . . 84
3.4.5 Characterizing the cellular RNA structures of the RNA-IN/OUT translational repressor . . . 84

3.4.6 Characterizing the cellular RNA interactions and function of the IS10 translational repressor . . . 88
3.4.7 Targeting Endogenous RNAs in *E. coli* . . . 90
3.4.8 Comparing *in vitro* and in-cell SHAPE-Seq reactivities . . . 95
3.4.9 Discussion . . . 96
3.5 Acknowledgements . . . 101
3.6 Funding . . . 101

4 Characterizing RNA structures *in vitro* and *in vivo* with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq) 103
4.1 Abstract . . . 103
4.2 Introduction . . . 104
4.3 SHAPE-Seq Background . . . 109
4.3.1 RNA modification . . . 109
4.3.2 Conversion of RNA to cDNA with reverse transcription . . . 110
4.3.3 Preparation for sequencing . . . 111
4.3.4 Bioinformatic read alignment and reactivity calculation . . . 112
4.3.5 RNA structure prediction using SHAPE-Seq reactivities . . . 112
4.4 Materials and Methods . . . 116
4.4.1 RNA folding and modification . . . 116
4.4.2 RNA linker ligation (skip for *in vivo* or direct priming experiments) 118
4.4.3 Reverse transcription . . . 119
4.4.4 Sequencing adapter ligation . . . 120
4.4.5 Quality analysis . . . 121
4.4.6 Library preparation for sequencing . . . 122
4.4.7 Illumina sequencing . . . 123
4.4.8 Data analysis with Spats . . . 124
4.4.9 SHAPE-directed computational RNA folding . . . 125
4.5 Results . . . 130
4.5.1 *in vitro* SHAPE-Seq analysis . . . 130
4.5.2 SHAPE-Seq v2.0 vs. v2.1 . . . 130
4.5.3 Using SHAPE-Seq v2.1 to observe ligand binding . . . 133
4.5.4 Inferring secondary structures with SHAPE-Seq data . . . 135
4.5.5 in-cell SHAPE-Seq analysis . . . 138
4.5.6 5S rRNA, expressed endogenously . . . 138
4.5.7 TPP riboswitch, expressed from a plasmid . . . 142
4.6 Experimental Considerations . . . 143
4.6.1 Effect of increasing PCR cycles . . . 143
4.6.2 RT primer length and library multiplexing . . . 146
4.6.3 Choosing SHAPE reagents and other chemical probes . . . 147
4.6.4 Factors influencing data quality and consistency . . . 150
4.6.5 Choosing an adapter trimming algorithm . . . 153
4.6.6 Measures of SHAPE reactivity . . . 153

4.7 Further potential improvements for SHAPE-Seq and restrained RNA folding . . . 154
4.7.1 Going transcriptome-wide . . . 154
4.7.2 Future directions for computational folding . . . 155
4.8 Conclusions . . . 156
4.9 Acknowledgements . . . 157

5 Cotranscriptional Folding of a Fluoride Riboswitch at Nucleotide Resolution 158
5.1 Abstract . . . 158
5.2 Introduction . . . 158
5.3 Results . . . 159
5.4 Conclusion . . . 172
5.5 Acknowledgments . . . 172

6 Measuring the cotranscriptional folding pathway of the pT181 transcriptional attenuator 174
6.1 Abstract . . . 174
6.2 Introduction . . . 175
6.3 Results and Discussion . . . 179
6.3.1 Cotranscriptional folding of the pT181 attenuator . . . 179
6.3.2 The pT181 attenuator refolds into an antiterminator late in transcription . . . 182
6.3.3 The pT181 antisense binds quickly and changes the attenuator's folding pathway . . . 185
6.3.4 Cotranscriptional vs. in-cell . . . 187
6.3.5 Cotranscriptional vs. equilibrium folding . . . 188
6.3.6 Deletion analysis and sequential minimization of the attenuator . 190
6.3.7 A minimized pT181 attenuator . . . 195
6.4 Conclusion . . . 196
6.5 Future work . . . 197
6.5.1 Determining the key motif . . . 197
6.5.2 Computational Modeling . . . 198
6.5.3 Creating orthogonal versions and building logics . . . 198
6.6 Acknowledgments . . . 199
6.7 Methods . . . 199
6.7.1 Plasmids . . . 199
6.7.2 Strains, growth media, and fluorescence assays . . . 200
6.7.3 RNA modification and extraction for in-cell SHAPE-Seq . . . 201
6.7.4 Template preparation for cotranscriptional SHAPE-Seq . . . 201
6.7.5 In vitro transcription for cotranscriptional SHAPE-Seq . . . 202
6.7.6 RNA modification for cotranscriptional SHAPE-Seq . . . 202
6.7.7 RNA Linker preparation . . . 203
6.7.8 Linker ligation for cotranscriptional SHAPE-Seq . . . 204
6.7.9 Reverse transcription . . . 204

6.7.10 DNA adapter ligation . . . 205
6.7.11 Quality analysis . . . 205
6.7.12 Library preparation and next generation sequencing . . . 206
6.7.13 Data analysis with Spats . . . 207

7 Design of a CRISPR sgRNA Derepressor Using Insight from SHAPE-Seq 208
7.1 Abstract . . . 208
7.2 Introduction . . . 209
7.3 Results and Discussion . . . 214
7.3.1 Examining sgRNAs with in-cell SHAPE-Seq . . . 215
7.3.2 Initial characterization of the sgRNA:dCas9 complex . . . 217
7.3.3 A deeper understanding of the sgRNA:dCas9 complex through mutational analysis coupled with in-cell SHAPE-Seq . . . 219
7.3.4 Initial designs for creating responsive sgRNAs . . . 226
7.3.5 Toehold-based sgRNA designs exhibit derepression . . . 228
7.3.6 Overall design considerations . . . 229
7.4 Future work . . . 230
7.4.1 Creating orthogonal variants . . . 231
7.4.2 Building complex logics . . . 232
7.4.3 Understanding the mechanism of derepression . . . 232
7.4.4 Expanding to higher organisms . . . 234
7.5 Acknowledgments . . . 234
7.6 Materials and Methods . . . 235
7.6.1 Plasmids . . . 235
7.6.2 Strains, growth media, and RNA expression . . . 236
7.6.3 Fluorescence assay . . . 237
7.6.4 RNA modification and extraction for in-cell SHAPE-Seq . . . 237
7.6.5 Reverse transcription . . . 238
7.6.6 DNA adapter ligation . . . 238
7.6.7 Quality analysis . . . 238
7.6.8 Library preparation and next generation sequencing . . . 239
7.6.9 Data analysis with Spats . . . 240

8 Structural Analysis of Cucumber Mosaic Virus RNA3 241
8.1 Abstract . . . 241
8.2 Introduction . . . 241
8.3 Results and Discussion . . . 245
8.3.1 5′ UTR structural features . . . 245
8.4 IGR structural features . . . 249
8.4.1 3′ UTR structural features and replicase binding . . . 250
8.4.2 Complete reactivity map of RNA3 . . . 252
8.5 Conclusion . . . 253
8.6 Acknowledgments . . . 253
8.7 Materials and Methods . . . 254

8.7.1 RNA preparation . . . 254
8.7.2 RNA modification and purification . . . 255
8.7.3 Reverse transcription . . . 255
8.7.4 Adapter ligation . . . 256
8.7.5 Quality analysis . . . 256
8.7.6 Library preparation and next generation sequencing . . . 257
8.7.7 Data analysis with Spats . . . 258

9 Structural Features of Protein Binding with RNase P and its Substrates 259
9.1 Abstract . . . 259
9.2 Introduction . . . 260
9.3 Results and Discussion . . . 266
9.3.1 L7Ae binds kink-turn motifs in archaeal RNase P . . . 266
9.3.2 Observing the independence of the S and C domains with SHAPE-Seq . . . 270
9.3.3 PRORP Binds the D-loop of pre-tRNAs . . . 273
9.4 Future work . . . 279
9.4.1 Examining other RPRs and RPPs with SHAPE-Seq . . . 279
9.4.2 Studying RNase P inside cells . . . 280
9.4.3 Obtaining a deeper understanding of PRORP binding . . . 281
9.4.4 Cotranscriptional Folding of RNase P ribozymes . . . 282
9.5 Acknowledgments . . . 282
9.6 Materials and Methods . . . 283
9.6.1 RNA folding and modification . . . 283
9.6.2 RNA Linker ligation . . . 284
9.6.3 Reverse transcription . . . 285
9.6.4 DNA adapter ligation . . . 285
9.6.5 Quality analysis . . . 286
9.6.6 Library preparation and next-generation sequencing . . . 286
9.6.7 Data analysis with Spats . . . 287

10 Conclusions and Perspectives 288
10.1 Conclusion . . . 288
10.2 Future Directions and Perspectives . . . 289
10.2.1 Genome-wide RNA structure probing techniques . . . 289
10.2.2 RNA folding dynamics . . . 291
10.2.3 Improvements in RNA structural prediction . . . 292
10.2.4 Combining structural data and RNA regulator design . . . 294
10.2.5 (Next-)next-generation sequencing technologies . . . 295
10.3 Improving the SHAPE-Seq Technique . . . 295
10.3.1 General improvements . . . 296
10.3.2 in-cell SHAPE-Seq improvements . . . 297
10.3.3 Cotranscriptional improvements . . . 298
10.4 Final Perspectives . . . 300

A Supporting Information for SHAPE-Seq 2.0: Systematic Optimization and Extension of High-Throughput Chemical Probing of RNA Secondary Structure with Next Generation Sequencing **301**
A.1 Supplementary Tables . . . 301
A.2 Supplementary Figures . . . 313

B Supplementary Information for Simultaneous characterization of cellular RNA structure and function with in-cell SHAPE-Seq **334**
B.1 Supplementary Equations . . . 334
B.2 Supplementary Tables . . . 336
B.3 Supplementary Figures . . . 348
B.4 Supplementary Methods. . . . 378
B.4.1 Materials . . . 378
B.4.2 Equipment . . . 379
B.4.3 Reagent Setup . . . 380
B.4.4 Procedure for in-cell SHAPE-Seq in *E. coli* . . . 381

C Supplementary Information for Cotranscriptional Folding of a Fluoride Riboswitch at Nucleotide Resolution **401**
C.1 Materials and Methods . . . 401
C.1.1 Plasmids . . . 401
C.1.2 Proteins . . . 401
C.1.3 Template preparation . . . 402
C.1.4 *in vitro* transcription (single length, radiolabeled) . . . 403
C.1.5 *in vitro* transcription (cotranscriptional SHAPE-Seq experiment) . 404
C.1.6 *in vitro* transcription (single length, unlabeled) . . . 404
C.1.7 RNA modification and purification . . . 405
C.1.8 Linker preparation . . . 406
C.1.9 Linker ligation . . . 406
C.1.10 Reverse transcription . . . 407
C.1.11 Adapter ligation . . . 407
C.1.12 Quality analysis . . . 408
C.1.13 Library preparation and next generation sequencing . . . 408
C.1.14 Data analysis with Spats . . . 409
C.2 Supplementary Text . . . 411
C.2.1 Analysis of the SRP Cotranscriptional Folding Pathway . . . 411
C.2.2 Mutant analysis of the *B. cereus* fluoride riboswitch . . . 411
C.2.3 Cotranscriptional SHAPE-Seq accesses non-equilibrium, kinetically trapped RNA structures . . . 414
C.3 Supplementary Figures . . . 416
C.4 Suppmentary Tables . . . 443

D Supplementary Information for Measuring the cotranscriptional folding pathway of the pT181 transcriptional attenuator **448**
D.1 Supplementary Figures 448

E Supplementary Information for Design of a CRISPR sgRNA Derepressor Using Insight from SHAPE-Seq **454**
E.1 Supplementary Figures 454

F Supplementary Information for Structural Analysis of Cucumber Mosaic Virus RNA3 **479**
F.1 Supplementary Figures 479
F.2 Supplementary Tables 497

G Supplementary Information for Structural Features of Protein Binding with RNase P and its Substrates **499**
G.1 Supplementary Figures 499

LIST OF TABLES

2.1 RNA structure prediction accuracy. 58

A.1 RNA sequences used in this study. 301
A.2 RNA folding buffer conditions and ligand concentrations 304
A.3 List of barcoded reverse transcription primers 305
A.4 RNA structure prediction accuracy with no SHAPE-Seq constraints . . 307
A.5 RNA structure prediction accuracy with default SHAPE-Seq constraints 308
A.6 RNA structure prediction accuracy with newly fit SHAPE-Seq constraints309
A.7 Data deposition table 310

B.1 List of terminators screened for building the antisense platform. 336
B.2 RNA sequences and plasmids used in this study. 337
B.3 Oligonucleotides used in this study. 344
B.4 RMDB data deposition table. 347

C.1 Sequences used for *in vitro* transcription templates. 443
C.2 Oligonucleotides used in this study. 445
C.3 RMDB data deposition table. 446

F.1 Oligonucleotides used in this study. 497

LIST OF FIGURES

1.1 Organization of RNA structure. 2
1.2 Selection of the major transcriptional and translational regulation mechanisms in prokaryotes. 6
1.3 RNA cotranscriptional folding. 11
1.4 Chemical structure probing experiment with next-generation sequencing. 13
1.5 SHAPE-Seq reactivities highlight nucleotides that are structurally flexible. 15
1.6 NGS-based RNA structure probing technologies. 17
1.7 Selective PCR prevents side product amplification. 22
1.8 Attenuation mechanism from the pT181 plasmid. 26

2.1 The basic SHAPE-Seq protocol. 34
2.2 A comparison of SHAPE-Seq v1.0 to adapter ligation variations. 47
2.3 Characterization of varying numbers of PCR cycles in SHAPE-Seq library construction. 50
2.4 Schematic of traditional SHAPE/SHAPE-Seq v1.0 RT priming strategies and the universal RT priming strategy of SHAPE-Seq v2.0. 52
2.5 SHAPE-Seq v2.0 vs. SHAPE-Seq v1.0 55

3.1 In-cell SHAPE-Seq overview. 65
3.2 Characterization of the cellular structures of the taR12/crR12 synthetic riboregulator RNA translational activator system. 77
3.3 In-cell structure-function characterization of the taR12/crR12 synthetic riboregulator RNA translational activator system. 81
3.4 In-cell structure-function characterization of the RNA-IN/OUT translational repressor system. 86
3.5 Structural characterization of three endogenously expressed RNAs in *E. coli* with in-cell SHAPE-Seq. 91

4.1 SHAPE-Seq workflow. 106
4.2 Comparison of SHAPE-Seq v2.0 vs. v2.1 *in vitro* reactivities. 132
4.3 SHAPE-Seq reveals reactivity changes in the presence of ligand for the *thiM* TPP riboswitch aptamer domain. 134
4.4 Incorporating SHAPE-Seq data improves computational folding accuracy. 136
4.5 *in vitro* vs. in-cell SHAPE-Seq reactivity map comparisons for 5S rRNA and the TPP riboswitch. 140
4.6 PCR does not bias SHAPE-Seq reactivity calculations. 144
4.7 Comparison of *in vitro* folded 5S rRNA reactivity maps from different amounts of starting RNA. 146

5.1 Cotranscriptional SHAPE-Seq overview. 161
5.2 SRP RNA cotranscriptional folding. 163
5.3 *B. cereus* fluoride riboswitch cotranscriptional SHAPE-Seq data. 166
5.4 Cotranscriptional folding pathway of the *B. cereus* fluoride riboswitch. . 168

6.1 General overview and structures of pT181 attenuator system. 176
6.2 Cotranscriptional SHAPE-Seq of the pT181 attenuator with and without antisense hairpin 2. 180
6.3 Folding pathway for the wt pT181 attenuator. 184
6.4 Equilibrium refolding of the pT181 attenuator with and without the antisense hairpin 2. 188
6.5 Minimizing the pT181 attenuator. 193
6.6 Minimized pT181 attenuator. 196

7.1 Overview of the Type II CRISPR system from *S. pyogenes*. 210
7.2 In-cell SHAPE-Seq characterization of the RR2 opt sgRNA. 216
7.3 Analysis of a selection of the mutants studied by Briner *et al.* 220
7.4 Analysis of additional sgRNA revised mutants. 222
7.5 Initial sgRNA designs containing RNA regulator motifs. 227
7.6 Toehold sgRNA designs derepress RFP expression. 229

8.1 Genome layout of the Cucumber Mosaic Virus. 243
8.2 Structures of the isolated untranslated regions of CMV RNA3. 246
8.3 Predicted secondary structure of CMV RNA3 using purified virions. . . 251

9.1 Characteristics of RNase P. 262
9.2 Proteinaceous RNase P structure. 265
9.3 Characterization of L7Ae binding in the two archaeal RPRs from *Pyrococcus furiosus* (*Pfu*) and *Methanocaldococcus jannaschii* (*Mja*). 268
9.4 Analysis of an *Mja* RPR circular permutant. 271
9.5 PRORP binds to the D-loop of pre-tRNA^{Cys}. 275
9.6 PRORP mutants exhibit similar pre-tRNA binding patterns as the wt. . 277
9.7 Direct priming two pre-tRNAs shows similar patterns of PRORP protection. 278

A.1 SHAPE-CE Flowchart 313
A.2 SHAPE-Seq v1.0 and Second Adapter Variation Library Construction Schematics 314
A.3 SHAPE-Seq v1.0 library indexing strategy. 316
A.4 SHAPE-Seq v2.0 Library Construction Schematic. 317
A.5 QuSHAPE vs. SHAPE-Seq (v1.0) detailed comparisons. 318
A.6 SHAPE-Seq v1.0 vs. Minimal or Inverted adapter variations for RNase P and tRNA. 321
A.7 Time course of CircLigase I ligation efficiency. 322
A.8 Ligase comparison for addition of SHAPE-Seq second adapter. 323
A.9 Modification and optimization of RT primer blocking groups for adapter ligation. 324
A.10 Effect of 3′ blocking group on second adapter concatemerization and ligation efficiency. 325

A.11 SHAPE-Seq v1.0 fragment distributions for different numbers of PCR cycles. . . . 326
A.12 SHAPE-Seq v2.0 vs. SHAPE-Seq v1.0. . . . 327
A.13 Choice of 5′ adenylated linker sequence. . . . 331
A.14 SHAPE-Seq v2.0 reactivities generated from the MiSeq and HiSeq platforms. . . . 333

B.1 Standardized platform for expressing sense/antisense regulatory RNA pairs in *E. coli.* . . . 348
B.2 In-cell SHAPE-Seq structural characterization overview. . . . 349
B.3 Selective PCR amplification strategy for cDNA libraries. . . . 351
B.4 Mechanism of the synthetic translation-activating riboregulator system. 352
B.5 Structural analysis of the synthetic riboregulator translational activation system using in-cell SHAPE-Seq data. . . . 353
B.6 Characterization of the cellular structures of the taR10/crR10 synthetic riboregulator RNA translational activator system. . . . 354
B.7 Characterization of the cellular structures of the taR12/crR12 synthetic riboregulator RNA translational activator system using in-cell dimethyl sulfate (DMS) probing. . . . 355
B.8 Structure-function characterization of the taR10/crR10 synthetic riboregulator RNA translational activator system. . . . 356
B.9 In-cell SHAPE-Seq reactivities for the trans-activating RNA variants. . . 357
B.10 Determining the dominant RBS sequence in crRNAs using in-cell SHAPE-Seq reactivities. . . . 358
B.11 Mechanism of the translation-repressing RNA-IN/OUT system. . . . 360
B.12 Characterization of the cellular structures of the S3/A3 RNA-IN/OUT translational repressor system. . . . 361
B.13 RNA-IN/OUT S4/A4 interaction complexes appear to be cleaved by a double-stranded RNase in the cell. . . . 362
B.14 Structure-function analysis of the S4/A3 interaction from the RNA-IN/OUT translational repressor. . . . 364
B.15 Structure-function analysis of the S3/A4 interaction from the RNA-IN/OUT translational repressor. . . . 365
B.16 RNA-IN mutations to resist RNase cleavage between nucleotides 25 and 26. . . . 366
B.17 RNA-IN S4 mutations resist cleavage and maintain functionality. . . . 367
B.18 Ribosome binding site analysis of RNA-IN S4 mutant C24A A25C. . . . 369
B.19 In-cell SHAPE-Seq reactivities for the antisense RNA-OUT A4 with RNA-IN variants. . . . 370
B.20 Increasing PCR selection does not affect reactivity calculation. . . . 371
B.21 Reactivity maps for endogenous RNA targets. . . . 372
B.22 In-cell vs. equilibrium *in vitro* refolding reactivity maps for the riboregulators. . . . 373

B.23 In-cell vs. equilibrium *in vitro* refolding reactivity maps for RNA-IN/OUT. 375
B.24 In-cell vs. equilibrium *in vitro* refolding reactivity maps of 5S rRNA from *E. coli*. 377
B.25 CE quality control example. 394

C.1 3D SRP RNA cotranscriptional SHAPE-Seq data. 416
C.2 SHAPE-Seq data for SRP RNA equilibrium refolded. 417
C.3 Transcription antitermination by the *B. cereus* crcB fluoride riboswitch. . 419
C.4 3D wt fluoride riboswitch cotranscriptional SHAPE-Seq data. 420
C.5 Folding of the P1 stem. 422
C.6 Folding of the P3 stem. 423
C.7 Reactivity profiles for A24, A25, and C27 of the wt fluoride riboswitch over the course of transcription. 424
C.8 *B. cereus* fluoride riboswitch mutants. 425
C.9 Cotranscriptional SHAPE-Seq data for fluoride riboswitch mutant M18. 427
C.10 Cotranscriptional SHAPE-Seq data for fluoride riboswitch mutant M19. 429
C.11 Cotranscriptional SHAPE-Seq data for fluoride riboswitch mutant M20. 430
C.12 Cotranscriptional SHAPE-Seq data for fluoride riboswitch mutant M21. 432
C.13 Cotranscriptional SHAPE-Seq data for fluoride riboswitch mutant M22. 434
C.14 Cotranscriptional SHAPE-Seq data for fluoride riboswitch mutant M23. 436
C.15 Reactivity profiles for A52, G53, U54, and A55 of the wt fluoride riboswitch over the course of transcription. 438
C.16 Reactivity profiles for G12, G13, A14, G15, and U16 of the wt fluoride riboswitch over the course of transcription. 439
C.17 SHAPE-Seq data for wt fluoride riboswitch equilibrium refolded. 440
C.18 Reactivity profiles for A67, G68, G69, A70, G71, and U72 of the wt fluoride riboswitch over the course of transcription. 442

D.1 In-cell SHAPE-Seq of the pT181 attenuator with and without antisense. 448
D.2 Potential sense-antisense interaction during equilibrium refolding. . . . 449
D.3 Removing the RBS from the pT181 attenuator. 450
D.4 Testing the impact of the RepC mRNA on gene expression. 451
D.5 Determining the minimal antisense length. 452
D.6 Increased polyU length does not improve termination. 453
D.7 Cotranscriptional SHAPE-Seq of the minimized pT181 attenuator. . . . 453

E.1 Moving the sgRNA to the in-SHAPE-Seq platform. 454
E.2 Secondary structures and in-cell SHAPE-Seq data for all 'v' series mutants. 455
E.3 Secondary structures and functional and in-cell SHAPE-Seq data for all 'r' series mutants. 459
E.4 Secondary structures and functional and in-cell SHAPE-Seq data for all 'd' series designs. 469

E.5 Secondary structures for the first four 't' series toehold designs. 475
E.6 Antisense orthogonality test of the sgRNA t4 design. 476
E.7 Assessing sgRNA target orthogonality in the sgRNA t4 design. 477
E.8 Potential OR gate design combining t5 and t6 designs. 478

F.1 SHAPE-Seq reactivity maps of the RNA3 5′ UTR. 479
F.2 SHAPE-Seq reactivity maps of the RNA3 IGR. 481
F.3 Alternate fold of the IGR. 483
F.4 SHAPE-Seq reactivity maps of the RNA3 3′ UTR. 484
F.5 SHAPE-Seq reactivities for the complete RNA3 from purified virions and infected cell lysates. 485
F.6 SHAPE-Seq reactivities for the complete RNA3 from using refolded *in vitro* transcripts. 491
F.7 Predicted secondary structure of CMV RNA3 using *in vitro* transcripts. 496

G.1 *Pfu* RPR reactivity maps with and without *Pfu* L7Ae. 499
G.2 *Mja* RPR reactivity maps with and without *Mja* L7Ae. 501
G.3 Reactivity maps of the *Pfu* RPR C domain. 503
G.4 Individual reactivity maps for pre-tRNA^{Cys} incubated with different PRORPs. 504
G.5 Replicate data of PRORP binding to pre-tRNA^{Cys} with a 12 nt leader and 23 nt trailer. 505
G.6 L- vs. λ-form structures of pre-tRNA^{Cys}. 506

CHAPTER 1

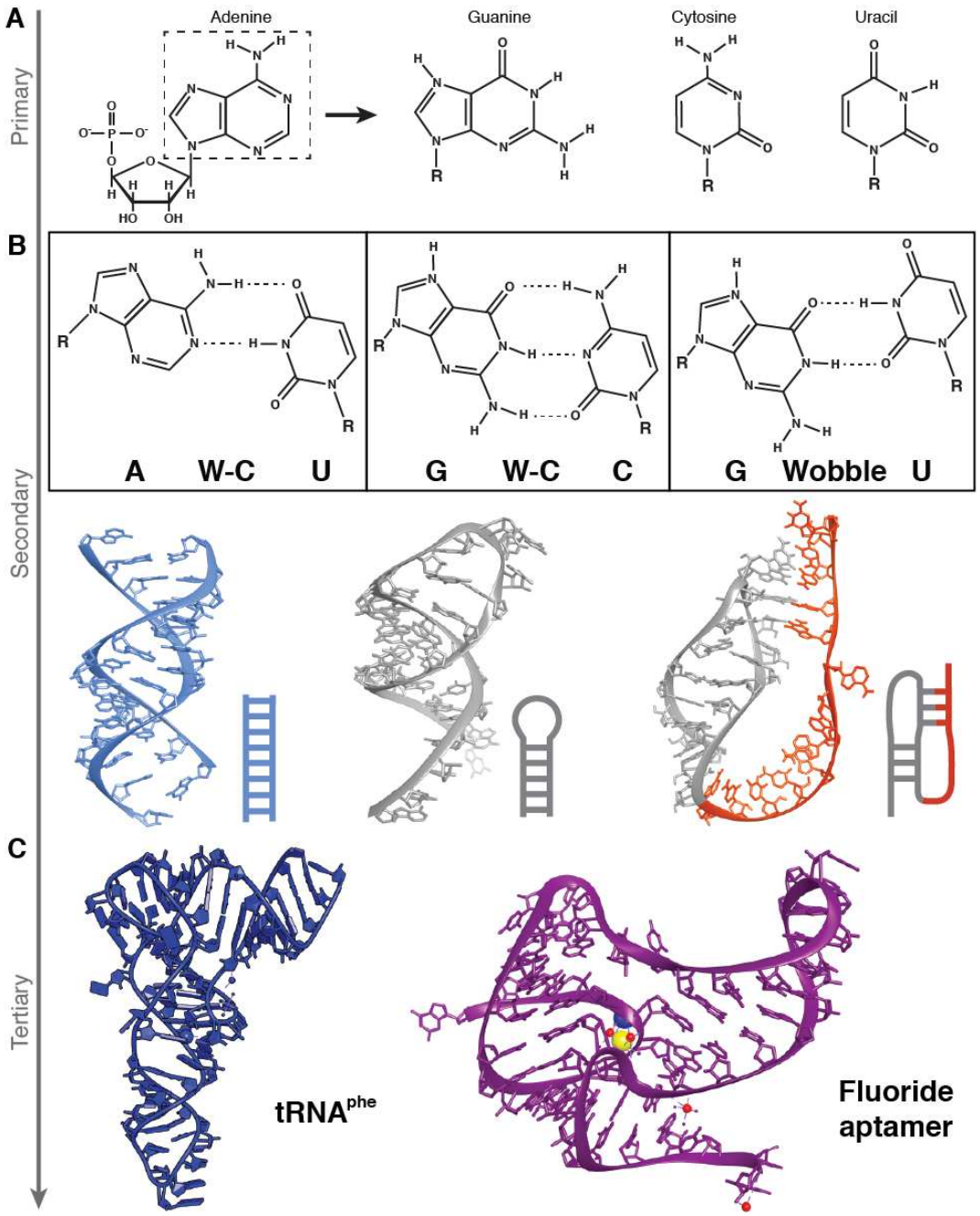
INTRODUCTION

1.1 The role of RNA structure in biology

RNA is one of the most versatile and abundant biological molecules in nature. The role of RNA in biology includes a vast array of functions critical to life that include, among others, encoding protein synthesis [1], catalyzing chemical reactions [2], storing genetic information [3], defending the cell [4, 5], and regulating gene expression [6, 7]. The incredible diversity of RNA functions has led many biologists to believe in an RNA World hypothesis that proposes that self-replicating RNAs were the precursors to terrestrial life [8].

A critical property of RNA is its ability to fold into intricate three-dimensional structures, despite being primarily composed of only four nucleotides: guanosine, adenosine, cytidine, and uridine. These three-dimensional structures are directly derived from the sequence of an RNA molecule and are generated by internal interactions between nucleotides that can include Watson-Crick base pairing, Hoogsteen base pairing, π-π stacking interactions, and other electrostatic interactions (Figure 1.1A) [9]. Together, these interactions confer upon each RNA sequence a folded structure, or ensemble of folds, that determine(s) the function of that RNA (Figure 1.1B, C). Thus, there is an intimate relationship between the sequence, structure, and function of an RNA.

Figure 1.1: Organization of RNA structure. RNA structure is organized in three levels: primary, secondary, and tertiary. (A) Primary RNA structure refers to the nucleotide sequence of an RNA. Each nucleotide is composed of a phosphate group, ribose sugar, and a nucleobase. (B) Nucleotides can form both canonical Watson-Crick (W-C) and non-canonical base pairs, including a fairly common G•U wobble pair. Secondary RNA structures include local structures such as a basic helix (left, blue), stem-loop (middle, gray), or pseudoknots (right, gray-orange). (C) At the tertiary level, interactions between secondary structures generate the overall structure of the RNA.


A
Primary
Adenine
Guanine
Cytosine
Uracil
B
Secondary
A W-C U
G W-C C
G Wobble U
C
Tertiary
tRNA^phe
Fluoride aptamer

RNA structures are crucial to many of the most conserved elements of life. For example, the core processes in protein translation are all RNA-based: messenger RNA (mRNA) bears protein sequence information, transfer RNA (tRNA) decodes the mRNA message, and ribosomes (as a multi-subunit RNA-protein complex) catalyze the polymerization of the nascent polypeptide chain [10]. Further, a pre-processing step of tRNA maturation involves an endoribonuclease, RNase P [11]. Even before translation, RNAs also affect the process transcription. In prokaryotes, RNA hairpin structures called intrinsic terminators signal end of transcription and cause RNA polymerase (RNAP) to dissociate from the template DNA [12]. In eukaryotes, RNA structures in the spliceosome direct alternative splicing of exons to generate new protein coding sequences [13]. In both domains of life, RNA structures can also resist, or promote, RNA degradation by cellular machinery [14–16]. RNA structures are also involved in cellular defense. In many prokaryotes, recently discovered CRISPRs (clustered regularly interspersed palindromic repeats) protect cells from invading plasmids and phages using structured RNAs within protein-RNA complexes to target and cleave the unwanted nucleic acids [4, 5]. Yet, one of the biggest roles of RNA structures in the cell is to help maintain cellular homeostasis by regulating the processes of transcription and translation.

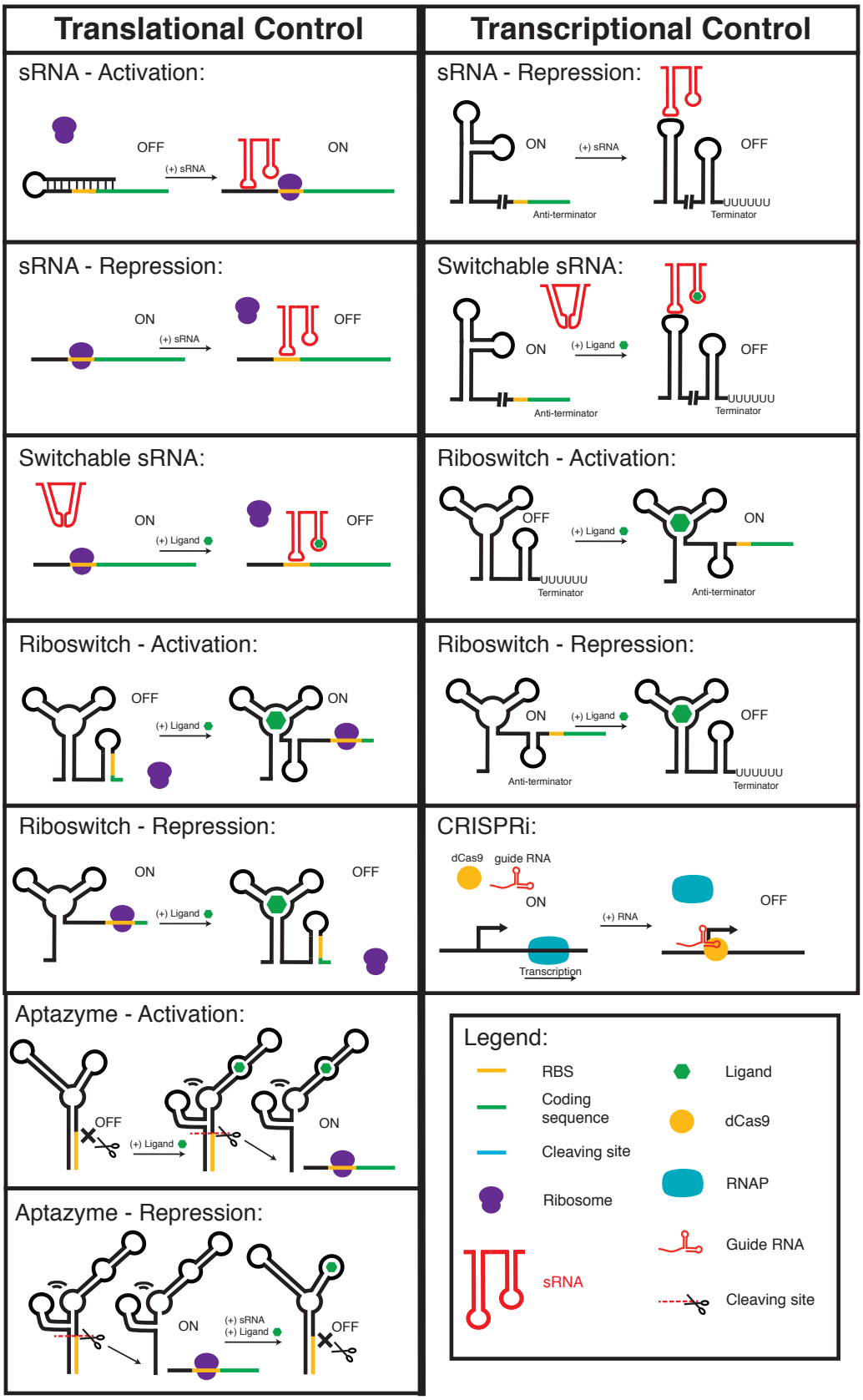
1.2 RNA structures that regulate translation in prokaryotes

There are many different types of RNA structures that can regulate translation. In prokaryotes (eukaryotic gene expression is not discussed in this thesis work), translation is first initiated by a ribosome binding to a sequence roughly 8-12 nts upstream of a start codon with the consensus sequence 'AGGAGG' [17]. The basal level of translation from a particular ribosome binding site (RBS) is dictated by two factors: 1) *trans*

interactions between the ribosome and the RBS and 2) the accessibility of the RBS in the mRNA [18, 19]. The most common method of prokaryotes use to regulate translation is to modulate RBS accessibility within mRNAs, although RNA stability and translation fidelity methods are also employed [18, 19]. Typically, access to the RBS is controlled by alternative RNA folding within the 5′ untranslated region (UTR) of an mRNA, such that one structure exposes the RBS to the translation machinery while another occludes the RBS, preventing ribosome binding [20].

One common method prokaryotes use to control RBS accessibility on a global level is through networks of sense/antisense RNA-RNA interactions (Figure 1.2) [21, 22]. In these networks, small RNAs (sRNA) bind to target mRNAs, mainly through Watson-Crick base pairing, to block or reveal an RBS [21, 22]. sRNAs are also used to control plasmid copy number, as observed in the ColE1 origin of replication and the *hok/sok* system of the R1 plasmid [23]. These types of sRNA regulators have inspired many RNA engineering efforts to devise synthetic versions of sRNA-mRNA pairs to control gene expression (Figure 1.2) [24]. The first such effort was published by Isaacs *et al.* in which the *hok/sok* was modified to create the first riboregulators; RNA hairpin structures in the 5′ UTR of mRNAs that would open to reveal an RBS after binding a trans-acting sRNA [25]. Others followed suit, designing RNA regulators that used RNA structural rearrangements triggered by an sRNA-mRNA interaction to reveal an RBS [24, 26].

Figure 1.2: Selection of the major transcriptional and translational regulation mechanisms in prokaryotes. Mechanistic representation of prokaryotic regulators of translation (left) and transcription (right). Adapted with permission from Chappell *et al.* 2013 [24].


Translational Control
Transcriptional Control
sRNA - Activation:
OFF
(+) sRNA
ON
sRNA - Repression:
ON
(+) sRNA
OFF
Switchable sRNA:
ON
(+) Ligand
OFF
Riboswitch - Activation:
OFF
(+) Ligand
ON
Riboswitch - Repression:
ON
(+) Ligand
OFF
Aptazyme - Activation:
OFF
(+) Ligand
ON
Aptazyme - Repression:
ON
(+) sRNA
(+) Ligand
OFF
sRNA - Repression:
ON
(+) sRNA
OFF
Anti-terminator
UUUUUU
Terminator
Switchable sRNA:
ON
(+) Ligand
OFF
Anti-terminator
UUUUUU
Terminator
Riboswitch - Activation:
OFF
(+) Ligand
ON
UUUUUU
Terminator
Anti-terminator
Riboswitch - Repression:
ON
(+) Ligand
OFF
Anti-terminator
UUUUUU
Terminator
CRISPRi:
dCas9
guide RNA
ON
(+) RNA
OFF
Transcription
Legend:
RBS
Coding sequence
Cleaving site
Ribosome
sRNA
Ligand
dCas9
RNAP
Guide RNA
Cleaving site

A second method prokaryotes employ to regulate gene expression is a class of structured RNAs called riboswitches (Figure 1.2) [27, 28]. Riboswitches are RNA structures that contain an aptamer, an RNA structure that can recognize and bind a small molecule ligand, and an expression platform. Riboswitches typically reside in the 5′ UTR of an mRNA and fold into one of two structures depending on whether the aptamer domain binds its ligand. One of its possible structures contains a stabilized, ligand-bound aptamer, while the alternative contains a disrupted aptamer [27–32]. The state of the aptamer affects the structure of the expression platform that controls the translation of a downstream open reading frame (ORF) by either exposing an RBS or occluding it [27, 28, 31–33].

Like the sRNA-mRNA type of regulation, synthetic biologists have undergone efforts to design novel synthetic riboswitches (Figure 1.2) [34, 35]. A typical approach involves taking a well-characterized aptamer domain from one riboswitch and fusing it with a ribozyme (aptazyme) [36, 37] or an entirely synthetic sequence [38–40]. Both translational repressors and activators have been designed with these strategies (Figure 1.2) [34–36, 41]. However, engineering synthetic riboswitches is more difficult compared to simple synthetic riboregulators due to the challenge of understanding how mutations in a riboswitch impact switching between its two structures, which typically contain hard to predict noncanonical interactions.

1.3 RNA structures that regulate transcription in prokaryotes

Many of the regulatory mechanisms used by prokaryotes to control translation are also used to control transcription (Figure 1.2) [6, 7]. However, instead of switching between RBS accessible/inaccessible structures, transcriptional attenuators typically

either form an intrinsic terminator or an antiterminator structure. An intrinsic terminator is an RNA hairpin loop that is followed by a 6-8 U-rich tract that causes the RNAP elongation complex to become unstable and dissociate from its template DNA [12]. While the characteristic inverted repeat motif of an intrinsic terminator is easily identified, fewer examples of genes regulated via intrinsic termination are known compared to those regulated by RBS accessibility [7].

Riboswitches can also control transcription [27, 42, 43]. Most of the riboswitch families that have been discovered so far contain variants that regulate either transcription or translation, or both [44–50]. Like the translational variants, the aptamer domain exists in one of two states: stably bound to the ligand or at least partially destabilized in favor of an alternative structure, one of which will allow an intrinsic terminator to fold [27, 28, 43].

sRNA interactions can also dictate the formation of an intrinsic terminator (Figure 1.2). While rarely in bacterial genomes, sRNA-mRNA sense/antisense interactions have been discovered to play a role in a number of plasmid copy-control mechanisms [51]. The first known example of an RNA-RNA interaction leading to transcription termination event was discovered in the plasmid pT181 [52–54]. In the pT181 replication system, an antisense RNA binds the 5′ UTR of an mRNA encoding a replication protein to prevent the formation of an antiterminator structure, allowing an intrinsic terminator to fold instead [53]. No known natural examples of transcriptional anti-attenuators (activators) exist, but a recent study demonstrated that intrinsic terminators could be converted into sRNA-responsive transcriptional activators [55].

Recently, the acquired immunity CRISPR system from *S. pyogenes* was adapted to create a mechanism of transcriptional regulation called CRISPR interference (CRISPRi; Figure 1.2) [56, 57]. In the wt CRISPR system, CRISPR associated (Cas) proteins, guided

by a structured RNA or pair of structured RNAs, bind to a nucleic acid target and cleave it [4, 5, 58, 59]. In CRISPRi, however, the endonucleolytic capabilities of the Cas protein(s) are removed, creating an RNA-guided, nucleic acid binding RNA-protein complex. The specific CRISPRi system derived from *S. pyogenes* includes a catalytically dead Cas9 (dCas9) protein that can be targeted to promoters or ORFs to prevent RNAP initiation or elongation, respectively [57]. The binding of dCas9 is very tight and slow to dissociate [60], making CRISPRi a very strong mechanism of repression.

1.4 Cotranscriptional RNA folding and its role in structure formation

The timescale of RNA folding is faster than RNA transcription, leading to the formation of RNA structures during the process of RNA synthesis (Figure 1.3) [61, 62]. For many regulatory RNA structures, the cotranscriptional nature of RNA folding is critical, especially for those structures that regulate transcription (discussed above), as they must be able to arrive at different final folding states before RNAP transcribes past them [63–65]. For some large and/or complex RNA structures, cotranscriptional folding can also be an important factor to ensure that the correct final structures are achieved by establishing kinetically trapped structures [66–70].

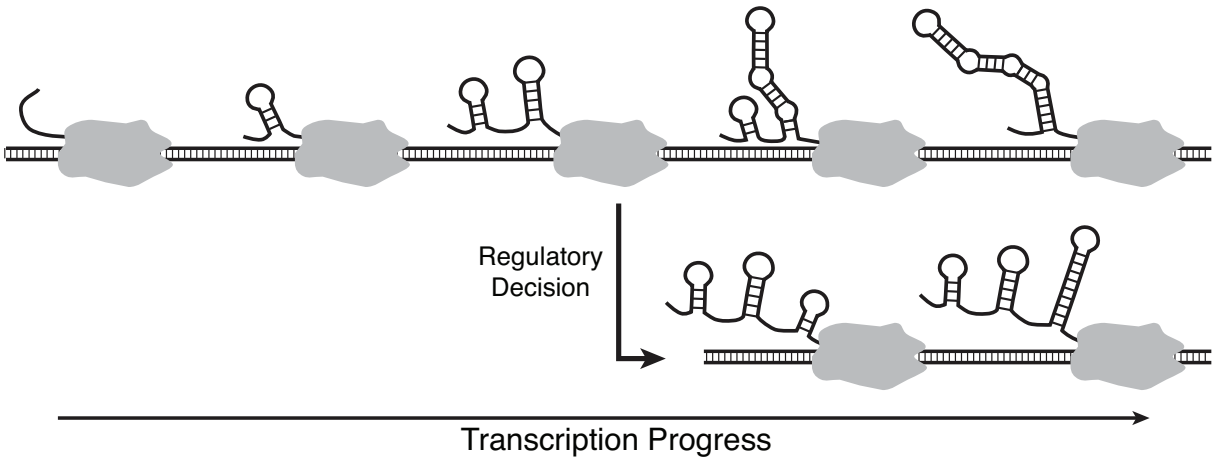


Figure 1.3: RNA cotranscriptional folding. As RNAs are transcribed by RNA polymerase (gray), they immediately begin to fold into secondary and tertiary structures. Some regulatory decisions change the cotranscriptional folding pathway, resulting in a different final structure.

Current studies aimed at understanding the contribution of cotranscriptional folding to RNA structure are largely limited to circular permutation studies [71–73], computational folding [74–77], endpoint structure probing assays [78], and single molecule pulling experiments [79–81], all of which are low-throughput. A few of these pioneering studies are beginning to identify how RNAs fold cotranscriptionally, but with fairly low resolution [79, 80, 82, 83]. Thus, there is a need for improved techniques that are able to both capture cotranscriptional folding events with high resolution and be high-throughput.

1.5 Understanding RNA function by measuring RNA structure

There are a number of experimental techniques available to measure RNA structure including crystallography [84], nuclear magnetic resonance (NMR) [85], and chemi-

The text in this section is in press at *Methods in Molecular Biology* to be published in August, 2016.

cal/enzymatic probing [86–88]. Of these, crystallographic methods and NMR produce structures with the highest resolution, yielding very high quality three-dimensional representations of RNA structure. However, obtaining good crystals for crystallographic methods can be challenging and/or time-consuming and does not work well for RNAs that have many disordered regions [84]. NMR experiments, on the other hand, are less technically challenging than crystallographic methods, but are difficult to solve and typically limited to smaller RNAs [85].

Mapping RNA structure with chemical probing or nuclease cleavage has become a powerful technique for uncovering RNA structure-function relationships in a broad array of contexts [86]. Chemical probing experiments use reagents that covalently modify RNAs in a structure-dependent fashion, allowing the structure of an RNA under study to be inferred once the locations of the modifications are determined. Similarly, nuclease cleavage experiments use ssRNA and dsRNA endonucleases to selectively cut at specific nucleotides depending on their structural context. Although chemical probing and nuclease cleavage structural information is lower resolution than that produced by crystallography or NMR [86, 89], the experimental speed, flexibility, and accessibility of RNA chemical probing and nuclease cleavage experiments have made them amenable to many different RNA structural biology studies.

RNA chemical structure probing consists of several distinct steps: preparation and folding of an RNA of interest, structure-dependent covalent modification of the RNA at the nucleotide level, and determination of the modification locations (Figure 1.4) [86]. The locations of the modifications can then be used to infer the underlying RNA structure, since many chemical probes preferentially modify nucleotides that are unstructured. Thus, a higher frequency of modification can be used to infer nucleotides that are present in single stranded regions, loops, or bulges [90–92].

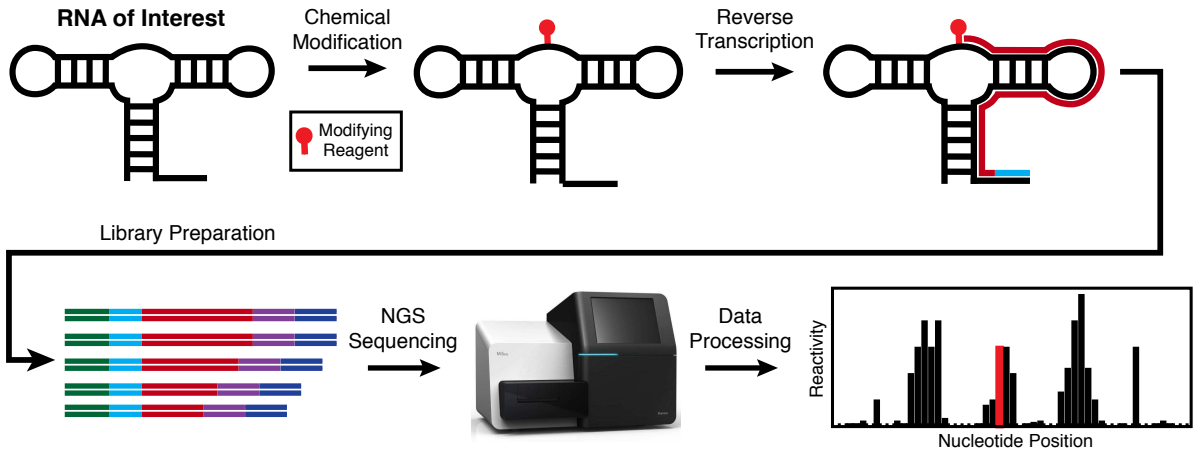


Figure 1.4: Chemical structure probing experiment with next-generation sequencing. A typical next-generation sequencing (NGS) chemical probing experiment begins with a folded RNA of interest that is then subject to structure-dependent covalent modification by a chemical reagent (light red). Modification positions are located by reverse transcription primer (blue) extension that stops one nucleotide before the modification position. The cDNA product (dark red) is then prepared for NGS (see Figure 1.6 for details). Last, the NGS sequencing reads are converted to a measure of reactivity that represents the relative frequency of modification, or structural flexibility, of a particular nucleotide. The red modification position is indicated in the hypothetical reactivity data.

While there are a wide variety of chemicals that can be used to probe RNA structure [86, 91, 93, 94], in this thesis work I will focus on the SHAPE class of chemical probes [86, 95]. SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension) reagents react with the 2'-OH group of the ribose backbone to form covalent adducts at nucleotides that are structurally flexible [90, 96]. The positions of these adducts are then detected with reverse transcriptase (RT) primer extension, which stops one nucleotide before the modification, to create a pool of cDNAs whose lengths reflect the location of SHAPE modification (+ channel) [92]. A control RT primer extension on an unmodified RNA (- channel) is also performed to identify locations where the RT has a natural propensity to abort extension and 'drop off'. The location and frequency of modifications and the natural RT drop off sites from the (+) and (-) channel cDNAs, re-

spectively, are used to estimate a 'reactivity' for each nucleotide in the RNA (Figure 1.5) [97, 98]. High SHAPE reactivities correspond to nucleotides that are unstructured, and are more likely to occur in single-stranded regions, loops, or bulges. Conversely, low reactivities can correspond to constrained nucleotides located in a double-stranded helix, bound to a protein or ligand, or involved in a non-canonical base pair [99]. In addition to qualitative interpretation, SHAPE reactivity data can also be quantitatively used to constrain secondary structure prediction algorithms to generate RNA structure models that are more consistent with experimental measurements [100–102].

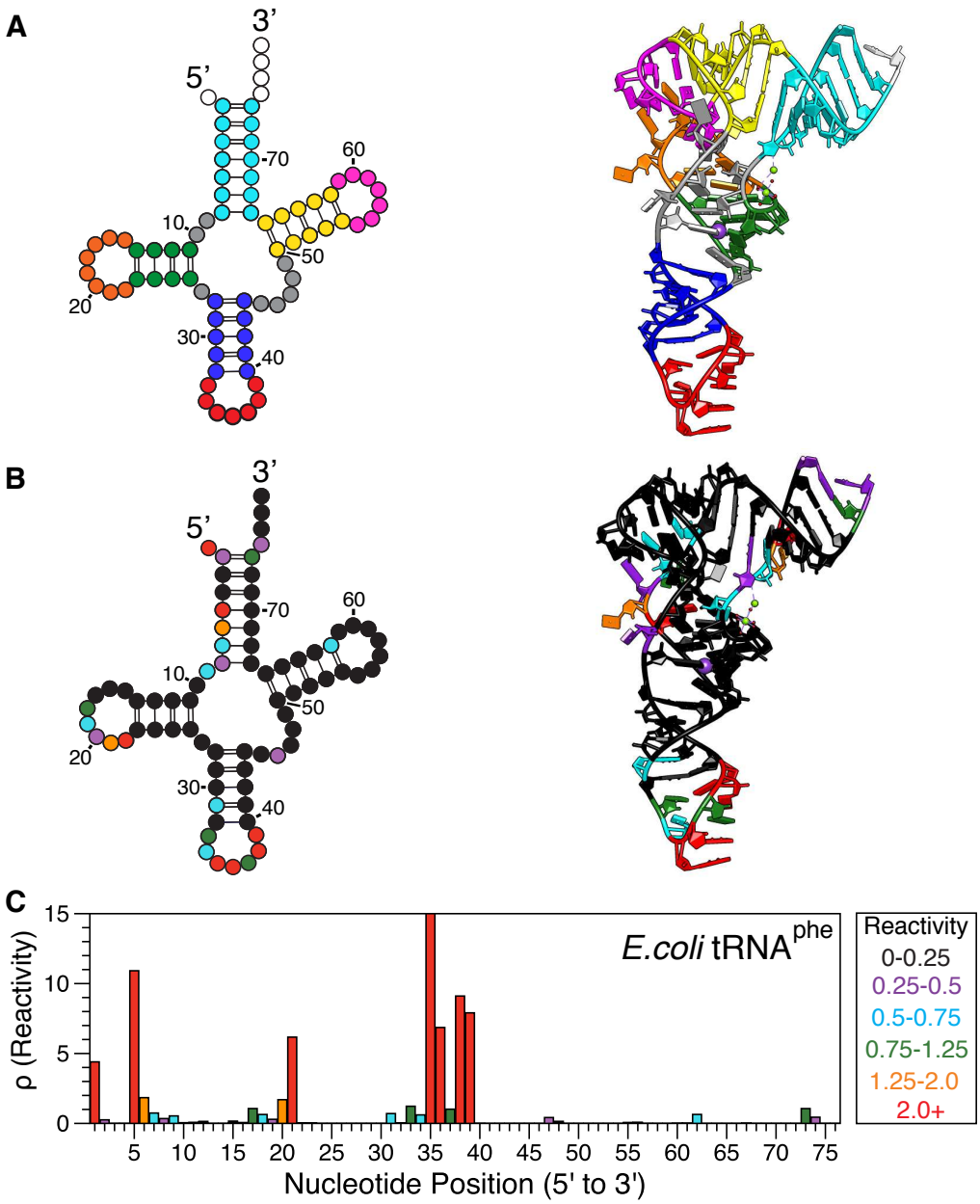
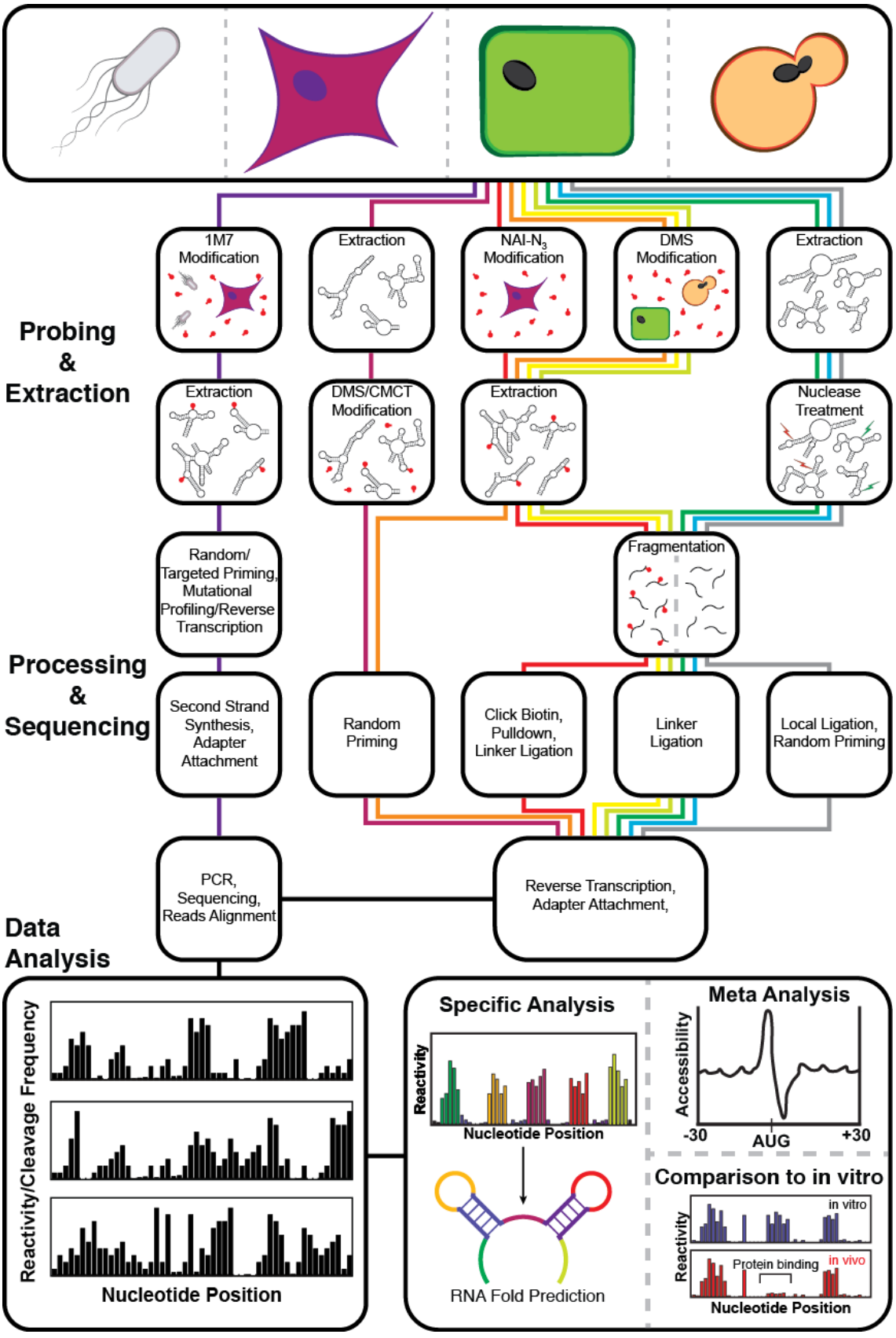


Figure 1.5: SHAPE-Seq reactivities highlight nucleotides that are structurally flexible. (A) Secondary structure model of *E. coli* $tRNA^{phe}$ (left) colored to show features of the tertiary structure of $tRNA^{phe}$ (right; PDB: 3L0U) [103]. Note that the secondary structure image does not capture the tertiary interaction between the hairpin 1 (orange) and hairpin 3 (magenta) loops. (B) Secondary and tertiary structures of $tRNA^{phe}$ colored to show SHAPE-Seq reactivity intensity (ρ) according to the bar chart in (C). The nucleotides involved in the tertiary interaction between hairpins 1 and 3 have low reactivities. (C) Bar chart showing a representative SHAPE-Seq reactivity spectrum for $tRNA^{phe}$, colored to reactivity intensity.

To estimate SHAPE reactivities, the location and frequency of SHAPE modifications and natural RT drop off sites need to be determined. Originally, this was done using gel electrophoresis with radiolabeled primers [92]. A sequencing ladder provided the location of the 3′ end of the cDNAs in each channel, revealing the modification location, and a comparison of the band intensities with or without the chemical probe provided the relative frequency of modification [99]. Fluorescently labeled primers were introduced later as an alternate readout method, which improved throughput with the use of capillary electrophoresis (CE) [104] and simplified data analysis with the SHAPEFinder and QuSHAPE software packages [105, 106]. However, both methods are limited in that they cannot be multiplexed and suffer from the noise associated with integrating analog signals to determine the abundance of each cDNA length. Ultimately, these problems were solved by coupling the chemical modification step to next-generation sequencing (NGS) to determine the modification and natural RT drop off positions, creating the SHAPE-Seq method [107, 108]. Many other NGS-based chemical probing techniques were soon developed that, together with a few NGS-based enzymatic methods, focused on examining RNA structures transcriptome-wide, providing a wealth of new RNA structural information from multiple species (Figure 1.6) [109–111].

Figure 1.6: NGS-based RNA structure probing technologies. RNA structure probing uses chemical probing to introduce covalent modifications or enzymatic cleavage or directly cleave RNA in a structure-dependent fashion. Modification or cleavage positions are detected through processing steps followed by next-generation sequencing. Computational analysis of the resulting sequencing reads yields a measure of chemical modification 'reactivity' or enzymatic cleavage frequency at each nucleotide. High enzymatic cleavage or chemical modification frequencies give information on structure depending on the characteristics of the nucleases or probes used. These values can be used for several types of specific analyses, such as restrained RNA folding, meta-analysis of the entire transcriptome, and comparisons between *in vitro* and *in vivo* probing data. An outline of the steps for SHAPE-MaP [112] and in-cell SHAPE-Seq [113] (purple), CIRS-Seq [114] (maroon), icSHAPE [115] (red), structure-seq [116] (orange), DMS-Seq [117] (yellow), Mod-Seq [118] (light green), PARS [119] (green), FragSeq [120] (blue), and RPL [121] (gray), is shown. Figure reproduced with permission from Strobel *et al.* 2016 [109].


Probing & Extraction
1M7 Modification
Extraction
NAI-N3 Modification
DMS Modification
Extraction
Extraction
DMS/CMCT Modification
Extraction
Nuclease Treatment
Fragmentation
Processing & Sequencing
Random/ Targeted Priming, Mutational Profiling/Reverse Transcription
Second Strand Synthesis, Adapter Attachment
Random Priming
Click Biotin, Pulldown, Linker Ligation
Linker Ligation
Local Ligation, Random Priming
Data Analysis
PCR, Sequencing, Reads Alignment
Reverse Transcription, Adapter Attachment,
Reactivity/Cleavage Frequency
Nucleotide Position
Specific Analysis
Reactivity
Nucleotide Position
RNA Fold Prediction
Meta Analysis
Accessibility
-30
AUG
+30
Comparison to in vitro
in vitro
Reactivity
Protein binding
in vivo
Nucleotide Position

1.6 Developing new SHAPE-Seq technologies for new contexts

The main goal of this thesis work was to greatly expand the capabilities of the SHAPE-Seq technique described above to be able to measure as many different types of RNA structures in as many different contexts as possible. The following subsections below first detail how the original *in vitro* SHAPE-Seq protocol was improved and extended to relax the sequence requirements for RT priming (Chapters 2 and 4), probe RNAs directly inside living cells (Chapter 3), and measure cotranscriptional folding pathways *in vitro* (Chapter 5). Then the application of these new techniques to measure RNA structures from all three domains of life is described in the remaining subsections, which analyze many of the types of RNA structures discussed above, including translational regulators (Chapter 3), transcriptional regulators (Chapters 5 to 7), RNA-protein complexes (Chapters 8 and 9), and ribozymes (Chapters 3 and 9). Each subsection describes a technical challenge or new structural context that had yet to be solved/explored during the time this thesis work was conducted, and a brief summary of the results is presented.

1.6.1 Chapter 2: Developing SHAPE-Seq v2.0

In the original implementation of SHAPE-Seq [107, 108], RNA hairpins were included in RNA of interest to facilitate RT. These extra hairpins added to the bulk of the RNA, already subject to a ~350 nt limit, and could potentially cause misfolding depending on the RNA of interest. Simply removing the hairpins, however, would require directly priming from the RNA, resulting in an information loss, as no structural data can be obtained from the primer binding site for RT. Thus, the first goal of this thesis research

The work described in this chapter includes contributions from David Loughrey, Alexander H. Settle, and Julius B. Lucks.

was to develop a new 'universal' ligation strategy that would allow indirect RT priming from an additional ligated sequence. Described in Chapter 2, the new SHAPE-Seq v2.0 method introduced a series of protocol optimizations as well as new ligation/RT steps to facilitate RT from the 3' end of any RNA, regardless of sequence [122]. After technique development, we compared SHAPE-Seq reactivity profiles obtained from the v2.0 with the original SHAPE-Seq technique on a panel of well-known RNAs to benchmark the relaxed sequence requirement and found excellent agreement between the two methods. We also refit the SHAPE restraint parameters m and b for RNAstructure [101, 102] for a newly established reactivity parameter ρ. The new SHAPE-Seq v2.0 technique that resulted from the work in Chapter 2 also served as a starting point for the development of SHAPE-Seq v2.1 [123], as discussed in Chapter 4.

1.6.2 Chapter 3: Characterizing Cellular RNA Structure and Function with in-cell SHAPE-Seq

Early NGS chemical probing methods were only suited for measuring RNA structures *in vitro* and could not capture effects of the cellular environment on RNA structure. In fact, when the beginning of the research discussed in Chapter 3 began, dimethyl sulfate (DMS) was the only RNA chemical probe known to penetrate into living cells [124, 125]. To be able to access RNA structural information inside cells using NGS, we undertook the research conducted in Chapter 3 to develop the in-cell SHAPE-Seq method [113] to examine RNA structures in living *E. coli* cells. we found, along with Tyrrell *et al.* [126], that the fast-acting SHAPE reagent 1-methyl-7-nitroisatoic anhydride (1M7) could modify RNAs in the cell. However, during the course of developing

The work described in this chapter includes contributions from Timothy R. Abbott and Julius B. Lucks.

in-cell SHAPE-Seq, we observed that reverse transcription from total RNA tended to generate poor cDNA yield, leading downstream ligation issues. In SHAPE-Seq [107, 108], and most NGS based probing methods [109], a DNA-DNA ligation step is required to introduce sequences required for NGS, but can also generate unwanted side products when unincorporated RT primer ligates to the DNA adapter. When RT yields are low, the unwanted side product accumulates to high levels, resulting in poor quality cDNA libraries.

To reduce the levels of the unwanted ligation side product, we borrowed a technique from a single nucleotide polymorphism (SNP) discovery method (Figure 1.7). To discover SNPs, two primers are annealed together across a base suspected to contain a SNP, then the ligase chain reaction (LCR) is used [127], which only amplifies perfect matches [128]. Inspired by the selective LCR method, we reconfigured the PCR steps to avoid amplification of the unwanted side product (Figure 1.7) [113].

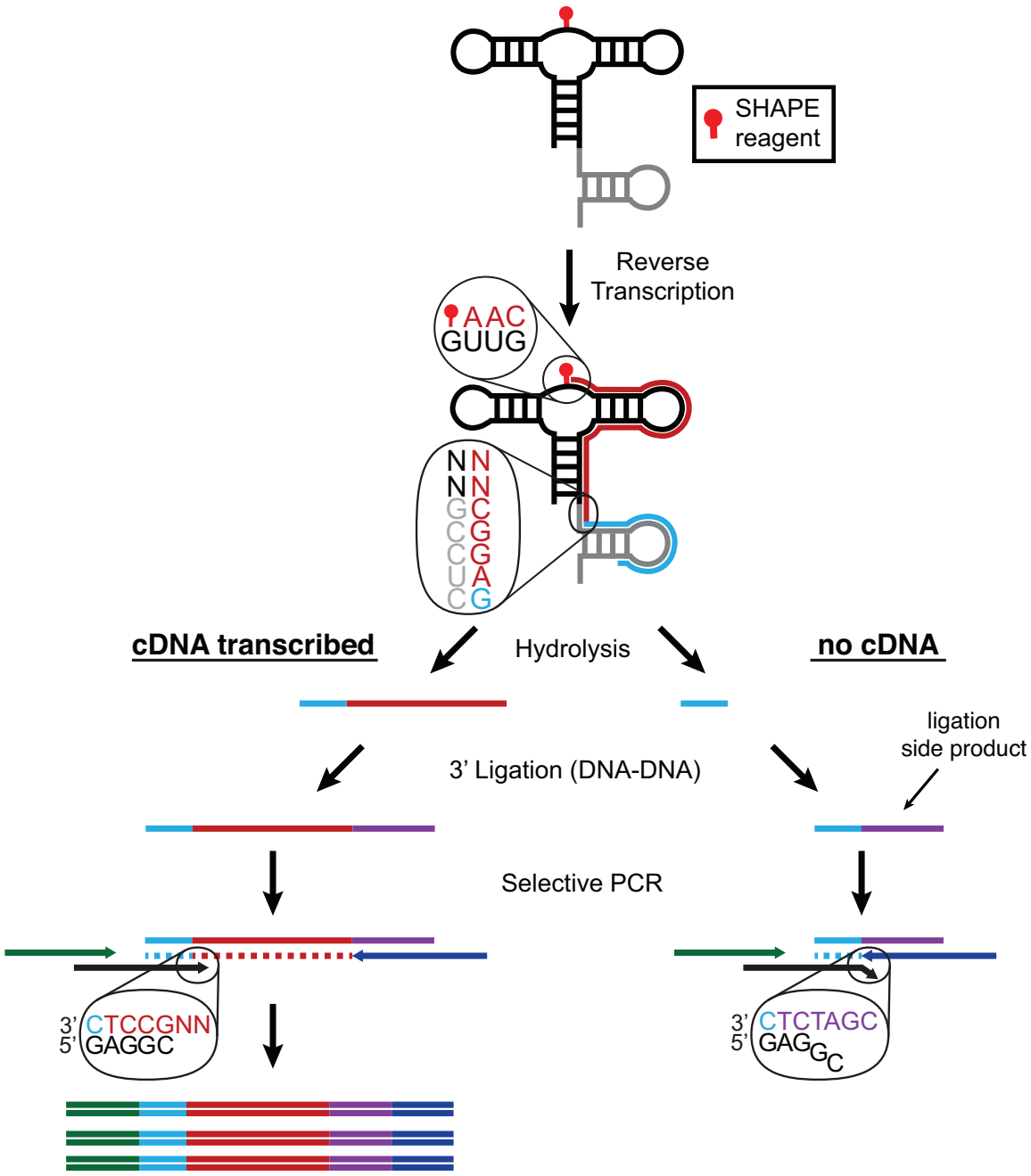


Figure 1.7: Selective PCR prevents side product amplification. When cDNA (red) yields are low, a high concentration of unextended RT primer (blue) remains that can be ligated to the DNA adapter (purple) to create an unwanted ligation side product. The introduction of a selective PCR step prevents amplification of the unwanted side product by creating a 3′ mismatch in the selection primer (black) if no cDNA was transcribed. If cDNA sequence is present, the PCR exponentially amplifies the ligated products. Adapted from Watters *et al.* [123].

Combining in-cell 1M7 modification with the new selective PCR steps allowed us to obtain highly reproducible in-cell SHAPE-Seq data for the first time. We used the new technique to examine two RNA-RNA interacting translational regulators and noted that RNA loops involved in sRNA-mRNA binding tend to very highly reactive when their binding partner is absent. we also observed some of the first evidence that long dsRNAs are cut in the cell by an endogenous dsRNase and that synthetic RNA structures look similar *in vitro* and in the cell [113].

The techniques developed in Chapter 3 will help shift the paradigm of RNA engineering to use more in-cell structural data to help guide synthetic RNA regulator design [26, 129]. Also, elements of the selective PCR method developed in Chapter 3 were used to improve *in vitro* SHAPE-Seq techniques, as discussed in Chapter 4.

1.6.3 Chapter 4: Further improvement of *in vitro* SHAPE-Seq

Despite the improvements made to *in vitro* SHAPE-Seq v2.0, the level of unwanted ligation side product (discussed above) was still very high, resulting in many lost reads during sequencing. In this portion of the thesis work, we add the selective PCR steps developed for in-cell SHAPE-Seq (Chapter 3) [113] to the SHAPE-Seq v2.0 technique, upgrading it to v2.1. With the additional selection steps, we observed a 10-40 fold reduction in the amount of unwanted side product sequenced. we also compared v2.1 to v2.0 and confirmed that the additional changes do not affect the structural probing results. Last, we discuss SHAPE-Seq techniques as a whole and a number of potential limitations, improvements, considerations associated with the SHAPE-Seq [123].

Chapter 4 also contains step-by-step instructions for performing both in-cell

The work described in this chapter includes contributions from Angela M Yu, Eric J. Strobel, Alex H. Settle, and Julius B. Lucks.

SHAPE-Seq and SHAPE-Seq v2.1, including a brief tutorial on restrained computational RNA folding with SHAPE reactivities [123]. The protocol improvements comprising SHAPE-Seq v2.1 were key to developing the cotranscriptional SHAPE-Seq technique described in Chapter 5.

1.6.4 Chapter 5: Cotranscriptional folding of a riboswitch at nucleotide resolution

The effects of cotranscriptional folding on RNA structure (Figure 1.3) are relatively unexplored because of the experimental difficulty of capturing structures of intermediate folding states [62]. Computational analysis has provided some insight [65, 74, 77], but lacks strong support from experimental data. Currently, cotranscriptional folding experiments are mainly limited to single molecule pulling experiments [79–81] and simple nuclease probing experiments of individual paused complexes [78, 83, 130]. However, these methods are time consuming and exhibit poor resolution. Obtaining structural information for intermediate structural states of transcribing RNAs is highly relevant to RNA biology, and motivates Chapter 5 of this thesis work.

In Chapter 5, we discuss the development of cotranscriptional SHAPE-Seq, which combines all of the technical advances of SHAPE-Seq v2.1 [123] with RNAP arrest to chemically probe active RNAP elongation complexes that are roadblocked by a catalytically dead EcoRI restriction enzyme [131]. By probing all of the structural intermediates of an RNA at once, cotranscriptional SHAPE-Seq captures single nucleotide resolution structural information for an entire RNA folding pathway.

The work described in this chapter includes contributions from Eric J. Strobel, Angela M Yu, and Julius B. Lucks.

We used the cotranscriptional SHAPE-Seq to examine the folding pathway of two RNAs, the signal recognition particle (SRP) RNA from *E. coli* and the crcB fluoride riboswitch from *B. cereus*. For both RNAs, we were able to observe the structural rearrangements that occur during transcription. As detailed in Chapter 5, we also determined the mechanism by which the fluoride riboswitch chooses between one of two folding paths that depend on the local concentration of the fluoride ion. Cotranscriptional SHAPE-Seq is an exciting new technique and promises to provide rapid insight into the effects of cotranscriptional folding on RNA structure, especially for RNA regulators.

1.6.5 Chapter 6: Measuring the cotranscriptional folding pathway of the pT181 transcriptional attenuator

With its basic function elucidated in 1985, the pT181 attenuator became the first known example of an RNA-regulated RNA transcriptional regular [54]. A series of enzymatic cleavage and deletion/mutations analyses then followed to try to understand the mechanism of attenuation in the RepC mRNA of pT181 [53, 132–134]. To date, it is generally understood that the attenuator can fold into a terminator (containing an intrinsic terminator sequence) or antiterminator structure, what the two structures likely look like, and where the antisense RNA binds (Figure 1.8) [53, 133]. However, the exact mechanism that links antisense RNA binding to the formation of the terminator or the antiterminated structure was never determined.

The work described in this chapter includes contributions from Katherine A. Berman, Alexandra M. Westbrook, Jane. B. Liao, Ruize Zhuang, Alex H. Settle, and Julius B. Lucks.

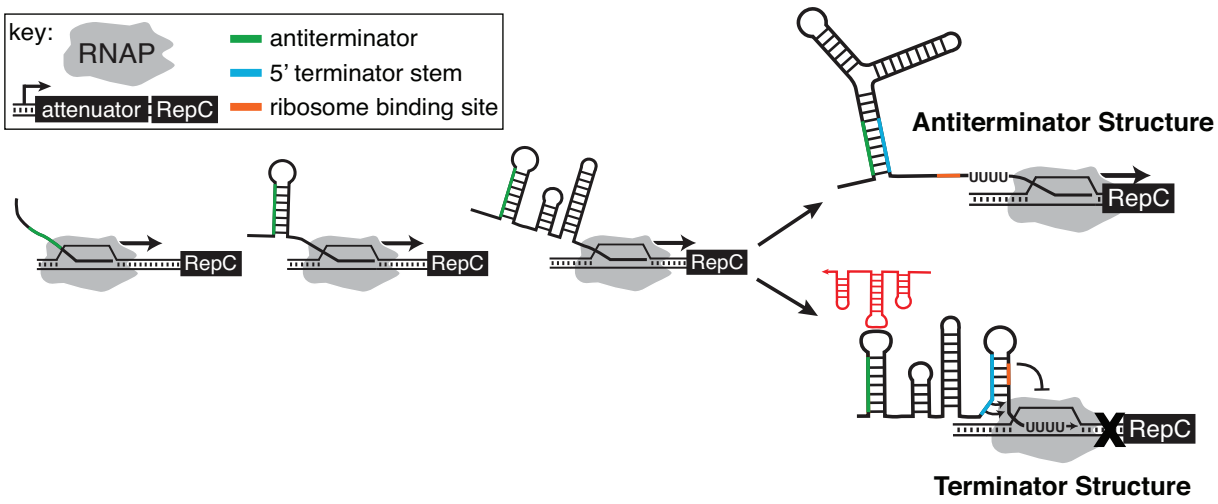


Figure 1.8: Attenuation mechanism from the pT181 plasmid. Transcription and translation of the pT181 replication protein (RepC) is controlled by an attenuator sequence in the 5′ UTR of its mRNA [54]. The attenuator folds into one of two structures: an antiterminator structure that permits transcription/translation of the RepC ORF or a terminator structure that prevents transcription/translation of RepC. The terminator structure is favored when an antisense RNA complementary to the beginning of the 5′ UTR is present.

One reason the mechanism could never be fully established for pT181 antitermination is the fact that cotranscriptional folding was not taken into account in previous experiments [133]. As an RNA transcriptional regulator, the pT181 attenuator must choose a structural form (terminator or antiterminator) before RNAP transcribes through the intrinsic terminator sequence, meaning that cotranscriptional folding must be important to the function of the attenuator. In fact, the study performed by Brantl and Wagner [133] likely contains flaws due to use of equilibrium refolding during the enzymatic cleavage assays, which would likely result in the terminated form based on our observations of other equilibrium refolded RNAs (Chapter 5).

The goal of the work presented in Chapter 6 is to apply the recently developed cotranscriptional SHAPE-Seq technique (Chapter 5) to determine the mechanism of

antitermination in the pT181 attenuator, over 30 years after its discovery. We compare cotranscriptional structural data of the attenuator with and without antisense and observe that few rearrangements are involved in antiterminator refolding. We also perform a series of mutations and deletions to support the conclusions drawn with cotranscriptional SHAPE-Seq. The mutants and deletions were also used work toward a minimized version of the attenuator for simplifying future engineering efforts with the pT181 attenuator, which have already yielded many successful designs [55, 135–137]. Ultimately, the work discussed in Chapter 6 resulted in a better mechanistic understanding of the pT181 attenuator and a smaller functioning version, both of which we expect to improve future engineering efforts to create RNA-only networks [26, 55, 135–138].

1.6.6 Chapter 7: Design of a CRISPR sgRNA Derepressor Using Insight from SHAPE-Seq

In the CRISPRi regulation system, a small guide RNA (sgRNA) targets the dCas9 protein to a segment of dsDNA that it binds tightly to, preventing RNAP initiation or elongation [56, 57]. The sgRNA:dCas9 RNA-protein complex is highly specific for its target, determined by the 5′ sequence of the sgRNA, and binds very strongly with a very slow dissociation constant [60]. Other than the requirement for an adjacent 'NGG' motif, target specificity is entirely determined by a 20 nt RNA sequence at the 5′ end of the sgRNA [139]. Thus, switching the DNA target sequence is a facile process, only requiring mutating the 5′ end of the sgRNA, which can be changed to any sequence. The ease of switching targets, of which there are potentially many in a given DNA

The work described in this chapter includes contributions from Jane B. Liao, Timothy R. Abbott, and Julius B. Lucks.

sequence, and the high repression performance of CRISPRi has prompted its rapid entry into the synthetic biology space [26, 56, 57, 140].

One drawback of CRISPRi regulation or synthetic biology applications, however, is the slow dissociation rate [60]. Because targeting is entirely RNA dependent, it is theoretically possible to construct RNA-only circuitry (with a dCas9 cofactor) with CRISPRi to take advantage of the fast degradation rate of RNA to yield biological circuits with fast dynamics [138, 141, 142]. However, because the sgRNA:dCas9 complex is highly stable in the cell and is slow to release its target [60], no dynamical advantage can be gained with current CRISPRi methodology as computational cycles would be slow to reset. Our goal in Chapter 6 is to change this paradigm by creating a sgRNA derepressor to free the sgRNA:dCas9 complex from its DNA target and restore the fast dynamics of an RNA-only circuit. Combined with the strong repression of CRISPRi, a sgRNA derepressor could produce an organism-independent, binary regulatory mechanism with very low background expression, a dream of the RNA synthetic biology community that has still yet to materialize [26].

In brief, our design approach was to first study a canonical sgRNA and a series of mutants with in-cell SHAPE-Seq [113, 123] to determine what structural elements of the non-targeting structures of the sgRNA were important for proper dCas9 loading and DNA targeting. Then, we added RNA regulatory structures to regions of the sgRNA deemed unimportant for dCas9 binding. Ultimately, we arrived at a split sgRNA design that exhibited levels of gene expression activation with very low background relative to other RNA regulatory mechanisms (Chapter 6).

1.6.7 Chapter 8: Structural Analysis of Cucumber Mosaic Virus RNA3

The cucumber mosaic virus (CMV) is a tripartite positive-sense RNA plant virus with one of the widest host ranges known [143]. Within its genome, a number of RNA structures critical for viral replication have been identified, including a highly conserved tRNA-like structure at the 3′ end of each genome segment that serves as the negative strand promoter region [144]. However, much of the CMV genome remains unexplored structurally.

In Chapter 8, we use SHAPE-Seq v2.1 [123] to delve into the structural features of RNA3, one of three genomic RNAs of CMV, both *in vitro* and in infected cell lysates. RNA3 contains three UTRs: a short 5′ UTR containing conserved repeat sequences, a 3′ UTR containing the tRNA-like structure, and an intergenic region (IGR) that separates the two ORFs of RNA3 [144]. We observe SHAPE reactivity data consistent with the tRNA-like structure and differences between the *in vitro* and lysate experiments that indicate binding of the viral replicase. We also examine the rest of the RNA3 sequence and propose possible secondary structures based on restrained secondary structure predictions with RNAstructure [101, 102]. we expect that our structural data will help inform the further discovery of RNA structures in RNA3 important to the CMV replication cycle.

The work described in this chapter includes contributions from Jeremy R. Thompson, Keith L. Perry, and Julius B. Lucks.

1.6.8 Chapter 9: Structural Features of Protein Binding on RNase P and its Substrates

In the last chapter of this thesis work, we use SHAPE-Seq v2.1 [123] to characterize two fundamentally different versions of the RNase P endoribonuclease. RNase P cleaves the 5′ leader sequence from precursor tRNAs (pre-tRNAs) as part of the tRNA maturation process [145] and is critical to cell viability in almost every species [146–148]. One version we studied, a ribozyme, is highly conserved across all species, although it takes on a few different forms [149]. The other version is a proteinaceous RNase P (PRORP) found in eukaryotes [150, 151]. Both versions perform the same cellular function.

In Chapter 9, we examine the reactivity map of two archaeal RNase P ribozymes and identify binding sites for one of their protein cofactors, L7Ae [152, 153]. we also examine the PRORP from the viewpoint of the tRNA to identify where PRORP binds pre-tRNAs prior to cleavage. In both cases, we find that distinct RNA structural motifs are involved in protein recognition. Relative to classic biochemical approaches, SHAPE-Seq was able to locate the protein binding sites quickly and easily, providing biologists a new and easily approachable tool for dissecting RNA-protein interactions.

The work described in this chapter includes contributions from Venkat Gopalan,Tien-Hao Chen, Lien Lai, Stella Lai, Ila Marathe, and Julius B. Lucks.

CHAPTER 2

SHAPE-SEQ 2.0: SYSTEMATIC OPTIMIZATION AND EXTENSION OF HIGH-THROUGHPUT CHEMICAL PROBING OF RNA SECONDARY STRUCTURE WITH NEXT GENERATION SEQUENCING

2.1 Abstract

RNA structure is a primary determinant of its function, and methods that merge chemical probing with next generation sequencing have created breakthroughs in the throughput and scale of RNA structure characterization. However, little work has been done to examine the effects of library preparation and sequencing on the measured chemical probe reactivities that encode RNA structural information. Here, we present the first analysis and optimization of these effects for selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). We first optimize SHAPE-Seq, and show that it provides highly reproducible reactivity data over a wide range of RNA structural contexts with no apparent biases. As part of this optimization, we present SHAPE-Seq v2.0, a 'universal' method that can obtain reactivity information for every nucleotide of an RNA without having to use or introduce a specific reverse transcriptase priming site within the RNA. We show that SHAPE-Seq v2.0 is highly reproducible, with reactivity data that can be used as constraints in RNA folding algorithms to predict structures on par with those generated using data from other SHAPE methods. We anticipate SHAPE-Seq v2.0 to be broadly applicable to understanding the RNA sequence-structure relationship at the heart of some of life's most fundamental

The work described in this chapter includes contributions from David Loughrey, Alexander H. Settle, and Julius B. Lucks. This work was originally published in *Nucleic Acids Research* and has been reproduced here with permission from Oxford University Press. David Loughrey*, Kyle E. Watters*, Alexander H. Settle and Julius B. Lucks, SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing, *Nucleic Acids Res*, 2014, **42**, (21), e165.

processes.

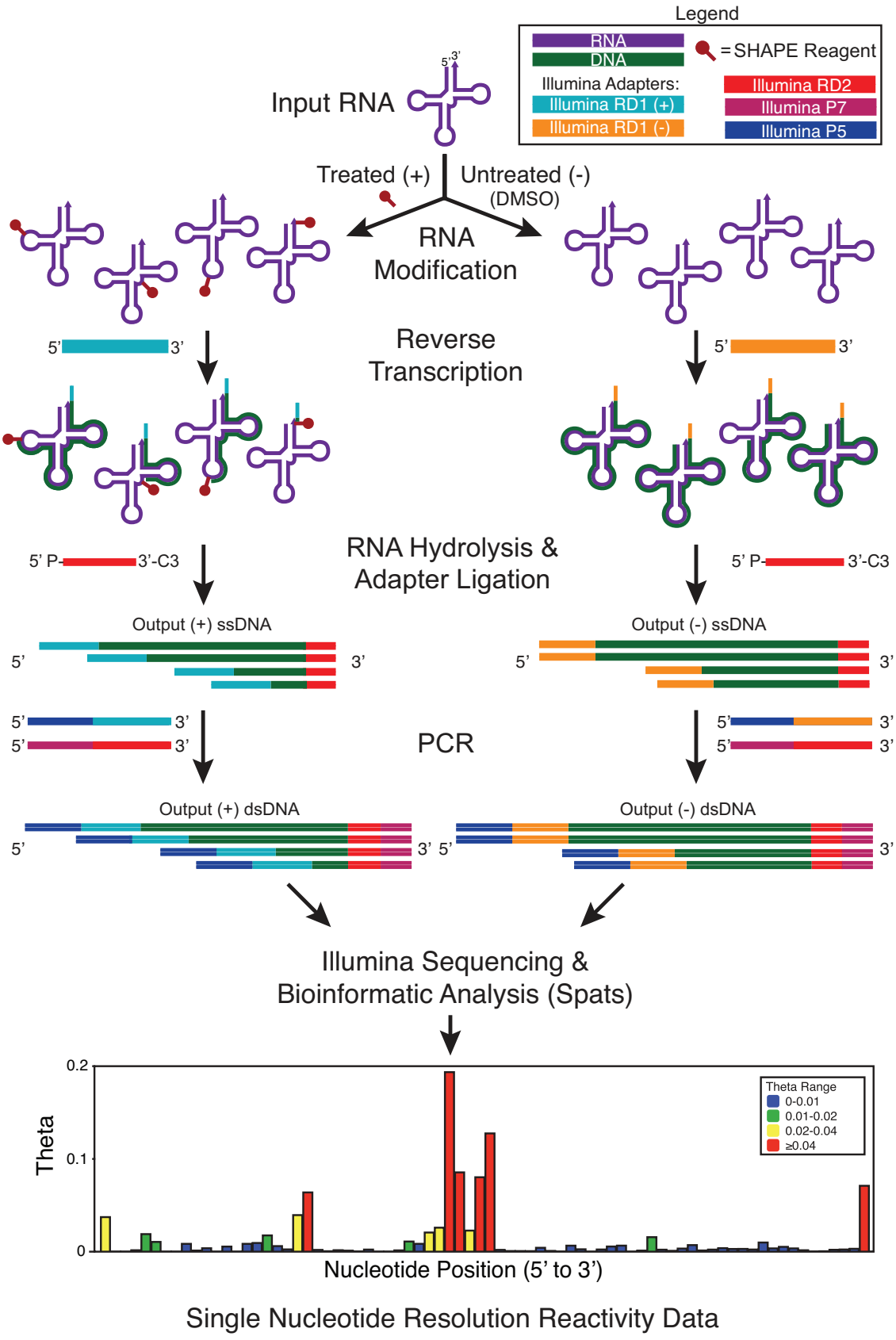
2.2 Introduction

RNAs play diverse functional roles in many natural cellular processes [154] and are being increasingly engineered to control these processes in many synthetic biology and biotechnology applications [24]. This diverse function of RNA is intimately connected to its ability to fold into intricate structures. Recently, high-throughput techniques that combine nuclease digestion [119, 120, 155] or chemical probing [107, 108, 116, 156] with next-generation sequencing have started to shed new light on the sequence/structure relationship of RNA. Because of the inherent multiplexing and enormous throughput offered by sequencing-based approaches, these techniques are providing some of the first 'genome-wide' snapshots of RNA structure [116, 117] - effectively bringing RNA structural biology into the 'omics' era [157].

Of these techniques, those that favor chemical probing over nuclease digests show the most promise because of the inherent versatility [86], higher resolution, and *in vivo* accessibility [116–118, 126, 158] of many chemical probes. Several such techniques have been developed (SHAPE-Seq [107, 108], DMS-Seq [86, 116, 117], MAP-Seq [156] that each follow the same general protocol consisting of: i) structure-dependent modification of the RNA *in vitro* or *in vivo*; ii) reverse transcription (RT) of the modified RNA into a cDNA pool whose length distribution reflects the location of modifications; iii) sequencing library construction, involving the addition of platform-specific adapter sequences to the cDNA pool, optional amplification with PCR, and quality control assessment steps; iv) sequencing of the library; and v) bioinformatic processing of sequencing reads and calculation of reactivity spectra for each RNA (Figure 2.1).

While proving to be powerful, these sequencing-based techniques are complex and involve many more steps, including ligation and PCR, than approaches that use electrophoretic analysis (Figure A.1) [105, 106, 159]. Very little work has been done to evaluate the impact of these extra steps.

Figure 2.1: In SHAPE-Seq [107, 108], RNAs are modified with chemical probes, such as 1M7 [104], or any probe that covalently modifies the RNA in a structure-dependent fashion [86]. RT of the RNAs creates a pool of cDNAs, whose length distribution reflects the distribution of modification positions. Control reactions are performed to account for RT fall-off at unmodified positions. RT primer tails contain a portion of one of the required Illumina sequencing adapters, while the other is added to the 3′ end of each cDNA through a single-stranded DNA ligation. A limited number of PCR cycles are used to both amplify the library and add the rest of the required adapters prior to sequencing. A freely available bioinformatic pipeline Spats [97, 98, 108] is then used to align sequencing reads, correct for biases due to RT-based signal decay [97, 98] and calculate reactivity spectra for each RNA. See Figures A.2, A.3, A.4 for protocol details.



Single Nucleotide Resolution Reactivity Data

In this work, we systematically analyze and optimize SHAPE-Seq, and present a new version, SHAPE-Seq v2.0, that can obtain reactivity information for every nucleotide of an RNA without requiring an internal RT priming site. We start by analyzing SHAPE-Seq in the context of a panel of RNAs used in previous benchmarking of chemical probing techniques [102, 160, 161]. Specifically, we systematically investigate steps of the SHAPE-Seq protocol that differ from more traditional methods that could affect measured reactivity data, including sequence context effects of adapter ligation, adapter ligation conditions, RT primer modifications, and PCR. During this process, we optimize several of these steps, and show that they do not appear to be a source of differences between SHAPE-Seq and electrophoresis-based SHAPE analysis. We also show that SHAPE-Seq is highly reproducible and report replicate reactivity spectra for each RNA in the panel.

Finally, we present SHAPE-Seq v2.0 and show that it recapitulates SHAPE-Seq reactivity spectra, but without requiring a specific RT priming site to be present on the target RNAs. SHAPE-Seq v2.0 significantly expands the capability of SHAPE-Seq by allowing it to be performed through a 'kit'-like protocol independent of the RNAs studied. We also show that SHAPE-Seq v2.0 reactivity data can be readily incorporated into RNA structure prediction algorithms to give experimentally-constrained predicted folds that are highly accurate, and on par with traditional SHAPE constrained folding [102].

2.3 Materials and Methods

2.3.1 RNA Preparation

For SHAPE-Seq v1.0, RNAs were generated through in vitro transcription reactions with T7 RNA polymerase. DNA templates consisted of a preceding 17-nucleotide T7 promoter, an optional 14-nucleotide 5′ structure cassette sequence [92], the desired RNA coding sequence, and an optional 43-nucleotide 3′ structure cassette sequence [92, 108] (Table A.1). DNA templates were generated by PCR [1 mL; containing 20 mM Tris (pH 8.4), 50 mM KCl, 2.5 mM $MgCl_2$, 200 μM each dNTP, 500 nM each forward and reverse primer, 5 pM template, and 0.025 U/μL Taq polymerase; denaturation at 94 °C, 45 s; annealing 55 °C, 30 s; and elongation 72 °C, 1 min; 34 cycles]. The PCR product was recovered by ethanol precipitation and resuspended in 150 $MgCl_2$L of TE [10 mM Tris (pH 8.0), 1 mM EDTA]. Transcription reactions (1.0 mL, 37 °C, 1214 h) contained 40 mM Tris (pH 8.0), 20 mM $MgCl_2$, 10 mM DTT, 2 mM spermidine, 0.01% (vol/vol) Triton X-100, 5 mM each NTP, 50 μL of PCR-generated template, 0.04 U/μL SuperaseIN RNase Inhibitor (Ambion), and 0.1 mg/mL of T7 RNA polymerase. The RNA products were purified by denaturing polyacrylamide gel electrophoresis (8% polyacrylamide, 7 M urea, 35 W, 3 h), excised from the gel using an appropriately placed sacrifice lane for UV shadowing, and recovered by passive elution and ethanol precipitation. The purified RNA was resuspended in 50 μL TE, and concentrations were measured with the Qubit fluorimeter. All of the RNAs with the flanking structure cassettes contained a unique four-nucleotide bar code to multiplex SHAPE-Seq v1.0 experiments as described previously [108] (Table A.1). In general, QuSHAPE and SHAPE-Seq v1.0 experiments were performed on RNAs containing the optional structure cassettes, while SHAPE-Seq v2.0 experiments were performed on RNAs without

these cassettes (Table A.1).

For SHAPE-Seq v2.0 RNAs, each RNA sequence was altered to begin with GG, and cloned between the T7 RNA promoter and Hepatitis δ (HepD) antigenomic ribozyme and PCR amplified. The resulting templates were transcribed in vitro using T7 RNA polymerase as above, and purified by standard gel excision methods [162]. For the HepC IRES and cyclic-di-GMP Riboswitch, standard run-off transcription was performed without the ribozyme to obtain higher yields. A table of all RNA sequences used in this study can be found in Table A.1.

2.3.2 RNA Modification

For the initial benchmarking and optimization studies, all RNAs were folded and modified individually with 1M7 (6.5 mM, final) in batches. The RNAs (50 pmol in 60 μL) were denatured by heating at 95 °C for 2 min, and snap-cooled on ice for 60 s before the addition of 30 μL of folding buffer. The sample was refolded in the 1X folding buffer (10 mM $MgCl_2$, 100 mM NaCl and 100 mM HEPES (pH 8.0)), in a total volume of 90 μL for 20 min at 37 °C. These reaction volumes were split and added to 5 μL 1M7 (10X, 65 mM in dry DMSO) and 5 μL dry DMSO to form the (+) and (-) reactions, respectively. Reactions were complete after 70 s at 37 °C. RNAs were recovered by the addition of 50 μL of water, 10 μL of 3 M NaOAc, 1 μL of 20 mg/mL glycogen and 300 μL of 100% ethanol, followed by incubation at -80 °C for 30 min and centrifugation (15k rpm) at 4 °C for 30 min. Multiple batches of the same RNAs were combined into an overall stock of 700 pmol RNA (350 pmol unmodified (+), 350 pmol modified (-), each in 70 μL TE), and stored at -20 °C until use ($<$3 weeks). For v1.0 replicate experiments, RNAs were folded as described above. For v2.0, RNAs were folded in 20 pmol batches in 12 μL,

with the addition of the folding buffer bringing the total volume to 18 μL. In addition, only 1 μL 1M7 and 1 μL dry DMSO were used for the (+) and (-) reactions and recovery required the addition of 90 μL H2O instead of 40 μL. Buffer conditions and ligand concentrations for each RNA are listed in Table A.2.

2.3.3 QuSHAPE

For QuSHAPE experiments, 10 pmol from the modified (+) and unmodified (-) batches of RNA were suspended in 10 μL of water. Two sequencing lanes were also established with 10 pmol of purified RNA in 9 μL of water. RT reaction mixtures were prepared with the addition of 3 μL of 0.3 μM Vic-labeled [(+) and one of the sequencing lanes, usually ddT] or Ned-labeled [(-) and the other sequencing lane, usually ddA] reverse transcription primer, with sequence GAACCGGACCGAAGCCCG. Primers were annealed following denaturation at 95 °C for 2 min and 65 °C for 5 min and immediate snap-cooling. Primer extension reactions were preformed by the addition of 1 μL of Superscript III, 4 μL of 5X Superscript First Strand Buffer, 0.4 μL of dNTPs at 10 mM each (dATP, dCTP, dTTP, dITP), 1 μL of 0.1 M DTT, and 0.6 μL of water. Following previous capillary electrophoresis methods, dITP was used instead of dGTP to reduce band compression and increases resolution of primer extension products by capiliary electrophoresis [104]. 1 μL of 10 mM pertinent di-deoxy stocks (usually ddATP or ddTTP) was added to the appropriate sequencing lanes as well. The reaction mixtures (total volume of 20 μL) were incubated at 45 °C for 1 min, 52 °C for 25 min, and 65 °C for 5 min. Then 4 μL of 50 mM EDTA (pH 8.0) was added to each sample, and oppositely labeled primers (i.e modified and ddA; unmodified and ddT) combined, precipitated with ethanol, resuspended in 10 μL of Hi-Di formamide and resolved on an Applied Biosystems 3730xl capillary electrophoresis instrument. Raw capillary electrophore-

sis traces were processed using QuSHAPE software as described in Karabiber *et al.* [106]. QuSHAPE reactivities were then converted to QuSHAPE θ's by dividing by a normalization factor so that they summed to 1 over the range of nucleotides for which reactivity data was obtained.

2.3.4 SHAPE-Seq Reverse Transcription

Reverse transcription (RT) conditions differed based on the particular library construction strategy. The details of reverse transcription primers and adapter configurations for SHAPE-Seq library preparation strategies can be found in Figures A.2-A.4. For SHAPE-Seq v1.0 (Figures A.2 and A.3), the general procedure for reverse transcription was carried out following the primer extension protocol in Mortimer, *et al.* [108]. The total amount of primer used in primer extension reactions was 9 pmol (3 μL of 3 μM primer), with the RNA concentration being 50 pmol in 10 μL. Primers were annealed by incubation at 95 °C for 2 min and at 65 °C for 5 min. Primer extension reactions were performed by the addition of 200 U of Superscript III, 4 μL of 5X Superscript First Strand Buffer, 0.4 μL of dNTPs at 10 mM each (dATP, dCTP, dTTP, dGTP), 1 μL of 100 mM DTT, and 0.6 μL of water. The reaction mixtures (total volume of 20 μL) were incubated at 45 °C for 1 min, 52 °C for 25 min, and 65 °C for 5 min. After primer extension, RNA was hydrolyzed by adding NaOH (1 μL, 4 M) and incubating for 5 min at 95 °C. cDNAs were ethanol precipitated and resuspended in 71 μL of nuclease-free water.

For SHAPE-Seq v2.0 (Figure A.4), RNAs were not purified after the modification step. Instead, a linker sequence was added to the 3' end of the RNA template via 5'-App ligation by adding 6.5 μL 50% PEG 8000, 2 μL 10X T4 RNA Ligase Buffer (NEB), 1 μM 5' App IDT miRNA linker 2 (5'App-CACTCGGGCACCAAGGAC-3'ddC), and 0.5

μL T4 RNA Ligase, truncated KQ (NEB) directly to the modified RNA. Samples were incubated overnight at room temperature, recovered by ethanol precipitation, and re-suspended in 10 μL of nuclease-free water. Reverse transcription was then carried out as above using 1.5 pmol RT primer (3 μL at 0.5 μM) complementary to the linker sequence, and containing flanking Illumina adapter sequence and custom internal barcodes to create unique 3'end alignments and increase randomness during Illumina sequencing (Figure A.4 & Table A.3). Hydrolysis was performed the same way, but partially quenched with 1.5 μL of 1 M HCl. After EtOH precipitation, the cDNAs were dissolved in 22.5 μL nuclease-free water.

2.3.5 SHAPE-Seq Second Adapter Ligation

Adapter ligations differed based on the particular library construction strategy (Figures A.2-A.4). In all SHAPE-Seq v1.0 cases, adapter sequences were ligated to each cDNA using a ssDNA ligase (CircLigase, Epicentre Biotechnologies) [100 μL; 50 mM MOPS (pH 7.5), 10 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 0.05 mM ATP, 2.5 mM $MnCl_2$, 5 μM adapter, and 200 U ligase] and incubating for 6 h at 68 °C in a thermal cycler. For SHAPE-Seq v2.0, the ligation was performed in 30 μL by mixing 50 mM MOPS (pH 7.5), 10 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 0.05 mM ATP, 2.5 mM $MnCl_2$, 1.67 μM adapter (Figure A.4), and 100 U ligase, incubated at 60 °C for 2 h. Separate ligation reactions were carried out for the (+) and (-) cDNA library pools. The ligation reactions were stopped by heating to 80 °C for 10 min, recovered by ethanol precipitation, and resuspended in 20 μL of nuclease-free water. Excess adapter was removed using Agencourt Ampure XP beads following the manufacturers protocol, eluting with 20 μL TE.

2.3.6 SHAPE-Seq PCR, Library QC, and Sequencing

Ligated libraries were then used as inputs into PCR reactions using 6, 9, 12 or 20 cycles of PCR amplification as indicated in Results. PCR primers contained sequences required for Illumina sequencing and index multiplexing as indicated in Figures A.2-A.4. A 50 μL PCR reaction contained 2.5 μL of cDNA template, 1 μL of 100 μM forward and reverse primers, 1 μL of 10 mM dNTPs, 10 μL 5X Phusion Buffer, 33.5 μL water, and 1 U Phusion DNA polymerase (NEB). Multiple reactions were made together and split before the pertinent number of PCR amplification cycles. PCR reactions were cleaned up with Agencourt Ampure XP beads following the manufacturers protocol, eluting with 20 μL TE. No direct size selection was performed on the resulting adapter-ligated library. Libraries were assayed for quality in one of two methods: i) using an Agilent Bioanalyzer 2100 high-sensitivity DNA chip to compare 9- and 12-cycle amplification, looking for characteristic peaks and peak enrichment as described in Mortimer *et al.* [108], or ii) using fluorescently labeled PCR primers to generate fragments to be analyzed by capillary electrophoresis, looking for the same features.

The 9-cycle PCR amplification products (unless otherwise indicated in Results) were then sequenced on an Illumina MiSeq or HiSeq platform following the manufacturers standard cluster generation and sequencing protocols. For runs that consisted of multiple RNAs, (+) and (-) channels were balanced for molarity, and loaded to reach a final concentration of at least 2 ng/μL. This was then diluted according to the standard Illumina sequencing protocols.

2.3.7 SHAPE-Seq Data Analysis

Fastq files generated from the Illumina sequencing process were analyzed using the freely-available software package Spats (`spats.sourceforge.net`), as described in [97, 98, 108]. Spats takes paired-end fastq files and performs bioinformatics read alignment and a maximum-likelihood-based signal decay correction [97, 98, 107, 108] to calculate SHAPE-Seq θ values for each nucleotide of an RNA. In SHAPE-Seq experiments, reactivity data are presented as $\Theta=\{\theta_i; i = 1...L\}$, which is a probability distribution over the length of an RNA, L, representing the probability, θ_i, that each nucleotide, i, is modified by the modification reagent [97, 98]. SHAPE-Seq θ's are similar to more traditional SHAPE reactivity data (typically referred to as a set of reactivity numbers for each nucleotide, $\{r_i\}$), except that θ's are constrained to sum to 1 over the length of the RNA since they represent a probability distribution, i.e. $\sum_{i=1}^{\infty} \theta_i = 1$. Thus θ's are independent of scale factors typically used to define SHAPE reactivities [101, 160, 161, 163], and can be rigorously calculated from the observed (+) and (-) fragment distributions in a SHAPE-Seq experiment [97, 98, 108].

For this work, we used Spats v0.8.0, which contains an updated adapter trimming algorithm. Once installed, a typical spats command was executed by entering:

python adapter_trimmer.py --A-b-sequence=<second_adapter_sequence> --A-t-seq uence=<first_adapter_sequence> –read-len=35 R1.fastq R2.fastq RNA_targets.fa

spats –num-mismatches 0 -o Output RNA_targets.fa RRRY YYYR combined_R1.fastq combined_R2.fastq

where <second_adapter_sequence> is the sequence of the second adapter, <first_a

dapter_sequence> is the sequence of the first adapter, R1.fastq and R2.fastq are the fastq sequencing files, and RNA_targets.fa is the FASTA formatted file containing the RNA sequences under study. Spats outputs text files containing (+) and (-) fragment counts and θ's, for each position in each RNA in the targets file. Fragment distributions were calculated by dividing the fragment counts at each position by the total number of fragments observed in the channel, so that the fragment distribution summed to 1 over the length of the RNA.

2.3.8 Computational Modeling

The *Fold* executable of the RNAstructure [164] software package (version 5.5) was used to calculate SHAPE-Seq-constrained RNA secondary structures. For each RNA, SHAPE-Seq v2.0 θ's were first converted into ρ values by multiplying by the length of the RNA, L (Equation (2.1)). Due to the inability to uniquely align reads containing a single nucleotide, L was one nucleotide less than the full length of the RNA used in experiments. The flags "-sh", "-sm", and "-si" were used to input the SHAPE-Seq ρ data file, slope m, and intercept b, respectively. The parameters m and b were used to define the pseudo free energy that was used in the minimum free energy structure calculation [101, 160, 163, 164]. The sensitivity and positive predictive value (PPV) of each predicted RNA structure was evaluated using the RNAstructure [164] *Scorer* executable (version 5.5), using the established crystallographic secondary structure as the accepted structure [102, 160, 161]. The entire RNA sequence was used in these calculations except for nucleotides at the very 5′ end of the RNAs that were included for RNA synthesis using T7 polymerase. A table of RNA sequences used can be found in Table A.1.

2.4 Results

2.4.1 Investigation and Optimization of SHAPE-Seq Library Preparation

There are two distinct protocol steps associated with sequencing library preparation that distinguish SHAPE-Seq from SHAPE analyzed by capillary electrophoresis (Figures 2.1 and A.1): adapter ligation and PCR. Here, we sought to investigate if these steps impact the measured fragment distributions and reactivities for a set of RNAs that have been used in other recent SHAPE benchmarking experiments [102, 107, 108, 160, 161]. The starting panel, chosen for the abundance of crystallographic and SHAPE structural data available in the literature, consisted of the RNase P specificity domain from *B. subtilis*, unmodified $tRNA^{phe}$ from *E. coli*, the P4-P6 domain of the Tetrahymena group I intron ribozyme, the 5S rRNA from *E. coli*, and the aptamer domain from the *V. vulnificus* adenine riboswitch (Table A.1).

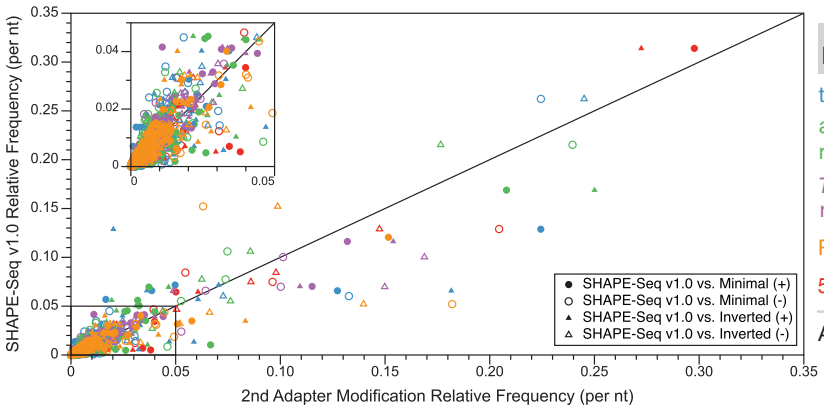
2.4.2 Adapter Ligation

One of the most distinct differences between SHAPE-Seq and capillary electrophoresis methods like QuSHAPE [106] is the ligation of a second adapter sequence required for Illumina sequencing (Figures 2.1 and A.1). The previously published SHAPE-Seq protocol (SHAPE-Seq v1.0) uses single-stranded DNA (ssDNA) ligation to add a 61 nt ssDNA oligonucleotide to the 3′ end of the cDNA required for sequencing. This oligo consists of the Illumina RD2 and P7 sequences used for priming sequencing and flow cell binding, respectively (Figure A.2) [108]. It was hypothesized that this ligation

step could introduce bias in SHAPE-Seq measurements due to sequence or structure-specific ligation efficiencies.

We began by comparing SHAPE-Seq v1.0 reactivity spectra to QuSHAPE reactivities for each panel RNA, using the same starting pools of modified (+) and unmodified (-) RNAs. Overall there was a strong degree of correlation between the two methods for each of the RNAs and we did not find any systematic differences in reactivity spectra that would suggest ligation was causing biases (Figure A.5).

We therefore tested two other adapter variants: a truncated 'Minimal' adapter (25 nt) (Figure A.2b) and an 'Inverted' layout where the RT primer tail and the adapter sequences were switched (Figure A.2c). These libraries tested whether the adapter length (Minimal), or sequence (Inverted) would contribute to potential ligation bias, while staying within the sequence constraints of the Illumina platform. After constructing and sequencing these libraries from the same modified and unmodified RNA pools, we compared cDNA fragment length distributions, since changes in ligation conditions only affect the sampling of the modified or unmodified RNA fragments (Figure 2.2).



R values:	vs. Minimal (+)	vs. Minimal (-)	vs. Inverted (+)	vs. Inverted (-)
tRNA	0.89	0.93	0.63	0.99
adenine riboswitch	0.91	0.96	0.93	0.98
Tetrahymena ribozyme	0.95	0.95	0.97	0.92
RNAse P	0.96	0.66	0.90	0.82
5S	0.98	0.93	0.99	0.99
All RNAs	0.93	0.89	0.89	0.93

Figure 2.2: A comparison of SHAPE-Seq v1.0 to adapter ligation variations. SHAPE-Seq libraries on the same pools of modified and unmodified RNAs were constructed and sequenced using the standard adapter configuration (v1.0), a shortened second adapter (Minimal), or an adapter layout where sequences were switched between RT primer tail and second ligation (Inverted) (Figure A.2). For each RNA and each configuration, the (+) and (-) fragment distributions (per nucleotide frequencies) are compared. Circles represent the v1.0 vs. Minimal comparison and triangles represent the v1.0 vs. Inverted comparison. Filled/open symbols are for (+)/(-) fragment distributions, respectively, and are color coded according to the RNA as indicated in the table, which contains Pearson correlation (R) values for the comparisons.

The aggregate Pearson correlation coefficients (R) across all RNAs in the panel show that these distributions are in good agreement, with R values ranging from 0.89-0.93. Comparisons between adapter variants for individual RNAs also showed a strong correlation, with values ranging from 0.82-0.99, except for two specific cases: RNase P v1.0 vs. Minimal (-) (R=0.66), and tRNA v1.0 vs. Inverted (+) (R=0.63) (Figure 2.2). These discrepancies for the RNase P case mostly arose from differences between the ends of the (-) distribution, though these differences do not cause major discrepancies in the overall calculation (R=0.92) (Figure A.6a). Like RNase P, the tRNA distributions only show major differences at the 5′ and 3′ ends, but these differences do cause a discrepancy in the calculation (R=0.68) (Figure A.6b).

While constructing these libraries, we also investigated the ligation conditions in general, including enzyme choice, ligation temperature, and specific blocking groups used to prevent adapter concatenation during ligation. Previous work by Kwok *et al.* demonstrated that there was much room for improvement in the ssDNA ligation step, and devised an alternate T4 DNA ligase-based method [165]. However, our concern that secondary structures could potentially cause uncharacterized ligation bias at the low temperatures required for T4 DNA ligase [165] led us to instead improve the CircLigase reaction originally used in SHAPE-Seq v1.0 [108]. We began by performing a time course assay on the SHAPE-Seq v1.0 ligation reaction. We ligated the 61 nt SHAPE-Seq v1.0 Illumina adapter (Figure A.2) to a 126 nt cDNA for varying amounts of time (Figure A.7) and found a saturation in ligated product after 1-2 hrs. We next compared CircLigase II to CircLigase I for potential improved ligation efficiency (Figure A.8). A two hour ligation of an RT primer to the Illumina 61 nt adapter at 68 °C was compared to ligation at 60 °C for each ligase. Both ligases performed better at 60 °C, with Circligase I at 60 °C being the optimal condition tested (Figure A.8).

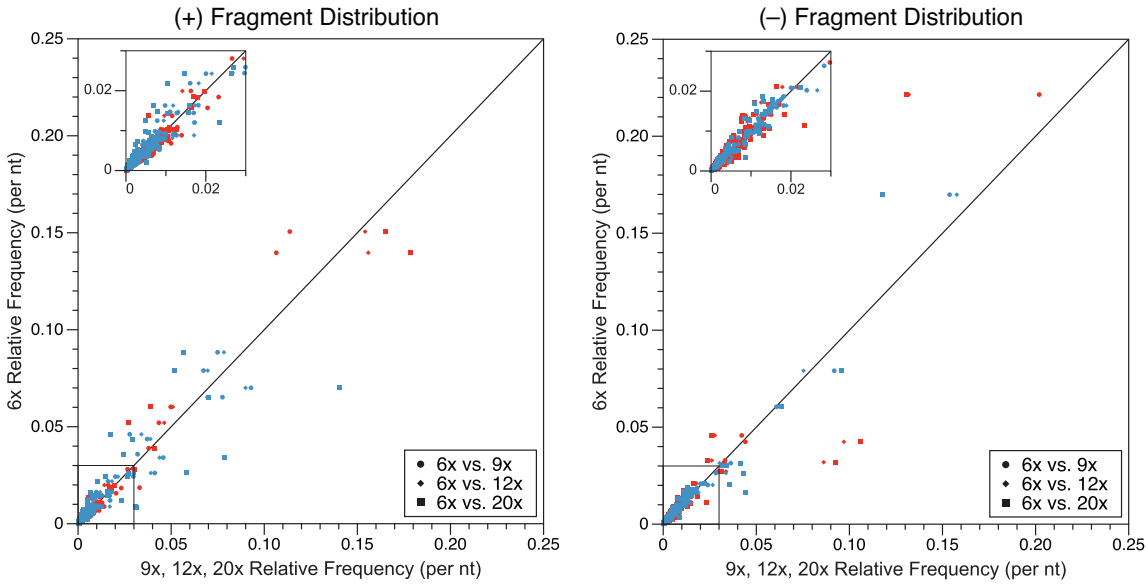
We also investigated blocking groups on the 5′ end of the RT primer, and 3′ end of the Illumina adapter (Figure A.2) for their importance in preventing concatemer formation during ligation (Figures A.9 and A.10). First, an unblocked or biotin-blocked RT primer was ligated to the 25 nt Minimal Illumina adapter (Figure A.9). Though the unblocked and blocked RT primers performed equivalently, we proceeded with the biotin-blocked primer (Figure A.9). We next tested 3′ di-deoxy-cytosine, phosphate, and 3-carbon spacer modifications to the Illumina second adapter designed to prevent adapter concatemerization (Figure A.10). We ligated the blocked adapter to an RT primer that was either 5′ blocked with biotin or unblocked, using CircLigase I. All 3′ modifications showed some degree of concatemer formation, with the phosphate modification producing the most concatemer (Figure A.10). We recommend the 3-carbon spacer modification, as it displayed lower amounts of concatemer formation and is less expensive than di-deoxy-cytosine.

2.4.3 PCR Amplification

Another distinct difference between SHAPE-Seq and SHAPE analyzed by capillary electrophoresis is the use of PCR to build and amplify SHAPE-Seq libraries before sequencing. PCR can be a powerful feature of SHAPE-Seq library construction, as it can amplify low signals [166], add custom barcodes or library indexes, or complete adapter sequences as in the Minimal library above (Figure A.2c). However, PCR could introduce a systematic bias in the SHAPE-Seq measurement since certain fragments can be preferentially amplified.

To test the effects of PCR on measured fragment distributions, we created SHAPE-Seq v1.0 libraries for tRNA and 5S rRNA using either 6, 9, 12 or 20 cycles of PCR

amplification before sequencing (Figure 2.3).



Frequencies:

R (6x vs.):	tRNA 9x	tRNA 12x	tRNA 20x	5S 9x	5S 12x	5S 20x
(+) Fragments	0.97	0.98	0.83	0.99	1.0	0.98
(–) Fragments	0.99	1.0	0.94	1.0	0.86	0.83

Thetas:

R (6x vs.):	tRNA 9x	tRNA 12x	tRNA 20x	5S 9x	5S 12x	5S 20x
	0.99	1.0	0.98	0.98	0.98	0.97

Figure 2.3: Characterization of varying numbers of PCR cycles in SHAPE-Seq library construction. SHAPE-Seq v1.0 libraries were constructed for tRNA and 5S rRNA using either 6x, 9x, 12x or 20x cycles of PCR before sequencing. (+) and (-) fragment distributions for each RNA were compared between 6x cycles and the other cycle numbers (top) as in Figure 2.2. Pearson correlation values (R) for individual comparisons between fragment distributions, and distributions, are shown on the bottom.

A comparison between the 6x and 9x (+) and (-) distributions for each RNA showed near perfect agreement, with R values ranging from 0.97-1.0 (Figure 2.3), indicating that the PCR cycles used in SHAPE-Seq v1.0 (9x) are not biasing these distributions. The comparison between 6x and 12x was similar, with the R values ranging from 0.86-1.0. Even the comparison between 6x and 20x showed strong agreement, with R values ranging from 0.83-0.98. For all of the worst comparisons, with R values of 0.83-0.86, the discrepancy stemmed mostly from a few positions rather than a systematic length bias across the lengths of the RNAs. Specifically, the shortest fragments for the 5S 12x (-) and 20x (-) fragment distributions differed slightly (Figure A.11). For the tRNA (+) distribution, there were six positions that showed a similar discrepancy (Figure A.11). Regardless of these discrepancies in the (+) and (-) distributions, the comparisons between values between the different PCR cycle conditions yielded R values in the range of 0.97-1.0 (Figure 2.3), indicating that up to 20 PCR cycles can be used without the introduction of systematic bias in SHAPE-Seq reactivities.

2.4.4 Assessing the Reproducibility of SHAPE-Seq

Replicate indexed SHAPE-Seq experiments were performed using a similar configuration as the minimal adapter library above, except that a 34 nt adapter sequence was used instead to take advantage of Illumina TruSeq multiplexing (Figure A.3). Replicates, measured from completely independent library preparations, were obtained for the five panel RNAs discussed above, as well as four additional RNAs — the cyclic di-GMP bacterial riboswitch from *V. cholerae*, the TPP riboswitch from *E. coli*, the SAM I riboswitch from *T. tencongensis* and the Hepatitis C virus IRES domain - all of which have been used in previous SHAPE benchmarking experiments [102, 160, 161]. As indicated in Section 2.3, folding conditions included ligands where appropriate (Table A.2).

The results showed that the SHAPE-Seq technique is highly reproducible for all nine RNAs (Figures 2.5 and A.12).

2.4.5 SHAPE-Seq v2.0: Removing RNA Sequence Requirements with Universal RT Priming

One major limitation to the SHAPE methods described above is the requirement of priming the RT step either within the RNA sequence itself, or by including structure cassette sequences that contain a 3′ RT primer binding site (Figure 2.4) [92].

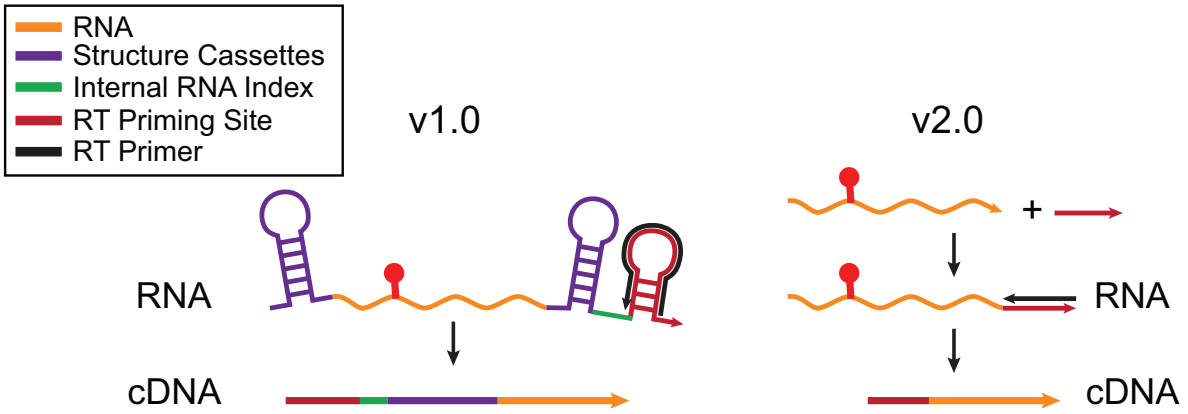


Figure 2.4: Schematic of traditional SHAPE/SHAPE-Seq v1.0 RT priming strategies and the universal RT priming strategy of SHAPE-Seq v2.0. Traditional strategies use sequences that are part of the RNA, or added structure cassette flanking sequences to prime RT reactions. In SHAPE-Seq v2.0, a linker sequence is added to the RNA post-modification, which serves as a priming site for the RT reaction (see Figure A.4).

In the case of priming within the RNA itself, custom primers must be used for each RNA, and structural information is lost at the site of RT priming. In the structure cassette case, flanking sequences must be added to the RNA itself, with the potential to alter the folded structures of the RNAs. SHAPE-Seq v2.0 creates a 'universal' RT primer binding site by ligating a linker sequence to the 3' end of the RNA after modification (Figures 2.4 and A.4).

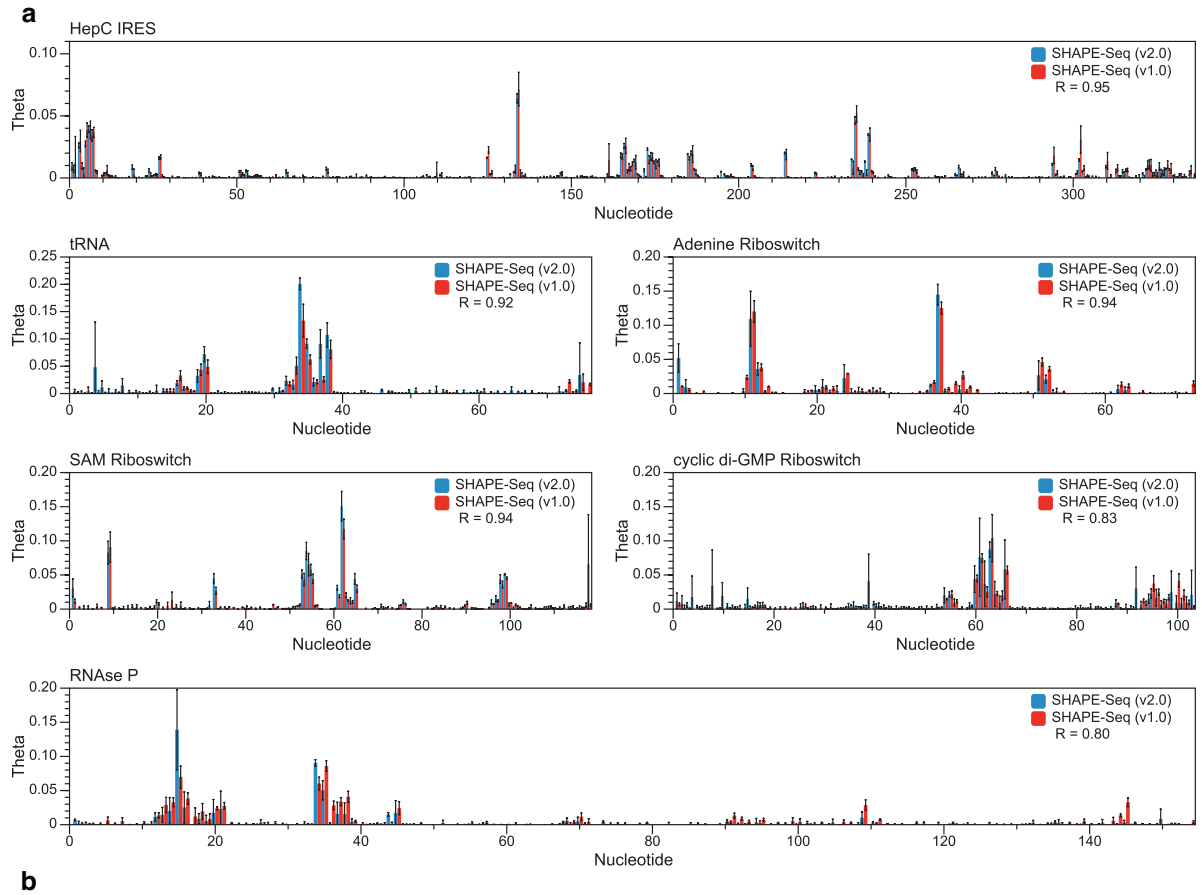
2.4.6 3' Ligation Methods and Other Protocol Adjustments

We focused specifically on ligating pre-adenylated linkers using truncated T4 RNA ligase 2 to prevent unwanted side reaction products. We tested ligation conditions using three previously designed and tested miRNA cloning linkers [167, 168] available from Integrated DNA Technologies. As shown in Figure A.13, each IDT linker was effective at ligating to the $tRNA^{phe}$ sequence. We chose linker 2 for having the highest melting temperature with respect to its complementary RT primer.

The addition of the 3' ligation step required a number of protocol adjustments to improve cDNA yield and reduce side products. A critical adjustment was reducing the amount of RT primer used (see Section 2.3). This had the effect of reducing the amount of unextended primer left over after RT, and thus reducing the amount of unwanted side product, which increases the usable signal from the sequencing run. As described above in Section 2.3, we also altered many other intermediate steps of SHAPE-Seq, culminating in SHAPE-Seq v2.0 (Figure A.4).

2.4.7 Comparison of SHAPE-Seq v2.0 and v1.0

Reactivity spectra generated with SHAPE-Seq v2.0 and v1.0 are compared in Figure 2.5. Pearson correlation coefficients were between 0.80-0.95, supporting a strong correlation between v1.0 and v2.0 results. Of the RNAs with weaker correlations (namely cyclic-di-GMP Riboswitch (R=0.83), RNase P (R=0.80), and 5S rRNA (R=0.85)), the main qualitative differences tend to be at the ends of the RNA molecule, with v2.0 tending toward higher θ values at the 5′ end, and v1.0 higher at the 3′ end (Figures 2.5 and A.12).



b

RNAs < 120 nt	tRNA	adenine riboswitch	TPP riboswitch	cyclic-di-GMP riboswitch	SAM riboswitch
R (v2.0 vs. v1.0)	0.92	0.94	0.91	0.83	0.94
RNAs ≥ 120 nt	5S	RNAse P	*Tetrahymena* ribozyme	HepC IRES	
R (v2.0 vs. v1.0)	0.85	0.80	0.89	0.95	

Figure 2.5: SHAPE-Seq v2.0 vs. SHAPE-Seq v1.0. (a) Reactivity spectra θs are plotted for each RNA, including error bars which are calculated as standard deviations of reactivities at each nucleotide from three independent replicate experiments for SHAPE-Seq v2.0 (blue) and SHAPE-Seq v1.0 (red). Pearson correlations from the comparisons of average reactivity spectra are shown in each plot and listed in the table in (b). Figure A.12 shows detailed comparisons for each RNA in the panel.

2.4.8 Using SHAPE-Seq Reactivities as Constraints in Thermodynamic RNA Folding Algorithms

While our primary goal was to assess the ability of SHAPE-Seq to generate accurate and reproducible reactivity data in a high-throughput manner, it is important to recognize downstream uses of this information. One common use is to constrain thermodynamic RNA folding algorithms [101, 102, 160, 161, 163, 169]. In this approach, each SHAPE reactivity, $\{r_i\}$, of an RNA is converted into a $\Delta G_{SHAPE,i} = m \ln(r_i + 1) + b$, which are then used to predict RNA secondary structural properties such as the Minimum Free Energy (MFE) structure. The parameters m and b are fit to produce the most accurate structural predictions over a benchmark set of RNAs for which reactivity information is available [163]. Secondary structure prediction accuracy is assessed by comparing the predicted MFE structure to the crystal structure using two representative statistical measures: sensitivity, or the fraction of base pairs in the accepted (crystal) structure predicted correctly; and positive predictive value (PPV), which is the fraction of predicted pairs that are correct [170]. Overall, the incorporation of SHAPE reactivity data into thermodynamic structure prediction algorithms has been shown to increase the accuracy of predictions [101, 102, 160, 161, 163].

To incorporate SHAPE-Seq reactivity data into folding algorithms, we converted θ's to a scale that is more similar to the reactivity scale typically used for SHAPE experiments. In traditional SHAPE data scaling, 'highly reactive' positions are set to a reactivity of ~1 by scaling to a normalization factor that averages the reactivities of these positions while excluding outliers [101]. In SHAPE-Seq, θ's are guaranteed to be <1 due to the constraint that they sum to 1 over the length of the RNA, thus SHAPE-Seq θ's are smaller than typical reactivities. However, θ's can be easily converted to a

similar scale by multiplying θ's by the length of the RNA, L. Defining

$$\rho_i = L\theta_i \tag{2.1}$$

ensures that the average ρ_i, $\bar{\rho}$, is

$$\bar{\rho} = \frac{\sum_{i=1}^{L} \rho_i}{L} = L\frac{\sum_{i=1}^{L} \theta_i}{L} = 1 \tag{2.2}$$

making ρ's on roughly the same scale as SHAPE r's. Therefore, we used SHAPE-Seq v2.0 ρ's to evaluate sensitivity and PPV predictions for each RNA in the panel (Table 2.1). As seen from Table 2.1, when we used the current recommended values of $m = 1.8$ and $b = -0.6$, we observe an increase in the total sensitivity and PPV values over unconstrained folds. These values are in fact comparable to results from recent studies that used QuSHAPE reactivity with these parameters [102]. However, since ρ's are slightly different than SHAPE r's, we anticipated that there could be room to adjust m and b values. We found that $m = 1.1$ and $b = -0.3$ gave total sensitivity and PPV values over the panel to be 84% and 89%, respectively, with many RNAs in the panel predicted to a very high sensitivity and PPV individually (Tables 2.1 and A.6). We emphasize that $m = 1.1$ and $b = -0.3$ should serve as a guide for incorporating SHAPE-Seq reactivity data into folding algorithms, though more work should be performed to refine these values over a broader context of RNA structures.

Table 2.1: RNA structure prediction accuracy using the RNAstructure [163] *Fold* algorithm and the SHAPE-Seq v2.0 reactivity data (ρ's) as constraints, with different m and b parameters. Sensitivity and PPV values for each RNA are in Tables A.4-A.6.

RNA	Total Sensitivity	Total PPV
No SHAPE data	228/360=63.3%	229/373=61.4%
SHAPE parameters (m = 1.8 and b = −0.6)	292/360=81.1%	293/351=83.5%
SHAPE-Seq updated parameters (m = 1.1 and b = −0.3)	303/360=84.2%	304/342=88.9%

2.5 Discussion

In this work, we present a systematic analysis and optimization of the SHAPE-Seq technique to structurally characterize RNAs in a high-throughput, multiplexed fashion. Overall, the above results demonstrate that our optimizations to the previously published SHAPE-Seq v1.0 technique [107, 108] provide highly reproducible nucleotide-resolution chemical reactivity data over the wide array of structural contexts present in our panel of benchmark RNAs (Figure 2.5).

Initial comparisons between reactivities generated from before SHAPE-Seq v1.0 optimization and QuSHAPE showed there was a strong overall correlation between the two methods, with specific differences highlighted by the inspection of individual reactivity spectra. As shown in Figure A.5, for most of the RNAs, SHAPE-Seq v1.0 and QuSHAPE capture the same clusters of reactive nucleotides, but differ in the specific θ value assigned to these positions. There does, however, appear to be a difference between the two techniques at low reactivity nucleotides. In particular, there are a large number of nucleotide positions (216 out of a total of 586) where the SHAPE-Seq θ is 0, while the QuSHAPE θ is a small, non-zero number. We hypothesize that the

analog nature of the capillary electrophoresis read-out in QuSHAPE experiments, and the Gaussian fitting algorithm used to quantify electopherogram peaks, can amplify baseline noise and make zero reactivity peaks appear to have a small but non-zero reactivity.

We systematically optimized each step of the SHAPE-Seq v1.0 protocol associated with library preparation and sequencing, which constitute the major differences between SHAPE-Seq and SHAPE analyzed by capillary electrophoresis. As shown above, steps such as adapter ligation (Figure 2.2) and PCR (Figure 2.3) do not appear to introduce a systematic bias in SHAPE-Seq reactivity data.

We also sequenced three SHAPE-Seq v2.0 libraries on both the HiSeq and MiSeq platforms. All three show a very strong correlation between the HiSeq and MiSeq for the (+) and (-) read distributions, with R values for these comparisons all in excess of 0.99 (Figure A.14). This indicates that the choice of sequencing platform has no effect on SHAPE-Seq data.

We then expanded the flexibility of the SHAPE-Seq technique by incorporating the standard Illumina library indexing strategy to sequence multiple libraries in the same lane (Figures A.3 and A.4). This can be used in a number of ways, for example, by using indexes to perform replicate experiments on the same RNA, or to run different groups of RNAs together as done in this work. In principle, this flexibility allows any number of experimental variations to be performed.

We have also significantly extended SHAPE-Seq by creating a universal RT priming strategy that does not require RNA-specific primers to be designed and used, or flanking sequence to be added to the RNA itself. This new technique, SHAPE-Seq v2.0, works by ligating a linker sequence after modification, which can then serve as

an RT priming site. With this innovation, SHAPE-Seq v2.0 can now be performed on unknown RNAs, without the need to deal with the complexities and biases associated with random RT primers [116, 117]. In addition, SHAPE-Seq v2.0 allows reactivity information to be characterized for almost the entire length of the RNA, without losing structural information at RT priming sites within the RNA sequence. Finally, SHAPE-Seq v2.0 should be equally applicable to naturally synthesized RNAs as it is to in vitro transcribed RNAs, and thus should allow the standardization of chemical probing experimental protocols and data analysis.

The data presented in this work also represents a high-quality benchmark SHAPE-Seq dataset for a panel of RNAs that are becoming the gold standard for technique comparison in the field [102, 160, 161]. As described in Table A.7, all data is freely available in the RNA Mapping Database [169]. This should serve as a useful resource for further experimental technique development, as well as to researchers interested in using SHAPE-Seq data to constrain computational RNA folding algorithms to give more accurate RNA structural models. In fact, as we have shown above SHAPE-Seq values can be used to make structural predictions that are as accurate as those from more traditional SHAPE experiments (Table 2.1). It is our hope that this dataset serves as a starting point for understanding how best to incorporate SHAPE-Seq data into computational structure prediction.

Finally, we note that while this technique was originally named 'SHAPE-Seq' after the SHAPE chemistry that was used in the first version of the technique, it is in fact applicable to any RNA structure-dependent chemical probe. With our innovation of the SHAPE-Seq v2.0 universal priming technique, and the already rigorous signal decay correction and accurate reactivity calculation offered by the Spats pipeline [97, 98, 108], we anticipate SHAPE-Seq to be continued to be used in a wide array of power-

ful techniques aimed at understanding the RNA sequence-structure relationship at the heart of some of life's most fundamental processes.

2.6 Acknowledgements

We thank Gary P. Schroth and Illumina, Inc. for advice and support during this work. We also thank Sharon Aviran for useful discussions on SHAPE-Seq data scaling. Finally, we thank Peter Schweitzer and the Cornell Life Sciences Core facility for sequencing support during this work.

2.7 Funding

National Science Foundation Graduate Research Fellowship Program [DGE-1144153 to K.E.W.]; Cornell University Center for Life Sciences Enterprises, a New York State Center for Advanced Technology supported by New York State and industrial partners [C110124 to J.B.L.]; New Innovator Award through the National Institute of General Medical Sciences of the National Institutes of Health [DP2GM110838 to J.B.L.]. J.B.L. is an Alfred P. Sloan Research Fellow. Funding for open access charge: National Institutes of Health [DP2GM110838].

CHAPTER 3

SIMULTANEOUS CHARACTERIZATION OF CELLULAR RNA STRUCTURE AND FUNCTION WITH IN-CELL SHAPE-SEQ

3.1 Abstract

Many non-coding RNAs form structures that interact with cellular machinery to control gene expression. A central goal of molecular and synthetic biology is to uncover design principles linking RNA structure to function to understand and engineer this relationship. Here we report a simple, high-throughput method called in-cell SHAPE-Seq that combines in-cell probing of RNA structure with a measurement of gene expression to simultaneously characterize RNA structure and function in bacterial cells. We use in-cell SHAPE-Seq to study the structure-function relationship of two RNA mechanisms that regulate translation in *Escherichia coli*. We find that nucleotides that participate in RNA-RNA interactions are highly accessible when their binding partner is absent and that changes in RNA structure due to RNA-RNA interactions can be quantitatively correlated to changes in gene expression. We also characterize the cellular structures of three endogenously expressed non-coding RNAs: 5S rRNA, RNase P, and the *btuB* riboswitch. Finally, a comparison between in-cell and *in vitro* folded RNA structures revealed remarkable similarities for synthetic RNAs, but significant differences for RNAs that participate in complex cellular interactions. Thus, in-cell SHAPE-Seq represents an easily approachable tool for biologists and engineers to uncover relationships between sequence, structure, and function of RNAs in the cell.

The work described in this chapter includes contributions from Timothy R. Abbott and Julius B. Lucks. This work was originally published in *Nucleic Acids Research* and has been reproduced here with permission from Oxford University Press. Kyle E. Watters, Timothy R. Abbott, and Julius B. Lucks, Simultaneous Characterization of Cellular RNA Structure and Function with in-cell SHAPE-Seq, *Nucleic Acids Res*, 2016, **44**, (2), e12.

3.2 Introduction

Non-coding RNAs (ncRNAs) have diverse functions, ranging from regulatory roles in transcription, translation, and messenger stability in prokaryotes [171, 172] to gene silencing, transcript splicing, and chromatin remodeling in eukaryotes [167, 173, 174]. This recognized importance of ncRNAs is accelerating, as high-throughput genomics techniques continue to discover new ncRNAs and their roles in globally tuning genome expression [175]. Synthetic biologists, in turn, have started to take advantage of this diversity of ncRNA mechanisms to design sophisticated RNA regulators that can precisely control gene expression [24, 25, 135–137, 176, 177]. Such widespread use of RNA-based gene regulation in both natural and engineered cellular systems has thus prompted a large effort to understand the relationship between RNA structure and function within the cell [27, 178, 179].

This effort has recently accelerated with the advent of high-throughput RNA structure characterization technologies that combine chemical probing with next-generation sequencing [107, 112, 114–118, 122]. In one such method, called selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq), SHAPE reagents [95] modify the 2′-OH of less-structured RNA nucleotides, which causes reverse transcription (RT) to halt one nucleotide before the modification [86, 104, 158]. Next-generation sequencing of the resulting cDNA fragments is then used to determine the location and frequency of modifications across each RNA under study. These modification frequencies are then used to estimate a 'reactivity' that quantifies the propensity of each nucleotide in an RNA to be modified by the chemical probe [97, 98]. High reactivities reflect nucleotides that are unstructured, while low reactivities suggest structural constraints such as base pairing, stacking, or RNA-ligand interactions [107, 122, 180].

The use of next-generation sequencing has allowed these methods to be highly multiplexed, which has offered some of the first 'transcriptome-wide' glimpses of RNA structure [114–118]. However, the current methods are designed for asking broad questions about cellular RNA structure and are not well suited for extensive structure-function analysis of specific RNA targets. Further, the current monetary costs and computational complexity of analyzing chemical probing data over the entire transcriptome are a significant barrier to overcome for studies requiring many replicates, such as mutational analysis of select RNAs. Yet, simpler methods based on capillary or gel electrophoresis cannot be multiplexed to characterize multiple RNAs at once or remove off-target cDNA products. In addition, other current techniques that use next-generation sequencing often rely on many time-consuming steps for sequencing library preparation [114–118], such as successive gel purifications that increase turnaround time, cost, and skill required to analyze RNA structures inside the cell. Finally, many current techniques focus on characterizing cellular RNA structures, without an explicit measurement of RNA function.

To address these issues for researchers interested in studying the structure-function relationship of select RNAs in depth, we have developed in-cell SHAPE-Seq. In-cell SHAPE-Seq combines in-cell probing of RNA structure with SHAPE-Seq [122] and a measurement of gene expression through fluorescent reporter assays to characterize RNA regulatory function. By measuring fluorescence and performing the chemical probing experiment on the exact same cell culture, in-cell SHAPE-Seq is able to link changes in cellular RNA structure to changes in gene expression (Figure 3.1). The use of a new selective PCR method during library construction further simplifies the experiment by removing gel-based purification steps. In-cell SHAPE-Seq thus provides nucleotide-resolution structural data for multiple RNAs at a time in a simple experiment that leverages many of the recent technical advances in SHAPE-Seq as well as

existing data analysis pipelines [97, 98, 122].

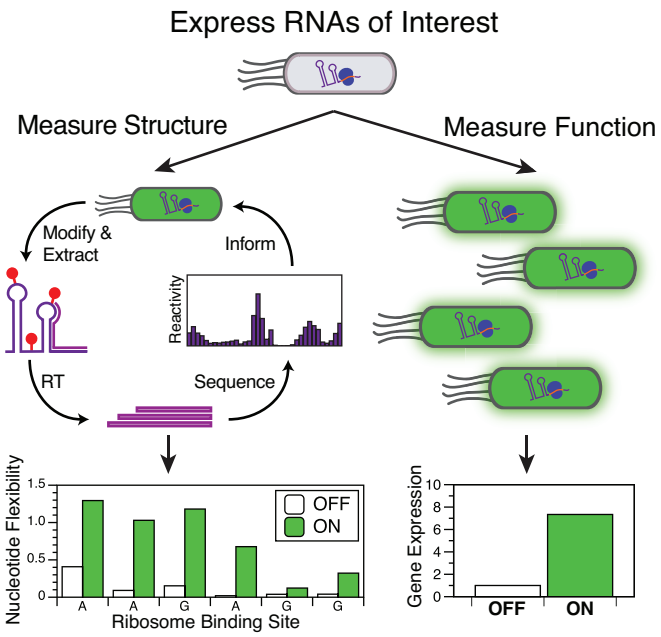


Figure 3.1: In-cell SHAPE-Seq overview. The in-cell SHAPE-Seq pipeline yields information about RNA structure within the cell by detecting regions of RNA flexibility using an in-cell chemical probe, reverse transcription (RT), next-generation sequencing, and bioinformatics. Gene expression measurements yield information about RNA function. Coupling the two measurements provides quantitative information about cellular RNA structure-function relationships. Coding sequence (CDS) is labeled.

In this work, we develop and apply in-cell SHAPE-Seq to study the structure-function relationship of two well-characterized RNA regulatory systems in *E. coli*: the synthetic RNA riboregulator translational activator system developed by Isaacs *et al.* [25], and the natural IS10 translational repressor system recently engineered by Mutalik *et al.* [176]. To perform these studies, we established a general two-plasmid expression system for studying RNA regulator pairs. Both plasmids contain convenient RT-priming sites that facilitate in-cell SHAPE-Seq measurements, as well as a fluorescent protein reporter on one of the plasmids for coupling to gene expression measurements (Figure B.1). Using the two-plasmid system, we show how in-cell SHAPE-Seq can be used to derive structural models of cellular RNA folding for these systems. We also show that in-cell SHAPE-Seq data can be used to generate quantitative links between RNA structural changes in the cell and their functional consequences. We then assess the sensitivity of the method by targeting three endogenously expressed RNAs in *E. coli*: 5S rRNA, RNase P, and the riboswitch domain of the *btuB* mRNA leader sequence. We show that in-cell SHAPE-Seq reactivity data can be used to corroborate and refine structural models of these functional ncRNAs. Next, we compare data from *in vitro* equilibrium refolding experiments to in-cell SHAPE-Seq reactivities and find intriguing similarities and differences between these folding environments. We end by discussing how in-cell SHAPE-Seq could be immediately applied to uncovering the cellular RNA structure-function relationship of a broad array of RNA regulatory systems.

3.3 Materials and Methods

See Appendix B.4 for a step-by-step protocol of the complete in-cell SHAPE-Seq experiment (Figure B.2).

3.3.1 Platform (plasmid) construction

Figure B.1 describes our standardized platform for expressing sense/antisense regulatory RNA pairs that are not endogenously expressed in *E. coli*. Specific primer designs and detailed cloning procedures to construct the plasmids used in this work, or plasmids for other RNA regulatory systems, can be found in Appendix B.4. The cis-repressed sense RNA (crRNA) and *trans*-activating RNA (taRNA) plasmids were generated by introducing the riboregulator sequences from Isaacs *et al.* [25] into the sense and antisense expression platforms with Gibson Assembly [181]. To create the RNA-IN sense plasmids, the original sequence from Mutalik *et al.* [176] of variant S1 was mutated using standard PCR amplification-ligation methods. The antisense RNA-OUT sequences from Mutalik *et al.* [176] were cloned into the antisense plasmid architecture with Gibson Assembly [181]. All plasmid sequences are listed in Table B.2.

3.3.2 Strains, growth media, and RNA expression

Each sense and antisense plasmid was transformed separately, or in combination, into chemically competent *E. coli* TG1 cells. Where indicated, a control antisense plasmid, lacking the antisense RNA sequence but containing a promoter and terminator (Figure B.1), was used to maintain a consistent cellular load to properly compare fluorescence levels with or without the antisense RNA present. Transformed cells were plated on LB+Agar media with 100 μg/mL carbenicillin, 34 μg/mL chloramphenicol, or both and incubated overnight at 37 °C. The next day, individual colonies were picked and grown in 1 mL of the appropriate LB+antibiotic in a 2 mL 96-well block (Costar) and grown approximately 17 hr overnight at 37 °C at 1,000 rpm in a VorTemp 56 (Labnet) benchtop shaker. Twenty-four microliters of this overnight culture was

then used to subculture into 1.2 mL of LB+antibiotic. *E. coli* TG1 cells without plasmids were prepared in the same way without antibiotic for probing endogenously expressed RNAs. The subculture was grown under the same conditions as the overnight culture for at least three hours before measuring fluorescence (for cultures containing the sense plasmid with SFGFP) and performing structure probing. See Appendix B.4 steps 17-21 for more details.

3.3.3 *in vitro* RNA purification

in vitro transcription templates for crR12, taR12, RNA-IN S3 C24A A25C, and RNA-OUT A3 were prepared using PCR with Taq polymerase (NEB), replacing the *E. coli* promoter with the T7 promoter (TAATACGACTCACTATAGG), followed by ethanol precipitation. The *in vitro* transcription reaction contained 5 μg of template, 40 mM tris-HCl pH 8.0, 20 mM $MgCl_2$, 10 mM DTT, 20 mM spermidine, 0.01% Triton X-100, 5 mM NTPs, 60 U of SUPERase-In, 20 μL of purified T7 RNA polymerase, brought to a total of 1 mL with MilliQ H_2O. The shorter RNAs (taR12 and RNA-OUT A3) were gel purified and passively eluted before ethanol precipitation. The longer RNAs containing the SFGFP sequence (crR12 and RNA-IN S3 C24A A25C) were purified from the transcription reaction using AMPure XP RNA beads according to the manufacturers instructions.

3.3.4 RNA modification and fluorescence assay

Fluorescence was measured after at least three hours of growth by pelleting 150 μL of each subculture and resuspending it in 200 μL PBS buffer with 100 μg/mL kanamycin

(to prevent further translation) and assaying for fluorescence (485/520 nm) and optical density (OD_{600}), which typically ranged from 0.1-0.5. Fluorescence and OD_{600} were first corrected by subtracting values measured for a media blank. Relative fluorescence levels of each culture (except those used for endogenous RNA characterization) were determined by normalizing the fluorescence readout by optical density (FL/OD) and subtracting the FL/OD of the antisense plasmid containing cultures (which did not contain superfolder GFP; SFGFP) to correct for cell autofluorescence.

For RNA modification with 1-methyl-7-nitroisatoic anhydride (1M7), 500 μL of each 3 hour subculture was modified in the 96-well block with 13.3 μL 250 mM 1M7 in DMSO (6.5 mM final) (+) or 13.3 μL DMSO (-) for 3 min before RNA extraction. For the DMS modification, the 1M7 was replaced with 27.75 μL of 13% DMS in ethanol and the DMSO replaced with 27.75 μL ethanol. After 3 min of incubation with DMS, the reaction was quenched with 240 μL 2-mercaptoethanol. See Appendix B.4 steps 22-32 for a more detailed in-cell modification protocol.

