ABSTRACT

Revealing Hidden Biology by Identifying Ligands for the *ykkC* Orphan Riboswitches

Madeline Elizabeth Sherlock

2018

Riboswitches are non-coding RNAs that bind a small molecule or ion ligand to regulate gene expression. The cognate ligands have been identified for dozens of riboswitch classes, but the ligand identities are unknown for many promising structured RNA motifs, which are known as orphan riboswitches. The *ykkC* RNA motif was discovered over a decade ago by using bioinformatics and, until recently, was the longest standing orphan riboswitch candidate. The cognate ligand of *ykkC* riboswitches remained unsolved mostly due to the seemingly disparate set of genes under its regulatory control. These genes commonly code for proteins annotated as urea carboxylases, multi-drug efflux transporters, sulfonate/nitrate/carbonate transporters, arginases, *de novo* purine biosynthesis enzymes, and branched-chain amino acid enzymes.

My colleagues found that guanidine induces expression of a riboswitch-*lacZ* reporter gene fusion in a large screen of various growth conditions. They collected additional evidence implicating guanidine as the elusive ligand for *ykkC* riboswitches by demonstrating that reporter gene expression is activated at a lower concentration of guanidine when the *ykkCD* operon is knocked out in *B. subtilis*. These gene products, which likely function as a selective guanidine transporter, are commonly associated with this riboswitch class. I used in-line probing as an *in vitro* biochemical approach to further evaluate guanidine as the cognate ligand and found that some *ykkC* motif RNAs from

distantly related bacteria become more structured at specific sites in a dose-dependent manner upon the addition of guanidine.

I subsequently determined that representatives of two other RNA motifs, called mini-*ykkC* and *ykkC*-III, also bind guanidine *in vitro*. All three guanidine-binding RNA motifs have similar downstream gene associations, and therefore these three distinct classes of RNA represent the guanidine-I, -II and -III riboswitch classes. I performed experiments on the protein family whose expression is most commonly regulated by guanidine riboswitches and demonstrated that these proteins are involved in breaking down guanidine into non-toxic molecules. Although the biological relevance of guanidine is still poorly understood, the wide distribution of these riboswitches and the genes they control suggest that free guanidine is a greatly underappreciated metabolite.

While the majority of *ykkC* RNAs selectively bind guanidine, I found that the remaining 30% of *ykkC* RNAs do not bind guanidine and carry nucleotide changes in their binding pockets to allow recognition of a different ligand. I analyzed the consensus sequence and structure models as well as the downstream gene associations for these variant *ykkC* RNAs, which revealed at least four distinct candidate riboswitch classes in addition to the guanidine-I riboswitch class. I have validated the cognate ligands for two of these variant *ykkC* riboswitch classes, both of which are RNA nucleotide derivatives.

I discovered that the first variant class selectively binds guanosine tetraphosphate (ppGpp), a bacterial alarmone derived from the ribonucleotide GTP. The ppGpp riboswitch class primarily regulates transcription of genes involved in branched-chain amino acid biosynthesis. Shortly thereafter, I determined that the second of these variant *ykkC* riboswitch classes binds a similar molecule called phosphoribosyl

pyrophosphate (PRPP), the initial substrate for the biosynthesis of RNA monomers, to regulate the expression of genes for purine nucleotide biosynthesis. Members of these newly validated riboswitch classes frequently form tandem arrangements between and other RNA-based regulatory elements. I demonstrated that these tandem arrangements accomplish more complex gene control systems than a typical singlet riboswitch system. The existence of sophisticated RNA regulatory systems that respond to a widespread RNA-derived signaling molecule and a fundamental RNA biosynthetic precursor supports the hypothesis that RNA World organisms could manage a complex metabolic state without the assistance of protein factors.

Revealing Hidden Biology by Identifying Ligands for the *ykkC* Orphan Riboswitches

A Dissertation

Presented to the Faculty of the Graduate School

of

Yale University

in Candidacy for the Degree of

Doctor of Philosophy

by

Madeline Elizabeth Sherlock

Dissertation Director: Ronald R. Breaker

May 2018

©2018 by Madeline Elizabeth Sherlock

All rights reserved.

TABLE OF CONTENTS

ABSTRACT i

TABLE OF CONTENTS vi

ACKNOWLEDGEMENTS xviii

CHAPTER ONE 1

Riboswitches 2

Ligands Sensed by Riboswitches 2

Biological processes controlled by riboswitch ligands 4

Riboswitch class abundance 5

Riboswitch Structure 6

Riboswitch aptamers 6

Variant riboswitch aptamers 8

Riboswitch expression platforms 9

Tandem Riboswitch Architectures 11

Cooperative riboswitches 11

Riboswitches as two-input logic gates 13

Riboswitch Discovery 15

Orphan Riboswitches 17

ykoK RNAs: Mg^{2+} riboswitches 18

ykvJ RNAs: $preQ_1$ riboswitches 19

GEMM RNAs: c-di-GMP riboswitches 20

crcB RNAs: F^- riboswitches 22

ydaO RNAs: c-di-AMP riboswitches 23

yybP RNAs: Mn^{2+} riboswitches 24

pfl RNAs: ZMP/ZTP riboswitches 25

yjdF RNAs: azaaromatic riboswitches 26

ykkC RNAs: the last great orphan? 27

References 29

Figures 37

Figure 1-1. Ligands of known riboswitch classes 37

Figure 1-2. Abundance of validated riboswitch classes 38

Figure 1-3. Consensus models for the guanine and adenine riboswitch class aptamers 39

Figure 1-4. Cartoon models of riboswitches acting at the transcriptional and translational level 40

CHAPTER TWO 41

Abstract 42

Introduction 43

Results 45

The *ykkC* Motif is a widespread orphan riboswitch candidate predicted to control the expression of a variety of transporter and metabolic genes 45

Addition of guanidine triggers gene expression mediated by a subtype 1 *ykkC* motif RNA 46

Molecular recognition determinants for guanidine binding by a riboswitch 48

Guanidine selectively induces riboswitch-mediated transcription 51

A urea carboxylase enzyme prefers guanidine over urea 52

A riboswitch-controlled transport protein selectively binds guanidine 54

Free guanidine is naturally produced by bacteria 56

Discussion 57

Materials and Methods 61

Chemicals and reagents 61

Bioinformatics 61

Reporter assays and knockout construction 62

Preparation of radiolabeled RNA constructs for in-line probing 63

Single-round transcription termination Assays 63

Preparation and analysis of an annotated urea carboxylase 65

Expression and purification of SugE 66

Tryptophan fluorescence assay 67

LC-MS analysis of bacterial cell extracts 67

Acknowledgments 68

References 69

Figures 75

Figure 2-1. The *ykkC* motif is widespread in bacteria 75

Figure 2-2. Subtype 2 *ykkC* motif RNAs associate with a distinct set of genes 76

Figure 2-3. A *ykkC* motif RNA responds to guanidine and controls the expression of a transporter that alleviates guanidine toxicity 77

Figure 2-4. Screening growth conditions for activation of a subtype 1 *ykkC* riboswitch reporter fusion yields guanidine as the sole valid hit 79

Figure 2-5. Guanidine selectively activates riboswitch reporter constructs 81

Figure 2-6. Ligand binding by a guanidine-I riboswitch aptamer 83

Figure 2-7. Selective recognition of guanidine by a representative guanidine-I riboswitch aptamer 85

Figure 2-8. Mutation of numerous individual conserved nucleotides disrupts guanidine binding 86

Figure 2-9. Ligand binding by a guanidine-I riboswitch aptamer from *K. pneumoniae* 87

Figure 2-10. Ligand binding by a guanidine-I riboswitch aptamer from *B. subtilis* . 88

Figure 2-11. Guanidine regulates riboswitch-dependent transcript termination 89

Figure 2-12. Guanidine regulates transcription termination of additional guanidine-I riboswitch representatives, but not a representative from another riboswitch class . 91

Figure 2-13. Guanidine-related functions for protein products of genes associated with guanidine-I riboswitches 92

Figure 2-14. Natural production of free guanidine by bacteria 94

Tables 96

Table 2-1. Sequences of synthetic DNAs used in Chapter 2. 96

CHAPTER THREE 109

Abstract 110

Introduction 111

Results and Discussion 113

Additional mini-*ykkC* motif RNA representatives form highly-conserved symmetrical loops 113

Genes associated with mini-*ykkC* are related to guanidine metabolism 114

Guanidine is bound by mini-*ykkC* motif RNAs with affinity similar to guanidine-I riboswitches 115

Mutation of conserved nucleotides confirms multiple guanidine binding sites 117

Both stem loops are necessary for guanidine binding 118

Guanidine-II binds its ligand selectively albeit with less stringency than guanidine-I 119

Concluding remarks 121

Materials and Methods 122

Chemicals and oligonucleotides 122

Bioinformatics analyses 123

RNA oligonucleotide preparation 123

RNA in-line probing analyses 124

Acknowledgements 125

References 125

Figures 127

Figure 3-1. Structure and function of guanidine-II riboswitches 127

Figure 3-2. Guanidine is bound by a guanidine-II riboswitch aptamer *in vitro* 128

Figure 3-3. Guanidine binding by the 60 *sugE* RNA from *E. coli* 130

Figure 3-4. Guanidine binding by the 60 *uca* RNA from *K. pneumoniae* 131

Figure 3-5. Mutations of conserved loop nucleotides affect structural modulation of both loops 132

Figure 3-6. Mutations in each stem loop affect guanidine binding independently at each binding site 133

Figure 3-7. Two singlet stem loops interact to accomplish guanidine binding 134

Figure 3-8. A guanidine-II RNA only recognizes chemical analogs with similar structures to its ligand 135

Figure 3-9. A guanidine-II riboswitch aptamer can recognize close structural analogs of guanidine 137

Tables 139

Table 3-1. Sequences of synthetic DNAs and RNA used in Chapter 3. 139

CHAPTER FOUR 144

Abstract 145

Introduction 146

Results and Discussion 147

Updated consensus model and expanded gene associations for *ykkC*-III motif RNAs 147

Guanidine is bound by *ykkC*-III motif RNAs with affinities similar to guanidine-I and -II riboswitch aptamers 149

Mutation of conserved nucleotides in guanidine-III riboswitch aptamers abolishes ligand binding 151

Selective ligand binding by guanidine-III aptamers 152

Concluding remarks 153

Materials and Methods 154

Chemicals and oligonucleotides 154

Bioinformatics analyses 154

RNA oligonucleotide preparation 154

RNA in-line probing analyses 154

Acknowledgements 155

References 155

Figures 157

Figure 4-1. Structure and function of guanidine-III riboswitches 157

Figure 4-2. Guanidine is bound by a guanidine-III RNA *in vitro* 158

Figure 4-3. Guanidine binding is unaffected by the deletion of the P0 stem 160

Figure 4-4. Guanidine binding by the 65 *arginase* RNA from *M. kansasii* 161

Figure 4-5. Guanidine binding by the 64 *smr* RNA from *S. bingchenggensis* 162

Figure 4-6. Guanidine binding by the 71 *emrE* RNA from *L. penumophila* 163

Figure 4-7. Mutations at conserved nucleotides prevent recognition of guanidine. 164

Figure 4-8. Guanidine-III aptamers can discriminate against even close analogs of guanidine 165

Figure 4-9. A guanidine-III RNA only recognizes chemical analogs with highly similar structures to its ligand 166

Tables 168

Table 4-1. Sequences of synthetic DNAs and RNA used in Chapter 4. 168

CHAPTER FIVE 174

Abstract 175

Introduction 176

Results and Discussion 177

Identification of ppGpp riboswitch candidates 177

Subtype 2a *ykkC* motif RNAs selectively bind ppGpp 178

Transcription termination regulated by a ppGpp riboswitch 180

Gene associations of ppGpp riboswitches 182
Some ppGpp riboswitches reside in tandem with T Box RNAs 184
Bioinformatic and biochemical data reveal additional riboswitch classes 185
Conclusions 185
Materials and Methods 186
Chemicals and reagents 186
Bioinformatics analyses 186
RNA oligonucleotide preparation 187
RNA in-line probing analyses 187
Single-round transcription termination assays 188
Acknowledgements 189
References 189
Figures 194
Figure 5-1. ppGpp riboswitches and their regulatory network 194
Figure 5-2. Ligand binding by a ppGpp riboswitch aptamer 195
Figure 5-3. Mutation of conserved nucleotides disrupt ppGpp binding 197
Figure 5-4. Recognition of pppGpp by a riboswitch aptamer 199
Figure 5-5. Binding selectivity of a representative ppGpp riboswitch 200
Figure 5-6. Mutation of a conserved nucleotide changes ligand specificity 202
Figure 5-7. ppGpp regulates riboswitch-dependent transcript termination 203

Figure 5-8. Ligand binding by an additional ppGpp riboswitch aptamer representative found upstream of glutamate synthase genes 205

Figure 5-9. Ligand binding by an additional ppGpp riboswitch aptamer representative found upstream of *natA* 207

Tables 209

Table 5-1. Sequences of synthetic DNAs used in Chapter 5. 209

CHAPTER SIX 215

Abstract 216

Introduction 217

Results 220

Variant riboswitch aptamers selectively bind the nucleotide precursor phosphoribosyl pyrophosphate (PRPP) 220

PRPP triggers riboswitch-mediated regulation of transcription termination 221

In vivo regulation of gene expression by a PRPP riboswitch 222

Natural examples of tandem riboswitch aptamers for guanine and PRPP bind these ligands to induce mutually-exclusive structures 223

Boolean logic function by a riboswitch carrying tandem guanine and PRPP aptamers 224

The biological utility of the opposing guanine and PRPP signals evaluated by tandem riboswitch aptamers 226

Expansion of natural Boolean logic devices made of RNA 228

Discussion 228

Materials and Methods 232

Chemicals and reagents 232

Bioinformatics analyses 232

RNA oligonucleotide preparation 233

RNA In-line Probing Analyses 233

Single-round transcription termination assays 233

Riboswitch reporter assays 234

Acknowledgements 235

References 235

Figures 240

Figure 6-1. PRPP riboswitches and their regulatory network 241

Figure 6-2. Ligand binding by a natural singlet PRPP riboswitch 242

Figure 6-3. Transcription regulation and reporter gene expression by a natural singlet PRPP riboswitch 244

Figure 6-4. Replicate *in vitro* transcription termination assays. 246

Figure 6-5. Ligand binding by tandem guanine and PRPP aptamers 248

Figure 6-6. Guanine and PRPP competitively regulate transcript termination via a riboswitch integrating two aptamers 249

Figure 6-7. Comprehensive dataset for the in vitro transcription termination assays with the *B. megaterium codA* riboswitch construct 251

Figure 6-8. Natural examples of tandem riboswitches that approximate Boolean logic functions 252

Tables 253

Table 6-1. Sequences of synthetic DNAs used in Chapter 6. 253

CHAPTER SEVEN 258

Introduction 259

Subtype 2c Riboswitch Candidates 261

Structural features of *ykkC* subtype 2c RNAs 261

Genes associated with *ykkC* subtype 2c RNAs 262

Speculations on a ligand for the *ykkC* subtype 2c candidate riboswitch class 263

Subtype 2d Riboswitch Candidates 264

Structural features of *ykkC* subtype 2d RNAs 264

Genes associated with *ykkC* subtype 2d RNAs 265

Speculations on a ligand for the *ykkC* subtype 2d candidate riboswitch class 266

Conclusions 267

References 273

Figures 277

Figure 7-1. Additional unsolved variant *ykkC* motif RNAs. 277

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my adviser Ronald Breaker. I'm sure Ron doesn't remember this, but I first met him when I interviewed at Yale for graduate school. His descriptions of riboswitch mechanisms and abundance, the variety of functions performed by RNAs, and his vision for the future of noncoding RNA discovery strengthened my desire to continue to pursue nucleic acids research. I greatly appreciate the feedback he has provided in terms of troubleshooting experiments as well as extensive editing of my writing and figures. The more valuable skills I have learned from Ron are how to prioritize the most important experiments, pursue the most interesting biology, and strive to answer scientific questions with the greatest possible significance.

I am thankful for my fellow Breaker lab members who played a huge role in my success as a graduate student. During my rotation and after I joined the lab, I received a huge amount of help from Christina Lünse in learning the technical aspects, experimental design, and trouble-shooting of in-line probing. Once I started producing data, I always looked to Adam Roth for insight and advice. Convincing Adam was usually a challenge, but I embraced the challenge and always aimed to produce enough data to get him on board. I am also extremely grateful to Narasimhan Sudarsan, whose amazing knowledge of bacterial metabolism immensely improved my understanding of the pathways I was interested in studying. I am extremely grateful for the advice and encouragement I have received from Kim Harris, who could provide an outside perspective on my project. I deeply value our friendship and her mentorship, which has been incredibly empowering to me at this early stage of my career.

Additionally, I would like to thank the Breaker lab members who directly contributed to the research presented in this thesis: Shira Stav, Ruben Atilho, Sarah Malkowski, and

James Nelson. Shira's bioinformatic expertise was imperative for the analyses that led to the breakthroughs in Chapters 5 and 6. Ruben was always willing to share his expertise in designing and creating bacterial reporters. Sarah was eager to help when I asked if she wanted to be involved in the project in Chapter 3, and together we accomplished a rapid (in fact, the fastest to my knowledge) validation and publication of a new riboswitch class. James took a chance in letting a brand-new graduate student take over some crucial experiments when the guanidine discovery was first made. I am extremely grateful to have been included in what I thought would be a short starter project which turned out to be the entirety of my thesis. I collaborated with or sought advice from a number of other awesome Breaker lab members over the years, including Zasha Weinberg, Gayan Mirihana Arachchilage, Kevin Perkins, Etienne Greenlee, and Keith Corbino.

My research has benefited tremendously from other professors and students in the MB&B department, which fosters a wonderfully collaborative community. In particular, Caroline Reiss and Chris Lim, with the guidance of their respective mentors Scott Strobel and Yong Xiong, performed experiments that were instrumental to our understanding of the riboswitch classes I worked on. The members of my thesis committee, Susan Baserga and Matt Simon, have helped at each stage of my graduate research and I am grateful for their input and advice as I transition to the next stage of my career. I know I can trust Susan to tell me what I need to hear, even if it may not be what I want to hear, and I value her candor. Matt has been such a great mentor to me in so many additional roles, including as a course instructor, rotation mentor, fellow instructor for my teaching duties, and has helped me become a more confident scientist.

My fellow MB&B classmates have been such an incredibly supportive and fun group. It seems like we all instantly bonded in our first year when we were in classes all the time and we have all stuck together through the years. Some of my favorite memories are the MB&B retreats at Woods Hole and Jiminy Peak, camping trips to Acadia, the Adirondacks, and the Catskills, skiing in Vermont, countless potluck brunches and dinners, and nights at GPSCY. We talk about science way too much outside of lab, but I wouldn't have it any other way.

Lastly, I want to thank my family members that have been there every step of the way for my entire life. My parents, Susie and Phil, are truly the most loving pair I could ask for and I know I have their trust and support to pursue anything I want to accomplish. My brother Nevin is the first person in my life who I watched develop into a scientist and pursue a Ph.D. and I admire his thoughtfulness and intelligence. My brother Carl, who survived and recovered from cancer and numerous complications, shows remarkable strength and determination that I can only attempt to strive for. My sisters-in-law, Melissa and Alyssa, both compliment my brothers so well to round out our family in the best way possible, and my nephew Sebastian and niece Scarlett bring so much happiness to our lives. I love you all!

CHAPTER ONE

Introduction

RIBOSWITCHES

Riboswitches reside in the non-coding portion of mRNAs and bind their corresponding ligand to control gene expression.[1-3] Riboswitch RNAs are typically located in the 5′ untranslated region (UTR) directly preceding the coding region(s) whose expression they regulate.[2,4] To accomplish this *cis*-regulatory role, the riboswitch sequence is always on the same strand and oriented in the same direction as the coding region it controls. Riboswitches occupy a substantial portion of the intergenic region in which they reside and typically begin at least a few nucleotides downstream of the promoter sequence (or upstream gene when located within an operon) and can stretch all the way to the start codon of the adjacent open reading frame (ORF).[5,6]

Riboswitches are unique compared to other regulatory RNAs, such as sRNAs, ribosomal protein leaders, and T box RNAs, in that they directly bind a small molecule or ion ligand to control gene expression levels.[7-10] The relative abundance of this ligand is responsible for determining the state of expression for downstream genes. As such, the identity of the cognate ligand for each riboswitch class is inherently linked to the genes under its regulation.[11]

LIGANDS SENSED BY RIBOSWITCHES

Ligands that have been identified for riboswitch classes include coenzymes, nucleotide derivatives, amino acids, sugars, and ions (Figure 1-1A).[12] More than half of the known and validated riboswitch ligands are nucleotide derivatives or contain nucleotide moieties (Figure 1-1B).[12] Coenzymes that have been validated as riboswitch ligands are coenzyme B_{12} (adenosylcobalamin, AdoCbl) and its derivative aquacobalamin (AqCbl),[13-15] thiamin

pyrophosphate (TPP),[16] flavin mononucleotide (FMN),[17] *S*-adenosylmethionine (SAM)[6,18-23] and *S*-adenosylhomocysteine (SAH),[23,24] molybdenum and tungsten cofactors (MoCo and Wco),[25] and tetrahydrofolate (THF).[26] All of these coenzyme ligands are synthesized from nucleotide precursors and some of these ligand molecules contain an adenosine moiety as part of their structure.

The signaling molecules sensed by riboswitches include cyclic di-adenosine monophosphate (c-di-AMP),[27] cyclic di-guanosine monophosphate (c-di-GMP),[28] and cyclic adenosine monophosphate-guanosine monophosphate (c-AMP-GMP),[29] which are all circular dinucleotide compounds with 3′-5′ phosphodiester linkages. In addition, the nucleotide precursor aminoimidazole carboxamide ribonucleotide (AICAR or ZMP),[30] the queuosine precursor $preQ_1$ (derived from GTP),[31-33] guanine,[34] adenine,[34,35] and deoxyguanosine (2′-dG)[36,37] complete the current list of ligands with RNA-derived components.

It has previously been hypothesized that this strong correlation between riboswitches and RNA-derived ligand partners is a result of their RNA World origins.[38-40] The RNA World theory includes the idea that life on earth passed through a stage wherein organisms were comprised predominantly of RNA.[41,42] If RNA World organisms could manage even a very simple metabolic state, they might have utilized ancient versions of modern riboswitches to accomplish regulatory tasks. These ligand-binding RNAs also could have functioned as the substrate- or cofactor-binding portion of a ribozyme complex. Although riboswitches are not assuredly relics of the RNA World, they are certainly very ancient in origin given the extremely widespread nature of the most abundant riboswitch classes as well as the nature of their corresponding ligands.[38,39,43] In some cases, riboswitch functions

may have been replaced by protein factors, and riboswitches could have been even more common in the ancestors of present-day bacteria.

Additional riboswitch classes sense ligands that are not structural derivatives of RNA molecules. While some of these compounds might have been relevant to RNA World organisms and their corresponding riboswitch classes might also be of ancient origin, others may have appeared later in evolutionary history. Elemental ions that have been validated as riboswitch ligands are Mg^{2+} (refs 44,45), Mn^{2+} (refs 46,47), Ni^{2+} and Co^{2+} (ref 48), and F^- (ref 49). Amino acids and sugars that have been validated as riboswitch ligands are lysine,[50] glycine,[51] glutamine,[52] and glucosamine 6-phosphate (GlcN6P).[53]

Biological processes controlled by riboswitch ligands. In many cases, riboswitches that bind a certain small-molecule ligand will control expression of biosynthesis genes for production of that ligand. For example, riboswitches for the coenzyme SAM are often upstream of the *metK* gene, which encodes SAM synthetase.[18-23] Accordingly, when sufficient SAM levels are achieved in the cell, the riboswitch binds its ligand and shuts off the SAM biosynthesis pathway. Riboswitches that sense signaling molecules might associate with a wide variety of genes related to the regulatory response network of that particular signaling molecule. In the case of riboswitches for the ions F^-, Ni^{2+}/Co^{2+}, and Mn^{2+}, which are lethal to bacteria at elevated concentrations, some of the downstream genes have functions in export to achieve detoxification.[46-49]

Due to the diversity in the processes controlled by riboswitches as well as in the identities and roles of riboswitch ligands, the concentration of each ligand necessary to alter gene expression varies greatly. Some ligands, such as glutamine, are highly abundant under most conditions with a cellular concentration of ~4 mM in *Escherichia coli* cells.[54]

On the other hand, a signaling molecule such as c-di-AMP is generally not present in cells unless its dedicated synthase is produced during various stress conditions.[55,56] The *in vitro* affinity of glutamine riboswitches for their ligand of ~1 mM[52] is vastly different that the picomolar affinity of c-di-AMP riboswitches for their ligand.[27] While *in vitro* riboswitch affinities vary greatly, the *in vivo* response for each class is tuned to accomplish regulation within the relevant cellular concentration range of each ligand.

Riboswitch class abundance. The vast diversity in the types of ligands sensed by riboswitches is matched by the variance in their commonality. The TPP riboswitch class is the most abundant, as well as the only class found in eukaryotic organisms, with >15,000 examples in the Reference Sequence (RefSeq) version 56 database (Figure 1-2).[12] Overall, the ligand with the most riboswitch examples is SAM, which is sensed by five different riboswitch classes.[18-23] Additionally, only the c-di-GMP riboswitch has been found in a bacteriophage[28] while F^- riboswitches have been identified in archaea.[49]

Most riboswitches sense compounds that are ubiquitous in bacteria, making it difficult to provide a definitive rationale for why certain classes are much more abundant than others. Evolutionary pressures, including the emergence of proteins, undoubtedly played a role in shaping the utility and abundance of modern riboswitch classes. Many more riboswitch classes are hypothesized to exist that have not yet been discovered due to their relative scarcity in nature.[1,12] Currently, most of the exceedingly rare known riboswitch classes are variants of more abundant known classes,[36,37] or control the same types of genes and therefore sense the same ligand as abundant known riboswitch classes.[19,23,33]

RIBOSWITCH STRUCTURE

Riboswitches can function as an "on" or "off" switch in a manner dictated by the ligand they sense and the nature of the downstream genes.[2] For example, high concentrations of SAM in the cell shuts off SAM biosynthesis genes[18] whereas high concentrations of fluoride in the cell turn on genes encoding fluoride exporters.[49] To regulate gene expression in a ligand-dependent manner, the riboswitch RNA must be able to specifically recognize the small molecule ligand and interact with the cellular machinery.[2] Riboswitches comprise two domains to accomplish these two tasks: a ligand-binding aptamer region and an expression platform that affects transcriptional or translational processes.

Riboswitch aptamers. The portion of the riboswitch RNA involved in ligand recognition is termed the aptamer domain.[2] Before the discovery of riboswitches, artificial RNA sequences were enriched from large libraries of sequences to isolate those with the ability to specifically bind a certain target molecule. Riboswitches were the first small-molecule binding RNAs to be found in nature. Riboswitch aptamers consist of conserved structural features as well as nucleotides with specific conserved identities.[59] Typically, the most highly-conserved nucleotides reside in and around the ligand binding pocket with roles ranging from hydrogen bonding with the ligand to coordinating metals for charge stabilization.[59,60]

In some cases, only a handful of nucleotides actually contact the ligand molecule in an aptamer consisting of >100 nucleotides. Creating a binding pocket specific enough to bind the cognate ligand with sufficient affinity while discriminating against highly similar molecules requires the support of these additional nucleotides. Although the identity of these nucleotides may not be conserved, they are crucial to the aptamer in that they ensure

the binding pocket forms with a specific geometry to solely accommodate the cognate ligand. Many of the nucleotides create the overall scaffold of the structure by forming a canonical helix, usually as part of a stem loop or pseudoknot. The identity of some nucleotides is unconstrained where the nucleotide plays no role other than to connect two stems or form a stem loop distal to the binding pocket. Riboswitch aptamers use some combination of multi-stem junctions, internal loops, pseudoknots, and tertiary interactions to set up and support the binding pocket. Therefore, highly-conserved nucleotides typically reside in loops, bulges, and multi-stem junctions.

The aptamer portion of a riboswitch can vary greatly in length and complexity.[12] Both AdoCbl[13] and lysine[50] riboswitch aptamers are typically around 200 nucleotides in length whereas preQ$_1$-I riboswitches can be as short as 35 nucleotides.[31] Accordingly, the preQ$_1$-I riboswitch class has a very simple structure involving just a single stem loop and pseudoknot. At the other extreme for complexity are FMN riboswitches,[17] which contain a multi-stem junction involving six stem loops, and AdoCbl riboswitches,[13] which contain three separate multi-stem junctions.

The above observations are based on the overall averages for each riboswitch class, but the individual representatives within each riboswitch class can also vary greatly. Some loops, stems, and bulges are highly variable in their lengths, leading to significant differences in the length of individual examples of the same riboswitch class aptamer in different organisms.[59] The overall primary sequence conservation of a riboswitch aptamer can be relatively low between examples from different species as long as the most highly-conserved nucleotides remain and the secondary and tertiary structural elements are able to form properly. Comparing all of the riboswitch representatives from a certain class at

the primary sequence level is only of limited utility whereas creating a consensus model to account for structural features is more informative of the architecture of the riboswitch class aptamer.

The consensus sequence and secondary structure model for the guanine riboswitch aptamer[34] is presented in Figure 1-3A. Of the ~70 nucleotide that typically comprise the aptamer, only 15 have greater than 97% conservation of their nucleotide identity (indicated by red letters). Approximately 30 nucleotides are found in conserved stem structures, termed pairing elements 1 through 3 (P1-P3). These stems are not conserved at the primary sequence level but rather require complementarity of bases at each position. This type of variability likely occurs through random changes through evolution such that when a nucleotide on one side of the stem mutated, its partner on the other side of the stem is forced to mutate as well to restore the base pairing interaction. Base pairs for which sequence variability is tolerated as long as complementarity is retained display a characteristic termed covariation, which is indicated on the consensus model by green shading. In other cases, such as the long-range interaction between the P2 and P3 loops of the guanine aptamer, the identities of the nucleotides involved in base pairing are highly conserved. The remaining nucleotides in an aptamer can also form base pairs, though sometimes non-canonical, or form non-conserved linkers and loops.

Variant riboswitch aptamers. Although each riboswitch class binds its cognate ligand with sufficient specificity as to exclude related compounds, some riboswitch class aptamers have evolved to switch their specificity and sense a new cognate ligand. Perhaps the best example of this type of specificity switch is the adenine riboswitch class (Figure 1-3B).[34,35] The overall structural architecture of the adenine riboswitch class (P1-P3) is identical to

the guanine riboswitch class and all except one of the highly-conserved nucleotides of the guanine aptamer are conserved in the adenine aptamer. This nucleotide position, denoted with an asterisk in Figure 1-3, is a cytidine for guanine aptamers and a uridine for adenine riboswitches. Accordingly, high-resolution structures of these aptamer RNAs solved by crystallography revealed that this nucleotide position in the aptamer is involved in Watson-Crick base pairing with the cognate ligand.

In all cases of variant aptamers with altered ligand specificity, the binding pocket changes are accompanied by a change in downstream gene association. Guanine riboswitches typically regulate purine biosynthesis genes while adenine riboswitches regulate genes coding for adenine-specific processes such as adenosine deaminase. Two additional variant classes of purine riboswitches exist that specifically sense 2′-dG,[36,37] with each containing separate sets of subtle mutations as compared to the binding pocket of guanine riboswitch aptamers. Additional examples of variant riboswitch class pairs include MoCo and WCo,[25] AdoCbl and AqCbl,[13-15] and c-di-GMP and c-AMP-GMP.[28-29]

Riboswitch expression platforms. The second domain of a riboswitch, termed the expression platform, dictates gene expression based on aptamer occupancy.[2] In contrast to the aptamer domain, riboswitch expression platforms are not always highly conserved. Riboswitch aptamers are stabilized by their cognate ligand, which can either promote or inhibit formation of structured elements in the expression platform. Riboswitch-mediated regulation typically occurs either during transcription, by influencing the stability of an intrinsic terminator stem, or during translation initiation, by exposing or sequestering the ribosome binding site and/or the start codon. Other rare gene control mechanisms also exist, including ribozyme-mediated transcript cleavage for the GlcN6P-sensing *glmS*

riboswitch-ribozyme[53] and allosteric c-di-GMP-I/group-II ribozyme[61] as well as alternative mRNA splicing for TPP riboswitches in eukaryotes.[62,63]

Transcriptional regulation by riboswitches. Bacteria commonly use intrinsic terminator stems to halt RNA polymerase at the desired end of a transcript.[64,65] Terminator stems, which are enriched for GC base pairs to maintain a high level of stability, are directly followed by multiple uridine nucleotides. The transcriptional machinery pauses upon encountering these consecutive uridines and formation of the terminator stem within the RNA causes the transcript to dissociate from the template DNA. These terminators have been usurped by riboswitches such that their formation depends on binding of the cognate ligand.[2,4,59] Transcriptional "on" switches create an interaction between the aptamer and expression platform such that the terminator stem cannot form in the ligand-bound form (Figure 1-4A). This is typically accomplished by involving part of the terminator sequence in base pairing within the aptamer in a stem termed the anti-terminator. Conversely, transcriptional "off" switches allow the terminator to form when the cognate ligand is bound (Figure 1-4B). In this case, an anti-terminator stem is only stable in the ligand-free state and is destabilized upon ligand binding. Gene expression is decreased by inhibiting the ability of RNA polymerase to transcribe beyond this intrinsic terminator stem in the 5′ UTR. Consequently, mRNA for the downstream gene is not produced and therefore the protein cannot be translated.

Translational regulation by riboswitches. Riboswitches also commonly affect gene expression at the post-transcriptional level by blocking or allowing access of the ribosome to its binding site and the start codon of the ORF.[2,4,59] This regulation is typically accomplished by selectively sequestering the purine-rich ribosome binding site (RBS)

sequence[66] based on the occupancy of the aptamer by its ligand. Translational "on" switches must create an interaction between the aptamer and expression platform such that the RBS is exposed in the ligand-bound form (Figure 1-4C). In this case, an anti-RBS stem is stable only in the ligand-free state and is destabilized upon ligand binding. Conversely, translational "off" switches allow the anti-RBS stem to form when the cognate ligand is bound (Figure 1-4D). This is typically accomplished by involving the RBS sequence and occasionally also the start codon in base pairing within the aptamer. In this case, the riboswitch only directly affects the ability of the cell to produce the protein. Within the lifetime of the transcript, ligand levels could fluctuate to switch between the two regulatory states.

TANDEM RIBOSWITCH ARCHITECTURES

A typical riboswitch binds a single ligand and exerts its regulation over approximately a 100-fold concentration range of its cognate ligand. This dynamic range is sufficient for many molecules and signals in the cell. However, in some cases, more sophisticated regulation is required. This can be accomplished through protein factors, but riboswitches and other RNA elements can also conspire to create more complex regulatory networks.

Cooperative riboswitches. Cellular concentrations of certain metabolites and signaling molecules experience subtle fluctuations and the requirement for a 100-fold change in concentration to achieve switching of gene expression is either undesirable or unobtainable. Therefore, the dynamic range of the riboswitch affinity for its ligand must decrease to switch over a narrower range of ligand concentration. The most efficient way to accomplish this is by binding multiple ligands cooperatively, resulting in a steeper transition between the unbound and bound state. Riboswitches can create cooperative

systems through three main strategies: 1) a single aptamer with multiple binding sites, 2) multiple identical aptamers with a shared expression platform, and 3) multiple adjacent independent riboswitches. Natural examples of all three of these strategies have been observed, although additional cooperative riboswitch systems likely exist in nature.

The first strategy is exploited by the c-di-AMP,[27] THF,[26] NiCo,[48] and Mg^{2+}-I[44] riboswitch classes, all of which have at least two ligand binding sites for the same ligand type. Both the THF[26,67] and c-di-AMP[27,68] riboswitches were revealed by their high-resolution crystal structure to contain two binding sites for their small molecule ligands. The metal-binding NiCo and Mg^{2+}-I riboswitch classes have four and six ligand binding sites, respectively, and display highly cooperative binding behavior.[44,48]

The second strategy is exploited by the glycine riboswitch class, which is found in a tandem aptamer arrangement with a shared expression platform in 96% of known examples.[51,69] These two highly-similar consecutive aptamers interact to accomplish interdependent structural stabilization, which supports the ligand-binding pocket in each aptamer. Singlet glycine aptamers have also been identified which contain a non-ligand binding structure that provides necessary supportive contacts for the single aptamer.[69]

The third strategy is exploited by many riboswitch classes, as it is the easiest arrangement to achieve since each riboswitch module functions independently. In this case, a narrower response range to the ligand is achieved at the level of overall gene expression, but each riboswitch itself does not display cooperative binding. Common riboswitch classes, such as TPP,[16] AdoCbl,[13] Mn^{2+} (ref 46,47), Mg^{2+}-I (ref 44), c-di-GMP,[28] and glutamine,[52] have been observed adjacent to one another in the same 5′-UTR.[70] In some

cases, the glutamine riboswitch can be found with three adjacent copies to sample an even narrower dynamic range of ligand concentration.[52]

For some of the ligands sensed by cooperative riboswitch systems, such as Mg^{2+}, cellular concentrations are very high and do not vary drastically.[71] Therefore, riboswitches that use these metabolites and ions as regulatory inputs have adapted to sense only modest fluctuations in ligand concentration. While cooperative ligand binding achieves a slightly more complex regulatory state, gene expression often relies on multiple regulatory factors that involve a variety of inputs and signals at the promoter, transcriptional, and translational level. Riboswitches and other regulatory RNAs can come together to create more complex systems that involve more than one ligand.[70]

Riboswitches as two-input logic gates. Two-input systems comprised solely of regulatory RNA elements can use the same three strategies as detailed above for cooperative systems. The first strategy, which would involve a single aptamer domain that senses two different ligands, has never been observed. It could be argued that Mg^{2+} ions are necessary for ligand recognition for the F^- and Mn^{2+} (and other) riboswitch classes,[47,72] but due to the high concentration and relatively limited fluctuation of Mg^{2+} levels *in vivo*, only one ligand is truly responsible for gene regulation. Although this strategy is theoretically possible to achieve and might exist in nature, its utility is limited compared to a typical riboswitch with one cognate ligand because only a subset of genes overlaps between the regulatory network of each of the two individual ligands. For this and perhaps other reasons, the other strategies are more practical and indeed have been observed in nature.

The second and third strategies involve two separate entities that sense the concentrations of two different ligands to make the singular decision of whether the downstream should be expressed or not. Boolean algebra defines all sixteen possible outcomes when two inputs are combined to produce a single output.[73,74] Boolean two-input logic gates are fundamental to circuitry and computational languages and can be depicted using truth tables. If one considers each riboswitch ligand to be an input, the pattern of gene expression output can be summarized using a truth table corresponding to one of the sixteen Boolean logic functions.[70]

The second strategy involves two tandem aptamers that share a single expression platform. Such an arrangement has been observed[70] for the guanine riboswitch[34] and the *ykkC* riboswitch candidate,[5] which will be described in great detail in Chapter 6. This tandem arrangement is predicted to conform to the Boolean IMPLY gate.[70] Additionally, examples of glycine riboswitches have been observed in which one aptamer is a canonical glycine-binding aptamer and the other is a variant glycine aptamer that does not bind glycine.[37] In this case, the system would no longer be cooperative because two different ligands are sensed. Although the ligand for the glycine variant is unknown at this time, it is likely that both ligands will be necessary to turn on gene expression, making a Boolean AND gate.

The third strategy is again the most common as it employs the use of completely separate RNA entities. The first natural tandem riboswitch system to be discovered[70] consists of two "off" switches: a SAM-I riboswitch[18] followed by a AdoCbl riboswitch.[13] In this case, the tandem arrangement ensures the gene is not expressed when either ligand is present, therefore constructing a NOR logic gate.[70] Another tandem arrangement

involves a ZTP riboswitch[30] ("on" switch) followed by a THF riboswitch[67] ("off" switch) which creates an "A AND NOT B" Boolean logic function. Additionally, tandem arrangements have been observed involving a known riboswitch class and an unsolved riboswitch candidate that are likely to mimic a Boolean function, but the unsolved candidates await ligand validation to evaluate the logic of the tandem system.

RIBOSWITCH DISCOVERY

The first natural riboswitches to be reported were identified using genetic evidence.[13,16] This approach relied on analyzing genes for which metabolite-dependent regulation involving the 5′ leader sequence had been demonstrated, but no protein factor had been identified. For example, the AdoCbl riboswitch class, the first to be reported, was identified after reports of unexplained gene control of the *btuB* mRNA in *E. coli*.[13,75] While this method was instrumental in identifying the first riboswitch classes, including TPP,[16,76] FMN,[17,76] SAM,[18,77,78] guanine,[34] and lysine,[50,79] there were also severe limitations to its utility. Identifying riboswitch classes in this manner requires them to be present in well-studied model organisms, such as *E. coli* or *Bacillus subtilis*, and furthermore requires prior knowledge of the regulation of specific genes. To overcome these limitations, bioinformatic strategies were developed to identify riboswitch sequences.[5]

Creating an alignment of multiple individual sequence representatives of a particular riboswitch class reveals its most vital sequence and structural components. Highly-conserved nucleotide positions as well as base pairing interactions can be modeled onto the primary sequence, which can then be used to establish a consensus sequence and secondary structure model. Subsequent to the identification of the first riboswitch examples, these

alignments and consensus models were used to search genome databases for additional representatives whose sequence conform to the consensus and genetic context was consistent with a *cis*-regulatory function. As search algorithms were improved, this method was used by Barrick *et. al.* to identify entirely new riboswitch classes by using comparative sequence analysis of notably long intergenic regions.[5] From that point on, almost all riboswitch class discoveries have relied on some form of comparative sequence analysis.[6,20,22,80,81] As mentioned above, some riboswitch classes have altered their ligand specificity by mutating specific nucleotides in and around the binding pocket.[35-37] As such, these relatively rare variant riboswitches were discovered through sequence or experimental analysis of the parent riboswitch class.

Perhaps the most notable finding from this first round of searches using comparative sequences analysis was the discovery of multiple riboswitch candidates for which the identity of their cognate ligand was not immediately clear.[5] From a structural perspective, the main criteria for an RNA motif to be considered a riboswitch candidate are conserved structural elements that exhibit covariation of base pairing, a mix of conserved nucleotide identities found in stems, loops, linkers, and bulges, and potentially some more complex structural features such as multi-stem junctions, pseudoknots, and conserved internal loops. However, these criteria are consistent with other types of structured RNAs, such as ribozymes and other large non-coding RNAs. Therefore, one of the key distinguishing factors as to whether an RNA is considered a riboswitch candidate is the genetic context. Riboswitches are *cis*-regulatory and should therefore consistently appear in the correct orientation upstream of the same genes or gene families in multiple organisms. Many *cis*-regulatory RNA elements are not riboswitches, but these motifs usually rely on interactions

with a protein factor and therefore do not exhibit the structural complexity necessary to specifically recognize a small-molecule ligand. Once a motif is determined to possess all of the necessary features to qualify as a riboswitch candidate, the remaining challenge is to identify its cognate ligand. Understanding the function of the genes associated with a particular riboswitch candidate as well as how different types of genes under its regulatory network relate to one another is crucial to determining the ligand identity.

For the first reported riboswitch classes, metabolite-based regulation was already known, making the ligand identity quite obvious.[13,16-18] For some riboswitch candidate motifs found by the first comparative sequence analyses,[5] such as the glycine riboswitch class,[51] a candidate ligand was chosen based on the involvement of the downstream genes in a particular metabolic pathway leading to a straightforward validation. However, other motifs proved to be extremely difficult to validate and the effort to solve the identity of the cognate ligand took many years or is still ongoing.[5,82,83] Because the ligand identity is unknown, each candidate riboswitch RNA motif is instead named after the *B. subtilis* gene(s) it is predicted to regulate. When a candidate riboswitch RNA does not have any representatives in *B. subtilis*, it is instead named after the gene, gene groups, or types of genes that are most commonly found downstream of the motif. RNA motifs with a high likelihood to function as a riboswitch but whose cognate ligands have not been identified are collectively termed orphan riboswitch candidates.[1,5,82,83,84]

ORPHAN RIBOSWITCHES

Each orphan riboswitch candidate presents unique challenges which, for the most part, can only be fully appreciated after the ligand has been identified. When the processes

predicted to be regulated by a riboswitch candidate are not well understood, identifying the correct ligand becomes extremely difficult. Orphan riboswitch candidates have largely remained unsolved due to a lack of information, or misleading information, about the function of their associated genes. However, experimental issues have prevented other orphan riboswitch candidates from successful ligand validation.

Only a handful of factors cause a promising RNA motif to become an unsolved orphan riboswitch candidate, but the strategies that led to successful identification of a ligand for former orphan riboswitches are much more diverse and unique to each class. No single strategy worked, or could have worked, for every orphan riboswitch class. In some cases, the key piece of missing knowledge in terms of the biology related to the regulon of the riboswitch class was gained before the ligand identity could finally be solved. However, in the cases of the classes for which the ligand was fortuitously discovered through other means, the key insight into these unknown pathways and processes was gained by solving the orphan riboswitch. The biology learned through the validation of orphan riboswitch classes provides ample motivation to continue studying these challenging and mysterious RNAs.

***ykoK* RNAs: Mg^{2+} riboswitches.** The *ykoK* riboswitch candidate originated from the first report of orphan riboswitches discovered by comparative sequence analysis by Barrick *et. al.* in 2004.[5] This riboswitch class primarily controls the expression of cation transporters, especially those for Mg^{2+}. The structure of this RNA motif is highly elaborate and its length rivals the length of the AdoCbl riboswitch class (>200 nucleotides).[13] In fact, the complexity of the *ykoK* RNA structure hindered its validation as a Mg^{2+} riboswitch in that it seemed unlikely for an RNA to require so many nucleotides to orchestrate a

structural reconfiguration caused by a cation that affects the structure of nearly all RNAs. In this case, the technical challenges of proving that Mg^{2+} binding alone causes these regulatory changes and that the RNA structure contains specific binding sites for this ion prevented immediate validation.

Most typical *in vitro* experiments used in riboswitch validation, including in-line probing[85,86] and *in vitro* transcription termination,[87] use high levels of Mg^{2+} due to its well-established role in RNA structure stabilization.[88] Nearly every RNA changes its shape in a test tube based on the Mg^{2+} concentration and therefore this fact alone is insufficient to conclusively validate a Mg^{2+}-sensing riboswitch. With this added burden of proof, the validation of the *ykoK* motif as the Mg^{2+}-I riboswitch class required a combination of many techniques including in-line probing, SHAPE probing, *in vitro* transcription termination, *in vivo* riboswitch-*lacZ* reporter fusions, analytical ultracentrifugation, and a high-resolution crystal structure.[44] These studies revealed that the aptamer binds many Mg^{2+} ions in a highly cooperative fashion, which helps explain the necessity for such a complex aptamer structure. Overall, the validation of Mg^{2+} as the ligand for the *ykoK* orphan riboswitch class provided the first proof of an elemental ion instead of a small molecule metabolite to be sensed by a riboswitch aptamer.

***ykvJ* RNAs: preQ$_1$ riboswitches.** Due to the relatively small size and simple structure of the *yvkJ* RNA motif, it was initially considered to only be a weak candidate for a riboswitch class.[5] At the time, most known riboswitch aptamers consisted of multiple stems and higher order structures and >100 nucleotides in length[2] whereas *ykvJ* RNAs can be just 34 nucleotides in length and consist of a single stem and pseudoknot. Despite its simple structure, the motif still possesses many qualities that made it a likely *cis*-regulatory

element. However, the unknown function of the genes it typically controls, which in most cases are homologs of the *B. subtilis ykvJKLM* operon, delayed the identification of a small molecule ligand for this RNA.

Riboswitch ligand validation is typically dependent on examination of the function of the downstream gene products. The lack of information available for the genes predicted to be controlled by *ykvJ* riboswitch candidates prevented the application of traditional lines of reasoning used to identify potential ligands for this promising RNA motif. The association of a riboswitch candidate with genes of unknown function almost invariably leads to orphan status. Fortunately for this candidate, the role of the *ykvJKLM* operon in the biosynthesis of the queuosine precursor $preQ_1$ from GTP was subsequently reported.[89] It was then demonstrated that *ykvJ* RNAs specifically recognize $preQ_1$ to turn off this biosynthetic operon when sufficient levels of the queuosine precursor are present.[31] This was the first of three known $preQ_1$ riboswitch classes to be identified.[31-33] The $preQ_1$-I riboswitch class is an additional example of the frequent riboswitch partnerships with ligands that are nucleotide-derived compounds and also provided evidence that some extremely simple RNA structural motifs can also selectively and independently bind a small-molecule ligand to regulate gene expression.

GEMM RNAs: c-di-GMP riboswitches. The GEMM (Genes for the Environment, for Membranes and for Motility) RNA motif was the first riboswitch candidate with one of the most confounding features of several orphan riboswitch classes: many different types of associated genes with a wide variety of functions that initially appear to be involved in unrelated processes. At the time of their discovery, GEMM RNAs were predicted to control the expression of genes encoding proteins with vastly different functions, although many

had likely roles related to the membrane or cellular motility.[80] The consensus model of the riboswitch aptamer as well as its predicted switching that involves a downstream terminator stem were all consistent with riboswitch function.

Although c-di-GMP was a known signaling molecule and some of its synthases, degradases, and downstream targets had been identified,[90-92] the c-di-GMP regulon was not well understood when the GEMM RNA motif was discovered. As such, the c-di-GMP riboswitch was the first receptor to be reported which acts as a global regulator using this second messenger.

Once the ligand had been identified for this riboswitch class, it immediately expanded our knowledge of c-di-GMP signaling. Bioinformatic data on the gene context of riboswitch examples is extremely powerful in that the gene directly downstream of each example of a riboswitch must somehow relate to the biological role of the cognate ligand. Additionally, each organism with an example of a certain riboswitch class must synthesize, import, or otherwise encounter the cognate ligand. In this case, solving the c-di-GMP riboswitch class provided insight into the numerous genes involved c-di-GMP signaling as well as the wide variety of bacterial species that utilize this signaling molecule.

In addition to the knowledge gained about bacterial c-di-GMP signaling, this riboswitch class was the first to be discovered in bacteriophages. This implies that the ability of viruses to sense the metabolites previously reported as riboswitch ligands is far less crucial or advantageous than their ability to sense a signaling molecule produced by the bacteria in relation to their external environment. It was also later discovered that rare variants of c-di-GMP riboswitches have altered their specificity to sense another type of circular nucleotide signaling molecule, cyclic-AMP-GMP.[29]

***crcB* RNAs: F^- riboswitches.** First reported in 2010, the *crcB* motif appeared to be a great candidate for a riboswitch aptamer with plenty of covarying base pairs, a pseudoknot, an optional stem, and a good mix of conserved nucleotides.[20] Previously, only TPP riboswitches were known to exist outside of the bacterial domain of life,[93] but *crcB* motif RNAs were found to be widespread in archaeal species as well as bacteria.[20] At first, it appeared that the major obstacle to identify the ligand for *crcB* RNAs would again be the wide variety of gene functions predicted to be under the regulatory network of this riboswitch candidate. However, the larger obstacle was one that plagued multiple former orphan riboswitches and likely continues to impede progress of riboswitch ligand validation for many current orphans: genes whose functions are mis-annotated or ambiguous.

The *crcB* motif was predicted to control the expression of genes encoding K^+ and Cl^- transporters as well as DNA repair proteins.[20] After the identification of the c-di-GMP riboswitch class, which also regulates a wide variety of processes, it seemed plausible that the *crcB* motif might also be involved in a signaling network mediated by a small-molecule second messenger. To test the hypothesis that this signaling molecule might be another nucleotide derivative, a range of synthetic dinucleotides were evaluated by in-line probing for binding by the *crcB* aptamer. Remarkably, many of these compounds modulated the RNA structure, but only when the molecules were obtained from one commercial source and not another.[49] Eventually, it was determined that the dinucleotide compounds themselves had no effect on the RNA, rather a contaminant produced during their chemical synthesis, tetra-n-butylammonium fluoride, was responsible for this signal. When tested in

other forms, such as KF and NaF, it was confirmed that fluoride was the cognate ligand for the *crcB* orphan riboswitch.[49]

In this case, the biological breakthrough from solving this RNA motif came after this serendipitous discovery. The *eriC* gene, which is commonly associated with the *crcB* RNA motif, encodes a protein that had been demonstrated to transport Cl^- in *E. coli*. However, the EriC transporters whose expression is regulated by F^- riboswitches have mutated certain amino acids in order to selectively transport F^- ions.[49] The mis-annotation of all EriC transporters as canonical Cl^- transporters, as well as a general lack of awareness of the role of fluoride and its toxic effects in biology, complicated the initial efforts to validate the *crcB* motif as the F^- riboswitch class. The discovery of fluoride as the ligand for this orphan riboswitch class contributed to our understanding of fluoride resistance mechanisms in a number of organisms,[94-96] and provided evidence that RNA can selectively bind a small, negatively-charged ion.[72]

***ydaO* RNAs: c-di-AMP riboswitches.** The *ydaO* riboswitch candidate is another of the original orphans from the first comparative sequence analyses published in 2004.[5] This RNA exhibits all of the hallmarks of a riboswitch class including a mix of highly-conserved nucleotides, a multi-stem junction, a pseudoknot, and a downstream terminator stem. A huge range of genes were predicted to be controlled by this riboswitch class including those encoding K^+ transporters, amino acid transporters, and other proteins involved in cell wall remodeling and osmotic shock response. The major obstacle to solving the *ydaO* motif was that the cognate ligand for this riboswitch class was unknown to biology at the time the RNA motif was discovered.

While a multitude of specific hypotheses were pursued over the years, many unbiased approaches were also used to identify the mystery ligand including testing cellular extracts as a rich source of small-molecule metabolites and ions. Indeed, yeast extract appeared to contain a compound that consistently caused structural modulation of the RNA in conserved regions when examined by in-line probing. Further fractionation of the extract followed by mass spectrometry analysis revealed that the compound responsible for these effects contains an AMP moiety. Although AMP itself is not recognized by the RNA, the signaling molecule cyclic di-AMP, a molecule first reported to be found in biology in 2008,[97] is bound by the RNA with sub-nanomolar affinity.[27]

In a close parallel to c-di-GMP riboswitches, the validation of the c-di-AMP riboswitch class as the first widespread receptor for this signaling molecule helped strengthen the connection between c-di-AMP and osmotic shock response in certain organisms.[27] In addition, the collection of c-di-AMP riboswitches helped identify a large variety of new gene functions associated with the c-di-AMP regulon and demonstrated the widespread use of this signaling molecule by a multitude of bacterial species.

***yybP* RNAs: Mn^{2+} riboswitches.** The *yybP* RNA motif spent a long time classified as an orphan riboswitch candidate.[5,82] This well-conserved, widespread, and extremely abundant RNA element is found upstream of various genes encoding ambiguously-annotated putative cation transporters. Similar to the case for fluoride riboswitches,[49] the limiting factor in identifying the ligand for this orphan riboswitch class was the undetermined specificity of the associated transporters. The key insight was gained when the one of the genes predicted to be regulated by this riboswitch candidate in *E. coli* was implicated as a Mn^{2+}exporter.[98] Once the substrate specificity for this transporter was

known, Mn^{2+} became a likely ligand candidate for this orphan riboswitch class. Thereafter, experimental evidence confirmed direct interactions between the *yybP* orphan riboswitch RNA and its ligand, divalent manganese.[46,47]

Mn^{2+} riboswitches are the most abundant metal ion-sensing riboswitch class. This riboswitch class demonstrates the ability of RNA to exhibit high specificity for one particular ion over all other ions of similar size and charge that are at higher concentrations in the cell. The validation of Mn^{2+} as the ligand for this riboswitch class also suggests other putative transporters in its regulatory network may transport Mn^{2+}.

***pfl* RNAs: ZMP/ZTP riboswitches.** Unlike many other orphan riboswitch candidates, the *pfl* RNA motif is found upstream of genes involved in well-established metabolic pathways.[20,82] The most commonly controlled genes encode enzymes at the intersection of *de novo* purine biosynthesis and 10-formyl-tetrahydrofolate (10f-THF) biosynthesis.[20] Positioned at this metabolic intersection is the molecule 5-aminoimidazole-4-carboxamide ribonucleotide (AICAR, also known as ZMP) which is formylated using 10f-THF as a donor to produce the next intermediate (FAICAR) in purine synthesis. Upon the discovery of the *pfl* motif, AICAR was examined by in-line probing to evaluate whether it could be directly recognized by a representative *pfl* RNA, but ligand-induced structural modulation was not observed.[20]

For this reason, the *pfl* RNA motif became an orphan riboswitch candidate despite the initial correct identification of its cognate ligand. Unfortunately, these first experiments[20] as well as further investigations[82] with numerous additional potential ligands used a representative RNA that was slightly misfolded and therefore did not sample the correct aptamer structure predicted by the consensus model, which is necessary for ligand

recognition. Years later, this technical issue involving construct design was remedied and validation of ZMP, or its triphosphate analog ZTP, as the cognate ligand(s) for the *pfl* orphan riboswitch class was successful.[30] The ZMP/ZTP riboswitch class and the genes under its control provide evidence for a hypothesis from decades earlier claiming that, in addition to its role as a purine biosynthesis intermediate, ZMP (and ZTP) can act as an alarmone for 10f-THF deficiency.[30,99]

***yjdF* RNAs: azaaromatic riboswitches.** The *yjdF* motif is almost exclusively found upstream of a gene of unknown function, termed *yjdF* (or DUF2992) in *B. subtilis*.[20] Genes of unknown function provide inadequate information to rationally determine a possible ligand identity for a riboswitch candidate. Therefore, two different approaches were used to overcome this lack of knowledge. First, extremely rare gene associations implicated the ligand might have some relation to NAD^+ and FMN metabolism.[20] As such, these coenzymes and their derivatives were examined for *in vitro* binding to *yjdF* RNAs.[100] In an unusual result, a range of flavin analogs were found to be bound by *yjdF* with relatively high affinity. However, this promiscuous binding did not appear to be non-specific in that point mutations to conserved nucleotides of the *yjdF* aptamer disrupt this binding event.

The second approach involved the unbiased screening of a chemical library to attempt to find cellular conditions that would activate the RNA, which was predicted to be an "on" switch based on potential interactions with a downstream intrinsic terminator stem. A strain of *B. subtilis* containing a *yjdF* RNA-*lacZ* reporter fusion was grown in ~2,000 conditions. This plate-based assay contained a huge variety of compounds related to bacteria including many metabolites, such as nucleotides, sugars, and amino acids, as well as various toxins and pH or salt conditions. This screen revealed only five compounds that specifically

activated the *yjdF* RNA-*lacZ* reporter, all of which were various large, planar, polycyclic molecules.[100] The hits from the screen as well as the flavin derivatives share little resemblance except that they all fall into the general category of azaaromatic molecules. While these approaches did not reveal a single cognate ligand, they did provide experimental support for the *yjdF* RNA as a gene control element and almost certainly as a small-molecule sensing riboswitch.

The *yjdF* riboswitch class may generally sense a wide variety of these azaaromatic compounds, many of which are toxic, to increase expression of an exporter protein for these toxins, YjdF. Broad substrate specificities have been observed for some transporter proteins, yet this phenomenon has not been observed for any previous riboswitch class. Indeed, most known riboswitch classes demonstrate impressive selectivity for their cognate ligand over highly similar molecules. The *yjdF* riboswitch class might be the first example of an RNA sensor with broad specificity. Alternatively, the possibility remains that one specific, biologically relevant compound is the true ligand for *yjdF* RNAs. For this reason, the azaaromatic riboswitch class still has lingering quasi-orphan status, which will likely only be resolved through experimental demonstration of the function of the YjdF protein or through high-resolution structural analysis of *yjdF* RNAs bound to various azaaromatic compounds.

***ykkC* RNAs: the last great orphan?** The *ykkC* RNA motif is the last unsolved candidate of the set first published by Barrick *et. al.* in 2004.[5] This orphan riboswitch class is relatively abundant with examples distributed among six bacterial phyla. The RNA structure based on the consensus model consisted mainly of two stems with conserved nucleotides in an internal bulge followed by a tail of many additional highly-conserved

nucleotides. Although there was no structure prediction for the nucleotides in this highly-conserved tail, in-line probing analyses showed that there were likely additional structured elements.[82] The *ykkC* motif aptamer is always located adjacent to a terminator stem that involves some of the conserved nucleotides in the tail region such that it would likely function as an "on" switch.

The eclectic mixture of gene products predicted to be under the regulatory network of this orphan riboswitch includes those annotated as urea carboxylases and their associated proteins, multi-drug efflux pumps, nitrate/sulfate/bicarbonate transporters, *de novo* purine biosynthesis enzymes, and branched-chain amino acid biosynthesis enzymes.[5,82] Additional motivation for solving the ligand for this orphan riboswitch was the discovery of two additional RNA motifs, termed mini-*ykkC*[80] and *ykkC*-III,[20] that also appear to be *cis*-regulatory elements found upstream of the same types of genes as the original *ykkC* motif.

A number of factors conspired to hinder validation of the *ykkC* riboswitch candidate, including the wide variety of genes controlled, gene mis-annotations, and hidden variant aptamers with substantial differences in their cognate ligand identity. The validation of ligands for the *ykkC*, mini-*ykkC*, and *ykkC*-III orphan riboswitch classes is the topic of the remainder of my thesis.

REFERENCES

(1) Breaker, R.R. (2011) Prospects for riboswitch discovery and analysis. *Mol Cell 43,* 867-879.

(2) Roth, A., Breaker R.R. (2009) The structural and functional diversity of metabolite-binding riboswitches. *Annu Rev Biochem 78*, 305-334.

(3) Serganov, A., Nudler, E. (2013) A decade of riboswitches. *Cell 152*, 17-24.

(4) Sherwood, A.V., Henkin, T.M. (2016) Riboswitch-mediated gene regulation: Novel RNA architectures dictate gene expression responses. *Annu Rev Microbiol 70,* 361-374.

(5) Barrick, J.E., Corbino, K.A., Winkler, W.C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I. (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc Natl Acad Sci USA 101,* 6421-6426.

(6) Corbino, K.A., Barrick, J.E., Lim, J., Welz, R., Tucker, B.J., Puskarz, I., Mandal, M., Rudnick. N.D., Breaker, R.R. (2005) Evidence for a second class of *S*-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. *Genome Biol 6,* R70.

(7) Gottesman, S., Storz, G. (2011) Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb Perspect Biol 3*, a003798.

(8) Waters, L.S., Storz, G. (2009) Regulatory RNAs in Bacteria. *Cell 136,* 615-628.

(9) Grundy, F.J., Henkin T.M. (2006) From ribosome to riboswitch: control of gene expression in bacteria by RNA structural rearrangements. *Crit Rev Biochem Mol Biol 41*, 329-338.

(10) Green, N.J., Grundy, F.J., Henkin, T.M. (2010) The T box mechanism: tRNA as a regulatory molecule. *FEBS Lett 584,* 318-324.

(11) Winkler, W.C., Breaker, R.R. (2005) Regulation of bacterial gene expression by riboswitches. *Annu Rev Microbiol 59,* 487-517.

(12) McCown, P.J., Corbino, K.A., Stav, S., Sherlock, M.E., Breaker, R.R. (2017) Riboswitch diversity and distribution. *RNA 23*, 995-1011.

(13) Nahvi, A., Sudarsan, N., Ebert, M.S., Zou, X., Brown, K.L., Breaker, R.R. (2002) Genetic control by a metabolite binding mRNA. *Chem Biol 9,* 1043-1049.

(14) Nahvi, A., Barrick, J.E., Breaker, R.R. (2004) Coenzyme B_{12} riboswitches are widespread genetic control elements in prokaryotes. *Nucleic Acid Res 32,* 143-150.

(15) Johnson, J.E., Reyes, F.E., Polaski, J.T., Batey, R.T. (2012) B_{12} cofactors directly stabilize an mRNA regulatory switch. *Nature 492,* 133-137.

(16) Winkler, W.C., Nahvi, A., Breaker, R.R. (2002) Thiamine derivatives that bind messenger RNAs directly to regulate bacterial gene expression. *Nature 419*, 952-956.

(17) Winkler, W.C., Cohen-Chalamish, S., Breaker, R.R. (2002) An mRNA structure that controls gene expression by binding FMN. *Proc Natl Acad Sci USA* 99, 15908-15913.

(18) Winkler, W.C., Nahvi, A., Sudarsan, N., Barrick, J.E., Breaker, R.R. (2003) An mRNA structure that controls gene expression by binding *S*-adenosylmethionine. *Nat Struct Biol 10*, 701-707.

(19) Weinberg, Z., Regulski, E.E., Hammond, M.C., Barrick, J.E., Yao, Z., Ruzzo, W.L., Breaker, R.R. (2008) The aptamer core of SAM-IV riboswitches mimics the ligand-binding site of SAM-I riboswitches. *RNA 14*, 822-828.

(20) Weinberg, Z., Wang, J.X., Bogue, J., Yang, J., Corbino, K., Moy, R.H., Breaker, R.R. (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol 11*, 31.

(21) Poiata, E., Meyer, M.M., Ames, T.D., Breaker, R.R. (2009) A variant riboswitch aptamer class for *S*-adenosylmethionine common in marine bacteria. *RNA 15*, 2046-2056.

(22) Fuchs, R.T., Grundy, F.J., Henkin, T.M. (2006) The S_{MK} box is a new SAM-binding RNA for translational regulation of SAM synthetase. *Nat Struct Mol Biol 13*, 226-233.

(23) Arachchilage, G.M., Sherlock, M.E., Weinberg, Z., Breaker, R.R. (2017) SAM-VI RNAs selectively bind *S*-adenosylmethionine and exhibit similarities to SAM-III riboswitches. *RNA Biol* (in press).

(24) Wang, J.X., Lee, E.R., Morales, D.R., Lim, J., Breaker, R.R. (2008) Riboswitches that sense *S*-adenosylhomocysteine and activate genes involved in coenzyme recycling. *Mol Cell 29*, 691-702.

(25) Regulski EE, Moy RH, Weinberg Z, Barrick, J.E., Yao, Z., Ruzzo, W.L., Breaker, R.R. (2008) A widespread riboswitch candidate that controls bacterial genes involved in molybdenum cofactor and tungsten cofactor metabolism. *Mol Microbiol 68*, 918-32.

(26) Ames, T.D., Rodionov, D.A., Weinberg, Z., Breaker, R.R. (2010) A Eubacterial Riboswitch Class That Senses the Coenzyme Tetrahydrofolate. *Chem Biol 17*, 681-685.

(27) Nelson, J.W., Sudarsan, N., Furukawa, K., Weinberg, Z., Wang, J.X., Breaker, R.R. (2013) Riboswitches in eubacteria that sense the second messenger c-di-AMP. *Nat Chem Biol 9*, 834-839.

(28) Sudarsan, N., Lee, E.R., Weinberg, Z., Moy, R.H., Kim, J.N., Link, K.H., Breaker, R.R. (2008) Riboswitches in eubacteia sense the second messenger cyclic di-GMP. *Science 321*, 411-413.

(29) Nelson, J.W., Sudarsan, N., Phillip, G.E., Stav, S., Lünse, C.E., McCown, P.J., Breaker, R.R. (2015) Control of bacterial exoelectrogenesis by c-AMP-GMP. *Proc Natl Acad Sci USA 112,* 5389-5394.

(30) Kim, P.B., Nelson, J.W., Breaker, R.R. (2015) An ancient riboswitch class in bacteria regulates purine biosynthesis and one-carbon metabolism. *Mol Cell 57*, 317-328.

(31) Roth, A., Winkler, W.C., Regulski, E.E., Lee, B.W., Lim, J., Jona, I., Barrick, J.E., Ritwik, A., Kim, J.N., Welz, R., Iwata-Reuyl, D., Breaker, R.R. (2007) A riboswitch selective for the queuosine precursor preQ$_1$ contains an unusually small aptamer domain. *Nat Struct Mol Biol 14*, 308-317.

(32) Meyer, M.M., Roth, A., Chervin, S.M., Garcia, G.A., Breaker, R.R. (2008) Confirmation of a second natural preQ$_1$ aptamer class in Streptococcaceae bacteria. *RNA 14*, 685–95.

(33) McCown, P.J., Liang, J.J., Weinberg, Z., Breaker, R.R. (2014) Structural, functional, and taxonomic diversity of three preQ$_1$ riboswitch classes. *Chem Biol 21*, 880–889.

(34) Mandal, M., Boese, B., Barrick, J.E., Winkler, W.C., Breaker, R.R. (2003) Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. *Cell 113*, 577-586.

(35) Mandal, M., Breaker, R.R. (2004) Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nat Struct Mol Biol 11*, 29–35.

(36) Kim, J.N., Roth, A., Breaker, R.R. (2007) Guanine riboswitch variants from *Mesoplasma florum* selectively recognize 2'-deoxyguanosine. *Proc Natl Acad Sci USA 104*, 16092-16097.

(37) Weinberg, Z., Nelson, J.W., Lünse, C.E., Sherlock, M.E., Breaker, R.R. (2017) Bioinformatic analysis of riboswitch structures uncovers variant classes with altered ligand specificity. *Proc Natl Acad Sci USA 114*, 2077-2085.

(38) Breaker, R.R. (2010) RNA second messengers and riboswitches: relics from the RNA World? *Microbe 5*, 13-20.

(39) Breaker, R.R. (2012) Riboswitches and the RNA World. *Cold Spring Harb Perspect Biol 4*, a003566.

(40) Nelson, J.W., Breaker, R.R. (2017) The lost language of the RNA World. *Sci Signal 483*, eaam8812.

(41) Joyce, G.F. (1991) The rise and fall of the RNA world. *New Biol 3*, 399-407.

(42) Gilbert, W. (1986) Origin of life: the RNA world. *Nature 319*, 618.

(43) Vitreschak, A.G., Rodionov, D.A., Mironov, A.A., Gelfand, M.S. (2004) Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet 20***,** 44-50.

(44) Dann III, C.E., Wakeman, C.A., Sieling, C.L., Baker, S.C., Irnov, I., Winkler, W.C. (2007) Structure and mechanism of a metal-sensing regulatory RNA. *Cell 130,* 878-892.

(45) Cromie, M.J., Shi, Y., Latifi, T., Groisman, E.A. (2006) An RNA sensor for intracellular Mg^{2+}. *Cell 125,* 71-84.

(46) Dambach. M., Sandoval, M., Updegrove, T.B., Anantharaman, V., Aravind, L., Waters, L.S., Storz, G. (2015) The ubiquitous *yybP-ykoY* riboswitch is a manganese-responsive regulatory element. *Mol Cell 57,* 1099-1109.

(47) Price, I.R., Gaballa, A., Ding, F., Helmann, J.D., Ke, A. (2015) Mn^{2+}-sensing mechanisms of *yybP-ykoY* orphan riboswitches. *Mol Cell 57*, 1110-1123.

(48) Furukawa, K., Ramesh, A., Zhou, Z., Weinberg, Z., Vallery, T., Winkler, W.C., Breaker, R.R. (2015) Bacterial riboswitches cooperatively bind Ni^{2+} or Co^{2+} ions and control expression of heavy metal transporters. *Mol Cell 57,* 1088-1098.

(49) Baker, J.L., Sudarsan, N., Weinberg, Z., Roth, A., Stockbridge, R.B., Breaker, R.R. (2012) Widespread genetic switches and toxicity resistance proteins for fluoride. *Science 335*, 233-235.

(50) Sudarsan, N., Wickiser, J.K., Nakamura, S., Ebert, M.S., Breaker, R.R. (2003) An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes Dev 17,* 2688–2697.

(51) Mandal, M., Lee, M., Barrick, J.E., Weinberg, Z., Emilsson, G.M., Ruzzo, W.L., Breaker, R.R. (2004) A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science 306*, 275–279.

(52) Ames, T.D., Breaker, R.R. (2011) Bacterial aptamers that selectively bind glutamine. *RNA Biol 8,* 82-89.

(53) Winkler, W.C., Nahvi, A., Roth, A., Collins, J.A., Breaker, R.R. (2004) Control of gene expression by a natural metabolite-responsive ribozyme. *Nature 428,* 281-286.

(54) Bennett, B.D., Kimball, E.H., Gao, M., Osterhout, R., Van Dien, S.J., Rabinowitz, J.D. (2009) Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nat Chem Biol 5*, 593–599.

(55) Witte, G., Hartung, S., Buttner, K. & Hopfner, K.P. (2008) Structural biochemistry of a bacterial checkpoint protein reveals diadenylate cyclase activity regulated by DNA recombination intermediates. *Mol Cell 30*, 167–178.

(56) Corrigan, R.M., Abbott, J.C., Burhenne, H., Kaever, V. & Gründling, A. (2011) c-di-AMP is a new second messenger in *Staphylococcus aureus* with a role in controlling cell size and envelope stress. *PLoS Pathog 7*, e1002217.

(57) Ellington, A.D., Szostak, J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature 346,* 818-22.

(58) Tuerk, C., Gold, L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science 249*, 505-10.

(59) Barrick, J.E., Breaker, R.R. (2007) The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biol 8,* R239.

(60) Serganov, A., Patel, D.J. (2012) Metabolite recognition principles and molecular mechanisms underlying riboswitch function. *Annu Rev Biophys 41,* 343-370.

(61) Lee, E.R., Baker, J.L., Weinberg, Z., Sudarsan, N., Breaker, R.R. (2010) An allosteric self-splicing ribozyme triggered by a bacterial second messenger. *Science 329,* 845-848.

(62) Cheah, M.T., Wachter, A., Sudarsan, N., Breaker, R.R. (2007) Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature 447,* 497-500.

(63) Wachter, A., Tunc-Ozdemir, M., Grove, B.C., Green, P.J., Shintani, D.K., Breaker, R.R. (2007) Riboswitch control of gene expression in plants by splicing and alternative 3′ end processing. *Plant Cell 19,* 3437-3450.

(64) Farnham, P.J., Platt, T. (1981) Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription *in vitro*. *Nucleic Acids Res 9,* 563–577.

(65) Wilson, K.S., von Hippel, P.H. (1995) Transcription termination at intrinsic terminators: the role of the RNA hairpin. *Proc Natl Acad Sci USA 92,* 8793–8797.

(66) Steitz, J.A., Jakes, K. (1975) How ribosomes select initiator regions in mRNA: base pair formation between the 3'-terminus of 16S rRNA and the mRNA during the initiation of protein synthesis in *Escherichia coli*. *Proc Natl Acad Sci USA 72*, 4734–4738.

(67) Trausch, J.J., Ceres, P, Reyes, F.E., Batey, R.T. (2011) The structure of a tetrahydrofolate-sensing riboswitch reveals two ligand binding sites in a single aptamer. *Structure 19*, 1413-1423.

(68) Ren, A., Patel, D.J. (2014) c-di-AMP binds the *ydaO* riboswitch in two pseudo-symmetry-related pockets. *Nat Chem Biol 10,* 780-786.

(69) Ruff, K.M., Muhammad, A., McCown, P.J., Breaker, R.R., Strobel, S.A. (2016) Singlet glycine riboswitches bind ligand as well as tandem riboswitches. *RNA 22,* 1728-1738.

(70) Sudarsan, N., Hammond, M.C., Block, K.F., Welz, R., Barrick, J.E., Roth, A., Breaker, R.R. (2006) Tandem riboswitch architectures exhibit complex gene control functions. *Science 13,* 300-304.

(71) Hurwitz, C., Rosano, C.L. (1967) The intracellular concentration of bound and unbound magnesium ions in *Escherichia coli*. *J Biol Chem 242,* 3719-3722.

(72) Ren, A., Rajashankar, K.R., Patel, D.J. (2013) Fluoride ion encapsulation by Mg^{2+} and phosphates in a fluoride riboswitch. *Nature 486*, 85-89.

(73) Nisan, N., Schocken, S., (2005) The elements of computing systems: building a modern computer from first principles. Cambridge, MA, USA:MIT Press.

(74) Simpson, R.E. (1974) Introductory electronics for scientists and engineers. Boston, MA, USA:Allyn and Bacon, Inc.

(75) Vitreschak, A.G., Rodionov, D.A., Mironov, A.A., Gelfand, M.S. (2003) Regulation of the vitamin B_{12} metabolism and transport in bacteria by a conserved RNA structural element. *RNA 9*, 1084-1097.

(76) Mironov, A.S., Gusarov, I., Rafikov, R., Lopez, L.E., Shatalin, K., Kreneva, R.A., Perumov, D.A., Nudler, E. (2002) Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell 111*, 747-756.

(77) Epshtein, V., Mironov, A.S., Nudler, E. (2003) The riboswitch-mediated control of sulfur metabolism in bacteria. *Proc Natl Acad Sci USA 100*, 5052-5056.

(78) McDaniel, B.A., Grundy, F.J., Artsimovitch, I., Henkin, T.M. (2003) Transcription termination control of the S box system: direct measurement of *S*-adenosylmethionine by the leader RNA. *Proc Natl Acad Sci USA 100*, 3083-3088.

(79) Grundy, F.J., Lehman, S.C., Henkin, T.M. (2003) The L box regulon: lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. *Proc Natl Acad Sci USA 100,* 12057-12062.

(80) Weinberg, Z., Barrick, J.E., Yao, Z., Roth, A., Kim, J.N., Gore, J., Wang, J.X., Lee, E.R., Block, K.F., Sudarsan, N., Neph, S., Tompa, M., Ruzzo, W.L., Breaker, R.R. (2007) Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res 35,* 4809-4819.

(81) Weinberg, Z., Lünse, C.E., Corbino, K.A., Ames, T.D., Nelson, J.W., Roth, A., Perkins, K.R., Sherlock, M.E., Breaker, R.R. (2017) Detection of 224 candidate structured RNAs by comparative analysis of specific subsets of intergenic regions. *Nucleic Acids Res 45,* 10811-10823.

(82) Meyer, M.M., Hammond, M.C., Salinas, Y., Roth, A., Sudarsan, N., Breaker, R.R. (2011) Challenges of ligand identification for riboswitch candidates. *RNA Biol 8,* 5-10.

(83) Greenlee, E.B., Stav, S., Atilho, R.M., Brewer, K.I., Harris, K.A., Malkowski, S.N., Arachchilage, G.M., Perkins, K.R., Sherlock, M.E., Breaker, R.R. (2017) Challenges of Ligand Identification for the Second Wave of Orphan Riboswitch Candidates. *RNA Biol* (in press).

(84) Dambach, M.D., Winkler, W.C. (2009) Expanding roles for metabolite-sensing regulatory RNAs. *Curr opin microbiol 12*, 161-169.

(85) Soukup, G.A., Breaker, R.R. (1999) Relationship between internucleotide linkage geometry and the stability of RNA. *RNA 5,* 1308-1325.

(86) Regulski, E.E., Breaker, R.R. (2008) In-line probing analysis of riboswitches. *Methods Mol Biol 419,* 53-67.

(87) Wickiser, J.K., Winkler, W.C., Breaker, R.R., Crothers, D.M. (2005) The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Mol Cell 18*,49-60.

(88) Misra, V.K., Draper, D.E. (1998) On the role of magnesium ions in RNA stability. *Biopolymers 48*, 113-135.

(89) Reader, J.S., Metzgar, D., Schimmel, P., de Crécy-Lagard, V. (2004) Identification of four genes necessary for biosynthesis of the modified nucleoside queuosine. *J Biol Chem 279*, 6280-6285.

(90) Ross, P., Weinhouse, H., Aloni, Y., Michaeli, D., Weinberger-Ohana, P., Mayer, R., Braun, S., de Vroom, E., van der Marel, G.A., van Boom, J.H., Benziman, M. (1987) Regulation of cellulose synthesis in *Acetobacter xylinum* by cyclic diguanylic acid. *Nature 325*, 279–281.

(91) Simm, R., Morr, M., Kader, A., Nimtz, M., Römling, U. (2004) GGDEF and EAL domains inversely regulate cyclic di-GMP levels and transition from sessility to motility. *Mol Microbiol 53,* 1123-1134.

(92) Ryjenkov, D.A., Simm, R., Römling, U., Gomelsky, M. (2006) The PilZ domain is a receptor for the second messenger c-di-GMP: the PilZ domain protein YcgR controls motility in enterobacteria. *J Biol Chem 281*, 30310-30314.

(93) Sudarsan, N., Barrick, J.E., Breaker, R.R. (2003) Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA 9,* 644-647.

(94) Li, S., Smith, K.D., Davis, J.H., Gordon, P.B., Breaker, R.R., Strobel, S.A. (2013) Eukaryotic resistance to fluoride toxicity mediated by a widespread family of fluoride export proteins. *Proc Natl Acad Sci USA 110,* 19018-19023.

(95) Smith, K.D., Gordon, P.B., Rivetta, A., Allen, K.E., Berbasova, T., Slayman, C., Strobel, S.A. (2015) Yeast Fex1p is a constitutively expressed fluoride channel with functional asymmetry of its two homologous domains. *J Biol Chem 290,* 19874-19887.

(96) Stockbridge, R.B., Kolmakova-Partensky, L., Shane, T., Koide, A., Koide, S., Miller, C., Newstead, S. (2015) Crystal structures of a double-barrelled fluoride ion channel. *Nature 525*, 548-551.

(97) Witte, G., Hartung, S., Buttner, K., Hopfner, K.P. (2008) Structural biochemistry of a bacterial checkpoint protein reveals diadenylate cyclase activity regulated by DNA recombination intermediates. *Mol Cell 30*, 167–178.

(98) Waters, L.S., Sandoval, M., Storz, G. (2011) The *Escherichia coli* MntR miniregulon includes genes encoding a small protein and an efflux pump required for manganese homeostasis. *J Bacteriol 193,* 5887–5897.

(99) Bochner, B.R., Ames, B.N. (1982) ZTP (5-amino 4-imidazole carboxamide riboside 50-triphosphate): a proposed alarmone for 10-formyl-tetrahydrofolate deficiency. *Cell 29*, 929–937.

(100) Li, S., Hwang, X.Y., Stav, S., Breaker, R.R. (2016) The yjdF riboswitch candidate regulates gene expression by binding diverse azaaromatic compounds. *RNA 22,* 530-541.

FIGURES

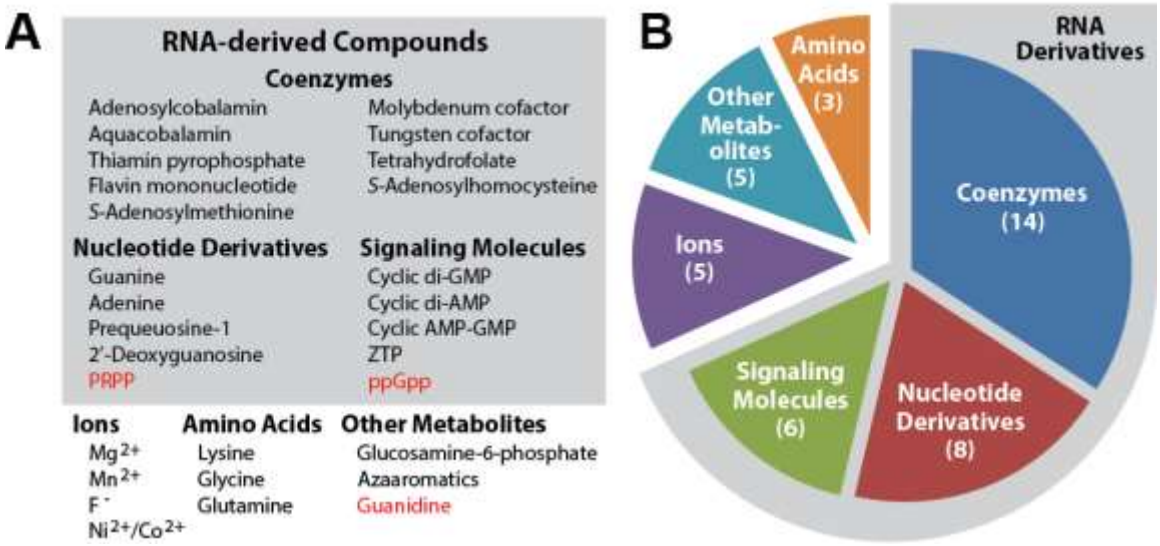

Figure 1-1. Ligands of known riboswitch classes. Adapted from McCown et. al.[12] (A) Small molecules and ions that are known ligands for validated riboswitch classes. Molecules are divided by their general identity, including coenzymes, nucleotide derivatives, signaling molecules, ions, amino acids, and other metabolites. Molecules in red were discovered and validated through experiments reported in this thesis. (B) The total number of riboswitch classes for ligands that fall into the six categories reported in panel A. It should be noted that some ligands are sensed by multiple riboswitch classes.

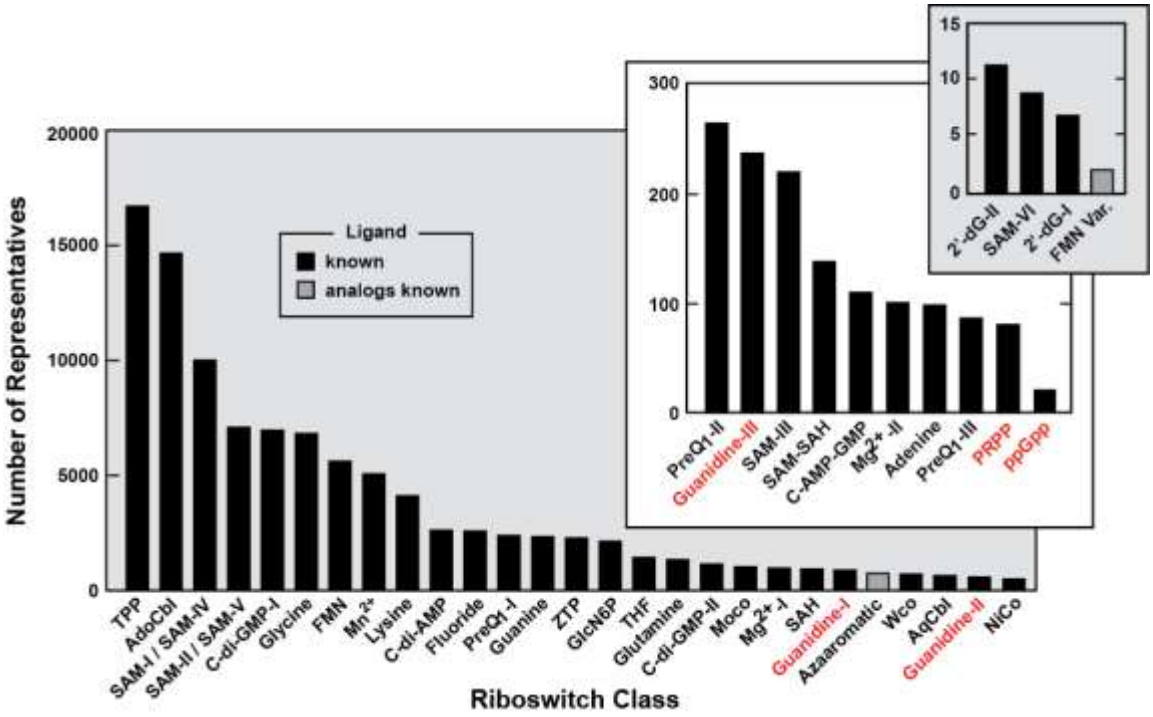

Figure 1-2. Abundance of validated riboswitch classes. Adapted from McCown *et. al.*[12] The number of unique representatives of each validated riboswitch class, based on the RefSeq version 56 database, in descending order of their relative abundance. Riboswitch classes indicated in red were discovered and validated through experiments reported in this thesis.

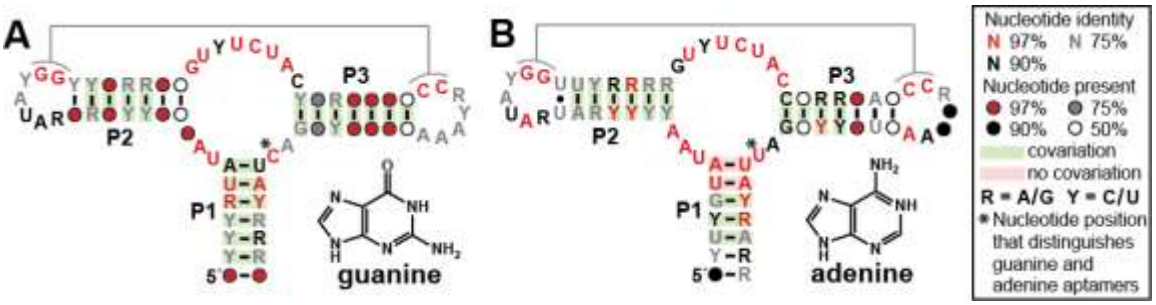

Figure 1-3. Consensus models for the guanine and adenine riboswitch class aptamers. (A) The consensus sequence and secondary structure model based on all representatives of the guanine riboswitch class in the RefSeq 56 database, and the structure of the guanine molecule. Green shading indicates covarying base pairing in the P1, P2, and P3 stems. Gray lines indicate a long-range base pairing interaction. The asterisk denotes the nucleotide position that distinguish guanine aptamers from adenine aptamers. (B) The consensus sequence and secondary structure model based on all representatives of the adenine riboswitch class in the RefSeq 56 database, and the structure of the adenine molecule. All other notations are as in panel A.

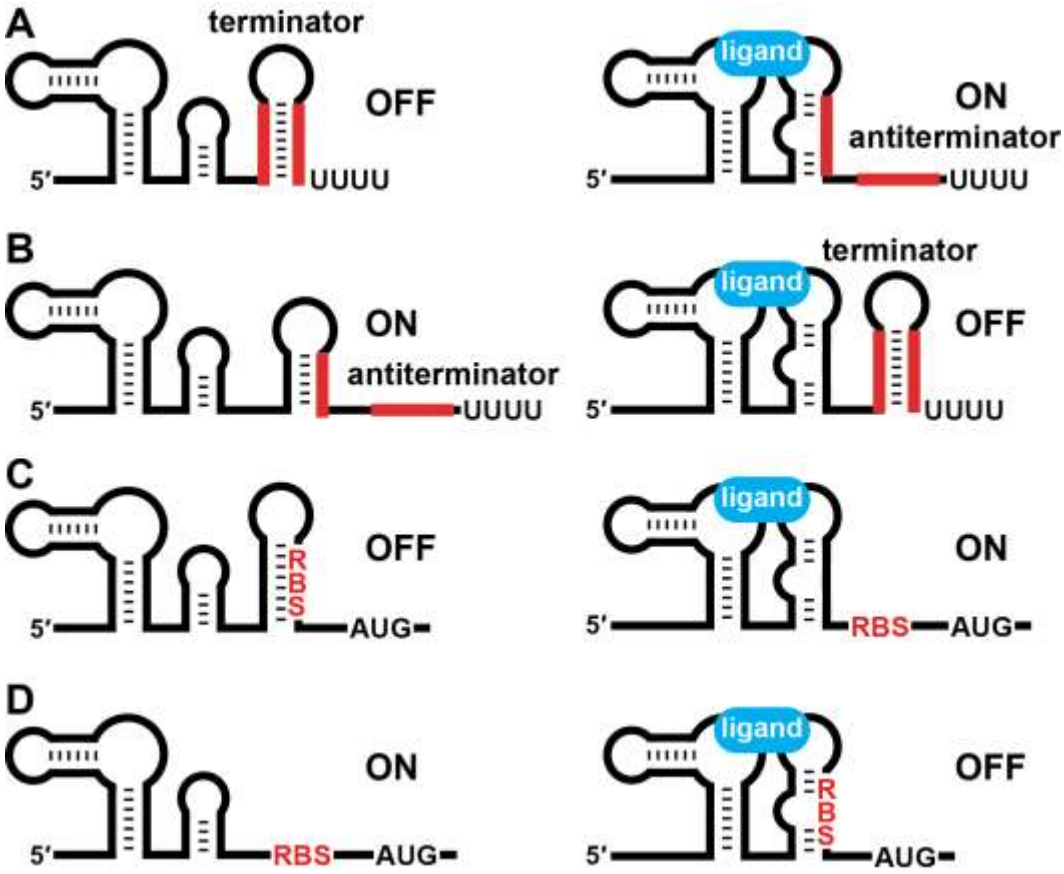

Figure 1-4. Cartoon models of riboswitches acting at the transcriptional and translational level. (A) A model for a transcriptional "on" switch, in which ligand binding prevents formation of an intrinsic terminator stem. (B) A model for a transcriptional "off" switch, in which ligand binding allows formation of an intrinsic terminator stem. (C) A model for a translational "on" switch, in which ligand binding exposes the ribosome binding site. (D) A model for a translational "off" switch, in which ligand binding prevents access to the ribosome binding site.

CHAPTER TWO

Metabolism of free guanidine in bacteria is regulated by a widespread riboswitch class

Largely adapted from the following publication:

Nelson, J.W., Atilho, R.M., Sherlock, M.E., Stockbridge, R.B., Breaker, R.R. (2017) Metabolism of free guanidine in bacteria is regulated by a widespread riboswitch class. *Mol Cell* 65, 220-230.

ABSTRACT

The guanidyl moiety is a component of fundamental metabolites including the amino acid arginine, the energy carrier creatine, and the nucleobase guanine. Curiously, reports regarding the importance of free guanidine in biology are sparse and no biological receptors that specifically recognize this compound have been previously identified. Many members of the *ykkC* motif RNA, the longest-unresolved riboswitch candidate, naturally sense and respond to guanidine. This RNA is found throughout much of the bacterial domain of life, where it commonly controls the expression of proteins annotated as urea carboxylases and multidrug efflux pumps. Analyses performed by me and my colleagues reveal that these proteins likely function as guanidine carboxylases and guanidine transporters, respectively. Furthermore, my colleagues demonstrated that bacteria are capable of endogenously producing guanidine. These and related findings demonstrate that free guanidine is a biologically relevant compound and several gene families exists that can alleviate guanidine toxicity.

INTRODUCTION

Guanidine was first discovered as a thermal decomposition product of the nucleobase guanine.[1] Industrial-scale production of guanidine has been exploited for the manufacture of explosives and other chemicals, but this compound is perhaps best known to the scientific community as a commonly-used reagent for protein denaturation.[2] Microbes capable of metabolizing guanidine have been previously studied, in part due to the interest in guanidine bioremediation.[3-8] Guanidine also has been proposed to be generated by human cells as part of the guanidine cycle via the cleavage of canavanine.[6,9] Despite these scattered findings, guanidine has never been considered to be a major cellular metabolite. Indeed, no biological receptor has previously been identified that specifically recognizes guanidine.

In this chapter, my colleagues and I determined that many bacteria employ riboswitches[10,11] to selectively bind guanidine (most likely as the guanidinium cation) and regulate genes encoding proteins that overcome its toxicity. Over the past fifteen years, more than 30 different riboswitch classes have been demonstrated to control a variety of metabolic or physiological processes, including previously underappreciated aspects of biology. Typically, widespread bacterial riboswitches recognize metabolites or signaling molecules that are themselves widespread and broadly important to life.

One of the most common uncharacterized RNA motifs identified in bacteria over the past two decades is the *ykkC* riboswitch candidate. This RNA motif was one of eight classes discovered in an earlier study using comparative sequence analysis to identify riboswitches.[12] Seven of the motifs identified in this previous study have since been characterized, and six of these were found to control important aspects of biology related

to cell wall biosynthesis[13,14] and sporulation,[14,15] Mg^{2+} homeostasis,[16] Mn^{2+} homeostasis,[17,18] glycine metabolism,[19] and biosynthesis of modified nucleotides.[20] Some of these RNA candidates were considered "orphan" riboswitches because their ligands had not immediately been validated experimentally.[21,22] The *ykkC* motif RNA is the last orphan remaining from this initial collection of RNA classes[12] to resist all previous efforts at experimental validation.[21]

Members of the *ykkC* class control an eclectic distribution of genes including those encoding predicted urea carboxylases and their associated proteins, as well as predicted allophanate hydrolases, arginases, creatininases, and nitrate/sulfate/bicarbonate transporters.[12,21] The RNA is also commonly found upstream of genes annotated as *emrE* or *sugE*, which are predicted to encode small multidrug resistance (SMR) efflux pumps.[23,24] For example, the *Escherichia coli* EmrE protein has been demonstrated to be highly indiscriminate and is thought to have evolved to protect the cell against a wide variety of small molecule toxins.[25] In addition, some *ykkC* motif RNAs are found upstream of genes involved in branched-chain amino acid biosynthesis and purine metabolism.[21] Frequently, these RNAs are predicted to control the entire *de novo* purine biosynthetic operon of their host organism. Due to the diversity of genes associated with this motif, no single compound emerged as a probable riboswitch ligand.

My colleagues hypothesized that *ykkC* RNAs might recognize some toxic ligand,[21] as they are associated with disparate metabolic genes and genes that encode numerous types of transporter proteins. Accordingly, my colleague James Nelson screened a wide variety of growth conditions, including exposure to numerous toxic compounds, for those that trigger riboswitch activation in cells. Of the roughly 2,000 conditions tested, only the

addition of guanidine hydrochloride led to riboswitch-mediated gene expression. I subsequently demonstrated that certain *ykkC* motif RNAs are capable of specifically recognizing guanidine *in vitro* at concentrations many orders of magnitude below those required for guanidine to non-specifically affect RNA structure. Furthermore, I performed biochemical analyses of a previously studied urea carboxylase[26] and determined that this protein prefers guanidine over urea as a substrate. Additionally, my colleague Randy Stockbridge examined of one of the SugE transporters whose expression is controlled by the riboswitch and collected evidence that it selectively recognizes guanidine. Finally, my colleague Ruben Atilho demonstrated that bacteria are capable of endogenously producing guanidine under laboratory conditions. These and other findings reveal that the majority of *ykkC* motif RNAs are members of a broadly distributed riboswitch class that specifically senses guanidine and controls the expression of a variety of genes that together constitute a previously unknown super-regulon for its metabolism and transport.

RESULTS

The *ykkC* Motif is a widespread orphan riboswitch candidate predicted to control the expression of a variety of transporter and metabolic genes. By performing a series of homology searches, James Nelson identified ~2000 distinct *ykkC* RNA sequences in Actinobacteria, Firmicutes, Proteobacteria and Cyanobacteria. Through additional experimental efforts as discussed below, I established that there are at least two distinct subtypes of *ykkC* motif RNAs. Most of these RNAs are of the predominant "subtype 1" (Figure 2-1A), whereas a rarer variant *ykkC* motif called "subtype 2" (Figure 2-2) carries prevalent alterations in three otherwise highly-conserved regions of the *ykkC* motif (Figure

2-2A). An adjoining expression platform consisting of a transcriptional terminator stem[27,28] is frequently located 3′ of this motif. Therefore, previous reports predicted[12,21] that the vast majority of *ykkC* motif representatives of both subtypes function as riboswitches by activating expression of downstream genes upon recognition of the cognate ligand by precluding formation of an intrinsic transcription terminator stem.

The identities of the ligands recognized by newly discovered riboswitch classes frequently can be inferred based on the functions of proteins whose genes reside immediately downstream. Unfortunately, the total collection of genes associated with *ykkC* motif RNAs (Figure 2-1B, Figure 2-2B) did not suggest any likely ligand candidates in addition to those that already had been examined.[21] Intriguingly, two other bacterial RNA motifs, called mini-*ykkC*[29] and *ykkC*-III,[30] are often found upstream of genes similar to those associated with subtype 1 RNAs. This suggests that these three RNA motifs might all respond to the same ligand, either directly or by employing an intermediary protein factor. Therefore, any insight gained by the study of subtype 1 RNAs would likely also advance our understanding of the biological and biochemical functions of these other two RNA classes as further discussed below.

Addition of guanidine triggers gene expression mediated by a subtype 1 *ykkC* motif RNA. James Nelson screened a larger diversity of compounds and growth conditions for any factor that might result in activation of a riboswitch reporter construct. He used a reporter created previously[21] by fusing the subtype 1 *ykkC* motif representative naturally associated with the *Bacillus subtilis ykkC* coding region to a β-galactosidase (*lacZ*) reporter gene (Figure 2-3A). Of the nearly 2,000 conditions examined, only the addition of

guanidine triggered reporter gene expression mediated by the *ykkC* motif RNA (Figure 2-4A).

To experimentally confirm this hit, Ruben Atilho grew the *B. subtilis* reporter strain with various concentrations of added guanidine (Figure 2-4B). Reporter gene expression progressively increases with increasing guanidine, beginning with concentrations as low as 100 μM. Since guanidine is an effective protein denaturant above 1 M,[2] this suggests that guanidine's influence on reporter gene expression is unlikely to be due to nonspecific denaturation of a riboswitch or protein regulatory factor.

Guanidine-dependent expression also occurs if the native promoter for the *ykkC* gene[21] is replaced with an unrelated promoter (Figure 2-3B), which indicates that the *ykkC* motif RNA is responsible for ligand-induced gene regulation. Furthermore, disruption of conserved RNA sequence or structural features, such as in reporter constructs M1 and M2, results in a loss of expression (Figure 2-3C). In contrast, incorporation of mutations in construct M3 that retain a conserved RNA structure also permits guanidine-dependent regulation.

To further demonstrate that guanidine selectively triggers RNA-mediated gene expression, Ruben Atilho created another reporter construct in *E. coli* cells based on a representative subtype 1 example associated with the *Klebsiella pneumoniae tauA* gene, which encodes a putative ATPase located near additional genes commonly associated with *ykkC* motif RNAs. Again, the addition of guanidine to the culture medium activates expression of the WT reporter construct, whereas mutation of a single conserved nucleotide in construct M4 disrupts regulation (Figure 2-5A-C). Furthermore, guanidine does not trigger expression of a member of an unrelated riboswitch class (Figure 2-5D-E). These

results are consistent with the hypothesis that subtype 1 representatives of *ykkC* motif RNAs function as guanidine-dependent riboswitches, and hereafter these RNAs will be called guanidine class I or 'guanidine-I' riboswitches.

The guanidine-I riboswitch representative examined from *B. subtilis* controls the expression of an apparent two-gene operon comprised of the *ykkC* and *ykkD* coding regions. The proteins encoded by these genes are predicted to form a heterodimer and function as an SMR efflux pump.[23] If guanidine is the biologically-relevant ligand for this riboswitch class, then this protein complex might function to alleviate guanidine toxicity by reducing its concentration in cells. Indeed, upon deleting the *ykkC* and *ykkD* genes, Ruben observed decreases in both the concentrations of guanidine required to inhibit bacterial growth (Figure 2-3D) and to activate riboswitch-mediated expression of *lacZ* (Figures 2.3E-F).

As noted earlier, guanidine is a well-known chaotropic agent, and therefore might adversely affect RNA structure in a manner similar to its protein denaturation mechanism. Alternatively, structural reorganization by ionic stabilization of the RNA could be mediated by guanidinium cations (Figure 2-3B), which represent the predominant form of guanidine at biologically-relevant pH conditions. Metal and nonmetal (e.g. amine and polyamine) cations are well known to interact non-specifically with RNA and induce structural changes.[31,32] Given these potential modes of non-specific interaction, my colleagues and I sought to provide several additional lines of evidence that guanidine is the biologically relevant ligand as described in the following sections.

Molecular recognition determinants for guanidine binding by a riboswitch. Ruben Atilho examined the determinants of guanidine recognition by testing a number of

chemically related compounds both for reporter gene regulation and for binding *in vitro*. Of the 16 guanidine analogs he examined using the *B. subtilis* riboswitch reporter (Figure 2-3A), none resulted in expression equal to that induced by guanidine (Figure 2-5F). Only the closest chemical analogs of guanidine, specifically hydroxy-, amino-, and methylguanidine, yielded reporter gene expression above background. Of these, hydroxyguanidine is unstable, and yields guanidine upon degradation.[33] Therefore, it is possible that the signal generated by this compound is not caused by the original analog. Strikingly, the addition of up to molar amounts of urea failed to yield any measurable increase in reporter gene expression. These results suggest that the *ykkC* motif RNAs examined above have evolved to serve as highly selective sensors for free guanidine, and to exclude even closely-related analogs such as urea that might otherwise inappropriately trigger gene expression. It seems unlikely that the riboswitch might preferentially respond to another natural compound that incorporates a guanidyl moiety since any modification to a nitrogen atom causes a loss of affinity.

My colleagues and I considered the possibility that guanidine might induce cells to produce a structurally unrelated signaling compound, or perhaps a metabolic byproduct of guanidine degradation, that then directly triggers riboswitch function. To determine whether the riboswitch aptamer directly recognizes guanidine, I conducted in-line probing assays[34,35] to monitor RNA structural changes that are indicative of riboswitch modulation by ligand binding. Initially, I examined a 104 nucleotide aptamer called 104 *uca* (Figure 2-6A) that was derived from the guanidine-I riboswitch associated with the putative urea carboxylase gene from the bacterium *Nitrosomonas europaea*. Guanidine causes decreased levels of spontaneous RNA strand scission at two sites (Figure 2-6B), indicating that

specific nucleotides within these conserved regions undergo structural reorganization. A dissociation constant (K_D) of ~60 μM for the guanidine-RNA complex is estimated from these data (Figure 2-6C). These results are typical of the structural changes that occur when representatives of other riboswitches classes are exposed to their cognate ligands. In contrast, 20 different compounds carrying a guanidyl moiety or a charged amine group were also examined by in-line probing (Figure 2-7), but only guanidine triggers changes in the shape of the RNA aptamer.

The structural changes observed with the 104 *uca* RNA are unlikely to be due to the chaotropic nature of guanidine. Spontaneous RNA strand scission is reduced by guanidine, which indicates that the RNA is becoming more rather than less structured. The latter effect would be expected if guanidine were functioning as a denaturant. Also, the K_D for RNA binding is more than five orders of magnitude lower than is necessary to cause protein denaturation.[2] Moreover, ligand-mediated modulation occurs with a dose-response curve consistent with a 1-to-1 binding interaction (Figure 2-6C). This suggests that the mechanism of riboswitch function is due to the docking of a single ligand in a single saturable binding site, rather than by non-specific neutralization of negatively charged phosphate groups on the RNA backbone by guanidinium.

For other riboswitch classes, many highly-conserved nucleotides within the aptamer domain establish critical structural contacts to form a selective ligand binding pocket. Therefore, if guanidine is being selectively bound by the RNA, then the mutation of any single conserved nucleotide is likely to adversely affect ligand-induced riboswitch function. To test this hypothesis, I examined a larger collection of mutant 104 *uca* aptamers (Figure 2-8, constructs M8 through M16) that individually alter nine highly-conserved

nucleotides. In each instance, the single mutation either reduces (M10 and M12) or eliminates (M8, M9, M11, M13-M16) guanidine binding. These results suggest that guanidine-I riboswitches exploit numerous conserved nucleotides to form a selective binding pocket for guanidine. Indeed, x-ray structural analysis of a guanidine-I riboswitch aptamer confirms the critical roles of all highly-conserved nucleotides to form a 1-to-1 complex between RNA and guanidine.[36]

Similar ligand-induced structural modulation of other guanidine-I riboswitch aptamers also is observed (Figures 2-9 and 2-10). In contrast, the addition of up to 1 mM guanidine to a subtype 2 *ykkC* motif RNA (data not shown) or 0.5 M guanidine to aptamers from four other riboswitch classes (azaaromatic, c-di-AMP, c-AMP-GMP, ZTP; data not shown) failed to induce structural changes. These results again indicate that guanidine, even at exceedingly high concentrations, does not cause RNA structural changes by non-specific interactions. Importantly, the addition of guanidine to representative mini-*ykkC* and *ykkC*-III motif RNAs also induces structure modulation (see Chapters 3 and 4). These preliminary findings are consistent with the original hypothesis that mini-*ykkC* and *ykkC*-III motif RNAs might represent additional distinct riboswitch classes that respond to the same ligand as *ykkC* motif RNAs with similar gene associations.[30]

Guanidine selectively induces riboswitch-mediated transcription. James Nelson assessed the ability of guanidine to selectively regulate transcription *in vitro* by measuring the extent of transcript elongation of several guanidine-I riboswitch constructs in the presence of guanidine or various analogs. Each construct has the potential to form a strong intrinsic terminator stem, but ligand binding presumably results in a stable aptamer structure that precludes formation of the terminator. In single-round transcription reactions

with a riboswitch construct derived from the region upstream of the *smr* gene from the bacterium *Desulfotomaculum ruminis* (Figure 2-11A), the inclusion of guanidine ranging from 1 to 100 mM yields progressively greater amounts of full-length RNA transcript (Figure 2-11B). James Nelson observed similar results for the guanidine-I riboswitches associated with the *Clostridium difficile emrE* (Figure 2-12A) and *B. subtilis ykkC* (Figure 2-12B) genes. Mutant construct M5, which carries a nucleotide change at a highly conserved position, exhibits a complete loss of guanidine responsiveness (Figure 2-11C). Likewise, mutations M6 and M7 predictably diminish and restore, respectively, the guanidine-responsive transcription of the *smr* construct.

Notably, guanidine has no effect on the extent of transcript elongation for a riboswitch that recognizes a different ligand[37] (Figure 2-12C). Finally, the addition of structural analogs of guanidine, including urea, does not lead to transcript elongation (Figure 2-11D). These data demonstrate that the riboswitches tested are capable of specifically recognizing guanidine *in vitro* and that ligand binding by the RNA is sufficient to control the expression of downstream genes.

A urea carboxylase enzyme prefers guanidine over urea. We speculated that the genes controlled by guanidine-I riboswitches represent a super-regulon that is capable of reducing guanidine toxicity, either by its export or degradation. Therefore, I chose to look more closely at predicted urea carboxylase genes,[26] which are frequently controlled by guanidine-I riboswitches (Figure 2-1B). Urea carboxylases and ureases have been proposed to catalyze the two major routes for the biological decomposition of urea. However, this redundancy seems unnecessary, and it has been previously hypothesized that one of these protein families might have an alternative purpose.[38] Intriguingly, urea

carboxylase genes are commonly associated with guanidine-I riboswitches, whereas ureases never appear to be controlled by this riboswitch class. Therefore, I speculated that urea carboxylases might prefer guanidine as a substrate (Figure 2-13A) and initiate its degradation in a pathway that is similar to that currently proposed for urea.

The only previously characterized prokaryotic urea carboxylase, from the Proteobacterium *Oleomonas sagaranensis,* is encoded by a gene located immediately downstream of a mini-*ykkC* RNA. As noted above, guanidine-I riboswitches and the noncoding mini-*ykkC* and *ykkC*-III motif RNAs might have similar gene control functions. I demonstrated that the *O. sagaranensis* urea carboxylase indeed can use both urea and guanidine as substrates (Figure 2-13B). Importantly, the K_M for urea is in the millimolar range (7.2 mM reported by Kanamori et al.[26]; 5.2 mM observed by me, Figure 2-13C), whereas the K_M for guanidine improves to 0.19 mM. Thus, the enzyme carboxylates guanidine with a catalytic efficiency roughly 40-fold better than with urea. Since the maximum velocity and k_{cat} are only marginally higher compared to urea, this increased efficiency is presumably due to the enhanced affinity of the enzyme for guanidine.

Urea carboxylases are thought to have been introduced to eukaryotes by a horizontal gene transfer event from Proteobacteria.[39] An x-ray structural model of the carboxylase domain of the urea amidolyase from the eukaryote *Kluyveromyces lactis* has been solved previously.[40] Based on this structure, Fan et. al. hypothesized that a protonated aspartic acid residue near the active site, which is conserved in the *O. sagaranensis* carboxylase, forms a hydrogen bond with the oxygen atom of urea. For this to be true, however, the side-chain carboxylic acid of aspartate would have to undergo a pK_a shift of nearly 4 units. In contrast, I hypothesize that this residue could more easily form an ionic interaction or a

hydrogen bond with guanidine without undergoing a shift in pK_a. This mechanism would explain the apparent enhanced affinity for guanidine over urea as described above. These observations suggest that diverse species of bacteria and eukarya express carboxylase enzymes that have evolved to degrade guanidine.

A riboswitch-controlled transport protein selectively binds guanidine. One of the most intriguing aspects of the guanidine-I riboswitch class is its association with genes predicted to code for small multidrug resistance (SMR) proteins. This family is subdivided into small multidrug proteins (SMP), suppressor of *groEL* mutation (SUG) proteins, and paired small multidrug resistance (PSMR) proteins. The latter group includes the proteins YkkC and YkkD in *B. subtilis*,[41] which are regulated by a guanidine-I riboswitch (Figures 2.3 and 2.9). To date, most studies have been performed either on the SMP *E. coli* EmrE protein, which has been found to transport a wide variety of positively charged small molecules,[42,43] or on the *E. coli* SugE protein, which has been suggested to transport considerably fewer compounds.[44]

To explore how SMR proteins encoded by genes controlled by guanidine-I riboswitches differ from the well-studied *E. coli* EmrE transporter, Randy Stockbridge constructed a phylogenetic tree of these proteins (Figure 2-13D). The riboswitch almost always controls a pair of SMR proteins, one of which appears to be EmrE-like and the other SugE-like. Very frequently, their genes are found in a single operon, which suggests that heterodimers of EmrE-like and SugE-like proteins might form naturally. Additionally, when a guanidine-I riboswitch controls only a single copy of an *emrE*- or *sugE*-like gene, its putative partner is always found elsewhere in the genome, most often controlled by another guanidine riboswitch. In rare instances, guanidine-I riboswitches also control a single SMR gene that

is highly similar to *E. coli sugE*. In contrast, genes that are highly similar to *E. coli emrE* are never controlled by the riboswitch.

We speculated that SMR proteins whose expression is controlled by guanidine-I riboswitches might naturally function as guanidine transporters. In a preliminary investigation of this hypothesis, Randy Stockbridge examined a representative SugE protein (called *Clo* SugE) that is derived from a *Clostridiales* bacterium (oral taxon 876) and controlled by a guanidine-I riboswitch. The distribution of positive charges on the loops suggested that this protein was likely to assemble as a homo-complex,[45] and size exclusion chromatography of the purified protein supported this inference. A dose-dependent increase in the intrinsic fluorescence of tryptophan residues within the *Clo* SugE protein was observed upon the addition of guanidine, demonstrating that it binds guanidine (Figure 2-13E). The K_D for this interaction is ~1 mM, as estimated by fitting the fluorescence data to a single-site binding model (Figure 2-13F). Importantly, a similar urea titration has no effect on *Clo* SugE fluorescence, and guanidine titration has no effect on an unrelated fluoride transporter with a similar number of tryptophan residues (data not shown).

Finally, Randy Stockbridge sought to determine whether *Clo* SugE was capable of selectively recognizing guanidine by monitoring the affinity of their interaction in the presence of 1 mM tetraphenylphosphonium (TPP^+), a typical substrate of the *E. coli* EmrE protein. A concentration of TPP^+ at ~1,000-fold higher than its K_D for an *E. coli* EmrE[46] fails to disrupt guanidine binding (Figure 2-13G). These results suggest that this riboswitch controlled SMR protein selectively recognizes guanidine and not the bulky, cationic substrates known to be transported by distal members of this general family of proteins.

Free guanidine is naturally produced by bacteria. The findings described above demonstrate the existence of riboswitches that selectively respond to free guanidine, and suggest that proteins are widely produced by bacteria that degrade and transport guanidine. However, there are no previous reports regarding the natural production of guanidine at biologically-relevant concentrations.[7,8] It seems unlikely that an ancient riboswitch and its associated super-regulon would respond to an unnatural compound. Therefore, we speculated that guanidine might be produced through natural metabolic processes under certain conditions, and that its presence had not yet been detected, perhaps due to its small size or its perceived irrelevance. Accordingly, we sought to identify conditions where a guanidine riboswitch reporter is expressed even though no external guanidine is supplied.

If bacteria experience conditions under which guanidine is produced, the compound might be ejected from cells into the growth medium. Therefore, Ruben Atilho chose an *E. coli* strain that lacks the *tolC* gene to examine this hypothesis because the TolC protein is a well-known exit channel for small toxic molecules in Gram-negative bacteria.[47] Indeed the Δ*tolC* strain, but not the WT strain, exhibits robust WT riboswitch reporter gene expression when grown overnight in minimal medium (Figure 2-14A). The M4 construct (Figure 2-5A), which carries a single mutation in the aptamer domain, exhibits lower induction by guanidine as expected. These data suggest that the riboswitch ligand is synthesized by *E. coli* under these conditions, and accumulates in cells that are deficient in exporting small toxic compounds.

James Nelson analyzed the contents of these bacteria via high-resolution liquid chromatography-mass spectrometry (LC-MS) and observed the presence of a compound with a mass-to-charge ratio equivalent to that of guanidine (Figure 2-14B). This compound

also had a similar retention time to that of guanidine (Figure 2-14C). James Nelson speculated that the modest difference in these retention times might be due to the intrinsic differences of a pure guanidine solution and a complex extract sample. To further confirm that this compound was, in fact, guanidine, he treated cell extracts with di-tert-butyl dicarbonate (Boc-anhydride) to convert guanidine into Boc-guanidine. A new compound with the identical mass-to-charge ratio and similar retention time as authentic Boc-guanidine was observed. Upon mixing the Boc-anhydride-treated extract and Boc-guanidine, a single LC-MS peak was observed (Figure 2-14C), indicating that these peaks both correspond to Boc-guanidine. These results demonstrate that *E. coli* is capable of endogenously producing guanidine.

DISCUSSION

It has been proposed that the cognate ligands for widespread orphan riboswitch classes are likely to be involved in underappreciated, but fundamental, aspects of biology.[10] The data described in this chapter supports the conclusion that free guanidine is the natural ligand for the predominant subtype of *ykkC* motif RNAs, the most common and longest unresolved orphan riboswitch candidate. However, it remains possible that guanidine-I riboswitches might have some additional function that would contribute to the extensive sequence conservation observed in this motif. Despite being generated over 150 years ago, only in rare instances has guanidine been considered to be a biologically relevant compound. However, Ruben Atilho's and James Nelson's results (Figure 2-14) demonstrate that guanidine is naturally produced by bacteria. Perhaps many bacteria

generate free guanidine as they utilize suboptimal energy sources when preferred sources such as glucose are in short supply.

Guanidine-I riboswitches, and the structurally unrelated mini-*ykkC* and *ykkC*-III motif RNAs[29,30] that likely function as guanidine riboswitches (see Chapters 3 and 4), control a super-regulon that includes genes for proteins that appear to detoxify free guanidine by degradation or export. Given the widespread distribution of guanidine riboswitches in bacteria, free guanidine might have been biologically relevant in early life forms. Notably, guanidine is one of a number of compounds that could possibly have been part of a prebiotic pathway for purine synthesis.[48] Guanidine degradation in bacteria could occur in the same manner proposed for urea, wherein previously annotated urea carboxylase enzymes preferentially use guanidine as a substrate (Figure 2-13A). Genes coding for such 'guanidine carboxylases' are almost always found associated with two families of uncharacterized proteins. It seems likely that these other proteins play some additional role in guanidine metabolism. Notably, the product of carboxylation of guanidine, carboxyguanidine, is a close analog of allophanate. Therefore, I hypothesize that proteins annotated as allophanate hydrolases whose expression levels are controlled by guanidine riboswitches presumably use carboxyguanidine as a substrate for subsequent decomposition.