For *in vitro* transcribed RNAs, 10 pmol of RNA total (1 pmol of sense, 9 pmol of antisense for bimolecular experiments) were diluted in 12 μL H_2O total before denaturing at 95 °C for 2 min. The RNAs were than snap-cooled on ice for 1 min before adding 6 μL 3.3X Folding Buffer (333 mM HEPES, 333 mM NaCl, 33 mM $MgCl_2$, pH 8.0) and incubated at 37 °C for 20 min. The resulting 18 μL were split and added to 1 μL 65 mM 1M7 (6.5 mM final) or 1 μL DMSO and incubated at 37 °C for 1 min. The modified *in vitro* RNAs were then ethanol precipitated and dissolved in 10 μL H_2O before reverse transcription.

3.3.5 RNA extraction

For in-cell probing experiments, both modified (+) and control (-) samples were pelleted, then resuspended in 100 μL of hot Max Bacterial Enhancement Reagent (Life Technologies) before extraction with TRIzol Reagent (Life Technologies) according to the manufacturers protocol. Extracted RNA was dissolved in 10 μL of water. See Appendix B.4 steps 33-39 for more details.

3.3.6 Reverse transcription

For each RNA sample, 3 μL of 0.5 μM oligonucleotide were added for reverse transcription (RT), except for the *btuB* riboswitch samples to which 3 μL of 50 nM oligonucleotide were added instead. Sense RNAs were extended with an RT primer for SFGFP, antisense RNAs were extended with a primer for the ECK120051404 terminator, and endogenously expressed RNAs were extended with an RNA-specific primer (Table B.3). For samples containing sense-antisense pairs, 1.5 μL of each primer were mixed together. All RNAs were denatured at 95 °C for 2 min, then 65 °C for 5 min. After denaturing, each RNA sample was then snap-cooled on ice for 1 min before extension with Superscript III (Life Technologies) at 52 °C for 25 min. After RT the RNA was hydrolyzed with 1 μL 10 M NaOH. The solution was then partially neutralized with 5 μL of 1 M hydrochloric acid and ethanol precipitated. See Appendix B.4 steps 40-51 for more details.

3.3.7 Adapter ligation

The cDNA from each RT reaction was separately ligated to a ssDNA adapter for Illumina sequencing with CircLigase I ssDNA ligase (Epicentre). Each ligation reaction was incubated at 60 °C for 2 hours, followed by deactivation at 80 °C for 10 min. The ligated cDNA was then ethanol precipitated and dissolved in 20 μL water. Unligated oligonucleotides were removed by purification with 36 μL of Agencourt AMPure XP beads (Beckman Coulter) according to manufacturers protocol. See Appendix B.4 steps 52-57 for details.

3.3.8 Quality control

Each single-stranded cDNA library from a highly expressed RNA was PCR amplified with Phusion polymerase (NEB) for 15 cycles with two forward primers, a selection primer (containing an RNA-specific sequence and part of the forward Illumina adapter) and a longer primer containing all of the forward Illumina adapter, and a fluorescent reverse primer that binds to the reverse Illumina adapter sequence as part of the ligated ssDNA adapter (Table B.3 and Figure B.3), Appendix B.4 step 58). Moderate to weakly expressed RNAs (RNase P and the *btuB* riboswitch) were amplified for 15 cycles without the complete forward Illumina adapter primer first, which was then added for a second set of 15 cycles. Libraries that were derived from cultures that contained both sense and antisense plasmids were amplified separately with one selection primer to visually separate the library qualities of the independent priming locations. The fluorescently tagged amplifications were run on an ABI 3730xl Analyzer with GeneScan 500 LIZ standard (Life Technologies) and checked for the correct full-length product (indicating good RT and PCR) and minimal side product formation.

See Appendix B.4 steps 58-67 for further details.

3.3.9 dsDNA sequencing library construction

Highly expressed RNA libraries passing quality analysis were PCR amplified with Phusion polymerase (NEB) for 15 cycles using three primers: a forward primer that contained an Illumina adapter, another RNA-specific forward selection primer, and a reverse primer that contained the other Illumina adapter and one of 24 TruSeq indexes (Table B.3 and Figure B.3). Moderate to weakly expressed RNAs (RNase P and the *btuB* riboswitch) were amplified for 15 cycles without the complete forward Illumina adapter primer first, before it was added for a second set of 15 cycles. Excess primer was removed with ExoI (NEB) before purification with 90 μL of Agencourt AMPure XP beads (Beckman Coulter) according to the manufacturers protocol. See Appendix B.4 steps 68-75 for more details.

3.3.10 Next-generation sequencing

The molarity of the individual libraries was estimated from the lengths and intensity of peaks in the fluorescent quality traces, and the concentration of each library measured with the Qubit fluorometer (Life Technologies). All libraries were mixed to have the same final molar concentration and sequenced with the Illumina MiSeq v3 kit or HiSeq 2500 rapid run using 2x35 bp paired end reads. Adapter trimming was turned off during Illumina post-sequencing processing.

3.3.11 Data analysis

Reactivity spectra were calculated using Spats v0.8.0 and a number of utility scripts to prepare the Illumina output for Spats following previous work [122]. Illumina adapter sequences were trimmed from each read using the FASTX toolkit [`http://hannonlab.cshl.edu/fastx_toolkit/`], then aligned to the target RNA sequences with Bowtie 0.12.8 [182] based on the input RNAs to determine locations of modifications. Spats separates the (+) and (-) channel reads according to the handle sequence, and calculates θ for each nucleotide using statistical corrections for RT drop-off, where θ represents the probability of modification at a particular nucleotide [97, 98]. Resulting θ values were then normalized to ρ values according to Appendices B.1 to B.1. Reactivities (ρ) greater than 1.25 are considered highly reactive, between 0.5-1.25 moderately reactive, and less than 0.5 weakly reactive. All data is freely accessible from the RNA Mapping Database (RMDB) (`http://rmdb.stanford.edu/repository/`) [169] using the IDs in Table B.4..

3.3.12 Structure folding predictions

RNA secondary structure predictions were performed using the RNAstructure web server [164]. In-cell SHAPE-Seq reactivities (ρ) were used to constrain predictions with the pseudo free energy parameters m (1.1) and b (-0.3) [122] where indicated (Appendix B.1). All computationally predicted folds shown represent the minimum free energy structure.

3.4 Results

3.4.1 A standardized platform for characterizing RNA structures, interactions, and regulatory function in cells

One goal of the in-cell SHAPE-Seq platform is to characterize cellular structural and functional states of regulatory RNAs simultaneously (Figure 3.1). Often, RNA regulatory function is mediated by structural changes in mRNA targets brought about by specific interactions with other cellular molecules such as ligands [27], small RNAs (sRNAs) [22], or ribosomes [183]. We began by first focusing on the natural IS10 and the synthetic riboregulator bacterial sRNA systems that regulate translation in response to RNA-RNA interactions that occur in *trans*. In these systems, translation is controlled by specific RNA structures in the 5′ untranslated region (5′ UTR) of a 'sense' target mRNA. Interaction with a *trans*-acting complementary 'antisense' RNA sequence causes structural rearrangements to occur, turning downstream gene expression ON in the case of riboregulators, or OFF in the case of the IS10 system.

To characterize RNA regulator function, we began by constructing a standardized platform to separately express both the sense and antisense RNAs of each system in *E. coli* (Figure B.1) [25, 176]. In this platform, the sense regulatory RNA sequences were placed downstream of a constitutive promoter and upstream of the superfolder GFP (SFGFP) coding sequence (CDS) [184] on a medium-copy plasmid. The antisense RNAs were placed on a separate high-copy plasmid downstream of the same constitutive promoter (Figure B.1). Gene expression was then characterized by measuring differences in fluorescence between cultures containing the sense plasmid with the antisense plasmid or an antisense control plasmid (See Section 3.3).

To characterize cellular RNA structures, we adapted the SHAPE-Seq experiment [107, 108, 122] to perform the chemical probing step on bacterial cell cultures rather than on *in vitro* pools of purified RNAs, using the ability of certain SHAPE reagents to penetrate living cells (Figure B.2) [126, 158, 185]. To directly couple RNA structure and function characterization, we added 1-methyl-7-nitroisatoic anhydride (1M7; (+) reaction), or the control solvent dimethyl sulfoxide (DMSO; (-) control), to the same *E. coli* cultures that were assayed for SFGFP fluorescence (Figure 3.1). While this probing step modifies all RNAs in the cell, our goal was to target the structural measurement to our regulatory RNAs. To do this, we designed highly specific RT primers that would not exhibit non-specific binding to other RNAs in the transcriptome. To target the sense RNA, we chose an RT primer binding site near the 5′ end of the SFGFP CDS from a set of four designed sequences. To target the antisense RNA, we tested a set of efficient transcription terminators (Table B.1) for specific RT priming capability and found that the synthetic ECK120051404 terminator [186] produced a good quantity of cDNA while remaining highly specific as an RT priming site. Thus, the antisense plasmid contained the ECK120051404 terminator at the 3′ end of the antisense RNA immediately followed by the t500 terminator [186] to improve termination efficiency (Figure B.1). After chemical probing and RNA extraction, reverse transcription was performed with primers targeting one or both of the priming sites described above, and the resultant cDNAs were input into the standard SHAPE-Seq experimental and data analysis pipelines (Figure B.2) [107, 108, 122].

While successful, initial versions of our protocol suffered from an excess of RT primer-sequencing adapter ligation dimers, making it difficult to accumulate enough sequencing reads with our libraries for computational reactivity analysis [97, 98]. In some cases, the amount of ligation dimer could exceed 90% of the total sequencing reads. To overcome this problem, we developed a simple method of selecting against

these unwanted ssDNA dimers by using a mismatch-based selective PCR amplification in place of the normal SHAPE-Seq PCR step (Figure B.3). By using this mismatch PCR as a filter, we removed the need for laborious gel purification steps typically used in other methods [114–118], and reduced amplification of potential off-target RT products. With selective PCR, we observed a 10-40-fold reduction in ligation dimer amplification, with a greater reduction observed for cases where higher quantities of cDNA were obtained. Typically, the PCR selection step reduced the amount of ligation dimer to less than 10% of the total sequencing reads. Together, the PCR selection step and the multiplexing capabilities of SHAPE-Seq allowed many in-cell SHAPE-Seq experiments, containing multiple RNAs probed simultaneously, to be sequenced in a single MiSeq run with deep read coverage.

3.4.2 Characterizing cellular RNA structures of synthetic riboregulators that activate translation

We first used in-cell SHAPE-Seq to examine a synthetic riboregulator system that activates translation in bacteria [25]. In the riboregulator system, the 5′ UTR of the sense mRNA is designed to form a hairpin structure that occludes the RBS and blocks translation (Figure B.4). This cis-repressed RNA (crRNA) is thus OFF in the basal state. To activate translation, a *trans*-activating antisense RNA (taRNA) is expressed that base pairs with the crRNA, causing structural rearrangements that expose the RBS and allow translation (ON state).

As the riboregulators were first designed *in silico* using computational models of RNA folding [25], we first sought to characterize the cellular structures of crRNAs and taRNAs individually using in-cell SHAPE-Seq. We began our analysis with the

taR12/crR12 antisense/sense pair (respectively), which had the highest fold activation of the original riboregulator designs [25]. The in-cell SHAPE-Seq reactivity spectra of crR12 and taR12 were largely consistent with the original designed structures, with several notable adjustments (Figures 3.2 and B.5).

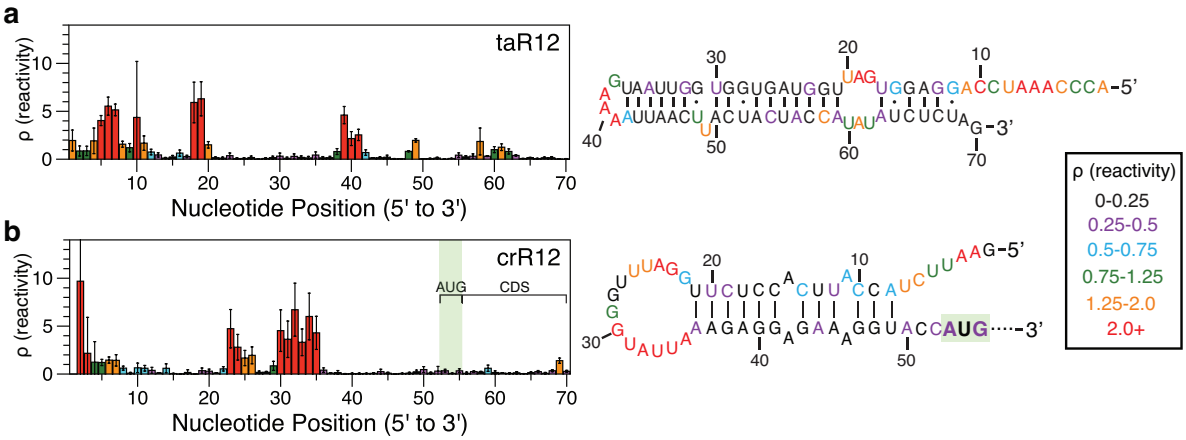


Figure 3.2: Characterization of the cellular structures of the taR12/crR12 synthetic riboregulator RNA translational activator system. Reactivity maps and constrained secondary structure folds are shown for taR12 (a) and crR12 (b). Color-coded reactivity spectra represent averages over three independent in-cell SHAPE-Seq experiments, with error bars representing one standard deviation of the reactivities at each position. RNA structures represent minimum free energy structures generated by RNAstructure [164] using average in-cell SHAPE-Seq reactivity data as constraints (see Section 3.3). Comparisons to the original structural designs from Isaacs *et al.* [25] are shown in Figure B.5. The crR12 structural model was generated from the first 70 nts of the sequence (55 nt shown). Similarly, the terminators following taR12 were not included in the structural analysis. The start codon (AUG) location is boxed and the coding sequencing (CDS) is labeled.

For taR12, designed to be a highly structured hairpin, we observed clusters of high reactivities in all nucleotide positions that were originally expected to be unpaired (Figures 3.2a and B.5). In particular, we were able to distinguish the highly unstructured 5′ tail designed to initiate interactions with the crR12 apical loop [25]. We could also clearly distinguish the hairpin loop, single nucleotide bulge, and inner loop structures within the hairpin. A model of the cellular secondary structure of taR12 generated using in-cell SHAPE-Seq reactivities to constrain computational folding with RNAstructure [164] corroborated these findings, but suggested a larger inner loop structure and an adjustment of the location of the single nucleotide bulge (Figure B.5).

For crR12, we observed a cluster of high reactivities at the 5′ end and in the middle of the molecule, consistent with the overall hairpin design (Figure 3.2b). The large cluster of highly reactive positions between nucleotides 22-35 suggested that crR12 contains a larger loop in cells than previously thought, as seen in the reactivity-constrained secondary structure model of the first 70 nts (Figure 3.2b and B.5). Notably, this loop structure begins at a designed G-A inner loop which was originally introduced to prevent RNAse cleavage and improve fold activation [25], but may also serve to open the upper portion of the hairpin into a larger loop to improve sense-antisense target recognition. Interestingly, nucleotides 27-29 have lower reactivities than the rest of the loop. These nucleotides are part of a YUNR (Y=pyrimidine, N=nucleotide, R=purine) RNA recognition motif that was included in the riboregulator design to facilitate interactions with the taRNA [25]. YUNR motifs are ubiquitous in natural sRNA systems that rely on RNA-RNA interactions to regulate gene expression [136, 187], and the lower reactivities could be reflective of stacking interactions between these nucleotides that can occur in this motif [180, 188].

When considering the designed structural model, two other regions of crR12 have

reactivities lower than expected (Figure B.5). The first region is the hairpin stem, which is predicted to contain multiple sets of inner loops. Low reactivities in inner loops are not uncommon with SHAPE reactivities [107, 180] and could be due to stacking constraints imposed upon the bulged nucleotides by their neighbors or non-canonical base pairing. The second region of low reactivity is from positions 50-70, the majority of which comprise the start of the SFGFP CDS. These low reactivities could be due to several factors, including the binding of cellular proteins, RNA-RNA interactions in the CDS, or ribosomes translating at low levels, preventing the chemical probe from accessing this region.

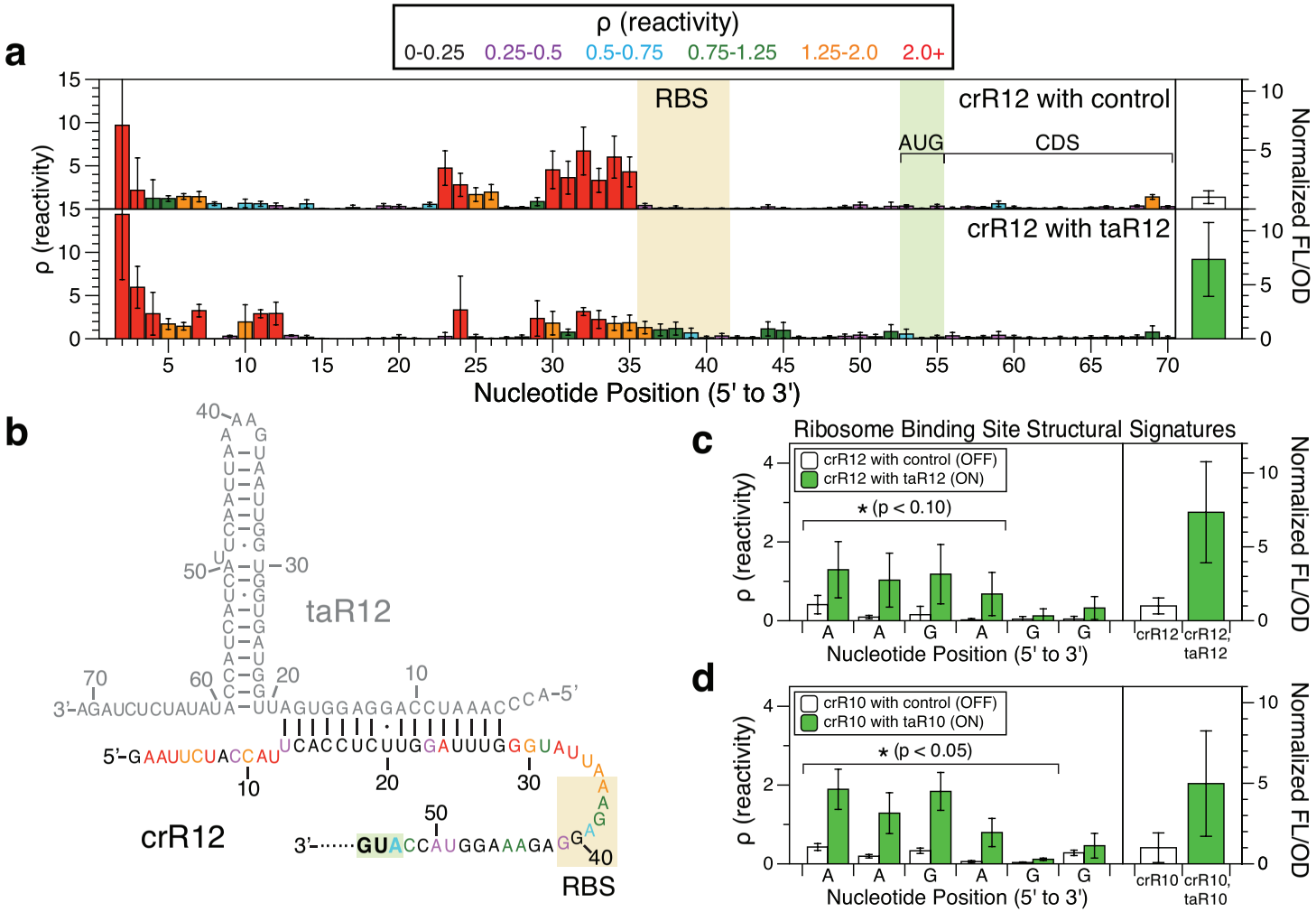
To corroborate our findings, we also examined the taR10/crR10 riboregulator variant, which has a similar overall design and was the second best riboregulator pair in terms of fold activation [25]. We repeated the same measurements and found that the in-cell SHAPE-Seq reactivity spectra and constrained structural models were consistent with the taR12/crR12 results (Figures B.5b and B.6).

Additionally, we compared our in-cell SHAPE-Seq results to an equivalent in-cell 'DMS-Seq-like' approach [117], where the 1M7 modification was replaced with a DMS modification (Figure B.7). Overall, we observed very similar reactivities between in-cell SHAPE-Seq and DMS-Seq at comparable nucleotide positions, corroborating our overall in-cell SHAPE-Seq structure probing approach. However, since DMS shows strong preferences for As and Cs [125] the DMS-Seq reactivities show many gaps, especially since the riboregulators are GU-rich. In fact, the DMS-Seq data was unable to uncover the highly reactive loop of crR12 because of its GU-rich nature, further highlighting the benefit of using SHAPE probes to characterize cellular RNA structures.

3.4.3 Characterizing the cellular RNA interactions and function of synthetic riboregulators that activate translation

We next sought to characterize the structural changes of crR12 that occur in the cell when taR12 activates its translation (Figures 3.3 and B.4). To do this, we performed the full in-cell SHAPE-Seq structure-function measurement in *E. coli* cells expressing both the crR12 sense construct and the taR12 antisense construct. We observed distinct in-cell SHAPE-Seq reactivity changes in several specific regions of crR12 caused by the addition of taR12 that lead to the observed 7.3-fold increase in gene expression (Figure 3.3a). For example, nucleotides in the 5′ half of the crR12 loop region (nts 22-28) generally decrease in reactivity except for nucleotide 24, which remains high but with large error. The observed reactivity changes in crR12 in the presence of taR12 are consistent with the designed taR12/crR12 structural interaction (Figure 3.3b) [25]. However, nucleotides 4-12, 29, and 30 of crR12 remain or become highly reactive, suggesting that these nucleotides are unbound in the taR12/crR12 complex in the cell. These results from in-cell SHAPE-Seq support a model of the taR12/crR12 complex where a 16 bp duplex forms rather than a 25 bp duplex as originally proposed [25].

Figure 3.3: In-cell structure-function characterization of the taR12/crR12 synthetic riboregulator RNA translational activator system. Reactivity maps (a) and a suggested RNA-RNA interaction structure (b) are shown for crR12 of the synthetic riboregulator activator system. (a) Color-coded reactivity spectra for crR12 expressed with taR12 or an antisense control plasmid. Reactivities represent averages over three independent in-cell SHAPE-Seq experiments, with error bars representing one standard deviation. Average fluorescence (FL/OD) values (normalized to the crR12 with antisense control plasmid FL/OD value) on the right show a 7.3-fold activation of gene expression when taR12 is expressed, with error bars representing one standard deviation. The ribosome binding site (RBS) (determined in Figure B.10) and start codon (AUG) locations are boxed. (b) Structural model of the taR12/crR12 binding complex derived from the mechanism proposed by Isaacs *et al.* [25] and refined with the average crR12 with taR12 reactivity data in (a). Nucleotides for crR12 are color-coded by reactivity intensity. (c) Reactivity and functional data of the RBS region show an increase in RBS reactivity (left) and fluorescence (right) when taR12 is co-expressed with crR12. Nucleotide positions that are significantly different ($p < 0.10$) according to a one-sided Welch's t-test are indicated with *. (d) RBS reactivity and functional data for the taR10/crR10 variant (see Figure B.8) Nucleotide positions that are significantly different ($p < 0.05$) are indicated with *.


a
ρ (reactivity)
0-0.25
0.25-0.5
0.5-0.75
0.75-1.25
1.25-2.0
2.0+
RBS
AUG
CDS
crR12 with control
crR12 with taR12
ρ (reactivity)
Normalized FL/OD
Nucleotide Position (5' to 3')
b
taR12
crR12
RBS
c
Ribosome Binding Site Structural Signatures
crR12 with control (OFF)
crR12 with taR12 (ON)
* (p < 0.10)
ρ (reactivity)
Normalized FL/OD
Nucleotide Position (5' to 3')
crR12
crR12, taR12
d
crR10 with control (OFF)
crR10 with taR10 (ON)
* (p < 0.05)
ρ (reactivity)
Normalized FL/OD
Nucleotide Position (5' to 3')
crR10
crR10, taR10

Similar features were observed when the structure-function relationship of the taR10/crR10 interaction was characterized with in-cell SHAPE-Seq (Figure B.8). One difference, however, was a change in the specific nucleotides that were observed to decrease in reactivity as a result of taR10 binding. Overall, more of the 5′ end of crR10 appeared to bind to the taR10 sequence relative to taR12/crR12, though there is a seven nucleotide region from positions 17-23 on crR10 which does not appear to bind as strongly, if at all. One possible explanation for the difference in the interacting structures of these variants is the relative stabilities of the terminal stem loops in taR12 (nt 19-61) and taR10 (nt 22-62). The taR12 hairpin is more stable (ΔG = -20.8 kcal/mol) than the taR10 hairpin (ΔG = -19.6 kcal/mol), as predicted by RNAstructure [164]. Therefore, it may be less energetically favorable for taR12 to unwind to the same extent as taR10 when interacting with crR12 or crR10, respectively (Figure 3.3b and B.8b). Despite these differences, we observed a similar level of activation of gene expression for each system, suggesting that multiple binding states can achieve the same functional consequence.

Unlike crR12 and crR10, no major reactivity changes were observed for either taR12 or taR10 when expressed with their corresponding crRNA targets (Figure B.9). Since these RNAs are expressed in excess of their targets, our in-cell SHAPE-Seq experiment is likely capturing a majority of non-interacting taRNA states, as they take up a large portion of the cellular population. We also note that we did not observe significant reactivity changes in either crRNAs CDS when the corresponding taRNA was present.

3.4.4 Quantitatively linking ribosome binding site reactivity with gene expression

Because the riboregulator mechanism is thought to functionally activate translation in bacteria by removing structural constraints in the crRNA RBS region, we sought to examine how changes in the in-cell SHAPE-Seq reactivities of the RBS region relate to changes in gene expression. However, the AG-rich region between nucleotides 36-46 in crR10 and crR12 has the potential to contain multiple Shine-Dalgarno (SD) sequences. Since the exposure of the RBS turns on gene expression, we hypothesized that the dominant six-nucleotide SD sequence would exhibit the largest reactivity increase. To find this sequence, we summed reactivities over a six nucleotide sliding window for the crR12/crR10 ON and OFF states and looked for the biggest difference between them (Figure B.10). We found that nucleotides 36-41 showed the largest overall increase, with the most notable increases occurring at nucleotides 36-39 in both crR12 and crR10 (Figure 3.3c,d and B.10). These increases correspond to a 6.2-fold and a 4.8-fold change in overall RBS reactivity for the taR12/crR12 and taR10/crR10 systems, respectively, and are linked to 7.3-fold and 5-fold changes in gene expression, respectively (Figure 3.3c,d).

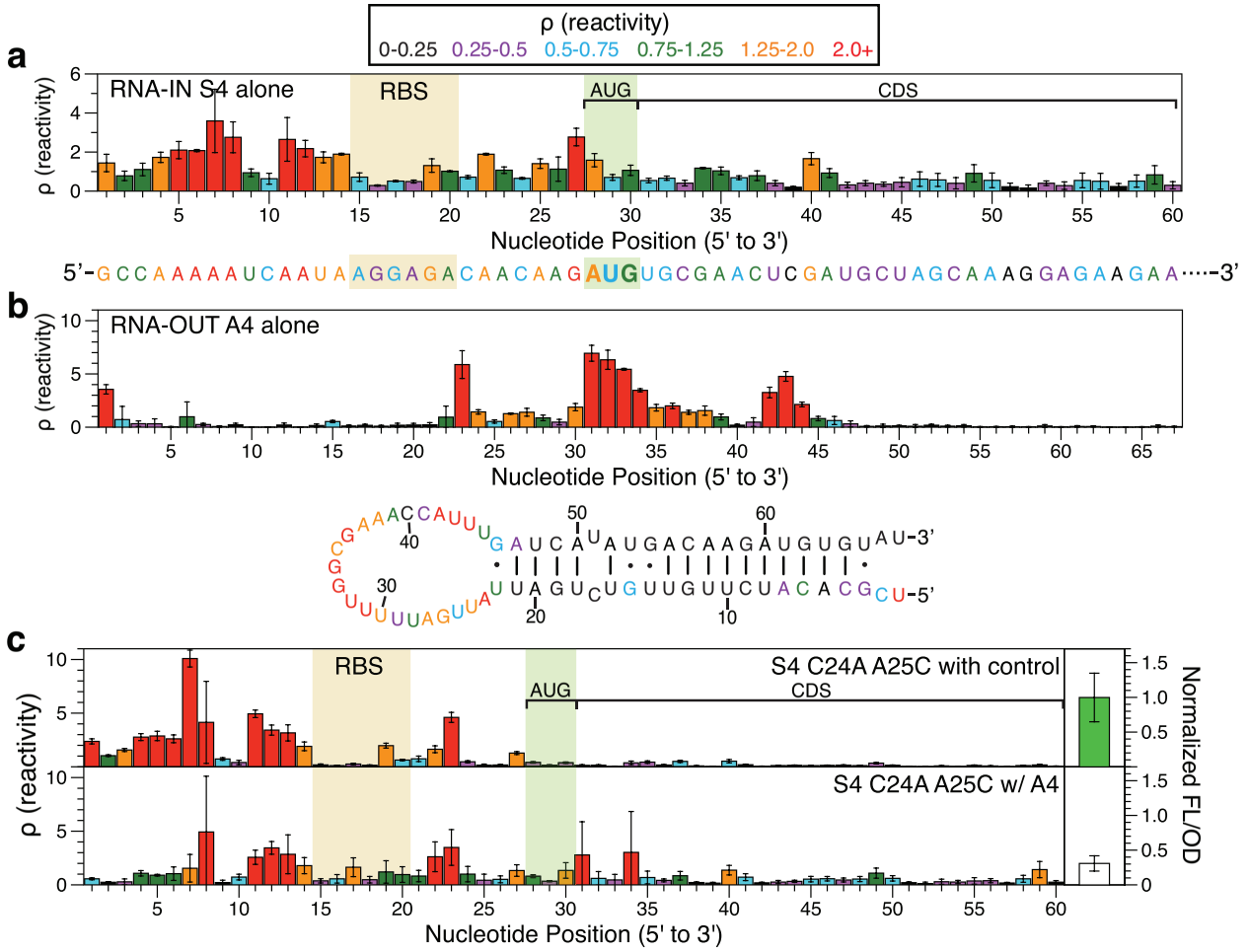
3.4.5 Characterizing the cellular RNA structures of the RNA-IN/OUT translational repressor

We next sought to use in-cell SHAPE-Seq to examine a modified version of the natural sRNA translation repression system from the insertion sequence 10 (IS10) transposon [176]. In the IS10, or RNA-IN/OUT, system the hairpin loop of an antisense RNA

called RNA-OUT initiates interaction with the unstructured 5′ tail of the sense mRNA (RNA-IN) to form a duplex that blocks the RBS and prevents translation in bacteria (Figure B.11) [189]. Recently, six pairs of RNA-IN/OUT variants were designed to be orthogonal, or independently acting, by rationally mutating the sequences that initiate binding [176]. We examined two of these pairs with a truncated form of RNA-OUT (first 67 nt) [189, 190] using in-cell SHAPE-Seq.

We began by characterizing the in-cell structures of RNA-IN and RNA-OUT individually. Our first observation was that the nucleotides in the RNA-IN S4 5′ UTR were highly reactive and likely unstructured in the cell (Figure 3.4a). In addition, the RBS region was found to have intermediate reactivities that were similar in magnitude to the riboregulator ON-state RBS reactivities (Figure 3.3c). For RNA-OUT A4, in-cell SHAPE-Seq reactivities clearly reflected a hairpin structure with a large, highly reactive, loop at the site of RNA-IN recognition (Figure 3.4b). As with the loops of crR10 and crR12, the secondary structure model of RNA-OUT constrained with in-cell reactivity data showed a much larger loop than previously suggested [190]. Similar results were obtained for the S3/A3 RNA-IN/OUT pair analyzed individually (Figure B.12).

Figure 3.4: In-cell structure-function characterization of the RNA-IN/OUT translational repressor system. Color-coded reactivity spectra of RNA-IN S4 (a), RNA-OUT A4 (b), and RNA-IN S4 C24A A25C with RNA-OUT A4 or the antisense control plasmid (c) represent averages over three independent in-cell SHAPE-Seq experiments. Error bars represent one standard deviation. All secondary structures are color-coded by reactivity intensity. (a) Reactivity spectrum of the first 60 nts of RNA-IN S4 (top), with nucleotides color-coded by reactivity on a single-stranded structural model of this region (bottom). RBS and start codon (AUG) are boxed. (b) Reactivity spectrum of RNA-OUT A4 (top), with a minimum free energy structure generated by RNAstructure [164] using in-cell SHAPE-Seq reactivity data as constraints (bottom; see Section 3.3). The terminators following RNA-OUT A4 were not included in structural analysis. (c) Reactivity maps of RNA-IN S4 C24A A25C expressed with RNA-OUT A4 or an antisense control plasmid are on the left. Average fluorescence (FL/OD) values (normalized to the S4 C24A A25C with antisense control plasmid FL/OD value) on the right show 69% repression of gene expression when RNA-OUT A4 is expressed, with error bars representing one standard deviation. The RBS and start codon (AUG) locations are boxed. CDS = coding sequence.


ρ (reactivity)
0-0.25 0.25-0.5 0.5-0.75 0.75-1.25 1.25-2.0 2.0+
a
RNA-IN S4 alone
RBS
AUG
CDS
ρ (reactivity)
Nucleotide Position (5' to 3')
5'-GCCAAAAAUCAAUAAGGAGACAACAAGAUGUGCGAACUCGAUGCUAGCAAAGGAGAAGAA·····-3'
b
RNA-OUT A4 alone
ρ (reactivity)
Nucleotide Position (5' to 3')
c
RBS
AUG
CDS
S4 C24A A25C with control
S4 C24A A25C w/ A4
ρ (reactivity)
Normalized FL/OD
Nucleotide Position (5' to 3')

3.4.6 Characterizing the cellular RNA interactions and function of the IS10 translational repressor

We then characterized how RNA-OUT binding to RNA-IN leads to translation repression by performing the full in-cell SHAPE-Seq structure-function measurement in *E. coli* cells expressing the RNA-IN reporter construct with the RNA-OUT antisense construct. Initially, we performed three replicate experiments with the S4/A4 pair, but observed varying RNA-IN reactivity patterns, despite each replicate exhibiting roughly the same level of translation repression (~85%) (Figure B.13). A closer analysis of the raw SHAPE-Seq (+) and (-) channel fragment distributions revealed large spikes at position 26 in both channels, suggesting that RNA-IN S4 was being cleaved between positions 25 and 26 when in complex with RNA-OUT A4. To further confirm this effect was due to cognate RNA-RNA interactions, we examined orthogonal pairs of RNA-IN/RNA-OUT (i.e., pairs S4/A3 and S3/A4) and found no spikes at position 26 or major changes in reactivity compared to the individually measured RNAs (Figures B.14 and B.15).

Previous work showed that the wild-type RNA-IN/RNA-OUT duplex is targeted by RNAse III between nucleotides 15-16 of RNA-IN and 22-23 of RNA-OUT for degradation in the cell [191]. However, we did not observe spikes at these positions due to mutations introduced at positions 16 and 17 of RNA-IN that form bulges in the RNA-IN/RNA-OUT complex and abolish RNAse III cleavage [176]. Given the propensity for the wild-type system to be cleaved by RNAse III, we hypothesized that a secondary RNAse III site was present between nucleotides 25 and 26 that gave rise to the observed spikes in the fragment distributions from the cognate complexes. To test this hypothesis, we made two different mutations (C24A and A25C) to RNA-IN S4 to prevent RNAse III cleavage (Figure B.16) and tested them using in-cell SHAPE-Seq.

We observed that both mutants were still functional and neither generated a fragment spike at position 26 when expressed with RNA-OUT A4, indicating that cleavage was abolished by these changes. We also tested a double mutant version that functioned similarly (Figure B.17).

To characterize the cellular RNA-RNA interactions that lead to translation repression, we performed replicate in-cell SHAPE-Seq experiments with the RNA-IN S4 C24A A25C double mutant and RNA-OUT A4 (Figure 3.4c). Several notable features are apparent when comparing the RNA-IN reactivity spectra with and without RNA-OUT. First, there is a drop in the reactivity spectrum for the first seven nucleotides of RNA-IN where RNA-OUT is predicted to initiate binding, similar to what we observed for the riboregulators (Figure 3.3a), corresponding to a 69% decrease in measured fluorescence. Second, we observed reactivity increases at positions 16 and 17 in the RBS of RNA-IN, which are predicted to form a bulge when in complex with RNA-OUT (Figure B.16). We also observed slight increases in reactivity across the CDS and start codon when translation is repressed (Figure 3.4c). Interestingly, we did not observe a drop in reactivity in the RBS in the presence of RNA-OUT as we might expect, but rather a few nucleotides that increase (Figure B.18). It could be the case that the interaction between the 5' end of RNA-IN with the loop of RNA-OUT brings the two RNAs close enough together to hinder ribosome access without directly binding the RBS. We also note that the consistently high reactivities in nucleotides 11-13 are unexpected, suggesting that the duplex between RNA-IN and RNA-OUT may not be as extensive in the cell as originally thought.

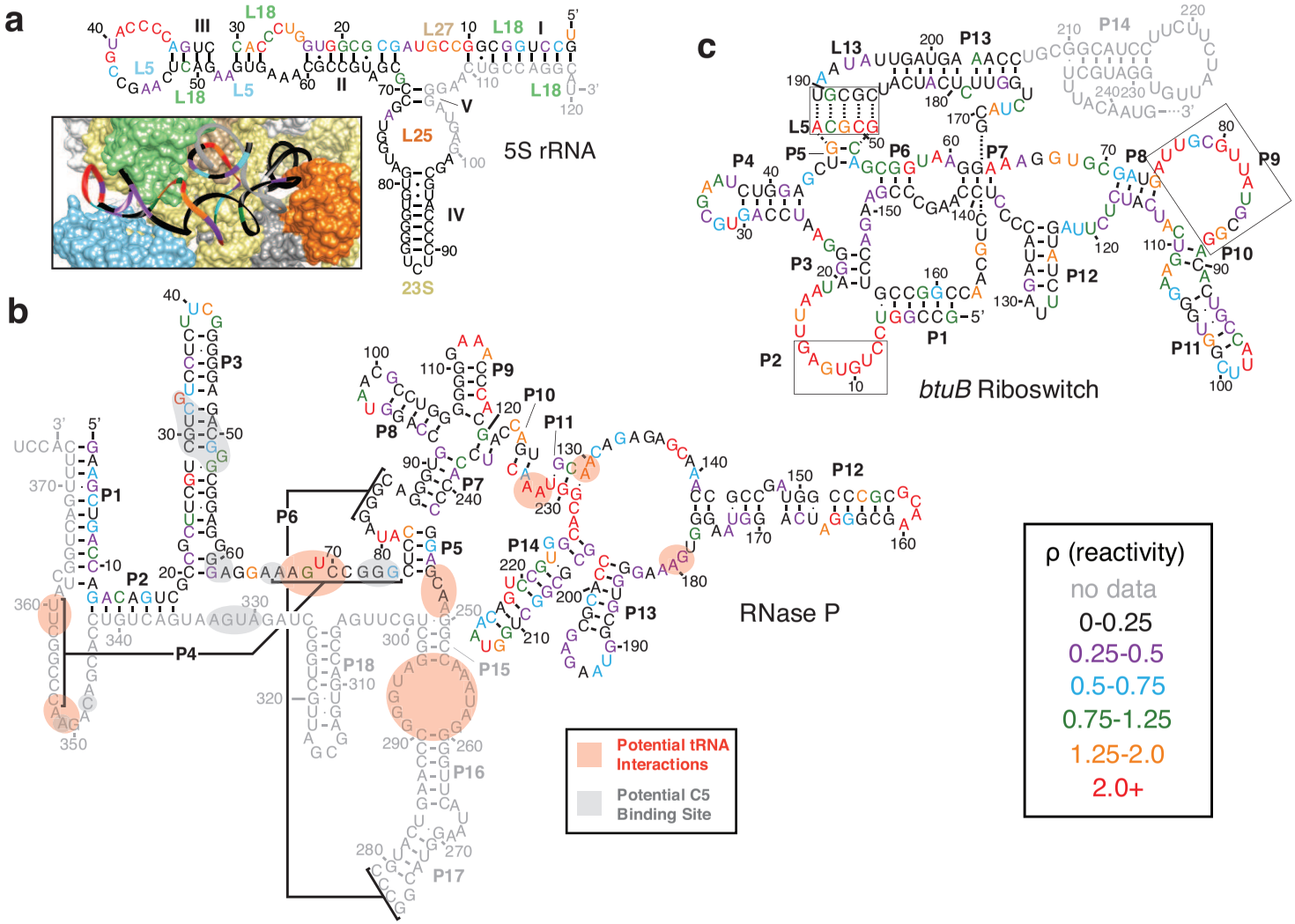
Finally, we examined reactivity changes from the perspective of the antisense RNA-OUT RNAs (Figure B.19). As expected, there are no major differences in the reactivity map of RNA-OUT A4 when the orthogonal RNA-IN S3 is present. However, unlike in

the riboregulator system, we did observe subtle reactivity changes in RNA-OUT A4 in the presence of RNA-IN S4 C24A A25C, despite the copy number difference.

3.4.7 Targeting Endogenous RNAs in *E. coli*

To further test the capabilities of in-cell SHAPE-Seq, we targeted three endogenously expressed functional RNAs that are present at varying levels in *E. coli* cells: 5S rRNA, RNase P, and the *btuB* mRNA riboswitch domain (Figure 3.5). 1M7 probing of *E. coli* cell cultures was performed as before, except that sequence specific RT primers were used for each endogenous target. For the highly abundant 5S rRNA the experiment was straightforward, as the level of cDNA obtained was similar to the synthetic RNAs expressed from plasmids. For the less abundant RNase P and *btuB* riboswitch RNAs, however, it was necessary to modify the PCR steps to prevent the amplification of unwanted side products that began accruing when the amount of correct cDNA product was low and more than 15 cycles of PCR were used. We determined that the side products were due to the Illumina forward primer (primer I in Table B.3). To remedy this, we first amplified the ssDNA libraries without this primer for 15 cycles to amplify the target of interest, then added primer I for another 15 cycles to build the rest of the adapter required for sequencing (see Section 3.3). We confirmed the additional cycles did not alter the resulting reactivities (Figure B.20).

Figure 3.5: Structural characterization of three endogenously expressed RNAs in *E. coli* with in-cell SHAPE-Seq. RNA secondary structures are color-coded by average in-cell SHAPE-Seq reactivity intensity according to the key in the lower right. Nucleotides not included in the reactivity calculation are marked in gray. Bar charts depicting the average reactivities of each RNA can be found in Figure B.21. (a) 5S rRNA. Reactivities overlaid on the accepted secondary structure [192] and an atomic resolution model of the ribosome derived from cryo-EM data fit with molecular dynamics simulations (inset; from PDB 4V69) [193]. Individual ribosomal proteins (L5, L18, L25, L27) and the 23S rRNA are labeled on the secondary structure near their approximate locations and are color-coded to match the three dimensional model. Helices are numbered I-V. (b) RNase P. Reactivities overlaid on the accepted secondary structure derived from comparative sequence analysis [194]. Potential interactions with tRNAs are highlighted with pink shading according to the crystal structure of the related A-type *T. maritima* RNase P in complex with $tRNA^{phe}$ [195]. Similarly, the expected interactions with the C5 protein measured from hydroxyl-radical mediated cleavage interactions [196] are indicated with gray shading. Helices P1-P18 are labeled. (c) *btuB* riboswitch domain. Reactivities overlaid on secondary structure model [197, 198]. Boxes indicate regions where the structural model was refined according to the high reactivities observed by opening base pairs in those regions. Dashed lines indicate a predicted pseudoknot between L5 and L13 according to the model, though high reactivities are observed in L5 in the cell.


a
5S rRNA
23S
b
RNase P
Potential tRNA Interactions
Potential C5 Binding Site
c
btuB Riboswitch
ρ (reactivity)
no data
0-0.25
0.25-0.5
0.5-0.75
0.75-1.25
1.25-2.0
2.0+

We first examined the highly abundant 5S rRNA [199]. As seen in Figure 3.5a, we observed strong agreement between in-cell SHAPE-Seq reactivities (Figure B.21a) and the accepted secondary structure and an atomic resolution model of 5S within the ribosome [192, 193, 200]. Reactivities for the 5S rRNA appeared high in loop regions as expected, except when in close proximity to, or bound locally by, ribosomal proteins such as L5, L18 and L25. Positions 70-99 were very low in reactivity, which is consistent with helices IV and V being threaded into the interior of the ribosome and the inner loop between helices IV and V interacting with protein L25. We did notice one discrepancy in which nucleotides 28-30 are observed to be highly reactive even though they are predicted to be paired with nucleotides 54-56. In this region however, the 5S rRNA appears to be distorted with nucleotides 54-56 positioned near the L5 protein.

We then characterized the reactivities of RNase P, a ribozyme that complexes with a protein cofactor (C5) to cleave the 5′-leader from precursor tRNAs (pre-tRNAs) [201]. The RNase P RNA (RPR) consists of two domains: a catalytic and a specificity domain. We largely focused our analysis on the latter. We found strong agreement between the measured in-cell SHAPE-Seq reactivities (Figure B.21b) and the secondary structure of the *E. coli* RPR derived from comparative sequence analysis [194] (Figure 3.5b). Specifically, there is concurrence between highly reactive positions and nucleotides expected to be unpaired in the secondary structure. Because the binding sites for the C5 protein are largely in structured regions or regions not probed (for instance, helices P3 and P4) [195, 196, 202], it is difficult to attribute low reactivities that arise in these regions specifically to protein-RPR interactions.

Also shown in Figure 3.5b are potential sites for tRNA recognition based on the crystal structure of the related A-type Thermotoga maritima RNAse P in complex with $tRNA^{phe}$ [195]. Interestingly, we observe several features in this region suggesting that

our experiments likely captured the substrate-bound form of RNase P *in vivo*. First, we observe very low reactivity at position A180, which is expected to stack directly with the nucleotides in the T-loop of the pre-tRNA to enable substrate recognition [195, 203]. Second, we observe low reactivity at position A248, which stabilizes the RPR-pre-tRNA complex through stacking interactions with the pre-tRNA [195]. Finally, we observe very high reactivity at position U69, a universally-conserved nucleotide, which is unstacked from pseudoknot P4 to coordinate one of the two divalent metal ions needed for pre-tRNA cleavage [195]. Collectively, these observations suggest that our probing experiments have captured the substrate-bound form of RNase P *in vivo*, which could be expected given the large number of pre-tRNAs that need to be processed by RNase P, a low copy-number enzyme [204].

To further test the sensitivity of in-cell SHAPE-Seq, we targeted the endogenously expressed riboswitch domain of the *btuB* mRNA, which regulates the translation of the cobalamin transport protein BtuB in bacteria by sequestering its RBS when adenosylcobalamin (AdoCbl) is present [197]. In-cell SHAPE-Seq reactivities (Figure B.21c) were largely consistent with a secondary structure model of the *btuB* riboswitch derived from comparative sequence analysis and structural probing [197] (Figure 3.5c). We did, however, observe high reactivities in several areas that are predicted to be paired according to the model. Specifically, the nucleotides comprising the P2 and P9 helices were observed to be highly reactive, indicating that they are unstructured in the cell. In the case of P9, this would suggest this region is disordered as was observed in the crystal structure of the *T. tengcongensis* AdoCbl riboswitch (TteAdoCbl) [198]. Most interesting are the high reactivities observed in the loop of P5 (L5), which is expected to form a kissing-loop (KL) interaction with the loop of P13 (L13). Recently, it was shown that this KL interaction is a critical regulatory feature of AdoCbl riboswitches, and crystal structures of the TteAdoCbl riboswitch showed that bound AdoCbl inter-

acts with the groove of the KL in a structure-specific way that promotes its formation [198]. While the in-cell SHAPE-Seq reactivities of L13 were observed to be low, the very high reactivities in L5 suggest that there is a significant population of *btuB* riboswitches that are unbound by AdoCbl, or that the KL interaction is flexible enough to allow the riboswitch to significantly sample the non-KL configuration. Additional in-cell SHAPE-Seq analysis on functionally variant mutants of this system would help shed further light on the cellular structural state of this riboswitch.

Overall, these results indicate that in-cell SHAPE-Seq can be used to obtain nucleotide-resolution reactivity maps for endogenous transcripts directly in *E. coli* cells. Our range of examples demonstrate that these reactivity spectra can be used to corroborate existing models of RNA folding and interactions, as well as suggest refinements to our understanding of RNA systems that are less well studied. We thus anticipate in-cell SHAPE-Seq to be useful for the study of a broad array of endogenous RNAs.

3.4.8 Comparing *in vitro* and in-cell SHAPE-Seq reactivities

Our ability to characterize cellular RNA structures with in-cell SHAPE-Seq gave us an opportunity to compare our results with reactivities generated with *in vitro* SHAPE-Seq experiments [122] to study how the cellular environment affects RNA structure. To begin this study, we performed equilibrium refolding SHAPE-Seq v2.0 experiments on the riboregulators and the RNA-IN/OUT systems following our previously published protocol using the same RT primers as the in-cell experiment [122]. Interestingly, we found remarkable agreement between in-cell and *in vitro* refolded SHAPE-Seq reactivities for the riboregulators (Figure B.22) and the RNA-IN/OUT systems (Figure B.23).

In many cases, the trends in reactivities across the molecules were consistent, with quantitative differences at isolated positions. The biggest deviations were seen when we examined the RNA-IN/OUT complex, which showed significantly lower in-cell reactivities in the region surrounding the RBS of RNA-IN (Figure B.23b). Overall, the similarity between the in-cell and *in vitro* refolded SHAPE-Seq reactivities suggests that for these regulatory RNAs the complex cellular environment does not play a significant role in altering structures from their equilibrium states.

Next, we performed similar in-cell vs. *in vitro* SHAPE-Seq experimental comparisons for 5S rRNA, which is routinely used as a benchmark for *in vitro* RNA folding (Figure B.24) [122]. In contrast to the above results, we observed dramatic differences in reactivities between these two conditions. In particular, large reactivity differences were observed at positions 35-54 near the site of L5 interactions (Figure 3.5a) [200]. In addition, almost all peaks that are highly reactive downstream of position 54 *in vitro* are near zero in-cell. All of these changes visible in the in-cell reactivity spectra reflect a structural state of the 5S rRNA docked into the ribosome (Figure 3.5a inset). It is thus clear that the cellular environment can significantly alter the folding of certain RNAs.

3.4.9 Discussion

In this work, we established in-cell SHAPE-Seq, which was designed to characterize the cellular RNA structure and function of a set of RNAs in a single experiment. With the coupling of structure and function measurements, we showed how we can use in-cell SHAPE-Seq to directly correlate changes in cellular RNA structure with changes in cellular function in bacteria. The development of in-cell SHAPE-Seq required a number of technical modifications of *in vitro* SHAPE-Seq, including the use of highly specific

reverse transcription priming sites to target select RNAs, PCR selection against ligation dimers and off-target cDNAs (Figure B.3), and a flexible platform for rapid functional characterization of RNA regulators in *E. coli*. All of these improvements enabled deep read coverage for many in-cell SHAPE-Seq experiments in a single MiSeq run with less effort than current in-cell next-generation sequencing-based techniques [114–118], partly because we removed the need for gel purification in the library construction process. We used these improvements to report some of the first detailed replicate in-cell RNA structure chemical probing data, which we anticipate will be important to the field for understanding the variability of cellular RNA structural states.

We demonstrated the capabilities of our in-cell SHAPE-Seq technique for studying RNA structure-function by applying it to two different RNA regulatory systems: the synthetic riboregulator translational activator [25] and the RNA-IN/OUT translational repressor [176]. Each system consists of a pair of RNAs - a sense 5′ UTR containing the RBS of a downstream gene and an antisense RNA that targets the sense RNA to cause structural rearrangements near the RBS, leading to changes in gene expression. In general, we observed that the in-cell SHAPE-Seq reactivity spectra of the isolated sense and antisense RNAs agreed well with the structural models for both systems. For example, the reactivity patterns clearly reflect the hairpin nature of the antisense taRNAs (Figures 3.2a and B.6a), the sense crRNAs (Figures 3.2b and B.6b), and RNA-OUT (Figures 3.4b and B.12b). Interestingly, the loops of the crRNA and RNA-OUT hairpins exhibited a larger span of high reactivities than expected. By constraining computational folding algorithms with in-cell SHAPE-Seq data, we generated structural models that suggested these loops are more unstructured in bacterial cells than originally predicted (Figures 3.2b and 3.4b) [25, 176]. The extensive clusters of high reactivities in these RNAs may actually be an important feature for RNA-RNA recognition, as both loops are involved in initiating interactions between the sense and antisense RNAs of

their respective systems.

We also observed low reactivities in the CDS of both sense RNAs in all conditions tested. However, there are many potential causes for low SHAPE reactivity values in these regions including: structures within the CDS, cellular protein binding, or the presence of translating ribosomes. In contrast, the transcriptome-wide structural analysis performed by Rouskin *et al.* indicated that translating ribosomes were not associated with lower reactivities, although their experiment was performed in *S. cerevisiae*, not *E. coli* [117]. Ding *et al.* alternatively observed a three-nucleotide periodic reactivity pattern in coding sequences across the *Arabidopsis* transcriptome. Although we did not observe any such periodic pattern, our experiments were performed in a different organism and we focused on specific RNAs rather than averaging reactivity signatures over large windows [116].

When antisense RNAs were co-expressed with the matching sense RNAs, we found substantial reactivity changes that could be directly linked to functional changes in gene expression. In the riboregulator system we observed reactivity increases in the RBS that correlated with an increase in SFGFP expression (Figure 3.3c, d). We also detected other changes in the crRNA reactivity map that led us to refine the model of taRNA/crRNA interactions (Figure 3.3b). In the RNA-IN/OUT system, this analysis was initially complicated by our discovery of a double-stranded RNAse cleavage site in RNA-IN based on analysis of the raw in-cell SHAPE-Seq fragment alignments (Figure B.13). Thus, we mutated RNA-IN to remove the cleavage site and performed the in-cell SHAPE-Seq experiment on the cleavage-resistant double mutant and found it exhibited a normal fragment distribution (Figure B.17). Structurally, we observed reactivity decreases that corresponded to RNA-IN/OUT complex formation, as well as reactivity increases that implied the complex is less structured in parts than the

proposed mechanism would suggest (Figure 3.4c) [176]. We note that changes in RBS reactivity between the two functional states of the RNA-IN/OUT system were not as clear as those for the riboregulators (Figure 3.4c and Figure B.18). However, we did detect an interaction at the 5′ end of RNA-IN, which could serve to bring RNA-OUT close enough to hinder ribosome access without directly binding the RBS. All together, our in-cell SHAPE-Seq reactivity data speaks to the fact that RNA structures typically exist in an ensemble and suggests that different RNA structural states can give rise to similar functional outputs.

We also demonstrated that in-cell SHAPE-Seq could be used to characterize endogenous bacterial RNAs expressed at a range of levels. In particular, we showed that in-cell SHAPE-Seq reactivities recapitulated many of the structural features and interactions of two well-studied RNAs that interact with known proteins: 5S rRNA and RNase P (Figure 3.5a, b). An additional study of the *btuB* riboswitch suggested interesting refinements to the covariation/*in vitro* probing-based structural model that could reflect differences in folding due to the cellular environment (Figure 3.5c). To obtain these reactivity spectra, we needed to modify the PCR steps of our library preparation strategy in order to improve selectivity and prevent undesired DNA from dominating the libraries. With minor modifications, we were able to obtain a robust in-cell SHAPE-Seq method that should be applicable to studying a broad range of endogenously expressed RNAs. This could be a particular advantage of the targeted in-cell SHAPE-Seq approach, especially for lowly expressed RNAs, since transcriptome-wide approaches do not capture low abundance transcripts as well, as they inherently distribute reads across a large number of targets. We note that both targeted and transcriptome-wide approaches have distinct advantages and can be viewed as complementary methods to study cellular RNA structure-function principles.

Finally, this work enabled us to study how the complex cellular environment can affect RNA folding. This was most clear in a comparison between SHAPE-Seq reactivities from *in vitro* equilibrium and in-cell experiments on 5S rRNA (Figure B.24), where a large number of changes were observed that matched well with the known interactions of 5S rRNA within the ribosome (Figure 3.5a). Thus, we found that the cellular environment can significantly affect RNA folding, even for highly expressed RNAs. A similar comparison for the synthetic riboregulator and RNA-IN/OUT systems showed the opposite, with strong agreement observed between *in vitro* and in-cell reactivities (Figures B.22 and B.23). While these systems are designed to interact with ribosomes in the cell, these interactions may be too fleeting, or not present at high enough abundance, to be detected within the population of RNAs probed in these experiments, as was the case with the antisense RNAs for these systems (Figures B.9 and B.19). Consistent with our results, while this manuscript was under review, a complementary in-cell SHAPE probing technique called icSHAPE was used to show that the agreement between *in vitro* and in-cell RNA folds was closer than previously expected, especially near translation initiation regions [115]. This intriguing agreement could reflect the robustness of the biophysics of RNA folding to environmental perturbations and warrants further study.

We anticipate in-cell SHAPE-Seq to be applicable to studying cellular RNA structure-function relationships within a broad array of mechanistic and cellular contexts, including other organisms beyond *E. coli* such as *S. cerevisiae*, *M. musculus*, or *A. thaliana* [114–116, 118]. While we focused on regulatory systems containing two RNAs and several endogenously expressed RNAs, the inherent multiplexing and accuracy of SHAPE-Seq [108, 122] allows many RNAs to be measured at once, enabling the study of larger mixed populations of cellular RNAs. In its current form, in-cell SHAPE-Seq could be immediately applied to study a host of RNA regulators including ligand-

sensing riboswitches, ribozymes, bacterial small RNAs, and other RNAs that affect aspects of gene expression [24]. In addition, performing in-cell SHAPE-Seq experiments alongside *in vitro* SHAPE-Seq experiments offers a way to reveal interactions and structural changes that may be present in the cellular environment as we demonstrated with 5S rRNA. Further, we have provided a detailed step-by-step protocol in the Appendix B.4 to facilitate the application of in-cell SHAPE-Seq to other systems, including RT primer design guidelines. We expect that in-cell SHAPE-Seq will be an easily approachable tool for biologists and engineers to uncover relationships between the sequence, structure, and function of RNAs in the cell.

3.5 Acknowledgements

We thank Peter Schweitzer and the Cornell Life Sciences Core facility for sequencing support during this work and Nicole Ricapito for assistance with synthesizing SHAPE reagents. We also thank Alfonso Mondragón (Northwestern University) for suggesting RNase P as an endogenous target and Venkat Gopalan and Lien Lai (Ohio State University) for critical help in interpreting the in-cell RNase P experimental data.

3.6 Funding

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program [Grant No. DGE-1144153 to K.E.W.], the Cornell University College of Engineering 'Engineering Learning Initiatives' Undergraduate Research Grant Program [to T.R.A.], the Defense Advanced Research Projects Agency Young Faculty Award (DARPA YFA) [N66001-12-1-4254 to J.B.L.], and a New Innova-

tor Award through the National Institute of General Medical Sciences of the National Institutes of Health [grant number 1DP2GM110838 to J. B. L.]. K.E.W. is a Fleming Scholar in the School of Chemical and Biomolecular Engineering at Cornell University. J.B.L. is an Alfred P. Sloan Research Fellow. Funding for open access charge: National Institutes of Health/1DP2GM110838.

CHAPTER 4

CHARACTERIZING RNA STRUCTURES *IN VITRO* AND *IN VIVO* WITH SELECTIVE 2′-HYDROXYL ACYLATION ANALYZED BY PRIMER EXTENSION SEQUENCING (SHAPE-SEQ)

4.1 Abstract

RNA molecules adopt a wide variety of structures that perform many cellular functions, including, among others, catalysis, small molecule sensing, and cellular defense. Our ability to characterize, predict, and design RNA structures are key factors for understanding and controlling the biological roles of RNAs. Fortunately, there has been rapid progress in this area, especially with respect to experimental methods that can characterize RNA structures in a high throughput fashion using chemical probing and next-generation sequencing. Here, we describe one such method, selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq), which measures nucleotide resolution flexibility information for RNAs *in vitro* and *in vivo*. We outline the process of designing and performing a SHAPE-Seq experiment and describe methods for using experimental SHAPE-Seq data to restrain computational folding algorithms to generate more accurate predictions of RNA secondary structure. We also provide a number of examples of SHAPE-Seq reactivity spectra obtained *in vitro* and *in vivo* and discuss important considerations for performing SHAPE-Seq experiments, both in terms of collecting and analyzing data. Finally we discuss improvements and extensions of these experimental and computational techniques that promise to deepen our

The work described in this chapter includes contributions from Angela M Yu, Eric J. Strobel, Alex H. Settle, and Julius B. Lucks. This work was originally published in *Methods* and has been reproduced here with permission from Elsevier. Kyle E. Watters, Angela M Yu, Eric J. Strobel, Alex H. Settle, and Julius B. Lucks, Characterizing RNA structures *in vitro* and *in vivo* with selective 2-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq), *Methods*, 2016, doi:10.1016/j.ymeth.2016.04.002.

knowledge of RNA folding and function.

4.2 Introduction

The dual informational/structural nature of RNA molecules allows them to simultaneously encode genetic information and actively direct cellular processes. Many RNAs assume highly sophisticated structures that mediate a diverse set of functions. These functions range from catalysis, as in the case of RNA enzymes like RNase P [205] and the ribosome [206], to a diverse and expanding array of regulatory mechanisms including riboswitches [27, 28], RNAi [207], RNA transcriptional attenuators [208], CRISPR [209], thermometers [210], and many others. A large number of these RNAs are non-coding (ncRNA) and function in a purely structural manner without carrying genetic information [22, 173, 211]. Our understanding of the importance of these functional ncRNAs is increasing and many more continue to be discovered at a rapid pace [211]. Thus, the development of tools to quickly and accurately characterize the structure-function relationships of ncRNAs is essential to advancing the field of RNA biology.

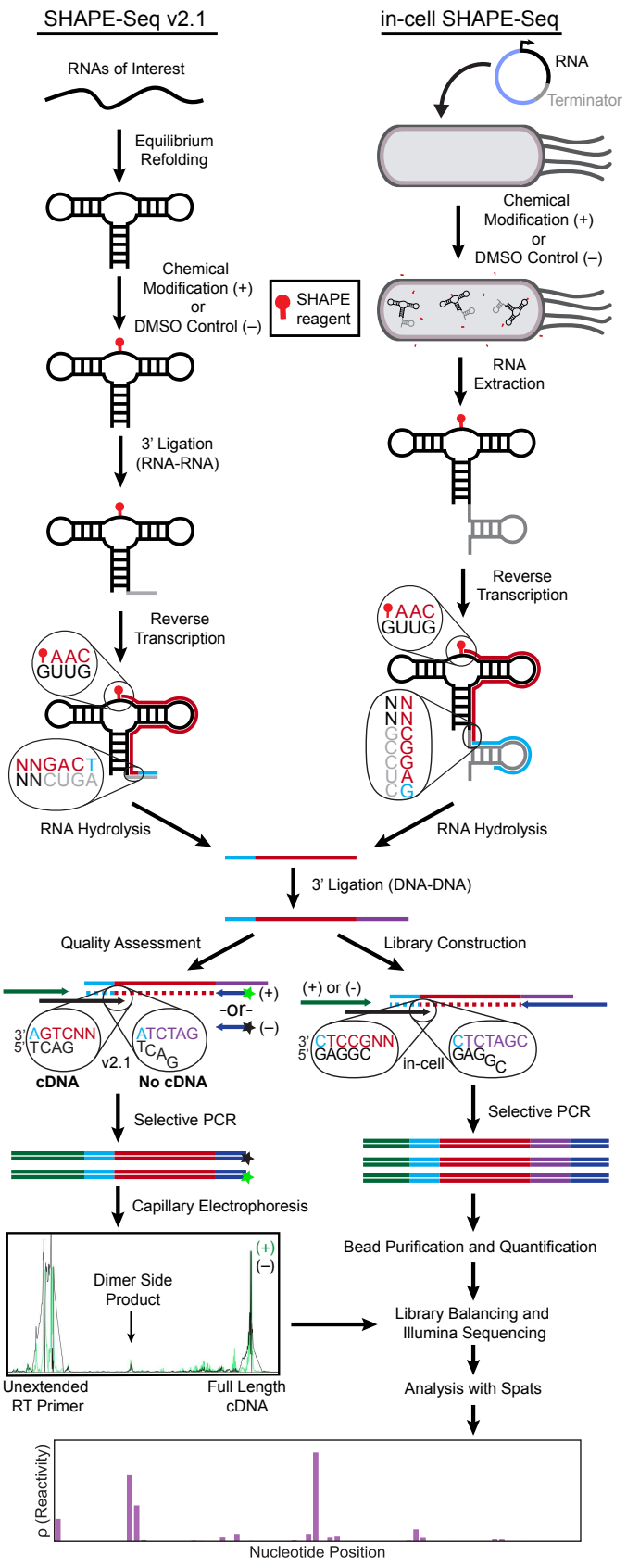
A common method of characterizing RNA structure is to isolate the RNA of interest *in vitro* and perform enzymatic or chemical probing experiments that reveal information about an RNA molecules secondary and tertiary structure [86, 87]. These experiments interrogate RNA structures by measuring nucleotide accessibility for RNase cleavage or chemical modification and can be used to infer whether a nucleotide within an RNA molecule is predominantly single- or double-stranded [86]. Chemical probes have become more frequently used due to their higher resolution and ability to transport across membranes to react with RNAs inside cells. These probes use a range of chemistries to covalently modify RNAs in a structure-dependent fashion and can be

roughly divided into three classes [86]: 1) base-specific probes such as dimethyl sulfate (DMS), 1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene sulfonate (CMCT), and kethoxal [212], 2) backbone-cleaving reagents such as hydroxyl radicals [93] and metal ions [213], and 3) SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension) reagents that modify the 2'-OH of the RNA backbone [86, 95].

Chemical probes of RNA structure react with specific nucleotide positions that are solvent-accessible, or flexible, to covalently modify them. Modification positions are typically mapped by reverse transcription (RT), which either stops at these positions [86, 212] or introduces a mutation into the cDNA [112]. An analysis of the resulting cDNAs can then be used to determine modification frequency at each nucleotide position. Modern approaches couple chemical probing and RT with next-generation sequencing (NGS) to directly sequence the cDNA products and determine modification positions [107, 112–118, 122, 156]. The use of NGS immediately enables these experiments to be highly multiplexed, allowing up to thousands of RNAs to be probed and analyzed in a single experiment. These extensive datasets [169] can be analyzed in many ways and are routinely used to restrain RNA structure prediction algorithms [214].

The recent transition to NGS-based methods has also been accompanied by a second transition in the field: the move from characterizing RNA structure *in vitro* in favor of *in vivo*. Chemical probing *in vivo* requires that the probe be able to quickly diffuse across membranes to modify RNAs inside the cell. While DMS has long been known to penetrate cells [215], it was also recently shown that some members of the SHAPE family of reagents can do so as well [126, 158]. Shortly after the first NGS-based chemical probing method was published [107], a number of different approaches combined NGS with *in vivo* or *in virio* chemical modification [112–118, 122, 156], many of which were to designed to probe the transcriptome of their respective organism [114–118].

Figure 4.1: SHAPE-Seq workflow. *In vitro* RNA structures are analyzed by first purifying RNAs of interest, refolding in an appropriate buffer with optional ligands, and modifying with a SHAPE reagent (+) or a control solvent (-). In-cell probing experiments modify RNAs within the cell after the SHAPE reagent or control solvent is added to the media. RT is initiated in one of two ways. In SHAPE-Seq v2.1, an RNA linker sequence is ligated to RNAs that serves as an RT priming site, whereas in-cell probed RNAs are extracted then primed directly with an internal RT priming site, such as the intrinsic terminator shown. Insets show specific 5′ and 3′ cDNA sequences, the latter of which is used to identify the SHAPE modification position. After RT, all SHAPE-Seq steps are similar. After RNA hydrolysis, a DNA adapter required for Illumina sequencing is ligated to the 3′ cDNA ends. Selective PCR is then used to generate quality assessment or sequencing libraries. The selective PCR uses a selection primer (black) designed to bridge the RT priming site and the 5′ end of the extended cDNA. This allows efficient PCR amplification only if the RT reaction created cDNA extensions. If no cDNA was synthesized, the selection primer cannot bind properly to the RT primer-adapter side product junction (inset), limiting amplification. Quality assessment libraries use a reverse primer that is fluorescently labeled for analysis with capillary electrophoresis, while sequencing libraries use the full Illumina adapter sequence. Subsequent sequencing and bioinformatic analysis (Spats) of the (+) and (-) libraries generates the characteristic SHAPE-Seq reactivity spectra.


SHAPE-Seq v2.1
in-cell SHAPE-Seq
RNAs of Interest
RNA
Terminator
Equilibrium Refolding
Chemical Modification (+) or DMSO Control (–)
SHAPE reagent
Chemical Modification (+) or DMSO Control (–)
RNA Extraction
3' Ligation (RNA-RNA)
Reverse Transcription
Reverse Transcription
AAC GUUG
NNGACT NNCUGA
NN NN GCGCUCG
RNA Hydrolysis
RNA Hydrolysis
3' Ligation (DNA-DNA)
Quality Assessment
Library Construction
(+)
-or-
(–)
(+) or (-)
3' AGTCNN 5' TCAG
ATCTAG TCAG
v2.1
cDNA
No cDNA
3' CTCCGNN 5' GAGGC
CTCTAGC GAGGC
in-cell
Selective PCR
Selective PCR
Capillary Electrophoresis
Bead Purification and Quantification
Dimer Side Product
(+)
(–)
Library Balancing and Illumina Sequencing
Unextended RT Primer
Full Length cDNA
Analysis with Spats
ρ (Reactivity)
Nucleotide Position

The growing number of NGS-based techniques have many basic steps in common including chemical modification, reverse transcription, and PCR for library preparation (Figure 4.1) [107, 112–118, 122, 156]. However, there are many details involved in several steps of these methods that present challenges. Beyond the continuous need for greater sequencing depth at lower costs, reducing the time, effort, complexity, and cost of the NGS library preparation steps are the biggest barriers to wide adoption of NGS-based chemical probing methods. For example, many NGS-based methods use gel purification steps that are typically time intensive and reduce cDNA yield. In previous work, we simplified the process for *in vitro* studies by describing SHAPE-Seq v2.0 [122], which reduced library preparation time and added a 'universal' RNA ligation method to characterize short RNA sequences by priming RT from a 3' ligated linker sequence post-modification. We also developed in-cell SHAPE-Seq to characterize the structures of small groups of targeted RNAs directly inside cells [113]. Although not originally designed to cover the entire transcriptome, in-cell SHAPE-Seq avoids gel purification steps, making it much quicker and more amenable to first time users of *in vivo* NGS-based chemical probing methods. Overall, in-cell SHAPE-Seq is an approachable technique for RNA biologists interested in studying a few select RNAs rather than the whole transcriptome [113].

In this work, we outline approaches for characterizing RNA structure *in vitro* and *in vivo* using SHAPE-Seq v2.1 and in-cell SHAPE-Seq, respectively. SHAPE-Seq v2.1 further upgrades the v2.0 technique to incorporate more flexible library barcoding capabilities and a re-engineered linker sequence to permit selective amplification of cDNA sequences as inspired by in-cell SHAPE-Seq [113]. In addition to describing experimental aspects of these techniques, we cover computational approaches that incorporate the resulting SHAPE-Seq reactivity data to improve the prediction of RNA structure models. We also highlight how comparing reactivity differences with and without

ligand or *in vitro* vs. *in vivo* can reveal information about the effects of the folding environment on RNA structure. We discuss important considerations for performing SHAPE-Seq experiments and restraining computational predictions. Finally, we also suggest future improvements for the SHAPE-Seq technique and the interpretation of its reactivity data.

4.3 SHAPE-Seq Background

The SHAPE-Seq protocol consists of several core steps. These include: RNA chemical modification, converting RNA into cDNA with reverse transcription, sequencing the cDNA, bioinformatically processing sequencing reads (Figure 4.1), and, optionally, using the reactivities to restrain computational folding algorithms. Here we discuss the approaches and relevant background for each step before covering them in greater detail in Section 4.4.

4.3.1 RNA modification

To begin, the RNA of interest and its proper folding environment need to be determined. For *in vitro* studies, this involves determining the proper folding buffer and conditions (times, temperatures, ligand concentrations, etc.). For *in vivo* studies, the main choices to consider are which organism is being examined and whether endogenous or exogenous RNAs (e.g. plasmids, etc.) are being targeted. Once determined, the RNA is first folded in the chosen environment then treated with a SHAPE reagent (+), or a solvent control (-), by adding it directly to the *in vitro* solution or the cell culture media (Figure 4.1). The modified RNAs are then prepared for RT according to the type

of experiment being performed. For example, *in vivo* studies, or *in vitro* studies with proteins, first require a two-phase extraction to remove proteins and/or contaminating DNA. For *in vitro* experiments where internal RT priming is not convenient, a ligation step introduces extra RNA sequence at the 3' end to serve as a priming site [122]. In all cases, the final RNA form is concentrated using ethanol precipitation before RT.

4.3.2 Conversion of RNA to cDNA with reverse transcription

For both *in vitro* and *in vivo* experiments, RT and the steps that follow are largely the same (Figure 4.1). An RT primer specific to either an internal RNA sequence or ligated linker is added for extension, which stops one nucleotide before the SHAPE modification position. The RNA is then hydrolyzed, followed by ethanol precipitation to remove the base. Next, a DNA adapter is added to the 3' end of the cDNA via a single stranded DNA-DNA ligation. The DNA adapter introduces one of the Illumina sequencing adapters required for DNA amplification and downstream sequencing.

This DNA-DNA ligation step is fairly standard among NGS-based chemical probing techniques [107, 113–118, 122, 156]. It is also one reason that many NGS-based methods require gel purification, as the ligation products tend to be a complex mixture of oligonucleotides that contains a large amount of unwanted side product or starting material that needs to be removed before sequencing library preparation. One recently described solution to the purification problem has been to include an azide group on the modifying reagent, which can then be covalently linked to a biotin moiety via a 'click' reaction for selective pull-down [115].

Alternatively, as part of our in-cell SHAPE-Seq technique, we developed a selective PCR step that only allows significant amplification of correctly ligated products con-

taining some length of transcribed cDNA (Figure 4.1) [113]. Selective PCR removes the need for gel purification and reduces the time, cost, and expertise required to prepare sequencing libraries. As described below, we also altered the SHAPE-Seq v2.0 linker to allow for selective PCR, which was not part of the original v2.0 protocol [122]. Finally, in all SHAPE-Seq methods a final bead purification step is employed to reduce the amount of unligated adapter present.

4.3.3 Preparation for sequencing

To prepare libraries for sequencing, the ssDNA libraries from Section 4.3.2 are amplified with PCR to add the complete Illumina TruSeq adapter sequences on each end. The adapters contain DNA sequences necessary for binding to the flow cell, priming the sequencing reactions, and barcoding the SHAPE-Seq libraries to sequence multiple libraries on a single flow cell. The PCR step requires three oligonucleotides (Figure 4.1). The first is the reverse primer that binds to the ligated DNA adapter and adds a TruSeq index and one of the flow cell binding sequences. The second primer contains the other flow cell binding sequence and the sequencing primer site for the first read of sequencing (RD1). The third, or selection primer, selects against unwanted DNA adapter-RT primer ligation side products during PCR and consists of a combination of the RD1 sequence and a designed sequence specific to the RT primer that extends 2-5 nt into the cDNA (Figure 4.1). In this PCR amplification scheme, the ligation side product formed between the unextended RT primer and the DNA adapter cannot be exponentially amplified due to a 3' overhanging mismatch, providing a mechanism of selection against this side product, which can be present at a high concentration.

After PCR library construction, DNA library quality can be determined using ei-

ther an Agilent BioAnalyzer or similar equipment to verify that the length distribution of the library matches the expected lengths for the RNA(s) tested. Alternatively, we prefer a more sensitive quality control reaction, where the reverse primer is replaced with a shorter, fluorescently labeled version for visualization with capillary electrophoresis (Figure 4.1).

Once the library quality is verified, individual libraries containing different TruSeq indexes are measured for concentration and pooled before sequencing with either an Illumina MiSeq or HiSeq instrument, using short paired-end reads.

4.3.4 Bioinformatic read alignment and reactivity calculation

Sequencing data is converted to chemical reactivity values by first aligning all of the sequences to the RNA(s) being studied to establish profiles of stop frequency for the modified RNA sample (+) and the unmodified control (-). These profiles are then used in a maximum likelihood estimation procedure to determine the relative modification frequency, or reactivity, of each nucleotide [97, 98] (see Section 4.6.6). High reactivities suggest unpaired nucleotides, while low reactivities are indicative of structural inflexibility due to base pairing, helical stacking, tertiary contacts, protein interactions, or other factors that can reduce nucleotide flexibility [180, 216, 217].

4.3.5 RNA structure prediction using SHAPE-Seq reactivities

A major application of RNA chemical probing is to use reactivity data as restraints in computational RNA folding algorithms to improve structural predictions *in silico*. In these methods, users supply an RNA sequence along with reactivity data as inputs

to generate predicted RNA structures that are more consistent with the experimental data. Several such methods incorporating reactivity values have shown that the use of SHAPE reactivities improves RNA structure prediction accuracy [102, 122, 163, 218, 219].

There are two main approaches for restraining computational RNA folding algorithms with SHAPE-Seq reactivities: 1) directly modifying the folding calculation or 2) selecting the structure from the results of the folding calculation that is most consistent with the experimental data. Both methods first calculate a partition function that describes how a population of RNA molecules partitions into an ensemble of different structures in equilibrium, with each structure occurring with a distinct probability [170]. Many properties can be determined from the partition function, including the minimum free energy (MFE) structure, which has the highest probability of occurring in the ensemble.

The first method to directly modify the RNA folding calculation was introduced by Deigan *et al.* as part of the RNAstructure suite of tools [163]. To use experimental SHAPE reactivities in the folding calculation, they are first converted into pseudo-free energy terms $\Delta G_{SHAPE}(i)$ that are included for each nucleotide i involved in base pair stacking in the RNA structures calculated overall free energy (ΔG) according to:

$$\Delta G_{SHAPE}(i) = m \ln [r(i) + 1] + b \tag{4.1}$$

where $r(i)$ is the reactivity at nucleotide i, and m and b are parameters that were originally fit by comparing a restrained prediction to the known crystal structure of the 16S rRNA from *E. coli* [163]. Thus, each paired nucleotide in a helix has two contributions of $\Delta G_{SHAPE}(i)$, for each of the two base pair stacking interactions, one above and one below. Two exceptions exist: 1) base pairs at the ends of a helix, which only have one stacking interaction and 2) single bulges, which are assumed to stack in within the he-

lix, and thus have two contributions of $\Delta G_{SHAPE}(i)$. By including $\Delta G_{SHAPE}(i)$ in the free energy calculation when the nucleotide i is paired, MFE structures are returned that are more consistent with the observed reactivity data [122, 163, 220]. Work has also been done to predict pseudoknot interactions using ShapeKnots, an algorithm that runs two folding stages that include $\Delta G_{SHAPE}(i)$ and another pseudo-free energy term for pseudoknots [102].

The second method uses the RNA structure partition function to generate a set of possible RNA structures [221] and then selects the structure(s) that most closely agree(s) with the experimental data. To perform this type of selection, the algorithm SeqFold first converts SHAPE reactivities to a "structural preference profile" (SPP), a normalized vector of reactivities restricted to [0,1] [222]. Each RNA structure is similarly converted into a binary vector such that 0 and 1 represent paired or unpaired bases, respectively. A minimum distance structure is then chosen by calculating the distance between the SHAPE SPP and each possible structure vector. Finally, the cluster of structures most closely related to this minimum distance structure is used to calculate a representative centroid structure [222]. Unlike MFE-based methods, this cluster-based approach provides more information about different structure sub-populations.

Recently, a method that combines both approaches was developed called restrained MaxExpect (RME). RME uses SHAPE reactivities to modify the partition function and selects a structure from it that best matches the experimental data [223]. First, RME calculates a partition function after adding a pseudo-free energy term for each nucleotide i using:

$$\Delta G = -RTm \ln \left[\frac{(q_i + \varepsilon)}{(1 - q_i + \varepsilon)} \right] \tag{4.2}$$

where m is the weighting parameter for the pseudo-free energies, ϵ is a small constant value to ensure a real answer, and q_i is the measured base pair probability obtained

from the experimental reactivities for position *i*, assigned to a 'low' or 'high' value of 0 or 1 based on a reactivity cutoff to represent paired or unpaired bases, respectively. Then, a predicted base pairing probability matrix is derived from the partition function that is then adjusted by the experimental data to account for discrepancies between the predicted base pairing matrix and the measured reactivities. This newly modified base pairing matrix is finally used to predict a structure that maximizes expected accuracy [223].

The concept of modifying the results of a partition function calculation with experimental data was introduced previously by Washietl *et al.* [224]. In their method, the free energy calculation for each RNA structure is perturbed with pseudo-free energy terms that are numerically determined instead of explicitly calculated (e.g. Equation (4.1)). The method first estimates an experimental base pairing vector using a reactivity cutoff as in RME. Then a partition function calculation is performed in which the free energy model is perturbed with a vector of pseudo-free energy terms for each nucleotide. This vector of pseudo-free energies is adjusted iteratively to minimize the differences between the predicted base pairing probability matrix and the estimated experimental base pairings [224]. After arriving at the final base pair probability matrix, it is used to predict a structure that maximizes expected accuracy.

Finally, Kutchko *et al.* took a different approach to RNA structure prediction by examining ensembles of structures through their subpopulations rather than picking or predicting a single structure for each set of SHAPE reactivities [225]. In this method, SHAPE reactivities are used to modify the partition function calculated by RNAstructure using the free energies introduced in Equation (4.1). The modified partition function is then used to generate many possible structures that are converted into binary vectors for structure clustering as described above. The resulting structures near clus-

ter centroids represent different RNA structure subpopulations in the ensemble that are more consistent with the experimental data [225]. The benefit of this approach is that the entire ensemble of structures is analyzed rather than focusing on a single 'best' predicted structure.

4.4 Materials and Methods

4.4.1 RNA folding and modification

The following steps outline RNA folding and modification with 1-methyl-7-nitroisatoic anhydride (1M7) [104], which can be synthesized in a one-step reaction [226]. Alternatively, N-methylisatoic anhydride (NMIA) can be commercially purchased and used in these steps with the noted modifications.

***in vitro* experiments**

1. Generate dsDNA templates with the promoter sequence for T7 RNA polymerase followed by the DNA sequence encoding the RNA of interest for run-off transcription. For the v2.1 method, where an RT priming site will be added later with ligation (Section 4.4.2 below), we recommend following the RNA of interest with the hepatitis δ ribozyme to produce the correct 3′ end after cleavage and 3′ end healing [122, 162]. Otherwise, the tendency of T7 RNA polymerase to add 1-3 spurious nucleotides to the 3′ end of the RNA can cause alignment issues downstream if not accounted for.
2. Set up an *in vitro* transcription reaction using standard methods [107, 108, 122].

3. Ethanol precipitate the transcription products to concentrate them.
4. Gel purify the RNA of interest. If UV shadowing is used, be careful to avoid directly shadowing the RNA being purified to avoid damage [227].
5. Check RNA integrity using a polyacrylamide gel.
6. Choose an appropriate folding buffer for the RNA of interest and prepare a 3.3X concentrated solution. A good starting point folding buffer (1X) is: 10 mM $MgCl_2$, 100 mM NaCl and 100 mM HEPES (pH 8.0) [108].
7. Dilute 1-20 pmol of RNA in 12 μL RNase-free H_2O. Denature at 95 °C for 2 min. Snap cool on ice for 1 min, then add 6 μL of 3.3X buffer. Incubate at 37 °C for 20 min. If adding an RNA-binding protein, do so after the second incubation, and follow with a third incubation to give the protein time to bind. This step can be adjusted based on the folding conditions desired.
8. Prepare a 65 mM solution of 1M7 in anhydrous DMSO. Aliquot 1 μL each of 65 mM 1M7 and anhydrous DMSO into different tubes. Upon completing the RNA folding incubation, add 9 μL to the 1M7-containing tube (+ sample) and mix. Do the same for the other 9 μL with DMSO (- sample). Incubate for 2 min at 37 °C to complete the reaction [122]. See Section 4.6.3 below for modifications if other folding conditions or chemical probes are used.

***in vivo* experiments**

1. If the RNA of interest is not endogenously expressed, clone it into an expression vector, being sure to include a priming site for reverse transcription. Two examples of convenient RNA expression vectors containing specific RT primers for superfolder green fluorescent protein mRNA and a synthetic intrinsic terminator can be found in Watters *et al.* [113]. These vectors are available on Addgene.

2. Grow 1 mL of cells into the desired growth phase, increasing the growth volume if more culture will be required for a functional assay. For *E. coli* grown at 37 °C, an OD_{600} value within 0.3-0.8 is recommended for exponential phase. Induce RNA expression if required and allow an appropriate amount of time for RNA synthesis.

3. Prepare a 250 mM solution of 1M7 in anhydrous DMSO. Aliquot 13.3 μL each of 250 mM 1M7 and DMSO into different tubes. See Section 4.6.3 below for modifications to this step if other chemical probes are used.

4. If taking a functional measurement (e.g. a fluorescence assay to determine regulatory function of a non-coding RNA [113]), set aside the volume of cell culture required for the assay, leaving 1 mL for RNA modification. For *E. coli*, we suggest pelleting the functional test aliquot (typically 150 μL) and resuspending in cold PBS with antibiotics to prevent further gene expression [113]. At the same time, add 500 μL of cell culture into each tube of 1M7 or DMSO and mix. Incubate at the culture growth conditions with shaking for 2-3 min to complete the reaction.

5. Extract the total RNA quickly to prevent degradation. Any RNA isolation method is acceptable. For *E. coli*, we recommend the TRIzol Max Bacterial RNA Isolation Kit (ThermoFisher) [113]. Dissolve/Elute the extracted total RNA with 10 μL of RNase-free H_2O.

4.4.2 RNA linker ligation (skip for *in vivo* or direct priming experiments)

1. Prepare 5'-adenylated linker by purchasing the phosphorylated RNA linker sequence (5'Phos-CUGACUCGGGCACCAAGGA-3'ddC) from an oligonucleotide

supplier and the 5′ DNA Adenylation Kit from New England BioLabs (NEB). Follow the manufacturer's instructions to adenylate 500 pmol of RNA linker in 50-100 μL reaction aliquots. Purify the RNA linker with a phenol-chloroform or TRIzol (ThermoFisher) extraction. Quantify the amount of RNA after purification and prepare a 20 μM solution.

2. Ethanol precipitate the modified RNA from the end of Step 8 of Section 4.4.1. If proteins and/or DNA were present in the folding conditions, perform a phenol-chloroform or TRIzol extraction first. Dissolve the pellet in 10 μL of 10% DMSO in RNase-free H_2O.

3. Mix the ligation reaction by adding: 0.5 μL of SuperaseIN (ThermoFisher), 6 μL 50% PEG 8000, 2 μL 10x T4 RNA Ligase Buffer (NEB), 1 μL of 20 μM 5′-adenylated RNA linker, and 0.5 μL T4 RNA Ligase, truncated KQ (NEB). Lower concentrations of the adenylated linker can be used as long as the linker is in at least 4-fold excess. Mix well and incubate overnight at room temperature.

4. Ethanol precipitate the RNA using glycogen as a carrier and dissolve the pellet in 10 μL of RNase-free H_2O.

4.4.3 Reverse transcription

1. Add 3 μL of 0.5 μM RT primer to the dissolved RNA. For *in vitro* experiments with ligation, the primer sequence should be GTCCTTGGTGCCCGAGT. For internally primed reactions, an appropriate RT primer should be designed prior to this step and used for this reaction.

2. Prepare the RT master mix by combining 0.5 μL of Superscript III (ThermoFisher), 4 μL 5X First Strand Buffer (ThermoFisher), 1 μL 100 mM dithiothreitol (DTT), 1

μL 10 mM dNTPs, and 0.5 μL RNase-free H_2O.

3. Incubate the RNA and RT primer at 95 °C for 2 min, followed by 65 °C for 5 min. Snap-cool the tubes for 30 seconds, then add 7 μL of the master mix from Step 2 and mix well.

4. Incubate at 52 °C for 25 min, followed by 65 °C for 5 min to inactivate the reverse transcriptase. These RT conditions may be adjusted if necessary for specific RNAs that are difficult to reverse transcribe.

5. Hydrolyze the RNA by adding 1 μL of 4 M NaOH to the RT reactions and incubate at 95 °C for 5 min (Replace 4 M NaOH with 10 M NaOH for *in vivo* experiments). Add either 2 μL or 5 μL of 1 M HCl for *in vitro* or *in vivo* experiments, respectively, to partially neutralize remaining base.

6. Ethanol precipitate the cDNA, washing the pellet thoroughly with 70% ethanol. Dissolve the pellet in 22.5 μL of nuclease-free H_2O.

4.4.4 Sequencing adapter ligation

1. To the cDNA, add 3 μL 10X CircLigase Buffer (Epicentre), 1.5 μL 50 mM $MnCl_2$, 1.5 μL 2.5 mM ATP, 0.5 μL 100 μM DNA adapter (5'-Phos-AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3C Spacer-3'), and 1 μL CircLigase I (Epicentre). Mix well. (Note: the DNA adapter should be PAGE purified before use).

2. Incubate at 60 °C for 2 hr, then 80 °C for 10 min to inactivate the ligase.

3. Ethanol precipitate the ligated cDNA using glycogen as a carrier and dissolve in 20 μL of nuclease-free H_2O.

4. Purify using 36 μL of Agencourt XP Beads (Beckman Coulter), according to manufacturers instructions to remove excess DNA adapter. Elute from beads with 20 μL TE buffer. ssDNA libraries can be stored until sequencing or quality analysis is performed.

4.4.5 Quality analysis

1. Design a pair of selection PCR primers that correspond to each RT primer used. To design the primer pairs, start with either the sequence CTTTCCCTACACGACGCTCTTCCGATCTYYYR (- samples) or CTTTCCCTACACGACGCTCTTCCGATCTRRRY (+ samples) as the 5′ end of the selection primer sequence. Then, append the desired RT sequence to the 3′ end of these sequences. Next, extend the RT primer sequence at the 3′ end to contain a few bases of the cDNA and protect these bases with phosphorothioate modifications to prevent 3′→5′ exonuclease degradation. As an example, the selection primer pair for samples that use the v2.1 *in vitro* linker ligation strategy are CTTTCCCTACACGACGCTCTTCCGATCT(RRRY/YYYR)GTCCTTGGTGCCCGAG*T*C*A*G, where * represents a phosphorothioate modification.

2. Mix a separate PCR reaction for each (+) and (-) sample by combining: 13.75 μL nuclease-free H_2O, 5 μL 5X Phusion Buffer (NEB), 0.5 μL 10 mM dNTPs, 1.5 μL of 1 μM labeling primer (5′-Fluor-GTGACTGGAGTTCAGACGTGTGCTC-3′; see below), 1.5 μL of 1 μM primer PE_F (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT), 1 μL of 0.1 μM selection primer (+ or -) from Step 1, 1.5 μL ssDNA library (+ or -), and 0.25 μL Phusion DNA polymerase (NEB). Use two different compatible fluorophores for the (+)

and (-) samples. We recommend VIC and NED, respectively. See Figure 4.1 for a schematic of this step.

3. Amplify PCRs for 15 cycles using an annealing temperature of 65 °C and an extension time of 15 seconds. More cycles can be used if the input RNA was low. However, when using a large amount of cycles, we recommend excluding primer PE_F from the reaction until the last 10-15 cycles to reduce side product formation [113].

4. Combine the (+) and (-) reactions, add 50 μL nuclease-free H_2O, and ethanol precipitate. Dissolve the combined reactions in formamide and examine with capillary electrophoresis, looking for good full-length RT extension and low dimer side product (Figure 4.1). See Watters *et al.* for more details [113].

5. (Alternate method) Skip Steps 2-4 above and follow the steps for sequencing library preparation in Section 4.4.6 below. Check libraries on an Agilent BioAnalyzer or similar instrument, looking for good full-length RT extension and low dimer side product.

4.4.6 Library preparation for sequencing

1. Assess whether the libraries are of sufficient quality to sequence.

2. Mix a separate PCR for each (+) and (-) sample by combining: 33.5 μL nuclease-free H_2O, 10 μL 5X Phusion Buffer (NEB), 0.5 μL 10 mM dNTPs, 0.25 μL of 100 μM TruSeq indexing primer (CAAGCAGAAGACGGCATACGAGATxxxxxxGTGACTGGAGTTCAGACGTGTGCTC; see below), 0.25 μL of 100 μM primer PE_F (Section 4.4.5 Step 2), 2 μL of 0.1 μM selection primer (+ or -, Section 4.4.5 Step 1), 3 μL ssDNA library (+ or -), and 0.5 μL Phusion DNA polymerase (NEB). Re-

place the 'xxxxxx' sequence with the appropriate six nucleotide TruSeq indexing barcode for Illumina sequencing. Additional barcoding can be added 5' of the RT primer sequence in the selection primer pair, although these barcodes will not be detected and split automatically by the Illumina sequencing processing pipeline. See Loughrey *et al.* for more details [122].

3. Amplify PCRs for 15 cycles using an annealing temperature of 65 °C and an extension time of 15 seconds. If a different number of cycles was used during quality analysis, use that PCR configuration instead.

4. Add 0.25 μL of Exonuclease I and incubate at 37 °C for 30 minutes to remove excess primer. Allow the PCRs from Step 3 to cool before adding the enzyme.

5. Purify reactions using 90 μL of Agencourt XP Beads (Beckman Coulter), according to manufacturer's instructions. Elute from the beads with 20 μL TE buffer. The dsDNA libraries are complete and ready for sequencing.

4.4.7 Illumina sequencing

1. Determine the mass concentration of all dsDNA libraries to be sequenced. We recommend using the Qubit high sensitivity DNA kit (ThermoFisher). Then calculate the molarity after determining the average molecular weight using the average length of the dsDNA library from the quality analysis traces [113].

2. Choose either the MiSeq or HiSeq platforms for sequencing. As a conservative rule of thumb, we suggest running one library pair (+ and -) for each million reads provided by the sequencing kit chosen. However, more libraries can be sequenced at once, provided that the amount of unwanted dimer side product is low.

3. Sequence the library mixture according to the manufacturer's instructions using 2x35 bp paired-end reads. Significantly longer read lengths are not recommended or necessary.

4.4.8 Data analysis with Spats

1. Obtain the compressed sequencing data on a Linux, Unix, or Mac OS X capable computer and extract the ".fastq.gz" files. Each TruSeq index will contain a pair of sequencing data files generated from the Illumina sequencing processing pipeline.
2. Create a fasta (.fa) formatted targets file that contains all of the RNA sequences that were measured in each TruSeq index. Include all of the linker sequences, internal barcodes, etc. if present.
3. Download and install the latest version of Spats (`https://github.com/LucksLab/spats`), including its utility scripts and dependent programs. Detailed instructions for running Spats and its utility scripts can be found in Watters *et al.* [113].
4. Run adapter_trimmer.py with the following command:

   adapter_trimmer.py <R1_seq.fastq> <R2_seq.fastq> <targets.fa>

   where <R1_seq.fastq> and <R2_seq.fastq> are the Illumina ".fastq" files for Read 1 (R1) and Read 2 (R2), respectively, and <targets.fa> is the fasta-formatted targets file created in Step 2.
5. Run Spats on the output file pair from adapter_trimmer.py using the following

command:

spats <targets.fa> RRRY YYYR combined_R1.fastq combined_R2.fastq

6. Normalize the output θ_i values to ρ_i values by multiplying all of the θ_i values by one less than the original RNA length [113, 122] (See Section 4.6.6 below). Do not include the linker or adapter sequences in this length. The ρ reactivities can be used to restrain secondary structure folding (See Section 4.4.9 below).

4.4.9 SHAPE-directed computational RNA folding

In the following steps we outline the process for restraining three different RNA folding algorithms with SHAPE-Seq reactivity data: RNAstructure [101, 102], restrained MaxExpect (RME) [223], and the Washietl *et al.* method (as part of the RNAprobing webserver) [224]. As discussed in Section 4.3.5, RNAstructure (containing *Fold* and *ShapeKnots*) can calculate the MFE structure directly as well as generate an ensemble of structures (with the partition and stochastic commands).

Algorithm 1: RNAstructure

1. Install the RNAstructure text interface program (`http://rna.urmc.rochester.edu/RNAstructure.html`) on a Linux, Unix, or Mac OS X operating system [101]. Alternatively, the webserver tools are available at `http://rna.urmc.rochester.edu/RNAstructureWeb/`.
2. Create a sequence file with a ".seq" extension that contains the following individual lines in order: 1) RNA name, 2) RNA sequence followed by the number '1'

to indicate the end of the file. Comment lines may be added as well, if preceded by a semicolon at the beginning of the file. Note that within the RNA sequence lowercase type forces the base to be single-stranded in the folding calculation. Uppercase denotes bases to include in the folding calculation. The letter 'T' is used as 'U' for RNA predictions.

3. Create a SHAPE reactivities file with a ".shape" extension that contains two tab-separated columns. The first column is the nucleotide number, starting with one, and the second column is the reactivity value for that position calculated by spats (ρ).

4. To run the *Fold* algorithm [228] (pseudoknots forbidden) use the command:

   Fold <.seq file> <.ct file> -sh <SHAPE file> -m 1 -sm 1.1 -si "-0.3"

   To run *ShapeKnots* [102] (up to one pseudoknot allowed) use the command:

   ShapeKnots <.seq file> <.ct file> -sh <SHAPE file> -m 1 -sm 1.1 -si "-0.3"

   For both commands, replace <.seq file> and <SHAPE file> with the files created in steps 2 and 3, respectively. In order, the options 'm', 'sm', and 'si' correspond to the number of structures drawn, the SHAPE slope parameter (m in Equation (4.1)), and the SHAPE intercept parameter (b in Equation (4.1)). The values of 1.1 for m and -0.3 for b were fit for SHAPE-Seq ρ inputs in Loughrey *et al.* [122]. The output is <.ct file>, which needs to be specified as a ".ct" filename in the command. Once generated by the algorithm, it contains the minimum free energy structure as predicted by *Fold* or *ShapeKnots*. The first line of the ".ct" file will contain the length of the sequence, the free energy of the structure, and the

title of the structure, respectively. The following lines contain, from left to right: the number of nucleotide i, its base, $i-1$, $i+1$, the number of its base pair partner (0 if unpaired), and the natural numbering (typically i).

5. As an alternative to *Fold* or *ShapeKnots*, RNAstructure can also sample structures from a partition function using the commands *partition* and *stochastic* [225]. Calculate the partition function based on SHAPE reactivities with the command:

   partition <.seq file> <.pfs file> -sh <.shape file> -sm 1.1 -si "-0.3"

   All of the options are the same as in Step 4 above, except <.pfs>, which is the calculated partition function output file. This partition function can then be used to sample structures using:

   stochastic <.pfs file> <.ct file> -e <#> -seed <random integer>

   The above command will sample the number of structures specified by the -e option and output them in a concatenated list to the <.ct file>. Changing the -seed option (default of 1234) will result in a different set of sampled structures.

Algorithm 2: RME

1. Install R version 3.2.2 (`https://www.r-project.org/`) and the packages rshape, mixtools, and evir. Also install Bioconductor (`http://bioconductor.org/install/`) and the RME source code, found at `https://github.com/lulab/RME` [223].
2. Create a FASTA file of the RNA sequence to analyze.

3. Create a ".ct" file for the RNA sequence being analyzed using *Fold* (Step 4 in Algorithm 1 above). Note that the base pairing information is not used during the RME calculation.

4. Prepare a tab-separated data file containing the reactivity information. The first line should read "RNA<tab>Index<tab>Reactivity<tab>Base". The following lines should contain the RNA name, index, reactivity, and base (e.g. TPP, 1, ρ_1, G; TPP, 2, ρ_2, C, etc.) for the entire length of the RNA.

5. Locate the SHAPE example training and test files in the /example/dat/data/ directory. They are required for calculation. Copy the ".ct" files from /example/dat/structure/ to the directory containing the '.ct" files generated in Step 3.

6. Pre-process the SHAPE data using the 23S rRNA training data according to:

   RME-Preprocess -d SHAPE -s <ct files directory> example/dat/data/SHAPE .train.23SrRNA.dataexample/dat/data/SHAPE.test.data<pre-processdirectory >

   where <pre-process directory> is a specified directory name to store the pre-processing data.

7. Predict SHAPE-directed structures using the following commands:

   RME -d SHAPE <pre-process directory>/SHAPE.for-test.txt SHAPE

   where <prediction directory> is a supplied directory name for the folding output and <pre-process directory> is the same as Step 6. The output will be in the in the form of a ".ct" file, which can be used in the same manner as the folding

results from Algorithm 1.

Algorithm 3: RNAprobing webserver (Washietl *et al.* method)

1. Go to the RNAprobing WebServer at `http://rna.tbi.univie.ac.at/cgi-bin/RNAprobing.cgi`.
2. Enter the RNA sequence either by copy-paste or uploading a fasta formatted file.
3. Upload a SHAPE reactivities file according to Step 3 of Algorithm 1. Note that SHAPE reactivities of 0 must be input as "0" and not "0.0" in order to be parsed properly.
4. Select "Washietl *et al.* 2012" as the "SHAPE method" and "Cutoff" from the drop-down window for "Method used to derive pairing probabilities". A cutoff threshold of 0.25 was used in Washietl *et al.* [224], although other values can be used.
5. Click "Proceed" to be redirected to an output page once the calculations are complete. The output will include the dot bracket notation of the predicted structure and an image of the predicted structure.

Drawing secondary structures

The ".ct" results file generated from the final steps of Algorithms 1 and 2 can be converted to a ".dbn" file that contains the structure in dot bracket notation. One way to do this is using the RNAstructure command:

ct2dot <ct file> 1 <bracket file>

where <ct file> is the ".ct" file being converted and <bracket file> is the output file

name ending in ".dbn". Note that *ct2dot* removes pseudoknots that may be generated by *ShapeKnots*. A number of different programs, such as VARNA (`http://varna.lri.fr/`) [229] can use dot bracket information to draw the secondary structure of an RNA sequence. The *draw* command in RNAstructure [164] can also draw secondary structures using ".ct" files.

4.5 Results

4.5.1 *in vitro* SHAPE-Seq analysis

SHAPE-Seq v2.0 consisted of several protocol optimizations to simplify and shorten the original SHAPE-Seq technique [107, 108]. In addition, it increased the techniques flexibility by adding a 3' linker ligation step after modification to remove RT priming site restrictions within the RNA [122]. In this work, we expand on these improvements by adapting the mismatch PCR selection developed in Watters *et al.* for in-cell SHAPE-Seq [113] to SHAPE-Seq v2.0 with a new redesigned linker for reduced dimer side product formation, thereby updating the *in vitro* experiment to SHAPE-Seq v2.1.

4.5.2 SHAPE-Seq v2.0 vs. v2.1

We compared SHAPE-Seq v2.1 to v2.0 using two well-benchmarked RNAs [102, 122, 163, 220]: 5S rRNA from *E. coli* and the *add* adenine riboswitch aptamer domain from *V. vulnificus*. For both RNAs, we folded and modified 40 pmol of RNA using the same buffer and ligand conditions as in Loughrey *et al.* [122] before splitting the (+) and (-) samples to process them individually with either the SHAPE-Seq v2.0 or v2.1 protocol.

One immediately observable difference in the v2.1 vs. v2.0 raw sequencing data was a 25-fold reduction in the amount of DNA ligation side product sequenced with the v2.1 improvements. Additionally, we observed only slight differences between the reactivities obtained using v2.0 or v2.1, as expected (Figure 4.2). Further, the reactivity maps agree well with previous measurements [102, 122, 220].

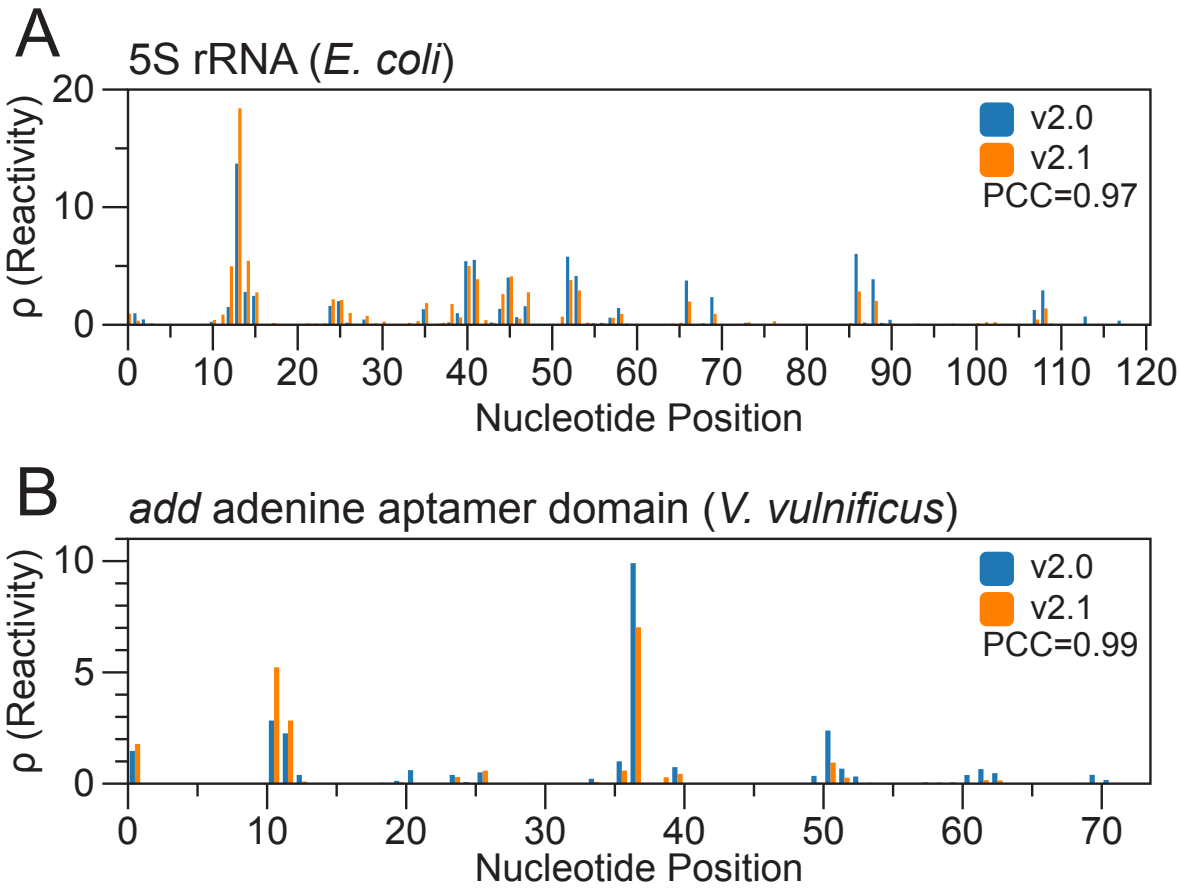


Figure 4.2: Comparison of SHAPE-Seq v2.0 vs. v2.1 *in vitro* reactivities. (A) Reactivity maps derived from the 5S rRNA from *E. coli* after equilibrium refolding and modification processed with either v2.0 (RMDB: 5SRRNA_1M7_0008) or v2.1 (RMDB: 5SRRNA_1M7_0009) library preparation steps. The reactivities closely agree, with a Pearson Correlation Coefficient (PCC) of 0.97. The 'GGA' sequence added to the 5' end of the 5S rRNA to aid in vitro transcription is not shown, although included with PCC analysis. (B) The same analysis for the *add* adenine aptamer domain from *V. vulnificus* shows a PCC of 0.99 (RMDB: ADDSC_1M7_0007 and ADDSC_1M7_0008). Like the 5S rRNA, a 'GG' sequence was added to aid in vitro transcription and is not shown on the graph, but is included in the PCC analysis.

In general, reactivities derived from the v2.1 experiment tend to be slightly shifted toward the 5′ end of the RNA relative to v2.0 reactivities. This difference is likely due to the reduction of reads in the v2.1 data that align immediately upstream of the RT priming site in the area where the unwanted dimer side product appears. Thus, the reduction of these reads may represent a slightly more accurate reactivity map. However, this difference was minor, as we observed that the Pearson Correlation Coefficients (PCC) comparing v2.0 vs. v2.1 were in the range of 0.97-0.99 (Figure 4.2).

4.5.3 Using SHAPE-Seq v2.1 to observe ligand binding

To demonstrate how SHAPE-Seq can be used to detect changes in RNA folding upon ligand binding, we examined the *thiM* thiamine pyrophosphate (TPP) riboswitch aptamer domain from *E. coli* using SHAPE-Seq v2.1. Comparing the reactivity maps with and without ligand for the TPP riboswitch aptamer domain shows a number of reactivity differences (Figure 4.3A). Specifically, we observe an interesting set of decreasing reactivities at positions 8-12 near the binding pocket of TPP, suggesting that part of the ligand binding pocket is first flexible, but becomes rigid after ligand binding, which agrees with previous observations (Figure 4.3B; inset) [230]. These decreases also come with reactivity increases in positions 14, 17-22, 34, and 36.

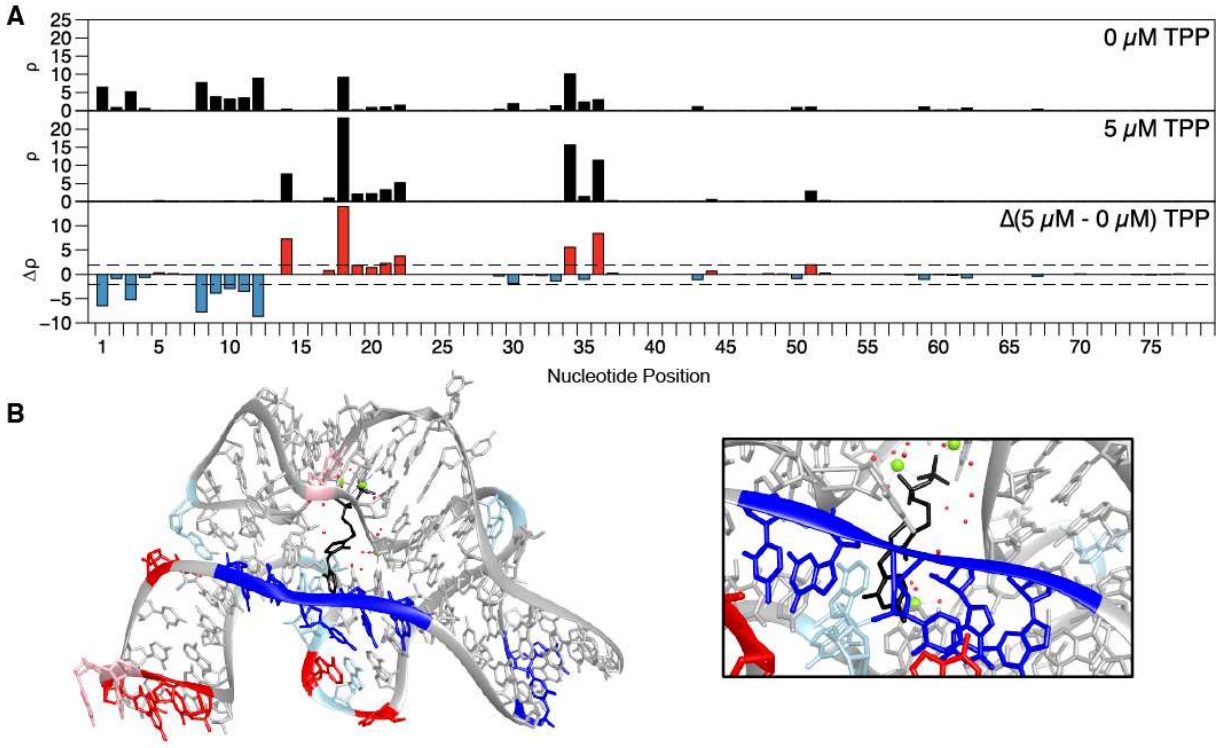
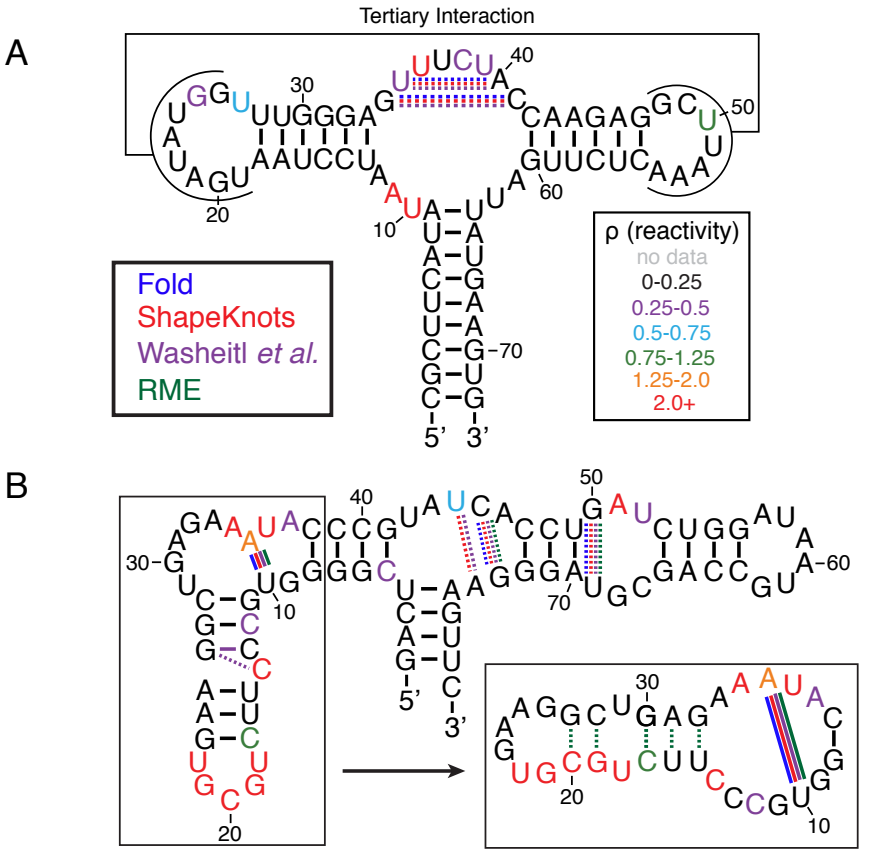


Figure 4.3: SHAPE-Seq reveals reactivity changes in the presence of ligand for the *thiM* TPP riboswitch aptamer domain. (A) *in vitro* reactivity maps for the *thiM* TPP aptamer domain with 0 μM or 5 μM TPP (RMDB: TPPSC_1M7_0005). The reactivity difference map (bottom) shows increases (red) and decreases (blue) in reactivity in the presence of ligand. (B) Crystal structure (PDB 2GDI) [231] of the *thiM* TPP aptamer domain with TPP (black) bound, colored by change in reactivity in the presence of ligand from part A. Magnesium ions are colored light green and the solvent is denoted as red dots. Nucleotides marked in red/blue show increases/decreases above $|\Delta\rho| \geq 2$ (dashed lines in part A). Light red and blue mark changes for which $1 < |\Delta\rho| < 2$. The region closing the TPP binding pocket shows a cluster of nucleotides that become less flexible upon ligand binding (inset). An extra 'G' was added to the 5′ end to aid in vitro transcription and is not displayed.

4.5.4 Inferring secondary structures with SHAPE-Seq data

We next investigated how incorporating SHAPE-Seq v2.1 reactivity values affects the calculated structures of the 5S rRNA, TPP riboswitch, and adenine riboswitch using the four methods described in Section 4.4.9. In general, we saw an improvement in both the percentage of known base pairs predicted correctly (sensitivity) and the percentage of predicted base pairs in the known structure (PPV; positive predictive value) for all of the methods used when SHAPE data was included (Figure 4.4), as has been shown multiple times [102, 163, 223, 224, 228]. The adenine riboswitch folded into the expected 'T' structure with or without SHAPE data, although three of the methods predicted two extra base pairs relative to the secondary structure representation of the ligand-bound aptamer domain crystal structure (Figure 4.4A) [232]. The TPP riboswitch aptamer was also folded into the correct general structure in three of the four methods, although with some differences from the crystal structure representation (Figure 4.4B) [231]. It is worth noting, however, that for both RNAs most of the incorrect predictions occur in regions known to be involved in non-canonical base pairing or protein/ligand interactions that are represented as unpaired bases in the 'accepted structure' for the purpose of calculating structural prediction accuracy.

Figure 4.4: Incorporating SHAPE-Seq data improves computational folding accuracy. (A) *add* adenine riboswitch aptamer domain ligand-bound secondary structure representation from Serganov *et al.* [232]. Dashed lines mark base pairs predicted by computational methods, as indicated by color, restrained with SHAPE-Seq v2.1 reactivities (Figure 4.2B) that do not exist in the crystal structure representation. Colored solid lines indicate base pairs that are present in the crystal structure, but are not predicted with v2.1 reactivities. Individual nucleotides are color-coded by reactivity intensity. (B) *thiM* TPP riboswitch aptamer domain ligand-bound secondary structure representation from Serganov *et al.* [231]. Tertiary interactions and non-canonical base pairings are not shown. Solid and dashed lines represent the same features as in part A and reactivities (Figure 4.3A) are color-coded the same way. The predicted structure from RME is visibly different for nucleotides 8-38 (boxed) as drawn on the right. (C) Table summarizing the folding accuracies for the four computational algorithms *Fold* [163, 228], *ShapeKnots* [102], Washietl *et al.* [224], and RME [223]. The Washietl *et al.* method was calculated using the RNAProbing webserver. No calculation could be performed for RME without SHAPE reactivities. sens. = sensitivity, PPV = positive predictive value [170].



	SHAPE data (v2.1)	Fold sens.	Fold PPV	ShapeKnots sens.	ShapeKnots PPV	RME sens.	RME PPV	Washietl *et al.* sens.	Washietl *et al.* PPV
5S rRNA	+	97.1	91.9	97.1	91.9	91.4	91.4	97.1	91.9
	-	28.6	25.0	28.6	25.0	N/A	N/A	28.6	23.8
TPP riboswitch	+	95.5	91.3	95.5	87.5	68.2	68.2	90.9	83.3
	-	77.3	85.0	77.3	85.0	N/A	N/A	77.3	85.0
adenine riboswitch	+	100	91.3	100	91.3	100	100	100	91.3
	-	100	100	100	100	N/A	N/A	100	100

Accuracy: 100 % 90 80 70 60 50 40 30 20 10 0

Generally, we found that all four methods performed similarly, although the dataset of RNAs is too small to be a fair comparison (Figure 4.4C). We also observed that v2.1 reactivities resulted in slightly higher accuracy for all four methods than v2.0 reactivities, which resulted in similar accuracies to those discussed in Loughrey *et al.* [122]. Thus all of the computational methods described in this work, coupled with SHAPE-Seq v2.1 reactivity data, can help guide researchers to more accurately model the structures of RNAs. This is particularly valuable for RNAs for which no crystal structure is available.

4.5.5 in-cell SHAPE-Seq analysis

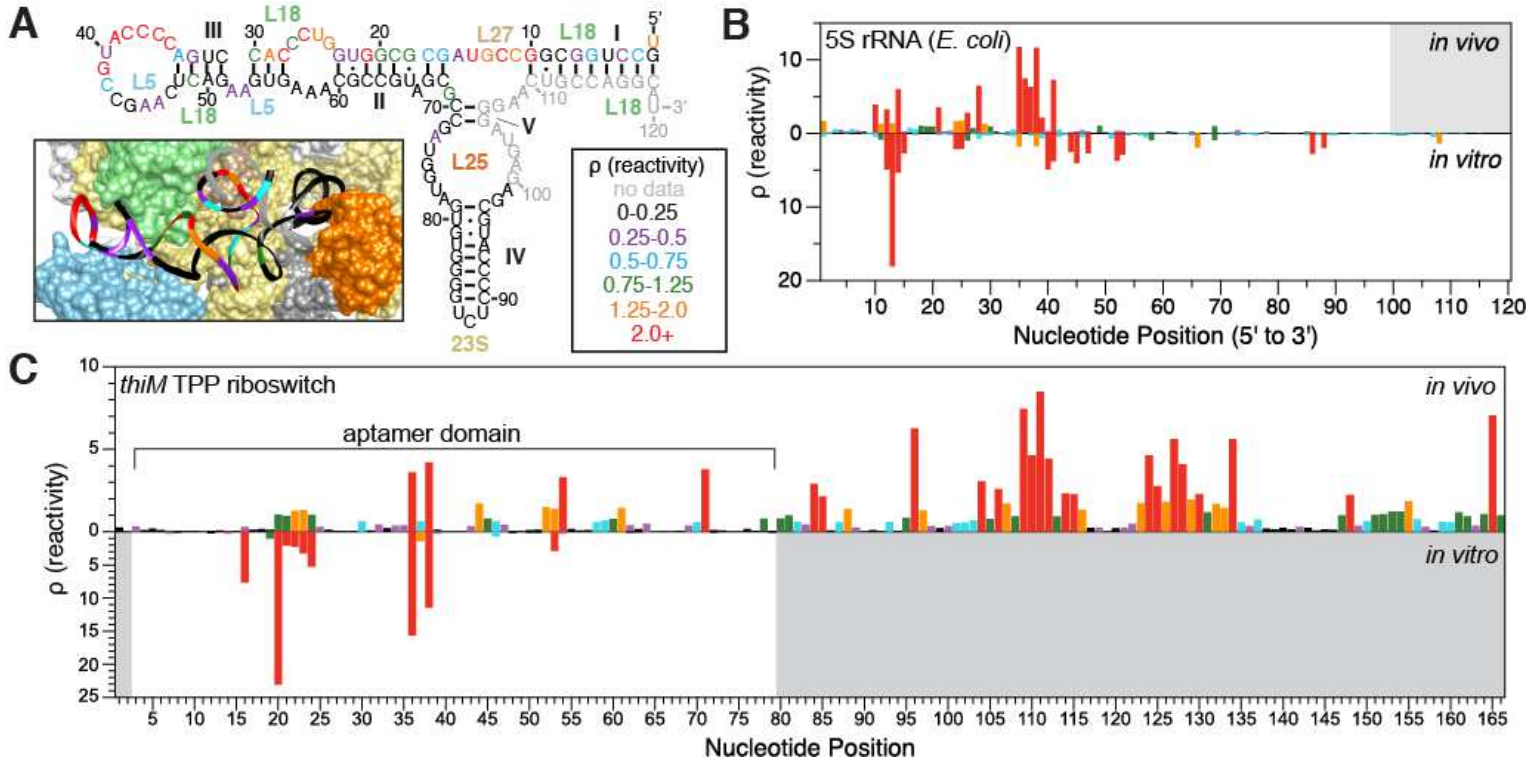
The in-cell SHAPE-Seq technique is closely related to SHAPE-Seq v2.1, as many of the improvements in v2.1 were derived from the in-cell method [113]. The main difference, as outlined in Section 4.4.1, is that the RNA modification step occurs *in vivo* to provide a more natural context for RNA folding to occur. Below, we present two different examples of in-cell SHAPE-Seq data for RNAs expressed in *E. coli*.

4.5.6 5S rRNA, expressed endogenously

To demonstrate how a combination of *in vitro* and *in vivo* SHAPE-Seq can provide information about how the cellular environment affects RNA folding, we used in-cell SHAPE-Seq to measure the structural characteristics of the *E. coli* 5S rRNA. We found that the reactivities we observed matched well to three-dimensional representations of the 5S rRNA from cryo-EM data fit with molecular dynamics simulations (Figure 4.5A) [113]. High reactivities tend to occur in unstructured loop regions, with the exception

of regions that are bound by proteins within the ribosome. Regions expected to be protein-bound appear lower in reactivity, suggesting that cellular 5S rRNA is predominantly contained within the ribosome during exponential growth.

Figure 4.5: *in vitro* vs. in-cell SHAPE-Seq reactivity map comparisons for 5S rRNA and the TPP riboswitch. (A) 5S rRNA in-cell reactivities overlaid on a predicted secondary structure [192] and a three dimensional model of the 5S rRNA within the entire ribosome (inset; from PDB 4V69) [193]. Individual ribosomal proteins (L5, L18, L25, L27) and the 23S rRNA are labeled on the secondary structure near their approximate locations; helices are numbered I-V. Reproduced from Watters *et al.*, 2015 [113] with permission from Oxford University Press. (B) Comparison of reactivities for the *E. coli* 5S rRNA measured in-cell (endogenous expression, top, RMDB: 5SRRNA_1M7_0007) vs. *in vitro* (bottom, RMDB: 5SRRNA_1M7_0009). Reactivities are color-coded according to (A). Clear differences in the endogenous 5S rRNA reactivities are apparent, especially for nucleotides 35-38 and 44-100, which increase and decrease, respectively relative to *in vitro*. (C) Comparison of the *E. coli thiM* TPP riboswitch measured *in vivo* (expressed from a plasmid vector, top, RMDB: TPPSC_1M7_0004) vs. an *in vitro* measurement of the adapter domain only with 5 μM TPP present (RMDB: TPPSC_1M7_0005). Comparing the reactivities in the 5′ half of the aptamer domain suggests that the riboswitch is primarily in the bound form in the cell, though differences in the 3′ half suggest that the cellular environment and the aptamer sequence context affect the RNA fold. Nucleotides that were not mapped in (B) and (C) are indicated with gray.


A
B
C
5S rRNA (E. coli)
in vivo
in vitro
ρ (reactivity)
Nucleotide Position (5' to 3')
thiM TPP riboswitch
aptamer domain
Nucleotide Position
L18
L27
L25
23S
ρ (reactivity)
no data
0-0.25
0.25-0.5
0.5-0.75
0.75-1.25
1.25-2.0
2.0+

We can also directly compare to *in vitro* 5S rRNA data to determine how the cellular environment changes the reactivity pattern (Figure 4.5B). For example, nucleotides 45-54 exhibit lower reactivities *in vivo*, roughly where the L5 protein is expected to bind (Figure 4.5B). Also, there are clusters of peaks downstream of nucleotide 54 that are near zero *in vivo*, but are highly reactive *in vitro*. Identifying these types of decreases, or increases, can reveal how different folding conditions (i.e., the cellular environment) affect RNA structure and function inside the cell.

4.5.7 TPP riboswitch, expressed from a plasmid

Next, we examined the TPP riboswitch with in-cell SHAPE-Seq. Unlike 5S rRNA, the TPP riboswitch was supplied exogenously as a translational fusion with superfolder green fluorescent protein (SFGFP) from a plasmid. Interestingly, we observed a reactivity pattern in the aptamer domain that matched well to the *in vitro* reactivity map of this region in the presence of ligand (Figures 4.3A and 4.5C). This suggests that the aptamer domain is predominantly in a ligand bound confirmation in the cell. However, there are also some slight differences between the *in vivo* and *in vitro* data. For example, positions 43-46, 58-61, and 70-71 exhibit higher reactivity *in vivo* (Figure 4.5C). In general, it appears that the TPP aptamer domain folded *in vitro* out of context of the expression platform captures most of the interesting reactivity clusters, but potentially misrepresents some details of the riboswitch. Such comparisons illustrate an advantage of in-cell SHAPE-Seq in that RNAs can be easily introduced with expression vectors to provide a more relevant picture of RNA folding in the cellular environment.

4.6 Experimental Considerations

4.6.1 Effect of increasing PCR cycles

In cases of low input RNA or poor cDNA yield, it may be advantageous to increase the number of PCR cycles to increase the amount of dsDNA available for sequencing. While there has been some concern that the increased number of cycles may introduce bias, we have shown in previous work that an increased number of PCR cycles does not cause substantial reactivity changes using either in-cell SHAPE-Seq (Figure 4.6A) or SHAPE-Seq v2.0 [113, 122]. However, SHAPE-Seq v2.1 inherently requires more cycles of PCR than v2.0 to reach an equivalent library concentration because v2.1, as well as in-cell SHAPE-Seq, uses multiple forward primers that require extra PCR cycles to build the complete Illumina adapter sequences in a stepwise fashion. Therefore, to confirm that the increased number of PCR cycles in SHAPE-Seq v2.1 does not bias reactivity calculation, we sequenced our v2.1 libraries using 15 (the v2.1 standard), 18, or 20 cycles of PCR (Figure 4.6B-D). As expected, there was little difference in the calculated reactivities due to increased cycling for 5S rRNA, the TPP riboswitch, and adenine riboswitch.

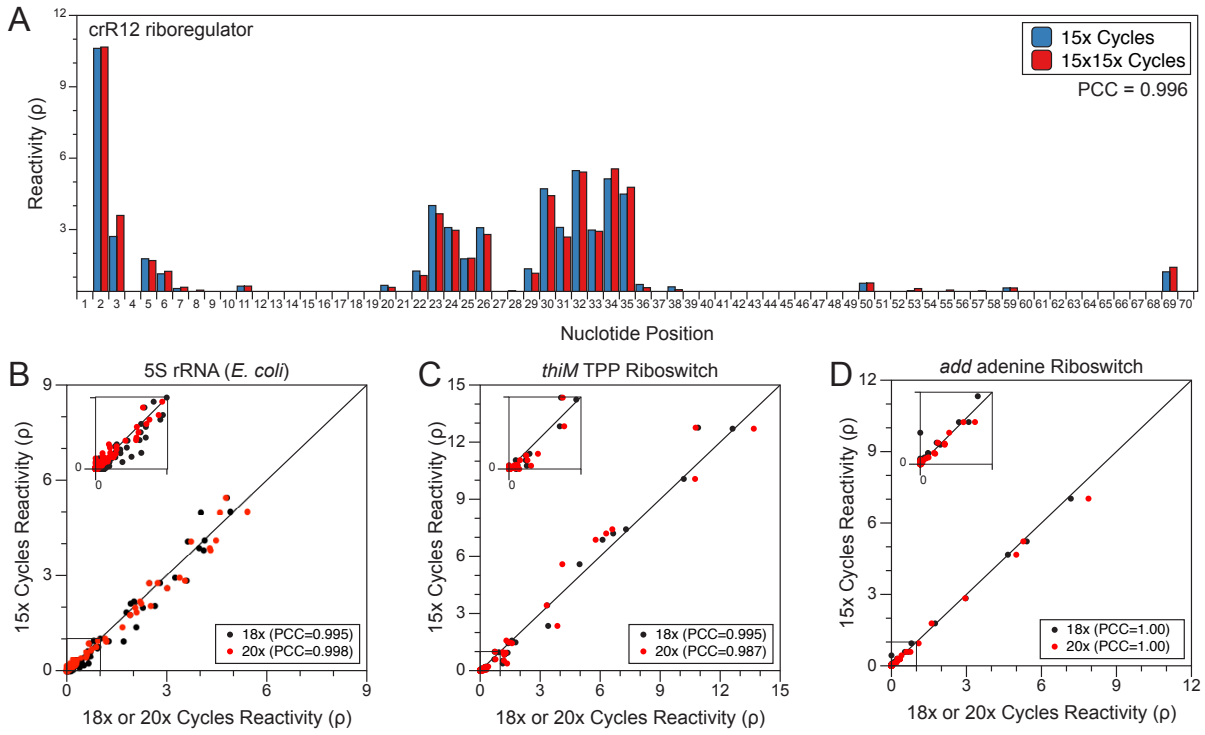


Figure 4.6: PCR does not bias SHAPE-Seq reactivity calculations. (A) Comparison of the crR12 riboregulator calculated reactivities from an in-cell SHAPE-Seq experiment [113] using either 15x cycles of PCR (15x; blue) vs. 15x cycles without PE_F, followed by another 15x cycles including PE_F (15x15x; red). A Pearson Correlation Coefficient (PCC) value of 0.996 suggests increased PCR cycling does not affect the reactivity calculation. (B) Base-wise comparison of reactivities calculated from 5S rRNA SHAPE-Seq libraries using 15x cycles of PCR vs. 18x (black) or 20x (red) cycles of PCR (RMDB: 5SRRNA_1M7_0009). A PCC near unity suggests little difference in the reactivity values calculated from libraries with increased PCR cycling. Similar analyses were done for the *thiM* TPP aptamer domain (RMDB: TPPSC_1M7_0005) (C) and *add* adenine aptamer domain (D) (RMDB: ADDSC_1M7_0008).

For experiments in which the amount of input RNA is particularly low, or improved selection against unwanted dimer side product is desired, we suggest performing the quality analysis and dsDNA library PCRs in a stepwise fashion. As shown in Watters *et al.*, the amount of unwanted dimer amplification can be further reduced by first cycling with only the inner selection primer before adding PE_F for the required subsequent cycles [113]. We also found that excessive cycling with PE_F beyond 15 rounds may cause an increase in off target and dimer side product amplification, especially when the in-cell target transcript is at low abundance. Thus, for sensitive applications we recommend splitting the reaction into two phases as described above and increasing the number of cycles in the first phase if increased sensitivity is required, with the second round limited to 15 or fewer cycles after PE_F is added to complete the reaction.

Because of the ability of SHAPE-Seq v2.1 and in-cell SHAPE-Seq to selectively amplify correctly extended cDNAs, both have the capability to analyze RNAs that are present at low concentrations without altering the RNA folding conditions. To demonstrate that SHAPE-Seq v2.1 provides consistent results over a range of relevant *in vitro* RNA concentrations, we compared reactivity maps obtained using four different starting amounts of 5S rRNA: 1, 5, 10, and 20 pmol (Figure 4.7). Across all concentrations, there is good agreement between the reactivity maps of the various starting amounts. There are several positions that exhibit slight differences, such as nucleotides 102-104, but they do not show a trend related to the starting amount of RNA. Interestingly, nucleotides in this region exhibit fewer aligned reads relative to other nearby positions in the RNA sequence. Ultimately, it is not the amount of starting RNA, but rather the number of aligned sequencing reads used for reactivity calculation that determines data quality and consistency.

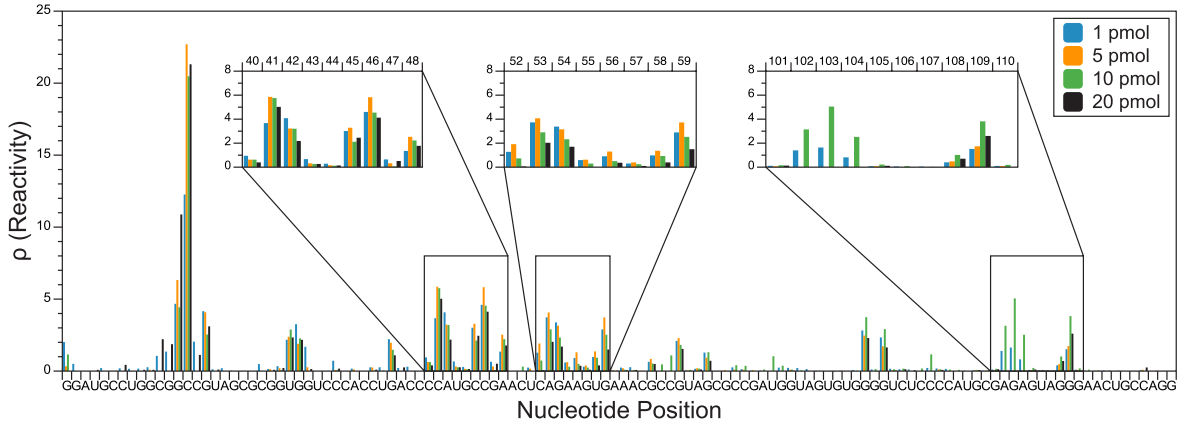


Figure 4.7: Comparison of *in vitro* folded 5S rRNA reactivity maps from different amounts of starting RNA. The same *in vitro* SHAPE-Seq experiment was performed using either 1, 5, 10, or 20 pmol of starting RNA (RMDB: 5SRRNA_1M7_0009). As expected, all of the reactivity maps are very similar, although there is some disagreement near positions 102-104 (right inset). However, these differences do not show a trend with increasing/decreasing starting RNA level and are thus likely experimental noise.

4.6.2 RT primer length and library multiplexing

One of the key features of the SHAPE-Seq improvements is the shortening of the RT primer. In its original conception, SHAPE-Seq v1.0 used long RT primers that contained the entire Illumina adapter sequence [107]. SHAPE-Seq v2.0 greatly shortened these primers and the DNA adapter sequence. It also included both TruSeq barcoding provisions and the potential for internal barcoding by using a pair (+ or -) of RT primers containing a pre-planned barcode [122]. The changes added to v2.0 improved dimer side product removal during bead purification and lowered oligonucleotide expenses.

In the current state-of-the-art SHAPE-Seq methods (v2.1 and in-cell), the RT primer is further shortened to only contain sequence that binds directly to an RNA of interest, thus requiring that all of the Illumina sequences are provided with PCR, except for

those included in the DNA adapter. By adding all of the Illumina sequences this way, much more customization during library preparation for sequencing is allowed. It is also a major advantage because it decouples the preparation of the ssDNA library from the preparation of the dsDNA sequencing library. Thus, libraries can be stored for long periods of time without the concern of potential sequencing incompatibilities, as ssDNA libraries can be easily converted to dsDNA libraries at a later date with the most current barcoding and sequencing configuration. In contrast, v2.0 barcodes had to be pre-selected and could not be changed at a later date, which could cause sequencing incompatibilities between certain libraries during sequencing. Last, the shortened primers, roughly 20-35 nucleotides compared to 55-70 for v2.0 or 80-90 for v1.0, produce shorter dimer side products that are more easily removed during the bead purification steps.

Thus, using short RT primers provides major advantages for SHAPE-Seq techniques. In fact, using longer RT primers with the most recent protocols will result in increased amounts of unwanted dimer ligation product, even with the improved PCR selection methods.

4.6.3 Choosing SHAPE reagents and other chemical probes

There are many choices of a chemical probe. For routine probing experiments we recommend 1M7, although there are many instances in which use of a different chemical is advantageous. The SHAPE reagents modify the 2'-OH of flexible nucleotides. Other chemicals such as DMS or CMCT directly modify the Watson-Crick face, though they are typically limited in their range of selectivity. Further, the reaction time scale of the chemical probe or its ability to enter living cells may have a factor in the choice of

probe to use. Below we describe some of the basic characteristics of the most common reagents to provide insight into choosing one over another for an experiment. If using a chemical other than 1M7, the reaction time and conditions may need to change relative to the described method in Section 4.4.1.

There are a number of SHAPE reagents that have slightly different modification properties. Three similar compounds (NMIA, 1M6, and 1M7) are based on the same anhydride scaffold and have increasingly shorter half-lives from 260 to 14 seconds for NMIA and 1M7, respectively [104, 216]. Differences in the reactivities measured for the same RNA with these reagents can yield information about the ribose sugar conformational sampling based on the dynamics of modification [216]. If using NMIA or 1M6 in place of 1M7 for *in vitro* experiments, increase the reaction incubation times to 22 min or 3 min, respectively, using the same concentration as 1M7 [216].

Benzoyl cyanide (BzCN) is another type of SHAPE reagent. It reacts with a very short half-life (250 ms), reacting to completion in seconds [233]. Historically, BzCN has been used when the modification needs to take place as quickly as possible, such as with time course assays [233–235]. Because of the increased difficulty of use and elevated safety considerations required for BzCN, we recommend using 1M7 instead unless a very fast modification time is needed. When fast modifications are desired, use BzCN at 400 mM (in place of 65 mM for 1M7) and incubate the reaction for 1-2 seconds to bring it to completion.

The last major class of SHAPE reagents, consisting of NAI and FAI, have hydrolysis half-lives in the middle of the NMIA-1M7 spectrum [158]. Recently, they were further functionalized to contain an azide group that allows the addition of a biotin moiety via a 'click' reaction for subsequent pull-down and selection of modified RNAs only, thus reducing the required sequencing depth downstream [115]. If using NAI to

modify RNA, replace the 65 mM 1M7 reagent with a 1-2 M stock of NAI or NAI-N_3 and incubate for 15 min. Quench using a two-phase extraction (e.g. TRIzol) to remove unreacted NAI.

NAI, FAI, and 1M7, were recently shown to diffuse into living cells to modify RNAs inside the cell [113, 115, 126, 158, 185]. For in-cell SHAPE-Seq in *E. coli* we recommend 1M7 because its half-life is on a shorter time scale than cell division and RNA degradation. All three SHAPE reagents usable *in vivo* can be synthetized in one (1M7 [226], NAI, and FAI [158]) or a few steps (NAI-N_3 [115]) from commercially available reagents. To use NAI instead of 1M7 for *in vivo* probing, replace the 13.3 μL of 250 mM 1M7 with 51.2 μL of NAI (or NAI-N_3) 1-2 M stock solution and incubate for 15 min in place of 2-3 min before two-phase extraction to quench the reaction.

There are also a number of chemical probes that directly modify base positions. The two most popular are DMS and CMCT, which are known to preferentially modify A/C or G/U positions, respectively, although not equivalently [236]. Others, such as DEPC (diethylpyrocarbonate) and kethoxal [86], are also base specific, but are used less frequently now, mainly due to the fact that DMS and CMCT cover all four bases together and react more consistently. Unlike CMCT, DMS can enter cells to modify RNAs directly inside without forcing them to be permeable. This property and DMS's longstanding use as a chemical probe led to its use in many of the recently published *in vivo* NGS-based probing methods [116–118].

To use DMS in place of 1M7, replace the 13.3 μL of 250 mM 1M7 with 27.75 μL of 13% DMS in ethanol (for *in vivo*) or the 1 μL of 65 mM 1M7 with 1 μL of 3.5% DMS in ethanol (for *in vitro*), replacing the DMSO control with ethanol. Incubate for 3 minutes before quenching with 240 μL or 2.4 μL 2-mercaptoethanol for *in vivo* or *in vitro* experiments, respectively. Use two-phase extraction to purify the RNA as suggested

in Section 4.4.1.

It should be noted that any of these chemicals should be cross-compatible with most NGS-based RNA probing methods, including SHAPE-Seq, given that most of the steps involved are for preparing the sequencing libraries. While differences in library preparation techniques do exist, most chemical probing methods, except for SHAPE-MaP [112], rely on the ability of modified nucleotides to block reverse transcription. Thus, by simply changing the RNA modification step, SHAPE-Seq [107, 108, 113, 122] could use DMS modification just as easily as DMS-Seq [117] could use SHAPE modification, as was done in Watters *et al.* [113].

4.6.4 Factors influencing data quality and consistency

There are a number of factors that have the potential to influence the final results that should be kept in mind while performing SHAPE-Seq experiments.

One of the biggest factors in collecting meaningful and consistent results is the importance of good RNA extractions and purifications. Poor recovery of RNA after extraction or precipitation will greatly lower the number and quality of reads aligned, mainly through increasing the amount of unwanted dimer product that is generated, as there will be less cDNA to ligate to the DNA adapter. This can be especially problematic for in-cell SHAPE-Seq during the initial total RNA extraction. We have found that extractions that become degraded, either by poor RNase-free technique or excessive delay in extracting the RNA, result in very poor yields. Thus, careful pipetting for precipitations and extractions as well as quickly extracting total RNA, if performing in-cell SHAPE-Seq, are crucial.

For SHAPE-Seq v2.1 experiments, a high-quality preparation of the 5'-adenylated linker is critical. In cases where the adenylation reaction is inefficient, the ligation reaction produces low yields, hindering downstream RT and increasing the levels of unwanted dimer side product. Another issue that can arise is the loss of 3' end blocking groups during the adenylation reaction. Unblocked linker molecules can ligate together and create a convoluted mix of RT products that are visible in the quality analysis as sets of peaks separated by ~20 nucleotides. In either case repeatedly arises, we recommend preparing a new batch of the adenylated linker.

A wealth of information can be gained about a library from the quality analysis steps (Figure 4.1). First, the rough percentage of the library that is composed of the unwanted dimer side product can be determined by observing the expected dimer peak that typically shows up around 100 nucleotides, depending on the RT primer length [113]. Second, the full-length peak can be used to ensure that the reverse transcriptase extended to the 5' end of the RNA. Also, the heights of all the peaks are indicative of the general library quality. Higher peaks suggest higher quality libraries that will need fewer cycles of PCR to prepare sequencing libraries. Last, the relative level of signal decay from reverse transcriptase stopping can be qualitatively estimated from the quality analysis and can help inform *a priori* how many reads may be required for an acceptable reactivity map.

Not surprisingly, more aligned reads generate a more accurate reactivity spectrum and reduce run-to-run variability, or noise, between individual experiments [237]. As a rule of thumb, we suggest a minimum of 50,000 reads aligned to be confident that the maximum likelihood estimation used by Spats [97, 98] generates a reliable reactivity map. However, this assumes that the reads are well distributed between the (+) and (-) samples and within the RNA, which is frequently not the case. Despite this, some

RNAs actually generate reliable maps with even fewer reads, although they are RNAs that tend to have fewer, highly reactive peaks rather than large clusters of intermediate reactivities. Another rule of thumb is that roughly 10-100 reads per nucleotide position should be aligned in both channels. These values, however, represent minimums. We suggest a few hundred thousand reads to generate the cleanest reactivity maps with the least amount of variability.

The last point of consideration is the level of signal decay that occurs within the RNA. As reverse transcriptase transcribes from the 3' end, it has a tendency to 'fall off', or stop transcribing, with some probability which is increased in the (+) channel due to the presence of the chemical modifications. Correcting this signal decay is performed by the maximum likelihood model used within Spats to calculate reactivity [97, 98]. However, certain nucleotide positions, either due to an inherent high 'fall off' rate or a high probability of chemical modification, greatly increase the signal decay rate [237]. Because Spats uses signal decay to calculate reactivity, it is resistant to errors that can occur in other analyses from rapid signal decay. However, these sharp drop-offs in read alignments can still affect Spats processing if the number of reads upstream (closer to the 5' end) of the drop-off becomes very small. RNAs that contain these extreme drop-offs are cases were the number of reads required for a reliable reactivity map is increased. One example is highlighted above for nucleotides 102-104 in the *in vitro* 5S rRNA reactivity spectrum, which exhibits run-to-run variation and occurs in a region of fewer read alignments (Figure 4.7).

4.6.5 Choosing an adapter trimming algorithm

In Loughrey *et al.*, we updated the Spats data analysis pipeline to include an improved adapter trimming algorithm based on the fastx toolkit named adapter_trimmer. More recently, we have also created a version of adapter_trimmer that is based on cutadapt Martin:2011eu as an alternative adapter clipping method. Further, we have relaxed the requirement that all reads aligned match perfectly to the target. The updated version of spats, as well as older versions, can be found at `https://github.com/LucksLab/spats/`.

4.6.6 Measures of SHAPE reactivity

The mapped read counts from the sequencing data are converted into a measure of reactivity using Spats called θ. Each θ_i value represents the relative probability that a modification within an RNA molecule occurs at nucleotide i. Values for θ are calculated by fitting the aligned reads to a Poisson model of reverse transcription 'fall off' at modification positions using a maximum likelihood estimation procedure to find the underlying θ that best explain the pattern of (+) and (-) read counts [97, 98].

Because θ is a distribution describing the relative probability of modification at each position within the RNA, it is dependent on the length of the RNA according to:

$$\sum_{i=1}^{l} \theta_i = 1 \tag{4.3}$$

where l is the length of the RNA molecule. To compare RNAs of different lengths, θ_i can be normalized to ρ_i by multiplying by length l [113, 122]. Using ρ_i in place of θ_i is also useful because it sets the reactivity values to the order of magnitude expected by secondary structure prediction algorithms as was shown in Loughrey *et al.* with

RNAstructure, using $m = 1.1$ and $b = -0.3$ as folding parameters [122].

4.7 Further potential improvements for SHAPE-Seq and restrained RNA folding

While we have continued to improve the SHAPE-Seq technique, there are a few areas where further improvements/extensions are still desired.

4.7.1 Going transcriptome-wide

While we generally advocate for targeted RNA structure analysis, there are many potential benefits to the recent innovations in transcriptome-wide probing techniques [114–118, 238] that have sparked a growing interest in examining RNA structure at a global level. Therefore, extending SHAPE-Seq to be able to optionally target the entire transcriptome would be valuable. To switch to total RNA structure probing, all that would be experimentally needed is an alternative RT priming step in place of the targeted approach we have chosen to follow up to this point. Random priming of total RNA is in principle easy to perform, but does not allow for selective PCR methods. Yet, it may not be required if good extension across the random primer set is achieved, leaving little to no unextended RT primer. An alternative approach is to fragment the total RNA and ligate an RT priming site, using a SHAPE-Seq v2.1-like approach. The drawback is that the added ligation step will negatively impact library generation efficiency, and additional methods will be needed to distinguish RT fall off due to fragmentation from modification.

4.7.2 Future directions for computational folding

A common application for SHAPE-Seq reactivity data is to restrain RNA folding algorithms, as discussed extensively above. Reactivity data has been repeatedly shown to improve secondary [101, 102, 122, 156, 163] and tertiary [239] structure predictions, with secondary structure algorithms being more popular. However, there are still many cases where reactivity information alone is not enough to obtain an accurate fold. Beyond improving free energy terms and including pseudoknots, there are two main ways in which structural prediction accuracy could be greatly improved: inclusion of non-canonical base pairing and better representation of RNA structure subpopulations.