Another class of genes commonly regulated by these riboswitches encode arginases, which typically catalyze the conversion of arginine into ornithine. These arginases are most likely activate gene expression in the presence of guanidine. When controlled by the riboswitch, these genes are always found in an operon with the genes *hypA* and *hypB*, which have been implicated in the insertion of Ni^{2+} into hydrogenases in *E. coli*.[49] Since arginases

can also utilize a Ni^{2+} ion at their active site, we speculate that *hypA* and *hypB* may be involved in the insertion of Ni^{2+} or some other cation into these enzymes. Perhaps guanidine-induced expression of this operon helps cells overcome inhibition of arginase activity by free guanidine. Alternatively, these enzymes could conceivably catalyze the difficult reverse reaction, where ornithine or 5-hydroxynorvaline is directly converted to arginine via the addition of guanidine.

Guanidine riboswitches also commonly control the expression of numerous transporters. Many of these are predicted to import nitrate, sulfate, or bicarbonate. I find it likely that these proteins might import bicarbonate to facilitate the carboxylation of guanidine and its subsequent metabolism. Indeed, genes encoding such transporters are often found in an operon with guanidine carboxylases genes. The majority of the remaining transporters whose expression levels are controlled by guanidine riboswitches are predicted SMR transporters. Certain members of this protein superfamily, such as the *E. coli* proteins EmrE and SugE have been extensively studied as model transporters due to their small size. EmrE in particular has been demonstrated to transport a wide variety of positively charged small molecules ranging from TPP^+ to glycine betaine.[42,43] Randy Stockbridge demonstrated that one of the SMR proteins whose expression is controlled by a guanidine riboswitch selectively recognizes guanidine. Her data further suggest that SMR transporters whose expression levels are controlled by guanidine-I riboswitches usually assemble as heterodimers of SugE- and EmrE-like subunits and that their biological role is to function as guanidine exporters. This hypothesis is supported by Ruben Atilho's observation that *B. subtilis* cells wherein this transporter is knocked out have a greater

sensitivity to guanidine (Figure 2-3D) and a higher cellular concentration of the riboswitch ligand (Figure 2-3E and 2.3F).

In *E. coli*, the SugE protein is controlled by the mini-*ykkC* motif, suggesting it might also transport guanidine. Interestingly, the *sugE* gene was first identified as a suppressor of a *groEL* mutant gene, a gene that encodes one of a family of molecular chaperones that assists in the proper folding of proteins.[50] An intriguing possibility is that SugE reduces the cellular concentration of guanidine, a known protein destabilization agent, thereby reducing the need for GroEL to chaperone protein folding.

It is also important to note that *emrE* genes not controlled by guanidine-I riboswitches are commonly associated with horizontal gene transfer elements and are relatively recently diverged.[41] In contrast, genes that encode SMR proteins that are controlled by these RNAs are only rarely found within 10 kb of an integrase or transposase. These observations suggest that a mutation may have occurred in an ancestral guanidine transporter that allowed it to transport a wider variety of cationic compounds, leading to its rapid transfer amongst bacteria and subsequent contribution to antibiotic resistance.

As noted earlier, I found that approximately 30% of *ykkC* motif RNAs conform to subtype 2 (Figure 2-2A). Members of this subtype are commonly associated with genes necessary for *de novo* purine metabolism and branched-chain amino acid biosynthesis. These RNAs are distinguished by nucleotide variations at the sites where the structure of guanidine-I riboswitches modulates in the presence of guanidine. Therefore, subtype 2 *ykkC* motif RNAs have adapted to recognize unknown ligands that relate to the metabolic processes relevant to their associated genes. Future experiments established some of the ligands for this remaining collection of orphan riboswitches (see Chapters 5 and 6).

MATERIALS AND METHODS

Chemicals and reagents. Guanidine was used in its hydrochloride salt form in all experiments and was purchased from Sigma-Aldrich. Other fine chemicals were also purchased from Sigma-Aldrich with the exceptions of hydroxyguanidine hemisulfate (Santa Cruz Biotech) and guanidinosuccinic acid (Alfa Chemistry). [γ-^{32}P]- and [α-^{32}P]-ATP were purchased from Perkin Elmer and used within two weeks of receipt. Bulk chemicals were purchased from J.T. Baker and enzymes from New England Biolabs, unless otherwise noted. All solutions were prepared using deionized water (dH_2O) and were either autoclaved or filter sterilized (using 0.22 μm filters, Millipore) prior to use. DNA oligonucleotides were purchased from Sigma-Aldrich and Integrated DNA Technologies. A list of oligonucleotides used in this chapter can be found in Table 2-1.

Bioinformatics. Additional homologs of *ykkC* motif RNAs were identified by using Infernal 1.1[51] based on the previously published *ykkC-yxkD* motif alignment.[12] Iterative searches and re-alignment of a library of genomic sequences obtained from bacterial sequences deposited in RefSeq version 63[52] and environmental sequences from the Human Microbiome Project,[53] IMG/M,[54] MG-RAST,[55] CAMERA,[56] and GenBank[57] resulted in the identification of a group of nearly 2000 hits. Genes were predicted to be controlled by the adjacent RNA motif if they were downstream and within 700 nucleotides, or if they were predicted to be in an operon controlled by the RNA motif. These genes were then sorted based on predicted function into various classes as presented in Figure 2-1B and Figure 2-2B.

SMR protein phylogenies were constructed from 221 bacterial sequences, including 182 riboswitch-controlled SMRs and ~39 SMRs that have been experimentally implicated in drug resistance.[41,58,59,60,61] Sequences were aligned using MUSCLE[62], and variable loop regions were trimmed from the alignment using JalView. The unrooted phylogeny was constructed using PhyML 3.0.[63]

Reporter assays and knockout construction. Initially, a riboswitch reporter construct consisting of the native *B. subtilis ykkC* gene promoter and riboswitch fused upstream of the *E. coli lacZ* gene (Figure 2-3) was created by using methods as described previously[64] and integrated into the *amyE* locus of *B. subtilis*.[21] The construct with the native promoter was used for growth condition screening (Figure 2-4), but the promoter was subsequently replaced with the *B. subtilis lysC* promoter via PCR, restriction digest, and subsequent cloning into the vector pDG1661. This latter construct, which avoids any possible complications from *ykkC* gene natural promoter regulation, was used for all other *B. subtilis* experiments. Incorporation of various mutations was carried out via PCR followed by restriction digestion and cloning. The *K. pneumoniae* riboswitch reporter construct (Figure 2-5A) was designed to include the *lysC* promoter from *E. coli* and the entire 5' untranslated region of the *tauA* gene fused directly to the start codon for the full coding region of the *lacZ* reporter gene. The resulting vector pRS414 construct was transformed into *E. coli* BW25113 cells. Screening of various conditions via a phenotype microarray was performed by Biolog, Inc. using proprietary solutions and media. Follow-up reporter experiments were performed by inoculating either *B. subtilis* strain 168 1A1 (*trpC2*) or *E. coli* BW25113 transformed with the riboswitch reporter construct into Lysogeny Broth (LB) or minimal medium (GMM for *B. subtilis* or M9 for *E. coli*) supplemented with

various concentrations of guanidine or other compounds as indicated. Bacteria were grown overnight at 37 ºC, with shaking if using the *lysC* promoter, or without if using the native promoter. Synthetic DNAs used in cloning are listed in Table 2-1.

Preparation of radiolabeled RNA constructs for in-line probing. Templates for RNA transcription were assembled via extension of overlapping synthetic DNAs (Table 2-1) using SuperScript II reverse transcriptase (Thermo Fisher Scientific). Each RNA construct was generated by transcription for four hours at 37 ºC in a 100 µL reaction containing 80 mM HEPES-KOH (pH 7.5 at 23 ºC), 24 mM $MgCl_2$, 2 mM spermidine, 40 mM DTT, and T7 RNA polymerase (16 U μL^{-1}). Each desired RNA product was purified via denaturing PAGE, the gel slice containing the full-length RNA was excised, and the RNA extracted via crush-soak treatment (200 mM NaCl, 10 mM Tris [pH 7.5 at 23 ºC], 1 mM EDTA; 400 µL) for 30 minutes at room temperature. The purified RNAs were subsequently precipitated by addition of 1 mL ethanol, pelleted via centrifugation, dried via rotary evaporation (DNA SpeedVac), and resuspended in 20 µL dH_2O. RNAs were dephosphorylated via rAPid alkaline phosphatase (Roche) treatment according to the manufacturer's instructions. 8 pmol of RNA was radiolabeled using T4 polynucleotide kinase in a 20 µL reaction volume containing 25 mM CHES [pH 9.0 at 23 ºC], 5 mM $MgCl_2$, 3 mM DTT, and 20 µCi [γ-^{32}P]-ATP. The resulting radiolabeled RNAs were purified again as described above before being subjected to in-line probing.

Single-round transcription termination Assays. DNA constructs for single-round *in vitro* transcription were designed to carry the riboswitch of interest spanning from the predicted natural transcription start site to 40 (*C. difficile* and *D. ruminis* constructs) or 50 (*B. subtilis* construct) nucleotides following the terminator stem. The promoter from the

pfl gene in *Clostridium beijerinckii* NCIMB 8052 was used due to its previously demonstrated function in similar assays.[37] A mutation was incorporated in the region of the *B. subtilis* construct before the *ykkC* aptamer so that it contained no cytidines, while the *C. difficile* and *D. ruminis* constructs were naturally cytidine-deficient. Finally, 36 additional nucleotides were added to the 3´ termini of the *C. difficile and D. ruminis* constructs to aid in both resolution of the full-length and terminated products and construction of transcription templates. Synthetic DNA templates based on these design features were then constructed by extension of overlapping, synthetic single-stranded oligonucleotides (Integrated DNA Technologies) followed by PCR amplification of the final product. Approximately 2 pmol of the resulting, purified DNA template was added to a transcription initiation mixture (20 mM Tris-HCl [pH 8.0 at 23 ºC], 20 mM NaCl, 7 mM $MgCl_2$, 0.1 mM EDTA, 10 μg mL^{-1} bovine serum albumin [BSA], 130 μM ApA dinucleotide, 1% glycerol, 0.04 U μL^{-1} *E. coli* RNA polymerase holoenzyme, 2.5 μM GTP, 2.5 μM UTP, and 1.5 μM ATP). Approximately 8 μCi [α-^{32}P]-ATP was added to the 90 μL transcription reaction and transcription was allowed to proceed at 37 ºC for 10 minutes, leading to formation of a stalled polymerase complex at the first cytidine nucleotide of each transcript. The reaction mixture was then distributed in 8 μL aliquots into separate tubes to which 1 μL of a 10x solution of the ligand of interest was added. The resulting solutions were gently mixed and centrifuged before being placed on ice. Subsequently, 1 μL of 10x elongation buffer (20 mM Tris-HCl [pH 8.0 at 23 ºC], 20 mM NaCl, 7 mM $MgCl_2$, 0.1 mM EDTA, 1 mg mL^{-1} heparin) was added. The 10x elongation buffer for the *C. difficile* and *D. ruminis* constructs also contained 1.5 mM each of ATP, GTP, and CTP, as well as 500 μM UTP. The 10x elongation buffer for the *B. subtilis* construct contained 3 mM each of ATP, GTP,

and CTP, and 1 mM UTP. The percentages of elongated RNA transcripts were calculated as previously described.[65] Experimental conditions for single-round transcription of the ZTP riboswitch from *C. beijerinckii* have been described previously.[37]

Preparation and analysis of an annotated urea carboxylase. The full-length urea carboxylase (*uca*) gene from *O. sagaranensis* was cloned into the multiple cloning site 1 of the pETDuet-1 vector (Novagen) using the BamHI and HindIII restriction sites. The recombinant protein was designed such that it contained a His_6 tag at its N terminus. The resulting plasmid was then transformed into *E. coli* BL21(DE3) and cultured in Terrific Broth (IBI Scientific) supplemented with 100 mg L^{-1} ampicillin. Once cells reached an OD_{600} of roughly 0.2, 15 mg L^{-1} D-biotin was added. At an OD_{600} of between 0.8 and 1.0, expression of the *uca* gene was induced by the addition of 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG) and incubated at 16 ºC for 18 h. Cells were harvested by centrifugation and lysed using a microfluidizer (Microfluidics) at 15,000 psi. The lysate was then clarified by ultracentrifugation at 40,000 RCF for 40 min at 4 ºC. His_6-tagged protein was purified first via nickel-affinity chromatography using nickel nitrolotriacetic acid (Ni-NTA) agarose (Anatrace) and then by strong anion exchange chromatography (HiTrap Q HP, GE Healthcare), both performed on an AKTA purification system. The purified protein was subsequently exchanged into buffer containing 25 mM Tris-HCl (pH 8.0 at 23 ºC), 100 mM NaCl, and 0.1 mM TCEP (tris-(2-carboxyethyl)phosphine). The resulting protein sample was concentrated via dialysis, flash-frozen in liquid nitrogen, and stored at -80 ºC.

Catalytic activity of the *O. sagaranensis* carboxylase enzyme was determined by coupling the carboxylation reaction with a lactate dehydrogenase and pyruvate kinase

system as described previously.[26] Briefly, conversion of ATP to ADP by carboxylase is coupled to conversion of NADH to NAD^+ by lactate dehydrogenase and pyruvate kinase, which can be followed by measuring the absorbance of NADH at 340 nm. Guanidine or urea was added at the indicated concentration to the reaction mixture [50 mM HEPES (pH 7.5 at 23 ºC), 50 mM KCl, 8 mM $MgCl_2$, 0.5 mM phosphoenolpyruvate, 8 mM $NaHCO_3$, 0.4 mM NADH, 5 U of pyruvate kinase (Sigma-Aldrich), and 1 U of lactate dehydrogenase (Sigma-Aldrich)] to yield a final reaction volume of 50 µL. The reactions were allowed to incubate at room temperature (~23 ºC) for 45 minutes. ATP was then added to a final concentration of 1 mM to reactions incubated at 30ºC, and absorbance measurements at 340 nm immediately commenced using a Biotek Synergy H1 plate reader. Three technical replicates were performed per trial for each concentration of substrate. Absorbance as a function of time was recorded and the slope of each data set was used to determine the rate of substrate carboxylation. Velocity is reported such that 1 U of activity is defined as 1 µmol NAD^+ formed per minute. The velocity for the no-substrate control (0.5 U mg^{-1}) was subtracted prior to determination of kinetic parameters. V_{max}, K_M, and k_{cat} were determined by fitting a plot of velocity vs. substrate concentration using GraphPad Prism 6 assuming Michaelis-Menten kinetics.

Expression and purification of SugE. *E. coli* transformed with the *sugE*-bearing plasmid were induced with 1 mM IPTG at an OD_{600} of 1.5, and protein expression was allowed to proceed for 3 hr at 37 °C. Cells were lysed by sonication and the resulting mixture was extracted with 60 mM n-decyl-β-D-maltopyranoside (DM) for 2 hr. Cell debris was pelleted, the supernatant was passed over a cobalt affinity resin (Talon, 1 mL per L of culture), and the resin was subsequently washed first with a mixture of 100 mM

NaCl, 20 mM Tris (pH 7.5 at 23 ºC), and 5 mM DM (column buffer, CB), then CB supplemented with 20 mM imidazole, and finally eluted with CB supplemented with 400 mM imidazole. Histidine tags were removed by 1 hr treatment with endoproteinase Lys-C (Roche, 0.2 U mg^{-1} protein). The desired protein was further purified using a Superdex200 gel filtration column with an aqueous solution containing 25 mM HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) (pH 7.5 at 23 ºC), 150 mM NaCl, and 5 mM DM.

Tryptophan fluorescence assay. Protein samples were titrated with guanidinium chloride dissolved in 150 mM NaCl and 25 mM HEPES (pH 7.5 at 23 ºC), with 5 mM DM, and incubated for 2 min to allow the sample to reach an equilibrium value before excitation at 280 nm. Emission was monitored from 200-400 nm, and titration curves were constructed from the emission spectrum peak (333 nm), which typically increased by ~50% over the course of the experiment. For experiments comparing binding in the presence and absence of 1 mM $TPP^{+,}$ the excitation wavelength was 293 nm. Each titration experiment was repeated in triplicate, and data were fit to a single-site binding model using SigmaPlot. For comparison between experiments, data were normalized to the maximum fluorescence value obtained from the regression analysis.

LC-MS analysis of bacterial cell extracts. Bacteria were harvested via centrifugation, the supernatant removed, and the pellet dried via lyophilization. The bacteria were lysed via the addition of 0.1 culture volumes of 50% aqueous methanol (v/v). Insoluble cell debris or salts were pelleted via centrifugation and the supernatant removed, lyophilized to dryness, and resuspended in 0.01 culture volumes of dH_2O.

Extract samples were analyzed via LC-MS first by injection onto a Hypercarb porous graphitic carbon column (2.1 mm x 100 mm, particle size 3 µm, Thermo Fisher Scientific). The column was pre-equilibrated with 2% acetonitrile (ACN) for five minutes prior to injection, followed by elution with 2% ACN over 15 minutes, gradually increasing to 99% ACN over a further 20 minutes. The mass-to-charge ratio of compounds within the eluent was determined using an Agilent 6550 qToF mass spectrometer. Data were analyzed using the Agilent Mass Hunter software. Peaks were considered to have the same mass-to-charge ratio as guanidine if they were within 10 ppm of the calculated ratio. To confirm that this peak corresponded to guanidine, cell extracts or authentic guanidine alone were treated with 10 mM t-butoxycarbonyl anhydride (Boc-anhydride) in the presence of 100 mM NaOH for three hours at room temperature to convert any guanidine into Boc-guanidine. The resulting reaction mixtures were lyophilized, extracted with 50% aqueous methanol (v/v), resuspended in dH_2O, and analyzed via LC-MS as described above. Peaks corresponding to the expected mass of Boc-guanidine were observed in both samples, but they had slightly different LC retention times. These peaks coalesced into a single peak upon mixing the two samples, indicating that the differences in retention times are due to sample effects from the separate preparations.

ACKNOWLEDGMENTS

My colleagues and I thank Adam Roth, Zasha Weinberg, Narasimhan Sudarsan, and other members of the Breaker Lab for helpful conversations. We also thank Shira Stav for assistance in assembling the dataset of RNA sequences, Sanshu Li for experimental reagents, and Christina Lünse and Takemasa Kawashima for assistance with literature

translations. We are grateful to Christopher Lim, Joshua Temple, and Yong Xiong of Yale University for assistance in the expression and purification of the guanidine carboxylase protein, and Rob Bjornson for assisting our use of the Yale Life Sciences High Performance Computing Center (NIH grant RR19895-02). Finally, we thank the Bacillus Genomic Stock Center at The Ohio State University and the Coli Genetics Stock Center at Yale University for supplying bacterial strains and plasmids. R.M.A. was supported by an NIH T32 Genomic and Proteomics Training Grant (3T32HG003198-10S2). M.E.S. and C.W.R were supported by an NIH Cellular and Molecular Biology Training Grant (T32GM007223). This work was also supported by NIH grants to R.B.S. (R00-GM111767) and to R.R.B. (GM022778 and DE022340). R.R.B. is also supported by the Howard Hughes Medical Institute.

REFERENCES

(1) Strecker, A. (1861) Untersuchungen über die chemischen Beziehungen zwischen Guanin, Xanthin, Theobromin, Caffeïn und Kreatinin. *Justus Liebigs Ann Chem 118*, 151-177.

(2) Greene, R.F., Pace, C.N. (1974) Urea and guanidine hydrocholoride denaturation of ribonuclease, lysozyme, α-chymotrypsin, and β-lactoglobulin. *J Biol Chem 249*, 5388-5393.

(3) Bierema, S. (1909) Die assimilation von ammon-, nitrat- und amidstickstoff durch mikroorganismen. *Centr Bakt Par 2*, 672-726.

(4) Ebisuno, T., Takimoto, M. (1981) Examination on the biodegradability of guanyl compounds. *Eisei Kagaku 27*, 156-162.

(5) Iwanoff, N.N., Awetissowa, A.N. (1931) Über die fermentative umwandlung des guanidins in harnstoff. *Biochem Z 231*, 67-78.

(6) Kihara, H., Prescott, J.M., Snell, E.E. (1955) The bacterial cleavage of canavanine to homoserine and guanidine. *J Biol Chem 217*, 497-503.

(7) Mitchell, W.R. (1985) Biodegradation of guanidinium by aquatic microorganisms. *US Army Med Res and Dev Command Tech Rep 8506*, 1-26.

(8) Mitchell, W.R. (1987) Biodegradation of guanidinium ion in aerobic soil samples. *Bull Environ Contam Toxicol 39*, 974-981.

(9) Natelson, S., Sherwin, J.E. (1979) Proposed mechanism for urea nitrogen re-utilization: relationship between urea and proposed guanidine cycles. *Clin Chem 25*, 1343.

(10) Breaker, R.R. (2011) Prospects for riboswitch discovery and analysis. *Mol Cell 43*, 867-879.

(11) Serganov, A., Nudler, E. (2013) A decade of riboswitches. *Cell 152*, 17-24.

(12) Barrick, J.E., Corbino, K.A., Winkler, W.C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I., Wickiser, J.K., Breaker, R.R. (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc Natl Acad Sci USA 101*, 6421-6426.

(13) Winkler, W.C., Nahvi, A., Roth, A., Collins, J.A., Breaker, R.R. (2004) Control of gene expression by a natural metabolite-responsive ribozyme. *Nature 428*, 281-286.

(14) Nelson, J.W., Sudarsan, N., Furukawa, K., Weinberg, Z., Wang, J.X., Breaker, R.R. (2013) Riboswitches in eubacteria sense the second messenger c-di-AMP. *Nat Chem Biol 9*, 834-839.

(15) Marchais, A., Duperrier, S., Durand, S., Gautheret, D., Stragier, P. (2011) CsfG, a sporulation-specific, small non-coding RNA highly conserved in endospore formers. *RNA Biol 8*, 358-364.

(16) Dann, C.E., Wakeman, C.A., Sieling, C.L., Baker, S.C., Irnov, I., and Winkler, W.C. (2007) Structure and mechanism of a metal-sensing regulatory RNA. *Cell 130*, 878-892.

(17) Dambach, M., Sandoval, M., Updegrove, T.B., Anantharaman, V., Aravind, L., Waters, L.S., Storz, G. (2015) The ubiquitous *yybP-ykoY* riboswitch is a manganese-responsive regulatory element. *Mol Cell 57*, 1099-1109.

(18) Price, I.R., Gaballa, R., Ding, F., Helmann, J.D., Ke, A. (2015) Mn^{2+}-sensing mechanisms of *yybP-ykoY* orphan riboswitches. *Mol Cell 57*, 1110-1123.

(19) Mandal, M., Lee, M., Barrick, J.E., Weinberg, Z., Emilsson, G.M., Ruzzo, W.L., Breaker, R.R. (2004) A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science 306*, 275-279.

(20) Roth, A., Winkler, W.C., Regulski, E.E., Lee, B.W., Lim, J., Jona, I., Barrick, J.E., Ritwik, A., Kim, J.N., Welz, R., *et al.* (2007) A riboswitch selective for the queuosine

precursor preQ$_1$ contains an unusually small aptamer domain. *Nat Struct Mol Biol 14*, 308-317.

(21) Meyer, M.M., Hammond, M.C., Salinas, Y., Roth, A., Sudarsan, N., Breaker, R.R. (2011) Challenges of ligand identification for riboswitch candidates. *RNA Biol 8*, 5-10.

(22) Block, K.F., Hammond, M.C., Breaker, R.R. (2010) Evidence for widespread gene control function by the *ydaO* riboswitch candidate. *J Bacteriol 192*, 3983-3989.

(23) Jack, D.L., Storms, M.L., Tchieu, J.H., Paulsen, I.T., Saier Jr., M.H. (2000) A broad-specificity multidrug efflux pump requiring a pair of homologous SMR-type proteins. *J Bacteriol 182*, 2311-2313.

(24) Schuldiner, S.S. (2009) EmrE, a model for studying evolution and mechanism of ion-coupled transporters. *Biochim Biophys Acta 1794*, 748-762.

(25) Schuldiner, S.S. (2012) Undecided membrane proteins insert in random topologies. Up, down and sideways: it does not really matter. *Trends Biochem Sci 37*, 215-219.

(26) Kanamori, T., Kanou, N., Atomi, H., Imanaka, T. (2004) Enzymatic characterization of a prokaryotic urea carboxylase. *J Bacteriol 186*, 2532-2539.

(27) Gusarov, I, Nudler, E. (1999). The mechanism of intrinsic transcription termination. *Mol Cell 3*, 495-504.

(28) Yarnell, W.S., Roberts, J.W. (1999) Mechanism of intrisnic transcription termination and antitermination. *Science 284*, 611-615.

(29) Weinberg, Z., Barrick, J.E., Yao, Z., Roth, A., Kim, J.N., Gore, J., Wang, J.X., Lee, E.R., Block, K.F., Sudarsan, N., *et al.* (2007) Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res 35*, 4809-4819.

(30) Weinberg, Z., Wang, J.X., Bogue, J., Yang, J., Corbino, K., Moy, R.H., Breaker, R.R. (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol 11*, R31.

(31) Hermann, T., Westhof, E. (1999) Docking of cationic antibiotics to negatively charged pockets in RNA folds. *J Med Chem 42*, 1250-1261.

(32) Pyle, A.M. (2002) Metal ions in the structure and function of RNA. *J Biol Inorg Chem 7*, 679-690.

(33) Walker, J.B., Walker, M.S. (1959) The enzymatic reduction of hydroxyguanidine. *J Biol Chem 234*, 1481-1484.

(34) Regulski, E.E., Breaker, R.R. (2008) In-Line Probing Analysis of Riboswitches. *Methods Mol Biol 419*, 53-67.

(35) Soukup, G.A., Breaker, R.R. (1999) Relationship between internucleotide linkage geometry and the stability of RNA. *RNA 5*, 1308-1325.

(36) Reiss, C.W., Xiong, Y., Strobel, S.A. (2017) Structural basis for ligand binding to the guanidine-I riboswitch. *Structure 25*, 195-202.

(37) Kim, P.B., Nelson, J.W., Breaker, R.R. (2015) An ancient riboswitch class in bacteria regulates purine biosynthesis and one-carbon metabolism. *Mol Cell 57*, 317-328.

(38) Hausinger, R.P. (2004) Metabolic versatility of prokaryotes for urea decomposition. *J Bacteriol 186*, 2520-2522.

(39) Strope, P.K., Nickerson, K.W., Harris, S.D. Moriyama, E.N. (2011) Molecular evolution of urea amidolyase and urea carboxylase in fungi. *BMC Evol Biol 11*, 80.

(40) Fan, C., Chou, C.Y., Tong, L., Xiang, S. (2012) Crystal structure of urea carboxylase provides insights into the carboxyltransfer reaction. *J Biol Chem 287*, 9389-9398.

(41) Bay, D.C., Rommens, K.L., Turner, R.J. (2008) Small multidrug resistance proteins: a multidrug transporter family that continues to grow. *Biochim Biophys Acta 1778*, 1814-1838.

(42) Bay, D.C., Turner, R.J. (2012) Small multidrug resistance protein EmrE reduces host pH and osmotic tolerance to metabolic quaternary cation osmoprotectants. *J Bacteriol 194*, 5941-5948.

(43) Yerushalmi, H., Lebendiker, M., Schuldiner, S. (1995) EmrE, an *Escherichia coli* 12-kDa multidrug transporter, exchanges toxic cations, and H^+ and is soluble in organic solvents. *J Biol Chem 270*, 6856-6863.

(44) Chung, Y.J., Saier Jr., M.H. (2002) Overexpression of the *Escherichia coli sugE* gene confers resistance to a narrow range of quaternary ammonium compounds. *J Bacteriol 184*, 2543-2545.

(45) Rapp, M., Granseth, E., Seppala, S., von Heijne, G. (2006) Identification and evolution of dual-topology membrane proteins. *Nat Struct Mol Biol 13*,112-116.

(46) Morrison, E.A. Henzler-Wildman, K.A. (2014) Transported substrate determines exchange rate in the multidrug resistance transporter EmrE. *J Biol Chem 289*, 6825-6836.

(47) Koronakis, V. (2003) TolC – the bacterial exit duct for proteins and drugs. *FEBS Lett 555*, 66-71.

(48) Becker, S., Thoma, I., Deutsch, A., Gehrke, T., Mayer, P., Zipse, H., Carell, T. (2016) A high-yielding, strictly regioselective prebiotic purine nucleoside formation pathway. *Science 352*, 833-836.

(49) Lutz, S., Jacobi, A., Schlensog, V., Böhm, R., Sawers, G., and Böck, A. (1991) Molecular characterization of an operon (*hyp*) necessary for the activity of the three hydrogenase isoenzymes in *Escherichia coli*. *Mol Microbiol 5*, 123-135.

(50) Greener, T., Govezensky, D., Zamir, A. (1993) A novel multicopy supressor of a groEL mutation includes two nested open reading frames transcribed from different promoters. *EMBO J 12*, 889-896.

(51) Nawrocki, E.P., Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics 29*, 2933-2935.

(52) Pruitt, K.D., Tatusova, T., Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts, and proteins. *Nucleic Acids Res 35*, D61-D65.

(53) Gevers, D. (2012) The human microbiome project: a community resource for the healthy human microbiome. *PLoS Biol 10*, e1001377.

(54) Markowitz, V.M., Ivanova, N.N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., Chen, I.M., Grechkin, Y., Dubchak, I., Anderson, I., *et al.* (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res 36*, D534-D538.

(55) Meyer, F. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinfo 9*, 386-393.

(56) Sun, S. (2011) Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res 39*, D546-D551.

(57) Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L. (2008) Genbank. *Nucleic Acids Res 36*, D25-D30.

(58) Cruz, A., Micaelo, N., Felix, V., Song, J.Y., Kitamura, S., Suzuki, S., Mendo, S. (2013) *sugE*: a gene involved in tributyltin (TBT) resistance of *Aeromonas molluscorum* Av27. *J Gen Appl Microbiol 59*, 39-47.

(59) He, G.X., Zhang, C., Crow, R.R., Thorpe, C., Chen, H., Kumar, S., Tsuchiya, T., Varela, M.F. (2011) SugE, a new member of the SMR family of transporters, contributes to antimicrobial resistance in *Enterobacter cloacae*. *Antimicrob Agents Chemother 55*, 3954-3957.

(60) Kovacevic, J., Ziegler, J., Walecka-Zacharska, E., Reimer, A., Kitts, D.D., Gilmour, M.W. (2016) Tolerance of *Listeria monocytogenes* to quaternary ammonium sanitizers is mediated by a novel efflux pump encoded by *emrE*. *Appl Environ Microbiol 82*, 939-953.

(61) Nasie, I., Steiner-Mordoch, S., Schuldiner, S. (2012) New substrates on the block: clinically relevant resistances for EmrE and homologues. *J Bacteriol 194*, 6766-6770.

(62) Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res 32*, 1792-1797.

(63) Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordjik, W., Gascuel, O. (2010) New algorithims and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol 59*, 307-321.

(64) Sudarsan, N., Wickiser, J.K., Nakamura, S., Ebert, M.S., Breaker, R.R. (2003) An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes Dev 17*, 2688-2697.

(65) Kim, J.N., Roth, A., Breaker, R.R. (2007) Guanine riboswitch variants from *Mesoplasma florum* selectively recognize 2′-deoxyguanosine. *Proc Natl Acad Sci USA 104*, 16092-16097.

FIGURES

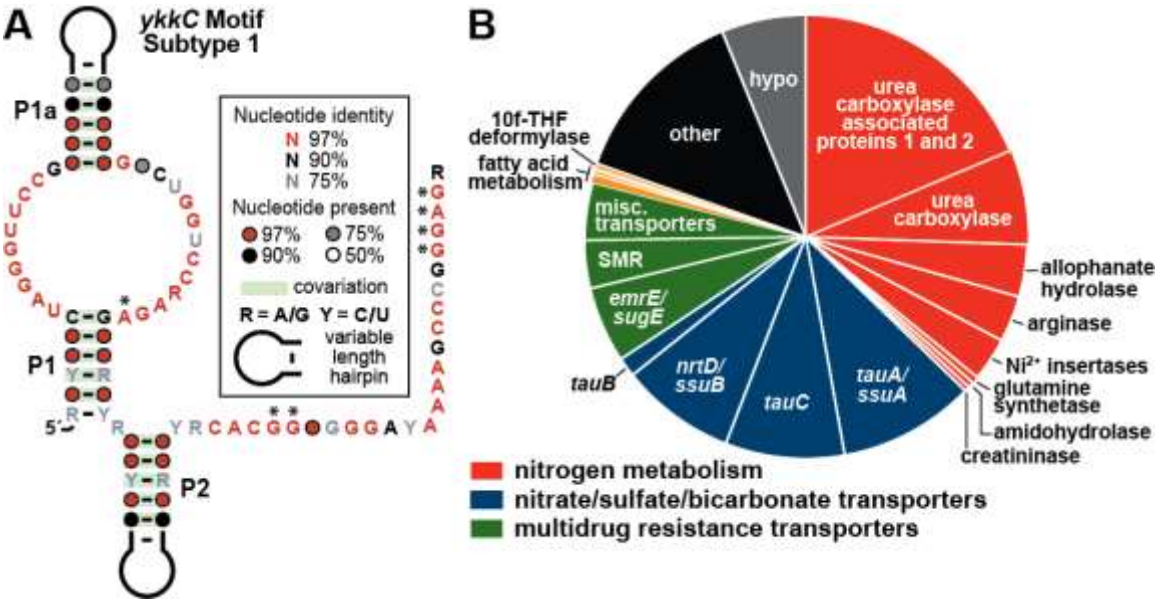

Figure 2-1. The *ykkC* motif is widespread in bacteria. (A) Consensus sequence and secondary structure model derived from ~1500 subtype 1 *ykkC* motif RNAs. Asterisks identify nucleotides that differ between subtypes 1 and 2. (B) Genes predicted to be controlled by subtype 1 *ykkC* motif RNAs. James Nelson and I performed bioinformatic analyses used to create this figure.

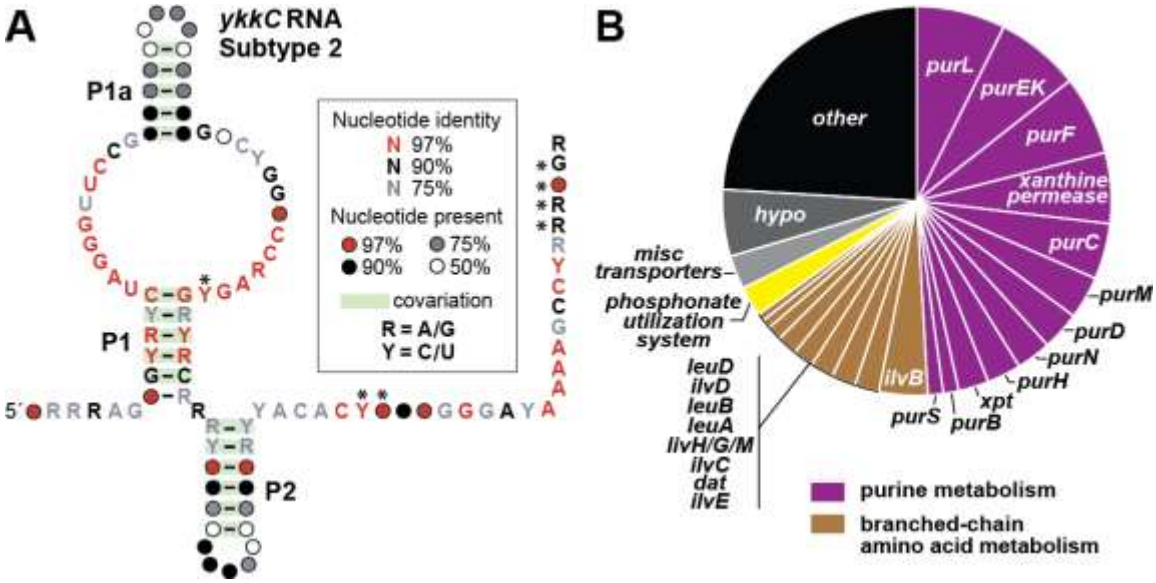

Figure 2-2. Subtype 2 *ykkC* motif RNAs associate with a distinct set of genes. (A) Consensus sequence and secondary structure of subtype 2 *ykkC* motif RNAs. Annotations are as described for Figure 2-1A. Asterisks designate the nucleotide positions that are most distinct from the predominant subtype 1 *ykkC* motif RNAs. (B) Genes associated with subtype 2 *ykkC* motif RNAs. I performed bioinformatic analyses used to create this figure.

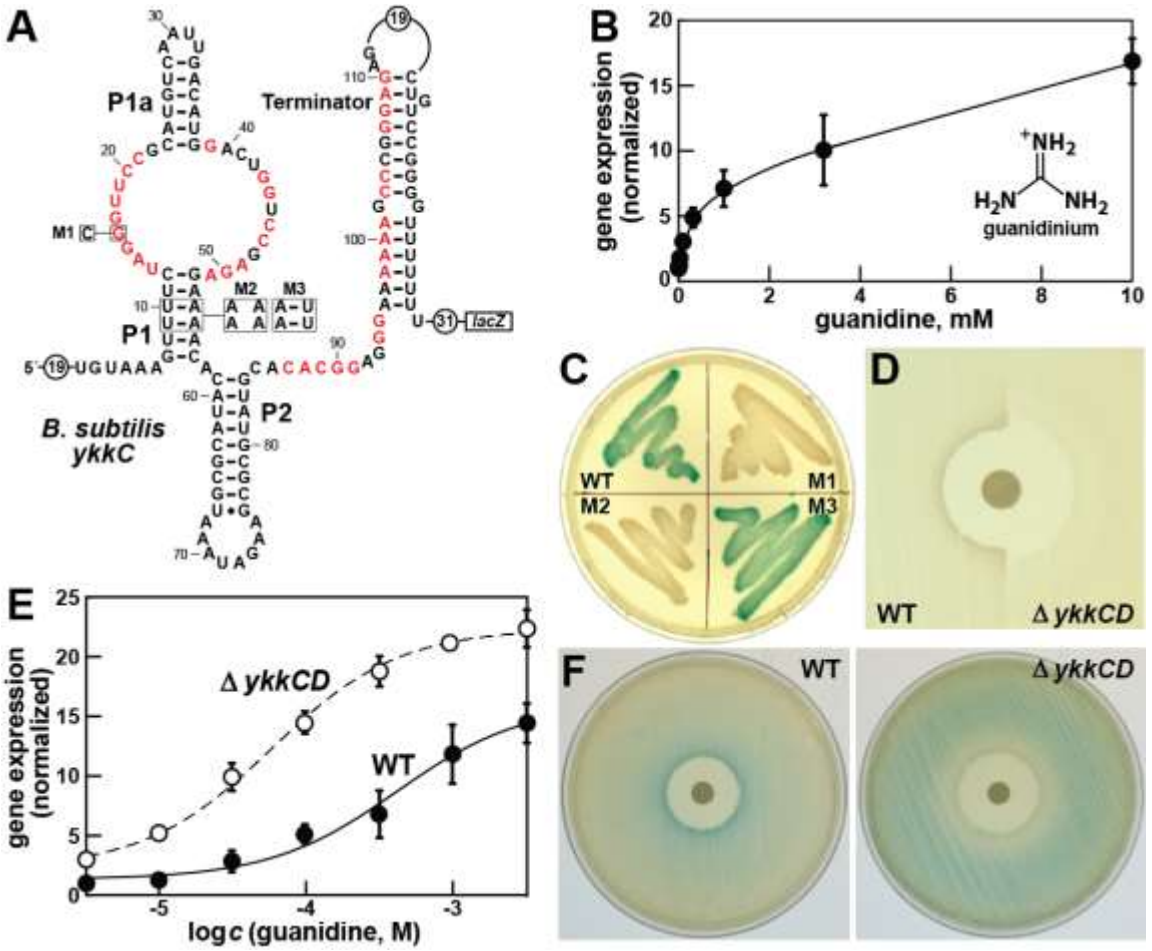

Figure 2-3. A *ykkC* motif RNA responds to guanidine and controls the expression of a transporter that alleviates guanidine toxicity. (A) Sequence and predicted secondary structure of the *ykkC*-reporter construct based on the *ykkC* motif representative associated with the *B. subtilis ykkCD* operon. Red nucleotides are >97% conserved. The secondary structure depicted is proposed to form in the absence of the riboswitch ligand. As with many *ykkC* motif RNAs, part of the predicted aptamer region overlaps a downstream terminator stem such that the presence of the riboswitch ligand will disrupt terminator stem formation, thereby resulting in increased reporter expression. Mutations of conserved nucleotides and secondary structures are designated M1 through M3. (B) Plot of β-galactosidase reporter gene expression in *B. subtilis* versus the concentration of guanidine,

wherein expression in the absence of added guanidine is set to 1. Data points are the average of three measurements and are representative of experiments performed on multiple days. Error bars indicate the standard deviation of the measurements, and when not present are smaller than the data point. Structure for guanidinium is depicted, as this is the predominant form of guanidine under the assay conditions. (C) Reporter gene expression of *B. subtilis* containing wild-type (WT) and mutant reporter constructs as described in A when grown on Lysogeny Broth (LB) agar containing 10 mM guanidine and 100 μg mL^{-1} X-gal. (D) Agar diffusion assay of wild-type and *ykkCD* knockout (Δ*ykkCD*) *B. subtilis* cells grown on the same LB agar plate. Paper disk (center) was spotted with 10 μL of a 6 M guanidine hydrochloride solution (pH 4.7 at 23 ºC). (E) Plot of the β-galactosidase expression levels for WT (set to 1 when measured in the lowest guanidine concentration) and Δ*ykkCD B. subtilis* cells at various concentrations of guanidine. Data are the average of three measurements and are representative of experiments performed on multiple days. Error bars represent the standard deviation of the measurement. Other annotations are as described in B. (F) Agar diffusion assays of *B. subtilis* WT and Δ*ykkCD* cells. Disks were spotted with 10 μL of a 6 M guanidine solution. Both plates contained LB medium supplemented with 100 μg mL^{-1} X-gal. Notably for Δ*ykkCD* cells, an inner halo of growing bacteria that do not exhibit reporter gene expression in the presence of high guanidine levels was observed. A similar phenomenon has been observed for other bacteria bearing riboswitch reporters in the presence of toxic compounds.[37] Ruben Atilho performed the experiments shown in this figure.

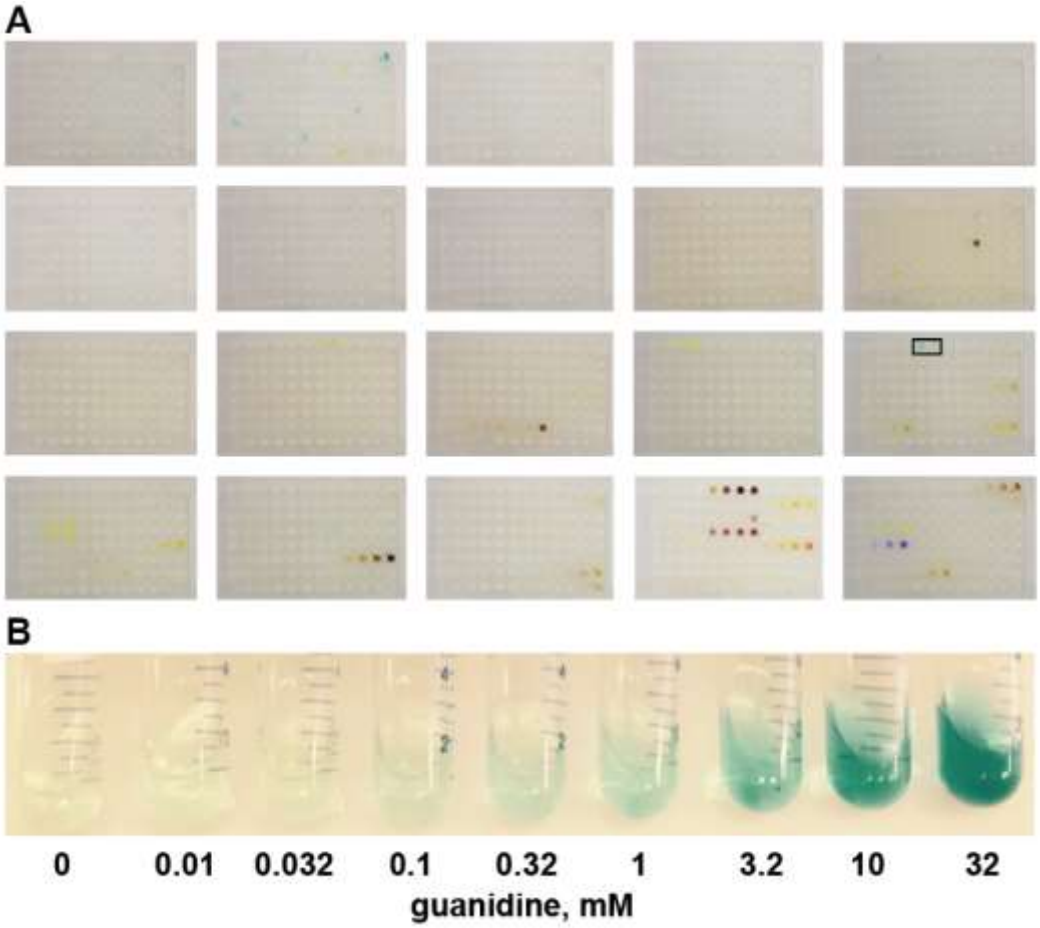

Figure 2-4. Screening growth conditions for activation of a subtype 1 *ykkC* riboswitch reporter fusion yields guanidine as the sole valid hit. (A) Activation of reporter expression in *B. subtilis* by guanidine hydrochloride in Biolog screening assays (see Experimental Procedures). Bacteria were grown either in 0.2x LB or proprietary Biolog media supplemented with various compounds and 1 mM X-gal. Two wells containing guanidine are identified by the box. Other conditions that resulted in activation of the *ykkC* riboswitch reporter (light blue color) also activate an azaaromatic riboswitch reporter fusion construct in the same genetic background, indicating that these conditions likely induce expression of the galactosidase enzyme whose gene is naturally present in the *B. subtilis* genome. Additional colored wells are the result of the inclusion of colored test

compounds in the Biolog screening library. (B) Activation of *ykkC* riboswitch reporter expression in *B. subtilis* by guanidine. Bacteria were grown in 0.2x LB supplemented with the indicated concentration of guanidine and 50 μg mL^{-1} X-gal overnight without shaking. James Nelson performed the experiments shown in this figure.

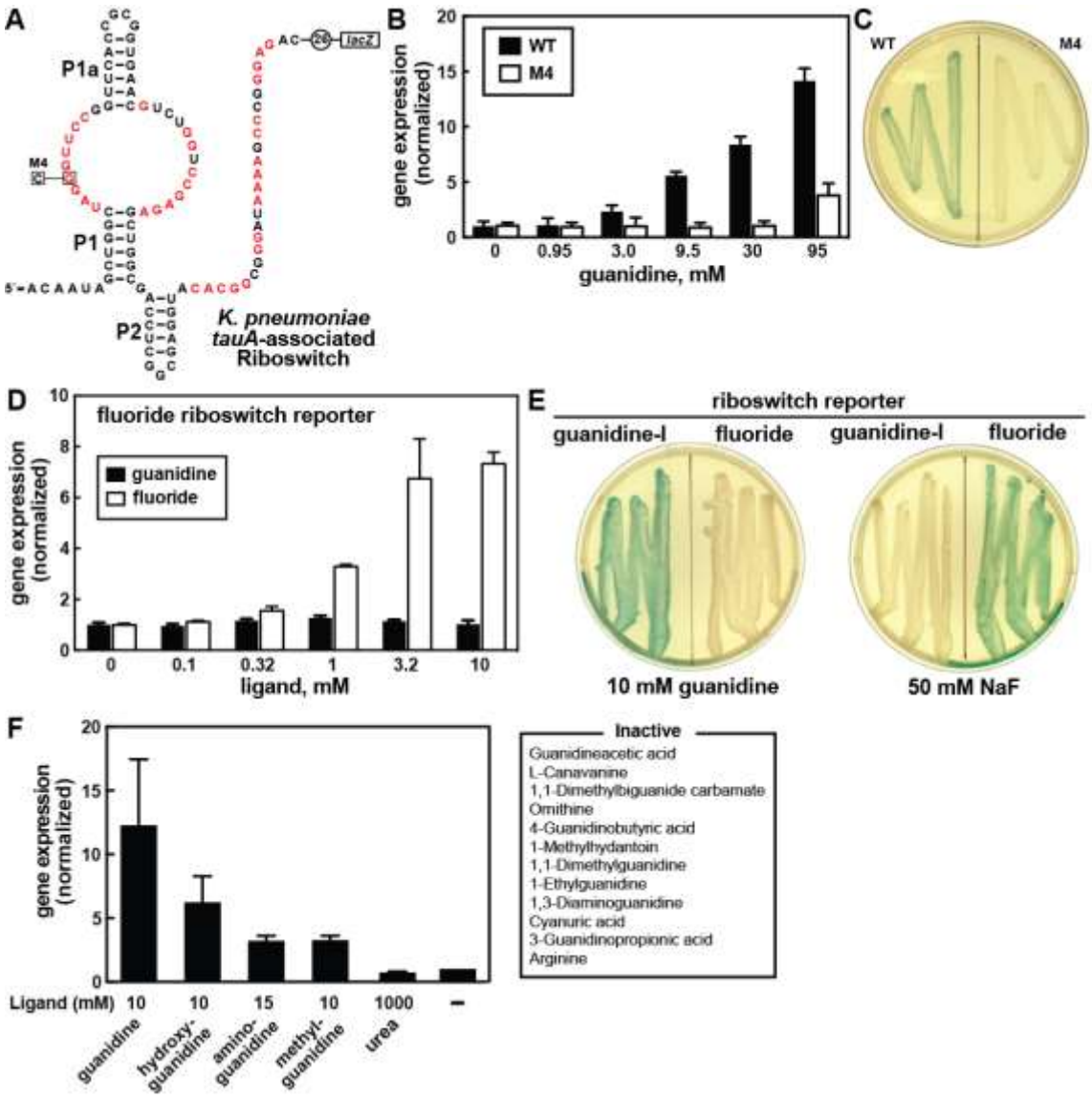

Figure 2-5. Guanidine selectively activates riboswitch reporter constructs. (A) Sequence and secondary structure of a riboswitch reporter construct derived from the subtype 1 *ykkC* motif RNA associated with the *K. pneumoniae tauA* gene (ATPase transporter). Red letters identify the most highly conserved nucleotides within the motif. The expression platform structure and mechanism are not known for this example. M4 designates a mutation at a conserved nucleotide within the riboswitch aptamer. (B) Plot of the β-galactosidase activity for *E. coli* cells containing either the WT or M4 *K. pneumoniae* riboswitch reporter construct versus the concentration of guanidine. Data are the average

of three measurements, and error bars represent the standard deviation of those measurements. Gene expression was normalized to WT *E. coli* grown in the absence of guanidine. (C) LB agar plates containing 100 μg mL^{-1} X-gal and 100 mM guanidine streaked with *E. coli* bearing either the WT or M4 mutant *K. pneumoniae* riboswitch reporter construct as indicated. (D) Plot of β-galactosidase expression of *B. subtilis* cells bearing a fluoride riboswitch reporter versus the addition of guanidine or sodium fluoride. Data are the average of three measurements, and error bars represent the standard deviation of those measurements. Gene expression was normalized to the level of β-galactosidase activity measured from host *B. subtilis* cells grown in the absence of guanidine or fluoride. (E) LB agar plates containing 100 μg mL^{-1} X-gal and either 10 mM guanidine or 50 mM sodium fluoride were streaked with *B. subtilis* containing either a guanidine-responsive (Figure 2-3A) or fluoride riboswitch reporter as indicated. (F) Plot of the relative β-galactosidase expression of *B. subtilis* cells containing the WT guanidine-I riboswitch reporter construct (Figure 2-3A) upon addition of various analogs of guanidine at the indicated concentrations using assay conditions as noted for Figure 2-3. All other compounds were examined at a concentration either equal to their solubility limit or one-half their MIC. (–): no added compound. Gene expression was normalized to the host *B. subtilis* bacteria grown in the absence of guanidine. Values shown are the average of three replicates. Error bars indicate the standard deviation of the measurement. Ruben Atilho performed the experiments shown in this figure.

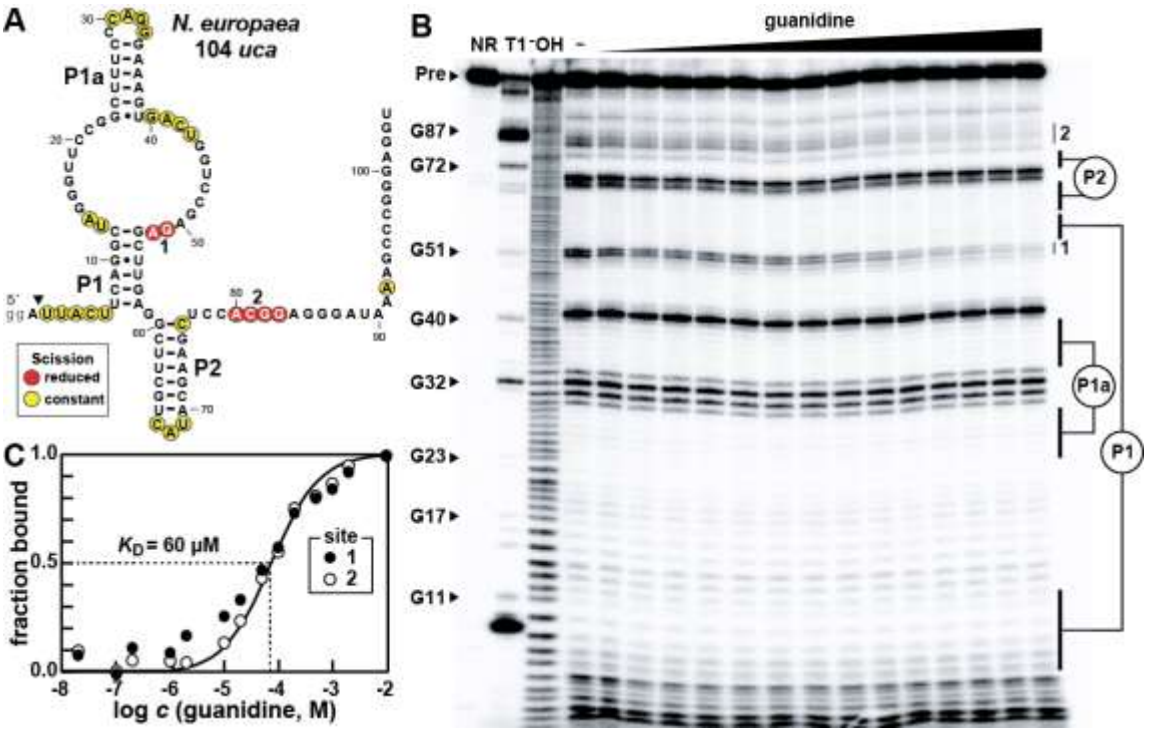

Figure 2-6. Ligand binding by a guanidine-I riboswitch aptamer. (A) Sequence and secondary structure of the 104 *uca* RNA construct derived from the *uca* gene of *N. europaea*. Lowercase letters are nucleotides added to promote efficient *in vitro* transcription with T7 RNA polymerase. Data from the in-line probing assay depicted in B were used to map regions of constant and reduced RNA scission. Arrowhead denotes the start of mapping of available data on the gel. (B) Polyacrylamide gel electrophoresis (PAGE) analysis of the products of in-line probing of 5′ ^{32}P-labeled 104 *uca* RNA in the presence of added guanidine hydrochloride. NR, T1, and ⁻OH, respectively, designate RNA undergoing no reaction, or RNAs that have been partially digested with ribonuclease T1 (cleaves after each G), or partially digested under alkaline conditions (cleaves after every nucleotide). Bands corresponding to cleavage after certain G residues in the sequence are annotated. Guanidine concentrations were progressively increased from no ligand added (–) to 10 mM (far right lane). Regions 1 and 2 undergo guanidine-dependent

structure stabilization. (C) Plot of the fraction of RNA bound by guanidine as inferred from the normalized fraction of spontaneous RNA scission at regions 1 and 2 (as depicted in A and B) relative to the logarithm of the molar concentration of guanidine. The solid line is an idealized binding curve expected for a 1-to-1 interaction between a receptor and its ligand. Data shown are representative of multiple experiments. I performed the experiments shown in this figure.

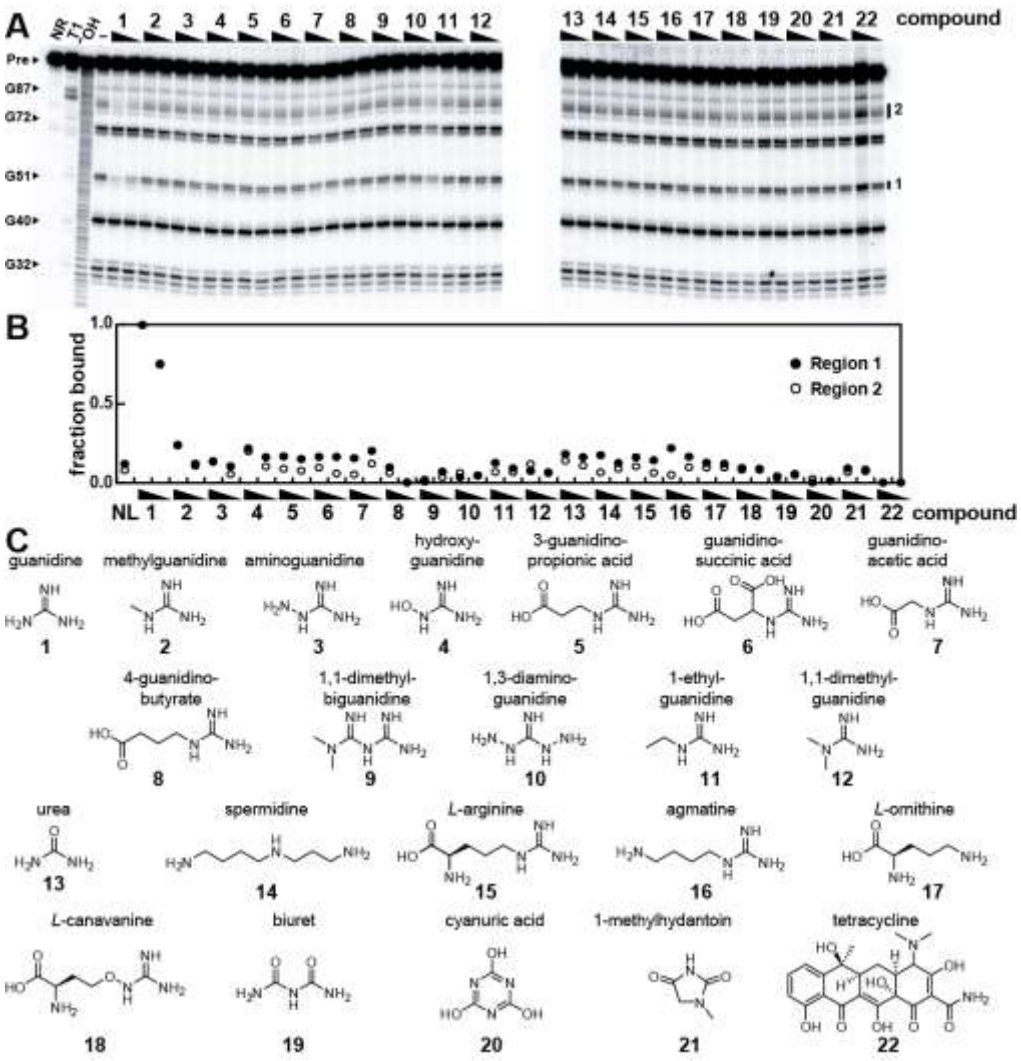

Figure 2-7. Selective recognition of guanidine by a representative guanidine-I riboswitch aptamer. (A) PAGE analysis of in-line probing reactions of the 104 *uca* aptamer from *N. europaea* in the presence of various analogs of guanidine (1 or 0.1 mM). Annotations are as described for Figure 2-6B. (B) Plot of the fraction of RNA bound at 1 or 0.1 mM for each compound tested. Data were calculated as described in Figure 2-6C using the data presented in A. (C) Chemical structures of compounds tested in A. Although tetracycline (22) is not an analog of guanidine, this compound was examined due to a recent claim that it serves as the ligand for *ykkC* RNAs. I performed the experiments shown in this figure.