One common cause of prediction inaccuracies is the presence of non-canonical base pairing, which is pervasive in RNA structural motifs. In many cases, non-canonical base pairs are in regions that are predicted to be in single-stranded, allowing the rest of the RNA to attain a fairly accurate folding prediction. However, bases that participate in non-canonical interactions are frequently incorporated into canonical Watson-Crick pairs during computational folding, which can generate nonsensical RNA structures that are misleading for *de novo* RNA structure modeling. SHAPE reactivities often reflect non-canonical structures well if knowledge of their presence is provided *a priori* via crystal structure data, etc. Thus, even small improvements in predicting non-canonical base pairing would be of great interest to the RNA community and would greatly aid SHAPE-directed structure folding accuracy.

Another common pitfall encountered when predicting RNA structures computationally is the focus on the minimum free energy (MFE) structure. Frequently, these structures may be misleading, especially in cases where non-canonical base pairing is present, as discussed above. Further, a population of identical RNA molecules is not

restrained to fold into only one structure. Rather, RNA structure is more accurately described as a combination of many structural subpopulations, which typically contain several different dominant structural motifs [240]. One method to address these subpopulations is to cluster predicted RNA structures and use these clusters to obtain a characteristic structure. Sfold and SeqFold both use this method [222, 240] as well as the approach taken by Kutchko *et al.* [225]. SeqFold chooses which characteristic structure best represents the SHAPE reactivity data, but this collapses the subpopulation information into one structure. A more powerful approach would be to use SHAPE reactivity data to not only predict which structures are likely present, but also at what level they exist in a structural population. Some preliminary work has been done to understand structural populations in this manner [241] but further improvement and adoption would be beneficial to understanding potential structures when studying a new RNA.

4.8 Conclusions

SHAPE-Seq is a rapidly improving and expanding technique for characterizing the RNA structure-function relationship both *in vitro* and *in vivo*. In this work, we have presented different experimental approaches for characterizing RNA structures both *in vitro* and *in vivo* and showed how to use the structural information obtained to computationally predict what RNA structures were present in the experimental conditions. As our data suggest, SHAPE-Seq is a robust technique that has been updated to be simpler to perform through the use of selective PCR and optimized library construction steps. Further, the wide variety of computational tools available for RNA secondary structure prediction can be used to help interpret SHAPE-Seq results, with many of them able to directly incorporate reactivity data to improve structure prediction ac-

curacy. We believe that the widespread adoption of SHAPE-Seq methods backed by computational tools will continue to drive the discovery of new insights in RNA structural biology.

4.9 Acknowledgements

This material is based upon work supported by the Tri-Institutional Training Program in Computational Biology and Medicine [NIH training grant T32GM083937 to AMY] and a New Innovator Award through the National Institute of General Medical Sciences of the National Institutes of Health [grant number 1DP2GM110838 to J. B. L.]. K.E.W. is a Fleming Scholar in the School of Chemical and Biomolecular Engineering at Cornell University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

CHAPTER 5

COTRANSCRIPTIONAL FOLDING OF A FLUORIDE RIBOSWITCH AT NUCLEOTIDE RESOLUTION

5.1 Abstract

As RNA polymerase transcribes a gene, the emerging RNA can fold into alternative structures that depend on internal base pairing and interactions with trans-acting ligands. To examine how RNA folding progresses during transcription at nucleotide resolution, we developed cotranscriptional selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq) and applied it to determine how the *B. cereus* crcB fluoride riboswitch cotranscriptionally bifurcates its folding pathway in response to ligand to make a gene regulatory decision. We observed that the riboswitch folds into a meta-stable aptamer before undergoing structural rearrangements that either delay or promote the formation of an intrinsic terminator hairpin in the presence or absence of fluoride, respectively. Our approach provides a new framework for analyzing cotranscriptional folding pathways of RNAs with single-nucleotide resolution.

5.2 Introduction

As nascent RNA molecules exit RNA polymerase (RNAP), they transition through multiple intermediate structural states that ultimately determine the final structure and function of an RNA [66, 68, 83, 242]. Because RNA folding generally occurs faster than transcription, the 5′ to 3′ polarity of RNA synthesis directs an order of folding,

The work presented in this chapter has been submitted at *Nature Structural and Molecular Biology* and contains contributions from Eric J. Strobel, Angela M Yu, and Julius B. Lucks.

or cotranscriptional 'folding pathway', that sets the structural stage for many types of interactions that govern cellular processes such as transcription, translation, and macromolecular assembly [61, 62, 243].

Cotranscriptional folding is predicted to be particularly important for bacterial riboswitches [43], a class of regulatory RNAs that control gene expression as a function of specific ligand concentration. Riboswitches contain two domains: a ligand-binding aptamer and an expression platform that makes regulatory decisions based on the structural state of the aptamer [28, 43]. For riboswitches that regulate transcription, ligand binding must influence folding pathways within a short time window in order to commit the riboswitch to one of two mutually exclusive pathways: promote intrinsic terminator hairpin formation or prevent it [44, 79, 130, 230, 244]. A number of structural studies have revealed the details of RNA-ligand interactions for many aptamers [43], and biochemical [49, 78] and biophysical [79, 230] studies using actively transcribing RNAP have observed distinct RNA structural transitions during transcription. However, we still lack a complete, nucleotide-resolution understanding of how ligand binding influences the folding pathway of an entire riboswitch and enables it to regulate gene expression.

5.3 Results

Here, we introduce cotranscriptional SHAPE-Seq (selective 2'-hydroxyl acylation analyzed by primer extension sequencing), a method that couples *in vitro* RNAP arrest with high-throughput structure probing to characterize the structures of nascent RNA transcripts at single-nucleotide resolution (Figure 5.1). We first perform *in vitro* transcription using a library of DNA templates that direct the synthesis of each interme-

diate length of a target RNA (Figure 5.1A). Each template contains an EcoRI site at the 3′ end that when bound by a catalytically inactive EcoRI Gln111 mutant [245] is used to establish a roadblock that halts RNAP 14 nt upstream of the EcoRI binding site [246]. Initiation of single-round transcription from this template library generates halted elongation complexes at all intermediate lengths of the target RNA that are rapidly modified with the fast-acting SHAPE reagent benzoyl cyanide (BzCN) [234]. Extracted RNAs are then processed for paired-end sequencing according to our previously developed SHAPE-Seq v2.1 protocol [123].

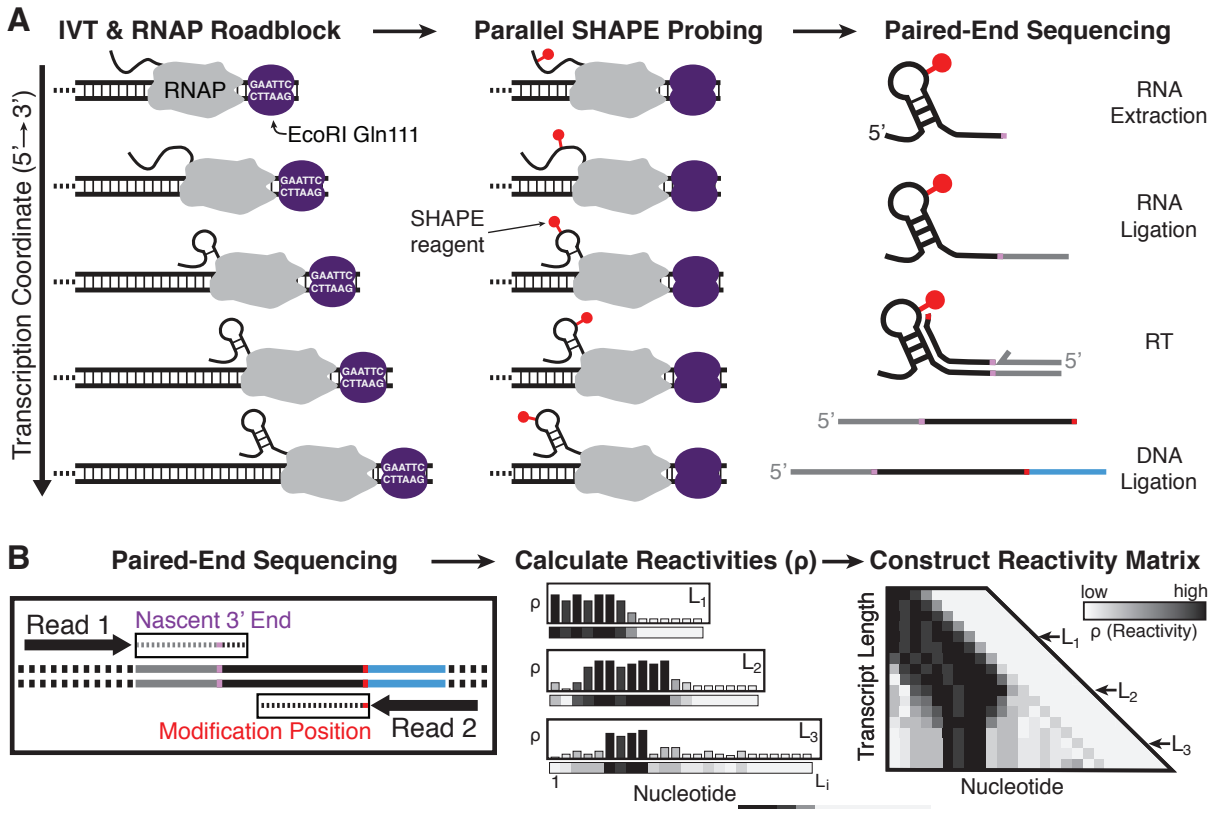
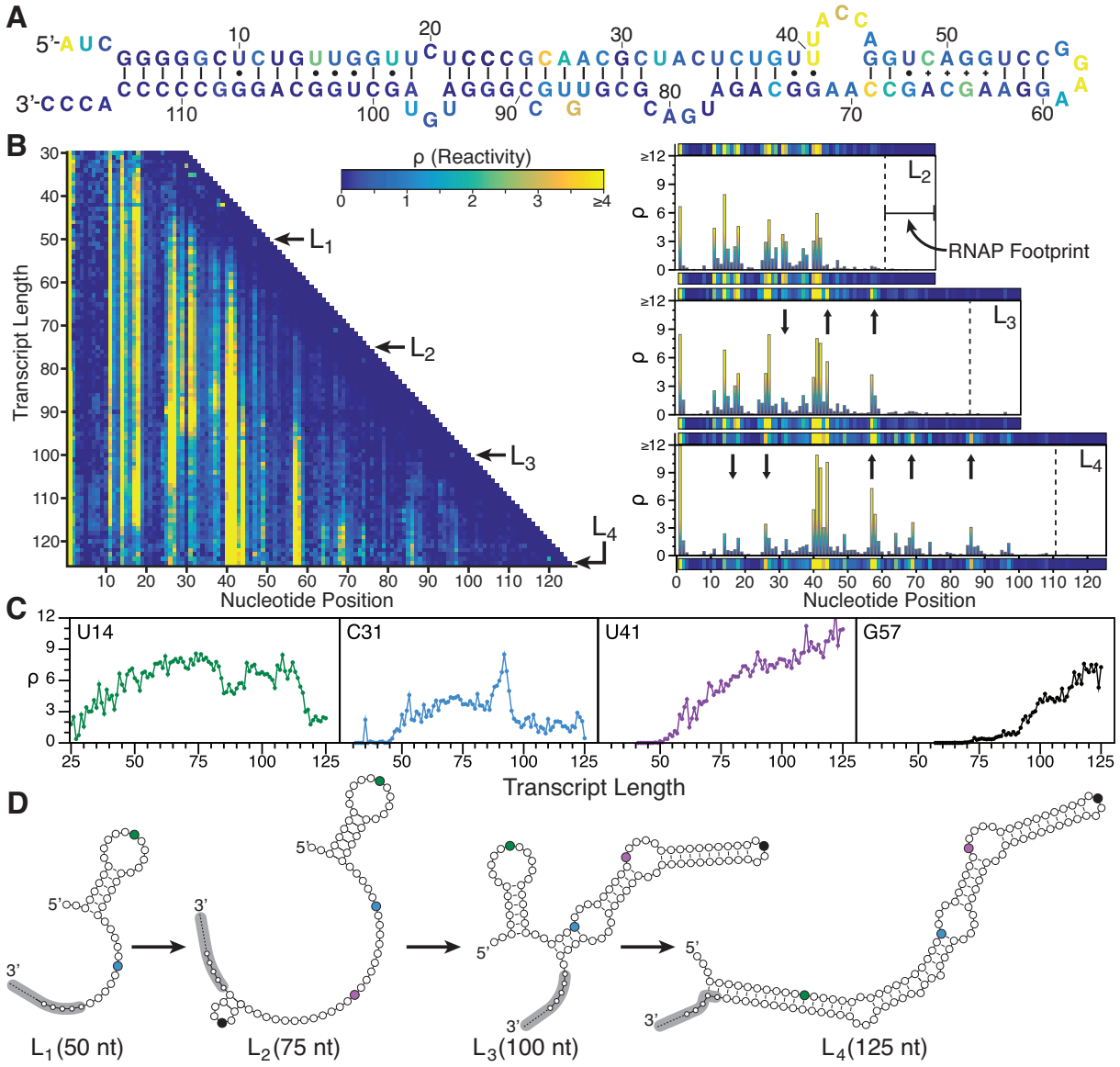


Figure 5.1: Cotranscriptional SHAPE-Seq overview. (A) A set of templates are generated that each contain an *E. coli* promoter, a variable length RNA template, and an EcoRI Gln111 roadblock site. Single-round *in vitro* transcription is performed using a template library containing a roadblock site at each intermediate transcript length, followed by simultaneous SHAPE probing of the arrested complexes and preparation for sequencing. (B) Paired-end sequencing reveals the SHAPE modification position and the 3′ end of each nascent RNA transcript. Reads are binned by transcript length and used to calculate SHAPE reactivity profiles that are stacked to generate the reactivity matrix. Increases or decreases in reactivity between transcript lengths (rows) at particular nts (columns) of this matrix reveal cotranscriptional folding events.

Each paired-end read contains two pieces of information: the locations of the halted RNAP (nascent RNA 3′ end) and the SHAPE modification position (Figure 5.1B). Reads are first bioinformatically binned by RNAP position and then used to calculate a SHAPE-Seq reactivity spectrum for each intermediate length of the RNA [98]. These reactivities represent flexibilities for every nucleotide within each nascent RNA transcript length. High reactivities are indicative of unpaired bases and low reactivities indicate bases that are potentially involved in base pairing, stacking, or ligand interactions [123, 180]. A comparison of reactivities at different points during transcription allows structural rearrangment events to be identified as transcription proceeds.

To validate cotranscriptional SHAPE-Seq, we first examined the signal recognition particle (SRP) RNA from *E. coli*. The final folded form of the SRP RNA is an extended helical structure containing interspersed inner loops (Figure 5.2A) [247]. Biochemical studies have suggested that the 5′ end first forms a labile hairpin structure early during transcription that rearranges into the extended helix only after the 3′ end is synthesized [83]. To test if we could observe this rearrangement, we used cotranscriptional SHAPE-Seq to obtain a matrix of reactivity spectra for the intermediate lengths of the nascent SRP RNA transcripts (Figures 5.2B and C.1).

Figure 5.2: SRP RNA cotranscriptional folding. (A) Secondary structure of the final SRP RNA fold colored by reactivity intensity at length L_4 (125 nts) drawn according to a crystal structure determined in Batey *et al.* [247]. (B) Cotranscriptional SHAPE-Seq reactivity matrix for the SRP folding pathway (left). A 3D representation can be viewed in Figure C.1. Lengths L_1, L_2, L_3, and L_4 correspond to 50, 75, 100, and 125 nts, respectively. Selected bar charts and corresponding matrix rows above and below (right) display reactivities for L_2 to L_4. Reactivity changes for $L_2 \rightarrow L_3$ and for $L_3 \rightarrow L_4$ are marked with arrows. (C) Reactivity values for positions U14, C31, U41, and G57 over the course of transcription. U14 undergoes a loop $\rightarrow$ helix transition at length 117. Similarly, C31 becomes paired at length 96. Plot colors correspond to the marked base positions in (D) outlining the folding pathway of the SRP RNA that is consistent with these transitions. The 14 nt RNAP footprint [246] for each length is indicated with gray, with the ~5 nts in the RNA exit channel marked as small circles.


A
5'-AUC GGGGGCUCUGUUGGUUCUCCCGCAACGCUACUCUGUU ACC A GGUCAGGUCCG G
3'-CCCA CCCCCGGGACGGUCGA AGGGCGUUGCG AGACGGAACCGACGAAGG A
B
ρ (Reactivity)
Transcript Length
Nucleotide Position
L1
L2
L3
L4
RNAP Footprint
C
U14
C31
U41
G57
ρ
Transcript Length
D
L1(50 nt)
L2(75 nt)
L3(100 nt)
L4(125 nt)

Over the course of transcription, changes in nucleotide reactivity patterns (Figure 5.2C) suggest a series of structural transitions that correspond to the early formation of a stem loop that ultimately rearranges into the elongated SRP RNA helical structure (Figure 5.2D; Appendix C.2). Formation of the early stem-loop structure can be seen as a cluster of highly reactive nucleotides (nts) across positions 11-18 in the loop. The pattern of high reactivity persists until the SRP RNA reaches a length of 117 nts at which point the sharp drop in reactivities at these positions indicate the formation of the elongated helical structure. We observed similar transitions when intermediate SRP RNA fragments were refolded at equilibrium (Figure C.2), although the cotranscriptional transitions occur later due to the 14 nt RNAP footprint protecting the RNA 3' ends. Interestingly, the transitions are sharper in the equilibrium refolding data, suggesting that the cotranscriptional experiment can capture RNA populations that are not at equilibrium.

Based on our SRP RNA results, we expected cotranscriptional SHAPE-Seq to possess the resolution necessary to reveal how particular alternative folding pathways are controlled during ligand regulation of a riboswitch. To test this, we examined the *B. cereus* crcB fluoride riboswitch, which controls transcription by preventing the formation of an intrinsic terminator hairpin in the presence of fluoride [50]. Covariation and equilibrium structural analyses have suggested how the fluoride bound and unbound forms of the aptamer domain may interact with the downstream expression platform sequence (Figure 5.3A) [50, 248]. However, the specific mechanism by which fluoride binding directs or prevents the folding of the intrinsic terminator during transcription has yet to be elucidated.

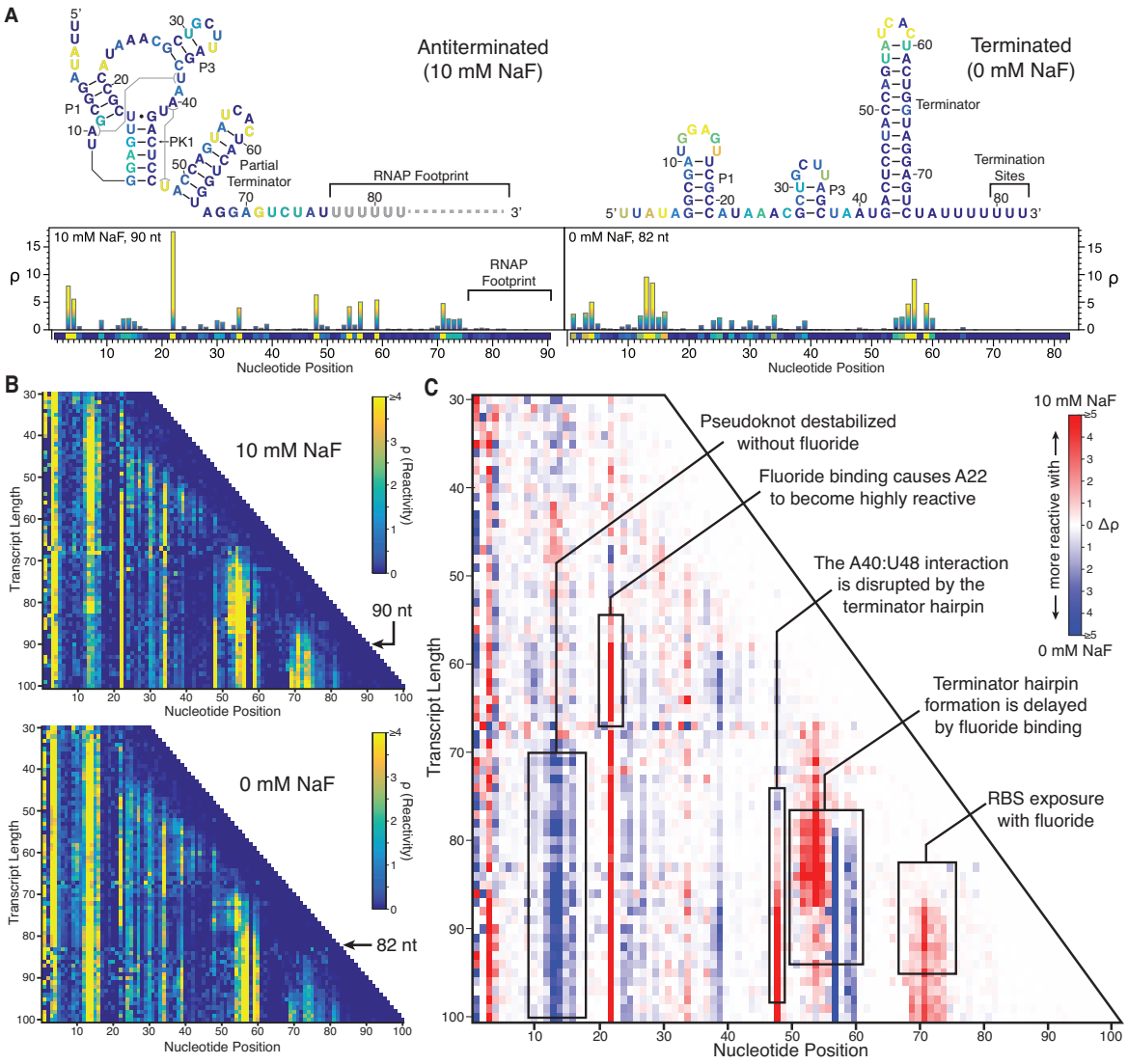
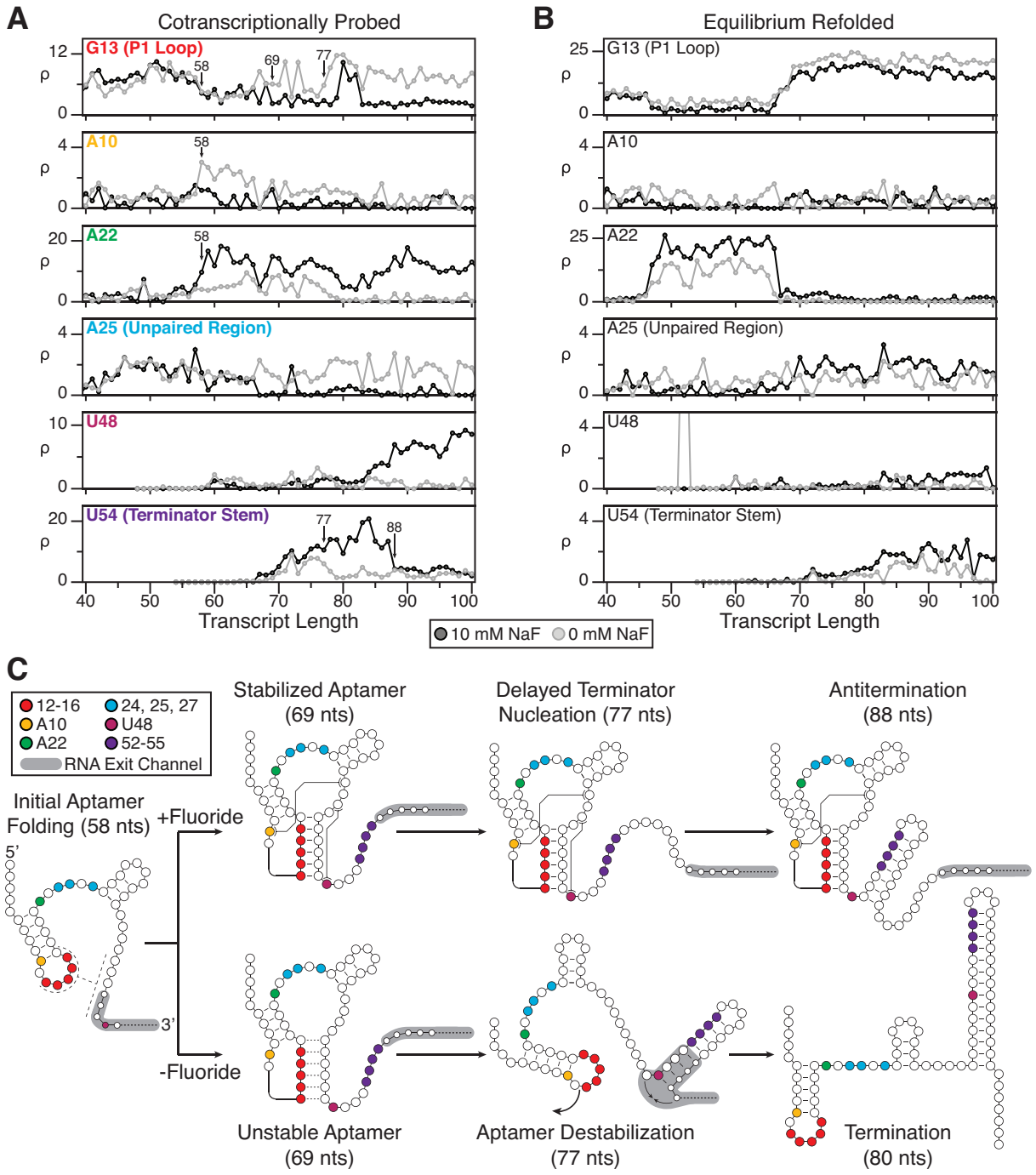


Figure 5.3: *B. cereus* fluoride riboswitch cotranscriptional SHAPE-Seq data. (A) The antiterminated and terminated folds of the fluoride riboswitch, colored by the reactivity values at lengths 90 nt and 82 nt, respectively (bottom). (B) Reactivity matrices for the fluoride riboswitch transcribed with 10 mM (top) or 0 mM NaF (bottom). A 3D representation can be viewed in Figure C.4. (C) Reactivity differences ($\Delta\rho$) between the matrices in (B) annotated according to folding events during transcription. The reactivity changes that occur over the course of transcription suggest that the fluoride riboswitch traverses two cotranscriptional folding pathways depending on the presence of ligand.

To probe the OFF (terminated) and ON (antiterminated) structural states of the *B. cereus* fluoride riboswitch, we generated cotranscriptional SHAPE-Seq reactivity matrices in the presence of either 0 mM or 10 mM NaF, respectively (Figures 5.3, C.3 and C.4). Comparison of the matrices reveals both ligand-independent similarities in the initial folding of the aptamer as well as distinct differences later in the folding of the riboswitch that accompany fluoride binding and fluoride-directed antitermination (Figures 5.3C and 5.4). Early in transcription, the *B. cereus* fluoride riboswitch folds into two hairpins that precede formation of the aptamer, regardless of fluoride concentration. The first hairpin forms within the first ~40 nts of transcription and is comprised of the P1 stem and a highly reactive loop between nts 11-16 (Figure C.5). The second hairpin forms shortly thereafter and is comprised of the P3 helix and a loop that exhibits a highly reactive position at U34 and low to moderate reactivities elsewhere (Figure C.6).

Figure 5.4: Cotranscriptional folding pathway of the *B. cereus* fluoride riboswitch. (A) Single nucleotide trajectories displaying changes in the reactivities of nucleotides involved in several key structural transitions when transcribed with either 0 mM (gray) or 10 mM NaF (black). The trajectories diverge at lengths were structural transitions occur. (B) Same as (A), except that the RNAs were synthesized, extracted, denatured, and equilibrium refolded in transcription buffer with either 0 mM (gray) or 10 mM NaF (black) prior to SHAPE modification. The lack of divergence between the trajectories indicates that cotranscriptional folding is required to obtain alternate ligand-dependent structures. (C) The folding pathway for the fluoride riboswitch begins with initial aptamer folding. If fluoride binds, the pre-folded aptamer then stabilizes through specific interactions, leading to delays in the early folding stages of the intrinsic terminator hairpin, which does not stabilize until RNAP has escaped the polyU tract. However, if there is no fluoride binding, the top of terminator hairpin quickly folds disrupting the pseudoknot and reaching into the RNA exit channel to allow the full terminator hairpin to trigger transcription termination. Intermediate structural states are inferred from cotranscriptional SHAPE-Seq reactivities, covariation analysis [50], and crystallographic data [248].


A
Cotranscriptionally Probed
G13 (P1 Loop)
A10
A22
A25 (Unpaired Region)
U48
U54 (Terminator Stem)
ρ
Transcript Length
B
Equilibrium Refolded
G13 (P1 Loop)
A10
A22
A25 (Unpaired Region)
U48
U54 (Terminator Stem)
Transcript Length
10 mM NaF
0 mM NaF
C
12-16
24, 25, 27
A10
U48
A22
52-55
RNA Exit Channel
Initial Aptamer Folding (58 nts)
+Fluoride
-Fluoride
5'
3'
Stabilized Aptamer (69 nts)
Delayed Terminator Nucleation (77 nts)
Antitermination (88 nts)
Unstable Aptamer (69 nts)
Aptamer Destabilization (77 nts)
Termination (80 nts)

The formation of the initial hairpins set the stage for folding of the aptamer domain. After they form, RNAP continues transcribing to a length of 58 nts, at which point we observe a fluoride-independent drop in the reactivity values at nts 12-16 in the P1 loop that signals the formation of pseudoknot PK1 between nts 12-17 and 42-47 as the latter emerge from the RNA exit channel (Figure 5.4). Once PK1 is formed the aptamer is complete [50, 248], demonstrating that it first enters a pre-organized state before ligand binding, as has been observed for other aptamers [80, 249, 250]. From here, the fate of the aptamer structure is determined by whether or not fluoride is present.

The first steps in the fluoride-dependent bifurcation of the folding pathway involve fluoride-mediated aptamer stabilization. In the presence of fluoride, the P1 loop reactivities continue to decrease until length 69, suggesting that fluoride binding stabilizes the pseudoknot (Figure 5.4A). Stabilization of the fluoride-binding pocket also requires a long-range non-canonical base pair between U38 and A10 (Figure 5.3A) [50, 248], the latter of which is paired in the P1 stem prior to PK1 formation. Thus, in the presence of fluoride we expect A10 to be consistently paired and maintain a low reactivity throughout the folding pathway (Figure 5.4A). In contrast, in the absence of fluoride an increased reactivity at A10 at transcript length 58 indicates PK1 formation disrupts its base pair within the P1 stem. Our observation of fluoride-induced aptamer stabilization is further supported by distinct reactivity changes in nts 22-27, which join the P1 and P3 helices but do not participate in any pairing interactions in the analogous *T. petrophila* aptamer domain [248]. While nts 24-27 display lower reactivities in the presence of fluoride (Figures 5.4 and C.7), A22 undergoes a dramatic reactivity spike upon PK1 formation when fluoride is present, but only a modest increase in the absence of fluoride (Figure 5.4A) revealing that A22 hyper-reactivity is a strong indicator of aptamer state. We observe that similar structural rearrangements occur for mutants capable of binding fluoride, but do not for mutants that are incapable (Figures C.8-C.14

and Appendix C.2). Taken together, these results support a model in which the pseudoknot forms the basis of the aptamer that can undergo further coordinated restructuring upon fluoride binding to form a more stable structure with additional interactions.

Following aptamer formation, the riboswitch follows one of two ligand-dependent folding trajectories that direct it to terminate transcription or antiterminate. When RNAP reaches length 71 the upper 3′ stem of the terminator hairpin begins to emerge from the RNA exit channel as the riboswitch prepares for the regulatory decision that occurs when RNAP reaches length 77. Without fluoride, the terminator hairpin nucleates at length 77, observed as a decrease in reactivity in the upper terminator stem (nts 52-55) (Figure C.15). Increased reactivity in the P1 loop (nts 12-16) (Figure C.16) and decreased reactivity at A22 occur concurrently, indicating that PK1 opens, thereby dissolving the meta-stable aptamer (Figure 5.4). Equilibrium refolding analysis reveals that once the transcript length reaches 68 nt PK1 can no longer form regardless of fluoride concentration (Figures 5.4B and C.17). Interestingly, position G68 exits the RNA:DNA hybrid at roughly length 77 during transcription, suggesting that without bound fluoride, the instability of PK1 enables terminator nucleation to extend into the RNA exit channel to allow C47 to pair with G68 at the upstream edge of the RNA:DNA hybrid and promote terminator formation [251]. Thus, terminator nucleation and the pairing of G68 and C47 at transcript length 77 direct the precise decision-making event of the riboswitch.

With fluoride, the stabilized aptamer promotes antitermination in two ways: 1) disfavoring complete terminator formation by sequestering part of the terminator hairpin (Figure 5.3A and 5.2) delaying initial terminator hairpin nucleation until length 88, by which point RNAP has transcribed past the polyU sequence (Figures 5.4A and C.15). When the terminator hairpin does begin to form, high reactivities at U48 and nts 69-74

indicate that only the top half of the terminator winds (Figures 5.3 and C.18), leaving the ribosome binding site (RBS, nts 67-72) accessible for translation of the downstream fluoride transporter [50]. Thus, fluoride binding fundamentally alters the cotranscriptional folding pathway of the fluoride riboswitch by stabilizing an RNA structure that promotes transcription via antitermination and translation by RBS exposure.

5.4 Conclusion

The work presented here helps answer longstanding questions about how ligands affect riboswitch cotranscriptional folding pathways to enable them to make regulatory decisions. We anticipate that techniques like cotranscriptional SHAPE-Seq will become increasingly important to understand the roles of cotranscriptional folding in regulating broader cellular processes.

5.5 Acknowledgments

We thank J. Roberts, J. Filter, and I. Artsimovitch for EcoRI Gln111 protein, its expression plasmid, and thoughtful discussions. We also thank J. Liao for plasmid construction. Chemical probing data is deposited in the RNA Mapping Database (RMDB; `http://rmdb.stanford.edu/`) with deposition numbers as listed in Table C.3. This work was supported by the Tri-Institutional Training Program in Computational Biology and Medicine [NIH training grant T32GM083937 to AMY], a New Innovator Award through the National Institute of General Medical Sciences of the National Institutes of Health [grant number 1DP2GM110838 to JBL and grant GM25232 to JTL], and a National Science Foundation Graduate Research Fellowship Program [grant num-

ber DGE-1144153 to KEW]. KEW is a Fleming Scholar in the School of Chemical and Biomolecular Engineering at Cornell University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

CHAPTER 6

MEASURING THE COTRANSCRIPTIONAL FOLDING PATHWAY OF THE PT181 TRANSCRIPTIONAL ATTENUATOR

6.1 Abstract

During transcription RNA polymerase transcribes new RNA molecules that can co-transcriptionally fold into intricate structures as they emerge from the polymerase. Co-transcriptional folding is especially important for RNAs that regulate transcription, as they must be able to rapidly form structures to affect RNA polymerase before it transcribes past the a regulatory region. The pT181 transcriptional attenuator is an RNA regulator that can fold into an intrinsic terminator or an antiterminator structure, depending on whether a complementary antisense RNA is present or absent. Here, we apply the newly developed cotranscriptional SHAPE-Seq (selective 2'-hydroxyl acylation analyzed by primer extension sequencing) technique to study the folding pathway of the pT181 attenuator. Using cotranscriptional SHAPE-Seq, we observe that the attenuator sequentially folds into three independent RNA stem-loops before it begins to transcribe sequence critical for its antitermination mechanism. If the antisense RNA is present, it binds to the first hairpin of the attenuator, stabilizing it and allowing a downstream terminator to fold. In the absence of the antisense RNA, the nascent RNA strand binds to the base of the first hairpin and quickly refolds into an elongated antiterminator structure. Following our analysis of the mechanism, we perform deletion/mutational analysis and derive a minimized version of the attenuator. This work signifies an important advance in the understanding of the effects of cotranscriptional folding on RNA regulators.

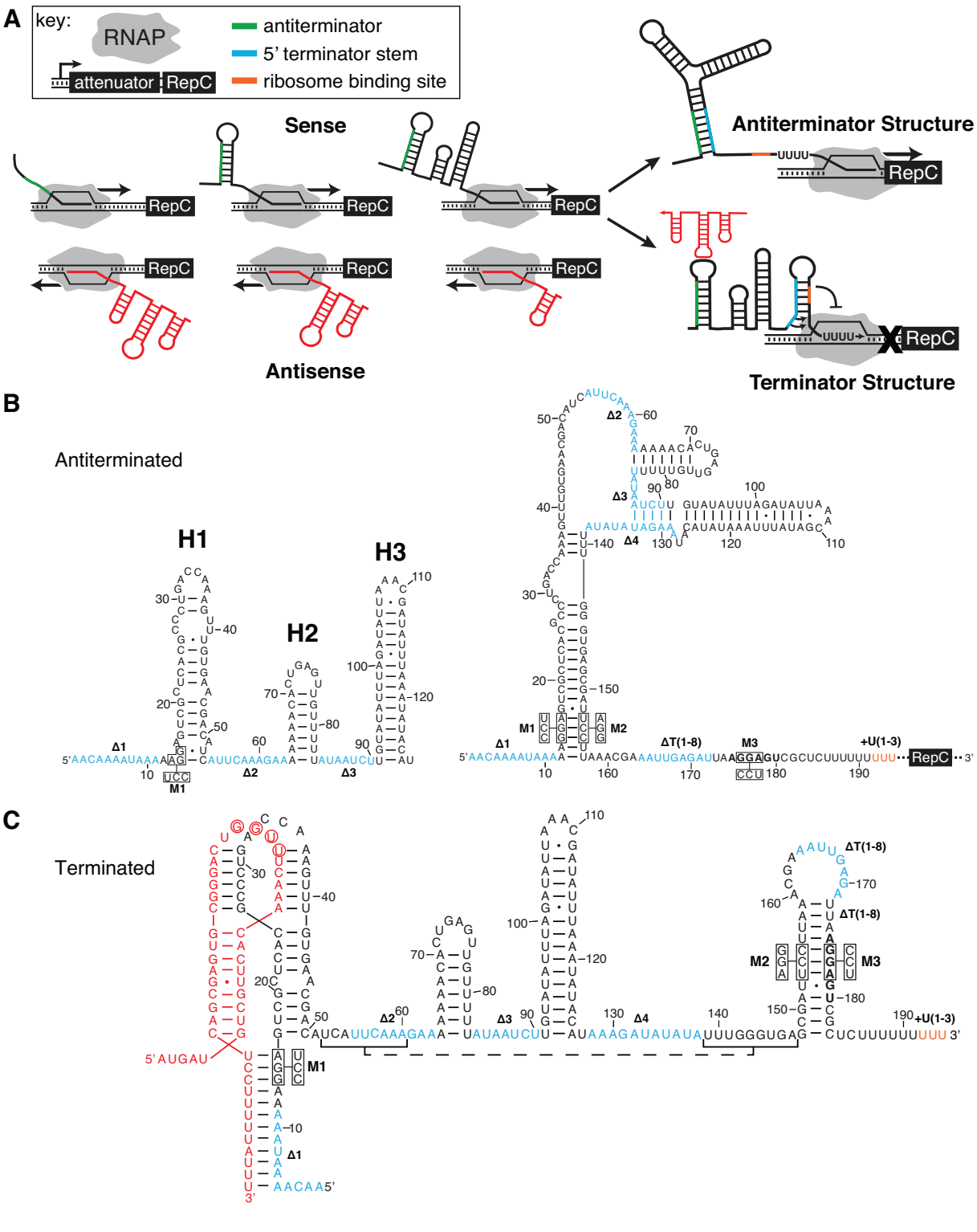
The work presented in this chapter is preparation for submission and contains contributions from Katherine A. Berman, Alexandra M. Westbrook, Jane B. Liao, Ruize Zhuang, and Alexander H. Settle.

6.2 Introduction

Prokaryotic cells employ many RNA regulators to control gene expression that include riboswitches [27, 109], thermometers [210], ribozymes [2, 109], and antisense small RNAs (sRNAs) [21, 22], many of which have served as starting points for further engineering [24, 26, 252]. For example, a number of concepts and platforms borrowed from the antisense-sRNA class of regulators have been used to engineer RNAs that regulate both translation [25, 37, 176, 177] and transcription [135, 136, 253].

In antisense sRNA regulation, RNA-RNA base pairing between an antisense RNA and its target leads to a conformational change in the target RNA that typically alters the structural context of a ribosome binding site (RBS) or intrinsic transcription terminator [21, 22, 24] to modulate gene expression. The first recognized case of antisense-mediated transcription termination, or attenuation, was discovered in the pT181 plasmid, within its copy-number regulation machinery [53]. Replication of pT181 requires a replication protein (RepC) [254] that is translated from the RepC mRNA (sense). Copy number control comes from the ability of a countertranscript (antisense) to bind to the untranslated region (UTR) of the RepC mRNA and favor the formation of an intrinsic terminator within the RepC mRNA that both prevents further transcription into the RepC coding sequence and occludes the RepC RBS (Figure 6.1) [52, 53, 133, 255].

Figure 6.1: General overview and structures of pT181 attenuator system. (A) The wt pT181 attenuation system involves two RNAs, a 'sense' RNA (black) that contains the RepC ORF and an 'antisense' (red) that is a reverse complement of the sense. Interaction between the sense and antisense causes the sense structure to fold into its terminator form, halting further transcription. In the absence of the antisense, the sense folds into a different antiterminator structure instead, allowing downstream transcription. (B) Structures derived from chemical probing analysis [133] of the sense antiterminator form after 127 nts of transcription (left) or 191 nts of transcription. (C) Predicted terminator structure deduced using SHAPE-Seq data. The sense/antisense interaction is likely a kissing-loop interaction that rearranges into a four-way junction, based on analogous RNA-RNA interactions in related copy-control mechanisms [256–258], but does not result in complete complementarity between the antisense and sense hairpins. In (B) and (C) the nucleotides deleted (Δ1-4, ΔT1-8), mutated (M1-M3), or inserted (+U1-3) during testing of the minimalized pT181 sense are marked with blue, boxes, and orange, respectively.


A
key:
RNAP
attenuator
RepC
antiterminator
5' terminator stem
ribosome binding site
Sense
Antisense
Antiterminator Structure
Terminator Structure
RepC
UUUU
B
Antiterminated
H1
H2
H3
Δ1
Δ2
Δ3
Δ4
M1
M2
M3
ΔT(1-8)
+U(1-3)
5'AACAAAAUAAAAAAGCAUUCAAAGAAAUAAUCUU
AUAUAUAGAA
CAUAUAAAUUUAUAGC
AUCUU GUAUAUUUAGAUAUUA
5'AACAAAAUAAAAUAAACGAAAUUGAGAUUAAGGAGUCGCUCUUUUUUUUUU
C
Terminated
5'AUGAU
AACAA5'
CAUCAUUCAAAGAAAUAAUCUUAUAAAGAUAUAUAUUUUGGGUGAGCUCUUUUUUUUUU 3'
AAUUG

Binding of the sense and antisense RNAs is mediated by a 'kissing loop' of Watson-Crick base pairs, a common mechanism of antisense regulation in prokaryotes [187]. The simplicity and ubiquity of these pairing rules led to two successful efforts to create orthogonal pT181 sense/antisense regulator pairs by mutating the sequences of the kissing loop partners or replacing them entirely with analogous RNA antisense hairpins from other RNA regulators entirely [135, 136]. However, despite some structural probing insight [129, 133] the details of antisense/sense binding are unknown as well as mechanistic details involving the influence of the sense hairpin 1 (H1; Figure 6.1) on formation of the terminator or antiterminator structures. Further, because the pT181 attenuator affects transcription, it must make its regulatory decision before the terminator sequence begins to form [259] and the aforementioned structural studies [129, 133] were unable capture this kinetic element.

Recently, we developed a new method to measure how RNAs fold cotranscriptionally *in vitro* at nucleotide resolution using the RNA secondary structure probing technique selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq) (Chapter 5) [107, 123]. Cotranscriptional SHAPE-Seq combines high-throughput SHAPE chemical probing with RNA polymerase arrest along a transcription template to characterize all of the intermediate structures in an RNA folding pathway (Chapter 5). By observing changes in each nucleotide's susceptibility to chemical modification by a SHAPE reagent over the course of transcription, we are able to build folding pathway models to better understand how structures evolve during transcription.

In this work, we examine the folding pathway of the pT181 attenuator system and the effect of sense-antisense binding on the final choice of terminator/antiterminator structure. We then mutate/delete regions of the attenuator sequence to better under-

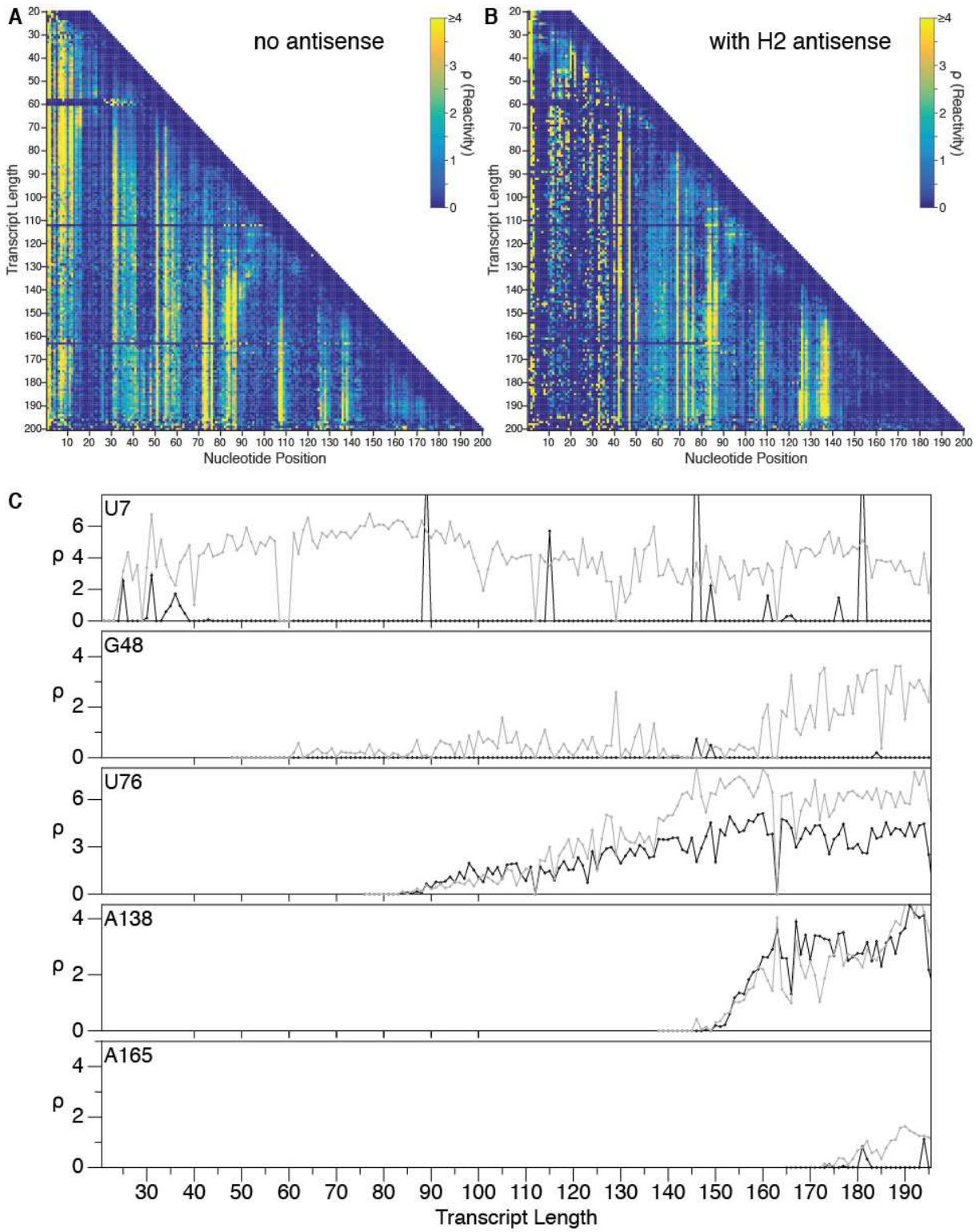
stand pT181 attenuator folding and generate a minimized version for simpler use in the synthetic biology community [24, 135, 136]. Last, we characterize the folding pathway of the minimized pT181 attenuator with cotranscriptional SHAPE-Seq and confirm that it functions the same way as the wild-type (wt) system.

6.3 Results and Discussion

6.3.1 Cotranscriptional folding of the pT181 attenuator

We began by using the recently developed cotranscriptional SHAPE-Seq technique to characterize the folding pathway of the pT181 attenuator RNA (sense). SHAPE-Seq captures structural information about RNA structure by identifying which nucleotides in an RNA molecule are susceptible to modification by a chemical probe. Positions that are less structured are more likely to be modified, allowing us to use this 'reactivity' data to infer structural features within in a RNA. In this work, we focused on the version of the attenuator RNA described in Lucks *et al.* [135] because it contains two mutations: one to prevent transcription of the antisense [51] and another to increase terminator efficiency (**??**) [135]. We also used a smaller version of the antisense RNA that only contained hairpin 2 (H2), containing the stem-loop that interacts with hairpin 1 (H1) of the sense RNA that has been shown to be sufficient for termination [138].

Figure 6.2: Cotranscriptional SHAPE-Seq of the pT181 attenuator with and without the antisense hairpin 2. Cotranscriptional SHAPE-Seq data for the pT181 attenuator are shown for transcript lengths 20-200 nts without antisense present (A) and with 10-fold excess of the antisense hairpin 2 (H2) (B). Single nucleotide trajectories for selected nucleotide positions over the course of transcription without antisense (gray) and with 10-fold excess of H2 (black).


A
no antisense
B
with H2 antisense
Transcript Length
Nucleotide Position
ρ (Reactivity)
C
U7
G48
U76
A138
A165
ρ
Transcript Length

We compared the reactivity matrices we obtained for the attenuator with and without the antisense H2 (Figure 6.2A,B) and immediately observed a number of differences between them. Despite these differences, however, the each respective attenuator reactivity matrix maintains the same qualitative reactivity patterns over increasing transcript lengths until 166 nts are transcribed where we observe the transition to the antiterminator form without antisense and no transition with antisense.

6.3.2 The pT181 attenuator refolds into an antiterminator late in transcription

In the absence of the antisense RNA, we observe step-wise folding of the pT181 attenuator up to transcript length 166 consistent with the formation of the three individual hairpins H1-H3 (Figure 6.2A). Early in the transcription, the folding of H1 can be observed in the first 70 nts of transcription (after 56 nts have existed RNAP). First, we observe a patch of mild reactivity at nts 20-23 that decreases shortly after their base pairing partners, nts 45-47, emerge from the polymerase. We also observe a short-lived cluster of highly reactive nucleotides spanning positions 14-16 at the 5' base of H1 that decreases to a lower value as the rest of H1 exits the polymerase. As transcription continues past 70 nts, a clear span of reactivity across nts 31-41 (H1 loop region) bracketed by low reactivity regions heralds the complete folding of H1. Interestingly, A51 exhibits high reactivity while A16 does not. Likely, A16 stably stacks between G15 and G17, lowering its reactivity [90, 180], but the neighboring purine content of A51 prevents the formation of a stable stack. All of these reactivity observations in H1 correlate well to previously collected in-cell SHAPE-Seq data for a truncated pT181 sense RNA [129].

Without antisense, the formation of H2 and H3 following H1 is clearly observed (Figures 6.2A and 6.3A). After approximately 130 nts have been transcribed, high reactivities across positions 51-61, 72-76, and 83-87 can be seen interspersed with regions of lower reactivity, corresponds very well to the expected structural context of H2: a hairpin with single-stranded regions on each side. The formation of H3 can also be observed by the reactivity decreases in positions 89-93 over lengths 133-143, corresponding to the emergence of the 3' base pairing partners of positions 89-93 from the RNAP footprint (~14 nt). The long, stable H3 helix is easily identified at longer lengths by the high reactivity at A107 and A108 that develops as transcription proceeds and the large surrounding areas of very low reactivity. At these longer lengths, we also observe a decrease in the first 11 nts that is potentially due to transient interactions with unpaired elongating RNA before the antiterminator refolding occurs.

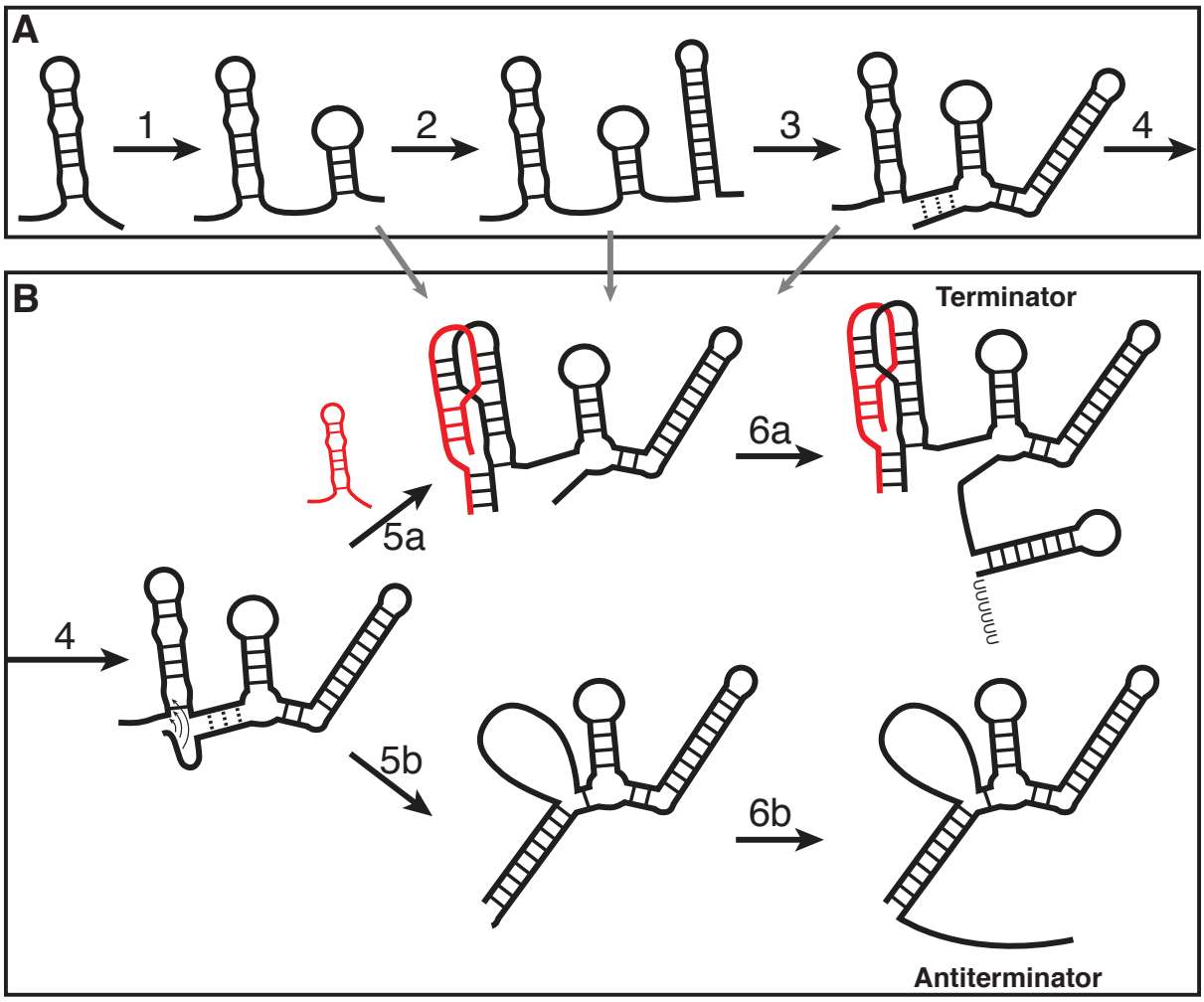


Figure 6.3: Folding pathway for the wt pT181 attenuator. (A) Antisense independent folding. Based on sequence analysis and cotranscriptional SHAPE-Seq data, the first four folded stages of the attenuator folding pathway are shown. Following the formation of H1, H2 and H3 fold individually during transitions 1 and 2. Then, during transition 3 the RNA sequence after H3 begins to fold back on itself, bringing the 3′ end of the RNA close to H1. Transition 4 marks the addition of extra single-stranded nucleotides that are complementary to the 5′ half of H1, but do not have a free binding partner. (B) As the RNA lengthens, it can begin to base pair with H1 and unwind it (5b), leading to a complete antiterminator (6b). If the antisense RNA is present, however, it can bind at any point before transition 5 to generate the structure depicted after transition 5a. The sense/antisense structure that forms prevents H1 from being unwound and allows the terminator stem to form.

When RNAP reaches 166 nts, the attenuator makes its regulatory decision. At this time, 153U emerges from the RNAP exit channel to pair with the stacked A16, triggering the displacement of the 3′ half of H1 (Figure 6.3). There are two pieces of evidence that support hypothesis. First, we see a spontaneous increase in the reactivities of nts 42-50, which are on the 3′ bottom half of H1 and are displaced by this binding event. These changes also persist to the end of transcription, indicating the new conformation is stable, even after the 3′ side of the terminator (nts 172-185) exits the polymerase at the end of transcription. Also, high reactivities at nts 126-130 and 136-138 match well to an expected three nucleotide bulge and unstructured region, respectively, that would form in the antiterminated structure. Further, we also observe a broad area of moderate reactivity across nts 156-175 that would indicate that the nucleotides preceding the ribosome binding site (RBS; 175-180) are likely interconverting between many unstable conformations. The observed increase in reactivities at nts 1-11 may be due to the combined stabilization of the rest of the RNA within the antiterminator structure as well as a physical separation of nts 1-11 from potential transient binding partners in unstructured region neighboring H2.

6.3.3 The pT181 antisense binds quickly and changes the attenuator's folding pathway

The presence of the antisense RNA changes the folding pathway of the attenuator. Unfortunately, we could not obtain a good set of reactivities for the ~41 nts and may be due to the presence of the antisense RNA blocking reverse transcription (RT), which is supported by the observation that many reads are aligned between nts 40-50. The antisense RNA also causes nts 42G and 47C in the sense H1 to become highly reactive.

Further, an interesting shift occurs over the course of transcription where 43U is highly reactive until nt 142 is synthesized at which point it becomes low and C50 increases until ~170 nts are transcribed.

The meaning of these highly reactive positions is still somewhat unclear, although we hypothesize that the antisense RNA is interacting with the sense RNA through a kissing-loop interaction that proceeds to a four-way junction where the highly reactive positions are due to the steric requirement that some nucleotides remain unbound. Similar RNA sense/antisense mechanisms in other plasmid copy-number control systems have been shown to interact as partially interacting loops as depicted in Figure 6.1C [256–258, 260], although it has never been demonstrated for pT181.

The reactivities of the attenuator with the antisense look largely the same as without except for three key differences. First, there is no increase in reactivity between nts 42-50 after nt 166 is synthesized. In that region, the reactivity of 43U and 47C were already high and do not significantly change. A second difference is in the region between nts 69-74 in the loop of the sense H2. The attenuator exhibits high reactivities in that region with or without the antisense present, but the pattern is different between the two (Figure 6.2). Last, the positions between 156-175 are poorly reactive in contrast to without the antisense. The reactivity differences in this region follow with the expectation that the RBS region is paired when the antisense is present (terminator form) and unpaired when absent (antiterminator form). Interestingly, however, we do not observe a cluster of highly reactive nucleotides between nts 160-171 that we might expect for the terminator loop.

When interacting with the antisense, the final structure that the sense adopts likely contains a four-way junction at the 5′ end with the antisense. However, direct evidence of that interaction is not currently available due to the scarcity of alignments

in that region, likely due to a duplex structure that forms during RT that blocks the reverse transcriptase from extending. By sequencing analysis, we would also expect an interaction between nts 52-60 and 139-147, however, nts 52-60 are reactive with or without the antisense present and neither region exhibits a reactivity change during cotranscriptional folding.

6.3.4 Cotranscriptional vs. in-cell

We next compared our results from the cotranscriptional SHAPE-Seq experiment to an in-cell SHAPE-Seq experiment [113]. To perform the in-cell SHAPE-Seq experiment, we first mutated the polyU to allow terminator readthrough in order to provide a convenient spot for RT priming. When we looked at reactivity maps for the fully transcribed pT181 attenuator in the cell (Figure D.1), we found that reactivity maps generally matched well to the *in vitro* experiments, although the AU-rich hairpins (H2 and H3) were somewhat less stable in the cell and exhibited higher reactivity. We were also able to observe a clear differentiation of terminator stem-loop vs. antiterminator for nts 155-195. With the antisense present, the attenuator exhibited one cluster of reactivity in this region, at the location of the terminator hairpin, while without antisense, reactivities were diffuse through the region, including the RBS, which was low with antisense (Figure D.1). The comparison between the in-cell and cotranscriptional SHAPE-Seq data shows good agreement, suggesting that the structures we observe in vitro are a good approximation of how pT181 cotranscriptionally folds in the cell.

6.3.5 Cotranscriptional vs. equilibrium folding

To analyze the effect of cotranscriptional folding on the pT181 attenuator, we performed an equilibrium refolding experiment. The equilibrium refolding data was collected by first generating all of the intermediate lengths without antisense then extracting the RNA and refolding it in 1X transcription buffer with or without the antisense present (Figure 6.4).

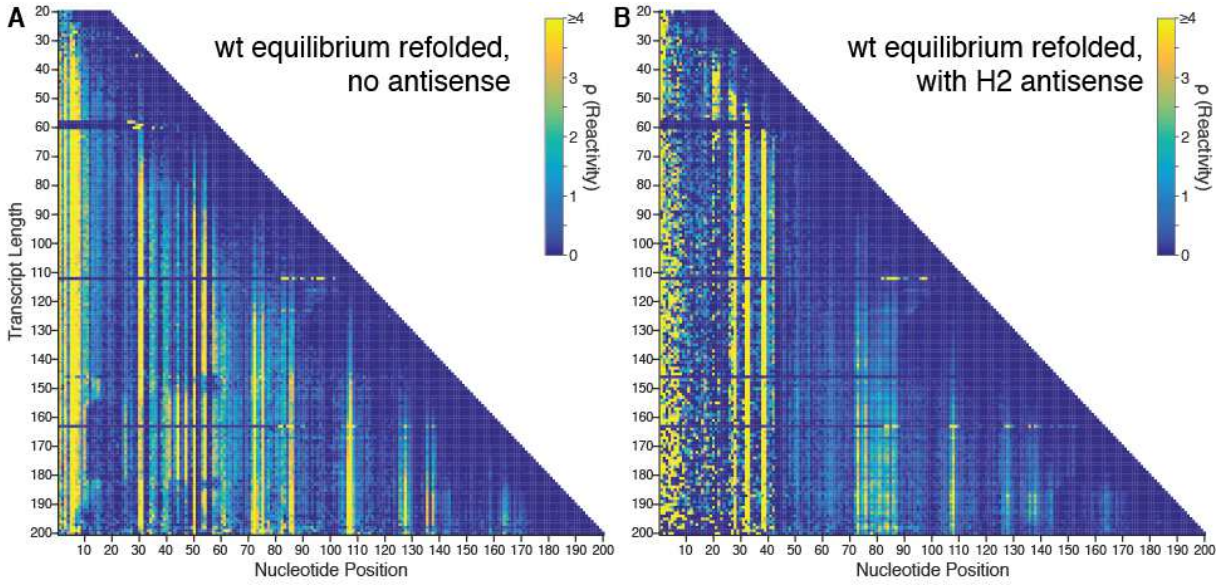


Figure 6.4: Equilibrium refolding of the pT181 attenuator with and without the antisense hairpin 2. After performing the transcription and RNAP arrest (without antisense), the RNA products were extracted and refolded in 1X transcription buffer without (A) or with (B) the antisense H2 present. (A) Without antisense, the reactivity map produced by the attenuator follows the proposed folding pathway to antitermination (Figure 6.3) very closely until length 182, when the terminator becomes more stable than the antiterminator when refolded. (B) When refolded with the antisense present, nts 29, 32, 33, 38, and 39 all exhibit very high reactivity [256], although in a different pattern than the cotranscriptional experiment. Also, the reactivities between H1 and H3 are markedly different, with the 5′ half of H2 and its preceding nts being largely unreactive, and the 3′ half of H2 being highly reactive.

When refolding without antisense, we clearly see the step-by-step evolution of the attenuator folding pathway as depicted in Figure 6.3. As the transcribed RNA lengths get longer, the reactivity traces are indicative of an ordered fold of H1→H3 (Figure 6.4). It also appears that the sequence downstream of H3 folds back and base pairs with the single-stranded region between H1 and H2 (Figure 6.3), easily observable as a drop in reactivities between nucleotides 53-60 and an increase in 14-16. Shortly after, the reactivities between nts 42-49 increase, as the antiterminator structure is more energetically favored. However, when folded in equilibrium the longest lengths show decreased reactivity at positions 42-49 and increased reactivities in the terminator loop nts, indicating that once the terminator is able to form (at 183 nts), it is the preferred thermodynamically stable structure.The preference for the terminator form is supported by minimum free energy predictions using Mfold [261], yielding ΔGs of -52.93 kcal/mol for the terminator structure vs. -45.83 kcal/mol for the antiterminator structure. Interestingly, the base pairs that can form between 52-60 and 139-147 only appear to form for short transcript length ranges during equilibrium refolding. Combined with the observation that these positions maintain moderate reactivity through cotranscriptional folding, it is likely that this interaction serves as a short-lived transition point during transcription such that base pairing of 52-60 to 139-147 immediately results in unwinding of the H1 helix and transition into the antiterminator structure.

When refolding with antisense, we observe a series of high reactivities between positions 25-39 that do not agree well with the four-way junction proposed above, suggesting that during equilibrium refolding the antisense binds to the sense differently and does not proceed to the extensively paired four-way junction, which may be dependent on cotranscriptional folding (Figure D.2). Thus, it seems unlikely that a full duplex is ever evolved between the sense and antisense, similar to what has been observed for antisense-mediated translational repression systems [256, 257]. It is also

interesting that positions 43-71 exhibit very low reactivities that were somewhat higher when cotranscriptionally folded. Further, the high reactivities between nts 72-87 are more consistently high than when cotranscriptionally folded.

Ultimately, the equilibrium folding results demonstrate the need for kinetics in order to obtain the correct antiterminator structure to allow for transcriptional read-through. Like the fluoride riboswitch (Chapter 5), equilibrium refolding analysis showed that the terminator form is more stable. Thus when performing basic refolding analysis, such as the original work by Brantl *et al.* [133], it is important to consider the potential impact of cotranscriptional folding on the final structure, especially for regulatory RNAs. Otherwise, incorrect structures may be mistakenly studied, while the relevant ones remain unobserved.

6.3.6 Deletion analysis and sequential minimization of the attenuator

Given the use of the pT181 attenuator in the synthetic biology space [55, 129, 135, 136, 138, 252, 262–264], there is a great interest in not only understanding the mechanism of the pT181 attenuator, but also in creating smaller versions with high signal-to-noise ratios to be easily networked together to form larger biological circuits [24, 26]. One key element to maintaining a high signal-to-noise ratio is designing repressors, like the pT181 attenuator, with the highest percent repression possible. Therefore, we analyzed a set of deletions and mutations with the combined goal of corroborating our SHAPE-Seq analysis and creating a minimized version of the attenuator.

With our secondary goal of creating an improved part for RNA synthetic biology,

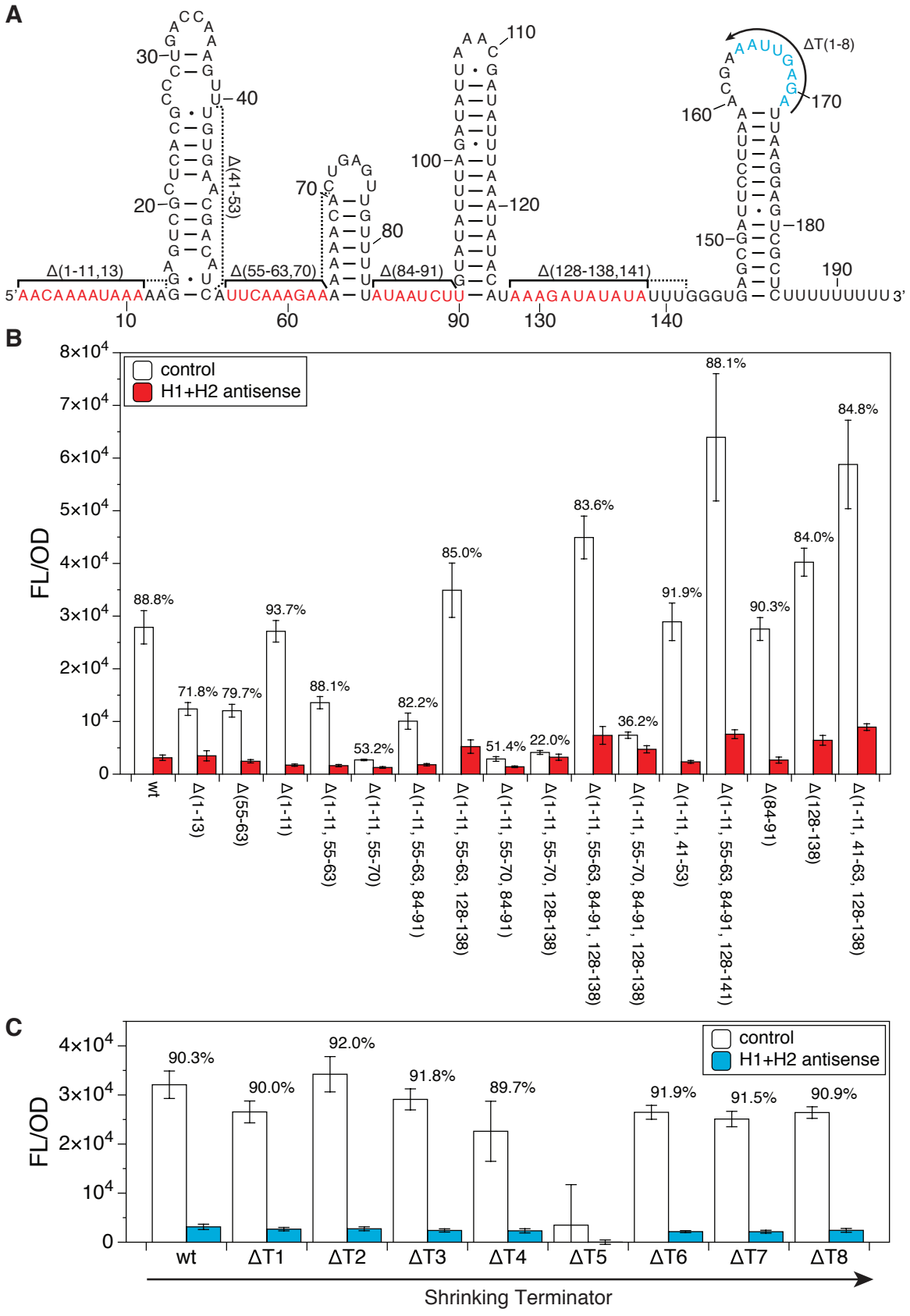
we began by mutating the RBS of the attenuator to disrupt the translational capabilities of the wt system. To disrupt the RBS, we swapped three base pairs in the middle of the terminator stem (M2 & M3: Figure 6.1) and made the compensatory mutation in H1 (M1) to allow the antiterminator to form as well (Figure D.3). As had been previously observed [53], disrupting terminator or antiterminator base pairing causes the attenuator to predominately adopt the antiterminator or terminator structures, respectively.

Current practice is to include a gene fragment of RepC after the attenuator sequence and create transcriptional fusions downstream. However, the RepC gene fragment is nearly 100 base pairs, roughly 50% the length of the attenuator, representing a large, potentially unnecessary sequence. Efforts to shorten the RepC gene fragment resulted in varying levels of gene expression that did not correlate to the amount of the gene fragment deleted (Figure D.4), likely due to changing structural contexts surrounding the transcriptional fusion RBS. Because the RepC sequence in question lies entirely downstream of the attenuator sequence, all the changes we observed making RepC deletions or mutations must be translational effects in the second reading frame. Thus, we deleted the RepC ORF in the minimized design to remove all translational elements from the regulator.

In an effort to minimize the regulator and better understand its function, we sequentially deleted regions that appeared unstructured in either the terminator or antiterminator form (Figure 6.5). We began with the 13 leading nucleotides. Deleting all 13 resulted in a decrease in the measured ON level, likely due to increased difficulty of antiterminator formation due to the missing contributions of A12 and A13 to the antiterminator structure (Figure 6.5A, B). Then we deleted the regions between the stem loops, finding that losing any of them, or all, did not result in significantly different levels of repression, although resulted in higher gene expression overall. Interestingly,

deletion of the nts between H1 and H2 (55-63) lowered %repression, but was largely rescued by the deletion of nts 1-11 (Figure 6.5A, B). We also sequentially deleted nucleotides from the 3' side of the terminator loop sequence (Figure 6.5C). Our goal in shrinking the terminator loop was to increase the speed and stability of terminator hairpin folding [265]. However, we found that the percent repression obtained after antisense addition varied little as the terminator loop size was decreased, with the exception of ΔT5, which exhibited low fluorescence regardless of antisense presence.

Figure 6.5: Minimizing the pT181 attenuator. (A) pT181 attenuator sequence. Deletions are indicated; nucleotides that are colored are removed in the final minimized structure. (B) pT181 attenuator deletions were tested for dynamic range (%repression) using the H1+H2 antisense. Many of the deletions are well tolerated, except for Δ(64-70) and Δ(1-13) as well Δ(55-63), unless accompanied with Δ(1-11). (C) The terminator loop was sequentially truncated to test if a smaller loop would improve termination efficiency. No significant changes in dynamic range were observed. Variant ΔT5 does not express GFP and appears unstable. Error bars represent one standard deviation of 12 replicates.


A
Δ(1-11,13)
Δ(55-63,70)
Δ(84-91)
Δ(128-138,141)
Δ(41-53)
ΔT(1-8)
5'AACAAAAUAAAAAG—CAUUCAAAGAAA—UAUAAUCUU—AUAAAGAUAUAUAUUUUGGGUG—CUUUUUUUUU3'
10
60
90
130
140
B
control
H1+H2 antisense
FL/OD
88.8%
71.8%
79.7%
93.7%
88.1%
53.2%
82.2%
85.0%
51.4%
22.0%
83.6%
36.2%
91.9%
88.1%
90.3%
84.0%
84.8%
wt
Δ(1-13)
Δ(55-63)
Δ(1-11)
Δ(1-11, 55-63)
Δ(1-11, 55-70)
Δ(1-11, 55-63, 84-91)
Δ(1-11, 55-63, 128-138)
Δ(1-11, 55-70, 84-91)
Δ(1-11, 55-70, 128-138)
Δ(1-11, 55-63, 84-91, 128-138)
Δ(1-11, 55-70, 84-91, 128-138)
Δ(1-11, 41-53)
Δ(1-11, 55-63, 84-91, 128-141)
Δ(84-91)
Δ(128-138)
Δ(1-11, 41-63, 128-138)
C
control
H1+H2 antisense
FL/OD
90.3%
90.0%
92.0%
91.8%
89.7%
91.9%
91.5%
90.9%
wt
ΔT1
ΔT2
ΔT3
ΔT4
ΔT5
ΔT6
ΔT7
ΔT8
Shrinking Terminator

As a complement to the sense deletion effort, we also examined the minimal sequence required in the antisense to achieve maximum repression. Nucleotides were deleted from the 5′ end of the H1+H2 antisense to get the H2 sequence used in the cotranscriptional folding experiments. Further deletions were added that H2 sequence and assayed for repression level (Figure D.5). We found that despite the ability of the attenuator to function without a complete H1 (Figure 6.5), the antisense required a complete hairpin H2 and only tolerated a few extra nucleotides being removed from flanking single-stranded regions before the repression level was reduced. These single-stranded ends likely help to bring the antisense in proximity to the sense RNA to allow time for the four-way junction to form, however, it was shown in the related CopA/CopT system that only the hairpin was needed [257].

6.3.7 A minimized pT181 attenuator

To generate the minimized pT181 sequence, we combined the internal deletions, terminator loops deletions, RBS swamp mutations, and RepC deletion to arrive a smaller version of the pT181 attenuator (Figure 6.6). All of the changes resulted in an attenuator that was ~150 nts shorter (including the RepC sequence) and exhibited similar levels of repression (Figure D.6). As a last attempt to try to improve termination efficiency further, we tried adding up to three Us to the end of the polyU sequence. However, no improvements in termination or repression level were gained (Figure D.5).

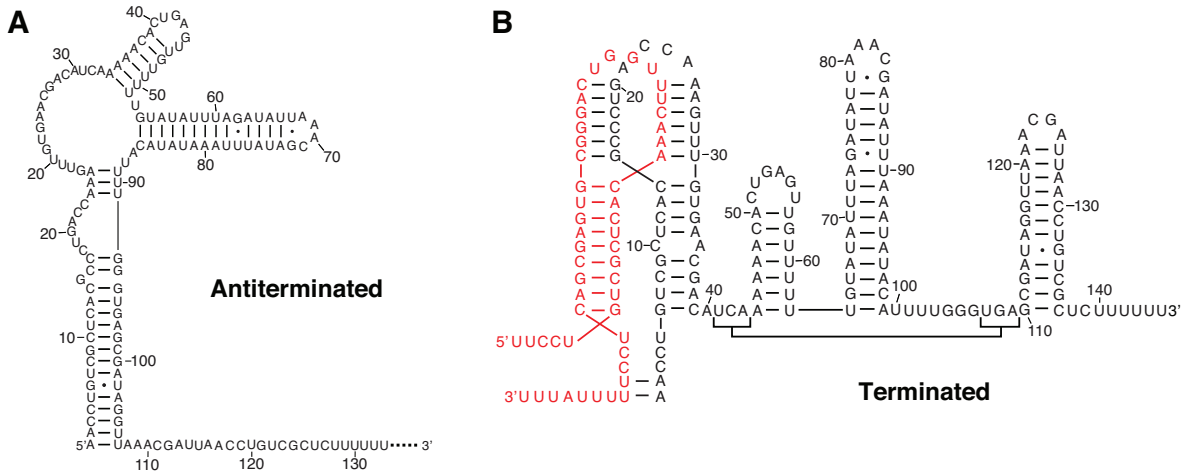


Figure 6.6: Minimized pT181 attenuator. Shown are the predicted structures for the minimized pT181 attenuator in its antiterminated and terminated forms. Antisense H2 (red) is shown interacting with the minimized attenuator H1.

To confirm that the minimized pT181 attenuator was still traversing the same folding pathway as the original attenuator, we performed another cotranscriptional SHAPE-Seq experiment on the minimized attenuator (Figure D.7). Ignoring regions that were removed in the minimized version, we observed very similar reactivity patterns and trends for both sizes of the attenuator, indicating that all of the deletions did not change how the minimized attenuator folded during transcription.

6.4 Conclusion

In this work, we apply the recently developed cotranscriptional SHAPE-Seq technique to answer an aging question about how RNA-mediated transcriptional attenuators sense and respond to antisense RNA and elicit a structural and functional change. Given the recent growing interest in RNA synthetic biology [26], and the firm rooting of the pT181 attenuator within it [55, 129, 135, 136, 138, 252, 264], we expect that

this mechanistic study of the pT181 will not only help guide the design of future pT181-based regulators, but also inspire a new approach to studying RNA regulators.

6.5 Future work

6.5.1 Determining the key motif

While the pT181 attenuator has served well in the synthetic biology space as an RNA-responsive transcriptional regulator, of which there are few natural examples known [23], there has been no comprehensive investigation into the principles of how the attenuator works. This and two other works [129, 133] are the only studies that have looked closely at the details of the mechanism of pT181 folding and antisense recognition.

Starting with the results discussed above, a greater library of deletions and mutations would greatly help determine what the true minimal elements of the attenuation mechanism are. We suspect, like the recently published STARs design [55], that only a fraction of the attenuator is required for function and the rest facilitates the nucleation of the antitermination interaction. With the minimal core of the attenuator completely understood, new designs could be much more easily designed *de novo* that would likely be orthogonal by design.

6.5.2 Computational Modeling

To support the model of antitermination, we propose modeling the antitermination point using oxRNA [266], a coarse-grained 3D simulation package for RNA. Due to the speed afforded by use of coarse-grained models, many different interaction scenarios could be tested to support or refute the pT181 folding pathway presented above.

To specifically study pT181 antiterminator formation, we propose first constraining the system to fold the structures present at the various stages in the proposed pT181 attenuator folding model then releasing the restraints to begin the simulation. The preferences for the terminator or antiterminator structure would then be assessed and all of the intermediate stages compared. Also, by comparing simulations of the minimal and wt pT181 attenuators, which should have the same folding pathway, a higher confidence level in the accuracy of the simulation can be established to support the cotranscriptional SHAPE-Seq data.

6.5.3 Creating orthogonal versions and building logics

The next step for adoption of the new minimalized pT181 attenuator into the synthetic biology community is to demonstrate orthogonality and composability. Currently, the minimized version of the attenuator demonstrates neither (data not shown). However, establishing the cause of the poor orthogonality by determining what deletion/mutation caused the decrease would shed light on what makes different antisense sequences orthogonal. Solving the minimized attenuator's orthogonality problem would also immediately make it useful to replace the current bulkier version immediately, assuming the composability issue can also be solved.

Concatenating multiple copies of the minimized attenuator together results in weak gene expression, suggesting that the first instance is interacting with downstream copies to cause them to adopt the terminator structure without the antisense present. We tried a version with the 3' half of the sense H1 deleted, hypothesizing that it was interacting with H1 of a downstream attenuator sequence. However, that did not solve the termination problem. Current efforts are underway to include ribozymes as spacer elements between adjacent attenuators, but no design has yet to exhibit promising results. Determining the cause of the self-attenuation would greatly help.

6.6 Acknowledgments

We thank Katherine Berman, Alexandra Westbrook, Jane Liao, Alexander Settle, and Ruize Zhuang for collecting gene expression functional data, Katherine Berman for extensive SHAPE-Seq work and data collection, and Jane Liao, Alexander Settle, and Ruize Zhuang for cloning many of the plasmids used in this study. We also thank Eric Strobel for lending reagents.

6.7 Methods

6.7.1 Plasmids

The 'wild-type' pT181 sequence was taken from Lucks *et al.* [135], replacing the TrrnB operon fragment used to terminate the superfolder GFP (SFGFP) sequence with a stronger double terminator. The antisense plasmid used originated from the same

work, but the terminator sequence was replaced with the double terminator from Watters *et al.* [113]. Mutations to the sense and antisense plasmids were introduced using a PCR-ligation strategy and maintained the rest of the vector sequences.

6.7.2 Strains, growth media, and fluorescence assays

To test the ON and OFF levels of the each sense/antisense pair, each sense plasmid was transformed into chemically competent *E. coli* TG1 cells with an antisense plasmid. Where indicated, a control antisense plasmid that did not contain antisense RNA sequence was used. Transformed cells were plated on LB/agar media with 100 g/mL carbenicillin and 34 g/ml chloramphenicol and incubated overnight at 37 °C. The next day, four colonies were picked and grown in 200 μL of LB with antibiotics in a 2 mL 96-well block (Costar) and grown approximately 17 h overnight at 37 °C at 1000 rpm in a VorTemp 56 (Labnet) benchtop shaker. Four microliters of this overnight culture was then used to subculture into 196 μL of freshly prepared M9 minimal media with antibiotics. Subcultures for in-cell SHAPE-Seq experiments were scaled up 6x in volume. The subcultures was grown for 4 h before measuring fluorescence by adding 100 μL culture to 100 μL PBS and assaying fluorescence intensity using wavelengths 485 nm and 528 nm for excitation and emission, respectively. Fluorescence was normalized by OD_{600} after subtracting a media blank. Relative fluorescence levels of each culture were determined by normalizing the fluorescence readout by optical density (FL/OD) and subtracting the FL/OD of cells transformed with control sense and antisense plasmids.

6.7.3 RNA modification and extraction for in-cell SHAPE-Seq

RNA was modified in cells with 1-methyl-7-nitroisatoic anhydride (1M7) by adding 500 μL of subculture to either 13.3 μL 250 mM 1M7 in DMSO (6.5 mM final) (+) or 13.3 μL DMSO (-) for 3 min before RNA extraction. Immediately following RNA modification, both modified (+) and control (-) samples were pelleted and resuspended in 100 μL of pre-heated (95 °C) Max Bacterial Enhancement Reagent (Life Technologies) and incubated for 4 min. Then TRIzol Reagent (Life Technologies) was used to extract the RNA according to the manufacturer's protocol and dissolved in 10 μL of water.

6.7.4 Template preparation for cotranscriptional SHAPE-Seq

DNA template libraries for each pT181 sense were prepared by combining individual PCR amplifications for each length of transcript from 5' to 3'. Each template length was amplified in a 25 μL PCR using Taq polymerase (New England Biolabs) that included 25 pM of the forward primer (ATAAGCTTCCGATGGCGCGC), 0.15 μL plasmid DNA template, and 25 pM reverse primer. The reverse primer was unique to each length generated, and incorporated an EcoRI site. Reaction mixes were run using a standard thermal cycle program consisting of 30 cycles of amplification using an annealing temperature of 52 °C. After thermal cycling, the PCRs were pooled and precipitated with EtOH before gel extraction on a 1% agarose gel using the QIAquick Gel Extraction Kit (Qiagen). The concentration of the purified template was measured using the Qubit Fluorometer (Life Technologies) and the molarity of the template was calculated using the median template length.

6.7.5 In vitro transcription for cotranscriptional SHAPE-Seq

50 μL total reaction mixtures containing 100 nM linear DNA template library (see above) and 4 U of *E. coli* RNAP holoenzyme (New England Biolabs) were incubated in transcription buffer (20 mM Tris-HCl pH 8.0, 0.1 mM EDTA, 1 mM DTT and 50 mM KCl), 0.2 mg/mL bovine serum albumin, and 500 μM NTPs for 7.5 min at 37 °C to form open complexes. When present, 10 pmol of antisense RNA (wt, H2 only) was first folded by denatured at 95 °C for 2 min, 1 min snap-cool on ice, and refolded in 1X transcription buffer before addition after open complex formation. After forming open complexes the EcoRI Gln111 dimer was added to a final concentration of 500 nM and incubated at 37 °C for another 7.5 min. Immediately following the second incubation, single-round transcription reactions were initiated by addition of $MgCl_2$ to 5 mM and rifampicin to 10 g/ml and proceeded for 30 seconds. Cotranscriptional experiments were then directly SHAPE modified (see RNA modification and purification below). Equilibrium refolding experiments were stopped by addition of 150 μL TRIzol solution (Life Technologies), purified, and equilibrium refolded in transcription buffer before SHAPE modification as described below (see RNA modification and purification).

6.7.6 RNA modification for cotranscriptional SHAPE-Seq

For cotranscriptional experiments the 30 second transcription products were immediately SHAPE modified by splitting the reaction and mixing half with 2.78 μL of 400 mM benzoyl cyanide (BzCN; Pfaltz & Bower) in anhydrous dimethyl sulfoxide (DMSO; (+) sample) or anhydrous DMSO only (Sigma Aldrich; (-) sample) for ~2 seconds before addition of 75 μL of TRIzol solution. Transcription products for equilibrium refolding

had 150 μL TRIzol added after in vitro transcription. Both were extracted according to the manufacturer's protocol and dissolved in 20 μL total of 1X DNase I buffer (New England Biolabs) containing 1 U of DNase I enzyme. Digestion proceeded at 37 °C for 30 min, after which 30 μL of RNase-free H_2O was added, followed by 150 μL TRIzol. The RNA samples were then extracted again according to the manufacturer's protocol and dissolved in either: 10 μL 10% DMSO in H_2O (cotranscriptional experiments) or 25 μL RNase-free H_2O (equilibrium refolding experiments). Equilibrium refolding experiment samples were then heated to 95 °C for 2 min, snap cooled on ice for 1 min, and refolded by adding 24 μL 2X folding buffer for 10 min at 37 °C (1X: 20 mM Tris-HCl pH 8.0, 0.1 mM EDTA, 1 mM DTT, 50 mM KCl, 0.2 mg/mL bovine serum albumin, and 500 μM NTPs). After 10 min, either 1 μL of RNase-free H_2O was added or 1 μL of 10 μM folded antisense (see above). RNA modification of the equilibrium refolding samples was performed as described above, followed by the addition of 30 μL RNase-free H_2O and 150 μL TRIzol and extracted a third time according to the manufacturers instructions. The resulting pellet was dissolved in 10 μL 10% DMSO in H_2O.

6.7.7 RNA Linker preparation

The phosphorylated linker (5'Phos-CUGACUCGGGCACCAAGGA-ddC-3') was purchased from Integrated DNA Technologies and adenylated with the 5' DNA Adenylation Kit (New England Biolabs) according to the manufacturer's protocol at a 10X scale, dividing the reactions into 50 μL aliquots. After completion the reactions were extracted using TRIzol and diluted to a 2 μM stock.

6.7.8 Linker ligation for cotranscriptional SHAPE-Seq

To the modified and unmodified RNAs in 10% DMSO (see RNA modification for cotranscriptional SHAPE-Seq above), 0.5 μL of SuperaseIN (Life Technologies), 6 μL 50% PEG 8000, 2 μL 10X T4 RNA Ligase Buffer (New England Biolabs), 1 μL of 2 μM 5'-adenylated RNA linker, and 0.5 μL T4 RNA Ligase, truncated KQ (200 U/μL; New England Biolabs) were added to bring the total reaction volume to 20 μL. The reactions were mixed well and incubated overnight (>10 h) at room temperature. The completed linker ligations were brought to 150 μL with RNase-free H_2O before addition of 15 μL 3 M NaOAc, 1 μL 20 mg/mL glycogen, and 450 μL EtOH for EtOH precipitation. Precipitated pellets were dissolved in 10 μL RNase-free H_2O.

6.7.9 Reverse transcription

To each dissolved RNA sample, 3 μL of 0.5 μM reverse transcription primer were added. For in-cell SHAPE-Seq, primers 5'-Biotin-TTTATCGGCCGAAGCAGGTAG (antisense) and 5'-Biotin-CAACAAGAATTGGGACAACTCCAGTG were added. For cotranscriptional SHAPE-Seq experiments, 5'-Biotin-GTCCTTGGTGCCCGAGT was added. The RNA samples mixed with primer were denatured at 95 °C for 2 min, then 65 °C for 5 min, and snap-cooled on ice for 1 min before addition of RT master mix [6 μL 5X First Strand Buffer (Life Technologies), 1 μL 10 mM dNTPs, 0.5 μL H_2O, and 0.5 μL Superscript III (Life Technologies)]. Primer extension was performed by incubating at 52 °C for 25 min followed by 65 °C for 5 min. The RNA was hydrolyzed with either 1 μL 10 M NaOH for in-cell experiments or 1 μL 4 M NaOH for cotranscriptional experiments then partially neutralized with 5 μL or 1 μL of 1 M hydrochloric acid, respectively, and ethanol precipitated by addition of 78 μL (69 μL for cotranscriptional)

of cold EtOH. The precipitated pellets were dissolved in 22.5 μL of nuclease-free H_2O.

6.7.10 DNA adapter ligation

To each sample, 3 μL 10X CircLigase Buffer (Epicentre), 1.5 μL 50 mM $MnCl_2$, 1.5 μL 1 mM ATP, 0.5 μL 100 μM DNA adapter (5'-Phos-AGATCGGAAGAGCACACGTC TGAACTCCAGTCAC-3CSpacer-3'), and 1 μL CircLigase I (Epicentre) were added. The reaction was incubated at 60 °C for 2 hr, then 80 °C for 10 min. The ligated DNA was EtOH precipitated, dissolved in 20 μL of nuclease-free H_2O, purified using 36 μL of Agencourt XP beads (Beckman Coulter; according to manufacturer's instructions), and eluted with 20 μL TE buffer.

6.7.11 Quality analysis

For quality analysis, a separate PCR reaction for each (+) and (-) sample was mixed by combining: 13.75 μL nuclease-free H_2O, 5 μL 5X Phusion Buffer (NEB), 0.5 μL 10 mM dNTPs, 1.5 μL of 1 μM labeling primer (Fluor-GTGACTGGAGTTCAGACGTGTGCTC), 1.5 μL of 1 μM primer PE_F (5'-AATGAT ACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'), 1 μL of 0.1 μM selection primer (5'-CTTTCCCTACACGACGCTCTTCCGATCT(RRRY/YYYR)-specific sequence-3'), 1.5 μL ssDNA library (+ or -), and 0.25 μL Phusion DNA polymerase (NEB). The specific sequences for the selection primer are as follows: TTTATCGGCCGAAGCAGGTAGA*G*G*C (in-cell, antisense), CAACAAGA ATTGGGACAACTCCAGT*G*A*A*A*A*G (in-cell, sense), and GTCCTTGGTGCCCG AG*T*C*A*G (cotranscriptional). The fluorophores for the labeling primer are VIC or

NED (Applied Biosystems) and were added to the (+) or (-) samples, respectively. The selection primers were purchased from Integrated DNA Technologies and contain an internal barcode, RRRY or YYYR, to indicate (+) or (-) sequences, respectively, during sequencing. Asterisks represent phosphorothioate modifications included to prevent exonuclease activity. Amplification was performed for 15 cycles, with an annealing temperature of 65 °C, and an extension time of 15 seconds. The completed (+) and (-) reactions for each sample were mixed together with an additional 50 μL nuclease-free H_2O and ethanol precipitated. The resulting pellet was dissolved in formamide and analyzed with an ABI 3730xl capillary electrophoresis device.

6.7.12 Library preparation and next generation sequencing

To construct sequencing libraries, a separate PCR for each (+) and (-) sample was mixed by combining: 33.5 μL nuclease-free H_2O, 10 μL 5X Phusion Buffer (NEB), 0.5 μL 10 mM dNTPs, 0.25 μL of 100 μM TruSeq indexing primer (5'-CAAGCAGAAGACGGCATA CGAGATxxxxxxGTGACTGGAGTTCAGACGTGTGCTC-3', 0.25 μL of 100 μM primer PE_F, 2 μL of 0.1 μM selection primer (+ or -, as noted above), 3 μL ssDNA library (+ or -), and 0.5 μL Phusion DNA polymerase (NEB). Multiple TruSeq indexes were used, were the 'xxxxxx' is replaced with the unique TruSeq barcoding sequence. Amplification was performed as indicated in 'Quality analysis' above, using the appropriate selection primer(s). Completed reactions were chilled at 4 °C for 2 min before addition of 5 U Exonuclease I (NEB) and subsequent incubation at 37 °C for 30 min to digest unextended primer. The libraries were purified with 90 μL of Agencourt XP beads (Beckman Coulter) according to manufacturer's instructions. The complete libraries were eluted with 20 μL TE buffer and quantified with the Qubit 2.0 Fluorometer (Life Technologies). Individual libraries were balanced and sequenced on either a MiSeq or

HiSeq (Illumina) using 2x35 bp paired end reads.

6.7.13 Data analysis with Spats

Reads analysis was performed with Spats v1.0.0 (`https://github.com/LucksLab/spats/releases/`), using cutadapt v1.5 [267] and Bowtie 0.12.8 [182] to do the adapter removal pre-processing. Each paired-end read was uniquely mapped to a targets file containing the sense/antisense RNA sequences, or all the transcript lengths enumerated for cotranscriptional SHAPE-Seq and appending the 3′ RNA linker sequence (CUGACUCGGGCACCAAGGA) to each, in order to generate a reactivity value θ_i, representing the probability of modification at nucleotide i relative to the rest of the nucleotides in the RNA. The θ_i values were normalized to ρ_i values according to previous SHAPE-Seq work (Chapter 5) [113, 122, 123].

CHAPTER 7

DESIGN OF A CRISPR SGRNA DEREPRESSOR USING INSIGHT FROM SHAPE-SEQ

7.1 Abstract

Over the last 15 years synthetic biology has seen an explosion in the number of characterized parts available for designing genetic circuitry. RNA-only systems, a subset of these parts, have been studied recently with the expectation that the high turnover rate of cellular RNAs will allow for faster dynamics in synthetic systems. However, many RNA regulators exhibit lower dynamic range and increased leakiness relative to their protein counterparts. To address that problem, we sought to reengineer the CRISPR interference (CRISPRi) system to allow for the dynamic repression of target genes that responds to the expression of non-CRISPR small RNAs. To that end, we used in-cell SHAPE-Seq (selective 2'-hydroxyl acylation analyzed by primer extension sequencing) RNA structural information to guide the design of a small guide RNA (sgRNA) to refold into a form incapable of dCas9 binding in the presence of an antisense RNA. We demonstrate that our RNA-responsive sgRNA design can achieve between 8- and 13-fold activation of gene expression in response to the expression of the antisense RNA. Further, the design is host-independent and potentially highly scalable and transferable between organisms. With further development, we expect that our RNA-responsive sgRNA design will allow for the future construction of robust, complex logics with higher dynamic range than current systems.

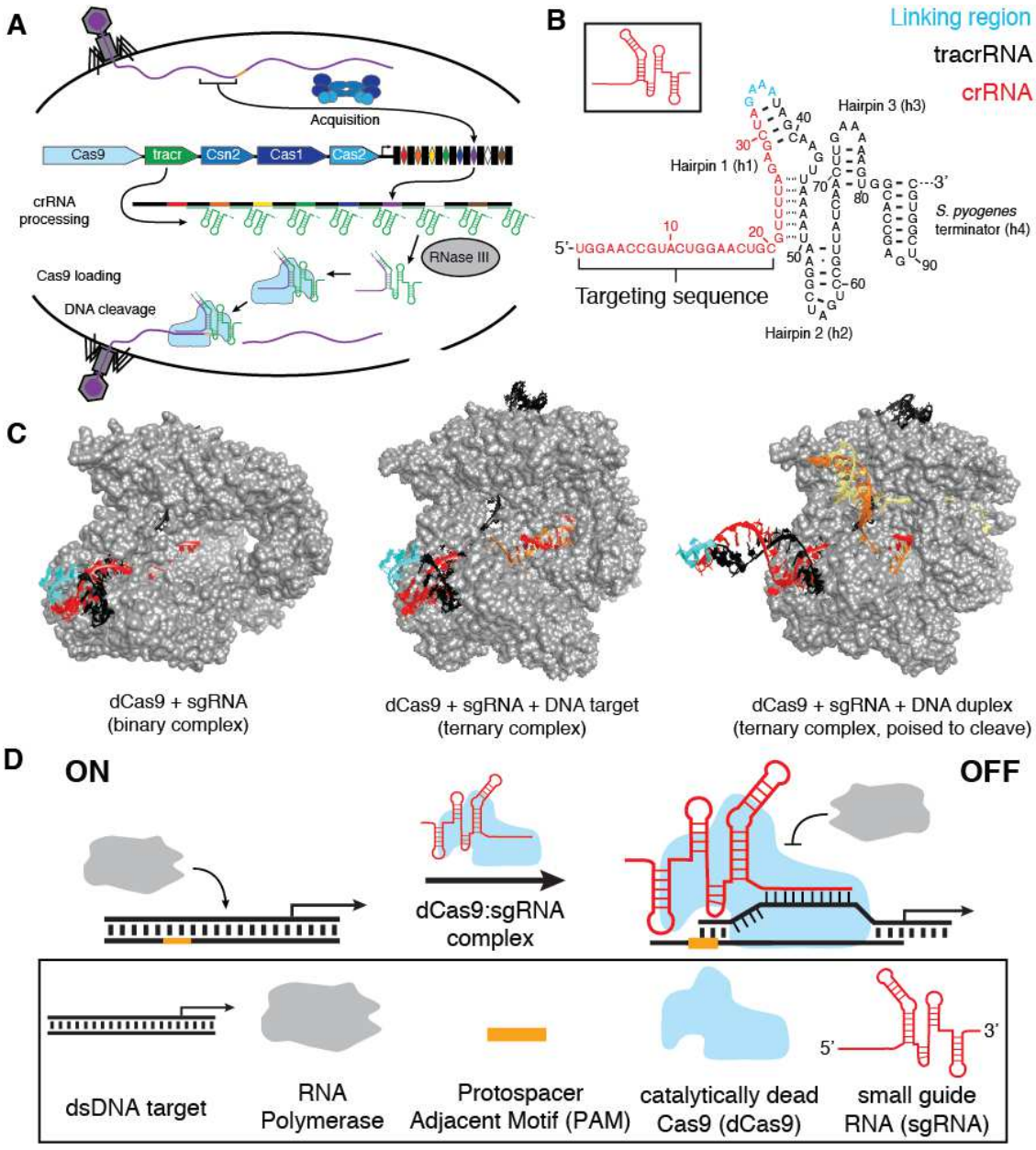
The work presented in this chapter is in preparation for submission and contains contributions from Jane B. Liao, Timothy R. Abbott, and Julius B. Lucks.

7.2 Introduction

The recent development of CRISPR technology has changed the face of biology [268]. Found in bacterial and archaeal genomes, CRISPR systems serve to defend prokaryotes from the constant threat of invasion from phages, parasitic plasmids, etc. by targeting the invading sequences for cleavage based on a genetic 'memory' of past invasion [4, 5, 269–272]. The ability of CRISPR systems to target highly specific stretches of sequence has led to an explosion in the number of applications for CRISPR technology, including gene expression control and the hotly debated field of genome editing, making CRISPR research an entirely new field of biology [26, 56, 58, 59, 140, 268, 273–301].

While there are currently five recognized Types of CRISPR systems [302, 303], the most utilized CRISPR system is Type II, with most of the technological applications stemming from developments involving the Type II-A system from *S. pyogenes* [58, 59, 139, 268]. Within the CRISPR locus of *S. pyogenes* there are four known proteins (Cas1, Cas2, Csn2, and Cas9), a trans-activating CRISPR RNA (tracrRNA), and an array of 20 nt spacer sequences that are acquired from invading DNA and separated by a repeating sequence (Figure 7.1A) [304]. In the natural context, these spacer sequences are acquired by a Cas1-Cas2-Csn2 complex and must contain a protospacer adjacent motif (PAM) sequence that is unique to each Type II CRISPR system immediately downstream [305]. In *S. pyogenes*, the consensus PAM sequence is NGG [139].

Figure 7.1: Overview of the Type II CRISPR system from *S. pyogenes*. (A) The CRISPR system in *S. pyogenes* is an adaptive immune system that fights off invading DNA. First, a complex containing the Cas1, Cas2, and Csn2 proteins recognizes an invading DNA sequence and incorporates it between two repeat sequences (black rectangles) as a 'spacer' in the CRISPR locus (colored diamonds). The CRISPR locus is transcribed as one long RNA. Multiple copies of the trans-activating CRISPR RNA (tracrRNA) base pair to the long RNA. RNase III cleaves the interacting RNAs into smaller CRISPR RNA (crRNA):tracrRNA pairs that are loaded into the Cas9 protein to target and cleave a specific sequence based on the spacer sequence in the crRNA. (B) The small guide RNA (sgRNA) is a direct fusion of the first 32 nt of a crRNA (red) with ~60 nts from the 3' end of the tracrRNA (black), using a GAAA tetraloop (blue). (C) Crystal structures of the catalytically dead Cas9 (dCas9) protein showing the structural changes that occur as the sgRNA is loaded (left), the DNA target is found (middle), and the DNA target is being cleaved (right) [306–309]. The pre-loaded structure (left) has regions of the sgRNA missing due to high flexibility. (D) Basic mechanism of CRISPR interference (CRISPRi). In CRISPRi, the dCas9 protein is loaded with a sgRNA targeting the non-template strand of a transcript to be silenced. In prokaryotes, sgRNA:dCas9 complex binds the target site and prevents either transcription or initiation by RNA polymerase (depicted). In eukaryotes, dCas9 is used to localize transcription activators or repressors rather than directly act on transcription itself.


A
Acquisition
Cas9
tracr
Csn2
Cas1
Cas2
crRNA processing
RNase III
Cas9 loading
DNA cleavage
B
Linking region
tracrRNA
crRNA
Hairpin 3 (h3)
Hairpin 1 (h1)
S. pyogenes terminator (h4)
5'-UGGAACCGUACUGGAACUGC
Targeting sequence
Hairpin 2 (h2)
C
dCas9 + sgRNA (binary complex)
dCas9 + sgRNA + DNA target (ternary complex)
dCas9 + sgRNA + DNA duplex (ternary complex, poised to cleave)
D
ON
OFF
dCas9:sgRNA complex
dsDNA target
RNA Polymerase
Protospacer Adjacent Motif (PAM)
catalytically dead Cas9 (dCas9)
small guide RNA (sgRNA)
5'
3'

DNA targeting and cleaving is performed by the Cas9 protein after it is loaded with two additional RNAs, the tracrRNA and a CRISPR RNA (crRNA). The crRNA includes within it the 20 nt spacer sequence that defines the target for Cas9 to find and cleave [139]. To prepare the RNA part of the complex, the tracrRNA is repeatedly synthesized while the CRISPR array, containing the spacer and repeat sequences, is transcribed into one long RNA (Figure 7.1A). Then, the tracrRNAs base pair to the transcribed CRISPR array RNA at each of the repeat units, generating dsRNA segments at each repeat that are processed by RNase III into individual tracrRNA:crRNA pairs [310]. Each of these pairs contains the same tracrRNA, but different crRNAs that have different 5′ 20 nt spacer sequences. Last, the tracrRNA:crRNA pairs are bound by Cas9 and complete the targeting complex, which searches the DNA within the cell for the spacer targeting sequence. If a DNA sequence found by the targeting complex matches the first 20 nts of the crRNA, and the PAM sequence of the adjacent DNA is correct, Cas9 will cleave both DNA strands [306, 308, 311, 312].

The ability of Cas9 to specifically target a wide range of DNA sequences as a single protein with a clearly defined RNA sequence has made gene editing much simpler and more precise than methods employing retroviruses, zinc-finger nucleases, or transcription activator-like effector nucleases [58, 268]. Further, Jinek *et al.* showed that the tracrRNA:crRNA pair can be directly fused and shortened down to an approximately 100 nt single guide RNA (sgRNA) that functions with the same efficacy as the wt tracrRNA:crRNA pair (Figure 7.1B) [139]. Thus, CRISPR-Cas9 allows for precise DNA binding and cleavage with a single RNA, single protein system (Figure 7.1C) and a massive list of potential targets, as the GG dinucleotide in the *S. pyogenes* PAM sequence should occur once for every 16 dinucleotide pairs on average. The high frequency of the PAM motif in genomes has lead to the immediate application of widespread, targeted gene modifications in a host of different organisms [268, 276, 281, 283, 284, 289].

Alongside the race to edit new and exciting genomes, a catalytically dead mutant of Cas9 (dCas9) was introduced, converting Cas9 from an endonuclease to a targeted DNA binding protein (Figure 7.1D) [57, 294]. It was shown that dCas9 binds tightly to DNA and can serve as a strong protein roadblock to transcription initiation or elongation. The dCas9 roadblock is highly efficient, allowing tight control over gene expression, and is referred to as CRISPR interference (CRISPRi) [26, 56, 57, 140]. Similarly, dCas9 has also been used to localize transcription activators and repressors to promoter regions to control gene expression [287, 292, 294, 298–301, 313, 314] as well as fluorescent proteins for microscope imaging [293].

The strong affinity of the dCas9 complex for its DNA target makes it an excellent repressor, but its slow off-rate means that all CRISPRi-based regulation suffers from very slow dynamics [60]. Dilution (via cell division) appears to be the predominating determinate of derepression after the expression of new dCas9 is stopped [57]. While the slow off-rate is beneficial for many applications, it prevents any sort of rapid execution of synthetic biological programs and is thought to contribute to off-target effects during gene editing. As leaky circuitry has been a major hindrance to the progress of synthetic biology, the field would greatly benefit from the extremely efficient repression that CRISPRi offers if its dynamics could be better controlled. Therefore, we sought to devise a method to trigger quick release of the bound dCas9 protein from its DNA target and remove the limitation of dCas9's slow off-rate.

Previous work by Briner *et al.* demonstrated that certain regions of the sgRNA are very sensitive to mutations that alter sgRNA structure, preventing efficient dCas9 repression or Cas9 cleavage [315]. They also identified regions of the sgRNA that were insensitive to mutation, retaining repression or cleavage levels near or matching the original sgRNA, suggesting that additional exogenous sequence could be added to the

sgRNA scaffold.

In this work, we use in-cell SHAPE-Seq (Chapter 3) [113, 123] to more deeply study the features of the sgRNA structure that are critical to dCas9 binding and function in *E. coli* and use those insights to guide the design of a switchable sgRNA that triggers the dissociation of the dCas9:sgRNA complex upon application of a stimulus. Using SHAPE-Seq on the original sgRNA design and a series of mutants, we identify structural features that are critical to maintain in order to achieve proper dCas9 binding as well as features that can be altered. We then added exogenous RNA regulator sequences in regions of high mutation tolerance to trigger refolding of nearby critical structures upon the addition of a stimulus. Ultimately, we arrived at a split sgRNA design that exhibits substantial levels of both repression and derepression. However, challenges remain to achieving orthogonality and fast dynamics with our sgRNA-based derepressor and are discussed thoroughly later in the text.

7.3 Results and Discussion

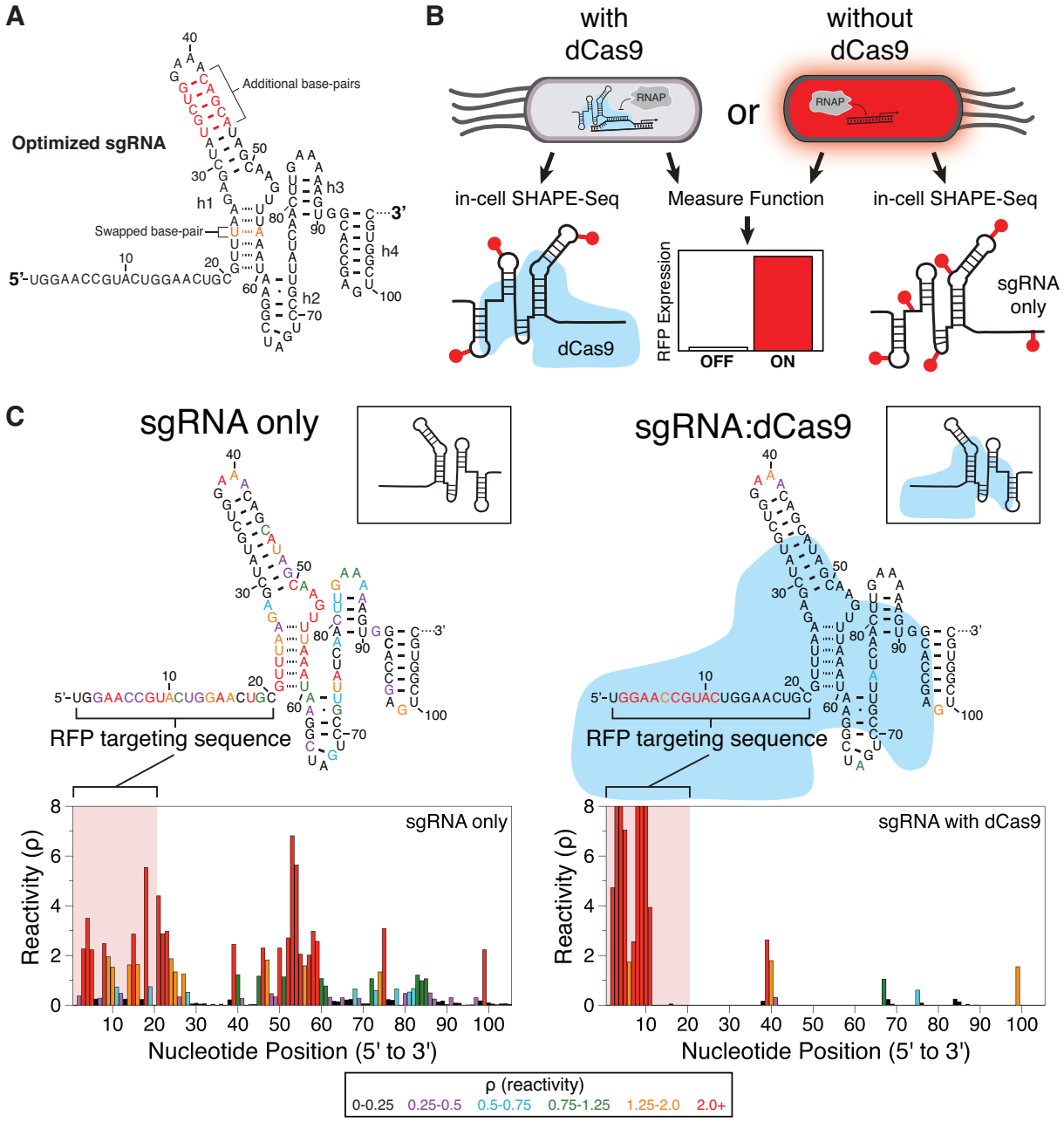
In the area of RNA synthetic biology a number of repressors have been designed that suppress gene expression, but many suffer from fairly high levels of transcription or translation leak when completely turned off [24, 26]. Protein regulators tend to repress gene expression at higher levels, but suffer the drawback of exhibiting slower dynamics than RNA [141, 262]. One way to marry the benefits of RNA and protein regulators would be to engineer a CRISPRi system that responds to stimuli in the timescale of RNA. In this work, we make progress toward engineering a sgRNA that changes structural folds in response to cellular stimuli to control dCas9 binding and targeting for transcriptional control via CRISPRi.

7.3.1 Examining sgRNAs with in-cell SHAPE-Seq

Throughout the work, we examine the structural characteristics of sgRNAs in *E. coli* using in-cell SHAPE-Seq [113]. In-cell SHAPE-Seq captures local nucleotide flexibility information in an RNA molecule by exposing it to a SHAPE reagent that preferentially reacts with positions that are more structurally flexible. The chemical modifications are detected using reverse transcription, which stops one nucleotide before the modification, and next-generation sequencing. The modification pattern is used to calculate a reactivity map that serves as a structural 'fingerprint' of the RNA and can be used to infer structural features [123].

Instead of focusing on the original sgRNA platform design by Jinek *et al.* (Figure 7.1B) [139], we instead chose to use an optimized version from Chen *et al.*, as it was shown to have better cross-species tolerance when over-expressed (Figure 7.2A) [293]. To study sgRNA:dCas9 complexes in the cell, we transformed a plasmid containing the sgRNA sequence along with a second plasmid with an open reading frame (ORF) that either contained dCas9 or no protein into *E. coli* cells with an RFP/GFP dual expression cassette integrated into the genome. Then, we simultaneously measured the level of RFP fluorescence and the structural features of the sgRNA, using the SHAPE chemical probe 1-methyl-7-nitroisatoic (1M7) anhydride (Figure 7.2B). By comparing how efficiently different sgRNAs repressed RFP and their reactivity maps, we were able to determine structural features of the sgRNA that are critical to its function.

Figure 7.2: In-cell SHAPE-Seq characterization of the RR2 opt sgRNA. (A) Optimized sgRNA design targeting RFP (RR2 opt), highlighted according the changes put forth in Chen *et al.* [293]. (B) Basic overview of the in-cell SHAPE-Seq experiment to characterize sgRNA structures. Each sgRNA sequence is expressed from a plasmid in *E. coli* along with a plasmid expressing dCas9 or an empty control. The resulting fluorescent output and in-cell structures are simultaneously measured with in-cell SHAPE-Seq. (C) Reactivity maps of the RR2 opt sgRNA with (left) and without (right) dCas9 binding. Without dCas9, highly reactive positions in the targeting sequence and the inner loop and lower stem of h1 indicate single strandedness. Upon dCas9 binding, reactivities decrease across the whole sgRNA, except for the first 11 nts of the targeting sequence and loops that stick out away from the protein. Nucleotides are colored according to reactivity intensity and RNA structures are drawn according to how they appear in in complex with dCas9.



7.3.2 Initial characterization of the sgRNA:dCas9 complex

To begin, we first converted an optimized sgRNA targeting RFP (RR2 opt) to be amenable to in-cell SHAPE reactivity measurements by removing the polyU sequence that follows the natural *S. pyogenes* terminator and appending the dual in-cell SHAPE-Seq terminators that are described in Watters *et al.* (Chapter 3) [113]. We also switched

from promoter J23119 (BioBricks), used in Qi *et al.* [57], in favor of the weaker J23150 to minimize the negative side effects of over-expressing the sgRNA on cell growth. We found that both of these changes had minimal impact on the overall level of repression by dCas9 (Figure E.1).

When we applied the in-cell SHAPE-Seq technique to the RR2 opt sgRNA design, we immediately saw a number of striking changes in the sgRNA reactivity map upon dCas9 binding in the cell (Figure 7.2C). When expressed without dCas9 present, the sgRNA exhibits high reactivity values in loops and single stranded regions, which would be expected based on the structure of the sgRNA bound within the sgRNA:dCas9 complex [306–309]. Namely, the apical loops of h1, h3, and h4 exhibit nucleotides with high reactivities as well as the single stranded targeting region and the nucleotides between h2 and h3. However, the expected helical region in the lower stem of h1 also appears very highly reactive, suggesting that that helix does not form when the sgRNA is present without dCas9.

When dCas9 is present, its effect on the sgRNA reactivity map upon binding is clear. Nucleotides throughout the sgRNA decrease in reactivity, except for those that are known to stick out away from the complex (Figures 7.1C and 7.2C) [306–309]. We also observe an increase in the first 11 nts of the sgRNA that correlates well to the observed crystal structure in which only nts 12-20 of the targeting region are bound by dCas9. We hypothesize that these two regions of the targeting region perform two different tasks. The region bound to the protein (nts 12-20) is more directly involved in the initial recognition of a correct target sequence and assists dCas9 in separating the target DNA strands. Then, the flexible part of the sgRNA targeting region (nts 1-11) can rapidly extend the nucleated RNA:DNA interaction.

7.3.3 A deeper understanding of the sgRNA:dCas9 complex through mutational analysis coupled with in-cell SHAPE-Seq

During the initial characterization of the sgRNA:dCas9 complex, Briner *et al.* published a mutational study of the *S. pyogenes* sgRNA [315]. Their results indicated that base pairing in the lower h1 stem and h2 were important for functional activity of the sgRNA:Cas9 complex, as well as the maintenance of the inner loop in h1 [315]. However, they did not examine the structural changes brought about by the various mutations and their connection to functional consequences.

Therefore, we examined a subset of their mutant variants ('v' designation) by mutating the RR2 opt sgRNA and used in-cell SHAPE-Seq to better understand what structural features of the sgRNA are important for function (Figure 7.3A). We quickly found that our functional results matched closely to the results observed in Briner *et al.*, immediately removing the possibility that the context of our assay would impact our observations (Figure 7.3B). We then went about collecting in-cell SHAPE-Seq data for a number of the mutants in variant series of mutations (Figures 7.3A and E.2). We found that sgRNAs that failed to generate any appreciable level of repression did not undergo the characteristic reactivity changes that we observed with the original RR2 opt design. However, we did observe partial reactivity decreases in the lower h1 stem that suggested that dCas9 may be partially binding the sgRNA, but not progressing to the completely loaded sgRNA:dCas9 stage (Figure 7.1C). These results, along with the observation that dCas9 is very sensitive to mutations and insertions/deletions in h2 [315], led to us hypothesize that dCas9 first recognizes the h2 structure then closes the lower h1 stem by sequentially binding each side of the forming helix.

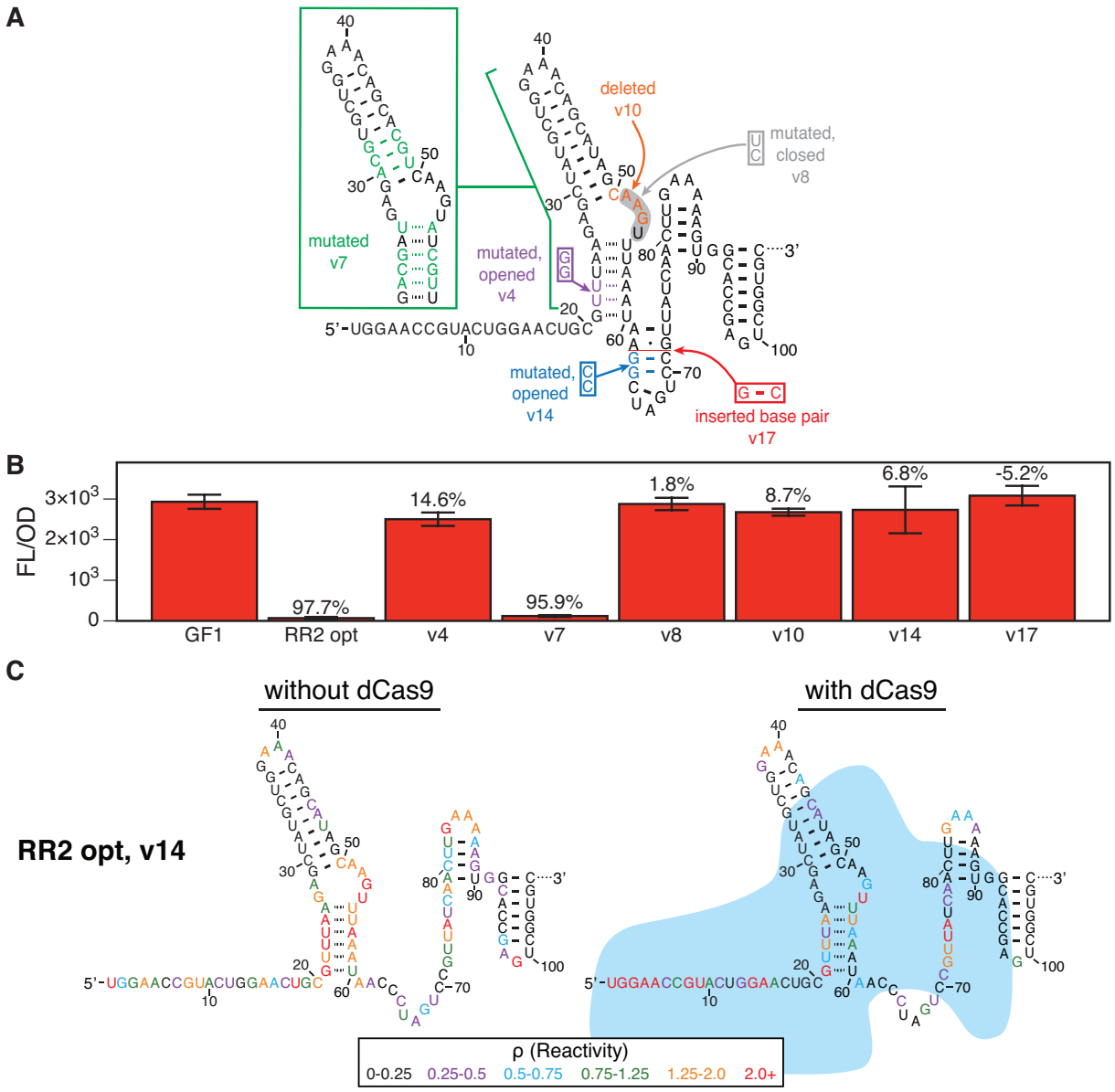
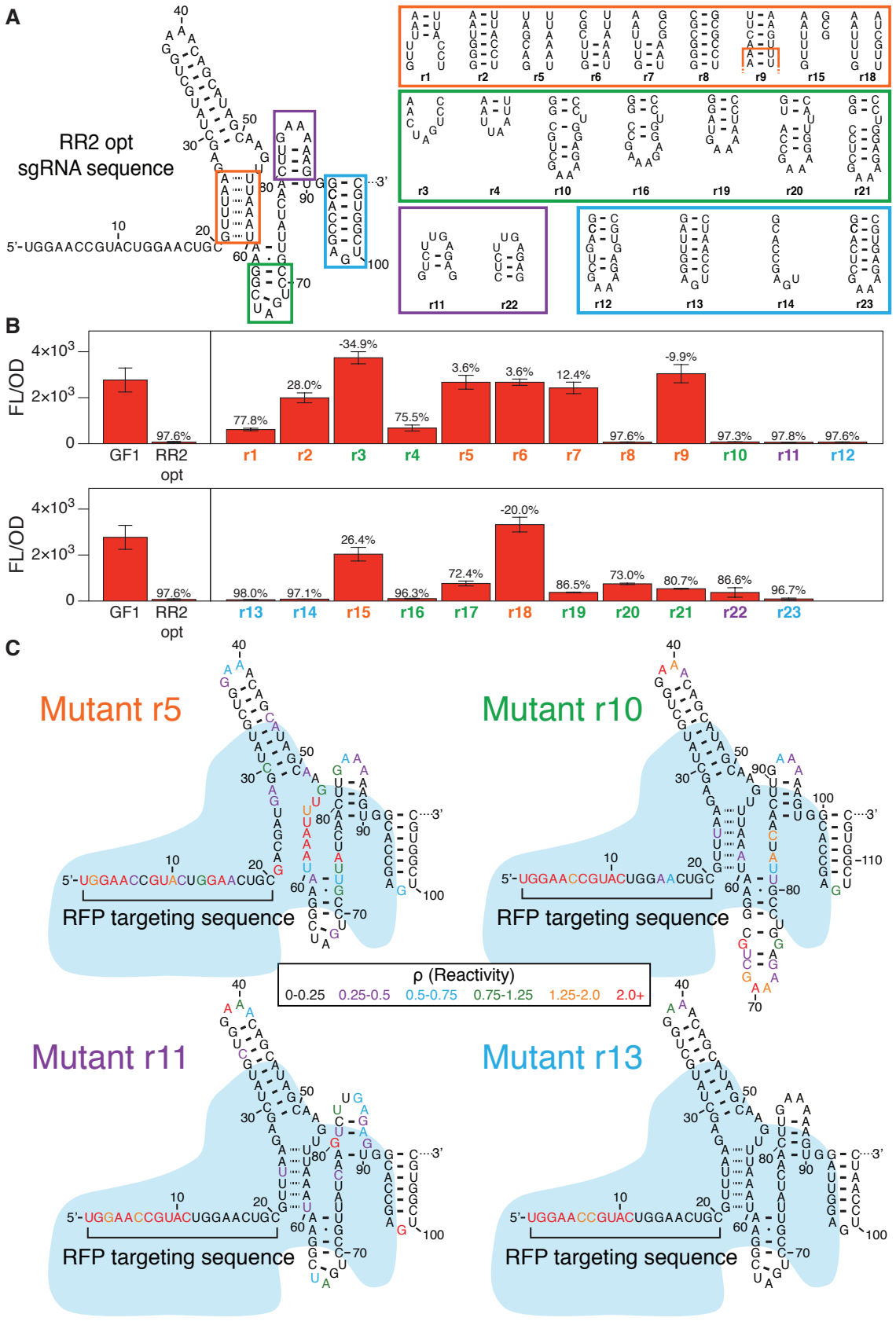


Figure 7.3: Analysis of a selection of the mutants studied by Briner *et al.* (A) Diagram depicting the individual mutations that were made to the RR2 opt sgRNA to generate each mutated version ('v' designation) from Briner *et al.* [315]. (B) RFP fluorescence normalized by culture optical density (FL/OD) for each mutant. Only mutant v7 demonstrates any appreciable repression. Error bars represent one standard deviation of four replicates. (C) In-cell SHAPE-Seq analysis of the sgRNA v14. The disrupted h2 hairpin results in incomplete binding of dCas9, leading to only partial reactivity decreases in the sgRNA. Individual nucleotides are colored according to reactivity intensity.

To delve further into our hypothesis of dCas9 binding, we created a series of revised mutants ('r' designation) that included changes that altered the pairing of the h1 lower stem and h2. We also examined mutations in h3 and h4 (and a few in h1 and h2) to locate regions that might be insensitive to local structural changes that could later be replaced with RNA regulator sequences to generate responsive sgRNAs (Figures 7.4A and E.3). We largely ignored the inner loop and upper stem of h1, as they were already shown to be critical and dispensable, respectively.

We quickly observed that hairpins h3 and h4 were fairly insensitive to mutation, as all of the mutants in these regions did not strongly affect repression and the reactivity changes associated with complete dCas9 binding were observed (Figure 7.4). However, as previously observed [315], the lower h1 stem and h2 were highly sensitive to mutations.

Figure 7.4: Analysis of additional sgRNA revised mutants. (A) Set of revised RR2 opt mutations ('r' designation). Colors indicate which hairpin the mutations were made in. (B) Relative RFP fluorescence normalized by cell density (FL/OD) measured for each mutant. Mutations in h3 (purple) and h4 (blue) are well tolerated, although mutations in h1 (orange) and h2 (green) depend on the structural context of the mutation. Error bars represent one standard deviation of four replicates. (C) Measured in-cell SHAPE-Seq reactivities for sgRNAs r5, r10, r11, and r13 when co-expressed with dCas9. Mutations that disrupt the lower h1 stem (r5) only allow dCas9 to bind to one strand. Reactivity data from r10 supports the hypothesis that inner loops or bulges are required near the bulged U in h2 to exhibit repression. sgRNAs containing mutations that are well tolerated (r11 and r13) exhibit the same reactivity pattern as the original RR2 opt with dCas9 present. Nucleotides are color-coded by reactivity intensity.


A
RR2 opt
sgRNA sequence
5'-UGGAACCGUACUGGAACUGC
r1
r2
r5
r6
r7
r8
r9
r15
r18
r3
r4
r10
r16
r19
r20
r21
r11
r22
r12
r13
r14
r23
B
FL/OD
GF1
RR2 opt
77.8%
28.0%
-34.9%
75.5%
3.6%
3.6%
12.4%
97.6%
-9.9%
97.3%
97.8%
97.6%
98.0%
97.1%
26.4%
96.3%
72.4%
-20.0%
86.5%
73.0%
80.7%
86.6%
96.7%
C
Mutant r5
Mutant r10
Mutant r11
Mutant r13
RFP targeting sequence
ρ (Reactivity)
0-0.25
0.25-0.5
0.5-0.75
0.75-1.25
1.25-2.0
2.0+

We analyzed the lower h1 stem and observed that destabilization of the potential Watson-Crick pairing results in greatly diminished repression. It appears that, in general, compensatory mutations to mismatched base pairs in this region restores the ability of the sgRNA to repress RFP. However, in the case of sgRNA r2, we only observe partial restoration despite complete pairing. We suspect the cause of r2's poor repression is the potential for a helical shift where G21 pairs with C59 and the base pairs above readjust to create a longer stem with U24 and A57 bulged out. The additional mutations in sgRNA r8 (exhibiting strong repression) would prevent a helical shift and rule out the changed base identity as the cause of failure of sgRNA r2 to repress RFP.

Like the mutations to the lower stem of h1, mutations in h2 that changed the structural context of h2 resulted in poor repression (Figures 7.4 and E.3). For example, mutating the two G:C base pairs to no longer form (r3) or to weaken them with A:U pairs [315] results in greatly reduced repression. Alternatively, we sought to add additional sequence to h2 by replacing the apical UA (nts 66 and 67) with additional base pairs capped with a 4 nt loop. However, we observed that while all of these types of h2 helix extension mutants (r10, r16, r19, r21) exhibited some form of repression, only two (r10 and r16) could repress RFP to the same degree as the original RR2 opt sequence. These differences were somewhat unexpected, as these sequences should not directly interact with dCas9 according to crystallographic data [306–309]. We observed, however, that sgRNAs r10 and r16 contained shorter continuous base pairing next to the bulged U69. We then determined that the likely cause of the lack of complete repression was the potential for a second helical shift in the sgRNA where G64 pairs with the bulged U, leading to G63 to pair with C70, ultimately destabilizing the non-canonical A62:G72 base pair that is critical for dCas9 binding [306]. Thus, the added extended helical sequences in h2 likely promote this helical shift by introducing a situation where the helical shift will create a longer, more stable h2 helix, one that is not amenable to

dCas9 binding. Thus, the G:G inner loop in sgRNA r10 likely serves to disrupt an elongated helix from forming, directly observable as a high reactivity as position G66 (Figure 7.4C).

Examining the in-cell SHAPE-Seq results also provides clues as to how sgRNA:dCas9 assembly occurs. The h1 lower stem mutants reveal that designs containing mismatches exhibit high reactivities on only one side of the h1 lower stem in the presence of dCas9. In sgRNA r5 for example, the 3′ side is highly reactive while the 5′ is not, suggesting that dCas9 binds one side of the helix first and brings the other side in close proximity in order completely form the h1 lower stem and subsequently bind it. However, other sgRNA mutants containing h1 mismatches exhibit high reactivities on the 5′ side and low reactivities on the 3′ side, such as sgRNA r1. Therefore, it remains unclear which side of the lower h1 stem is bound first.

Our in-cell SHAPE-Seq data also suggests that the h2 loop is one of the earlier RNA elements recognized by dCas9. In mutants where the h2 structure is greatly disrupted (e.g. sgRNAs r3, r4, and r19) we do not observe major decreases in the lower h1 stem (maintained in the original RR2 opt context), which serves as a good indicator of dCas9 binding (Figure E.3). Thus, it appears that h2 is bound by dCas9 before h1. While our mutants cannot rule out that h3 and h4 bind before h2, recent results studying a split Cas9 protein suggest that removing h1 and h2 only increases the equilibrium dissociation constant of the sgRNA:Cas9 complex by two-fold, while removing h3 and h4 increases it by nearly 10-fold [316]. Our results in combination with the works of Briner *et al.* and Wright *et al.* would preliminarily suggest a model by which dCas9 first weakly binds h3/h4, followed by h2 recognition, and ending with h1 binding and lower helix closure. In this model, Cas9 binds the sgRNA in the 3′→5′ direction, beginning with the least buried interactions and ending in the most buried.

7.3.4 Initial designs for creating responsive sgRNAs

Having initially characterized the structural elements of the RR2 opt sgRNA, we next turned toward applying our structural insights to add RNA regulator sequences into the sgRNA scaffold. Our goal was to include RNA structures in the sgRNA that would alter the overall sgRNA fold in response to a small molecule ligand, a protein, or another RNA to switch between two different states: one that would permit dCas9 binding, and therefore gene repression, and another that would prevent dCas9 binding and allow gene expression.

We began with three well-characterized RNA regulatory structures: the MS2 coat protein aptamer [317], the theophylline aptamer [40, 318, 319], and the pT181 recognition hairpin [133]. Each of the RNA structures was incorporated into h1, h2, or h3 in a variety of different sequence contexts and each design ('d' designation) was tested for its ability to repress RFP (Figures 7.5 and E.4). As expected, we found that most additions to h1 or h3 to be fairly straightforward, yielding high levels of repression. However, additions to h2 were more difficult due to the helical shifting issue described above, although two designs using the theophylline aptamer were shown to exhibit a high level of repression. However, none of the tested designs demonstrated significant levels of derepression, although only the theophylline designs were tested thoroughly. The MS2 protein designs were abandoned early since the main goal of the study was to operate at the speed of RNA dynamics and the pT181 designs showed no sign of derepression when co-transformed with antisense RNAs on agar plates (data not shown). In-cell SHAPE-Seq analysis of sgRNAs d1 and d2 (containing the pT181 recognition hairpin) revealed no reactivity changes associated with antisense RNA binding (Figures 7.5 and E.4).

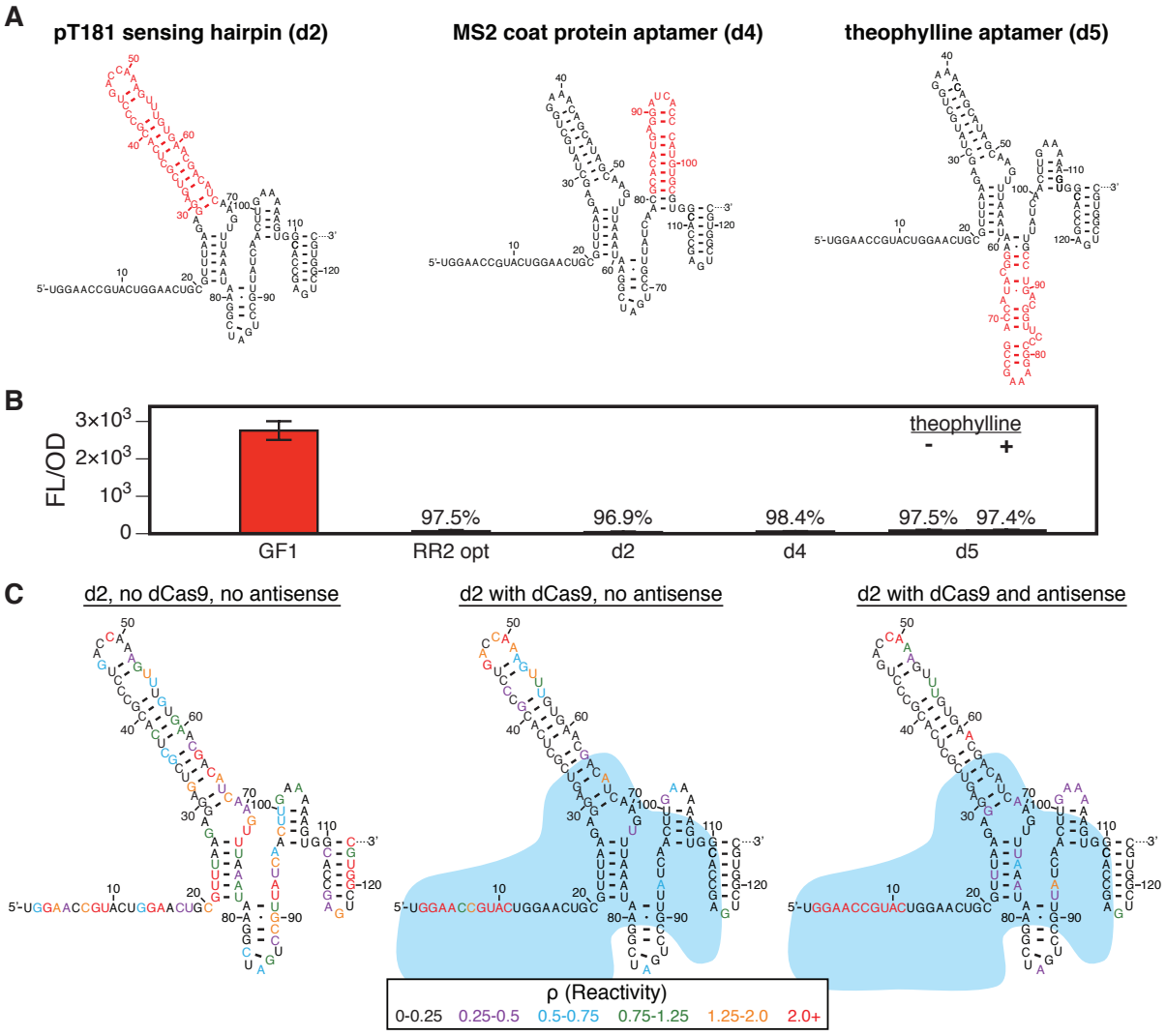


Figure 7.5: Initial sgRNA designs containing RNA regulator motifs. (A) A selection of three representative sgRNA designs ('d' designation) that included the pT181 recognition hairpin (d2), MS2 coat protein aptamer (d4) or the theophylline aptamer (d5). RNA regulator sequences were inserted in various contexts within h1, h2, or h3. A complete list can be found in Figure E.4. (B) Relative RFP fluorescence normalized by cell density (FL/OD) measured for each design. All RNA regulator sequences can be inserted into the sgRNA platform and retain high levels of RFP expression, although no design exhibited derepression when the activator RNA, protein, or ligand was added. Error bars represent one standard deviation of four replicates. (C) In-cell SHAPE-Seq analysis of sgRNA d2. The reactivity changes match well to the unaltered RR2 opt, suggesting that the pT181 recognition hairpin or its antisense have no effect on dCas9 binding. Nucleotides are color-coded according to reactivity intensity.

7.3.5 Toehold-based sgRNA designs exhibit derepression

With the first three RNA regulatory structures proving to be largely unresponsive within sgRNAs, we turned to a toehold design that has been shown to exhibit high dynamic ranges for RNA translational systems in the synthetic biology community [26, 37, 177, 320–322]. Toehold designs work via strand exchange to selectively expose one half of a helix by displacement. A toehold design consists of helical RNA structure in which the strand that will not be displaced has an extra, unpaired extension, or toehold, for an invading RNA to bind to. The invading 'antisense' RNA binds to the toehold, creating a double-stranded region that subsequently grows into a longer extended helix by displacing the other (non-toehold) strand, generating a more thermodynamically stable, longer helix [26].

In the context of the sgRNA, we applied the toehold design strategy ('t' designation) by splitting the sgRNA into two halves at h1 or h2 that base pair with each other and interact with dCas9 in order to repress RFP (Figures 7.6 and E.5). As before, we found that maintaining the bulged U in h2 critical for good repression of the split design, requiring a few rounds of optimization, while the additional sequence in h1 had essentially no restrictions. Ultimately, we arrived at the sgRNA t5 and t6 designs that exhibited roughly 13.3-fold and 8.0-fold activation of RFP expression upon overexpression of their respective antisense RNAs (Figure 7.6), roughly on par with current published bacterial transcriptional activators [55, 263] and 2-3 times better than another recently published sgRNA derepression system [323]. However, it has yet to be tested what timescale the derepression occurs on, although early indications would suggest that it dominated by cell division (i.e., sgRNA:dCas9 dilution rate).

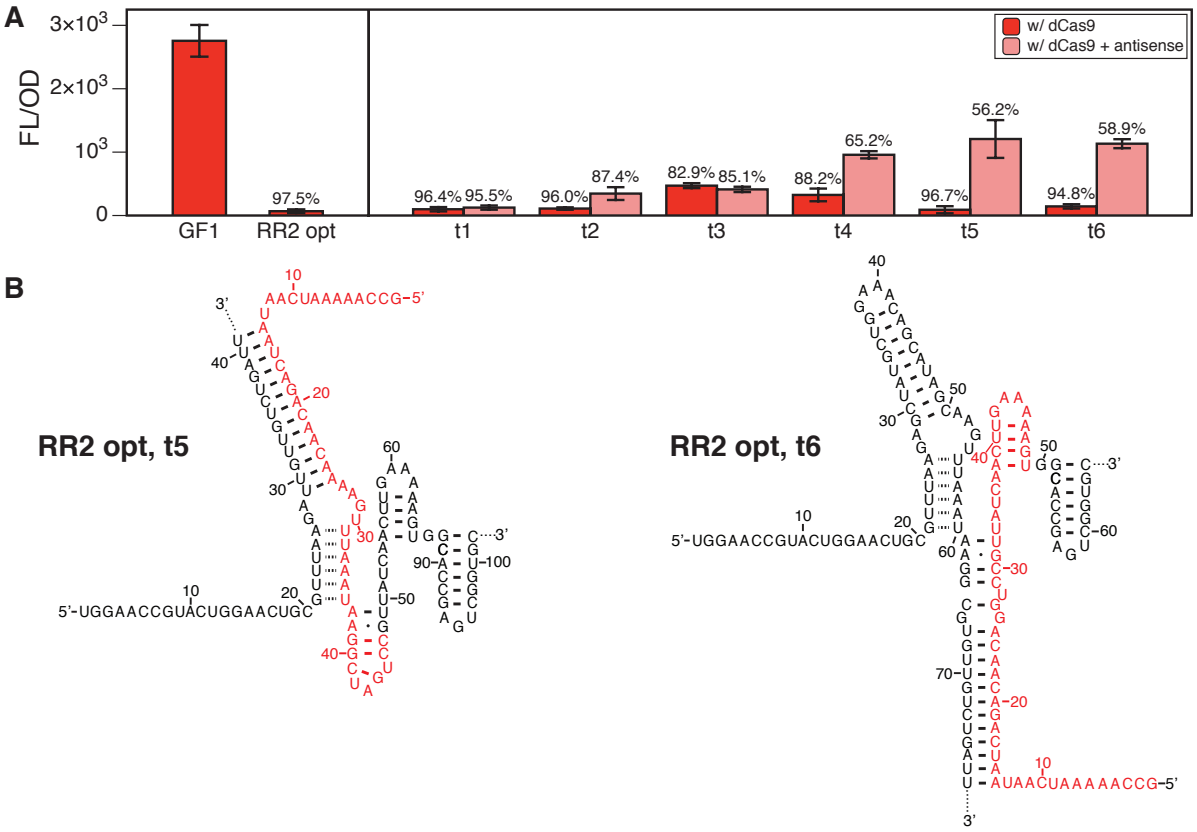


Figure 7.6: Toehold sgRNA designs derepress RFP expression. (A) Relative RFP fluorescence normalized by cell density (FL/OD) measured for each toehold design mutant ('t' designation). Many of the toehold designs show good to excellent repression, although early designs did not show much derepression in response to antisense expression. A complete list of the toehold designs can be found in Figure E.5. Error bars represent one standard deviation of four replicates. (B) The best toehold designs that split the sgRNA at h1 (left; t5) or h2 (right; t6), exhibited 8.0- and 13.3-fold activation in response to antisense expression, respectively.

7.3.6 Overall design considerations

Over the course of the sgRNA design process we observed a number of interesting structural consequences of mutating the sgRNA sequence/structure, including the potential helical shifts in h1 and h2 as described above. Generally, it appears that most

mutations in the lower h1 stem are well tolerated as long as base pairing in that region is preserved. The only two exceptions appear to be mutations that cause a potential helical shift, and mutations that disrupt the bottom G:U base pair as it is known to play a base-specific role in dCas9 binding [307, 309, 315]. In hairpin h2, maintenance of the AAGG:CCGU helix and neighboring bulged U is critical for dCas9 binding. Additional sequence can be placed above the bulged U, but there must be unpaired bases directly above the bulged U, otherwise the h2 helical shift occurs, greatly lowering the ability of dCas9 to bind the sgRNA.

We also observed that long dsRNA helices tend to be cleaved in the cell according to sequence alignments in the in-cell SHAPE-Seq data (sgRNAs v8 and r9). We have observed this phenomenon before [113], and it appears that helices that are 18 bps are cleaved near the middle of the helix, probably by RNase III. Further, we suspect that such RNase III cleavage is responsible for the overall failure of sgRNAs t1 and t2 to derepress when their respective antisenses are expressed (Figure E.5) by cleaving off their toeholds.

7.4 Future work

The current state of this work is unfinished. While we have made excellent progress toward responsive-sgRNA designs, there remains more work to be done to finalize the designs and gather more in-cell SHAPE-Seq data to support the conclusions. Below, we discuss steps to take to finalize the designs as well as the future outlook for responsive sgRNAs.

7.4.1 Creating orthogonal variants

One of the current drawbacks of the sgRNA t5 and t6 designs is a lack of both antisense and targeting orthogonality. During development of the t series of designs, we tested two potential variants of the t4 design, variants KW1 and KW2, by changing the sequence of the toehold (Figure E.6). However, regardless of the t4 variant sgRNA tested, each responded roughly the same to each of the antisense variants. We suspect that the lack of orthogonality is due the interaction of the antisense RNA with the toehold-containing half of the sgRNA before the two halves of the sgRNA can interact with each other. A possible solution to this problem is to alter the helical region above h2 as well as the toehold sequence to provide a greater difference between orthologs.

The second orthogonality issue stems from difficulties in easily transferring the t4 design to a different target sequence. We tested the sgRNA t4 design using the GF1 GFP-targeting sequence and found that it was a poor repressor (Figure E.7). Likely, the GFP targeting sequence is base pairing with the 3' tail of the added helical region, preventing the necessary interstrand sgRNA duplex from forming properly. In fact, there is indeed a potential for over 15 base pairs to form. Thus, testing a set of differently pairing helices as described above for antisense orthogonality will likely solve these issues for target orthogonality as well. The best case scenario is to develop a set of interaction helices and unique toeholds that can be selected from and computationally screened with software such has NUPACK [324] to ensure that the target sequence and the added orthogonal sequences are cross-compatible.

7.4.2 Building complex logics

One of the long term goals of this work is to able to create more complex logics and circuitry using the responsive sgRNAs. Because CRISPRi is such an efficient repressor and can be easily targeted to a wide selection of PAM sites, the potential for creating biological circuitry with responsive sgRNAs is huge. However, in order to be able to create higher-order functions operating at the RNA timescale, we need a diverse set of RNA regulators. So far, RNA-only NOR [135, 136] and AND [55] logic gates have been constructed in small orthogonal libraries.

By combining the sgRNA t5 and t6 designs, where three RNAs would interact together to form the complete sgRNA, we would be able to create an additional OR gate. In the three sgRNA fragment configuration, the expression of an antisense targeting either h1 or h2 would dissociate the sgRNA and prevent it from binding dCas9. Thus, gene expression is permitted if the antisense to h1 or h2 is present. Even without combining the sgRNA t5 and t6 designs, individual sgRNA variants could be layered to repress each others antisense RNAs or constituent sgRNA fragments, creating complex circuitry, including the potential for an RNA-only oscillator that has yet to be achieved. However, tuning the expression levels of the different sgRNA fragments will likely be an important parameter when designing high-order logics, as the responsive sgRNAs have yet to be tested targeting a multi-copy target such as a plasmid.

7.4.3 Understanding the mechanism of derepression

While we have observed derepression with our t series sgRNA designs, we have yet to elucidate the exact mechanism by which they function. Our original goal was to produce a sgRNA design that would remove the sgRNA from within a complex with

dCas9, however that may not be occuring with the current t series designs. Our observation of poor orthogonality between the t4 orthologs (Figure E.6) suggests that antisense-sgRNA fragment pairing occurs before the complete sgRNA forms. Thus, the antisense appears to sequester the sgRNA fragment and prevent it from forming the complete sgRNA, rather than dissociating the completed sgRNA either inside or outside a complex with dCas9. However, the ability of the antisense to prevent sgRNA association does not preclude it from dissociating the complete sgRNA. Because the antisense is expressed in great excess over the sgRNA fragments, it is in fact more likely that an antisense RNA would interact with the sgRNA fragment before its matching half.

To discover the mechanism of complex formation, we will employ three strategies. First, SHAPE-Seq analysis of the complexes *in vitro* using all combinations of sgRNA fragments, antisense, and dCas9 would reveal which RNAs are interacting and if dCas9 is bound. By forming the complete sgRNA:dCas9, then adding the antisense RNA before probing with SHAPE reagents, we would be able to determine if the dCas9 was dissociated by the appearance of the characteristic high reactivities in the targeting sequence and h1. Second, gel shift assays performed in parallel with the SHAPE-Seq experiments would allow for the direct visualization of complex formation/dissociation and support the SHAPE-Seq results. Last, establishing a time course of transcription repression/derepression would reveal the timescale that the dCas9 binding changes occur on, providing the most direct evidence of achieving the goal of fast dynamics we set at the beginning of this work. Testing the timescale could be done by inducing expression of the antisense RNA and measuring the change in fluorescence or mRNA concentration (via qPCR) over time. If the change occurs in the timescale of RFP maturation or mRNA degradation, we could expect that the sgRNA:dCas9 is being dissociated, but if the change occurs on the timescale of dCas9

dilution, the sgRNA t5/t6 designs likely only function to prevent complete sgRNA formation.

7.4.4 Expanding to higher organisms

The toehold designs operate using RNA-RNA interactions and dCas9 repression, both of which function in all types of organisms, unlike the current CRISPR derepressor design that relies on *E. coli* cellular machinery [323]. Therefore, we could expect that the sgRNA t series designs should be transferrable to higher organisms to interface with already published eukaryotic CRISPR gene regulation methods [58, 59]. Transferring the design would require recloning the sgRNA into eurkaryotic expression vectors, but would be aided by the fact that the current design is already based on the eukaryotic optimized version.

7.5 Acknowledgments

We thank Stanley Qi at Stanford University for the gift of the dCas9 and sgRNA plasmids as well as helpful insight. We also deeply thank Timothy Abbott and Jane Liao for their tireless work cloning and testing all of the sgRNA designs as well as Paul Carlson for early work testing the polyU knockouts and aTc concentration for dCas9 expression.

7.6 Materials and Methods

7.6.1 Plasmids

Plasmids for dCas9 and sgRNA expression were kindly provided by Stanley Qi (Stanford University) and have been previously described in Qi et al. [57]. In brief, the plasmid expressing dCas9 contained the *S. pyogenes* dCas9 coding sequence (CDS) downstream of a tet-inducible promoter within a vector containing the p15A origin of replicate and a chloramphenicol acetyltransferase gene for antibiotic resistance. The sgRNA plasmid contained the sgRNA sequences described in Qi *et al.* [57] downstream of a strong constitutive promoter (J23119), terminated with the TrrnB operon fragment [135], within a vector containing the ColE1 origin of replication (high copy) and an ampicillin resistance gene. As described in Section 7.3.2, the sgRNA was altered to replace the strong J23119 promoter with the weaker J23150 variant and the TrrnB operon fragment was replaced with the in-cell SHAPE-Seq amenable dual terminators described in Chapter 3 [113]. The dCas9 expression plasmid was unaltered throughout the study.

To generate mutants for the v, d, and r series of sgRNAs, primers were purchased for inverse PCR (iPCR) followed by ligation to introduce mutations, deletions, or new sequence into the sgRNA plasmid. To clone the plasmids for the t series, iPCR was used to create two different plasmids for the sgRNA halves; one that replaced the 5′ half of the sgRNA with the first 13 nts of RNA-IN variant S4 (Chapter 3) [176] or replaced the 3′ half of the sgRNA with the reverse complement of the first 13 nts of RNA-IN S4. Then, PCR was used to linearize one of plasmids and amplify the promoter-sgRNA fragment-terminator fragment, which there then stitched together using Gibson Assembly [181]. The t series antisense vector was created by stitching together a

kanamycin resistance gene, the BBR1 origin of replication, and a mobility gene using Gibson Assembly. To generate the t series antisenses, iPCR was used to first introduce the antisense sequence desired into sgRNA expression plasmid (containing the J23119 promoter and SHAPE-Seq terminators) before transfer into the antisense vector via Gibson Assembly.

7.6.2 Strains, growth media, and RNA expression

For all experiments except the t series the sgRNA expression plasmid and the dCas9 expression plasmid (or a control lacking its CDS) were transformed into chemically competent *E. coli* TG1 or MG1655 cells. The t series experiments also included the sgRNA antisense plasmid (or a control lacking an antisense sequence) in the transformation. Transformed cells were plated on LB+Agar media containing 100 μg/mL carbenicillin, 34 μg/mL chloramphenicol, and 100 μg/mL kanamycin (for t series experiments) and incubated overnight at 37 °C. The next day, individual colonies were picked and grown in 200 μL of LB (in-cell SHAPE-Seq experiments) or EZ rich MOPS media (Teknova; functional assays) with the appropriate antibiotics in a 2 mL 96-well block (Costar) and grown approximately 17 hr overnight at 37 °C at 1,000 rpm in a VorTemp 56 (Labnet) benchtop shaker.

At this point, three different subculture methods were used. For in-cell SHAPE-Seq experiments 24 μL of the LB overnight culture was added to 1.2 mL of M9 media (1:50 dilution) containing antibiotics and 1 μM aTc to induce dCas9 expression, and grown for 6 hours. Functional assays were carried out different depending on which series the sgRNA belonged to. The sgRNA 'd' and 't' series were by adding 4 μL of the EZ rich media overnight culture to 196 μL of EZ rich media (1:50 dilution) with antibiotics

and 1 μM aTc and grown for 3 hours. The sgRNA 'v' and 'r' series functional assays were instead performed by adding 1.2 μL of the EZ rich media overnight culture to 1.2 mL (1:1000 dilution) of EZ rich media with antibiotics and 1 μM aTc and grown for 6 hours. Upon complete the of subculture growth period, 100 μL of subculture was mixed with 100 μL PBS for fluorescence measurements.

7.6.3 Fluorescence assay

Fluorescence intensity was assayed using wavelengths 575 nm and 610 nm for excitation and emission, respectively. Fluorescence was normalized by OD_{600} after subtracting a media blank. Relative fluorescence levels of each culture were determined by normalizing the fluorescence readout by optical density (FL/OD) and subtracting the FL/OD of cells transformed with control sgRNA and dCas9 expression plasmids (that did not contain an sgRNA or dCas9 mRNA sequence) to correct for cell autofluorescence.

7.6.4 RNA modification and extraction for in-cell SHAPE-Seq

RNA was modified in cells with 1-methyl-7-nitroisatoic anhydride (1M7) by adding 500 μL of subculture to either 13.3 μL 250 mM 1M7 in DMSO (6.5 mM final) (+) or 13.3 μL DMSO (-) for 3 min before RNA extraction. Immediately following RNA modification, both modified (+) and control (-) samples were pelleted and resuspended in 100 μL of Max Bacterial Enhancement Reagent (Life Technologies) preincubated at 95 °C. All samples were then extracted with TRIzol Reagent (Life Technologies) according to the manufacturer's protocol and dissolved in 10 μL of water.

7.6.5 Reverse transcription

To each dissolved RNA sample, 3 μL of 0.5 μM reverse transcription primer (5' Biotin-TTTATCGGCCGAAGCAGGTAG) were added. Then the samples were denatured at 95°C for 2 min, then 65°C for 5 min. After denaturing, each RNA sample was then snap-cooled on ice for 1 min before addition of RT master mix [6 μL 5X First Strand Buffer (Life Technologies), 1 μL 10 mM dNTPs, 0.5 μL H_2O, and 0.5 μL Superscript III (Life Technologies)] and incubation at 52 °C for 25 min followed by 65 °C for 5 min. After RT the RNA was hydrolyzed with 1 μL 10 M NaOH then partially neutralized with 5 μL of 1 M hydrochloric acid and ethanol precipitated by addition of 78 μL of cold EtOH. The precipitated pellets were dissolved in 22.5 μL of nuclease-free H_2O.

7.6.6 DNA adapter ligation

To the solution containing the cDNA 3 μL 10X CircLigase Buffer (Epicentre), 1.5 μL 50 mM $MnCl_2$, 1.5 μL 1 mM ATP, 0.5 μL 100 μM DNA adapter (5'-Phos-AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3C Spacer-3'), and 1 μL CircLigase I (Epicentre) were added. The reaction was incubated at 60 °C for 2 hr, then 80 °C for 10 min. The ligated DNA was EtOH precipitated, dissolved in 20 μL of nuclease-free H_2O, purified using 36 μL of Agencourt XP beads (Beckman Coulter; according to manufacturer's instructions), and eluted with 20 μL TE buffer.

7.6.7 Quality analysis

For quality analysis, a separate PCR reaction for each (+) and (-) sample was mixed by combining: 13.75 μL nuclease-free H_2O, 5 μL 5X Phusion

Buffer (NEB), 0.5 μL 10 mM dNTPs, 1.5 μL of 1 μM labeling primer (Fluor-GTGACTGGAGTTCAGACGTGTGCTC), 1.5 μL of 1 μM primer PE_F (5'-AATGATA CGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'), 1 μL of 0.1 μM selection primer (5'-CTTTCCCTACACGACGCTCTTCCGATCT(RRRY /YYYR)TTTATCGGCCGAAGCAGGTAGA*G*G*C-3'), 1.5 μL ssDNA library (+ or -), and 0.25 μL Phusion DNA polymerase (NEB) Two different fluorescent primers were purchased from Applied Biosystems containing either the VIC or NED fluorophores (replacing 'Fluor' above), and added for either (+) or (-) samples, respectively. The selection primers were purchased from Integrated DNA Technologies and contain an internal barcode, RRRY or YYYR, to indicate (+) or (-) sequences, respectively, during sequencing. Asterisks represent phosphorothioate modifications included to prevent the 3'→5' exonuclease activity of Phusion polymerase. The quality analysis library reactions were amplified for 15 cycles according NEBs recommendations, using an annealing temperature of 65 °C and an extension time of 15 seconds. To the completed reactions, 50 μL nuclease-free H_2O was added and ethanol precipitated. The resulting pellet was dissolved in formamide and analyzed with an ABI 3730xl capillary electrophoresis device.

7.6.8 Library preparation and next generation sequencing

To construct sequencing libraries, a separate PCR for each (+) and (-) sample was mixed by combining: 33.5 μL nuclease-free H_2O, 10 μL 5X Phusion Buffer (NEB), 0.5 μL 10 mM dNTPs, 0.25 μL of 100 μM TruSeq indexing primer (5'-CAAGCAGAAGACGGCATAC GAGATxxxxxxGTGACTGGAGTTCAGACGTGTGCTC-3'), 0.25 μL of 100 μM primer PE_F, 2 μL of 0.1 μM selection primer (+ or -, as noted above), 3 μL ssDNA library (+ or -), and 0.5 μL Phusion DNA polymerase (NEB). Multiple TruSeq indexes were used,

were the 'xxxxxx' is replaced with the unique TruSeq barcoding sequence. Amplification was performed as indicated in 'Quality analysis' above. Completed reactions were chilled at 4 °C for 2 min before addition of 5 U exonuclease I (NEB) to remove unextended primer when subsequently incubated at 37 °C for 30 min. Following incubation, 90 μL of Agencourt XP beads (Beckman Coulter) were added for purification according to manufacturer's instructions. The complete libraries were eluted with 20 μL TE buffer and quantified with the Qubit 2.0 Fluorometer (Life Technologies).

To prepare the libraries for Illumina sequencing, the average length of each sample was determined using the results from the quality analysis in order to calculate the molarity of each (+) or (-) separately. Sequencing pools were mixed to be equimolar, such that all of the sequencing libraries were present in the solution at the same level. Sequencing was performed using the Illumina MiSeq v3 kit with 2x35 bp paired end reads.

7.6.9 Data analysis with Spats

Reads analysis was performed with Spats v1.0.0 (`https://github.com/LucksLab/spats/releases/`), using cutadapt v1.5 [267] and Bowtie 0.12.8 [182] to do the adapter removal pre-processing. Unique mapping of each paired-end read to the targets file generated a reactivity value θ_i, representing the probability of modification at nucleotide i relative to the rest of the nucleotides in the RNA. The θ_i values were normalized to ρ_i values according to previous SHAPE-Seq work [113, 122, 123] such that the average value of ρ_i across an entire RNA molecule is one. The normalization to ρ_i values also allowed for fair comparison between sgRNAs of different lengths.

CHAPTER 8

STRUCTURAL ANALYSIS OF CUCUMBER MOSAIC VIRUS RNA3

8.1 Abstract

Cucumber mosaic virus (CMV) has one of the widest known host ranges of a virus, infecting over 1,200 species of more than 100 plant families. Composed of a tripartite positive-sense RNA genome, CMV contains a number of characterized RNA structures in all three genome segments that are important to its viral life cycle. Here, we explore the structural features of the third genome segment, RNA3, using selective 2'-hyroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). RNA3 contains three non-coding regions that separate two opening reading frames encoding the virus's movement and capsid proteins. We use our obtained structural information to examine the global fold of RNA3 and the isolated structures of the non-coding regions. We comment on length-dependent structural differences in the 5' UTR that suggest a potential long-range interaction and observe structural changes that suggest replicase binding to conserved structures in the 3' UTR. Our SHAPE-Seq data provides a new dataset to help researchers uncover the structural features of CMV and related viruses that are important to its replication cycle.

8.2 Introduction

The RNA structure of viral RNAs (vRNA) is intimately related to essential viral functions including replication, translation, and encapsidation. Recent efforts to determine

The work presented in this chapter is being prepared as a manuscript in conjunction with Jeremy Thompson and Keith Perry in Plant Pathology at Cornell University.

vRNA structures have led to a broader understanding of the relationship between viral RNA structure and function and how it pertains to the virus infection cycle [325–329].

Plant viruses have played a central role in expanding our knowledge of vRNA structure. Structural predictions of the 3′ terminus of the Turnip yellow mosaic virus revealed the existence of a structure similar to that of tRNA that was later termed a pseudoknot [330–332]. Similar structures were also confirmed for Brome mosaic virus (BMV) [333–335] and Tobacco mosaic virus [336–338] leading to the dissection of their function by mutational analyses [339–343].

Cucumber mosaic virus (CMV) likely has the widest known host range of any known virus, infecting over 1,200 species in more 100 plant families [143]. It is a tripartite positive sense single stranded RNA (ssRNA) virus encoding five genes (Figure 8.1). The longest genome segment, RNA1 (~3,200 nts) encodes the 1a protein that contains methyltransferase and helicase domains. RNA2 (~3,000 nts) encodes two proteins: protein 2a that contains RNA-dependent RNA polymerase domain and protein 2b silencing suppressor, expressed from an overlapping reading frame. RNA3 is bicistronic and encodes the movement protein (MP) and coat protein (CP), respectively, separated by the intergenic region (IGR). CP is translated from subgenomic RNA4, whose promoter lies in the IGR [344].

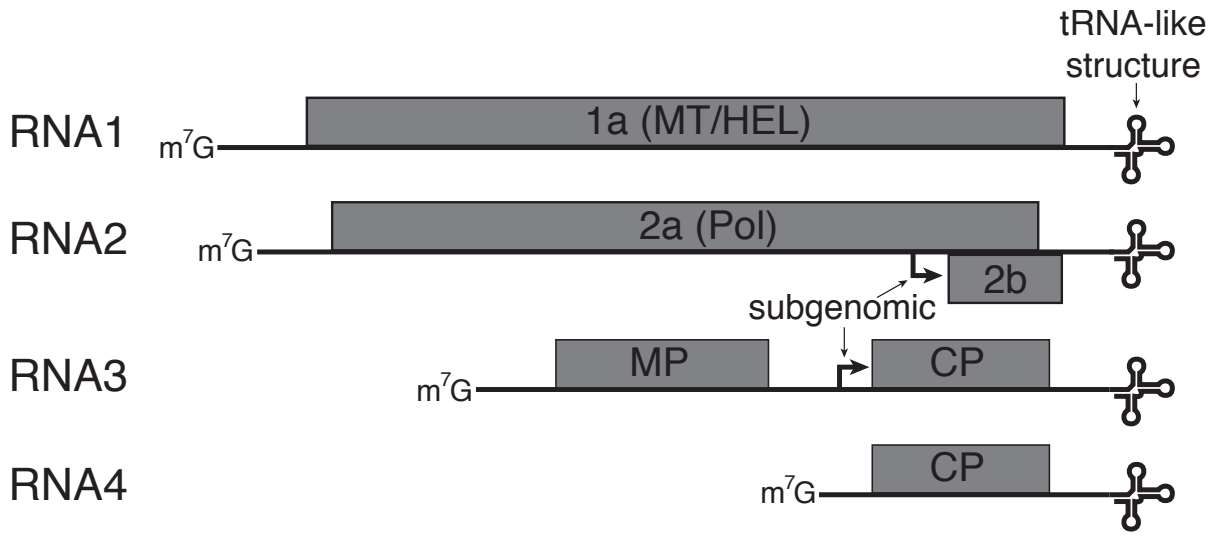


Figure 8.1: Genome layout of the Cucumber Mosaic Virus. Cucumber Mosaic Virus (CMV) is a tripartite (+)-sense ssRNA virus made up of RNAs 1-3. RNA1 encodes protein 1a that has methyltransferase (MT) and helicase (HEL) activities. RNA2 encodes protein 2a containing the RNA-dependent RNA polymerase domain and protein 2b, a silencing suppressor. RNA3 contains two ORFs, one for the movement protein (MP) and one for the coat protein (CP). CP is expressed via a subgenomic RNA (RNA4). All of the genome segments have a 5′ m^7G cap and a tRNA-like structure at the 3′ end.

Each RNA segment is capped at the 5′ end and has a highly conserved tRNA-like pseudoknot structure at the 3′ terminus that is used to initiate negative strand synthesis [144, 345, 346]. It was determined that the last 135 nts of the tRNA-like structure are sufficient to promote viral accumulation [347, 348]. Within the 135 nts, one stem-loop structure (SLC) was shown to interact with the viral replicase and is required for RNA synthesis [349, 350], with a CA dinucleotide in the trinucleotide loop of SLC being essential for replicase interaction [341]. The conserved structures upstream of the tRNA-like structure, including a complete map of 3′ untranslated region (UTR) of CMV RNA3, have been identified using a combination of enzymatic probing and co-variation analyses [335, 351, 352]. However, beyond being associated with high levels of recombination, the function of the rest of the 3′ UTR is not known [351–353].

Another well-studied structure is the IGR of RNA3. The RNA3 IGR contains a highly conserved box-B motif involved in Pol III transcription that, for BMV, has been shown in yeast to form a hairpin loop analogous to the TΨC-stem loop in tRNA [354]. Box-B motifs are also present in the 5′ UTRs of RNA1 and RNA2 in CMV and other related viruses and have been demonstrated to have a role in replication and interact with the 1a protein in BMV [355–357].

Much of the structural work performed to date, however, only focuses on small, isolated regions of the viral genome. Further, little to no study on RNA structure in open reading frames (ORFs) has been performed. However, the recent development of many high-throughput probing techniques [109] is allowing for the complete coverage of viral genome structures in a single experiment. To date, structural information exists for four ssRNA positive sense viral genomes: hepatitis C [328], human immunodeficiency virus [326], satellite tobacco mosaic virus [358], and tomato bushy stunt virus [327]. By structurally characterizing large regions of viral genomes at once, potentially interesting structures can be discovered much more rapidly and in the proper context.

In this study, we use selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq) v2.1 (Chapter 4) [123] to analyze the complete sequence of RNA3 from CMV strain Bn57 infected cell lysates [359]. We directly compare the structural features observed in lysates to those observed when refolding purified viral RNA *in vitro* and propose a secondary structure map of RNA3 consistent with our SHAPE-Seq data. We also observe structural features in the 3′ UTR consistent with the tRNA-like structure and replicase binding in cell lysates. This work is the first application of high-throughput chemical probing to plant viruses and the first demonstration of direct modification in plant cell lysates.

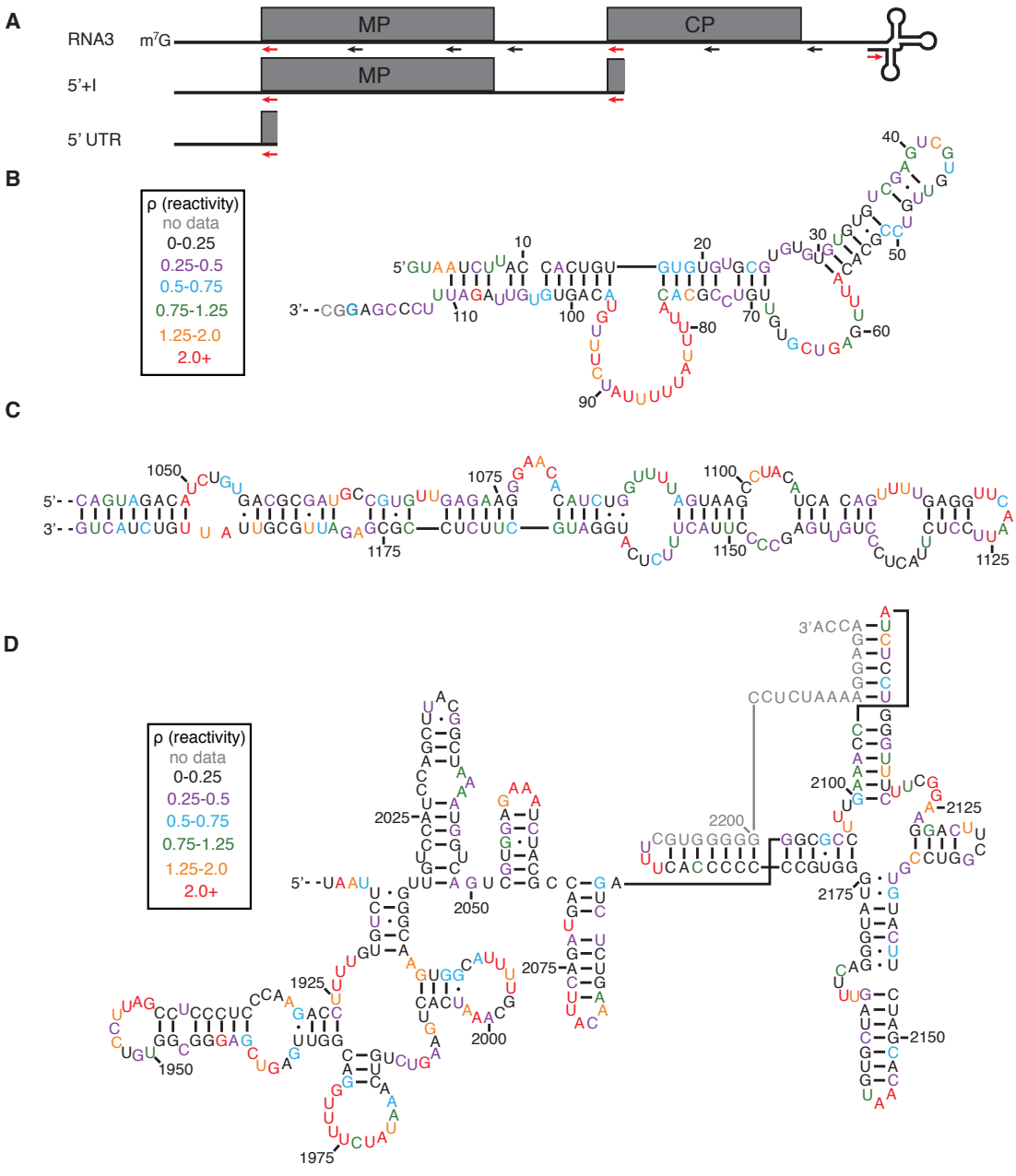
8.3 Results and Discussion

To obtain structural information for the entire RNA3 of CMV strain Bn57, we used the SHAPE-Seq v2.1 technique as described in Watters *et al.* [123] to analyze eight regions approximately 250-400 nts each (Figure 8.2A). SHAPE-Seq provides structural information about an RNA by covalently modifying positions that are structurally flexible. The modifications are detected using reverse transcriptase, which stops one nt before the modified position. Next-generation sequencing and bioinformatics then determine the frequency and position of modifications across the entire RNA to determine each nucleotides reactivity. The calculated reactivity map serves as a structural 'fingerprint' that can be used to restrain secondary structure predictions algorithms [123]. Below, we compare the structural elements of RNA3 in three different contexts that include *in vitro* transcribed RNA3 subgenomic segments, RNA from purified virions, and infected cell lysates.

8.3.1 5′ UTR structural features

To date, little to no work has been done to characterize the structure of the RNA3 5′ UTR, which lacks the box-B motif present in the 5′ UTRs of RNA1 and RNA2. However, the appearance of many elements of the RNA3 5′ UTR are consistent across many species, including a 'UG' repeat present in all subgroups and a direct repeat in subgroup I [360], suggesting that the 5′ UTR is important for the viral life cycle.

Figure 8.2: Structures of the isolated untranslated regions of CMV RNA3. (A) RNAs used in this study. Priming locations for reverse transcription are indicated with arrows. Red arrows indicate the priming site used to analyze each UTR. (B) Secondary structure of the 5′ UTR folded using ShapeKnots [102] with SHAPE restraints obtained from the average of refolded in vitro transcript probing experiments with the full RNA3 (m = 1.1, b = -0.3 [122]). (B) Secondary structure from the middle of the IGR folded with SHAPE restraints as described in (A). The displayed structure agrees well with the proposed structure in [354]. (C) Secondary structure of the 3′ UTR folded with SHAPE restraints as described in (A), including a forced pseudoknot interaction at the 3′ end [334] The structure matches well to previously observed 3′ UTR structures [352].


A
RNA3
m7G
MP
CP
5'+I
MP
5' UTR
B
ρ (reactivity)
no data
0-0.25
0.25-0.5
0.5-0.75
0.75-1.25
1.25-2.0
2.0+
C
D
ρ (reactivity)
no data
0-0.25
0.25-0.5
0.5-0.75
0.75-1.25
1.25-2.0
2.0+

We performed SHAPE-Seq analysis on the RNA3 5′ UTR using a priming site starting at the MP start codon (Figure 8.2A). We first began by analyzing *in vitro* purified transcripts of the just the 5′ UTR, the 5′ UTR through the end of the IGR (5+I), and the full length transcript to look for potential long-range interactions within the virus. In general, the reactivity maps for the 5′ UTR are similar between the *in vitro* contexts, however, the reactivity of the GUCGUGUUG nucleotide sequence (nts 40-48 and 62-70) decreases when the MP ORF and IGR are present (Figure F.1). Interestingly, the reactivity pattern of that sequence is similar between both repeats, suggesting it is in the same structural context in both. These results may hint at a long-range interaction between the 5′ UTR and either the MP ORF or IGR, although we did not observe any direct evidence of one.

We also examined the 5′ UTR SHAPE-Seq reactivities in the context of the full RNA3 sequence either purified from virions or probed directly in infected cell lysates, although little difference was observed between the 5+I reactivity map and those for the full RNA3. Next, we folded the 5′ UTR sequence with ShapeKnots [102] using the SHAPE reactivities as restraints, resulting it the structure show in Figure 8.2B. The same structure was observed when folding the entire RNA3 at once (Figure 8.3), indicating that if there is a long-range interaction present it may involve non-canonical base pairing or an unusual motif. Last, the lack of reactivity, or changes therein, of the 'UG' repeats was surprising given that such a strongly conserved sequence would be expected to participate in a defining structure or protein binding. However, there is no clear evidence for an obvious role of the 'UG' repeat sequence from the SHAPE reactivity data alone.

8.4 IGR structural features

We next examined the IGR with SHAPE-Seq. First, we compared the 5+I RNA to the full RNA3 sequence probed in lysates or *in vitro*, using RNA3 from purified virions or *in vitro* transcripts. We observed generally similar reactivity trends between the purified virions and infected cell lysates, as well as the 5+I RNA, except for ~40 nts near the 3′ end (Figure F.2). There are two possible causes: 1) the 3′ end of the IGR interacts with either the CP ORF or the 3′ UTR, or 2) there is high abundance of RNA4 in the purified virus and lysate, which has a different structural context in that region. The 5′ end of RNA4 (nt 1184 predicted) is at the beginning of the difference between the reactivity maps, leading to the possibility that RNA4, if highly transcribed, has an entirely different structural context than in the entire RNA3, as has been suggested by Kwon and Chug [361]. However, because of the sequence homology between RNA3 and RNA4, SHAPE-Seq would be unable to measure individual structural motifs, and would instead report a population average. To assess the possibility of structural interference by RNA4, we also examined refolded full length RNA3 in vitro transcripts. Interestingly, we observed that the 5′ and 3′ ends of the reactivity map of the refolded transcripts did not match to either the 5+I RNA or the other full length RNAs. Instead, the refolded transcripts had low reactivities in the 5′ end and high reactivities at the 3′ end. These differences may point to a possible interaction between RNA3 and RNA4 during folding.

Using the reactivities from experiments with refolded full length *in vitro* transcripts to restrain ShapeKnots with the isolated IGR sequence yields the structure depicted in Figure 8.2C. The structure we obtained contains an elongated stem-loop structure that was originally proposed by Baumstark *et al.* in the brome mosaic virus [354]. They observed that the apical loop resembles the TΨC loop from $tRNA^{Asp}$. However,

restraining the IGR fold with the reactivity data from purified virions or infected cell lysates results in an alternative branched structure (Figure F.3), but may be a result of RNA4 altering the SHAPE-Seq results as discussed above.

8.4.1 3′ UTR structural features and replicase binding

Of the three UTRs in RNA3, the 3′ UTR is the most well-studied [331, 333, 335–337, 347, 349, 351, 352, 362], consisting of a series of stem-loops, some of which are thought to contain pseudoknots, followed by the tRNA-like structure described above [331, 333, 335, 336, 351, 352]. Again, using ShapeKnots, we restrained the fold of the 3′ UTR the refolded *in vitro* transcript RNA reactivity data and obtained the structure shown in Figure 8.2D, which matches well to literature.

When comparing the lysate and purified virions *in vitro* refolding results, we observed a few regions where the reactivities differed (Figure F.3). In the first 100 nts, positions 1944, 1945, 1947, 1971, and 2003 of the isolated 3′ UTR appear higher *in vitro*, while positions 1920, 1921, 1930, 1998, and 1999 appear higher in the lysate. All of these bases appear in single-stranded regions according to our predicted structure (Figure 8.3D), but the lack of larger windows of reactivity differences in the first 100 nts of the 3′ UTR make it difficult to assign specific meaning to the observed differences. We also observed higher reactivities in the *in vitro* data at positions 2154, 2155, 2164, and 2165 of the isolated 3′ UTR. Of these nucleotides, A2154 and A2155 correspond to the three-nucleotide apical loop of the SLC structure, with A243 being part of the conserved CA dinucleotide motif. Positions U2164 and U2165 are in the nearby inner loop of the SLC. The low reactivities in the lysate correspond directly to the binding sites of the replicase [349, 362].

Our observation of replicase binding is one of the first examples of chemical probing being used to identify how proteins interact with viral RNA. By comparing chemical probing data from infected cell lysates and *in vitro* refolding experiments, we can obtain extra information about the role of RNA structure in viral replication.

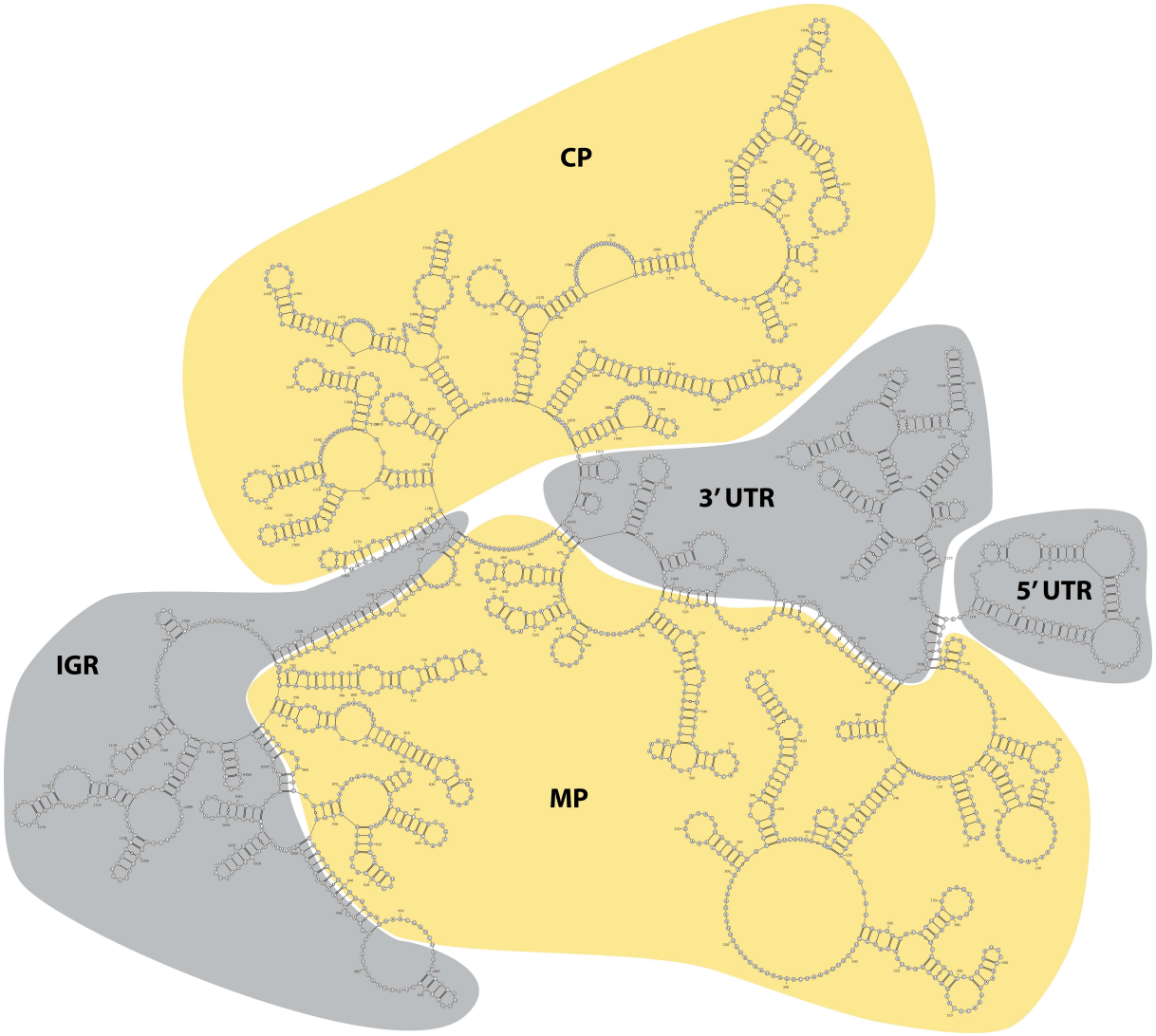


Figure 8.3: Predicted secondary structure of CMV RNA3 using purified virions. Secondary structure prediction of the CMV RNA3 using the reactivities obtained with the purified viral RNA to restrain the *Fold* algorithm of RNAstructure [101] (m = 1.1 and b = −0.3). Coding regions are indicated with yellow and non-coding regions with gray. As depicted the 3′ UTR tRNA-like structure is incorrect due to the inability of *Fold* to include pseudoknots in structural predictions.

8.4.2 Complete reactivity map of RNA3

Last, we used SHAPE-Seq to examine the entire RNA3 structure of refolded *in vitro* using purified viral RNA and *in vitro* transcripts as well as infected cell lysates using eight priming locations. The window of reactivities measured from each priming location was normalized such that the average reactivity value across it was one. The windows were then assembled together, and overlaps were resolved by selecting whichever priming location generated more sequencing reads (Figures F.5 and F.6).

We then used the complete RNA3 reactivity map from purified virions to restrain a secondary structure fold of RNA3 using the *Fold* algorithm of RNAstructure (Figure 8.3) [101]. The RNA3 structure appears highly branched with many short stem-loops. When folded the context of the entire RNA3 with the purified virion data we observed that the IGR and 3′ UTR did not fold independently, but instead interact with the coding regions. Both the purified virion and infected cell lysate data generated the same restrained fold (Figure 8.3). As before, we also did not observe the elongated stem-loop observed in Baumstark *et al.* [354]. We also observed an incorrect fold for the 3′ UTR lacking the pseudoknot, but it is mainly due to the inability of *Fold* to find pseudoknotted structures. Interestingly, folding the entire RNA3 with the reactivity data from the *in vitro* transcripts yields a different structure than the RNAs generated from lysate or virions (Figure F.7). The alternate structure from the *in vitro* transcript data does exhibit the expected elongated stem-loop structure in the IGR as well as the correct structure for the tRNA-like structure in the 3′ UTR (except for the pseudoknot interaction that is outside the capabilities of *Fold*). The differences between the two structures are likely due to the presence of the rest of the viral RNAs (RNA1, RNA2, RNA4), which may interact with each other in a previously unknown fashion.

8.5 Conclusion

The high-throughput nature of techniques like SHAPE-Seq is making genome-level structural information increasingly accessible for RNA viruses. In this work, we have provided structural information for CMV RNA3 that will assist other researchers to unlock the remaining secrets of the CMV replication life cycle. We have also outlined a general approach for looking for important structural features by comparing *in vitro* structural measurements to those made directly in cell lysates. We expect that similar structural studies like ours will greatly speed up the rate of discovery of new and exciting RNA viral replication mechanisms.

8.6 Acknowledgments

We thank Keith Perry and Jeremy Thompson (Cornell University) for all of their hard work and assistance in devising and maintaining a fruitful collaboration. We especially thank Jeremy Thompson for help preparing and critical reading of this chapter, as well as for providing all of the virally derived RNA used in this study. We also thank Alexander Settle and Timothy Abbott for early work on *in vitro* structure characterization.

8.7 Materials and Methods

8.7.1 RNA preparation

Run-off DNA templates were constructed for the isolated UTRs by amplifying the Bn57-CMV RNA3 UTR sequences in pBn57-3 [359] with *Taq* polymerase (New England Biolabs) and primers that bracket the UTR sequences. The forward primers contained a 17 nt T7 RNA polymerase promoter. Transcription reactions (1.0 mL, 37 °C, 1214 h) contained 40 mM Tris (pH 8.0), 20 mM $MgCl_2$, 10 mM DTT, 2 mM spermidine, 0.01% (vol/vol) Triton X-100, 5 mM each NTP, 50 μL of PCR-generated template, 0.04 U/μL SuperaseIN RNase Inhibitor (Ambion), and 0.1 mg/mL of T7 RNA polymerase. The RNA products were purified by denaturing polyacrylamide gel electrophoresis, excised from the gel via UV shadowing, and recovered by passive elution and ethanol precipitation. The purified RNA was dissolved in 50 μL TE buffer pH 8.0. Concentrations were measured with the Qubit Fluorometer (Life Technologies).

Cell lysates were prepared by grinding 20 mg of symptomatic *N. tabacum* leaf, inoculated two weeks prior with Bn57-CMV, in lysis buffer (150 mM KCl, 25 mM Tris pH 7.5, 5 mM EDTA, 5 mM $MgCl_2$, 0.5% NP-40, 1X HALT protease inhibitor cocktail (ThermoFisher), 0.5 mM DTT, 100 U/ml RNase OUTTM (ThermoFisher) and snap frozen in liquid nitrogen. Virion RNA was isolated by taking 100 μL of purified virions and adding 400 μL phenol, 100 μL chloroform, and 100 μL disruption buffer (200 mM Tris pH 8.5, 1M NaCl, 2 mM EDTA, 1% SDS). This mix was then vortexed for 30 seocnds and centrifuged at 13,000 x g for 5 min. The supernatant was then transferred to a new tube along with 100 μL phenol and 50 μL chloroform, vortexed again for 30 seconds, and centrifuged at 13,000 x g for 5 min. Nucleic acids in the resulting supernatant were then ethanol precipitated and resuspended in 50μL sterile distilled water.

8.7.2 RNA modification and purification

Frozen cell lysates were first thawed slowly on ice, then quickly spun to pellet any cell debris. The cleared lysates were then incubated at 20 °C for 15 min. Then 180 μL of the cleared lysates were added to either 20 μL dimethyl sulfoxide (DMSO; (-) control) or 65 mM 1-methyl-7-nitroisatoic anhydride (1M7; (+) sample) in DMSO and incubated for 5 min at 20 °C to complete modification. Then 600 μL of TRIzol reagent (Life Technologies) was added to the lysate samples and extracted according to the manufacturer's protocol using 20 mg of glycogen as a carrier. The resulting pellet was dissolved in 10 μL RNase-free H_2O. For purified RNA samples 3 μg of viral RNA, or 10 pmol of *in vitro* transcribed RNA, were dissolved in 12 μL of RNase-free H_2O then incubated for 95 °C for 2 min and snap-cooled on ice for 1 min. The viral RNA was refolded by adding 6 μL of 3.3X folding buffer (333 mM HEPES, 333 mM NaCl, 33 mM $MgCl_2$) and incubating at 20 °C for 20 min. Then, the 18 μL of folded viral RNA were split, adding 9 μL of RNA to 1 μL of either DMSO or 65 mM 1M7. The RNAs were modified for 3 min at 20 °C then ethanol precipitated, using glycogen as a carrier, and dissolved in 10 μL RNase-free H_2O.

8.7.3 Reverse transcription

To the 10 μL of modified RNA (or unmodified control), 3 μL reverse transcription primer mix was added, containing 0.5 μM of each oligonucleotide A-H (Table F.1), spanning the length of RNA3. The resulting mix was heated to 95 °C for 2 min, then incubated at 65 °C for 5 min before placing on ice for ~30 seconds. Next, 7 μL of SSIII master mix was added, containing: 0.5 μL of Superscript III (Life Technologies), 4 μL 5X First Strand Buffer (Life Technologies), 1 μL 100 mM (DTT), 1 μL 10 mM dNTPs,

and 0.5 μL RNase-free H_2O. The complete reaction mix was then incubated at 42 °C for 1 min, followed by extension at 52 °C for 25 min and deactivation at 65 °C for 5 min. The RNA was hydrolyzed by addition of 1 μL of 4 M NaOH solution and heating to 95 °C for 5 min. The basic solution containing the cDNA was partially neutralized with 2 μL of 1 M HCl and precipitated with 69 μL cold EtOH with thorough washing with 70% EtOH. The washed pellet, free of base, was dissolved in 22.5 μL of nuclease-free H_2O.

8.7.4 Adapter ligation

To the cDNA, 3 μL 10X CircLigase Buffer (Epicentre), 1.5 μL 50 mM $MnCl_2$, 1.5 μL 1 mM ATP, 0.5 μL 100 μM DNA adapter (oligonucleotide I; Table F.1), and 1 μL CircLigase I (Epicentre) were added. The reaction was incubated at 60 °C for 2 hr, then 80 °C for 10 min to inactivate the ligase. The ligated DNA was EtOH precipitated with 20 mg glycogen as a carrier and dissolved in 20 μL of nuclease-free H_2O. Then the cDNA was purified using 36 μL of Agencourt XP beads (Beckman Coulter), according to manufacturer's instructions and eluted with 20 μL TE buffer.

8.7.5 Quality analysis

For quality analysis (QA), a separate PCR reaction for each (+) and (-) sample was mixed by combining: 13.75 μL nuclease-free H_2O, 5 μL 5X Phusion Buffer (New England Biolabs), 0.5 μL 10 mM dNTPs, 1.5 μL of 1 μM labeling primer (oligonucleotides J/K; Table F.1), 1.5 μL of 1 μM primer PE_F (oligonucleotide L; Table F.1), 1 μL of 0.1 μM selection primer mix (0.1 μM each of oligonucleotides M-T or U-AB; Table F.1), 1.5 μL

ssDNA library (+ or -), and 0.25 μL Phusion DNA polymerase (New England Biolabs). Both fluorescent primers were purchased from Applied Biosystems and the selection primers were purchased from Integrated DNA Technologies. Phosphorothioate modifications were added to prevent the 3′→5′ exonuclease activity of Phusion polymerase (Table F.1). Amplification was performed for 15 cycles, then 50 μL nuclease-free H_2O was added, and the diluted reaction was ethanol precipitated. The resulting pellet was dissolved in formamide and analyzed with an ABI 3730xl capillary electrophoresis device.

8.7.6 Library preparation and next generation sequencing

To construct sequencing libraries, a separate PCR for each (+) and (-) sample was mixed by combining: 33.5 μL nuclease-free H_2O, 10 μL 5X Phusion Buffer (New England Biolabs), 0.5 μL 10 mM dNTPs, 0.25 μL of 100 μM TruSeq indexing primer (oligonucleotide AC; (Table F.1)), 0.25 μL of 100 μM primer PE_F, 2 μL of 0.1 μM selection primer mix (+ or -, as noted above), 3 μL ssDNA library (+ or -), and 0.5 μL Phusion DNA polymerase (New England Biolabs). Amplification was performed as indicated in 'Quality analysis' above. Completed reactions were chilled at 4 °C for 2 min before addition of 5 U exonuclease I (New England Biolabs) to remove unextended primer. The reactions were then incubated at 37 °C for 30 min. After incubation, the libraries were purified using 90 μL of Agencourt XP beads (Beckman Coulter) according to manufacturer's instructions. The complete libraries were eluted with 20 μL TE buffer and quantified with the Qubit 2.0 Fluorometer (Life Technologies).

To prepare the libraries for sequencing, the average length of each sample was determined using the results from the quality analysis in order to calculate the molarity

of each (+) or (-) separately. Sequencing pools were mixed to be equimolar, such that all of the sequencing libraries were present in the solution at the same level. Sequencing was performed on the Illumina MiSeq using 2x35 bp paired end reads.

8.7.7 Data analysis with Spats

Reads analysis was performed with Spats v1.0.0 (`https://github.com/LucksLab/spats/releases/`), using cutadapt v1.5 [267] and Bowtie 0.12.8 [182] to do the adapter removal pre-processing. Because of the length of RNA3 and the tendency of reverse transcriptase to 'fall off' during RT, the output of Spats contains distinct windows of ~300-400 nt of meaningful reactivity data upstream of each priming site. To combine all of the reactivity data together the calculated β_i reactivity values for each nucleotide i, representing the probability that nucleotide i was modified, were first normalized to B_i values within each reactivity window according to:

$$B_i = \sum_{j=1}^{L} \frac{\beta_i L}{\beta_j} \tag{8.1}$$

where L is the length of the reactivity window. At overlapping positions between reactivity windows, the number of good alignments present in each window determined where the boundary lines were drawn. If both windows exhibited similar read alignments, the window of the RT priming site closer to the 3' end of RNA3 was selected.

CHAPTER 9

STRUCTURAL FEATURES OF PROTEIN BINDING WITH RNASE P AND ITS SUBSTRATES

9.1 Abstract

RNase P is an ancient enzyme present in all three domains of life, best known for its role in cleaving the 5′ leader of precursor tRNAs (pre-tRNAs) during tRNA maturation. Two types of RNase P exist in nature: a well-studied ribozyme, which functions as part of a ribonucleoprotein (RNP), and a proteinaceous version (PRORP) that exists in eukaryotes. In this work, we used SHAPE-Seq to study how protein cofactor binding to the archaeal RNase P ribozyme alters the RNA's structural features. Specifically, the binding of the L7Ae protein to kink-turns present in the substrate specificity domains of *Methanocaldococcus jannaschii* (*Mja*) and *Pyrococcus furiosus* (*Pfu*) RNase P RNA led to pronounced decrease in the reactivity of sites interacting with the L7Ae protein. We also examined changes that occur in a pre-tRNA when it binds a proteinaceous RNase P. We also provide experimental evidence to show that PRORP binds to nucleotides in the D-loop of pre-tRNA and loosens the D stem. Collectively, our results demonstrate how SHAPE-Seq can be used as a tool to infer protein-RNA interactions, which determine the structure and function of RNase P in different evolutionary contexts.

The work presented in this chapter was done in collaboration with the Gopalan Lab at the Ohio State University.

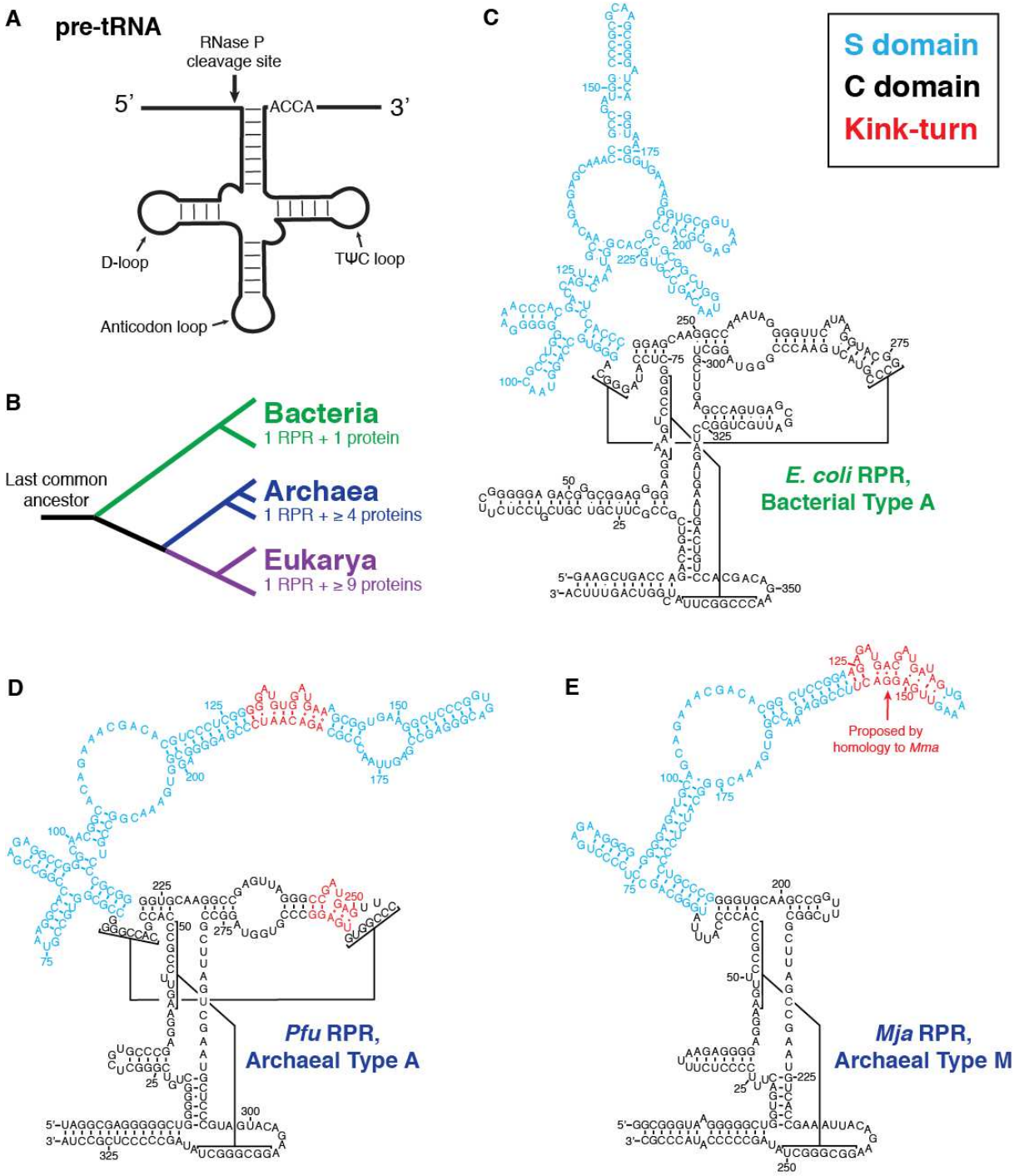
9.2 Introduction

RNase P was one of the first ribozymes ever discovered [363, 364]. Ribozymes are RNAs that perform catalysis [2], and their discovery began a new chapter in our biological understanding of the roles RNA plays in the cell [154]. RNase P cleaves a host of RNAs including 4.5S RNA (in bacteria), precursor tmRNA, mRNA transcripts, riboswitches, and precursor tRNAs (pre-tRNAs) [365–368]. However, RNase P is most commonly known for its major cellular function of cleaving the 5′ leader sequences from pre-tRNAs (Figure 9.1A) [201, 205, 369, 370]. Mature tRNAs are critical for cell viability [371], as they are key adaptors for translation of the triplet code. However, they are typically transcribed in a pre-tRNA form that is not easily recognized by aminoacyl transferases, the proteins that 'charge' tRNAs for translation. RNase P binds these pre-tRNAs and cleaves the 5′ leader sequence as part of the tRNA maturation process [201, 205, 369, 370]. Interestingly, all forms of life have some type of RNase P, making it one of the most conserved RNAs known (Figure 9.1B) [149].

Most structures of bacterial and archaeal RNase P RNA fall into one of a few types. Bacterial RPRs can be classified as type A (ancestral), B (*Bacillus*), or C (Chloroflexi) [372, 373]. Meanwhile, archaeal RPRs are classified as type A, M (Methanococci), or T (*Pyrobaculum*) [374–378]. Most of the five unique types maintain a pair of conserved motifs that are important for substrate recognition (the 'S' domain) and catalysis (the 'C' domain) (Figure 9.1C-E). The S and C domains fold independently to form the conserved core of RNase P and are sufficient for activity with certain RPRs without assistance from protein cofactors [370]. In different species the structural periphery, which includes regions for protein cofactor binding, differs in how it forms around the core. In bacteria, a single protein interacts with the RPR to improve catalysis, but further in the evolutionary timeline the number of RNase P protein cofactors tends to

increase while the catalytic activity of the isolated RPR decreases [11, 205]. Archaeal RNase P has ≥5 or more protein cofactors (including one ribosomal protein), while the human RNase P holoenzyme contains at least 10 proteins [11]. One theory suggests that the increasing number of proteins provides a higher level of processing flexibility and fidelity [379].

Figure 9.1: Characteristics of RNase P. (A) RNase P cleaves the 5′ leader of a precursor tRNA (pre-tRNA) at the site indicated as part of the tRNA maturation process. (B) RNase P is a well-conserved RNA that spans all three domains of life. Over evolutionary history, bacterial RNase P RNAs (RPRs) evolved to work in concert with 1 protein cofactor. Archaeal and Eukaryal RPRs evolved to work with multiple proteins. Three example RPRs are shown for *Escherichia coli* (*E. coli*), *Pyrococcus furiosus* (*Pfu*), and *Methanocaldococcus jannaschii* (*Mja*) in (C), (D), and (E) as prototypes for bacterial type A and archaeal types A and M, respectively. The specificity domain (S domain) and catalytic domain (C domain) are indicated with light blue and black, respectively. Predicted (Mja) [153] and discovered (*Pfu*) [152] kink-turn motifs in the archaeal RPRs are marked in red in (D) and (E).


A
pre-tRNA
RNase P
cleavage site
5'
ACCA
3'
D-loop
TΨC loop
Anticodon loop
B
Last common
ancestor
Bacteria
1 RPR + 1 protein
Archaea
1 RPR + ≥ 4 proteins
Eukarya
1 RPR + ≥ 9 proteins
C
S domain
C domain
Kink-turn
E. coli RPR,
Bacterial Type A
D
Pfu RPR,
Archaeal Type A
E
Proposed by
homology to Mma
Mja RPR,
Archaeal Type M

The structures have been solved for the proteins that interact with the bacterial and archaeal RPRs [380–384], leading to a great deal of insight into how the bacterial protein interacts with its RPR to improve substrate recognition [205, 370, 373]. However, the roles that the protein components serve in archaeal and eurkaryal RNase P remains poorly understood [373], although it is known that two of the archaeal proteins improve substrate affinity [385], and two others increase turnover rate [386]. The fifth archaeal RPR protein cofactor was recently discovered to be the ribosomal protein L7Ae [153], known to interact with an RNA motif known as a kink-turn [387]. The binding of L7Ae to kink-turns within both M type and A type archaeal RPRs was demonstrated using *Methanococcus maripaludis* (*Mma*) [153] and *Pyrococcus furiosus* (*Pfu*) [152]. The addition of L7Ae along with the other four aforementioned proteins increased the optimal reaction temperature and cleavage rate to nearly the activity observed with the native holoenzyme [152].

The discovery of a proteinaceous RNase P (PRORP) in *Arabidopsis thaliana* [388] and human [389] that did not contain an RPR showed that RNase P could use either an RNA- or protein-based active site. In *Arabidopsis*, three different PRORP proteins function as single proteins to process pre-tRNAs [150], unlike the human PRORP that exists as part of a three-protein complex [390]. One of the *Arabidopsis* PRORP variants (PRORP1), which was localized to mitochondria and chloroplasts, has the ability to rescue an RNase P knockout in *E. coli*, demonstrating the same functional equivalency as the bacterial RNase P [150, 391]. Interestingly, PRORP does appear to maintain the general organization of two domains that separately manage RNA binding and catalysis (Figure 9.2) [150, 201, 392, 393].

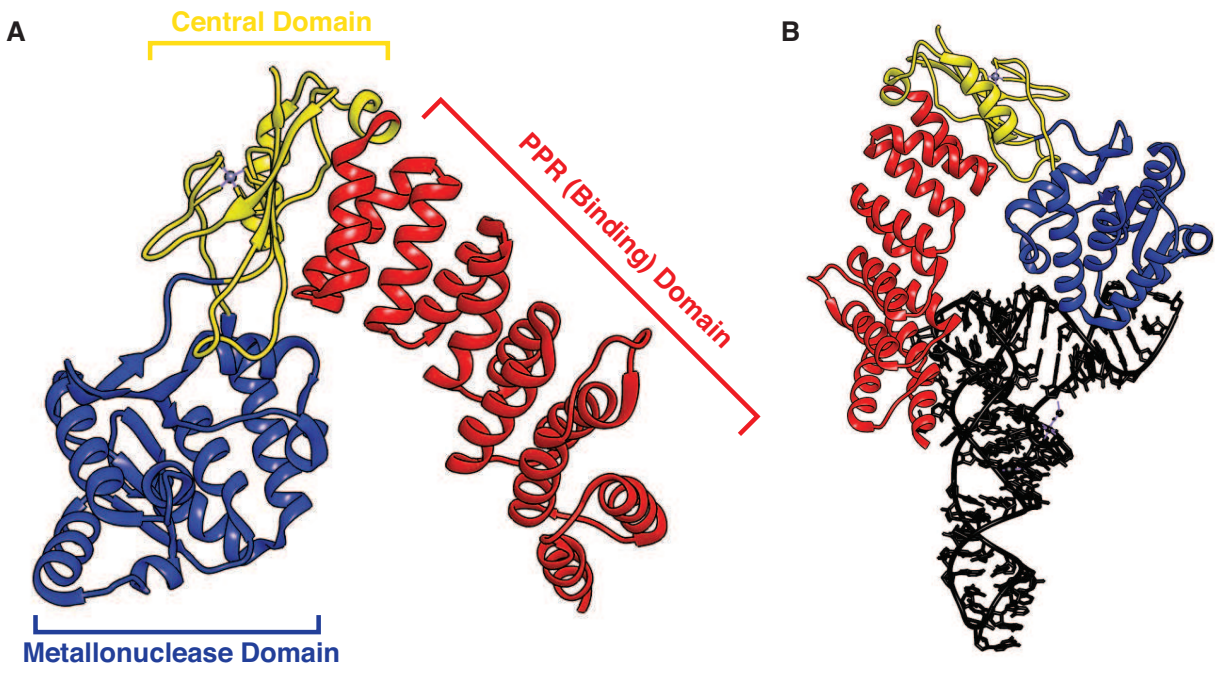


Figure 9.2: Proteinaceous RNase P structure. (A) Structure of the proteinaceous RNase P from *A. thaliana* (PRORP1), colored by domain. The PPR domain (red) is responsible for pre-tRNA recognition, the metallonuclease domain (blue) cleaves the 5′ leader sequence, and the central domain (yellow) links the PPR and metallonuclease domains. (B) Proposed structural docking of PRORP1 on to a tRNA structure. The PDBs used were originally deposited by Howard *et al.* [394] (PRORP1; 4G23) and Byrne *et al.* [103] ($tRNA^{phe}$; 3L0U).

In this work, we examine the role that protein binding has on RNA structure in terms of RNase P substrate recognition and cleavage, both from the viewpoint of the 'classical' ribozyme RNase P, via the archaeal *Mja* (archaeal type M) and *Pfu* RPRs, as well as PRORP1 from *A. thaliana* organelles (herein referred to as simply 'PRORP'). By examining changes in the nucleotide flexibilities before and after protein addition, we locate protein-binding sites within the RPRs or pre-tRNAs (for PRORP experiments) and suggest how the bound proteins might alter local structures to promote substrate recognition and/or cleavage.

9.3 Results and Discussion

9.3.1 L7Ae binds kink-turn motifs in archaeal RNase P

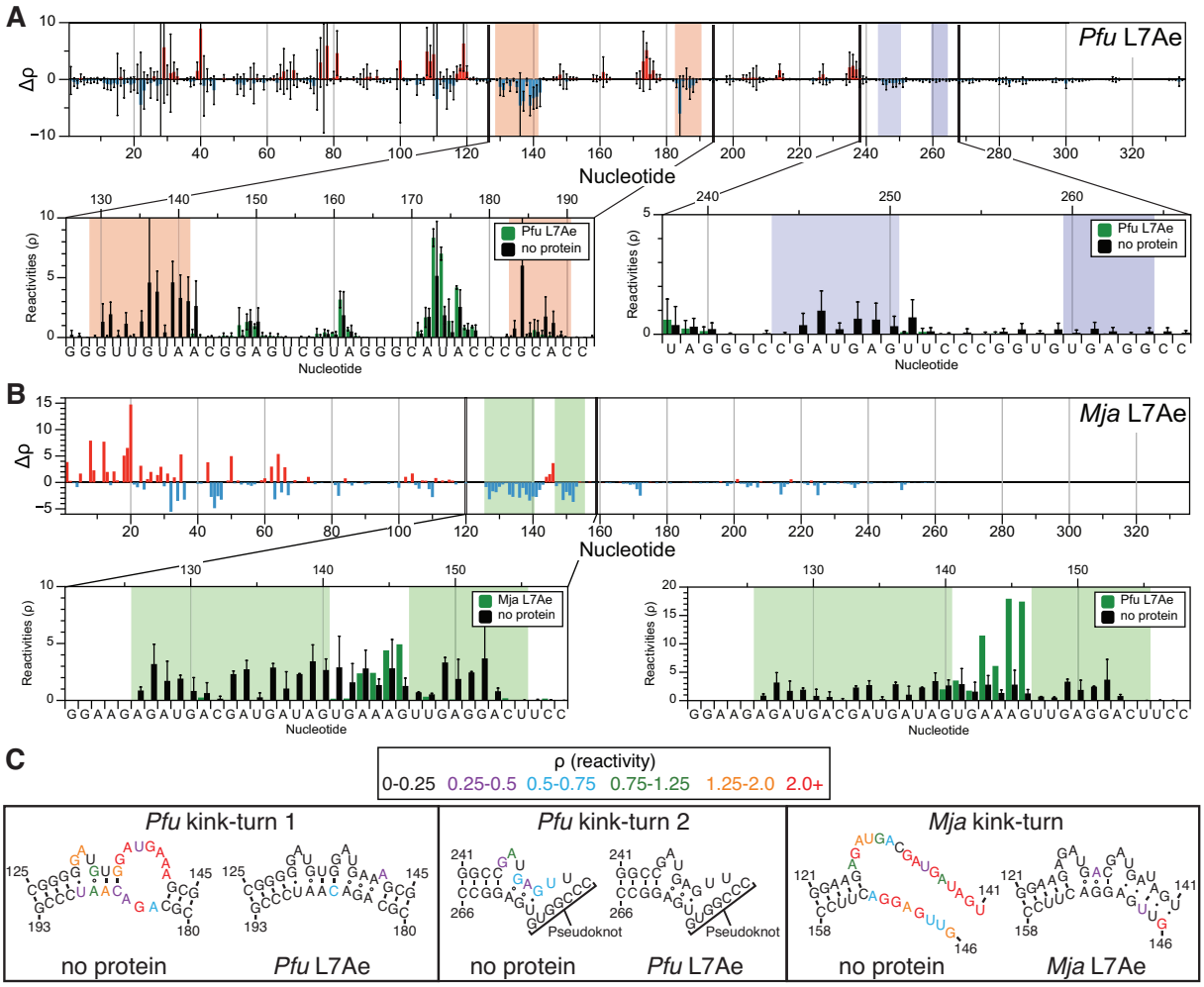
We began our study on the role of protein binding on RNase P structure and function with two archaeal ribozymes from *Pyrococcus furiosus* (*Pfu*) and *Methanocaldococcus jannaschii* (*Mja*) (Figure 9.1D,E). In both species, four RNase P proteins (RPPs) have been identified [11] that bind their RPRs as two distinct heterodimers to improve substrate recognition and catalysis [385, 386, 395–398]. Recently, a fifth protein, L7Ae, a well-characterized protein class known to bind RNA kink-turn motifs in a variety of cellular contexts [399, 400], was identified as an archaeal protein cofactor [153, 401]. The L7Ae binding site within the *Mma* RPR was found by searching within the RPR sequence for potential kink-turn motif sites and either functionally testing each potential site by mutational analysis [153] or directly modifying the L7Ae protein with an iron complex to locally cleave the RPR via generated hydroxyl radicals [152]. However, both methods are time/labor intensive and not likely to be used to experimentally validate kink-turns in a large set of RPRs or other RNAs.

To rapidly identify L7Ae binding sites, we applied the recently developed SHAPE-Seq v2.1 technique [123] to search for regions in the RPRs that exhibit decreased RNA flexibility upon binding to L7Ae. We first examined the *Pfu* RPR, as two different kink-turn motifs had been previously identified within it, and both serve as L7Ae binding sites [152].

By comparing the relative reactivity of each nucleotide within the *Pfu* RPR with and without the addition of the *Pfu* L7Ae protein, we were able to observe how L7Ae binding affects RPR structure (Figures 9.3A and G.1). Notably, we see reactivity de-

creases in both sites previously identified to contain kink-turns. Interestingly, while the reactivity drops are clear for first kink-turn (indicated by red shading), the second kink-turn (indicated with blue shading) exhibits a weaker change in reactivity, mainly due to the overall lower flexibility in the region before L7Ae addition. The proximity of a pseudoknot within the C-domain kink-turn 2 to may serve to stabilize the sheared G•A base pairs, which are not as stable in kink-turn 1 (Figure 9.1C). In fact, the presence of L7Ae may be required to stabilize kink-turn 1 for its binding or, alternatively, the properly formed kink-turn 1 exists rarely in the RPR structural population, but long enough for L7Ae to bind and drive the equilibrium of the structural population as a whole toward the L7Ae-bound form.

Figure 9.3: Characterization of L7Ae binding in the two archaeal RPRs from *Pyrococcus furiosus* (*Pfu*) and *Methanocaldococcus jannaschii* (*Mja*). (A) Differences in *Pfu* RPR reactivity after *Pfu* L7Ae is added during *in vitro* folding. The two kink-turn motifs present in *Pfu* RPR are highlighted in red and blue. When L7Ae is added, kink-turn reactivities decrease (insets). Side by side bars for the entire *Pfu* RPR can be found in Figure G.1. Data represent an average of two or four replicates with or without L7Ae present, respectively. (B) Same analysis as (A) for the *Mja* RPR, which contains only 1 kink-turn motif highlighted in green. Zoomed regions surrounding the kink-turn demonstrate that both *Mja* and *Pfu* L7Ae can bind the *Mja* RPR kink-turn and lower reactivity. Side by side bars for the entire *Mja* RPR can be found in Figure G.2. Data represent an average of two replicates for *Mja* RPR without L7Ae, one replicate each with L7Ae. All error bars in (A) and (B) represent one standard deviation. (C) Secondary structures of the kink-turns highlighted red, blue, and green above for *Pfu* kink-turn 1, *Pfu* kink-turn 2, and the *Mja* kink-turn, respectively. Nucleotides are colored according to reactivity intensity with and without the indicated L7Ae protein. High reactivities in certain kink-turn positions suggest single-strandedness and are thus drawn as such.


A
Pfu L7Ae
Δρ
Nucleotide
Reactivities (ρ)
Pfu L7Ae
no protein
B
Mja L7Ae
C
ρ (reactivity)
0-0.25 0.25-0.5 0.5-0.75 0.75-1.25 1.25-2.0 2.0+
Pfu kink-turn 1
Pfu kink-turn 2
Mja kink-turn
Pseudoknot
no protein
Pfu L7Ae
Mja L7Ae

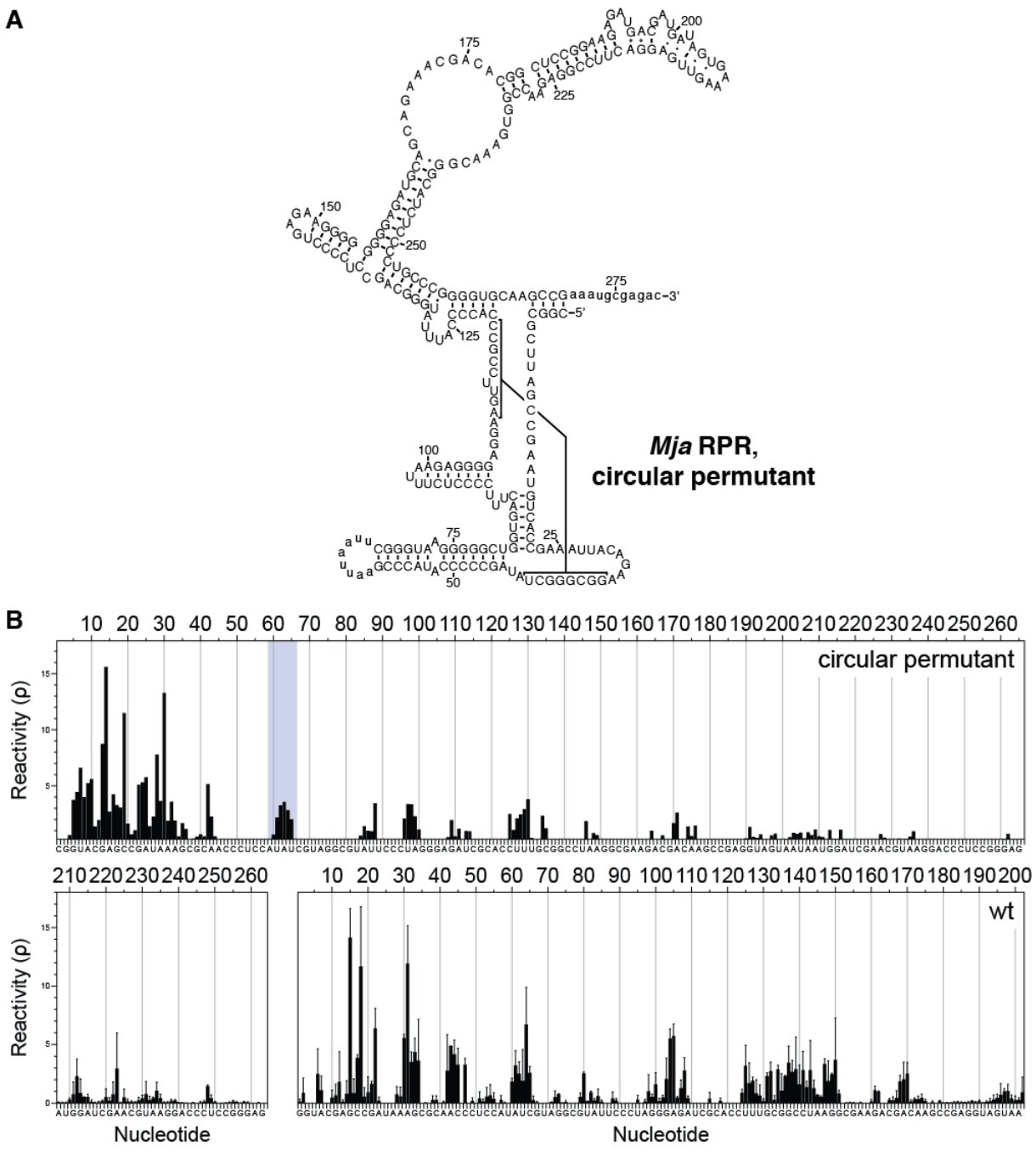
We also examined the *Mja* RPR with SHAPE-Seq to locate a potential L7Ae binding site, as suggested by comparison to the known binding site in the closely related Mma RPR [153]. Upon examining the reactivity maps of the *Mja* RPR with and without L7Ae, we were quickly able to locate a large region of decreased reactivity in the presence of L7Ae between nucleotides (nts) 126-141 and 147-153 (Figures 9.3B,C and G.2). Like kink-turn 1 in the *Pfu* RPR, the kink-turn motif in the *Mja* RPR appears to be fairly unstructured without protein binding (Figure 9.1C). We also observed that the stabilizing of the *Mja* RPR kink-turn led to increased reactivities in the loop between nts 141-146 (Figure 9.1E). Interestingly, we found that L7Ae from *Pfu* could enact the same structural changes in the *Mja* RPR, leading to even higher reactivities in the aforementioneD-loop.

By observing reactivity changes in RPR with SHAPE-Seq, we can gain insight into how the RPPs and L7Ae remodel RNase P for faster and more accurate pre-tRNA processing [402].

9.3.2 Observing the independence of the S and C domains with SHAPE-Seq

Past work has demonstrated that the S and C domains of RNase P fold independently [403, 404], making it possible to generate functional circular permutants [71, 405]. These circular permutants are created by disrupting the RPR sequence in a loop to create new 5′ and 3′ ends while also linking the original 5′ and 3′ ends together to create a new loop (Figure 9.4A). Changing the RPR termini generates new contexts for RNA folding, potentially aiding overcoming folding traps [71].

Figure 9.4: Analysis of an *Mja* RPR circular permutant. (A) Structure of the *Mja* RPR circular permutant (cp) studied. Bases in lowercase were artificially added to create the cp. (B) Alignment of the equivalent nucleotides between the wt and cp *Mja* RPR, numbered according to each nucleotide's position in its respective RPR context. The reactivity maps match well between the wt and cp, except for nts 210-250 (wt numbering), suggesting that either the circular permutant introduces some instability in that region, or that the data is noisy in that region. The highlighted region represents nts 59-66 in the artificially added sequence to connect the wt 5′ and 3′ ends. Two additional 5′ Gs were included in the *Mja* RPRs that are not shown. The reactivity data shown for the wt is an average of 2 measurements with error bars that represent one standard deviation.


A
Mja RPR,
circular permutant
B
circular permutant
wt
Reactivity (ρ)
Nucleotide

To examine the modularity of the RNase P domains, we collected SHAPE-Seq data on a circular permutant (cp) of the *Mja* RPR and compared it to the wt *Mja* RPR (Figure 9.3). Across the *Mja* RPR, we found remarkable similarity in the reactivity patterns between the wt and cp. One major difference did exist, however, as the first 45 nts of the cp construct exhibited much higher reactivities than the wt. It is unclear whether this difference was due to a lack of experimental replicates or a destabilization effect introduced by the circular permutation. Due to the averaging nature of the ρ reactivity measurement (as the sum of all ρ values equals the length of the RNA), the increased reactivity at the 5′ end of the cp RPR leads to lower reactivities downstream. Further experiments would need to be performed to discern the cause of the high reactivities.

We also examined the C domain of the *Pfu* RPR in isolation relative to the complete RPR (Figure G.3). As expected, the reactivity traces of the isolated C domain match very well to the complete RPR, except across nts 223-229 (in the wt numbering) where reactivities are higher for the isolated domain. However, these nts directly abut the junction point between nts 63 and 223 (wt numbering) that distinguishes the C domain. Likely, these nucleotides lose some of their structure, and thus increase in reactivity, to allow the nearby pseudoknot to form.

9.3.3 PRORP Binds the D-loop of pre-tRNAs

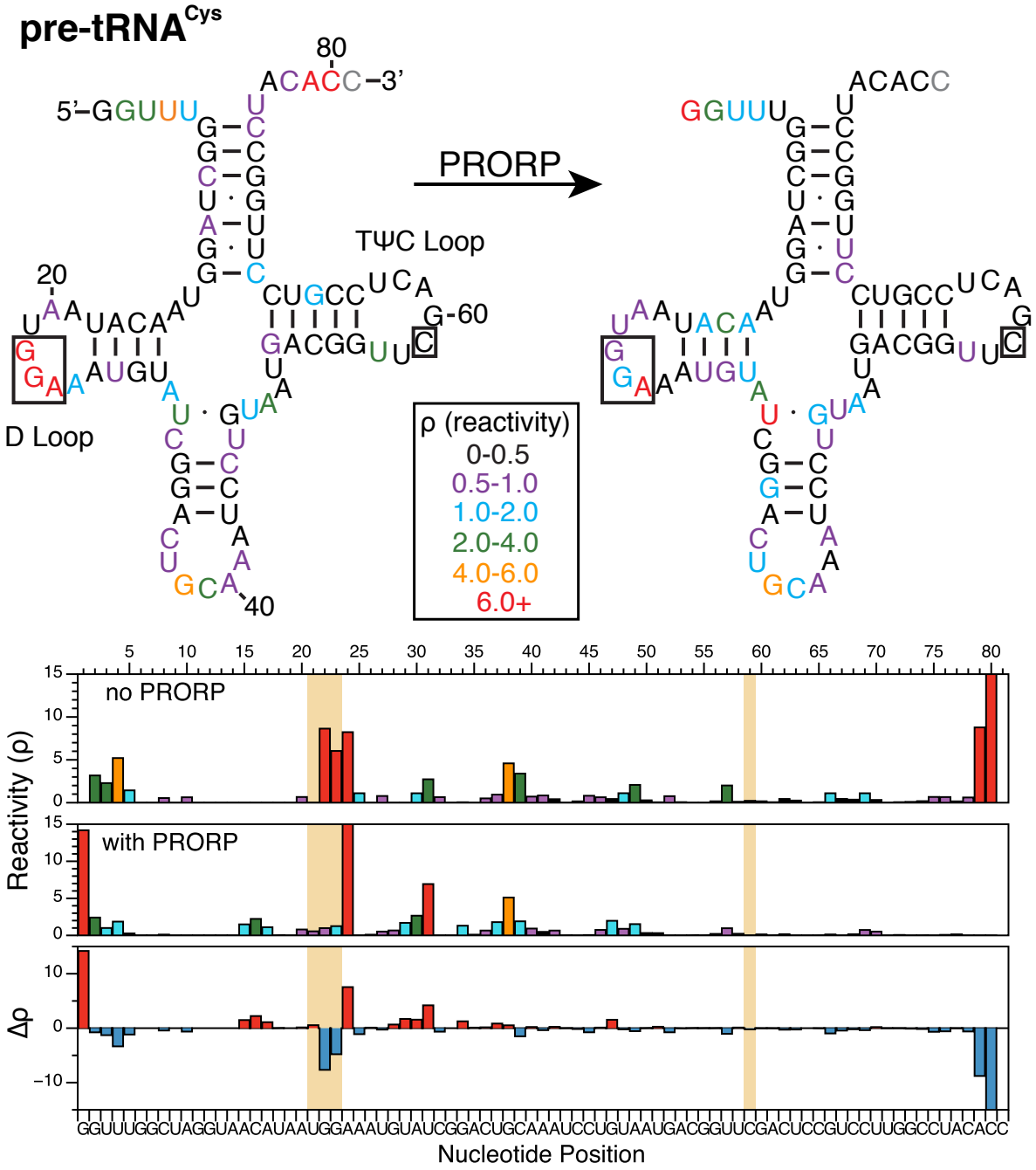
While there is an expansive amount of literature discussing the RNase P ribozyme, the proteinaceous versions of RNase P, which were discovered only eight years ago, are not as well understood [150, 388–393, 406–409]. Conservation analysis and crystal structure data have indicated that PRORP1 in plant mitochondria and chloroplasts is organized into an RNA binding domain containing multiple pentatricopeptide re-

peat (PPR) motifs and a metallonuclease (NYN) domain [150, 409]. Interaction studies between PRORP and its pre-tRNA substrate have suggested that the PPR domains recognize the D-loop/TΨC loop stack to bring the NYN domain in proximity to the 5′ leader sequence for cleavage [393, 408, 409]. However, these studies relied on the use of local protection assays that are subject to solvent accessibility biases and therefore exhibit too much noise to clearly identify single-base interactions between PRORP and its pre-tRNA substrate.

To gain a better understanding of the interactions occuring between PRORP and its substrate, we again turned to SHAPE-Seq v2.1 [123] to identify reactivity changes within the pre-tRNA upon PRORP binding. We began by analyzing the *Arabidopsis* pre-tRNA^{Cys} in isolation before adding the wt PRORP, also from *Arabidopsis* (PRORP1), in calcium buffer to prevent cleavage. Upon PRORP addition to 5-fold excess, we observed a drop in the reactivities of G22 and G23 in the D-loop, as well as an increase in A24 (Figure 9.5). Nucleotides in the D stem also appeared to show a modest increase in reactivity, although the signal/noise ratios are not high. Interestingly, no clear reactivity changes are observed in the TΨC loop, although the loop is generally unreactive to the SHAPE probe without PRORP [122], a fact that makes mapping changes difficult.

Next, we examined the effect of the K101A and K439A mutations previously described in Chen *et al.* on the ability of PRORP to bind pre-tRNA [393]. The K101A mutation occurs in the PPR domain, and the K439A mutation in the NYN domain. Comparing the wt PRORP binding to the mutants revealed very similar patterns (Figures 9.6 and G.4), with all three variants exhibiting protection of G22 and G23. Also, slight reactivity increases in the D stem were again observed for all variants. These results, in combination with those from Chen *et al.* [393], suggest that while K101 is localized to the D/TΨC loop stack, it is not likely interacting with the pre-tRNA.

Figure 9.5: PRORP binds to the D-loop of pre-tRNA^{Cys}. Shown at the top is secondary structure of pre-tRNA^{Cys} from *Arabidopsis* with a 5 nt leader sequence. The nucleotides are colored by reactivity intensity (bars, bottom) after refolding *in vitro* with Ca^{2+} buffer in the absence (left) or presence (right) of 5-fold excess PRORP. Nucleotides G22 and G23 show clear decreases in reactivity upon PRORP binding, while neighboring nucleotides in the D stem exhibit increases. The boxes and tan shading represent the same nucleotide positions, which were observed to exhibit protection from PRORP in Gobert *et al.* [392]. No decreases in the TΨC loop were observed in the presence of PRORP, however the TΨC loop exhibited low reactivity in the absence of PRORP, making changes hard to detect.


pre-tRNA^Cys
5'–GGUUU
ACACC–3'
80
20
40
D Loop
TΨC Loop
G-60
PRORP
ρ (reactivity)
0-0.5
0.5-1.0
1.0-2.0
2.0-4.0
4.0-6.0
6.0+
no PRORP
with PRORP
Reactivity (ρ)
Δρ
GGUUUGGCUAGGUAACAUAAUGGAAAUGUAUCGGACUGCAAAUCCUGUAAUGACGGUUCGACUCCGUCCUUGGCCUACACC
Nucleotide Position

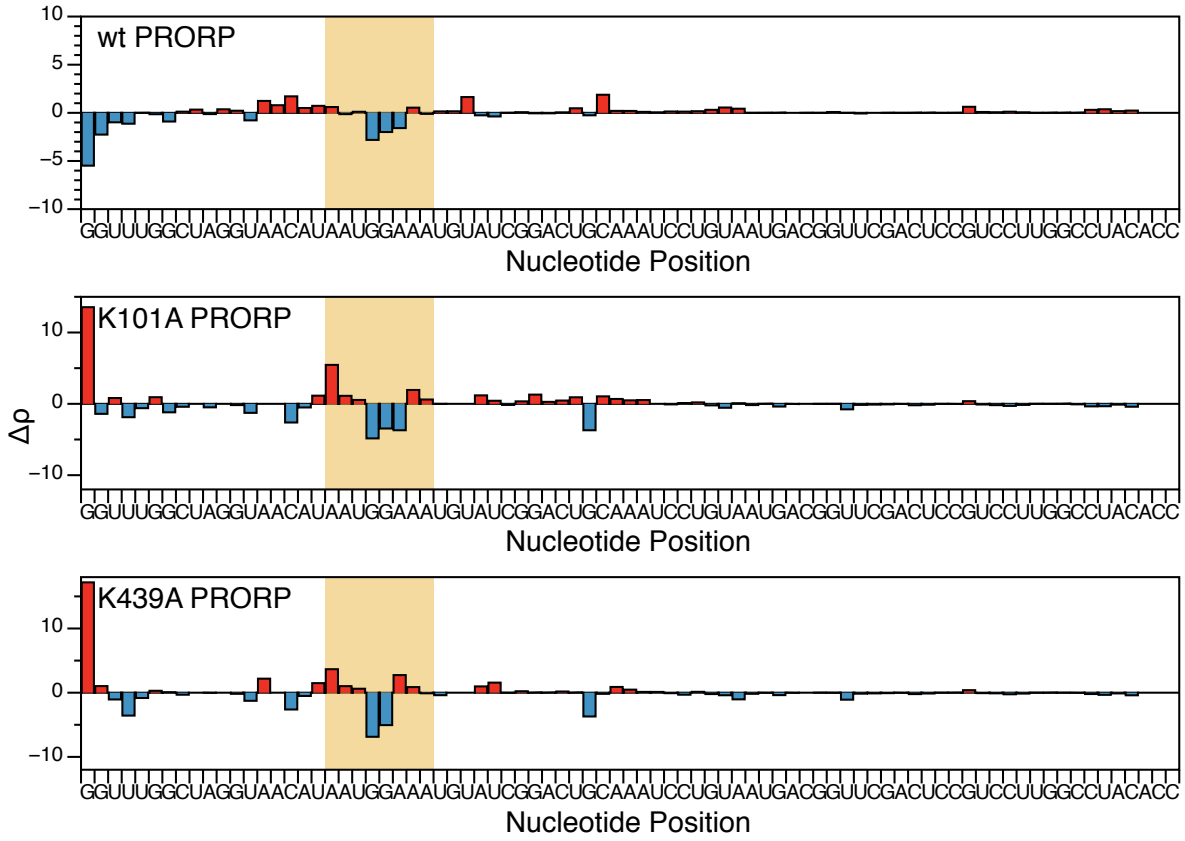


Figure 9.6: PRORP mutants exhibit similar pre-tRNA binding patterns as the wt. The difference in reactivity that occurs upon addition of 20-fold excess of PRORP is shown for the wt PRORP as well as two mutants: K101A and K439A from Chen *et al.* [393]. The D-loop is highlighted with tan. The K101A mutation occurs in a PPR domain within PRORP important for substrate binding (and exhibited a high degree of protection with pre-tRNA [393]), while the K439A mutation is in the NYN domain. However, when either mutant PRORP is incubated with pre-tRNA^{Cys} roughly the same changes in reactivity are observed as the wt. The consistency of this result suggests that while K101 is likely near a critical residue in PRORP for pre-tRNA binding, K101 itself is not crucial for substrate binding. Data represent an average of two replicates for wt PRORP and one replicate each for the mutants. The individual reactivity maps can be found in Figure G.4.

We also examined an alternate form of the pre-tRNA containing longer leader and trailer sequences with the expectation that the increased number of bases may reduce the run-to-run variability that we observed in SHAPE-Seq data of the pre-tRNA/PRORP binding experiments. We performed the same PRORP binding experiment in triplicate, observing the same features as described above, although one replicate contributed to a considerable amount of noise (Figure G.5).

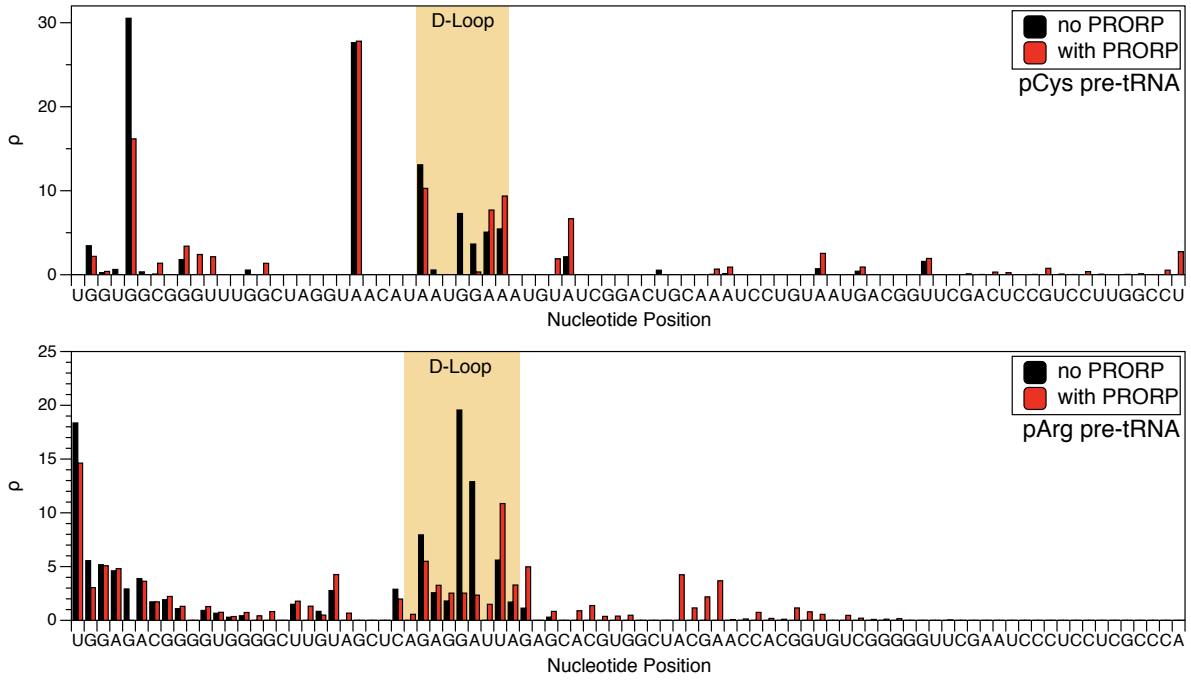


Figure 9.7: Direct priming two pre-tRNAs shows similar patterns of PRORP protection. In both pCys and pArg pre-tRNAs, we observe protection of two bases in the D-loop when PRORP is present. In pCys, it is positions 4 and 5 in the D-loop that are protected, while positions 5 and 6 are protected for pArg.

Last, we examined the cysteine and arginine pre-tRNAs, pCys and pArg, respectively, with longer 3′ tails in order to directly prime the RT reaction, thus removing the need for ligation. Again, we observed protections in the D-loop from the addition of PRORP in both pre-tRNAs, providing evidence that PRORP recognizes the D-loop, regardless of the anticodon sequence. Interestingly, however, the protected bases were different in the two pre-tRNAs, with positions 4 and 5 protected in pCys and 5 and 6 in pArg. The difference in the protected positions may be due to the local 3D context of the loop nts, given that both dinucleotide pairs are in the center of their respective D-loops.

9.4 Future work

The preliminary data collected in this work demonstrates that SHAPE-Seq is a useful tool to examine protein binding in the context of RNase P biochemistry. We observed clear reactivity changes associated with L7Ae binding to RPRs as well as decreases in D-loop reactivities in pre-tRNAs bound by PRORP. We outline future directions that are likely to be productive.

9.4.1 Examining other RPRs and RPPs with SHAPE-Seq

Our initial studies of RPRs contained a larger set than just *Mja* and *Pfu*, and suggested that similar SHAPE-Seq measurements could be made for other RPRs with variable sizes and GC content. By extension, RPRs from other domains would be amenable to SHAPE-Seq analysis and could provide useful insight into RPR structure and catalysis. This would be particularly valuable for eukaryal RPRs, as little is known about them.

Our promising results with L7Ae highlight the potential to gather information about RPP binding via SHAPE-Seq. While less useful for bacterial RPRs, as the crystal structure of the RPP-RPR complex has been solved [195], the archaeal and eukaryal RPPs are less well characterized. For example, the four archaeal RPPs have been assigned functions [385, 386, 395–398], but the exact binding sites are unknown and no crystal structure exists. The difficulty in determining the locations of RPP binding sites in a wide variety of species is compounded are new families RPRs are constantly being discovered with different structural motifs [375, 378].

While studies to assay the effect of RPP binding on RPR catalytic activity are fairly straightforward, determining a binding site could require time-consuming mutagenesis and RNA cleavage studies [152]. Using the SHAPE-Seq approach, the wt RPPs and RPRs can be used directly in the experiment without modification, removing the concern of complicating side effects from potentially altering the RNA or protein form and/or function. It is conceivable that the entire RPR/RPP assembly process can be monitored with SHAPE-Seq reactivity changes. The one drawback, however, is the inability of SHAPE-Seq to resolve protein binding in structured regions, where reactivities are very low and do not change upon binding to RPPs.

9.4.2 Studying RNase P inside cells

Much of the work on RNase Ps has been performed *in vitro*, outside of the cellular context. In another study (Chapter 3), we demonstrated that in-cell SHAPE-Seq can probe endogenous RNAs in living cells and obtained a reactivity map for the *E. coli* RPR in the cell [113]. Comparing *in vitro* and in-cell reactivity maps of RPRs could yield information about cellular state and tRNA processing and potentially help identify

new protein cofactors. It would also be exciting to see in-cell SHAPE-Seq data collected for archaeal RPRs, as little to no work has been done to examine RNA structures inside living archaea.

9.4.3 Obtaining a deeper understanding of PRORP binding

The method by which PRORP binds pre-tRNAs is still fairly unknown. To improve our understanding of PRORP binding, we are currently measuring pre-tRNAs with long trailers in order to use the SHAPE-Seq technique with direct priming for reverse transcription, as opposed to ligation. We expect to map reactivity changes that occur upon PRORP binding with greater accuracy than we have achieved up to now (Figure G.5).

Beyond just observing PRORP binding in the D-loop, understanding what the reactivity changes in the pre-tRNA physically mean may reveal functional roles of PRORP. Given the observed increase in reactivities in the D stem, one possibility is that either PRORP binds pre-tRNAs when they are in a λ conformation, or that PRORP binding rearranges the pre-tRNA into the λ-form (Lai, Chen, and Gopalan, personal communication; Figure G.6) [410]. If indeed the λ-form is preferred by PRORP, there could be greater implications for how tRNA modification enzyme bind and process tRNAs, especially when considering the unusual forms some tRNAs maintain in non-nuclear organelles, where PRORP is found. The λ-form binding preference of PRORP might distinguish it from the RPR, and suggest that despite convergent evolution the RNA- and protein-based active sites may recognize different pre-tRNA conformations.

9.4.4 Cotranscriptional Folding of RNase P ribozymes

Many RPRs can function as true ribozymes, not requiring protein cofactors to cleave pre-tRNA 5′ leader sequences. However, *in vitro* assays are constantly plagued with misfolded structures that are not catalytically active. The order of folding of the RNase P domains has been implicated as the culprit since *in vitro* refolding does not capture the cotranscriptional folding elements of RPR synthesis, and circularly permutant RPR variants exhibit different levels of correct folding [71, 393]. If cotranscriptionally folded, one would expect the entire S domain to fold first before the C domain because the C domain would not have been synthesized yet.

To test the effects of the order of folding, a cotranscriptional SHAPE-Seq experiment (Chapter 5) could be done on the wt RPR, as well as a few circular permutants. It would also be interesting to combine elements of the RPP binding studies proposed above to understand when and where RPPs bind during folding. These experiments represent a significant challenge to the cotranscriptional SHAPE-Seq technique (as it exists at the time this passage was written) in terms of length, as some RPRs can be much longer than RNAs studied so far. However, we are making progress in that area (Chapter 6).

9.5 Acknowledgments

We thank the Gopalan Lab at Ohio State University for their help and insight into making this work possible. Specifically, Tien-Hao Chen, Lien Lai, Stella Lai and Ila Marathe for thoughtful discussions and preparation of RPRs, pre-tRNAs, L7Ae, and PRORPs. We also thank Venkat Gopalan for his time and effort performing experiments with us

in the lab, as well as his thoughtful discussions and proposals and for critical reading of this document.

9.6 Materials and Methods

9.6.1 RNA folding and modification

All purified RNAs and proteins were provided by the Gopalan Lab at Ohio State University.

Experiments examining RNase P RNA structures were performed by first dissolving 1 pmol of the RNase P RNA in 12 μL H_2O and incubating it at 50 °C for 50 min followed by 37 °C for 10 min. Subsequently, 6 μL of 3.3X acetate buffer (50 mM Tris-acetate (pH 8), 800 mM NH_4OAc, and 10 mM $Mg(OAc)_2$ or $Ca(OAc)_2$) were mixed with the RNA solution before addition of either 1 μL 10 μM L7Ae protein (in 1X acetate buffer) or 1X acetate buffer only. The RNA was then incubated for another 10 min at 55 °C. Modification was performed by adding 9.5 μL of the folded RNA to 0.5 μL of either dimethyl sulfoxide [DMSO; (-) control] or 130 mM 1-methyl-7-nitroisatoic anhydride in DMSO [1M7; (+) sample] and incubating for 2 min at 55 °C. To both (+) and (-) reactions, 40 μL of nuclease-free H_2O were added followed by 150 μL of TRIzol reagent (Life Technologies). The RNA was extracted according to the manufacturer's protocol and precipitated using glycogen as a carrier. The resulting pellets were dissolved in 10 μL of 10% DMSO.

Experiments examining proteinaceous RNase P (PRORP) entailed dissolving 1 pmol of pre-tRNA in 12 μL H_2O and incubating it at 95 °C for 5 min before cooling

to 37 °C (0.1 °C/s ramp rate) where it was held for an additional 10 min. Subsequently, 6 μL of 3.3X PRORP binding buffer (1X: 20 mM HEPES-KOH pH 7.2, 100 mM ammonium acetate, 10 mM $Ca(OAc)_2$, 4 mM DTT, and 5% glycerol) were added, followed by an incubation step of 30 min at 37 °C. Then, 1 μL of either 5 μM or 20 μM PRORP protein solution (in 1X PRORP buffer) or 1X PRORP buffer was added. The resulting RNA-protein mixture was incubated at 22 °C for 15 min. Modification was performed by adding 9.5 μL of sample to 0.5 μL of either dimethyl sulfoxide [DMSO; (-) control] or 130 mM 1-methyl-7-nitroisatoic anhydride in DMSO [1M7; (+) sample] and incubating for 3 min at 22 °C. The RNA was then extracted and precipitated as described above. The resulting pellets were also dissolved in 10 μL of 10% DMSO.

9.6.2 RNA Linker ligation

Linker ligation and all further steps used to prepare sequencing libraries have been extensively described in Watters *et al.* [123] for SHAPE-Seq v2.1. Therefore, only a brief description is provided here.

RNA linker was prepared by adenylating the oligonucleotide 5'-Phos-CUGACU CGGGCACCAAGGA-ddC-3' using the DNA adenylation kit sold by New England Biolabs (NEB). The adenylated linker was then ligated to the modified and unmodified RNA samples by adding 0.5 μL of SuperaseIN (Life Technologies), 6 μL 50% PEG 8000, 2 μL 10X T4 RNA Ligase Buffer (NEB), 1 μL of 2 μM 5'-adenylated RNA linker, and 0.5 μL T4 RNA Ligase, truncated KQ (200 U/μL; NEB) to the RNA mixture and incubated overnight (>10 hrs) at room temperature.

9.6.3 Reverse transcription

Upon completion of the RNA ligation reaction, the RNA was ethanol precipitated and dissolved in 10 μL RNase-free H_2O. Next, 3 μL of 0.5 μM reverse transcription primer (5'-Biotin-GTCCTTGGTGCCCGAGT-3') were added. The resulting mix was then denatured completely by heating to 95°C for 2 min, followed by an incubation at 65°C for 5 min before placing on ice for ~30 seconds before addition of 7 μL of SSIII master mix, containing: 0.5 μL of Superscript III (Life Technologies), 4 μL 5X First Strand Buffer (Life Technologies), 1 μL 100 mM (DTT), 1 μL 10 mM dNTPs, and 0.5 μL RNase-free H_2O. The reaction mix was further incubated at 42 °C for 1 min, then 52 °C for 25 min and deactivated by heating at 65 °C for 5 min. The RNA was then hydrolyzed by the addition of 1 μL of 4 M NaOH solution and heating at 95 °C for 5 min. The basic solution containing the cDNA was partially neutralized, precipitated, and dissolved in 22.5 μL of nuclease-free H_2O.

9.6.4 DNA adapter ligation

To the cDNA, 3 μL 10X CircLigase Buffer (Epicentre), 1.5 μL 50 mM $MnCl_2$, 1.5 μL 1 mM ATP, 0.5 μL 100 μM DNA adapter (5'-Phos-AGATCGGAAGAGCACACGTC TGAACTCCAGTCAC-3CSpacer-3'), and 1 μL CircLigase I (Epicentre) were added. The reaction was incubated at 60 °C for 2 hrs, then 80 °C for 10 min to inactivate the ligase. The ligated DNA was precipitated with ethanol, dissolved in 20 μL of nuclease-free H_2O, purified using 36 μL of Agencourt XP beads (Beckman Coulter; according to manufacturer's instructions), and eluted with 20 μL TE buffer.

9.6.5 Quality analysis

For quality analysis (QA), a separate PCR reaction for each (+) and (-) sample was mixed as described in Watters *et al.* [123] to fluorescently label the PCR amplicons with the VIC and NED (ABI) fluorophores, respectively. Amplification was performed for 15 cycles, then 50 μL nuclease-free H_2O was added and the diluted reaction was ethanol precipitated. The resulting pellet was dissolved in formamide and analyzed with an ABI 3730xl capillary electrophoresis device at the Cornell Biotechnology Resource Center.

9.6.6 Library preparation and next-generation sequencing

To construct sequencing libraries, separate PCRs for each (+) and (-) sample was mixed using the three oligonucleotide strategy according to the protocol in Watters *et al.* [123]. Amplification was performed as indicated in 'Quality analysis' above. Completed reactions were digested with 5 U exonuclease I (NEB) at 37 °C for 30 min to remove excess primer. After incubation, the libraries were purified using 90 μL of Agencourt XP beads (Beckman Coulter), eluted with 20 μL TE buffer, and quantified with the Qubit 2.0 Fluorometer (Life Technologies).

To prepare the libraries for sequencing, the average length of each sample was determined using the results from the quality analysis in order to calculate the molarity of each (+) or (-) separately. Sequencing pools were mixed to be equimolar, such that all of the sequencing libraries were present in the solution at the same level. Sequencing was performed on the Illumina MiSeq using 2x35 bp paired end reads.

9.6.7 Data analysis with Spats

Reads analysis was performed with Spats v1.0.0 (`https://github.com/LucksLab/spats/releases/`), using cutadapt v1.5 [267] and Bowtie 0.12.8 [182] to remove the sequencing adapters as part of the Spats pre-processing steps. Unique mapping of each paired-end read to the targets file generated a reactivity value θ_i, representing the probability of modification at nucleotide i relative to the rest of the nucleotides in the RNA. The θ_i values were normalized to $\rho_i$values according to previous SHAPE-Seq work (Chapter 4) [113, 122, 123] such that the average value of ρ_i across an entire RNA molecule is one. The normalization to ρ_i values also allowed for meaningful comparison between RNAs of different lengths.

CHAPTER 10

CONCLUSIONS AND PERSPECTIVES

10.1 Conclusion

The work detailed in the previous chapters has resulted in significant advances in the SHAPE-Seq chemical probing technique. Along with other contributors, I not only developed a method to accurately and repeatedly measure RNA structures inside living cells, but we also made significant improvements to the *in vitro* SHAPE-Seq protocol. Additionally, we created a new protocol to measure RNA cotranscriptional folding events *in vitro* with unprecedented resolution and simplicity.

I then applied these new techniques to study the RNA structure-function relationship in many new contexts that were previously unexplored. Our examination of the structure and function of two synthetic translational regulators in *E. coli* cells was the first of its kind, directly linking a functional outcome to changes at the structural level. Similarly, my contributors and I proposed mechanisms for how cotranscriptional folding of two transcriptionally acting RNA regulators affected the final functional decisions of the regulators. Our work demonstrates the importance of RNA cotranscriptional folding; an element of RNA folding that has been underappreciated in RNA biology for some time.

The work describing RNA structures in the CMV genome was also pioneering and represents the first application of chemical probing techniques to plant cell lysates. Probing directly in cell lysates allowed us to note differences between *in vitro* and cellular probing conditions, leading to our observation of replicase binding in the tRNA-like structure of CMV. I also saw structural differences between RNAs probed *in vitro*

and in *E. coli* cells. Our comparisons add to the recent drive in the field to determine what the effects of the cellular folding environment are on RNA structure.

Last, I studied RNA-protein interactions in the contexts of PRORP and the RNase P ribozyme as well as sgRNA:dCas9 binding in the Type II CRISPR system. For RNase P, clear drops in reactivity showed that the L7Ae protein cofactor binds to kink-turn motifs in the ribozyme and that PRORP binds to the D-loop of pre-tRNAs. dCas9 binding triggered reactivity changes across the entire sgRNA, interpreted with the help of published crystal structures that the SHAPE-Seq data agreed well with. My study of the sgRNA:dCas9 complex also aided engineering efforts to create a new type of transcriptional control, CRISPRi derepression, that could potentially eliminate the problems of high background noise observed in many types of RNA regulators.

Altogether, I expect that all of the SHAPE-Seq improvements presented in this work will greatly assist future researchers to understand, discover, and engineer the structure-function relationship of many more RNAs, including those that have yet to be found.

10.2 Future Directions and Perspectives

10.2.1 Genome-wide RNA structure probing techniques

In the last few years, there have been a plethora of *in vivo* NGS-based structure probing techniques that have been developed. During in-cell SHAPE-Seq development, a string of NGS-based techniques using dimethyl sulfate (DMS) were developed for transcriptome level studies, as well as one SHAPE-based technique [109]. However, many

RNA biologists are unsure what to do with transcriptome-level data, especially given the tendency of less abundant RNAs to be poorly covered. Further, DMS only reacts with adenosine or cytidine. For the field to advance, a better measure of transcriptome coverage is a must to ensure confidence in probing accuracy, especially where poor coverage of rare RNAs is concerned.

The utility of transcriptome-wide studies is generally questionable for studying a few RNAs at a time, which is the reason that in-cell SHAPE-Seq was not originally developed as a transcriptome-wide technique. Rather, they are best suited for database-style data collection. Currently, structure probing information is fairly scattered in the literature, and the few databases that exist are quite small [169, 411], although they are slowly growing in size. The RNA community as a whole would greatly benefit from a sequence-and-store approach akin to efforts underway with whole genome sequencing and RNA-seq. One particularly interesting application of this sequence-and-store approach relates to the fields of cancer and genetic disease. Comparing the structurome of many normal and 'disease' cell lines may reveal new causes of cancer or genetic disease that are caused by RNA structural defects. Such an idea is promising, but would require mobilization of and collaboration in the structure probing field that does not currently exist.

Last, in terms of SHAPE-Seq itself, movement to a transcriptome-wide style protocol would be fairly straightforward and is discussed in Section 4.7.1 of Chapter 4. In short, the only change that is required is the method in which cDNA is generated, as the techniques discussed in this work are focused on one or a few RNAs at a time. One of two approaches to reverse transcribe total RNA would suffice. The first is to use random priming with some sequence bias corrections as done in Siegfried *et al.* [112]. The second is to fragment the total RNA and ligate to the 3' end of each fragment using

SHAPE-Seq v2.1. Advantages of random priming include a simpler protocol and data analysis, but uneven transcriptome coverage (due to primer binding bias) and a lack of PCR selection present drawbacks. Conversely, fragmentation/ligation has the advantage of PCR selection and limited sequence bias, but requires extra steps and analysis software updates to remove the linker from sequencing reads, a difficult task.

10.2.2 RNA folding dynamics

Despite having been studied for over 30 years, the dynamics of RNA folding remains a poorly understood topic in biology. The difficulty in understanding the process of cotranscriptional folding mainly stems from the difficulty in capturing structural intermediates during the process of transcription, as there are currently no techniques that can do so. The best we can currently do to observe the folding process outside of transcription is time-resolved NMR [412]. Unfortunately, current NMR-based techniques cannot capture cotranscriptional events due to limitations in the number of atoms in the system (due to the presence of RNAP).

Another way of measuring RNA folding as it occurs is using time-resolved Förster energy transfer (FRET) [413], which can be used to measure the distance between two points in an RNA over time. However, FRET pairs cannot be incorporated cotranscriptionally, although it would be technically possible to covalently add a FRET pair to a halted elongation complex and then restart transcription to measure structural changes. Because of the need for modified nucleotides to add the labels, however, this would only be reasonably possible for an elongation complex that was assembled with pre-transcribed RNA, somewhat defeating the purpose of measuring cotranscriptional folding. Thus, only techniques remaining for studying cotranscriptional folding are

chemical probing or single-molecule pulling experiments.

Our development of cotranscriptional SHAPE-Seq represents a huge step forward in our ability to measure cotranscriptionally folded RNA structures. One of the biggest benefits to the technique is the simplicity of the experiment, as it only requires standard equipment that can be found in any biochemistry lab. Therefore, I suspect that the cotranscriptional SHAPE-Seq approach will quickly become a cornerstone technique for researchers interested in cotranscriptional folding. There are also many immediate applications of the technique to understanding RNA regulators, especially riboswitches and termination/antitermination mechanisms. One particularly interesting example would be measurement of tandem riboswitches [414, 415]. Cooperativity between the aptamers has been reported [416], but the reason tandem aptamers evolved rather than an improved single aptamer is not understood. Application of the cotranscriptional SHAPE-Seq technique may help alleviate some of the confusion associated with these riboswitches. It would also be interesting to see cotranscriptional SHAPE-Seq applied to the classically studied RNA models used to examine cotranscriptional folding: the Tetrahymena ribozyme and RNase P.

10.2.3 Improvements in RNA structural prediction

A frequent use of structural probing data is to try to improve computational predictions of RNA structure. A number of secondary structure predictions algorithms exist that can incorporate reactivity data and have been repeatedly shown to increase prediction accuracy [123]. However, a number of shortcomings exist that hinder the utility of these secondary structural predictors. First, there are no energy rules associated with noncanonical base pairs as there are for canonical RNA base pairs [417, 418].

Thus, noncanonical base pairing is frequently ignored in secondary structure predictions, although it occurs at a fairly high frequency in nature. Second, many algorithms are unable to predict pseudoknots, and those that do are limited to one [102]. Third, kinetic elements of RNA folding are poorly modeled. A number of computational cotranscriptional folding models have been devised [65], but they are still under heavy development and are not well supported by experiment.

RNA secondary structure prediction algorithms are still useful, however, and provide a good starting point for deeper investigation. Research into 3D structure prediction is making headway as well, but the computational requirements are intense and more involved than for protein structures, limiting the approachability of 3D prediction to the greater community. Computational time also limits pseudoknot discovery as RNA sequences get longer and more possible pseudoknots must be vetted.

Despite all of the roadblocks to improving computational RNA structure prediction, there are a number of directions the field could go in to begin making progress. Inclusion of Hoogsteen base pairing in the nearest neighbor rules [418] would be the easiest first step toward a more expanded set of noncanonical parameters and would help improve energy models for predicted structures. Also, cotranscriptional structure prediction models would greatly benefit from the incorporation of SHAPE reactivities to restrain folding or at the very least confirm or refute results from current unrestrained models to better improve them. The application of coarse-grained 3D models like oxRNA [266] to the cotranscriptional folding problem could also help improve prediction accuracy and provide more prediction capabilities to the RNA biology community.

10.2.4 Combining structural data and RNA regulator design

With the development of in-cell SHAPE-Seq, my contributors and I opened the door to many possibilities for a combined structure and function approach to designing new synthetic RNA regulators. By combining structural level information and functional consequences from different regulator designs, more insight can be obtained about the rules governing a particular design than by simple mutational analysis alone, as discussed in Chapter 7. However, there is an inherent time cost to performing this type of combined analysis in terms of experimental preparation and sequencing turnaround. For well-established design platforms, or designs expected to have a longer timeframe, the time cost of performing structural analysis is usually well worth it, resulting in insights that greatly speed development. Yet, the same time investment may present a high barrier to quick, fail-fast design ideas.

Thus, automating SHAPE-Seq analysis to limit the amount of time spent collecting structural data that could be instead focused on the next rounds of design would speed up regulator development. As there is little remaining in the in-cell protocol left to optimize, adapting the in-cell SHAPE-Seq technique for use in a microfluidics layout would be highly beneficial. In an ideal scenario, a researcher interested in collecting in-cell SHAPE-Seq data would first measure the functional result of the regulator in question, then perform the in-cell SHAPE-Seq steps automatically in a microfluidics device. Combined with automated cloning, as used by the Voigt lab [419], time spent performing bench experiments could be replaced with data interpretation and faster RNA circuitry design.

10.2.5 (Next-)next-generation sequencing technologies

Nanopore sequencing has been hailed as the potential next generation in low-cost high-throughput sequencing. To sequence a DNA molecule, electrical resistance is measured across each base as it passes through a protein pore. Each base generates a unique resistance, allowing a string of measured resistances to be converted to a sequence. SHAPE-Seq could benefit greatly from nanopore sequencing, as the low complexity issues that plague it are a result of the fluorescence microscopy used in modern sequence-by-synthesis technologies. More interesting perhaps is the possibility of using nanopore sequencing to directly detect modifications in the RNA by reading the modification as a different resistance. However, if DNA conversion were still required, no additional handle sequences should be required for sequencing, meaning that the adapter ligation step would no longer be necessary, eliminating the problems associated with the unwanted ligation side product. Thus, nanopore sequencing could potentially solve both problems of low complexity and side product amplification while reducing the time and cost to perform SHAPE-Seq.

10.3 Improving the SHAPE-Seq Technique

Even with all of the improvements added to the SHAPE-Seq technique in this work, there is still much that can be improved. Below, a number of potential improvements that could be added to the SHAPE-Seq protocol are discussed, as well as a few areas where better characterization of the technique would be beneficial.

10.3.1 General improvements

One of the biggest challenges when using SHAPE-Seq is keeping the amount of unwanted ligation side product to a minimum. The selective PCR protocol introduced in Chapter 3 greatly reduced the amount of side product that appears in the final dsDNA sequencing library, but falls short of completely resolving the issue. Low abundance targets still generate high levels of unwanted side product. Lowering oligonucleotide concentrations helps somewhat, but is a poorly characterized solution. A better solution would be to selectively cleave the dimer products using a nuclease with the capability of resolving a single base mismatch. For example, Cas9 could be used to cleave the dimer product if an oligonucleotide is introduced that is a reverse complement to the unwanted dimer product. The potential drawbacks of this method, however, are that commercially prepared Cas9 is fairly expensive, the technique introduces another oligonucleotide, and a PAM sequence is still required. Alternatively a lower fidelity method could be employed using the same reverse complement oligonucleotide and a thermostable restriction enzyme. However, a restriction enzyme, TaqI, was tested before PCR selection was developed and did not yield satisfactory results.

A second problem that current SHAPE-Seq methods face is the prevalence of low complexity regions in sequencing libraries. These low complexity reads increase sequencing errors and require extra DNA be added to improve the library randomness, which reduces the total sequencing read counts obtained. A general solution to the low complexity problem is to move toward random priming, which is not ideal for measuring a few short RNAs. Alternatively, libraries with variable numbers of random bases between the RRRY and YYYR handles and the RT priming sequence would help stagger individual sequencing reading frames, but would require software upgrades that would be laborious. In the absence of a better solution, a deeper investigation into

the unaligned sequencing reads from different experiments is warranted to determine if upgrades to the Spats pipeline could reduce the number of reads lost to sequencing errors.

Improvements to the Spats pipeline would also greatly benefit those interested in using SHAPE-Seq both in the Lucks Lab and elsewhere. First, removing the dependence of the fastx_tools and boost would greatly simplify the installation process which should be more user-friendly in future versions. Second, as mentioned above, the ability to include sequencing reads with 'harmless' mutations or insertions/deletions in the reactivity calculation would greater reduce the number of sequencing reads required to obtain good reactivity maps. Last, with Illumina's addition of a second set of barcodes, separating the positive and negative channels with internal indexes may no longer be necessary if future priming schemes are revamped. In that case, allowing users to indicate positive and negative channel reads by inputting four fastq files instead of two would shorten the time required to process sequencing data and allow for more custom sequencing layouts.

10.3.2 in-cell SHAPE-Seq improvements

While much of the in-cell SHAPE-Seq protocol has been heavily optimized (Chapter 3), there are still a few avenues worth exploration. First, in-cell SHAPE-Seq has a tendency to exhibit run-to-run variability in terms of obtaining good quality cDNA libraries. The most likely cause is inconsistency in the RNA extraction efficiency from sample-to-sample. Yet, exactly what factors are most important to getting an optimal extraction have not been well explored. While I suspect that the time between modification and the addition of the TRIzol reagent is critical, no concrete evidence has been collected

to support this conclusion. The observation that running more samples at once tends to lower cDNA quality agrees with this hypothesis, but a better characterization of the extraction process would be enlightening.

In-cell SHAPE-Seq would also benefit from a standardized RT primer concentration that is based on the total RNA concentration. As developed, in-cell SHAPE-Seq uses standard oligonucleotide concentrations that assume high quality extractions. Instead, I would propose determining an optimal total RNA/RT primer concentration ratio that is adjusted based on an additionally included step to measure the total RNA concentration. However, an adjustment may be required to correct for differential levels of expression between different RNA species being characterized.

10.3.3 Cotranscriptional improvements

The cotranscriptional SHAPE-Seq protocol that was developed in Chapter 5 is still in its early stages. As currently designed, paused complexes for each intermediate RNA length are generated by exhaustive PCR to create a library of intermediate DNA template lengths that include a roadblock binding site on each end. For shorter RNAs, exhaustive PCR is burdensome, but manageable. However, as the RNA of interest gets longer, the cost of the required oligonucleotides increases as well as the difficulty of ensuring every template length was amplified correctly.

During development of the exhaustive PCR/Gln111 roadblocking scheme, other roadblocking ideas were tested that could be re-explored to reduce the burden of generating the DNA template. Namely, randomly incorporated biotin moieties in the DNA template could serve as flexible roadblocks after the addition of streptavidin and would only require a single PCR to generate a template library, simplifying mutational

studies. Further development of the biotin-streptavidin method is currently underway and appears promising.

The most difficult step in the cotranscriptional SHAPE-Seq protocol involves initiating transcription and modifying the RNA. The step is most easily performed with two people, but inconsistencies in timing, etc. add variability to experiments, especially those where transcription is limited to 15 seconds or less. I propose to instead move toward a stop-flow kinetic apparatus or microfluidics to allow for faster kinetic measurements and more consistent timing of transcription and modification. The main challenge with adopting a microfluidics platform, other than initial setup costs, is that the scale of the cotranscriptional SHAPE-Seq reactions are smaller than a microfluidic device is typically used for. Droplet mixing expertise would be required, but could allow for a great deal of automation, especially if downstream processing steps could be automated as well.

The inherent randomness of chemical modification position and transcript 3′ end position could allow cotranscriptional SHAPE-Seq to avoid problems with low complexity during sequencing. However, the introduction of the linker sequence in every read generates a low complexity region that reduces the number of usable sequencing reads for reactivity calculation. A potential solution to this problem is to alternatively barcode the positive and negative reads (using the proposed software update suggested above) and use a custom sequencing primer that binds the linker region to skip over the low complexity region during sequencing. Solving the low complexity issue for cotranscriptional SHAPE-Seq would provide an immediate boost in good sequencing reads and greatly improve the data quality of the calculated reactivity matrices.

10.4 Final Perspectives

Recent technological advances in RNA structure characterization and computational modeling are changing the way we think about RNA structures and their role in biology. The last five years have witnessed a surge in the popularity of RNA structure probing techniques, fuelled in part by the development of many new methods that capture structures transcriptome-wide. Other advances in the understanding of RNA cotranscriptional folding, led by a combination of single-molecule pulling experiments, chemical probing, and new computational folding approaches, are beginning to shed light on its widespread importance in RNA folding. All of these new techniques promise to change the way we as biologists and engineers study the RNA structure-function relationship so that we may better understand and engineer the immense variety of RNAs that govern life's processes.

APPENDIX A

SUPPORTING INFORMATION FOR SHAPE-SEQ 2.0: SYSTEMATIC OPTIMIZATION AND EXTENSION OF HIGH-THROUGHPUT CHEMICAL PROBING OF RNA SECONDARY STRUCTURE WITH NEXT GENERATION SEQUENCING

A.1 Supplementary Tables

Table A.1: Sequences are listed as DNA sequence for convenient use in Spats. Blue lettering indicates the structure cassette sequence initially developed for SHAPE analysis using capillary electrophoresis [92], or the GG sequence required for T7 RNA polymerase in vitro transcription. Green lettering indicates an RNA-specific barcode used for multiplexing SHAPE-Seq v1.0 experiments [107, 108]. Red lettering indicates reverse transcriptase (RT) priming site. Note that the SHAPE-Seq v2.0 uses an RT priming site that is introduced after linker ligation (Figure A.4), so no explicit RT priming site is present in the RNA, although it is appended to the RNA sequence before Spats analysis.

Name	Sequence	Experiments	Figure
5S rRNA, *E. coli*	GGCCTTCGGGCCAAATGCCTGGCGGCC GTAGCGCGGTGGTCCCACCTGACCCCA TGCCGAACTCAGAAGTGAAACGCCGTA GCGCCGATGGTAGTGTGGGGTCTCCCC ATGCGAGAGTAGGGAACTGCCAGGCAT CCGATCCGCTTCGGCGGATCCAAATAAA TCGGGCTTCGGTCCGGTTC	QuSHAPE, SHAPE-Seq v1.0	Figure 2, Figure 3, Figure 5, SI Figure 5, SI Figure 11, SI Figure 12, SI Figure 14
Adenine riboswitch, *V. vulnificus*	GGCCTTCGGGCCAAACGCTTCATATAAT CCTAATGATATGGTTTGGGAGTTTCTAC CAAGAGCCTTAAACTCTTGATTATGAAGT GCCGATCCGCTTCGGCGGATCCAAACA AATCGGGCTTCGGTCCGGTTC	QuSHAPE, SHAPE-Seq v1.0	Figure 2, Figure 5, SI Figure 5, SI Figure 12
Cyclic di-GMP riboswitch, *V. cholerae*	GGTGTCACGCACAGGGCAAACCATTCG AAAGAGTGGGACGCAAAGCCTCCGGCC TAAACCAGAAGACATGGTAGGTAGCGG GGTTACCGATGGCAAAATGCATACCCGA TCCGCTTCGGCGGATCCAAATCGGGCTT CGGTCCGGTTC	QuSHAPE, SHAPE-Seq v1.0	Figure 5, SI Figure 12

Table A.1 (Continued)

P4-P6, *Tetrahymena* group I intron ribozyme	GGCCTTCGGGCCAAGAATTGCGGGAAA GGGGTCAACAGCCGTTCAGTACCAAGT CTCAGGGGAAACTTTGAGATGGCCTTGC AAAGGGTATGGTAATAAGCTGACGGACA TGGTCCTAACCACGCAGCCAAGTCCTAA GTCAACAGATCTTCTGTTGATATGGATG CAGTTCAAAACCCCGATCCGCTTCGGCG GATCCAATAAAATCGGGCTTCGGTCCGG TTC	QuSHAPE, SHAPE-Seq v1.0	Figure 2, Figure 5, SI Figure 5, SI Figure 12, SI Figure 14
RNAse P, specificity domain, *B. subtilis*	GGTCGTGCCTAGCGAAGTCATAAGCTAG GGCAGTCTTTAGAGGCTGACGGCAGGA AAAAAGCCTACGTCTTCGGATATGGCTG AGTATCCTTGAAAGTGCCACAGTGACGA AGTCTCACTAGAAATGGTGAGAGTGGAA CGCGGTAAACCCCTCGACCGATCCGCT TCGGCGGATCCCTTGAAATCGGGCTTC GGTCCGGTTC	QuSHAPE, SHAPE-Seq v1.0	Figure 2, Figure 5 SI Figure 5, SI Figure 6, SI Figure 12
$tRNA^{phe}$, *E. coli*	GGCCTTCGGGCCAAGCGGATTTAGCTC AGTTGGGAGAGCGCCAGACTGAAGATC TGGAGGTCCTGTGTTCGATCCACAGAAT TCGCACCACCGATCCGCTTCGGCGGAT CCAAAGAAATCGGGCTTCGGTCCGGTTC	QuSHAPE, SHAPE-Seq v1.0	Figure 2, Figure 3, Figure 5, SI Figure 5 SI Figure 6, SI Figure 11, SI Figure 12, SI Figure 14
Hepatitis C virus IRES domain	GGCCTTCGGGCCAACCATGAATCACTCC CCTGTGAGGAACTACTGTCTTCACGCAG AAAGCGTCTAGCCATGGCGTTAGTATGA GTGTCGTGCAGCCTCCAGGACCCCCCC TCCCGGGAGAGCCATAGTGGTCTGCGG AACCGGTGAGTACACCGGAATTGCCAG GACGACCGGGTCCTTTCTTGGATTAACC CGCTCAATGCCTGGAGATTTGGGCGTG CCCCCGCGAGACTGCTAGCCGAGTAGT GTTGGGTCGCGAAAGGCCTTGTGGTAC TGCCTGATAGGGTGCTTGCGAGTGCCC CGGGAGGTCTCGTAGACCGTGCATCAT GAGCACGAATCCTAAACCTCAACCGATC CGCTTCGGCGGATCCAAGCAAATCGGG CTTCGGTCCGGTTC	QuSHAPE, SHAPE-Seq v1.0	Figure 5, SI Figure 12
SAM I riboswitch, *T. tencongensis*	GGCCTTCGGGCCAATTCTTATCAAGAGA AGCAGAGGGACTGGCCCGACGAAGCTT CAGCAACCGGTGTAATGGCGATCAGCC ATGACCAAGGTGCTAAATCCAGCAAGCT CGAACAGCTTGGAAGATAAGAACCGATC CGCTTCGGCGGATCCAACCAAATCGGG CTTCGGTCCGGTTC	QuSHAPE, SHAPE-Seq v1.0	Figure 5, SI Figure 12
TPP riboswitch, *E.*	GGCCTTCGGGCCAAGACTCGGGGTGCC CTTCTGCGTGAAGGCTGAGAAATACCCG	QuSHAPE, SHAPE-Seq v1.0	Figure 5, SI Figure 12

Table A.1 (Continued)

TPP riboswitch, *E. coli*	GGCCTTCGGGCCAAGACTCGGGGTGCC CTTCTGCGTGAAGGCTGAGAAATACCCG TATCACCTGATCTGGATAATGCCAGCGT AGGGAAGTTCCCGATCCGCTTCGGCGG ATCCATAAAAATCGGGCTTCGGTCCGGT TC	QuSHAPE, SHAPE-Seq v1.0	Figure 5, SI Figure 12
5S rRNA, *E. coli*	GGATGCCTGGCGGCCGTAGCGCGGTG GTCCCACCTGACCCCATGCCGAACTCA GAAGTGAAACGCCGTAGCGCCGATGGT AGTGTGGGGTCTCCCCATGCGAGAGTA GGGAACTGCCAGGCAT	SHAPE-Seq v2.0	Figure 5, SI Figure 12
Adenine Riboswitch, *V. vulnificus*	GGACGCTTCATATAATCCTAATGATATG GTTTGGGAGTTTCTACCAAGAGCCTTAA ACTCTTGATTATGAAGTG	SHAPE-Seq v2.0	Figure 5, SI Figure 12
Cyclic di-GMP riboswitch, *V. cholerae*	GGTGTCACGCACAGGGCAAACCATTCG AAAGAGTGGGACGCAAAGCCTCCGGCC TAAACCAGAAGACATGGTAGGTAGCGG GGTTACCGATGGCAAAATGCATAC	SHAPE-Seq v2.0	Figure 5, SI Figure 12
P4-P6, *Tetrahymena* group I intron ribozyme	GGAATTGCGGGAAAGGGGTCAACAGCC GTTCAGTACCAAGTCTCAGGGGAAACTT TGAGATGGCCTTGCAAAGGGTATGGTAA TAAGCTGACGGACATGGTCCTAACCACG CAGCCAAGTCCTAAGTCAACAGATCTTC TGTTGATATGGATGCAGTTCAAAACC	SHAPE-Seq v2.0	Figure 5, SI Figure 12
RNAse P, specificity domain, *B. subtilis*	GGTCGTGCCTAGCGAAGTCATAAGCTAG GGCAGTCTTTAGAGGCTGACGGCAGGA AAAAAGCCTACGTCTTCGGATATGGCTG AGTATCCTTGAAAGTGCCACAGTGACGA AGTCTCACTAGAAATGGTGAGAGTGGAA CGCGGTAAACCCCTCGA	SHAPE-Seq v2.0	Figure 5, SI Figure 12
$tRNA^{phe}$, *E. coli*	GGCGGATTTAGCTCAGTTGGGAGAGCG CCAGACTGAAGATCTGGAGGTCCTGTGT TCGATCCACAGAATTCGCACCA	SHAPE-Seq v2.0	Figure 5, SI Figure 12
Hepatitis C virus IRES domain	GGCCATGAATCACTCCCCTGTGAGGAAC TACTGTCTTCACGCAGAAAGCGTCTAGC CATGGCGTTAGTATGAGTGTCGTGCAGC CTCCAGGACCCCCCCTCCCGGGAGAGC CATAGTGGTCTGCGGAACCGGTGAGTA CACCGGAATTGCCAGGACGACCGGGTC CTTTCTTGGATTAACCCGCTCAATGCCT GGAGATTTGGGCGTGCCCCCGCGAGAC TGCTAGCCGAGTAGTGTTGGGTCGCGA AAGGCCTTGTGGTACTGCCTGATAGGGT GCTTGCGAGTGCCCCGGGAGGTCTCGT AGACCGTGCATCATGAGCACGAATCCTA AACCTCAA	SHAPE-Seq v2.0	Figure 5, SI Figure 12
SAM I riboswitch, *T. tencongensis*	GGTTCTTATCAAGAGAAGCAGAGGGACT GGCCCGACGAAGCTTCAGCAACCGGTG TAATGGCGATCAGCCATGACCAAGGTGC	SHAPE-Seq v2.0	Figure 5, SI Figure 12

Table A.2: RNA folding buffer conditions and ligand concentrations used in replicate experiments.

RNA	Buffer / Ligand	Reference
5S rRNA, *E. coli*	10 mM $MgCl_2$, 100 mM NaCl and 100 mM HEPES (pH 8.0)	[160]
Adenine riboswitch, *V. vulnificus*	10 mM $MgCl_2$, 100 mM NaCl and 100 mM HEPES (pH 8.0), 5 μM Ligand	[160]
Cyclic di-GMP riboswitch, *V. cholerae*	10 mM $MgCl_2$, 100 mM NaCl and 100 mM HEPES (pH 8.0), 5 μM Ligand	[160]
P4-P6, Tetrahymena group I intron ribozyme	10 mM $MgCl_2$, 100 mM NaCl and 100 mM HEPES (pH 8.0)	[160]
RNAse P, specificity domain, *B. subtilis*	10 mM $MgCl_2$, 100 mM NaCl, and 100 mM HEPES (pH 8.0)	[107]
$tRNA^{p}he$, *E. coli*	10 mM $MgCl_2$, 100 mM NaCl and 100 mM HEPES (pH 8.0)	[160]
Hepatitis C virus IRES domain	10 mM $MgCl_2$, 100 mM NaCl and 100 mM HEPES (pH 8.0)	[102]
SAM I riboswitch, *T. tencongensis*	10 mM $MgCl_2$, 100 mM NaCl and 100 mM HEPES (pH 8.0), 5 μM Ligand	[102]
TPP riboswitch, *E. coli*	10 mM $MgCl_2$, 100 mM NaCl and 100 mM HEPES (pH 8.0), 5 μM Ligand	[102]

Table A.3: List of barcoded reverse transcription primers used during the SHAPE-Seq v2.0 library generation. Illumina adapter sequences are in black, the (+/−) handles are in green, internal barcodes are in red, and the RT priming sites are in purple. Illumina TruSeq indexes can be found at: `http://supportres.illumina.com/documents/myillumina/6378de81-c0cc-47d0-9281-724878bb1c30/2012-09-18_illuminacustomersequenceletter.pdf`

RNA	RT Primer Sequences	Illumina Index# (TruSeq)
$tRNA^{phe}$, *E. coli*	/5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTYYYRgtccttggtgcccgagtg /5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTRRRYgtccttggtgcccgagtg	1,2,3
Adenine riboswitch, *V. vulnificus*	/5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTYYYRgtccttggtgcccgagtg /5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTRRRYgtccttggtgcccgagtg	4,5,6
5S rRNA, *E. coli*	/5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTYYYRgtccttggtgcccgagtg /5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTRRRYgtccttggtgcccgagtg	1,2,3
P4-P6, Tetrahymena group I intron ribozyme	/5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTYYYRCCgtccttggtgcccgagtg /5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTRRRYCCgtccttggtgcccgagtg	1,2,3
RNAse P, specificity domain, *B. subtilis*	/5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTYYYRAgtccttggtgcccgagtg /5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTRRRYAgtccttggtgcccgagtg	4,5,6
TPP riboswitch, *E. coli*	/5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTYYYRAgtccttggtgcccgagtg /5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTRRRYAgtccttggtgcccgagtg	4,5,6
SAM I riboswitch, *T. tencongensis*	/5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTYYYRAgtccttggtgcccgagtg /5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTRRRYAgtccttggtgcccgagtg	10,11,12
Cyclic di-GMP riboswitch, *V. cholerae*	/5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTYYYRCCgtccttggtgcccgagtg /5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTRRRYCCgtccttggtgcccgagtg	7,8,9

Table A.3 (Continued)

RNA	RT Primer Sequences	Illumina Index# (TruSeq)
Hepatitis C virus IRES domain	/5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTYYYRCCgtccttggtgcccgagtg /5Biosg/CTTTCCCTACACGACGCTCTTCCGATCTRRRYCCgtccttggtgcccgagtg	10,11,12

Table A.4: RNA structure prediction accuracy using the RNAstructure [164] Fold algorithm with no SHAPE-Seq constraints.

RNA	Sensitivity	PPV
5S rRNA	10/35 = 28.6%	10/40 = 25.0%
Adenine Riboswitch	21/21 = 100%	21/21 = 100%
Cyclic di-GMP riboswitch	21/28 = 75.0%	21/27 = 77.8%
P4-P6, Tetrahymena group I intron ribozyme	46/48 = 95.8%	47/55 = 85.5%
RNAse P	23/42 = 54.8%	23/46 = 50.0%
$tRNA^{phe}$	20/21 = 95.2%	20/20 = 100%
Hepatitis C IRES	41/104 = 39.4%	41/108 = 38.0%
SAM I riboswitch	29/39 = 74.4%	29/36 = 80.6%
TPP riboswitch	17/22 = 77.3%	17/20 = 85.0%
Total	228/360 = 63.3%	229/373 = 61.4%

Table A.5: RNA structure prediction accuracy using SHAPE-Seq v2.0 reactivity data (ρ's) as constraints in the RNAstructure [164] Fold algorithm with m = 1.8 and b = -0.6.

RNA	Sensitivity	PPV
5S rRNA	34/35 = 97.1%	34/37 = 91.9%
Adenine Riboswitch	21/21 = 100%	21/21 = 100%
Cyclic di-GMP riboswitch	19/28 = 67.9%	19/24 = 79.2%
P4-P6, Tetrahymena group I intron ribozyme	45/48 = 93.8%	46/51 = 90.2%
RNAse P	22/42 = 52.4%	22/44 = 50.0%
$tRNA^{phe}$	21/21 = 100%	21/21 = 100%
Hepatitis C IRES	81/104 = 77.9%	81/96 = 84.4%
SAM I riboswitch	32/39 = 82.1%	32/37 = 86.5%
TPP riboswitch	17/22 = 77.3%	17/20 = 85.0%
Total	292/360 = 81.1%	293/351 = 83.5%

Table A.6: RNA structure prediction accuracy using SHAPE-Seq v2.0 reactivity data (ρ's) as constraints in the RNAstructure [164] Fold algorithm with m = 1.1 and b = -0.3.

RNA	Sensitivity	PPV
5S rRNA	34/35 = 97.1%	34/37 = 91.9%
Adenine Riboswitch	21/21 = 100%	21/21 = 100%
Cyclic di-GMP riboswitch	21/28 = 75.0%	21/28 = 75.02%
P4-P6, Tetrahymena group I intron ribozyme	44/48 = 91.7%	45/49 = 91.8%
RNAse P	33/42 = 78.6%	33/40 = 82.5%
$tRNA^{phe}$	20/21 = 95.2%	20/20 = 100%
Hepatitis C IRES	81/104 = 77.9%	81/90 = 90.0%
SAM I riboswitch	32/39 = 82.1%	32/37 = 86.5%
TPP riboswitch	17/22 = 77.3%	17/20 = 85.0%
Total	303/360 = 84.2%	304/342 = 88.9%

Table A.7: Data deposition table. SHAPE-Seq reactivity spectra generated in this work are freely available from the RNA Mapping Database (RMDB) [169] (`http://rmdb.stanford.edu/repository/`), accessible using the RMDB ID numbers indicated below.

Name	Library Type	RMDB ID	Figure
5S rRNA, *E. coli*	SHAPE-Seq v1.0	5SSC_1M7_0001	Figure 2.2, Figure A.5
5S rRNA, *E. coli*	SHAPE-Seq v1.0 Minimal	5SSC_1M7_0002	Figure 2.2
5S rRNA, *E. coli*	SHAPE-Seq v1.0 Inverted	5SSC_1M7_0003	Figure 2.2
5S rRNA, *E. coli*	SHAPE-Seq v1.0 Indexed Replicates 1-3	5SSC_1M7_0004, 5SSC_1M7_0005, 5SSC_1M7_0006	Figure 2.3, Figure 2.5, Figure A.11, Figure A.12
5S rRNA, *E. coli*	SHAPE-Seq v2.0 Indexed Replicates 1-3	5SRRNA_1M7_0001, 5SRRNA_1M7_0002, 5SRRNA_1M7_0003	Figure 2.5, Figure A.12, Figure A.14
Adenine riboswitch, *V. vulnificus*	SHAPE-Seq v1.0	ADDSC_1M7_0001	Figure 2.2, Figure A.5
Adenine riboswitch, *V. vulnificus*	SHAPE-Seq v1.0 Minimal	ADDSC_1M7_0002	Figure 2.2
Adenine riboswitch, *V. vulnificus*	SHAPE-Seq v1.0 Inverted	ADDSC_1M7_0003	Figure 2.2
Adenine riboswitch, *V. vulnificus*	SHAPE-Seq v1.0 Indexed Replicates 1-3	ADDSC_1M7_0004 ADDSC_1M7_0005, ADDSC_1M7_0006	Figure 2.5, Figure A.12
Adenine riboswitch, *V. vulnificus*	SHAPE-Seq v2.0 Indexed Replicates 1-3	ADDRSW_1M7_0001, ADDRSW_1M7_0002, ADDRSW_1M7_0003	Figure 2.5, Figure A.12
Cyclic di-GMP riboswitch, *V. cholerae*	SHAPE-Seq v1.0 Indexed Replicates 1-3	GMPSC_1M7_0001 GMPSC_1M7_0002, GMPSC_1M7_0003	Figure 2.5, Figure A.12
Cyclic di-GMP riboswitch, *V. cholerae*	SHAPE-Seq v2.0 Indexed Replicates 1-3	CIDGMP_1M7_0001, CIDGMP_1M7_0002, CIDGMP_1M7_0003	Figure 2.5, Figure A.12
P4-P6, Tetrahymena group I intron ribozyme	SHAPE-Seq v1.0	TRIBSC_1M7_0001	Figure 2.2, Figure A.5
P4-P6, Tetrahymena group I intron ribozyme	SHAPE-Seq v1.0 Minimal	TRIBSC_1M7_0002	Figure 2.2

Table A.7 (Continued)

Name	Library Type	RMDB ID	Figure
P4-P6, Tetrahymena group I intron ribozyme	SHAPE-Seq v1.0 Inverted	TRIBSC_1M7_0003	Figure 2.2
P4-P6, Tetrahymena group I intron ribozyme	SHAPE-Seq v1.0 Indexed Replicates 1-3	TRIBSC_1M7_0004 TRIBSC_1M7_0005, TRIBSC_1M7_0006	Figure 2.5, Figure A.12
P4-P6, Tetrahymena group I intron ribozyme	SHAPE-Seq v2.0 Indexed Replicates 1-3	TRP4P6_1M7_0001, TRP4P6_1M7_0002, TRP4P6_1M7_0003	Figure 2.5, Figure A.12, Figure A.14
RNAse P, specificity domain, *B. subtilis*	SHAPE-Seq v1.0	RNPSC_1M7_0001	Figure 2.2, Figure A.5, Figure A.6
RNAse P, specificity domain, *B. subtilis*	SHAPE-Seq v1.0 Minimal	RNPSC_1M7_0002	Figure 2.2, Figure A.6
RNAse P, specificity domain, *B. subtilis*	SHAPE-Seq v1.0 Inverted	RNPSC_1M7_0003	Figure 2.2, Figure A.6
RNAse P, specificity domain, *B. subtilis*	SHAPE-Seq v1.0 Indexed Replicates 1-3	RNPSC_1M7_0004, RNPSC_1M7_0005, RNPSC_1M7_0006	Figure 2.5, Figure A.12
RNAse P, specificity domain, *B. subtilis*	SHAPE-Seq v2.0 Indexed Replicates 1-3	RNASEP_1M7_0001, RNASEP_1M7_0002, RNASEP_1M7_0003	Figure 2.5, Figure A.12
$tRNA^{phe}$, *E. coli*	SHAPE-Seq v1.0	TRNASC_1M7_0001	Figure 2.2, Figure A.5, Figure A.6
$tRNA^{phe}$, *E. coli*	SHAPE-Seq v1.0 Minimal	TRNASC_1M7_0002	Figure 2.2, Figure A.6
$tRNA^{phe}$, *E. coli*	SHAPE-Seq v1.0 Inverted	TRNASC_1M7_0003	Figure 2.2, Figure A.6
$tRNA^{phe}$, *E. coli*	SHAPE-Seq v1.0 Indexed Replicates 1-3	TRNASC_1M7_0004, TRNASC_1M7_0005, TRNASC_1M7_0006	Figure 2.5, Figure A.12
$tRNA^{phe}$, *E. coli*	SHAPE-Seq v2.0 Indexed Replicates 1-3	TRNAPH_1M7_0001, TRNAPH_1M7_0002, TRNAPH_1M7_0003	Figure 2.3, Figure 2.5, Figure A.11, Figure A.12, Figure A.14
Hepatitis C virus IRES domain	SHAPE-Seq v1.0 Indexed Replicates 1-3	HEPCSC_1M7_0001, HEPCSC_1M7_0002, HEPCSC_1M7_0003	Figure 2.5, Figure A.12

Table A.7 (Continued)

Name	Library Type	RMDB ID	Figure
Hepatitis C virus IRES domain	SHAPE-Seq v2.0 Indexed Replicates 1-3	HCIRES_1M7_0001, HCIRES_1M7_0002, HCIRES_1M7_0003	Figure 2.5, Figure A.12
SAM I riboswitch, *T. tencongensis*	SHAPE-Seq v1.0 Indexed Replicates 1-3	SAMSC_1M7_0001, SAMSC_1M7_0002, SAMSC_1M7_0003	Figure 2.5, Figure A.12
SAM I riboswitch, *T. tencongensis*	SHAPE-Seq v2.0 Indexed Replicates 1-3	SAMRSW_1M7_0001, SAMRSW_1M7_0002, SAMRSW_1M7_0003	Figure 2.5, Figure A.12
TPP riboswitch, *E. coli*	SHAPE-Seq v1.0 Indexed Replicates 1-3	TPPSC_1M7_0001, TPPSC_1M7_0002, TPPSC_1M7_0003	Figure 2.5, Figure A.12
TPP riboswitch, *E. coli*	SHAPE-Seq v2.0 Indexed Replicates 1-3	TPPRSW_1M7_0001, TPPRSW_1M7_0002, TPPRSW_1M7_0003	Figure 2.5, Figure A.12

A.2 Supplementary Figures

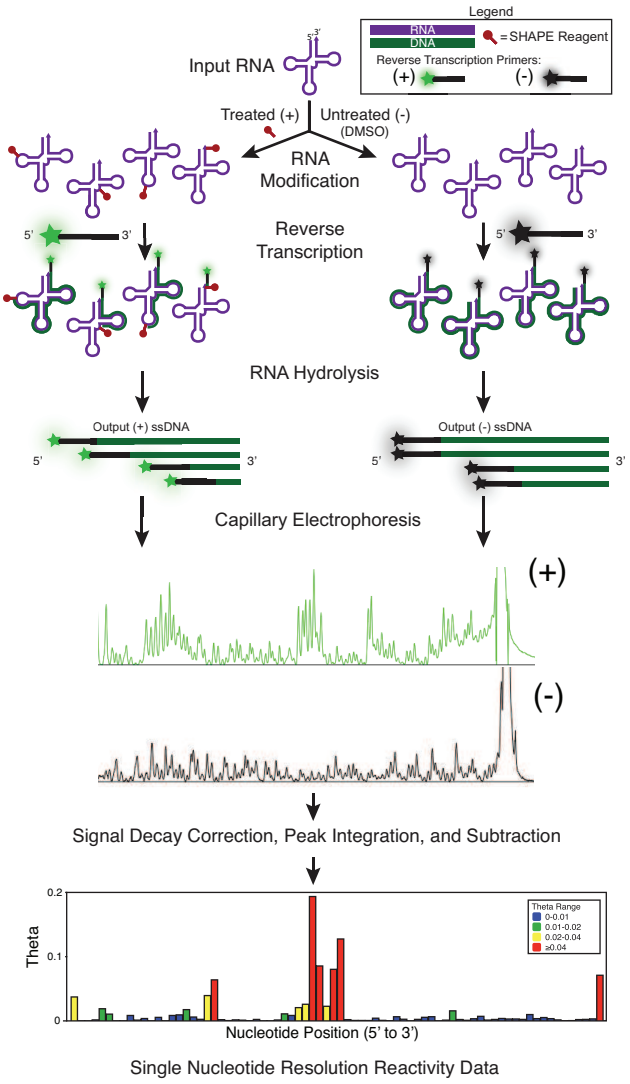
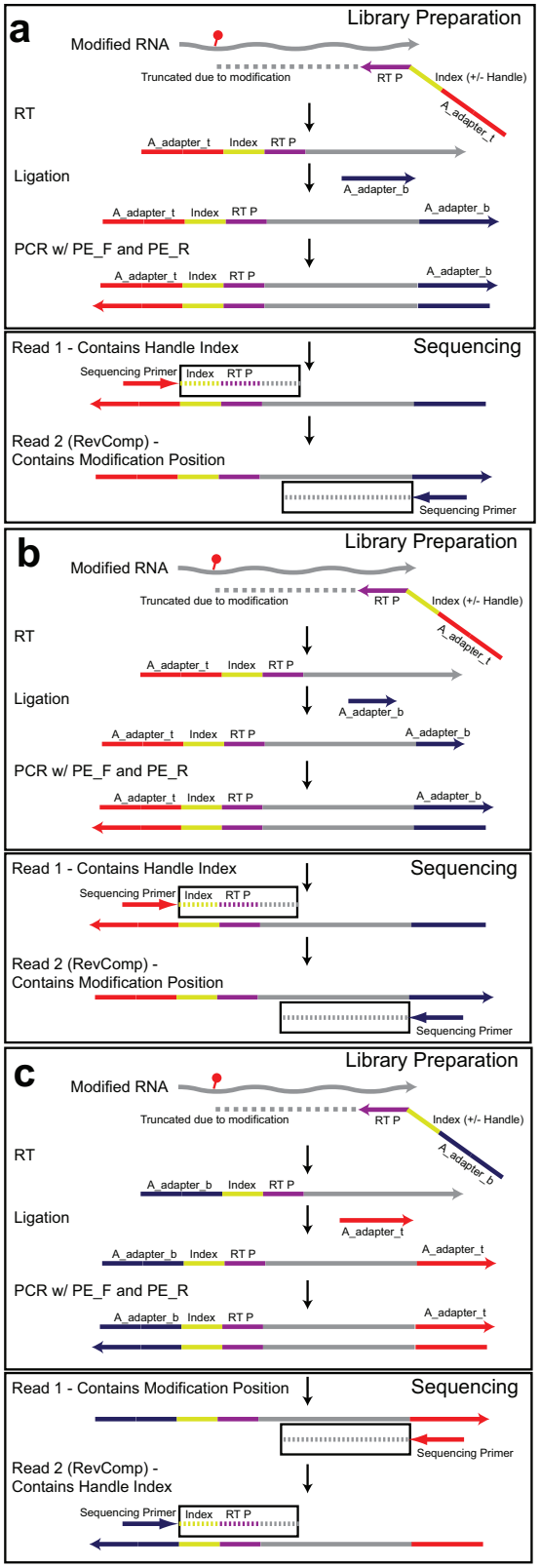


Figure A.1: SHAPE-CE Flowchart. Like SHAPE-Seq, SHAPE-CE begins by modifying RNAs with a SHAPE reagent such as 1M7 [92]. However, unlike SHAPE-Seq, reverse transcription (RT) is performed with fluorescent primers to create a pool of cDNAs (whose length distribution reflects the distribution of modification positions, like SHAPE-Seq). Two different fluorophores distinguish the modified and control reactions, which are detected with capillary electrophoresis (CE). The resulting CE traces are manually integrated and subtracted to obtain a reactivity spectrum for an RNA. An example with *E. coli* $tRNA^{phe}$ is shown. Not shown are additional di-deoxy-terminated sequencing reactions that are used to align CE peaks to the RNA sequence.