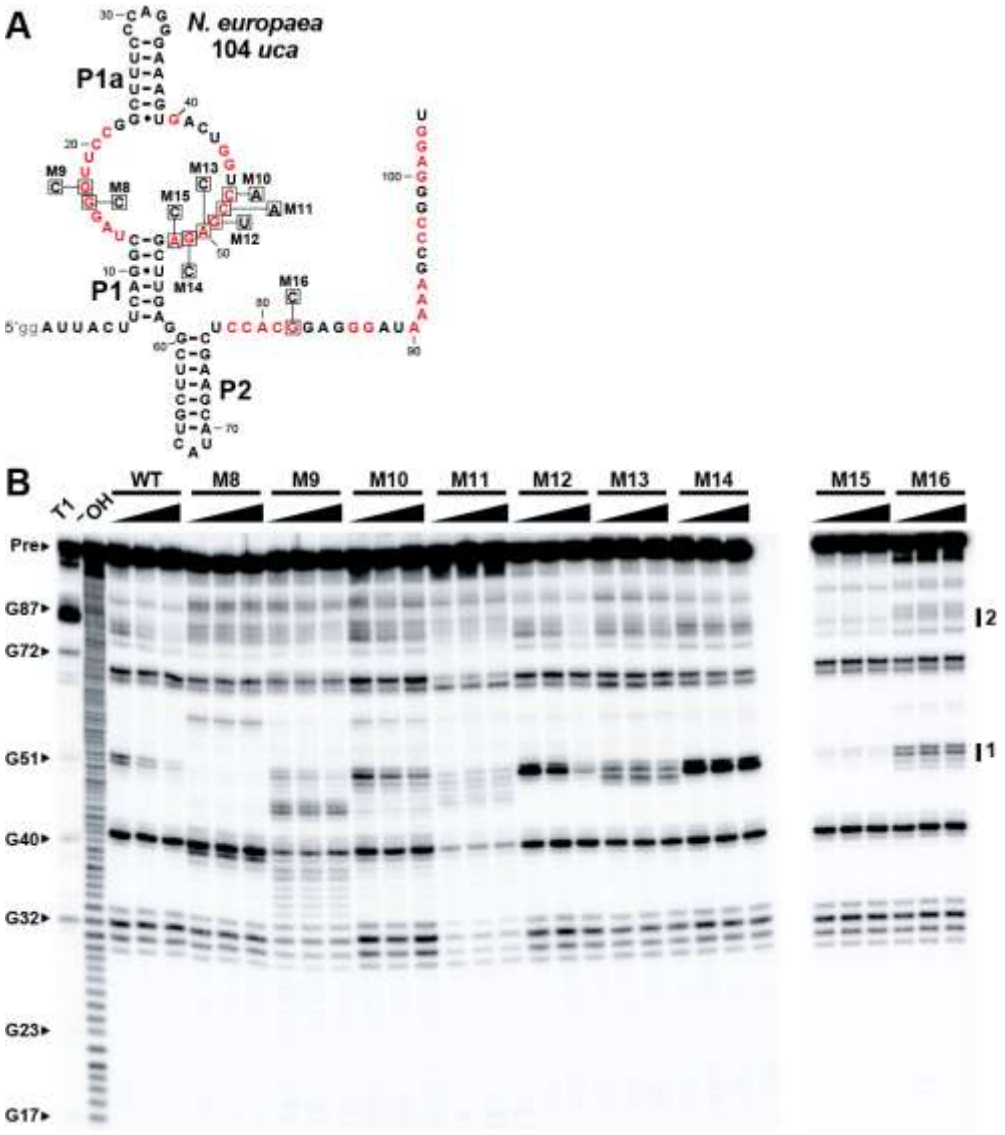

Figure 2-8. Mutation of numerous individual conserved nucleotides disrupts guanidine binding. (A) Location of mutations M8 through M16 within the 104 *uca* RNA construct (Figure 2-6A) derived from a guanidine-I riboswitch from *N. europaea*. Annotations are as described for Figure 2-6A. (B) PAGE analysis of in-line probing reactions of the 5′ ^{32}P-labeled 104 *uca* RNA from *N. europaea* containing various mutations to conserved nucleotides or secondary structures in the presence of 0, 0.1 mM, or 1 mM guanidine. Additional annotations are as described for Figure 2-6B. I performed the experiments shown in this figure.

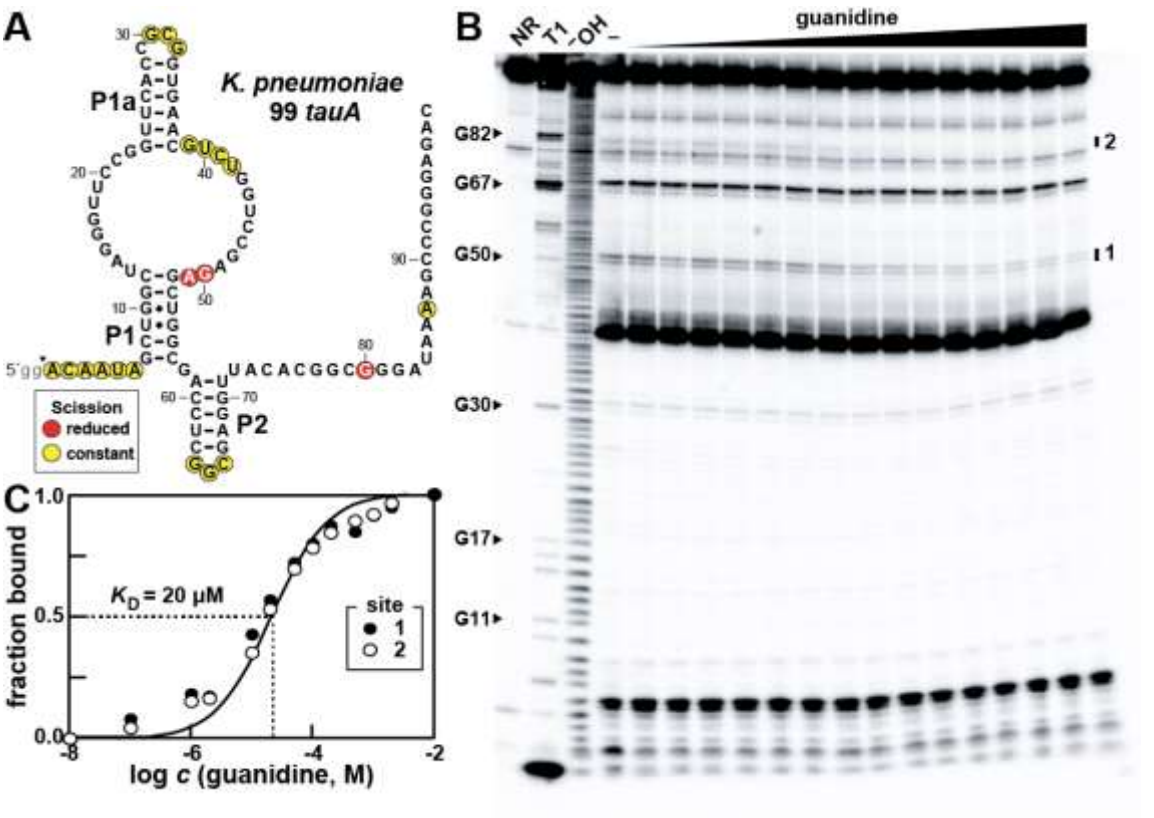

Figure 2-9. Ligand binding by a guanidine-I riboswitch aptamer from *K. pneumoniae*. (A) Sequence and secondary structure of the 99 *tauA* RNA from *K. pneumoniae*. This is the aptamer domain of the riboswitch used in the reporter system described in Figure 2-5. Annotations are as described for Figure 2-6A. Sites of ligand-induced structural modulation were determined using the data in B. (B) PAGE analysis of in-line probing reactions with 99 *tauA* in the presence of various guanidine concentrations. Annotations are as described for Figure 2-6B. (C) Plot of the fraction of RNA bound by guanidine derived from the normalized fraction of spontaneous RNA scission relative to the logarithm of the molar concentration of guanidine in panel B. The solid line is an idealized binding curve expected for a 1-to-1 binding interaction. Data are representative of multiple experiments. I performed the experiments shown in this figure.

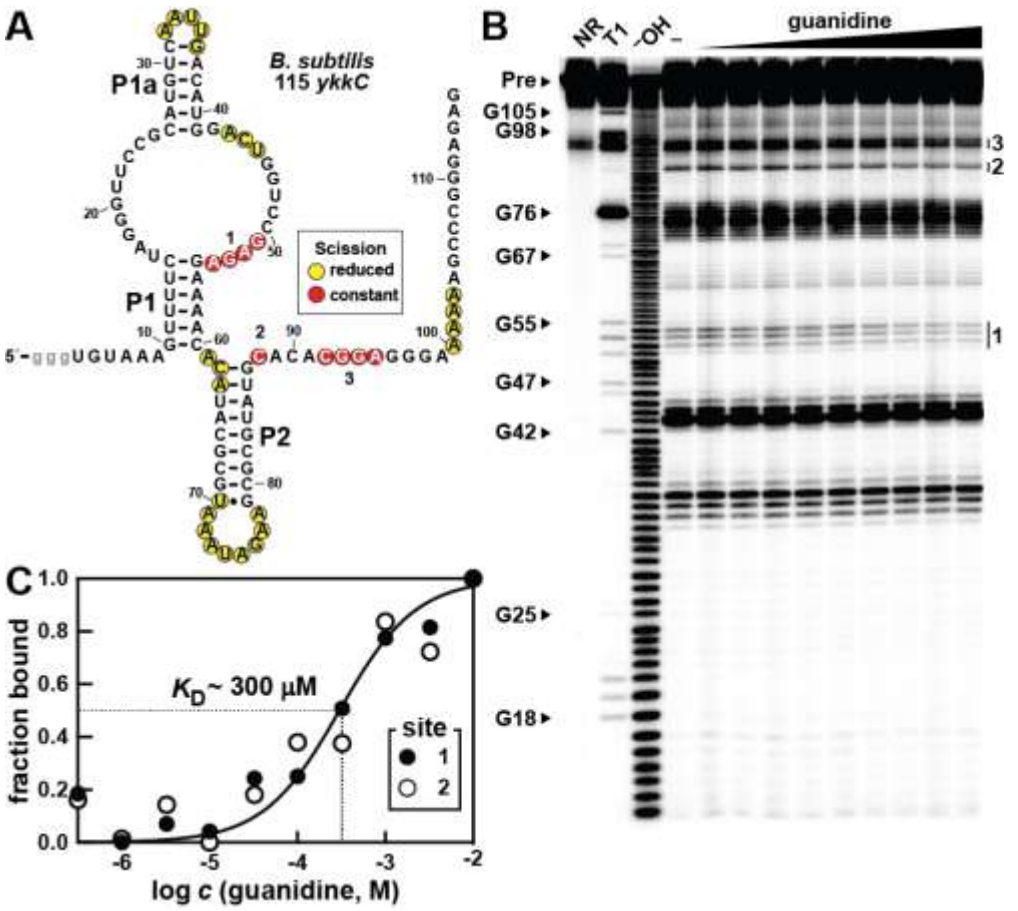

Figure 2-10. Ligand binding by a guanidine-I riboswitch aptamer from *B. subtilis*. (A) Sequence and secondary structure of the 115 *ykkC* RNA from *B. subtilis*. This is the aptamer domain of the riboswitch used in the reporter system described in Figure 2-3A. Annotations are as described for Figure 2-6A. Sites of ligand-induced structural modulation were determined using the data in B. (B) PAGE analysis of in-line probing reactions with 115 *ykkC* in the presence of various guanidine concentrations. Annotations are as described for Figure 2-6B. (C) Plot of the fraction of RNA bound by guanidine derived from the normalized fraction of spontaneous RNA scission relative to the logarithm of the molar concentration of guanidine in panel B. The solid line is an idealized binding curve expected for a 1-to-1 binding interaction. Data are representative of multiple experiments. James Nelson performed the experiments shown in this figure.

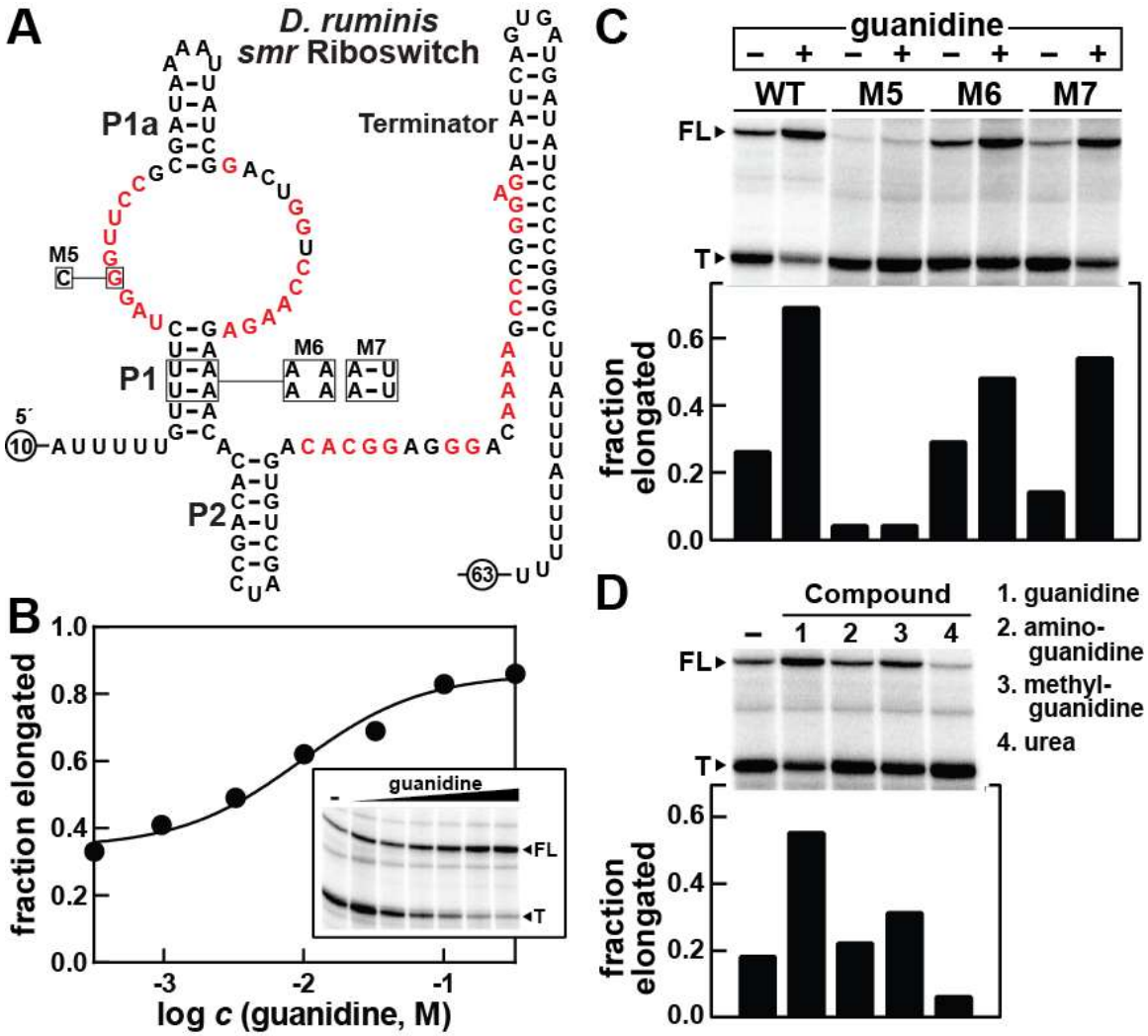

Figure 2-11. Guanidine regulates riboswitch-dependent transcript termination. (A) Sequence and secondary structure model of the guanidine-I riboswitch construct derived from the *smr* gene of *D. ruminis* and used for *in vitro* transcription assays. Annotations are as described in the legend for Figure 2-3. Terminated transcripts end within the U-rich tract immediately following the terminator stem, whereas full-length transcripts include the 63 nucleotides following the U-rich tract. (B) Plot of the fraction of full-length (FL) *smr* riboswitch transcripts contributing to the total number transcripts [FL plus terminated (T)] generated by single-round transcription reactions versus the logarithm of the molar

concentration of guanidine. The curve is an idealized representation of the results expected for a 1-to-1 interaction between guanidine and the riboswitch given the observed minimum and maximum values for fraction elongated. Inset: PAGE analysis used to generate the data plotted. Data shown here and in parts C and D are representative of repeated experiments performed on different days. (C) (Top) PAGE analysis of single-round transcription termination assays of WT and various mutant *smr* riboswitches. (Bottom) Plot of the fraction of FL RNA generated as derived from the PAGE data. (D) (Top) PAGE analysis of single-round transcription termination assays of WT *smr* riboswitches in the absence of ligand (–), in the presence of guanidine or its analogs (10 mM), or in the presence of urea (0.5 M). (Bottom) Plot of the fraction of FL RNA generated as derived from the PAGE data. James Nelson performed the experiments shown in this figure.

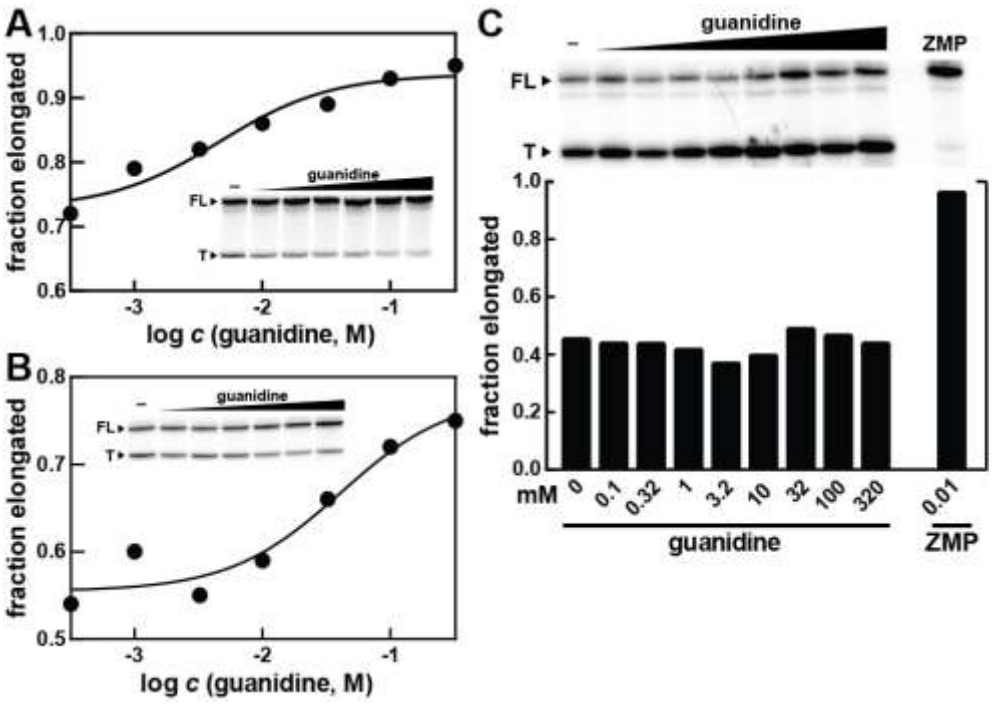

Figure 2-12. Guanidine regulates transcription termination of additional guanidine-I riboswitch representatives, but not a representative from another riboswitch class. (A) Plot of the fraction of FL RNA transcripts in single-round transcription assays of the guanidine-I riboswitch associated with the *emrE* gene from *C. difficile* versus the molar concentration of guanidine. Inset: PAGE analysis used to generate the data plotted. Annotations are as described for Figure 2-11B. Data shown here and in parts B and C are representative of experiments performed on different days. (B) Plot of the fraction of FL RNA transcripts in single-round transcription assays of the guanidine-I riboswitch associated with the *ykkC* gene from *B. subtilis* versus the molar concentration of guanidine. Inset: PAGE analysis used to generate the data plotted. Annotations are as described for Figure 2-11B. (C) (Top) PAGE analysis of single-round transcription assays of the ZTP riboswitch from *Clostridium beijerinckii* in the absence (–) or presence of guanidine or ZMP. (Bottom) Plot of the fraction of FL RNA generated as derived from the PAGE data. James Nelson performed the experiments shown in this figure.

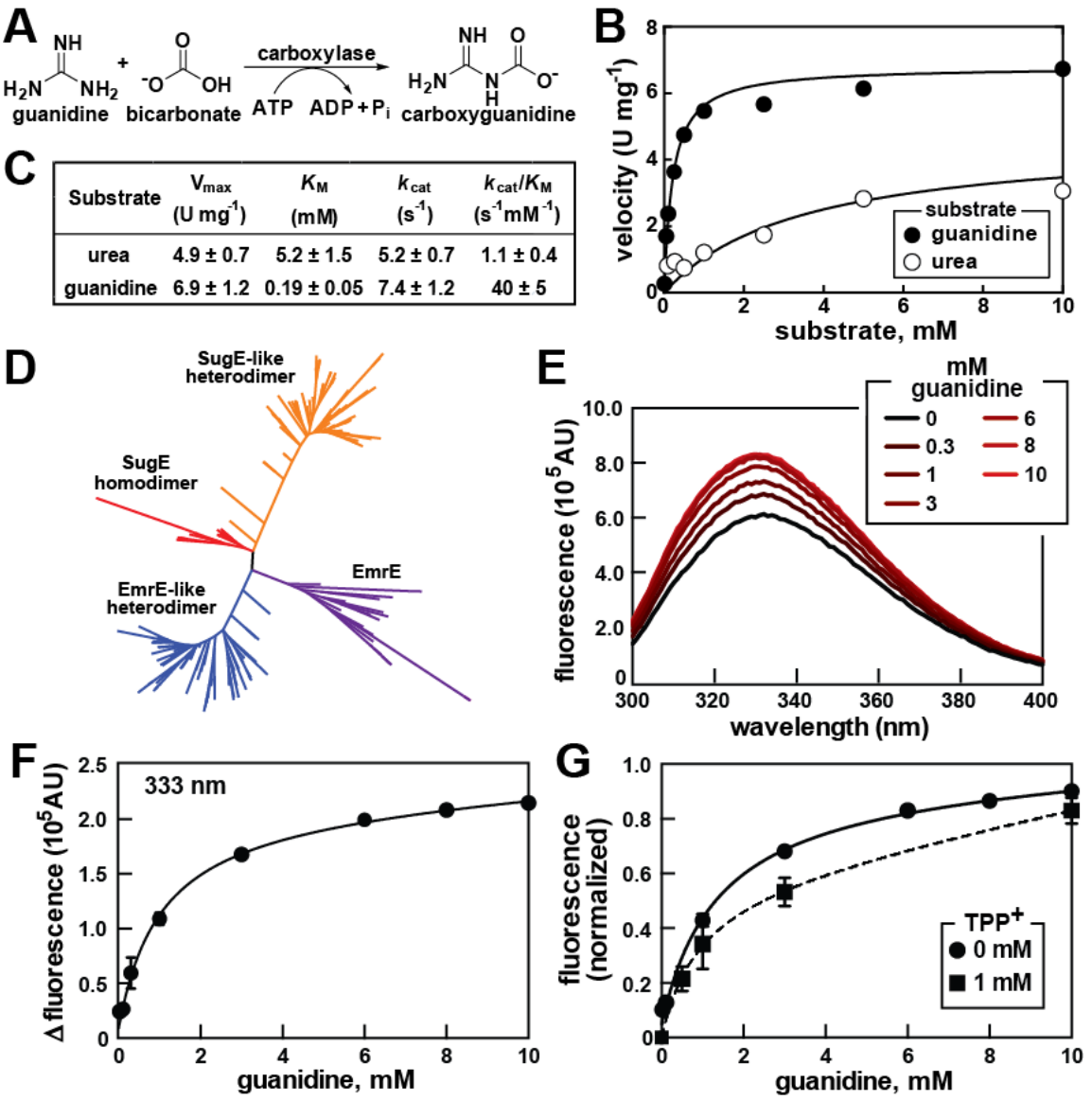

Figure 2-13. Guanidine-related functions for protein products of genes associated with guanidine-I riboswitches. (A) Proposed reaction scheme for the carboxylation of guanidine by an enzyme previously proposed to function as a urea carboxylase.[26] (B) Plot of the velocities of a previously defined urea carboxylase versus the concentration of guanidine or urea. Data points are the average of three technical replicates. (C) List of the enzymatic properties of a carboxylase enzyme with guanidine and urea substrates. Values shown are the averages and standard deviations of three independent experiments. (D)

Phylogenetic tree showing the genetic variability within SMR transporters controlled by guanidine-I riboswitches. Selected experimentally validated EmrE proteins that are known not to be controlled by the riboswitch are shown in purple. (E) Recognition of guanidine by predicted SMR transporters controlled by guanidine riboswitches evaluated by tryptophan emission spectra of *Clo* SugE at pH 7.5 in the presence of increasing guanidine. (F) Plot of the increase in fluorescence of tryptophan residues within *Clo* SugE at 333 nm versus the concentration of guanidine. Background fluorescence in the absence of guanidine is subtracted from each data point. Solid line represents the fit to a 1-to-1 binding model with a K_D of 1 mM. Error bars indicate the standard error of the experiment. When not visible, error bars are smaller than the data point. (G) Plot of the intrinsic fluorescence of tryptophan residues within *Clo* SugE at 333 nm versus the concentration of guanidine in the presence and absence of 1 mM TPP^+. Other annotations are as described for F. Data are the average of three measurements. Error bars indicate the standard error of the experiment. When not visible, error bars are smaller than the data point. I performed the experiments shown in panels A-C and Randy Stockbridge performed the experiments shown in this panels D-G.

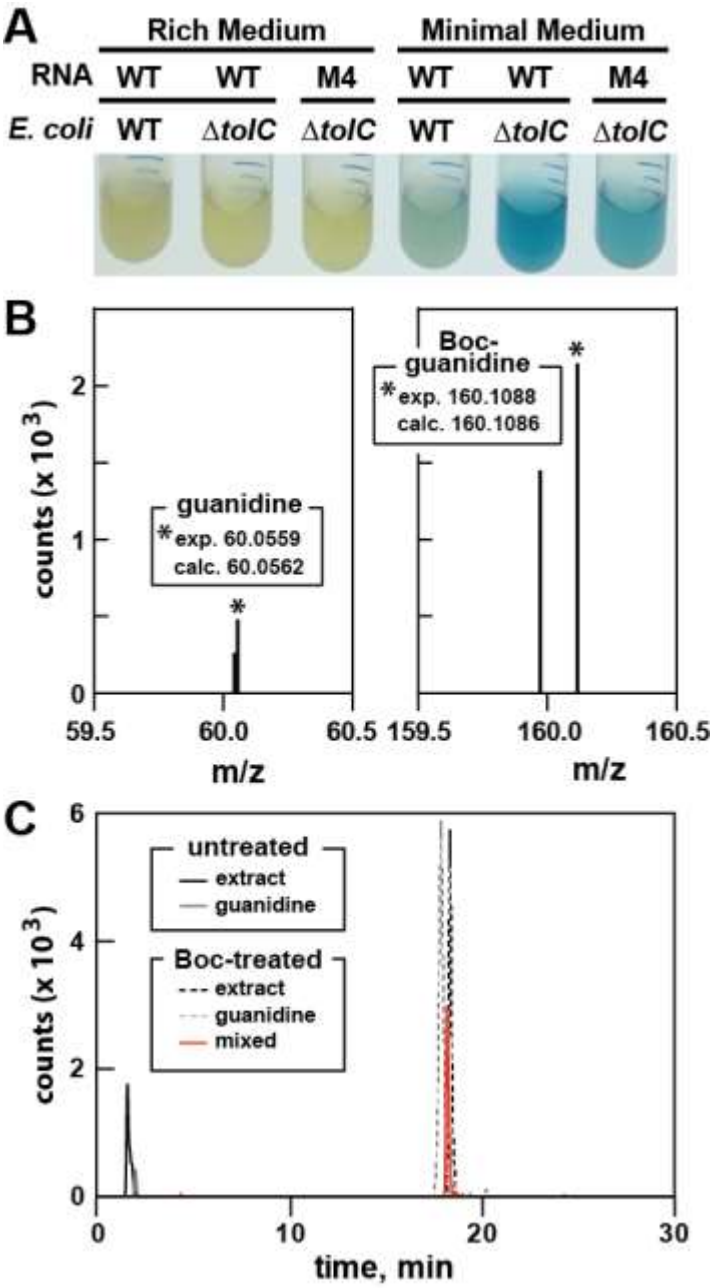

Figure 2-14. Natural production of free guanidine by bacteria. (A) *E. coli* WT and *tolC* knockout (Δ*tolC*) cells carrying the *K. pneumoniae* WT or M4 riboswitch reporter constructs (Figure 2-5) grown in rich (LB) or minimal (GMM) liquid media. (B) Plots of the mass-to-charge ratios (m/z) of compounds with similar retention times to authentic guanidine and authentic Boc-guanidine. Asterisks identify peaks from the extract LC-MS spectrum within 10 ppm of the calculated mass of each authentic compound as designated.

(C) LC-MS analysis for guanidine from the Δ*tolC* strain of *E. coli* grown in minimal medium as depicted in A. Extract and authentic guanidine samples were analyzed directly or were treated with Boc-anhydride to alter the mobility and mass of free guanidine as noted. Ruben Atilho performed the experiments shown in panel A and James Nelson performed the experiments shown in panels B-C.

TABLES

Table 2-1. Sequences of synthetic DNAs used in Chapter 2.

Name	Sequence	Annotation
ykkC001	TAATACGACTCACTATAGGGTGTAAAGTTTTCTAGGGTTCCGCATGTCAATTGACATGGACTGGTCCGAGAGAAAAC	Forward primer for the construction of a template for transcription of the *B. subtilis* 100 *ykkC* RNA
ykkC002	CTCTCCCGGGCTTTTTTCCCTCCGTGTGCATACGCGCTTCTATTTACGCGTATGTGTTTTCTCTCGGACCAGTCC	Reverse primer for the construction of a template for transcription of the *B. subtilis* 100 *ykkC* RNA
ykkC003	TAATACGACTCACTATAGGACAATAGCTGGCTAGGGTTCCGGTTCACCGCGGTGAACGTCTGGTCCGAGAGCTGGCGAC	Forward primer for the construction of a template for transcription of the *K. pneumoniae* 99 *tauA* RNA
ykkC004	GTCTCCCGGGCTTTTATCCCGCCGTGTAACCTCGCCGAGGTCGCCAGCTCTCGGACCAG	Reverse primer for the construction of a template for transcription of the *K. pneumoniae* 99 *tauA* RNA

ykkC005	TAATACGACTCACTATAGGATTACTTCAGGCTAGGGTTCCGGCTTTCCCAGGGAAAGTGACTGGTCCG	Forward primer for the construction of a template for transcription of the *N. europaea* 104 *uca* RNA
ykkC006	ACCTCCCGGGCTTTTATCCCTCCGTGGAGCTTCGTATGACGAAGCCTCAAGCTCTCGGACCAGTCACTTTCCCTG	Reverse primer for the construction of a template for transcription of the *N. europaea* 104 *uca* RNA
ykkC007	TAATACGACTCACTATAGGATTACTTCAGGCTAGCGTTCCGGCTTTCCCAGGGAAAGTGACTGGTCCG	Forward primer for the construction of a template for transcription of the *N. europaea* 104 *uca* RNA containing M8
ykkC008	TAATACGACTCACTATAGGATTACTTCAGGCTAGGCTTCCGGCTTTCCCAGGGAAAGTGACTGGTCCG	Forward primer for the construction of a template for transcription of the *N. europaea* 104 *uca* RNA containing M9
ykkC009	TAATACGACTCACTATAGGATTACTTCAGGCTAGGGTTCCGGCTTTCCCAGGGAAAGTGACTGGTACG	Forward primer for the construction of a template for transcription of the *N. europaea* 104 *uca* RNA containing M10

ykkC010	ACCTCCCGGGCTTTTATCCCTCC GTGGAGCTTCGTATGACGAAGCC TCAAGCTCTCGTACCAGTCACTT TCCCTG	Reverse primer for the construction of a template for transcription of the *N. europaea* 104 *uca* RNA containing M10
ykkC011	TAATACGACTCACTATAGGATTA CTTCAGGCTAGGGTTCCGGCTTT CCCAGGGAAAGTGACTGGTCAG	Forward primer for the construction of a template for transcription of the *N. europaea* 104 *uca* RNA containing M11
ykkC012	ACCTCCCGGGCTTTTATCCCTCC GTGGAGCTTCGTATGACGAAGCC TCAAGCTCTCTGACCAGTCACTT TCCCTG	Reverse primer for the construction of a template for transcription of the *N. europaea* 104 *uca* RNA containing M11
ykkC013	TAATACGACTCACTATAGGATTA CTTCAGGCTAGGGTTCCGGCTTT CCCAGGGAAAGTGACTGGTCCT	Forward primer for the construction of a template for transcription of the *N. europaea* 104 *uca* RNA containing M12
ykkC014	ACCTCCCGGGCTTTTATCCCTCC GTGGAGCTTCGTATGACGAAGCC TCAAGCTCTAGGACCAGTCACTT TCCCTG	Reverse primer for the construction of a template for transcription of the *N. europaea* 104 *uca* RNA containing M12

ykkC015	ACCTCCCGGGCTTTTATCCCTCC GTGGAGCTTCGTATGACGAAGCC TCAAGCTCGCGGACCAGTCACTT TCCCTG	Reverse primer for the construction of a template for transcription of the *N. europaea* 104 *uca* RNA containing M13
ykkC016	ACCTCCCGGGCTTTTATCCCTCC GTGGAGCTTCGTATGACGAAGCC TCAAGCTGTCGGACCAGTCACTT TCCCTG	Reverse primer for the construction of a template for transcription of the *N. europaea* 104 *uca* RNA containing M14
ykkC017	ACCTCCCGGGCTTTTATCCCTCC GTGGAGCTTCGTATGACGAAGCC TCAAGCGCTCGGACCAGTCACTT TCCCTG	Reverse primer for the construction of a template for transcription of the *N. europaea* 104 *uca* RNA containing M15
ykkC018	ACCTCCCGGGCTTTTATCCCTCG GTGGAGCTTCGTATGACGAAGCC TCAAGCTCTCGGACCAGTCACTT TCCCTG	Reverse primer for the construction of a template for transcription of the *N. europaea* 104 *uca* RNA containing M16
ykkC019	TAATACGACTCACTATAGGCAAT GATATGCTATAAAATAATCTCCG GACGAGACGCGCCGTACCCGGTT AAGACAAGACGACGCCTTTA	Forward template for transcription of the *L. pneumophila* 71 *ykkC*–III RNA

ykkC020	TAAAGGCGTCGTCTTGTCTTAAC CGGGTACGGCGCGTCTCGTCCGG AGATTATTTTATAGCATATCATT GCCTATAGTGAGTCGTATTA	Reverse template for transcription of the *L. pneumophila* 71 *ykkC*–III RNA
ykkC021	ATAAGTGTAAAATTATTCTTATT GACAAACTAAATGGTTTGGTATT ATTCTTATATTAGATATGATGAT ATCATGTAAAGTTTTCTAGGGTT CCGCATGTCAATTGACATGGACT GGTCCGAGAGAAAACACATACG CGTAAATAGAAGCGCGTATG	Forward primer for the construction of single-round transcription termination template from *B. subtilis*
ykkC022	TTTGAGTCATTCCTTTCTATCTAT TAAGATTCTCCGAAAACAAAAA AACCCCGGACAGTCGTCTCACAT GAGATTGACTCTCCCGGGCTTTT TTCCCTCCGTGTGCATACGCGCT TCTATTTACGCGTATGTGTTTTCT CTCGG	Reverse primer for the construction of single-round transcription termination template from *B. subtilis*
ykkC023	ATAAGTGTAAAATTATTCTTATT GACAAACTAAATGGTTTGGTATT ATTCTTATATTAGATAGAATAAT TAATAGAGTTTTATAGGGTTCCG	Forward primer for the construction of single-round

	CAATTATTAATGATTGGTTTGGT CCGAGATAAAACCCACAACATA GTTTTTGTTGTGAACACGGAAGG	transcription termination template from *C. difficile*
ykkC024	TTTGAGTCATTCCTTTCTATCTAT TAAGATTCTCCGATCCCTCTTTTC AAATATGTCTCTTATTTTTCCAAT AAAATAACCTGAGAAATATTATA AAATATCTCCCAGGTTTTTGTCCT TCCGTGTTCACAACAAAAACTAT GTTG	Reverse primer for the construction of single-round transcription termination template from *C. difficile*
ykkC025	ATAAGTGTAAAATTATTCTTATT GACAAACTAAATGGTTTGGTATT ATTCTTATATTAGATAAAATATA TATTTTTGTTTTCTAGGGTTCCGC GATAAAATTATCGGACTGGTCCA AGAGAAAACACACAGCCTAGCT GTGACACGGAGGGACAAA	Forward primer for the construction of single-round transcription termination template from *D. ruminis*
ykkC026	TTTGAGTCATTCCTTTCTATCTAT TAAGATTCTCCGGAAAGTCATGT CTACTTCCTCCTTATTAAAAAAT AAATAAGCCCGGGGATATCATCA CTGATATCTCCCGGGCTTTTGTCC	Reverse primer for the construction of single-round transcription termination template from *D. ruminis*

	CTCCGTGTCACAGCTAGGCTGTG TGTTTT	
ykkC027	TTTGAGTCATTCCTTTCTATCTAT TAAGATTCTCCGGAAAGTCATGT CTACTTCCTCCTTATTAAAAAAT AAATAAGCGGTTTTATATCATCA CTGATATCTCCCGGGCTTTTGTCC CTCCGTGTCACAGCTAGGCTGTG TGTTTT	Reverse primer for the construction of single-round transcription termination template from *D. ruminis* containing a mutation in the terminator stem
ykkC028	ATAAGTGTAAAATTATTCTTATT GACAAACTAAATGGTTTGGTATT ATTCTTATATTAGATAAAATATA TATTTTTGTTTTCTAGCGTTCCGC GATAAAATTATCGGACTGGTCCA AGAGAAAACACACAGCCTAGCT GTGACACGGAGGGACAAA	Forward primer for the construction of single-round transcription termination template from *D. ruminis* with a mutation at a conserved nucleotide
ykkC029	ATAAGTGTAAAATTATTCTTATT GACAAACTAAATGGTTTGGTATT ATTCTTATATTAGATAAAATATA TATTTTTGTAATCTAGGGTTCCGC GATAAAATTATCGGACTGGTCCA	Forward primer for the construction of single-round transcription termination template from *D. ruminis* with a mutation in the P1 stem

	AGAGAAAACACACAGCCTAGCTGTGACACGGAGGGACAAA	
ykkC030	ATAAGTGTAAAATTATTCTTATTGACAAACTAAATGGTTTGGTATTATTCTTATATTAGATAAAATATATATTTTTGTAATCTAGGGTTCCGCGATAAAATTATCGGACTGGTCCAAGAGATTACACACAGCCTAGCTGTGACACGGAGGGACAAA	Forward primer for the construction of single-round transcription termination template from *D. ruminis* with a compensatory mutation in the P1 stem
ykkC031	ATAAGTGTAAAATTATTCTTATTGACA	Forward primer for the amplification of all termination constructs
ykkC032	TTTGAGTCATTCCTTTCTATC	Reverse primer for the amplification of all termination constructs
ykkC033	TACGACGAATTCCAAAAATAATG	Forward primer for the construction of *B. subtilis ykkC* reporter construct containing the *B. subtilis lysC* promoter containing sites for cloning into pDG1661

ykkC034	TCCTTTGGATCCTATTAAGATTCTCCG	Reverse primer for the construction of *B. subtilis ykkC* reporter construct containing the *B. subtilis lysC* promoter, containing sites for cloning into pDG1661
ykkC035	TACGACAAATTGCAAAAATAATGTTGTCCTTTTAAATAAGATCTGATAAAATGTGAACTAATATGATGATATCATGTAAAGTTTTCTA	Forward primer for the construction of *B. subtilis ykkC* termination construct containing the *B. subtilis lysC* promoter, used as a template for construction of the *B. subtilis ykkC* reporter construct
ykkC036	TTTGGAGTCATTCCTTTCTATCTATTAAG	Reverse primer for the construction of *B. subtilis ykkC* termination construct containing the *B. subtilis lysC* promoter, used as a template for construction of the *B. subtilis ykkC* reporter construct

ykkC037	GATGATATCATGTAAAGTAATCTAGG	Forward primer for construction of the *B. subtilis ykkC* reporter with the *lysC* promoter containing M2 via QuikChange mutagenesis
ykkC038	CCTAGATTACTTTACATGATATCATC	Reverse primer for construction of the *B. subtilis ykkC* reporter with the *lysC* promoter containing M2 via QuikChange mutagenesis
ykkC039	CATGTAAAGTTTTCTAGCGTTC	Forward primer for construction of the *B. subtilis ykkC* reporter with the *lysC* promoter containing M1 via QuikChange mutagenesis
ykkC040	GAACGCTAGAAAACTTTACATG	Reverse primer for construction of the *B. subtilis ykkC* reporter with the *lysC* promoter containing M1 via QuikChange mutagenesis
ykkC041	GGTCCGAGAGATTACACATACG	Forward primer for construction of the *B. subtilis ykkC* reporter with the *lysC* promoter containing M3 via QuikChange mutagenesis

ykkC042	CGTATGTGTAATCTCTCGGACC	Reverse primer for construction of the *B. subtilis ykkC* reporter with the *lysC* promoter containing M3 via QuikChange mutagenesis
ykkC043	CTGCCAGGAATTGGGGATCGGA ATTCGCATGACCCCATCGTTGAC AACCGCCCCGCTCACCCTTTATT TATAAATGTACCAATTCAGATCA CACAATAGCTGGCTAGGGTTCCG GTTCACCGCGGTGAACGTCTGGT CCGAGAGCTGGCGACCTCGGCG AGGTTACACGGCGGGATAAAAG CCCGGGAGACAGCAGCACCGTT GGGTGTCGCGCTGCTCCGTTTCT CGTGCTAACCAGAGGTCCACCAT GACCGTGATTACGGATTCAGGGG ATCCCGTCGTTTTACAACGTCG	IDT G-block containing the wild-type *K. pneumoniae tauA* riboswitch and first 8 codons of *lacZ* gene controlled by the *E. coli lysC* promoter, for cloning into pRS414
ykkC044	CTGGCTAGCGTTCCGGTTC	Forward primer for construction of the *K. pneumoniae tauA* reporter with the *lysC* promoter containing M4 via QuikChange mutagenesis

ykkC045	GAACCGGAACGCTAGCCAG	Reverse primer for construction of the *K. pneumoniae tauA* reporter with the *lysC* promoter containing M4 via QuikChange mutagenesis
ykkC046	TTGGTGATGAGCTCGTTAGGAGG GGAACTGAACATGAAG	Forward primer for amplification of the 5' flank of the *ykkCD* KO cassette, for cloning into pDG1515 and subsequent construction of the knockout via homologous recombination
ykkC047	ATCACCAACTGCAGCTTTCAATT GTAAGATTTTCTTTCATATACCG C	Reverse primer for amplification of the 5' flank of the *ykkCD* KO cassette, for cloning into pDG1515 and subsequent construction of the knockout via homologous recombination
ykkC048	TTGGTGATGTCGACGCTCAGCAG TTGGTTTAAAAATTCTGTC	Forward primer for amplification of the 3' flank of the *ykkCD* KO cassette, for cloning into pDG1515 and subsequent

		construction of the knockout via homologous recombination
ykkC049	ATCACCAAGGTACCGGGCCTCTTGCAGTCTAGTTAAAGAC	Reverse primer for amplification of the 3' flank of the *ykkCD* KO cassette, for cloning into pDG1515 and subsequent construction of the knockout via homologous recombination

CHAPTER THREE

Biochemical validation of a second guanidine riboswitch class in bacteria

Largely adapted from the following publication:

Sherlock, M.E., Malkowski, S.N., Breaker, R.R. (2017) Biochemical validation of a second guanidine riboswitch class in bacteria. *Biochemistry 56*, 352-358.

ABSTRACT

In Chapter 2, my colleagues and I determined that representatives of the riboswitch candidates called *ykkC* directly bind free guanidine. Guanidine-binding *ykkC* motif RNAs, now renamed guanidine-I riboswitches, commonly regulate the expression of genes encoding guanidine carboxylases, as well as others encoding guanidine efflux proteins such as EmrE and SugE. Prior to the validation of guanidine as the ligand for these newly-established riboswitch classes, another RNA motif was discovered by comparative genomic analysis and termed mini-*ykkC* due to its small size and gene associations similar to the original *ykkC* motif. It was hypothesized that these distinct RNA structures might respond to the same ligand. However, the small size and repetitive nature of mini-*ykkC* RNAs suggested that it might respond to ligand via the action of a protein factor. Herein my colleagues and I demonstrate that, despite its extremely simple architecture, mini-*ykkC* motif RNAs constitute a distinct class of guanidine-sensing RNAs, called guanidine-II riboswitches. Surprisingly, each of the two stem loops that comprise the mini-*ykkC* motif appears to directly bind free guanidine in a cooperative manner. These findings reveal that bacteria make extensive use of diverse guanidine-responsive riboswitches to overcome the toxic effects of this compound.

INTRODUCTION

The *ykkC* motif is a conserved and highly-structured bacterial noncoding RNA that was first reported in 2004.[1] For over a decade, the cognate ligand for the *ykkC* riboswitch remained unidentified, mostly due to the apparent lack of connection amongst genes regulated by this RNA. Despite the challenges associated with this orphan riboswitch candidate, the search for the ligand for the putative *ykkC* riboswitch class continued. This effort was driven by the fact that the unidentified ligand would have an important, but likely unknown or underappreciated, role in bacteria because this riboswitch candidate is so common and widespread.[2]

The importance of this mystery ligand was further implicated by the discovery of two additional RNA motifs with highly similar gene associations, termed mini-*ykkC*[3] and *ykkC*-III.[4] It was predicted that all three types of *ykkC* RNA motifs might respond to this same, yet-to-be-identified compound. Recently, my colleagues and I experimentally validated representatives of the *ykkC* candidate riboswitch class to sense and respond to free guanidine.[5] Consequently, the subset of *ykkC* motif RNAs that have been demonstrated or are predicted to respond to this ligand have been renamed guanidine-I riboswitches. This work further revealed that free guanidine is naturally produced by bacteria and that multiple pathways exist to minimize its toxic effects. Similarly, I provide evidence in Chapter 4 that members of the *ykkC*-III class also function as guanidine-sensing RNAs, named guanidine-III riboswitches.

Upon the experimental validation of guanidine-I riboswitches, I sought to assess the hypothesis that mini-*ykkC* motif RNAs might function as a distinct class of guanidine-sensing riboswitches. However, the mini-*ykkC* motif was initially considered a poor

candidate for a riboswitch class due to its low structural complexity and its limited number of conserved nucleotides.[3] Additionally, repetitive stem loops with identical sequences were likely to be more characteristic of binding sites for RNA-binding protein homo-complexes. Thus, it seemed more plausible that mini-*ykkC* RNAs would require a protein partner to perform their regulatory role.

The mini-*ykkC* motif commonly controls the expression of small multidrug resistance (SMR) family efflux pumps including those annotated as EmrE and SugE. In our previous work, we show that a representative of the SMR proteins typically controlled by guanidine-I riboswitches is able to bind guanidine selectively.[5] Additionally, guanidine-I riboswitches and mini-*ykkC* RNAs are commonly located adjacent to genes originally annotated as urea carboxylase and associated proteins, and genes for allophanate hydrolase. My previous work revealed that a representative urea carboxylase protein from *O. sagaranensis* prefers guanidine as a substrate, with a 40-fold higher catalytic efficiency for guanidine versus urea.[5] Thus, mini-*ykkC* motif RNAs remained intriguing candidates as guanidine-responsive regulatory elements despite their simple architecture.

In this chapter, my colleagues and I present results confirming that mini-*ykkC* RNAs directly and selectively respond to guanidine. Also described are additional structural and biochemical attributes of mini-*ykkC* motif RNAs, which we rename as guanidine-II riboswitches. Most importantly, guanidine-II riboswitches appear to bind two guanidine molecules cooperatively, and therefore exhibit unusually complex characteristics for such a small and simple motif.

RESULTS AND DISCUSSION

Additional mini-*ykkC* motif RNA representatives form highly-conserved symmetrical loops. Initial bioinformatics revealed 208 representatives of the mini-*ykkC* motif RNA.[3] I conducted an updated homology search using Infernal[7] to further expand the number of representatives. I found approximately 800 unique examples mainly in Proteobacteria, with additional examples in Actinobacteria, Chrysiogenetes, Cyanobacteria, Planctomycetes, and Verrucomicrobia.

I used the updated homology search results to refine the consensus sequence and secondary structure model for this motif (Figure 3-1A). Overall, mini-*ykkC* motif RNAs are composed of two GC-rich stems, called pairing element 1 and 2 (P1 and P2), that are capped with 'ACGR' (R represents either G or A) tetraloops. These hairpins are connected by a non-conserved linker formed by a minimum of seven nucleotides. The presence of the repetitive loop sequences flanked by GC-rich stem sequences was suggestive of the presence of two binding sites for a possible riboswitch ligand or for a dimeric protein factor.

Examples of mini-*ykkC* motif RNAs frequently occur in close proximity (typically ~30 nucleotides) to the ribosome binding site (RBS) and start codon of each adjoining open reading frame (ORF). This arrangement is characteristic of regulatory RNA domains such as riboswitches. Moreover, if mini-*ykkC* motif RNAs enact their regulatory response in a manner consistent with guanidine-I riboswitches, then genes that alleviate guanidine toxicity should be expressed when guanidine concentrations are high. Accordingly, guanidine should interact with the ligand-binding aptamer domain in a manner that exposes the RBS and allows the translational machinery to initiate protein production.

Genes associated with mini-*ykkC* are related to guanidine metabolism. The genes associated with mini-*ykkC* motif RNAs (Figure 3-1B) largely overlap with those controlled by guanidine-I riboswitches, although the SMR transporters are more highly represented. Comparative analysis suggests that these predicted SMR proteins fall into the category of SugE-like heterodimers (Randy Stockbridge, personal communication). The protein products from these genes are distinct from the well-characterized canonical EmrE protein from *E. coli*, which has broad specificity for small toxic compounds.[12] Members of mini-*ykkC* are also associated with genes annotated as urea carboxylases, which are also commonly controlled by guanidine-I riboswitches. As noted above, I recently demonstrated that a representative of this enzyme class favors guanidine as a substrate over urea by 40 fold.[5]

Other gene associations, such as those encoding urea carboxylase associated proteins 1 and 2, and allophanate hydrolases, are also consistent with our hypothesis that mini-*ykkC* motif RNAs regulate gene expression in response to guanidine. An additional gene, *chaA*, whose expression is occasionally associated with mini-*ykkC*, encodes a predicted Ca^{2+}/H^{+} antiporter. Perhaps this activity assists in maintaining the membrane potential of the cell as it ejects positively charged guanidinium from the cytoplasm. The remaining genes associated with mini-*ykkC* motif RNAs are either very rarely affiliated, or have unknown functions (Figure 3-1B).

Intriguingly, mini-*ykkC* motif RNAs are occasionally encoded within apparent operons for a putative guanidine degradation pathway. Such genomic arrangements typically consist of genes annotated as encoding nitrate/sulfate/bicarbonate transporters followed by urea carboxylase associated proteins, urea carboxylase and finally allophanate hydrolase.

In many cases a guanidine-I riboswitch, which is a transcriptional 'ON' switch, appears in the 5′ UTR between the promoter and the first gene. Further into the operon, a mini-*ykkC* element is found in the intergenic region between the urea carboxylase associated genes and urea carboxylase. I predict that mini-*ykkC* is a translational 'ON' switch that regulates expression of the urea carboxylase gene immediately downstream by controlling access to the RBS and thereby translation initiation.

Initially, this seems like either an unnecessarily complex arrangement of two regulatory RNAs that respond to the same ligand, or hints at the existence of two distinct ligands for these two regulatory RNA structures. However, the RNA elements might demark the beginning of two different transcripts, or bacteria might be employing the mini-*ykkC* motif RNA as a secondary checkpoint before investing in the translation of the massive 125 kDa urea carboxylase protein long after the mRNA for the full operon has been transcribed.

Guanidine is bound by mini-*ykkC* motif RNAs with affinity similar to guanidine-I riboswitches. The experimental validation of guanidine as the ligand for the majority of *ykkC* motif RNAs[5] strongly implicated this same ligand as the trigger for the predicted gene regulation mediated by mini-*ykkC* RNAs. The major question was whether members of this RNA class act independently to directly sense guanidine or whether they require an additional factor, protein or otherwise, to respond to guanidine.

To assess ligand-binding potential, I prepared an RNA construct encompassing 60 nucleotides centered on the mini-*ykkC* motif of the *emrE* gene of *Gloeobacter violaceus* (Figure 3-2A). Note that, although this gene is annotated as *emrE*, it actually codes for a protein that is SugE-like, as do all genes regulated by guanidine riboswitches. This small RNA construct, called '*G. violaceus* 60 *emrE*', was subjected to in-line probing assays[10,11]

to monitor changes in spontaneous RNA degradation in response to ligand binding. Surprisingly, guanidine directly causes robust modulation of spontaneous RNA cleavage throughout the RNA construct (Figure 3-2B), indicating that the RNA substantially alters its shape upon ligand binding.

Most notably, both conserved ACGR loops of this RNA construct undergo dramatic reductions in spontaneous cleavage, which is consistent with guanidine-induced structure formation. Given the near identical sequence and structural features of P1 and P2, we considered the possibility that these repeated substructures might collaborate to form two binding pockets for guanidine. Some structural modulation is also observed in the non-conserved linker between the P1 and P2 stems, which might be required to allow crosstalk between the two, nearly-identical stem-loop substructures.

By plotting the relative increase or decrease in RNA scission at positions in four distinct regions of the construct versus the concentration of guanidine, I observed a guanidine-dependent dose-response curve with a Hill coefficient of ~1.4 (Figure 3-2C). A value greater than 1 suggests that multiple processes are being coupled, which is likely to be two or more ligand binding events. I therefore anticipate the existence of two binding sites for guanidine that function cooperatively such that binding at the first site improves binding affinity at the second site, and vice versa. It seems likely that an individual guanidine binding pocket will be formed in large part by a single highly-conserved ACGR loop sequence, presumably with some contact with the neighboring stem-loop structure. The apparent overall dissociation constant (K_D) of 50 μM also compares favorably to those exhibited by guanidine-I[5] and guanidine-III riboswitches.[6] These findings are consistent

with riboswitch function for mini-*ykkC* motif RNAs, and hereafter I call members of this class 'guanidine-II' riboswitches.

I selected two additional representative guanidine-II riboswitch aptamer sequences to confirm that mini-*ykkC* sequences from a distantly related organism, or with different gene associations, also respond cooperatively to guanidine. Both the 60-nucleotide aptamer associated with the *sugE* gene from *Escherichia coli* (*E. coli* 60 *sugE*, Figure 3-3) and the 60-nucleotide aptamer associated with the urea carboxylase gene from *Klebsiella pneumoniae* (60 *uca*, Figure 3-4) exhibit similar guanidine-dependent patterns of structure modulation. Specifically, each construct undergoes dramatic stabilization of both conserved ACGR loops with Hill coefficients greater than 1. Moreover, the *E. coli* 60 *sugE* and the 60 *uca* constructs exhibit K_D values for guanidine of 300 μM and 200 μM, respectively. Again, the binding affinities for these diverse guanidine-II riboswitch aptamers are similar those of guanidine-I[5] and guanidine-III[6] riboswitches. My results demonstrate that guanidine-II riboswitches are present in many bacterial species where they sense and respond cooperatively to guanidine.

Mutation of conserved nucleotides confirms multiple guanidine binding sites. To assess the importance of highly conserved nucleotides, and to further evaluate the multiple binding site hypothesis, my colleague Sarah Malkowski and I designed a series of mutant *G. violaceus* 60 *emrE* RNA constructs (Figure 3-5A) and assessed the effects of these mutations on guanidine binding. Specifically, Sarah Malkowski prepared eight mutant constructs (M1 through M8) that each carry a single nucleotide change in one of the two ACGR loops. A ninth construct (M9) was prepared that carries a single mutation in both of the ACGR loops.

Mutations M1 through M3, which are located in the loop of P1, completely eliminate guanidine-dependent structural modulation in this loop (Figure 3-5). Moreover, structural modulation is greatly diminished in the unaltered loop of P2, even at a guanidine concentration as high as 1 mM, which would be expected if two cooperative ligand binding sites are present. Reduced structure modulation is detected at 1 mM guanidine for M4 versus WT, although it is worth noting that this mutation also weakens the apparent affinity for guanidine at the second putative binding site. Additionally, this fourth position of the tetraloop is the least conserved, in that either purine nucleotide is acceptable.

Accordingly, Sarah Malkowski observed near symmetrical results for mutations M5 through M7, which exhibit little or no modulation in the loop of P2 upon guanidine addition. However, these mutations more modestly weaken ligand binding at the opposing site than do direct mutations in the loop of P1. Further analysis of M1 and M5 revealed that a mutation in one binding site decreases the affinity at that site by ~100 fold, but the other site only experiences a ~10-fold decrease in affinity for guanidine binding (Figure 3-6). Additionally, the Hill slope for these interactions is approximately 1, indicating that a single binding event at each site is being observed. When a mutation is simultaneously introduced into both loops (M9), binding is not observed at either site (Figure 3-6B).