Figure A.2: SHAPE-Seq v1.0 and Second Adapter Variation Library Construction Schematics. The left hand side shows adapter and primer orientations for library preparation (top), and which pieces of information are obtained from the library during the sequencing process (bottom), for (a) SHAPE-Seq v1.0, (b) the Minimal adapter configuration, and (c) the Inverted adapter configuration. The right hand side shows DNA sequences of primers and adapters (5′ to 3′ orientation), color-coded to match the schematic.



SHAPE-Seq v1.0

{Seq} = Reverse Complement of Seq

RT P = Reverse Transcriptase Primer

GAACCGGACCGAAGCCCG for priming off of the 3' structure cassette

or

CUSTOM SEQUENCE for priming off of the RNA sequence

Index = (+/-) Handle

(+) RRRY
(-) YYYR

A_adapter_t

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

A_adapter_b

/5Phos/AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG/3SpC3/

PE_F = A_adapter_t

PE_R = {A_adapter_b}

CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

SHAPE-Seq v1.0 Minimal

{Seq} = Reverse Complement of Seq

RT P = Reverse Transcriptase Primer

GAACCGGACCGAAGCCCG for priming off of the 3' structure cassette

or

CUSTOM SEQUENCE for priming off of the RNA sequence

Index = (+/-) Handle

(+) RRRY
(-) YYYR

A_adapter_t

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

A_adapter_b (Minimal)

/5Phos/AGATCGGAAGAGCGGTTCAGCAGGA/3SpC3/

PE_F = A_adapter_t

PE_R = {Full A_adapter_b from SHAPE-Seq v1.0}

CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

SHAPE-Seq v1.0 Inverted

{Seq} = Reverse Complement of Seq

RT P = Reverse Transcriptase Primer

GAACCGGACCGAAGCCCG for priming off of the 3' structure cassette

or

CUSTOM SEQUENCE for priming off of the RNA sequence

Index = (+/-) Handle

(+) RRRY
(-) YYYR

A_adapter_b (Inverted)

CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

A_adapter_t (Inverted)

/5Phos/AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT/3SpC3/

PE_F = A_adapter_b (without modifications)

PE_R = {A_adapter_t}

CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

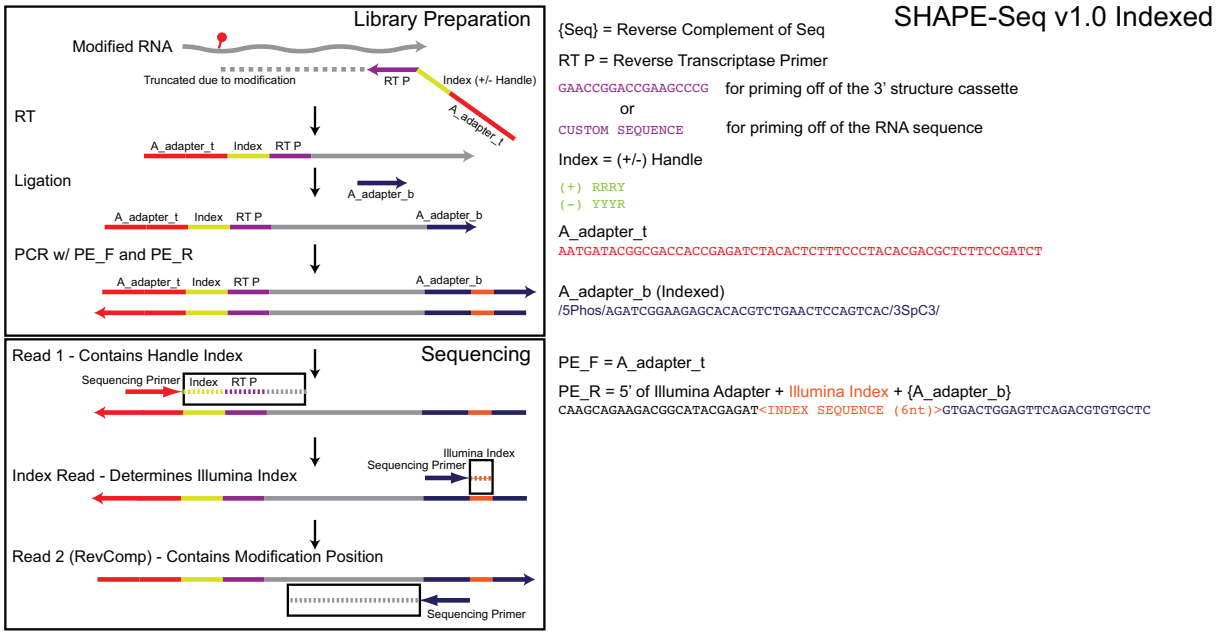
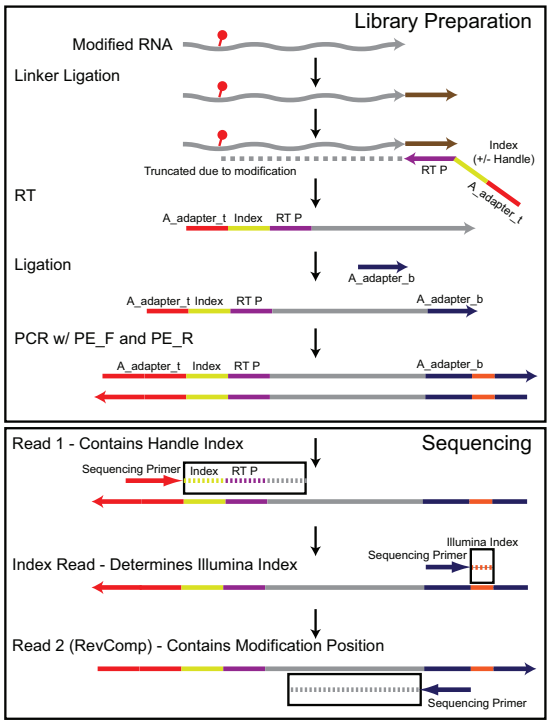


Figure A.3: SHAPE-Seq v1.0 library indexing strategy. The left hand side shows adapter and primer orientations for library preparation (top), and which pieces of information are obtained from the library during the sequencing process (bottom) for the SHAPE-Seq v1.0 Indexed library preparation strategy (compare to Figure A.2A). The difference between SHAPE-Seq v1.0 and SHAPE-Seq v1.0 Indexed is the presence of Illumina indexing sequences that are added during PCR and sequenced separately during the Index Read. This allows multiple SHAPE-Seq libraries to be sequenced in the same lane following standard Illumina indexing strategies. The right hand side shows DNA sequences of primers and adapters (5′ to 3′ orientation), color-coded to match the schematic.



SHAPE-Seq v2.0

{Seq} = Reverse Complement of Seq

Linker = 5' Adenylated Linker (DNA)
/5rApp/CACTCGGGCACCAAGGA/3ddC/

RT P = Reverse Transcriptase Primer
GTCCTTGGTGCCCGAGTG

Index = (+/-) Handle

(+) RRRY
(-) YYYR

A_adapter_t (v2.0)
CTTTCCCTACACGACGCTCTTCCGATCT

A_adapter_b (Indexed)
/5Phos/AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC/3SpC3/

PE_F = Rest of Illumina Adapter + A_adapter_t
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

PE_R = Rest of Illumina Adapter + Illumina Index + {A_adapter_b}
CAAGCAGAAGACGGCATACGAGAT<INDEX SEQUENCE (6nt)>GTGACTGGAGTTCAGACGTGTGCTC

Figure A.4: SHAPE-Seq v2.0 Library Construction Schematic. The left hand side shows adapter and primer orientations for library preparation (top), and which pieces of information are obtained from the library during the sequencing process (bottom) for the Indexed library preparation strategy. The right hand side shows DNA sequences of primers and adapters (5′ to 3′ orientation), color-coded to match the schematic. SHAPE-Seq v2.0 uses a 'universal' RT priming strategy as well as the standard Illumina library indexing strategy.

Figure A.5: QuSHAPE vs. SHAPE-Seq (v1.0) detailed comparisons. For each RNA, SHAPE-Seq and QuSHAPE θ's are plotted on top, with a zoomed window on portions of the comparison shown on the bottom left. The bottom right shows SHAPE-Seq vs. QuSHAPE θ's plotted as a scatter plot, from which the Pearson' s correlation (R) between the two techniques is calculated. Gray boxes represent regions for which no QuSHAPE data is available. Due to difficulties encountered in the alignment step of the QuSHAPE data analysis pipeline, we often found that a single QuSHAPE experiment yielded only a fraction of the reactivities for an individual RNA. To remedy this, we performed replicate QuSHAPE experiments for each RNA, and calculated the average and standard deviation of the QuSHAPE reactivities for each nucleotide position. These QuSHAPE reactivities were then converted to QuSHAPE θ's by dividing by a normalization factor so that they summed to 1 over the range of nucleotides for which reactivity data was obtained. Overall, there was a strong degree of correlation between the two methods for each of the RNAs. Specifically, the Pearson correlations between θ's for each RNA were: 0.88 ($tRNA^{phe}$) 0.70 (RNase P), 0.62 (ribozyme) and 0.95 (adenine riboswitch aptamer). Only the historically difficult 5S rRNA had a poor correlation (0.34).

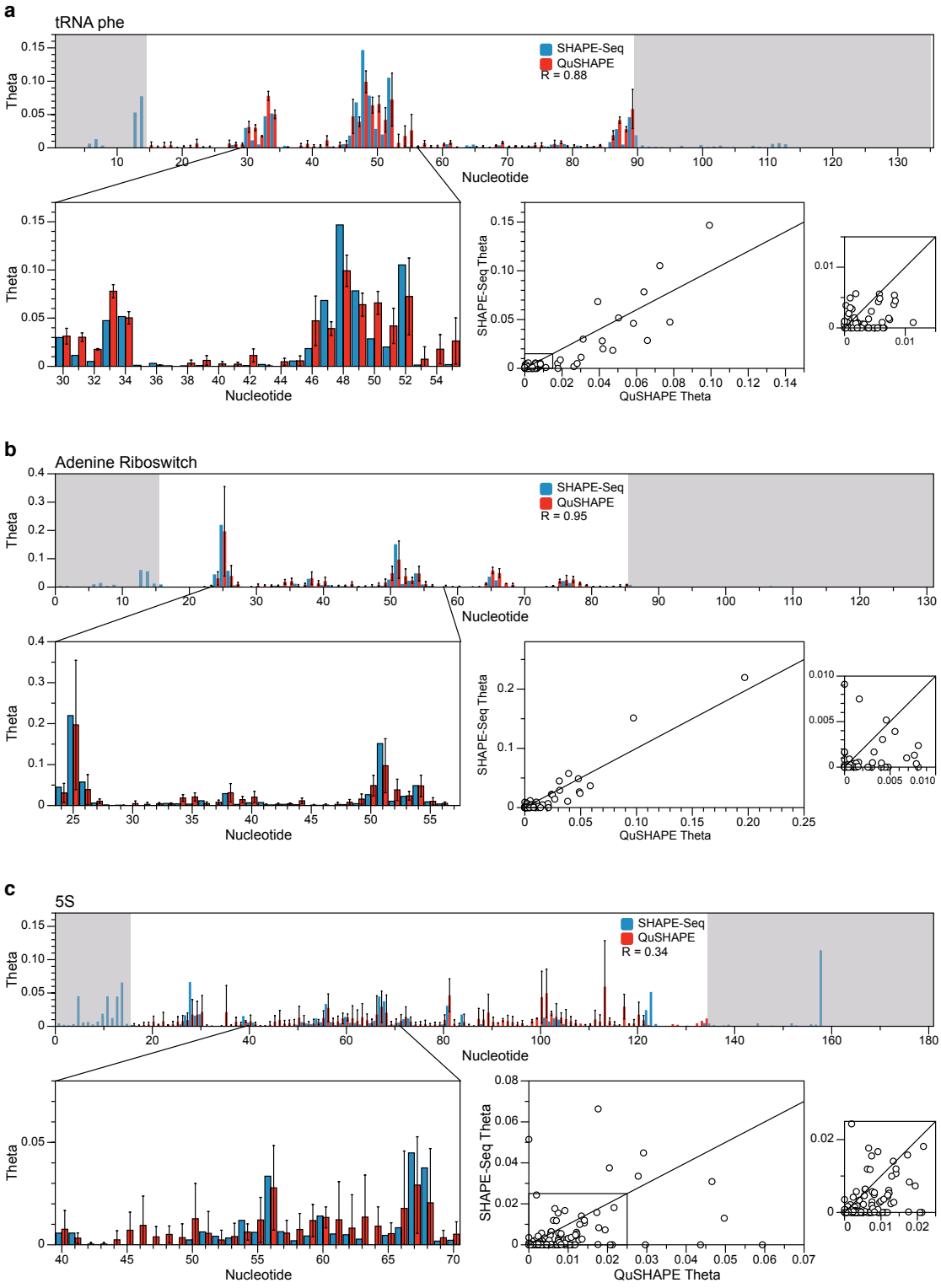
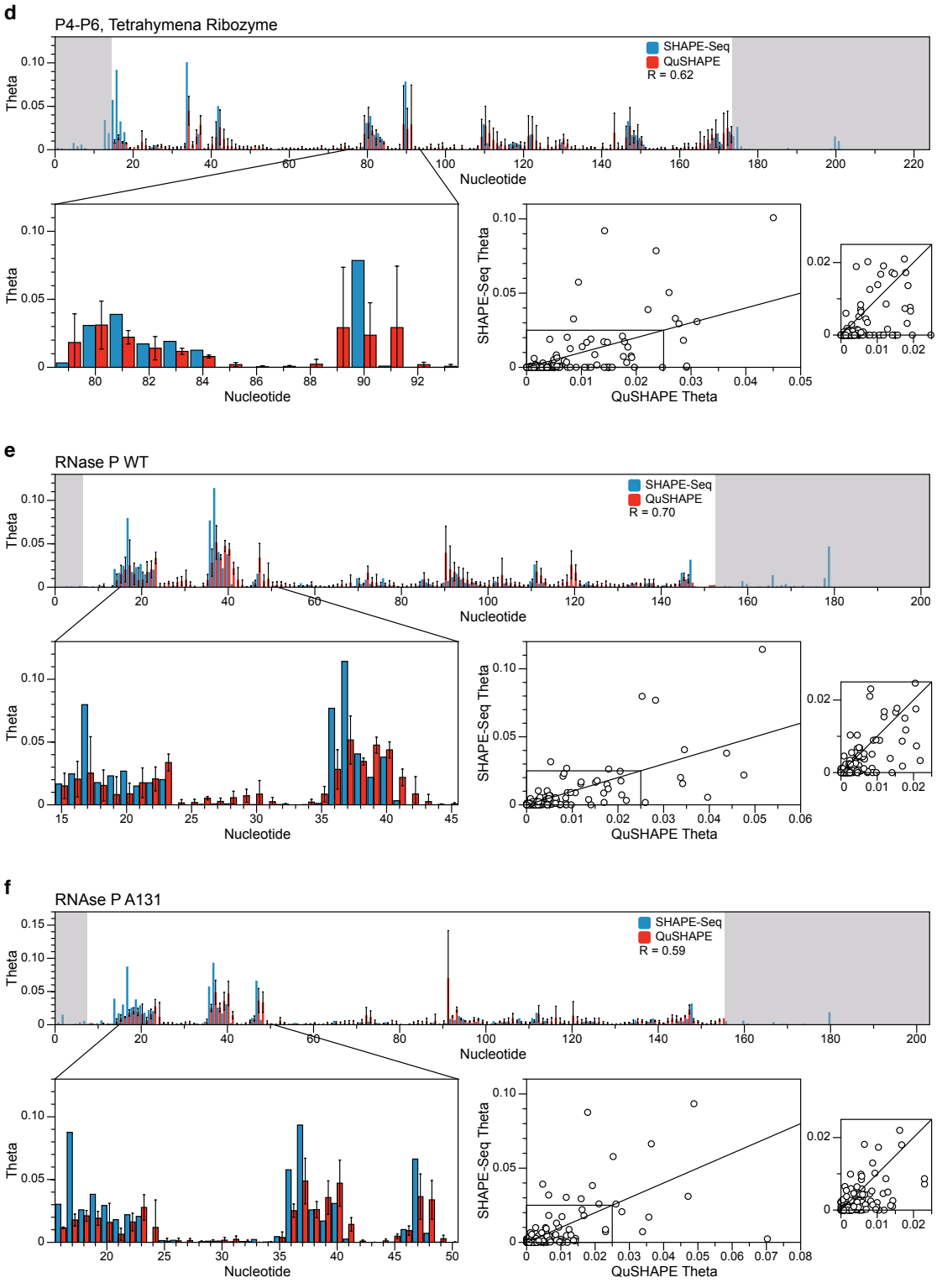

a
tRNA phe
SHAPE-Seq
QuSHAPE
R = 0.88
Theta
Nucleotide
SHAPE-Seq Theta
QuSHAPE Theta
b
Adenine Riboswitch
SHAPE-Seq
QuSHAPE
R = 0.95
Theta
Nucleotide
SHAPE-Seq Theta
QuSHAPE Theta
c
5S
SHAPE-Seq
QuSHAPE
R = 0.34
Theta
Nucleotide
SHAPE-Seq Theta
QuSHAPE Theta

Figure A.5 (Continued)



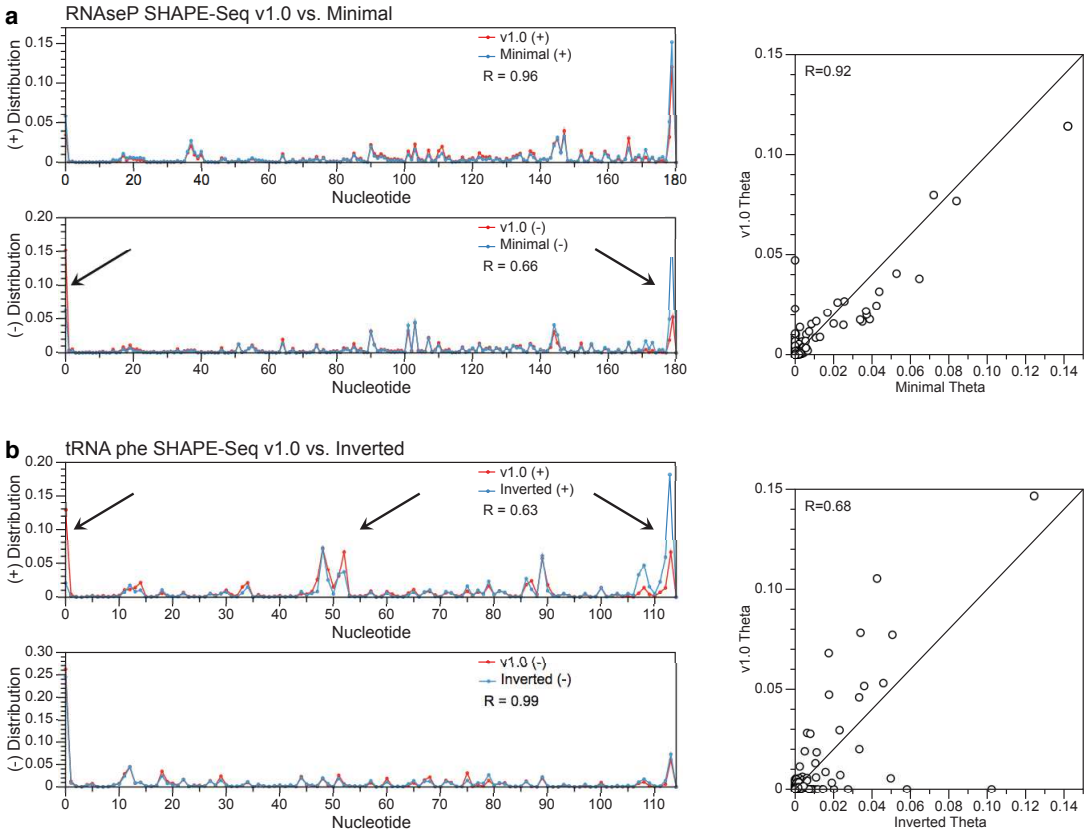


Figure A.6: SHAPE-Seq v1.0 vs. Minimal or Inverted adapter variations for RNase P and tRNA. (+) and (-) fragment distributions are plotted for (a) v1.0 vs. Minimal for RNase P, and for (b) v1.0 vs. Inverted for tRNA. Pearson correlation values for these comparisons are summarized in Figure 2.2. Arrows denote specific places of discrepancy discussed in the text. Scatter plots show θ value comparisons between the two libraries for the RNAs with Pearson correlation values shown in the plots.

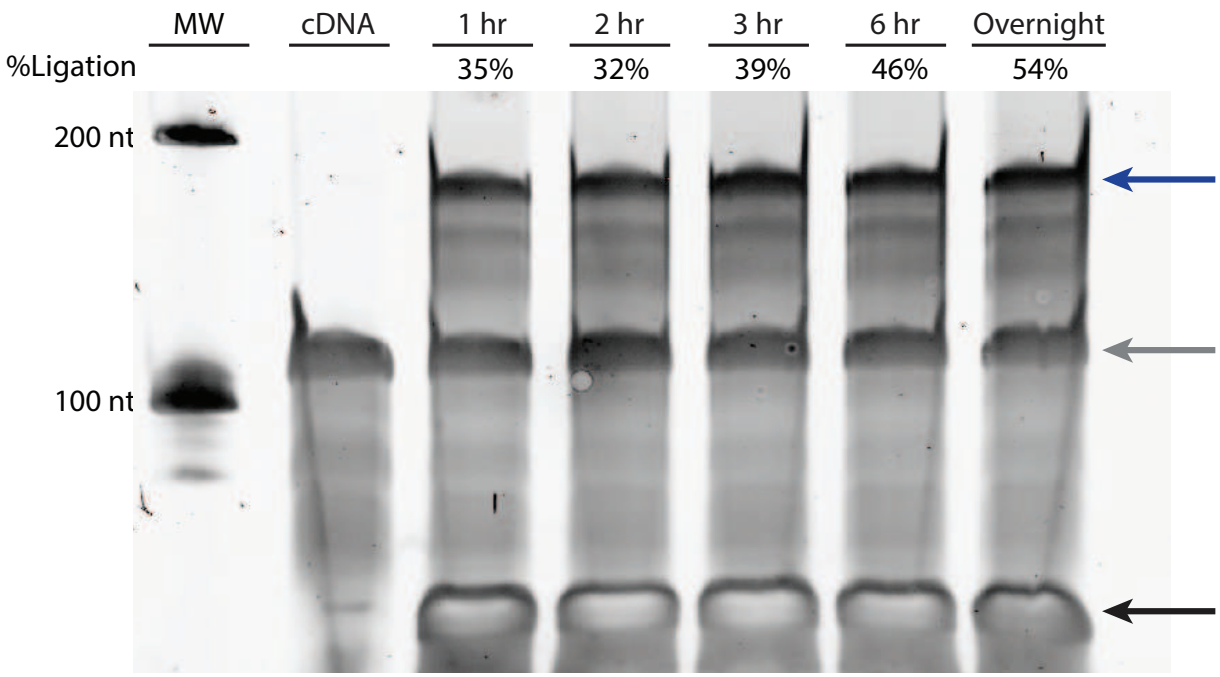


Figure A.7: Time course of CircLigase I ligation efficiency. A 126 nt cDNA (gray arrow) was ligated to the 61 nt SHAPE-Seq v1.0 Illumina adapter (Figure A.1, black arrow) for one, two, three, or six hours or overnight at 68 °C to generate a 187 nt ligation product (blue arrow). Slight ligation improvement is observable over time by integration of the disappearance of the primer band. Note, however, that the gel is stained with SYBR and only provides general trends and is not absolutely quantitative. Ligation was halted after 2 hrs because the improvement gained by further incubation was less important relative to the increase in protocol time.

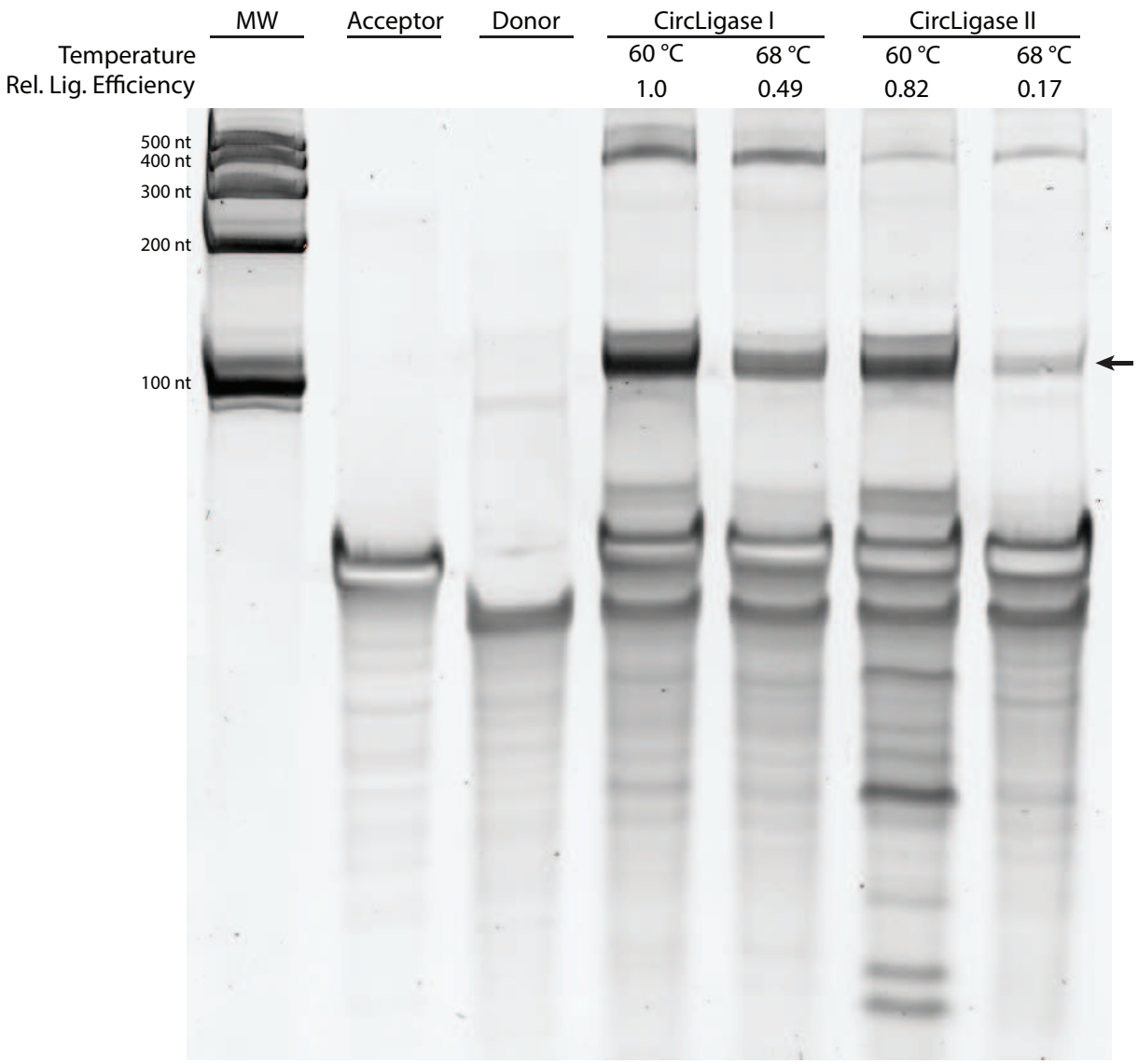


Figure A.8: Ligase comparison for addition of SHAPE-Seq second adapter. CircLigase I and CircLigase II (Epicentre) were used to ligate the 50 nt RT primer (donor) to the 61 nt Illumina adapter (acceptor) at both 68 °C and 60 °C (manufacturer's suggestion) for two hours. The expected ligation product is denoted by the black arrow (101 nt). We found the optimum ligation condition to be CircLigase I at 60 °C for two hours by determining the relative intensity of the ligated product band. Note, however, that the gel is stained with SYBR gold and is not absolutely quantitative.

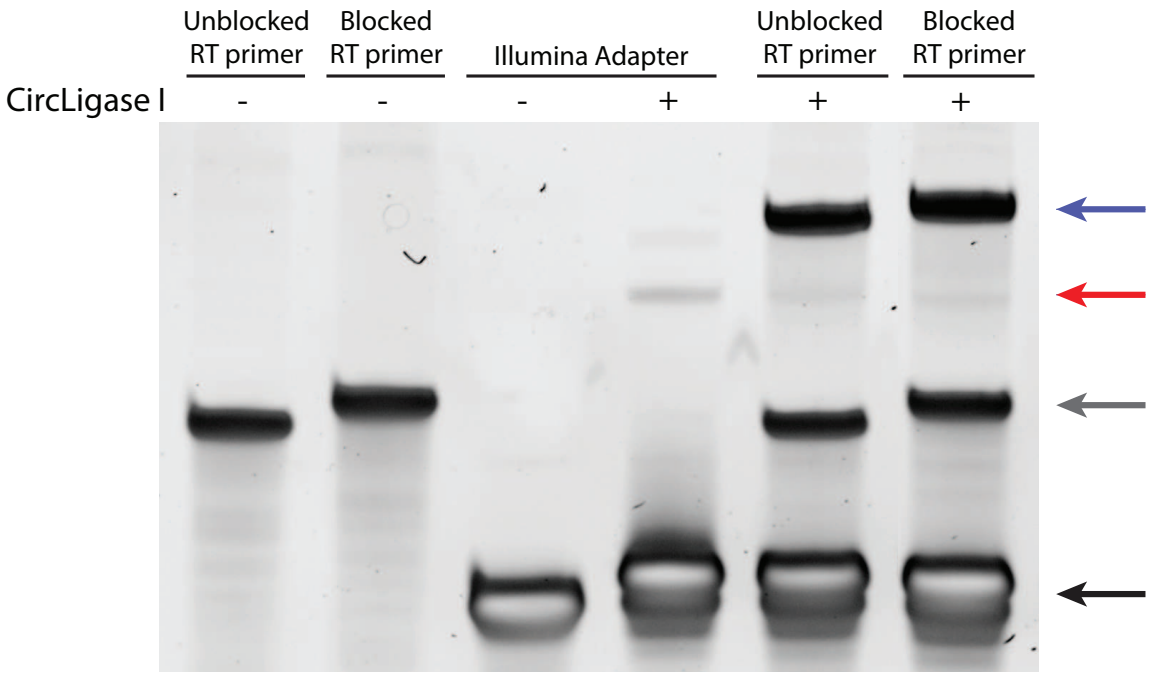


Figure A.9: Modification and optimization of RT primer blocking groups for adapter ligation. Comparison of effect of blocking groups on the possible formation of RT primer concatemers. The top (blue) arrow shows the successfully ligated RT primer (gray arrow) + adapter (black arrow) (60 °C for 2 hours with Circligase I). The red arrow marks the location of adapter dimer formation (studied further below in Figure A.10). No bands corresponding to RT primer concatemers were observed.

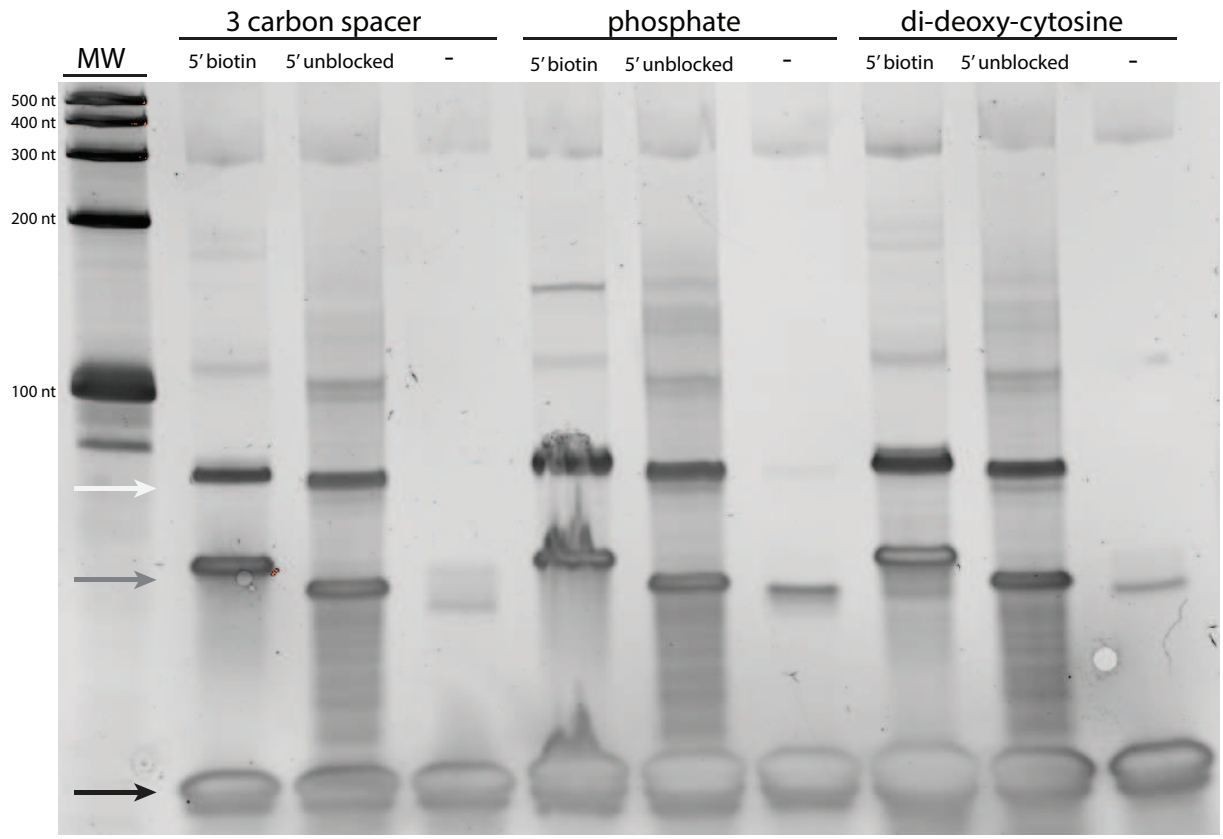


Figure A.10: Effect of 3′ blocking group on second adapter concatemerization and ligation efficiency. Three modifications (3 carbon spacer, phosphate, and di-deoxy-cytosine) were added to the 3′ end of the adapter (black arrow), which was then ligated (60 °C for 2 hours with Circligase I) to an RT primer (gray arrow) that was either 5′ blocked (57 nt) with biotin or unblocked (54 nt). A third control lane for each 3′ modification consisting of the ligation of the adapter without RT primer is shown (-). Regardless of 3′ modification, some adapter form concatemers during the ligation, with concatemers showing up at 50 nt, near the size of the RT primers (gray arrow). However, this effect is the weakest for the 3 carbon spacer modification, which also shows the least amount of adapter concatemer ligated to the RT primer, seen as bands longer than the expected ligated product which is 79 nt or 82 nt (white arrow). While no modification stood out as the clear best, the 3′ 3-carbon spacer modification was chosen for its relative cleanliness and lower cost than di-deoxy-cytosine.

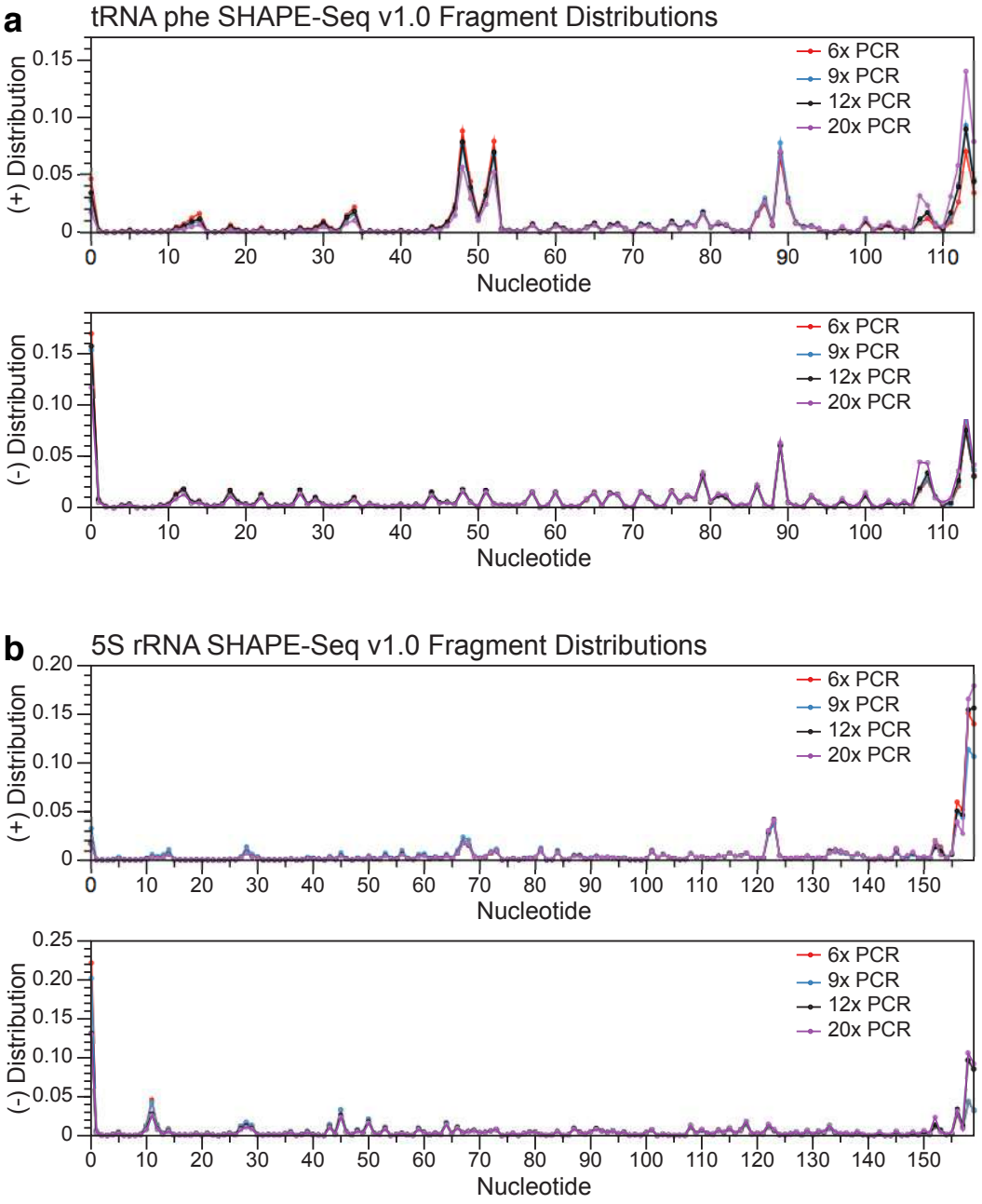


Figure A.11: SHAPE-Seq v1.0 fragment distributions for different numbers of PCR cycles. (+) and (-) fragment distributions are plotted for SHAPE-Seq v1.0 with varying numbers of PCR cycles for (a) $tRNA^{phe}$ and (b) 5S rRNA (see Figure 2.2). Pearson correlation values for specific comparisons are shown in Figure 2.3.

Figure A.12: SHAPE-Seq v2.0 vs. SHAPE-Seq v1.0. For each RNA, SHAPE-Seq v2.0 (blue) and SHAPE-Seq v1.0 (red) θ's are plotted on top, with a zoomed window on portions of the comparison shown on the bottom left. Error bars are calculated as standard deviations of reactivities at each nucleotide from three independent replicate experiments. The bottom right shows SHAPE-Seq v2.0 vs. v1.0 θ's plotted as a scatter plot, from which the Pearsons correlation (R) between the two techniques is calculated. Gray boxes represent flanking structure cassette regions included in the RNAs for SHAPE-Seq v1.0, but not present in the RNAs for SHAPE-Seq v2.0 (see Table A.1).

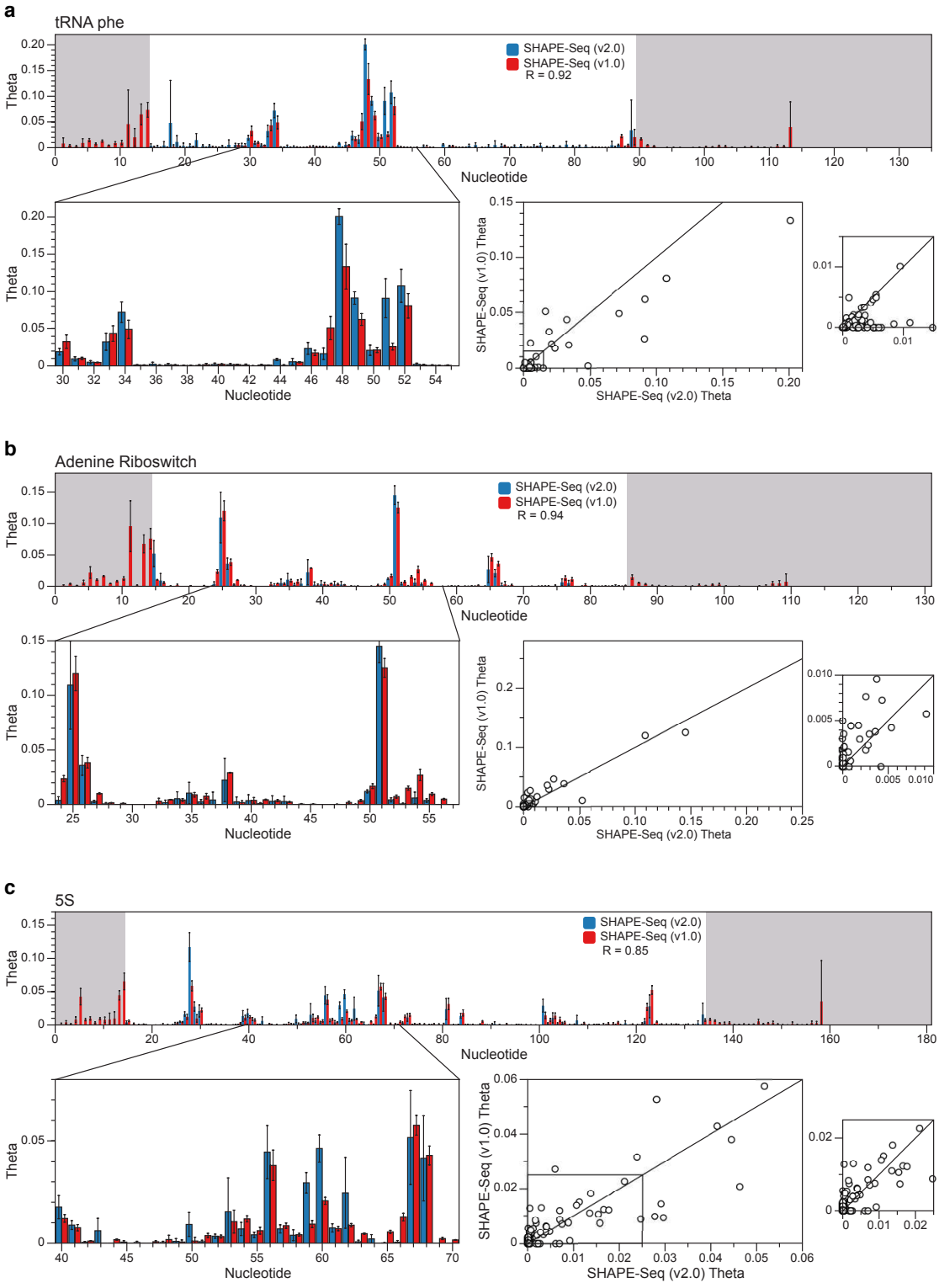

a
tRNA phe
SHAPE-Seq (v2.0)
SHAPE-Seq (v1.0)
R = 0.92
Theta
Nucleotide
SHAPE-Seq (v1.0) Theta
SHAPE-Seq (v2.0) Theta
b
Adenine Riboswitch
SHAPE-Seq (v2.0)
SHAPE-Seq (v1.0)
R = 0.94
Theta
Nucleotide
SHAPE-Seq (v1.0) Theta
SHAPE-Seq (v2.0) Theta
c
5S
SHAPE-Seq (v2.0)
SHAPE-Seq (v1.0)
R = 0.85
Theta
Nucleotide
SHAPE-Seq (v1.0) Theta
SHAPE-Seq (v2.0) Theta

Figure A.12 (Continued)

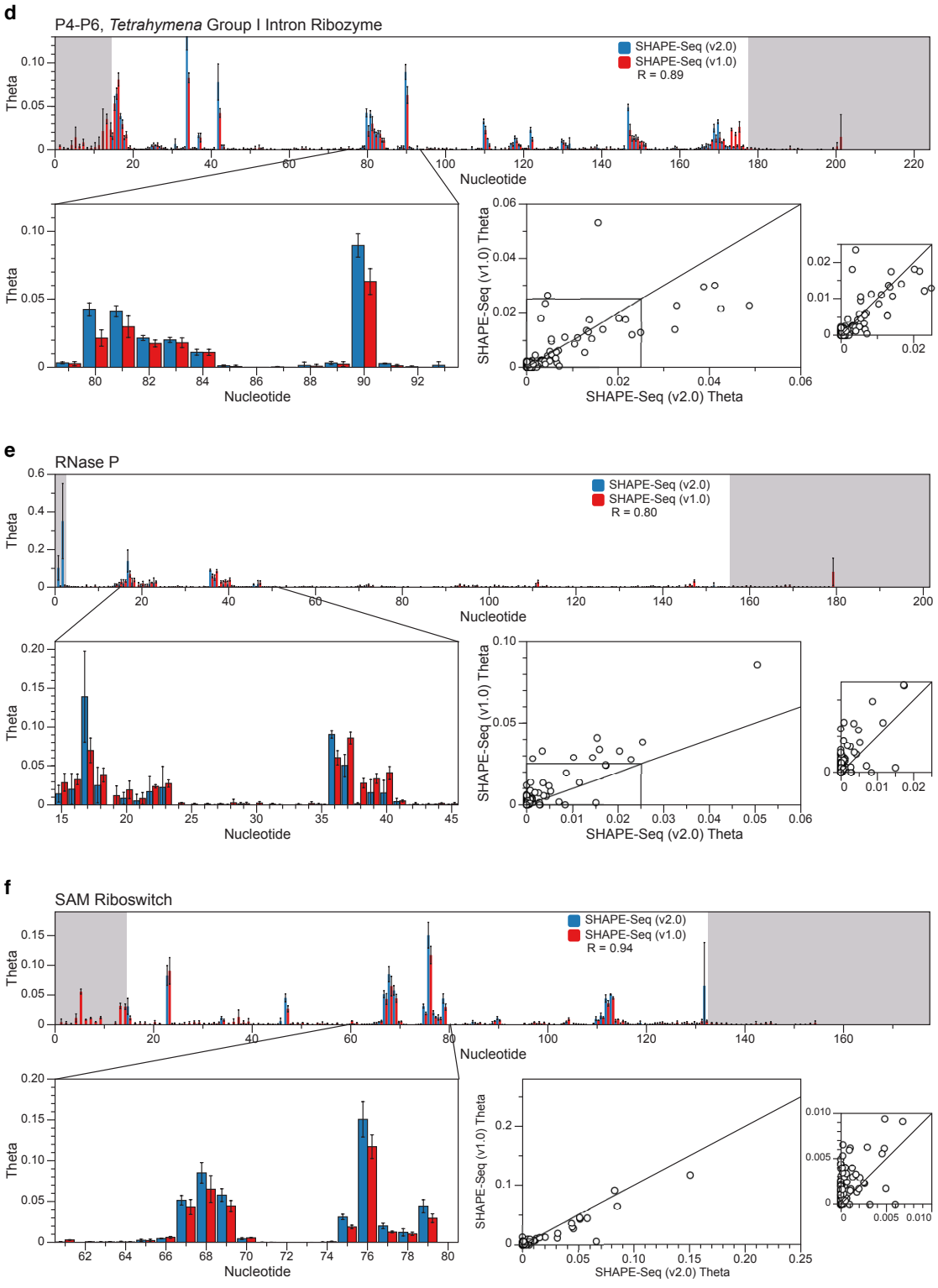


Figure A.12 (Continued)

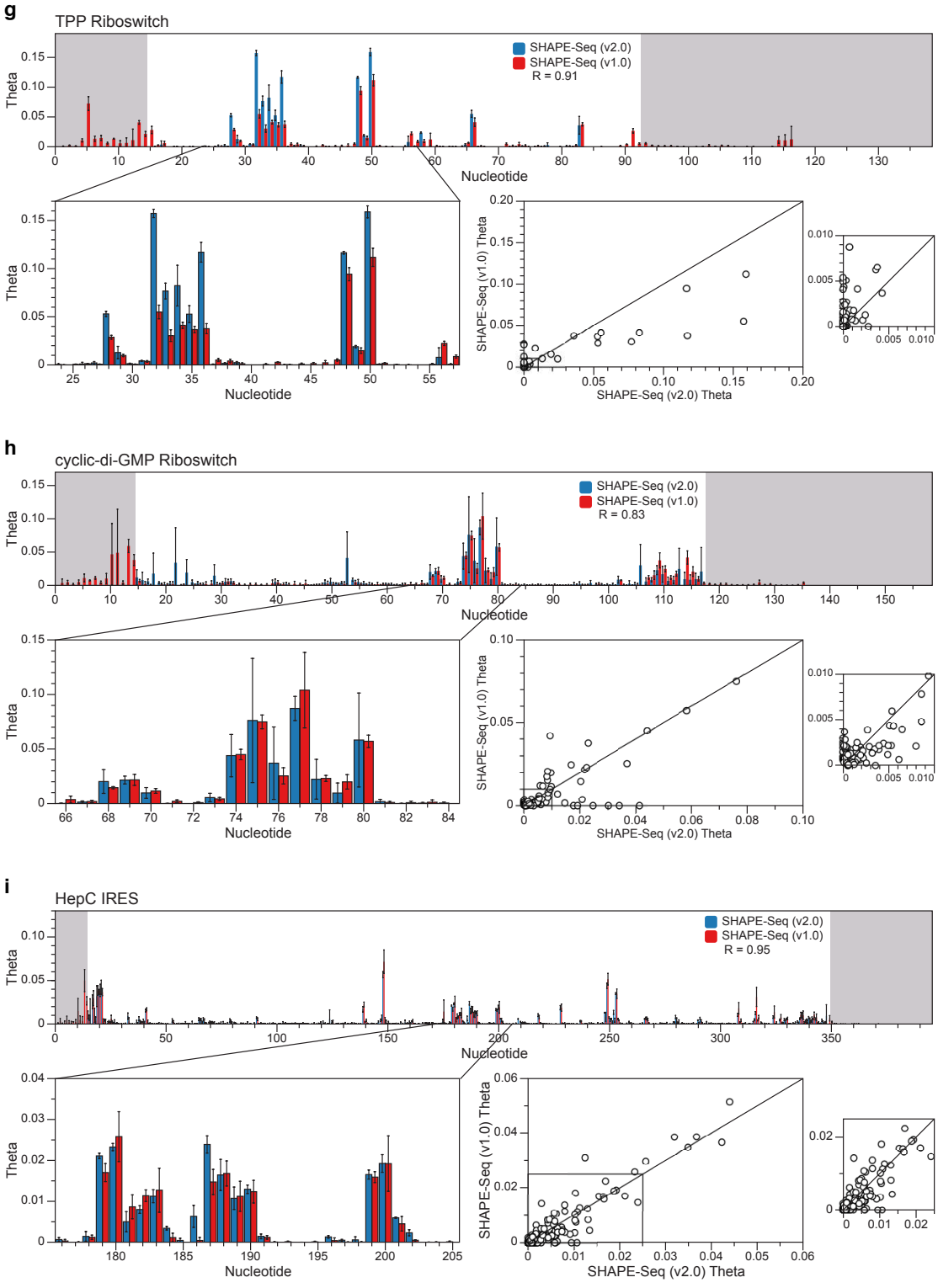
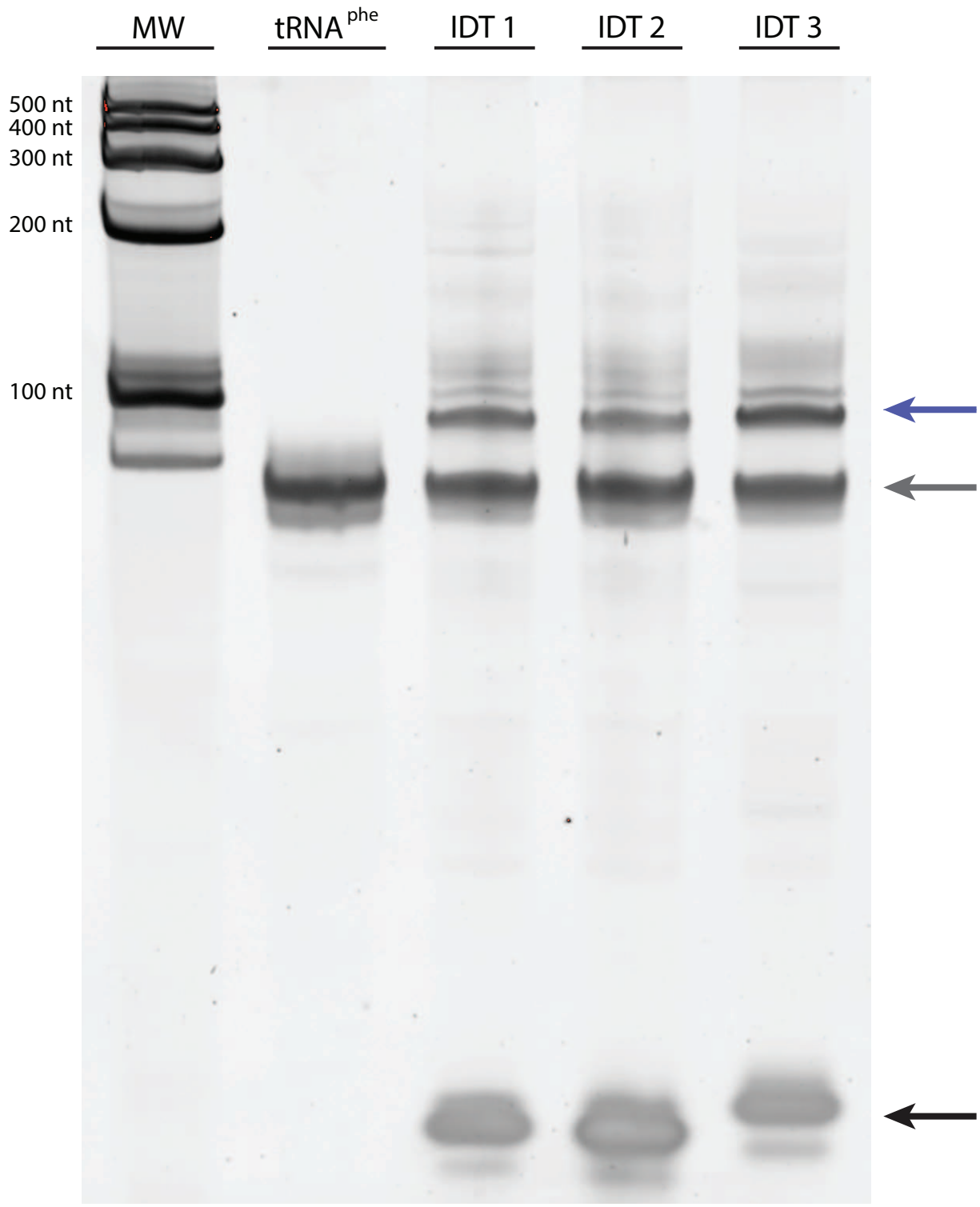
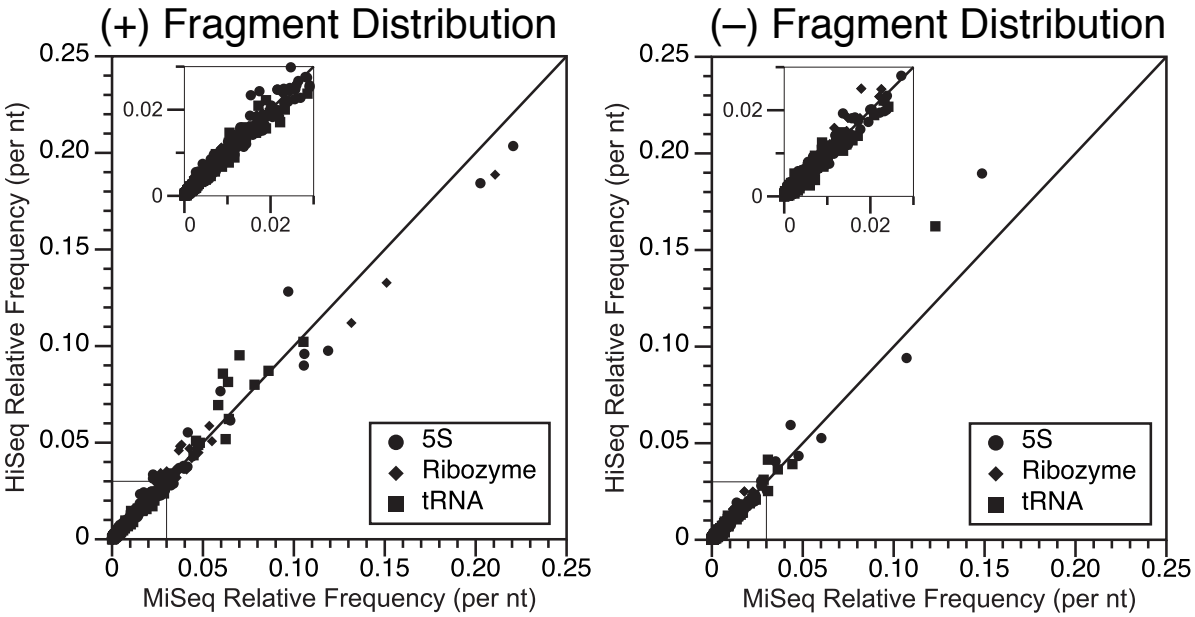


Figure A.13: Choice of 5 adenylated linker sequence for SHAPE-Seq v2.0 universal priming strategy. Each linker choice was ligated onto an unmodified strand of $tRNA^{phe}$ using T4 RNA Ligase 2, truncated KQ overnight at room temperature. The gray arrow indicates the bands that correspond to the full-length, unligated tRNA, the blue arrow above shows the bands corresponding to the successfully ligated tRNA+linker, and the bands at the very bottom of the figure correspond to unligated linker. IDT1, IDT2 and IDT3 sequences are commercially available from Integrated DNA Technologies, Inc. at: `http://www.idtdna.com/pages/products/mirna/mirna-cloning-products`. IDT2 was chosen due to it having the highest melting temperature with its RT primer (Figure A.4).


MW
tRNA phe
IDT 1
IDT 2
IDT 3
500 nt
400 nt
300 nt
200 nt
100 nt



R values	5S rRNA	*Tetrahymena* Ribozyme	$tRNA^{phe}$	All
(+) Fragments	0.99	0.99	1.0	0.99
(–) Fragments	0.99	1.0	1.0	1.0

Figure A.14: SHAPE-Seq v2.0 reactivities generated from the MiSeq and HiSeq platforms. (+) and (-) fragment distributions for each RNA were compared between sequencing platforms as in Figure 2.2. Pearson correlation values for individual comparisons between fragment distributions, are shown on the bottom.

APPENDIX B

SUPPLEMENTARY INFORMATION FOR SIMULTANEOUS CHARACTERIZATION OF CELLULAR RNA STRUCTURE AND FUNCTION WITH IN-CELL SHAPE-SEQ

B.1 Supplementary Equations

Normalization of reactivities and constrained structure prediction. The Spats software package (`http://spats.sourceforge.net/`) uses input sequencing reads [108, 122] to determine the θ_i value for each nucleotide, i, where θ_i represents the fraction of modifications occurring at position i in an RNA of length L [97]. However, because θ_i is dependent on L, we introduced a new normalization method in Loughrey *et al.* [122] to convert θ_i to ρ_i, which is the normalized reactivity intensity for a nucleotide i in an RNA of length L (which does not include the RT priming site), where $\langle \rho_i \rangle = 1$. A brief explanation of the normalization method is presented below:

First, determine L, the length of the RNA region under study that does not include the bases specific to the RNA of interest in the selection primer. These bases constitute the RT priming site and the extension into the cDNA that provides selection against the dimer side products. For example, the sequence GCCTCTACCTGCTTCGGCC-GATAAA should be excluded from the end of libraries that are generated from the ECK120051404 RT priming site. Next, determine a normalization constant, n, such that all of the θ_i values in L sum to 1 (Appendix B.1). These steps are necessary because proper alignment of the sequencing reads to the RNA target requires the inclusion of the RT priming site sequence, since it is read by the Illumina platform. However, the RT priming site and the selection primer extension into the cDNA contain no structural information although they are automatically included in the calculation of θ_i by Spats.

Thus, subtle differences in the (+) and (-) samples can cause nucleotides in the RT priming site to appear reactive, although they are simply an artifact of the data processing steps and contain no structural information. To exclude these spurious nucleotides from the original calculation of θ_i generated by Spats, n is calculated as:

$$n = \sum_{i=1}^{L} \theta_i \tag{B.1}$$

where L is the length of the RNA exclusive of the bases specific to the RNA of interest in the selection primer used to generate the dsDNA libraries. Here, an index of 1 refers to the 5′ end of the RNA. Next, θ_i is converted to ρ_i by multiplying by the length, L, and dividing by the normalization factor (Appendix B.1) to set the total average of ρ to 1 (Appendix B.1).

$$\rho_i = \frac{\theta_i L}{n} \tag{B.2}$$

$$\bar{\rho} = \frac{\sum_{i=1}^{L} \rho_i}{L} = \frac{\sum_{i=1}^{L} \frac{\theta_i L}{n}}{L} = \frac{1}{n} \sum_{i=1}^{L} \theta_i = \frac{\sum_{i=1}^{L} \theta_i}{\sum_{i=1}^{L} \theta_i} = 1 \tag{B.3}$$

The normalized ρ_i values can be used to constrain the RNAstructure [100, 122] secondary structure predication program using a pseudo free energy term, ΔG (Appendix B.1). Slope and intercept values were fit in Loughrey *et al.* [122] to be $m = 1.1$, $b = -0.3$, which are the values used in this work.

$$\Delta G = m \ln(\rho_i + 1) + b \tag{B.4}$$

B.2 Supplementary Tables

Table B.1: List of terminators screened for building the antisense platform. During design of the antisense platform, we sought a small intrinsic terminator that would serve as both an RT priming site and an efficient terminator of *E. coli* transcription. Ten different terminators with varying stem-loop properties were tested for RT priming site capability and the maintenance of ON/OFF functionality of the sense-antisense pairs. A double terminator strategy was ultimately used, combining the high termination efficiency of t500 with the RT priming capability of ECK120051404.

Terminator	Source	Sequence
t500	Phage 82 mut	CAAAGCCCGCCGAAAGGCGGGCTTTTTTTT
T7 RNA pol gene	T7 phage	CTGGCTCACCTTCGGGTGGGCCTTTCTGCGTTTATAAGG
T7 RNA pol	T7 phage	CCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTT
T3 RNA pol	T3 phage	CTGGCTCACCTTCACGGGTGGGCCTTTCTTCGTTCCGGGCA
pT181	*S. aureus*	CGATTCCTTAAACGAAATTGAGATTAAGGAGTCGCTCTTTTTT
ECK120051404	Synthetic ([186])	CCTCTACCTGCTTCGGCCGATAAAGCCGACGATAATACTCC
ECK120010812-R	Synthetic ([186])	AACGGTTTATTAGTCTGGAGACGGCAGACTATCCTCTTCCC
ECK120010840	Synthetic ([186])	CGTACCAGGCCCCTGCAATTTCAACAGGGGCCTTTTTTTATCC
tryptophan attenuator L126	*E. coli* ([186])	ACCCAGCCCGCCTAATAAGCGGGCTTTTTTTTGAACA
p81	*S. aureus*	GCGGGGAATGTATACAGTTCATGTATATATTCCCCGCTTTTTTTTT

Table B.2: RNA sequences and plasmids used in this study. The sense RNAs (crRNAs and RNA-IN) are expressed before the superfolder GFP (SFGFP) sequence [184]. The ribosome binding sites (RBS) within the regulator sequences and the sense RT primer binding site are near the 5′ end of the SF-GFP sequence. The antisense RNAs (taRNAs and RNA-OUT) are terminated by a double terminator composed of the ECK120051404 terminator [186], which contains the antisense RT primer binding site, and the t500 terminator. See Figure B.1 for plasmid construction order. For the endogenously expressed RNAs the entire transcript is shown, with each specific RT priming site bolded and underlined.

Name	Sequence
btuB mRNA	GCCGGTCCTGTGAGTTAATAGGGAATCCAGTGCGAATCTGGAGCTGACGCG CAGCGGTAAGGAAAGGTGCGATGATTGCGTTATGCGGACACTGCCATTCG GTGGGAAGTCATCATCTCTTAGTATCTTAGATACCCCTCCAAGCCCGAAGA CCTGCCGGCCAACGTCGCATCTGGTTCTCATCATCGCGTAATATTGATGAA ACCTGCGGCATCCTTCTTCTATTGTGGATGCTTTACAATGATTAAAAAAGCT TCGCTGCTGACGGCGTGTTCCGTCACGGCATTTTCCGCTTGGGCACAGGAT ACCAGCCCGGATACTCTCGTCGTTACTGCTAACCGTTTTGAACAGCCGCGC AGCACTGTGCTTGCACCAACCACCGTTGTGACCCGTCAGGATATCGACCGC TGGCAGTCGACCTCGGTCAATGATGTGCTGCGCCGTCTTCCGGGCGTCGAT ATCACCCAAAACGGCGGTTCAGGTCAGCTCTCATCTATTTTTATTCGCGGTA CAAATGCCAGTCATGTGTTGGTGTTAATTGATGGCGTACGCCTGAATCTGG CGGGGGTGAGTGGTTCTGCCGACCTTAGCCAGTTCCCTATTGCGCTTGTCCA GCGTGTTGAATATATCCGTGGGCCGCGCTCCGCTGTTTATGGTTCCGATGCA ATAGGCGGGGTGGTGAATATCATCACGACGCGCGATGAACCCGGAACGGA AATTTCAGCAGGGTGGGGAAGCAATAGTTATCAGAACTATGATGTCTCTAC GCAGCAACAACTGGGGGATAAGACACGGGTAACGCTGTTGGGCGATTATG CCCATACTCATGGTTATGATGTTGTTGCCTATGGTAATACCGGAACGCAAG CGCAGACAGATAACGATGGTTTTTTAAGTAAAACGCTTTATGGCGCGCTGG AGCATAACTTTACTGATGCCTGGAGCGGCTTTGTGCGCGGCTATGGCTATG ATAACCGTACCAATTATGACGCGTATTATTCTCCCGGTTCACCGTTGCTCGA TACCCGTAAACTCTATAGCCAAAGTTGGGACGCCGGGCTGCGCTATAACGG CGAACTGATTAAATCACAACTCATTACCAGCTATAGCCATAGCAAAGATTA CAACTACGATCCCCATTATGGTCGTTATGATTCGTCGGCGACGCTCGATGA GATGAAGCAATACACCGTCCAGTGGGCAAACAATGTCATCGTTGGTCACG GTAGTATTGGTGCGGGTGTCGACTGGCAGAAACAGACTACGACGCCGGGT ACAGGTTATGTTGAGGATGGATATGATCAACGTAATACCGGCATCTATCTG ACCGGGCTGCAACAAGTCGGCGATTTTACCTTTGAAGGCGCAGCACGCAGT GACGATAACTCACAGTTTGGTCGTCATGGAACCTGGCAAACCAGCGCCGGT TGGGAATTCATCGAAGGTTATCGCTTCATTGCTTCCTACGGGACATCTTATA AGGCACCAAATCTGGGGCAAC

Table B.2 (Continued)

Name	Sequence
btuB mRNA (cont.)	TGTATGGCTTCTACGGAAATCCGAATCTGGACCCGGAGAAAAGCAAACAG TGGGAAGGCGCGTTTGAAGGCTTAACCGCTGGGGTGAACTGGCGTATTTCC GGATATCGTAACGATGTCAGTGACTTGATCGATTATGATGATCACACCCTG AAATATTACAACGAAGGGAAAGCGCGGATTAAGGGCGTCGAGGCGACCG CCAATTTTGATACCGGACCACTGACGCATACTGTGAGTTATGATTATGTCG ATGCGCGCAATGCGATTACCGACACGCCGTTGTTACGCCGTGCTAAACAGC AGGTGAAATACCAGCTCGACTGGCAGTTGTATGACTTCGACTGGGGTATTA CTTATCAGTATTTAGGCACTCGCTATGATAAGGATTACTCATCTTATCCTTA TCAAACCGTTAAAATGGGCGGTGTGAGCTTGTGGGATCTTGCGGTTGCGTA TCCGGTCACCTCTCACCTGACAGTTCGTGGTAAAATAGCCAACCTGTTCGA CAAAGATTATGAGACAGTCTATGGCTACCAAACTGCAGGACGGGAATACA CCTTGTCTGGCAGCTACACCTTCTGA
5S rRNA	TGCCTGGCGGCCGTAGCGCGGTGGTCCCACCTGACCCCATGCCGAACTCAG AAGTGAAACGCCGTAGCGCCGATGGTAGTGTGGGGTCTCCCCATGCGAGA GTAGGGAACTGCCAGGCAT
RNase P	GAAGCTGACCAGACAGTCGCCGCTTCGTCGTCGTCCTCTTCGGGGGAGACG GGCGGAGGGGAGGAAAGTCCGGGCTCCATAGGGCAGGGTGCCAGGTAAC GCCTGGGGGGGAAACCCACGACCAGTGCAACAGAGAGCAAACCGCCGAT GGCCCGCGCAAGCGGGATCAGGTAAGGGTGAAAGGGTGCGGTAAGAGCG CACCGCGCGGCTGGTAACAGTCCGTGGCACGGTAAACTCCACCCGGAGCA AGGCCAAATAGGGGTTCATAAGGTACGGCCCGTACTGAACCCGGGTAGGC TGCTTGAGCCAGTGAGCGATTGCTGGCCTAGATGAATGACTGTCCACGACA GAACCCGGCTTATCGGTCAGTTTCACCT
crR10	GAATTCTACCTATCTGCTCTTGAATTTGGGTATTAAAGAGGAGAAAGGTAC C SFGFP - TrrnB
crR12	GAATTCTACCATTCACCTCTTGGATTTGGGTATTAAAGAGGAGAAAGGTAC C - SFGFP - TrrnB
taR10	ACACCCAAATTCATGAGCAGATTGGTAGTGGTGGTTAATGAAAATTAACTT ACTACTACCTTTCTCTAGAG - ECK404 - t500
taR12	ACCCAAATCCAGGAGGTGATTGGTAGTGGTGGTTAATGAAAATTAACTTAC TACTACCATATATCTCTAGAG - ECK404 - t500
RNA-IN S3	GGGAAAAATCAATAAGGAGACAACAAGATGTGCGAACTCGATGCT - SFGFP (w/o AUG) - Double terminator
RNA-IN S4	GCCAAAAATCAATAAGGAGACAACAAGATGTGCGAACTCGATGCT - SFGFP (w/o AUG) - Double terminator
RNA-IN S4 A25C	GCCAAAAATCAATAAGGAGACAACCAGATGTGCGAACTCGATGCT - SFGFP (w/o AUG) - Double terminator
RNA-IN S4 C24A	GCCAAAAATCAATAAGGAGACAAAAAGATGTGCGAACTCGATGCT - SFGFP (w/o AUG) - Double terminator
RNA-IN S4 C24A A25C	GCCAAAAATCAATAAGGAGACAAACAGATGTGCGAACTCGATGCT - SFGFP (w/o AUG) - Double terminator
RNA-OUT A3	TCGCACATCTTGTTGTCTGATTATTGATTTTTCCCGAAACCATTTGATCATAT GACAAGATGTGTATG - ECK404 - t500
RNA-OUT A4	TCGCACATCTTGTTGTCTGATTATTGATTTTTGGCGAAACCATTTGATCATAT GACAAGATGTGTATG - ECK404 - t500

Table B.2 (Continued)

Name	Sequence
Superfolder GFP (SFGFP)	ATGAGCAAAGGAGAAGAACTTTTCACTGGAGTTGTCCCAATTCTTGTTGAA TTAGATGGTGATGTTAATGGGCACAAATTTTCTGTCCGTGGAGAGGGTGAA GGTGATGCTACAAACGGAAAACTCACCCTTAAATTTATTTGCACTACTGGA AAACTACCTGTTCCGTGGCCAACACTTGTCACTACTCTGACCTATGGTGTTC AATGCTTTTCCCGTTATCCGGATCACATGAAACGGCATGACTTTTTCAAGAG TGCCATGCCCGAAGGTTATGTACAGGAACGCACTATATCTTTCAAAGATGA CGGGACCTACAAGACGCGTGCTGAAGTCAAGTTTGAAGGTGATACCCTTGT TAATCGTATCGAGTTAAAGGGTATTGATTTTAAAGAAGATGGAAACATTCT TGGACACAAACTCGAGTACAACTTTAACTCACACAATGTATACATCACGGC AGACAAACAAAAGAATGGAATCAAAGCTAACTTCAAAATTCGCCACAAC GTTGAAGATGGTTCCGTTCAACTAGCAGACCATTATCAACAAAATACTCCA ATTGGCGATGGCCCTGTCCTTTTACCAGACAACCATTACCTGTCGACACAA TCTGTCCTTTCGAAAGATCCCAACGAAAAGCGTGACCACATGGTCCTTCTT GAGTTTGTAACTGCTGCTGGGATTACACATGGCATGGATGAGCTCTACAAA TAA
TrrnB operon fragment	GGATCTGAAGCTTGGGCCCGAACAAAAACTCATCTCAGAAGAGGATCTGA ATAGCGCCGTCGACCATCATCATCATCATCATTGAGTTTAAACGGTCTCCA GCTTGGCTGTTTTGGCGGATGAGAGAAGATTTTCAGCCTGATACAGATTAA ATCAGAACGCAGAAGCGGTCTGATAAAACAGAATTTGCCTGGCGGCAGTA GCGCGGTGGTCCCACCTGACCCCATGCCGAACTCAGAAGTGAAACGCCGT AGCGCCGATGGTAGTGTGGGGTCTCCCCATGCGAGAGTAGGGAACTGCCA GGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTA TCTGTTGTTTGTCGGTGAACTGGATCCTTACTCGAGTCTAGA
ECK1200-51404 (ECK404) terminator	CCTCTACCTGCTTCGGCCGATAAAGCCGACGATAATACTCC
t500 terminator	CAAAGCCCGCCGAAAGGCGGGCTTTTTTTT
Double terminator	GGATCCAAACTCGAGTAAGGATCTCCAGGCATCAAATAAAACGAAAGGCT CAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTC TCTACTAGAGTCACACTGGCTCACCTTCGGGTGGGCCTTTCTGCGTTTATAC CTAGGGTACGGGTTTTGC

Table B.2 (Continued)

Name	Sequence
ColE1 ori fragment	GGATCCTTACTCGAGTCTAGACTGCAGGCTTCCTCGCTCACTGACTCGCTGC GCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAA TACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCA AAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTT TTTCCACAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAG TCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCC CTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATA CCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGC TGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGC ACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTC TTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTG GTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTG AAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGC GCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCC GGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAG ATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACG GGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATG A
p15A ori fragment	GGTGAGAATCCAAGCCTCCGATCAACGTCTCATTTTCGCCAAAAGTTGGCC CAGGGCTTCCCGGTATCAACAGGGACACCAGGATTTATTTATTCTGCGAAG TGATCTTCCGTCACAGGTATTTATTCGGCGCAAAGTGCGTCGGGTGATGCTG CCAACTTACTGATTTAGTGTATGATGGTGTTTTTGAGGTGCTCCAGTGGCTT CTGTTTCTATCAGCTGTCCCTCCTGTTCAGCTACTGACGGGGTGGTGCGTAA CGGCAAAAGCACCGCCGGACATCAGCGCTAGCGGAGTGTATACTGGCTTA CTATGTTGGCACTGATGAGGGTGTCAGTGAAGTGCTTCATGTGGCAGGAGA AAAAAGGCTGCACCGGTGCGTCAGCAGAATATGTGATACAGGATATATTC CGCTTCCTCGCTCACTGACTCGCTACGCTCGGTCGTTCGACTGCGGCGAGC GGAAATGGCTTACGAACGGGGCGGAGATTTCCTGGAAGATGCCAGGAAGA TACTTAACAGGGAAGTGAGAGGGCCGCGGCAAAGCCGTTTTTCCATAGGC TCCGCCCCCCTGACAAGCATCACGAAATCTGACGCTCAAATCAGTGGTGGC GAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGCGGCTCC CTCGTGCGCTCTCCTGTTCCTGCCTTTCGGTTTACCGGTGTCATTCCGCTGTT ATGGCCGCGTTTGTCTCATTCCACGCCTGACACTCAGTTCCGGGTAGGCAGT TCGCTCCAAGCTGGACTGTATGCACGAACCCCCCGTTCAGTCCGACCGCTG CGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGAAAGACATGCAAA AGCACCACTGGCAGCAGCCACTGGTAATTGATTTAGAGGAGTTAGTCTTGA AGTCATGCGCCGGTTAAGGCTAAACTGAAAGGACAAGTTTTGGTGACTGCG CTCCTCCAAGCCAGTTACCTCGGTTCAAAGAGTTGGTAGCTCAGAGAACCT TCGAAAAACCGCCCTGCAAGGCGGTTTTTTCGTTTTCAGAGCAAGAGATTA CGCGCAGACCAAAACGATCTCAAGAAGATCATCTTATTAATCAGATAAAA TATTTCTAGATTTCAGTGCAATTTATCTCTTCAAATGTAGCACCTGAAGTCA GCCCCATACGATATAAGTTGTAATTCTCATGTTTGACAGCTTATCATCGATA AGCTTCCGATGGCGCGCCGAGAGGCTTTACACTTTATGCTTCCGGCT

Table B.2 (Continued)

Name	Sequence
AmpR fragment	GATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTT TTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAAT GCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCAT AGTTGCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACC ATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCC AGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTG GTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGC TAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCT ACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCG GTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAG CGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGT GTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCA TCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAG AATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATA ATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTT CTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGA TGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAG CGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGA ATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATT ATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATG TATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAG TGCCACCTGACGTCTAAGAAACCATTATTATCATGACATTAACCTATAAAA ATAGGCGTATCACGAGGCAGAATTTCAGATAAAAAAAATCCTTAGCTTTCG CTAAGGATGATTTCTG
CmR fragment	CTGCAGTTGATCGGGCACGTAAGAGGTTCCAACTTTCACCATAATGAAATA AGATCACTACCGGGCGTATTTTTTGAGTTATCGAGATTTTCAGGAGCTAAG GAAGCTAAAATGGAGAAAAAAATCACTGGATATACCACCGTTGATATATC CCAATGGCATCGTAAAGAACATTTTGAGGCATTTCAGTCAGTTGCTCAATG TACCTATAACCAGACCGTTCAGCTGGATATTACGGCCTTTTTAAAGACCGT AAAGAAAAATAAGCACAAGTTTTATCCGGCCTTTATTCACATTCTTGCCCG CCTGATGAATGCTCATCCGGAATTTCGTATGGCAATGAAAGACGGTGAGCT GGTGATATGGGATAGTGTTCACCCTTGTTACACCGTTTTCCATGAGCAAACT GAAACGTTTTCATCGCTCTGGAGTGAATACCACGACGATTTCCGGCAGTTT CTACACATATATTCGCAAGATGTGGCGTGTTACGGTGAAAACCTGGCCTAT TTCCCTAAAGGGTTTATTGAGAATATGTTTTTCGTCTCAGCCAATCCCTGGG TGAGTTTCACCAGTTTTGATTTAAACGTGGCCAATATGGACAACTTCTTCGC CCCCGTTTTCACCATGGGCAAATATTATACGCAAGGCGACAAGGTGCTGAT GCCGCTGGCGATTCAGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGC AGAATGCTTAATGAATTACAACAGTACTGCGATGAGTGGCAGGGCGGGGC GTAATTTGATATCGAGCTCGCTTGGACTCCTGTTGATAGATCCAGTAATGAC CTCAGAACTCCATCTGGATTTGTTCAGAACGCTCGGTTGCCGCCGGGCGTTT TTTATT
J23119 promoter variants	GAATTC(TAAAGATC/GAA)TTTGACAGCTAGCTCAGTCCTAGGTATAAT(AC TAGT/GAATTC) (riboregulators/RNA-OUT-RNA-IN)

Table B.2 (Continued)

Name	Sequence
Example plasmid (crR10)	GAATTCTAAAGATCTTTGACAGCTAGCTCAGTCCTAGGTATAATACTAGTG AATTCTACCTATCTGCTCTTGAATTTGGGTATTAAAGAGGAGAAAGGTACC ATGAGCAAAGGAGAAGAACTTTTCACTGGAGTTGTCCCAATTCTTGTTGAA TTAGATGGTGATGTTAATGGGCACAAATTTTCTGTCCGTGGAGAGGGTGAA GGTGATGCTACAAACGGAAAACTCACCCTTAAATTTATTTGCACTACTGGA AAACTACCTGTTCCGTGGCCAACACTTGTCACTACTCTGACCTATGGTGTTC AATGCTTTTCCCGTTATCCGGATCACATGAAACGGCATGACTTTTTCAAGAG TGCCATGCCCGAAGGTTATGTACAGGAACGCACTATATCTTTCAAAGATGA CGGGACCTACAAGACGCGTGCTGAAGTCAAGTTTGAAGGTGATACCCTTGT TAATCGTATCGAGTTAAAGGGTATTGATTTTAAAGAAGATGGAAACATTCT TGGACACAAACTCGAGTACAACTTTAACTCACACAATGTATACATCACGGC AGACAAACAAAAGAATGGAATCAAAGCTAACTTCAAAATTCGCCACAAC GTTGAAGATGGTTCCGTTCAACTAGCAGACCATTATCAACAAAATACTCCA ATTGGCGATGGCCCTGTCCTTTTACCAGACAACCATTACCTGTCGACACAA TCTGTCCTTTCGAAAGATCCCAACGAAAAGCGTGACCACATGGTCCTTCTT GAGTTTGTAACTGCTGCTGGGATTACACATGGCATGGATGAGCTCTACAAA TAAGGATCTGAAGCTTGGGCCCGAACAAAAACTCATCTCAGAAGAGGATC TGAATAGCGCCGTCGACCATCATCATCATCATCATTGAGTTTAAACGGTCT CCAGCTTGGCTGTTTTGGCGGATGAGAGAAGATTTTCAGCCTGATACAGAT TAAATCAGAACGCAGAAGCGGTCTGATAAAACAGAATTTGCCTGGCGGCA GTAGCGCGGTGGTCCCACCTGACCCCATGCCGAACTCAGAAGTGAAACGC CGTAGCGCCGATGGTAGTGTGGGGTCTCCCCATGCGAGAGTAGGGAACTG CCAGGCATCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGT TTTATCTGTTGTTTGTCGGTGAACTGGATCCTTACTCGAGTCTAGACTGCAG TTGATCGGGCACGTAAGAGGTTCCAACTTTCACCATAATGAAATAAGATCA CTACCGGGCGTATTTTTTGAGTTATCGAGATTTTCAGGAGCTAAGGAAGCT AAAATGGAGAAAAAAATCACTGGATATACCACCGTTGATATATCCCAATG GCATCGTAAAGAACATTTTGAGGCATTTCAGTCAGTTGCTCAATGTACCTAT AACCAGACCGTTCAGCTGGATATTACGGCCTTTTTAAAGACCGTAAAGAAA AATAAGCACAAGTTTTATCCGGCCTTTATTCACATTCTTGCCCGCCTGATGA ATGCTCATCCGGAATTTCGTATGGCAATGAAAGACGGTGAGCTGGTGATAT GGGATAGTGTTCACCCTTGTTACACCGTTTTCCATGAGCAAACTGAAACGTT TTCATCGCTCTGGAGTGAATACCACGACGATTTCCGGCAGTTTCTACACATA TATTCGCAAGATGTGGCGTGTTACGGTGAAAACCTGGCCTATTTCCCTAAA GGGTTTATTGAGAATATGTTTTTCGTCTCAGCCAATCCCTGGGTGAGTTTCA CCAGTTTTGATTTAAACGTGGCCAATATGGACAACTTCTTCGCCCCCGTTTT CACCATGGGCAAATATTATACGCAAGGCGACAAGGTGCTGATGCCGCTGG CGATTCAGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCT TAATGAATTACAACAGTACTGCGATG

Table B.2 (Continued)

Name	Sequence
Example plasmid (crR10) (cont.)	TGGGCAAATATTATACGCAAGGCGACAAGGTGCTGATGCCGCTGGCGATT CAGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTAATG AATTACAACAGTACTGCGATGAGTGGCAGGGCGGGGCGTAATTTGATATC GAGCTCGCTTGGACTCCTGTTGATAGATCCAGTAATGACCTCAGAACTCCA TCTGGATTTGTTCAGAACGCTCGGTTGCCGCCGGGCGTTTTTTATTGGTGAG AATCCAAGCCTCCGATCAACGTCTCATTTTCGCCAAAAGTTGGCCCAGGGC TTCCCGGTATCAACAGGGACACCAGGATTTATTTATTCTGCGAAGTGATCTT CCGTCACAGGTATTTATTCGGCGCAAAGTGCGTCGGGTGATGCTGCCAACT TACTGATTTAGTGTATGATGGTGTTTTTGAGGTGCTCCAGTGGCTTCTGTTTC TATCAGCTGTCCCTCCTGTTCAGCTACTGACGGGGTGGTGCGTAACGGCAA AAGCACCGCCGGACATCAGCGCTAGCGGAGTGTATACTGGCTTACTATGTT GGCACTGATGAGGGTGTCAGTGAAGTGCTTCATGTGGCAGGAGAAAAAAG GCTGCACCGGTGCGTCAGCAGAATATGTGATACAGGATATATTCCGCTTCC TCGCTCACTGACTCGCTACGCTCGGTCGTTCGACTGCGGCGAGCGGAAATG GCTTACGAACGGGGCGGAGATTTCCTGGAAGATGCCAGGAAGATACTTAA CAGGGAAGTGAGAGGGCCGCGGCAAAGCCGTTTTTCCATAGGCTCCGCCC CCCTGACAAGCATCACGAAATCTGACGCTCAAATCAGTGGTGGCGAAACC CGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGCGGCTCCCTCGTGC GCTCTCCTGTTCCTGCCTTTCGGTTTACCGGTGTCATTCCGCTGTTATGGCCG CGTTTGTCTCATTCCACGCCTGACACTCAGTTCCGGGTAGGCAGTTCGCTCC AAGCTGGACTGTATGCACGAACCCCCCGTTCAGTCCGACCGCTGCGCCTTA TCCGGTAACTATCGTCTTGAGTCCAACCCGGAAAGACATGCAAAAGCACC ACTGGCAGCAGCCACTGGTAATTGATTTAGAGGAGTTAGTCTTGAAGTCAT GCGCCGGTTAAGGCTAAACTGAAAGGACAAGTTTTGGTGACTGCGCTCCTC CAAGCCAGTTACCTCGGTTCAAAGAGTTGGTAGCTCAGAGAACCTTCGAAA AACCGCCCTGCAAGGCGGTTTTTTCGTTTTCAGAGCAAGAGATTACGCGCA GACCAAAACGATCTCAAGAAGATCATCTTATTAATCAGATAAAATATTTCT AGATTTCAGTGCAATTTATCTCTTCAAATGTAGCACCTGAAGTCAGCCCCAT ACGATATAAGTTGTAATTCTCATGTTTGACAGCTTATCATCGATAAGCTTCC GATGGCGCGCCGAGAGGCTTTACACTTTATGCTTCCGGCT

Table B.3: Oligonucleotides used in this study. Below is a table of oligonucleotides used during the in-cell SHAPE-Seq experiments with the platform described in Figure B.1 and the endogenously expressed RNAs. Primers used during platform construction are not included. This table is mainly meant to serve as a reference. Abbreviations within primer sequences are as follows: '/5Biosg/' is a 5' biotin moiety, '/5Phos/' is a 5' monophosphate group, '/3SpC3/' is a 3' 3-carbon spacer group, VIC and NED are fluorophores (ABI), and asterisks indicate a phosphorothioate backbone modification. These abbreviations were used for compatibility with the Integrated DNA Technologies ordering notation.

Name	Sequence	Abbr.
Reverse Transcription		
ECK120051404 Terminator (ECK404)	/5Biosg/TTTATCGGCCGAAGCAGGTAG	A
Super Folder GFP (SFGFP)	/5Biosg/CAACAAGAATTGGGACAACTCCAGTG	B
5S rRNA (*E. coli*)	ATGCCTGGCAGTTCCCTA	C
RNase P specificity region (*E. coli*)	CCGTACCTTATGAACCCCTATTTGG	D
btuB riboswitch 5 UTR (*E. coli*)	GCATCCACAATAGAAGAAGGATGC	E
Adapter Ligation		
A_adapter_b (A_b) (ssDNA adapter)	/5Phos/AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC/3SpC3/	F
Fluorescent Quality Analysis		
Reverse QA primer (+)	VIC-GTGACTGGAGTTCAGACAAGCAGAACGTGTGCTC	G
Reverse QA primer (-)	NED-GTGACTGGAGTTCAGACAAGCAGAACGTGTGCTC	H
Primers for Building dsDNA Libraries		
PE_forward†	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	I
ECK404 (+) selection primer (forward)	CTTTCCCTACACGACGCTCTTCCGATCTRRRYTTTATCGGCCGAAGCAGGTAgA*G*G*C	J
ECK404 (-) selection primer (forward)	CTTTCCCTACACGACGCTCTTCCGATCTYYYRtTTATCGGCCGAAGCAGGTAgA*G*G*C	K
SF-GFP (+) selection primer (forward)	CTTTCCCTACACGACGCTCTTCCGATCTRRRYCAACAAGAATTGGGACAACTCCAGT*G*A*A*A*A*G	L

Table B.3 (Continued)

Name	Sequence	Abbr.
SF-GFP (-) selection primer (forward)	CTTTCCCTACACGACGCTCTTCCGATCTYYYRCA ACAAGAATTGGGACAACTCCAGT*G*A*A*A*A*G	M
5S rRNA (+) selection primer (forward)	CTTTCCCTACACGACGCTCTTCCGATCTRRRYAT GCCTGGCAGTTCCCTA*C*T*C	N
5S rRNA (-) selection primer (forward)	CTTTCCCTACACGACGCTCTTCCGATCTYYYRAT GCCTGGCAGTTCCCTA*C*T*C	O
RNase P (+) selection primer (forward)	CTTTCCCTACACGACGCTCTTCCGATCTRRRYCC GTACCTTATGAACCCCTATTTGG*C*C*T	P
RNase P (-) selection primer (forward)	CTTTCCCTACACGACGCTCTTCCGATCTYYYRCC GTACCTTATGAACCCCTATTTGG*C*C*T	Q
btuB (+) selection primer (forward)	CTTTCCCTACACGACGCTCTTCCGATCTRRRYGC ATCCACAATAGAAGAAGGATGC*C*G*C*A	R
btuB (-) selection primer (forward)	CTTTCCCTACACGACGCTCTTCCGATCTYYYRGC ATCCACAATAGAAGAAGGATGC*C*G*C*A	S
Illumina Multiplexing Primers		
Illumina Index #1†	CAAGCAGAAGACGGCATACGAGATCGTGATGT GACTGGAGTTCAGACGTGTGCTC	T
Illumina Index #2†	CAAGCAGAAGACGGCATACGAGATACATCGGT GACTGGAGTTCAGACGTGTGCTC	U
Illumina Index #3†	CAAGCAGAAGACGGCATACGAGATGCCTAAGT GACTGGAGTTCAGACGTGTGCTC	V
Illumina Index #4†	CAAGCAGAAGACGGCATACGAGATTGGTCAGT GACTGGAGTTCAGACGTGTGCTC	W
Illumina Index #5†	CAAGCAGAAGACGGCATACGAGATCACTGTGT GACTGGAGTTCAGACGTGTGCTC	X
Illumina Index #6†	CAAGCAGAAGACGGCATACGAGATATTGGCGT GACTGGAGTTCAGACGTGTGCTC	Y
Illumina Index #7†	CAAGCAGAAGACGGCATACGAGATGATCTGGT GACTGGAGTTCAGACGTGTGCTC	Z
Illumina Index #8†	CAAGCAGAAGACGGCATACGAGATTCAAGTGT GACTGGAGTTCAGACGTGTGCTC	AA
Illumina Index #9†	CAAGCAGAAGACGGCATACGAGATCTGATCGT GACTGGAGTTCAGACGTGTGCTC	AB
Illumina Index #10†	CAAGCAGAAGACGGCATACGAGATAAGCTAGT GACTGGAGTTCAGACGTGTGCTC	AC

Table B.3 (Continued)

Name	Sequence	Abbr.
Illumina Index #11†	CAAGCAGAAGACGGCATACGAGATGTAGCCGT GACTGGAGTTCAGACGTGTGCTC	AD
Illumina Index #12†	CAAGCAGAAGACGGCATACGAGATTACAAGGT GACTGGAGTTCAGACGTGTGCTC	AE
Illumina Index #13†	CAAGCAGAAGACGGCATACGAGATTTGACTGT GACTGGAGTTCAGACGTGTGCTC	AF
Illumina Index #14†	CAAGCAGAAGACGGCATACGAGATGGAACTGT GACTGGAGTTCAGACGTGTGCTC	AG
Illumina Index #15†	CAAGCAGAAGACGGCATACGAGATTGACATGT GACTGGAGTTCAGACGTGTGCTC	AH
Illumina Index #16†	CAAGCAGAAGACGGCATACGAGATGGACGGGT GACTGGAGTTCAGACGTGTGCTC	AI
Illumina Index #18†	CAAGCAGAAGACGGCATACGAGATGCGGACGT GACTGGAGTTCAGACGTGTGCTC	AJ
Illumina Index #19†	CAAGCAGAAGACGGCATACGAGATTTTCACGT GACTGGAGTTCAGACGTGTGCTC	AK
Illumina Index #20†	CAAGCAGAAGACGGCATACGAGATGGCCACGT GACTGGAGTTCAGACGTGTGCTC	AL
Illumina Index #21†	CAAGCAGAAGACGGCATACGAGATCGAAACGT GACTGGAGTTCAGACGTGTGCTC	AM
Illumina Index #22†	CAAGCAGAAGACGGCATACGAGATCGTACGGT GACTGGAGTTCAGACGTGTGCTC	AN
Illumina Index #23†	CAAGCAGAAGACGGCATACGAGATCCACTCGT GACTGGAGTTCAGACGTGTGCTC	AO
Illumina Index #25†	CAAGCAGAAGACGGCATACGAGATATCAGTGT GACTGGAGTTCAGACGTGTGCTC	AP
Illumina Index #27†	CAAGCAGAAGACGGCATACGAGATAGGAATGT GACTGGAGTTCAGACGTGTGCTC	AQ

Table B.4: RMDB data deposition table. SHAPE-Seq reactivity spectra generated in this work are freely available from the RNA Mapping Database (RMDB) (`http://rmdb.stanford.edu/repository/`) [169], accessible using the RMDB ID numbers indicated in the table below. Any other reactivity data from this work can be provided upon request.

RMDB ID	**Contents**	**Figure(s) used in**
CRR10_1M7_0001	Triplicate data of crR10 co-expressed with the antisense control plasmid	Figures 3.3, B.5, B.6, B.8 and B.10
CRR10_1M7_0002	Triplicate data of crR10 co-expressed with taR10	Figures 3.3, B.8 and B.10
CRR12_1M7_0001	Triplicate data of crR12 co-expressed with the antisense control plasmid	Figures 3.2, 3.3, B.5, B.10 and B.22
CRR12_1M7_0002	Triplicate data of crR12 co-expressed with taR12	Figures 3.3, B.10 and B.22
TAR10_1M7_0001	Triplicate data of taR10 expressed alone	Figures B.5, B.6 and B.9
TAR12_1M7_0001	Triplicate data of taR12 expressed alone	Figures 3.2, B.5, B.9 and B.22
RNAINS3_1M7_0001	Triplicate data of RNA-IN S3 expressed alone	Figures B.12 and B.15
RNAINS4_1M7_0001	Triplicate data of RNA-IN S4 expressed alone	Figures 3.4 and B.14
INS4DBL_1M7_0001	Triplicate data of RNA-IN S4 C24A A25C double mutant co-expressed with the antisense control plasmid	Figures 3.4, B.18 and B.23
INS4DBL_1M7_0002	Triplicate data of RNA-IN S4 C24A A25C double mutant co-expressed with RNA-OUT A4	Figures 3.4, B.18 and B.23
RNAOUT3_1M7_0001	Triplicate data of RNA-OUT A3 expressed alone	Figures B.12 and B.23
RNAOUT4_1M7_0001	Triplicate data of RNA-OUT A4 expressed alone	Figures 3.4 and B.19
5SRRNA_1M7_0007	Triplicate data of endogenous *E. coli* 5S rRNA	Figures 3.5, B.21 and B.24
BTUBR_1M7_0001	Five replicates of endogenous *E. coli* btuB riboswitch leader sequence	Figures 3.5 and B.21
RNASEP_1M7_0001	Triplicate data of endogenous *E. coli* RNase P specificity region	Figures 3.5 and B.21

B.3 Supplementary Figures

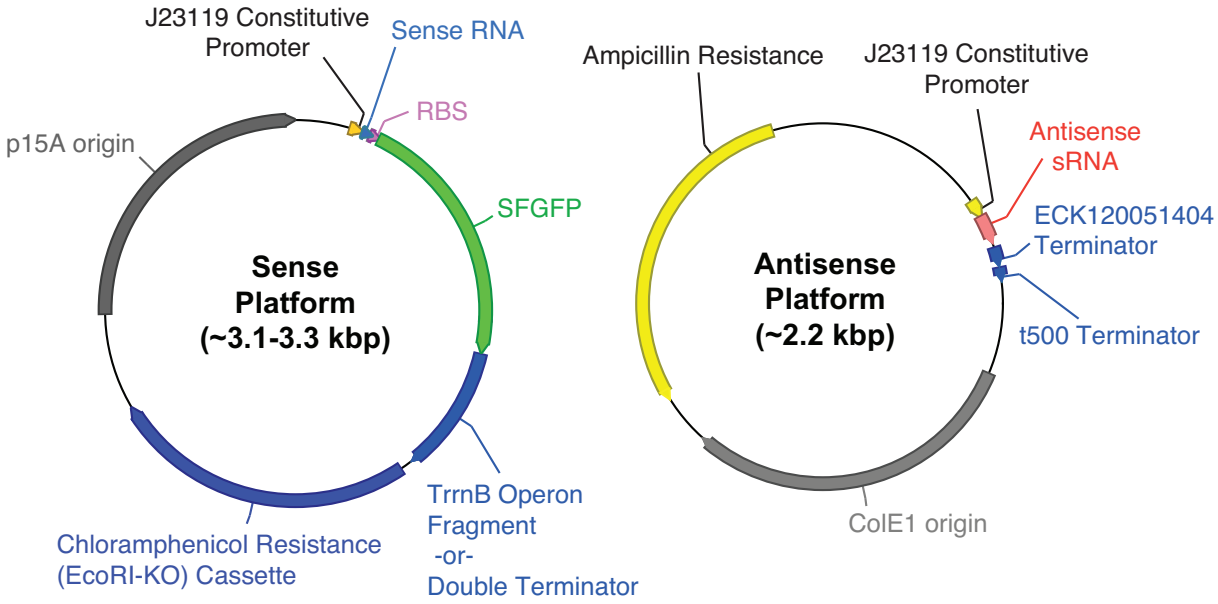
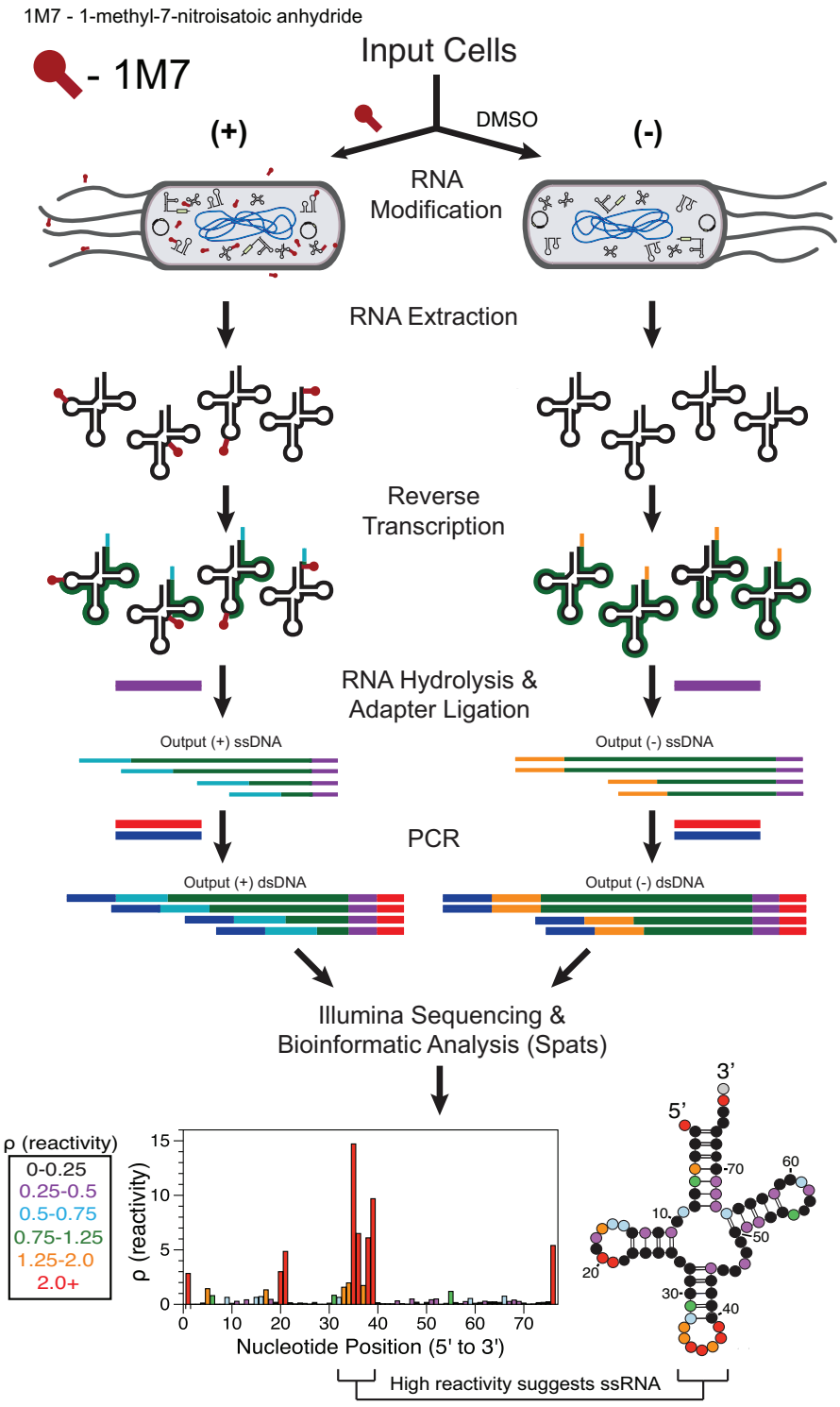


Figure B.1: Standardized platform for expressing sense/antisense regulatory RNA pairs in *E. coli*. Sense RNAs that control the translation of a downstream superfolder GFP (SFGFP) sequence [184] are expressed in *E. coli* using a constitutive promoter (σ^{70}) as part of the Sense Platform. The Antisense Platform similarly expresses an antisense RNA that can target the sense RNA. The origins of replications were selected such that the antisense is always in molar excess of the sense to facilitate RNA binding. Specific sequences of RNA regulators are found in Table B.2. Note that the sense platform shown below is for the riboregulators crR10 and crR12. The TrrnB operon fragment of the sense platform is replaced with a double terminator for the RNA-IN/OUT system. The Antisense Platform contains the ECK120051404 and t500 intrinsic terminators [186]. The control antisense plasmid lacks the antisense sRNA sequence, but is otherwise the same as the Antisense Platform vector. A selection of the plasmids used in this study was deposited in the addgene database and can be found by searching for this paper. Other plasmids can be provided upon request.

Figure B.2: In-cell SHAPE-Seq structural characterization overview. To perform in-cell SHAPE-Seq, cells are grown (potentially with a plasmid, or combination of plasmids, containing the RNA(s) of interest) then split for fluorescence measurement and SHAPE probing (Figure 3.1). Cultures for SHAPE probing are subjected to modification with 1M7 (+) or a DMSO control (-). Subsequent RNA extraction, reverse transcription (halting at modifications), and PCR prepare the cDNA fragments for next-generation sequencing. Bioinformatic analysis of sequenced reads with Spats (`http://spats.sourceforge.net/`) generates reactivity maps representing the cellular flexibility of each nucleotide in an RNA [97, 98]. $tRNA^{phe}$ from *E. coli* is shown as a hypothetical example.



E. coli cells are subcultured for 3+ hours

1M7 modifies RNAs in-cell in a structurally dependent manner (favors ssRNA)

RNAs are extracted with TRIzol (tRNA example)

Reverse transcription halts at modifications

cDNA is isolated and ligated to ssDNA adapter

PCR adds Illumina sequencing adapters

Illumina sequencing quantitatively determines number and location of modifications

Reactivity maps represent modification distributions (tRNA example)

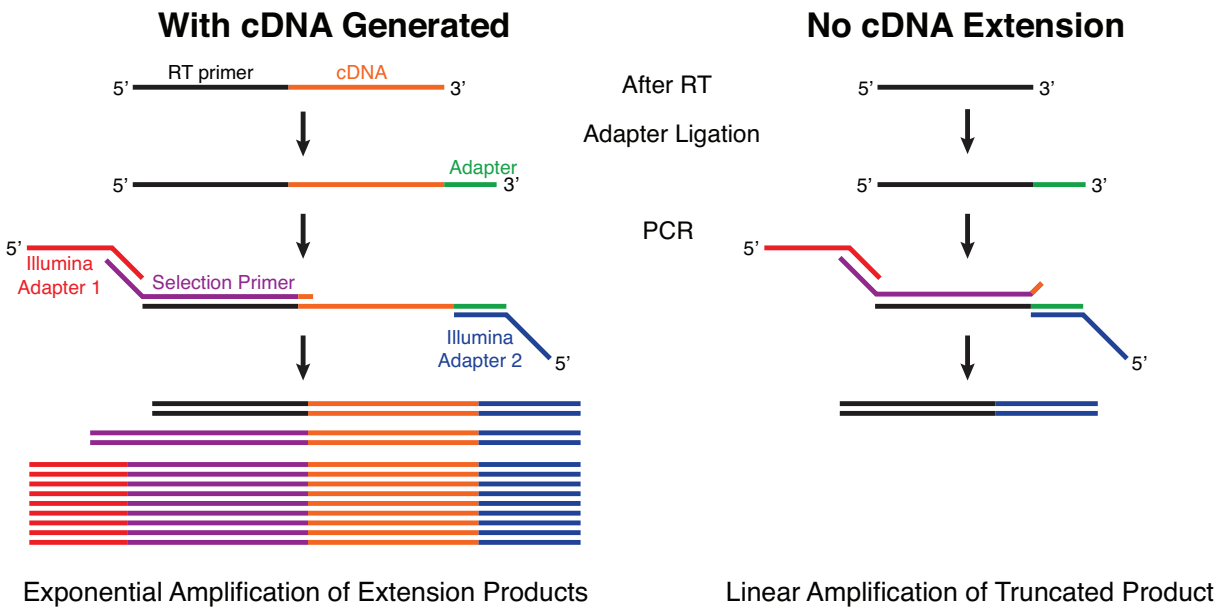


Figure B.3: Selective PCR amplification strategy for cDNA libraries. The cDNA generated from the reverse transcription step of in-cell SHAPE-Seq is ligated to an adapter that contains part of the sequence required for Illumina TruSeq barcoding (Figure B.2). To amplify the ssDNA library, primers containing the rest of the sequences for Illumina TruSeq barcoding are added to a PCR reaction (red, blue), along with a selection primer (purple). The selection primer contains several nucleotides on its 3′ end that extend into the expected cDNA sequence. In this way, unextended RT primer ligated to excess adapter (dimer side product) is unable to extend due to a 3′ mismatch (right). This selective PCR greatly reduces the amount of dimer side product that is amplified and removes the need for gel extraction.

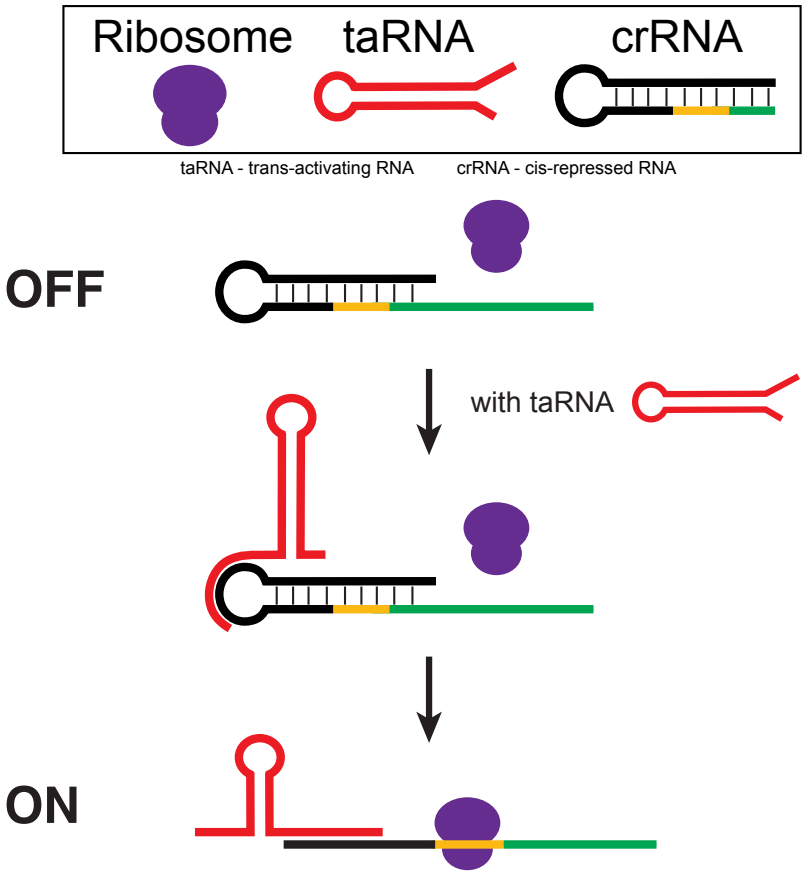


Figure B.4: Mechanism of the synthetic translation-activating riboregulator system. In the riboregulator system, the cis-repressed sense RNA (crRNA) is designed to form a hairpin structure that occludes the RBS, blocking translation. The riboregulator is thus in the OFF state when expressed by itself. A trans-activating antisense RNA (taRNA) is designed to base pair with the 5′ region of the crRNA through a loop-linear interaction intermediate to expose the RBS on the crRNA and allow translation. The riboregulator switches to the ON state in the presence of a cognate taRNA [25].

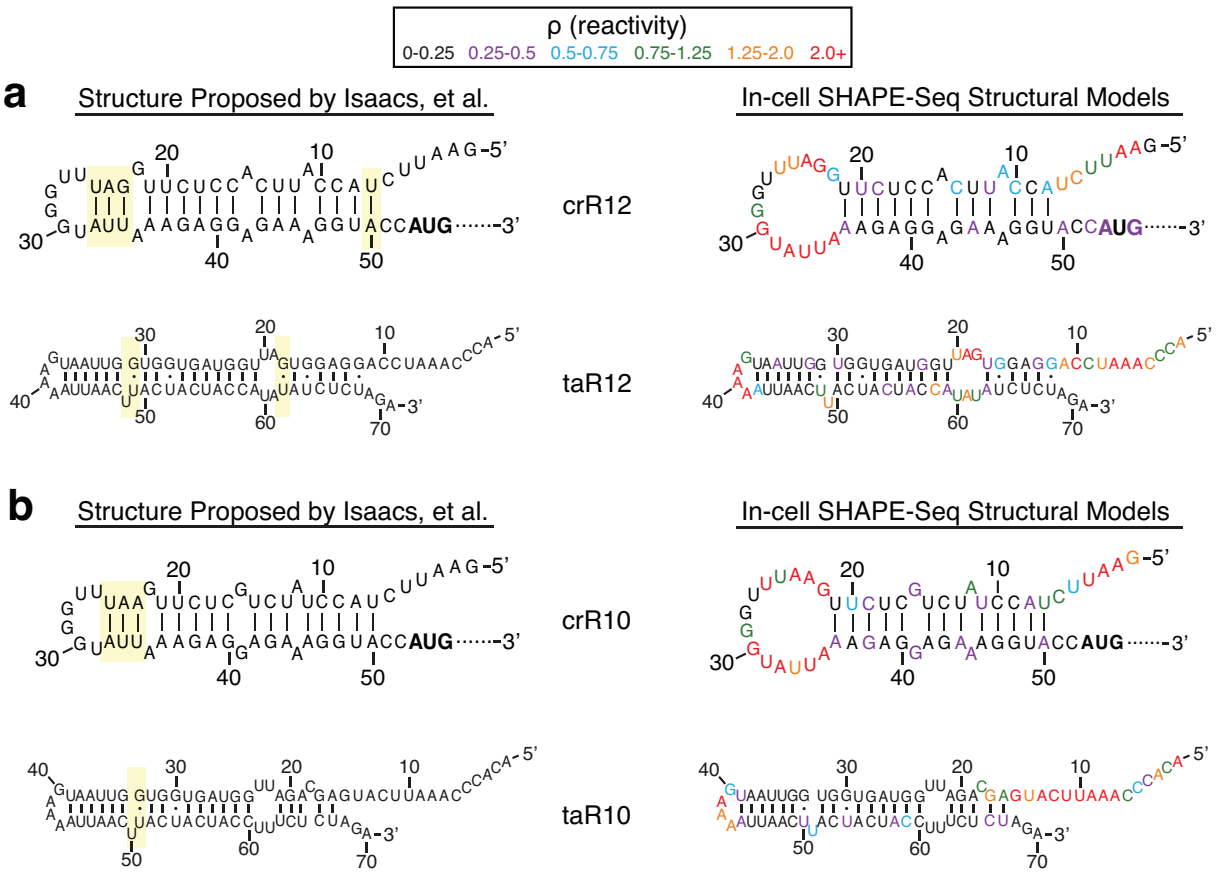


Figure B.5: Structural analysis of the synthetic riboregulator translational activation system using in-cell SHAPE-Seq data. Structures originally designed by Isaacs *et al.* [25] for riboregulator sets taR12/crR12 and taR10/crR10 compared to models of the cellular structural states generated from in-cell SHAPE-Seq reactivities. The reactivities were used to constrain the secondary structure prediction program RNAstructure [164]. Nucleotide locations where the structures differ are highlighted in yellow (left). (a) Constraining the crR12 fold with the average reactivities calculated from in-cell SHAPE-Seq indicates that the apical loop may be larger than originally predicted in cells. There are few differences between the predicted and constrained folds for the highly structured taR12: the U bulge near position 50 is in a different position in the cell, and the inner loop near position 20 is predicted to be larger in cells. (b) The same analysis for taR10/crR10. This analysis also found that the apical loop of crR10 was likely less structured in the cell than originally predicted and the same U bulge in taR10 near position 50 was in a different location. However, the taR10 inner loop near positions 22-23 matched the structures predicted by Isaacs *et al.* [25, 164].

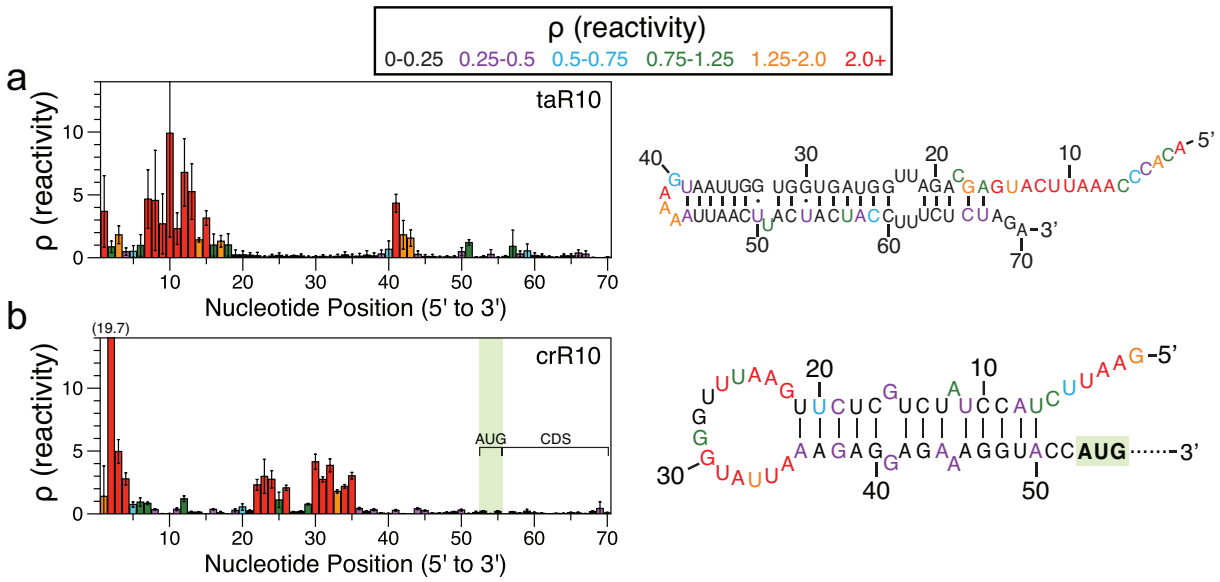


Figure B.6: Characterization of the cellular structures of the taR10/crR10 synthetic riboregulator RNA translational activator system. Reactivity maps and constrained secondary structure folds are shown for both taR10 (a) and crR10 (b) of the synthetic riboregulator activator system. Color-coded reactivity spectra represent averages over three independent in-cell SHAPE-Seq experiments, with error bars representing one standard deviation of the reactivities at each position. RNA structures represent minimum free energy structures generated by RNAstructure [25, 164] using in-cell SHAPE-Seq reactivity data as constraints (see Section 3.3). Nucleotides are color-coded by reactivity intensity. Comparisons to original structural designs from Isaacs *et al.* [25] are shown in Figure B.5. The crR10 structural model was generated from the first 70 nts of the sequence (55 nt shown), and the terminators following taR10 were not included in the structural analysis. The start codon (AUG) location is boxed and the coding sequence (CDS) is labeled.

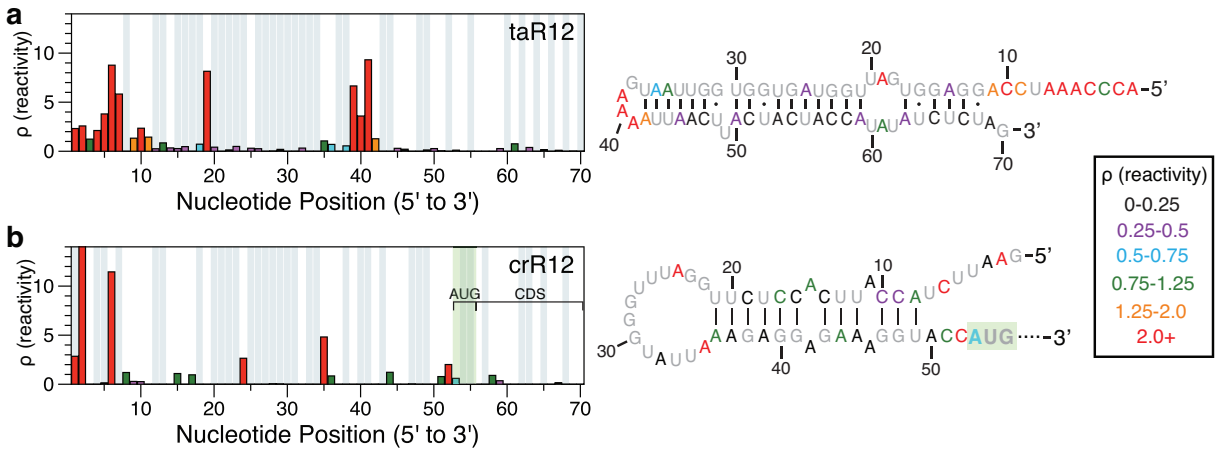


Figure B.7: Characterization of the cellular structures of the taR12/crR12 synthetic riboregulator RNA translational activator system using in-cell dimethyl sulfate (DMS) probing. Reactivity maps for an in-cell dimethyl sulfate (DMS) probing experiment (see Section 3.3) are shown for taR12 (a) and crR12 (b) of the synthetic riboregulator system. Color-coded reactivity spectra represent the reactivity level at each position, calculated in the same way as in-cell SHAPE-Seq reactivity data. G and U positions are marked with gray, as DMS reacts with strong preference for As and Cs [125]. The structures presented are the same from Figure 3.2 with DMS reactivities overlaid and Gs and Us marked with gray. The start codon (AUG) in crR12 is boxed and the coding sequence (CDS) is labeled. In general, there is good agreement between the in-cell DMS reactivities shown here and the in-cell SHAPE reactivities displayed in Figure 3.2 at A and C positions.

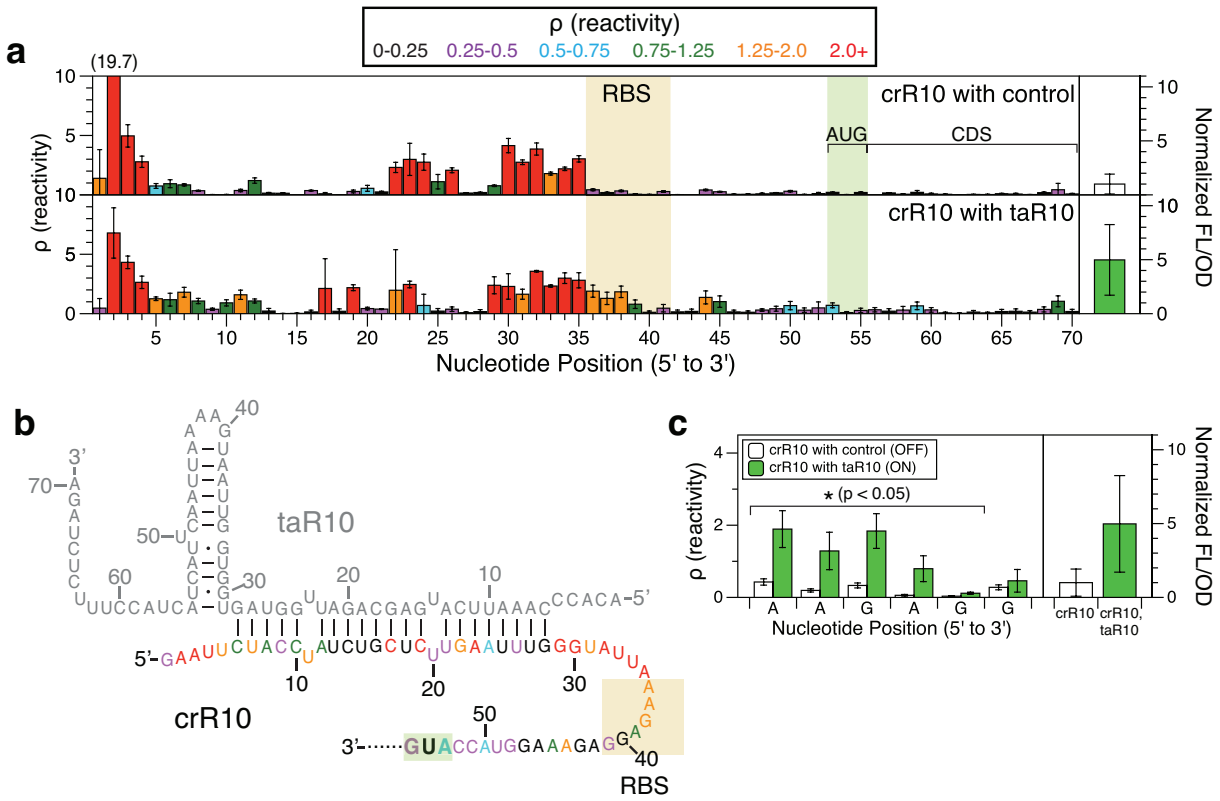


Figure B.8: Structure-function characterization of the taR10/crR10 synthetic riboregulator RNA translational activator system. Reactivity maps (a) and a suggested RNA-RNA interaction structure with taR10 (b) are shown for crR10 of the synthetic riboregulator activator system. (a) Color-coded reactivity spectra for crR10 expressed in conjunction with the control plasmid or taR10. Reactivities represent averages over three independent in-cell SHAPE-Seq experiments, with error bars representing one standard deviation. Average fluorescence (FL/OD) values (normalized to crR10 expressed with the control antisense plasmid) on the right show a 5-fold activation of gene expression when taR10 interacts with crR10, with error bars representing one standard deviation. The RBS (determined in Figure B.10) and start codon (AUG) locations are boxed. (b) Structural model of the crR10/taR10 binding complex derived from the mechanism proposed by Isaacs *et al.* [25], refined using the average crR10 reactivities with taR10 present in (a). Nucleotides for crR10 are color coded by reactivity intensity. (c) Reactivity and functional data of the RBS region, showing an increase in RBS reactivity (left) and fluorescence (right) when taR10 is co-expressed with crR10. Positions that are statistically significantly different ($p < 0.05$) according to a one-sided Welchs t-test are indicated with *.

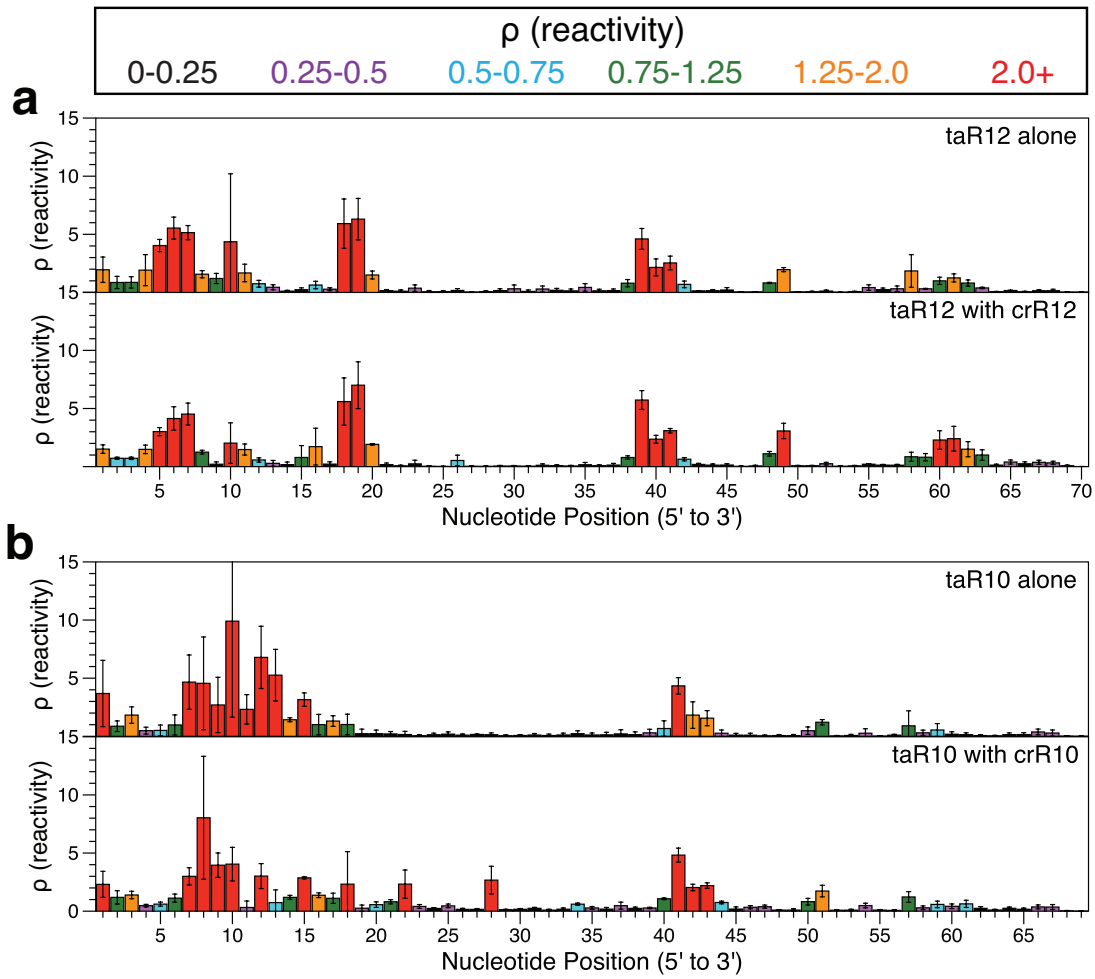
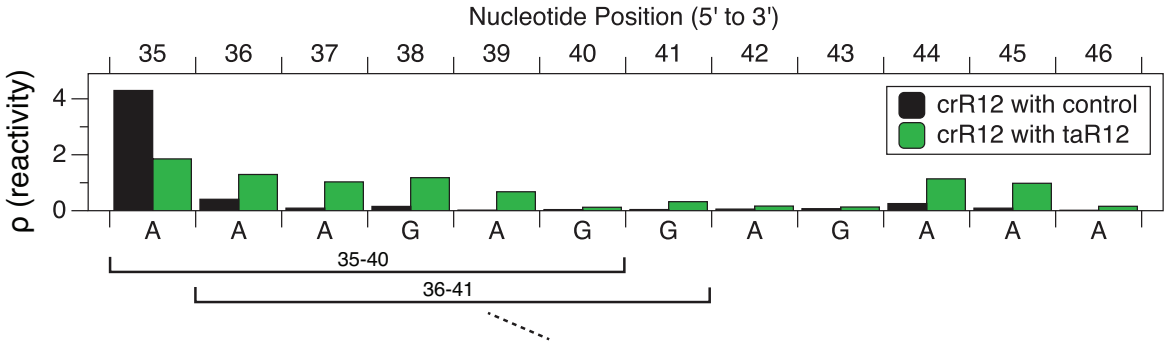


Figure B.9: In-cell SHAPE-Seq reactivities for the trans-activating RNA variants. In-cell SHAPE-Seq reactivity spectra for variants taR12 and taR10 of the trans-activating RNAs from the riboregulator system [25]. Color-coded reactivity spectra represent averages over three independent in-cell SHAPE-Seq experiments, with error bars representing one standard deviation of the reactivities at each position. (a) The reactivity spectra for the taR12 variant with and without its cognate crR12 show clusters of high reactivity in the GAAA loop (nt 39-43) and at the 5′ end, which is predicted to be single-stranded when expressed alone (Figure B.5) and interact with the crR12 apical loop when it is expressed with crR12. Yet, no major changes in reactivity are observed when crR12 is present, likely due to the higher copy number of taR12 in the cell relative to crR12. (b) Reactivity spectra for the taR10 variant with and without the cognate crR10. The pattern is similar to that of taR12 above and also shows little change in the presence of the sense RNA crR10. The GAAA tetraloop is in between nucleotides 40-44 in taR10.

Figure B.10: Determining the dominant RBS sequence in crRNAs using in-cell SHAPE-Seq reactivities. Each crRNA contains an AG-rich region with the potential to contain multiple six-nucleotide Shine-Dalgarno (SD) sequences [420]. (a) To determine which six nucleotides comprise the dominant SD sequence in crR12, a sliding window was used to analyze reactivity changes in the AG-rich region when crR12 is interacting with taR12. The reactivity of each window was calculated by summing the reactivities for the six nucleotides in the window. Windows were calculated separately for the OFF (crR12 with control plasmid) and ON (crR12 and taR12) states. The differences between the ON and OFF state for each window were used to determine which window displayed the largest change in reactivity. Nucleotides 36-41 of crR12 demonstrated the largest increase in reactivity (3.88) and are one nucleotide away from the consensus AGGAGG, suggesting 36-41 is likely the dominant SD sequence. (b) The same analysis was used to find the dominant SD sequence for crR10. The 36-41 nucleotide window was again found to have the highest increase in reactivity (5.07).

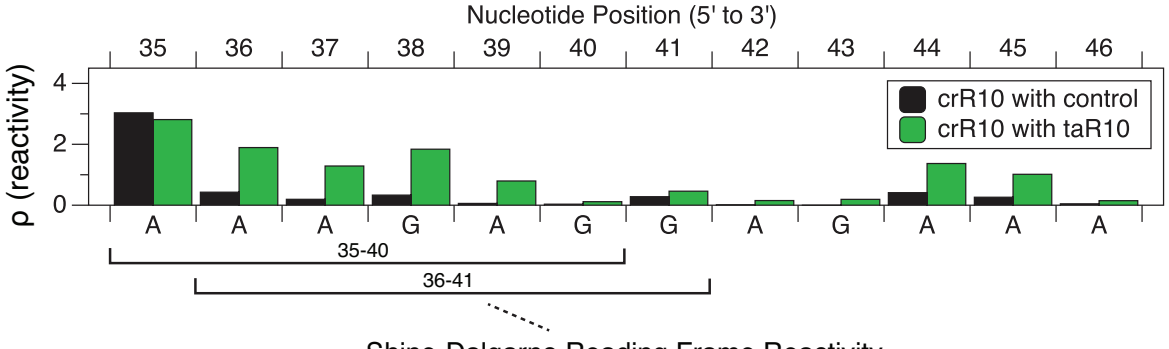
a



Shine-Dalgarno Reading Frame Reactivity

	35-40	36-41	37-42	38-43	39-44	40-45	41-46
crR12 with control	5.01	0.75	0.40	0.38	0.48	0.55	0.53
crR12 with taR12	6.16	4.63	3.50	2.60	2.56	2.86	2.90
Difference in ρ	1.15	**3.88**	3.10	2.22	2.08	2.31	2.37

b



Shine-Dalgarno Reading Frame Reactivity

	35-40	36-41	37-42	38-43	39-44	40-45	41-46
crR10 with control	4.08	1.32	0.91	0.72	0.80	1.00	1.01
crR10 with taR10	8.74	6.39	4.65	3.55	3.08	3.30	3.34
Difference in ρ	4.66	**5.07**	3.74	2.84	2.29	2.30	2.33

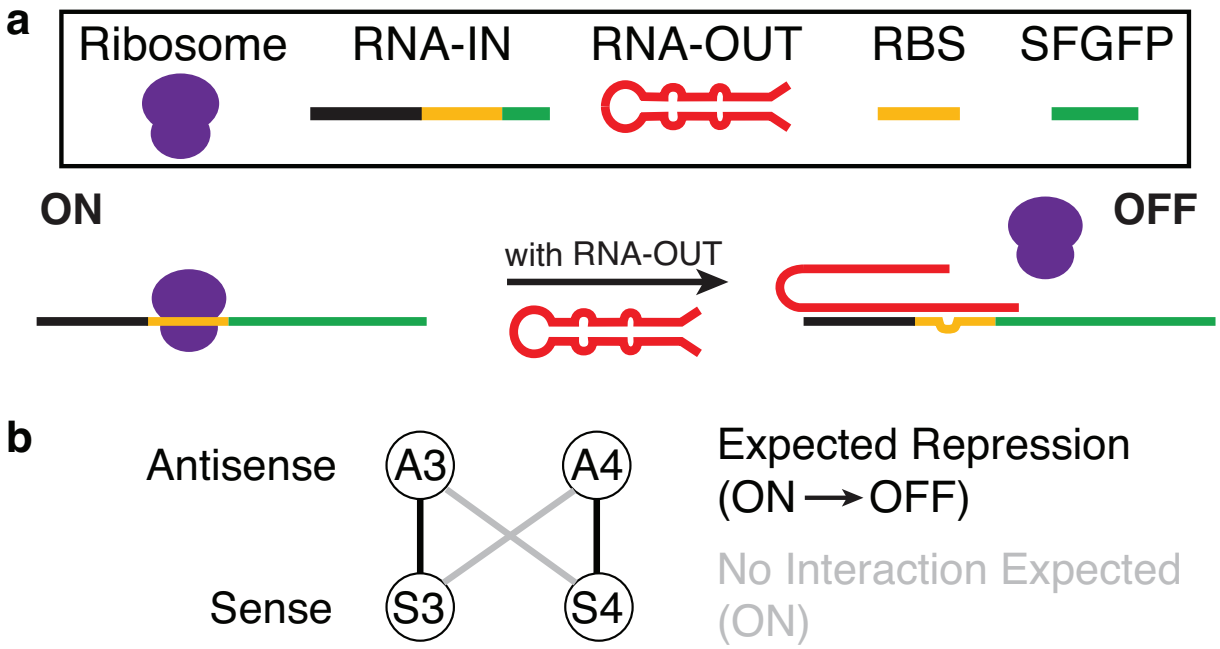


Figure B.11: Mechanism of the translation-repressing RNA-IN/OUT system. (a) In the RNA-IN/OUT system modified by Mutalik *et al.* [176], the sense RNA (RNA-IN) is expressed upstream of SFGFP and contains an exposed RBS that allows the translation of SFGFP in the ON state. When the antisense RNA (RNA-OUT) is present, RNA-IN is predicted to base-pair to the 5′ half of RNA-OUT, causing the RBS to become double-stranded and block the translation of SFGFP, switching to the OFF state. As depicted, RNA-OUT is predicted to be an extended stem-loop structure with two inner bulges [176]). (b) By mutating the interaction region between RNA-IN (first 5 nt of 5′ end) and RNA-OUT (apical loop), Mutalik *et al.* [176] generated many orthogonal, or independently acting, pairs of regulators. For example, the RNA-IN variant S3 will be repressed by A3, but the mutations between A3 and A4 do not allow A4 to significantly repress S3.

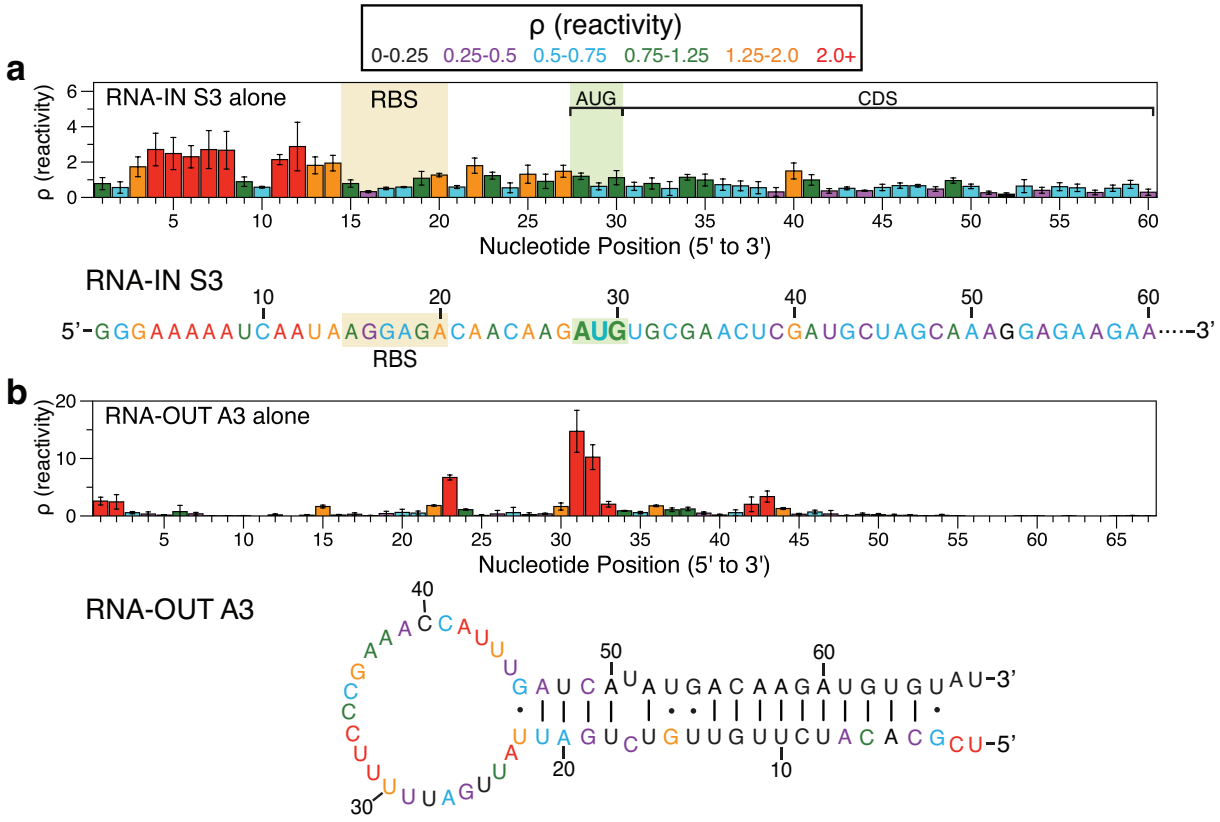
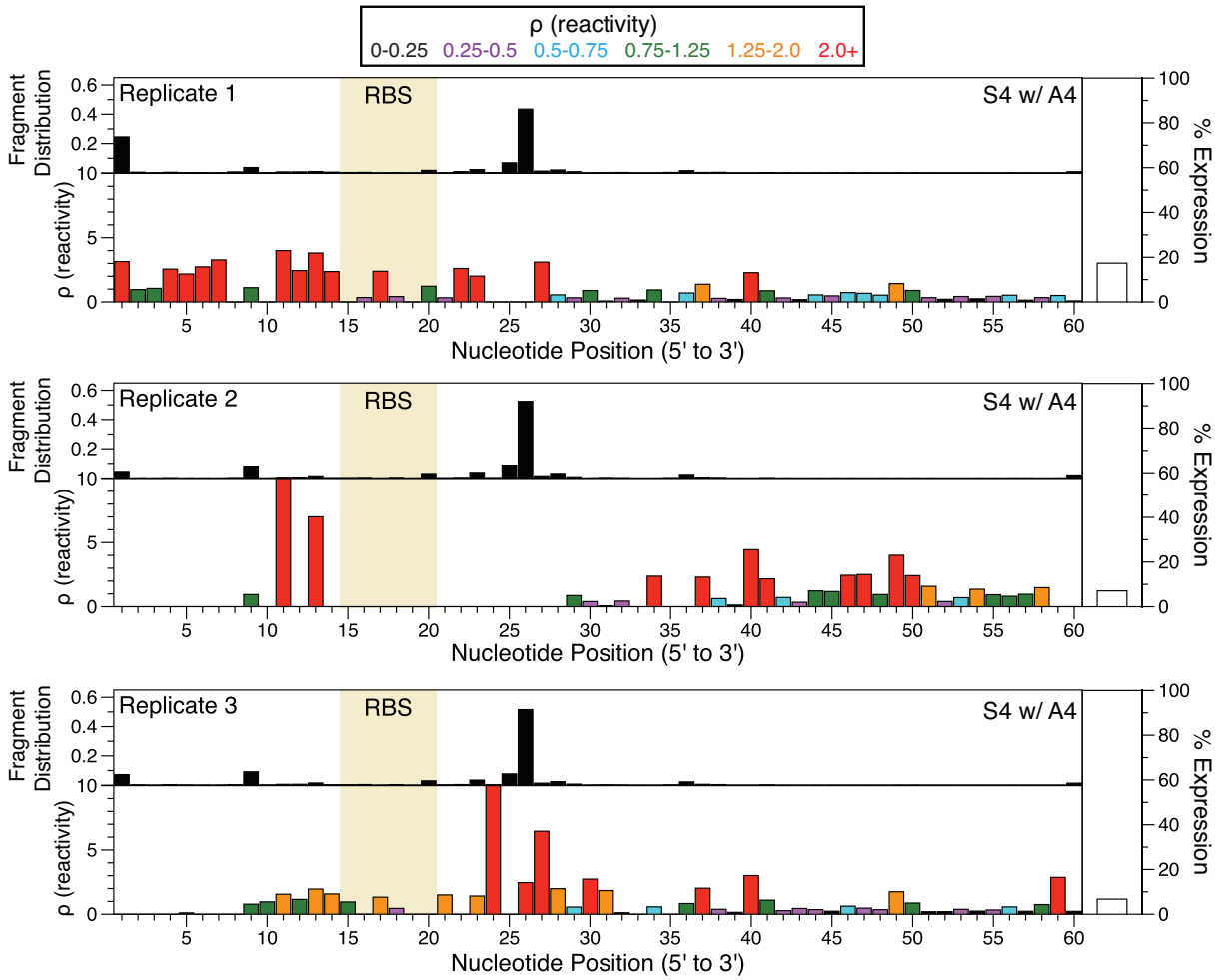


Figure B.12: Characterization of the cellular structures of the S3/A3 RNA-IN/OUT translational repressor system. Color-coded reactivity spectra are averaged over three independent in-cell SHAPE-Seq experiments, with error bars representing one standard deviation of the reactivities at each position. RNA structures represent minimum free energy structures generated by RNAstructure [164] using in-cell SHAPE-Seq reactivity data as constraints (see Section 3.3). Nucleotides are color-coded by reactivity intensity. (a) Reactivity spectrum of the first 60 nts of RNA-IN S3 (top), with nucleotides color-coded by reactivity on a single-stranded structural model of this region (bottom). RBS and start codon (AUG) are boxed. (b) Reactivity spectra (top) and reactivity-constrained secondary structure model (bottom) for RNA-OUT A3 (antisense). As observed with RNA-OUT A4 in Figure 4b, the interacting loop is predicted to be larger than originally proposed by Mutalik *et al.* [176, 191] when SHAPE constraints are applied. The terminators following RNA-OUT were not included in the structural analysis. CDS = coding sequence.

Figure B.13: RNA-IN/OUT S4/A4 interaction complexes appear to be cleaved by a double-stranded RNase in the cell. Reactivity maps (colored bars), (-) control fragment distributions (black bars), and percent expression (white bars) for three independent replicates of in-cell SHAPE-Seq experiments co-expressing both RNA-IN S4 and RNA-OUT A4. Despite all of the replicates having a similar level of translational repression, the reactivity maps have different patterns. We hypothesized that this was caused by the large peak aligning to nucleotide 26 in the (-) control fragment length distributions in each replicate. The wt RNA-IN/OUT interacting duplex is known to contain an RNase III cut site before nucleotide 15 in RNA-IN [176, 191], although this cut site was removed in our system by Mutalik *et al.* by mutating nucleotides 15 and 16 to GG, causing a local mismatch with RNA-OUT (see Figure B.16) [176, 191]. We hypothesized that there may be a second cut site for a double-stranded RNase between nucleotides 25 and 26, which would prevent reverse transcriptase from reading all the way to the 5′ end of RNA-IN. If true, this would suggest that our measurement would be probing two major states of the interacting RNAs between nucleotides 1-25 at once: the uncut and cut states. Variability in the relative level of dsRNA cleavage would then cause the variability in the reactivity maps.


ρ (reactivity)
0-0.25 0.25-0.5 0.5-0.75 0.75-1.25 1.25-2.0 2.0+
Replicate 1
RBS
S4 w/ A4
Fragment Distribution
ρ (reactivity)
% Expression
Nucleotide Position (5' to 3')
Replicate 2
RBS
S4 w/ A4
Fragment Distribution
ρ (reactivity)
% Expression
Nucleotide Position (5' to 3')
Replicate 3
RBS
S4 w/ A4
Fragment Distribution
ρ (reactivity)
% Expression
Nucleotide Position (5' to 3')

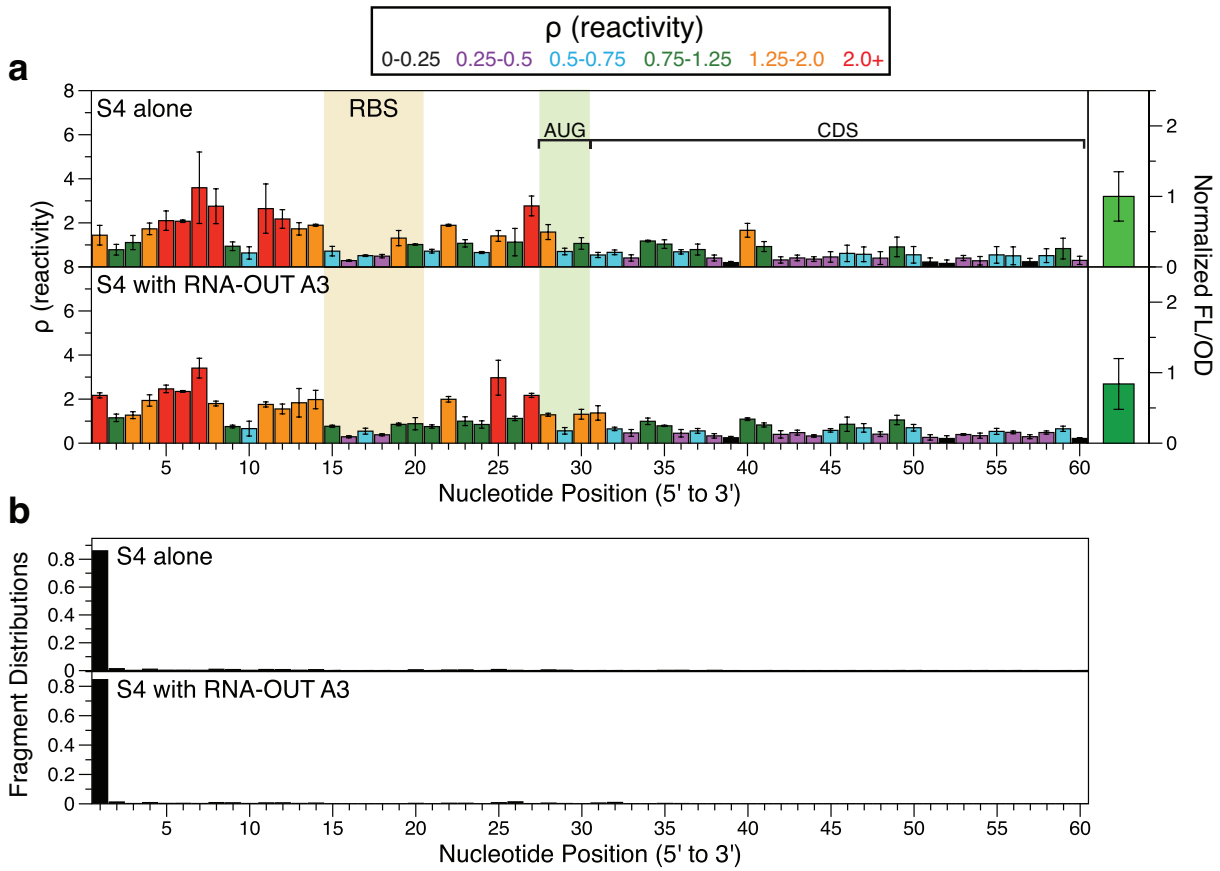


Figure B.14: Structure-function analysis of the S4/A3 interaction from the RNA-IN/OUT translational repressor. Color-coded reactivity spectra represent averages over three independent in-cell SHAPE-Seq experiments, with error bars representing one standard deviation of the reactivities at each position. (a) Reactivity maps and fluorescence (FL/OD) of RNA-IN S4 (normalized to average S4 alone fluorescence) expressed alone (top) and with RNA-OUT A3 (bottom). RBS and start codon (AUG) are boxed. Note that RNA-OUT A3 is orthogonal to S4 and should therefore not cause a change in gene expression. As expected, the reactivity patterns and fluorescence of S4 do not show major changes between the two conditions, indicating that the two RNAs are indeed not interacting. (b) (-) control fragment distributions for one replicate of the libraries sequenced to produce the reactivity maps in (a). Note that the degradation peak observed at nucleotide 26 (Figure B.13) is not observed for the non-interacting RNAs and thus is not due to the fact that both RNA-IN and RNA-OUT variants were co-expressed. CDS = coding sequence.

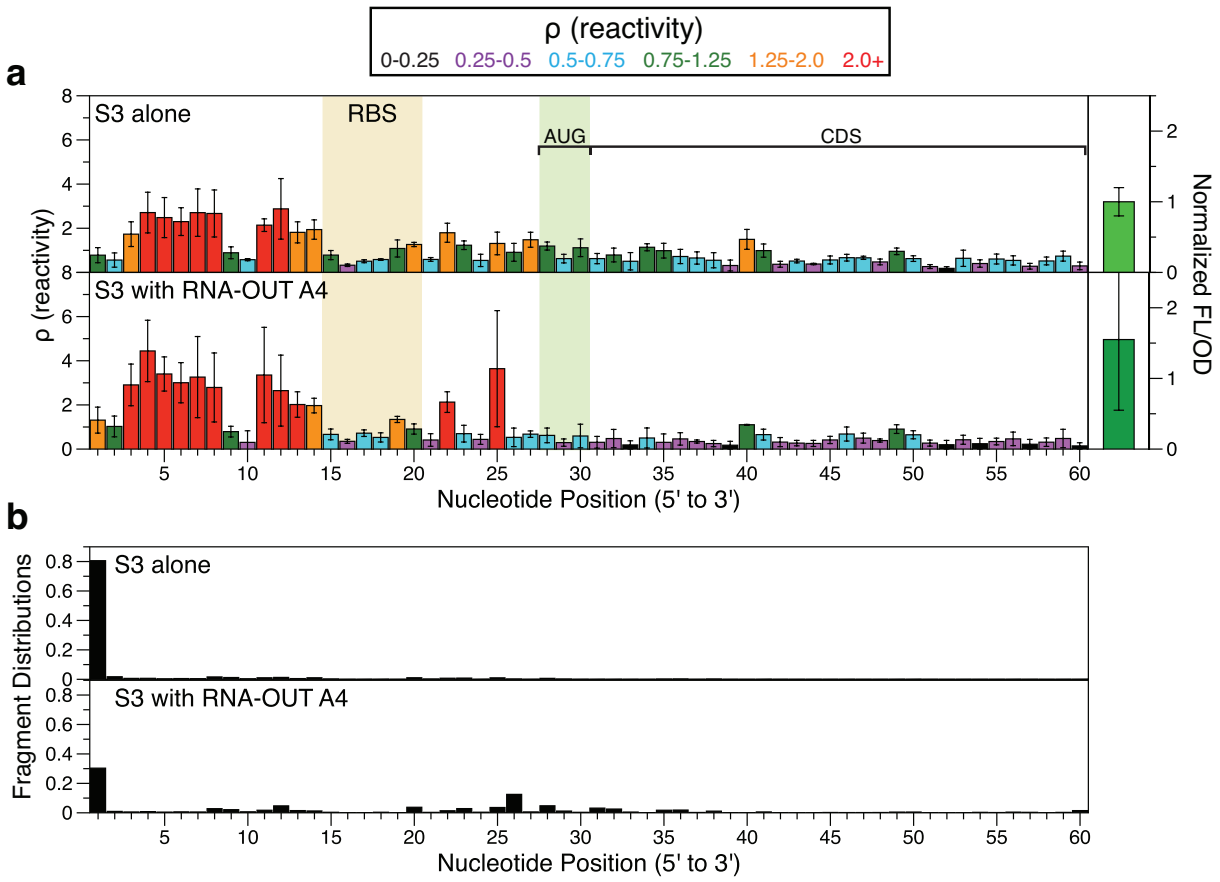


Figure B.15: Structure-function analysis of the S3/A4 interaction from the RNA-IN/OUT translational repressor. Color-coded reactivity spectra represent averages over three independent in-cell SHAPE-Seq experiments, with error bars representing one standard deviation of the reactivities at each position. (a) Reactivity maps and fluorescence (FL/OD) for RNA-IN S3 (normalized to average of S3 fluorescence) expressed alone (top) and with RNA-OUT A4 (bottom). RBS and start codon (AUG) are boxed. Unlike for the S4/A3 pair (Figure B.14), the S3/A4 pair shows small changes in the reactivity maps of RNA-IN S3 and in the expression of SFGFP in the presence of RNA-OUT A4, although both exhibit a higher level of noise. (b) (-) control fragment distributions for one replicate of the libraries sequenced to produce the reactivity maps in (a). The fragment distribution shown for the S3/A4 interaction is for a replicate deviating somewhat from the other two in terms of nucleotide reactivities and fluorescent output, generating most of the observed noise. CDS = coding sequence.

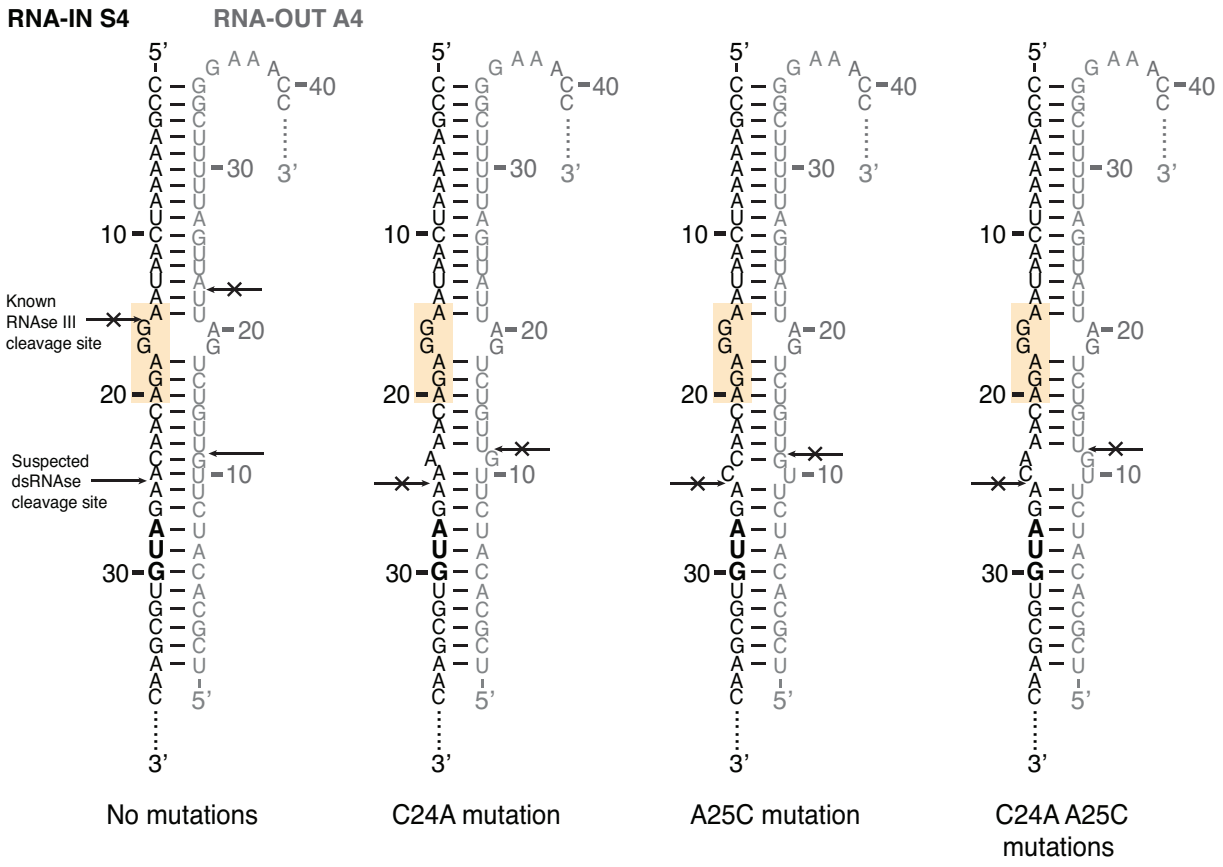
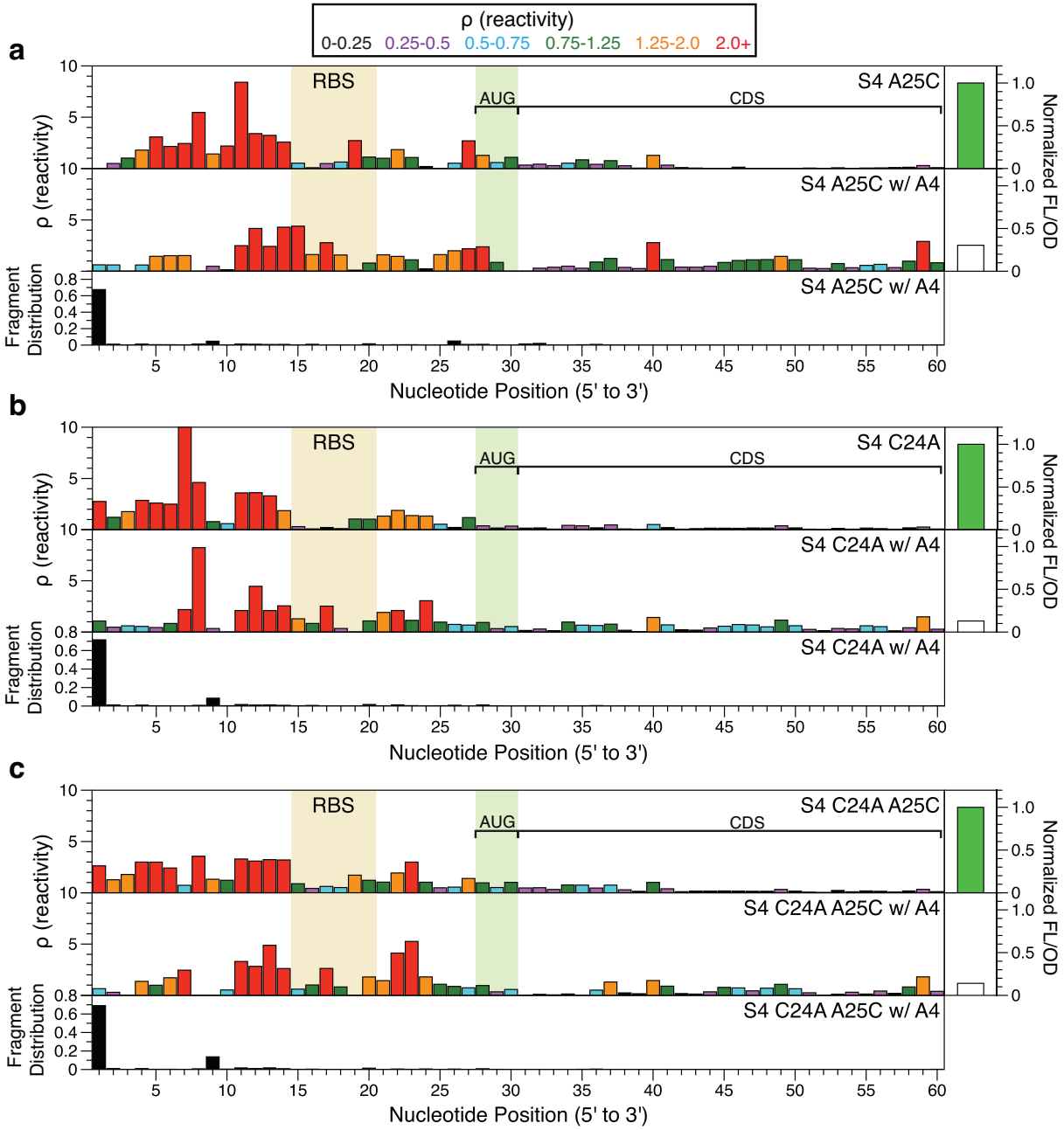


Figure B.16: RNA-IN mutations to resist RNase cleavage between nucleotides 25 and 26. The structure on the far left depicts the interaction between RNA-IN S4 and RNA-OUT A4 as proposed by Mutalik *et al.* after they blocked a known RNase III cleavage site with an interior loop [176, 191]. Suspecting that RNase III may also be involved in the observed cleavage between RNA-IN nucleotides 25 and 26 (Figure B.13), we mutated the positions that would most directly block RNase III cleavage between RNA-IN and RNA-OUT. We introduced the C24A and A25C mutations, tested them (Figure B.17), and ultimately characterized the double mutant in triplicate with in-cell SHAPE-Seq (Figures 3.4c and B.18).

Figure B.17: RNA-IN S4 mutations resist cleavage and maintain functionality. Color-coded reactivity spectra represent one independent in-cell SHAPE-Seq experiment with fluorescence (FL/OD) measurements on the right (ON state normalized to one). (a) The RNA-IN S4 A25C mutant exhibits 70% repression, which is lower than the original S4. However, mutation A25C does not exhibit a peak in the fragment distribution at nucleotide 26 in the presence of RNA-OUT A4, as was observed with the original S4 sequence in Figure B.13. Reactivities decrease near the 5′ end, and increase modestly between nucleotides 21-27 when S4 A25C interacts with A4. (b) RNA-IN S4 mutant C24A shows a similar reactivity decrease at the 5′ end as S4 A25C in (a) and the absence of a peak in the fragment distribution at nucleotide 26 in the presence of RNA-OUT A4. However, the C24A mutant exhibited better repression (87%). (c) The C24A A25C double mutant shows the same general features of (a) and (b), and was used for replicate in-cell SHAPE-Seq experiments (Figure 3.4c). RBS and start codon (AUG) are boxed. CDS = coding sequence.


ρ (reactivity)
0-0.25 0.25-0.5 0.5-0.75 0.75-1.25 1.25-2.0 2.0+
a
RBS
AUG
CDS
S4 A25C
S4 A25C w/ A4
S4 A25C w/ A4
ρ (reactivity)
Fragment Distribution
Normalized FL/OD
Nucleotide Position (5' to 3')
b
S4 C24A
S4 C24A w/ A4
S4 C24A w/ A4
c
S4 C24A A25C
S4 C24A A25C w/ A4
S4 C24A A25C w/ A4

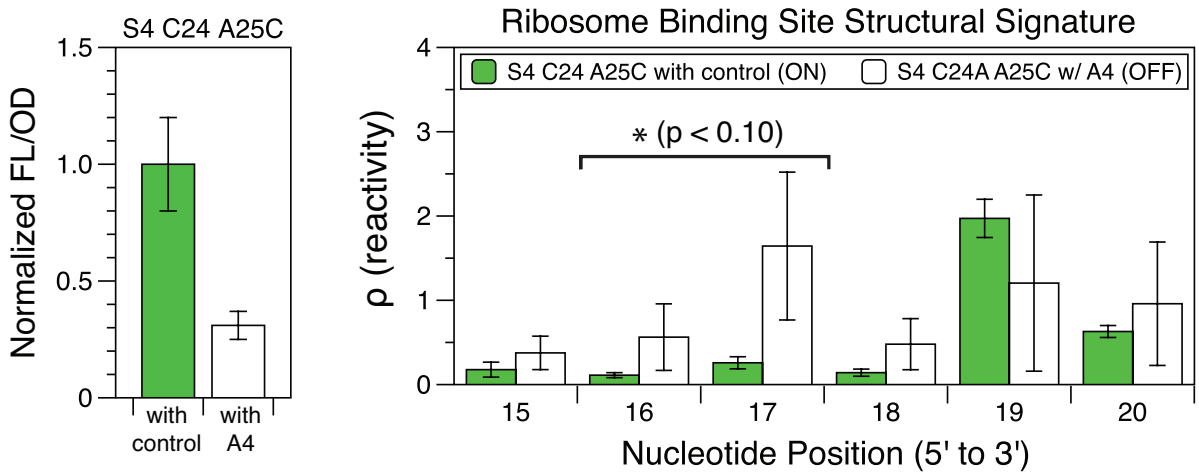


Figure B.18: Ribosome binding site analysis of RNA-IN S4 mutant C24A A25C. The fluorescence measured (FL/OD; left) and ribosome binding site (RBS) reactivities (right) of RNA-IN S4 mutant C24A A25C represent the average of three independent in-cell SHAPE-Seq experiments, with error bars representing one standard deviation. The S4 C24A A25C RBS nucleotides show significant reactivity changes in the presence of RNA-OUT A4 at nucleotides 16 and 17 ($p < 0.10$), which increase in response to antisense A4 binding. The increase in nucleotides 16 and 17 is expected due to the designed double bulge at the previously known RNase III cut site [176, 191]. Positions that are significantly different ($p < 0.10$) according to a one-sided Welchs t-test are indicated with *.

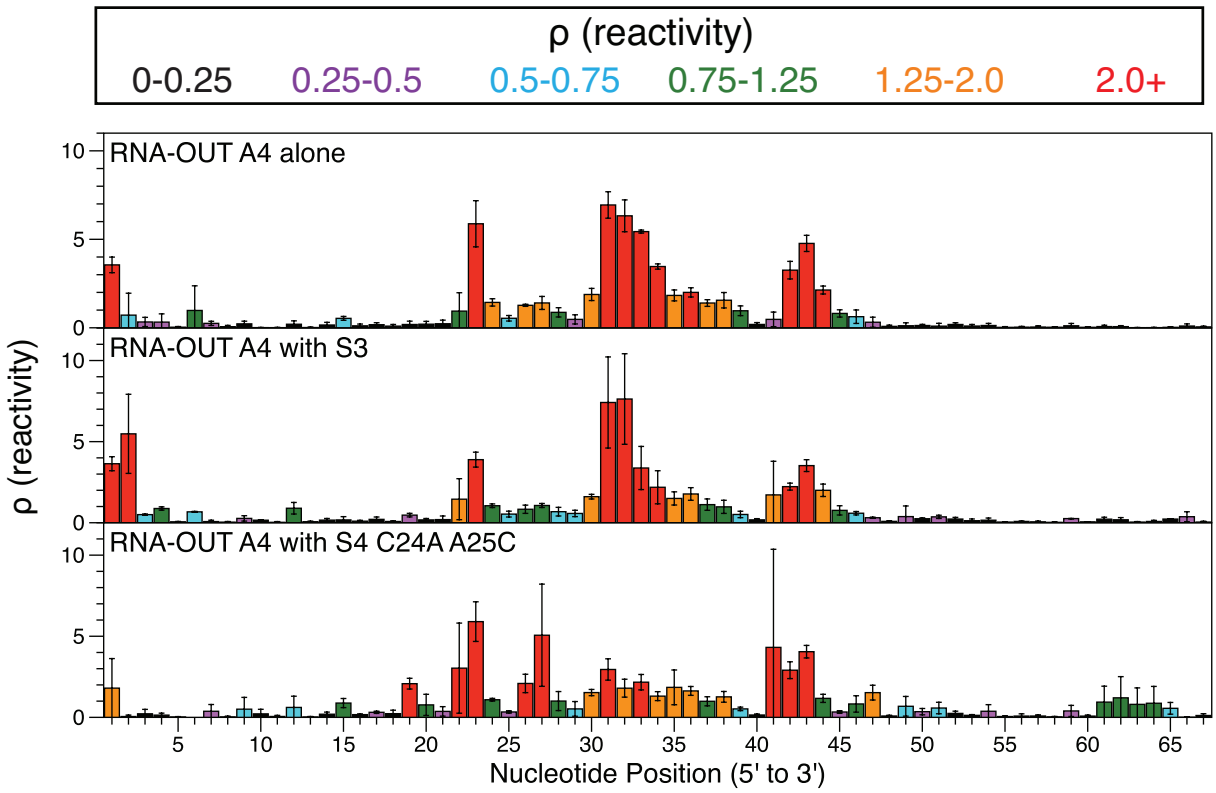


Figure B.19: In-cell SHAPE-Seq reactivities for the antisense RNA-OUT A4 with RNA-IN variants. Color-coded reactivity spectra represent averages over three independent in-cell SHAPE-Seq experiments, with error bars representing one standard deviation of the reactivities at each position. The SHAPE-Seq reactivities of RNA-OUT A4 are shown without RNA-IN, with RNA-IN S3 (remains ON), and with RNA-IN S4 mutant C24A A25C (turns OFF). There are no major differences between A4 alone or with the orthogonal S3. When S4 C24A A25C is present to interact with A4, there are decreases in the reactivities of nucleotides 1, 2, and 31-34 and increases in nucleotides 19, 26, 27, and 47. The nucleotides with decreasing reactivities correspond well to regions of A4 expected to interact with S4 C24A A25C (Figure B.16). However, the locations of increasing reactivities are generally unexpected except for nucleotide 19, which forms part of the double bulge originally designed to prevent RNase III cleavage [176]. It should be noted that the copy number difference between the RNA-OUT and the RNA-IN plasmids should cause RNA-OUT to be present in excess, such that the dominating structure is likely the non-interacting one, lessening the changes that are observable for RNA-OUT binding to RNA-IN.

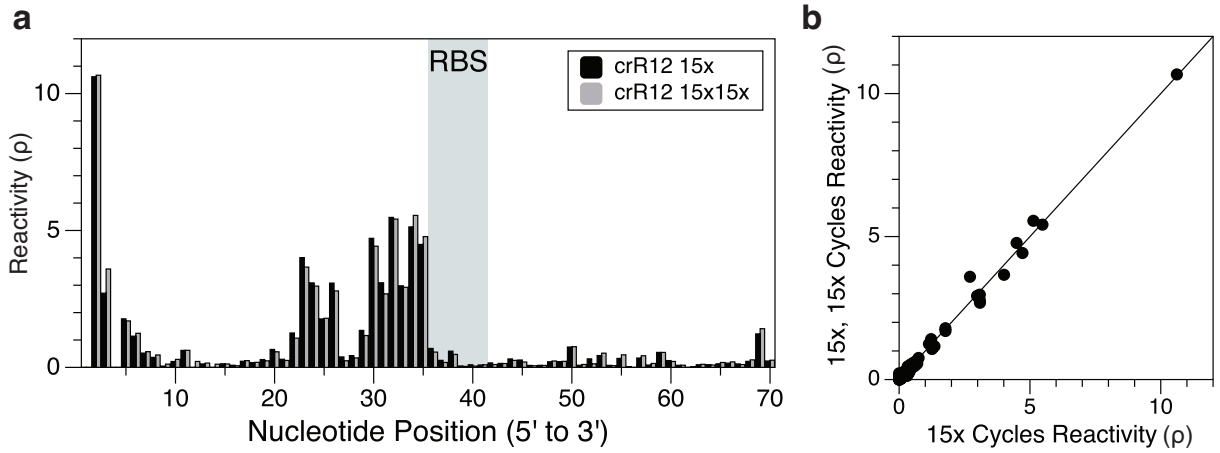


Figure B.20: Increasing PCR selection does not affect reactivity calculation. (a) Reactivity map comparing the same crR12 replicate (expressed alone) sequenced using 15 cycles of PCR with all oligonucleotides included (15x) or 15 cycles of PCR without oligonucleotide I (Table B.3) followed by another 15 cycles after adding oligonucleotide I. The reactivity maps show very close agreement as depicted in the correlation plot in (b), with a Pearson Correlation Coefficient of 0.996.

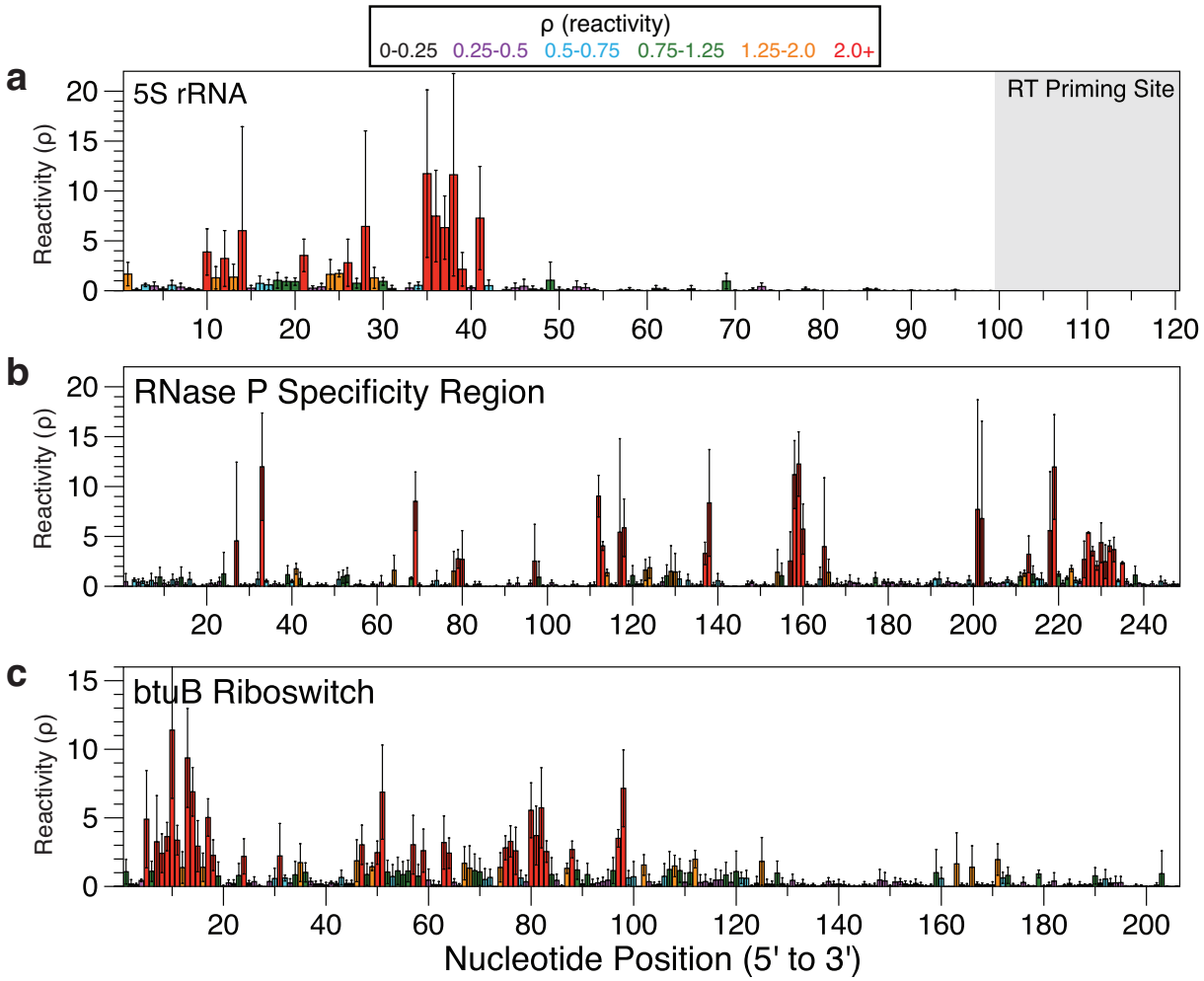


Figure B.21: Reactivity maps for endogenous RNA targets. Bar charts depicting the reactivity values for the 5S rRNA, RNase P specificity region, and btuB riboswitch domain used to color Figure 3.5. Color-coded reactivity spectra represent averages over three (five for btuB riboswitch) independent in-cell SHAPE-Seq experiments, with error bars representing one standard deviation of the reactivities at each position. (a) 5S rRNA shows clusters of reactivity in the 5′ half when probed within the cell in contrast to equilibrium folded RNA *in vitro* (Figure B.24). The RT priming site on the 3′ end, where no reactivity information was obtained, is marked. (b) The specificity region of RNase P shows well-distributed, highly reactive peak clusters. (c) Reactivity map for the btuB adenosylcobalamin riboswitch, which displays more unstructured nucleotides in the 5′ half.

Figure B.22: In-cell vs. equilibrium *in vitro* refolding reactivity maps for the riboregulators. Bar charts comparing reactivities measured for the riboregulator variant 12 RNAs in-cell (measured in triplicate) and an equilibrium refolding experiment with in vitro purified RNA. Error bars for the in-cell measurements represent one standard deviation. (a) Few differences are observed for the crR12 RNA between in-cell (with control antisense plasmid) and equilibrium measurements. The most notable difference is near the 5′ end where the intensities differ somewhat and the most 5′ nucleotide is reactive in vitro. (b) Like the crR12 alone case in (a), there are few observable differences in the crR12 with taR12 case. We observe that the average reactivity between nucleotides 24-40 is somewhat higher for in-cell reactivities in some positions. Also, there is a reactive spike at position 10 *in vitro*, but it is likely a spurious peak and not representative. (c) taR12 has a similar reactivity profile *in vitro* and in-cell. Nucleotides 4, 9, and 18-19 are somewhat higher in-cell while nucleotides 41, 44, 49, and some within 60-68 are higher *in vitro*.

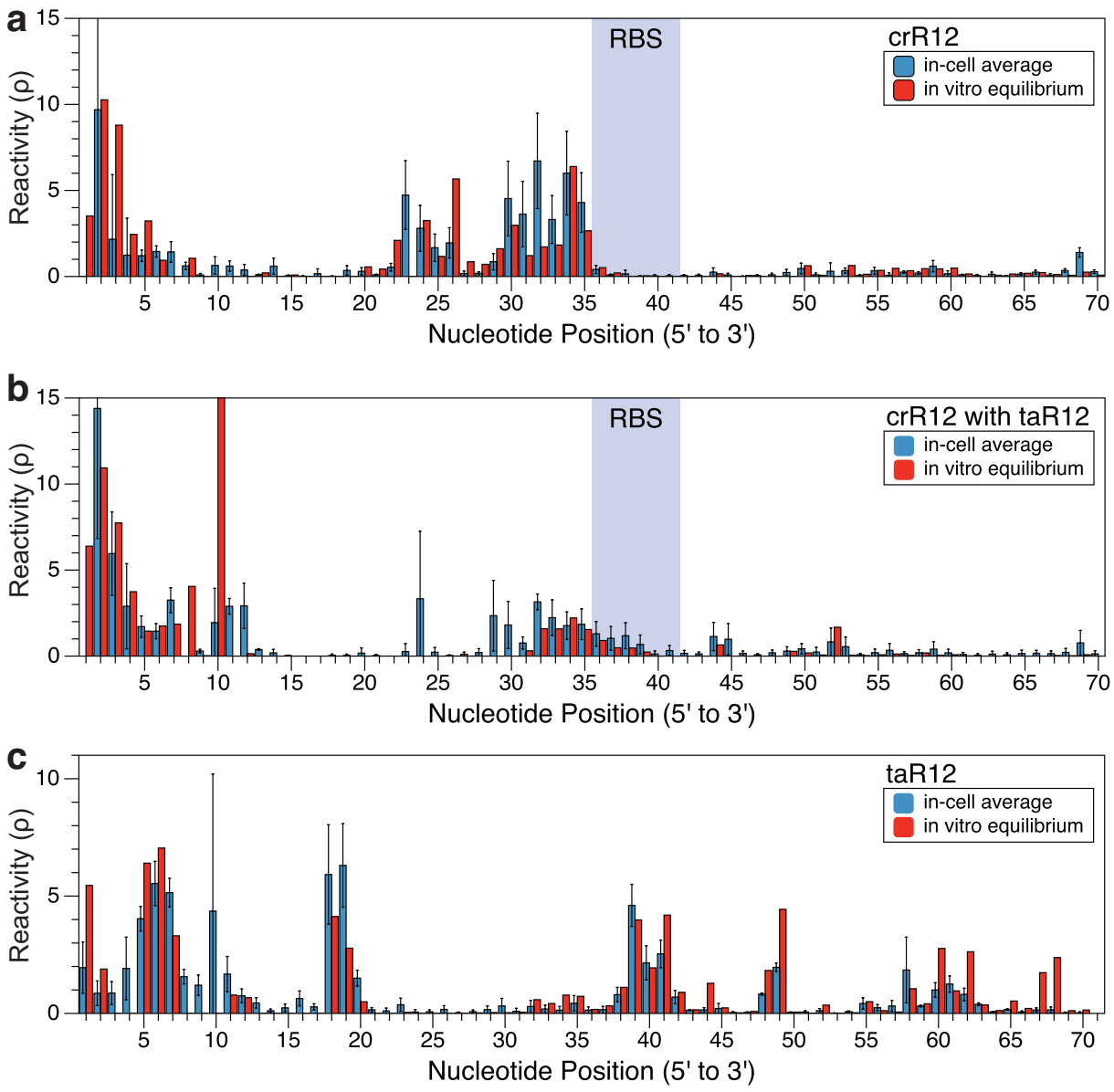
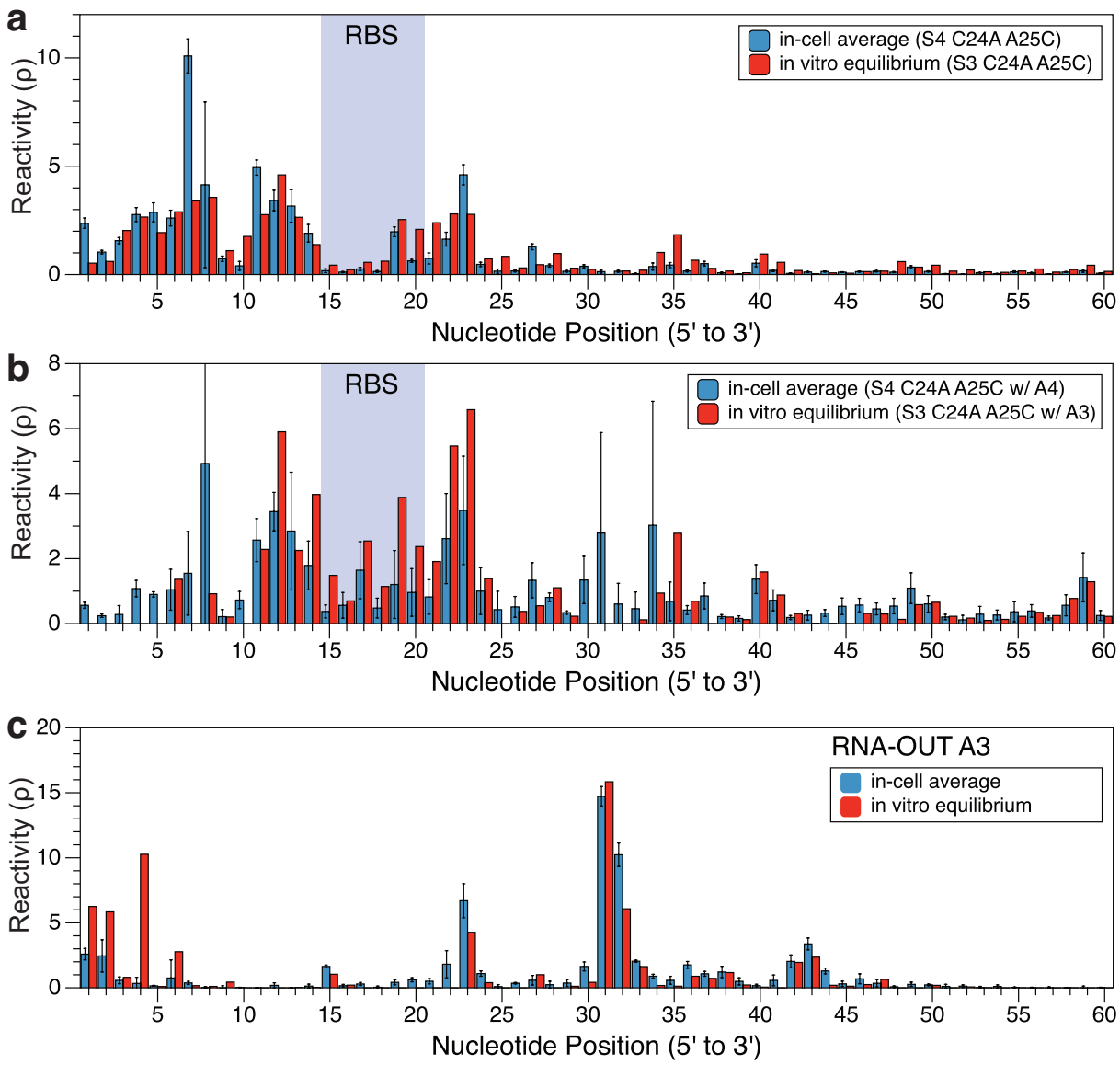

a
Reactivity (ρ)
RBS
crR12
in-cell average
in vitro equilibrium
Nucleotide Position (5' to 3')
b
Reactivity (ρ)
RBS
crR12 with taR12
in-cell average
in vitro equilibrium
Nucleotide Position (5' to 3')
c
Reactivity (ρ)
taR12
in-cell average
in vitro equilibrium
Nucleotide Position (5' to 3')

Figure B.23: In-cell vs. equilibrium *in vitro* refolding reactivity maps for RNA-IN/OUT. Bar charts comparing reactivities measured in-cell (RNA-IN S4 C24A A25C) and in vitro (RNA-IN S3 C24A A25C) for the RNA-IN double mutants and RNA-OUT A3. Error bars for the in-cell measurements represent one standard deviation. RNA-IN S3 C24A A25C was used in place of S4 C24A A25C in refolding experiments to ease in vitro transcription with T7 polymerase. (a) Few differences are observed for RNA-IN between in-cell (with control antisense plasmid) and *in vitro* equilibrium measurements. Nucleotides 6 and 7 are higher in-cell, while nucleotides around positions 20 and 25 are higher *in vitro*. (b) The differences between RNA-IN in-cell and *in vitro* are more pronounced when the cognate RNA-OUT is added. The reactivities around the RBS region are generally higher *in vitro*, while other nucleotides in RNA-IN are higher in-cell. (c) The reactivity pattern for RNA-OUT A3 matches well between in-cell and *in vitro* except for nucleotides in the 5′ end, which appear more reactive *in vitro*.


a
RBS
in-cell average (S4 C24A A25C)
in vitro equilibrium (S3 C24A A25C)
Reactivity (ρ)
Nucleotide Position (5' to 3')
b
RBS
in-cell average (S4 C24A A25C w/ A4)
in vitro equilibrium (S3 C24A A25C w/ A3)
Reactivity (ρ)
Nucleotide Position (5' to 3')
c
RNA-OUT A3
in-cell average
in vitro equilibrium
Reactivity (ρ)
Nucleotide Position (5' to 3')

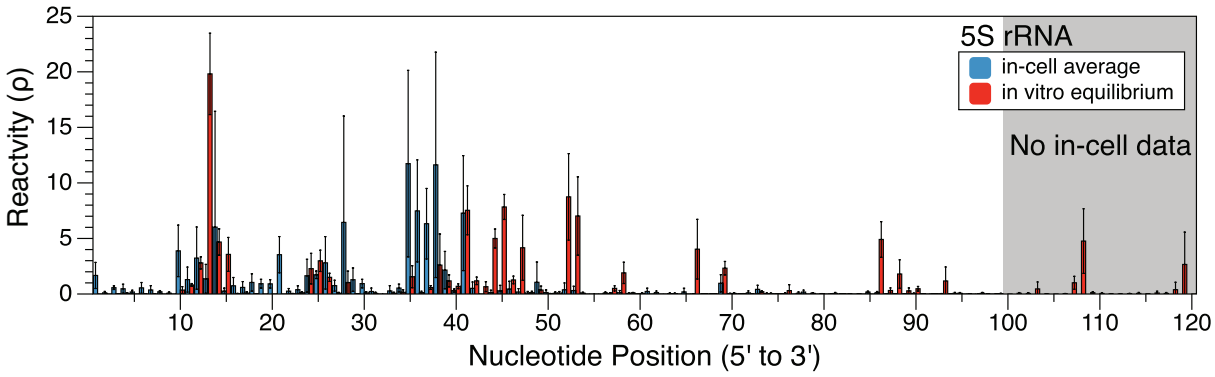


Figure B.24: In-cell vs. equilibrium *in vitro* refolding reactivity maps of 5S rRNA from *E. coli*. Reactivity maps comparing the 5S rRNA measured from in-cell and *in vitro* experiments. Error bars represent one standard deviation. *in vitro* data is taken from Loughrey *et al.* [122] and can be found in the RMDB [169] with IDs: 5SRRNA_1M7_0001, 5SRRNA_1M7_0002, and 5SRRNA_1M7_0003. There are many reactivity differences between the in-cell and *in vitro* conditions. Most notably, nucleotides 35-38 have greatly increased reactivity in the cell, while nucleotides from position 44 to the 3′ end all greatly decrease in the cell. In contrast, we observe that the reactivity clusters near positions 12 and 25 are similar between the *in vitro* and in-cell conditions. Few reactivity changes in those regions would be expected because those nucleotides are in loop regions of the 5S rRNA that face away from the ribosome into the solvent (Figure 3.5a). The region around nucleotides 35-55 fits in a groove near the L5 protein [193, 200] and is likely remodeled as it fits into the ribosome, which could cause the differences in the reactivities in that region of the reactivity map.

B.4 Supplementary Methods.

Detailed in-cell SHAPE-Seq protocol.

B.4.1 Materials

- *E. coli* strain NEBTurbo (cloning strain)
- *E. coli* strain TG1 (experimental strain)
- Sense and antisense platform vectors - See Figure B.1
- Oligonucleotides, see Table B.3
- Cloning Reagents
  - 1.33X Gibson Assembly master mix (See below) [181]
  - Phusion High-Fidelity PCR kit (NEB, cat. no. E0553S)
  - Restriction enzyme DpnI (NEB, cat. no. R0176S)
  - T4 DNA ligase (NEB, cat. no. M0202S)
  - QIAprep spin Miniprep kits (Qiagen, cat. no. 27104)
  - QIAquick PCR purification kit (Qiagen cat. no. 28104)
  - Qiagen plasmid mini kit (Qiagen cat. no. 12123)
  - LB medium (Life Technologies, cat. no. 12795-084)
  - LB agar (BD, cat. no. 214530)
  - 5X KCM solution (0.5 M KCl, 0.15 M $CaCl_2$, 0.25 $MgCl_2$)
  - 2-YT medium (Invitrogen, cat. no. 22712-020)
  - Agarose gel electrophoresis reagents
  - Carbenicillin (Sigma-Aldrich, cat. no. C3416)
  - Chloramphenicol (Sigma-Aldrich, cat. no. C1919)
- 1-methyl-7-nitroisatoic anhydride - see [226]
- Anhydrous dimethyl sulfoxide (Sigma-Aldrich, cat. no. D8418)
- PBS solution (pH 7.4)
- Kanamycin sulfate (Sigma-Aldrich, cat. no. K1377)
- TRIzol Max Bacterial Enchancement Reagent (Ambion, cat. no. 16096-040)
- TRIzol reagent (Ambion, cat. no. 15596-026)

- Chloroform (Sigma-Aldrich, cat. no. 372978)
- Isopropyl alcohol (Sigma-Aldrich, cat. no. I9516)
- Ethanol (Sigma-Aldrich, cat. no. E7023)
- 10 mM dNTPs (NEB, cat no. N0447S)
- 0.1 M dithiothreitol (DTT) (Sigma-Aldrich, cat. no. D0632)
- Super Script III (SSIII) reverse transcriptase (Invitrogen, cat. no. 18080093)
- 5X SSIII First Strand Buffer (supplied with SSIII)
- Sodium Hydroxide (Sigma-Aldrich, cat. no. S8045)
- Hydrochloric Acid (Sigma-Aldrich, cat. no. H1758)
- CircLigase ssDNA Ligase, 10x CircLigase reaction buffer, 1 mM ATP, and 50 mM $MnCl_2$ (Epicentre, cat. no. CL4111K)
- 20 mg/mL glycogen (Invitrogen, cat. no. 10814-010)
- 3.0 M NaOAc (pH 5.5)
- Agencourt AMPure XP beads (Beckman Coulter, cat. no. A63880)
- TE Buffer (10 mM Tris-HCl, 1 mM EDTA, pH 7.5)
- GeneScan 500 LIZ standard (Applied Biosystems, cat. no. 4322682)
- Deionized formamide (Ambion, cat. no. AM9342)
- Exonuclease I (ExoI) (NEB, cat. no. M0293S)
- Qubit dsDNA High Sensitivity Kit (Invitrogen, cat. no. Q32854)

B.4.2 Equipment

Exact items used in this study in parentheses.

- 100 mm x 15 mm Petri dishes
- 0.7 mL, 1.5 mL microcentrifuge tubes
- QIAquick purification columns (Qiagen) or similar silica purification columns
- Thin-walled PCR tubes
- Optically clear 96-well reaction plates (Applied Biosystems, cat. no. N8010560)
- Clear bottom, black-sided 96-well plates (Corning, cat. no. 3631)
- 2 mL 96-well culture block (Corning, cat. no. 29445-164)
- Breathable adhesive cover for 96-well culture blocks (Sigma-Aldrich, cat. no. Z763624)

- Refrigerated microcentrifuge (to 4 °C)
- 96-well capable fluorescent spectrophotometer
- Thermal cycler
- Agarose gel electrophoresis setup
- Dry heating block
- Magnetic 96-well plate stand
- Shaking Incubator (96 deep-well compatible)
- Qubit 2.0 Fluorometer (Invitrogen, cat. no. Q32866)
- Unix, Linux or Mac OS X system for data analysis

B.4.3 Reagent Setup

1X Kanamycin in PBS: Prepare a 1000X solution of 100 mg/mL (working solution 100 μg/mL) kanamycin by dissolving in water. Add 10 μL of 1000X kanamycin to 10 mL of PBS. Refrigerate dilution until ready to use.

Gibson assembly master mix: Mix 320 μL isothermal reaction buffer (25% PEG-8000, 500 mM Tris-HCl, 50 mM $MgCl_2$, 50 mM DTT, 1 mM dNTPs, 5 mM NAD, pH 7.5), 0.64 μL of 10 U/μL T5 exonuclease, 20 μL of 2 U/μL Phusion DNA polymerase, 160 μL of 40 U/μL Taq DNA Ligase, and enough water to bring the total volume to 1.2 mL. Aliquot 15 μL at a time into microcentrifuge tubes and store at -20 °C. Pre-made Gibson assembly master mix can be alternatively be purchased from New England Biolabs (cat. no. E2611S).

Super Script III master mix: Mix 4 volumes of 5X SSIII First Strand Buffer (supplied with SSIII), 1 volume of 0.1 M DTT, and 1 volume of 10 mM dNTPs. Store at -20 °C.

B.4.4 Procedure for in-cell SHAPE-Seq in *E. coli*

Cloning

(~1 week) This section should be skipped unless adding a new sense/antisense RNA to platform for expression in *E. coli*.

1. Design primers suitable for Gibson Assembly of the sense/antisense insert and platform vectors.

   (a) To insert an RNA sequence in the antisense platform, amplify the antisense vector with 5′-GCCTCTACCTGCTTCGGCCG-3′ and 5′-ACTAGTATTATA CCTAGGACTGAGCTAGC-3′. Design primers for the insert using 5′-GC TCAGTCCTAGGTATAATACTAGTxxxxx-3′ for the forward primer and 5 ′-GCCGAAGCAGGTAGAGGCzzzzzz-3′ for the reverse (replace x's with primer sequence specific to 5′ end of antisense RNA and z's with reverse complement of 3′ end, both with a T_m of about 58 °C).

   (b) To insert an RNA sequence in the sense platform, amplify the sense vector with 5′- ATGCTAGCAAAGGAGAAGAACTTTTC-3′ and 5′-ACTAGTATTATACCTAGGACTGAGCTAGC-3′. Design primers for the insert using 5′-GCTCAGTCCTAGGTATAATACTAGTxxxxx-3′ for the forward primer and 5′-GAAAAGTTCTTCTCCTTTGCTAGCATzzzzzz-3′ for the reverse (replace x's with primer sequence specific to 5′ end of sense RNA and z's with reverse complement of 3′ end, both with a T_m of about 58 °C). Note that the overlap for SFGFP contains the start codon - modify these overlaps if other sequences are required near the start codon.

2. Perform PCR using the Phusion PCR kit (NEB) according to the manufacturer's

instructions, using the following thermal cycling conditions:

Cycle Number	Denature	Anneal	Extend	Hold
1	98 °C, 30 s			
2-25	98 °C, 15 s	3 °C above primer's T_m	72 °C, 15 s/kb	
26			72 °C, 5 min	
				4 °C

3. Digest PCR products with 1 μL DpnI at 37 °C for 30 min.
4. Check on agarose gel to verify the size of the PCR products.
5. Purify the PCR products using the QIAquick PCR purification kit.
6. Measure the DNA concentration of the DNA fragments to be assembled with the Qubit or Nanodrop.
7. Thaw a tube of 1.33X Gibson assembly master mix on ice for each construct to assemble. Keep on ice until use.
8. Add approximately 3:1 (molar ratio) insert:platform vector to be assembled to a total volume of 5 μL, and a to 15 uL of 1.33X Gibson assembly master mix.
9. Incubate at 50 °C for 1 hr.
10. Place assembly products on ice to cool to at least room temperature.
11. Transform assembled DNA into chemically competent *E. coli* cells (we use KCM competent NEBTurbo).
12. Incubate on ice for 20 min. Heat shock at 42 °C for 1.5 min and then place back on ice for 1 min.

13. Add 50 μL 2YT medium (or other rich media such as SOB or SOC) and incubate the cells with vigorous shaking at 37°C for 1 hr. While shaking, warm agar plates with appropriate antibiotic at 37 °C.
14. Plate the cells and incubate overnight at 37 °C.
15. Grow and extract plasmids for sequencing, using the QIAQuick plasmid purification kit.
16. Store sequence confirmed plasmids at -20 °C until ready to use.

Preparation for in-cell probing

(3 days, ~3 hours of active effort) - Skip to Step 19 if targeting endogenously expressed RNAs.

17. Determine which combinations of sense and antisense to test.
    (a) Due to how intensive future steps are, it is advised to not do more than 10-12 total in-cell SHAPE-Seq experiments at a time. We recommend 1-4 for those new to the protocol.
        i. If characterizing more than 6 samples, it may be helpful to get a second person to help during the modification and extraction steps.
18. (Co-)Transform the sense/antisense plasmid(s) into the strain being used for testing functionality.
    (a) Include empty plasmid controls to correct for autofluorescence later.
    (b) We recommend co-transforming sense plasmids with a control antisense plasmid when only the sense RNA is being studied to maintain consistency between fluorescence measurements.

19. Plate transformed cells on an LB+Agar plate with the appropriate antibiotics and grow overnight at 37 °C.

    (a) If targeting endogenously expressed RNAs, plate cells directly with no antibiotics. Alternatively, this step could be skipped and cells could be grown directly from glycerol stocks, etc.

20. To an autoclaved 96-well culture block, add 1 mL of LB with antibiotic (if required). Pick individual colonies from the plates and inoculate the media with the corresponding antibiotic resistance(s). Grow the cells overnight on an incubator shaker at 37 °C and 1,000 rpm.

    (a) It is recommended to start this growth late in the day.

    (b) Replace LB with other medium if desired.

21. The next day, add 24 μL of the primary inoculum to 1.2 mL fresh, pre-warmed LB with antibiotic(s) (if required). Allow to grow for at least 3 hours at 37 °C and 1,000 rpm on the shaker.

    (a) The time for growth can be adjusted based on the desired functional testing assay and speed of growth of the cells.

    (b) An OD_{600} of 0.3 or higher is recommended for probing, but not necessarily required.

    (c) Replace LB with other medium if desired.

    (d) Alternative dilution factors and growing times can be used instead so long as a suitable OD is reached.

RNA Modification & Fluorescence Assay

(10-25 min)

22. During the 3 hour (or greater) subculture, prepare 15 μL of 250 mM 1M7 in neat DMSO for each in-cell SHAPE-Seq experiment being performed. Also preheat 200 μL of TRIzol Max Bacterial Enhancement reagent at 95 °C for each in-cell SHAPE-Seq experiment.
23. After the 3 hour (or greater) incubation is complete, take 150 μL of each culture and put into labeled 1.5 mL microcentrifuge tubes. Spin these tubes at ≥ 15,000 x g for 1 min and discard the used media. Set aside until Step 26.
24. Add 13.3 μL of 250 mM 1M7 to clean, empty adjacent wells for each culture. Also add 13.3 uL of anhydrous DMSO to a second empty well for each culture. The modification experiment will be performed in these wells. Using a multichannel pipet, quickly add 500 μL of culture to the wells containing either 1M7 (positive (+) channel) or DMSO (negative (-) channel).
25. Return the 96-well culture block to the shaker for 3 min.
26. During the 3 min incubation, resuspend the cultures from Step 23 in 200 μL cold PBS with 1X kanamycin. Store these at 4 °C until Step 31.
27. Remove the block from the shaker and pipet the (+) and (-) samples into individual, labeled 1.5 mL microcentrifuge tubes.
28. Spin the tubes at 15,000 x g for 1 min and aspirate the used media. Resuspend each pellet in 100 μL of the preheated TRIzol Max Bacterial Enhancement reagent.
29. Heat the samples in the TRIzol Max Bacterial Enhancement reagent at 95 °C for 4 min.

30. Remove from heat and add 500 μL TRIzol reagent to each. Shake vigorously for $\geq$ 5 s to mix.

    (a) Be careful that tubes are well sealed when shaking.

    (b) CAUTION TRIzol contains phenol, which causes burns, lung edema, and kidney damage.

    (c) CAUTION TRIzol contains chloroform, a known carcinogen, and must be handled with care.

    (d) **(Pause Point)**

31. At this time, pipet the cells in PBS (from Step 26) into a 96-well plate with a clear bottom. Include a well containing 200 μL of PBS with 1X kanamycin as a blank.

    (a) Skip Steps 31-32 if not measuring a fluorescent output.

32. Measure OD_{600} and fluorescence (485 nm excitation, 520 nm emission - this may be adjusted if you are using a different reporter system). Calculate the fold activation/repression by subtracting the PBS blank from the OD and fluorescence measurements for all cells. Normalize all of the wells for growth rate by dividing fluorescence intensity (FL) by OD. Then subtract cell autofluorescence from all wells using a non-fluorescent cell control wells FL/OD. The resulting FL/OD values can be used to compare ON and OFF levels between different conditions.

RNA Extraction

(~1.5 hours)

33. Incubate the TRIzol-containing tubes for $\geq$ 5 min at room temperature.

34. Add 100 μL chloroform to each tube, shake vigorously for 10 s, and incubate at room temperature for another 2 min.

    (a) Be careful that tubes are well sealed when shaking.

35. Centrifuge the TRIzol-containing tubes at 12,000 x g and 4 °C for 15 min.

36. Carefully transfer the clear aqueous phases (up to 350 μL) to new 0.7 mL micro-centrifuge tubes.

37. Add 250 μL cold 100% isopropyl alcohol and mix by inverting the tubes 6-8 times. Incubate at room temperature for 10 min.

    (a) Add 1 μL of 20 mg/mL glycogen if RNA yield is expected to be low.

38. Spin the tubes at ≥ 15,000 x g and 4 °C for 10 min. Aspirate the alcohol and wash with 500 μL 70% EtOH. Respin the tubes for 2 min and aspirate the EtOH. Respin again for 2 min and aspirate the remaining ethanol, allowing traces of remaining EtOH to air dry.

39. Resuspend each resulting RNA sample (will be 2 for each in-cell SHAPE-Seq experiment) in 10 μL RNase-free water.

    (a) **(Pause Point)** The RNAs can be stored at -20 °C overnight.

Reverse Transcription

(1-1.5 hours)

40. Prepare a thermal cycler and start the following protocol (RT) to preheat the block:

Step Number	Step Name	Temperature	Time
1	Hot start	95 °C	∞
2	Denature	95 °C	2 min
3	Denature	65 °C	5 min
4	Cooling	45 °C	∞
5	Pre-heat	45 °C	1 min
6	Extend	52 °C	25 min
7	Inactivate	65 °C	5 min
8	Infinite hold	4 °C	∞

41. Add 3 μL of 0.5 μM reverse transcription primer. If using the two-plasmid system, use primer A for antisense RNA only, primer B for sense RNA only, or a mixture of both primers A and B for sense and antisense mixtures. If targeting endogenously expressed RNAs, use primers designed specifically for those RNAs, such as primers C-E.

    (a) For rare or weakly expressed RNAs, reduce the primer concentration to 50 nM.

    (b) The primers in Table B.3 are for the RNAs described in this work, other sequences can also be used.

    (c) Be mindful of plasmid copy numbers when mixing RT primers, in the work presented we used a 50-50 mixture. However, balancing the relative amount of RT primer for different target abundances is recommended for future work.

42. Start the RT protocol in Step 40.

43. When the thermal cycler block is preheated to 95 °C, place each tube on the thermal cycler, and advance the steps.

44. During the first 7 min of incubation (2x Denature), mix 6 μL of SSIII Master Mix and 1 μL of 0.5X SSIII reverse transcriptase (diluted in its storage buffer) for each RNA sample in a microcentrifuge tube.

    (a) We recommend mixing some extra SSIII Master Mix and enzyme to avoid running short during Step 45.

45. At the end of the first 65 °C step (2nd Denature), move all the tubes to ice for 30 s. Then add 7 μL of the SSIII Master Mix/reverse transcriptase mix (from Step 44) to each tube, mix well, and return to ice. Once all of the RNAs have received the 7 μL, return all the tubes to the thermal cycler and continue the thermal cycler procedure.

    (a) IMPORTANT - When adding the Master Mix/reverse transcriptase mix, make sure to knock down the condensed water from the sides of the tubes to maintain the reaction volume.

46. Upon completion of the RT protocol, remove all the tubes to ice and add 1 μL of 10 M NaOH.

47. Incubate the tubes at 95 °C for 5 min to hydrolyze the RNA.

48. Partially neutralize the solutions by adding 5 μL of 1 M HCl.

49. Ethanol precipitate by adding 78 μL of ice-cold 100% ethanol to each tube, inverting 6-8 times, and incubating at -80 °C for 15 min.

    (a) **(Pause Point)** The cDNA can be stored at -80 °C.

50. Centrifuge the tubes at ≥ 15,000 x g and 4 °C for 15 min. Aspirate the ethanol and wash with 500 μL of 70% ethanol. Respin for 2 min, aspirate, respin for 2 min again, and aspirate the remaining ethanol. Allow to air dry.

51. Resuspend each cDNA in 22.5 μL of RNase-free water.

(a) **(Pause Point)** The cDNA can be stored at -20 °C.

A_adapter_b Ligation & Purification

(4 hours)

52. Add 3 μL of 10x CircLigase reaction buffer, 1.5 μL 50 mM $MnCl_2$, 1.5 μL 1 mM ATP, 1 μL CircLigase, and 0.5 μL of 100 μM oligonucleotide F to each tube with cDNA.

    (a) For samples with reduced RT primer concentrations (from Step 41), reduce oligonucleotide F concentration to 10 μM.

53. Incubate at 60 °C for 2 hrs, followed by a 10 min incubation at 80 °C to inactivate CircLigase I.

    (a) Meanwhile, thaw a tube of 20 mg/mL glycogen.

54. After the ligation is complete, add 70 μL RNase-free water, 10 μL 3.0 M sodium acetate pH 5.5, 1 μL 20 mg/mL glycogen (for visualization), and 300 μL of ice-cold 100% ethanol. Mix.

    (a) **(Pause Point)** The cDNA can be stored at -80 °C.

55. Incubate at -80 °C for 30 min then spin at ≥ 15,000 x g at 4 °C for 30 min. Aspirate the ethanol, re-spin for 2 min, and aspirate any remaining ethanol. Resuspend each cDNA in 20 μL water.

56. Resuspend a bottle of Agencout AMPure XP beads. Add 36 μL of bead slurry to each tube containing cDNA and mix well by pipetting up and down.

57. Continue purification of ssDNA libraries according to manufacturer's protocol, eluting in 20 μL TE buffer.

    (a) We recommend using the 96-well plate method for a large number of samples.

    (b) **(Pause Point)** The eluted cDNA can be stored at -20 °C.

Quality Control (QC)

(1 hour + CE time). An alternative method is mentioned in Step 76

58. Mix the following PCR or each sample: 1.5 μL ssDNA library, 1 μL of 1 μM primer G (for (+) samples) or primer H (for (-) samples), 1 μL of 0.1 μM selection primer (ex.: primer J (antisense samples), L (sense samples), or N, P, or R (endogenous)], 1.5 μL of 1 μM primer I, 5 μL of 5X Phusion reaction buffer (NEB), 0.5 μL of 10 mM dNTPs, 0.25 μL (0.5 U) Phusion Polymerase, and H_2O to 25 μL.

    (a) The selection primer chosen should match the RT primer sequence and contain a further 3' extension.

    (b) For rare transcripts, exclude primer I in the initial reaction mix (see Step 59b).

    (c) If using priming sites other than those designed in Table B.3, you will need to design an alternative set of mismatch primers instead of primers J-S. To do this, add the sequence CTTTCCCTACACGACGCTCTTCCGATCTYYYR to the 5' end of the RT primer for (-) samples, and CTTTCCCTACACGACGCTCTTCCGATCTRRRY for (+) samples. Then, extend the 3' end of primer a few nucleotides into the cDNA sequence such that it mismatches

the 5′ end of oligonucleotide F. Protect this mismatch from 3′→5′ exonuclease activity with phosphorothioate modifications.

(d) Either positive or negative primer can be used as they are the same length and will appear the same on the electropherogram output.

(e) For cDNAs containing more than 1 RT primer, mix a separate QC library for each RT primer. For example, if using both sense and antisense, mix 2 QC reactions, one with primer J and one with primer L. If using more than 3-4 RT primers, you may want to combine some of the selection primers together when doing QC and look for multiple full length peaks instead.

(f) When analyzing low abundance transcripts (not those used in the platforms from Figure B.1), the number of cycles can be increased. However, non-specific amplification will start to occur at significant levels after 15 cycles of amplification when primer I is included in libraries derived from low abundance transcripts.

(g) For rare transcripts, perform 15 cycles of amplification without primer I (more cycles can be done if necessary). Then, add the 1.5 μL of 1 μM primer I directly to the reaction and perform another 15 cycles of amplification to finish building the libraries.

59. Run the LIB2CE15 protocol on a thermal cycler:

Cycle Number	Denature	Anneal	Extend	Hold
1	98 °C, 30 s			
2-16	98 °C, 15 s	63 °C, 30 s	72 °C, 30 s	
17			72 °C, 5 min	
18				4 °C

60. Add 50 μL H_2O to the (+) reaction, then mix the (+) and (-) reaction products together.

61. Ethanol precipitate each (+/-) reaction mixture by adding 10 μL 3.0 M NaOAc pH 5.5 and 300 μL of ice-cold EtOH to each combined reaction pair.

62. Incubate the mixtures at -80 °C for 15 min.

63. Centrifuge each at ≥ 15,000 x g and 4 °C for 15 min.

64. Aspirate the EtOH, re-spin for 2 min, and aspirate the remaining EtOH. Air dry the pellet.

65. Dissolve each pellet in 10 μL of deionized formamide and add 0.2-0.3 μL of GeneScan 500 LIZ standard.

    (a) We recommend heating at 85-95 °C for 5-10 min to aid dissolving.

66. Run each sample on a capillary electrophoresis (CE) machine.

67. Use the LIZ standard to identify peak lengths (We recommend SHAPEFinder [105] for easy viewing). There should be a full-length peak clearly visible at the length of the cDNA expected + RT primer length + 96 bp for the PCR overhangs for quality analysis. Peaks at RT primer length + 96 bp are indicative of RT primer-A_adapter_b dimers. A good library trace should show considerable peak height and a good full length:RT-A_adapter_b dimer ratio, with minor peaks in between for RT stops.

    (a) The antisense libraries have a dimer (no cDNA) length of 117 bp, while the sense libraries have a dimer length of 122 bp. The fully extended peak (to the 5′ end of the RNA) should be expected at 96 nt + length of the RNA extended, including the RT primer length.

(b) Shown in Figure B.25 is a good QC trace for the antisense taR12 from the riboregulator system (96 nt long). Note that the dimer side product peak is low, but the full-length cDNA peak is high. A large concentration of unextended primer is normal, as the PCR described above does not consume all the fluorescent primer.

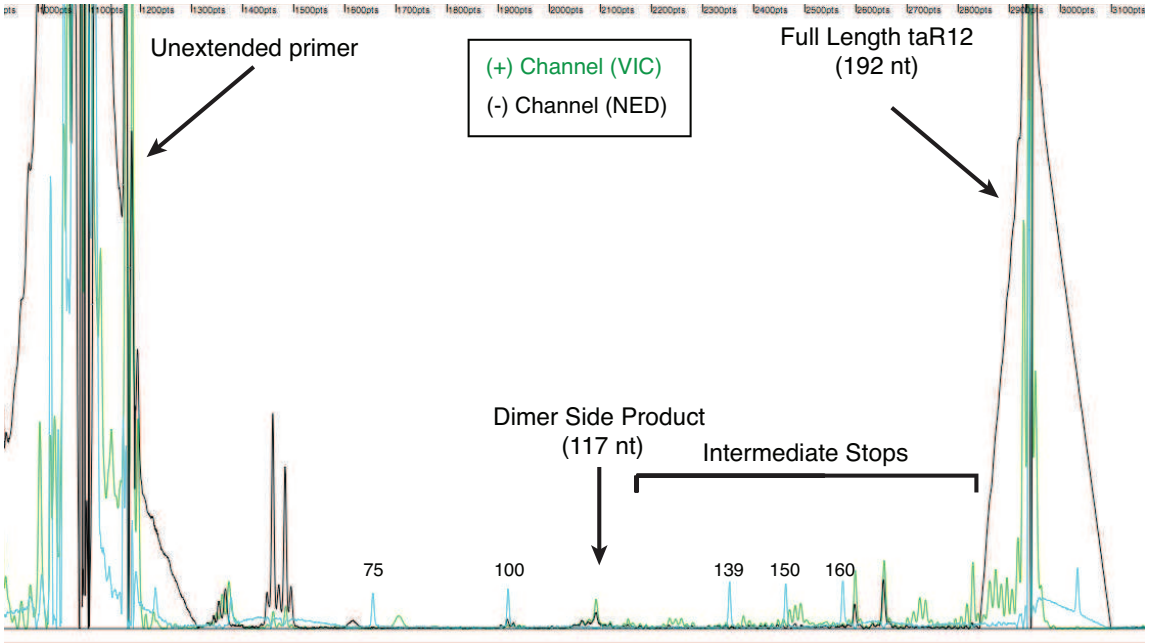


Figure B.25: CE quality control example. Positive (+; green) and negative (-; black) channels quality analysis of a run using taR12.

dsDNA Library Construction

(1-2 hours)

68. Mix the following PCR for each sample: 3 μL ssDNA library, 2 μL of 0.1 μM primer J (+) or K (-) (antisense samples) or primer L (+) or M (-) (sense samples), 0.25 μL of 100 μM primer I, 0.25 μL of 100 μM Illumina indexing primer T-AQ (choose which one depending on indexes being used), 10 μL of 5X Phusion reaction buffer (NEB), 0.5 μL of 10 mM dNTPs, 0.5 μL (1 U) Phusion Polymerase, and H_2O to 50 μL.

    (a) As discussed in Step 58b, use the appropriate mismatch selection primers if not using RT primers A or B.

    (b) The relative amounts of primers or ssDNA library can be increased to increase yields if desired or initial yields are low.

    (c) If characterizing low abundance or rare transcripts and the split PCR method from Steps 58-59 was used, similarly exclude primer I from the first round of amplification, then add for the second as done in those steps (at the concentration listed in this step).

69. Run the SEQPHU15 protocol on a thermal cycler:

    (a) If a different number of cycles, other than 15, were used in Step 59, use that number of cycles here as well. Also see Note 68c.

Cycle Number	Denature	Anneal	Extend	Hold
1	98 °C, 30 s			
2-16	98 °C, 15 s	65 °C, 30 s	72 °C, 30 s	
17			72 °C, 5 min	
18				4 °C

70. Chill the reactions at 4 °C for 5 min.

71. Add 0.25 μL (5 U) of ExoI to each library reaction.

72. Incubate at 37 °C for 30 min.

73. Resuspend a bottle of Agencout AMPure XP beads. Add 90 μL of bead slurry to each tube containing cDNA and mix well by pipetting up and down.

74. Continue purification of dsDNA libraries according to manufacturers protocol, eluting in 20 μL TE buffer. We recommend using a 96-well plate method for a large number of samples.

75. Measure the concentration of each dsDNA library using Qubit or another preferred method for quantifying concentration of DNA.

    (a) **(Pause Point)** The eluted libraries can be stored at -20 °C.

76. *(Optional) Alternative Quality Control Method*: Run 1 μL of the dsDNA libraries on a BioAnalyzer high sensitivity dsDNA chip. Look for the same features described in Step 67, except that the cDNA full length should appear at 125 bp + RT primer length + the cDNA length (instead of cDNA expected + RT primer length + 96 bp).

    (a) The dimer (no cDNA) length of dsDNA for the antisense is 146 bp, while the sense is 151 bp for the platform used in this work.

Next-Generation Sequencing

(1 day - 2 weeks, depending on resources available).

77. Use the concentration measurements from each library and the quality control traces to estimate the molarity of each library.
78. Mix a few microliters of each library together to pool all of the SHAPE-Seq libraries such that they are all balanced by mole.
79. Run on the Illumina platform, using 2x35 bp paired end reads. We recommend the MiSeq v3 kit for less than 30-40 libraries that are properly balanced between all indexes and RT primers.
    (a) Longer read lengths can be used, but are generally unnecessary. The read length only needs to be long enough to uniquely align the 3′ ends of all RNAs analyzed within a TruSeq index.
    (b) Do not process the data using the native Illumina adapter trimming; it will be removed by downstream in the data processing steps.

Data Processing

(~0.5-3 hours, if software installed) - Requires Unix, Linux, or Mac OS X

80. If not previously installed, download and install Spats and its associated software from `spats.sourceforge.net`, using the instructions provided there.
    (a) Other software required for running Spats includes: bowtie (`http://bowtie-bio.sourceforge.net`), the fastx toolkit and libgtextutils

(http://hannonlab.cshl.edu/fastx_toolkit/download.html), boost (www.boost.org), and Python.

(b) Note: Future releases of Spats can be found at: http://spats.sourceforge.net/

81. Obtain the sequencing data from a local store or BaseSpace and unzip the files to get the .fastq files.

    (a) Future releases of Spats may contain library analysis tools.

82. Create a fasta (.fa) style formatted targets file as indicated in the Spats documentation. For each .fastq pair (Read 1 and Read 2), there should only be one .fa file, that contains all of the target RNAs.

    (a) For each RNA, create a new line beginning with a caret '>' followed by the RNA name. The line beneath should then contain your sequence of interest from the 5' end to the 3' end of the cDNA product.

83. Run adapter_trimmer.py. By default, the script will determine a number of the parameters for Spats automatically by analyzing the targets input. However, some values must be set if differing from the default. adapter_trimmer.py will find and remove sequences containing Illumina sequencing adapters.

    (a) If using something other than 2x35 bp sequencing, be sure to include the flag '–read-len XX' where XX is the read length from the 2xXX bp paired end sequencing. adapter_trimmer.py assumes that both read lengths are equal.

84. Run Spats with the .fastq output from adapter_trimmer.py, where <rna.fasta> is the targets file, RRRY and YYYR are the treated and untreated handles (respectively). If adapter_trimmer.py was run, the adapter trimming capabilities in Spats

itself are not necessary to use. The output directory will contain a text file named reactivities.out, which are the results.

(a) We recommend always using adapter_trimmer.py, not the trimming algorithm in Spats itself.

85. Normalize the output θ_i values to ρ_i values according to Appendix B.1. These ρ_i values can be plotted to obtain reactivity maps or used as secondary structure prediction constraints.

RNA Structure Prediction

(20 minutes)

86. Download and install RNAstructure (`http://rna.urmc.rochester.edu/RNAstructure.html`) or use the webserver (`http://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Predict1/Predict1.html`)

87. Create a .shape text file containing the ρ_i values for each position. To do this, create a tab separated text file where the first column is 1→L, where L is the length of the RNA, and the second column is the ρ_i value for that position (5'→3'). For nucleotides to be analyzed that do not have ρ_i values, enter '-999' instead.

(a) Note: The length of the RNA analyzed will need to match the number of positions there is reactivity information for in the .shape file.

88. If using the webserver, copy and paste the RNA sequence to be analyzed in all caps into the 'Sequence' box, adjust any parameters as desired, then choose the .shape file to upload under 'Select SHAPE Constraints File:' Adjust the SHAPE

Intercept (b) to be -0.3 and the SHAPE slope (m) to be 1.1 and submit the query. The minimum free energy structure appears after calculations complete.

(a) Make sure the sequence is in all caps, lower case is forced to be single-stranded.

89. If using the GUI version, create a new .seq file using the RNA sequence of interest in all caps. Then choose "RNA..Fold RNA Single Strand" and select the sequence file for the RNA of interest. Then select "Force..Read SHAPE Reactivity Pseudo-Energy Constraints" and choose the .shape file containing the appropriate ρ_i values. Adjust the SHAPE Intercept to be -0.3 and the SHAPE slope to be 1.1 and hit 'OK'. Run the calculations by hitting 'Start'. The minimum free energy structure will appear after calculations are complete.

(a) Make sure the sequence is in all caps, lower case is forced to be single-stranded.

(b) The command line version also accepts SHAPE reactivity constraints. See `http://rna.urmc.rochester.edu/Text/Fold.html` for instructions.

APPENDIX C

SUPPLEMENTARY INFORMATION FOR COTRANSCRIPTIONAL FOLDING OF A FLUORIDE RIBOSWITCH AT NUCLEOTIDE RESOLUTION

C.1 Materials and Methods

C.1.1 Plasmids

Plasmids used for DNA template synthesis contained a chloramphenicol resistance gene, the p15A origin of replication, and a consensus *E. coli* σ^{70} promoter followed by a sequence encoding the RNA under study. The *E. coli* SRP RNA sequence was cloned upstream of the antigenomic hepatitis δ ribozyme. The *B. cereus* crcB fluoride riboswitch (Genbank AE017194.1, bases 4763724 to 4763805) was cloned upstream of a consensus ribosome binding site and the superfolder green fluorescent protein (SFGFP) sequence. These non-native downstream sequences were used to allow transcription to proceed far enough such that the full length RNA of interest would emerge from RNA polymerase (RNAP). These sequences do not influence cotranscriptional SHAPE-Seq interpretations, as lengths where non-native RNA had emerged from RNAP were not used to draw conclusions. The fluoride riboswitch mutants were derived from the plasmid described above.

C.1.2 Proteins

EcoRI E111Q (Gln111) was a generous gift from Jeffrey Roberts and Joshua Filter (Cornell University, Ithaca NY).

C.1.3 Template preparation

DNA template libraries for cotranscriptional SHAPE-Seq were prepared by combining individual PCR amplifications of each RNA template length. Each 25 μL PCR reaction included 20.4 μL H_2O, 2.5 μL 10X Standard Taq Reaction Buffer (New England Biolabs), 0.5 μL 10 mM dNTPs, 0.25 μL 100 μM oligo J (forward primer, Table C.1), 0.15 μL plasmid DNA template, 0.25 μl *Taq* DNA polymerase (New England Biolabs), and 1 μL 25 μM reverse primer. The reverse primer incorporated an EcoRI site. Reaction mixes were run using a standard thermal cycle program consisting of 30 cycles of amplification using an annealing temperature of 55 °C. After thermal cycling, PCR reactions were pooled, mixed, and split into 500 μL aliquots before addition of 50 μL 3 M NaOAc pH 5.5 and 1 mL of 100% EtOH for EtOH precipitation. Precipitated pellets were dried using a SpeedVac and pooled by dissolving all pellets in 30 μL H_2O. The template was then run on a 1% agarose gel and extracted using the QIAquick Gel Extraction Kit (Qiagen). The concentration of the purified template was measured using the Qubit Fluorometer (Life Technologies) and the molarity of the template was calculated using the median template length.

Single-length DNA templates were prepared by performing five 100 μL PCR reactions including 82.25 μL H_2O, 10 μL 10X Standard Taq Reaction Buffer (New England Biolabs), 1.25 μL 10 mM dNTPs, 2.5 μL 10 μM oligo J (forward primer, Table C.2), 2.5 ul 10 μM oligo K (Table C.2), 0.5 μL plasmid DNA template, and 0.5 μL Taq DNA polymerase (New England Biolabs). Reactions were run with the thermal cycling program described above. After thermal cycling, reactions were pooled before addition of 50 μL 3M NaOAc pH 5.5 and 1 mL of 100% EtOH for EtOH precipitation. The precipitated pellet was dried using a SpeedVac and dissolved in 30 μL H_2O. The template was then run on a 1% agarose gel and extracted using the QIAquick Gel Extraction Kit

(Qiagen). The concentration of the purified template was measured using the Qubit 2.0 Fluorometer (Life Technologies).

C.1.4 *in vitro* transcription (single length, radiolabeled)

25 μL reaction mixtures containing 5 nM linear DNA template (see above) and 0.5 U of *E. coli* RNAP holoenzyme (New England Biolabs) were incubated in transcription buffer (20 mM Tris-HCl pH 8.0, 0.1 mM EDTA, 1 mM DTT and 50 mM KCl), 0.1 mg/mL bovine serum albumin, 200 μM ATP, GTP, CTP and 50 μM UTP containing 0.5 μCi/μL [α-^{32}P]-UTP for 10 min at 37 °C to form open complexes. When present, NaF was included to a final concentration of 1 μM, 10 μM, 100 μM, 1 mM or 10 mM as indicated in **??**. Single-round transcription reactions were initiated by addition of $MgCl_2$ to 5 mM and rifampicin to 10 μg/mL. Transcription was stopped by adding 125 μL of stop solution (0.6 M Tris pH 8.0, 12 mM EDTA, 0.16 mg/mL tRNA).

RNA from stopped transcription reactions was purified by addition of 150 μL of phenol/chloroform/isoamyl alcohol (25:24:1), vortexing, centrifugation, and collection of the aqueous phase that was then ethanol precipitated by adding 450 μL of 100% ethanol to each reaction and storage at -20 °C overnight. Precipitated RNA was resuspended in transcription loading dye (1X transcription buffer, 80% formamide, 0.05% bromophenol blue and xylene cyanol). Reactions were fractionated by electrophoresis using 12% denaturing polyacrylamide gels containing 7.5 M urea (National Diagnostics, UreaGel). Reactive bases were detected using an Amersham Biosciences Typhoon 9400 Variable Mode Imager. Quantification of bands was performed using ImageQuant. For all experiments, individual bands were normalized for incorporation of [α-^{32}P]-UTP by dividing band intensity by the number of Us in the transcript.

%Readthrough was calculated by dividing the sum of run-off RNAs by the sum of all terminated and run-off products.

C.1.5 *in vitro* transcription (cotranscriptional SHAPE-Seq experiment)

50 μL total reaction mixtures containing 100 nM linear DNA template library (see above) and 4 U of *E. coli* RNAP holoenzyme (New England Biolabs) were incubated in transcription buffer (20 mM Tris-HCl pH 8.0, 0.1 mM EDTA, 1 mM DTT and 50 mM KCl), 0.2 mg/mL bovine serum albumin, and 500 μM NTPs for 7.5 min at 37 °C to form open complexes. When present, NaF was included to a final concentration of 10 mM. Following the first incubation, EcoRI Gln111 dimer was added to a final concentration of 500 nM and incubated at 37 °C for another 7.5 min. Immediately following the second incubation, single-round transcription reactions were initiated by addition of $MgCl_2$ to 5 mM and rifampicin to 10 μg/ml. All transcription reactions were allowed to proceed for 30 seconds. Cotranscriptional experiments were then directly SHAPE modified (see RNA modification and purification below). Equilibrium refolding experiments were stopped by addition of 150 μL TRIzol solution (Life Technologies), purified, and equilibrium refolded in transcription buffer before SHAPE modification as described below (see RNA modification and purification).

C.1.6 *in vitro* transcription (single length, unlabeled)

in vitro transcription of single length, unlabeled RNA was performed as described above for cotranscriptional SHAPE-Seq, except at 25 μL total volume with 2 U of *E. coli*

RNAP holoenzyme (New England Biolabs) and without Gln111 addition or SHAPE modification. The resulting RNAs were purified as described for cotranscriptional SHAPE-Seq and fractionated using a 10% denaturing polyacrylamide gel containing 7.0 M urea. The resulting gel was stained using SYBR Gold (Life Technologies) and imaged using a Bio-Rad ChemiDoc MP system and quantified using Image Lab (Bio-Rad). %Readthrough was calculated as described above.

C.1.7 RNA modification and purification

For cotranscriptional experiments the 30 second transcription products were immediately SHAPE modified by splitting the reaction and mixing half with 2.78 μL of 400 mM benzoyl cyanide (BzCN; Pfaltz & Bower) dissolved in anhydrous dimethyl sulfoxide (DMSO; (+) sample) or anhydrous DMSO only (Sigma Aldrich; (-) sample) for ~2 seconds before addition of 75 μL of TRIzol solution. Transcription products for equilibrium refolding had 150 μL TRIzol added after *in vitro* transcription. Both were extracted according to the manufacturers protocol and dissolved in 20 μL total of 1X DNase I buffer (New England Biolabs) containing 1 U of DNase I enzyme. Digestion proceeded at 37 °C for 30 min, after which 30 μL of RNase-free H_2O was added, followed by 150 μL TRIzol. The RNA samples were then extracted again according to the manufacturer's protocol and dissolved in either: 10 μL 10% DMSO in H_2O (cotranscriptional experiments) or 25 μL RNase-free H_2O (equilibrium refolding experiments). Equilibrium refolding experiment samples were then heated to 95 °C for 2 min, snap cooled on ice for 1 min, and refolded in 1X folding buffer for 20 min at 37 °C (20 mM Tris-HCl pH 8.0, 0.1 mM EDTA, 1 mM DTT, 50 mM KCl, 0.2 mg/mL bovine serum albumin, and 500 μM NTPs), optionally containing 10 mM fluoride. RNA modification of the equilibrium refolding samples was performed as described above, followed by

the addition of 30 μL RNase-free H_2O and 150 μL TRIzol and extracted a third time according to the manufacturers instructions. The resulting pellet was dissolved in 10 μL 10% DMSO in H_2O.

C.1.8 Linker preparation

The phosphorylated linker, oligonucleotide A (Table C.2), was purchased from Integrated DNA Technologies and adenylated with the 5′ DNA Adenylation Kit (New England Biolabs) according to the manufacturers protocol at a 20X scale, dividing the reactions into 50 μL aliquots. After completion of the reaction, 150 μL TRIzol was added and the reactions were extracted according to the manufacturers instructions, dissolving the products in 20 μL RNase-free H_2O. The concentration of purified linker was measured using the Qubit Fluorometer (Life Technologies) and the molarity of the RNA was calculated using 6782.1g/mol as the molecular weight. The adenylation reaction was assumed to be 100% efficient. The linker was diluted to a 2 μM stock to be used later.

C.1.9 Linker ligation

To the modified and unmodified RNAs in 10% DMSO (see RNA modification and purification above), 0.5 μL of SuperaseIN (Life Technologies), 6 μL 50% PEG 8000, 2 μL 10X T4 RNA Ligase Buffer (New England Biolabs), 1 μL of 2 μM 5′-adenylated RNA linker, and 0.5 μL T4 RNA Ligase, truncated KQ (200 U/μL; New England Biolabs) were added to bring the total reaction volume to 20 μL. The reactions were mixed well and incubated overnight (>10 hours) at room temperature.

C.1.10 Reverse transcription

The completed linker ligations were brought to 150 μL with RNase-free H_2O before addition of 15 μL 3 M NaOAc, 1 μL 20 mg/mL glycogen, and 450 μL EtOH for EtOH precipitation. Precipitated pellets were dissolved in 10 μL RNase-free H_2O. Then, 3 μL of 0.5 μM reverse transcription primer, oligonucleotide B (Table C.2), were added. The resulting mix was then denatured completely by heating to 95 °C for 2 min, followed by an incubation at 65 °C for 5 min before placing on ice for ~30 seconds. Then, 7 μL of SSIII master mix was added, containing: 0.5 μL of Superscript III (Life Technologies), 4 μL 5X First Strand Buffer (Life Technologies), 1 μL 100 mM (DTT), 1 μL 10 mM dNTPs, and 0.5 μL RNase-free H_2O. The reaction mix was further incubated at 42 °C for 1 min, then 52 °C for 25 min and deactivated by heating at 65 °C for 5 min. The RNA was then hydrolyzed by the addition of 1 μL of 4 M NaOH solution and heating at 95 °C for 5 min. The basic solution containing the cDNA was partially neutralized with 2 μL of 1 M HCl, then precipitated with 69 μL cold EtOH, using 15 min at -80 °C and 15 min of spinning at 4 °C at max speed to pellet the RNA before washing the pellet with 70% EtOH. The washed pellet, free of base, was dissolved in 22.5 μL of nuclease-free H_2O.

C.1.11 Adapter ligation

To the cDNA, 3 μL 10X CircLigase Buffer (Epicentre), 1.5 μL 50 mM $MnCl_2$, 1.5 μL 1 mM ATP, 0.5 μL 100 μM DNA adapter, oligonucleotide C (Table C.2), and 1 μL CircLigase I (Epicentre) were added. The reaction was incubated at 60 °C for 2 hr, then 80 °C for 10 min to inactivate the ligase. The ligated DNA was EtOH precipitated with 1 μL 20 mg/mL glycogen as a carrier and dissolved in 20 μL of nuclease-free H_2O. Then the cDNA was purified using 36 μL of Agencourt XP beads (Beckman Coulter), according

to manufacturer's instructions and eluted with 20 μL TE buffer.

C.1.12 Quality analysis

For quality analysis (QA), a separate PCR reaction for each (+) and (-) sample was mixed by combining: 13.75 μL nuclease-free H_2O, 5 μL 5X Phusion Buffer (New England Biolabs), 0.5 μL 10 mM dNTPs, 1.5 μL of 1 μM labeling primer (oligonucleotides D/E (Table C.2)), 1.5 μL of 1 μM primer PE_F (oligonucleotide F (Table C.2)), 1 μL of 0.1 μM selection primer (oligonucleotides G/H (Table C.2)), 1.5 μL ssDNA library (+ or -), and 0.25 μL Phusion DNA polymerase (New England Biolabs). Both fluorescent primers were purchased from Applied Biosystems and the selection primers were purchased from Integrated DNA Technologies. Asterisks represent phosphorothioate modifications to prevent the 3'→5' exonuclease activity of Phusion polymerase. Amplification was performed for 15 cycles first, using an annealing temperature of 65 °C and an extension time of 15 seconds, excluding the PE_F primer. Then, the PE_F primer was added for an additional 10 cycles of amplification. To the complete reactions 50 μL nuclease-free H_2O was added, and the diluted reaction was ethanol precipitated. The resulting pellet was dissolved in formamide and analyzed with an ABI 3730xl capillary electrophoresis device.

C.1.13 Library preparation and next generation sequencing

To construct sequencing libraries, a separate PCR for each (+) and (-) sample was mixed by combining: 33.5 μL nuclease-free H_2O, 10 μL 5X Phusion Buffer (New England Biolabs), 0.5 μL 10 mM dNTPs, 0.25 μL of 100 μM TruSeq indexing primer (oligonucleotide

I (Table C.2)), 0.25 μL of 100 μM primer PE_F, 2 μL of 0.1 μM selection primer (+ or -, as noted above), 3 μL ssDNA library (+ or -), and 0.5 μL Phusion DNA polymerase (New England Biolabs). Amplification was performed as indicated in Quality analysis above. Completed reactions were chilled at 4 °C for 2 min before addition of 5 U exonuclease I (New England Biolabs) to remove unextended primer. The reactions were then incubated at 37 °C for 30 min. Following incubation, 90 μL of Agencourt XP beads (Beckman Coulter) were added for purification according to manufacturer's instructions. The complete libraries were eluted with 20 μL TE buffer and quantified with the Qubit 2.0 Fluorometer (Life Technologies). To prepare the libraries for sequencing, the average length of each sample was determined using the results from the quality analysis in order to calculate the molarity of each (+) or (-) separately. Sequencing pools were mixed to be equimolar, such that all of the sequencing libraries were present in the solution at the same level. Sequencing was performed on the Illumina HiSeq 2500 in either rapid run or high output mode, using 2x35 bp paired end reads. To help overcome the low-complexity of the linker region during sequencing 10-20% PhiX DNA was included.

C.1.14 Data analysis with Spats

All of the cotranscriptional SHAPE-Seq computational tools used in this study can be found on github at: `https://github.com/LucksLab/Cotrans_SHAPE-Seq_Tools`. Reactivity spectra were calculated using Spats v1.0.0 (`https://github.com/LucksLab/spats/releases/`) and a number of utility scripts to prepare the Illumina HiSeq output for Spats. First, target fasta files were prepared for each RNA sequence by enumerating all of its transcript intermediate lengths (beginning with 20 nt) and appending the 3' RNA linker sequence (CUGACUCGGGCACCAAGGA) to

each. Reads were then mapped and processed for Spats v1.0.0 as described in Watters *et al.* [123]. First, Illumina adapter sequences were trimmed from each read using cutadapt v1.5 (`https://cutadapt.readthedocs.org/en/stable/`), as part of the adapter_trimmer script, available as part of the Spats package (see above). Then the paired end reads were aligned to the enumerated target RNA sequences with Bowtie 0.12.8, with the (+) and (-) sample reads divided according to the SHAPE-Seq handle sequences [123]. In this setup, the SHAPE modification position was determined by the 5′ end of Read 2 (the 3′ end of the cDNA) and the 3′ end of the intermediate length transcript was determined by the sequence of Read 1 (containing the 3′ linker sequence and 3′ end of the RNA) (Figure 5.1B). Unique mapping of each paired-end read to the targets file containing all possible 3′ ends binned reads by RNA length before reactivity profile calculation of each length. Each length i contained a set of $\theta_{i,j}$ values, where $\theta_{i,j}$ is the probability of modification at a particular nucleotide j relative to the rest of the nucleotides in sequence i. The calculated $\theta_{i,j}$ values were then converted to $\rho_{i,j}$ values according to:

$$\rho_{i,j} = \theta_{i,j} L_i \tag{C.1}$$

where L_i is the length of the intermediate transcript i. By definition, $\theta_{i,j}$ must sum to 1 for length i, since it is a relative probability of modification within the sequence i [97]. We converted reactivities to $\rho_{i,j}$ to set the average relative reactivity to 1 across each sequence i [123] in order to better compare reactivities between RNAs of different transcript lengths by mitigating the influence of RNA length on reactivity.

To create the reactivity matrix, the $\rho_{i,j}$ were plotted using Matlab's pcolor with the Cotrans_matrix_rhos_processing_2D.m (2D matrices), Cotrans_matrix_rhos_processing_differences.m ($\Delta\rho$ matrices), or Cotrans_matrix_rhos_processing_3D.m (3D matrices) files in the Cotrans_SHAPE-Seq_Tools repository (see above). All data are freely accessible from the RNA Mapping Database (RMDB) (`http://rmdb.stanford.edu/`

repository/) using the IDs in Table C.3.

C.2 Supplementary Text

C.2.1 Analysis of the SRP Cotranscriptional Folding Pathway

Early in transcription, our reactivity data are consistent with the formation of a small hairpin structure near the 5′ end of the SRP RNA (Figure 5.2). As the nucleotides that comprise the apical loop of the final SRP RNA structure begin to exit the polymerase, the elongated helical structure begins to form, resulting in sequentially decreasing reactivity values for nucleotides (nts) 25-39 on the 5′ side of the SRP RNA as nts 92-108 are added to the 3′ end of the growing RNA. There are also increased reactivities at positions 41-44 and 57-58 that correspond to the formation of the inner and apical loops in the SRP helical structure as it stabilizes during elongation. Last, we observe reactivity decreases in nts 11-18 of the early-formed hairpin loop that indicate the cotranscriptional rearrangement into the extended helical structure occurs after the SRP RNA reaches a length of 117 nt. We observed similar transitions when intermediate SRP RNA fragments were refolded at equilibrium (Figure C.2), although the cotranscriptional transitions occur later due to the 14 nt RNAP footprint protecting the RNA 3′ ends [246].

C.2.2 Mutant analysis of the *B. cereus* fluoride riboswitch

Previous work on the *B. cereus* fluoride riboswitch examined a family of mutations in the aptamer domain and the terminator stem to evaluate the riboswitch expression

platform [50]. Each mutant showed distinct changes in ligand binding and termination capability that correspond to its location in the riboswitch (Figure C.8). Therefore, we used the previously studied mutants to both corroborate our interpretations of the wt riboswitch cotranscriptional SHAPE-Seq data (Figures 5.3 and 5.4) and uncover details of how individual mutations impact the cotranscriptional folding and function of the riboswitch.

Mutant M20 (G69A/A70U) contains base substitutions in the 3' terminator stem that render the intrinsic terminator nonfunctional, but maintain the wt aptamer domain (Figure C.8). As anticipated, the M20 mutant undergoes all wt transitions associated with aptamer formation and fluoride binding (Figure C.11). In both the presence and absence of fluoride, we observe initial PK1 folding as a decrease in P1 loop reactivity across transcript lengths 59-61. Without fluoride, PK1 folding occurs alongside increased reactivity at A10. A10 reactivity remains low with fluoride however, likely due to a long-range non-canonical base pair with U38. Similarly, in the presence of fluoride the reactivity values for A22 increase dramatically as PK1 folds. Without fluoride the increase in reactivity of A22 is smaller and more gradual across transcript lengths 59 to 74.

The terminator stem mismatches in M20 have a distinct effect on transcription termination in the absence of fluoride. In the wt riboswitch, PK1 opening begins at length 77 concurrently with the closure of nts 52-55 in the terminator stem and is completely open by length 80. In M20, PK1 opening is more gradual, beginning at length 77 and reaching completion at length 84. This effect is consistent with mismatches in the terminator stem interrupting terminator hairpin winding. The M20 mutation does not alter the transitions that lead to antitermination. In the presence of fluoride, low P1 loop reactivity and stable high reactivity at A22 indicate the persistence of PK1. Further, high

reactivities at nts 52-55 across transcript lengths 68 to 87 suggest that terminator stem closure is delayed (Figure C.11). In contrast to wt, M20 does not show any signature of termination at the end of the polyU tract (nts 80-82) because the M20 terminator is nonfunctional (Figure C.8).

Mutants M18 (G13A/A14U) (Figure C.9) and M19 (U45A/C46U) (Figure C.10) interrupt base pairing in PK1, thereby preventing its folding. M22 (Figure C.13) combines M19 and M20 to restore complementarity in the terminator stem but remains defective in PK1 folding. Consequently, M18, M19, and M22 cannot properly form the fluoride aptamer and are therefore fluoride-insensitive. We observe this defect in aptamer formation as the lack of any fluoride-dependent transitions, the most notable of which are the absence of decreasing P1 loop reactivities, indicating that PK1 does not form, and low reactivity at A22 across all transcript lengths. The distinct reactivity patterns that can be observed at A22 when: 1) PK1 cannot form (M18, M19, and M22), 2) PK1 forms without fluoride (wt; Figure 5.4A), and 3) PK1 forms with fluoride (wt; Figure 5.4A) correlate strongly with fluoride binding and further support the conclusion that the reactivity at A22 is a strong indicator of aptamer state.

The M21 (M18 and M19) (Figure C.12) and M23 (M18, M19, and M20) (Figure C.14) mutants restore complementarity in PK1, and therefore the capability of the aptamer to bind fluoride. As expected, we observe similar reactivity patterns within the matrices of M21, M23, and wt during the formation of the aptamer domain (Figures C.12 and C.14). Interestingly, differences in the P1 loop reactivity patterns between these mutants and the wt riboswitch indicate that PK1 does not stabilize as readily in M21 and M23, likely due to the replacement of a G:C base pair in PK1 with A:U. This effect is particularly visible for M21, where the P1 loop remains highly reactive across all transcript lengths in the absence of fluoride. However, the characteristic increase in

A22 still occurs, suggesting that PK1 still forms, but for a smaller fraction of the folded population. Likely, the weaker base pairing of PK1 in M21 causes frequent breaking and reforming of PK1, resulting in higher reactivities. The binding of fluoride could then trap the aptamer while interconverting the two forms, resulting in a trapped, fluoride-bound form.

C.2.3 Cotranscriptional SHAPE-Seq accesses non-equilibrium, kinetically trapped RNA structures

Equilibrium analysis was performed in order to assess whether cotranscriptional SHAPE-Seq accesses non-equilibrium, kinetically trapped folded states of nascent RNAs. To equilibrium refold all intermediate transcript lengths of the fluoride riboswitch, we first generated all of the intermediate transcript lengths as done for cotranscriptional SHAPE-Seq, but extracted, denatured, and equilibrium refolded the RNAs in transcription buffer prior to chemical probing (Figures 5.4B and C.17). Several distinct differences between cotranscriptional and equilibrium reactivity profiles indicate that cotranscriptional experiments probe non-equilibrium folding states of the nascent RNAs.

Because cotranscriptional SHAPE-Seq experiments probe RNAs that exist as part of a transcription elongation complex, only folding events that involve nts that have left, or are in, the RNA exit channel of RNAP can be observed. During equilibrium refolding however, RNAP is not present to prevent the 3′ end of the transcript from folding with the rest of the RNA, causing the structural transitions that we observe with cotranscriptional SHAPE-Seq to occur at shorter transcript lengths. Specifically, we observe ligand-independent folding of PK1 (via decreases in P1 loop reactivities) at

transcript length 47 with equilibrium refolding (Figure C.17) as opposed to length 58 (Figures 5.3 and 5.4) when probing the arrested elongation complexes cotranscriptionally. In cotranscriptional SHAPE-Seq experiments, length 58 corresponds to a point when nt 47 is expected to be leaving the RNA exit channel. Shifts in structural transitions to earlier lengths during equilibrium refolding were also observed for the SRP RNA (Figures 5.2 and C.2).

The most striking evidence that cotranscriptional SHAPE-Seq probes non-equilibrium RNA structures is the distinct fluoride-independent behavior of the P1 loop at later transcript lengths. Cotranscriptional SHAPE-Seq experiments show low P1 loop reactivities following aptamer formation in the presence of fluoride, indicating the formation and persistence of PK1 (Figure 5.4A). In contrast, equilibrium refolding produces a sharp, fluoride-independent rise in reactivities of the P1 loop nucleotides over transcript lengths 66-69, indicating the complete opening of PK1 (Figure 5.4B). The opening of PK1 at these lengths occurs simultaneously with the formation of the terminator stem-loop (nts 49-66) that grows to incorporate nts 46 and 47 (to pair with nts 68 and 69, respectively), thereby precluding them from participating in PK1. However, in the cotranscriptional SHAPE-Seq experiment PK1 remains stabilized at longer transcript lengths after binding fluoride. The observed deviation from equilibrium structures indicates that the RNAs probed are out of equilibrium. Thus, equilibrium refolding analysis does not permit meaningful analysis of RNAs beyond length 66 because terminator hairpin formation dominates the structural population regardless of fluoride concentration. These results highlight the importance of cotranscriptional folding to the function of riboswitches.

C.3 Supplementary Figures

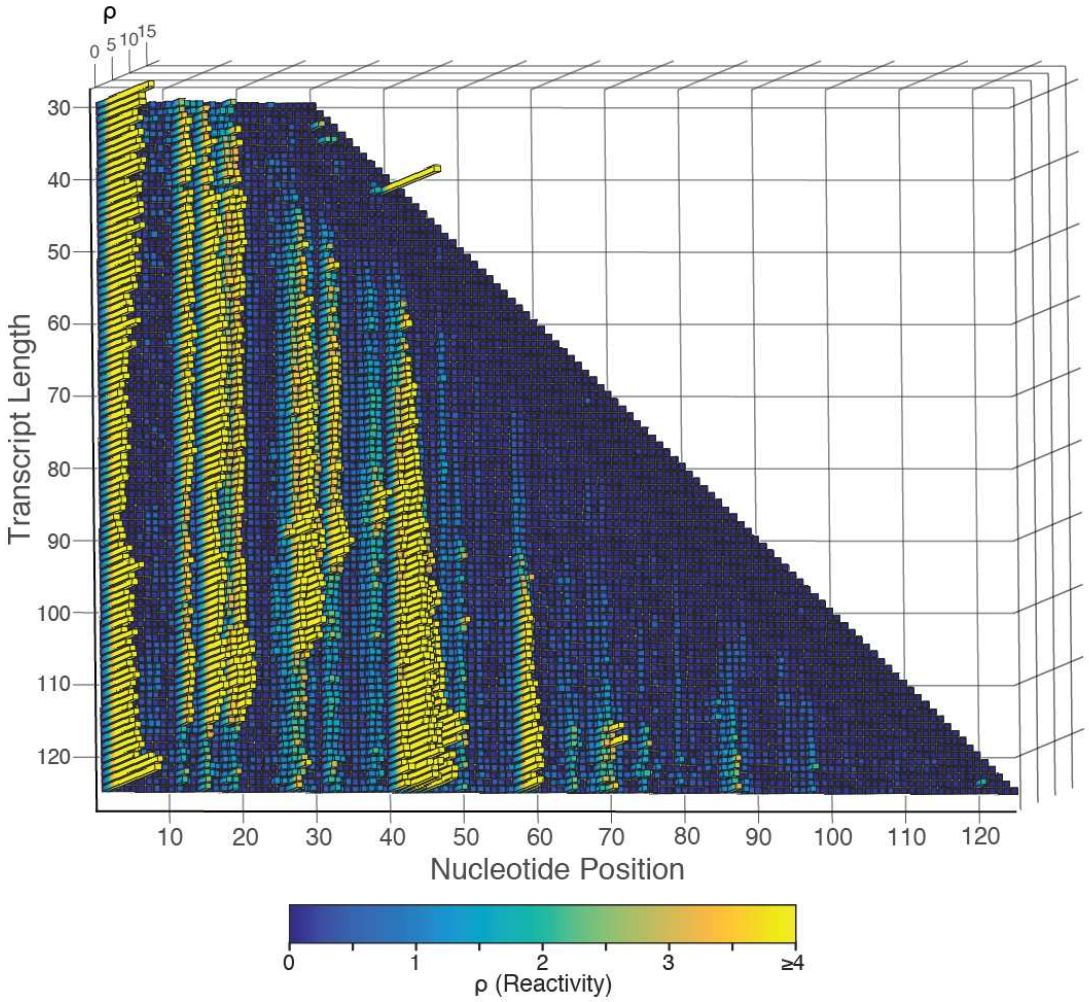
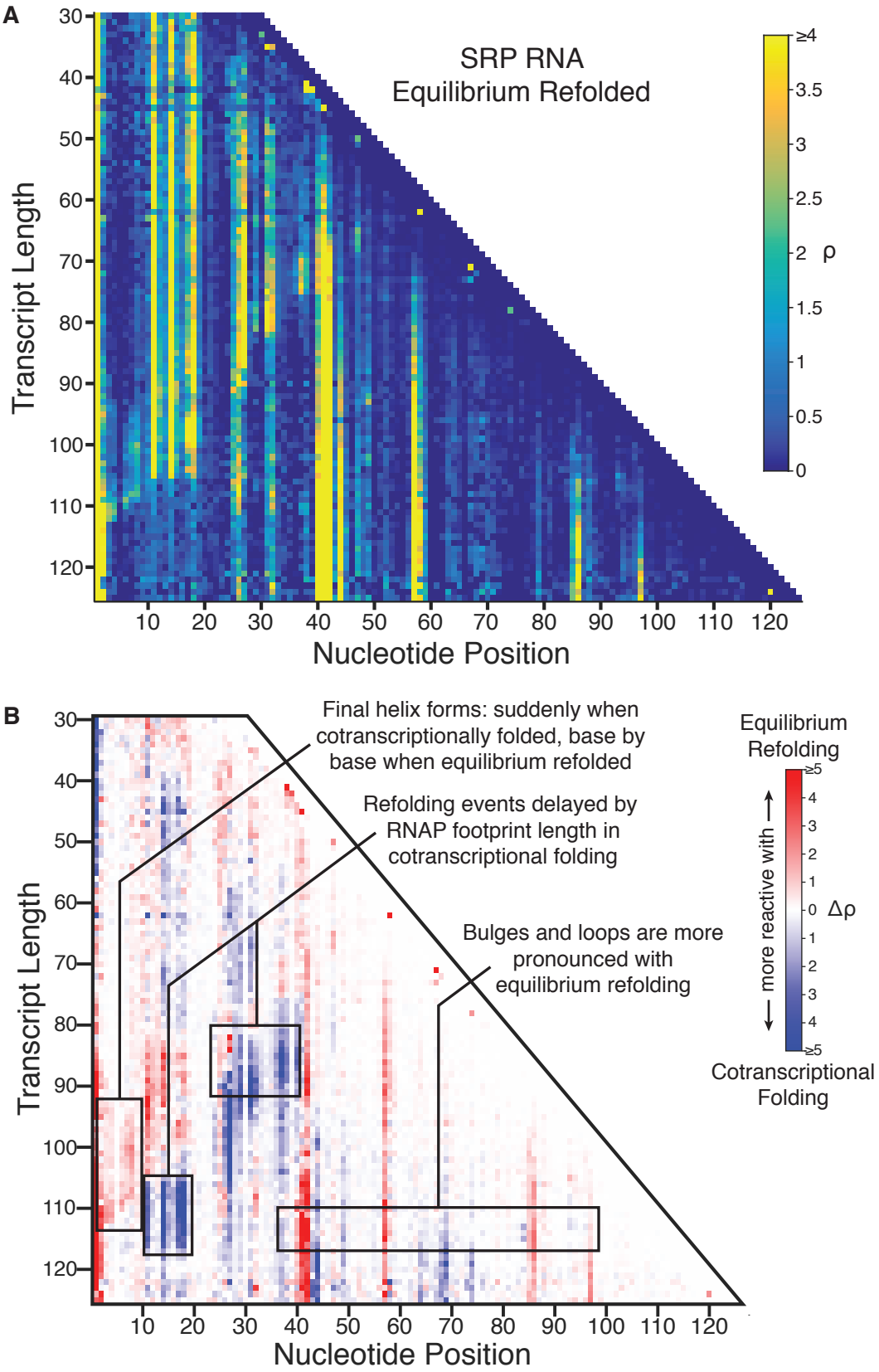


Figure C.1: 3D SRP RNA cotranscriptional SHAPE-Seq data. Three-dimensional representation of the cotranscriptional SHAPE-Seq reactivity data for the SRP RNA. Bar heights represent reactivity intensity at each nucleotide across all transcript lengths. The reactivity gradient (bottom) colors the bars. Explanations of the data and a top-down, two-dimensional view are available as part of Figure 5.2.

Figure C.2: SHAPE-Seq data for SRP RNA equilibrium refolded. (A) SHAPE-Seq reactivity matrix for all of the intermediate transcript lengths of the SRP RNA that were synthesized via *E. coli* RNAP and Gln111 roadblocking, extracted, denatured, and equilibrium refolded in transcription buffer before SHAPE modification. (B) Reactivity differences ($\Delta\rho$) between the cotranscriptionally and equilibrium folded SRP RNA transcript lengths. Red nucleotides are more reactive when equilibrium refolded, blue nucleotides are more reactive when cotranscriptionally folded. Due to the lack of an RNAP footprint at the 3' ends of the transcripts, the equilibrium refolded lengths display transitions ~14 nt earlier than with cotranscriptional folding [246].


A
SRP RNA
Equilibrium Refolded
Transcript Length
Nucleotide Position
ρ
≥4
3.5
3
2.5
2
1.5
1
0.5
0
B
Final helix forms: suddenly when cotranscriptionally folded, base by base when equilibrium refolded
Refolding events delayed by RNAP footprint length in cotranscriptional folding
Bulges and loops are more pronounced with equilibrium refolding
Equilibrium Refolding
more reactive with
Δρ
Cotranscriptional Folding

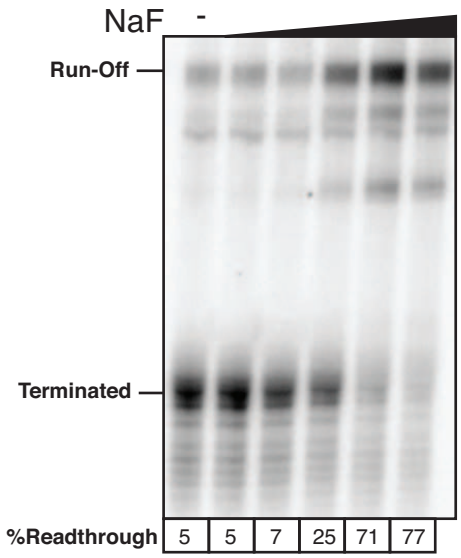
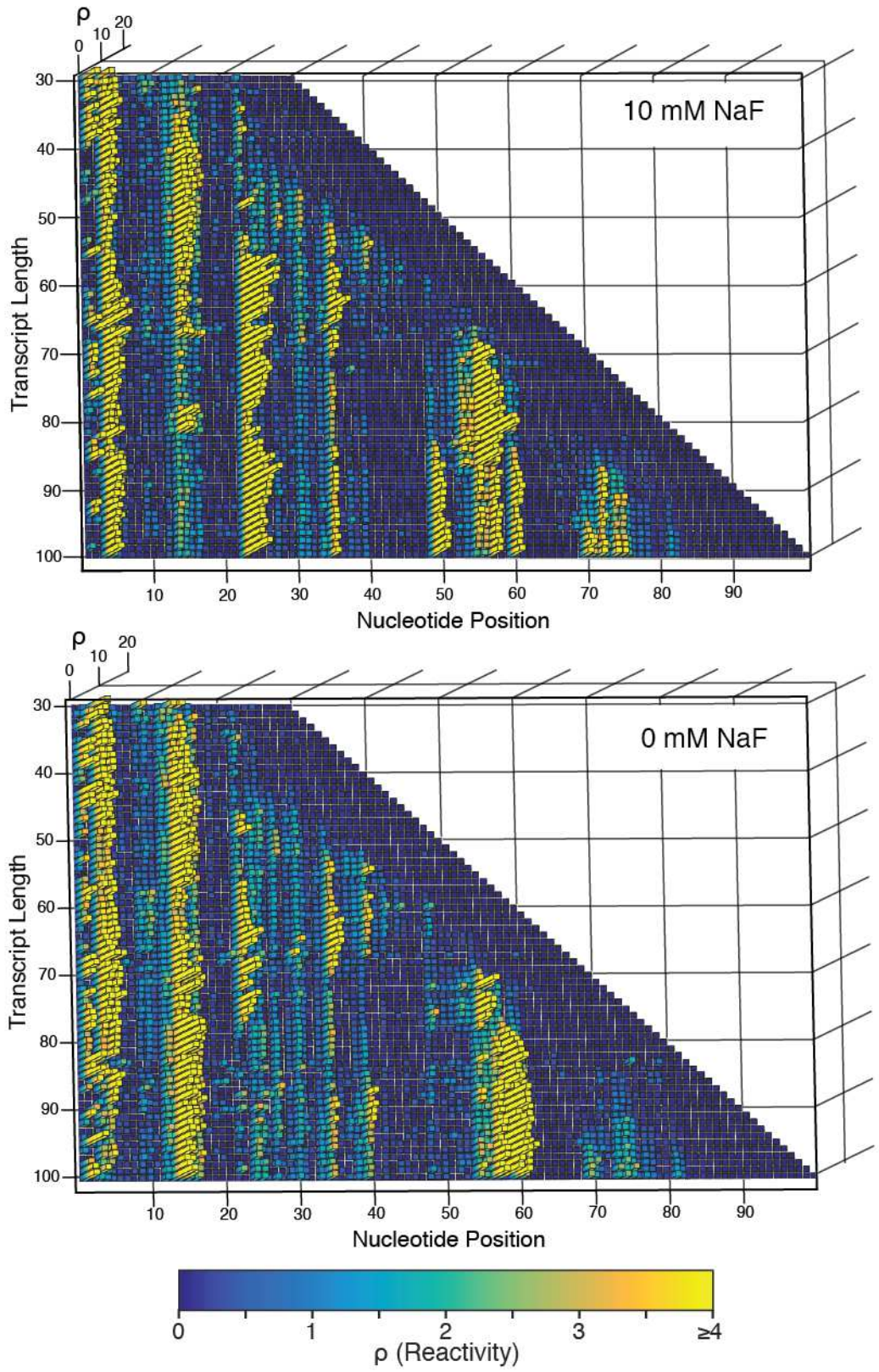


Figure C.3: Transcription antitermination by the *B. cereus* crcB fluoride riboswitch. Single-round *in vitro* transcription of the *B. cereus* crcB fluoride riboswitch. NaF was included at concentrations of 0 mM, 0.001 mM, 0.01 mM, 0.1 mM, 1 mM, and 10 mM. Terminator readthrough increases in response to fluoride.

Figure C.4: 3D wt fluoride riboswitch cotranscriptional SHAPE-Seq data. Three-dimensional representation of the cotranscriptional SHAPE-Seq reactivity data for the wt *B. cereus* fluoride riboswitch, transcribed with either 10 mM (top) or 0 mM NaF (bottom). Bar heights represent reactivity intensity at each nucleotide across all transcript lengths. The reactivity gradient (bottom) colors the bars. The reactivity cluster at nts 12-16 for transcript lengths 79-82 transcribed with 10 mM NaF is likely due to fluoride-independent termination of complexes stalled over the terminator polyU sequence. Top-down, two-dimensional views displayed in Figure 5.3.


ρ
0
10
20
10 mM NaF
Transcript Length
30
40
50
60
70
80
90
100
10
20
30
40
50
60
70
80
90
Nucleotide Position
ρ
0
10
20
0 mM NaF
Transcript Length
Nucleotide Position
0
1
2
3
≥4
ρ (Reactivity)

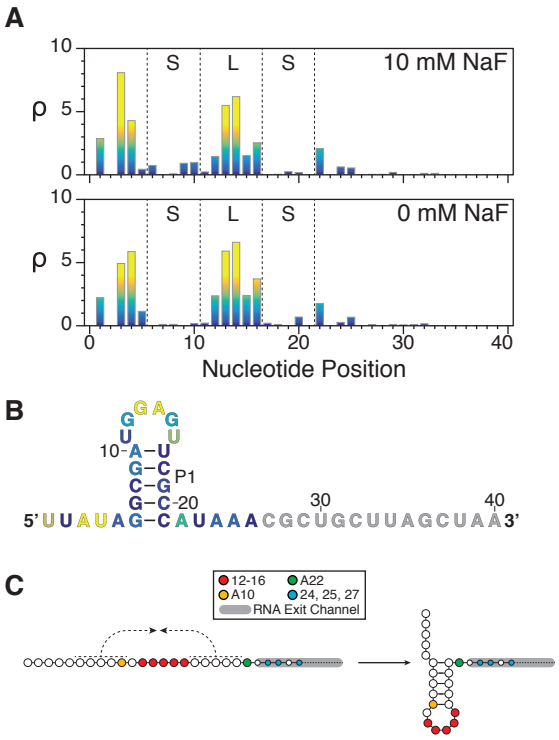


Figure C.5: Folding of the P1 stem. (A) Cotranscriptional SHAPE-Seq reactivities (from Figure 5.3) for transcript length 40 of the *B. cereus* fluoride riboswitch are shown in the presence (top) and absence (bottom) of fluoride. S and L indicate P1 stem and loop nts, respectively. (B) The secondary structure of transcript length 40 is inferred from cotranscriptional SHAPE-Seq reactivities, covariation analysis (24), and crystallographic data [248]. Nucleotide colors indicate cotranscriptional SHAPE-Seq reactivities with 10 mM NaF from (A). Gray nucleotides exist within the RNAP footprint at this length [246]. (C) Cartoon representation of P1 stem folding. Nucleotide colors correspond to key sequence elements as described in the main text and Figure 5.4.

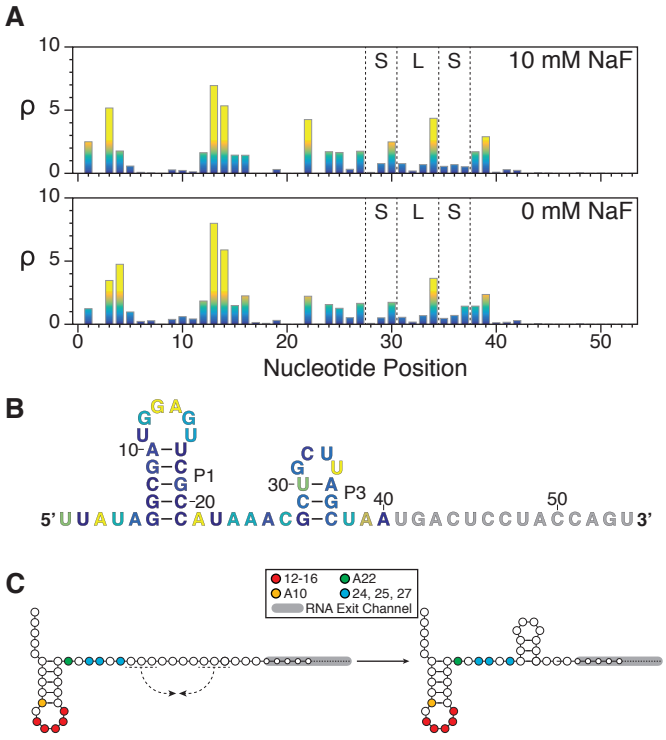


Figure C.6: Folding of the P3 stem. (A) Cotranscriptional SHAPE-Seq reactivities (from Figure 5.3B) for transcript length 54 of the *B. cereus* fluoride riboswitch are shown in the presence (top) and absence (bottom) of fluoride. S and L indicate P3 stem and loop nts, respectively. (B) The secondary structure of transcript length 54 is inferred from cotranscriptional SHAPE-Seq reactivities, covariation analysis [50], and crystallographic data [248]. Nucleotide colors indicate cotranscriptional SHAPE-Seq reactivities with 10 mM NaF from (A). Gray nucleotides exist within the RNAP footprint at this length [246]. (C) Cartoon representation of P3 stem folding is shown. Nucleotide colors correspond to key sequence elements as described in the main text and Figure 5.4.

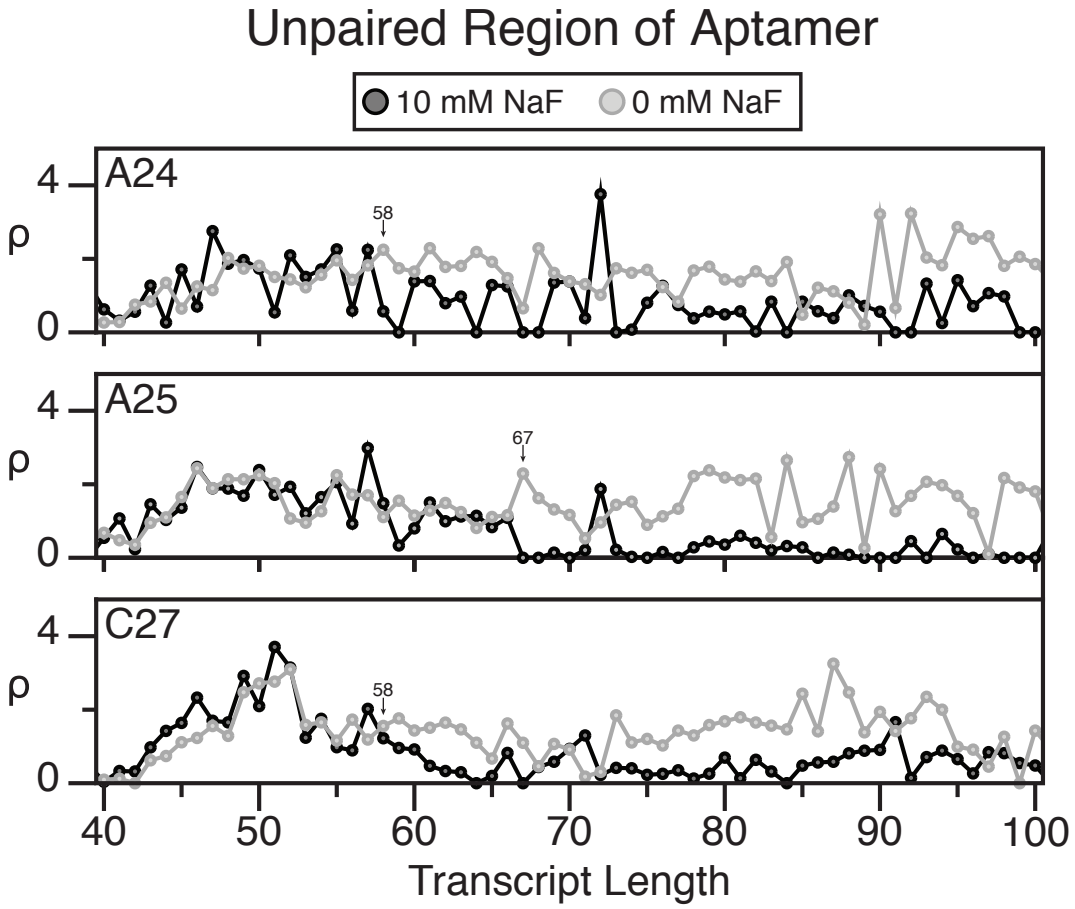
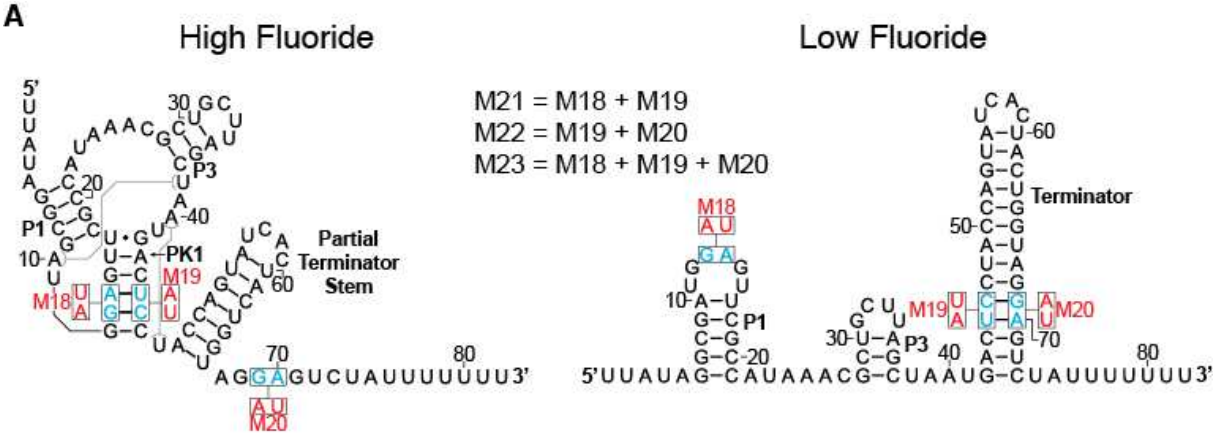


Figure C.7: Reactivity profiles for A24, A25, and C27 of the wt fluoride riboswitch over the course of transcription. Single nucleotide trajectories displaying changes in the reactivities of nucleotides A24, A25, and C27 of the wt fluoride riboswitch when transcribed with either 0 mM (gray) or 10 mM NaF (black). Data is taken from the cotranscriptional SHAPE-Seq matrices in Figure 5.3B. These nucleotides are found between P1 and P3 in the fluoride riboswitch and exhibit lower reactivities after fluoride binds the aptamer around length 70. Positions U23 and A26 in this region show persistently low reactivities throughout transcription.

Figure C.8: *B. cereus* fluoride riboswitch mutants. (A) The locations of mutations M18-M23 from Baker *et al.* [50] within the antiterminated (high fluoride) and terminated (low fluoride) secondary structures. (B) Table of mutant properties. Aptamer folding was determined from cotranscriptional SHAPE-Seq trajectories. Terminator activity was determined from part (C) and is consistent with reporter fusion assays performed in [50]. (C) Single-round *in vitro* transcription of the mutants shown in (A) in the presence and absence of 10 mM NaF. Transcription reactions were performed in conditions matching those used for cotranscriptional SHAPE-Seq. Lane M is the RNA Century Marker (Ambion) with 100 nt and 200 nt bands shown. Templates with a functional terminator (wt, M18, M22, and M23) show a higher basal level of terminator readthrough in cotranscriptional SHAPE-Seq conditions than in conditions used for radiolabeling (Figure C.3), which contain a 10-fold lower UTP concentration. Both the wt riboswitch and the mutants follow previously described trends [50]. Mutations that disrupt aptamer formation but allow terminator formation (M18 and M22) terminate at comparable efficiencies to wt without fluoride, regardless of fluoride concentration. Mutations that disrupt the terminator hairpin (M19, M20, and M21) do not terminate regardless of fluoride concentration. The combination of the M18, M19, and M20 mutations in M23 restores both aptamer and terminator hairpin base pairing and behaves as the wt riboswitch.

A



B

	M18	M19	M20	M21	M22	M23
Aptamer Folding	No	No	Yes	Yes	No	Yes
Terminator Active	Yes	No	No	No	Yes	Yes

C

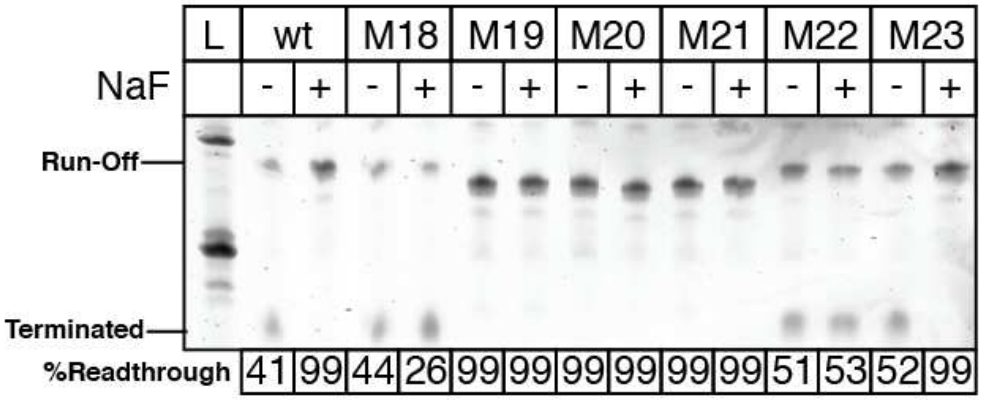
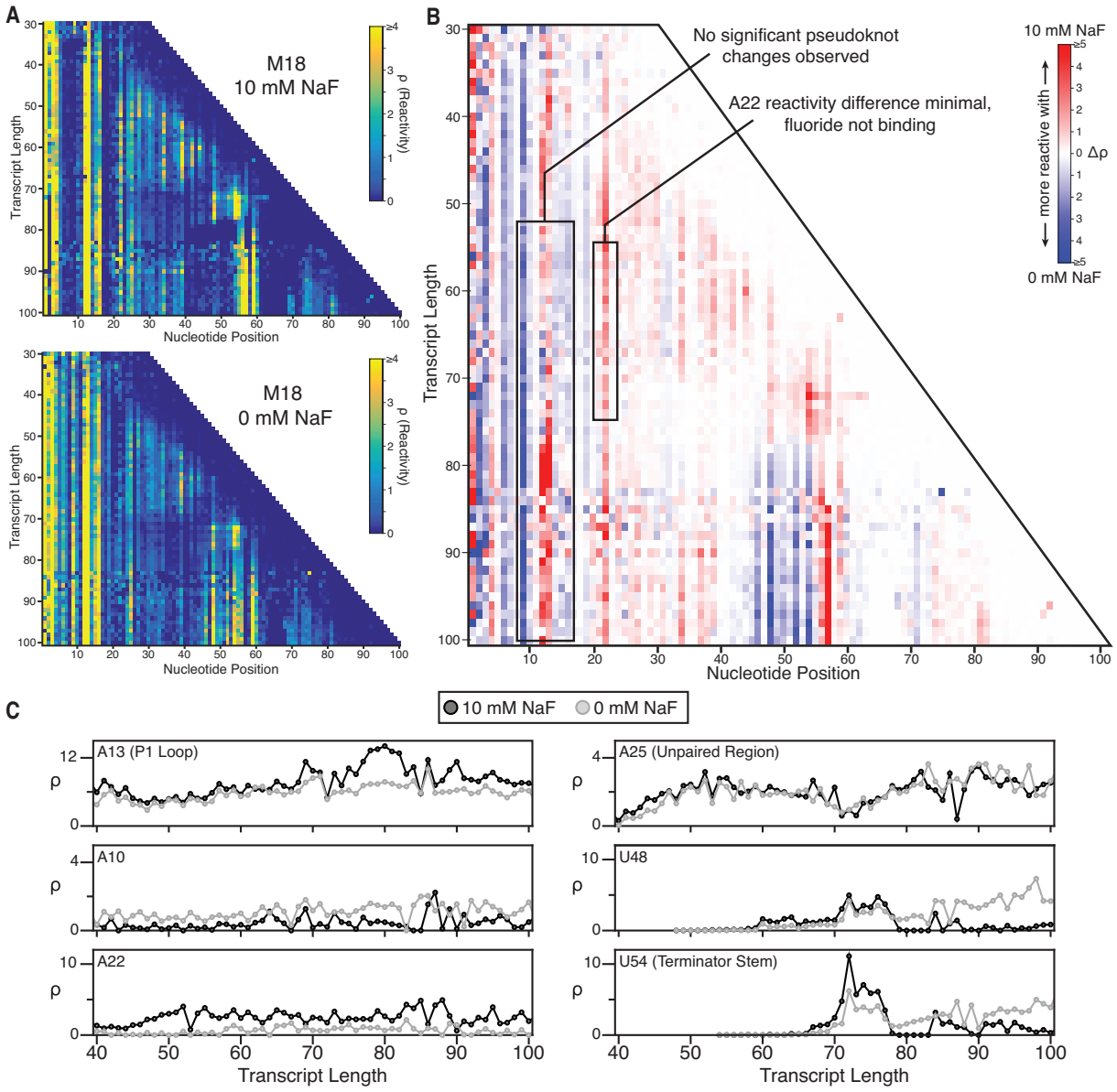


Figure C.9: Cotranscriptional SHAPE-Seq data for fluoride riboswitch mutant M18. (A) Reactivity matrices of the M18 mutant (pseudoknot disrupted; Figure C.8) transcribed with either 10 mM (top) or 0 mM NaF (bottom). Nucleotides are colored by reactivity intensity. (B) Reactivity differences ($\Delta\rho$) between the matrices in (A). Positions that are red are more reactive when fluoride is present, while positions that are blue are more reactive without fluoride. (C) Single nucleotide trajectories displaying changes in the reactivities of key nucleotides, or nucleotides representing key regions, when transcribed with either 0 mM (gray) or 10 mM NaF (black). The lack of bifurcation in the trajectories of nts A13, A10, A22 and A25 compared to the wt riboswitch (Fig. 4A) agree with the inability of the M18 mutant to respond to fluoride (Figure C.8). The transitions in U54 are consistent with terminator formation both with and without fluoride. Overall, the trajectories observed suggest that regardless of fluoride concentration, the M18 mutant only follows the terminator forming folding pathway.


A
M18
10 mM NaF
M18
0 mM NaF
Transcript Length
Nucleotide Position
ρ (Reactivity)
B
No significant pseudoknot changes observed
A22 reactivity difference minimal, fluoride not binding
10 mM NaF
more reactive with
Δρ
0 mM NaF
Transcript Length
Nucleotide Position
C
10 mM NaF
0 mM NaF
A13 (P1 Loop)
A25 (Unpaired Region)
A10
U48
A22
U54 (Terminator Stem)
ρ
Transcript Length

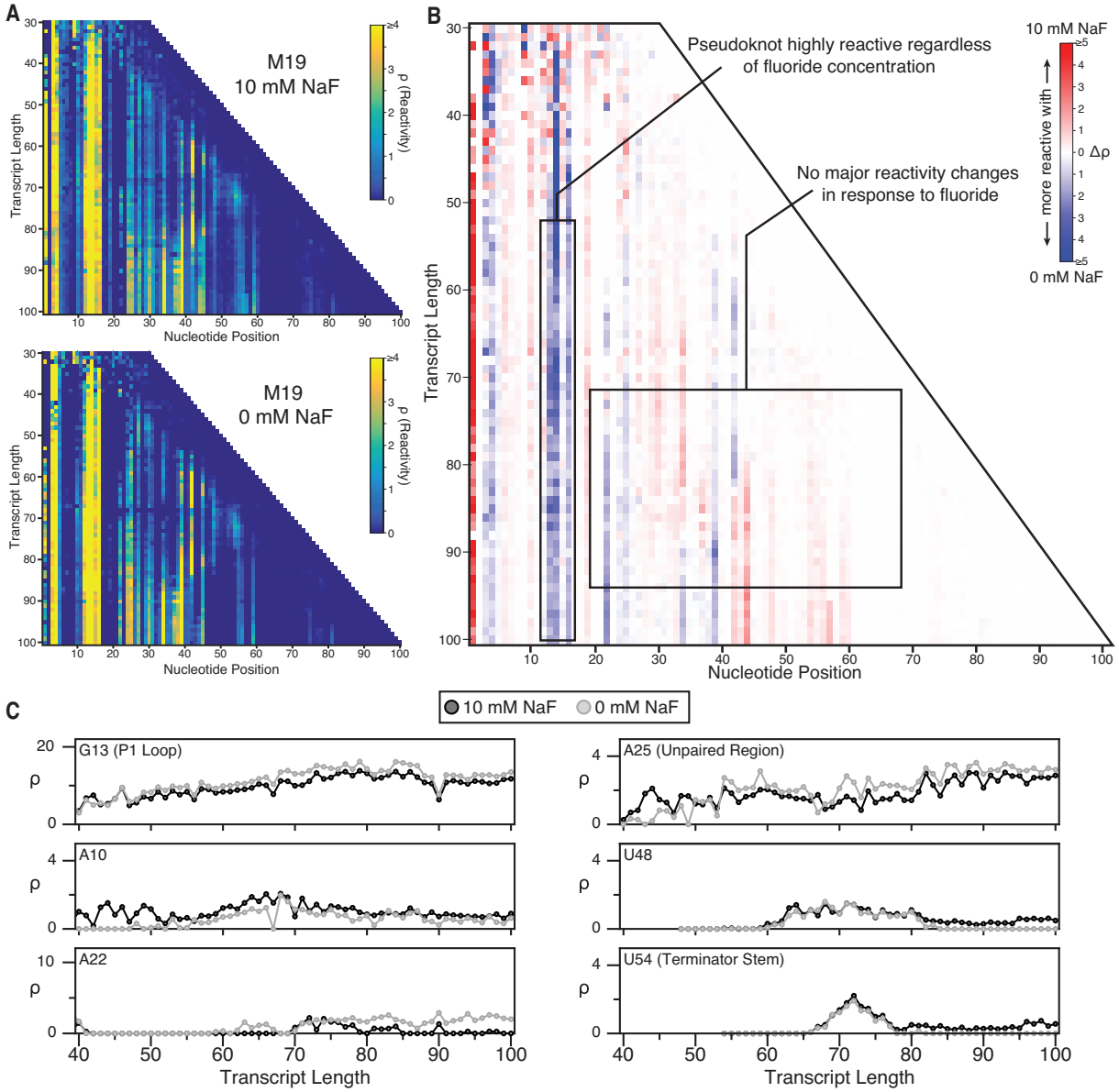


Figure C.10: Cotranscriptional SHAPE-Seq data for fluoride riboswitch mutant M19. (A) Reactivity matrices of the M19 mutant (pseudoknot and lower terminator stem disrupted; Figure C.8) as described in Figure C.9A. (B) Reactivity differences ($\Delta\rho$) between the matrices in (A) as described in Figure C.9B. (C) Single nucleotide reactivity trajectories as described in Figure C.9C. The lack of bifurcation in the trajectories of nts G13, A10, A22, and A25 compared to the wt riboswitch (Figure 5.4A) agree with the inability of the M19 mutant to respond to fluoride. The transitions in U54 are consistent with upper terminator stem formation both with and without fluoride.

Figure C.11: Cotranscriptional SHAPE-Seq data for fluoride riboswitch mutant M20. (A) Reactivity matrices of the M20 mutant (lower terminator stem disrupted; Figure C.8) as described in figure Figure C.9A. (B) Reactivity differences ($\Delta\rho$) between the matrices in (A) as described in Figure C.9B. (C) Single nucleotide reactivity trajectories as described in Figure C.9C. The data in (B) and (C) show clear differences in the reactivities of the PK1 nucleotides (nts 12-16), position A22, and the terminator hairpin due to the presence of fluoride that match well to the differences observed for the wt riboswitch (Figures 5.3 and 5.4). Since the M20 mutant and the wt contain the same aptamer domain (nts 6-48), M20 would be expected to display the same fluoride-mediated transitions as observed for the wt. Similar upper terminator stem transitions are observed for M20 as for wt (Figures 5.3 and 5.4) despite the fact that the M20 mutation disrupts the lower terminator stem (Figure C.8). However, the increase in reactivities observed in the terminator loop during the later stages of folding the wt riboswitch (Figure 5.3) are not observed for M20.

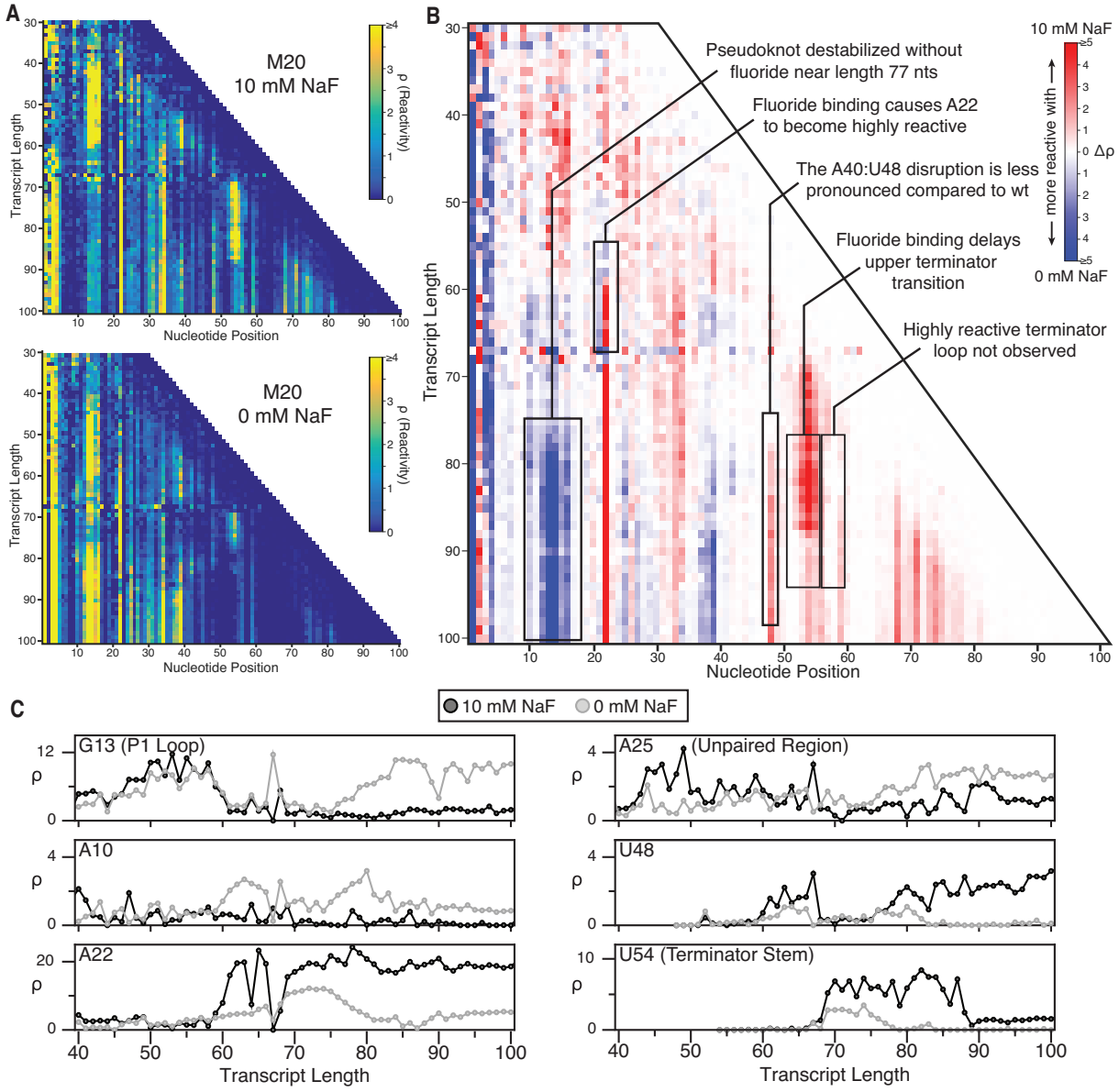

A
M20
10 mM NaF
M20
0 mM NaF
Transcript Length
Nucleotide Position
ρ (Reactivity)
B
Pseudoknot destabilized without fluoride near length 77 nts
Fluoride binding causes A22 to become highly reactive
The A40:U48 disruption is less pronounced compared to wt
Fluoride binding delays upper terminator transition
Highly reactive terminator loop not observed
10 mM NaF
more reactive with
Δρ
0 mM NaF
C
10 mM NaF
0 mM NaF
G13 (P1 Loop)
A25 (Unpaired Region)
A10
U48
A22
U54 (Terminator Stem)
ρ
Transcript Length

Figure C.12: Cotranscriptional SHAPE-Seq data for fluoride riboswitch mutant M21. (A) Reactivity matrices of the M21 mutant (compensatory mutations in the pseudoknot;Figure C.8) as described in Figure C.9A. (B) Reactivity differences ($\Delta\rho$) between the matrices in (A) as described in Figure C.9B. (C) Single nucleotide reactivity trajectories as described in figure S9C. Overall, the M21 mutant behaves similarly to the M20 mutant (Figure C.11), as both mutations produce functional aptamers but nonfunctional terminators (Figure C.8). Differences in the reactivities of M21 with 0 mM or 10 mM NaF match well to the analogous differences observed for the wt, supporting the conclusion that the M21 mutant can bind fluoride. M21 differs from M20 in two notable aspects however. First, the P1 loop does not readily decrease in reactivity at length ~58 without fluoride, suggesting that PK1 formation is less stable in M21 than in M20. Second, nts 43-44 are highly reactive in the presence of fluoride for M21. Both of these differences can be interpreted as changes of PK1 structure in M21 due to the weaker base pairing within PK1 compared to the M20 or wt aptamers.

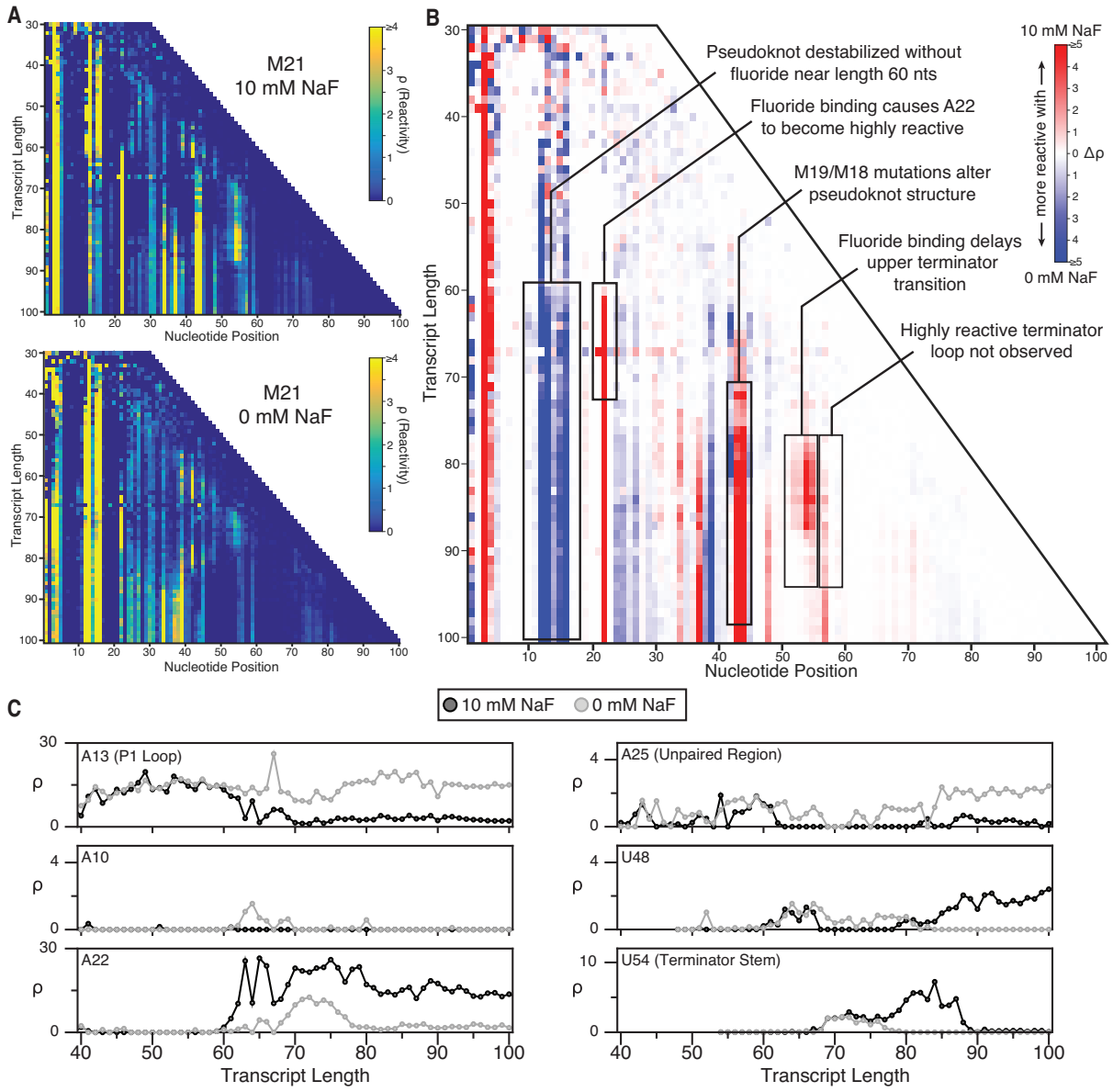

A
M21
10 mM NaF
M21
0 mM NaF
Transcript Length
Nucleotide Position
ρ (Reactivity)
B
Pseudoknot destabilized without fluoride near length 60 nts
Fluoride binding causes A22 to become highly reactive
M19/M18 mutations alter pseudoknot structure
Fluoride binding delays upper terminator transition
Highly reactive terminator loop not observed
10 mM NaF
more reactive with
Δρ
0 mM NaF
Transcript Length
Nucleotide Position
C
10 mM NaF
0 mM NaF
A13 (P1 Loop)
A10
A22
A25 (Unpaired Region)
U48
U54 (Terminator Stem)
ρ
Transcript Length

Figure C.13: Cotranscriptional SHAPE-Seq data for fluoride riboswitch mutant M22. (A) Reactivity matrices of the M22 mutant (pseudoknot disrupted; Figure C.8) as described in Figure C.9A. (B) Reactivity differences ($\Delta\rho$) between the matrices in (A) as described in Figure C.9B. (C) Single nucleotide reactivity trajectories as described in Figure C.9C. Overall, the M22 mutant behaves similarly to the M18 mutant (Figure C.9) as expected, since both mutations produce nonfunctional aptamers but functional terminators (Figure C.8). The trajectories observed suggest that regardless of fluoride concentration, the M22 mutant only follows the terminator forming folding pathway, like the M18 mutant. Higher reactivities were observed for A22 in the presence of fluoride, although the level of reactivity was much lower and more uniform across all lengths relative to the length-dependent increases observed at this position in the wt riboswitch in the presence of fluoride.