Both stem loops are necessary for guanidine binding. The findings described above are consistent with the presence of two distinct guanidine binding pockets, and indicate that these two sites are intimately linked to cooperatively affect ligand binding. We considered the possibility that a single stem loop might be able to bind guanidine when tested in the context of a truncated RNA that lacks the second site and any possible

negative-acting structures that compete with the ligand-bound architecture of the riboswitch aptamer.

To examine this possibility, Sarah Malkowski created a truncated 30 *emrE* RNA from *G. violaceus* and conducted in-line probing with guanidine. Ligand binding is evident only at relatively high guanidine concentrations when the RNA concentration in the in-line probing reaction is ~100 nM (Figure 3-7). However, increasing the RNA concentration in the in-line probing reaction to 2 μM improves the affinity for guanidine by ~5 fold, which is consistent with a model wherein two hairpin substructures need to assemble to achieve high-affinity ligand binding. The natural tandem arrangement of the stem loops in the guanidine-II RNA greatly increases the effective local concentration of the two hairpin substructures, and thereby allows the riboswitch to achieve the desired affinity and cooperativity characteristics for guanidine binding.

Guanidine-II binds its ligand selectively albeit with less stringency than guanidine-I. A remarkable feature of the guanidine-I riboswitch class is its ability to strongly discriminate against a diversity of close analogs of guanidine, and generate differences in affinity that are typically greater than 100 fold.[5] Based on the atomic-resolution x-ray structural model for a representative guanidine-I RNA,[13] this high level of selectivity is accomplished by utilizing complex tertiary structures, cation-π interactions, and a putative Mg^{2+} ion binding site to prevent recognition of any compound with modifications to the guanidine molecule. The guanidine-II riboswitch aptamers described in this chapter employ less than half the number of conserved nucleotides compared to guanidine-I riboswitches, yet likely form double the number of binding pockets. Therefore, I

anticipated that the binding pockets of guanidine-II riboswitches would not be as proficient at discriminating against certain guanidine analogs.

Sarah Malkowski investigated ligand binding specificity of the *G. violaceus* 60 *emrE* RNA by conducting in-line probing assays using a variety of guanidine analogs (Figure 3-8). Guanidine analogs that carry small additions to the guanidine moiety core, such as the addition of a single methyl or amino group, induce structural modulation in a manner that is similar that observed for guanidine. However, the apparent affinity for these analogs is modestly poorer than for the parent compound. Specifically, methyl- and aminoguanidine bind the RNA nearly as well as guanidine, and exhibit dissociation constants of 80 μM and 200 μM, respectively (Figure 3-9). However, analogs with larger substituents exhibit far poorer binding affinity, suggesting that guanidine-II riboswitches might use steric effects to block larger natural compounds that carry a guanidyl moiety.

It is unlikely that these guanidine-dependent structural changes are merely due to a general RNA-cation interaction, since compounds such as spermidine have no significant effect. Furthermore, it was previously shown that guanidine does not induce structural modulation in RNAs from other riboswitch classes.[5] Therefore it is unlikely that guanidine is non-specifically altering the structure of guanidine-II aptamers.

Perhaps the most important compound for a guanidine riboswitch to discriminate against is urea, which has numerous roles in metabolism. Urea differs from guanidine in several ways that should make it possible for an RNA to strongly favor guanidine binding and exclude urea from the binding pocket. Importantly, urea carries a keto oxygen atom in place of an imine nitrogen atom. This difference converts a hydrogen bond donor location to a hydrogen bond acceptor. In addition, urea is neutral, whereas the imine nitrogen of the

protonated form of guanidine (the predominant form under biological pH conditions) is positively charged. This makes possible several mechanisms by which a riboswitch can selectively bind guanidine even in the presence of high concentrations of urea. Sarah Malkoswki observed no evidence of RNA structure modulation even when 10 mM urea is present, indicating that this highly similar molecule is actively discriminated against by the binding pocket.

Concluding remarks. The individual aptamer domains of guanidine-II riboswitches are the simplest of all known riboswitch aptamers both in terms of structural complexity and number of conserved nucleotides. The minimal aptamer length of a functional guanidine-II aptamer substructure is no longer than the 30 *emrE* construct (Figure 3-7), and can probably be less than 20 nucleotides. For comparison, the next-simplest natural riboswitch aptamer is ~34 nucleotides and belongs to the preQ$_1$-I class.[14] However, preQ$_1$-I aptamers exploit a pseudoknot architecture that is more complex that a simple hairpin,[15,16] and can function without another aptamer as a partner.

Although the single aptamer construct present in the 30 *emrE* RNA does respond to guanidine, likely when forming a homodimer, this bimolecular complex must have an affinity that is far poorer than the natural tandem aptamers. At the time of its discovery,[3] it was proposed that single hairpins similar to those in mini-*ykkC* motif RNAs might exist, but that it would be challenging to sort false positives from functional examples given that the motif is so simple. Now that it is apparent that the tandem hairpin arrangement of mini-*ykkC* motif RNAs cooperatively binds multiple guanidine molecules, it seems possible that sequence or structural adaptations have occurred to permit variants of this riboswitch class to sense a single guanidine molecule.

A simple adaptation might be for a hairpin closed by an ACGR loop to reside near another hairpin that lacks this conserved loop sequence. This arrangement might provide the accessory contacts that permit tight binding of guanidine by the conserved hairpin, in a manner analogous to the singlet versions of glycine riboswitches.[17] Intriguingly, the *ykkC*-III motif that forms the aptamer of guanidine-III riboswitches might represent a distinctly-structured version of this proposed arrangement because it also contains an 'ACGR' loop, and appears to form only a single guanidine binding site.[6] The proposed pseudoknot structure of guanidine-III riboswitches might use this more complex structure to permit the ACGR sequence to contribute to guanidine binding without forming a tandem arrangement.

The validation of mini-*ykkC* motif RNAs as members of a novel guanidine-binding riboswitch class called guanidine-II further highlights the importance of guanidine metabolism in many bacterial species. Also, this small yet surprisingly sophisticated riboswitch class provides evidence that riboswitches and other noncoding RNAs can be formed from unusually short and simple structures formed by very few conserved nucleotides. Thus, some small functional RNA domains might easily escape detection by existing bioinformatics approaches that rely on the conservation of extensive sequence and structural domains.

MATERIALS AND METHODS

Chemicals and oligonucleotides. All chemicals and chemically synthesized oligonucleotides were purchased from Sigma-Aldrich with the exception of guanidinosuccinic acid (Alfa Chemistry). [γ-^{32}P]-ATP was purchased from PerkinElmer.

Enzymes were purchased from New England BioLabs, unless otherwise noted. A complete list of DNA and RNA oligonucleotides used in this chapter can be found in Table 3-1.

Bioinformatics analyses. Additional examples of mini-*ykkC* motif RNAs were identified using Infernal 1.1[7] to search RefSeq version 63 plus additional environmental microbial databases as described previously.[8] Iterative searches for new sequences were performed based on the previously published alignment of the mini-*ykkC* motif,[3] revealing over 800 unique examples. The consensus sequence and secondary structure model was constructed using R2R software.[9]

RNA oligonucleotide preparation. Synthetic DNA oligonucleotides (Table 3-1) containing a 5′-terminal T7 RNA polymerase (T7 RNAP) promoter followed by a template region coding for the aptamer of interest were combined in equimolar amount with their complementary strand to form a double-stranded DNA (dsDNA) template. The desired RNA constructs were prepared from these dsDNAs by in vitro transcription for four hours at 37 ºC in a 25 or 50 µL reaction volume containing laboratory-prepared T7 RNAP (2 U μL^{-1}), 80 mM HEPES-KOH (pH 7.5 at 23 ºC), 24 mM $MgCl_2$, 2 mM spermidine, 40 mM DTT, and 2 mM of each NTP.

The desired RNA products were purified via denaturing (8 M urea) 10% PAGE (National Diagnostics). For the truncated 30 *emrE* RNA construct, the synthetic RNA was chemically synthesized and purified via denaturing PAGE. The gel slice containing the appropriately-sized band was excised and the RNA extracted via the crush-soak method for 30 min at 23 ºC in 350 µL containing 200 mM NaCl, 10 mM Tris-HCl (pH 7.5 at 23 ºC) and 1 mM EDTA. RNA in the resulting supernatant was precipitated by the addition of 20 µL 3 M sodium acetate (pH 5.5 at 23 ºC) and 750 µL cold 100% ethanol, and

incubated for up to 30 min at -20 ºC. The RNA was subsequently pelleted via centrifugation, the supernatant removed, and the remaining pellet was dried and resuspended in 20 μL of deionized water (dH_2O). To generate 5′ ^{32}P-labeled RNAs, the 5′-terminal phosphate was removed via rAPid alkaline phosphatase (Roche Life Sciences) according to manufacturer instructions. Subsequently, 20 pmol of dephosphorylated RNA was radiolabeled using T4 polynucleotide kinase in a 20 μL reaction containing 25 mM CHES (pH 9.0 at 23ºC), 5 mM $MgCl_2$, 3 mM DTT, and 20 μCi [γ-^{32}P]-ATP and incubated for one hour. The resulting radiolabeled RNAs were purified before use by denaturing 10% PAGE as described above.

RNA in-line probing analyses. In-line probing assays were performed largely as previously described.[10,11] Briefly, 5′-labeled RNAs were incubated at 23ºC either with or without candidate ligands for 42 to 48 hours in 10 μL reactions containing 20 mM $MgCl_2$, 100 mM KCl, 50 mM Tris-HCl (pH 8.3 at 23ºC). The products of spontaneous cleavage were subsequently analyzed via denaturing 10% PAGE and visualized using a phosphorimager (GE Healthcare Life Sciences).

Fraction bound values were calculated by varying the ligand concentration in separate reactions and quantifying the changes in band intensity at nucleotides that exhibit structural modulation. Band intensities were corrected for loading differences based on the analysis of a band that does not alter its intensity upon ligand addition. The resulting band intensity values for specific regions were normalized to a fractional value between 0 and 1, the values for regions with band intensities that decrease with increasing ligand concentration were inverted, and all resulting values were plotted as 'fraction bound' using GraphPad Prism 7. The K_D and Hill coefficients were determined by plotting the average of the

fraction bound at the indicated sites of modulation as a function of the logarithm of ligand concentration and using a sigmoidal four parameter logistic fit.

ACKNOWLEDGEMENTS

My colleagues and I thank Adam Roth, Shira Stav, Zasha Weinberg, and other members of the Breaker laboratory as well as Caroline Reiss of Yale University and Randy Stockbridge of the University of Michigan for helpful discussions. M.E.S. was supported by an NIH Cellular and Molecular Biology Training Grant (T32GM007223). S.N.M. was supported by the NSF Graduate Research Fellowship Program (DGE1122492). This work was also supported by NIH grants (GM022778, DE022340) as well as funds from Howard Hughes Medical Institute to R.R.B.

REFERENCES

(1) Barrick, J.E., Corbino, K.A., Winkler, W.C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I. (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc Natl Acad Sci USA 101,* 6421-6426.

(2) Meyer, M.M., Hammond, M.C., Salinas, Y., Roth, A., Sudarsan, N., Breaker, R.R. (2011) Challenges of ligand identification for riboswitch candidates. *RNA Biol 8,* 5-10.

(3) Weinberg, Z., Barrick, J.E., Yao, Z., Roth, A., Kim, J.N., Gore, J., et al. (2007) Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res 35,* 4809–4819.

(4) Weinberg, Z., Wang, J.X., Bogue, J., Yang, J., Corbino, K., Moy, R.H., Breaker, R.R. (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol 11,* 31.

(5) Nelson, J.W., Atilho, R.M., Sherlock, M.E., Stockbridge, R.B., Breaker, R.R. (2017) Metabolism of free guanidine in bacteria is regulated by a widespread riboswitch class. *Mol Cell 65,* 1-11.

(6) Sherlock, M.E., Breaker, R.R. (2017) Biochemical validation of a third guanidine riboswitch class in bacteria. *Biochemistry 56,* 359–363.

(7) Nawrocki, E.P., Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics 29,* 2933-2935.

(8) Weinberg, Z., Kim, P.B., Chen, T.H., Li, S., Harris, K.A., Lünse, C.E., Breaker, R.R. (2015) New classes of self-cleaving ribozymes revealed by comparative genomics analysis. *Nat Chem Biol 11,* 606-610.

(9) Weinberg, Z., Breaker, R.R. (2011) R2R-software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics 12,* 1.

(10) Regulski, E.E., Breaker, R.R. (2008) In-line probing analysis of riboswitches. *Methods Mol Biol 419,* 53-67.

(11) Soukup, G.A., Breaker, R.R. (1999) Relationship between internucleotide linkage geometry and the stability of RNA. *RNA 5,* 1308-1325.

(12) Schuldiner, S.S. (2009) EmrE, a model for studying evolution and mechanism of ion-coupled transporters. *Biochim Biophys Acta 1794,* 748-762.

(13) Reiss, C.W., Xiong, Y., Strobel, S.A. (2017) Structural basis for ligand binding to the guanidine-I riboswitch. *Structure 25*, 195-202.

(14) Roth, A., Winkler, W.C., Regulski, E.E., Lee, B.W., Lim, J., Jona, I., Barrick, J.E., Ritwik, A., Kim, J.N., Welz, R., et al. (2007) A riboswitch selective for the queuosine precursor preQ$_1$ contains an unusually small aptamer domain. *Nat Struct Mol Biol 14,* 308-317.

(15) Klein, D.J., Edwards, T.E., Ferre-D'Amare, A.R. (2009) Cocrystal structure of a class I preQ$_1$ riboswitch reveals a pseudoknot recognizing an essential hypermodified nucleobase. *Nat Struct Mol Biol 16,* 343–344.

(16) Spitale, R.C., Torelli, A.T., Krucinska, J., Bandarian, V., Wedekind, J.E. (2009) The structural basis for recognition of the PreQ$_0$ metabolite by an unusually small riboswitch aptamer domain. *J Biol Chem 284,* 11012–11016.

(17) Ruff, K.M., Muhammad, A., McCown, P.J., Breaker, R.R., Strobel, S.A. (2016) Singlet glycine riboswitches bind ligand as well as tandem riboswitches. *RNA 22,* 1728-1738.

FIGURES

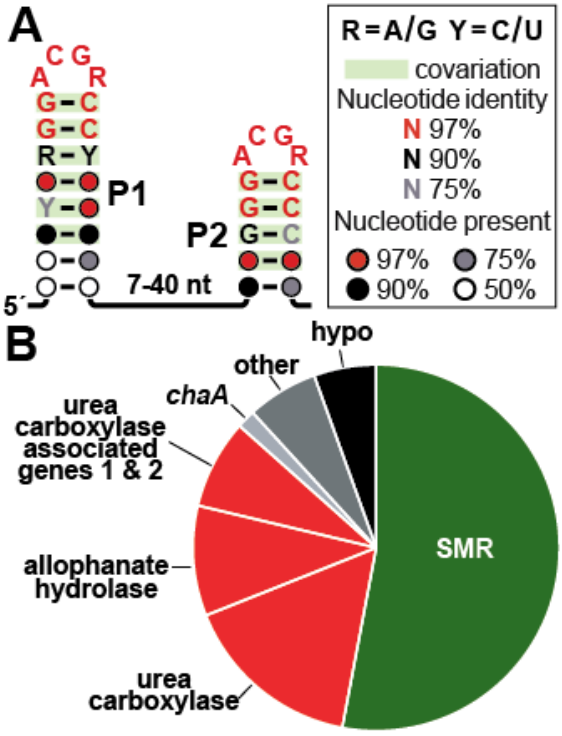

Figure 3-1. Structure and function of guanidine-II riboswitches. (A) The consensus sequence and structural model for the ~850 unique examples of mini-*ykkC* motif (guanidine-II) RNAs. Covarying base-pairs (indicated by green shading) are found in the P1 and P2 stems. Each stem has an extremely highly conserved 'ACGR' loop. (B) Genes whose expression is controlled by guanidine-II riboswitches, including those encoding predicted small multidrug resistance (SMR) transporters (green), urea (guanidine) carboxylases and their associated proteins (red), and other (gray) or hypothetical (hypo, black) genes. I performed bioinformatic analyses used to create this figure.

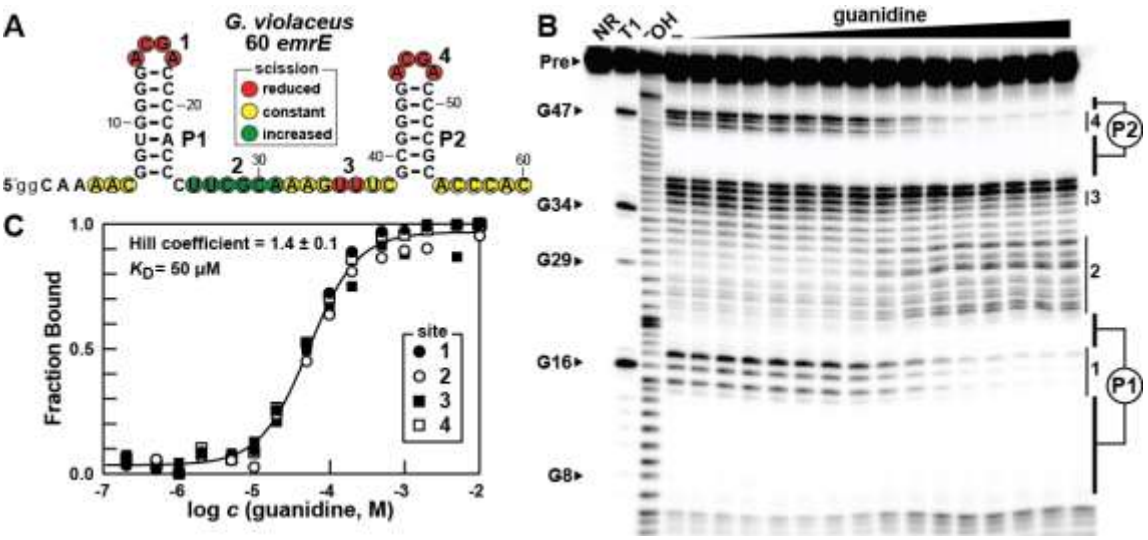

Figure 3-2. Guanidine is bound by a guanidine-II riboswitch aptamer *in vitro*. (A) The sequence and observed secondary structure of the 60 *emrE* RNA from *G. violaceus*. Lowercase nucleotides on the 5′ terminus designate G residues added to enable transcription by T7 RNA polymerase. Red, yellow, and green circles indicate nucleotides that undergo less, constant, or more spontaneous cleavage, respectively, upon addition of the ligand as determined by their relative band intensities in B. Four regions of guanidine-mediated structural modulation as identified in B are numbered accordingly. (B) Polyacrylamide gel electrophoresis (PAGE) analysis of in-line probing reactions of 60 *emrE G. violaceus* RNA in the absence (–) and presence of guanidine ranging from 200 nM to 10 mM. NR, T1, and $^{-}$OH indicate no reaction, RNA partially digested with T1 ribonuclease (cleaves after each G), and RNA partially digested under alkaline conditions (cleaves after every nucleotide). Precursor RNA (Pre) as well as certain G residues within the RNA sequence are annotated, and regions of modulation (vertical bars) are numbered as designated in A. Regions that are devoid of spontaneous cleavage and that correspond to base-paired substructures P1 and P2 stems are identified. (C) Plot of the fraction of RNA bound to ligand versus the logarithm of the guanidine concentration as determined by

quantification of the PAGE data in B (see Materials and Methods for details). The K_D and Hill coefficient values were determined by the sigmoidal curve fit, and the reported error represents the standard deviation reported in the fitting parameters. I performed the experiments shown in this figure.

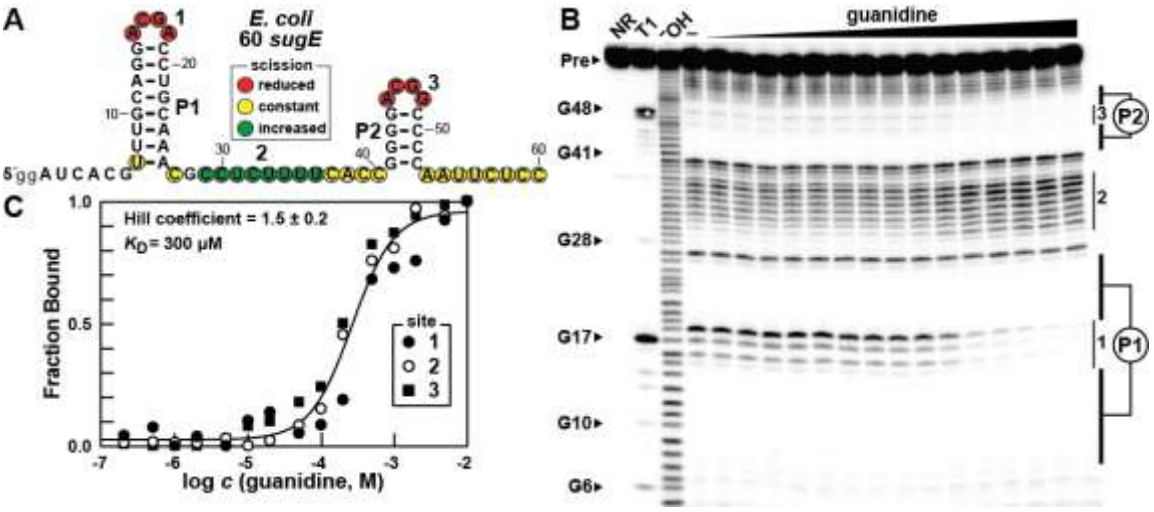

Figure 3-3. Guanidine binding by the 60 *sugE* RNA from *E. coli*. (A) The sequence and observed secondary structure of the 60 *sugE E. coli* RNA construct. Additional annotations and labels are as described in the legend to Figure 3-2A. (B) PAGE analysis of in-line probing reactions of 60 *sugE* RNA in the absence (–) and presence of guanidine ranging from 200 nM to 10 mM. Additional annotations are as described in the legend to Figure 3-2B. (C) Plot of the fraction of RNA bound to ligand versus the logarithm of the guanidine concentration as determined by quantification of the PAGE data in B, as described in the legend to Figure 3-2C. The K_D and Hill coefficient were determined by the sigmoidal fit and the reported error represents the standard deviation reported in the fitting parameters. I performed the experiments shown in this figure.

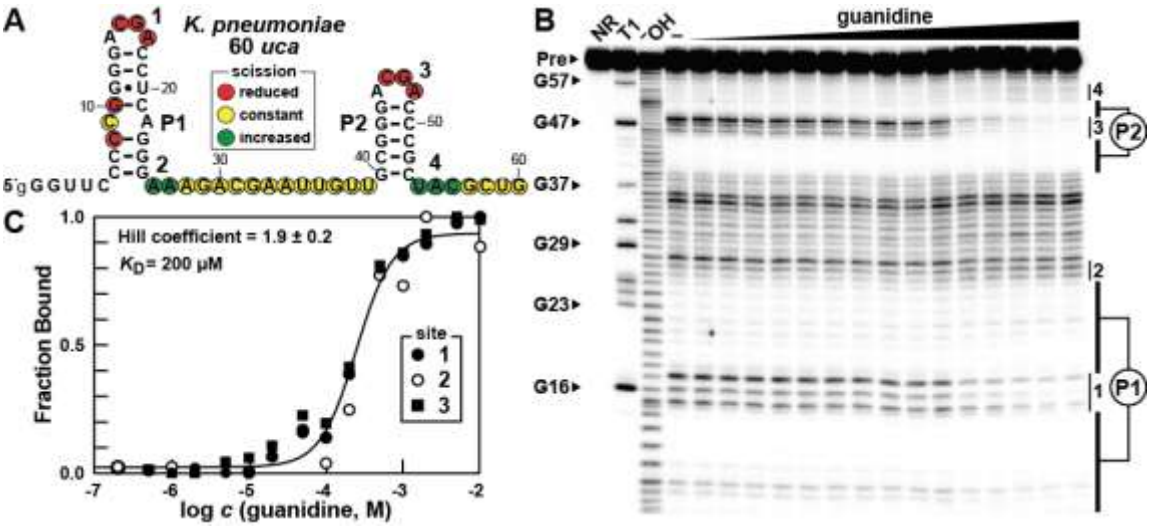

Figure 3-4. Guanidine binding by the 60 *uca* RNA from *K. pneumoniae*. (A) The sequence and observed secondary structure of the 60 *uca K. pneumoniae* RNA. Additional annotations and labels are as described in the legend to Figure 3-2A. (B) PAGE analysis of in-line probing reactions of 60 *uca* RNA in the absence (–) and presence of guanidine ranging from 200 nM to 10 mM. Additional annotations are as described in the legend to Figure 3-2B. (C) Plot of the fraction of RNA bound to ligand versus the logarithm of the guanidine concentration as determined by quantification of the PAGE data in B, as described in the legend to Figure 3-2C. The K_D and Hill coefficient were determined by the sigmoidal fit and the reported error represents the standard deviation reported in the fitting parameters. I performed the experiments shown in this figure.

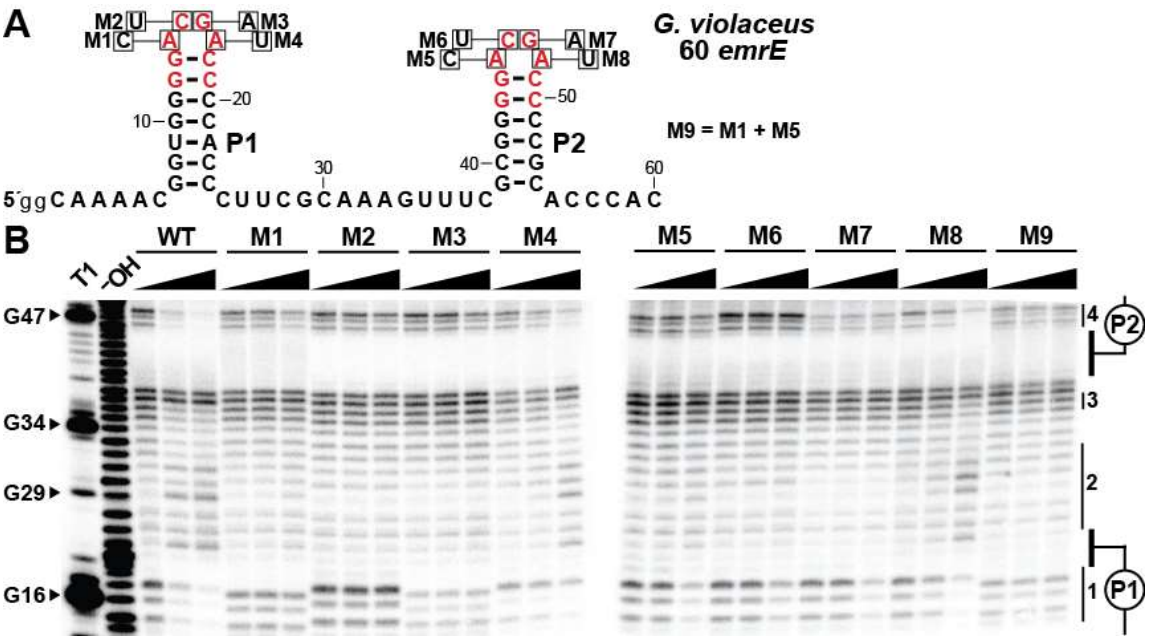

Figure 3-5. Mutations of conserved loop nucleotides affect structural modulation of both loops. (A) Sequence and secondary structure of the 60 *emrE* RNA from *G. violaceus*. Red letters denote highly conserved (>97%) nucleotide positions of the consensus sequence. Single nucleotide positions altered to create each RNA are identified with boxes, annotated with the mutant nucleotide identity, and labeled M1 through M8. M9 contains mutations at both positions specified by M1 and M5. (B) PAGE analysis of in-line probing assays with WT and mutated sequences of the 60 *emrE* RNA in the presence of 0, 0.1, and 1 mM guanidine. The cropped gel image depicts the regions spanning from loop 1 (nucleotides 14-17) to loop 2 (nucleotides 45-48). Additional annotations are as described in the legend to Figure 3-2B. Sarah Malkowski and I performed the experiments shown in this figure.

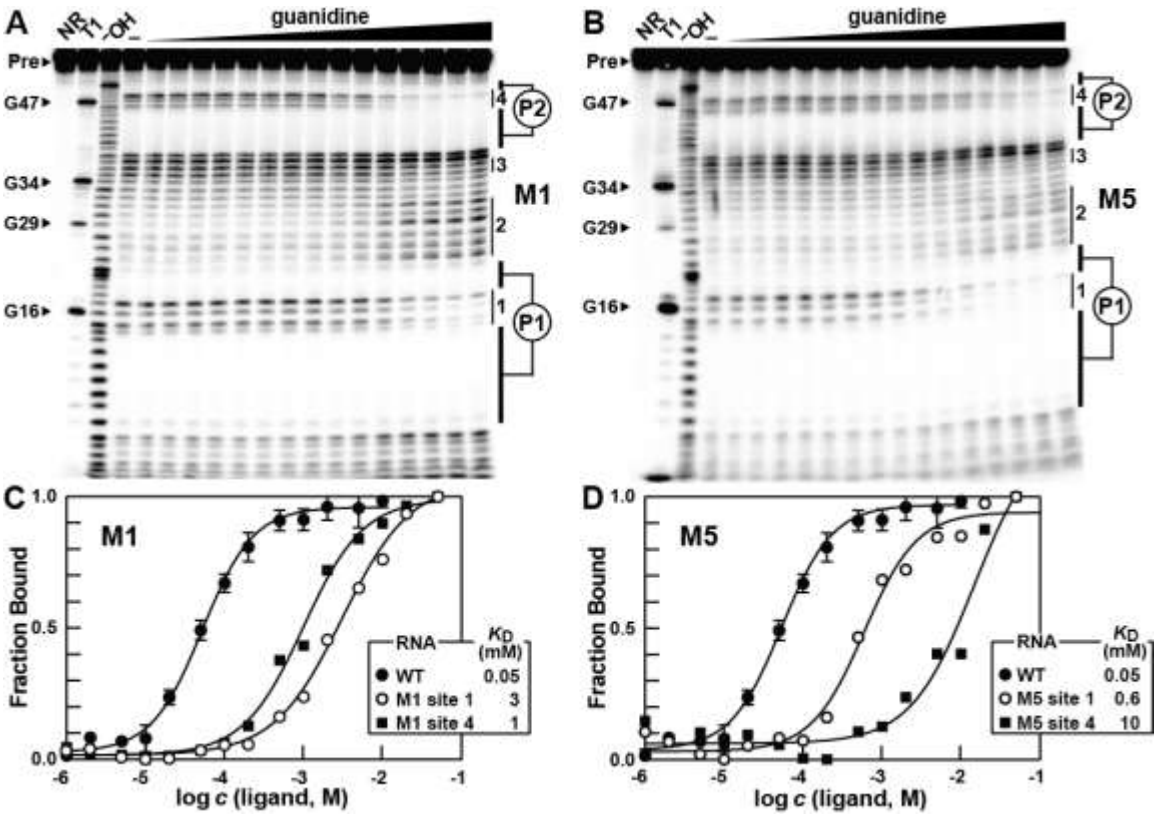

Figure 3-6. Mutations in each stem loop affect guanidine binding independently at each binding site. (A,B) PAGE analysis of in-line probing reactions of the M1 (A) or M5 (B) RNA in the presence of guanidine ranging from 1 μM to 50 mM. Additional annotations are as described in the legend to Figure 3-2B and the identity of M1 and M5 are as describe in Figure 3-6A. (C,D) Plot of the fraction of WT and M1 (C), or M5 (D) RNA bound to ligand by analysis of sites 1 and 4 versus the logarithm of the guanidine concentration. The data was obtained by quantification of the PAGE data in Figure 3-2B, A or B, respectively, as described in the legend to Figure 3-2C. K_D values were determined by curve fitting, wherein the sigmoidal curves are consistent with 1:1 binding. Sarah Malkowski performed the experiments shown in this figure.

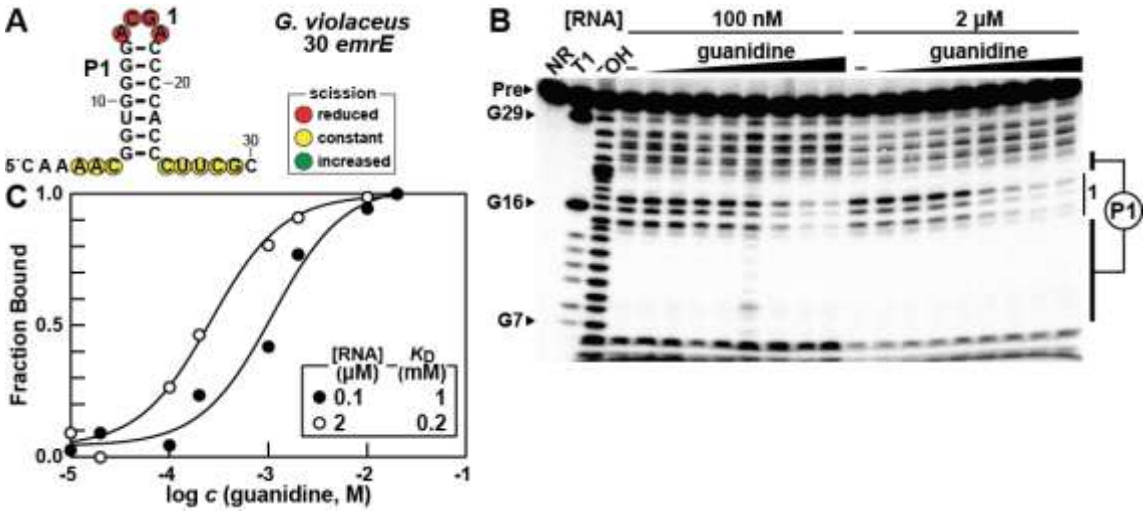

Figure 3-7. Two singlet stem loops interact to accomplish guanidine binding. (A) The sequence and observed secondary structure of the truncated 30 *emrE* RNA from *G. violaceus*, which lacks nucleotides that otherwise would form the P2 stem. Additional annotations are as described in Figure 3-2A. (B) PAGE analysis of in-line probing reactions of 30 *emrE* RNA in the absence (–) and presence of guanidine ranging from 10 µM to 20 mM at ~100 nM or 2 µM RNA concentration. (C) Plot of the fraction of RNA bound to ligand at site 1 versus the logarithm of the guanidine concentration as determined by quantification of the PAGE data in B, as described in the legend to Figure 3-2C. K_D values were determined by curve fitting, wherein the sigmoidal curves are consistent with 1:1 binding. Sarah Malkowski performed the experiments shown in this figure.

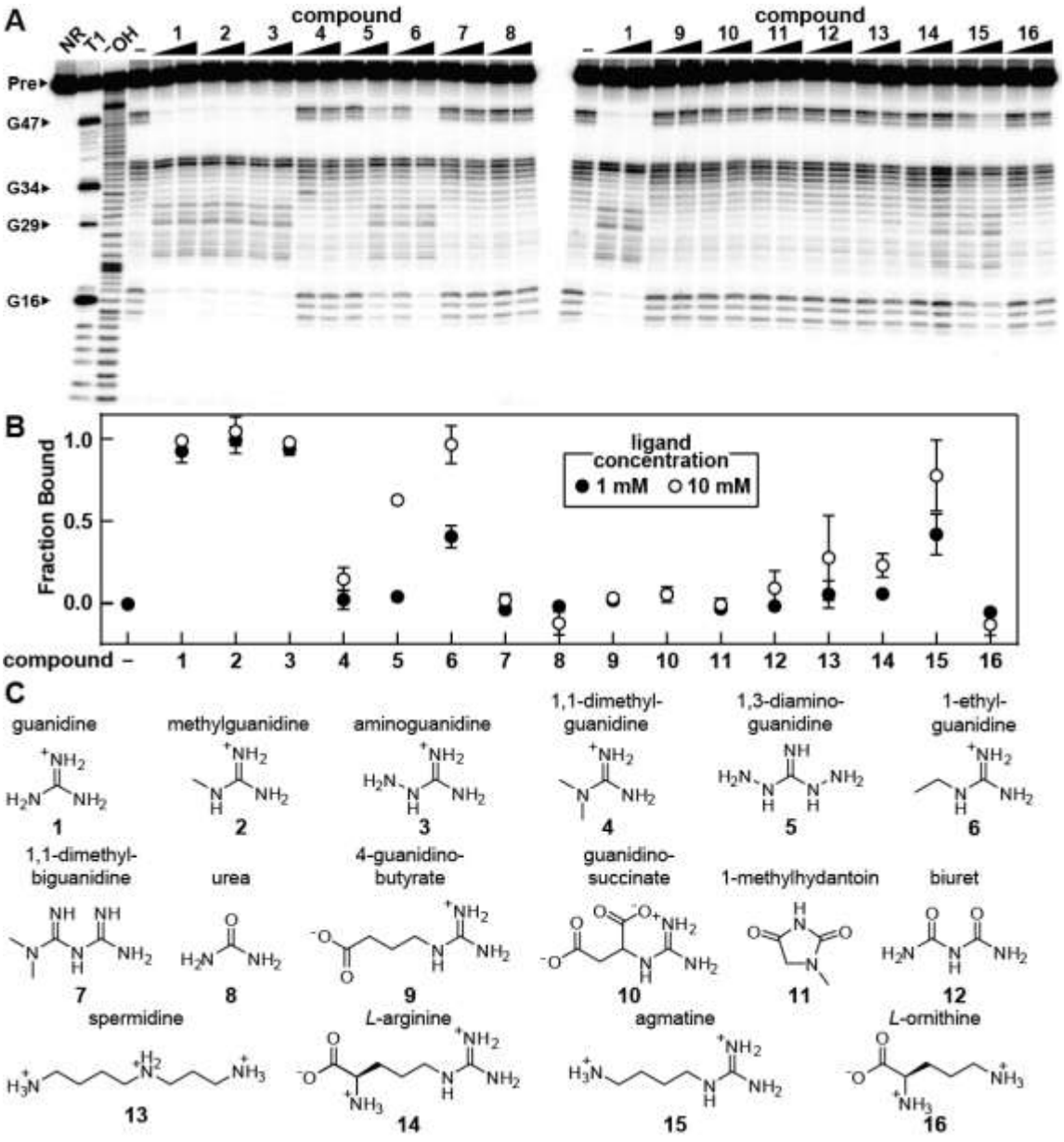

Figure 3-8. A guanidine-II RNA only recognizes chemical analogs with similar structures to its ligand. (A) PAGE analysis of in-line probing reactions of 60 *emrE* RNA from *G. violaceus* in the absence (–) or presence of guanidine or its structural analogs at 1 and 10 mM concentrations. Additional annotations are as described in the legend to Figure 3-2B. (B) Plot of the fraction of RNA bound to ligand for each compound at 1 and 10 mM concentrations as determined by quantification of the PAGE data in B, as described in the

legend to Figure 3-2C. For compounds that are not saturating, fraction bound values are derived by comparison to the data for guanidine. Each point represents the average and standard deviation of the fraction bound at three sites of modulation (regions 1-3, as indicated in Figure 3-2A). (C) Chemical structures of each analog in their predominant state at pH 7. Sarah Malkowski performed the experiments shown in this figure.

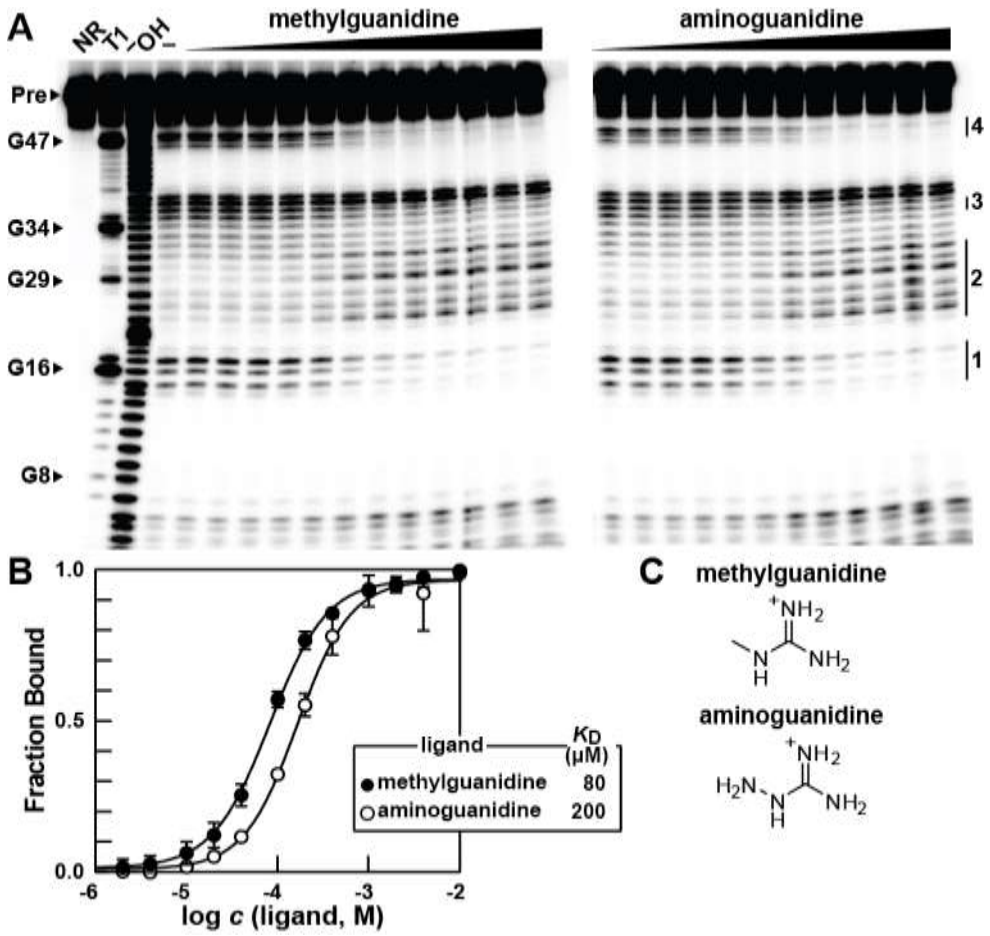

Figure 3-9. A guanidine-II riboswitch aptamer can recognize close structural analogs of guanidine. (A) PAGE analysis of in-line probing reactions of 60 *emrE* RNA from *G. violaceus* in the absence (–) or presence of methylguanidine or aminoguanidine ranging from 2 μM to 10 mM. Additional annotations are as described in the legend to Figure 3-2B. (B) Plot of the fraction of RNA bound to ligand versus the logarithm of the methyl- or aminoguanidine concentration as determined by quantification of the PAGE data in B, as described in the legend to Figure 2C. Each point represents the average and standard deviation of the fraction bound at four sites of modulation (regions 1-4, as indicated in Figure 3-2A). The K_D values were determined by curve fitting, wherein the sigmoidal

curves are consistent with 1:1 binding. (C) Chemical structures of methylguanidine and aminoguanidine. Sarah Malkowski performed the experiments shown in this figure.

TABLES

Table 3-1. Sequences of synthetic DNAs and RNA used in Chapter 3.

Name	Sequence	Annotation
MES103	TAATACGACTCACTATAGGATCAGCTTTGCAGGACGACCTGCAAACGCCTCTTTTCACCGGGGACGGCCCCAATTCTCC	Forward template for transcription of the WT *E. coli* 60 *sugE* RNA
MES104	GGAGAATTGGGGCCGTCCCCGGTGAAAAGAGGCGTTTGCAGGTCGTCCTGCAAAGCTGATCCTATAGTGAGTCGTATTA	Reverse template for transcription of the WT *E. coli* 60 *sugE* RNA
MES105	TAATACGACTCACTATAGGGTTCCCCCGGGGACGACCTCAGGGAAAGACGAATTGTTGCGGGACGACCCGCTACGCTG	Forward template for transcription of the WT *K. pneumoniae* 60 *uca* RNA
MES106	CAGCGTAGCGGGTCGTCCCGCAACAATTCGTCTTTCCCTGAGGTCGTCCCCGGGGGAACCCTATAGTGAGTCGTATTA	Reverse template for transcription of the WT *K. pneumoniae* 60 *uca* RNA
MES107	TAATACGACTCACTATAGGCAAAACGGTGGGGACGACCCCACCCTTCGCAAAGTTTCGCGGGGACGACCCCGCACCCAC	Forward template for transcription of the WT *G. violaceus* 60 *emrE* RNA

MES108	GTGGGTGCGGGGTCGTCCCCGCG AAACTTTGCGAAGGGTGGGGTCG TCCCCACCGTTTTGCCTATAGTGA GTCGTATTA	Reverse template for transcription of the WT *G. violaceus* 60 *emrE* RNA
MES109	CAAAACGGUGGGGACGACCCCAC CCUUCGC	Synthetic RNA of the truncated *G. violaceus* 30 *emrE*
MES110	TAATACGACTCACTATAGGCAAA ACGGTGGGGCCGACCCCACCCTT CGCAAAGTTTCGCGGGGACGACC CCGCACCCAC	Forward template for transcription of the M1 *G. violaceus* 60 *emrE* RNA
MES111	GTGGGTGCGGGGTCGTCCCCGCG AAACTTTGCGAAGGGTGGGGTCG GCCCCACCGTTTTGCCTATAGTGA GTCGTATTA	Reverse template for transcription of the M1 *G. violaceus* 60 *emrE* RNA
MES112	TAATACGACTCACTATAGGCAAA ACGGTGGGGATGACCCCACCCTT CGCAAAGTTTCGCGGGGACGACC CCGCACCCAC	Forward template for transcription of the M2 *G. violaceus* 60 *emrE* RNA
MES113	GTGGGTGCGGGGTCGTCCCCGCG AAACTTTGCGAAGGGTGGGGTCA TCCCCACCGTTTTGCCTATAGTGA GTCGTATTA	Reverse template for transcription of the M2 *G. violaceus* 60 *emrE* RNA

MES114	TAATACGACTCACTATAGGCAAA ACGGTGGGGACAACCCCACCCTT CGCAAAGTTTCGCGGGGACGACC CCGCACCCAC	Forward template for transcription of the M3 *G. violaceus* 60 *emrE* RNA
MES115	GTGGGTGCGGGGTCGTCCCCGCG AAACTTTGCGAAGGGTGGGGTTG TCCCCACCGTTTTGCCTATAGTGA GTCGTATTA	Reverse template for transcription of the M3 *G. violaceus* 60 *emrE* RNA
MES116	TAATACGACTCACTATAGGCAAA ACGGTGGGGACGTCCCCACCCTT CGCAAAGTTTCGCGGGGACGACC CCGCACCCAC	Forward template for transcription of the M4 *G. violaceus* 60 *emrE* RNA
MES117	GTGGGTGCGGGGTCGTCCCCGCG AAACTTTGCGAAGGGTGGGGACG TCCCCACCGTTTTGCCTATAGTGA GTCGTATTA	Reverse template for transcription of the M4 *G. violaceus* 60 *emrE* RNA
MES118	TAATACGACTCACTATAGGCAAA ACGGTGGGGACGACCCCACCCTT CGCAAAGTTTCGCGGGGCCGACC CCGCACCCAC	Forward template for transcription of the M5 *G. violaceus* 60 *emrE* RNA
MES119	GTGGGTGCGGGGTCGGCCCCGCG AAACTTTGCGAAGGGTGGGGTCG	Reverse template for transcription of the M5 *G. violaceus* 60 *emrE* RNA

	TCCCCACCGTTTTGCCTATAGTGA GTCGTATTA	
MES120	TAATACGACTCACTATAGGCAAA ACGGTGGGGACGACCCCACCCTT CGCAAAGTTTCGCGGGGATGACC CCGCACCCAC	Forward template for transcription of the M6 *G. violaceus* 60 *emrE* RNA
MES121	GTGGGTGCGGGGTCATCCCCGCG AAACTTTGCGAAGGGTGGGGTCG TCCCCACCGTTTTGCCTATAGTGA GTCGTATTA	Reverse template for transcription of the M6 *G. violaceus* 60 *emrE* RNA
MES122	TAATACGACTCACTATAGGCAAA ACGGTGGGGACGACCCCACCCTT CGCAAAGTTTCGCGGGGACAACC CCGCACCCAC	Forward template for transcription of the M7 *G. violaceus* 60 *emrE* RNA
MES123	GTGGGTGCGGGGTTGTCCCCGCG AAACTTTGCGAAGGGTGGGGTCG TCCCCACCGTTTTGCCTATAGTGA GTCGTATTA	Reverse template for transcription of the M7 *G. violaceus* 60 *emrE* RNA
MES124	TAATACGACTCACTATAGGCAAA ACGGTGGGGACGACCCCACCCTT CGCAAAGTTTCGCGGGGACGTCC CCGCACCCAC	Forward template for transcription of the M8 *G. violaceus* 60 *emrE* RNA

MES125	GTGGGTGCGGGGACGTCCCCGCG AAACTTTGCGAAGGGTGGGGTCG TCCCCACCGTTTTGCCTATAGTGA GTCGTATTA	Reverse template for transcription of the M8 *G. violaceus* 60 *emrE* RNA
MES126	TAATACGACTCACTATAGGCAAA ACGGTGGGGCCGACCCCACCCTT CGCAAAGTTTCGCGGGGCCGACC CCGCACCCAC	Forward template for transcription of the M9 *G. violaceus* 60 *emrE* RNA
MES127	GTGGGTGCGGGGTCGGCCCCGCG AAACTTTGCGAAGGGTGGGGTCG GCCCCACCGTTTTGCCTATAGTGA GTCGTATTA	Reverse template for transcription of the M9 *G. violaceus* 60 *emrE* RNA

CHAPTER FOUR

Biochemical validation of a third guanidine riboswitch class in bacteria

Largely adapted from the following publication:

Sherlock, M.E., Breaker, R.R. (2017) Biochemical validation of a third guanidine riboswitch class in bacteria. *Biochemistry 56*, 359–363.

ABSTRACT

In Chapters 2 and 3, my colleagues and I determined that representatives of the riboswitch candidates called *ykkC* and mini-*ykkC* directly bind free guanidine. These riboswitches regulate the expression of genes whose protein products are implicated in overcoming the toxic effects of this ligand. Thus, the relevant *ykkC* motif and mini-*ykkC* motif RNAs have been classified as guanidine-I and guanidine-II riboswitch RNAs, respectively. Moreover, it had previously been noted that a third candidate riboswitch class, called *ykkC*-III, was associated with a distribution of genes similar to those of the other two motifs. Therefore, it was predicted that *ykkC*-III motif RNAs would sense and respond to the same ligand. In this chapter, I present biochemical data supporting the hypothesis that *ykkC*-III RNAs represent a third class of guanidine-sensing RNAs called guanidine-III riboswitches. Members of the guanidine-III riboswitch class bind ligand with an affinity similar to that observed for members of the other two classes. Notably, there are some sequence similarities between guanidine-II and guanidine-III riboswitches. However, the characteristics of ligand discrimination by guanidine-III RNAs are different from those of the other guanidine-binding motifs, suggesting that the binding pockets have distinct features among the three riboswitch classes.

INTRODUCTION

Orphan riboswitch candidates are well-conserved RNA motifs that have been proposed to be riboswitches, but whose ligands remain to be identified.[1,2] The experimental validation of several orphan riboswitches recently has led to the discovery of novel riboswitch ligands, and has revealed new and unexpected areas of bacterial metabolism,[3] cell signaling,[4] metal ion homeostasis,[5,6] and toxicity response mechanisms.[7] Given the potential for novel biological insight to be gained, we became particularly interested in experimentally validating riboswitch candidates whose ligands have long remained undiscovered.

Recently, we determined that certain members of the *ykkC* orphan riboswitch class[2,8] selectively recognize free guanidine and regulate various genes whose protein products overcome the toxic effects of this compound.[9] As the number of distinct riboswitch classes has grown, it has become helpful to name riboswitch classes for the ligand they sense. Therefore, members of this newly-validated RNA class have been renamed 'guanidine-I' riboswitches. Moreover, we had previously speculated that two additional RNA motifs, initially called mini-*ykkC*[10] and *ykkC*-III,[11] might also function as riboswitches that respond to the same ligand as guanidine-I RNAs.[2,11] Indeed, in Chapter 3 I provide evidence that that mini-*ykkC* motif RNAs also function as guanidine-sensing RNAs, hereafter called 'guanidine-II' riboswitches.[12]

Although *ykkC*-III RNAs share no sequence or structural characteristics with the ligand-binding aptamer domain of guanidine-I riboswitches, *ykkC*-III RNAs are commonly associated with the same types of genes. For example, the genes commonly associated with guanidine-I riboswitches and *ykkC*-III RNAs encode 'small multidrug resistance' (SMR)

transporters, including those annotated as EmrE[13] and SugE,[14] that have been implicated as guanidine exporters.[9] Furthermore, there are some similarities in the conserved sequences of *ykkC*-III motif RNAs and guanidine-II riboswitches. Specifically, many guanidine-II riboswitches are formed by two near-identical hairpins with conserved 'ACGR' loops. This same sequence also occurs once in the most highly-conserved portion of *ykkC*-III motif RNAs, although the sequence is not present in a hairpin loop.

Given the recent advances in guanidine riboswitch validation, and given the similarities noted above, I chose to further investigate the hypothesis that *ykkC*-III motif RNAs function as guanidine-sensing riboswitches. In this chapter, I present my findings that confirm *ykkC*-III motif RNAs selectively bind guanidine, which indicates that they function as a distinct riboswitch class called 'guanidine-III'. In addition, I compare and contrast the biochemical attributes of *g*uanidine-III riboswitches with representatives of guanidine-I and guanidine-II riboswitches.

RESULTS AND DISCUSSION

Updated consensus model and expanded gene associations for *ykkC*-III motif RNAs. I used Infernal, a homology search method,[15] to identify examples of *ykkC*-III motif RNAs within fully-sequenced genomes and environmental sequences made available since the initial report of its discovery.[11] This search revealed over 400 additional unique examples, with representatives mainly residing within Actinobacteria as previously reported. However, some representatives were also found in Proteobacteria, Deinococcus-Thermus, Planctomycetes, and Firmicutes.

I created a revised consensus sequence and structural model for the *ykkC*-III motif (Figure 4-1A), which is very similar to the model from the original report, by using the updated collection of representatives. As observed previously,[11] the variable length pairing element (called P0) at the 5′ terminus of the motif lacks sequence conservation, and I identified numerous examples that appear to lack this substructure. Therefore, this variable hairpin is unlikely to be critical for the function of these RNAs. The majority of the nucleotides initially observed to be present in 97% of the representatives remain highly conserved in the updated model. Specifically, 18 very highly conserved nucleotides reside in three separate segments of the RNA, and 10 of these conserved nucleotides are predicted to reside in bulged regions. Most notably, the highly conserved ACGA sequence that is similar to the ACGR hairpin loop sequence in guanidine-II riboswitches resides in an unpaired region that joins stems P1 and P2.

Similar to guanidine-II riboswitches, the conserved *ykkC*-III motif usually resides in the 5′-untranslated region (UTR) in close proximity to the ribosome binding site (RBS) and start codon of the adjacent open reading frame (ORF). Specifically, the purine-rich RBS is typically found just a few nucleotides downstream of the right shoulder of the P2 stem. Consequently, I predict that both guanidine-II riboswitches and the *ykkC*-III motif function as genetic 'ON' switches that control the translation of the ORF by allowing ribosomes to access the RBS in the presence of ligand.

Genes encoding annotated EmrE and SugE transporters are most commonly associated with *ykkC*-III motif RNAs (Figure 4-1B). These transporters resemble those controlled by both guanidine-I and guanidine-II riboswitches, and fall into the category of SugE-like heterodimers (Randy Stockbridge, personal communication). Curiously, *ykkC*-III motif

RNAs never associate with genes encoding the guanidine carboxylase protein, which was previously annotated as a urea carboxylase but that I demonstrated prefers guanidine as a substrate.[9] The genes for guanidine carboxylase and its associated proteins are the most common types of protein-coding regions controlled by guanidine-I riboswitches, and are also frequently regulated by guanidine-II riboswitches.

Although *ykkC*-III motif RNAs lack associations with guanidine carboxylase genes, all three RNA classes share other gene associations that are consistent with the hypothesis that they sense the same ligand. One example is arginase, which is predicted to convert arginine to urea and ornithine. The role played by this enzyme in guanidine metabolism is not clear, but there are multiple possibilities.[9] Finally, both guanidine-I riboswitches and *ykkC*-III motif RNAs are occasionally associated with a class of genes belonging to the superfamily of N-acyltransferases, again suggesting that the ligand for these RNA classes might be identical.

Guanidine is bound by *ykkC*-III motif RNAs with affinities similar to guanidine-I and -II riboswitch aptamers. To assess the ability of *ykkC*-III motif RNAs to bind guanidine, I created an RNA construct carrying 69 nucleotides encompassing the representative located upstream of the *emrE* gene from *Nocardia farcinica* (Figure 4-2A). Although this and other *ykkC*-III motif RNAs are associated with genes annotated as *emrE*, the genes actually code for SugE-like members of the SMR protein class. This RNA construct, called 69 *emrE*, was subjected to in-line probing[18,19] to identify any changes in structure upon incubation with guanidine. Robust structural modulation is observed (Figure 4-2B) primarily within the regions that also exhibit high sequence conservation, whereas

regions that lack sequence or structure conservation remain largely unstructured, or even become less structured, in the presence of guanidine.

A plot of the relative increase or decrease in spontaneous RNA scission at positions in three of the areas of modulation reveals a curve that is consistent with a 1:1 binding interaction and a dissociation constant (K_D) of 60 μM (Figure 4-2C). These are characteristics similar to those observed for guanidine-I riboswitches, and are consistent with our hypothesis that *ykkC*-III motif RNAs represent a third guanidine riboswitch class.

Interestingly, nucleotides spanning position 8 to 20, which have the potential to form part of the P0 hairpin, exhibit an RNA cleavage pattern that is not consistent with hairpin formation as presented in the secondary structure model. This result, along with bioinformatics data indicating that the P0 stem is sometimes absent, suggests that this stem is not necessary for ligand binding. To investigate this possibility, I examined a truncated version of the 69 *emrE* RNA, called 45 *emrE* (Figure 4-3A). The 45 *emrE* construct, which begins just one nucleotide prior to the P1 stem, is able to bind guanidine and modulates in similar regions as the full length construct (Figure 4-3B). P0 deletion only slightly diminishes the binding affinity, with a dissociation constant of 200 μM (Figure 4-3C). Thus, it appears that the P0 stem of this guanidine-III aptamer only modestly enhances the affinity for the ligand, which is similar to that observed when a poorly conserved P0 stem of the pre-Q_1-I riboswitch is removed.[20] However, this stem might be more important to other guanidine-III riboswitch representatives, or might serve other biological or biochemical roles in its natural setting.

To further validate the guanidine-III riboswitch class, I evaluated three additional representative sequences for guanidine binding. Aptamers associated with the arginase

gene from *Mycobacterium kansasii* (65 *arg*, Figure 4-4), the *smr* gene from *Streptomyces bingchenggensis* (64 *smr*, Figure 4-5), and the *emrE* gene from *Legionella pneumophila* (71 *emrE*, Figure 4-6) all exhibit structural modulation in similar regions with 1:1 binding characteristics and dissociation constants of 100, 200, and 20 μM, respectively. Again, the binding affinities of these guanidine-III riboswitches are similar to those of guanidine-I[9] and guanidine-II RNAs.[12]

Mutation of conserved nucleotides in guanidine-III riboswitch aptamers abolishes ligand binding. I introduced single mutations separately at nine highly-conserved positions in the 69 *emrE* RNA (Figure 4-7A) to investigate the importance of these nucleotides for guanidine recognition. Each mutation causes the loss of the ability of the RNA to recognize guanidine at concentrations as high as 1 mM (Figure 4-7B). Some mutations, such as M1 and M9, appear to vastly alter the RNA structure, which explains the loss of ligand binding function. However, other mutations such as M2 through M6 yield folded RNAs that appear to have the same overall architecture, but these RNAs are unable to recognize guanidine at concentrations as much as 20-fold above the K_D of the unaltered aptamer.