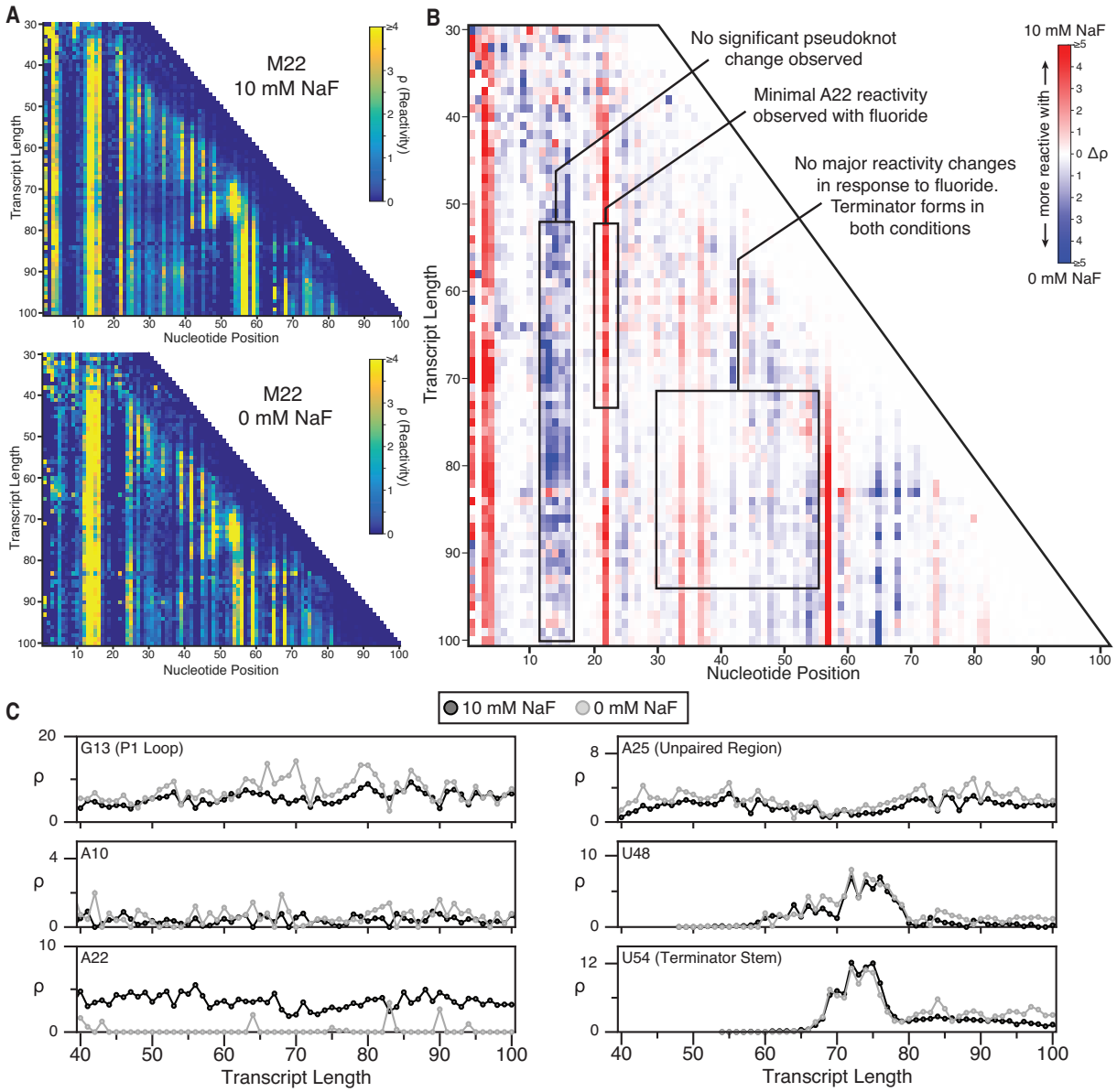
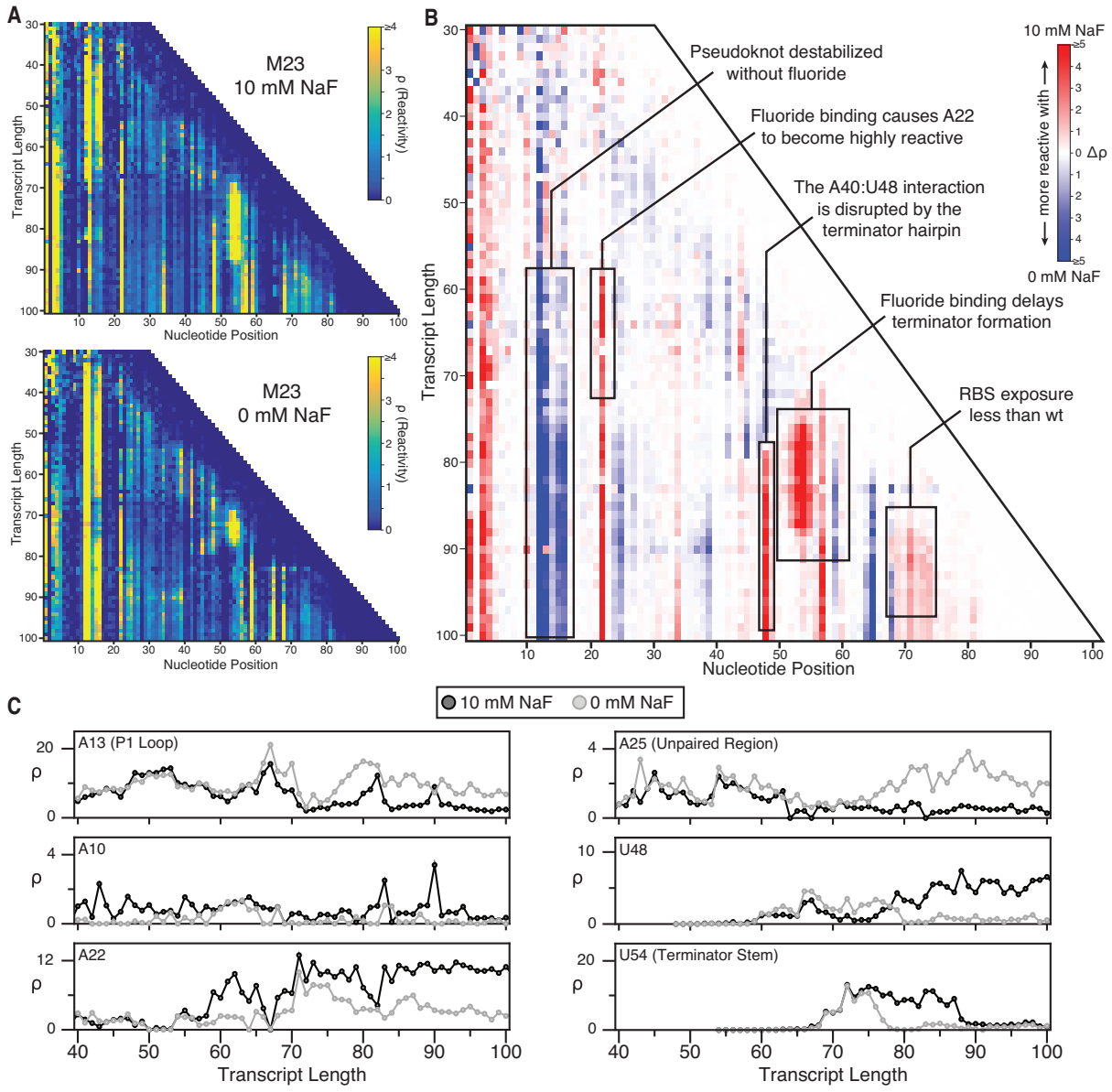

A
M22
10 mM NaF
M22
0 mM NaF
Transcript Length
Nucleotide Position
ρ (Reactivity)
B
No significant pseudoknot change observed
Minimal A22 reactivity observed with fluoride
No major reactivity changes in response to fluoride. Terminator forms in both conditions
10 mM NaF
more reactive with
Δρ
0 mM NaF
C
10 mM NaF
0 mM NaF
G13 (P1 Loop)
A10
A22
A25 (Unpaired Region)
U48
U54 (Terminator Stem)
ρ
Transcript Length

Figure C.14: Cotranscriptional SHAPE-Seq data for fluoride riboswitch mutant M23. (A) Reactivity matrices of the M23 mutant (compensatory mutations in the pseudoknot and terminator; Figure C.8) as described in Figure C.9A. (B) Reactivity differences ($\Delta\rho$) between the matrices in (A) as described in Figure C.9B. (C) Single nucleotide reactivity trajectories as described in Figure C.9C. The M23 mutant behaves similarly to the wt (Figures 5.3 and 5.4) as expected, since the M23 mutations produce a functional riboswitch (Figure C.8). Differences in the aptamer region in part (B) with either 0 mM or 10 mM NaF present match well to analogous differences for the wt (Figure 5.3), M20 (Figure C.11), and M21 (Figure C.12), supporting the conclusion that the M23 mutant can bind fluoride (Figure C.8). The trajectories in (C) and their similarity to the wt (Figure 5.4A) support the conclusion that M23 can prevent terminator formation like the wt. The replacement of a G:C base pair with an A:U base pair in PK1 lowers the relative decrease in G13 for the M23 mutant upon PK1 formation, suggesting that pseudoknot formation in M23 is weaker than the wt riboswitch, potentially leading to weaker antitermination (Figure C.8).


A
M23
10 mM NaF
M23
0 mM NaF
Transcript Length
Nucleotide Position
ρ (Reactivity)
B
Pseudoknot destabilized
without fluoride
Fluoride binding causes A22
to become highly reactive
The A40:U48 interaction
is disrupted by the
terminator hairpin
Fluoride binding delays
terminator formation
RBS exposure
less than wt
10 mM NaF
more reactive with
Δρ
0 mM NaF
Transcript Length
Nucleotide Position
C
10 mM NaF
0 mM NaF
A13 (P1 Loop)
A10
A22
A25 (Unpaired Region)
U48
U54 (Terminator Stem)
ρ
Transcript Length

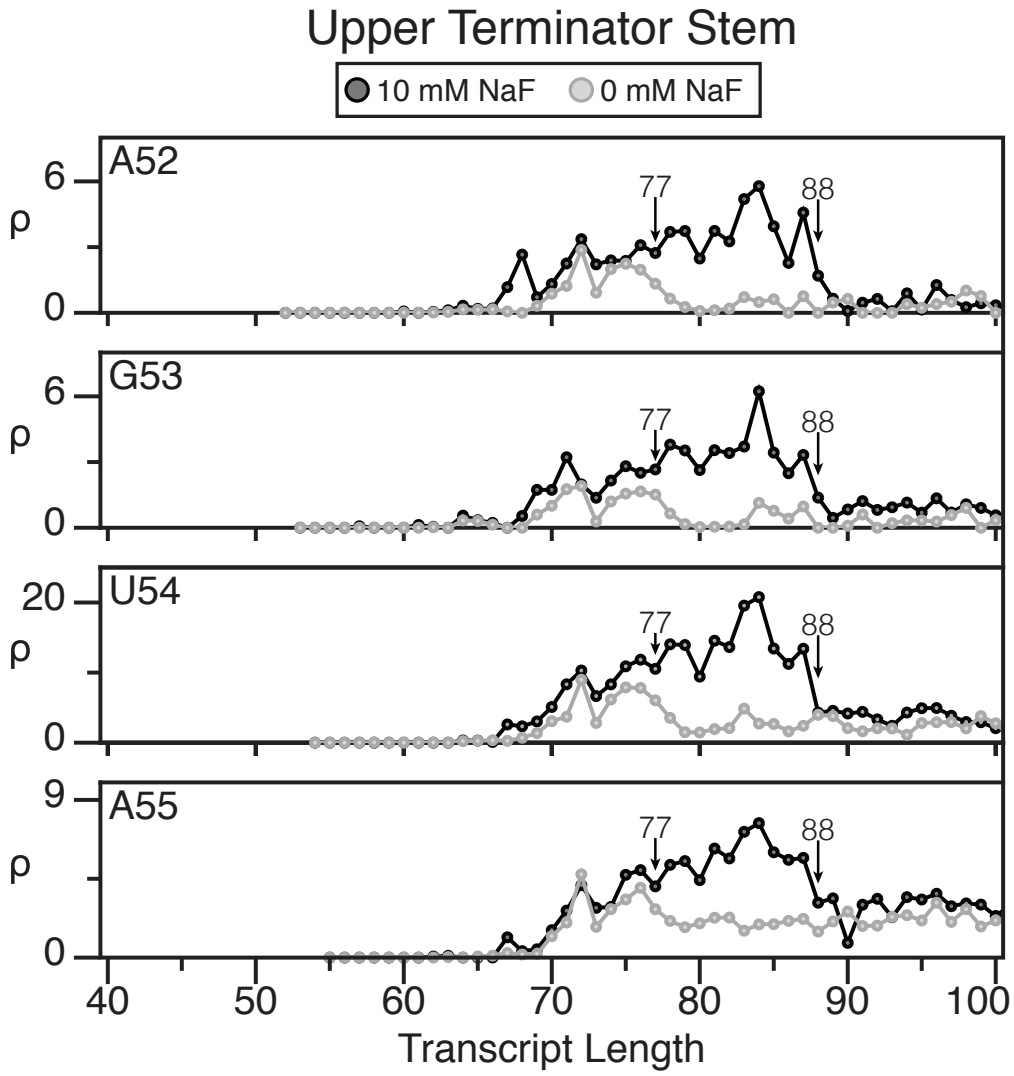


Figure C.15: Reactivity profiles for A52, G53, U54, and A55 of the wt fluoride riboswitch over the course of transcription. Single nucleotide trajectories displaying changes in the reactivities of nucleotides A52, G53, U54, and A55 of the wt fluoride riboswitch when transcribed with either 0 mM (gray) or 10 mM NaF (black). Data is taken from the cotranscriptional SHAPE-Seq matrices in Figure 5.3B. These nucleotides are found on the upper 5′ side of the terminator stem and exhibit higher reactivities after length 77 when fluoride is present, indicating that the terminator hairpin has not yet started to form at these lengths. When fluoride is present, these nucleotides only decrease in reactivity after length 88, when the entire terminator hairpin sequence has emerged from the polymerase and then stably folds. In the absence of fluoride, these positions decrease in reactivity and become weakly reactive by length 79, indicating that terminator formation and transcription of the polyU tract occur concurrently.

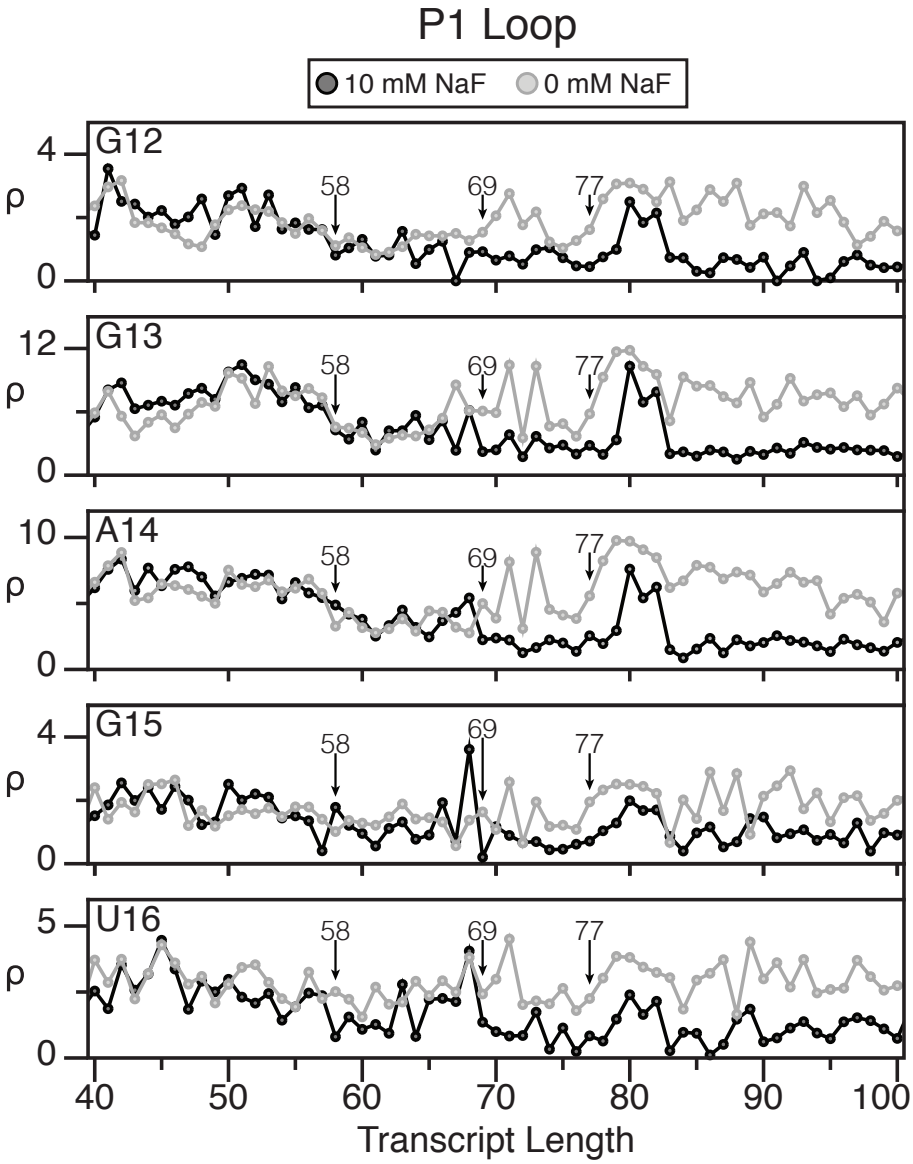
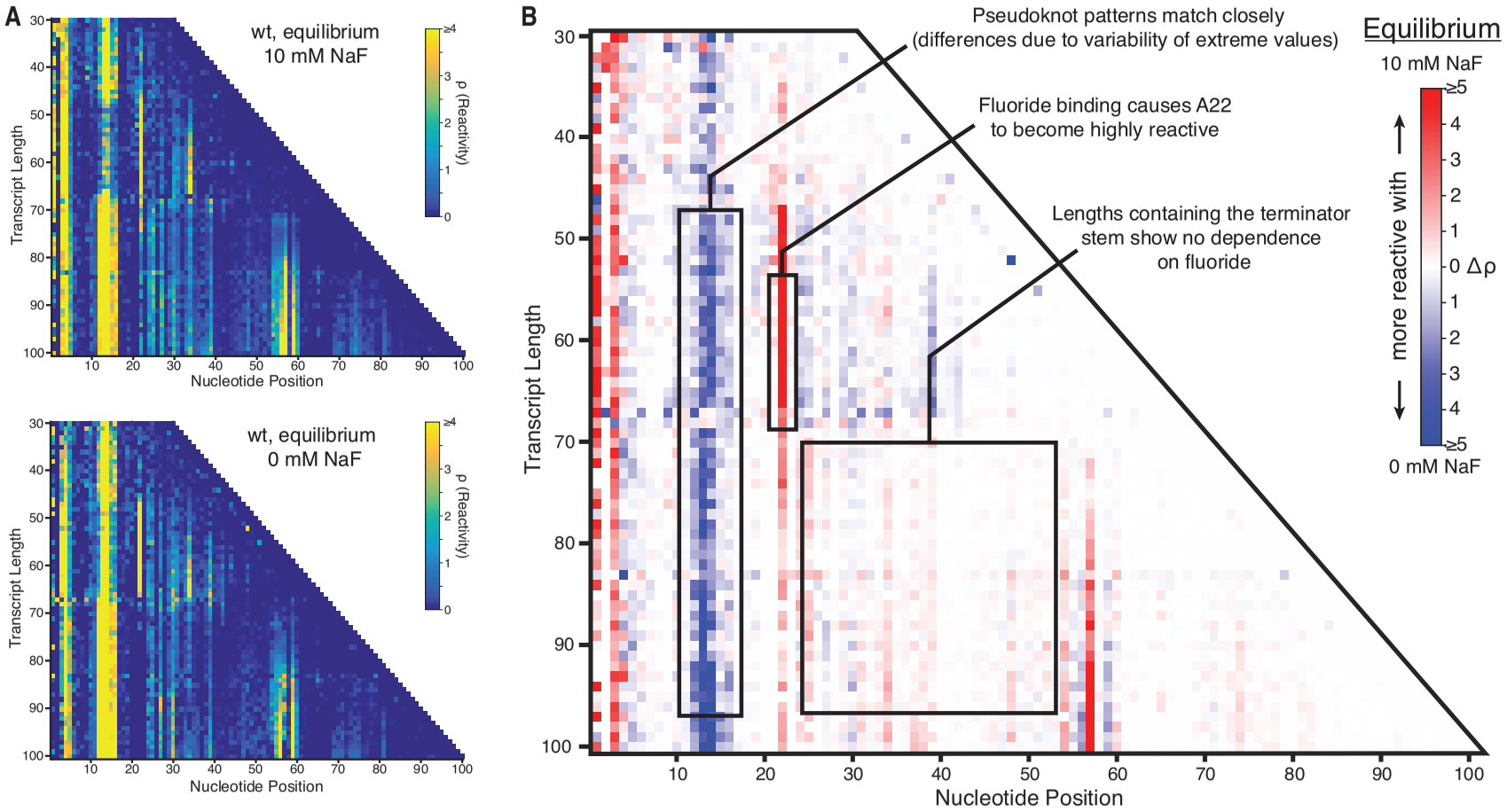


Figure C.16: Reactivity profiles for G12, G13, A14, G15, and U16 of the wt fluoride riboswitch over the course of transcription. Single nucleotide trajectories displaying changes in the reactivities of nucleotides G12, G13, A14, G15, and U16 of the wt fluoride riboswitch when transcribed with either 0 mM (gray) or 10 mM NaF (black). Data is taken from the cotranscriptional SHAPE-Seq matrices in Figure 5.3B. These nucleotides are found in the P1 loop and exhibit low reactivities after length 58 when PK1 of the aptamer domain forms. These reactivities stay low in the presence of fluoride, as the aptamer is stabilized by fluoride binding. However, if fluoride is absent, the growing terminator stem disrupts PK1 after length 77, causing an increase in reactivity.

Figure C.17: SHAPE-Seq data for wt fluoride riboswitch equilibrium refolded. (A) SHAPE-Seq reactivity matrix for all of the intermediate transcript lengths of the wt fluoride riboswitch that were first transcribed with 10 mM NaF, then extracted, denatured, and equilibrium refolded in transcription buffer containing either 10 mM (top) or 0 mM NaF before SHAPE modification. (B) Reactivity difference ($\Delta\rho$) matrix between the matrices in (A) as described in Figure 5.3C. Few differences are observed between the refolding conditions, except for generally higher reactivities in the P1 loop (nts 12-16) without ligand (although both are very high) and the appearance of high reactivity at position A22 across lengths 47-66. A sharp increase in reactivity is observed in the P1 loop at length 68 indicating that the terminated structure (PK1 open) is thermodynamically more stable in equilibrium, independent of the presence of fluoride. The differences observed between the cotranscriptional and equilibrium refolding matrices when fluoride is present provides strong evidence that the fluoride riboswitch can only function cotranscriptionally. Because the RNAP footprint is absent in equilibrium refolding, all of the transitions are observed ~14 nt earlier [246].


A
wt, equilibrium
10 mM NaF
wt, equilibrium
0 mM NaF
Transcript Length
Nucleotide Position
ρ (Reactivity)
≥4
B
Pseudoknot patterns match closely
(differences due to variability of extreme values)
Fluoride binding causes A22
to become highly reactive
Lengths containing the terminator
stem show no dependence
on fluoride
Equilibrium
10 mM NaF
more reactive with
Δρ
≥5
0 mM NaF

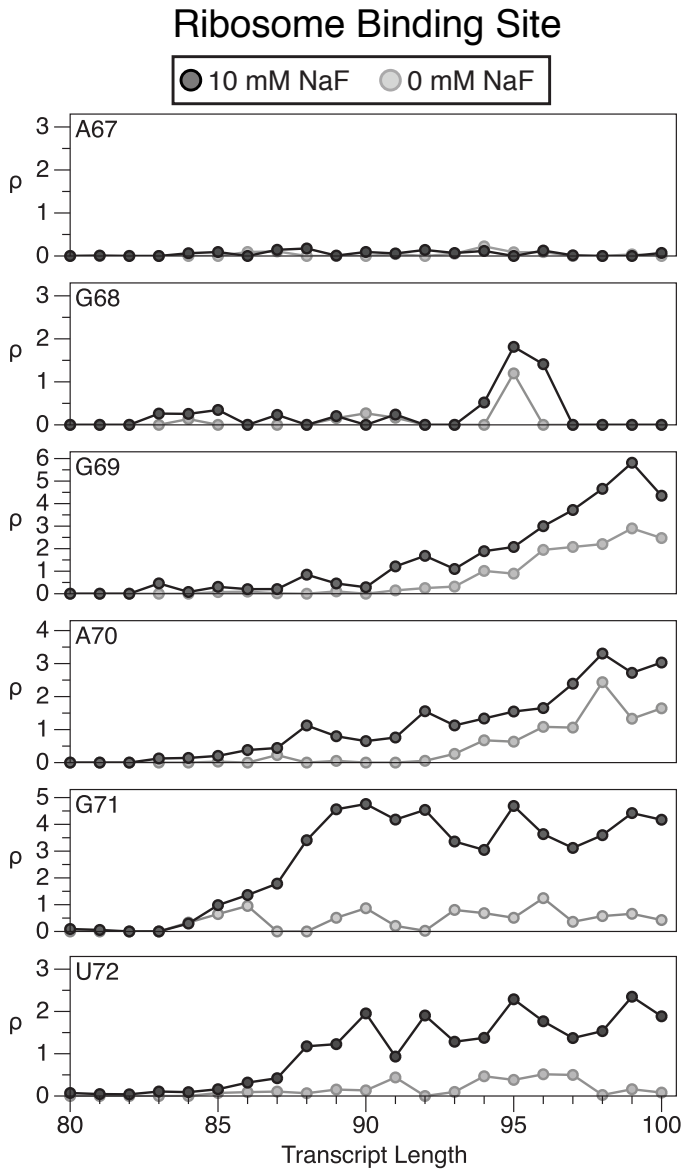


Figure C.18: Reactivity profiles for A67, G68, G69, A70, G71, and U72 of the wt fluoride riboswitch over the course of transcription. Single nucleotide trajectories displaying changes in the reactivities of nucleotides A67, G68, G69, A70, G71, and U72 of the wt fluoride riboswitch when transcribed with either 0 mM (gray) or 10 mM NaF (black). Data is taken from the cotranscriptional SHAPE-Seq matrices in Figure 5.3B. These nucleotides comprise the ribosome binding site (RBS), found in the 3' side of the terminator stem. Without fluoride, the terminator loop forms, simultaneously preventing further transcription and sequestering the RBS. With fluoride however, the lower terminator stem does not form, causing many of the RBS nts to be reactive, indicating that they are free to interact with a ribosome to initiate translation.

C.4 Suppmentary Tables

Table C.1: Sequences used for *in vitro* transcription templates. Summary of the sequences used for generating *in vitro* transcription templates with *E. coli* polymerase. For sequences that were studied by generating libraries of DNA intermediate length templates, only the longest sequence is shown. To extend beyond the length of the RNA being studied, excess sequence was added to allow the entire RNA to leave the polymerase and provide adequate space for protein footprints. For transcription of the Signal Recognition Particle (SRP) RNA (sequence used in Wong *et al.* [83], templates were extended with part of the antigenomic hepatitis δ ribozyme, while a ribosome binding site (RBS) and superfolder GFP (SFGFP) sequence was used for the fluoride riboswitch sequences. Fluoride riboswitch 'M' mutants are taken from Baker *et al.* [50]. An intermediate template length as used for cotranscriptional SHAPE-Seq can be assembled by concatenating, in order, the proper promoter/leader + the RNA length + the EcoRI binding site.

Description	**Sequence**
Promoter and leader region (for SRP)	ATAAGCTTCCGATGGCGCGCCGAGAGGCTTTACACTT TATGCTTCCGGCTTGATTCTAAAGATCTTTGACAGCT AGCTCAGTCCTAGGTATAATGAATTC
Promoter and leader region (for fluoride riboswitches)	ATAAGCTTCCGATGGCGCGCCGAGAGGCTTTACACTT TATGCTTCCGGCTTGATTCTAAAGATCTTTGACAGCT AGCTCAGTCCTAGGTATAATACTAGT
Signal Recognition Particle RNA, *E. coli*	ATCGGGGGCTCTGTTGGTTCTCCCGCAACGCTACTCT GTTTACCAGGTCAGGTCCGGAAGGAAGCAGCCAAGG CAGATGACGCGtGtGCCGGGATGTAGCTGGCAGGGCC CCCACCC
Antigenomic hepatitis δ ribozyme	GGGTCGGCATGGCATCTCCACCTCCTCGCGGTCCGAC CTGGGCATCC
EcoRI binding site	GAATTCaaaaaaa
Fluoride riboswitch, *B. cereus*, wt	TTATAGGCGATGGAGTTCGCCATAAACGCTGCTTAGC TAATGACTCCTACCAGTATCACTACTGGTAGGAGTCT ATTTTTTT
RBS and SFGFP	AGGAGGAAGGATCTATGAGCAAAGGAGAAGAACTT TTCACTGGAGTTGTC
Fluoride riboswitch, *B. cereus*, M18 mut	TTATAGGCGATGGTTCGCCATAAACGCTGCTTAGCTA ATGACTCCTACCAGTATCACTACTGGTAGGAGTCTAT TTTTTT
Fluoride riboswitch, *B. cereus*, M19 mut	TTATAGGCGATGGAGTTCGCCATAAACGCTGCTTAGC TAATGACATCTACCAGTATCACTACTGGTAGGAGTCT ATTTTTTT

Table C.1 (Continued)

Description	**Sequence**
Fluoride riboswitch, *B. cereus*, M20 mut	TTATAGGCGATGGAGTTCGCCATAAACGCTGCTTAGC TAATGACTCCTACCAGTATCACTACTGGTAGATGTCT ATTTTTTT
Fluoride riboswitch, *B. cereus*, M21 mut	TTATAGGCGATGATGTTCGCCATAAACGCTGCTTAGC TAATGACATCTACCAGTATCACTACTGGTAGGAGTCT ATTTTTTT
Fluoride riboswitch, *B. cereus*, M22 mut	TTATAGGCGATGGAGTTCGCCATAAACGCTGCTTAGC TAATGACATCTACCAGTATCACTACTGGTAGATGTCT ATTTTTTT
Fluoride riboswitch, *B. cereus*, M23 mut	TTATAGGCGATGATGTTCGCCATAAACGCTGCTTAGC TAATGACATCTACCAGTATCACTACTGGTAGATGTCT ATTTTTTT
Example template for fluoride riboswitch, wt, 50 nt	ATAAGCTTCCGATGGCGCGCCGAGAGGCTTTACACTT TATGCTTCCGGCTTGATTCTAAAGATCTTTGACAGCT AGCTCAGTCCTAGGTATAATACTAGTTTATAGGCGAT GGAGTTCGCCATAAACGCTGCTTAGCTAATGACTCCT ACGAATTCaaaaaaa

Table C.2: Oligonucleotides used in this study. Below is a table of oligonucleotides used during the cotranscriptional SHAPE-Seq experiments. The 'xxxxxx' sequence in the Illumina primer represents any TruSeq index. For the reverse template amplification primers, the sequence 'NNNNNN' is the reverse complement from the 3' end of the intermediate length being amplified. This table is meant to serve as a reference for Appendix C.1. Abbreviations within primer sequences are as follows: '/5Biosg/' is a 5' biotin moiety, '/5Phos/' is a 5' monophosphate group, '/3SpC3/' is a 3' 3-carbon spacer group, VIC and NED are fluorophores (ABI), and asterisks indicate a phosphorothioate backbone modification. These abbreviations were used for compatibility with the Integrated DNA Technologies ordering notation.

Description	Sequence	
RNA linker	/5Phos/CUGACUCGGGCACCAAGGA/3ddC/	A
RT primer	/5Biosg/GTCCTTGGTGCCCGAGT	B
DNA adapter	/5Phos/AGATCGGAAGAGCACACGTCTGAACTCCAG TCAC/3SpC3/	C
QA primer (+)	VIC-GTGACTGGAGTTCAGACGTGTGCTC	D
QA primer (-)	NED-GTGACTGGAGTTCAGACGTGTGCTC	E
PE_F†	AATGATACGGCGACCACCGAGATCTACACTCTTTCC CTACACGACGCTCTTCCGATCT	F
Selection primer (+)	CTTTCCCTACACGACGCTCTTCCGATCTRRRYGCATC CACAATAGAAGAAGGATGC*C*G*C*A	G
Selection primer (-)	CTTTCCCTACACGACGCTCTTCCGATCTYYYRGCATC CACAATAGAAGAAGGATGC*C*G*C*A	H
Illumina reverse primers (TruSeq)†	CAAGCAGAAGACGGCATACGAGATxxxxxxGTGACTG GAGTTCAGACGTGTGCTC	I
Forward Template amplification primer	/5Biosg/ATAAGCTTCCGATGGCGCGC	J
Reverse Template primer for single length fluoride riboswitch templates	/5Biosg/CAACAAGAATTGGGACAACTCCAGTG	K
Reverse Template amplification primers	TTTTTTTGAATTCNNNNNNNNNNNNNN	

Table C.3: RMDB data deposition table. SHAPE-Seq reactivity spectra generated in this work are freely available from the RNA Mapping Database (RMDB) (`http://rmdb.stanford.edu/repository/`), accessible using the RMDB ID numbers indicated in the table below. Any other reactivity data from this work can be provided upon request.

RMDB ID	**RNA**	**Experiment**	**Figure(s)**
FLUORSW_1M7_0001	F- riboswitch, wt	no ligand, cotranscriptional	Figures 5.2, 5.3, C.4 to C.7, C.15, C.16 and C.18
FLUORSW_1M7_0002	F- riboswitch, wt	10 mM ligand, cotranscriptional	Figures 5.2, 5.3, C.4 to C.7, C.15, C.16 and C.18
FLUORSW_1M7_0005	F- riboswitch, wt	transcribed with 10 mM fluoride, extracted, and equilibrium refolded without ligand	Figures 5.4 and C.17
FLUORSW_1M7_0006	F- riboswitch, wt	transcribed with 10 mM fluoride, extracted, and equilibrium refolded with 10 mM fluoride	Figures 5.4 and C.17
FLUORSW_1M7_0007	F- riboswitch, M18	no ligand, cotranscriptional	Figure C.9
FLUORSW_1M7_0008	F- riboswitch, M18	10 mM ligand, cotranscriptional	Figure C.9
FLUORSW_1M7_0009	F- riboswitch, M19	no ligand, cotranscriptional	Figure C.10
FLUORSW_1M7_0010	F- riboswitch, M19	10 mM ligand, cotranscriptional	Figure C.10
FLUORSW_1M7_0011	F- riboswitch, M20	no ligand, cotranscriptional	Figure C.11
FLUORSW_1M7_0012	F- riboswitch, M20	10 mM ligand, cotranscriptional	Figure C.11
FLUORSW_1M7_0013	F- riboswitch, M21	no ligand, cotranscriptional	Figure C.12
FLUORSW_1M7_0014	F- riboswitch, M21	10 mM ligand, cotranscriptional	Figure C.12

Table C.3 (Continued)

RMDB ID	**RNA**	**Experiment**	**Figure(s)**
FLUORSW_1M7_0015	F- riboswitch, M22	no ligand, cotranscriptional	Figure C.13
FLUORSW_1M7_0016	F- riboswitch, M22	10 mM ligand, cotranscriptional	Figure C.13
FLUORSW_1M7_0017	F- riboswitch, M23	no ligand, cotranscriptional	Figure C.14
FLUORSW_1M7_0018	F- riboswitch, M23	10 mM ligand, cotranscriptional	Figure C.14
SRPECLI_1M7_0001	SRP RNA	cotranscriptional	Figures 5.2 and C.1
SRPECLI_1M7_0002	SRP RNA	equilibrium refolded after transcription	Figure C.2

APPENDIX D

SUPPLEMENTARY INFORMATION FOR MEASURING THE COTRANSCRIPTIONAL FOLDING PATHWAY OF THE PT181 TRANSCRIPTIONAL ATTENUATOR

D.1 Supplementary Figures

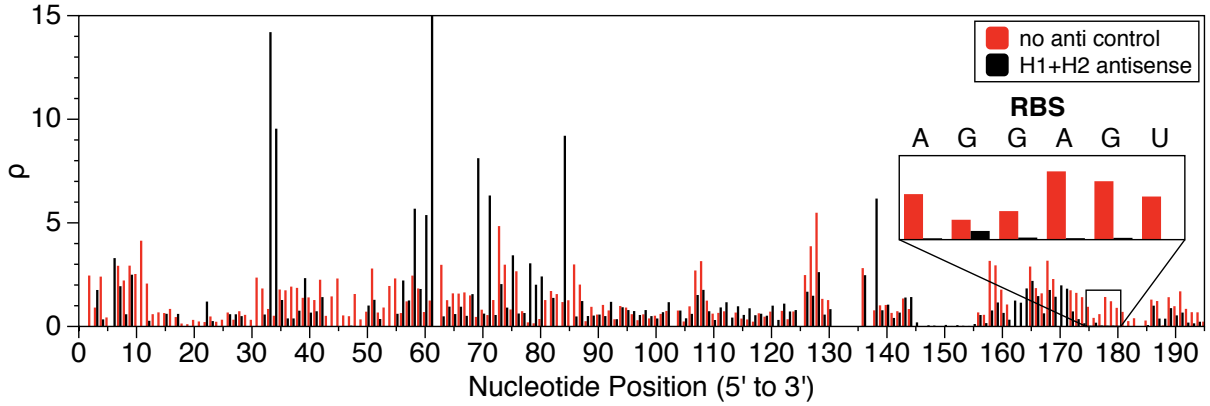


Figure D.1: In-cell SHAPE-Seq of the pT181 attenuator with and without antisense. A polyU mutant pT181 (to lower termination efficiency) was measured using in-cell SHAPE-Seq with (black) and without (red) the H1+H2 antisense present. A number of similarities exist, however, there are a number of highly reactive peaks appearing with the antisense and nts 155-195 in the terminator region match well to large terminator with antisense or an extended single-stranded region with an accessible RBS without antisense.

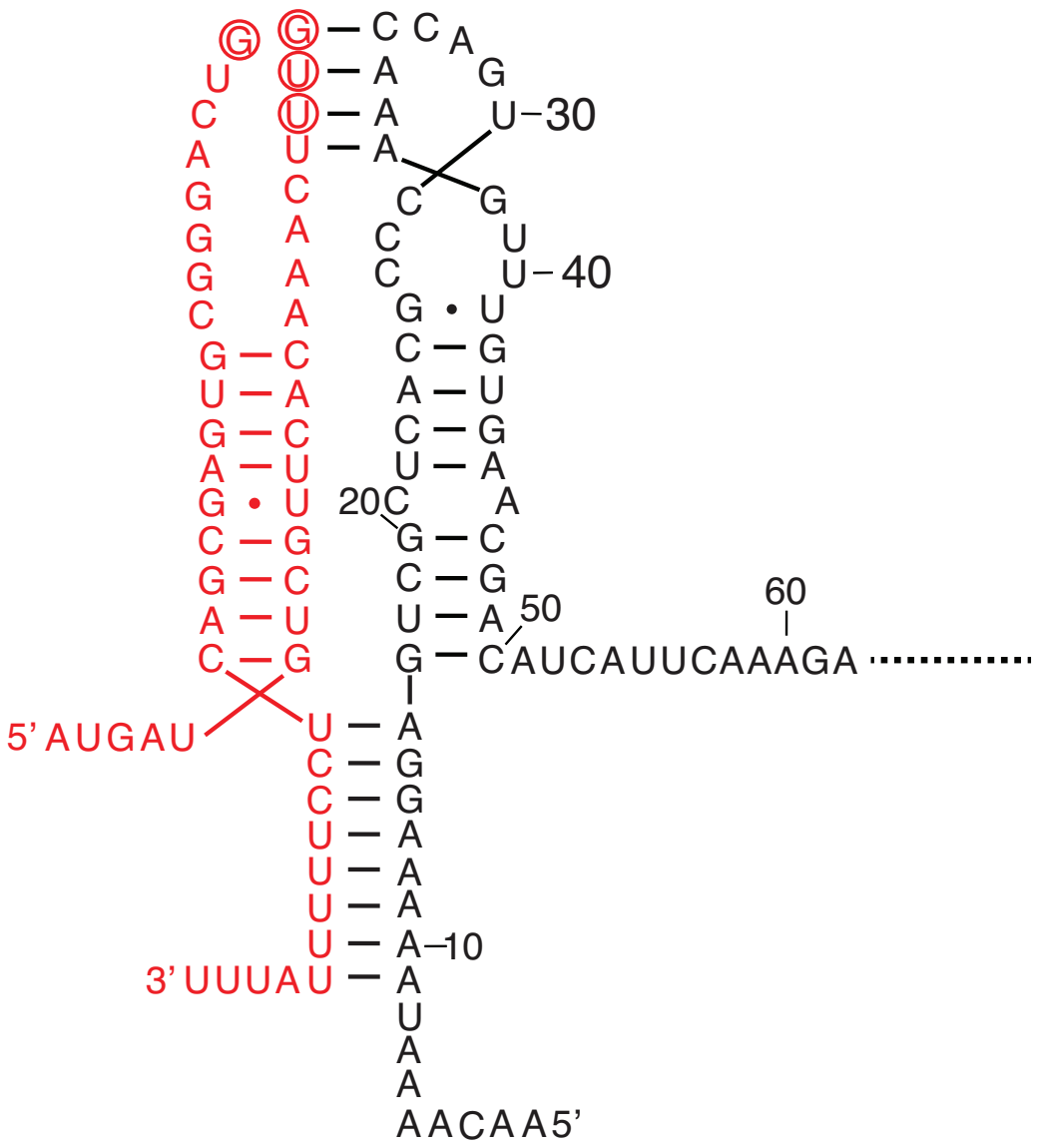


Figure D.2: Potential sense-antisense interaction during equilibrium refolding. In an equilibrium refolding experiment where the pre-folded H2 antisense was added after sense folding, the sense H1 reactivities exhibit a different pattern from those observed when cotranscriptionally folded with H2 present (Figures 6.2 and 6.4). Nucleotides 28-40 contain a number of highly reactive bases that are reminiscent of the no-antisense cotranscriptional fold of the sense RNA, suggesting that the presence of the antisense serves to stabilize the sense H1 hairpin, but with minimal contacts. The proposed structure contains mainly interactions between the single-stranded regions of the sense and antisense and include pairing between the YUNR motif of the antisense and the complementary bases in the sense loop, where the antisense is expected to dock [188].

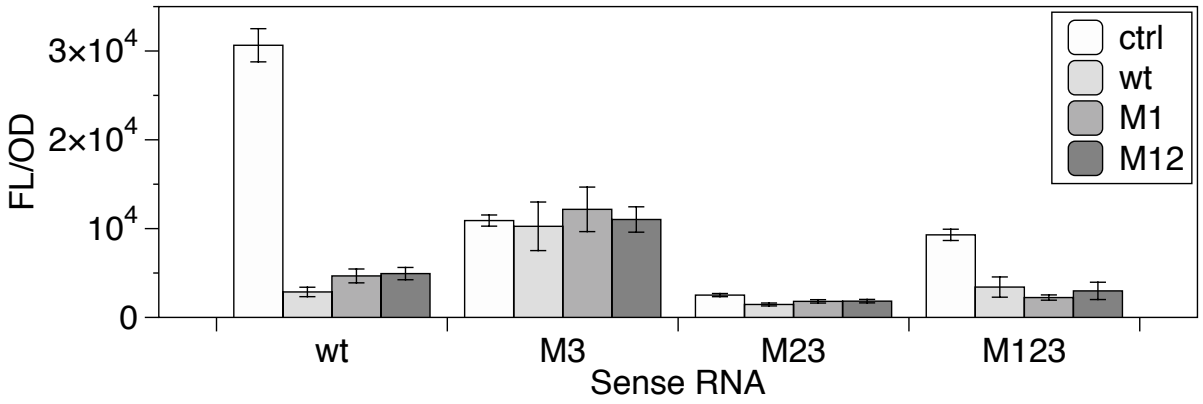


Figure D.3: Removing the RBS from the pT181 attenuator. A set of three mutations (M1, M2, and M3; see Figure 6.1) were made to swap the GGA (nts 176-178) sequence in the attenuator RBS to prevent translation initiation from the pT181 attenuator. Multiple mutations are indicated by M1+M2=M12, etc. The M3 mutant remains on due to a mismatch in the terminator hairpin, while M23 remains off due to correcting the M3 mismatch, but creating a mismatch in the antiterminator. Normal function is restored by mutating the equivalent positions in H1. Error bars represent one standard deviation of 12 replicates.

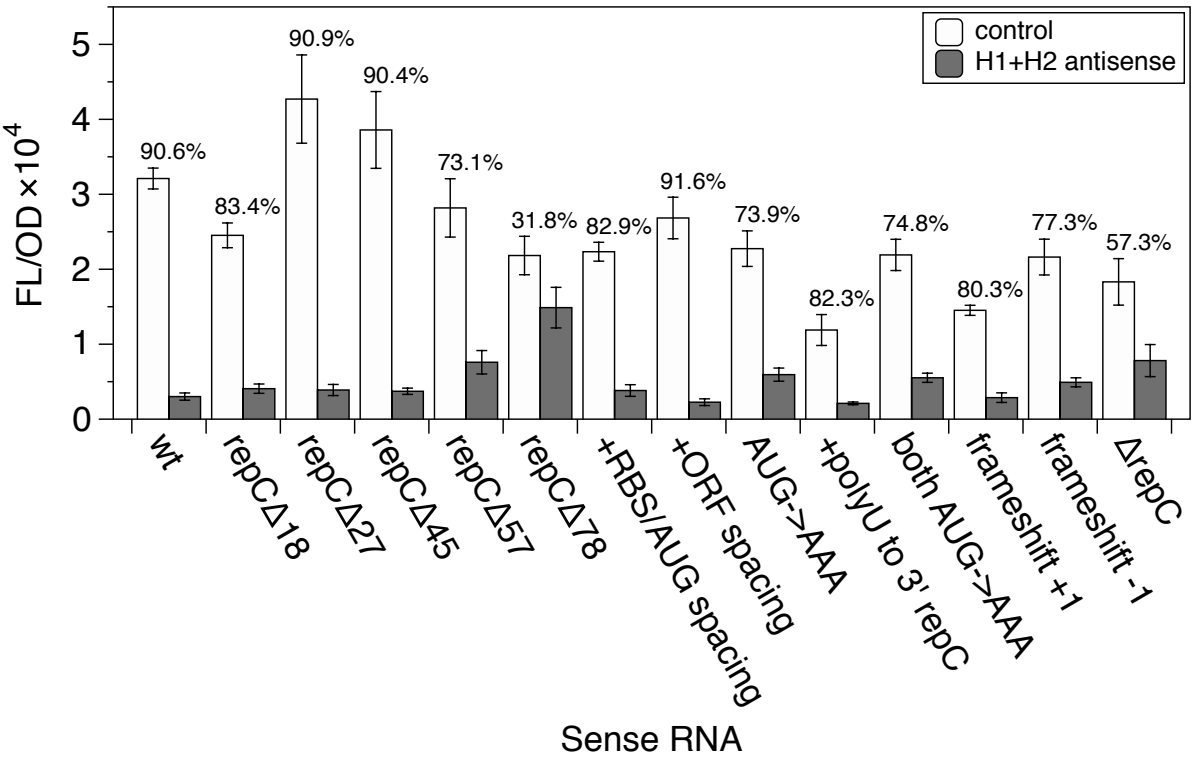


Figure D.4: Testing the impact of the RepC mRNA on gene expression. A series of deletions, mutations, and insertions were made to the RepC open reading frame as indicated to try to determine the effect of the repC sequence on the attenuator function. Spurious results suggest that the results observed are due to changing structural contexts within the RBS of the transcriptionally fused SFGFP sequence. The entire RepC sequence was later removed in the minimized version. Error bars represent one standard deviation of 12 replicates.

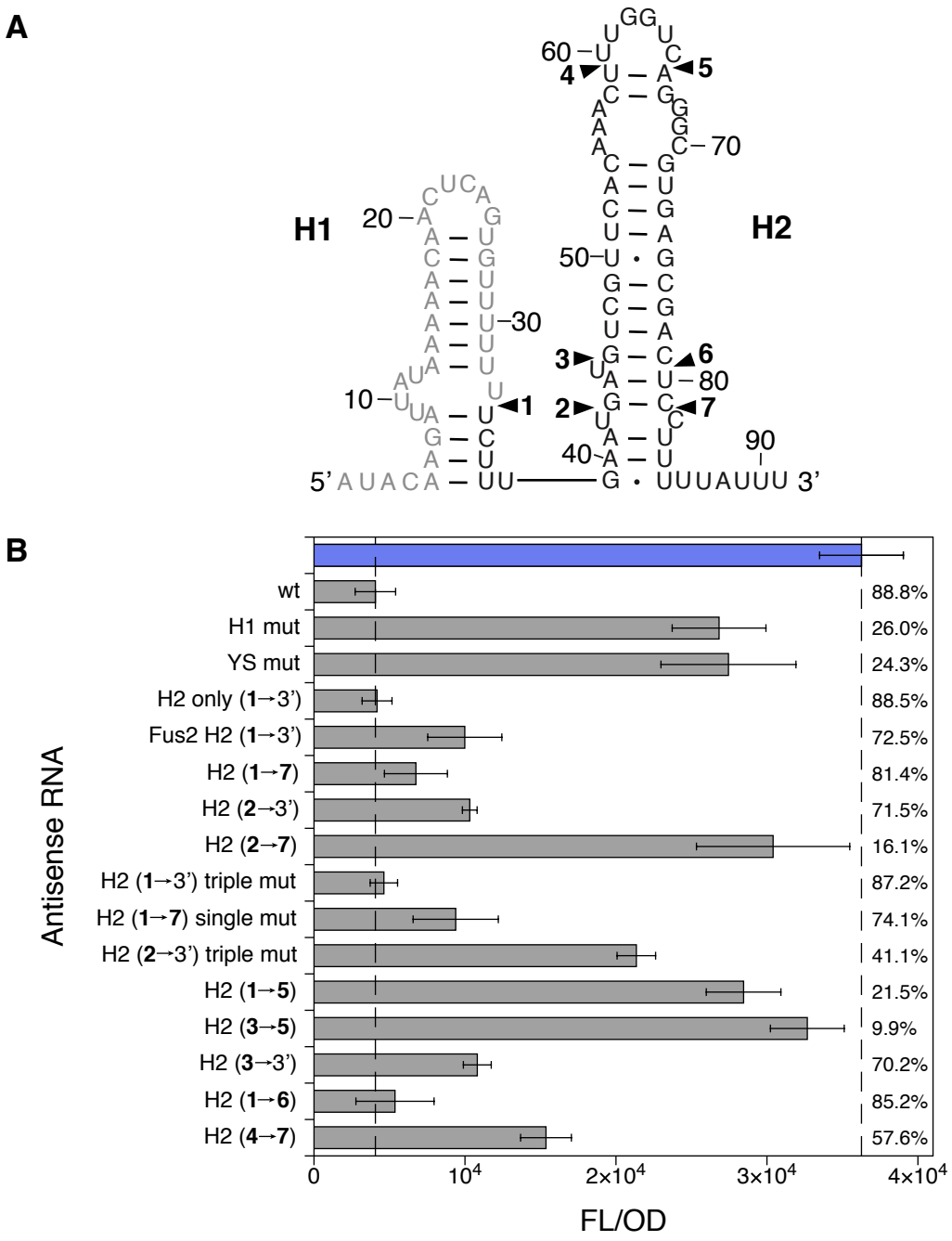


Figure D.5: Determining the minimal antisense length. Different antisenses were tested for their ability to switch the pT181 attenuator into the terminator form. As shown, the antisense H2 (H2 only) repressed transcription of the wt pT181 sequence as well as the H1+H2 antisense. Further small deletions from the 3' end of the antisense was tolerated, all other deletions resulted in a loss of termination efficiency. Error bars represent one standard deviation of 12 replicates.

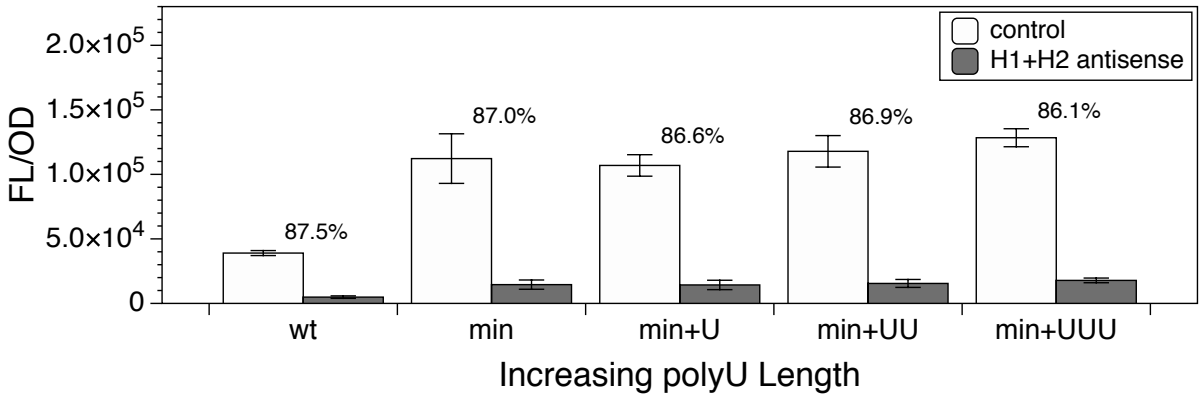


Figure D.6: Increased polyU length does not improve termination. Increasing the polyU tail length on the minimized pT181 attenuator (min) does not show any significant termination improvements. Error bars represent one standard deviation of 12 replicates.

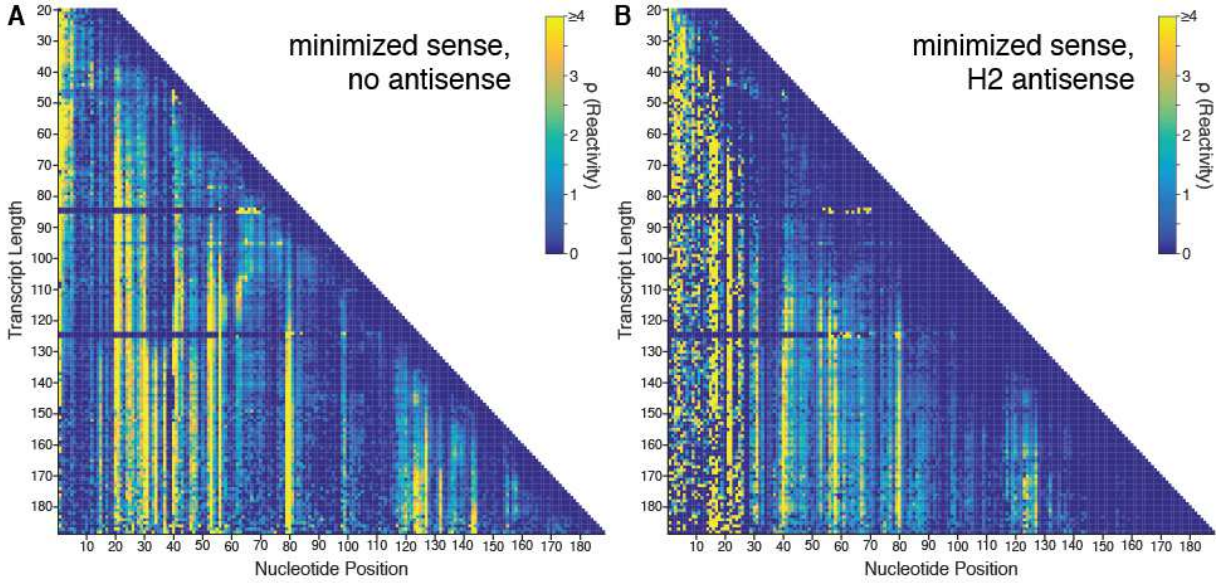


Figure D.7: Cotranscriptional SHAPE-Seq of the minimized pT181 attenuator. Cotranscriptional SHPAE-Seq data for the minimized pT181 transcribed without (A) or with (B) antisense H2. The results observed match well to the wt pT181 matrices when comparing equivalent nucleotides (Figure 6.2).

APPENDIX E

SUPPLEMENTARY INFORMATION FOR DESIGN OF A CRISPR SGRNA DEREPRESSOR USING INSIGHT FROM SHAPE-SEQ

E.1 Supplementary Figures

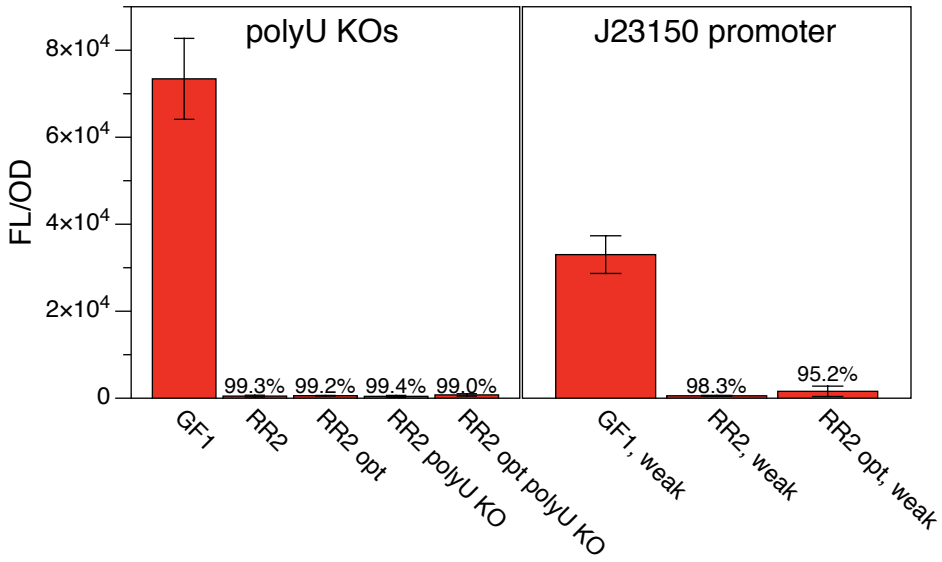
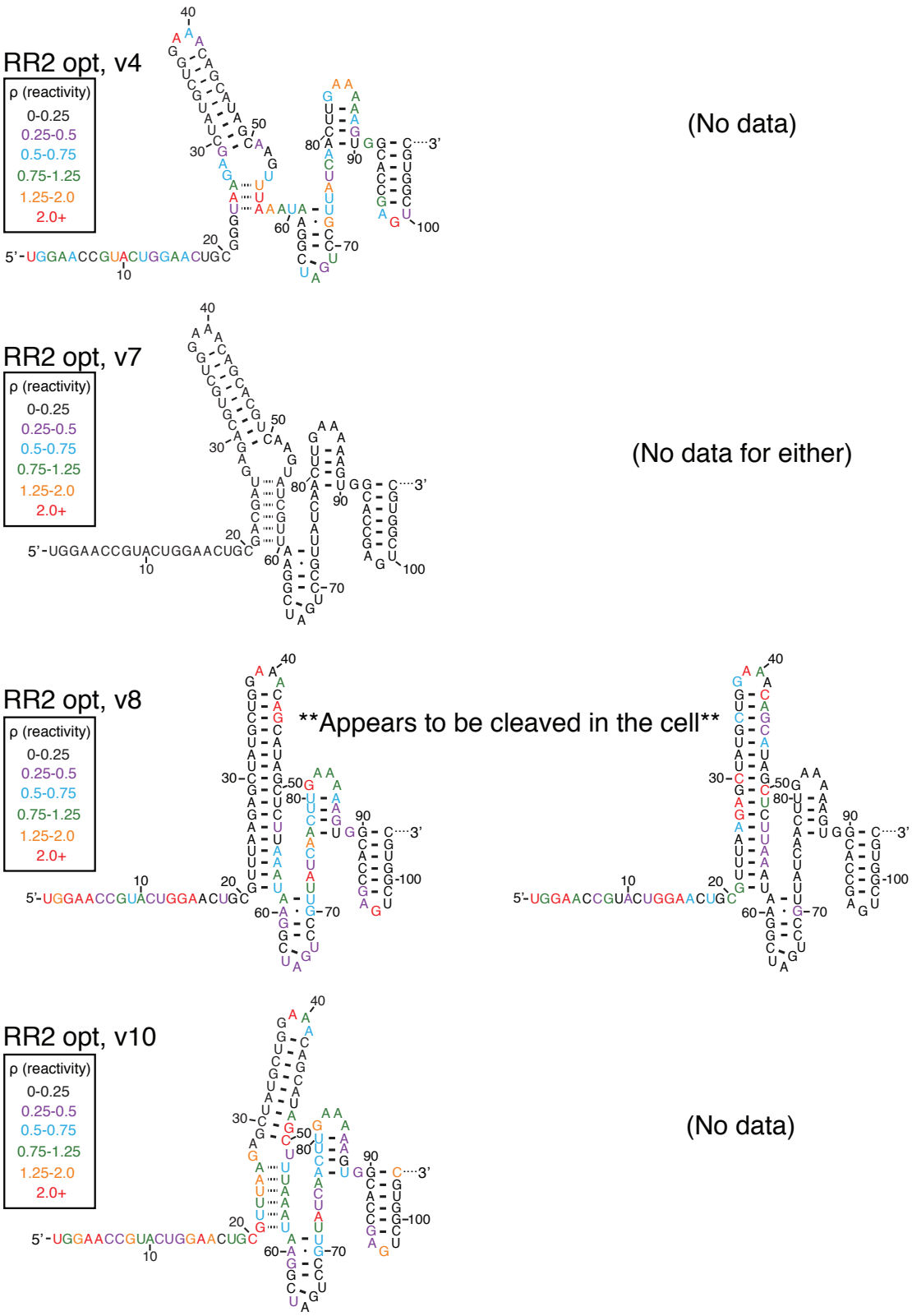


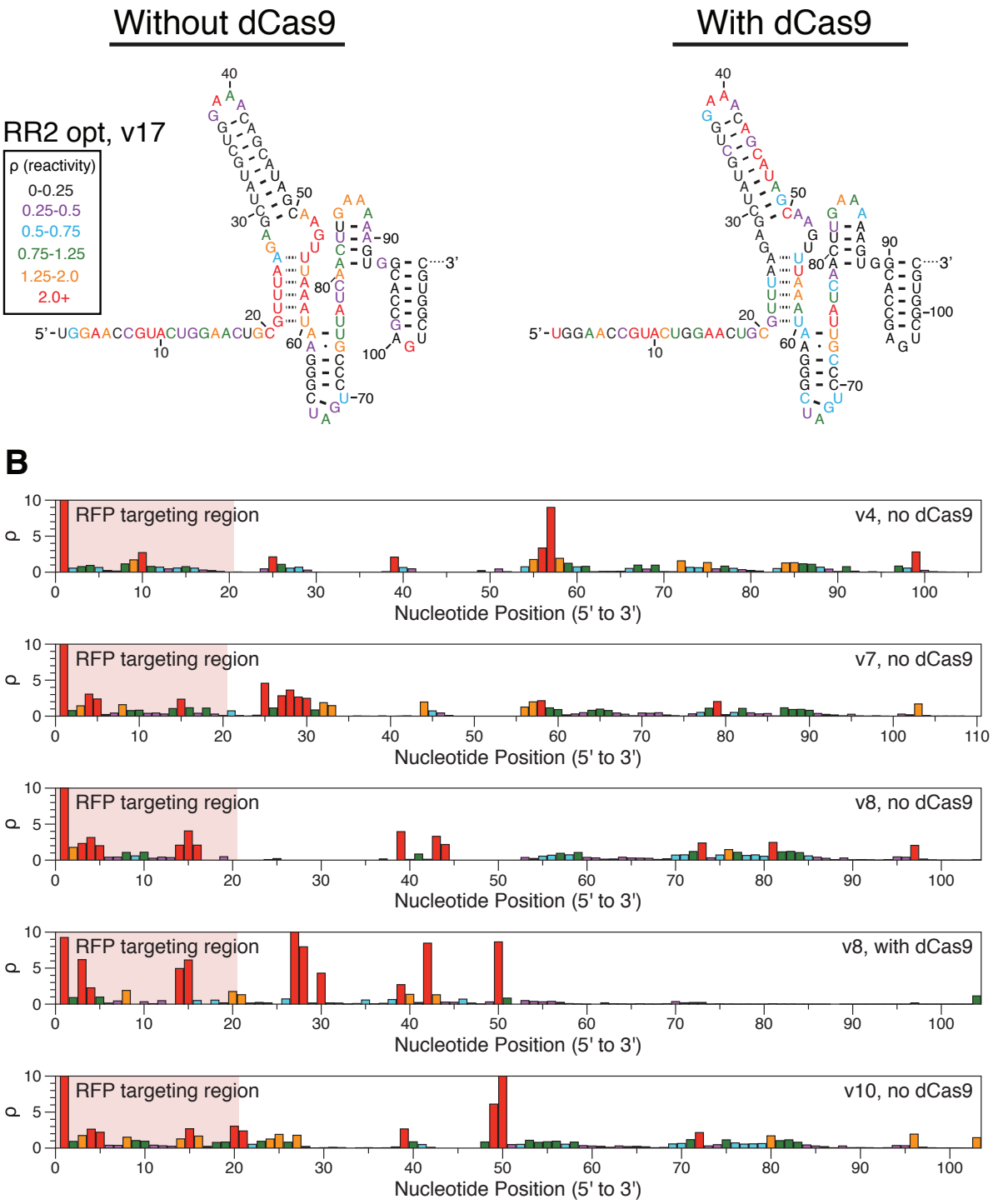
Figure E.1: Moving the sgRNA to the in-SHAPE-Seq platform. In order to make the sgRNAs amenable to SHAPE-Seq, we needed to remove the natural polyU sequences after the *S. pyogenes* terminator and replace them with the double terminator described in Watters *et al.* (Chapter 6) [113]. Changing the terminators did not affect RFP repression (left). We also tested a weakened constitutive promoter by replacing the original J23119 promoter (BioBricks) with J23150 and showed that repression was not strongly affected. Error bars represent one standard deviation of three replicates.

Figure E.2: (A) Secondary structures overlaid with in-cell SHAPE-Seq data for all 'v' series mutants. Each mutant tested is shown with nucleotides color-coded by reactivity intensity if in-cell SHAPE-Seq data was collected. The placeholder 'No data' indicates no SHAPE-Seq data was collected for that mutant condition (with or without dCas9). Mutant sgRNA v8 appeared to be cleaved in the cell by a dsRNase according to the sequencing read alignments. (B) Individual reactivity maps for 'v' series mutants that had data collected. The reactivity color scale is the same as in (A)

A


Without dCas9
With dCas9
RR2 opt, v4
ρ (reactivity)
0-0.25
0.25-0.5
0.5-0.75
0.75-1.25
1.25-2.0
2.0+
5'-UGGAACCGUACUGGAACUGC
(No data)
RR2 opt, v7
(No data for either)
RR2 opt, v8
Appears to be cleaved in the cell
RR2 opt, v10
(No data)

(Figure E.2 cont.)



(Figure E.2 cont.)

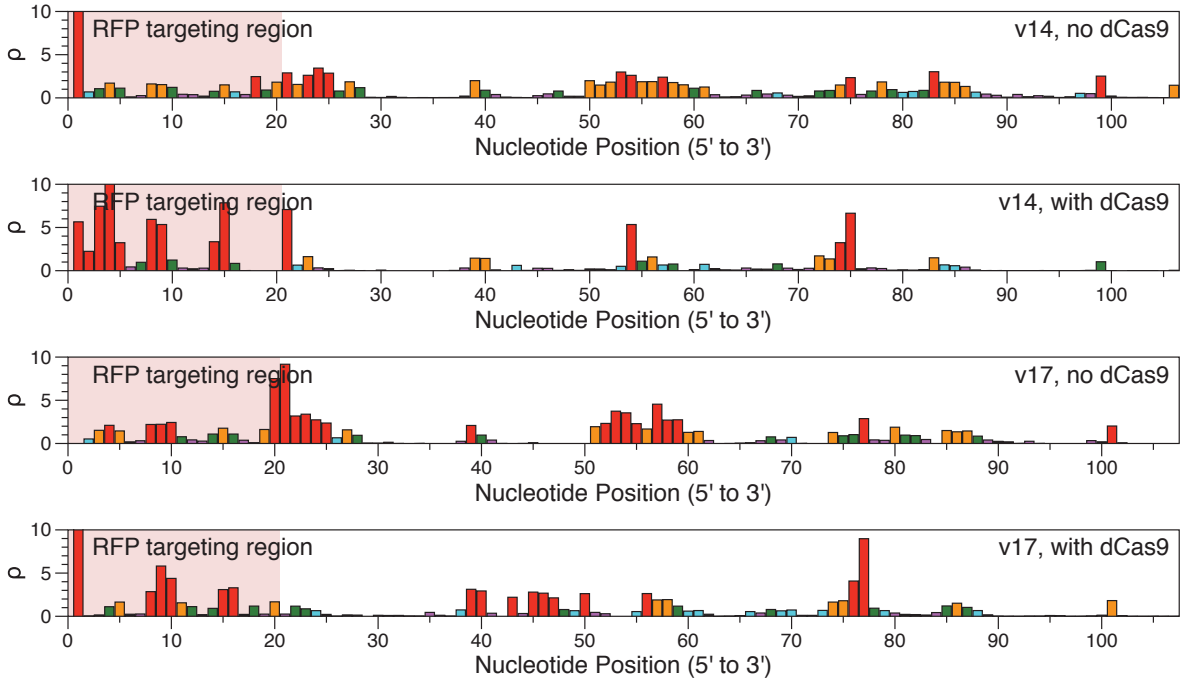
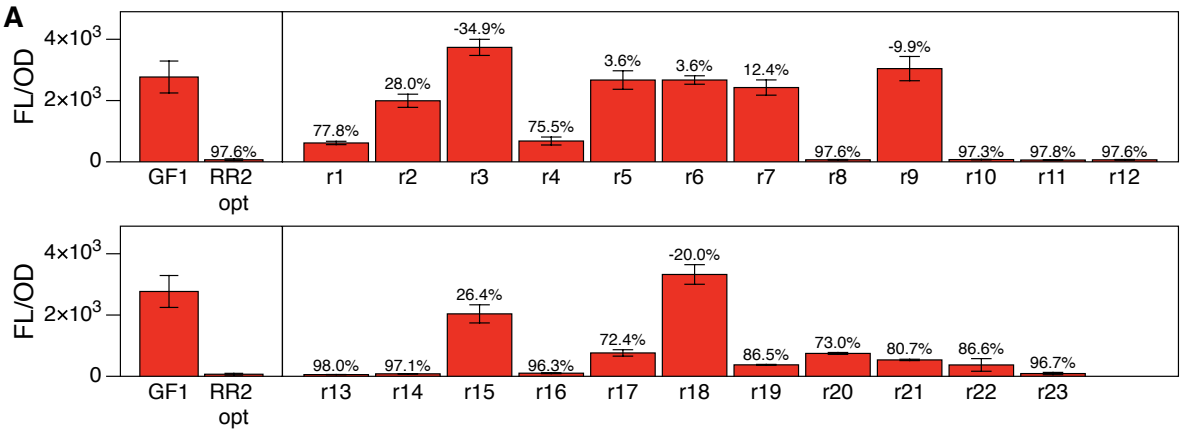
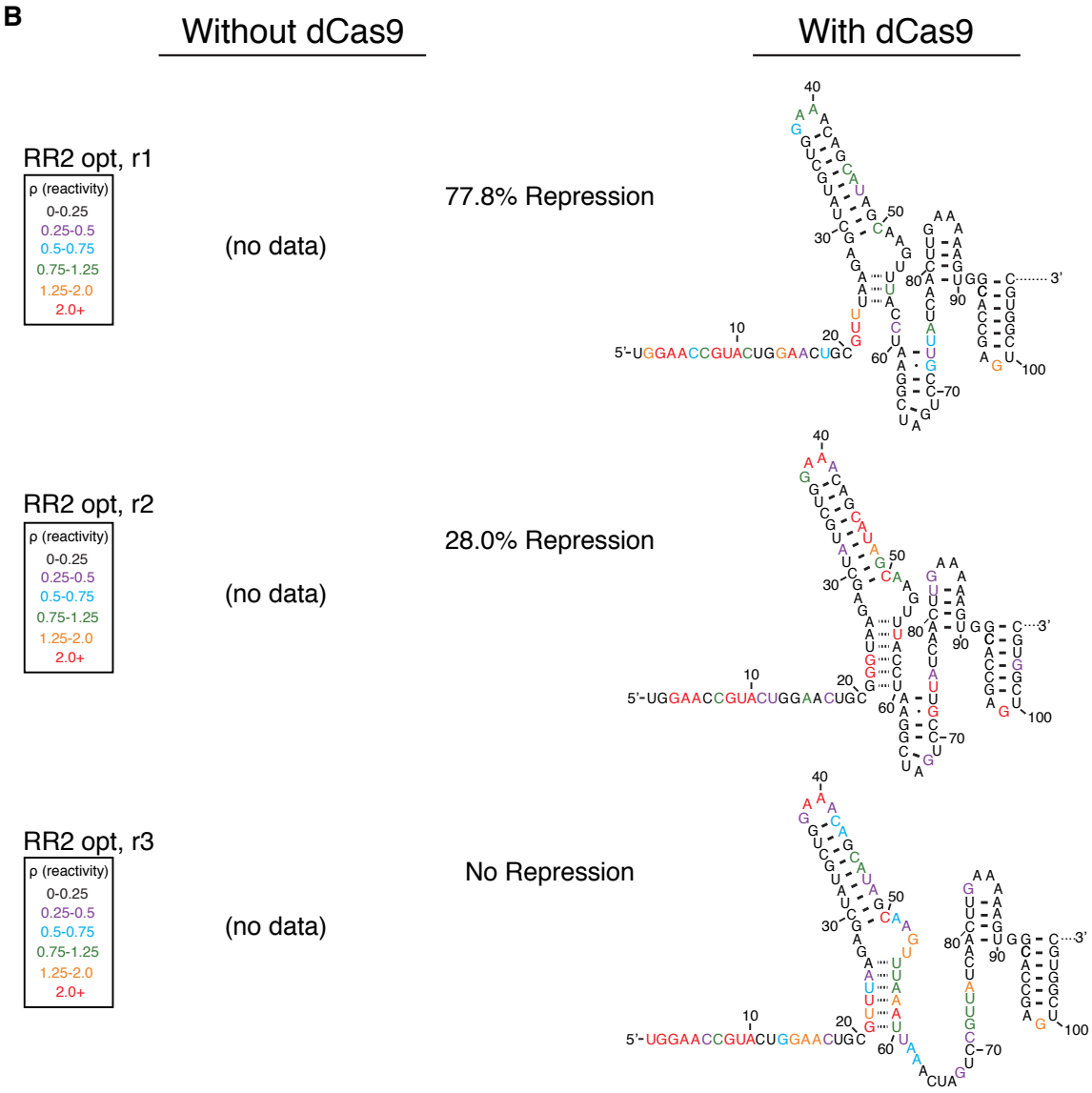
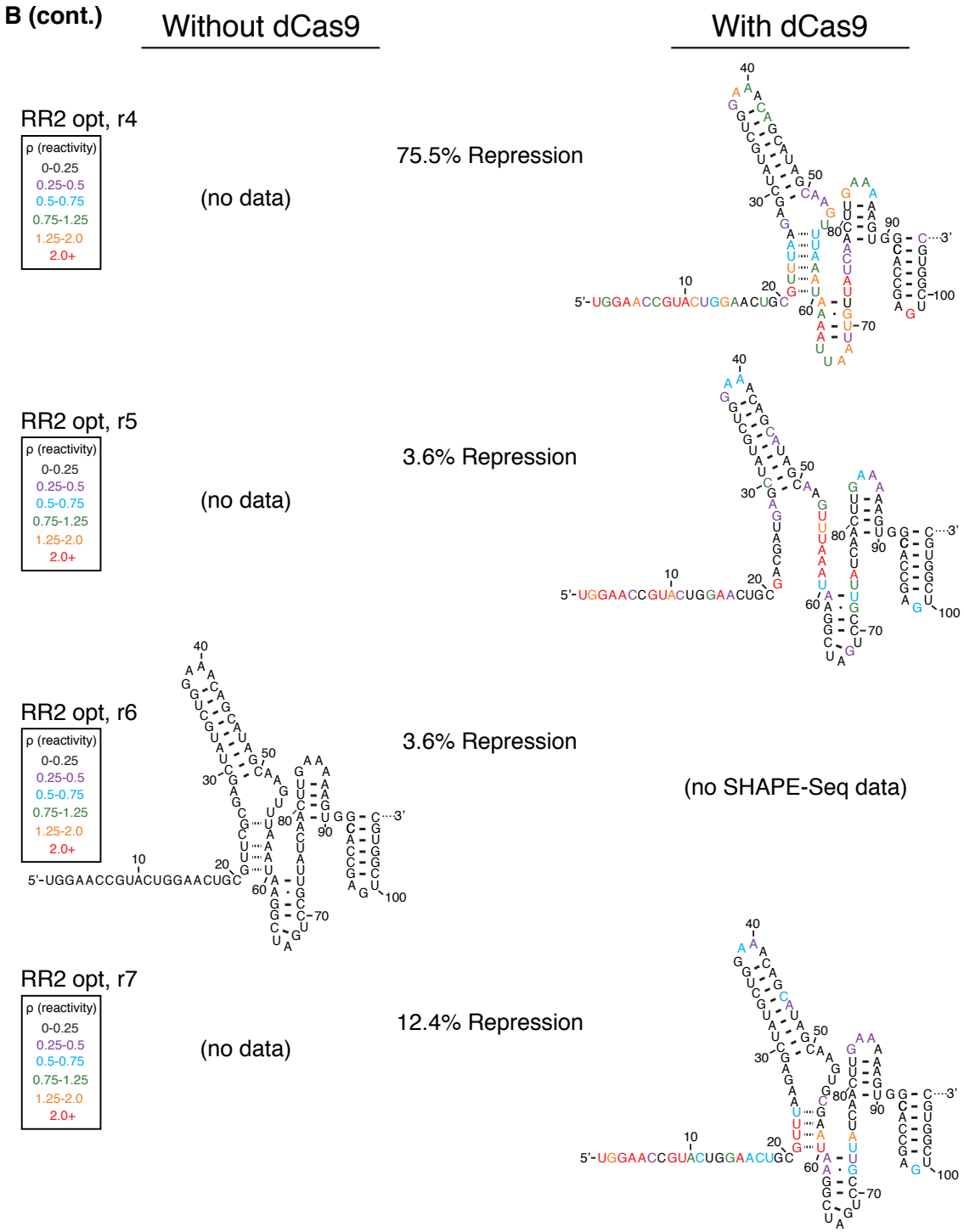


Figure E.3: Secondary structures and functional and in-cell SHAPE-Seq data for all 'r' series mutants. (A) Functional data collected for each sgRNA 'r' mutant. RFP fluorescence was measured and normalized by optical density (FL/OD). Error bars represent one standard deviation of four replicates. (B) 'r' series mutant structures are depicted with nucleotides color-coded by reactivity intensity if in-cell SHAPE-Seq data was collected. The placeholder 'no data' indicates no SHAPE-Seq data was collected for that mutant condition (with or without dCas9). Mutant sgRNA r9 appeared to be cleaved in the cell by a dsRNase according to the sequencing read alignments. (C) Individual reactivity maps for 'r' series mutants that had data collected. The reactivity color scale is the same as in (B)

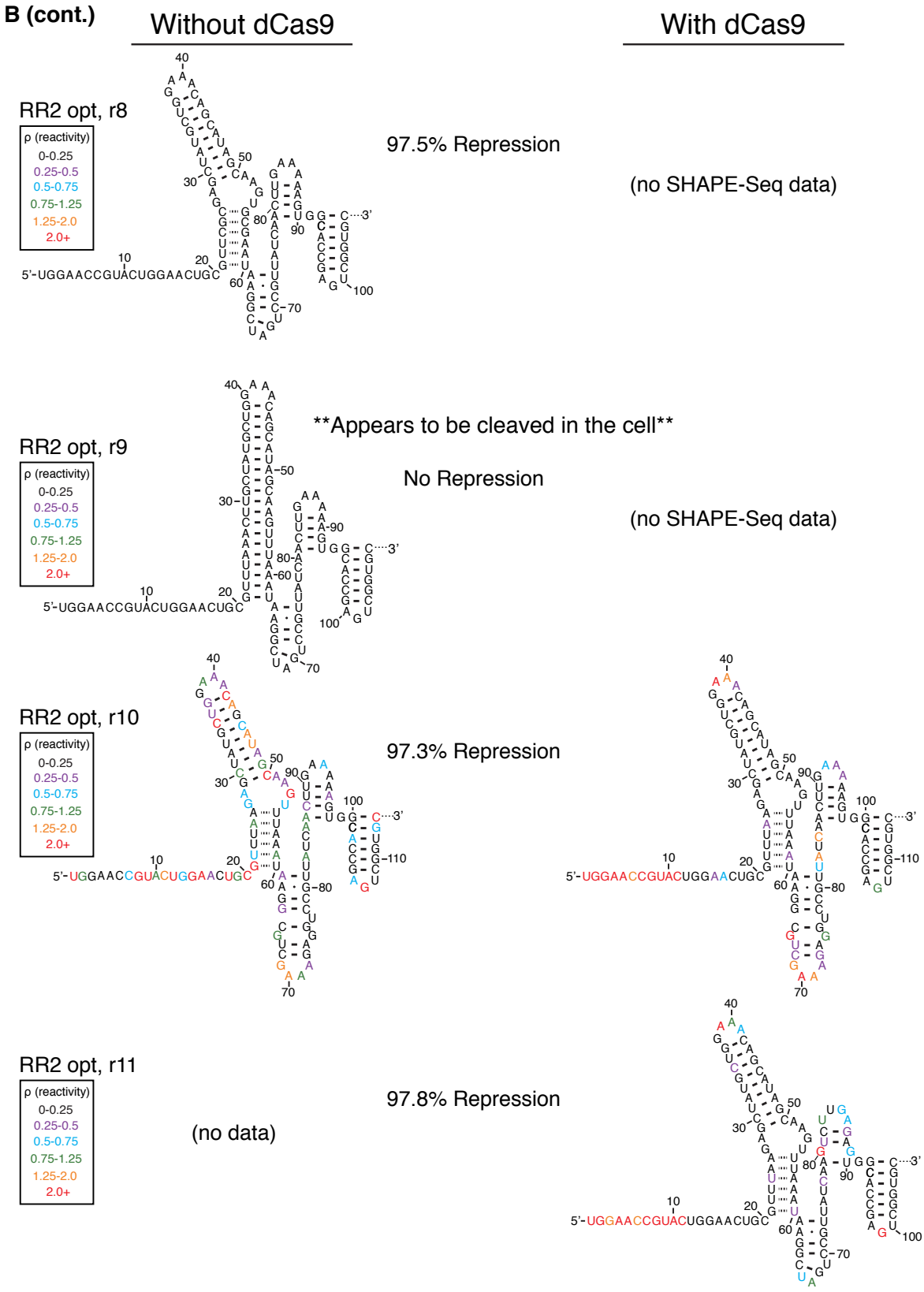

A
FL/OD
4×10^3
2×10^3
0
GF1
RR2 opt
97.6%
77.8%
28.0%
-34.9%
75.5%
3.6%
3.6%
12.4%
97.6%
-9.9%
97.3%
97.8%
97.6%
r1
r2
r3
r4
r5
r6
r7
r8
r9
r10
r11
r12
98.0%
97.1%
26.4%
96.3%
72.4%
-20.0%
86.5%
73.0%
80.7%
86.6%
96.7%
r13
r14
r15
r16
r17
r18
r19
r20
r21
r22
r23


B
Without dCas9
With dCas9
RR2 opt, r1
ρ (reactivity)
0-0.25
0.25-0.5
0.5-0.75
0.75-1.25
1.25-2.0
2.0+
(no data)
77.8% Repression
5'-UGGAACCGUACUGGAACUGC
RR2 opt, r2
(no data)
28.0% Repression
5'-UGGAACCGUACUGGAACUGC
RR2 opt, r3
(no data)
No Repression
5'-UGGAACCGUACUGGAACUGC

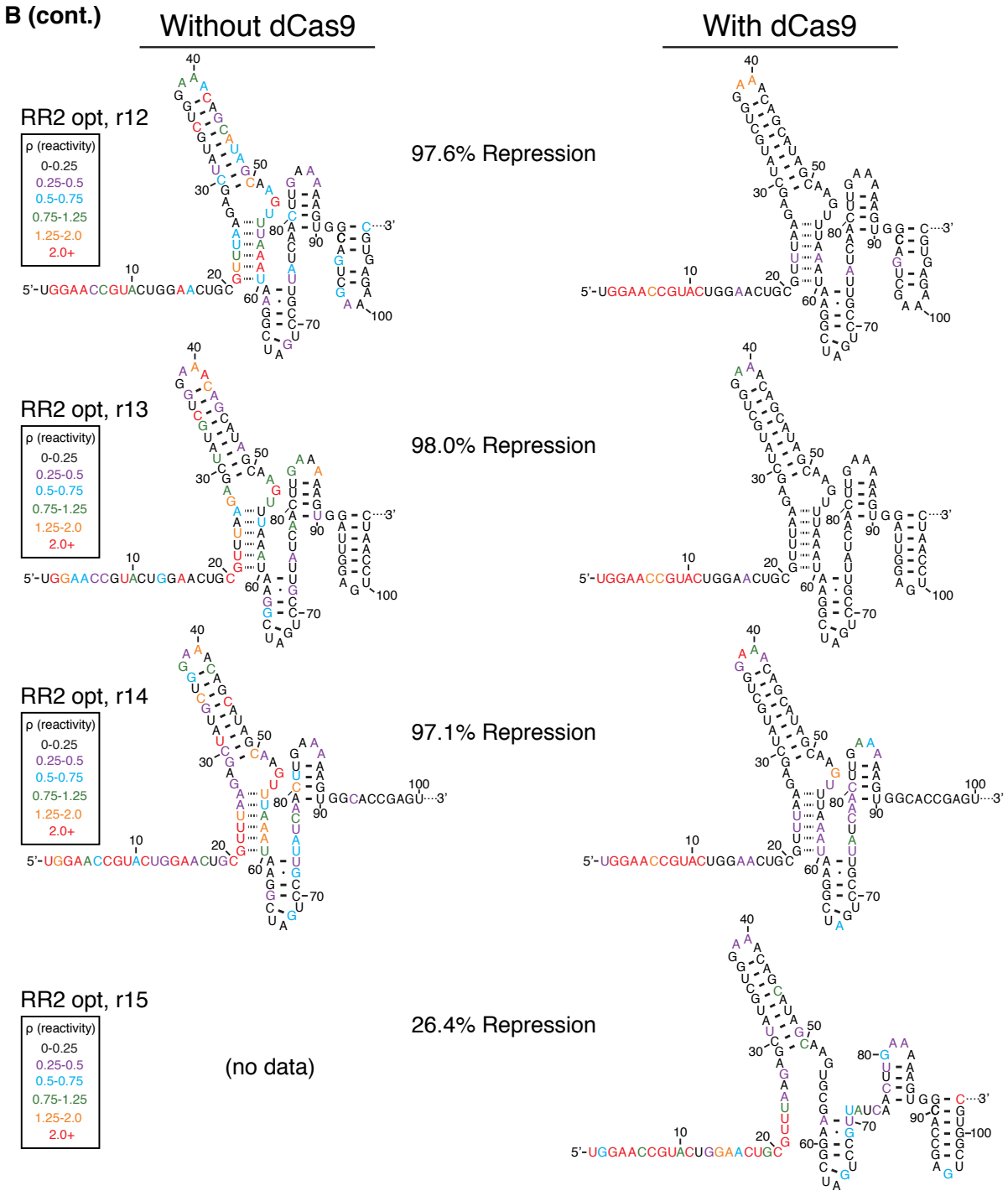
(Figure E.3 cont.)



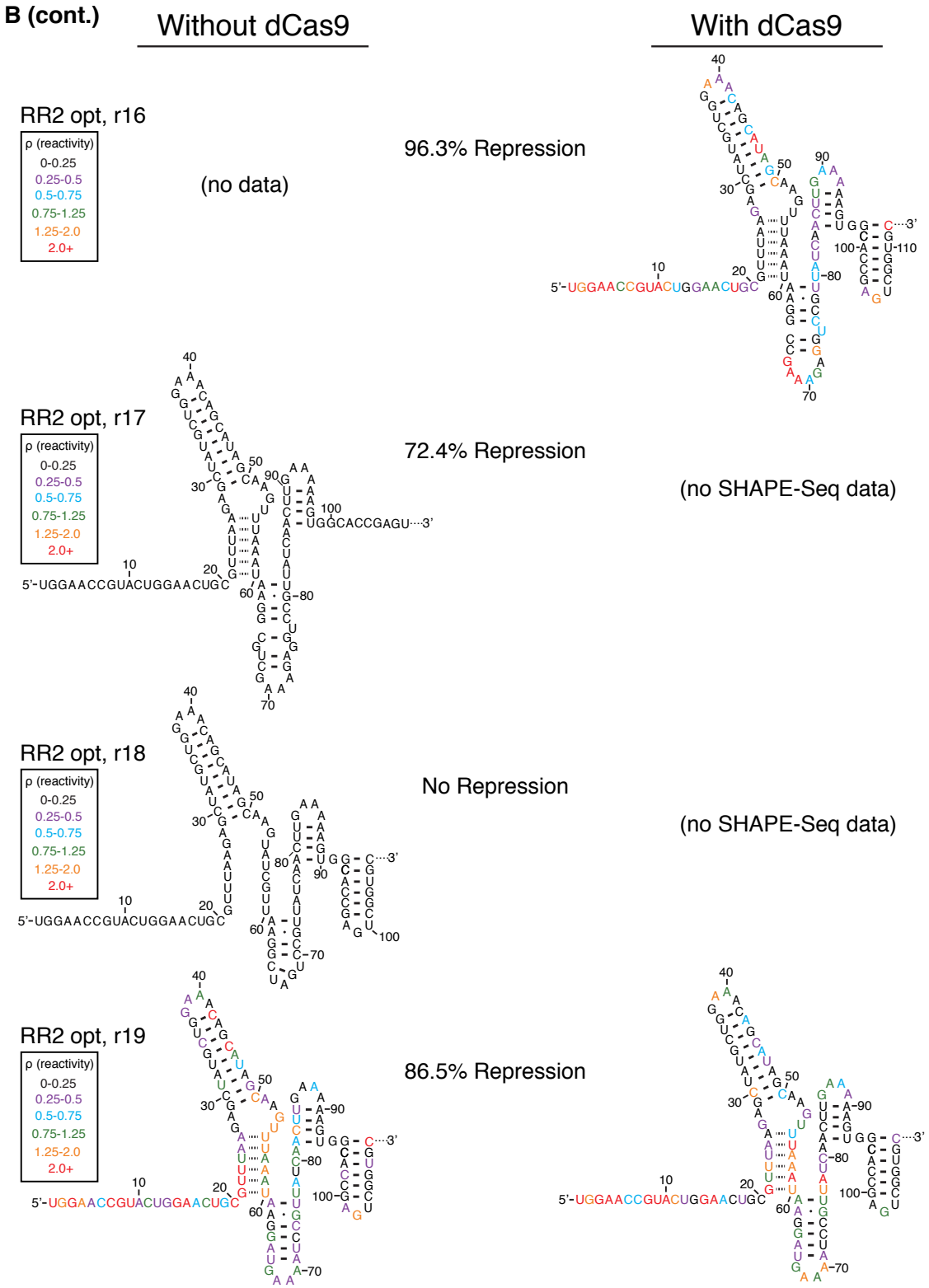
(Figure E.3 cont.)



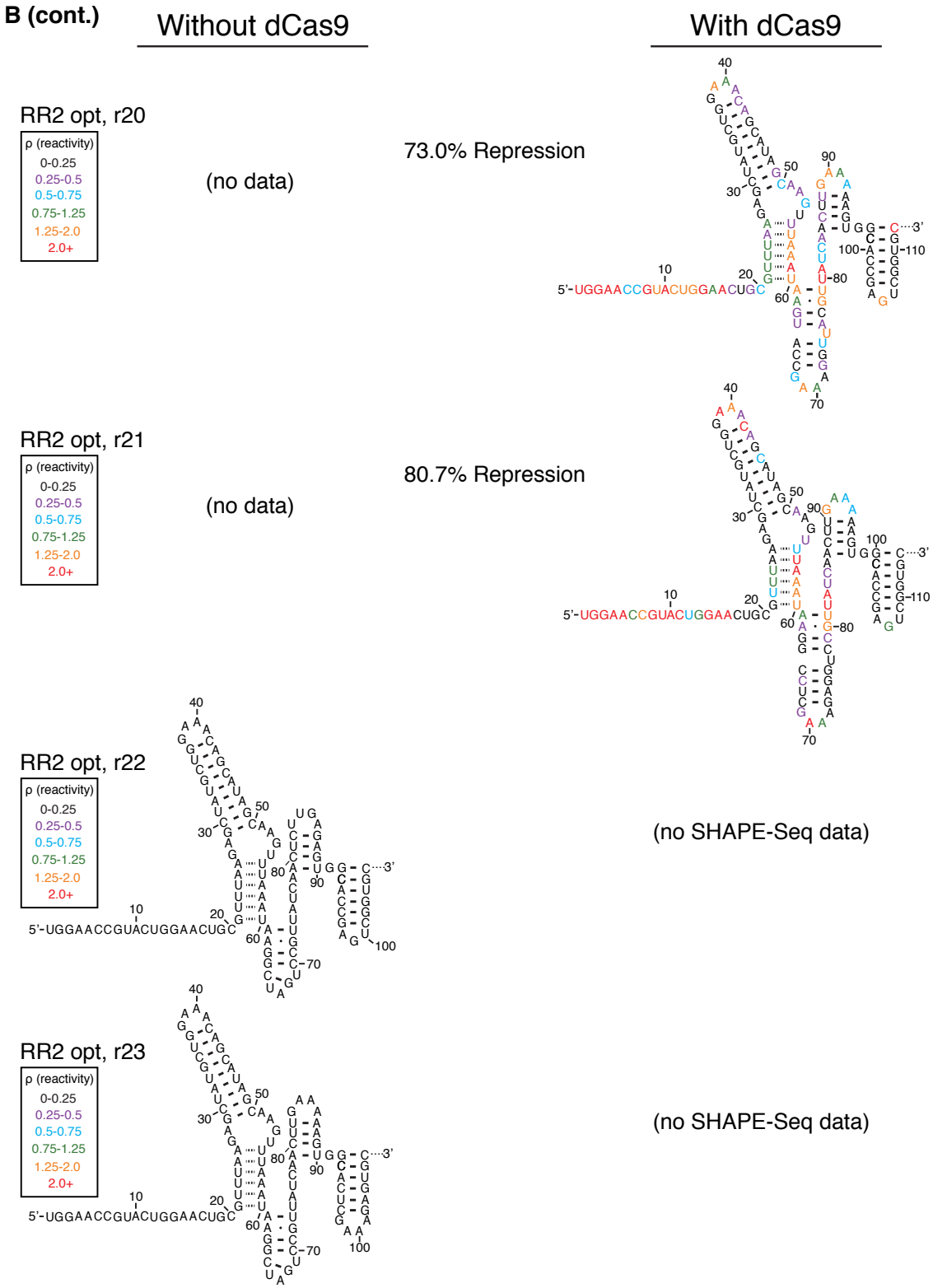
(Figure E.3 cont.)



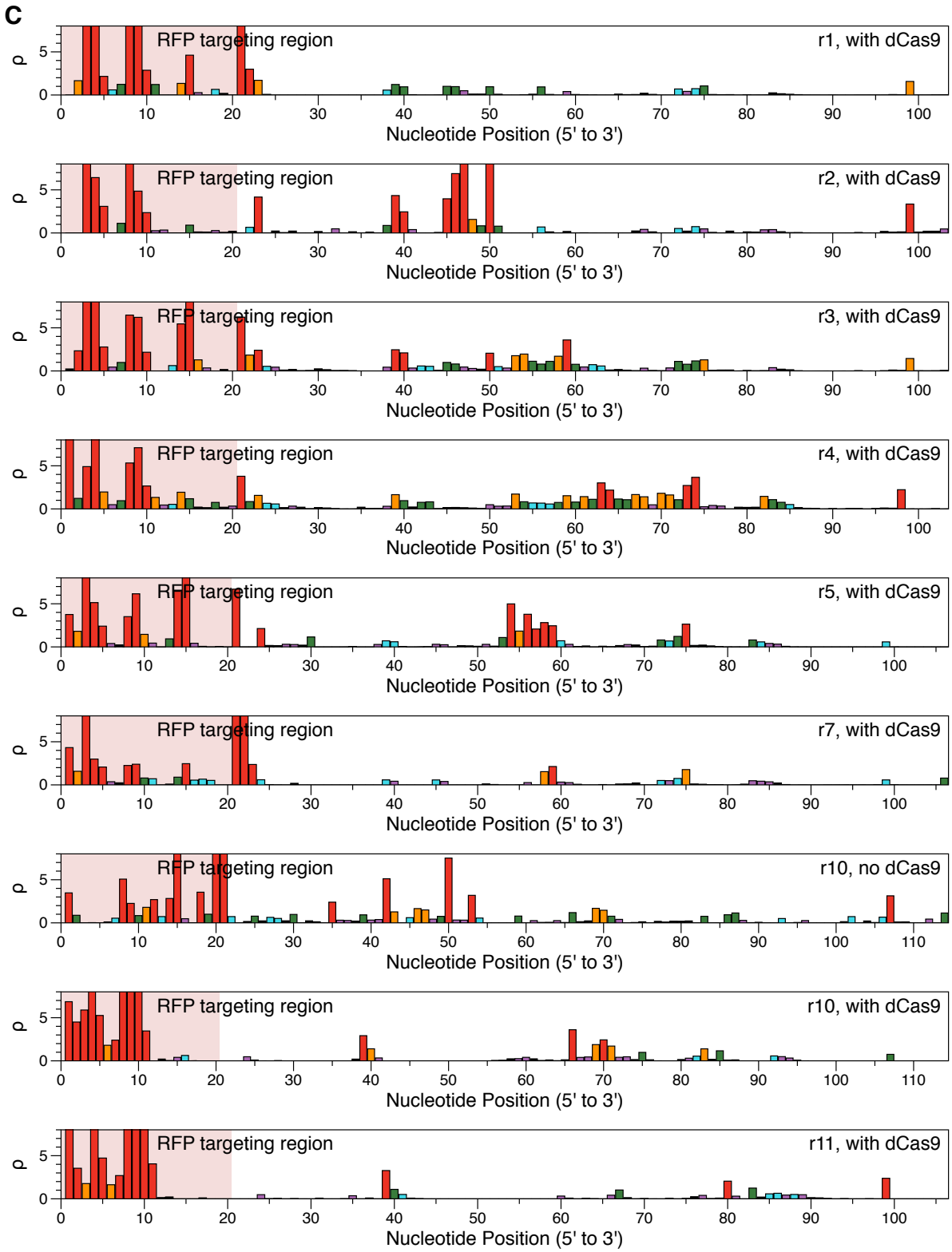
(Figure E.3 cont.)



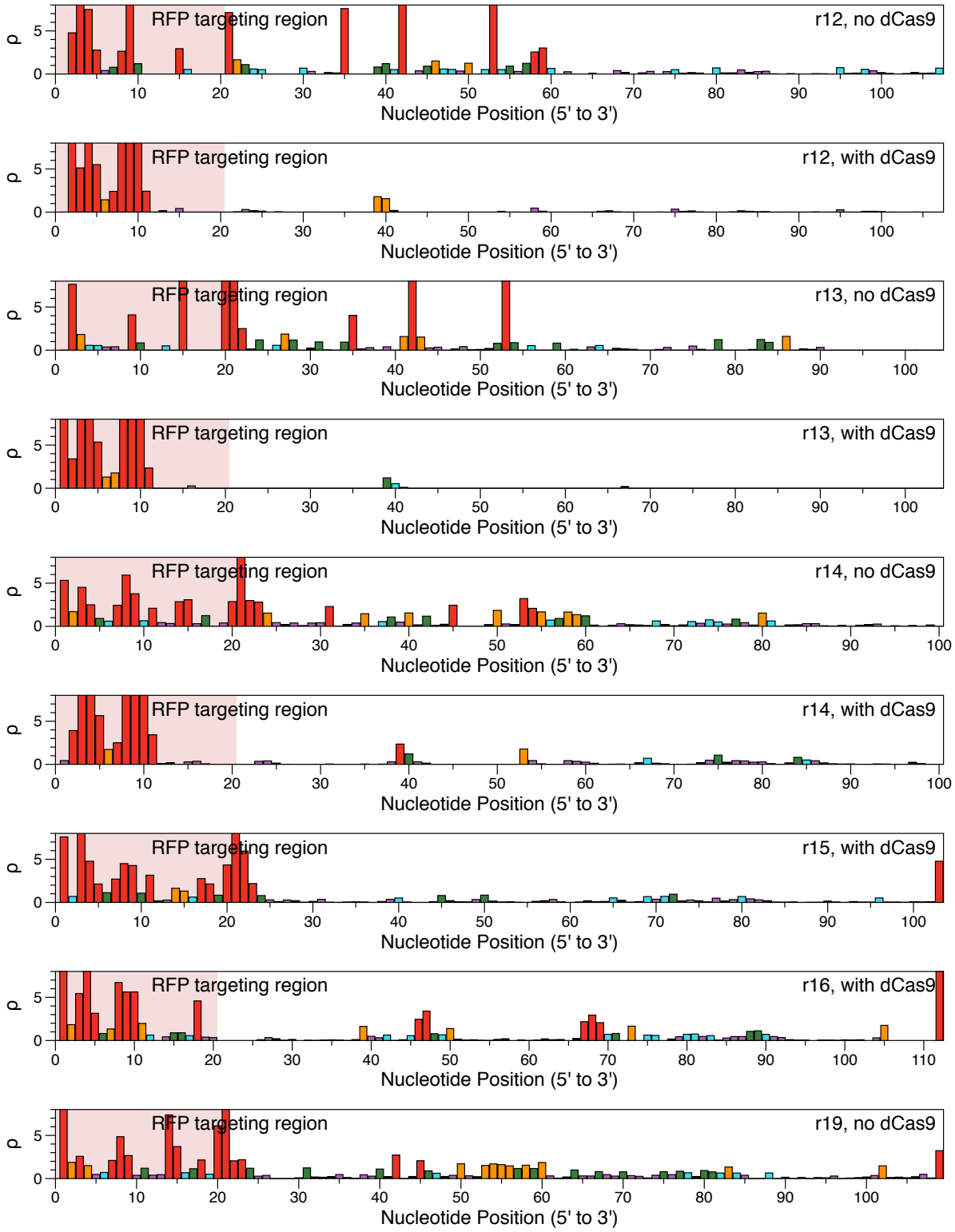
(Figure E.3 cont.)



(Figure E.3 cont.)



(Figure E.3 cont.)



(Figure E.3 cont.)

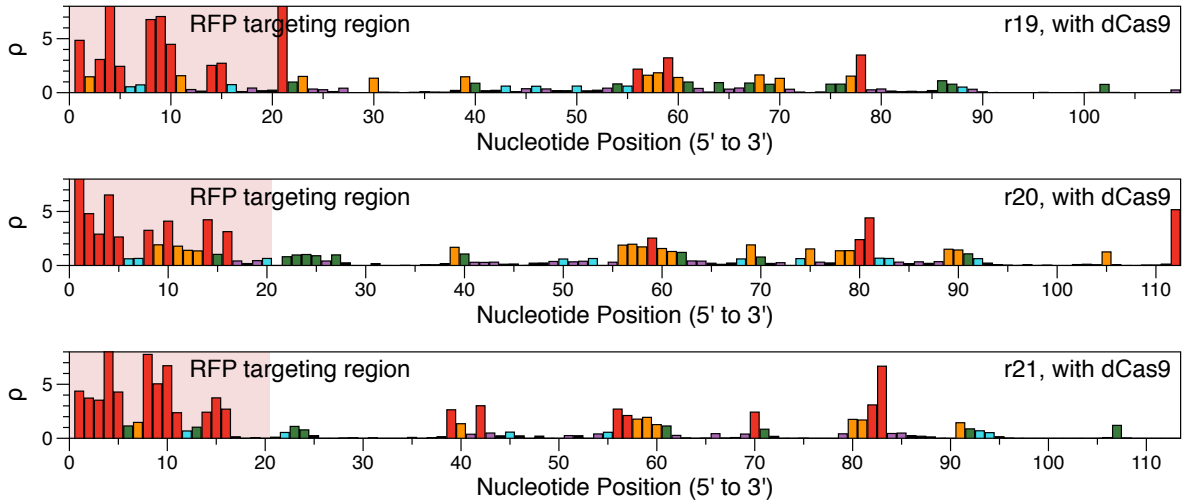
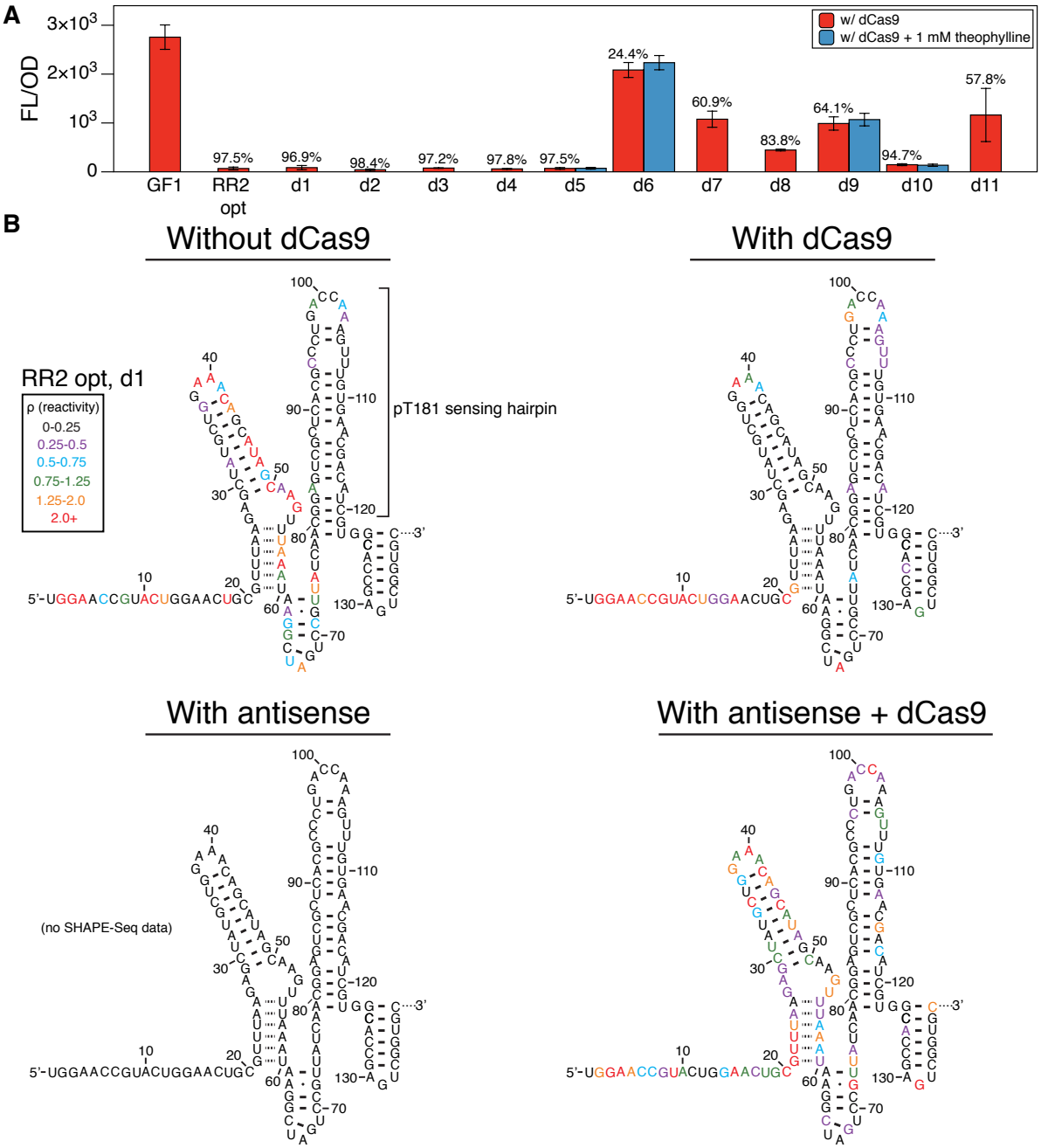
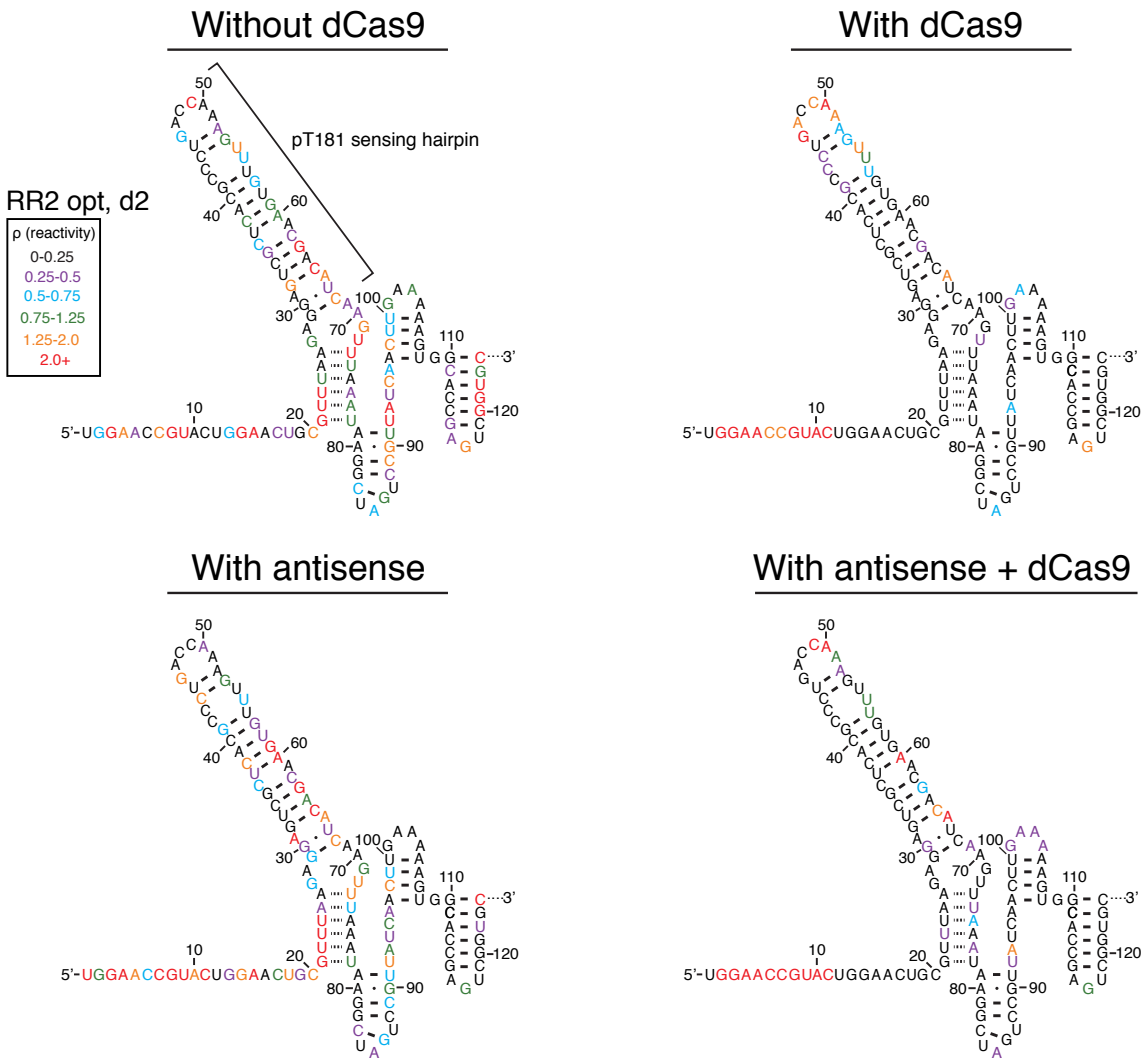


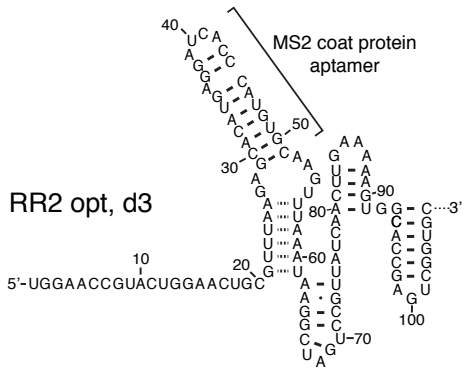
Figure E.4: Secondary structures and functional and in-cell SHAPE-Seq data for all 'd' series designs. (A) Functional data collected for each sgRNA 'd' design. RFP fluorescence was measured and normalized by optical density (FL/OD). Theophylline was added to 1 mM final during subculture to test sgRNA designs containing the theophylline aptamer. Error bars represent one standard deviation of four replicates. (B) 'd' series mutant structures are depicted with nucleotides color-coded by reactivity intensity if in-cell SHAPE-Seq data was collected. Inserted RNA regulator sequences are indicated in each structure. (C) Individual reactivity maps for d1 and d2 designs (those that had data collected). The reactivity color scale is the same as in (B).

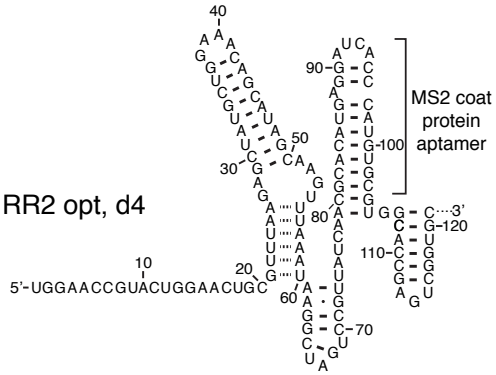

A
FL/OD
3x10^3
2x10^3
10^3
0
w/ dCas9
w/ dCas9 + 1 mM theophylline
97.5%
96.9%
98.4%
97.2%
97.8%
97.5%
24.4%
60.9%
83.8%
64.1%
94.7%
57.8%
GF1
RR2 opt
d1
d2
d3
d4
d5
d6
d7
d8
d9
d10
d11
B
Without dCas9
With dCas9
RR2 opt, d1
ρ (reactivity)
0-0.25
0.25-0.5
0.5-0.75
0.75-1.25
1.25-2.0
2.0+
pT181 sensing hairpin
5'-UGGAACCGUACUGGAACUGC
With antisense
With antisense + dCas9
(no SHAPE-Seq data)

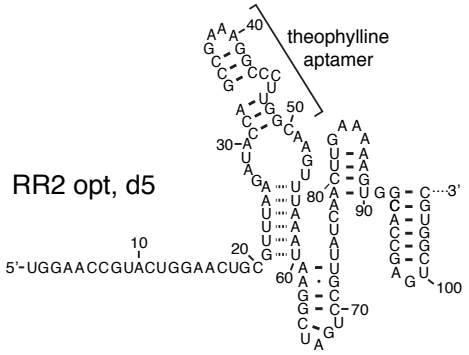
(Figure E.4 cont.)

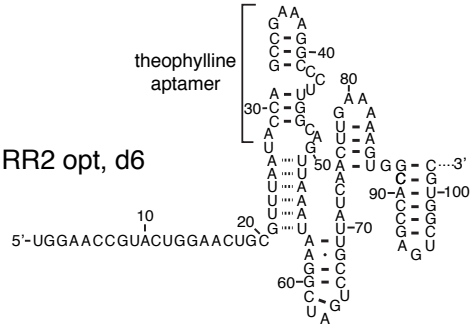


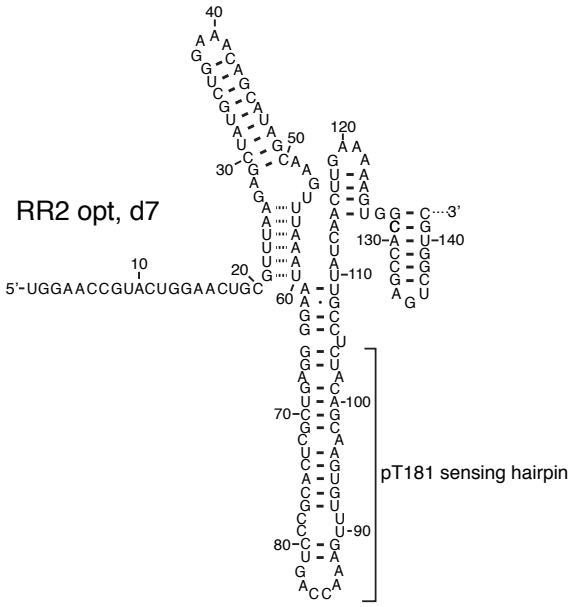
(Figure E.4 cont.)

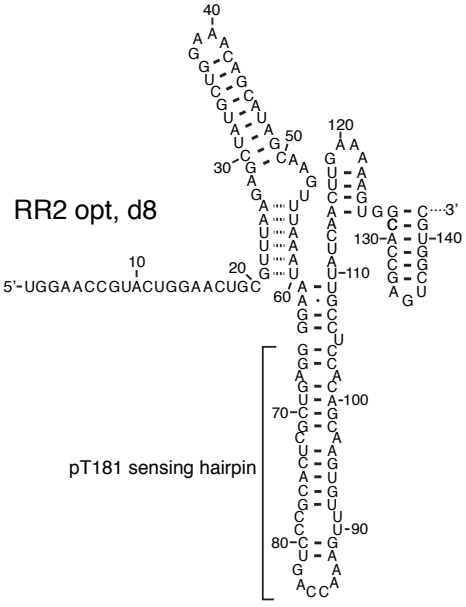




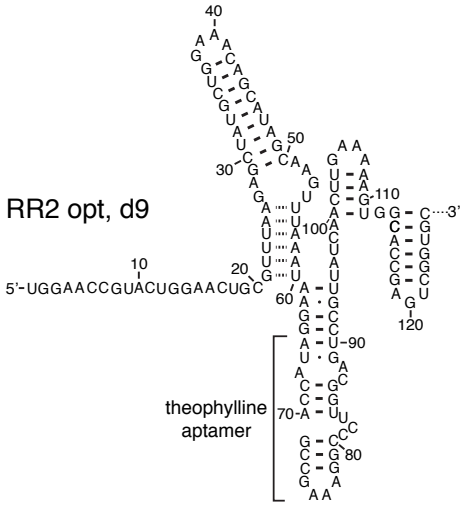


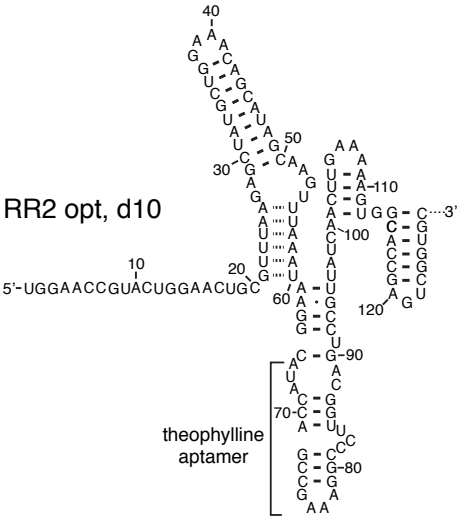


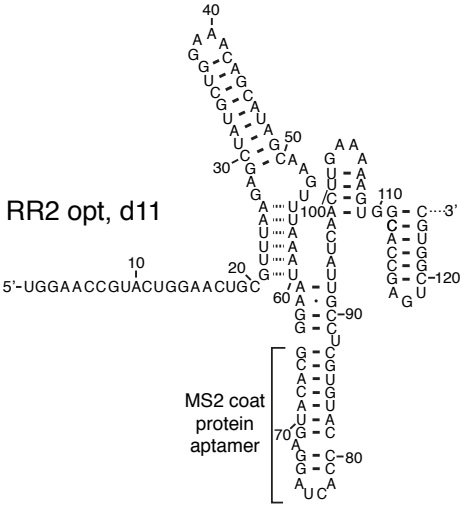




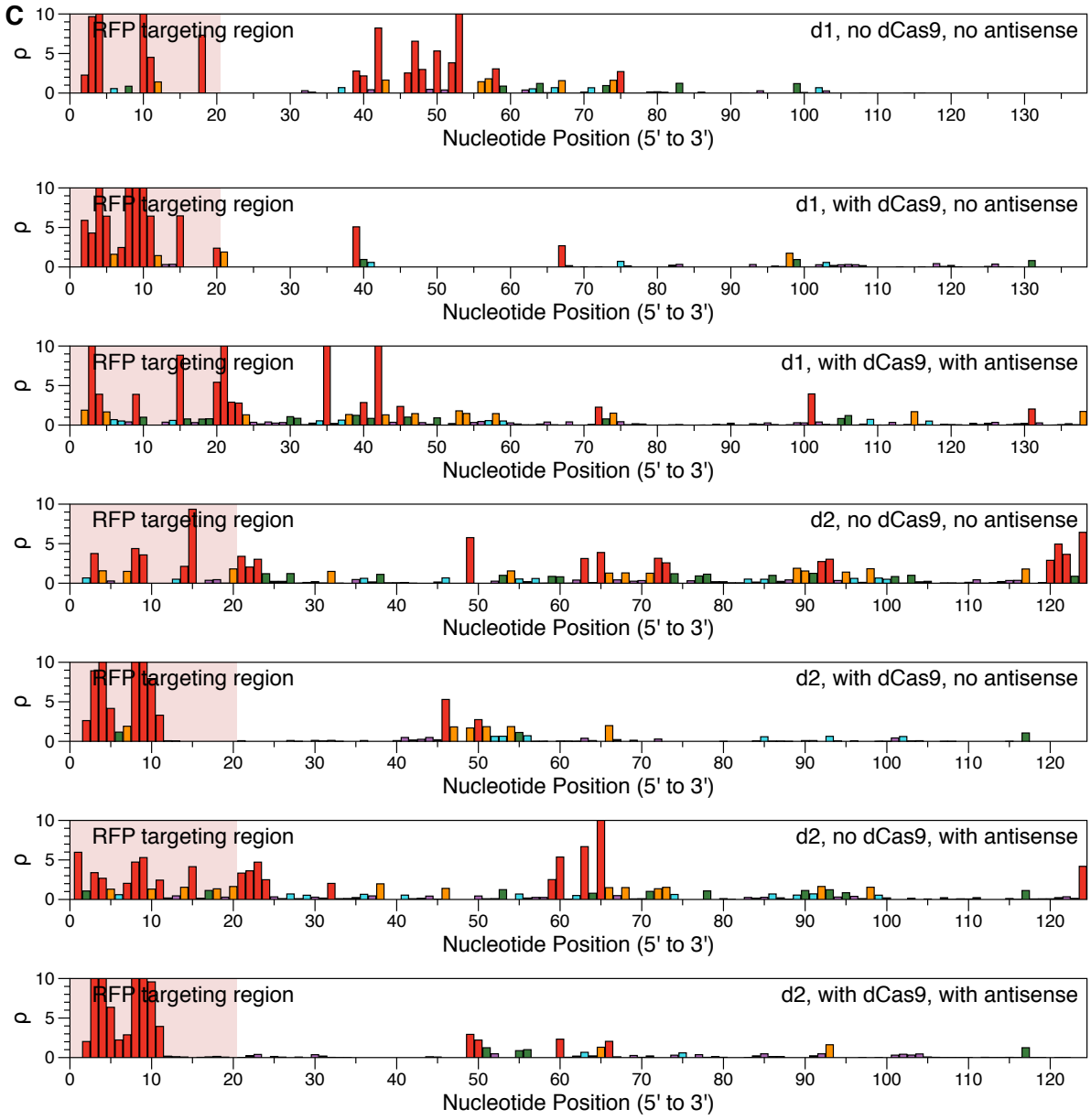
(Figure E.4 cont.)







(Figure E.4 cont.)



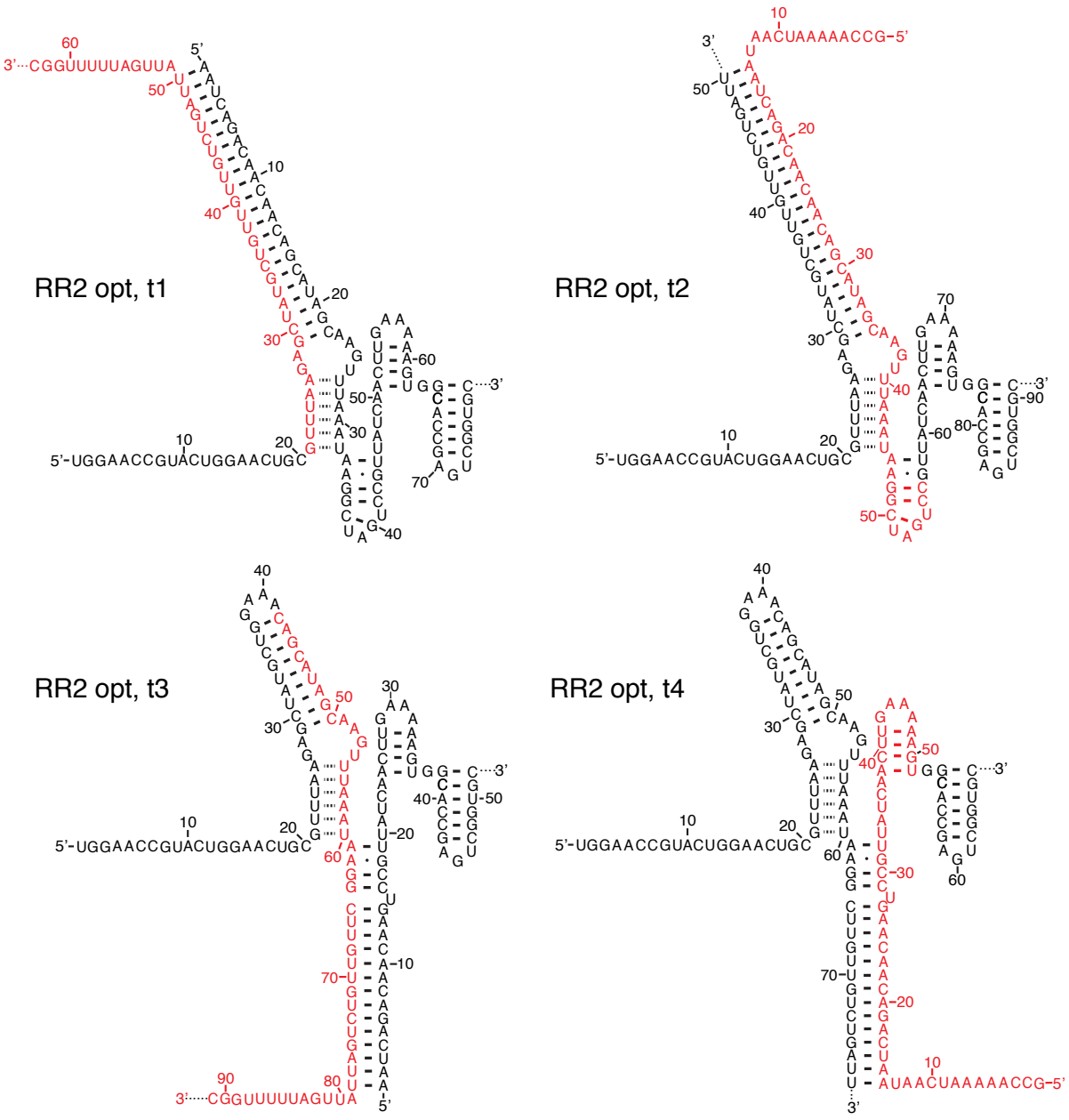


Figure E.5: Secondary structures for all 't' series toehold designs. Each toehold design structure is depicted as the complete sgRNA interaction. Nucleotides that base pair with the antisense RNA are colored red.

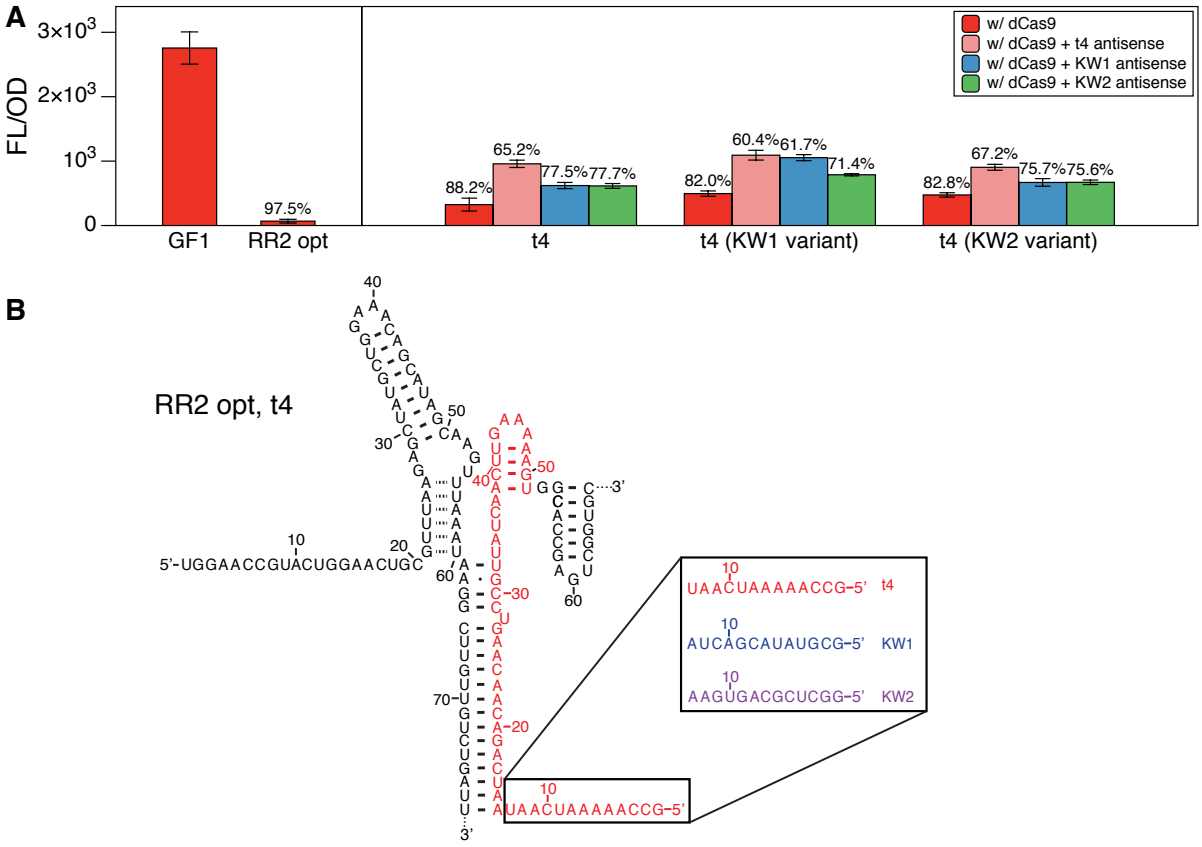


Figure E.6: Antisense orthogonality test of the sgRNA t4 design. (A) RFP fluorescence normalized by optical density (FL/OD) was measured for three orthogonal variants of the sgRNA t4 design depicted in (B). (B) Each variant of the sgRNA t4 design was generated by changing the sequence of the toehold (inset). All of the sgRNA variants are activated by every antisense variant, showing no orthogonality. Error bars represent one standard deviation of four replicates.

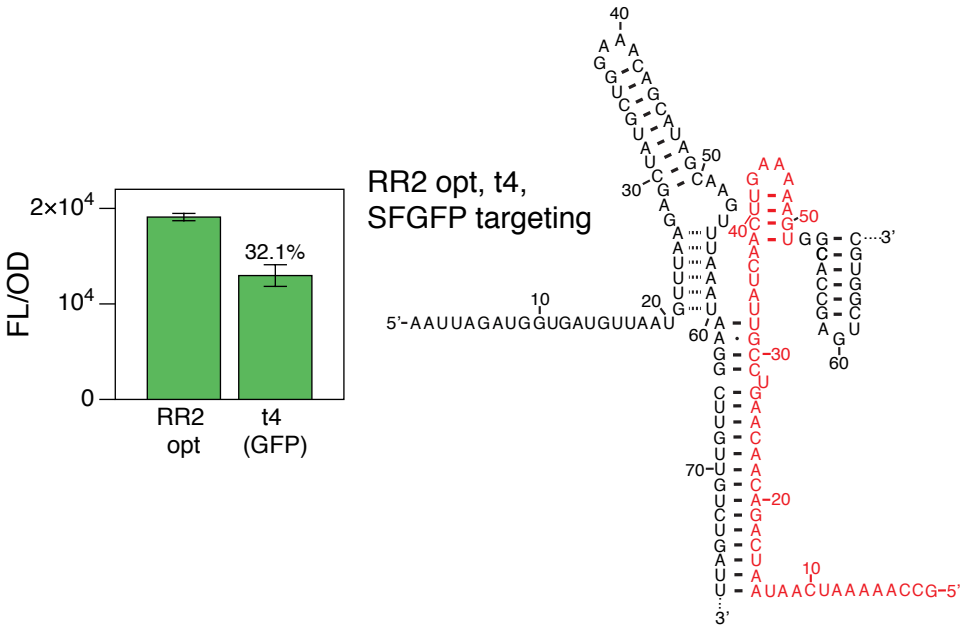


Figure E.7: Assessing sgRNA target orthogonality in the sgRNA t4 design. GFP fluorescence normalized by optical density (FL/OD; left) was measured for the sgRNA t4 design with its targeting sequenced changed from RFP to GFP (right) with and without dCas9 expression. Poor repression indicates that the sgRNA t4 design is sensitive to the sequence of the targeting region.

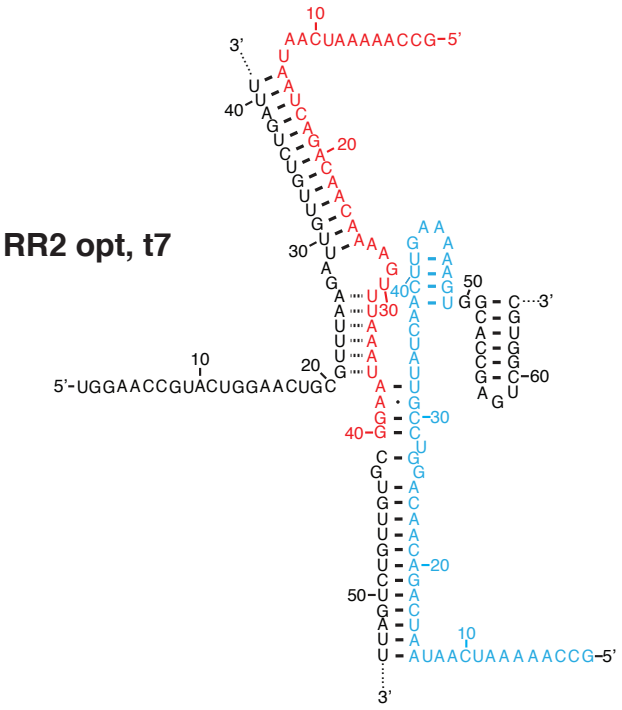


Figure E.8: Potential OR gate design combining t5 and t6 designs. By combining the toeholds introduced in the first two hairpins by the t5 and t6 designs into a t7 design, a sgRNA that derepresses in response to one of two antisenses (red and blue) could be created. Shown above is one potential design that targets mRFP1 that could serve as an OR gate.

APPENDIX F

SUPPLEMENTARY INFORMATION FOR STRUCTURAL ANALYSIS OF CUCUMBER MOSAIC VIRUS RNA3

F.1 Supplementary Figures

Figure F.1: SHAPE-Seq reactivity maps of the RNA3 5′ UTR. Reactivity data for the RNA3 5 UTR in four different contexts: 5′ UTR only (blue), 5+I (red), purified viral RNA (green), and in infected cell lysates (purple), and refolded full length *in vitro* transcripts (gray). The reactivity maps are very similar, except for the GUCGUGUUG nucleotide sequence repeats (nts 40-48 and 62-70) that have a different pattern when the MP ORF and IGR are not present. Both repeats have the same pattern when only the 5′ UTR is present and similar patterns in all the conditions where the IGR is included. Error bars represent one standard deviation of three replicates, except for the infected lysate experiment that contains five replicates.

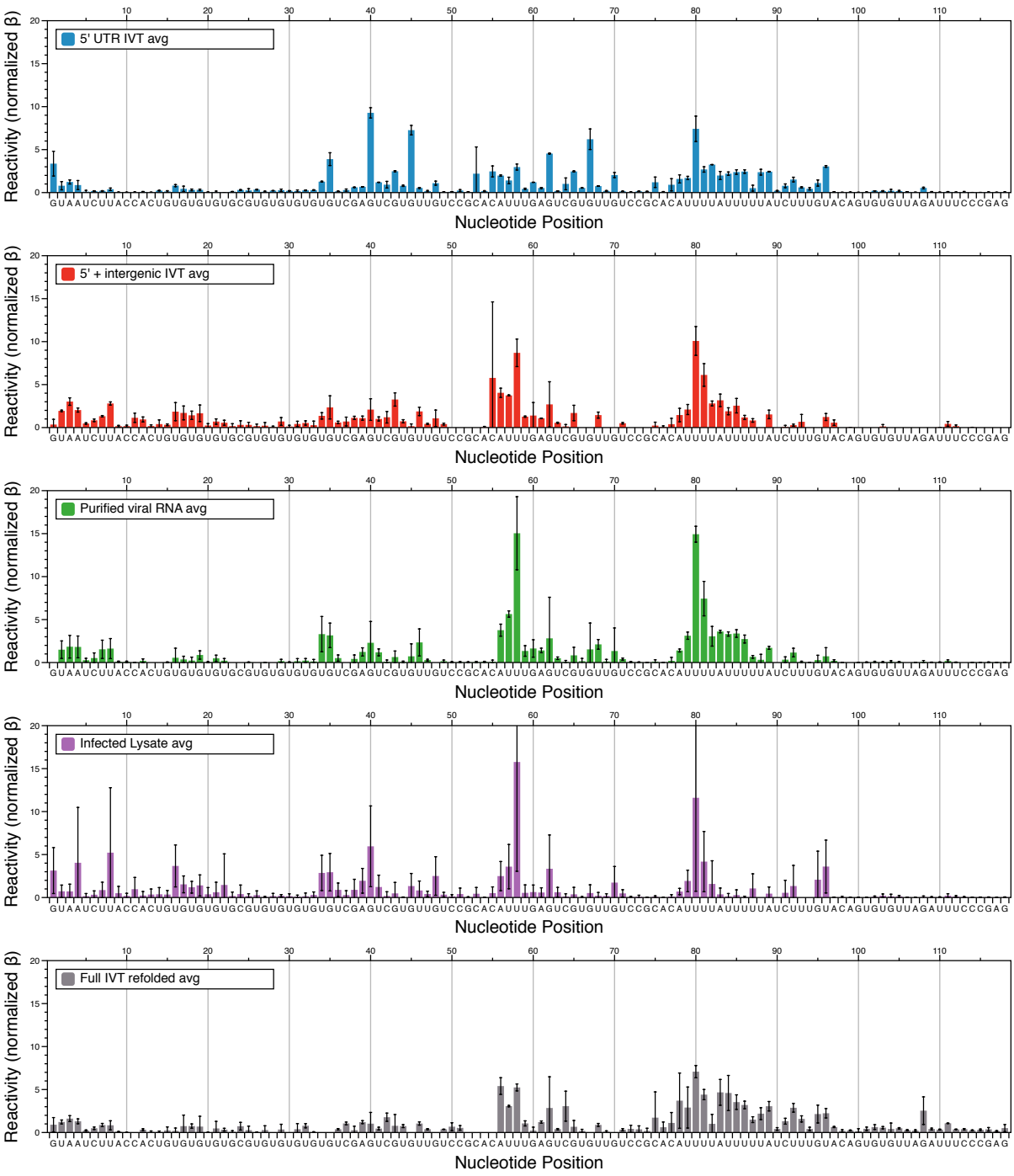
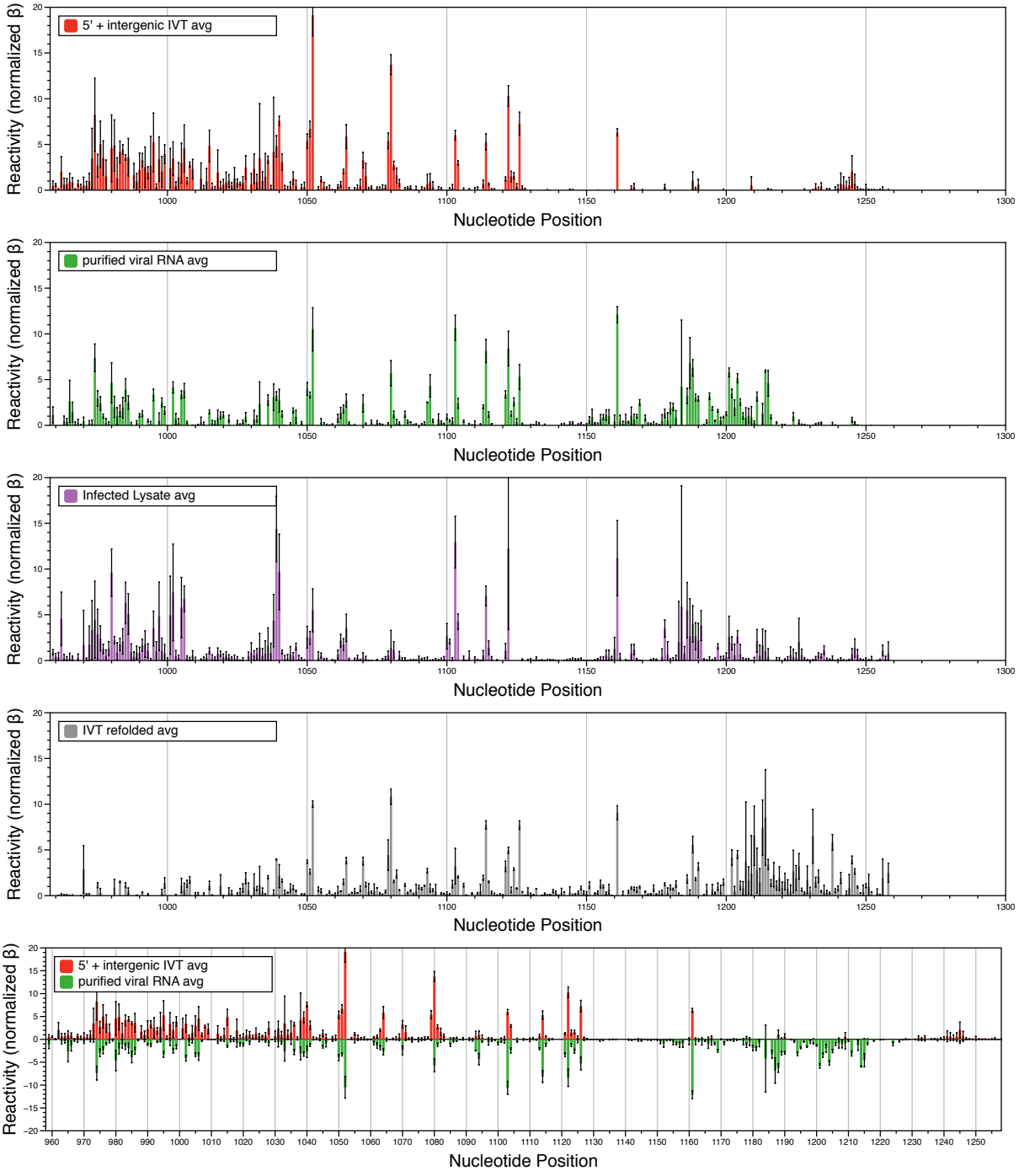

5' UTR IVT avg
5' + intergenic IVT avg
Purified viral RNA avg
Infected Lysate avg
Full IVT refolded avg
Reactivity (normalized β)
Nucleotide Position
GUAAUCUUACCACUGUGUGUGUGCGUGUGUGUGUGUCGAGUCGUGUUGUCCGCACAUUUGAGUCGUGUUGUCCGCACAUUUUAUUUUUAUCUUGUACAGUGUGUUAGAUUUCCGAG

Figure F.2: SHAPE-Seq reactivity maps of the RNA3 IGR. Reactivity data for the RNA3 IGR in four different contexts: 5+I (red), purified viral RNA (green), infected cell lysates (purple), and refolded *in vitro* transcripts (gray). The reactivity maps for the lysates and purified virions match closely, also matching the 5+I map as well, except that nts 1180-1220 are not reactive in the 5+I RNA (bottom). The refolded *in vitro* transcript maps differ the most, exhibiting low reactivities in the first 50 nts, and high reactivities in the last 50. The observed reactivity differences between maps representing subsections of the full length RNA3 could be a result of a different structural context of RNA4 that cannot be distinguished from the end of the IGR for the purified virions and lysate samples. The other observed differences could be due to unidentified long-range interactions. Error bars represent one standard deviation of three replicates, except for the infected lysate experiment that contains five replicates.


5' + intergenic IVT avg
purified viral RNA avg
Infected Lysate avg
IVT refolded avg
5' + intergenic IVT avg
purified viral RNA avg
Reactivity (normalized β)
Nucleotide Position

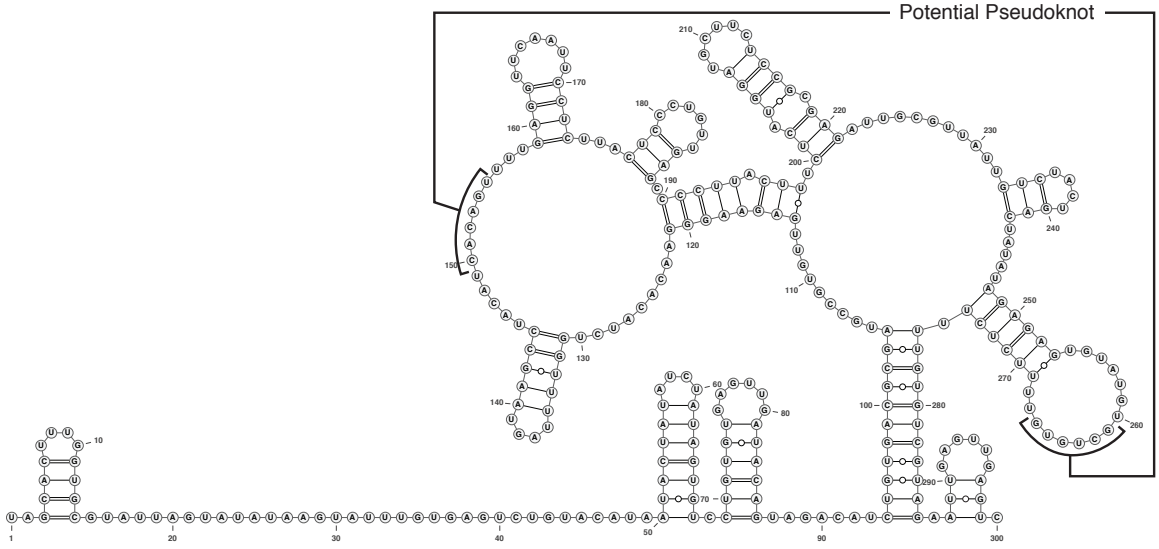


Figure F.3: Alternate fold of the IGR. When folding the IGR with reactivities obtained from purified virions or infected lysates, a branched, pseudoknotted structure is obtained (shown above) instead of the elongated stem-loop structure depicted in Figure 8.2 and proposed by Baumstark *et al.* [354].

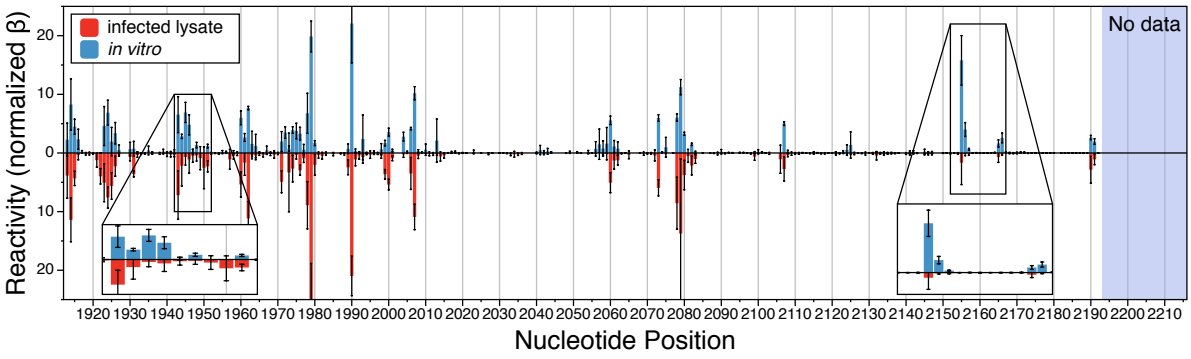
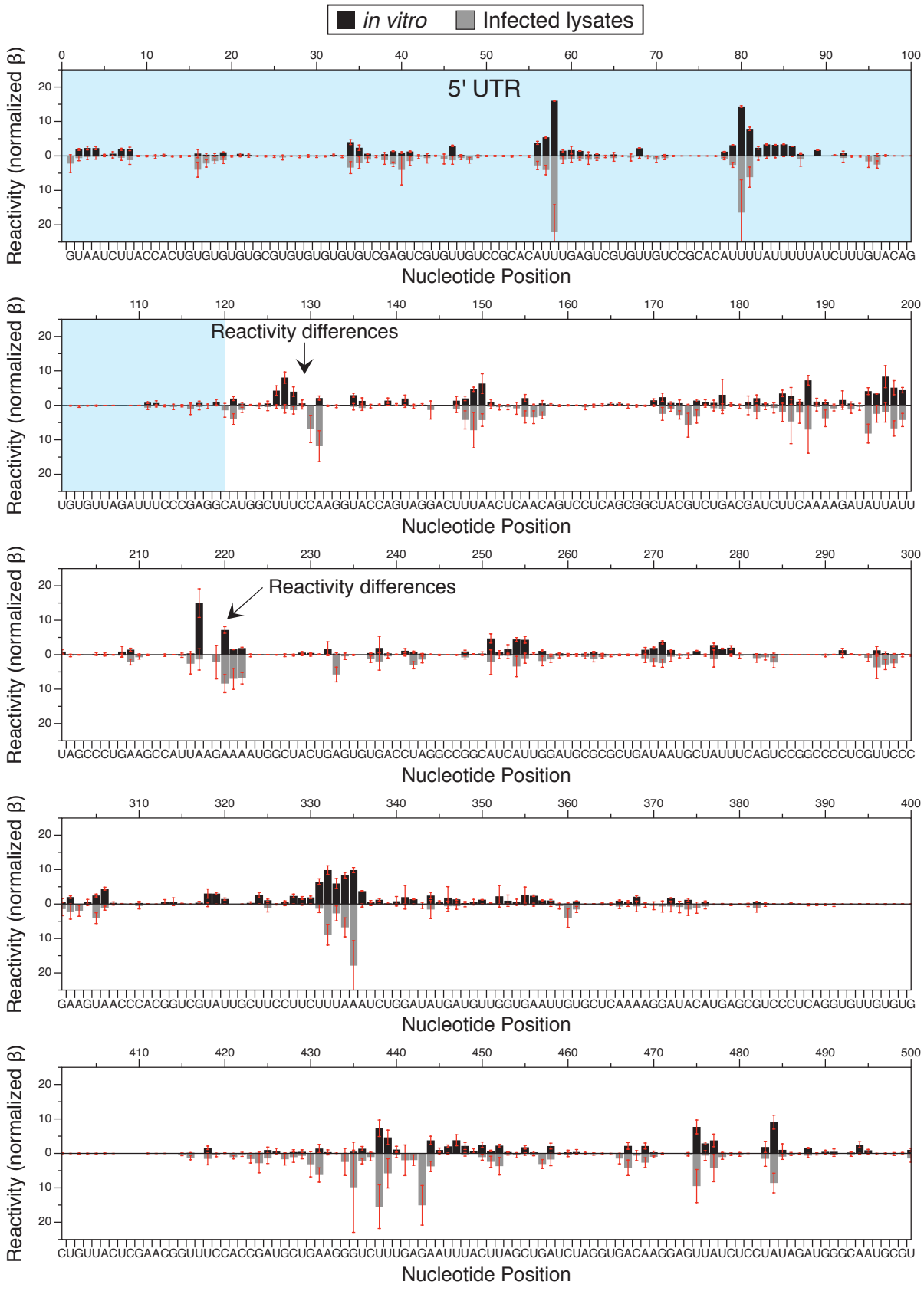
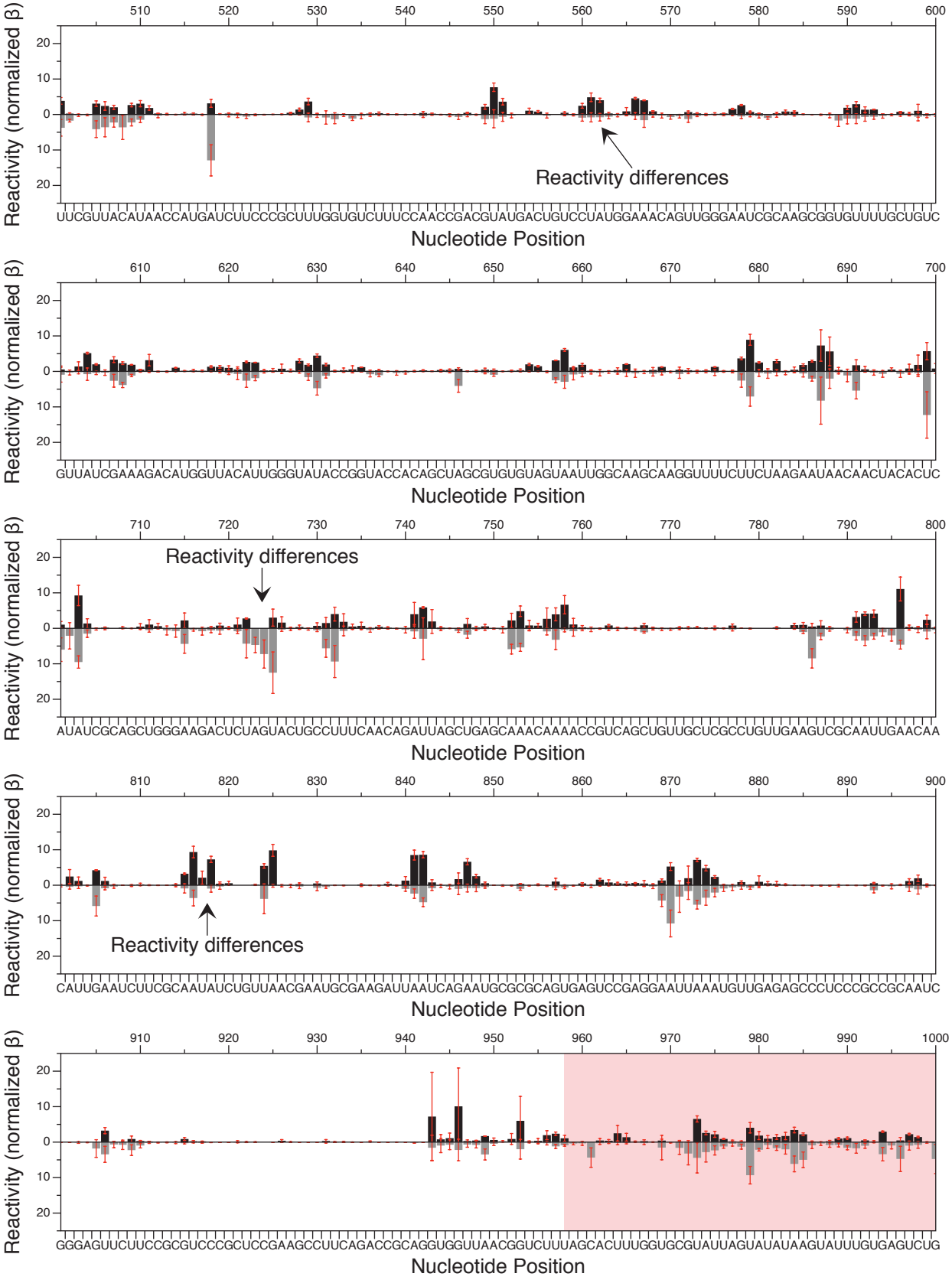


Figure F.4: SHAPE-Seq reactivity maps of the RNA3 3′ UTR. Reactivity data for the RNA3 3′ UTR in the full RNA3 measured using purified virions via *in vitro* refolding (blue) or directly in infected cell lysates (red). In the first ~240 nts the reactivity maps look very similar in both contexts except for positions 1944, 1945, 1947, 1971, and 2003 (relative to the start of the 3′ UTR) that appear higher *in vitro* and positions 1920, 1921, 1930, 1998, and 1999 that appear higher in infected lysate. At positions 2154, 2155, 2164, and 2165, the measured reactivity is higher *in vitro* than the lysates due to replicase binding in the trinucleotide loop (nts 2154 and 2155) and inner bulge (nts 2164 and 2165) of SLC [349, 362]. Error bars represent one standard deviation of three replicates for the *in vitro* experiments and five replicates for the infected cell lysate experiments.

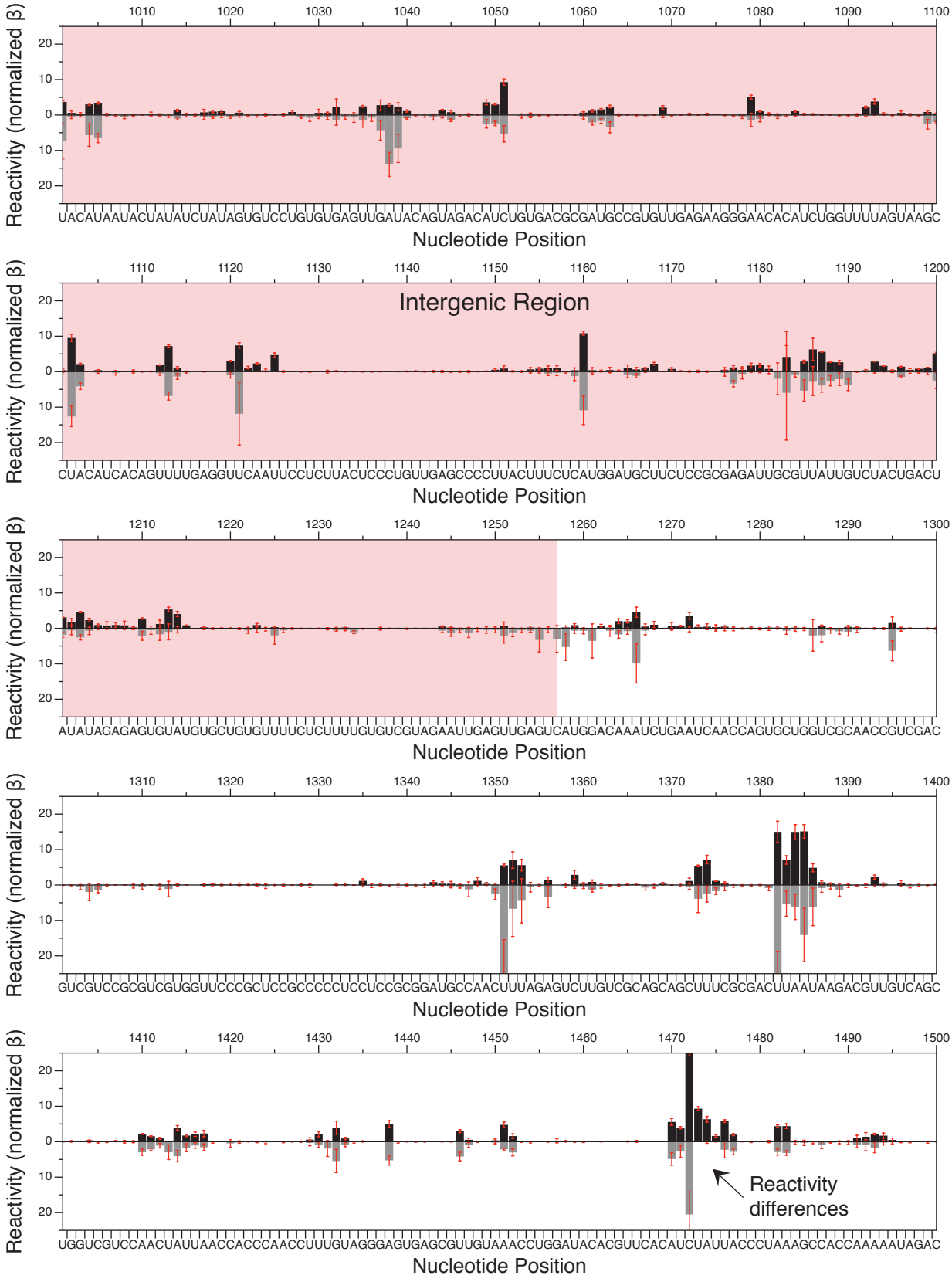
Figure F.5: SHAPE-Seq reactivities for the complete RNA3 from purified virions and infected cell lysates. SHAPE-Seq reactivities are displayed for the entire RNA3 genome segment from CMV strain Bn57. A number of observed differences between *in vitro* measurements from purified virions (black) and infected cell lysates (gray) are indicated.


in vitro
Infected lysates
5' UTR
Reactivity differences
Reactivity differences
Reactivity (normalized β)
Nucleotide Position
GUAAUCUUACCACUGUGUGUGUGCGUGUGUGUGUGUCGAGUCGUGUUGUCCGCACAUUUGAGUCGUGUUGUCCGCACAUUUUAUUUUUAUCUUUGUACAG
UGUGUUAGAUUUCCCGAGGCAUGGCUUUCCAAGGUACCAGUAGGACUUUAACUCAACAGUCCUCAGCGGCUACGUCUGACGAUCUUCAAAAGAUAUUAUU
UAGCCCUGAAGCCAUUAAGAAAAUGGCUACUGAGUGUGACCUAGGCCGGCAUCAUUGGAUGCGCGCUGAUAAUGCUAUUUCAGUCCGGCCCCUCGUUCCC
GAAGUAACCCACGGUCGUAUUGCUUCCUUCUUUAAAUCUGGAUAUGAUGUUGGUGAAUUGUGCUCAAAAGGAUACAUGAGCGUCCCUCAGGUGUUGUGUG
CUGUUACUCGAACGGUUUCCACCGAUGCUGAAGGGUCUUUGAGAAUUUACUUAGCUGAUCUAGGUGACAAGGAGUUAUCUCCUAUAGAUGGGCAAUGCGU

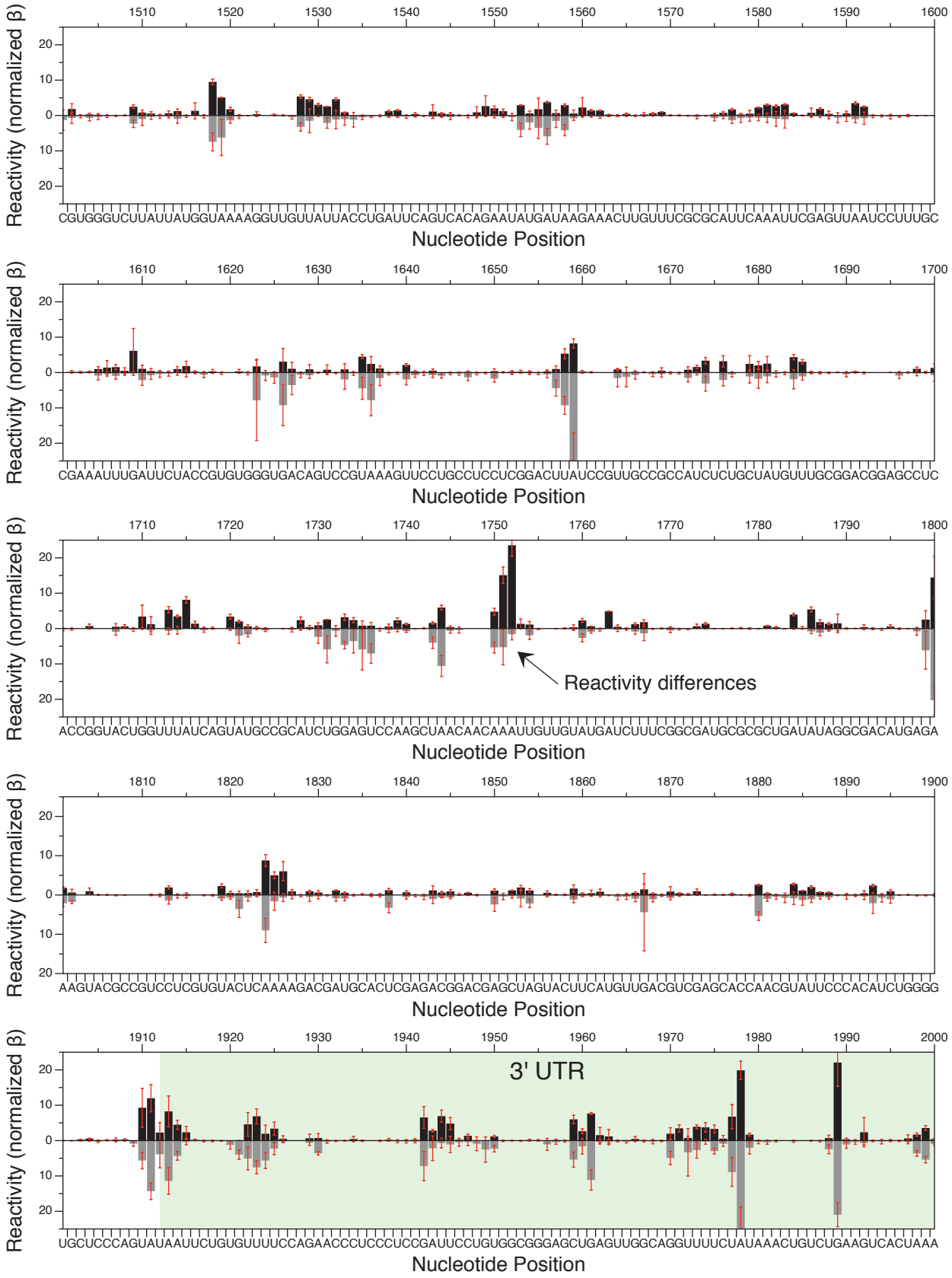
(Figure F.5 cont.)



(Figure F.5 cont.)



(Figure F.5 cont.)



(Figure F.5 cont.)

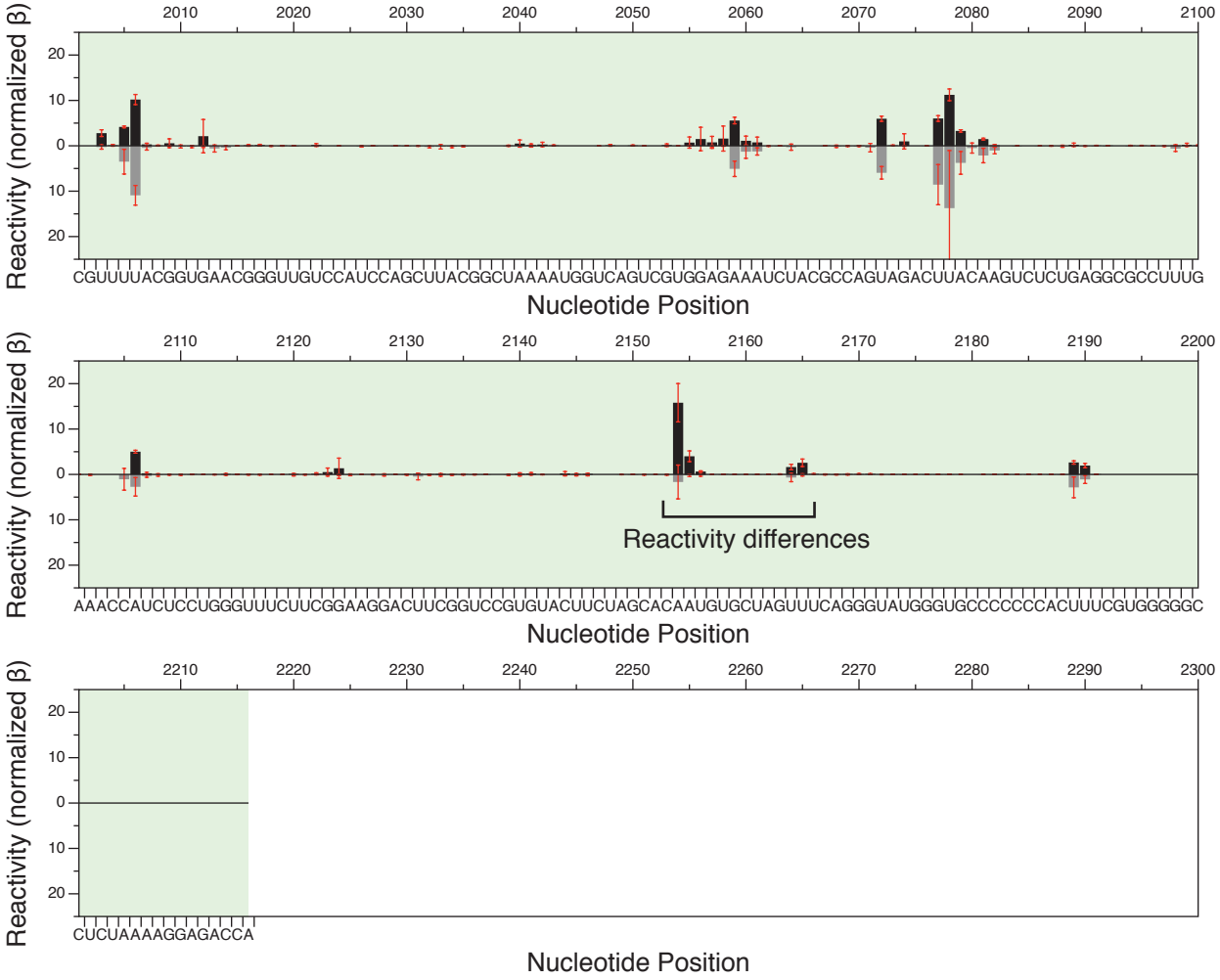
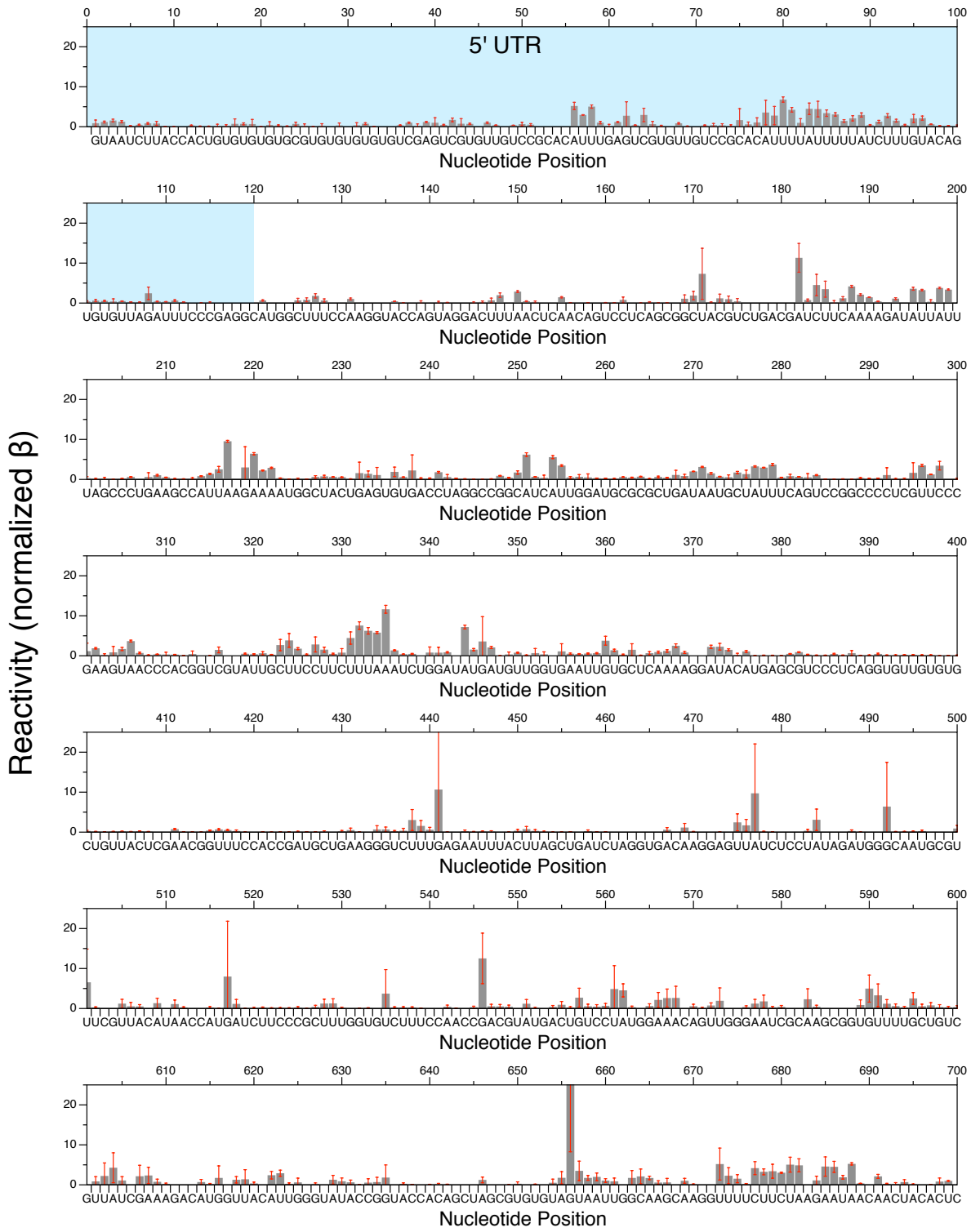
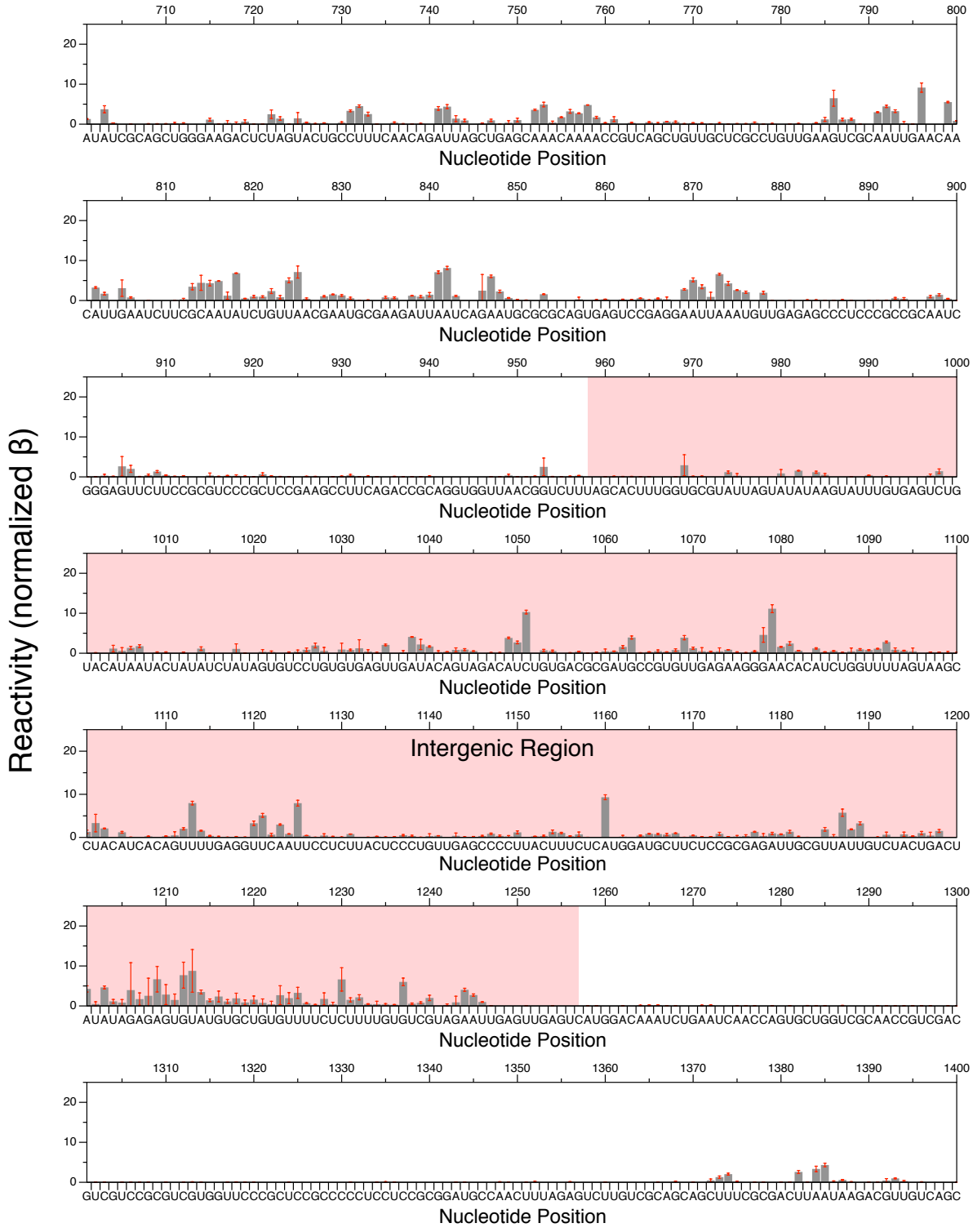


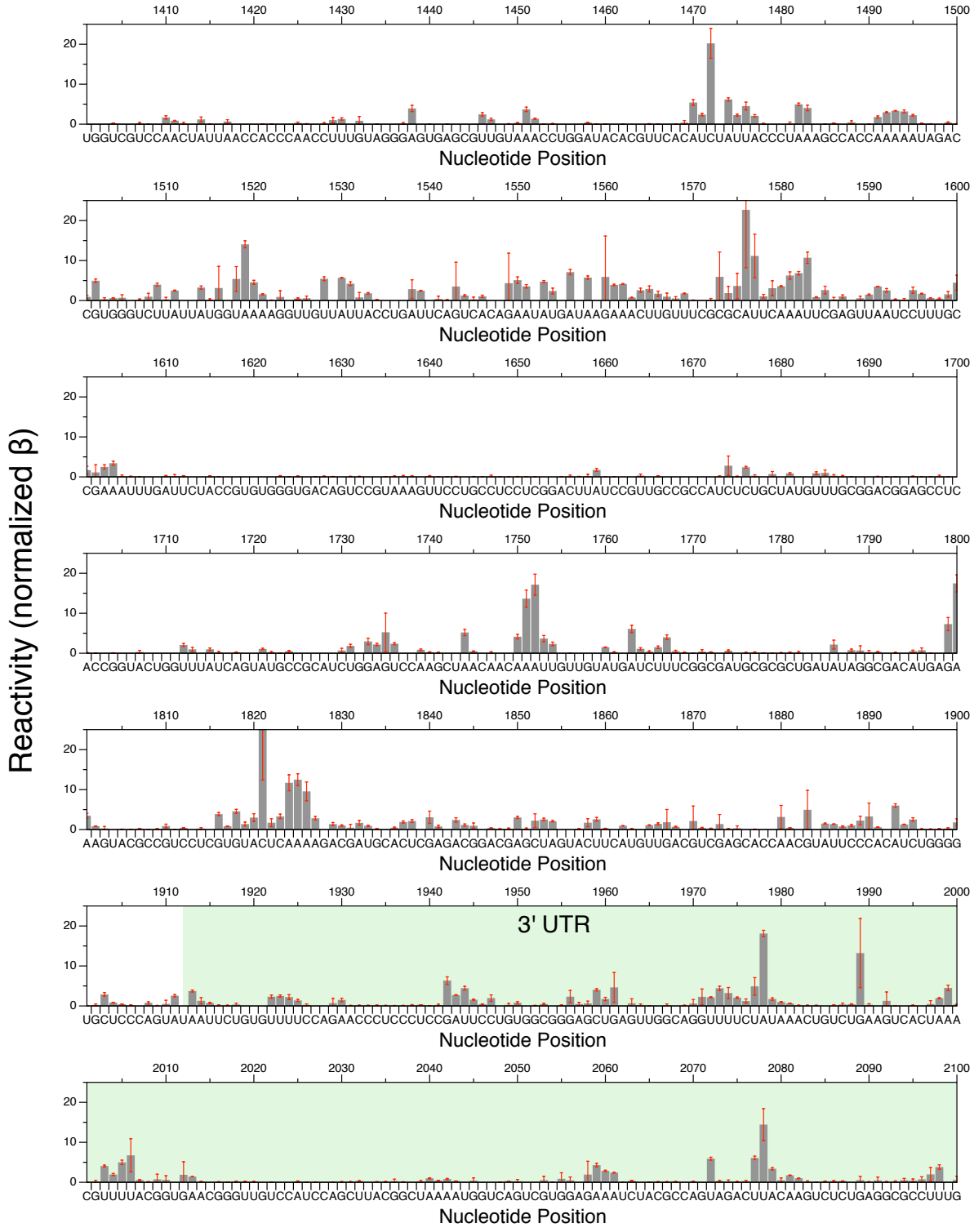
Figure F.6: SHAPE-Seq reactivities for the complete RNA3 from using refolded *in vitro* transcripts. SHAPE-Seq reactivities from refolded *in vitro* transcripts are displayed for the entire RNA3 genome segment from CMV strain Bn57.


Reactivity (normalized β)
5' UTR
GUAAUCUUACCACUGUGUGUGUGCGUGUGUGUGUGUCGAGUCGUGUUGUCCGCACAUUUGAGUCGUGUUGUCCGCACAUUUUAUUUUUAUCUUUGUACAG
UGUGUUAGAUUUCCCGAGGCAUGGCUUUCCAAGGUACCAGUAGGACUUUAACUCAACAGUCCUCAGCGGCUACGUCUGACGAUCUUCAAAAGAUAUUAUU
UAGCCCUGAAGCCAUUAAGAAAAUGGCUACUGAGUGUGACCUAGGCCGGCAUCAUUGGAUGCGCGCUGAUAAUGCUAUUUCAGUCCGGCCCCUCGUUCCC
GAAGUAACCCACGGUCGUAUUGCUUCCUUCUUUAAAUCUGGAUAUGAUGUUGGUGAAUUGUGCUCAAAAGGAUACAUGAGCGUCCCUCAGGUGUUGUGUG
CUGUUACUCGAACGGUUUCCACCGAUGCUGAAGGGUCUUUGAGAAUUUACUUAGCUGAUCUAGGUGACAAGGAGUUAUCUCCUAUAGAUGGGCAAUGCGU
UUCGUUACAUAACCAUGAUCUUCCCGCUUUGGUGUCUUUCCAACCGACGUAUGACUGUCCUAUGGAAACAGUUGGGAAUCGCAAGCGGUGUUUUGCUGUC
GUUAUCGAAAGACAUGGUUACAUUGGGUAUACCGGUACCACAGCUAGCGUGUGUAGUAAUUGGCAAGCAAGGUUUUCUUCUAAGAAUAACAACUACACUC
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300
310
320
330
340
350
360
370
380
390
400
410
420
430
440
450
460
470
480
490
500
510
520
530
540
550
560
570
580
590
600
610
620
630
640
650
660
670
680
690
700
Nucleotide Position

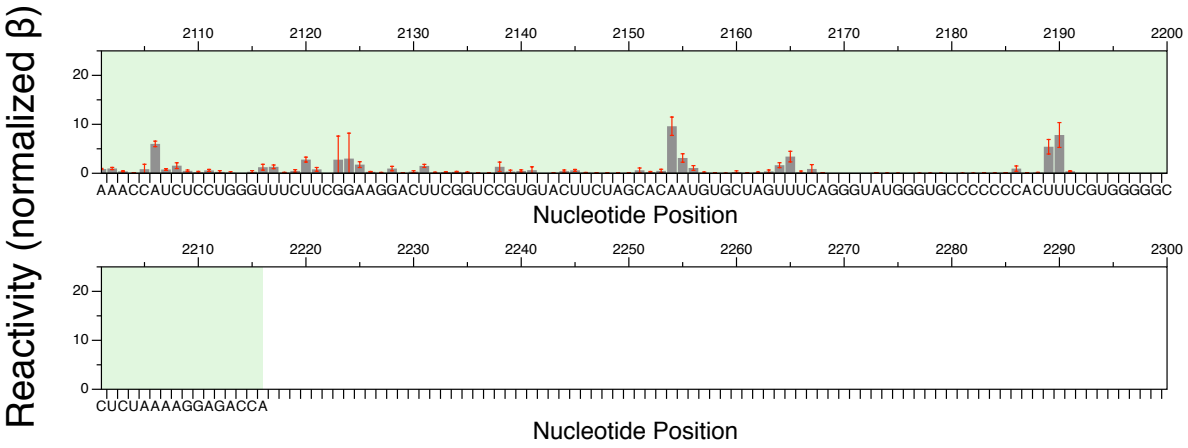
(Figure F.6 cont.)



(Figure F.6 cont.)



(Figure F.6 cont.)



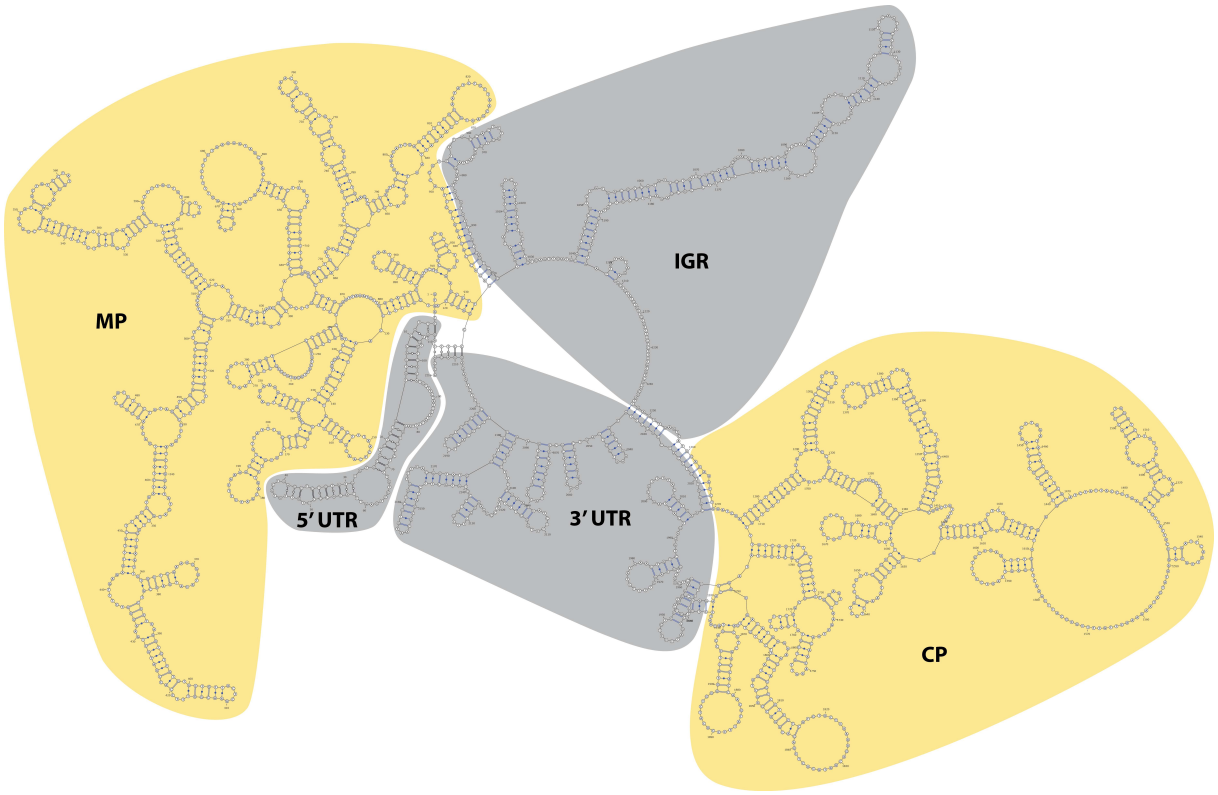


Figure F.7: Predicted secondary structure of CMV RNA3 using *in vitro* transcripts. Secondary structure prediction of the CMV RNA3 using the reactivities obtained from refolded in vitro transcripts to restrain the *Fold* algorithm of RNAstructure [101] (m = 1.1 and b = -0.3). Coding regions are indicated with yellow and non-coding regions with gray as in Figure 8.3. As depicted the 3′ UTR tRNA-like structure is incorrect due to the inability of *Fold* to include pseudoknots in structural predictions.

F.2 Supplementary Tables

Table F.1: Oligonucleotides used in this study. Below is a table of oligonucleotides used during the SHAPE-Seq experiments. Bases indicated by lower case match the sequence of Bn57 RNA3. The Illumina indexing primers (denoted as AC) contain a six nucleotide internal barcode marked as 'xxxxxx'. Any sequencing index is sufficient. Abbreviations within primer sequences are as follows: '/5Phos/' is a 5' monophosphate group, '/3SpC3/' is a 3' 3-carbon spacer group, VIC and NED are fluorophores (ABI), and asterisks indicate a phosphorothioate backbone modification. These abbreviations were used for compatibility with the Integrated DNA Technologies ordering notation.

Name	Sequence	Abbr.
Reverse Transcription		
Binds nts 121-146 of CMV Bn57 RNA3	GGcctactggtaccttggaaagccat	A
Binds nts 423-443	ctcaaagacccttcagcatcg	B
Binds nts 719-747	agctaatctgttgaaaggcagtactagag	C
Binds nts 962-996	ctcacaaatacttatatactaatacgcaccaaagt	D
Binds nts 1264-1288	gaccagcactggttgattcagattt	E
Binds nts 1623-1646	gcaggaactttacggactgtcacc	F
Binds nts 1943-1959	gctcccgccacaggaat	G
Binds nts 2196-2216	tggtctccttttagaggcccc	H
Adapter Ligation		
ssDNA adapter	/5Phos/AGATCGGAAGAGCACACGTCTGAACTC CAGTCAC/3SpC3/	I
Fluorescent Quality Analysis		
Reverse QA primer (+)	VIC-GTGACTGGAGTTCAGACGTGTGCTC	J
Reverse QA primer (-)	NED-GTGACTGGAGTTCAGACGTGTGCTC	K
Primers for Building dsDNA Libraries		
PE_forward	AATGATACGGCGACCACCGAGATCTACACTCTT TCCCTACACGACGCTCTTCCGATCT	L
Selection primer for A (+)	CTTTCCCTACACGACGCTCTTCCGATCTRRRYctca aagacccttcagcatcg*G*T*G*G	M

Table F.1 (Continued)

Name	Sequence	Abbr.
Selection primer for B (+)	CTTTCCCTACACGACGCTCTTCCGATCTRRRYagc taatctgttgaaaggcagtactagag*T*C*T	N
Selection primer for C (+)	CTTTCCCTACACGACGCTCTTCCGATCTRRRYctca caaatacttatatactaatacgcaccaaagt*G*C*T*A	O
Selection primer for D (+)	CTTTCCCTACACGACGCTCTTCCGATCTRRRYgac cagcactggttgattcagattt*G*T*C*C	P
Selection primer for E (+)	CTTTCCCTACACGACGCTCTTCCGATCTRRRYgca ggaactttacggactgtcacc*C*A*C*A	Q
Selection primer for F (+)	CTTTCCCTACACGACGCTCTTCCGATCTRRRYgct cccgccacaggaat*C*G*G*A	R
Selection primer for G (+)	CTTTCCCTACACGACGCTCTTCCGATCTRRRYtgg tctccttttagaggcccc*C*A*C*G	S
Selection primer for H (+)	CTTTCCCTACACGACGCTCTTCCGATCTRRRYagc taatctgttgaaaggcagtactagag*T*C*T	T
Selection primer for A (-)	CTTTCCCTACACGACGCTCTTCCGATCTYYYRctca aagacccttcagcatcg*G*T*G*G	U
Selection primer for B (-)	CTTTCCCTACACGACGCTCTTCCGATCTYYYRagc taatctgttgaaaggcagtactagag*T*C*T	V
Selection primer for C (-)	CTTTCCCTACACGACGCTCTTCCGATCTYYYRctca caaatacttatatactaatacgcaccaaagt*G*C*T*A	W
Selection primer for D (-)	CTTTCCCTACACGACGCTCTTCCGATCTYYYRgac cagcactggttgattcagattt*G*T*C*C	X
Selection primer for E (-)	CTTTCCCTACACGACGCTCTTCCGATCTYYYRgca ggaactttacggactgtcacc*C*A*C*A	Y
Selection primer for F (-)	CTTTCCCTACACGACGCTCTTCCGATCTYYYRgct cccgccacaggaat*C*G*G*A	Z
Selection primer for G (-)	CTTTCCCTACACGACGCTCTTCCGATCTYYYRtgg tctccttttagaggcccc*C*A*C*G	AA
Selection primer for H (-)	CTTTCCCTACACGACGCTCTTCCGATCTYYYRagc taatctgttgaaaggcagtactagag*T*C*T	AB
Illumina Multiplexing Primers		
Illumina Indexing primer†	CAAGCAGAAGACGGCATACGAGATxxxxxxGTG ACTGGAGTTCAGACGTGTGCTC	AC

† Oligonucleotide sequences © 2007-2013 Illumina, Inc. All rights reserved.

APPENDIX G

SUPPLEMENTARY INFORMATION FOR STRUCTURAL FEATURES OF PROTEIN BINDING WITH RNASE P AND ITS SUBSTRATES

G.1 Supplementary Figures

Figure G.1: *Pfu* RPR reactivity maps with and without *Pfu* L7Ae. (A) Secondary structure of *Pfu* RPR. Nucleotides are colored by reactivity intensity using the reactivity values obtained with 10-fold excess *Pfu* L7Ae. (B) Individual reactivity maps for *Pfu* RPR in the absence (red) and presence (black) of 10-fold excess *Pfu* L7Ae. The reactivity maps shown are an average of four or two measurements for without or with L7Ae, respectively, and error bars represent one standard deviation.

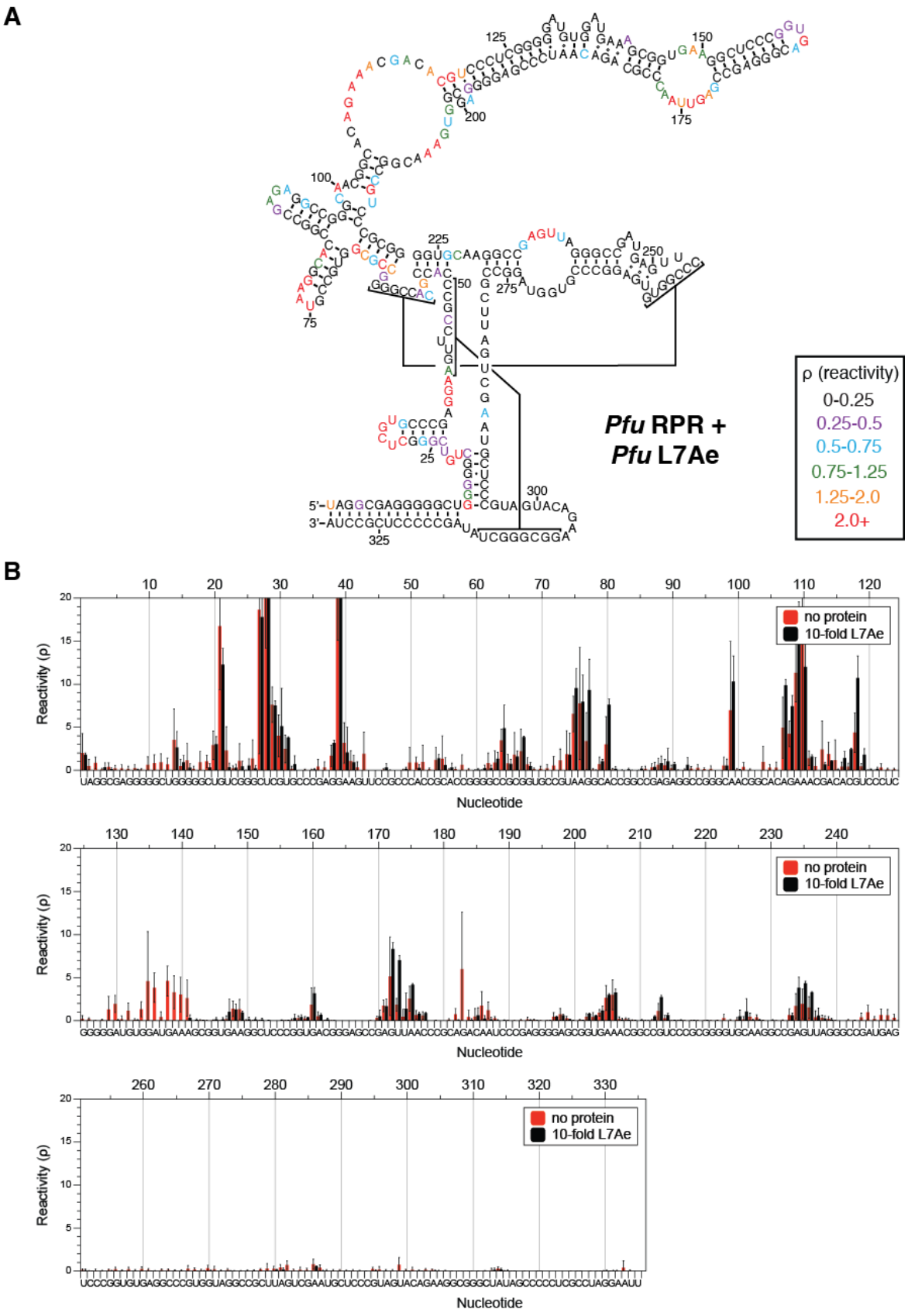
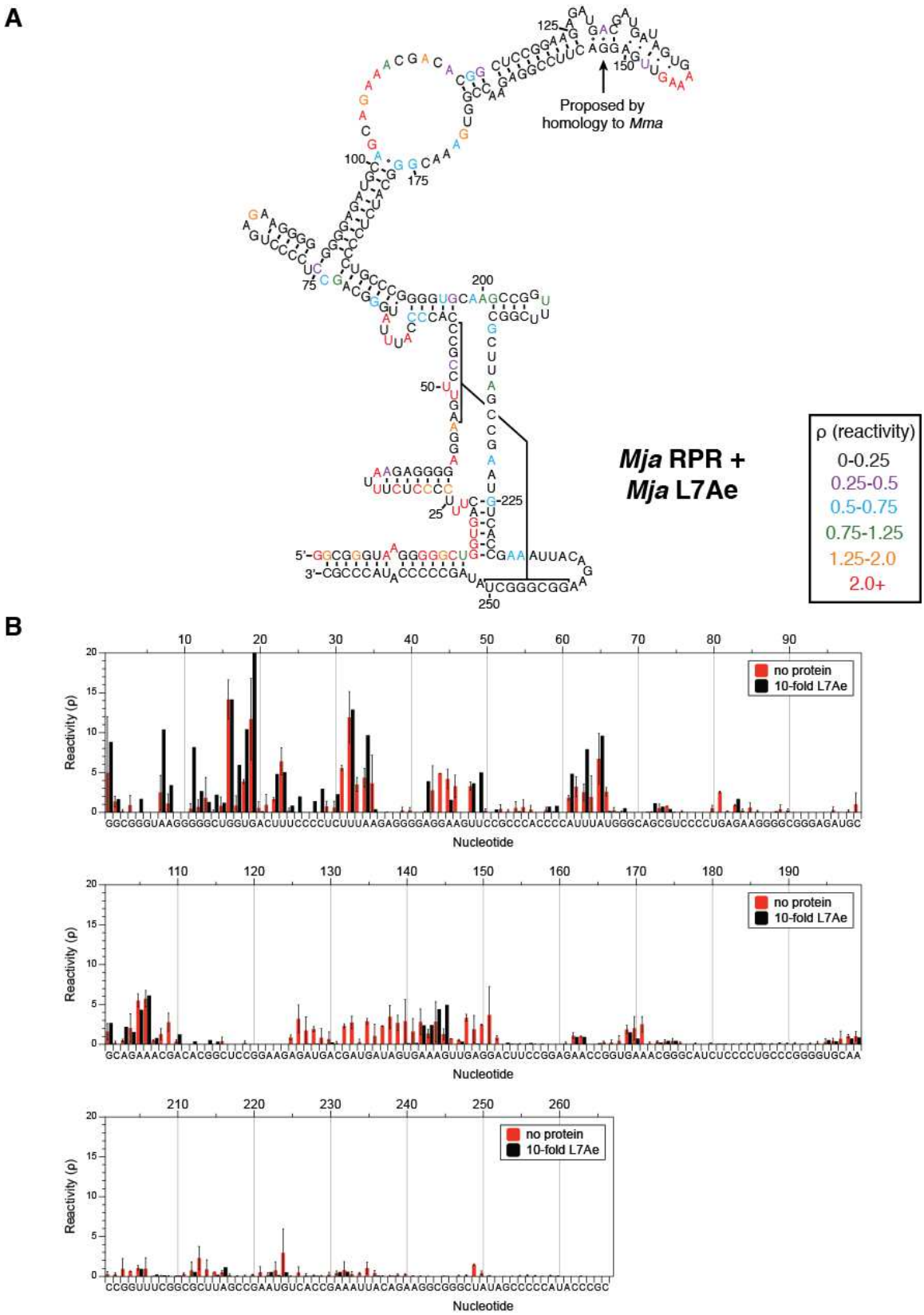

A
Pfu RPR +
Pfu L7Ae
ρ (reactivity)
0-0.25
0.25-0.5
0.5-0.75
0.75-1.25
1.25-2.0
2.0+
B
Reactivity (ρ)
Nucleotide
no protein
10-fold L7Ae

Figure G.2: *Mja* RPR reactivity maps with and without *Mja* L7Ae. (A) Secondary structure of *Mja* RPR. Nucleotides are colored by reactivity intensity using the reactivity values obtained with 10-fold excess *Mja* L7Ae. (B) Individual reactivity maps for *Mja* RPR in the absence (red) and presence (black) of 10-fold excess *Mja* L7Ae. The reactivity map of Mja without L7Ae is an average of two measurements, with error bars that represent one standard deviation.


A
Proposed by homology to Mma
Mja RPR + Mja L7Ae
ρ (reactivity)
0-0.25
0.25-0.5
0.5-0.75
0.75-1.25
1.25-2.0
2.0+
B
no protein
10-fold L7Ae
Reactivity (ρ)
Nucleotide
GGCGGGUAAGGGGGCUGGUGACUUUCCCCUCUUUAAGAGGGGAGGAAGUUCCGCCCACCCCAUUUAUGGGCAGCGUCCCCUGAGAAGGGGCGGGAGAUGC
GCAGAAACGACACGGCUCCGGAAGAGAUGACGAUGAUAGUGAAAGUUGAGGACUUCCGGAGAACCGGUGAAACGGGCAUCUCCCCUGCCCGGGGUGCAA
CCGGUUUCGGCGCUUAGCCGAAUGUCACCGAAAUUACAGAAGGCGGGCUAUAGCCCCCAUACCCGC

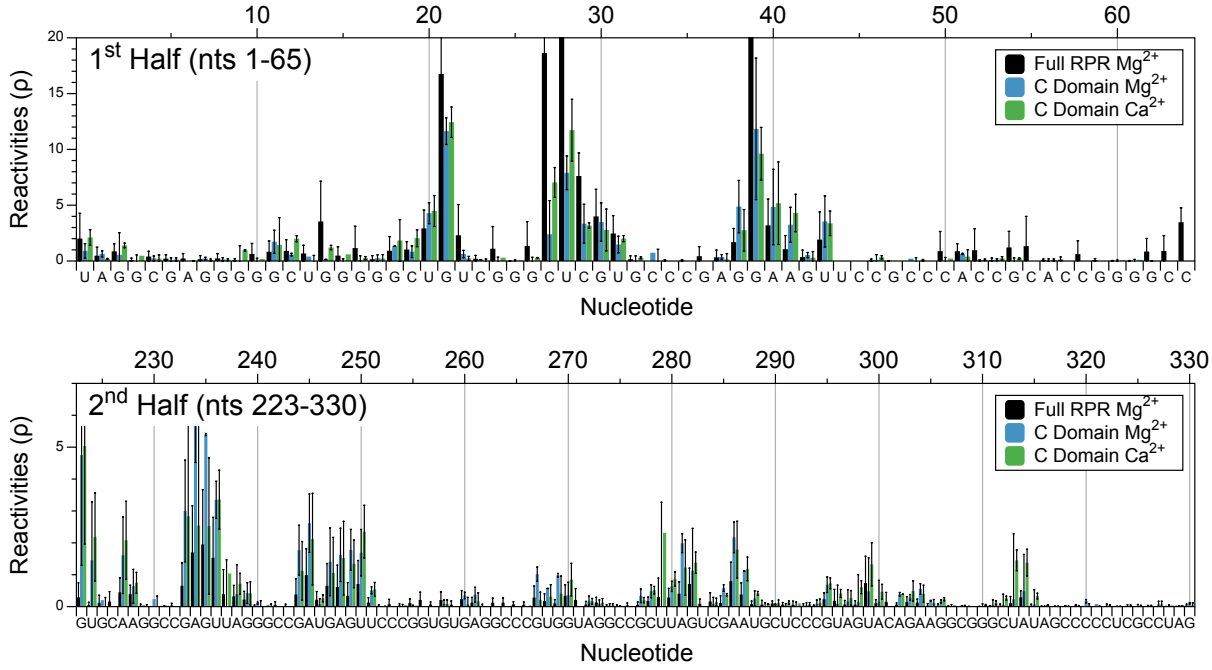


Figure G.3: Reactivity maps of the *Pfu* RPR C domain. Three reactivity maps for the *Pfu* RPR C domain from the full wt RPR refolded in Mg^{2+} containing buffer (black), the C domain only in Mg^{2+} buffer (blue), and the C domain only in Ca^{2+} buffer (green) are plotted side by side. Nucleotides are numbered according to the wt nucleotide positions and divided according to sequence of the wt, which is interrupted by the S domain. The isolated C domain RNA was a direct fusion of nts 65 and 223 (in the wt numbering). Few differences are observed between the reactivity maps of the wt and isolated C domains, regardless of divalent cation, with the exception of nts 223-228 in the isolated C domain that exhibit higher reactivity than the full RPR. Likely, however, this difference is due to the artificial linking of the two halves of the C domain in that region, which is also near a pseudoknot (see Figure 9.1). The reactivity map for the full RPR is an average of four measurements, while the C domains are an average of two measurements. All error bars represent one standard deviation.

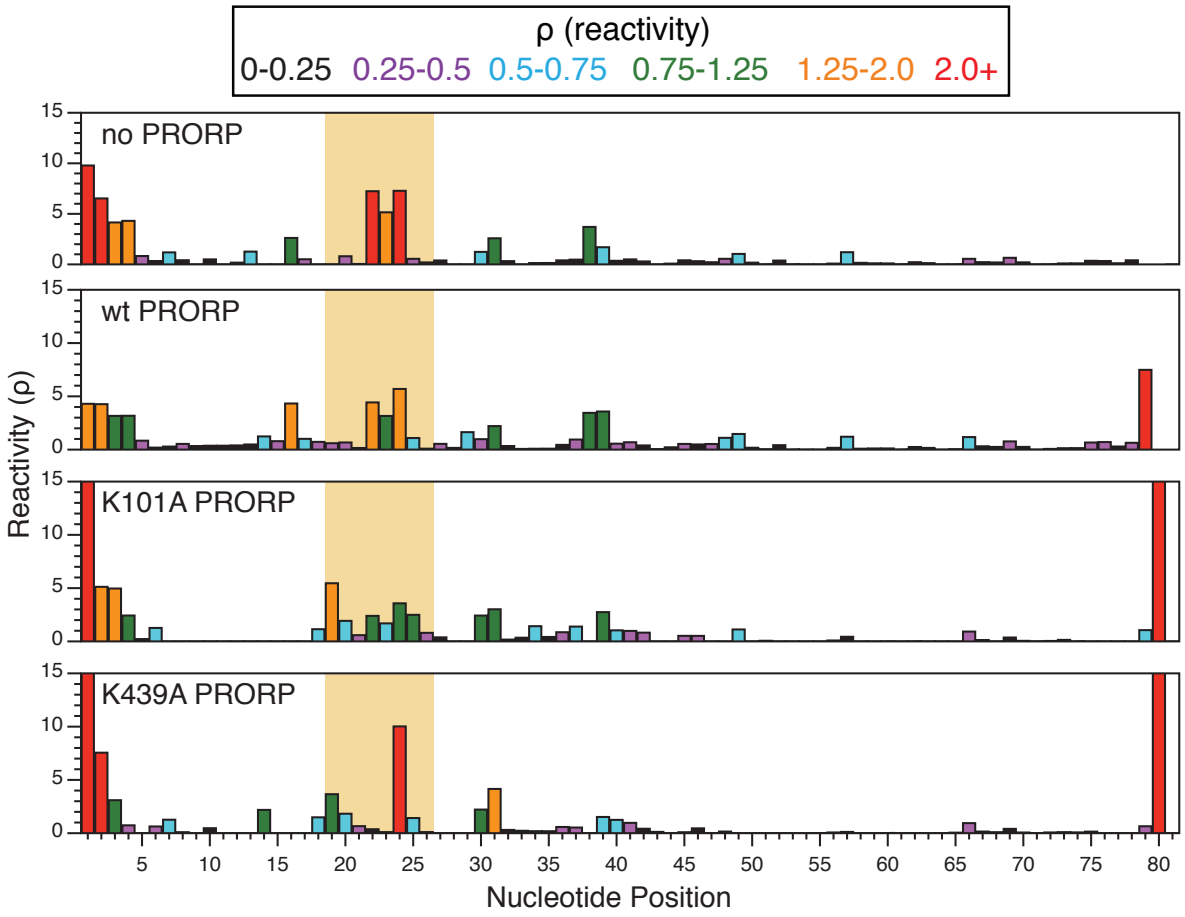


Figure G.4: Individual reactivity maps for pre-tRNA^{Cys} incubated with different PRORPs. D-loop nts are indicated with tan shading. Bars are colored by the indicated reactivity intensity. The addition of any of the three PRORPs causes a reduction in the reactivity of nts G22 and G23. Data represent an average of two replicates for no PRORP and wt PRORP and one replicate each for the mutants.

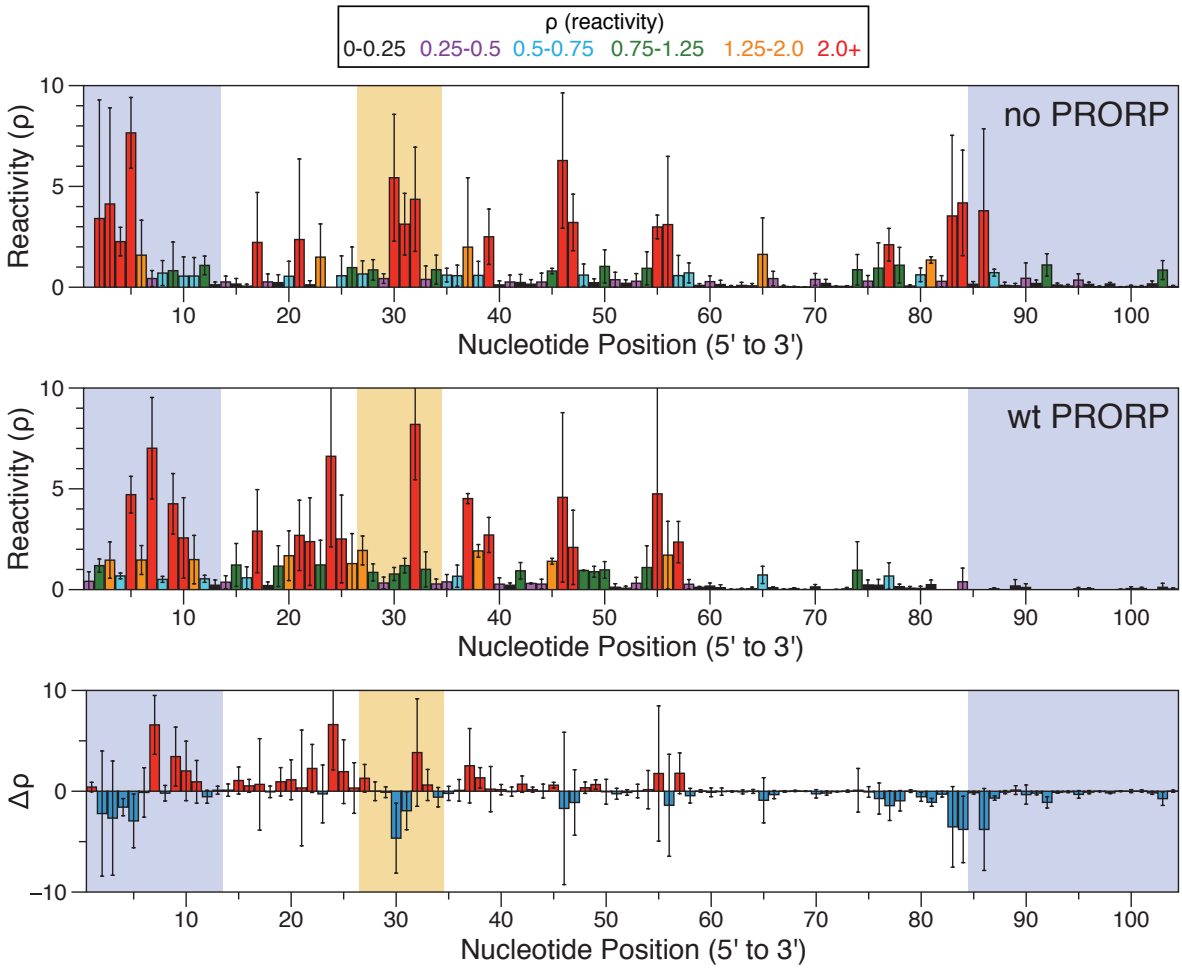


Figure G.5: Replicate data of PRORP binding to pre-tRNA^{Cys} with a 12 nt leader and 23 nt trailer. Shown are three replicate datasets for the longer pre-tRNA^{Cys} in the presence or absence of 20-fold excess PRORP. The D-loop is indicated by tan shading and the leader and trailer are indicated with blue shading. Bars are colored by reactivity intensity and error bars represent one standard deviation. The difference map (bottom) reveals a similar reactivity drops for the G residues in the D-loop as previously observed using a pre-tRNA with a 5 nt leader and no trailer (Figures 9.4 and 9.5). However, the excessive noise in these replicates makes it hard to draw conclusions.

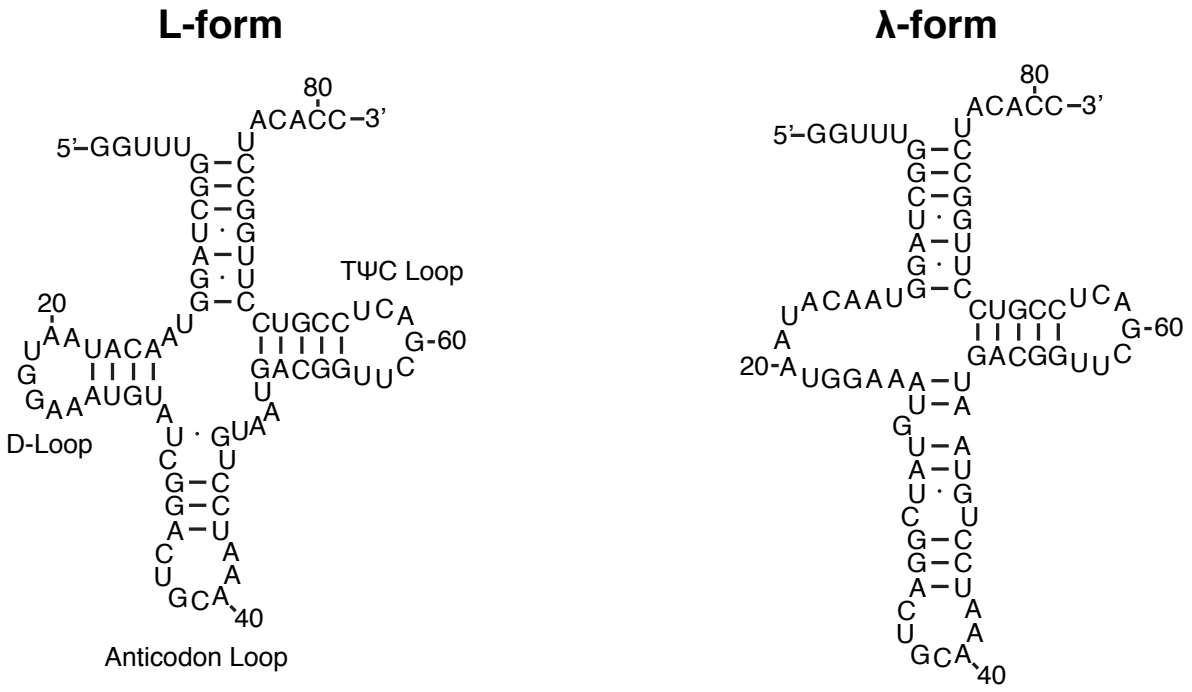


Figure G.6: L- vs. λ-form structures of pre-tRNA^{Cys}. The canonical L-form of tRNA (left) contains a D-loop, TΨC loop, and anticodon loop and folds on itself to form the 'L' shape. In the λ-form the 'core' region is rearranged, disrupting the stem-loop structure that produces the D-loop. The structure shown on the right is a proposed λ-form for pre-tRNA^{Cys}.

BIBLIOGRAPHY

1. Crick, F. H., Barnett, L, Brenner, S & Watts-Tobin, R. J. General nature of the genetic code for proteins. *Nature* **192,** 1227–1232 (1961).

2. Doudna, J. A. & Cech, T. R. The chemical repertoire of natural ribozymes. *Nature* **418,** 222–228 (2002).

3. Flint, S. J., Enquist, L. W., Racaniello, V. R. & Skalka, A. M. *Principles of Virology* (Amer Society for Microbiology, 2008).

4. Marraffini, L. A. CRISPR-Cas immunity in prokaryotes. *Nature* **526,** 55–61 (2015).

5. Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327,** 167–170 (2010).

6. Morris, K. V. & Mattick, J. S. The rise of regulatory RNA. *Nature Reviews Genetics* **15,** 423–437 (2014).

7. Waters, L. S. & Storz, G. Regulatory RNAs in bacteria. *Cell* **136,** 615–628 (2009).

8. Robertson, M. P. & Joyce, G. F. The origins of the RNA world. *Cold Spring Harbor Perspectives in Biology* **4** (2012).

9. Batey, R., Rambo, R. & Doudna, J. Tertiary Motifs in RNA Structure and Folding. *Angewandte Chemie (International ed. in English)* **38,** 2326–2343 (1999).

10. Steitz, T. A. A structural understanding of the dynamic ribosome machine. *Nature Reviews Molecular Cell Biology* **9,** 242–253 (2008).

11. Esakova, O. & Krasilnikov, A. S. Of proteins and RNA: the RNase P/MRP family. *RNA* **16,** 1725–1747 (2010).

12. Peters, J. M., Vangeloff, A. D. & Landick, R. Bacterial transcription terminators: the RNA 3′-end chronicles. *Journal of Molecular Biology* **412,** 793–813 (2011).

13. Lee, Y. & Rio, D. C. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry* **84,** 291–323 (2015).

14. Deutscher, M. P. Degradation of RNA in bacteria: comparison of mRNA and stable RNA. *Nucleic Acids Research* **34,** 659–666 (2006).

15. Garneau, N. L., Wilusz, J. & Wilusz, C. J. The highways and byways of mRNA decay. *Nature Reviews Molecular Cell Biology* **8,** 113–126 (2007).

16. Houseley, J. & Tollervey, D. The many pathways of RNA degradation. *Cell* **136,** 763–776 (2009).

17. Shine, J & Dalgarno, L. Terminal-sequence analysis of bacterial ribosomal RNA. Correlation between the 3'-terminal-polypyrimidine sequence of 16-S RNA and translational specificity of the ribosome. *European journal of biochemistry / FEBS* **57,** 221–230 (1975).

18. Goldman, E. Translation control by RNA. *eLS* (2008).

19. Kozak, M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361,** 13–37 (2005).

20. Marintchev, A. & Wagner, G. Translation initiation: structures, mechanisms and evolution. *Quarterly reviews of biophysics* **37,** 197–284 (2004).

21. Storz, G., Vogel, J. & Wassarman, K. M. Regulation by small RNAs in bacteria: expanding frontiers. *Molecular Cell* **43,** 880–891 (2011).

22. Gottesman, S. & Storz, G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harbor Perspectives in Biology* **3,** 1–16 (2011).

23. Brantl, S. Plasmid Replication Control by Antisense RNAs. *Microbiology spectrum* **2,** PLAS–0001–2013 (2014).

24. Chappell, J., Takahashi, M. K., Meyer, S., Loughrey, D., Watters, K. E. & Lucks, J. The centrality of RNA for engineering gene expression. *Biotechnology journal* **8,** 1379–1395 (2013).

25. Isaacs, F. J., Dwyer, D. J., Ding, C., Pervouchine, D. D., Cantor, C. R. & Collins, J. J. Engineered riboregulators enable post-transcriptional control of gene expression. *Nature Biotechnology* **22,** 841–847 (2004).

26. Chappell, J., Watters, K. E., Takahashi, M. K. & Lucks, J. B. A renaissance in RNA synthetic biology: new mechanisms, applications and tools for the future. *Current Opinion in Chemical Biology* **28,** 47–56 (2015).

27. Montange, R. K. & Batey, R. T. Riboswitches: emerging themes in RNA structure and function. *Annual Review of Biophysics* **37,** 117–133 (2008).

28. Serganov, A. & Nudler, E. A decade of riboswitches. *Cell* **152,** 17–24 (2013).

29. Mandal, M., Boese, B., Barrick, J. E., Winkler, W. C. & Breaker, R. R. Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. *Cell* **113,** 577–586 (2003).

30. Sudarsan, N, Lee, E. R., Weinberg, Z, Moy, R. H., Kim, J. N., Link, K. H. & Breaker, R. R. Riboswitches in eubacteria sense the second messenger cyclic di-GMP. *Science* **321,** 411–413 (2008).

31. Caron, M.-P., Bastet, L., Lussier, A., Simoneau-Roy, M., Massé, E. & Lafontaine, D. A. Dual-acting riboswitch control of translation initiation and mRNA decay. *Proceedings of the National Academy of Sciences of the United States of America* **109,** E3444–53 (2012).

32. Batey, R. T. Structure and mechanism of purine-binding riboswitches. *Quarterly reviews of biophysics* **45,** 345–381 (2012).

33. Nechooshtan, G., Elgrably-Weiss, M., Sheaffer, A., Westhof, E. & Altuvia, S. A pH-responsive riboregulator. *Genes & Development* **23,** 2650–2662 (2009).

34. Groher, F. & Suess, B. Synthetic riboswitches - A tool comes of age. *Biochimica et biophysica acta* **1839,** 964–973 (2014).

35. Wittmann, A. & Suess, B. Engineered riboswitches: Expanding researchers' toolbox with synthetic RNA regulators. *FEBS Letters* **586,** 2076–2083 (2012).

36. Ausländer, S., Ketzer, P. & Hartig, J. S. A ligand-dependent hammerhead ribozyme switch for controlling mammalian gene expression. *Molecular BioSystems* **6,** 807–814 (2010).

37. Rodrigo, G., Landrain, T. E. & Jaramillo, A. De novo automated design of small RNA circuits for engineering synthetic riboregulation in living cells. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 15271–15276 (2012).

38. Mishler, D. M. & Gallivan, J. P. A family of synthetic riboswitches adopts a kinetic trapping mechanism. *Nucleic Acids Research* **42,** 6753–6761 (2014).

39. Topp, S. & Gallivan, J. P. Guiding Bacteria with Small Molecules and RNA. *Journal of the American Chemical Society* **129,** 6807–6811 (2007).

40. Suess, B., Fink, B., Berens, C., Stentz, R. & Hillen, W. A theophylline responsive riboswitch based on helix slipping controls gene expression in vivo. *Nucleic Acids Research* **32,** 1610–1614 (2004).

41. Topp, S. & Gallivan, J. P. Riboswitches in unexpected places–a synthetic riboswitch in a protein coding region. *RNA* **14,** 2498–2503 (2008).

42. Lemay, J.-F., Desnoyers, G., Blouin, S., Heppell, B., Bastet, L., St-Pierre, P., Massé, E. & Lafontaine, D. A. Comparative Study between Transcriptionally- and Translationally-Acting Adenine Riboswitches Reveals Key Differences in Riboswitch Regulatory Mechanisms. *PLoS genetics* **7,** e1001278 (2011).

43. Garst, A. D., Edwards, A. L. & Batey, R. T. Riboswitches: Structures and Mechanisms. *Cold Spring Harbor Perspectives in Biology* **3,** a003533–a003533 (2011).

44. Rieder, U., Kreutz, C. & Micura, R. Folding of a transcriptionally acting preQ1 riboswitch. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 10804–10809 (2010).

45. Sudarsan, N., Wickiser, J. K., Nakamura, S., Ebert, M. S. & Breaker, R. R. An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes & Development* **17,** 2688–2697 (2003).

46. Furukawa, K., Ramesh, A., Zhou, Z., Weinberg, Z., Vallery, T., Winkler, W. C. & Breaker, R. R. Bacterial riboswitches cooperatively bind Ni(2+) or Co(2+) ions and control expression of heavy metal transporters. *Molecular Cell* **57,** 1088–1098 (2015).

47. Winkler, W. C., Cohen-Chalamish, S. & Breaker, R. R. An mRNA structure that controls gene expression by binding FMN. *Proceedings of the National Academy of Sciences* **99,** 15908–15913 (2002).

48. Wachsmuth, M., Findeiß, S., Weissheimer, N., Stadler, P. F. & Mörl, M. De novo design of a synthetic riboswitch that regulates transcription termination. *Nucleic Acids Research* **41,** 2541–2551 (2013).

49. Wickiser, J. K. Kinetics of riboswitch regulation studied by in vitro transcription. *Methods in molecular biology* **540,** 53–63 (2009).

50. Baker, J. L., Sudarsan, N., Weinberg, Z., Roth, A., Stockbridge, R. B. & Breaker, R. R. Widespread genetic switches and toxicity resistance proteins for fluoride. *Science* **335,** 233–235 (2012).

51. Brantl, S. & Wagner, E. G. H. An Antisense RNA-Mediated Transcriptional Attenuation Mechanism Functions in Escherichia coli. *Journal of Bacteriology* **184,** 2740–2747 (2002).

52. Novick, R. P., Adler, G. K., Projan, S. J., Carleton, S, Highlander, S. K., Gruss, A, Khan, S. A. & Iordanescu, S. Control of pT181 replication I. The pT181 copy control function acts by inhibiting the synthesis of a replication protein. *The EMBO journal* **3,** 2399–2405 (1984).

53. Novick, R. P., Iordanescu, S, Projan, S. J., Kornblum, J & Edelman, I. pT181 plasmid replication is regulated by a countertranscript-driven transcriptional attenuator. *Cell* **59,** 395–404 (1989).

54. Kumar, C. C. & Novick, R. P. Plasmid pT181 replication is regulated by two countertranscripts. *Proceedings of the National Academy of Sciences* **82,** 638–642 (1985).

55. Chappell, J., Takahashi, M. K. & Lucks, J. B. Creating small transcription activating RNAs. *Nature Chemical Biology* **11,** 214–220 (2015).

56. Larson, M. H., Gilbert, L. A., Wang, X., Lim, W. A., Weissman, J. S. & Qi, L. S. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature Protocols* **8,** 2180–2196 (2013).

57. Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P. & Lim, W. A. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* **152,** 1173–1183 (2013).

58. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology* **32,** 347–355 (2014).

59. Terns, R. M. & Terns, M. P. CRISPR-based technologies: prokaryotic defense weapons repurposed. *Trends in genetics : TIG* **30,** 111–118 (2014).

60. Knight, S. C., Xie, L., Deng, W., Guglielmi, B., Witkowsky, L. B., Bosanac, L., Zhang, E. T., El Beheiry, M., Masson, J.-B., Dahan, M., Liu, Z., Doudna, J. A. & Tjian, R. Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science* **350,** 823–826 (2015).

61. Al-Hashimi, H. M. & Walter, N. G. RNA dynamics: it is about time. *Current Opinion in Structural Biology* **18,** 321–329 (2008).

62. Pan, T. & Sosnick, T. RNA folding during transcription. *Annual Review of Biophysics and Biomolecular Structure* **35,** 161–175 (2006).

63. Garst, A. D. & Batey, R. T. A switch in time: detailing the life of a riboswitch. *Biochimica et biophysica acta* **1789,** 584–591 (2009).

64. Zhao, P., Zhang, W. & Chen, S.-J. Cotranscriptional folding kinetics of ribonucleic acid secondary structures. *The Journal of chemical physics* **135,** 245101 (2011).

65. Lai, D., Proctor, J. R. & Meyer, I. M. On the importance of cotranscriptional RNA structure formation. *RNA* **19,** 1461–1473 (2013).

66. Heilman-Miller, S. L. & Woodson, S. A. Effect of transcription on folding of the Tetrahymena ribozyme. *RNA* **9,** 722–733 (2003).

67. Nechooshtan, G., Elgrably-Weiss, M. & Altuvia, S. Changes in transcriptional pausing modify the folding dynamics of the pH-responsive RNA element. *Nucleic Acids Research* **42,** 622–630 (2014).

68. Pan, T, Artsimovitch, I, Fang, X. W., Landick, R & Sosnick, T. R. Folding of a large ribozyme during transcription and the effect of the elongation factor NusA. *Proceedings of the National Academy of Sciences* **96,** 9545–9550 (1999).

69. Sosnick, T. R. & Pan, T. RNA folding: models and perspectives. *Current Opinion in Structural Biology* **13,** 309–316 (2003).

70. Treiber, D. K. & Williamson, J. R. Beyond kinetic traps in RNA folding. *Current Opinion in Structural Biology* **11,** 309–314 (2001).

71. Pan, T, Fang, X & Sosnick, T. Pathway modulation, circular permutation and rapid RNA folding under kinetic control. *Journal of Molecular Biology* **286,** 721–731 (1999).

72. Wong, T., Sosnick, T. R. & Pan, T. Mechanistic insights on the folding of a large ribozyme during transcription. *Biochemistry* **44,** 7535–7542 (2005).

73. Chadalavada, D. M., Senchak, S. E. & Bevilacqua, P. C. The folding pathway of the genomic hepatitis delta virus ribozyme is dominated by slow folding of the pseudoknots. *Journal of Molecular Biology* **317,** 559–575 (2002).

74. Xayaphoummine, A, Bucher, T & Isambert, H. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Research* **33,** W605–10 (2005).

75. Proctor, J. R. & Meyer, I. M. COFOLD: an RNA secondary structure prediction method that takes co-transcriptional folding into account. *Nucleic Acids Research* **41,** e102–e102 (2013).

76. Meyer, I. M. & Miklós, I. Co-transcriptional folding is encoded within RNA genes. *BMC Molecular Biology* **5,** 10 (2004).

77. Dykeman, E. C. An implementation of the Gillespie algorithm for RNA kinetics with logarithmic time update. *Nucleic Acids Research* **43,** 5708–5715 (2015).

78. Wong, T. N. & Pan, T. RNA folding during transcription: protocols and studies. *Methods in enzymology* **468,** 167–193 (2009).

79. Frieda, K. L. & Block, S. M. Direct Observation of Cotranscriptional Folding in an Adenine Riboswitch. *Science* **338,** 397–400 (2012).

80. Greenleaf, W. J., Frieda, K. L., Foster, D. A. N., Woodside, M. T. & Block, S. M. Direct observation of hierarchical folding in single riboswitch aptamers. *Science* **319,** 630–633 (2008).

81. Neupane, K., Foster, D. A. N., Dee, D. R., Yu, H., Wang, F. & Woodside, M. T. Direct observation of transition paths during the folding of proteins and nucleic acids. *Science* **352,** 239–242 (2016).

82. Perdrizet, G. A., Artsimovitch, I., Furman, R., Sosnick, T. R. & Pan, T. Transcriptional pausing coordinates folding of the aptamer domain and the expression platform of a riboswitch. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 3323–3328 (2012).

83. Wong, T. N., Sosnick, T. R. & Pan, T. Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 17995–18000 (2007).

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

85. Latham, M. P., Brown, D. J., McCallum, S. A. & Pardi, A. NMR methods for studying the structure and dynamics of RNA. *Chembiochem : a European journal of chemical biology* **6,** 1492–1505 (2005).

86. Weeks, K. M. Advances in RNA structure analysis by chemical probing. *Current Opinion in Structural Biology* **20,** 295–304 (2010).

87. Knapp, G. Enzymatic approaches to probing of RNA secondary and tertiary structure. *Methods in enzymology* **180,** 192–212 (1989).

88. Kubota, M., Tran, C. & Spitale, R. C. Progress and challenges for chemical probing of RNA structure inside living cells. *Nature Chemical Biology* **11,** 933–941 (2015).

89. Butcher, S. E. & Pyle, A. M. The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Accounts of Chemical Research* **44,** 1302–1311 (2011).

90. McGinnis, J. L., Dunkle, J. A., Cate, J. H. D. & Weeks, K. M. The Mechanisms of RNA SHAPE Chemistry. *Journal of the American Chemical Society* **134,** 6617–6624 (2012).

91. Peattie, D. A. & Gilbert, W. Chemical probes for higher-order structure in RNA. *Proceedings of the National Academy of Sciences* **77,** 4679–4682 (1980).

92. Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *Journal of the American Chemical Society* **127,** 4223–4231 (2005).

93. Culver, G. M. & Noller, H. F. Directed hydroxyl radical probing of 16S ribosomal RNA in ribosomes containing Fe(II) tethered to ribosomal protein S20. *RNA* **4,** 1471–1480 (1998).

94. Brunel, C & Romby, P. Probing RNA structure and RNA-ligand complexes with chemical probes. *Methods in enzymology* **318,** 3–21 (2000).

95. Spitale, R. C., Flynn, R. A., Torre, E. A., Kool, E. T. & Chang, H. Y. RNA structural analysis by evolving SHAPE chemistry. *Wiley interdisciplinary reviews. RNA* **5,** 867–881 (2014).

96. Wilkinson, K. A., Vasa, S. M., Deigan, K. E., Mortimer, S. A., Giddings, M. C. & Weeks, K. M. Influence of nucleotide identity on ribose 2'-hydroxyl reactivity in RNA. *RNA* **15,** 1314–1321 (2009).

97. Aviran, S., Trapnell, C., Lucks, J. B., Mortimer, S. A., Luo, S., Schroth, G. P., Doudna, J. A., Arkin, A. P. & Pachter, L. Modeling and automation of sequencing-based characterization of RNA structure. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 11069–11074 (2011).

98. Aviran, S., Lucks, J. B. & Pachter, L. RNA structure characterization from chemical mapping experiments. *49th Annual Allerton Conference,* 1743–1750 (2011).

99. Wilkinson, K. A., Merino, E. J. & Weeks, K. M. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature Protocols* **1,** 1610–1616 (2006).

100. Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M. & Turner, D. H. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences* **101,** 7287–7292 (2004).

101. Low, J. T. & Weeks, K. M. SHAPE-directed RNA secondary structure prediction. *Methods* **52,** 150–158 (2010).

102. Hajdin, C. E., Bellaousov, S., Huggins, W., Leonard, C. W., Mathews, D. H. & Weeks, K. M. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 5498–5503 (2013).

103. Byrne, R. T., Konevega, A. L., Rodnina, M. V. & Antson, A. A. The crystal structure of unmodified tRNAPhe from Escherichia coli. *Nucleic Acids Research* **38,** 4154–4162 (2010).

104. Mortimer, S. A. & Weeks, K. M. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *Journal of the American Chemical Society* **129,** 4144–4145 (2007).

105. Vasa, S. M., Guex, N., Wilkinson, K. A., Weeks, K. M. & Giddings, M. C. ShapeFinder: a software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA* **14,** 1979–1990 (2008).

106. Karabiber, F., McGinnis, J. L., Favorov, O. V. & Weeks, K. M. QuShape: rapid, accurate, and best-practices quantification of nucleic acid probing information, resolved by capillary electrophoresis. *RNA* **19,** 63–73 (2013).

107. Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A. & Arkin, A. P. Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences of the United States of America* **108,** 11063–11068 (2011).

108. Mortimer, S. A., Trapnell, C., Aviran, S., Pachter, L. & Lucks, J. B. SHAPE-Seq: High-Throughput RNA Structure Analysis. *Current protocols in chemical biology* **4,** 275–297 (2012).

109. Strobel, E. J., Watters, K. E., Loughrey, D. & Lucks, J. B. RNA systems biology: uniting functional discoveries and structural tools to understand global roles of RNAs. *Current Opinion in Biotechnology* **39,** 182–191 (2016).

110. Kwok, C. K., Tang, Y., Assmann, S. M. & Bevilacqua, P. C. The RNA structurome:transcriptome-wide structure probingwith next-generation sequencing. *Trends in Biochemical Sciences* **40,** 221–232 (2015).

111. Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics* **15,** 469–479 (2014).

112. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature Methods* **11,** 959–965 (2014).

113. Watters, K. E., Abbott, T. R. & Lucks, J. B. Simultaneous characterization of cellular RNA structure and function with in-cell SHAPE-Seq. *Nucleic Acids Research* **44,** e12–e12 (2016).

114. Incarnato, D., Neri, F., Anselmi, F. & Oliviero, S. Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. *Genome biology* **15,** 491 (2014).

115. Spitale, R. C., Flynn, R. A., Zhang, Q. C., Crisalli, P., Lee, B., Jung, J.-W., Kuchelmeister, H. Y., Batista, P. J., Torre, E. A., Kool, E. T. & Chang, H. Y. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519,** 486–490 (2015).

116. Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C. & Assmann, S. M. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505,** 696–700 (2014).

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

118. Talkish, J., May, G., Lin, Y., Woolford, J. L. & McManus, C. J. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA* **20,** 713–720 (2014).

119. Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y. & Segal, E. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467,** 103–107 (2010).

120. Underwood, J. G., Uzilov, A. V., Katzman, S., Onodera, C. S., Mainzer, J. E., Mathews, D. H., Lowe, T. M., Salama, S. R. & Haussler, D. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature Methods* **7,** 995–1001 (2010).

121. Ramani, V., Qiu, R. & Shendure, J. High-throughput determination of RNA structure by proximity ligation. *Nature Biotechnology* **33,** 980–984 (2015).

122. Loughrey, D., Watters, K. E., Settle, A. H. & Lucks, J. B. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Research* **42** (2014).

123. Watters, K. E., Yu, A. M., Strobel, E. J., Settle, A. H. & Lucks, J. B. Characterizing RNA structures in vitro and in vivo with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Methods* **103,** 34–48 (2016).

124. Wildauer, M., Zemora, G., Liebeg, A., Heisig, V. & Waldsich, C. Chemical probing of RNA in living cells. *Methods in molecular biology* **1086,** 159–176 (2014).

125. Wells, S. E., Hughes, J. M., Igel, A. H. & Ares, M. Use of dimethyl sulfate to probe RNA structure in vivo. *Methods in enzymology* **318,** 479–493 (2000).

126. Tyrrell, J., McGinnis, J. L., Weeks, K. M. & Pielak, G. J. The cellular environment stabilizes adenine riboswitch RNA structure. *Biochemistry* **52,** 8777–8785 (2013).

127. Wiedmann, M, Wilson, W. J., Czajka, J, Luo, J, Barany, F & Batt, C. A. Ligase chain reaction (LCR)–overview and applications. *PCR methods and applications* **3,** S51–64 (1994).

128. Xerri, L, Parc, P, Bouabdallah, R, Camerlo, J & Hassoun, J. PCR-mismatch analysis of p53 gene mutation in Hodgkin's disease. *The Journal of pathology* **175,** 189–194 (1995).

129. Takahashi, M. K., Watters, K. E., Gasper, P. M., Abbott, T. R., Carlson, P. D., Chen, A. A. & Lucks, J. B. Using in-cell SHAPE-Seq and simulations to probe structure-function design principles of RNA transcriptional regulators. *RNA* **22,** 920–933 (2016).

130. Wickiser, J. K., Winkler, W. C., Breaker, R. R. & Crothers, D. M. The Speed of RNA Transcription and Metabolite Binding Kinetics Operate an FMN Riboswitch. *Molecular Cell* **18,** 49–60 (2005).

131. Pavco, P. A. & Steege, D. A. Elongation by Escherichia coli RNA polymerase is blocked in vitro by a site-specific DNA binding protein. *The Journal of biological chemistry* **265,** 9960–9969 (1990).

132. Khan, S. A., Murray, R. W. & Koepsel, R. R. Mechanism of plasmid pT181 DNA replication. *Biochimica et biophysica acta* **951,** 375–381 (1988).

133. Brantl, S. & Wagner, E. Antisense RNA-mediated transcriptional attenuation: an in vitro study of plasmid pT181. *Molecular microbiology* **35,** 1469–1482 (2000).

134. Highlander, S. K. & Novick, R. P. Mutational and physiological analyses of plasmid pT181 functions expressing incompatibility. *Plasmid* **23,** 1–15 (1990).

135. Lucks, J. B., Qi, L., Mutalik, V. K., Wang, D. & Arkin, A. P. Versatile RNA-sensing transcriptional regulators for engineering genetic networks. *Proceedings of the National Academy of Sciences* **108,** 8617–8622 (2011).

136. Takahashi, M. K. & Lucks, J. B. A modular strategy for engineering orthogonal chimeric RNA transcription regulators. *Nucleic Acids Research* **41,** 7577–7588 (2013).

137. Qi, L., Lucks, J. B., Liu, C. C., Mutalik, V. K. & Arkin, A. P. Engineering naturally occurring trans-acting non-coding RNAs to sense molecular signals. *Nucleic Acids Research* **40,** 5775–5786 (2012).

138. Takahashi, M. K., Chappell, J., Hayes, C. A., Sun, Z. Z., Kim, J., Singhal, V., Spring, K. J., Al-Khabouri, S., Fall, C. P., Noireaux, V., Murray, R. M. & Lucks, J. B. Rapidly characterizing the fast dynamics of RNA genetic circuitry with cell-free transcription-translation (TX-TL) systems. *ACS Synthetic Biology* **4,** 503–515 (2015).

139. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. & Charpentier, E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821 (2012).

140. Nielsen, A. A. K. & Voigt, C. A. Multi-input CRISPR/Cas genetic circuits that interface host regulatory networks. *Molecular Systems Biology* **10,** 763 (2014).

141. Mehta, P., Goyal, S. & Wingreen, N. S. A quantitative comparison of sRNA-based and protein-based gene regulation. *Molecular Systems Biology* **4,** 221 (2008).

142. Shimoni, Y., Friedlander, G., Hetzroni, G., Niv, G., Altuvia, S., Biham, O. & Margalit, H. Regulation of gene expression by small non-coding RNAs: a quantitative view. *Molecular Systems Biology* **3,** 138 (2007).

143. Edwardson, J. R. & Christie, R. G. in *CRC Handbook of Viruses Infecting Legumes* 293–309 (1991).

144. Palukaitis, P. & García-Arenal, F. Cucumoviruses. *Advances in virus research* **62,** 241–323 (2003).

145. Phizicky, E. M. & Hopper, A. K. tRNA biology charges to the front. *Genes & Development* **24,** 1832–1860 (2010).

146. Altman, S, Baer, M. F., Bartkiewicz, M, Gold, H, Guerrier-Takada, C, Kirsebom, L. A., Lumelsky, N & Peck, K. Catalysis by the RNA subunit of RNase P–a minireview. *Gene* **82,** 63–64 (1989).

147. Evans, D., Marquez, S. M. & Pace, N. R. RNase P: interface of the RNA and protein worlds. *Trends in Biochemical Sciences* **31,** 333–341 (2006).

148. Randau, L., Schröder, I. & Söll, D. Life without RNase P. *Nature* **453,** 120–123 (2008).

149. Liu, F. & Altman, S. *Ribonuclease P* (Springer Science & Business Media, 2009).

150. Gobert, A., Gutmann, B., Taschner, A., Gößringer, M., Holzmann, J., Hartmann, R. K., Rossmanith, W. & Giegé, P. A single Arabidopsis organellar protein has RNase P activity. *Nature Structural & Molecular Biology* **17,** 740–744 (2010).

151. Goldfarb, K. C., Borah, S & Cech, T. R. RNase P branches out from RNP to protein: organelle-triggered diversification? *Genes & Development* **26,** 1005–1009 (2012).

152. Lai, S. M., Lai, L. B., Foster, M. P. & Gopalan, V. The L7Ae protein binds to two kink-turns in the Pyrococcus furiosus RNase P RNA. *Nucleic Acids Research* **42,** 13328–13338 (2014).

153. Cho, I.-M., Lai, L. B., Susanti, D., Mukhopadhyay, B. & Gopalan, V. Ribosomal protein L7Ae is a subunit of archaeal RNase P. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 14573–14578 (2010).

154. Sharp, P. A. The centrality of RNA. *Cell* **136,** 577–580 (2009).

155. Wan, Y., Qu, K., Zhang, Q. C., Flynn, R. A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R. C., Snyder, M. P., Segal, E. & Chang, H. Y. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505,** 706–709 (2014).

156. Seetin, M. G., Kladwang, W., Bida, J. P. & Das, R. Massively parallel RNA chemical mapping with a reduced bias MAP-seq protocol. *Methods in molecular biology* **1086,** 95–117 (2014).

157. Wan, Y., Kertesz, M., Spitale, R. C., Segal, E. & Chang, H. Y. Understanding the transcriptome through RNA structure. *Nature Reviews Genetics* **12,** 641–655 (2011).

158. Spitale, R. C., Crisalli, P., Flynn, R. A., Torre, E. A., Kool, E. T. & Chang, H. Y. RNA SHAPE analysis in living cells. *Nature Chemical Biology* **9,** 18–20 (2012).

159. Yoon, S., Kim, J., Hum, J., Kim, H., Park, S., Kladwang, W. & Das, R. HiTRACE: high-throughput robust analysis for capillary electrophoresis. *Bioinformatics* **27,** 1798–1805 (2011).

160. Kladwang, W., VanLang, C. C., Cordero, P. & Das, R. Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry* **50,** 8049–8056 (2011).

161. Leonard, C. W., Hajdin, C. E., Karabiber, F., Mathews, D. H., Favorov, O. V., Dokholyan, N. V. & Weeks, K. M. Principles for understanding the accuracy of SHAPE-directed RNA structure modeling. *Biochemistry* **52,** 588–595 (2013).

162. Avis, J. M., Conn, G. L. & Walker, S. C. Cis-acting ribozymes for the production of RNA in vitro transcripts with defined 5′ and 3′ ends. *Methods in molecular biology* **941,** 83–98 (2012).

163. Deigan, K., Li, T., Mathews, D. & Weeks, K. Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences* **106,** 97–102 (2009).

164. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics* **11,** 129 (2010).

165. Kwok, C. K., Ding, Y., Sherlock, M. E., Assmann, S. M. & Bevilacqua, P. C. Analytical Biochemistry. *Analytical biochemistry* **435,** 181–186 (2013).

166. Kwok, C. K., Ding, Y., Tang, Y., Assmann, S. M. & Bevilacqua, P. C. Determination of in vivo RNA structure in low-abundance transcripts. *Nature Communications* **4,** 2971 (2013).

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

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

169. Cordero, P., Lucks, J. B. & Das, R. An RNA Mapping DataBase for curating RNA structure mapping experiments. *Bioinformatics* **28,** 3006–3008 (2012).

170. Seetin, M. G. & Mathews, D. H. RNA structure prediction: an overview of methods. *Methods in molecular biology* **905,** 99–122 (2012).

171. Gottesman, S. Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends in genetics : TIG* **21,** 399–404 (2005).

172. Gottesman, S. The small RNA regulators of Escherichia coli: roles and mechanisms. *Annual review of microbiology* **58,** 303–328 (2004).

173. Costa, F. F. Non-coding RNAs: Meet thy masters. *BioEssays* **32,** 599–608 (2010).

174. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annual Review of Biochemistry* **81,** 145–166 (2012).

175. ENCODE Project Consortium *et al.* Identification and analysis of functional elements in 1the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).

176. Mutalik, V. K., Qi, L., Guimaraes, J. C., Lucks, J. B. & Arkin, A. P. Rationally designed families of orthogonal RNA regulators of translation. *Nature Chemical Biology* **8,** 447–454 (2012).

177. Green, A. A., Silver, P. A., Collins, J. J. & Yin, P. Toehold switches: de-novo-designed regulators of gene expression. *Cell* **159,** 925–939 (2014).

178. Rana, T. M. Illuminating the silence: understanding the structure and function of small RNAs. *Nature Reviews Molecular Cell Biology* **8,** 23–36 (2007).

179. Cheng, A. A. & Lu, T. K. Synthetic biology: an emerging engineering discipline. *Annual review of biomedical engineering* **14,** 155–178 (2012).

180. Bindewald, E., Wendeler, M., Legiewicz, M., Bona, M. K., Wang, Y., Pritt, M. J., Le Grice, S. F. J. & Shapiro, B. A. Correlating SHAPE signatures with three-dimensional RNA structures. *RNA* **17,** 1688–1696 (2011).

181. Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A. & Smith, H. O. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* **6,** 343–345 (2009).

182. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10,** 1–10 (2009).

183. Kolter, R & Yanofsky, C. Attenuation in amino acid biosynthetic operons. *Annual review of genetics* **16,** 113–134 (1982).

184. Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T. C. & Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nature Biotechnology* **24,** 79–88 (2006).

185. McGinnis, J. L. & Weeks, K. M. Ribosome RNA assembly intermediates visualized in living cells. *Biochemistry* **53,** 3237–3247 (2014).

186. Chen, Y.-J., Liu, P., Nielsen, A. A. K., Brophy, J. A. N., Clancy, K., Peterson, T. & Voigt, C. A. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nature Methods* **10,** 659–664 (2013).

187. Franch, T, Petersen, M, Wagner, E. G., Jacobsen, J. P. & Gerdes, K. Antisense RNA regulation in prokaryotes: rapid RNA/RNA interaction facilitated by a general U-turn loop structure. *Journal of Molecular Biology* **294,** 1115–1125 (1999).

188. Franch, T & Gerdes, K. U-turns and regulatory RNAs. *Current opinion in microbiology* **3,** 159–164 (2000).

189. Kittle, J. D., Simons, R. W., Lee, J & Kleckner, N. Insertion sequence IS10 anti-sense pairing initiates by an interaction between the 5′ end of the target RNA and a loop in the anti-sense RNA. *Journal of Molecular Biology* **210,** 561–572 (1989).

190. Ross, J. A., Ellis, M. J., Hossain, S. & Haniford, D. B. Hfq restructures RNA-IN and RNA-OUT and facilitates antisense pairing in the Tn10/IS10 system. *RNA* **19,** 670–684 (2013).

191. Case, C. C., Simons, E. L. & Simons, R. W. The IS10 transposase mRNA is destabilized during antisense RNA control. *The EMBO journal* **9,** 1259–1266 (1990).

192. Szymanski, M., Barciszewska, M. Z., Erdmann, V. A. & Barciszewski, J. 5S Ribosomal RNA Database. *Nucleic Acids Research* **30,** 176–178 (2002).

193. Villa, E., Sengupta, J., Trabuco, L. G., LeBarron, J., Baxter, W. T., Shaikh, T. R., Grassucci, R. A., Nissen, P., Ehrenberg, M., Schulten, K. & Frank, J. Ribosome-induced changes in elongation factor Tu conformation control GTP hydrolysis. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 1063–1068 (2009).

194. Massire, C, Jaeger, L & Westhof, E. Derivation of the three-dimensional architecture of bacterial ribonuclease P RNAs from comparative sequence analysis. *Journal of Molecular Biology* **279,** 773–793 (1998).

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

196. Biswas, R, Ledman, D. W., Fox, R. O., Altman, S & Gopalan, V. Mapping RNA-protein interactions in ribonuclease P from Escherichia coli using disulfide-linked EDTA-Fe. *Journal of Molecular Biology* **296,** 19–31 (2000).

197. Nahvi, A., Barrick, J. E. & Breaker, R. R. Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes. *Nucleic Acids Research* **32,** 143–150 (2004).

198. Johnson, J. E., Reyes, F. E., Polaski, J. T. & Batey, R. T. B12 cofactors directly stabilize an mRNA regulatory switch. *Nature* **492,** 133–137 (2012).

199. Neidhardt, F. C. *Escherichia coli and Salmonella typhimurium* 2nd ed. (ASM Press, 1996).

200. Yusupov, M. M., Yusupova, G. Z., Baucom, A, Lieberman, K, Earnest, T. N., Cate, J. H. & Noller, H. F. Crystal structure of the ribosome at 5.5 A resolution. *Science* **292,** 883–896 (2001).

201. Lai, L. B., Vioque, A., Kirsebom, L. A. & Gopalan, V. Unexpected diversity of RNase P, an ancient tRNA processing enzyme: challenges and prospects. *FEBS Letters* **584,** 287–296 (2010).

202. Tsai, H.-Y., Masquida, B., Biswas, R., Westhof, E. & Gopalan, V. Molecular Modeling of the Three-dimensional Structure of the Bacterial RNase P Holoenzyme. *Journal of Molecular Biology* **325,** 661–675 (2003).

203. Chan, C. W., Chetnani, B. & Mondragón, A. Structure and function of the T-loop structural motif in noncoding RNAs. *Wiley interdisciplinary reviews. RNA* **4,** 507–522 (2013).

204. Dong, H, Kirsebom, L. A. & Nilsson, L. Growth rate regulation of 4.5 S RNA and M1 RNA the catalytic subunit of Escherichia coli RNase P. *Journal of Molecular Biology* **261,** 303–308 (1996).

205. Jarrous, N. & Gopalan, V. Archaeal/eukaryal RNase P: subunits, functions and RNA diversification. *Nucleic Acids Research* **38,** 7885–7894 (2010).

206. Nissen, P, Hansen, J, Ban, N, Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289,** 920–930 (2000).

207. Hannon, G. J. RNA interference. *Nature* **418,** 244–251 (2002).

208. Brantl, S. Regulatory mechanisms employed by cis-encoded antisense RNAs. *Current opinion in microbiology* **10,** 102–109 (2007).

209. Wiedenheft, B., Sternberg, S. H. & Doudna, J. A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* **482,** 331–338 (2012).

210. Kortmann, J. & Narberhaus, F. Bacterial RNA thermometers: molecular zippers and switches. *Nature Reviews Microbiology* **10,** 255–265 (2012).

211. Cech, T. R. & Steitz, J. A. The Noncoding RNA Revolution— Trashing Old Rules to Forge New Ones. *Cell* **157,** 77–94 (2014).

212. Behm-Ansmant, I., Helm, M. & Motorin, Y. Use of specific chemical reagents for detection of modified nucleotides in RNA. *Journal of nucleic acids* **2011,** 408053 (2011).

213. Forconi, M. & Herschlag, D. Metal ion-based RNA cleavage as a structural probe. *Methods in enzymology* **468,** 91–106 (2009).

214. Ge, P. & Zhang, S. Computational analysis of RNA structures with chemical probing data. *Methods* **79-80,** 60–66 (2015).

215. Nick, H & Gilbert, W. Detection in vivo of protein-DNA interactions within the lac operon of Escherichia coli. *Nature* **313,** 795–798 (1985).

216. Steen, K.-A., Rice, G. M. & Weeks, K. M. Fingerprinting noncanonical and tertiary RNA structures by differential SHAPE reactivity. *Journal of the American Chemical Society* **134,** 13160–13163 (2012).

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

218. Lorenz, R., Luntzer, D., Hofacker, I. L., Stadler, P. F. & Wolfinger, M. T. SHAPE directed RNA folding. *Bioinformatics* **32,** 145–147 (2016).

219. Rice, G. M., Leonard, C. W. & Weeks, K. M. RNA secondary structure modeling at consistent high accuracy using differential SHAPE. *RNA* **20,** 846–854 (2014).

220. Kladwang, W., VanLang, C. C., Cordero, P. & Das, R. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nature Chemistry* **3,** 954–962 (2011).

221. Ding, Y & Lawrence, C. E. A bayesian statistical algorithm for RNA secondary structure prediction. *Computers & chemistry* **23,** 387–400 (1999).

222. Ouyang, Z, Snyder, M. P. & Chang, H. Y. SeqFold: Genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Research* **23,** 377–387 (2013).

223. Wu, Y., Shi, B., Ding, X., Liu, T., Hu, X., Yip, K. Y., Yang, Z. R., Mathews, D. H. & Lu, Z. J. Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. *Nucleic Acids Research* **43,** 7247–7259 (2015).

224. Washietl, S., Hofacker, I. L., Stadler, P. F. & Kellis, M. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Research* **40,** 4261–4272 (2012).

225. Kutchko, K. M., Sanders, W., Ziehr, B., Phillips, G., Solem, A., Halvorsen, M., Weeks, K. M., Moorman, N. & Laederach, A. Multiple conformations are a conserved and regulatory feature of the RB1 5′ UTR. *RNA* **21,** 1274–1285 (2015).

226. Turner, R., Shefer, K. & Ares, M. Safer one-pot synthesis of the 'SHAPE' reagent 1-methyl-7-nitroisatoic anhydride (1m7). *RNA* **19,** 1857–1863 (2013).

227. Kladwang, W., Hum, J. & Das, R. Ultraviolet shadowing of RNA can cause significant chemical damage in seconds. *Scientific reports* **2,** 517 (2012).

228. Low, J. T., Knoepfel, S. A., Watts, J. M., ter Brake, O., Berkhout, B. & Weeks, K. M. SHAPE-directed Discovery of Potent shRNA Inhibitors of HIV-1. *Molecular Therapy* **20,** 820–828 (2012).

229. Darty, K., Denise, A. & Ponty, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25,** 1974–1975 (2009).

230. Haller, A., Altman, R. B., Soulière, M. F., Blanchard, S. C. & Micura, R. Folding and ligand recognition of the TPP riboswitch aptamer at single-molecule resolution. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 4188–4193 (2013).

231. Serganov, A., Polonskaia, A., Phan, A. T., Breaker, R. R. & Patel, D. J. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature* **441,** 1167–1171 (2006).

232. Serganov, A., Yuan, Y.-R., Pikovskaya, O., Polonskaia, A., Malinina, L., Phan, A. T., Hobartner, C., Micura, R., Breaker, R. R. & Patel, D. J. Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chemistry & biology* **11,** 1729–1741 (2004).

233. Mortimer, S. A. & Weeks, K. M. Time-Resolved RNA SHAPE Chemistry. *Journal of the American Chemical Society* **130,** 16178–16180 (2008).

234. Mortimer, S. A. & Weeks, K. M. Time-resolved RNA SHAPE chemistry: quantitative RNA structure analysis in one-second snapshots and at single-nucleotide resolution. *Nature Protocols* **4,** 1413–1421 (2009).

235. Grohman, J. K., Gorelick, R. J., Lickwar, C. R., Lieb, J. D., Bower, B. D., Znosko, B. M. & Weeks, K. M. A guanosine-centric mechanism for RNA chaperone function. *Science* **340,** 190–195 (2013).

236. Peattie, D. A. Direct chemical method for sequencing RNA. *Proceedings of the National Academy of Sciences* **76,** 1760–1764 (1979).

237. Aviran, S. & Pachter, L. Rational experiment design for sequencing-based RNA structure mapping. *RNA* **20,** 1864–1877 (2014).

238. Vandivier, L. E., Li, F. & Gregory, B. D. High-throughput nuclease-mediated probing of RNA secondary structure in plant transcriptomes. *Methods in molecular biology* **1284,** 41–70 (2015).

239. Miao, Z. *et al.* RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA* **21,** 1066–1084 (2015).

240. Ding, Y., Chan, C. Y. & Lawrence, C. E. Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Research* **32,** W135–41 (2004).

241. Homan, P. J., Tandon, A., Rice, G. M., Ding, F., Dokholyan, N. V. & Weeks, K. M. RNA tertiary structure analysis by 2′-hydroxyl molecular interference. *Biochemistry* **53,** 6825–6833 (2014).

242. Kramer, F. R. & Mills, D. R. Secondary structure formation during RNA synthesis. *Nucleic Acids Research* **9,** 5109–5124 (1981).

243. Russell, R., Zhuang, X., Babcock, H. P., Millett, I. S., Doniach, S., Chu, S. & Herschlag, D. Exploring the folding landscape of a structured RNA. *Proceedings of the National Academy of Sciences* **99,** 155–160 (2002).

244. Mandal, M. & Breaker, R. R. Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nature Structural & Molecular Biology* **11,** 29–35 (2004).

245. Wright, D. J., King, K & Modrich, P. The negative charge of Glu-111 is required to activate the cleavage center of EcoRI endonuclease. *The Journal of biological chemistry* **264,** 11816–11821 (1989).

246. Komissarova, N & Kashlev, M. Functional topography of nascent RNA in elongation intermediates of RNA polymerase. *Proceedings of the National Academy of Sciences* **95,** 14699–14704 (1998).

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

248. Ren, A., Rajashankar, K. R. & Patel, D. J. Fluoride ion encapsulation by Mg2+ ions and phosphates in a fluoride riboswitch. *Nature* **486,** 85–89 (2012).

249. Ali, M., Lipfert, J., Seifert, S., Herschlag, D. & Doniach, S. The Ligand-Free State of the TPP Riboswitch: A Partially Folded RNA Structure. *Journal of Molecular Biology* **396,** 153–165 (2010).

250. Ottink, O. M., Rampersad, S. M., Tessari, M., Zaman, G. J. R., Heus, H. A. & Wijmenga, S. S. Ligand-induced folding of the guanine-sensing riboswitch is controlled by a combined predetermined induced fit mechanism. *RNA* **13,** 2202–2212 (2007).

251. Ray-Soni, A., Bellecourt, M. J. & Landick, R. Mechanisms of Bacterial Transcription Termination: All Good Things Must End. *Annual Review of Biochemistry* (2016).

252. Qi, L. S. & Arkin, A. P. A versatile framework for microbial engineering using synthetic non-coding RNAs. *Nature Reviews Microbiology* **12,** 341–354 (2014).

253. Liu, C. C., Qi, L., Lucks, J. B., Segall-Shapiro, T. H., Wang, D., Mutalik, V. K. & Arkin, A. P. An adaptor from translational to transcriptional control enables predictable assembly of complex regulation. *Nature Methods* **9,** 1088–1094 (2012).

254. Manch-Citron, J. N., Gennaro, M. L., Majumder, S & Novick, R. P. RepC is rate limiting for pT181 plasmid replication. *Plasmid* **16,** 108–115 (1986).

255. Carleton, S, Projan, S. J., Highlander, S. K., Moghazeh, S. M. & Novick, R. P. Control of pT181 replication II. Mutational analysis. *The EMBO journal* **3,** 2407–2414 (1984).

256. Kolb, F. A., Westhof, E., Ehresmann, B., Ehresmann, C., Wagner, E. G. H. & Romby, P. Four-way Junctions in Antisense RNA-mRNA Complexes Involved in Plasmid Replication Control: A Common Theme? *Journal of Molecular Biology* **309,** 605–614 (2001).

257. Kolb, F., Malmgren, C, Westhof, E., Ehresmann, C., Ehresmann, B., Wagner, E. G. & Romby, P. An unusual structure formed by antisense-target RNA binding involves an extended kissing complex with a four-way junction and a side-by-side helical alignment. *RNA* **6,** 311–324 (2000).

258. Kolb, F., Engdahl, H., Slagter-Jäger, J., Ehresmann, B., Ehresmann, C., Westhof, E., Wagner, E. & Romby, P. Progression of a loop–loop complex to a four-way junction is crucial for the activity of a regulatory antisense RNA. *The EMBO journal* **19,** 5905–5915 (2000).

259. Brantl, S. & Wagner, E. G. Antisense RNA-mediated transcriptional attenuation occurs faster than stable antisense/target RNA pairing: an in vitro study of plasmid pIP501. *The EMBO journal* **13,** 3599–3607 (1994).

260. Siemering, K. R., Praszkier, J & Pittard, A. J. Mechanism of binding of the antisense and target RNAs involved in the regulation of IncB plasmid replication. *Journal of Bacteriology* **176,** 2677–2688 (1994).

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

262. Takahashi, M. K., Hayes, C. A., Chappell, J., Sun, Z. Z., Murray, R. M., Noireaux, V. & Lucks, J. B. Characterizing and prototyping genetic networks with cell-free transcription-translation reactions. *Methods* **86,** 60–72 (2015).

263. Meyer, S., Chappell, J., Sankar, S., Chew, R. & Lucks, J. B. Improving fold activation of small transcription activating RNAs (STARs) with rational RNA engineering strategies. *Biotechnology and Bioengineering* **113,** 216–225 (2016).

264. Hu, C. Y., Varner, J. D. & Lucks, J. B. Generating Effective Models and Parameters for RNA Genetic Circuits. *ACS Synthetic Biology* **4,** 914–926 (2015).

265. Antao, V. P., Lai, S. Y. & Tinoco, I. A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Research* **19,** 5901–5905 (1991).

266. Šulc, P., Romano, F., Ouldridge, T. E., Doye, J. P. K. & Louis, A. A. A nucleotide-level coarse-grained model of RNA. *The Journal of chemical physics* **140,** 235102 (2014).

267. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).

268. Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346,** 1258096 (2014).

269. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A. & Horvath, P. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315,** 1709–1712 (2007).

270. Garneau, J. E., Dupuis, M.-È., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A. H. & Moineau, S. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468,** 67–71 (2010).

271. Nuñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N. & Doudna, J. A. Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature* **527,** 535–538 (2015).

272. Brouns, S. J. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J. H., Snijders, A. P. L., Dickman, M. J., Makarova, K. S., Koonin, E. V. & van der Oost, J. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321,** 960–964 (2008).

273. Jusiak, B., Cleto, S., Perez-Pinera, P. & Lu, T. K. Engineering Synthetic Gene Circuits in Living Cells with CRISPR Technology. *Trends in biotechnology* **34,** 535–547 (2016).

274. Kleinstiver, B. P., Pattanayak, V., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Zheng, Z. & Joung, J. K. High-fidelity CRISPR–Cas9 nucleaseswith no detectable genome-wideoff-target effects. *Nature* **529,** 490–495 (2016).

275. Dahlman, J. E., Abudayyeh, O. O., Joung, J., Gootenberg, J. S., Zhang, F. & Konermann, S. Orthogonal gene knockout and activation with a catalytically active Cas9 nuclease. *Nature Biotechnology* **33,** 1159–1161 (2015).

276. Wang, Y., Zhang, Z.-T., Seo, S.-O., Lynn, P., Lu, T., Jin, Y.-S. & Blaschek, H. P. Bacterial Genome Editing with CRISPR-Cas9: Deletion, Integration, Single Nucleotide Modification, and Desirable "Clean" Mutant Selection in Clostridium beijerinckii as an Example. *ACS Synthetic Biology* **5,** 721–732 (2016).

277. Nuñez, J. K., Harrington, L. B. & Doudna, J. A. Chemical and Biophysical Modulation of Cas9 for Tunable Genome Engineering. *ACS Chemical Biology* **11,** 681–688 (2016).

278. Yao, L., Cengic, I., Anfelt, J. & Hudson, E. P. Multiple Gene Repression in Cyanobacteria Using CRISPRi. *ACS Synthetic Biology* **5,** 207–212 (2016).

279. Borchardt, E. K., Vandoros, L. A., Huang, M., Lackey, P. E., Marzluff, W. F. & Asokan, A. Controlling mRNA stability and translation with the CRISPR endoribonuclease Csy4. *RNA* **21,** 1921–1930 (2015).

280. Hemphill, J., Borchardt, E. K., Brown, K., Asokan, A. & Deiters, A. Optical Control of CRISPR/Cas9 Gene Editing. *Journal of the American Chemical Society* **137,** 5642–5645 (2015).

281. Kiani, S. *et al.* Cas9 gRNA engineering for genome editing, activation and repression. *Nature Methods* **12,** 1051–1054 (2015).

282. Kiani, S., Beal, J., Ebrahimkhani, M. R., Huh, J., Hall, R. N., Xie, Z., Li, Y. & Weiss, R. CRISPR transcriptional repression devices and layered circuits in mammalian cells. *Nature Methods* **11,** 723–726 (2014).

283. Chang, N., Sun, C., Gao, L., Zhu, D., Xu, X., Zhu, X., Xiong, J.-W. & Xi, J. J. Genome editing with RNA-guided Cas9 nuclease in Zebrafish embryos. *Cell Research* **23,** 465–472 (2013).

284. Woo, J. W., Kim, J., Kwon, S. I., Corvalán, C., Cho, S. W., Kim, H., Kim, S.-G., Kim, S.-T., Choe, S. & Kim, J.-S. DNA-free genome editing in plants with preassembled CRISPR-Cas9 ribonucleoproteins. *Nature Biotechnology* **33,** 1162–1164 (2015).

285. Esvelt, K. M., Mali, P., Braff, J. L., Moosburner, M., Yaung, S. J. & Church, G. M. Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nature Methods* **10,** 1116–1121 (2013).

286. Davis, K. M., Pattanayak, V., Thompson, D. B., Zuris, J. A. & Liu, D. R. Small molecule-triggered Cas9 protein with improved genome-editing specificity. *Nature Chemical Biology* **11,** 316–318 (2015).

287. Zetsche, B., Volz, S. E. & Zhang, F. A split-Cas9 architecture for inducible genome editing and transcription modulation. *Nature Biotechnology* **33,** 139–142 (2015).

288. Nihongaki, Y., Kawano, F., Nakajima, T. & Sato, M. Photoactivatable CRISPR-Cas9 for optogenetic genome editing. *Nature Biotechnology* **33,** 755–760 (2015).

289. Hou, Z., Zhang, Y., Propson, N. E., Howden, S. E., Chu, L.-F., Sontheimer, E. J. & Thomson, J. A. Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 15644–15649 (2013).

290. Slaymaker, I. M., Gao, L., Zetsche, B., Scott, D. A., Yan, W. X. & Zhang, F. Rationally engineered Cas9 nucleases with improved specificity. *Science* **351,** 84–88 (2016).

291. Ho, T.-T., Zhou, N., Huang, J., Koirala, P., Xu, M., Fung, R., Wu, F. & Mo, Y.-Y. Targeting non-coding RNAs with the CRISPR/Cas9 system in human cell lines. *Nucleic Acids Research* **43,** e17 (2015).

292. Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F. & Marraffini, L. A. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Research* **41,** 7429–7437 (2013).

293. Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G.-W., Park, J., Blackburn, E. H., Weissman, J. S., Qi, L. S. & Huang, B. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155,** 1479–1491 (2013).

294. Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., Stern-Ginossar, N., Brandman, O., Whitehead, E. H., Doudna, J. A., Lim, W. A., Weiss-

man, J. S. & Qi, L. S. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154,** 442–451 (2013).

295. Luo, M. L., Mullis, A. S., Leenay, R. T. & Beisel, C. L. Repurposing endogenous type I CRISPR-Cas systems for programmable gene repression. *Nucleic Acids Research* **43,** 674–681 (2015).

296. Nissim, L., Perli, S. D., Fridkin, A., Perez-Pinera, P. & Lu, T. K. Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells. *Molecular Cell* **54,** 698–710 (2014).

297. Perez-Pinera, P., Kocak, D. D., Vockley, C. M., Adler, A. F., Kabadi, A. M., Polstein, L. R., Thakore, P. I., Glass, K. A., Ousterout, D. G., Leong, K. W., Guilak, F., Crawford, G. E., Reddy, T. E. & Gersbach, C. A. RNA-guided gene activation by CRISPR-Cas9–based transcription factors. *Nature Methods* **10,** 973–976 (2013).

298. Polstein, L. R. & Gersbach, C. A. A light-inducible CRISPR-Cas9 system for control of endogenous gene activation. *Nature Chemical Biology* **11,** 198–200 (2015).

299. Nihongaki, Y., Yamamoto, S., Kawano, F., Suzuki, H. & Sato, M. CRISPR-Cas9-based photoactivatable transcription system. *Chemistry & biology* **22,** 169–174 (2015).

300. Maeder, M. L., Linder, S. J., Cascio, V. M., Fu, Y., Ho, Q. H. & Joung, J. K. CrIsPr rna–guided activation of endogenous human genes. *Nature Methods* **10,** 977–981 (2013).

301. Farzadfard, F & Lu, T. K. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* **346,** 1256272–1256272 (2014).

302. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nature Reviews Microbiology* **13,** 722–736 (2015).

303. Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J. M., Wolf, Y. I., Yakunin, A. F., van der Oost, J. & Koonin, E. V. Evolution and classification of the CRISPR-Cas systems. *Nature Reviews Microbiology* **9,** 467–477 (2011).

304. Chylinski, K., Makarova, K. S., Charpentier, E. & Koonin, E. V. Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Research* **42,** 6091–6105 (2014).

305. Heler, R., Samai, P., Modell, J. W., Weiner, C., Goldberg, G. W., Bikard, D. & Marraffini, L. A. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519,** 199–202 (2015).

306. Jiang, F., Zhou, K., Ma, L., Gressel, S. & Doudna, J. A. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* **348,** 1477–1481 (2015).

307. Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., Ishitani, R., Zhang, F. & Nureki, O. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156,** 935–949 (2014).

308. Jiang, F., Taylor, D. W., Chen, J. S., Kornfeld, J. E., Zhou, K., Thompson, A. J., Nogales, E. & Doudna, J. A. Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **351,** 867–871 (2016).

309. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513,** 569–573 (2014).

310. Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., Eckert, M. R., Vogel, J. & Charpentier, E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471,** 602–607 (2011).

311. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–67 (2014).

312. Jinek, M., Jiang, F., Taylor, D. W., Sternberg, S. H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., Kaplan, M., Iavarone, A. T., Charpentier, E., Nogales, E. & Doudna, J. A. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343,** 1247997 (2014).

313. Cheng, A. W., Wang, H., Yang, H., Shi, L., Katz, Y., Theunissen, T. W., Rangarajan, S., Shivalila, C. S., Dadon, D. B. & Jaenisch, R. Multiplexed activation of endoge-

nous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell Research* **23,** 1163–1171 (2013).

314. Gilbert, L. A., Horlbeck, M. A., Adamson, B., Villalta, J. E., Chen, Y., Whitehead, E. H., Guimaraes, C., Panning, B., Ploegh, H. L., Bassik, M. C., Qi, L. S., Kampmann, M. & Weissman, J. S. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159,** 647–661 (2014).

315. Briner, A. E., Donohoue, P. D., Gomaa, A. A., Selle, K., Slorach, E. M., Nye, C. H., Haurwitz, R. E., Beisel, C. L., May, A. P. & Barrangou, R. Guide RNA functional modules direct Cas9 activity and orthogonality. *Molecular Cell* **56,** 333–339 (2014).

316. Wright, A. V., Sternberg, S. H., Taylor, D. W., Staahl, B. T., Bardales, J. A., Kornfeld, J. E. & Doudna, J. A. Rational design of a split-Cas9 enzyme complex. *Proceedings of the National Academy of Sciences of the United States of America* **112,** 2984–2989 (2015).

317. Peabody, D. S. & Lim, F. Complementation of RNA binding site mutations in MS2 coat protein heterodimers. *Nucleic Acids Research* **24,** 2352–2359 (1996).

318. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346,** 818–822 (1990).

319. Jenison, R. D., Gill, S. C., Pardi, A & Polisky, B. High-resolution molecular discrimination by RNA. *Science* **263,** 1425–1429 (1994).

320. Shen, S, Rodrigo, G, Prakash, S, Majer, E, Landrain, T. E., Kirov, B, Daros, J. A. & Jaramillo, A. Dynamic signal processing by ribozyme-mediated RNA circuits to control gene expression. *Nucleic Acids Research* **43,** 5158–5170 (2015).

321. Kushwaha, M., Rostain, W., Prakash, S., Duncan, J. N. & Jaramillo, A. Using RNA as Molecular Code for Programming Cellular Function. *ACS Synthetic Biology* (2016).

322. Pardee, K. *et al.* Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular Components. *Cell* **165,** 1255–1266 (2016).

323. Lee, Y. J., Hoynes-O'Connor, A., Leong, M. C. & Moon, T. S. Programmable control of bacterial gene expression with the combined CRISPR and antisense RNA system. *Nucleic Acids Research* **44,** 2462–2473 (2016).

324. Zadeh, J. N., Wolfe, B. R. & Pierce, N. A. Nucleic acid sequence design via efficient ensemble defect optimization. *Journal of Computational Chemistry* **32,** 439–452 (2010).

325. Nicholson, B. L. & White, K. A. Functional long-range RNA-RNA interactions in positive-strand RNA viruses. *Nature Reviews Microbiology* **12,** 493–504 (2014).

326. Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess, J. W., Swanstrom, R., Burch, C. L. & Weeks, K. M. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460,** 711–716 (2009).

327. Wu, B., Grigull, J., Ore, M. O., Morin, S. & White, K. A. Global organization of a positive-strand RNA virus genome. *PLoS Pathogens* **9,** e1003363 (2013).

328. Mauger, D. M., Golden, M., Yamane, D., Williford, S., Lemon, S. M., Martin, D. P. & Weeks, K. M. Functionally conserved architecture of hepatitis C virus RNA genomes. *Proceedings of the National Academy of Sciences of the United States of America* **112,** 3692–3697 (2015).

329. Pirakitikulr, N., Kohlway, A., Lindenbach, B. D. & Pyle, A. M. The Coding Region of the HCV Genome Contains a Network of Regulatory RNA Structures. *Molecular Cell,* 1–11 (2016).

330. Florentz, C, Briand, J. P., Romby, P., Hirth, L, Ebel, J. P. & Glegé, R. The tRNA-like structure of turnip yellow mosaic virus RNA: structural organization of the last 159 nucleotides from the 3′ OH terminus. *The EMBO journal* **1,** 269–276 (1982).

331. Pleij, C. W., Rietveld, K & Bosch, L. A new principle of RNA folding based on pseudoknotting. *Nucleic Acids Research* **13,** 1717–1731 (1985).

332. Rietveld, K, Van Poelgeest, R, Pleij, C. W., Van Boom, J. H. & Bosch, L. The tRNA-like structure at the 3′ terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA. *Nucleic Acids Research* **10,** 1929–1946 (1982).

333. Ahlquist, P, Dasgupta, R & Kaesberg, P. Near identity of 3- RNA secondary structure in bromoviruses and cucumber mosaic virus. *Cell* **23,** 183–189 (1981).

334. Joshi, R. L., Joshi, S., Chapeville, F. & Haenni, A. L. tRNA-like structures of plant viral RNAs: conformational requirements for adenylation and aminoacylation. *The EMBO journal* **2,** 1123–1127 (1983).

335. Felden, B, Florentz, C, Giege, R & Westhof, E. Solution structure of the 3'-end of brome mosaic virus genomic RNAs. Conformational mimicry with canonical tRNAs. *Journal of Molecular Biology* **235,** 508–531 (1994).

336. Felden, B, Florentz, C, Giege, R & Westhof, E. A central pseudoknotted three-way junction imposes tRNA-like mimicry and the orientation of three 5' upstream pseudoknots in the 3' terminus of tobacco mosaic virus RNA. *RNA* **2,** 201–212 (1996).

337. Rietveld, K, Linschooten, K, Pleij, C. W. & Bosch, L. The three-dimensional folding of the tRNA-like structure of tobacco mosaic virus RNA. A new building principle applied twice. *The EMBO journal* **3,** 2613–2619 (1984).

338. Van Belkum, A, Abrahams, J. P., Pleij, C. W. & Bosch, L. Five pseudoknots are present at the 204 nucleotides long 3' noncoding region of tobacco mosaic virus RNA. *Nucleic Acids Research* **13,** 7673–7686 (1985).

339. Gallie, D. R., Feder, J. N., Schimke, R. T. & Walbot, V. Functional analysis of the tobacco mosaic virus tRNA-like structure in cytoplasmic gene regulation. *Nucleic Acids Research* **19,** 5031–5036 (1991).

340. Osman, T. A., Hemenway, C. L. & Buck, K. W. Role of the 3' tRNA-like structure in tobacco mosaic virus minus-strand RNA synthesis by the viral RNA-dependent RNA polymerase In vitro. *Journal of Virology* **74,** 11671–11680 (2000).

341. Sivakumaran, K, Bao, Y, Roossinck, M. J. & Kao, C. C. Recognition of the core RNA promoter for minus-strand RNA synthesis by the replicases of Brome mosaic virus and Cucumber mosaic virus. *Journal of Virology* **74,** 10323–10331 (2000).

342. Takamatsu, N, Watanabe, Y, Meshi, T & Okada, Y. Mutational analysis of the pseudoknot region in the 3′ noncoding region of tobacco mosaic virus RNA. *Journal of Virology* **64,** 3686–3693 (1990).

343. Zeenko, V. V., Ryabova, L. A., Spirin, A. S., Rothnie, H. M., Hess, D., Browning, K. S. & Hohn, T. Eukaryotic elongation factor 1A interacts with the upstream pseudoknot domain in the 3′ untranslated region of tobacco mosaic virus RNA. *Journal of Virology* **76,** 5678–5691 (2002).

344. Chen, M. H., Roossinck, M. J. & Kao, C. C. Efficient and specific initiation of subgenomic RNA synthesis by cucumber mosaic virus replicase in vitro requires an upstream RNA stem-loop. *Journal of Virology* **74,** 11201–11209 (2000).

345. Jacquemond, M. *Cucumber Mosaic Virus* 1st ed. (Elsevier Inc., 2012).

346. Palukaitis, P, Roossinck, M. J., Dietzgen, R. G. & Francki, R. I. Cucumber mosaic virus. *Advances in virus research* **41,** 281–348 (1992).

347. Chapman, M. R., Rao, A. L. & Kao, C. C. Sequences 5′ of the conserved tRNA-like promoter modulate the initiation of minus-strand synthesis by the brome mosaic virus RNA-dependent RNA polymerase. *Virology* **252,** 458–467 (1998).

348. Miller, W. A., Bujarski, J. J., Dreher, T. W. & Hall, T. C. Minus-strand initiation by brome mosaic virus replicase within the 3′ tRNA-like structure of native and modified RNA templates. *Journal of Molecular Biology* **187,** 537–546 (1986).

349. Chapman, M. R. & Kao, C. C. A minimal RNA promoter for minus-strand RNA synthesis by the brome mosaic virus polymerase complex. *Journal of Molecular Biology* **286,** 709–720 (1999).

350. Dreher, T. W., Bujarski, J. J. & Hall, T. C. Mutant viral RNAs synthesized in vitro show altered aminoacylation and replicase template activities. *Nature* **311,** 171–175 (1984).

351. Thompson, J. R., Buratti, E, de Wispelaere, M & Tepfer, M. Structural and functional characterization of the 5′ region of subgenomic RNA5 of cucumber mosaic virus. *Journal of General Virology* **89,** 1729–1738 (2008).

352. Thompson, J. R. & Tepfer, M. The 3′ untranslated region of cucumber mosaic virus (CMV) subgroup II RNA3 arose by interspecific recombination between CMV and tomato aspermy virus. *Journal of General Virology* **90,** 2293–2298 (2009).

353. Morroni, M., Thompson, J. R. & Tepfer, M. Analysis of recombination between viral RNAs and transgene mRNA under conditions of high selection pressure in favour of recombinants. *Journal of General Virology* **90,** 2798–2807 (2009).

354. Baumstark, T & Ahlquist, P. The brome mosaic virus RNA3 intergenic replication enhancer folds to mimic a tRNA TpsiC-stem loop and is modified in vivo. *RNA* **7,** 1652–1670 (2001).

355. Chen, J, Noueiry, A & Ahlquist, P. Brome Mosaic Virus Protein 1a Recruits Viral RNA2 to RNA Replication through a 5′ Proximal RNA2 Signal. *Journal of Virology* **75,** 3207–3219 (2001).

356. Sullivan, M. L. & Ahlquist, P. A brome mosaic virus intergenic RNA3 replication signal functions with viral replication protein 1a to dramatically stabilize RNA in vivo. *Journal of Virology* **73,** 2622–2632 (1999).

357. Yi, G. & Kao, C. cis- and trans-acting functions of brome mosaic virus protein 1a in genomic RNA1 replication. *Journal of Virology* **82,** 3045–3053 (2008).

358. Athavale, S. S., Gossett, J. J., Bowman, J. C., Hud, N. V., Williams, L. D. & Harvey, S. C. In vitro secondary structure of the genomic RNA of satellite tobacco mosaic virus. *PLoS ONE* **8,** e54384 (2013).

359. Thompson, J. R., Langenhan, J. L., Fuchs, M. & Perry, K. L. Genotyping of Cucumber mosaic virus isolates in western New York State during epidemic years: Characterization of an emergent plant virus population. *Virus Research* **210,** 169–177 (2015).

360. Roossinck, M. J., Zhang, L & Hellwald, K. H. Rearrangements in the 5′ nontranslated region and phylogenetic analyses of cucumber mosaic virus RNA 3 indicate radial evolution of three subgroups. *Journal of Virology* **73,** 6752–6758 (1999).

361. Kwon, C. S. & Chung, W. A single-stranded loop in the 5′ untranslated region of cucumber mosaic virus RNA 4 contributes to competitive translational activity. *FEBS Letters* **462,** 161–166 (1999).

362. Sivakumaran, K, Bao, Y, Roossinck, M. J. & Kao, C. C. Recognition of the core RNA promoter for minus-strand RNA synthesis by the replicases of Brome mosaic virus and Cucumber mosaic virus. *Journal of Virology* **74,** 10323–10331 (2000).

363. Robertson, H. D., Altman, S & Smith, J. D. Purification and properties of a specific Escherichia coli ribonuclease which cleaves a tyrosine transfer ribonucleic acid presursor. *The Journal of biological chemistry* **247,** 5243–5251 (1972).

364. Guerrier-Takada, C, Gardiner, K, Marsh, T, Pace, N & Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35,** 849–857 (1983).

365. Liu, F & Altman, S. Differential evolution of substrates for an RNA enzyme in the presence and absence of its protein cofactor. *Cell* **77,** 1093–1100 (1994).

366. Li, Y. & Altman, S. A specific endoribonuclease, RNase P, affects gene expression of polycistronic operon mRNAs. *Proceedings of the National Academy of Sciences* **100,** 13213–13218 (2003).

367. Altman, S., Wesolowski, D., Guerrier-Takada, C. & Li, Y. RNase P cleaves transient structures in some riboswitches. *Proceedings of the National Academy of Sciences* **102,** 11284–11289 (2005).

368. Mohanty, B. K. & Kushner, S. R. Rho-independent transcription terminators inhibit RNase P processing of the secG leuU and metT tRNA polycistronic transcripts in Escherichia coli. *Nucleic Acids Research* **36,** 364–375 (2008).

369. Kazantsev, A. V. & Pace, N. R. Bacterial RNase P: a new view of an ancient enzyme. *Nature Reviews Microbiology* **4,** 729–740 (2006).

370. Mondragón, A. Structural studies of RNase P. *Annual Review of Biophysics* **42,** 537–557 (2013).

371. Waugh, D. S. & Pace, N. R. Complementation of an RNase P RNA (rnpB) gene deletion in Escherichia coli by homologous genes from distantly related eubacteria. *Journal of Bacteriology* **172,** 6316–6322 (1990).

372. Haas, E. S. & Brown, J. W. Evolutionary variation in bacterial RNase P RNAs. *Nucleic Acids Research* **26,** 4093–4099 (1998).

373. Walker, S. C. & Engelke, D. R. Ribonuclease P: the evolution of an ancient RNA enzyme. *Critical reviews in biochemistry and molecular biology* **41,** 77–102 (2006).

374. Ellis, J. C. & Brown, J. W. The evolution of RNase P and its RNA. *Ribonuclease P*, 17–40 (2010).

375. Lai, L. B., Chan, P. P., Cozen, A. E., Bernick, D. L., Brown, J. W., Gopalan, V. & Lowe, T. M. Discovery of a minimal form of RNase P in Pyrobaculum. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 22493–22498 (2010).

376. Harris, J. K., Haas, E. S., Williams, D, Frank, D. N. & Brown, J. W. New insight into RNase P RNA structure from comparative analysis of the archaeal RNA. *RNA* **7,** 220–232 (2001).

377. Chan, P. P., Brown, J. W. & Lowe, T. M. Modeling the Thermoproteaceae RNase P RNA. *RNA Biology* **9,** 1155–1160 (2012).

378. Lai, L. B., Cho, I. M., Chen, W. Y. & Gopalan, V. Archaeal RNase P: a mosaic of its bacterial and eukaryal relatives. *Ribonuclease P* (2010).

379. Chen, W. Y., Singh, D, Lai, L. B., Stiffler, M. A., Lai, H. D., Foster, M. P. & Gopalan, V. Fidelity of tRNA 5′-maturation: a possible basis for the functional dependence of archaeal and eukaryal RNase P on multiple protein cofactors. *Nucleic Acids Research* **40,** 4666–4680 (2012).

380. Boomershine, W. P., McElroy, C. A., Tsai, H.-Y., Wilson, R. C., Gopalan, V. & Foster, M. P. Structure of Mth11/Mth Rpp29, an essential protein subunit of archaeal and eukaryotic RNase P. *Proceedings of the National Academy of Sciences* **100,** 15398–15403 (2003).

381. Sidote, D. J., Heideker, J. & Hoffman, D. W. Crystal structure of archaeal ribonuclease P protein aRpp29 from Archaeoglobus fulgidus. *Biochemistry* **43,** 14128–14138 (2004).

382. Sidote, D. J. & Hoffman, D. W. NMR structure of an archaeal homologue of ribonuclease P protein Rpp29. *Biochemistry* **42,** 13541–13550 (2003).

383. Wilson, R. C., Bohlen, C. J., Foster, M. P. & Bell, C. E. Structure of Pfu Pop5, an archaeal RNase P protein. *Proceedings of the National Academy of Sciences* **103,** 873–878 (2006).

384. Amero, C. D., Boomershine, W. P., Xu, Y. & Foster, M. Solution Structure of Pyrococcus furiosusRPP21, a Component of the Archaeal RNase P Holoenzyme, and Interactions with Its RPP29 Protein Partner †. *Biochemistry* **47,** 11704–11710 (2008).

385. Chen, W. Y., Pulukkunat, D. K., Cho, I. M., Tsai, H. Y. & Gopalan, V. Dissecting functional cooperation among protein subunits in archaeal RNase P, a catalytic ribonucleoprotein complex. *Nucleic Acids Research* **38,** 8316–8327 (2010).

386. Pulukkunat, D. K. & Gopalan, V. Studies on Methanocaldococcus jannaschii RNase P reveal insights into the roles of RNA and protein cofactors in RNase P catalysis. *Nucleic Acids Research* **36,** 4172–4180 (2008).

387. Klein, D. J., Schmeing, T. M., Moore, P. B. & Steitz, T. A. The kink-turn: a new RNA secondary structure motif. *The EMBO journal* **20,** 4214–4221 (2001).

388. Gegenheimer, P. Structure, mechanism and evolution of chloroplast transfer RNA processing systems. *Molecular biology reports* **22,** 147–150 (1995).

389. Rossmanith, W & Karwan, R. M. Characterization of human mitochondrial RNase P: novel aspects in tRNA processing. *Biochemical and Biophysical Research Communications* **247,** 234–241 (1998).

390. Holzmann, J., Frank, P., Löffler, E., Bennett, K. L., Gerner, C. & Rossmanith, W. RNase P without RNA: identification and functional reconstitution of the human mitochondrial tRNA processing enzyme. *Cell* **135,** 462–474 (2008).

391. Thomas, B. C., Li, X & Gegenheimer, P. Chloroplast ribonuclease P does not utilize the ribozyme-type pre-tRNA cleavage mechanism. *RNA* **6,** 545–553 (2000).

392. Gobert, A., Pinker, F., Fuchsbauer, O., Gutmann, B., Boutin, R. e., Roblin, P., Sauter, C. & eacute, P. G. Structural insights into protein-only RNase P complexed with tRNA. *Nature Communications* **4,** 1353–8 (2013).

393. Chen, T.-H., Tanimoto, A., Shkriabai, N., Kvaratskhelia, M., Wysocki, V. & Gopalan, V. Use of chemical modification and mass spectrometry to identify substrate-contacting sites in proteinaceous RNase P, a tRNA processing enzyme. *Nucleic Acids Research* **44,** 5344–5355 (2016).

394. Howard, M. J., Lim, W. H., Fierke, C. A. & Koutmos, M. Mitochondrial ribonuclease P structure provides insight into the evolution of catalytic strategies for precursor-tRNA 5′ processing. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 16149–16154 (2012).

395. Sinapah, S, Wu, S, Chen, Y, Pettersson, B. M. F., Gopalan, V & Kirsebom, L. A. Cleavage of model substrates by archaeal RNase P: role of protein cofactors in cleavage-site selection. *Nucleic Acids Research* **39,** 1105–1116 (2011).

396. Chen, W.-Y., Xu, Y., Cho, I.-M., Oruganti, S. V., Foster, M. P. & Gopalan, V. Cooperative RNP assembly: complementary rescue of structural defects by protein and RNA subunits of archaeal RNase P. *Journal of Molecular Biology* **411,** 368–383 (2011).

397. Crowe, B. L., Bohlen, C. J., Wilson, R. C., Gopalan, V. & Foster, M. P. Assembly of the Complex between Archaeal RNase P Proteins RPP30 and Pop5. *Archaea* **2011,** 1–12 (2011).

398. Tsai, H.-Y., Pulukkunat, D. K., Woznick, W. K. & Gopalan, V. Functional reconstitution and characterization of Pyrococcus furiosus RNase P. *Proceedings of the National Academy of Sciences* **103,** 16147–16152 (2006).

399. Shi, X., Huang, L., Lilley, D. M. J., Harbury, P. B. & Herschlag, D. The solution structural ensembles of RNA kink-turn motifs and their protein complexes. *Nature Chemical Biology* **12,** 146–152 (2016).

400. Huang, L. & Lilley, D. M. J. The Kink Turn, a Key Architectural Element in RNA Structure. *Journal of Molecular Biology* **428,** 790–801 (2016).

401. Fukuhara, H., Kifusa, M., Watanabe, M., Terada, A., Honda, T., Numata, T., Kakuta, Y. & Kimura, M. A fifth protein subunit Ph1496p elevates the optimum temperature for the ribonuclease P activity from Pyrococcus horikoshii OT3. *Biochemical and Biophysical Research Communications* **343,** 956–964 (2006).

402. Wang, J., Fessl, T., Schroeder, K. T., Ouellet, J., Liu, Y., Freeman, A. D. J. & Lilley, D. M. J. Single-molecule observation of the induction of k-turn RNA structure on binding L7Ae protein. *Biophysical Journal* **103,** 2541–2548 (2012).

403. Pan, T. Higher order folding and domain analysis of the ribozyme from Bacillus subtilis ribonuclease P. *Biochemistry* **34,** 902–909 (1995).

404. Loria, A & Pan, T. Domain structure of the ribozyme from eubacterial ribonuclease P. *RNA* **2,** 551–563 (1996).

405. Frank, D. N., Harris, M. E. & Pace, N. R. Rational design of self-cleaving pre-tRNA-ribonuclease P RNA conjugates. *Biochemistry* **33,** 10800–10808 (1994).

406. Gutmann, B., Gobert, A. & Giegé, P. PRORP proteins support RNase P activity in both organelles and the nucleus in Arabidopsis. *Genes & Development* **26,** 1022–1027 (2012).

407. Zhou, W., Karcher, D., Fischer, A., Maximova, E., Walther, D. & Bock, R. Multiple RNA processing defects and impaired chloroplast function in plants deficient in the organellar protein-only RNase P enzyme. *PLoS ONE* **10,** e0120533 (2015).

408. Imai, T., Nakamura, T., Maeda, T., Nakayama, K., Gao, X., Nakashima, T., Kakuta, Y. & Kimura, M. Pentatricopeptide repeat motifs in the processing enzyme PRORP1 in Arabidopsis thaliana play a crucial role in recognition of nucleotide bases at TψC loop in precursor tRNAs. *Biochemical and Biophysical Research Communications* **450,** 1541–1546 (2014).

409. Howard, M. J., Klemm, B. P. & Fierke, C. A. Mechanistic Studies Reveal Similar Catalytic Strategies for Phosphodiester Bond Hydrolysis by Protein-only and

RNA-dependent Ribonuclease P. *Journal of Biological Chemistry* **290,** 13454–13464 (2015).

410. Ishitani, R., Nureki, O., Nameki, N., Okada, N., Nishimura, S. & Yokoyama, S. Alternative tertiary structure of tRNA for recognition by a posttranscriptional modification enzyme. *Cell* **113,** 383–394 (2003).

411. Berkowitz, N. D., Silverman, I. M., Childress, D. M., Kazan, H., Wang, L.-S. & Gregory, B. D. A comprehensive database of high-throughput sequencing-based RNA secondary structure probing data (Structure Surfer). *BMC bioinformatics* **17,** 215 (2016).

412. Fürtig, B., Buck, J., Manoharan, V., Bermel, W., Jäschke, A., Wenter, P., Pitsch, S. & Schwalbe, H. Time-resolved NMR studies of RNA folding. *Biopolymers* **86,** 360–383 (2007).

413. Keller, B. G., Kobitski, A., Jäschke, A., Nienhaus, G. U. & Noé, F. Complex RNA folding kinetics revealed by single-molecule FRET and hidden Markov models. *Journal of the American Chemical Society* **136,** 4534–4543 (2014).

414. Ruff, K. M. & Strobel, S. A. Ligand binding by the tandem glycine riboswitch depends on aptamer dimerization but not double ligand occupancy. *RNA* **20,** 1775–1788 (2014).

415. Lipfert, J., Das, R., Chu, V. B., Kudaravalli, M., Boyd, N., Herschlag, D. & Doniach, S. Structural transitions and thermodynamics of a glycine-dependent riboswitch from Vibrio cholerae. *Journal of Molecular Biology* **365,** 1393–1406 (2007).

416. Kwon, M & Strobel, S. A. Chemical basis of glycine riboswitch cooperativity. *RNA* **14,** 25–34 (2007).

417. Allawi, H. T. & Santalucia, J. Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry* **36,** 10581–10594 (1997).

418. Turner, D. H. & Mathews, D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research* **38,** D280–2 (2010).

419. Patrick, W. G., Nielsen, A. A. K., Keating, S. J., Levy, T. J., Wang, C.-W., Rivera, J. J., Mondragón-Palomino, O., Carr, P. A., Voigt, C. A., Oxman, N. & Kong, D. S. DNA Assembly in 3D Printed Fluidics. *PLoS ONE* **10,** e0143636 (2015).

420. Steitz, J. A. & Jakes, K. How ribosomes select initiator regions in mRNA: base pair formation between the 3′ terminus of 16S rRNA and the mRNA during initiation of protein synthesis in Escherichia coli. *Proceedings of the National Academy of Sciences* **72,** 4734–4738 (1975).