These results suggest that guanidine binding is due to specific interactions, rather than nonspecific interactions between the negatively-charged RNA and positively-charged guanidinium ions. Similarly, the conserved nucleotides present in guanidine-I riboswitch aptamers appear to serve important roles in forming the guanidine binding pocket, either by establishing the global architecture of the RNA or by directly interacting with the cognate ligand. Therefore, I predict that the highly-conserved nucleotides of guanidine-III

riboswitches will be engaged in directly binding the ligand, or in establishing the global architecture necessary to form the guanidine-binding pocket.

Selective ligand binding by guanidine-III aptamers. I also evaluated the ability of guanidine-III RNAs to discriminate against chemical analogs of guanidine by using in-line probing (Figure 4-8, Figure 4-9). Small substituents on a single nitrogen atom of guanidine are reasonably well tolerated, as is evident by examining the binding characteristics of methylguanidine and aminoguanidine (Figure 4-8). These two analogs modulate RNA structure in a manner that is identical to that observed for guanidine. Moreover, both analogs exhibit a K_D of 200 μM, which is only ~3-fold poorer than the value measured for guanidine. Larger substituents or modifications to multiple nitrogen atoms are strongly rejected. These results suggest that the RNA forms a binding pocket and exploits a steric strategy to limit the size of any substituent attached to the guanidine moiety. Most notably, urea (K_D poorer than 10 mM) is rejected by the ligand binding pocket with a discrimination factor of greater than 100 fold. This finding suggests that the presence of all three nitrogen atoms and their associated hydrogens, and perhaps also the positive charge of guanidinium, are key ligand recognition determinants whose absence in other natural guanidine analogs such as urea is deleterious to binding.

These and other characteristics of the 69 *emrE* guanidine-III RNA are somewhat different than those observed for a guanidine-II riboswitch RNA. Although similarities in nucleotide sequences exist between guanidine-II and -III riboswitches, namely the conserved ACGR segments, high-resolution structural models will be needed to determine if the two classes form guanidine-binding pockets that are similar in structure. In contrast, ligand binding by guanidine-I riboswitches is much more selective than these other two

classes. I demonstrated that a representative of guanidine-I riboswitches more strongly discriminates against every guanidine analog tested.[9] These differences in ligand selectivity between class I versus class II and III guanidine riboswitches might explain why guanidine-I riboswitches are more common in bacteria.

Concluding remarks. The guanidine-III riboswitch class, frequently found in species of Actinobacteria, is abundant but less common than the guanidine-I and -II riboswitch classes. The genes most commonly controlled by guanidine-III RNAs mostly overlap with those of guanidine-I RNAs. However, some rarer and many hypothetical genes are uniquely associated with guanidine-III riboswitches. This fact prompts interesting questions for researchers who seek a deeper understanding of proteins that affect the metabolism of free guanidine. For example, one gene type associated with guanidine-III riboswitches is *rhtA*, which is annotated as encoding a threonine/homoserine efflux transporter. However, it seems possible that transporter genes associated with this riboswitch might have altered their original ligand specific to now recognize guanidine, as was the case for some transporters associated with fluoride riboswitches that converted from chloride to fluoride selectivity.[7] To expand insight regarding guanidine metabolism, one could further examine the biochemical functions of the genes associated with all guanidine riboswitches.

Experimental validation of the *ykkC*-III motif as a third class of guanidine riboswitches again indicates that free guanidine is an important metabolite in bacteria, but its role is almost completely unknown. Every cognate ligand for the known riboswitches to date represents a fundamentally important molecule or ion for bacteria and often also for organisms from all three domains of life. A major hurdle in identifying the cognate ligand

for all three *ykkC* motif RNAs was the lack of known roles for free guanidine in biology. Thus, the discovery and validation of three distinct classes of guanidine-responsive riboswitches has also revealed free guanidine as an important, but largely unexplored, compound in many species of bacteria.

MATERIALS AND METHODS

Chemicals and oligonucleotides. All chemicals and chemically synthesized oligonucleotides were purchased from Sigma-Aldrich with the exception of guanidinosuccinic acid (Alfa Chemistry). [γ-^{32}P]-ATP was purchased from PerkinElmer. Enzymes were purchased from New England BioLabs, unless otherwise noted. A complete list of DNA and RNA oligonucleotides used in this chapter can be found in Table 4.1.

Bioinformatics analyses. Additional examples of *ykkC*-III RNAs were identified using Infernal 1.1[15] to search the RefSeq version 63 plus additional environmental microbial databases.[16] Iterative searches for new sequences were performed based on the previously published alignment of the *ykkC*-III motif,[11] revealing over 400 new examples. The consensus sequence and secondary structure model was constructed using R2R.[17]

RNA oligonucleotide preparation. 5′ ^{32}P-labeled RNA oligonucleotides were prepared as previously described.[12] The only exception is the truncated 45 *emrE* RNA, for which synthetic RNA was chemically synthesized and subsequently purified and 5′ ^{32}P labeled as previously described.

RNA in-line probing analyses. In-line probing assays were performed largely as previously described.[18,19] Additional details can be found in Chapter 3.[12]

ACKNOWLEDGEMENTS

We thank Adam Roth, Zasha Weinberg, Ruben Atilho, James Nelson and other members of the Breaker laboratory as well as Randy Stockbridge of the University of Michigan for helpful discussions. M.E.S. was supported by an NIH Cellular and Molecular Biology Training Grant (T32GM007223). This work was also supported by NIH grants (GM022778, DE022340) as well as funds from Howard Hughes Medical Institute to R.R.B.

REFERENCES

(1) Breaker, R.R. (2011) Prospects for riboswitch discovery and analysis. *Mol Cell 43,* 867-879.

(2) Meyer, M.M., Hammond, M.C., Salinas, Y., Roth, A., Sudarsan, N., Breaker, R.R. (2011) Challenges of ligand identification for riboswitch candidates. *RNA Biol 8,* 5-10.

(3) Kim, P.B., Nelson, J.W., Breaker, R.R. (2015) An ancient riboswitch class in bacteria regulates purine biosynthesis and one-carbon metabolism. *Mol Cell 57,* 317-328.

(4) Nelson, J.W., Sudarsan, N., Furukawa, K., Weinberg, Z., Wang, J.X., Breaker, R.R. (2013) Riboswitches in eubacteria sense the second messenger c-di-AMP. *Nat Chem Biol 9,* 834-839.

(5) Dambach, M., Sandoval, M., Updegrove, T.B., Anantharaman, V., Aravind, L., Waters, L.S., Storz, G. (2015) The ubiquitous *yybP-ykoY* riboswitch is a manganese-responsive regulatory element. *Mol Cell 57,* 1099-1109.

(6) Price, I.R., Gaballa, A., Ding, F., Helmann, J.D., Ke, A. (2015) Mn^{2+}-sensing mechanisms of *yybP-ykoY* orphan riboswitches. *Mol Cell 57,* 1110-1123.

(7) Baker, J.L., Sudarsan, N., Weinberg, Z., Roth, A., Stockbridge, R.B., Breaker, R.R. (2012) Widespread genetic switches and toxicity resistance proteins for fluoride. *Science 335,* 233-235.

(8) Barrick, J.E., Corbino, K.A., Winkler, W.C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I. (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc Natl Acad Sci USA 101,* 6421-6426.

(9) Nelson, J.W., Atilho, R.M., Sherlock, M.E., Stockbridge, R.B., Breaker, R.R. (2017) Metabolism of free guanidine in bacteria is regulated by a widespread riboswitch class. *Mol Cell* 65, 220-230.

(10) Weinberg, Z., Barrick, J.E., Yao, Z., Roth, A., Kim, J.N., Gore, J., et al. (2007) Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res 35*, 4809–4819.

(11) Weinberg, Z., Wang, J.X., Bogue, J., Yang, J., Corbino, K., Moy, R.H., Breaker, R.R. (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol 11*, 31.

(12) Sherlock, M.E., Malkowski, S.N., Breaker, R.R. (2017) Biochemical validation of a second guanidine riboswitch class in bacteria. *Biochemistry 56*, 352-358.

(13) Schuldiner, S.S. (2009) EmrE, a model for studying evolution and mechanism of ion-coupled transporters. *Biochim Biophys Acta 1794*, 748-762.

(14) Jack, D.L., Storms, M.L., Tchieu, J.H., Paulsen, I.T., Saier Jr., M.H. 2000. A broad-specificity multidrug efflux pump requiring a pair of homologous SMR-type proteins. *J Bacteriol 182*, 2311-2313.

(15) Nawrocki, E.P., Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics 29*, 2933-2935.

(16) Weinberg, Z., Kim, P.B., Chen, T.H., Li, S., Harris, K.A., Lünse, C.E., Breaker, R.R. (2015) New classes of self-cleaving ribozymes revealed by comparative genomics analysis. *Nat Chem Biol 11*, 606-610.

(17) Weinberg, Z., Breaker, R.R. (2011) R2R-software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics 12*, 1.

(18) Regulski, E.E., Breaker, R.R. (2008) In-line probing analysis of riboswitches. *Methods Mol Biol 419*, 53-67.

(19) Soukup, G.A., Breaker, R.R. (1999) Relationship between internucleotide linkage geometry and the stability of RNA. *RNA 5*, 1308-1325.

(20) Roth, A., Winkler, W.C., Regulski, E.E., Lee, B.W., Lim, J., Jona, I., Barrick, J.E., Ritwik, A., Kim, J.N., Welz, R., et al. (2007) A riboswitch selective for the queuosine precursor preQ$_1$ contains an unusually small aptamer domain. *Nat Struct Mol Biol 14*, 308-317.

FIGURES

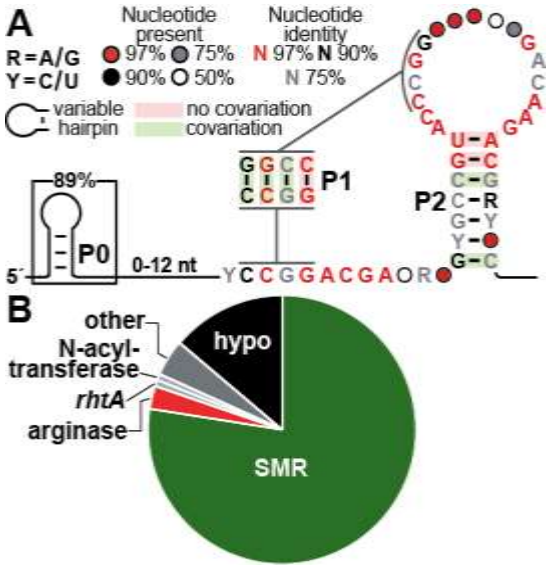

Figure 4-1. Structure and function of guanidine-III riboswitches. (A) The consensus sequence and structural model for the ~460 examples of *ykkC*-III motif (guanidine-III) RNAs. Covarying base pairs (indicated by green shading) are found in the P1 and P2 stems which form a pseudoknot as indicated by the gray lines. Many highly conserved (>97%) nucleotides are found in the loop regions between the stems. (B) Genes controlled by guanidine-III riboswitches, including a large number of predicted SMR transporters (green), arginases (red), and other (gray) or hypothetical (hypo, black) genes.

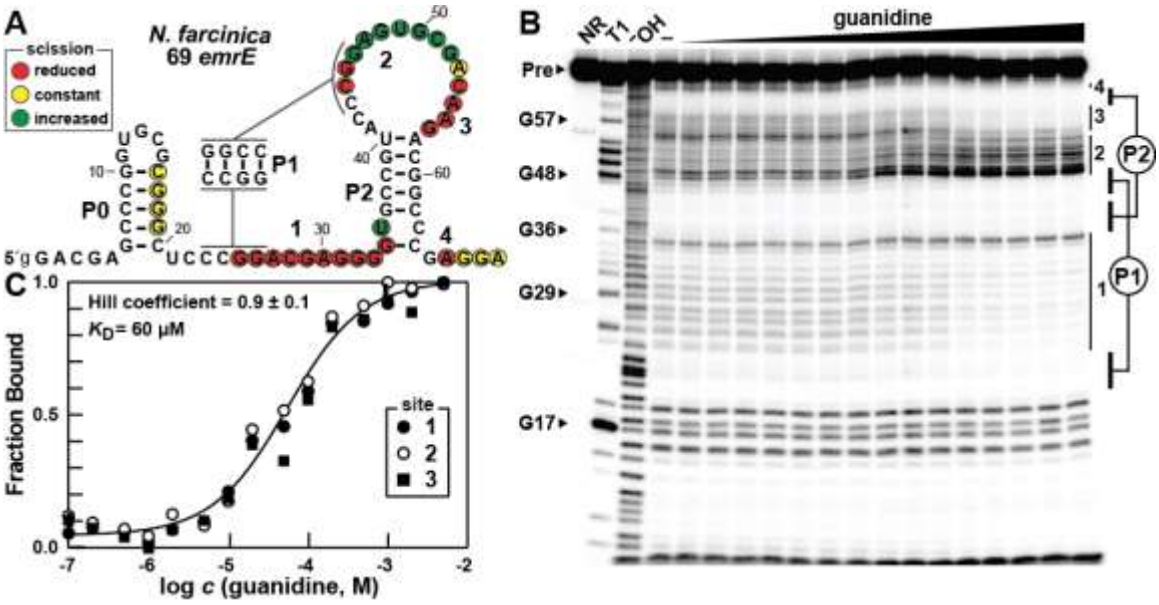

Figure 4-2. Guanidine is bound by a guanidine-III RNA *in vitro*. (A) The sequence and observed secondary structure of the 69 *emrE* RNA from *N. farcinica*. Lowercase 5′ nucleotides indicate G residues added to enable transcription by T7 RNA Polymerase. Red, yellow, and green circles indicate nucleotides that undergo decreased, constant, or increased scission, respectively, upon addition of the ligand as determined by their relative band intensities in B. Regions of guanidine-mediated structural modulation numbered to correspond to those in B and C. (B) Polyacrylamide gel electrophoresis (PAGE) analysis of in-line probing reactions of 69 *emrE* RNA in the absence (–) and presence of guanidine ranging from 100 nM to 5 mM. NR, T1, and $^{-}$OH indicate no reaction, RNA partially digested with T1 ribonuclease (cleavage after each G), and RNA partially digested under alkaline conditions (cleavage after every nucleotide). Precursor RNA (Pre) as well as certain G residues within the RNA sequence are indicated and regions of modulation are numbered and indicated with vertical bars. Base pairing of the pseudoknotted P1 and P2 stems is depicted. (C) Plot of the fraction of RNA bound to ligand versus the logarithm of the guanidine concentration as determined by quantification of B (see reference 12 for

details). The K_D and Hill coefficient were determined by the sigmoidal fit and the reported error represents the standard deviation reported in the fitting parameters.

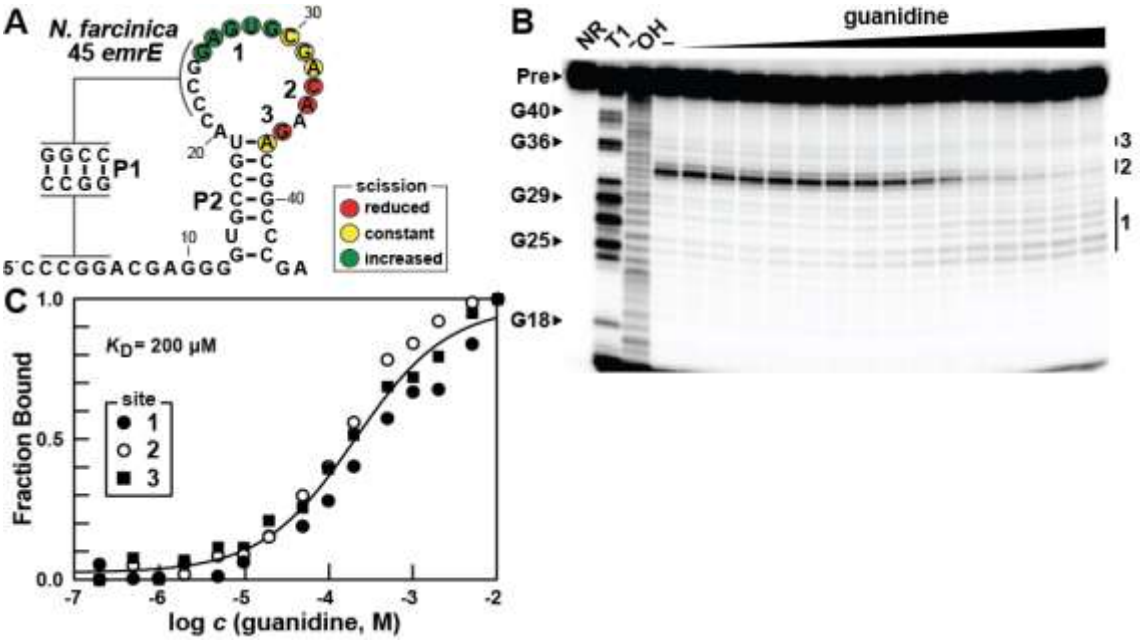

Figure 4-3. Guanidine binding is unaffected by the deletion of the P0 stem. (A) The sequence and observed secondary structure of the 45 *emrE* RNA from *N. farcinica*, which removes nucleotides including the P0 stem in the 5′ region. Additional annotations are as described in the legend to Figure 4-2A. (B) Polyacrylamide gel electrophoresis (PAGE) analysis of in-line probing reactions of 45 *emrE* RNA in the absence (–) and presence of guanidine ranging from 200 nM to 10 mM. (C) Plot of the normalized fraction of RNA bound to ligand versus the logarithm of the guanidine concentration as determined by quantification of the PAGE data in B. The K_D was determined by the sigmoidal fit.

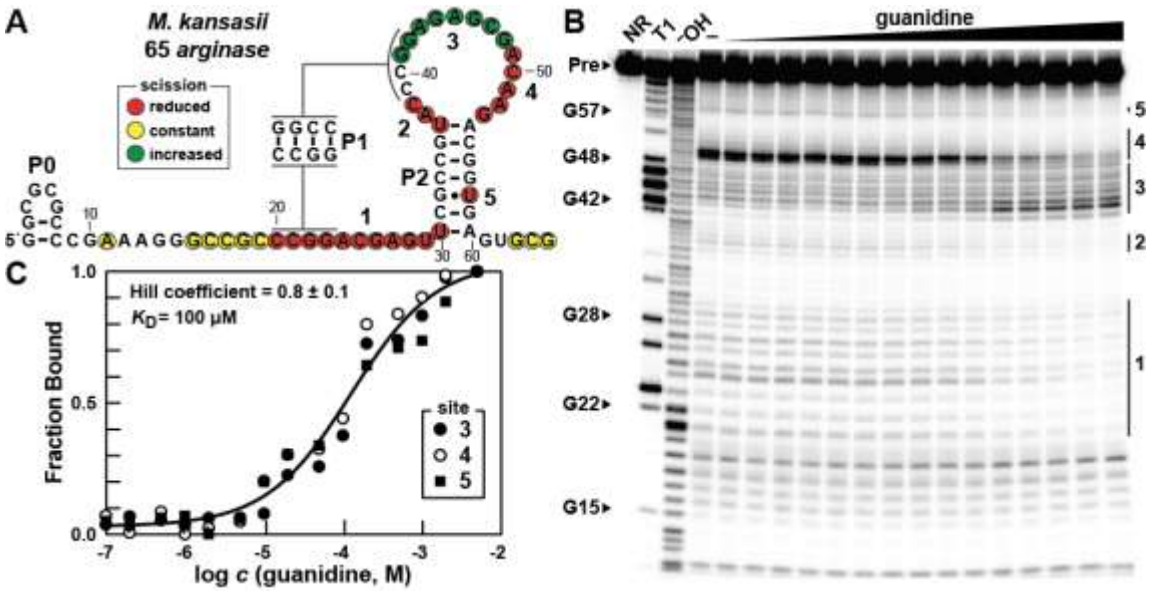

Figure 4-4. Guanidine binding by the 65 *arginase* RNA from *M. kansasii*. (A) The sequence and observed secondary structure of the 65 *arginase M. kansasii* RNA. Additional annotations and labels are as described in the legend to Figure 4-2A. (B) PAGE analysis of in-line probing reactions of 65 *arginase* RNA in the absence (–) and presence of guanidine ranging from 100 nM to 5 mM. Additional annotations are as described in the legend to Figure 4-2B. (C) Plot of the normalized fraction of RNA bound to ligand versus the logarithm of the guanidine concentration as determined by quantification of the PAGE data in B. The K_D and Hill coefficient were determined by the sigmoidal fit and the reported error represents the standard deviation reported in the fitting parameters.

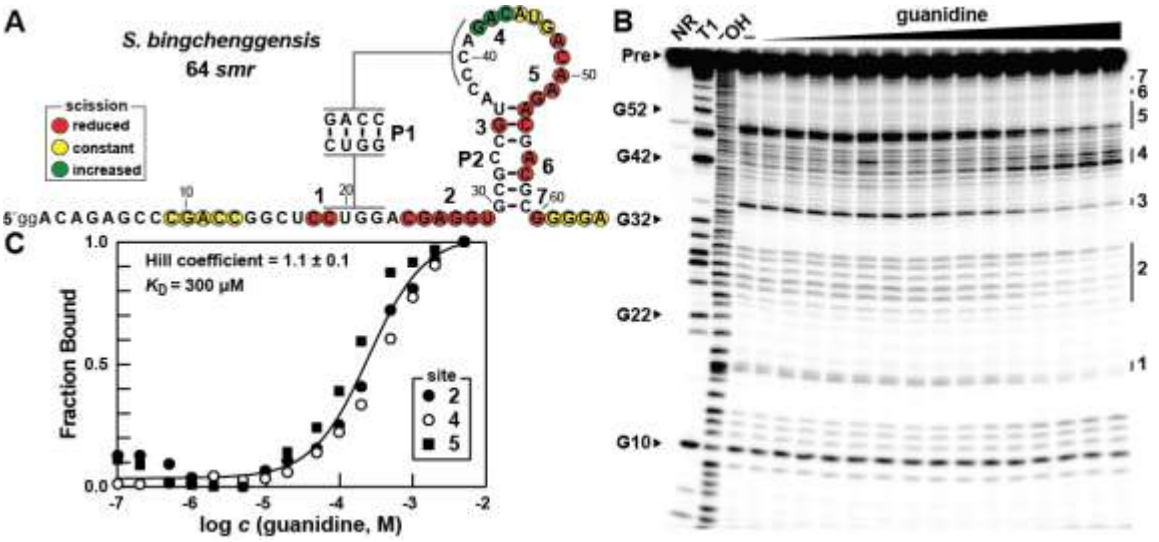

Figure 4-5. Guanidine binding by the 64 *smr* RNA from *S. bingchenggensis*. (A) The sequence and observed secondary structure of the 64 *smr S. bingchenggensis* RNA. Additional annotations and labels are as described in the legend to Figure 4-2A. (B) PAGE analysis of in-line probing reactions of 64 *smr* RNA in the absence (–) and presence of guanidine ranging from 100 nM to 5 mM. Additional annotations are as described in the legend to Figure 4-2B. (C) Plot of the normalized fraction of RNA bound to ligand versus the logarithm of the guanidine concentration as determined by quantification of the PAGE data in B. The K_D and Hill coefficient were determined by the sigmoidal fit and the reported error represents the standard deviation reported in the fitting parameters.

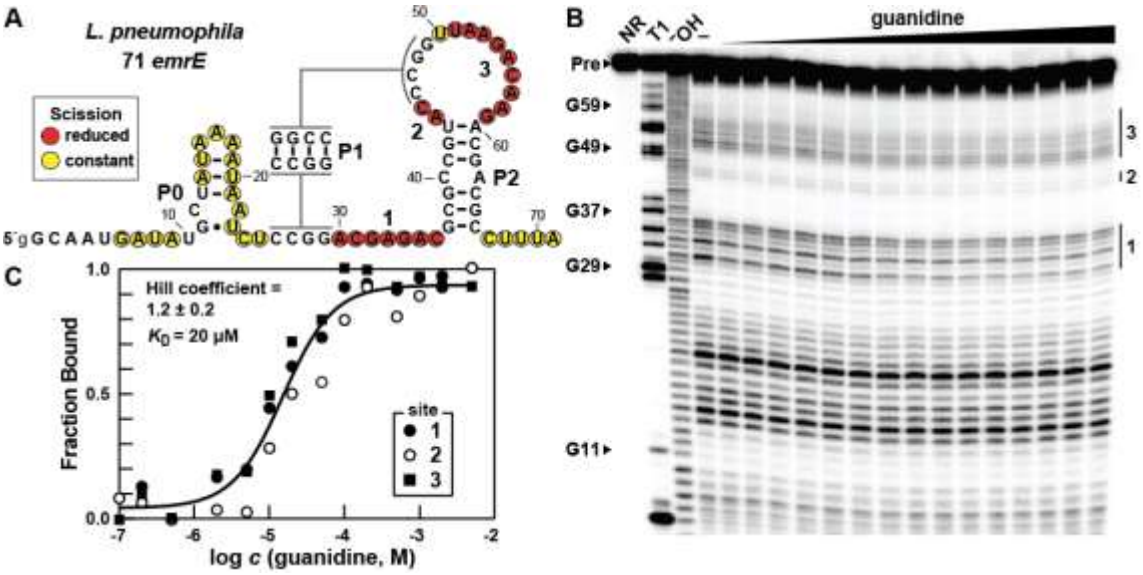

Figure 4-6. Guanidine binding by the 71 *emrE* RNA from *L. penumophila*. (A) The sequence and observed secondary structure of the 71 *emrE L. penumophila* RNA. Additional annotations and labels are as described in the legend to Figure 4-2A. (*B*) PAGE analysis of in-line probing reactions of 71 *emrE* RNA in the absence (–) and presence of guanidine ranging from 100 nM to 5 mM. Additional annotations are as described in the legend to Figure 4-2B. (C) Plot of the normalized fraction of RNA bound to ligand versus the logarithm of the guanidine concentration as determined by quantification of the PAGE data in B. The K_D and Hill coefficient were determined by the sigmoidal fit and the reported error represents the standard deviation reported in the fitting parameters.

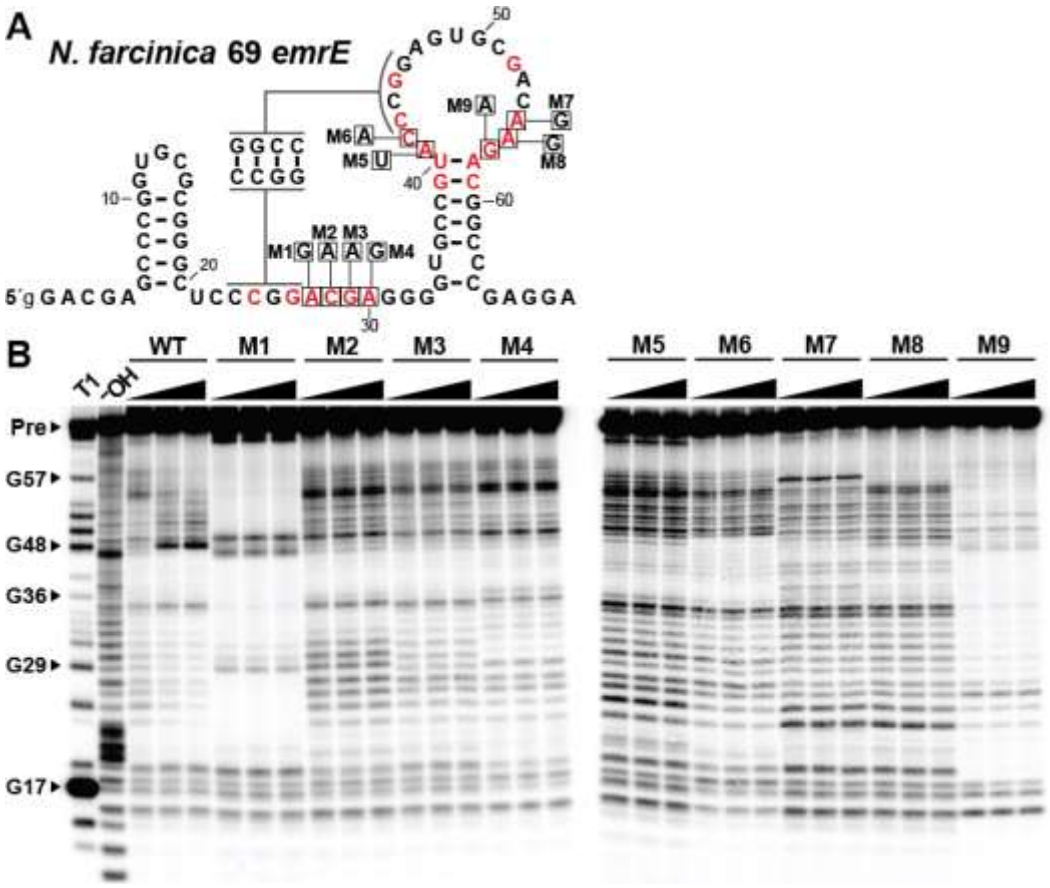

Figure 4-7. Mutations at conserved nucleotides prevent recognition of guanidine. (A) The sequence and secondary structure of the 69 *emrE* RNA from *N. farcinica*. Red letters denote highly conserved (>97%) nucleotides within the consensus *ykkC*-III motif. Nucleotides that are altered in each mutated RNA are indicated with boxes and labeled. (B) PAGE analysis of in-line probing reactions of WT and mutated sequences of the 69 *emrE* RNA in the presence of 0, 0.1, and 1 mM guanidine. Additional annotations are as described in the legend to Figure 4-2B.

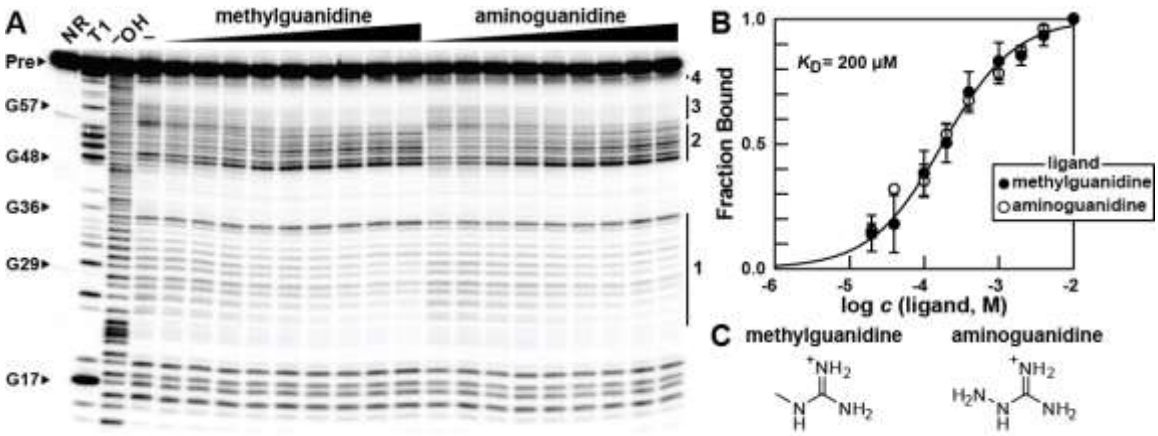

Figure 4-8. Guanidine-III aptamers can discriminate against even close analogs of guanidine. (A) PAGE analysis of in-line probing reactions of *69 emrE* RNA from *N. farcinica* in the absence (–) or presence of methylguanidine or aminoguanidine ranging from 1 μM to 40 mM. Additional annotations are as described in the legend to Figure 4-2B. (B) Plot of the fraction of RNA bound to ligand versus the logarithm of the methyl- or aminoguanidine concentration as determined by quantification of the data in A (see reference 12 for details). Each point represents the average and standard deviation of the fraction bound by assessing three sites of structure modulation (regions 1, 2, and 3, as indicated in Figure 4-2A). Only the fit calculated for methylguanidine binding is shown for clarity, although both fits yield an approximately identical K_D. The resulting curves are consistent with a 1:1 binding interaction. (C) Chemical structures of methylguanidine and aminoguanidine.

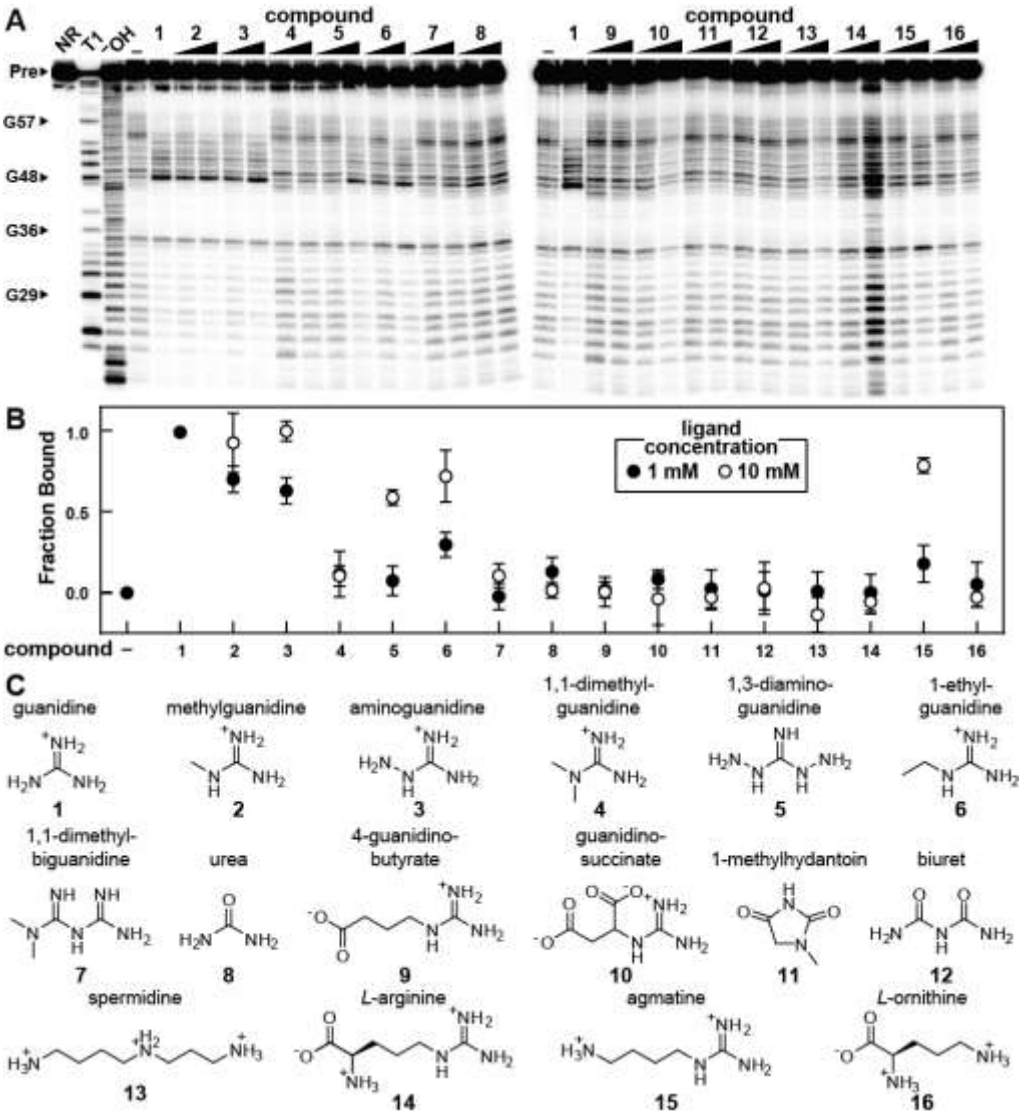

Figure 4-9. A guanidine-III RNA only recognizes chemical analogs with highly similar structures to its ligand. (A) PAGE analysis of in-line probing reactions of 69 *emrE* RNA from *N. farcinica* in the absence (–) or presence of guanidine or its structural analogs at 1 and 10 mM concentrations. Additional annotations are as described in the legend to Figure 4-2B. (B) Plot of the normalized fraction of RNA bound to ligand for each compound at 1 and 10 mM concentrations as determined by quantification of the PAGE data in B. Each point represents the average and standard deviation of the fraction bound based on in-line

probing data at three sites of modulation (regions 1, 2, and 3, as indicated in Figure 4-2A). (C) Chemical structures of each analog in their predominant state at pH 7.

TABLES

Table 4-1. Sequences of synthetic DNAs and RNA used in Chapter 4.

Name	Sequence	Annotation
MES055	TAATACGACTCACTATAGGACGA GCCCGGTGCGCGGGCTCCCGGAC GAGGGGTGCCGTACCCGGAGTGC GACAAGACGGCCCGAGGA	Forward template for transcription of the WT *N. farcinica* 69 *emrE* RNA
MES056	TCCTCGGGCCGTCTTGTCGCACTC CGGGTACGGCACCCCTCGTCCGG GAGCCCGCGCACCGGGCTCGTCC TATAGTGAGTCGTATTA	Reverse template for transcription of the WT *N. farcinica* 69 *emrE* RNA
MES061	TAATACGACTCACTATAGGACGA GCCCGGTGCGCGGGCTCCCGGGC GAGGGGTGCCGTACCCGGAGTGC GACAAGACGGCCCGAGGA	Forward template for transcription of the M1 *N. farcinica* 69 *emrE* RNA
MES062	TCCTCGGGCCGTCTTGTCGCACTC CGGGTACGGCACCCCTCGCCCGG GAGCCCGCGCACCGGGCTCGTCC TATAGTGAGTCGTATTA	Reverse template for transcription of the M1 *N. farcinica* 69 *emrE* RNA

MES063	TAATACGACTCACTATAGGACGA GCCCGGTGCGCGGGCTCCCGGAA GAGGGGTGCCGTACCCGGAGTGC GACAAGACGGCCCGAGGA	Forward template for transcription of the M2 *N. farcinica* 69 *emrE* RNA
MES064	TCCTCGGGCCGTCTTGTCGCACTC CGGGTACGGCACCCCTCTTCCGG GAGCCCGCGCACCGGGCTCGTCC TATAGTGAGTCGTATTA	Reverse template for transcription of the M2 *N. farcinica* 69 *emrE* RNA
MES065	TAATACGACTCACTATAGGACGA GCCCGGTGCGCGGGCTCCCGGAC AAGGGGTGCCGTACCCGGAGTGC GACAAGACGGCCCGAGGA	Forward template for transcription of the M3 *N. farcinica* 69 *emrE* RNA
MES066	TCCTCGGGCCGTCTTGTCGCACTC CGGGTACGGCACCCCTTGTCCGG GAGCCCGCGCACCGGGCTCGTCC TATAGTGAGTCGTATTA	Reverse template for transcription of the M3 *N. farcinica* 69 *emrE* RNA
MES067	TAATACGACTCACTATAGGACGA GCCCGGTGCGCGGGCTCCCGGAC GGGGGGTGCCGTACCCGGAGTGC GACAAGACGGCCCGAGGA	Forward template for transcription of the M4 *N. farcinica* 69 *emrE* RNA

MES068	TCCTCGGGCCGTCTTGTCGCACTC CGGGTACGGCACCCCCCGTCCGG GAGCCCGCGCACCGGGCTCGTCC TATAGTGAGTCGTATTA	Reverse template for transcription of the M4 *N. farcinica* 69 *emrE* RNA
MES069	TAATACGACTCACTATAGGACGA GCCCGGTGCGCGGGCTCCCGGAC GAGGGGTGCCGTTCCCGGAGTGC GACAAGACGGCCCGAGGA	Forward template for transcription of the M5 *N. farcinica* 69 *emrE* RNA
MES070	TCCTCGGGCCGTCTTGTCGCACTC CGGGAACGGCACCCCTCGTCCGG GAGCCCGCGCACCGGGCTCGTCC TATAGTGAGTCGTATTA	Reverse template for transcription of the M5 *N. farcinica* 69 *emrE* RNA
MES071	TAATACGACTCACTATAGGACGA GCCCGGTGCGCGGGCTCCCGGAC GAGGGGTGCCGTAACCGGAGTGC GACAAGACGGCCCGAGGA	Forward template for transcription of the M6 *N. farcinica* 69 *emrE* RNA
MES072	TCCTCGGGCCGTCTTGTCGCACTC CGGTTACGGCACCCCTCGTCCGG GAGCCCGCGCACCGGGCTCGTCC TATAGTGAGTCGTATTA	Reverse template for transcription of the M6 *N. farcinica* 69 *emrE* RNA

MES073	TAATACGACTCACTATAGGACGA GCCCGGTGCGCGGGCTCCCGGAC GAGGGGTGCCGTACCCGGAGTGC GACGAGACGGCCCGAGGA	Forward template for transcription of the M7 *N. farcinica* 69 *emrE* RNA
MES074	TCCTCGGGCCGTCTCGTCGCACTC CGGGTACGGCACCCCTCGTCCGG GAGCCCGCGCACCGGGCTCGTCC TATAGTGAGTCGTATTA	Reverse template for transcription of the M7 *N. farcinica* 69 *emrE* RNA
MES075	TAATACGACTCACTATAGGACGA GCCCGGTGCGCGGGCTCCCGGAC GAGGGGTGCCGTACCCGGAGTGC GACAGGACGGCCCGAGGA	Forward template for transcription of the M8 *N. farcinica* 69 *emrE* RNA
MES076	TCCTCGGGCCGTCCTGTCGCACTC CGGGTACGGCACCCCTCGTCCGG GAGCCCGCGCACCGGGCTCGTCC TATAGTGAGTCGTATTA	Reverse template for transcription of the M8 *N. farcinica* 69 *emrE* RNA
MES077	TAATACGACTCACTATAGGACGA GCCCGGTGCGCGGGCTCCCGGAC GAGGGGTGCCGTACCCGGAGTGC GACAAAACGGCCCGAGGA	Forward template for transcription of the M9 *N. farcinica* 69 *emrE* RNA

MES078	TCCTCGGGCCGTCCTGTCGCACTC CGGGTACGGCACCCCTCGTCCGG GAGCCCGCGCACCGGGCTCGTCC TATAGTGAGTCGTATTA	Reverse template for transcription of the M9 *N. farcinica* 69 *emrE* RNA
MES079	TAATACGACTCACTATAGGCGCG CCGAAAGGGCCGCCCGGACGAGT TCGCCGTACCCGGAGAGCGACAA GACGGTGAGTGCG	Forward template for transcription of the WT *M. kansasii* 65 *arginase* RNA
MES080	CGCACTCACCGTCTTGTCGCTCTC CGGGTACGGCGAACTCGTCCGGG CGGCCCTTTCGGCGCGCCTATAGT GAGTCGTATTA	Reverse template for transcription of the WT *M. kansasii* 65 *arginase* RNA
MES087	CCCGGACGAGGGGUGCCGUACCC GGAGUGCGACAAGACGGCCCGA	Synthetic RNA of the truncated *N. farcinica* 45 *emrE*
MES090	TAATACGACTCACTATAGGACAG AGCCCGACCGGCTCCTGGACGAG GTGCGCCGTACCCAGACATGACA AGACGACGCGGGGA	Forward template for transcription of the WT *S. bingchenggensis* 64 *smr* RNA

MES091	TCCCCGCGTCGTCTTGTCATGTCT GGGTACGGCGCACCTCGTCCAGG AGCCGGTCGGGCTCTGTCCTATAG TGAGTCGTATTA	Reverse template for transcription of the WT *S. bingchenggensis* 64 *smr* RNA
MES092	TAATACGACTCACTATAGGCAAT GATATGCTATAAAATAATCTCCG GACGAGACGCGCCGTACCCGGTT AAGACAAGACGACGCCTTTA	Forward template for transcription of the WT *L. pneumophila* 71 *emrE* RNA
MES093	TAAAGGCGTCGTCTTGTCTTAACC GGGTACGGCGCGTCTCGTCCGGA GATTATTTTATAGCATATCATTGC CTATAGTGAGTCGTATTA	Reverse template for transcription of the WT *L. pneumophila* 71 *emrE* RNA

CHAPTER FIVE

Riboswitches for the alarmone ppGpp expand the collection of RNA-based signaling systems

Largely adapted from the following manuscript in preparation:

Sherlock, M.E., Sudarsan, N., Breaker, R.R. Riboswitches for the alarmone ppGpp expand the collection of RNA-based signaling systems (submitted).

ABSTRACT

Riboswitches are noncoding portions of certain mRNAs that bind metabolite, coenzyme, signaling molecule, or inorganic ion ligands and regulate gene expression. Most known riboswitches sense ligands that are derivatives of RNA monomers. This bias in ligand chemical composition is consistent with the hypothesis that widespread riboswitch classes first emerged during the RNA World, which is proposed to have existed before proteins were present. In this chapter, I report the discovery and validation of a natural riboswitch class that selectively binds guanosine tetraphosphate (ppGpp), a widespread signaling molecule and bacterial 'alarmone' derived from the ribonucleotide GTP. Riboswitches for ppGpp regulate genes involved in branched-chain amino acid biosynthesis and transport, as well as other gene classes that previously had not been implicated to be part of its signaling network. This newfound riboswitch-alarmone partnership supports the hypothesis that prominent RNA World signaling pathways have been retained by modern cells to control key biological processes.

INTRODUCTION

Among the 38 distinct riboswitch[1-3] classes validated previously,[4] five respond to RNA-derived signaling molecules that might have emerged in an RNA World, but have been retained to trigger regulatory processes in modern cells. It has previously been proposed[5,6] that riboswitches for the RNA-derived molecules c-di-GMP,[7,8] c-di-AMP,[9] ZTP,[10] and c-AMP-GMP[11,12] represent modern versions of primordial signaling processes built entirely from RNA.

A widely-distributed RNA signaling molecule that is essential for many bacteria is ppGpp (Figure 5-1A), a guanosine tetraphosphate alarmone originally called magic spot.[13] This compound, along with its natural precursor pppGpp, is widely implicated in regulating RNA polymerase and other protein factors to control metabolic and physiological responses to various stresses[14-16] including the long-studied stringent response. The stringent response is a major starvation-induced adaptation that can be triggered by amino acid depletion.[17-19] In Proteobacteria, this adaptation is accomplished at least in part via the direct inhibition of RNA polymerase by ppGpp binding.[16]

In Firmicutes, transcription regulation by ppGpp is known to act indirectly. GTP levels drop as a result of (p)ppGpp synthesis to specifically downregulate transcription of genes that initiate with a G nucleotide, which affects the production of certain mRNAs and also rRNAs. GTP depletion still permits transcription of genes that initiate with an A nucleotide, such as those for branched-chain amino acid (BCAA) biosynthesis.[20,21] Specific gene activation also occurs by the dissociation of the CodY transcriptional repressor protein when GTP levels drop as a result of (p)ppGpp synthesis.[22] This protein-based system thus allows ppGpp to indirectly activate CodY-controlled genes. In this chapter, I describe

another mechanism for ppGpp-triggered regulation of gene expression in Firmicutes mediated by numerous examples of selective ppGpp-sensing riboswitches.

RESULTS AND DISCUSSION

Identification of ppGpp riboswitch candidates. The widespread distribution of ppGpp in biology, its role in regulating fundamental biological processes, and its construction from an RNA mononucleotide are consistent with an ancient RNA World origin.[5] Given these characteristics, it had previously been speculated that ppGpp would be an ideal ligand candidate for a riboswitch class that had yet to be discovered.[5,6,10,12] Thus, I was well positioned to test ppGpp as a candidate ligand when I identified a riboswitch subtype (Figure 5-1B) of unknown function that commonly associates with genes for BCAA metabolism (Figure 5-1C). The ppGpp riboswitch candidates were recognized as 'subtype 2a' variants of *ykkC* motif RNAs.[24]

In Chapter 2, my colleagues and I determined that the predominant members (subtype 1) of the *ykkC* motif collection function as guanidine-sensing RNAs, which are collectively called guanidine-I riboswitches.[23] I also discovered that certain *ykkC* motif RNAs (subtype 2) carry distinctive mutations at nucleotide positions found directly in the binding pocket.[25,26] Members of the *ykkC* subtype 2 collection do not bind guanidine, confirming that the binding site differences indeed alter riboswitch specificity.

Although ppGpp had previously been tested for binding to *ykkC* RNAs,[27] I determined in retrospect that the specific RNA sequences tested belong to the guanidine-I riboswitch class. I subsequently further separated subtype 2 into four additional classes (subtype 2a through 2d) based on distinctive gene associations and aptamer sequence variations (see

Chapters 6 and 7 for additional details). I identified ppGpp as a likely candidate ligand because of the association of many subtype 2a representatives with genes for BCAA biosynthesis. I developed this hypothesis because BCAA biosynthesis is activated during the stringent response in Firmicutes, which are the predominant organisms that contain subtype 2a *ykkC* motif variants.

Subtype 2a *ykkC* motif RNAs selectively bind ppGpp. To test for ppGpp riboswitch function, I first examined a BCAA-associated candidate RNA construct from the *ilvE* gene of *Thermosediminibacter oceani* (Figure 5-2A) by using in-line probing. This method exploits the differential instability of RNA phosphoester bonds in unstructured versus structured regions, to provide evidence that ligand docking alters the folding state of the ligand-binding aptamer.[28,29] I observed a pattern of ppGpp-dependent RNA structural modulation with this construct (Figure 5-2B) to yield a binding curve consistent with 1:1 complex formation and an apparent dissociation constant (K_D) of ~10 nM (Figure 5-2C). Nucleotides that exhibit structural modulation are located in regions that are known to form the binding pocket for guanidine-I riboswitches.[25,26] Remarkably, these two riboswitch classes utilize the same general architecture but have altered key nucleotides to create binding pockets capable of recognizing completely different ligands.

Individual mutation of six highly-conserved nucleotides (M1 through M6) within the *T. oceani* RNA (Figure 5-3A) predicted to be directly involved in ligand binding each cause at least a three order of magnitude loss in ppGpp binding affinity (Figure 5-3B). I speculate that these nucleotide positions might form the binding pocket for ppGpp because their equivalent positions are directly involved in contacting the ligand in the highly-similar members of the guanidine riboswitch class. Reversing the nucleotide identity of a

conserved base-pair (M7) similarly reduces ligand affinity by about one order of magnitude. These findings demonstrate that conserved nucleotides are critical for optimal binding by the aptamer.

I assessed riboswitch binding specificity via in-line probing by using a collection of ppGpp analogs comprised of guanosine nucleotide derivatives. In *Bacillus subtilis*, the most abundant natural version of the signaling molecule is the pentaphosphate species (pppGpp), which is directly synthesized when GTP is used as the biosynthetic precursor. The riboswitch displays equal affinities for ppGpp (Figure 5-2C) and pppGpp (Figure 5-4), indicating they are both likely to trigger gene expression in cells. At higher concentrations, similar compounds such as pGp, GTP, and GDP modulate the RNA structure in the same regions as pppGpp and ppGpp (Figure 5-5A). Both GDP and GTP require >1,000-fold higher concentration than ppGpp to half-maximally saturate the riboswitch aptamer (Figure 5-5B), whereas other nucleotides have no effect regardless of the concentrations tested (Figure 5-5C).

I hypothesized that the ppGpp aptamer RNA might recognize the guanine base of its ligand through Watson-Crick pairing and sought to identify a conserved cytosine residue that might perform this function. I identified the conserved cytosine between the P2 and P3 stems as a likely candidate due to its role in the guanidine-I riboswitch structure.[25,26] Specifically, this C forms a long-range base pair with a conserved G in the internal loop of the P3 stem. This base pair appears to be absent in ppGpp riboswitches because this corresponding G nucleotide is no longer conserved. It seemed plausible that the C nucleotide (C71 in the *T. oceani* 112 *ilvE* RNA), which is in a region of reduced scission upon ligand binding in-line probing reactions (Figure 5-2A), might be available to base

pair to the nucleobase of ppGpp. Indeed, introducing a C to U mutation at this position (M8) switches the specificity of the aptamer from a G-containing ligand to an A-containing ligand, as demonstrated by ability of the mutant RNA to preferentially recognize pAp over pGp or ppGpp (Figure 5-6). These data provide support for the role of this C nucleotide in the aptamer structure as well as for ppGpp as the cognate ligand. Currently, there is no evidence for a naturally-occurring aptamer for ppApp (or a similar adenosine-containing molecule) because the nucleotide in this position of the aptamer is conserved as a C across all examples of the riboswitch in nature.

Transcription termination regulated by a ppGpp riboswitch. To determine whether ppGpp binding by a member of this riboswitch class directly affects gene expression, I performed single turnover *in vitro* transcription termination assays[30] as previously described[31] with a representative from the 5′-untranslated region (UTR) of the *ilvE* gene of *Desulfitobacterium hafniense* (Figure 5-7). I chose this construct because the expression platform exhibited an architecture that suggested the *D. hafniense* riboswitch uses a transcription termination mechanism, and this construct was confirmed to operate with RNA polymerase from *Escherichia coli*. Specifically, I predicted that the P3 helix of the aptamer acts as an antiterminator stem, and so transcription of the full-length mRNA should increase when ppGpp is bound. As expected, the wild-type (WT) riboswitch permits RNA polymerase to proceed through the stop site of the adjoining intrinsic terminator stem with greater efficiency only when sufficient ppGpp is added to the reaction (Figure 5-7B). An RNA (M9) with a single mutation at a nucleotide position that disrupts the binding pocket terminates with the same frequency regardless of the addition of ppGpp. This finding demonstrates that transcription read-through requires ppGpp binding by the

riboswitch, rather than the binding of this signaling molecule to the RNA polymerase protein complex.

The concentration of ppGpp needed to achieve half-maximal termination (T_{50}) is 6 μM for the *D. hafniense ilvE* RNA (Figure 5-7C,D). This is nearly 1000-fold higher than the K_D value measured for direct binding of ppGpp to the *T. oceani* 112 *ilvE* RNA construct (Figure 5-2C). This difference might be partly due to the distinct sequences of the two ppGpp riboswitches, or to the different conditions in which the two types of assays are conducted. However, it is well established[31-34] that some riboswitches operate under kinetic control[31] (likely experienced during transcription termination assays), rather than reaching thermodynamic equilibrium (as attained during in-line probing assays). Riboswitches operating under kinetic control have been shown to require higher concentrations of ligand to trigger gene control compared to K_D values of the same constructs measured under thermodynamic equilibrium. In other words, K_D values measured in a test tube do not reflect the concentration of ligand naturally required to trigger large changes in gene expression because each riboswitch aptamer does not have time to reach thermodynamic equilibrium before RNA polymerase transcribes beyond the intrinsic terminator stem.

The amount of full-length transcript is quite high even when no ppGpp is present and the terminator is expected to form (Figure 5-7B). This is again likely due to kinetic aspects such as the speed of RNA polymerase, which can influence the amount of ligand needed to regulate riboswitch-mediated termination, and the absence of the transcription factor NusA, which can increase the dynamic range of transcription termination yields.[31,35] Also, the modest dynamic range of the *D. hafniense ilvE* riboswitch *in vitro* might be due in part to the fact that the run of U residues traditionally following a terminator stem is interrupted

with a C nucleotide. A construct wherein this C residue is altered to a U (M11) exhibits a modestly improved dynamic range (Figure 5-7C,D). Regardless of the factors affecting the dynamic range of the riboswitch when measured *in vitro*, similar dynamic ranges for other riboswitch classes examined by *in vitro* transcription assays have been observed previously.[36-38]

To further assess ppGpp as the natural ligand for this riboswitch class, I tested other guanine nucleotide compounds using *in vitro* transcription termination assays. Although pGp and GDP cause modulation of the RNA under in-line probing conditions, albeit with decreased affinity compared to ppGpp, neither is able to cause a shift from terminated to full-length product even at a concentration of 1 mM (Figure 5-7E). GTP was not tested because it is already present at 150 μM in the transcription reaction. These results demonstrate that these RNAs selectively regulate gene expression in response to ppGpp or pppGpp, but not any other guanine nucleotide derivative.

Gene associations of ppGpp riboswitches. The *in vitro* and *in vivo* gene control function of ppGpp riboswitch constructs described above are entirely consistent with the predominant gene associations observed for members of this riboswitch class. The link between branched-chain amino acid metabolism and ppGpp signaling is well established,[13-18] and my observations that ppGpp and pppGpp are bound by members of this riboswitch class and respond to changing ppGpp levels *in vivo* are consistent with this natural gene association.

However, ppGpp riboswitches are also commonly associated with members of two other gene classes, which are connections that were not established previously. The first includes genes encoding various glutamate synthase domains including *gltS*, *gltB*, *gltD* and

glxB. The second includes operons containing three ATP-binding cassette (ABC) transporters, one of which is predicted to be *natA* while the other two (COG3694 and COG4587) have unknown functions. Representative RNAs associated with glutamate synthase (Figure 5-8) and *natA* (Figure 5-9) gene classes bind ppGpp with dissociation constants similar to that observed for the *T. oceani* 112 *ilvE* riboswitch aptamer (Figure 5-2).

Organisms that use ppGpp riboswitches to control glutamate synthase genes typically also use ppGpp riboswitches to regulate BCAA biosynthesis-related gene expression, suggesting that these two metabolic processes are linked. Glutamate levels are known to be elevated during the stringent response,[39] and this riboswitch discovery provides a direct mechanism for upregulation of glutamate biosynthesis during this stress response. Moreover, the known connection between these biochemical responses reinforces my hypothesis that the natural ligand for some *ykkC* subtype 2 RNAs indeed is ppGpp and/or its immediate biosynthetic precursor pppGpp.

There was no prior established link between ppGpp signaling and the ABC transporters associated with ppGpp riboswitches. Organisms with ppGpp riboswitches upstream of these transporter genes are mutually exclusive to organisms with ppGpp riboswitches upstream of BCAA and glutamate synthesis genes. A possible explanation for this gene association pattern comes from the fact that one of these transporters shares homology with NatA, which has a predicted cation transport function and is known to be expressed under membrane stress conditions.[40] Various cell envelope stresses are also known to activate the *relP* ppGpp synthase gene in *B. subtilis*.[41-43] Therefore, it seems logical for these transporters, if they are indeed NatA homologs, to be expressed when ppGpp levels are

high, although in this case ppGpp production is not directly related to the stringent response.

Some ppGpp riboswitches reside in tandem with T Box RNAs. In ~40% of the instances when the ppGpp riboswitch is found in the 5′-UTR of BCAA biosynthesis genes, a leucine T box RNA is found either directly upstream or downstream of the riboswitch in the same UTR. T box RNAs are regulatory domains that control gene expression in response to binding certain uncharged tRNAs, thereby indirectly detecting a deficiency of an amino acid.[44] These two RNA elements presumably function independently because each has its own expression platform. When found in tandem, they function as a two-input Boolean AND logic gate, which would be similar to riboswitch logic gate architectures reported previously.[45] Specifically, both uncharged $tRNA^{Leu}$ and ppGpp would be necessary for transcription of the downstream protein coding regions.

This tandem arrangement likely exploits two required inputs to ensure that the expression of associated genes is reduced when nutrients are plentiful. Additionally, because RelA is allosterically regulated by any uncharged tRNA, the ppGpp riboswitch will turn on gene expression when any amino acid is limiting, whereas each T box RNA indirectly detects a limitation only of its cognate amino acid. At the promoter level, BCAA biosynthesis genes are regulated by the CodY transcriptional repressor, which uses GTP and isoleucine as regulatory inputs.[22,46] It is unclear whether all three of these mechanisms are utilized simultaneously because while organisms with ppGpp riboswitches typically have the *codY* gene, the canonical CodY binding site consensus sequence[46] was not found upstream of ppGpp riboswitch sequences.

Bioinformatic and biochemical data reveal additional riboswitch classes. All of the other RNAs I examined that were grouped with *ykkC* subtype 2 fail to bind either ppGpp or guanidine (data not shown), and therefore represent distinct riboswitch classes. I identified subtypes 2a through 2d by sorting RNA representatives based on their distinct gene associations, and then by comparative sequence analysis.[23] I also determined that some of these RNAs (subtype 2b) bind phosphoribosyl pyrophosphate (PRPP), a precursor of nucleotide biosynthesis (see Chapter 6). However, I speculate that at least two additional candidate riboswitch classes (subtypes 2c and 2d) whose ligands remain undiscovered are represented among the remaining *ykkC* RNA variants.

CONCLUSIONS

A riboswitch class for ppGpp has long been anticipated,[5,6,10,12] but representatives of this particular class were challenging to identify due to their striking similarity to riboswitches for guanidine.[23] At one point early in evolutionary history, ppGpp riboswitches might have been extremely widespread, but it appears that their function has been replaced by the CodY repressor protein factor in many organisms.

Notably, the discovery of ppGpp riboswitches in tandem with T box regulators serves as an additional demonstration of how complex molecular devices can be formed purely from RNA. The Boolean AND gate architecture described herein involves a tRNA sensed by a T box RNA and a riboswitch RNA that senses the RNA-derived alarmone ppGpp. This all-RNA system is similar to that for the two-input NOR logic gate formed by tandem riboswitches for *S*-adenosylmethionine and adenosylcobalamin.[45] Likewise, it has previously been demonstrated that a riboswitch located in tandem with a group I self-

splicing ribozyme functions as a two-input AND gate that requires both the signaling molecule c-di-GMP and the nucleotide GTP to properly splice and express a messenger RNA from *Clostridium difficile*.[8,47] Such complex all-RNA systems suggest that riboswitch-mediated responses to nucleotide-like signaling molecules such as ppGpp could be of ancient RNA World origin.

MATERIALS AND METHODS

Chemicals and reagents. Chemicals were purchased from Sigma-Aldrich with the exceptions of guanosine 5′,3′-bisdiphosphate (ppGpp), guanosine pentaphosphate (pppGpp) (Jena Bioscience), and guanosine 5′,3′-bisphosphate (pGp) (TriLink Biotech). [γ-^{32}P]-ATP was purchased from Perkin Elmer and used within two weeks of delivery. All bulk chemicals were purchased from J.T. Baker and enzymes were purchased from New England Biolabs, unless otherwise noted. All solutions were prepared using deionized water (dH_2O) and were either autoclaved or filter sterilized (using 0.22 μm filters, Millipore) prior to use. DNA oligonucleotides used in this chapter (Table 5-1) were purchased from Sigma-Aldrich and Integrated DNA Technologies. Strains of *B. subtilis* were obtained from the Bacillus Genetic Stock Center at Ohio State University.

Bioinformatics analyses. Additional examples of *ykkC* motif RNAs were identified using Infernal 1.1[48] to search RefSeq version 76 plus additional environmental microbial databases.[49] Iterative searches for new sequences were performed based on the previously published alignment of *ykkC* subtype 2 RNAs.[23] These sequences were sorted by nucleotide identity at a specific position as previously indicated[23] to exclude guanidine-I riboswitches, then further manually sorted by downstream gene association. This analysis identified 105

examples of subtype 2a RNAs (ppGpp riboswitches), of which 78 are upstream of BCAA metabolism genes, 15 are upstream of various glutamate synthase genes, and 12 are upstream of *natA* and other ABC transporter genes. It should be noted that this is an underestimate of the true number of representatives of this riboswitch class in this dataset, as examples with incomplete gene contexts were not evaluated. The consensus sequence and secondary structure model of these 105 ppGpp riboswitch RNAs was constructed using R2R software.[50]

RNA oligonucleotide preparation. Preparation of RNAs for use in in-line probing experiments was performed as previously described[23,51] with the following modifications. All double-stranded DNA (dsDNA) templates used for RNA transcription were assembled via extension of overlapping synthetic DNAs (Table 5-1) using SuperScript II reverse transcriptase (Thermo Fisher Scientific). Forward primers contain the T7 RNA polymerase (T7 RNAP) promoter followed by guanosine residues (when not present in the wild-type sequence) to enable transcription. Thereafter, RNA constructs were prepared from these dsDNAs by *in vitro* transcription, purified, and subsequently 5′ ^{32}P-labeled as previously described.[23,51]

RNA in-line probing analyses. In-line probing assays were performed as previously described.[28,29] The values for the amount of RNA aptamer bound to ligand or "fraction bound" were estimated based on the ligand-mediate changes in the band intensities from in-line probing assays also as reported previously.[23,51] Briefly, quantified changes in band intensities at designated sites reflect the fraction of RNAs that are bound to ligand, such that half-maximal increases or decreases in band intensities represent the K_D for the RNA-ligand interaction.

Single-round transcription termination assays. Template dsDNA constructs for single-round *in vitro* transcription were designed to include the aptamer and expression platform of the *D. hafniense* riboswitch starting at the predicted natural transcription start site and ending 28 nucleotides beyond the terminator stem. The promoter used to enable transcription was derived from the *B. subtilis lysC* gene, which is compatible with *E. coli* RNA polymerase. The described WT and mutant templates were purchased as synthetic double-stranded DNAs and subsequently amplified by PCR. Approximately 2 pmol of the resulting, purified DNA template was added to a transcription initiation mixture (20 mM Tris-HCl [pH 7.5 at 23ºC], 75 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 10 µg mL^{-1} bovine serum albumin [BSA], 130 µM ApA dinucleotide, 1% glycerol, 0.04 U $µL^{-1}$ *E. coli* RNA polymerase holoenzyme, 2.5 µM GTP, 2.5 µM ATP, and 1 µM UTP). Approximately 1 µCi [α-^{32}P]-UTP was added per 8 µL reaction and transcription was allowed to proceed at 37ºC for 10 minutes, leading to formation of a stalled polymerase complex at the first cytidine nucleotide of each transcript. For each 8 µL transcription reaction, 1 µL of 10x elongation buffer (20 mM Tris-HCl [pH 7.5 at 23ºC], 75 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 2 mg mL^{-1} heparin, 1.5 mM ATP, 1.5 mM GTP, 1.5 mM CTP, and 250 µM UTP) as well as 1 µL of a 10x solution of the ligand of interest were added sequentially. Because ppGpp is a known inhibitor of *E. coli* RNA polymerase initiation,[16] the holoenzyme was first allowed to assemble into a stalled, stable ternary complex before ppGpp was added concurrently with the elongation mixture. Transcription reactions were then incubated at 37ºC for an additional 60 minutes.

The transcription products were subsequently analyzed via denaturing 10% PAGE and visualized using a phosphorimager (GE Healthcare Life Sciences). Fraction full length

values were calculated by varying the ligand concentration in separate reactions and quantifying the changes in band intensity of both full length (FL) and terminated (T) transcription products using the formula (FL)/(FL+T). The T_{50} values were determined by plotting the fraction FL as a function of the logarithm of ligand concentration and using a sigmoidal four parameter logistic fit in GraphPad Prism 7.

ACKNOWLEDGEMENTS

We thank Shira Stav, Adam Roth, Kimberly Harris, and other members of the Breaker laboratory for helpful discussions. M.E.S. was supported by an NIH Cellular and Molecular Biology Training Grant (T32GM007223). This work was also supported by NIH grants to R.R.B. (GM022778 and DE022340). R.R.B. is also supported by the Howard Hughes Medical Institute.

REFERENCES

(1) Roth, A., Breaker, R.R. (2009) The structural and functional diversity of metabolite-binding riboswitches. *Annu Rev Biochem 78*, 305-334.

(2) Serganov, A., Nudler, E. (2013) A decade of riboswitches. *Cell 152*, 17-24.

(3) Sherwood, A.V., Henkin, T.M. (2016) Riboswitch-mediated gene regulation: Novel RNA architectures dictate gene expression responses. *Annu Rev Microbiol 70*, 361-374.

(4) McCown, P.J., Corbino, K.A., Stav, S., Sherlock, M.E., Breaker, R.R. (2017) Riboswitch diversity and distribution. *RNA 23*, 995-1011.

(5) Nelson, J.W., Breaker, R.R. (2017) The lost language of the RNA World. *Sci Signal 10*, 1-10.

(6) Breaker, R.R. (2010) Riboswitches: From ancient gene control systems to modern drug targets. *Microbe 5*, 13-20.

(7) Sudarsan, N., Lee, E.R., Weinberg, Z., Moy, R.H., Kim, J.N., Link, K.H., Breaker, R.R. (2008) Riboswitches in eubacteia sense the second messenger cyclic di-GMP. *Science 321,* 411-413.

(8) Lee, E.R., Baker, J.L., Weinberg, Z., Sudarsan, N., Breaker, R.R. (2010) An allosteric self-splicing ribozyme triggered by a bacterial second messenger. *Science 329*, 845-848.

(9) Nelson, J.W., Sudarsan, N., Furukawa, K., Weinberg, Z., Wang, J.X., Breaker, R.R. (2013) Riboswitches in eubacteria that sense the second messenger c-di-AMP. *Nat Chem Biol 9,* 834-839.

(10) Kim, P.B., Nelson, J.W., Breaker, R.R. (2015) An ancient riboswitch class in bacteria regulates purine biosynthesis and one-carbon metabolism. *Mol Cell 57*, 317-328.

(11) Kellenberger, C.A., Wilson, S.C., Hickey, S.F., Gonzalez, T.L., Su, Y., Hallberg, Z.F., Brewer, T.F., Iavarone, A.T., Carlson, H.K., Hsieh, Y.F., Hammond, M.C. (2015) GEMM-I riboswitches from *Geobacter* sense the bacterial second messenger cyclic AMP-GMP. *Proc Natl Acad Sci USA* 112, 5383-5388.

(12) Nelson, J.W., Sudarsan, N., Phillip, G.E., Stav, S., Lünse, C.E., McCown, P.J., Breaker, R.R. (2015) Control of bacterial exoelectrogenesis by c-AMP-GMP. *Proc Natl Acad Sci USA 112,* 5389-5394.

(13) Cashel, M. (1969) The control of ribonucleic acid synthesis in *Escherichia coli*. *J Biol Chem 244*, 3133-3141.

(14) Potrykus, K., Cashel, M. (2008) (p)ppGpp: still magical? *Annu Rev Microbiol 62*, 35-51.

(15) Dalebroux, Z.D., Swanson, M.S. (2012) ppGpp: magic beyond RNA polymerase. *Nat Rev Microbiol 10*, 203-212.

(16) Hauryliuk, V., Atkinson, G.C., Murakami, K.S., Tenson, T., Gerdes, K. (2015) Recent functional insights into the role of (p)ppGpp in bacterial physiology. *Nat Rev Microbiol 13*, 298-309.

(17) Cashel, M., Rudd, K. (1987) In *Escherichia coli and Salmonella typhimurium*: *Cellular and Molecular Biology*, (ed. Neidhardt F. C.) 1410-1438 (American Society for Microbiology Publishers, Washington DC, USA).

(18) Traxler, M.F., Summers, S.M., Nguyen, H.T., Zacharia, V.M., Hightower, G.A., Smith, J.T., Conway, T. (2008) The global, ppGpp-mediated stringent response to amino acid starvation in *Escherichia coli*. *Mol Microbiol 68*, 1128-1148.

(19) Shivers, R.P., Sonenshein, A.L. (2004) Activation of the *Bacillus subtilis* global regulator CodY by direct interaction with branched-chain amino acids. *Mol Micriobiol 53*, 599-611.

(20) Krásný, L., Tiserová, H., Jonák, J., Rejman, D., Sanderová, H. (2008) The identity of the transcription +1 position is crucial for changes in gene expression in response to amino acid starvation in *Bacillus subtilis*. *Mol Microbiol 69,* 42–54.

(21) Wolz, C., Geiger, T., Goerke, C. (2010) The synthesis and function of the alarmone (p)ppGpp in firmicutes. *Int J Med Microbiol 300*, 142-147.

(22) Sonenshein, A.L. (2005) CodY, a global regulator of stationary phase and virulence in gram- positive bacteria. *Curr Op Microbiol 8*, 203-207.

(23) Nelson, J.W., Atilho, R.M., Sherlock, M.E., Stockbridge, R.B., Breaker, R.R. (2017) Metabolism of free guanidine in bacteria is regulated by a widespread riboswitch class. *Mol Cell* 65, 220-230.

(24) Barrick, J.E., Corbino, K.A., Winkler, W.C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I. (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc Natl Acad Sci USA 101,* 6421-6426.

(25) Reiss, C.W., Xiong, Y., Strobel, S.A. (2017) Structural basis for ligand binding to the guanidine-I riboswitch. *Structure 25*, 195-202.

(26) Battaglia, R.A., Price, I.R., Ke, A. (2017) Structural basis for guanidine sensing by the *ykkC* family of riboswitches. *RNA 23*, 578-585.

(27) Meyer, M.M., Hammond, M.C., Salinas, Y., Roth, A., Sudarsan, N., Breaker, R.R. (2011) Challenges of ligand identification for riboswitch candidates. *RNA Biol 8,* 5-10.

(28) Soukup, G.A., Breaker, R.R. (1999) Relationship between internucleotide linkage geometry and the stability of RNA. *RNA 5*, 1308-1325.

(29) Regulski, E.E., Breaker, R.R. (2008) In-line probing analysis of riboswitches. *Methods Mol Biol 419*, 53-67.

(30) Landick, R., Wang, D., Chan, C.L. (1996) Quantitative analysis of transcriptional pausing by *Escherichia coli* RNA polymerase: *his* leader pause site as a paradigm. *Methods Enzymol 274*, 334-353.

(31) Wickiser, J.K., Winkler, W.C., Breaker, R.R., Crothers, D.M. (2005) The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Mol Cell 18*, 49-60.

(32) Wickiser, J.K., Cheah, M.T., Breaker, R.R., Crothers, D.M. (2005) The kinetics of ligand binding by an adenine-sensing riboswitch. *Biochemistry 44*, 13404-13414.

(33) Gilbert, S.D., Stoddard, C.D., Wise, S.J., Batey, R.T. (2006) Thermodynamic and kinetic characterization of ligand binding to the purine riboswitch aptamer domain. *J Mol Biol 359*, 754-768.

(34) Haller, A., Soulière, M.F., Micura, R. (2011) The dynamic nature of RNA as a key to understanding riboswitch mechanisms. *Acc Chem Res 44*, 1339-1348.

(35) Lemay, J.F., Desnoyers, G., Blouin, S., Heppell, B., Bastet, L., St-Pierre, P., Massé, E., Lafontaine, D.A. (2011) Comparative study between transcriptionally- and translationally-acting adenine riboswitches reveals key differences in riboswitch regulatory mechanisms. *PLoS Genetics 7*, e1001278.

(36) Winkler, W.C., Cohen-Chalamish, S., Breaker, R.R. (2002) An mRNA structure that controls gene expression by binding FMN. *Proc Natl Acad Sci USA 99*, 15908-15913.

(37) McDaniel, B.A.M., Grundy, F.J., Artsimovitch, I., Henkin, T.M. (2003) Transcription termination control of the S box system: Direct measurement of *S*-adenosylmethionine by the leader RNA. *Proc Natl Acad Sci USA 100*, 3083-3088.

(38) Winkler, W.C., Nahvi, A., Sudarsan, N., Barrick, J.E., Breaker, R.R. (2003) An mRNA structure that controls gene expression by binding *S*-adenosylmethionine. *Nat Struct Mol Biol 10*, 701-707.

(39) Imaizumi, A., Kojima, H., Matsui, K. (2006) The effect of intracellular ppGpp levels on glutamate and lysine overproduction in *Escherichia coli*. *J Biotechnol 125*, 328-337.

(40) Cheng, J., Guffanti, A.A., Krulwich, T.A. (1997) A two-gene ABC-type transport system that extrudes Na^+ in *Bacillus subtilis* is induced by ethanol or protonophore. *Mol Microbiol 23*, 1107-1120.

(41) Nanamiya, H., Kasai, K., Nozawa, A., Yun, C.S., Narisawa, T., Murakami, K., Natori, Y., Kawamura, F., Tozawa, Y. (2008) Identification and functional analysis of novel (p)ppGpp synthetase genes in *Bacillus subtilis*. *Mol Microbiol 67*, 291–304.

(42) Geiger, T., Wolz, C. (2014) Intersection of the stringent response and the CodY regulon in low GC gram-positive bacteria. *Int J Med Microbiol 304*, 150-155.

(43) Geiger, T., Kästle, B., Gratani, F.L., Goerke, C., Wolz, C. (2014) Two Small (p)ppGpp synthases in *Staphylococcus aureus* mediate tolerance against cell envelope stress conditions. *J Bacteriol 196*, 894–902.

(44) Green, N.J., Grundy, F.J., Henkin, T.M. (2010) The T box mechanism: tRNA as a regulatory molecule. *FEBS Lett 584*, 318-324.

(45) Sudarsan, N., Hammond, M.C., Block, K.F., Welz, R., Barrick, J.E., Roth, A., Breaker, R.R. (2006) Tandem riboswitch architectures exhibit complex gene control functions. *Science 13*, 300-304.

(46) den Hengst, C.D., van Hijum, S.A., Geurts, J.M., Nauta, A., Kok, J., Kuipers, O.P. (2005) The *Lactococcus lactis* CodY regulon: identification of a conserved cis-regulatory element. *J Biol Chem 280*, 34332-34342.

(47) Chen, A.G., Sudarsan, N., Breaker, R.R. (2011) Mechanism for gene control by a natural allosteric group I ribozyme. *RNA 17*, 1967-1972.

(48) Nawrocki, E.P., Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics 29*, 2933-2935.

(49) Weinberg, Z., Kim, P.B., Chen, T.H., Li, S., Harris, K.A., Lünse, C.E., Breaker, R.R. (2015) New classes of self-cleaving ribozymes revealed by comparative genomics analysis. *Nat Chem Biol 11*, 606-610.

(50) Weinberg, Z., Breaker, R.R. (2011) R2R - software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinfo 12*, 3.

(51) Sherlock, M.E., Malkowski, S.N., Breaker, R.R. (2017) Biochemical validation of a second guanidine riboswitch class. *Biochemistry 56*, 352-358.

FIGURES

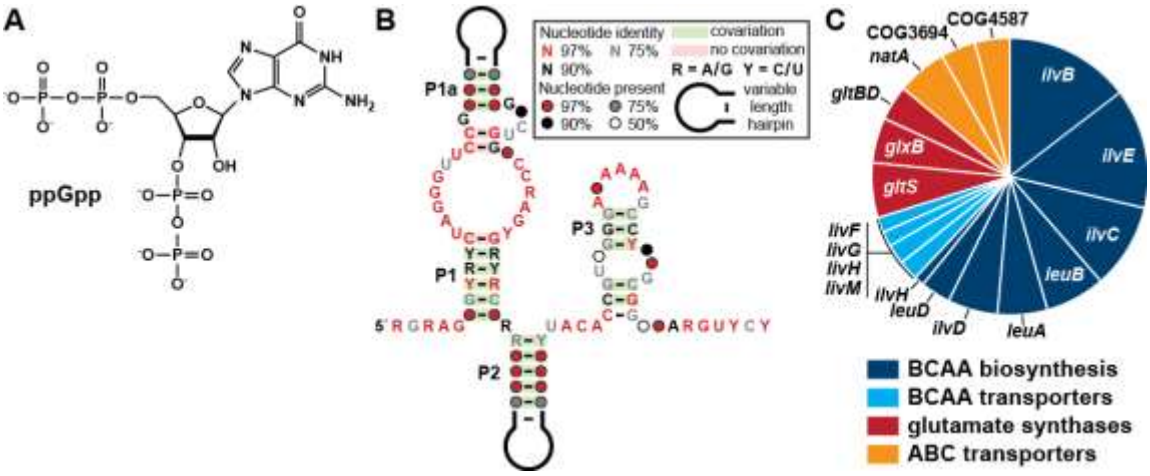

Figure 5-1. ppGpp riboswitches and their regulatory network. (A) The chemical structure of ppGpp (guanosine-3′,5′-bisdiphosphate). (B) Consensus sequence and secondary structure model for ppGpp riboswitches derived from 115 unique representatives. (C) Genes predicted to be regulated by ppGpp riboswitches. I performed bioinformatic analyses used to create this figure.

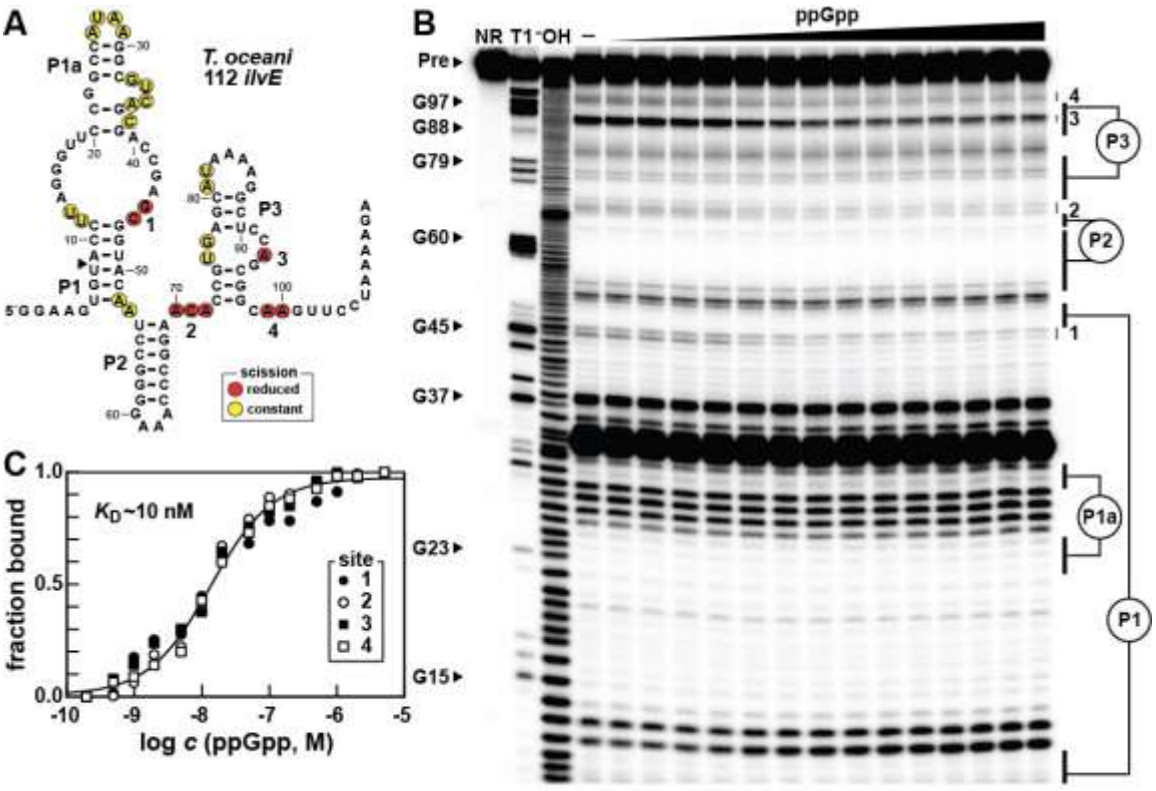

Figure 5-2. Ligand binding by a ppGpp riboswitch aptamer. (A) Sequence and secondary structure of the 112 nucleotide RNA derived from the *ilvE* gene of *T. oceani*. Data collected in panel B were used to determine regions of constant or reduced scission upon the addition of ppGpp to in-line probing reactions. (B) Polyacrylamide gel electrophoresis (PAGE) analysis of the products of in-line probing reactions with 5′-^{32}P-labeled 112 *ilvE* RNA in the presence of increasing ppGpp concentrations. NR, T1 and ⁻OH represent RNA undergoing no reaction, partial digest with RNase T1 (cleaves after guanosine nucleotides), or partial digest under alkaline conditions (cleaves after every nucleotide), respectively. Selected bands corresponding to RNA strand scission after G residues are annotated. In-line probing reactions contained either no ligand (–) or ppGpp ranging from 200 pM to 5 μM. Regions 1-4 of the RNA exhibit structural stabilization in a ppGpp-dependent manner. (C) Plot of the fraction of RNA bound by ppGpp, as

determined by the normalized fraction of RNA scission at regions 1-4 as a function of the logarithm of ppGpp concentration (c). The dissociation constant (K_D) was estimated by a sigmoidal fit of the average of the data points from all regions of RNA structural modulation using a procedure described in the Materials and Methods section. Data shown are representative of at least three separate in-line probing experiments. I performed the experiments shown in this figure.

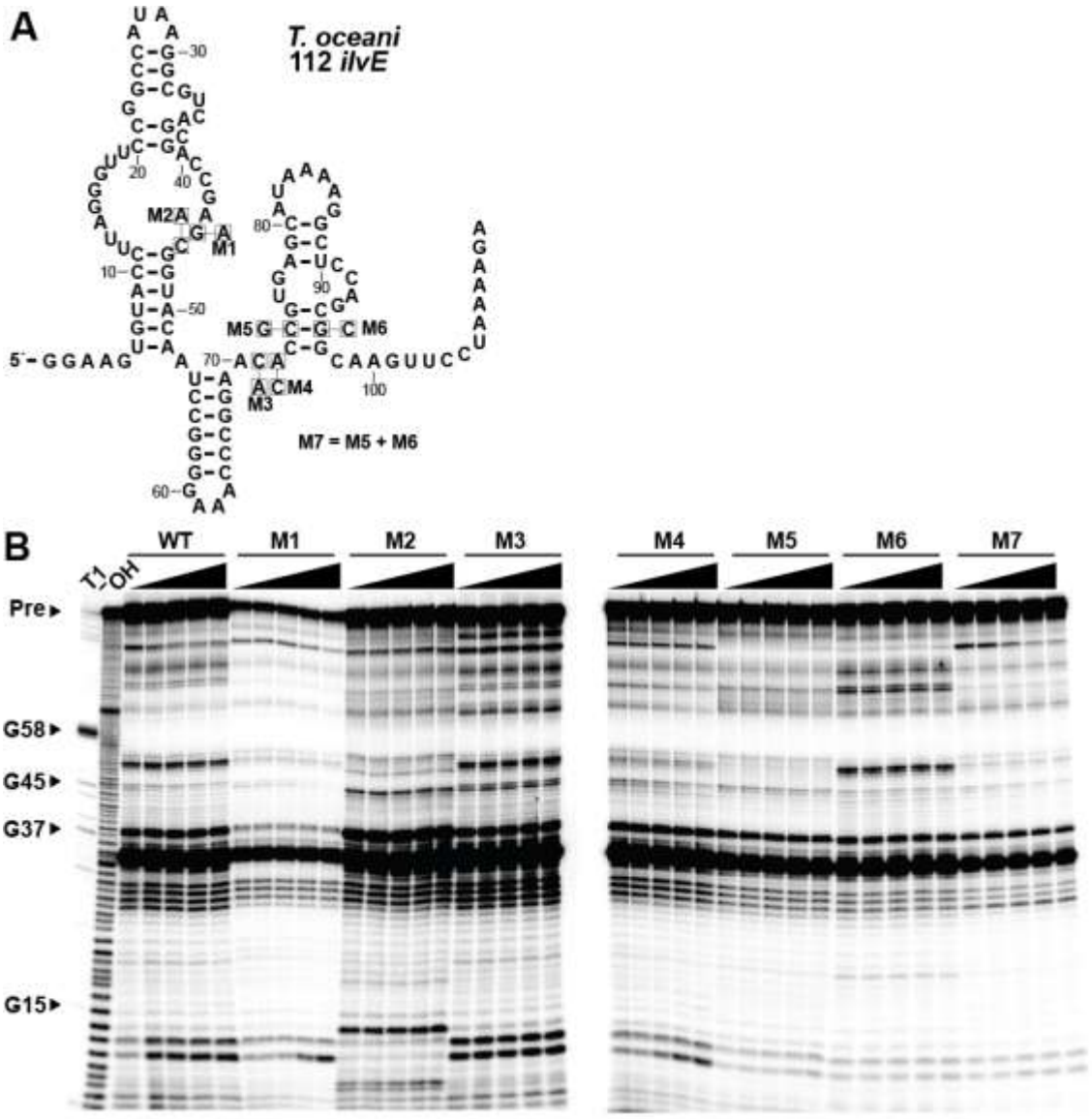

Figure 5-3. Mutation of conserved nucleotides disrupt ppGpp binding. (A) Sequence and secondary structure of the 112 *ilvE* RNA from *T. oceani*. Single nucleotide positions that were altered to produce each mutant RNA are identified with boxes, annotated with the nucleotide identity of each mutation, and labeled M1 through M6. M7 contains mutations at both positions specified by M5 and M6. (B) PAGE analysis of in-line probing assays with WT and mutated sequences of the 112 *ilvE* RNA in the presence of 0, 100 nM, 1 µM, 10 µM, or 100 µM ppGpp. Additional annotations are as described in the legend to

Figure 5-2B. Note that, for the mutant constructs, either no modulation (M2, M3, M5, M6) or very little modulation (M1, M4) is observed only at the highest ppGpp concentrations tested (10 μM and100 μM). Because the K_D measured for the WT construct is 10 nM, this means that most mutant constructs experience at least three orders of magnitude loss of affinity. I performed the experiments shown in this figure.

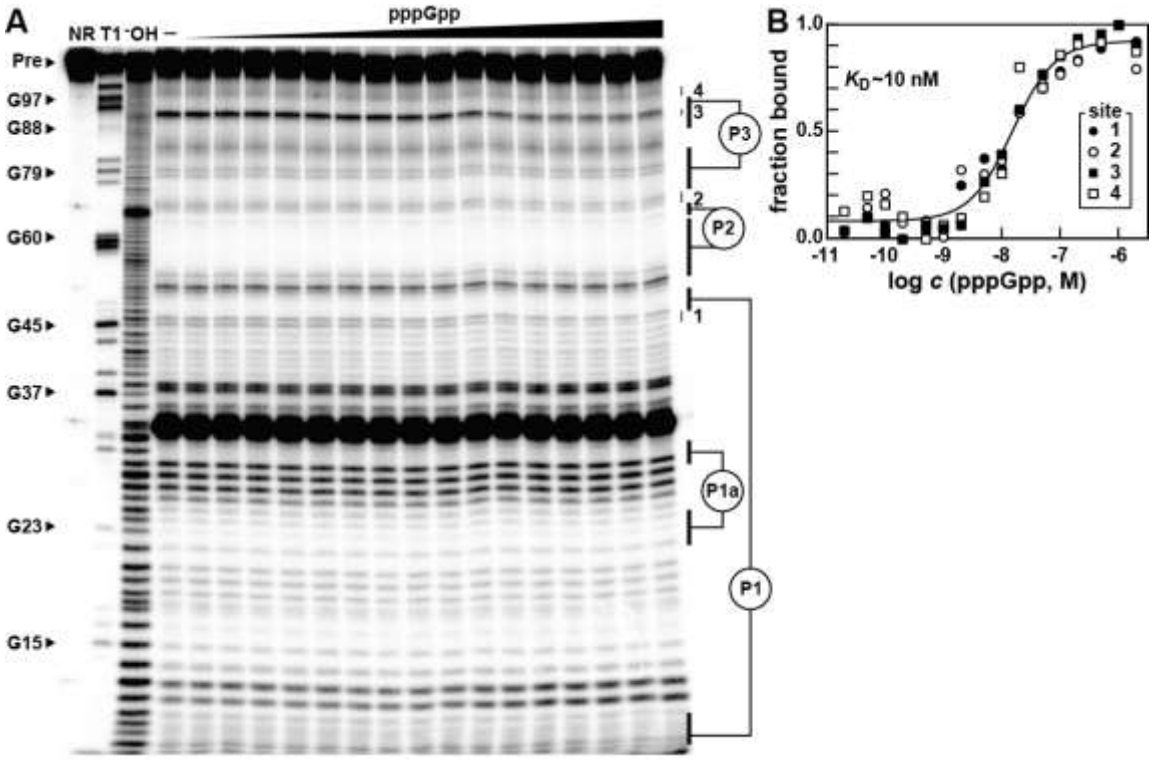

Figure 5-4. Recognition of pppGpp by a riboswitch aptamer. (A) PAGE analysis of the products of in-line probing of 5′-^{32}P-labeled 112 *ilvE* RNA in the presence of increasing pppGpp concentrations. In-line probing experiments contained either no ligand (–) or pppGpp ranging from 20 pM to 2 µM. Annotations are as described for Figure 5-2B. (B) Plot of the fraction of RNA bound by pppGpp, as determined by the normalized fraction of RNA scission at regions 1 through 4 as a function of the logarithm of pppGpp concentration. Additional details are as describe in the legend to Figure 5-2C. I performed the experiments shown in this figure.

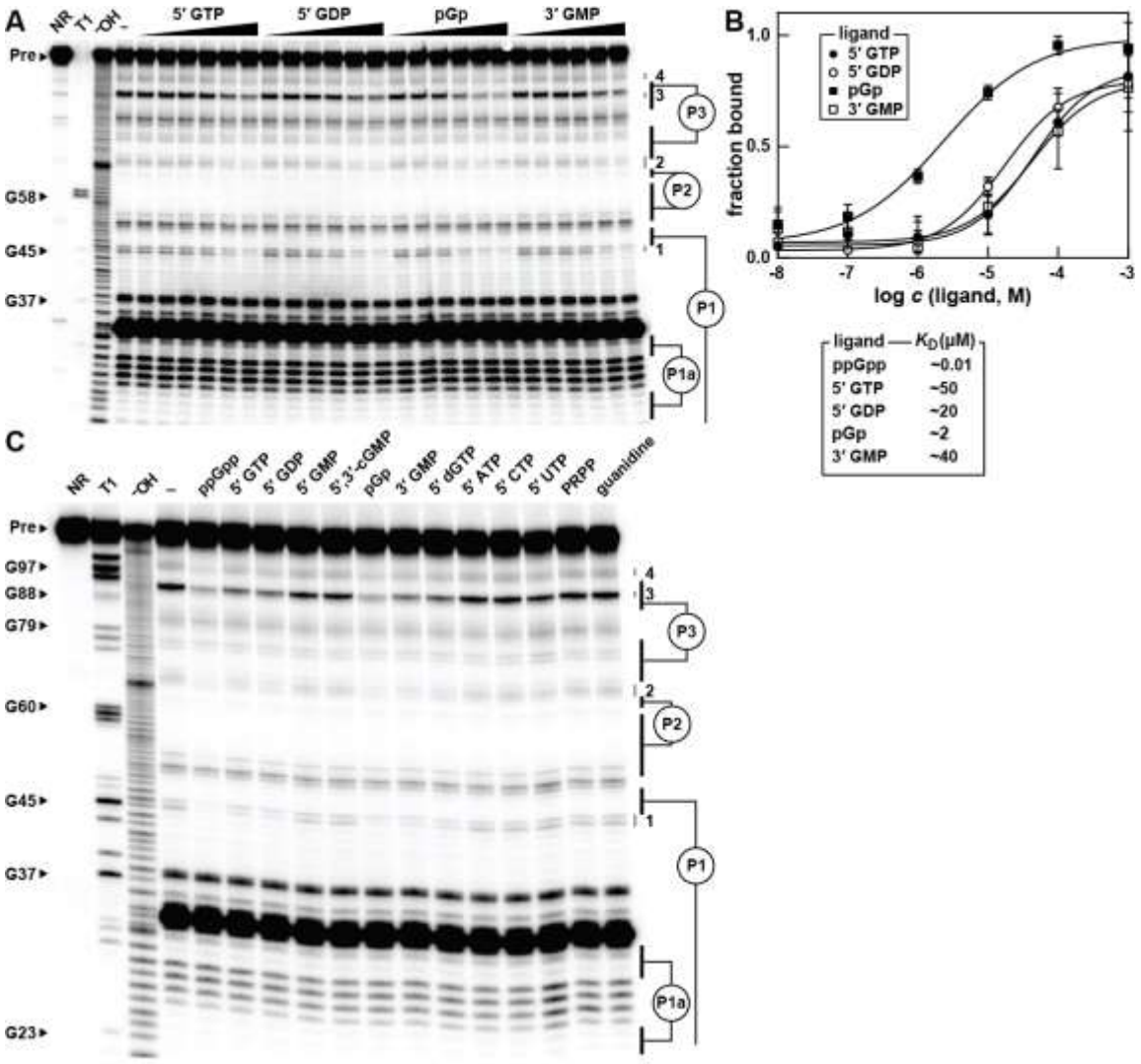

Figure 5-5. Binding selectivity of a representative ppGpp riboswitch. (A) PAGE analysis of in-line probing reactions of the 112 *ilvE* aptamer from *T. oceani* (see Figure 5-2A) in the presence of structural analogs of ppGpp that exhibited structural modulation. Each compound was tested at 10 nM, 100 nM, 1 μM, 10 μM, 100 μM, and 1 mM. Annotations are as described for Figure 5-2B. (B) Plot of the fraction of RNA bound by each ligand, as determined by the average and standard deviation of the normalized fraction of RNA scission at regions 1-4 as a function of the logarithm of ligand concentration. Additional details are as describe in the legend to Figure 5-2C. (C) PAGE analysis of in-

line probing reactions of the 112 *ilvE* aptamer in the presence of 100 μM concentration of various analogs of ppGpp, nucleotides, or the known ligands of other *ykkC* motif RNA subtypes (PRPP and guanidine). I performed the experiments shown in this figure.

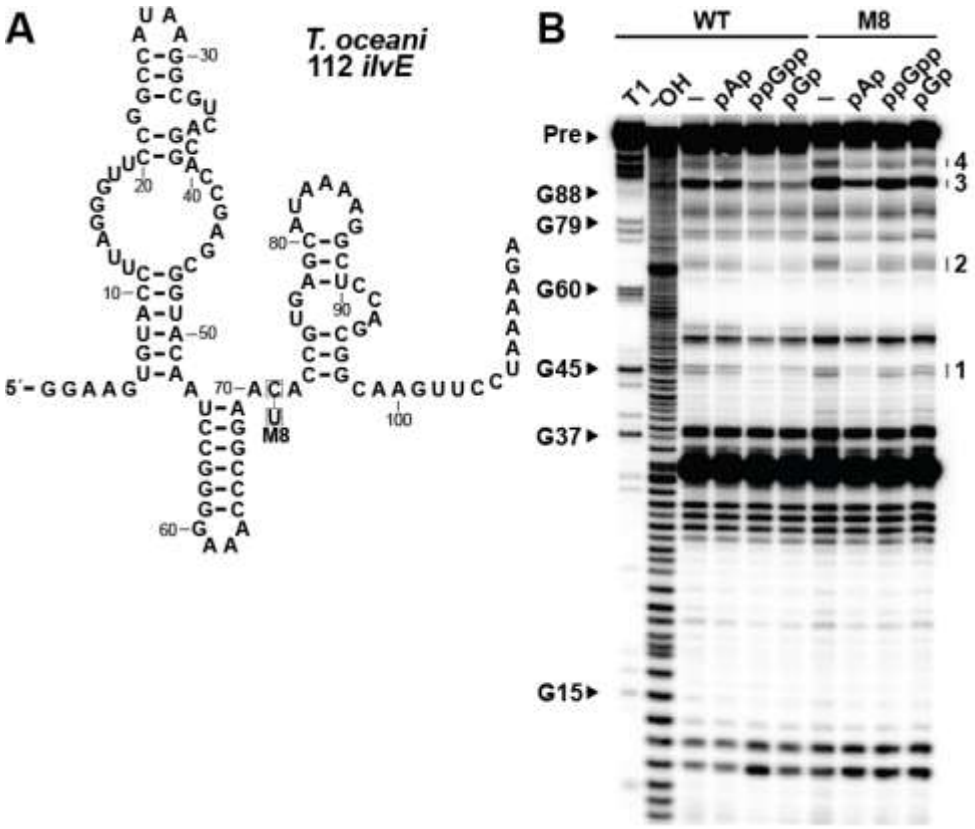

Figure 5-6. Mutation of a conserved nucleotide changes ligand specificity. (A) Sequence and secondary structure of the 112 *ilvE* RNA from *T. oceani* with the M8 RNA identified with a box and annotated with the nucleotide identity of the mutation. (B) PAGE analysis of in-line probing assays with WT and M8 112 *ilvE* RNA in the absence (–) or presence of 1 mM pAp, ppGpp, or pGp. Additional annotations are as described in the legend to Figure 5-2B. I performed the experiments shown in this figure.

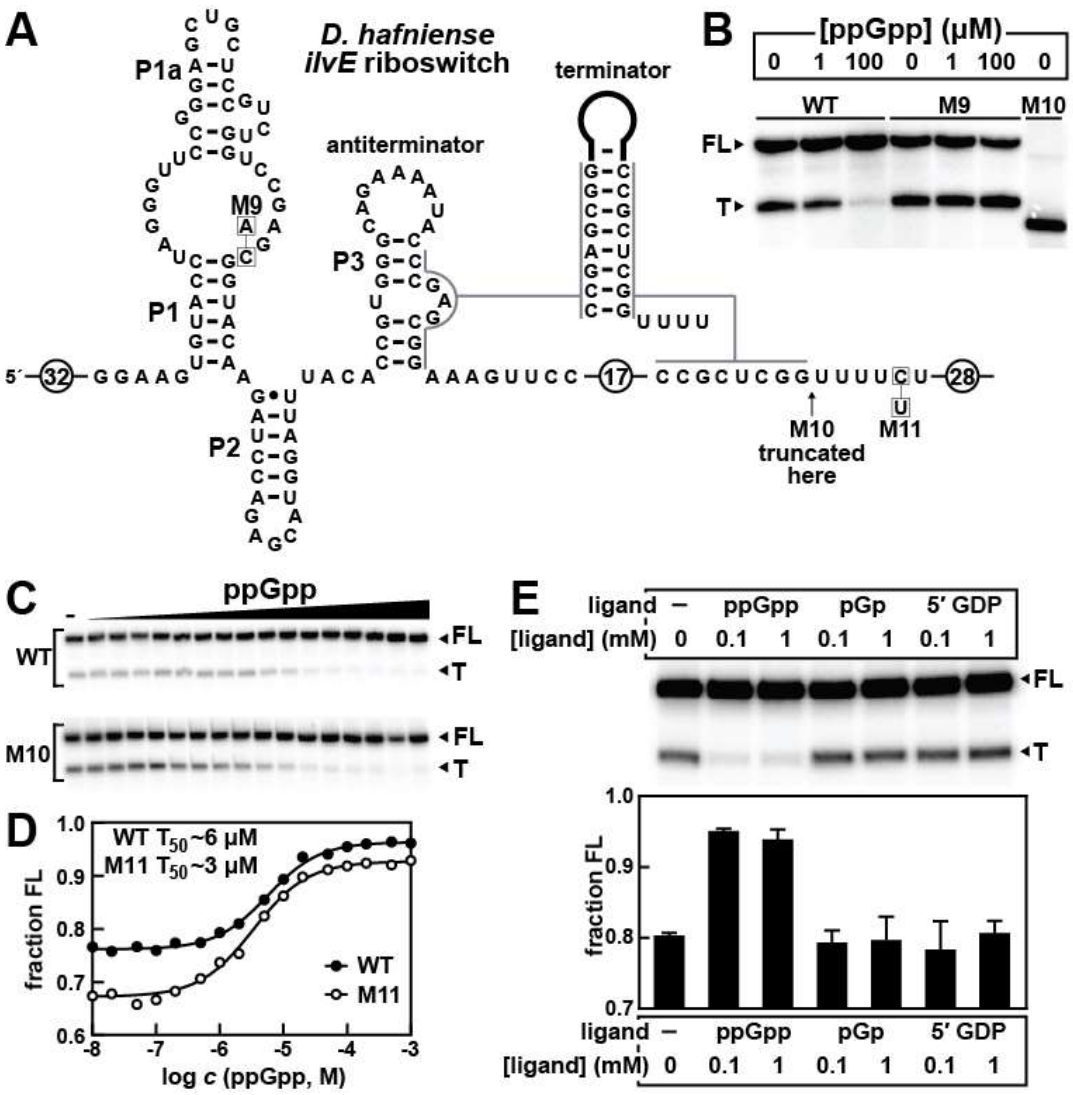

Figure 5-7. ppGpp regulates riboswitch-dependent transcript termination. (A) Sequence and secondary structure of the ppGpp riboswitch derived from the *ilvE* gene of *D. hafniense*. An alternative 'terminator' RNA structure, which only forms in the absence of ppGpp, is depicted wherein a hairpin forms followed by a U-rich tract. Note that the constructs used for *in vivo* reporter gene expression are longer than those used for *in vitro* transcription termination assays. These longer constructs ensure that the entire natural mRNA leader sequence is fused to the coding region of the *lacZ* reporter gene. (B) PAGE analysis of single-round transcription termination assays with WT, M9 and M10 *D.*

hafniense ilvE riboswitches conducted in the absence or presence of ppGpp as indicated. Construct sequences are as depicted in panel A. FL and T denote full length and terminated transcripts, respectively. Construct M10 provides a size marker for a transcript that terminates several nucleotides before the product generated when the intrinsic terminator forms. (C) PAGE analysis of single-round transcription termination assays of the WT and M11 *D. hafniense ilvE* RNAs with either no ligand (–) or ppGpp ranging from 10 nM to 1 mM. (D) Plot of the fraction of full length WT or M11 *D. hafniense ilvE* riboswitch transcripts contributing to the total number of transcripts (FL plus T) as a function of the logarithm of ppGpp molar concentration. The concentration of ppGpp required to cause half-maximal change in termination (T_{50}) was determined by a sigmoidal curve fit. (E) (Top) PAGE analysis of single-round transcription termination assays of the WT *D. hafniense ilvE* RNA with either no ligand (–) or each ligand as indicated at a concentration of 0.1 or 1 mM. Data are representative of multiple trials. (Bottom) Plot of the fraction of full length WT *D. hafniense ilvE* riboswitch transcripts contributing to the total number of transcripts (FL plus T) in the absence of ligand (–) or with each ligand as indicated. Values reported are the average and standard deviation of three separate trials, including the PAGE image depicted above. I performed the experiments shown in this figure.

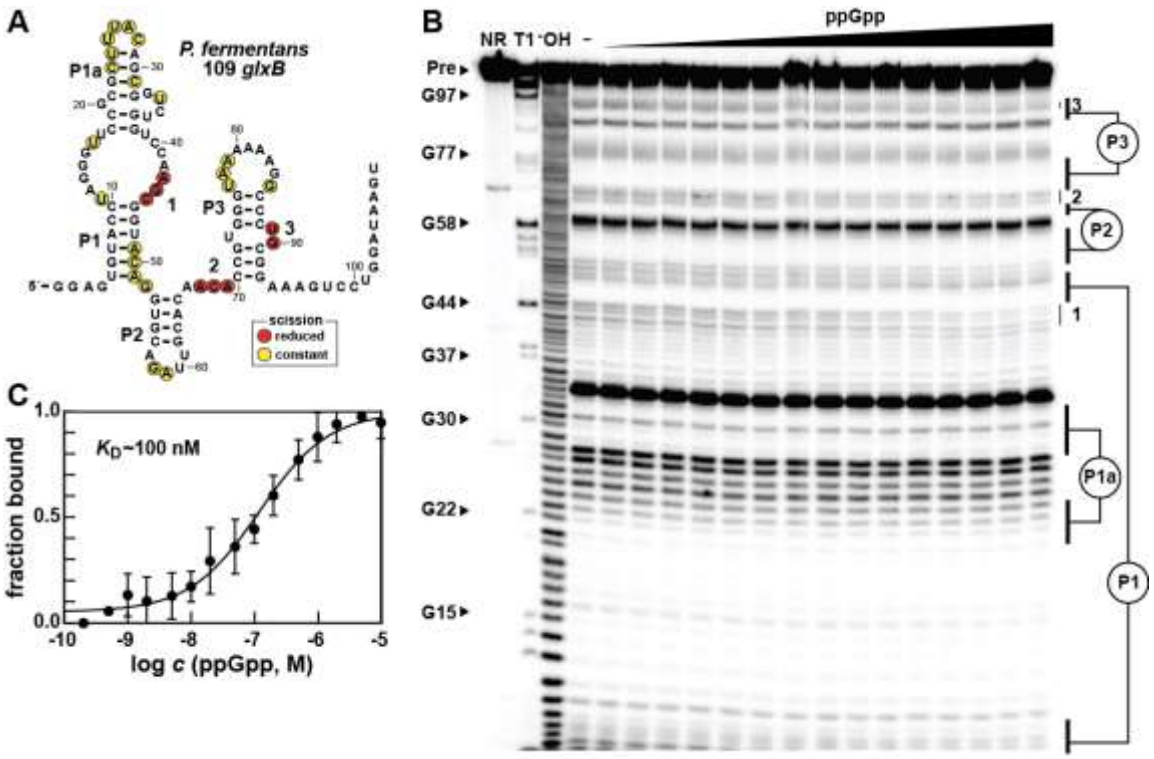

Figure 5-8. Ligand binding by an additional ppGpp riboswitch aptamer representative found upstream of glutamate synthase genes. (A) Sequence and secondary structure of the 109 nucleotide RNA derived from the *glxB* gene of *Pelosinus fermentans*. Data collected in panel B were used to determine regions of constant or reduced scission upon the addition of ppGpp to in-line probing assays. (B) PAGE analysis of the products of in-line probing of 5′-^{32}P-labeled 109 *glxB* RNA in the presence of increasing ppGpp concentrations. Annotations are as described for Figure 5-2B. In-line probing experiments contained either no ligand (–) or ppGpp ranging from 200 pM to 10 µM. Regions 1-3 of the RNA exhibit structural stabilization in a ppGpp-dependent manner. Note that this banding pattern and the pattern of changes resulting from in-line probing assay reactions are not identical to those observed for the original *T. oceani ilvE* construct presented in Figure 5-2. The distinct nucleotide sequences of these representatives will

exhibit patterns specific to these individual constructs. (C) Plot of the fraction of 109 *glxB* RNA bound to ligand as a function of the logarithm of ppGpp concentration. The fraction bound value plotted at each ppGpp concentration is the average of the values derived from band intensities at regions 1, 2 and 3 after normalization to a value between 1 (band intensity in the absence of ppGpp) and 0 (band intensity at maximum ppGpp concentration). The fraction bound value equals 1 minus the band intensity value. Error bars indicate standard deviation. Additional details are as describe in the legend to Figure 5-2C. I performed the experiments shown in this figure.

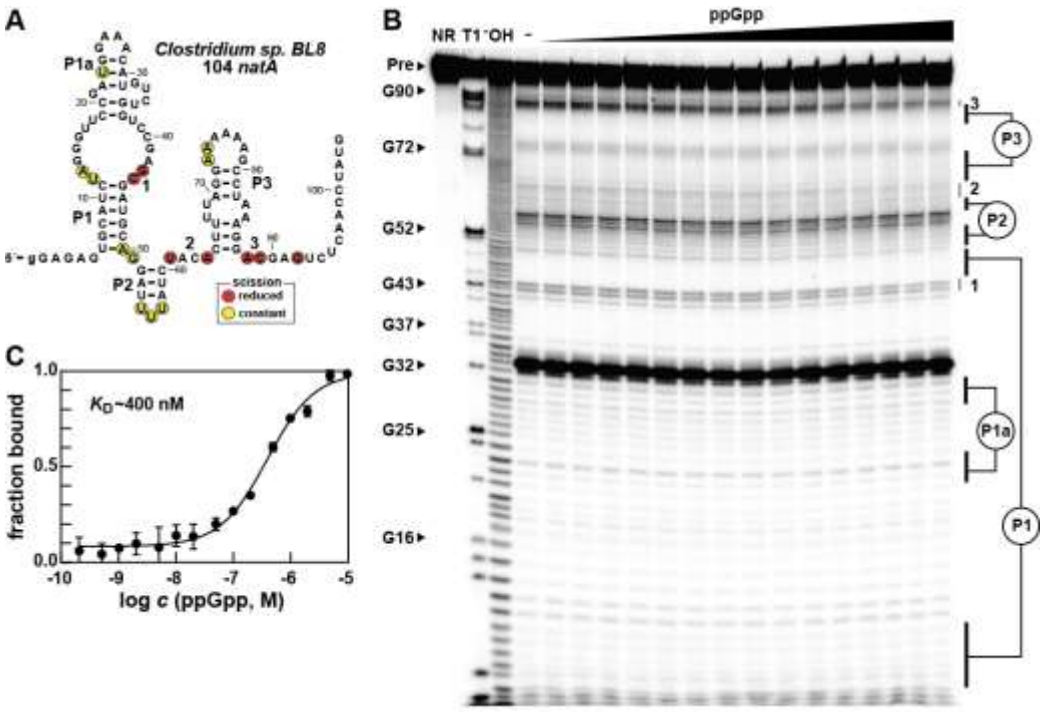

Figure 5-9. Ligand binding by an additional ppGpp riboswitch aptamer representative found upstream of *natA*. (A) Sequence and secondary structure of the 104 nucleotide RNA derived from the *natA* gene of *Clostridium sp. BL8*. Data collected in panel B were used to determine regions of constant or reduced scission upon the addition of ppGpp to in-line probing assays. (B) PAGE analysis of the products of in-line probing of 5′-^{32}P-labeled 104 *natA* RNA in the presence of increasing ppGpp concentrations. Annotations are as described for Figure 2B. In-line probing experiments contained either no ligand (–) or ppGpp ranging from 200 pM to 10 µM. Regions 1-3 of the RNA exhibit structural stabilization in a ppGpp-dependent manner. Note that this banding pattern and the pattern of changes resulting from in-line probing assay reactions are not identical to those observed for the original *T. oceani ilvE* construct presented in Figure 5-2. The distinct nucleotide sequences of these representatives will exhibit patterns specific to these individual constructs. (C) Plot of the fraction of 104 *natA* RNA bound to ligand as a

function of the logarithm of ppGpp concentration. The fraction bound value plotted at each ppGpp concentration was derived as described in the legend to Figure 5-8. Additional details are as describe in the legend to Figure 5-2C. I performed the experiments shown in this figure.

TABLES

Table 5-1. Sequences of synthetic DNAs used in Chapter 5.

Name	Sequence	Annotation
MES137	TAATACGACTCACTATAGGAAGTGTACCTTAGGGTTCCGGCCATAAGGCGTCAGCGACCGAGCGGTACAATCCGGGG	Forward template for transcription of the WT *T. oceani* 112 *ilvE* RNA
MES138	TCTTTTAGGAACTTGCCGCTGGAGCCTTTTATGCTCACGGTGTTCCGGGTTTCCCCGGATTGTACCGCTCGG	Reverse template for transcription of the WT *T. oceani* 112 *ilvE* RNA
MES177	TAATACGACTCACTATAGGAAGTGTACCTTAGGGTTCCGGCCATAAGGCGTCAGCGACCGAACGGTACAATCCGGGG	Forward template for transcription of the M1 *T. oceani* 112 *ilvE* RNA
MES178	TCTTTTAGGAACTTGCCGCTGGAGCCTTTTATGCTCACGGTGTTCCGGGTTTCCCCGGATTGTACCGTTCGG	Reverse template for transcription of the M1 *T. oceani* 112 *ilvE* RNA
MES179	TAATACGACTCACTATAGGAAGTGTACCTTAGGGTTCC	Forward template for transcription of the M2 *T. oceani* 112 *ilvE* RNA

	GGCCATAAGGCGTCAGCG ACCGAGAGGTACAATCCG GGG	
MES180	TCTTTTAGGAACTTGCCGC TGGAGCCTTTTATGCTCAC GGTGTTCCGGGTTTCCCCG GATTGTACCTCTCGG	Reverse template for transcription of the M2 *T. oceani* 112 *ilvE* RNA
MES181	TCTTTTAGGAACTTGCCGC TGGAGCCTTTTATGCTCAC GGTTTTCCGGGTTTCCCCG GATTGTACCGCTCGG	Reverse template for transcription of the M3 *T. oceani* 112 *ilvE* RNA
MES182	TCTTTTAGGAACTTGCCGC TGGAGCCTTTTATGCTCAC GGGGTTCCGGGTTTCCCCG GATTGTACCGCTCGG	Reverse template for transcription of the M4 *T. oceani* 112 *ilvE* RNA
MES183	TCTTTTAGGAACTTGCCGC TGGAGCCTTTTATGCTCAC CGTGTTCCGGGTTTCCCCG GATTGTACCGCTCGG	Reverse template for transcription of the M5 *T. oceani* 112 *ilvE* RNA
MES184	TCTTTTAGGAACTTGCGGC TGGAGCCTTTTATGCTCAC GGTGTTCCGGGTTTCCCCG GATTGTACCGCTCGG	Reverse template for transcription of the M6 *T. oceani* 112 *ilvE* RNA

MES144	TCTTTTAGGAACTTGCGGC TGGAGCCTTTTATGCTCAC CGTGTTCCGGGTTTCCCCG GATTGTACCGCTCGG	Reverse template for transcription of the M7 *T. oceani* 112 *ilvE* RNA
MES207	TCTTTTAGGAACTTGCCGC TGGAGCCTTTTATGCTCAC GGTATTCCGGGTTTCCCCG GATTGTACCGCTCGG	Reverse template for transcription of the M8 *T. oceani* 112 *ilvE* RNA
MES208	TAATACGACTCACTATAGG AGTGTACCTAGGGTTCCGC GCTTTACAGCGGTCTGGTC CAAGCGGTACAGGTGC	Forward template for transcription of the WT *P. fermentans* 109 *glxB* RNA
MES209	ACTTATCCAGGACTTTCCG CAGGGCCTTTTTTACCCAC GGTGTTGTGCAATCTGCAC CTGTACCGCTTGGACC	Reverse template for transcription of the WT *P. fermentans* 109 *glxB* RNA
MES212	TAATACGACTCACTATAGG AGAGTGCATCTAGGGTTCC GATGGAAACATGTCTGGTC CGAGCGATGCAGG	Forward template for transcription of the WT *Clostridium sp. BL8* 104 *natA* RNA
MES213	CATAGGTTGAGACTCGTCC TTTAGGCTTTTTTCCTAAA	Reverse template for transcription of the WT *Clostridium sp. BL8* 104 *natA* RNA

	AGTGTAGATAAAATCCTGC ATCGCTCGGACCAGAC	
MES153	GCGCACATGAGAATTCCAG CGACGCTGTTGATCCTTTT AAATAAGTCTGATAAAATG TGAACTAAAGTTATAAATG AAAGAATGTATTATAACAG CAGGAAGTGTACCTAGGGT TCCGGGAGCTGCTCCGTCT GGTCCGAGCGGTACAAAA TCCAGAGCATGGATTTACA CCGTGGGCAGAAAATACC CGAGCGGAAAGTTCCTCGA GAGGGTAGGAACACCGCT CGGTTTTCTATAATATTCA GAGGATCCCCAGCTGCGC	IDT G-block containing the wild-type *D. hafniense ilvE* riboswitch and controlled by the *B. subtilis lysC* promoter, serving as a template for the dsDNA template for *in vitro* transcription termination
MES154	GCGCACATGAGAATTCCAG CGACGCTGTTGATCCTTTT AAATAAGTCTGATAAAATG TGAACTAAAGTTATAAATG AAAGAATGTATTATAACAG CAGGAAGTGTACCTAGGGT TCCGGGAGCTGCTCCGTCT	IDT G-block containing the mutation of a conserved nucleotide (M9) *D. hafniense* riboswitch and controlled by the *B. subtilis lysC* promoter, serving as a template for the dsDNA template for *in vitro* transcription termination.

	GGTCCGAGAGGTACAAAA TCCAGAGCATGGATTTACA CCGTGGGCAGAAAATACC CGAGCGGAAAGTTCCTCGA GAGGGTAGGAACACCGCT CGGTTTTCTATAATATTCA GAGGATCCCCAGCTGCGC	
MES157	GCGCACATGAGAATTCCAG CGACG	Forward primer for amplification of all *D. hafniense ilvE* reporter constructs and single-round transcription termination assay templates
MES158	GCGCAGCTGGGGATCCTCT G	Reverse primer for amplification of the WT and M9 *D. hafniense ilvE* single-round transcription termination assay templates
MES200	CCGAGCGGTGTTCCTACCC TC	Reverse primer for amplification of the M10 *D. hafniense ilvE* single-round transcription termination assay template
MES204	GCGCAGCTGGGGATCCTCT GAATATTATAAAAAACCG AGCGGTGTTCCTACCC	Reverse primer for amplification of the M11 *D. hafniense ilvE* single-

		round transcription termination assay template

CHAPTER SIX

Tandem riboswitches form a natural Boolean logic gate to control purine metabolism in bacteria

Largely adapted from the following publication:

Sherlock, M.E., Sudarsan, N., Stav, S., Breaker, R.R. (2018) Tandem riboswitches form a natural Boolean logic gate to control purine metabolism in bacteria. *eLife 7*, e33908.

ABSTRACT

Gene control systems sometimes interpret multiple signals to set the expression levels of the genes they regulate. In rare instances, ligand-binding riboswitch aptamers form tandem arrangements to approximate the function of specific two-input Boolean logic gates. In this chapter, I report the discovery of riboswitch aptamers for phosphoribosyl pyrophosphate (PRPP) that naturally exist either in singlet arrangements or occur in tandem with guanine aptamers. Tandem guanine-PRPP aptamers can bind the target ligands, either independently or in combination, to approximate the function expected for an IMPLY Boolean logic gate to regulate transcription of messenger RNAs for *de novo* purine biosynthesis in bacteria. The existence of sophisticated all-RNA regulatory systems that sense two ancient ribonucleotide derivatives to control synthesis of RNA molecules supports the hypothesis that RNA World organisms could manage a complex metabolic state without the assistance of protein regulatory factors.

INTRODUCTION

A variety of riboswitch classes regulate gene expression in response to the binding of specific metabolite or inorganic ion ligands.[1-3] Previously, 25 riboswitch classes were known that selectively respond to ligands that are chemical derivatives of RNA monomers or their precursors.[4] This strong bias in favor of RNA-like ligands strengthens the hypothesis[5,6] that many of the widely-distributed riboswitch classes are direct descendants from the RNA World – a time before genetically-encoded protein biosynthesis.[7,8] If this speculation is generally correct, some of the most abundant riboswitch classes remaining to be discovered also will regulate gene expression in response to the binding of other ancient RNA derivatives.[9]

With these considerations in mind, I evaluated the possible functions of variants of recently validated riboswitches for guanidine.[10] Collectively, guanidine-I riboswitches along with differently-functioning variants were originally called the *ykkC* orphan riboswitch candidate because no ligand had been identified.[11,12] After removing guanidine-I riboswitches (subtype 1), which constituted ~70% of the original *ykkC* motif RNAs, I further separated the remaining collection of *ykkC* RNA variants (subtype 2). Upon further investigation of the RNA sequences, the phylogeny of organisms containing these RNAs, and the genes associated with each instance of the RNA, I concluded that *ykkC* subtype 2 RNAs were populated by multiple riboswitch classes that sensed distinct ligands.

With help from my colleague Shira Stav, I sorted subtype 2 representatives according to the gene located immediately downstream from the *ykkC* RNA motif to create revised consensus models for the RNAs within each gene association group. In some cases, additional conserved nucleotides were observed immediately 5′ and 3′ of the three hairpin

structures (called P1, P2 and P3) that are characteristic of *ykkC* motif RNAs. This region located downstream of the guanidine binding site of subtype 1 RNAs were found to be unnecessary for guanidine-I riboswitches. In total, I sorted these distinct representatives into four candidate riboswitch classes called subtypes 2a through 2d.

All subtype 2 RNAs appear to be extremely similar because they share many of the same highly-conserved nucleotides and secondary structure features. I was fortunate to have high-resolution crystal structures[13,14] of the guanidine-I riboswitch (subtype 1, Figure 6-1A) in hand when evaluating each of these variant subtypes. Nearly all highly-conserved nucleotides shared by *ykkC* RNAs are involved in forming tertiary contacts within the larger folded structure that positions other nucleotides to form the ligand binding pocket. However, nucleotides that change identity and are characteristic of each subtype are almost exclusively found in close proximity to the known binding pocket of guanidine-I riboswitches (indicated by black asterisks in Figure 6-1A). For subtypes 2a, 2b, and 2c, I speculated that the additional stretches of conserved nucleotides preceding the P1 stem and following the P3 stem likely play a role in ligand recognition, or perhaps provide structural support for an expanded binding pocket. Unfortunately, because these nucleotides are not conserved in guanidine-I riboswitches, I did not have a good understanding of their 3D positioning relative to the rest of the aptamer.

It quickly became clear that sequences associated with branched-chain amino acid metabolism appeared highly similar to those associated with glutamate synthases and *natA* transporters, and so these were collectively termed *ykkC* subtype 2a (Figure 6-1A). Shortly thereafter, I experimentally verified this variant motif as a riboswitch class that responds to the bacterial alarmone ppGpp (Chapter 5). In this chapter, I describe the function of *ykkC*

subtype 2b RNAs (Figure 6-1A,B), whereas the ligands for subtype 2c and 2d RNAs have not yet been identified (see Discussion and Chapter 7).

I exploited three major clues to identify phosphoribosyl pyrophosphate (PRPP; 5-phospho-α-D-ribose 1-diphosphate) (Figure 6-1B) as the ligand for riboswitches represented by *ykkC* subtype 2b RNAs. First, representatives of this riboswitch class commonly associate with genes for *de novo* purine biosynthesis (Figure 6-1C). In addition, cytosine deaminase (*codA*), a gene involved in pyrimidine salvage that is known to be repressed by the purine guanine,[15] is also commonly located downstream of subtype 2b RNAs. These associated genes, which are either related to the production of purines or are regulated by a purine, suggested that the ligand would most likely be a compound related to purine metabolism.

Second, there are 257 examples of *ykkC* subtype 2b RNAs distributed throughout bacteria of the phylum Firmicutes. Of these, 127 examples occur in a tandem arrangement[16] with a guanine aptamer[17] located directly upstream (Figure 6-1B). Such tandem arrangements are likely to function in concert as two-input molecular logic gates.[16,18,19] Guanine riboswitches typically turn off expression of downstream purine biosynthesis genes when sufficient guanine is present,[17] while we predicted that representatives of *ykkC* subtype 2b riboswitches turn on expression of downstream genes. The physical arrangement of each aptamer in the tandem arrangement suggests that they share a single expression platform. Guanine binding to its aptamer likely would promote the formation of an adjoining intrinsic transcription terminator stem, which would halt transcription. It is typical for guanine riboswitches to suppress expression of genes involved in purine biosynthesis.[17] In opposition, binding of ligand to the adjacent subtype 2b RNA would

preclude terminator stem formation and thereby override the typical function of the guanine riboswitch. If this speculation is correct, then the ligand for subtype 2b RNAs should trigger cells to make more purines even when one of the major purine compounds, guanine, is plentiful.

Third, I noticed that the consensus sequence and structural models for ppGpp riboswitch aptamers (subtype 2a) and subtype 2b RNAs are remarkably alike, including many similarities at the ligand binding site, suggesting that their ligands might also share some chemical features. Particularly notable was my observation that a long-range C-G base-pair present in guanidine-I riboswitches[13,14] is absent in ppGpp riboswitches, but appears to again be present in subtype 2b RNAs. I speculated that the structural role of the G nucleotide missing in ppGpp aptamers might be performed by the guanine base of the ligand. However, because the highly-similar subtype 2b RNA retains this G nucleotide (indicated by a red asterisk in Figure 6-1A) in its aptamer, its ligand might resemble an apurinic version of ppGpp.

RESULTS

Variant riboswitch aptamers selectively bind the nucleotide precursor phosphoribosyl pyrophosphate (PRPP). To evaluate PRPP binding and riboswitch aptamer function, I examined a 106 nucleotide singlet aptamer derived from the *purC* gene of the bacterium *Facklamia ignava* (Figure 6-2A) by using in-line probing.[20,21] This method takes advantage of the inherent chemical instability of phosphodiester linkages in unstructured regions of an RNA and the folding differences in aptamers brought about by ligand binding to provide details on riboswitch function.

The 106 *purC* RNA construct exhibits structural modulation at highly conserved nucleotide positions upon the addition of PRPP (Figure 6-2B). In contrast, I observe no structural modulation of the RNA when the known ligands for the other *ykkC* subtypes (guanidine and ppGpp) or for the guanine riboswitch (guanine) are added. Importantly, the structural changes selectively brought about by PRPP largely occur in regions equivalent to those previously implicated in ligand binding by guanidine-I[10,13,14] and ppGpp riboswitch classes. This observation strongly suggests that these binding site nucleotides have changed to alter ligand specificity, and that PRPP is recognized by this converted binding pocket.

PRPP triggers riboswitch-mediated regulation of transcription termination. Due to the well-characterized instability of the PRPP molecule,[22,23] especially under alkaline conditions and high Mg^{2+}, PRPP extensively degrades into multiple breakdown products over the course of our binding experiments. As a result, in-line probing assays will not yield accurate dissociation constants. Therefore, to evaluate the effects of PRPP on riboswitch function, I performed single-turnover *in vitro* transcription termination experiments,[24] which involve briefer incubations under milder conditions.

The P3 helix of the PRPP aptamer (Figure 6-1A) is predicted to act as an antiterminator stem, which should increase the number of full-length mRNA transcripts when PRPP is bound. Transcription of a singlet PRPP riboswitch derived from the *purB* gene of *Heliobacterium modesticaldum* (Figure 6-3A) produces a greater yield of full-length RNA when PRPP is added to the reaction (Figure 6-3B). The concentration of PRPP necessary to achieve half maximal termination (T_{50}) is less than 100 μM (Figure 6-3C and Figure 6-4). In contrast, an RNA construct (M1) with disruptive mutations within the highly-

conserved aptamer sequence yields the same fraction of terminated and full-length transcripts regardless of PRPP concentration (Figure 6-3B).

Importantly, T_{50} values should not be considered equivalent to K_D values. Riboswitches evaluated by transcription termination assays operate under kinetic limitations, whereas K_D values established by in-line probing can reach thermodynamic equilibrium.[24-27] Thus, the K_D for PRPP binding to the *H. modesticaldum* aptamer is likely to be far better than the T_{50} values we measured for the *H. modesticaldum purB* construct, which range between 40 and 90 μM (Figure 6-3C and Figure 6-4).

***In vivo* regulation of gene expression by a PRPP riboswitch.** My colleague Narasimhan Sudarsan and I created a fusion of the same singlet PRPP riboswitch sequence from *H. modesticaldum* (Figure 6-3A) with *lacZ* in wild-type (WT) and Δ*purF Bacillus subtilis* cells. Deletion of the *purF* gene, which encodes the enzyme that catalyzes the first committed step of *de novo* purine biosynthesis (Figure 6-1C), causes PRPP to accumulate when cells are starved for purines. Accordingly, the *lacZ* gene is expressed (blue) in glucose minimal medium in the Δ*purF* strain, but not (white) in the parental *B. subtilis* strain with a functional *purF* gene (Figure 6-3D).

Notably, the cells range from completely white to a robust blue under these two conditions, indicating that gene expression is fully suppressed when no ligand is present. Thus, the riboswitch appears to sample a much greater dynamic range for gene expression *in vivo* than is observed with *in vitro* transcription termination assays. A riboswitch-reporter fusion carrying a mutation (M2) in the terminator stem expresses *lacZ* under both conditions. These results demonstrate that PRPP accumulation in cells induce gene

expression, which parallels our findings *in vitro* that PRPP binding by *ykkC* subtype 2b RNAs promotes transcription elongation (Figure 6-3B,C).

Natural examples of tandem riboswitch aptamers for guanine and PRPP bind these ligands to induce mutually-exclusive structures. I next sought to characterize the tandem arrangement of the guanine and PRPP aptamers by in-line probing analysis of the 208 nucleotide RNA derived from the *codA* gene of *Bacillus megaterium* (Figure 6-5A). Each aptamer appears to independently fold to form its characteristic structure and each exhibits typical structural modulation (Figure 6-5B) upon addition of its cognate ligand as previously observed for singlet guanine[17] or PRPP (Figure 6-2B) aptamers.

I only examined the aptamer regions of the RNA by these in-line probing experiments, but the logic gate function of the tandem riboswitch is accomplished through interactions between the two aptamers and the adjoining expression platform. Consequently, terminator and antiterminator stems are stabilized or destabilized depending on the relative concentrations of both guanine and PRPP.

The precise mechanism by which the guanine and PRPP aptamers work in opposition to control the formation of the terminator stem has yet to be fully experimentally evaluated. However, bioinformatic and in-line probing data already might provide clues regarding the structural interplay between the two aptamers. First, guanine aptamers are always located immediately adjacent to the conserved nucleotides at the 5′ terminus of the PRPP aptamer in tandem arrangements (Figure 6-1B). Specifically, there are typically only 5 to 7 nucleotides separating the P1 stems of the guanine and PRPP aptamers, suggesting this juxtaposition is important for logic gate function. Second, the addition of PRPP to the tandem construct causes a reduction in spontaneous RNA cleavage within the conserved

nucleotides at the 5′ terminus of the PRPP aptamer (Figure 6-5B, region 5), whereas guanine addition causes this same region to experience a dramatic increase in scission. These results suggest that guanine binding to its aptamer destabilizes an important structural feature necessary for ligand binding by the PRPP aptamer. Thus, each aptamer might interfere with its neighboring aptamer upon ligand binding, perhaps by causing steric clashes due to their proximity. Additional experiments will need to be conducted to more fully evaluate this mechanistic hypothesis.

Boolean logic function by a riboswitch carrying tandem guanine and PRPP aptamers. To evaluate the proposed two-input logic gate function, I subjected the same guanine-PRPP tandem aptamer system including its associated expression platform from *B. megaterium* (Figure 6-6A) to *in vitro* transcription termination assays. Consistent with results from the singlet aptamer, the addition of PRPP increases the fraction of RNA transcripts that pass beyond the intrinsic terminator stem and reach full length (Figure 6-6B). Also as predicted, the addition of guanine causes an increase in the fraction of terminated transcripts. When both guanine and PRPP are at high levels, the predicted genetic override occurs, and the fraction of terminated products approaches zero.

Separate and simultaneous mutations to each aptamer (constructs M3 through M5) confirm the OFF and ON function of the guanine and PRPP aptamers, respectively. Specifically, the mutations in M3 alter two highly-conserved nucleotides in the core of the guanine aptamer that are known to be essential for forming its ligand-binding core.[28,29] As expected, the addition of guanine does not increase transcription termination. Similarly, the mutations in M4 alter two highly-conserved nucleotides in PRPP aptamer, which I expect to disrupt PRPP binding. Again as expected, the amount of transcription termination

product increases, and PRPP addition has no effect on the distribution of transcription products. Construct M5 combines the M3 and M4 mutations to generate an RNA that should fail to bind either ligand. Indeed, the M5 construct appears to be unaffected by the addition of either guanine or PRPP. (Figure 6-6B). I observed similar results *in vivo* by fusing the WT and M3 through M5 tandem guanine-PRPP riboswitch constructs from *B. megaterium* to the *lacZ* reporter gene and transforming these constructs into WT *B. subtilis* cells (Figure 6-6C).

The versatility of this tandem riboswitch system is best demonstrated by varying the concentrations of both guanine and PRPP in a series of *in vitro* transcription termination assays. The data I collected from these reactions yields a 3D landscape demonstrating again that the relative concentrations of guanine and PRPP dictate the fraction of RNA transcripts that reach full length (Figure 6-6D and Figure 6-7). Thus, the tandem riboswitch arrangement permits cells to tune the expression levels of associated genes based on the relative concentrations of guanine and PRPP, such that high PRPP levels promote purine production despite the abundance of guanine, a prominent purine.

At the single molecule level, each riboswitch aptamer in this tandem construct favors either transcription termination or elongation based on the binding of its cognate ligand. In this regard, each RNA molecule carrying this tandem aptamer arrangement functions as a Boolean IMPLY logic gate. The truth table for an IMPLY logic gate (Figure 6-8A) indicates that gene expression should be on when neither ligand is present, and when PRPP is bound by the RNA – even when guanine is available. Finally, gene expression should be off when guanine is bound by the RNA, which should occur when guanine is abundant while PRPP is not. This regulatory pattern could be accomplished by the tandem RNA

arrangement through formation of an antiterminator stem by alternative base pairing interactions between the two aptamers, similar to those observed for an allosterically-regulated tandem riboswitch-ribozyme described previously.[19,30] Alternatively, as noted earlier, the aptamers might compete sterically due to the proximity of their conserved sequences and structures. Regardless of the precise mechanism, we hypothesize that an unoccupied PRPP aptamer permits formation of the terminator stem, whereas an occupied guanine aptamer destabilizes the PRPP aptamer. Additional detailed kinetic and structural studies will be necessary to reveal the precise molecular interactions that yield IMPLY gene control logic.

When transcription termination or gene expression is observed as the average output from a population of RNA constructs, the resulting data do not represent a perfect Boolean function because the output yields intermediate values between zero and maximal values. Whereas each of the individual RNAs operates as a binary logic gate that is either on or off, a population of these RNAs exceeds the function of an individual Boolean logic gate. Together, the output of the tandem riboswitch population functions more like a molecular rheostat to tune gene expression based on relative ligand concentrations, rather than as an all-or-nothing binary switch. Regardless, when ligand concentrations are at their extreme values, the output of the tandem system (Figure 6-6D) approaches the binary logic of an IMPLY gate.

The biological utility of the opposing guanine and PRPP signals evaluated by tandem riboswitch aptamers. A situation in cells in which both guanine and PRPP are scarce would indicate an extremely starved state. Ordinarily, one might expect guanine and PRPP levels to be negatively correlated, such that an excess of guanine would deplete

PRPP as a result of the synthesis of 5′-guanosine monophosphate, and vice versa, either through *de novo* purine biosynthesis or the purine salvage pathway (Figure 6-1B). If true, then why is it necessary to form this complex tandem arrangement to override the gene control function of a guanine riboswitch? This apparent regulatory paradox is addressed in more detail below.

During the stringent response in Firmicutes, nutrient starvation is signaled by the presence of uncharged tRNAs that activate the ppGpp synthase RelA.[31] ppGpp synthesis causes GTP levels to drop and the transcriptional repressor CodY, which requires GTP and isoleucine to bind its target sites, to release and enable transcription of branched-chain amino acid (BCAA) biosynthesis genes.[32] The resulting increased ppGpp concentration promotes binding of this signaling molecule as an inhibitor of enzymes such as GuaB, GMK, and HprT (Figure 6-1C) to suppress GTP production,[33] causing both guanine and PRPP to accumulate.

Under these conditions the guanine-PRPP tandem riboswitch still turns on expression of its downstream genes, despite the mechanisms in place to keep GTP levels low. At first this appears counterintuitive, but the genes associated with the tandem riboswitch system only include those for catalyzing the steps of *de novo* purine synthesis up to the production of inosine monophosphate (IMP), the branch point between AMP and GMP production (Figure 6-1C). Thus, the riboswitch promotes AMP synthesis, but will not turn on genes specific for GDP or GTP synthesis that are already being inhibited by ppGpp. ATP is known to be present in elevated quantities during the stringent response because it is used as the initiating nucleotide for transcription of genes under the ppGpp superregulon,

whereas rRNA transcripts utilize GTP as the initiating nucleotide to suppress protein production.[34]

Expansion of natural Boolean logic devices made of RNA. As noted above, the guanine-PRPP tandem riboswitch arrangement at its functional limits most closely approximates an IMPLY logic gate (Figure 6-8A). This all-RNA system joins the NOR gate formed by tandem *S*-adenosylmethionine (SAM) and adenosylcobalamin (AdoCbl or coenzyme B_{12}) riboswitches[16] (Figure 6-8B) and the AND gate formed by tandem T box and ppGpp riboswitch RNAs (Figure 6-8C) as sophisticated regulatory RNAs that approximate Boolean logic functions.

Similarly, the tandem riboswitch-ribozyme arrangement formed by a c-di-GMP-II riboswitch and a group I self-splicing ribozyme senses both c-di-GMP and guanosine (or any of its 5′-phosphorylated derivatives) to approximate the function of a Boolean AND gate.[19,30] In this example, one chemical input is sensed by a conventional riboswitch aptamer, whereas the second chemical input is sensed by the active site of a ribozyme. These more complex riboswitch and ribozyme arrangements are notable because they demonstrate how sensory and regulatory systems comprised entirely of RNA can perform sophisticated decision-making in modern cells.

DISCUSSION

Most experimentally validated riboswitch classes sense and respond to ligands originating from RNA metabolism pathways, and therefore it seemed reasonable to consider the likelihood that riboswitches for PRPP would exist. PRPP is an essential precursor for the modern biosynthesis of RNA monomers in organisms from all three

domains of life and is likely to have been important for the production of ribonucleotides very early in evolution. If PRPP was used by RNA World organisms to generate the building blocks of this major biopolymer, then RNA aptamers that selectively bound this biosynthetic precursor in ancient times might have persisted to serve as riboswitch aptamers in modern cells.

Intriguingly, representatives of PRPP riboswitches were on the list of riboswitch candidates over a decade ago,[11] but due to their striking similarity to riboswitches for guanidine[10] and ppGpp, they were unusually difficult to identify and experimentally validate. Members of other riboswitch classes have previously been found to exhibit altered ligand specificities by accruing key mutations in nucleotides of their binding pockets. For example, guanine riboswitches have undergone one or a few key mutations to change their ligand specificity to adenine[35] or 2′-deoxyguanosine.[36,37] Also, subtle binding site variants of FMN riboswitches appear to bind a closely related ligand,[37,38] and riboswitches for molybdenum cofactor (MoCo) are proposed to have modest binding site changes to alter their ligand specificity to bind tungsten cofactor.[39] In nearly all these previous examples, there are relatively small changes in the chemical structure of the ligand that need to be accommodated by the riboswitch binding pocket changes.

For the subtypes of *ykkC* motif RNAs whose ligands have been identified, the chemical differences in the ligands can be substantial. The ppGpp (subtype 2a) and PRPP (subtype 2b) riboswitch ligands differ by the presence or absence of a guanine nucleobase and by the location and number of phosphate moieties. Guanidinium, the ligand for guanidine-I (subtype 1) riboswitches is entirely distinct from PRPP, and this moiety forms only a portion of the purine ring of ppGpp. The ligands sensed by subtypes 2c and 2d, which are

associated with genes encoding nucleoside diphosphate linked to X (NUDIX) hydrolases of unknown specificity and a transporter of unknown function associated with phosphonate utilization (phn_DUF6), respectively, still remain to be discovered. It will be interesting to compare the chemical structures of all of the *ykkC* family riboswitch ligands once those for the subtype 2c and 2d riboswitch candidates have been validated. Nonetheless, the collection of *ykkC* motif RNAs already demonstrates the ability of RNA to recognize a variety of distinct small molecules, all while utilizing the same general structural scaffold.

Furthermore, the discovery of tandem guanine-PRPP riboswitches serve as additional demonstrations of how complex molecular devices can be formed purely from RNA. The *in vitro* transcription data for the logic gate molecules (Figure 6-6B,C) represent the combined output of a population of RNA transcripts. Although the transcript length trends are as predicted for an IMPLY gate, these outputs do not match the perfect binary output states of the corresponding electronic logic gates. However, each individual RNA molecule indeed has a binary output just like an individual electronic logic gate. Each tandem riboswitch aptamer transcript is either terminated (OFF) or fully transcribed (ON), but the band intensities from analytical gel electrophoresis report on the results of the population of transcripts.

The outputs of the riboswitch logic gate can be further complicated by at least two issues. First, an individual transcript might fail to function properly, say by misfolding of an aptamer domain, such that it cannot respond to the presence of a ligand. This failure mode might help explain the imperfect yields of terminated and full-length transcripts observed when examining riboswitches by *in vitro* transcription assays. It seems likely that folding problems of RNAs prepared *in vitro* might be less prominent when the RNAs are

produced by cells. This means that high levels of transcription read-through that is non-responsive to the absence of ligand are not evident in reporter gene assays because the riboswitch aptamers have evolved to fold properly and on a relevant timescale in the cellular milieu.

Second, the presence or absence of the two ligand inputs will change the function of the logic gate system at the individual transcript level, and therefore change the frequency of their choice of these two states rather than change the number of output states possible. Unlike electronic logic gates that use electrical current at only two values (off and on), ligand concentrations for riboswitches can be highly variable. Therefore, given the nature of ligand binding to receptors (each described by an equilibrium constant equation), differences in ligand concentrations will result in differences in the probabilities that the two riboswitch aptamers will be occupied. This occupancy distribution will be reflected in the transcriptional output of the RNA population, but each individual transcript still yields a binary output (terminated or full-length product) that is biased by the concentrations of ligands present.

The IMPLY gate described in this chapter involves a riboswitch RNA with aptamers for the RNA nucleobase guanine and the RNA precursor PRPP. While tandem riboswitch systems have previously been described (Figure 6-8), the guanine-PRPP system is the only example of two tandem aptamers that share the same expression platform, allowing for a new type of logic function to be implemented. All of these RNA systems demonstrate complex chemical sensing and regulatory functions, which perhaps reflects the complexity of systems used for metabolite monitoring and biological regulation in ancient RNA World organisms. Furthermore, it seems likely that additional variations of two-input Boolean

logic gates will be represented by tandem riboswitches that have yet to be discovered in modern cells.

MATERIALS AND METHODS

Chemicals and reagents. Chemicals were purchased from Sigma-Aldrich with the exceptions of phosphoribosyl pyrophosphate (Santa Cruz Biotech) and guanosine 5′,3′-bisdiphosphate (TriLink Biotech). [γ-^{32}P]-ATP and [α-^{32}P]-UTP were purchased from Perkin Elmer and used within two weeks of receipt. Bulk chemicals were purchased from J.T. Baker and enzymes from New England Biolabs, unless otherwise noted. All solutions were prepared using deionized water (dH_2O) and were either autoclaved or filter sterilized (using 0.22 μm filters, Millipore) prior to use. DNA oligonucleotides were purchased from Sigma-Aldrich and Integrated DNA Technologies. A list of oligonucleotides used in this chapter can be found in Table 6-1.

Bioinformatics analyses. Additional examples of *ykkC* motif RNAs were identified using Infernal 1.1[40] to search RefSeq version 76 plus additional environmental microbial databases as described previously.[41] Iterative searches for new sequences were performed based on the previously published alignment of *ykkC* subtype 2 RNAs.[10] These sequences were sorted by nucleotide identity at a specific position as previously indicated[10] to exclude guanidine-I riboswitches, then further manually sorted by downstream gene association, as described above. This revealed 257 unique examples of subtype 2b RNAs (PRPP riboswitches), 127 of which are found in tandem with a guanine aptamer. The consensus sequence and secondary structure model was constructed using R2R software.[42]

RNA oligonucleotide preparation. Double-stranded DNA (dsDNA) templates for RNA transcription were produced either by extension of overlapping synthetic DNAs (Table 6-1) using SuperScript II reverse transcriptase (Thermo Fisher Scientific) for the *F. ignava* construct or by PCR from genomic DNA for *B. megaterium* construct. Primers were designed to contain a 5′-terminal T7 RNA polymerase (T7 RNAP) promoter to enable transcription. The desired RNA constructs were prepared from these dsDNAs by *in vitro* transcription, purified, and subsequently 5′ ^{32}P-labeled as previously described.[10,43]

RNA In-line Probing Analyses. In-line probing assays[20,21] were performed precisely as described previously.[10,43]

Single-round transcription termination assays. DNA constructs for single-round *in vitro* transcription were designed to carry the riboswitch of interest spanning from the predicted natural transcription start site to 55 (*H. modesticaldum* construct) or 26 (*B. megaterium* construct) nucleotides following the terminator stem. The promoter from the *B. subtilis lysC* gene, which is compatible with *E. coli* RNA polymerase, was used for both *in vitro* transcription termination as well as the *lacZ* reporter experiments detailed below. Mutations were incorporated in the region of the *B. megaterium* construct before the guanine aptamer so that it contained no cytidines, while the *H. modesticaldum* construct was naturally cytidine-deficient. Synthetic double-stranded DNA templates were designed, purchased (Integrated DNA Technologies) and subsequently amplified by PCR for the *H. modesticaldum* constructs whereas constructs for the WT and M3 through M5 *B. megaterium* riboswitch were amplified from the plasmids containing those inserts for the reporter gene fusion, as described below.

Approximately 2 pmol of the resulting, purified DNA template was added to a transcription initiation mixture (20 mM Tris-HCl [pH 7.5 at 23ºC], 75 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 10 µg mL^{-1} bovine serum albumin [BSA], 130 µM ApA dinucleotide, 1% glycerol, 0.04 U $µL^{-1}$ *E. coli* RNA polymerase holoenzyme, 2.5 µM GTP, 2.5 µM ATP, and 1 µM UTP). Approximately 1 µCi [α-^{32}P]-UTP was added per 8 µL reaction and transcription was allowed to proceed at 37ºC for 10 minutes, leading to formation of a stalled polymerase complex at the first cytidine nucleotide of each transcript. For each 8 µL transcription reaction, 1 µL of 10x elongation buffer (20 mM Tris-HCl [pH 7.5 at 23ºC], 75 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 2 mg mL^{-1} heparin, 1.5 mM ATP, 1.5 mM GTP, 1.5 mM CTP, and 250 µM UTP) as well as 1 µL of a 10x solution of the ligand of interest were added sequentially. Transcription reactions were then incubated at 37ºC for an additional 60 minutes.

The transcription products were subsequently analyzed via denaturing 10% PAGE and visualized using a phosphorimager (GE Healthcare Life Sciences). Fraction full length values were calculated by varying the ligand concentration in separate reactions and quantifying the changes in band intensity of both full length (FL) and terminated (T) transcription products using the formula (FL)/(FL+T). The T_{50} values were determined by plotting the fraction FL as a function of the logarithm of ligand concentration and using a sigmoidal four parameter logistic fit in GraphPad Prism 7.

Riboswitch reporter assays. Riboswitch-reporter constructs consisting of the *B. subtilis lysC* gene promoter and either the *H. modesticaldum* or *B. megaterium* riboswitch fused upstream of the *E. coli lacZ* gene were inserted into the vector pDG1661 as described previously[10,44] and integrated into the *amyE* locus of wild-type WT (1A1 strain 168 Δ*trp*)

or Δ*purF* (strain BKE06490, obtained from the Bacillus Genetic Stock Center, Ohio State University) *B. subtilis*.[12] Incorporation of mutations for the *B. megaterium* reporter were carried out via Quikchange site-directed mutagenesis (Agilent Technologies). Reporter experiments were performed by inoculating either *B. subtilis* strain 168 WT or Δ*purF* transformed with the riboswitch reporter construct into Lysogeny Broth (LB) with appropriate antibiotics and growing overnight at 37 ºC. For experiments with the singlet *H. modesticaldum* reporter, the bacteria were diluted into lysogeny broth (LB), grown for an additional 4-6 hours until culture growth was visible, centrifuged, supernatant removed, the resuspended in glucose minimal media (GMM). For experiments with the tandem *B. megaterium* reporter, bacteria were diluted directly into either LB or GMM, as indicated, and grown for 6 hours with or without supplemented guanine. Synthetic DNAs used in cloning are listed in Table 6-1.

ACKNOWLEDGEMENTS

We thank Adam Roth, Ruben Atilho, Kimberly Harris, and other members of the Breaker laboratory for helpful discussions. M.E.S. was supported by an NIH Cellular and Molecular Biology Training Grant (T32GM007223). This work was also supported by NIH grants to R.R.B. (GM022778 and DE022340). R.R.B. is also supported by the Howard Hughes Medical Institute.

REFERENCES

(1) Roth, A., Breaker, R.R. (2009) The structural and functional diversity of metabolite-binding riboswitches. *Annu Rev Biochem 78*, 305-334.

(2) Serganov, A., Nudler, E. (2013) A decade of riboswitches. *Cell 152*, 17-24.

(3) Sherwood, A.V., Henkin, T.M. (2016) Riboswitch-mediated gene regulation: Novel RNA architectures dictate gene expression responses. *Annu Rev Microbiol 70*, 361-374.

(4) McCown, P.J., Corbino, K.A., Stav, S., Sherlock, M.E., Breaker, R.R. (2017) Riboswitch diversity and distribution. *RNA 23*, 995-1011.

(5) Nelson, J.W., Breaker, R.R. (2017) The lost language of the RNA World. *Sci Signal 10*, 1-10.

(6) Breaker, R.R. (2010) Riboswitches: From ancient gene control systems to modern drug targets. *Microbe 5*, 13-20.

(7) Gilbert, W. (1986) Origin of life: The RNA world. *Nature 319*, 618.

(8) Benner, S.A., Ellington, A.D., Tauer, A. Modern metabolism as a palimpsest of the RNA world. *Proc Natl Acad Sci USA 86*, 7054-7058 (1989).

(9) Breaker, R.R. (2011) Prospects for riboswitch discovery and analysis. *Mol Cell 43*, 867-879.

(10) Nelson, J.W., Atilho, R.M., Sherlock, M.E., Stockbridge, R.B., Breaker, R.R. (2017) Metabolism of free guanidine in bacteria is regulated by a widespread riboswitch class. *Mol Cell* 65, 220-230.

(11) Barrick, J.E., Corbino, K.A., Winkler, W.C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I. (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc Natl Acad Sci USA 101*, 6421-6426.

(12) Meyer, M.M., Hammond, M.C., Salinas, Y., Roth, A., Sudarsan, N., Breaker, R.R. (2011) Challenges of ligand identification for riboswitch candidates. *RNA Biol 8*, 5-10.

(13) Reiss, C.W., Xiong, Y., Strobel, S.A. (2017) Structural basis for ligand binding to the guanidine-I riboswitch. *Structure 25*, 195-202.

(14) Battaglia, R.A., Price, I.R., Ke, A. (2017) Structural basis for guanidine sensing by the *ykkC* family of riboswitches. *RNA 23*, 578-585.

(15) Kilstrup, M., Meng, L.M., Neuhard, J., Nygaard, P. (1989) Genetic evidence for a repressor of synthesis of cytosine deaminase and purine biosynthesis enzymes in *Escherichia coli*. *J Bacteriology 171*, 2124-2127.

(16) Sudarsan, N., Hammond, M.C., Block, K.F., Welz, R., Barrick, J.E., Roth, A., Breaker, R.R. (2006) Tandem riboswitch architectures exhibit complex gene control functions. *Science 13*, 300-304.

(17) Mandal, M., Boese, B., Barrick, J.E., Winkler, W.C., Breaker, R.R. (2003) Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell 113*, 577-586.

(18) Stoddard, C.D., Batey, R.T. (2006) Mix and match riboswitches. *ACS Chem Biol 1*, 751-754.

(19) Lee, E.R., Baker, J.L., Weinberg, Z., Sudarsan, N., Breaker, R.R. (2010) An allosteric self-splicing ribozyme triggered by a bacterial second messenger. *Science 329*, 845-848.

(20) Soukup, G.A., Breaker, R.R. (1999) Relationship between internucleotide linkage geometry and the stability of RNA. *RNA 5*, 1308-1325.

(21) Regulski, E.E., Breaker, R.R. (2008) In-line probing analysis of riboswitches. *Methods Mol Biol 419*, 53-67.

(22) Dennis, A.L., Puskas, M., Stasaitis, S., Sandwick, R.K. (2000) The formation of a 1–5 phosphodiester linkage in the spontaneous breakdown of 5-phosphoribosyl-α-1-pyrophosphate. *J Inorg Biochem 81*, 73-80.

(23) Meola, M., Yamen, B., Weaver, K., Sandwick, R.K. (2003) The catalytic effect of Mg^{2+} and imidazole on the decomposition of 5-phosphoribosyl-α-1-pyrophosphate in aqueous solution. *J Inorg Biochem 93*, 235-242.

(24) Wickiser, J.K., Winkler, W.C., Breaker, R.R., Crothers, D.M. (2005) The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Mol Cell 18*, 49-60.

(25) Wickiser, J.K., Cheah, M.T., Breaker, R.R., Crothers, D.M. (2005) The kinetics of ligand binding by an adenine-sensing riboswitch. *Biochemistry 44*, 13404-13414.

(26) Gilbert, S.D., Stoddard, C.D., Wise, S.J., Batey, R.T. (2006) Thermodynamic and kinetic characterization of ligand binding to the purine riboswitch aptamer domain. *J Mol Biol 359*, 754-768.

(27) Rieder, R., Lang, K., Graber, D., Micura, R. (2007) Ligand-induced folding of the adenosine deaminase A-riboswitch and implications on riboswitch translational control. *ChemBioChem 8*, 896-902.

(28) Batey, R.T., Gilbert, S.D., Montange, R.K. (2004) Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature 432*, 411-415.

(29) Serganov, A., Yuan, Y.R., Pikovskaya, O., Polonskaia, A., Malinina, L., Phan, A.T., Hobartner, C., Micura, R., Breaker, R.R., Patel, D.J. (2004) Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chemical Biology 11*, 1729-1741.

(30) Chen, A.G., Sudarsan, N., Breaker, R.R. (2011) Mechanism for gene control by a natural allosteric group I ribozyme. *RNA 17*, 1967-1972.

(31) Geiger, T., Wolz, C. (2014) Intersection of the stringent response and the CodY regulon in low GC gram-positive bacteria. *Int J Med Microbiol 304*, 150-155.

(32) Sonenshein, A.L. (2005) CodY, a global regulator of stationary phase and virulence in gram- positive bacteria. *Curr Op Microbiol 8*, 203-207.

(33) Kriel, A., Bittner, A.N., Kim, S.H., Liu, K., Tehranchi, A.K., Zou, W.Y., Rendon, S., Chen, R., Tu, B.P., Wang, J.D. (2012) Direct regulation of GTP homeostasis by (p)ppGpp: a critical component of viability and stress resistance *Mol Cell* 48, 231-241.

(34) Krásný, L., Tiserová, H., Jonák, J., Rejman, D., Sanderová, H. (2008) The identity of the transcription +1 position is crucial for changes in gene expression in response to amino acid starvation in *Bacillus subtilis*. *Mol Microbiol 69*, 42–54.

(35) Mandal, M., Breaker, R.R. (2004) Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nat Struct Mol Biol 11*, 29–35.

(36) Kim, J.N., Roth, A., Breaker, R.R. (2007) Guanine riboswitch variants from *Mesoplasma florum* selectively recognize 2'-deoxyguanosine. *Proc Natl Acad Sci USA 104*, 16092-16097.

(37) Weinberg, Z., Nelson, J.W., Lünse, C.E., Sherlock, M.E., Breaker, R.R. (2017) Bioinformatic analysis of riboswitch structures uncovers variant classes with altered ligand specificity. *Proc Natl Acad Sci USA 114*, 2077-2085.

(38) Blount, K.F. (2013) Methods for treating or inhibiting infection by *Clostridium difficile*. U.S. Patent 13/576,989.

(39) Regulski EE, Moy RH, Weinberg Z, Barrick, J.E., Yao, Z., Ruzzo, W.L., Breaker, R.R. (2008) A widespread riboswitch candidate that controls bacterial genes involved in molybdenum cofactor and tungsten cofactor metabolism. *Mol Microbiol 68*, 918-32.

(40) Nawrocki, E.P., Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics 29*, 2933-2935.

(41) Weinberg, Z., Kim, P.B., Chen, T.H., Li, S., Harris, K.A., Lünse, C.E., Breaker, R.R. (2015) New classes of self-cleaving ribozymes revealed by comparative genomics analysis. *Nat Chem Biol 11*, 606-610.

(42) Weinberg, Z., Breaker, R.R. (2011) R2R - software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinfo 12*, 3.

(43) Sherlock, M.E., Malkowski, S.N., Breaker, R.R. (2017) Biochemical validation of a second guanidine riboswitch class. *Biochemistry 56*, 352-358.

(44) Sudarsan, N., Wickiser, J.K., Nakamura, S., Ebert, M.S., Breaker, R.R. (2003) An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes Dev 17*, 2688-2697.

FIGURES

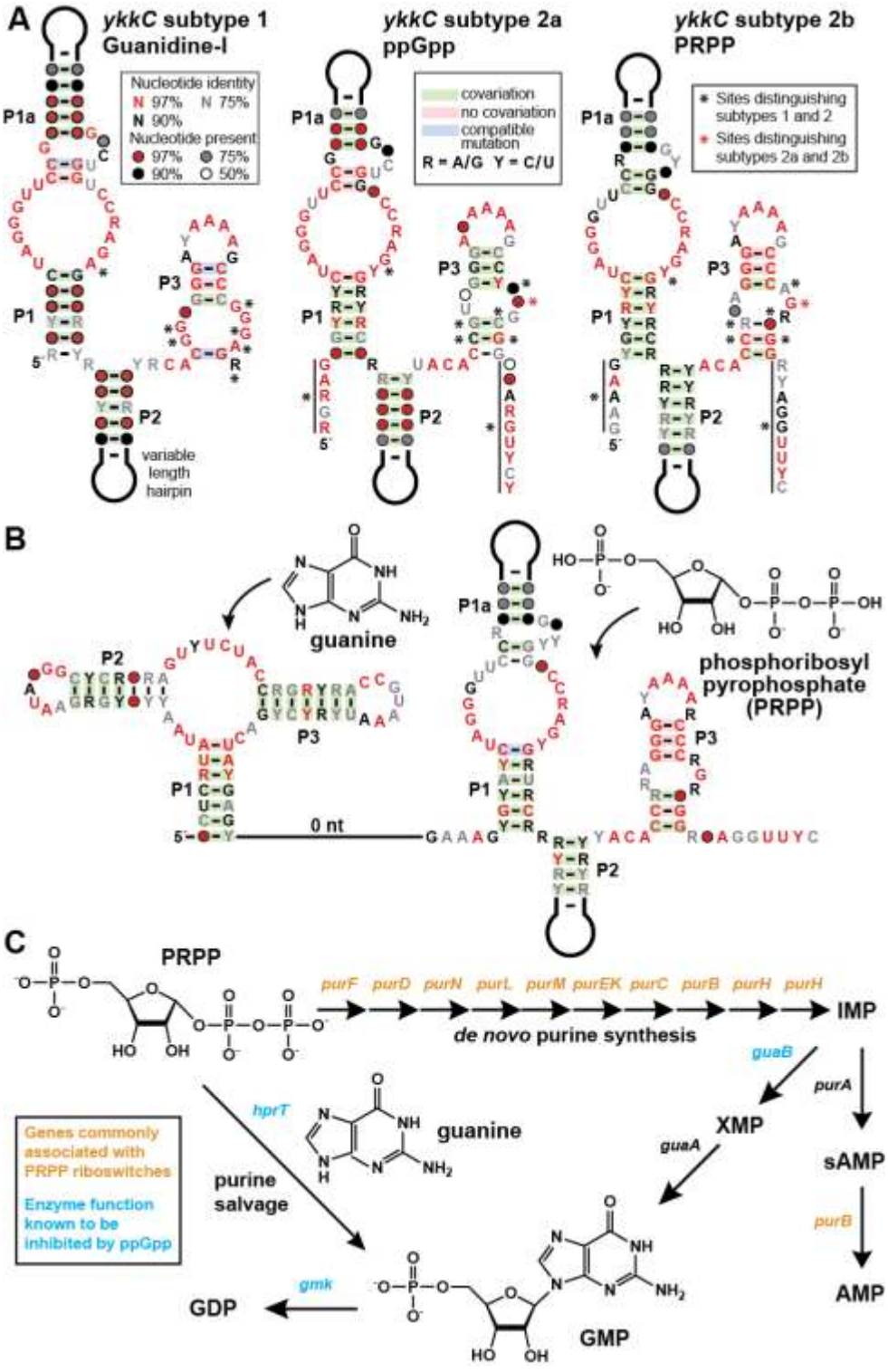
A
ykkC subtype 1
Guanidine-I
ykkC subtype 2a
ppGpp
ykkC subtype 2b
PRPP
Nucleotide identity
N 97% N 75%
N 90%
Nucleotide present
97% 75%
90% 50%
covariation
no covariation
compatible mutation
R = A/G Y = C/U
Sites distinguishing subtypes 1 and 2
Sites distinguishing subtypes 2a and 2b
P1a
P1
P2
P3
variable length hairpin
B
guanine
phosphoribosyl pyrophosphate (PRPP)
0 nt
C
PRPP
purF purD purN purL purM purEK purC purB purH purH
IMP
de novo purine synthesis
guaB
purA
XMP
sAMP
guaA
purB
AMP
hprT
guanine
purine salvage
Genes commonly associated with PRPP riboswitches
Enzyme function known to be inhibited by ppGpp
gmk
GDP
GMP

Figure 6-1. PRPP riboswitches and their regulatory network. (A) Comparison of the consensus sequence and secondary structure models for the guanidine-I riboswitch aptamer (subtype 1, left) and the aptamers for ppGpp (Chapter 5), and PRPP (subtype 2b, right). (B) Consensus sequence and secondary structure model for the 127 unique examples of adjacent guanidine-PRPP aptamers that form the corresponding tandem riboswitch architecture. (C) Only genes in the purine biosynthetic pathway that are necessary for the production of adenosine nucleotides are commonly associated with PRPP riboswitches. Shira Stav and I performed bioinformatic analyses used to create this figure.

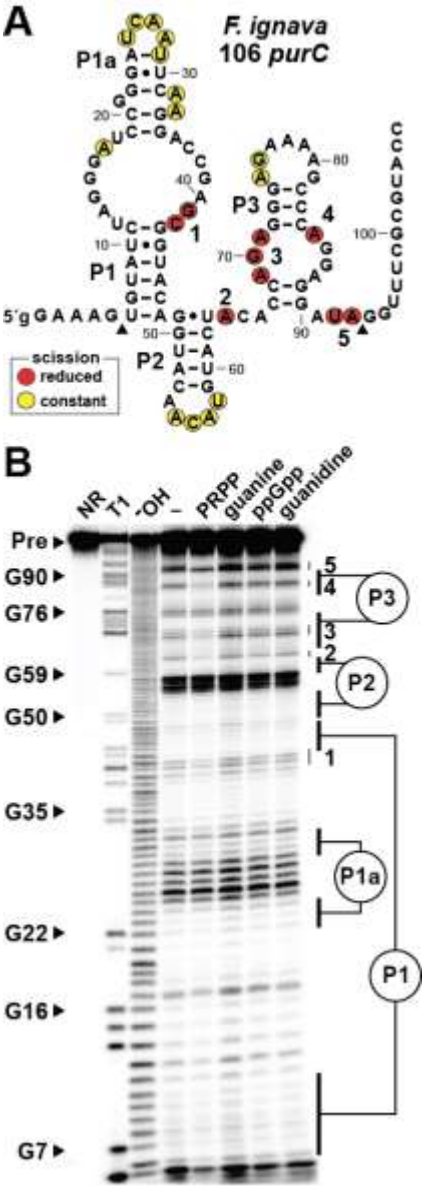

Figure 6-2. Ligand binding by a natural singlet PRPP riboswitch. (A) Sequence and secondary structure of the 106 nucleotide RNA derived from the *purC* gene of *F. ignava*. Data collected in b were used to determine regions of constant or reduced scission upon the addition of PRPP to in-line probing assays. Lowercase g indicates a guanosine nucleotide added to enable efficient transcription by T7 RNA polymerase. Arrowheads demark the approximate range of nucleotides with in-line probing data available as depicted in panel B. (B) Polyacrylamide gel electrophoresis (PAGE) analysis of the

spontaneous RNA cleavage products generated during in-line probing of 5′-^{32}P-labeled 106 *purC* RNA. NR, T1 and $^{-}$OH represent RNA undergoing no reaction, partial digest with RNase T1 (cleaves after guanosine nucleotides), or partial digest under alkaline conditions (cleaves after every nucleotide), respectively. Bands corresponding to selected G residues are annotated. In-line probing experiments contained either no ligand (–) or 1 mM of the compound indicated except guanine which was tested at 10 µM. Annotations 1 through 5 identify prominent regions of the RNA that undergo structural stabilization in a PRPP-dependent manner. Data shown are representative of multiple experiments. I performed the experiments shown in this figure.

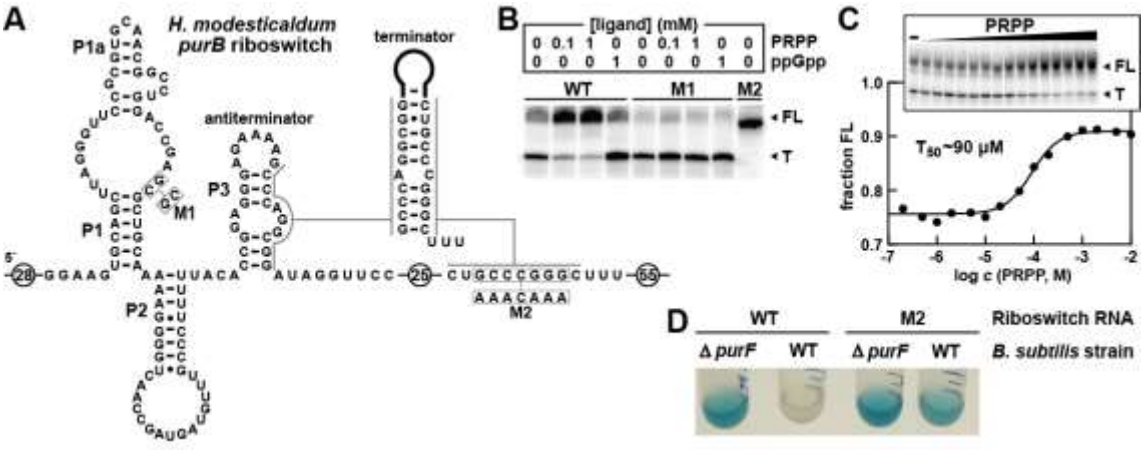

Figure 6-3. Transcription regulation and reporter gene expression by a natural singlet PRPP riboswitch. (A) Sequence and secondary structure of the natural PRPP singlet riboswitch derived from the *purB* gene from *H. modesticaldum*. An alternative RNA structure is depicted in which the terminator stem forms followed by a U-rich tract, which only forms in the absence of PRPP. (B) PAGE analysis of a single-round transcription termination assay of WT and mutant *H. modesticaldum purB* riboswitches at the indicated ligand concentration. FL and T denote full length and terminated transcripts, respectively. Values for the fraction of full-length transcripts relative to the total transcription yield is listed for each reaction. (C) Plot of the fraction of full length WT *H. modesticaldum purB* RNA riboswitch transcripts contributing to the total number of transcripts (FL plus T) as a function of the logarithm (base 10) of the molar PRPP concentration. The concentration of PRPP required to cause half-maximal termination efficiency (T_{50}) was determined by a sigmoidal curve fit. Inset: PAGE analysis of single round transcription termination assays of the WT *H. modesticaldum purB* RNA with either no ligand (–) or PRPP ranging from 200 nM to 10 mM. Data for replicates of this transcription termination assay are presented in Figure 6-4. (D) Reporter gene expression of WT and Δ*purF B. subtilis* cells containing wild-type (WT) or mutant (M2) *H. modesticaldum purB* riboswitch-*lacZ* reporter fusion

constructs as described in panel A. Cells were grown in glucose minimal medium (GMM) containing 50 µg mL^{-1} X-gal. I performed the experiments shown in panels A-C and Narasimhan Sudarsan performed the experiments shown in panel D.

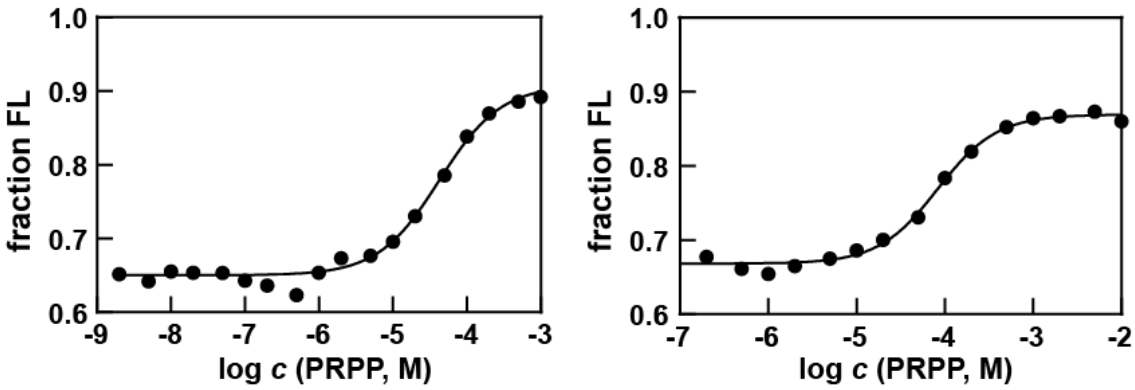

Figure 6-4. Replicate *in vitro* transcription termination assays. Termination assays with the *H. modesticaldum purB* riboswitch construct as depicted in Figure 6-3A. Assays are conducted as described for the data generated in Figure 6-3C. T_{50} values are estimated to be 40 μM (left) and 80 μM (right). I performed the experiments shown in this figure.

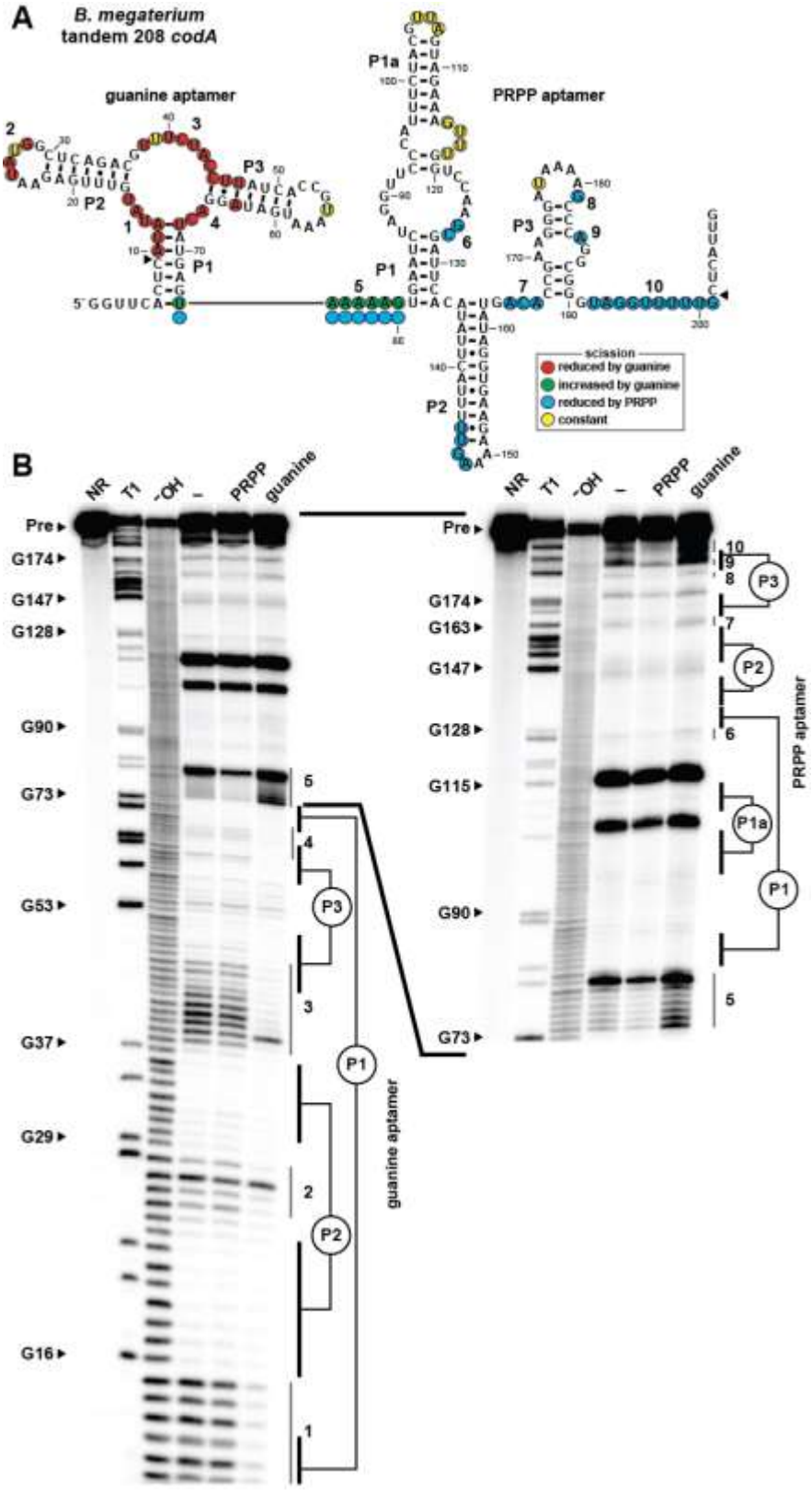
A
B. megaterium
tandem 208 codA
guanine aptamer
PRPP aptamer
P1a
P1
P2
P3
5' GGUUCA
scission
reduced by guanine
increased by guanine
reduced by PRPP
constant
B
NR
T1
-OH
–
PRPP
guanine
Pre
G174
G147
G128
G90
G73
G53
G37
G29
G16
G163
G115
guanine aptamer
PRPP aptamer

Figure 6-5. Ligand binding by tandem guanine and PRPP aptamers. (A) Sequence and secondary structure of the 208 nucleotide RNA derived from the *codA* gene of *B. megaterium*. Data collected in panel B were used to determine regions of constant, increased, or reduced scission upon the addition of guanine or PRPP to in-line probing assays. Additional annotations are as described for Figure 6-3A. Note that the precise locations of regions experiencing strand scission are approximated after nucleotide 74, as guided by the related in-line probing data generated for the singlet PRPP aptamer depicted in Figure 6-2. (B) PAGE analysis of the spontaneous RNA cleavage products generated during in-line probing of 5′-^{32}P-labeled 208 *codA* RNA. Annotations are described in the legend of Figure 6-2B. In-line probing experiments contained either no ligand (–), 1 mM PRPP or 10 μM guanine. Annotations 1 through 10 identify prominent regions of the RNA that undergo structural stabilization from either guanine or PRPP. Both gel images contain RNA from the same in-line probing reactions, but the gel in the image on the right was run for a longer time to increase resolution of the 3′ region of the RNA, which contains the regions of the PRPP aptamer involved in ligand recognition. I performed the experiments shown in this figure.

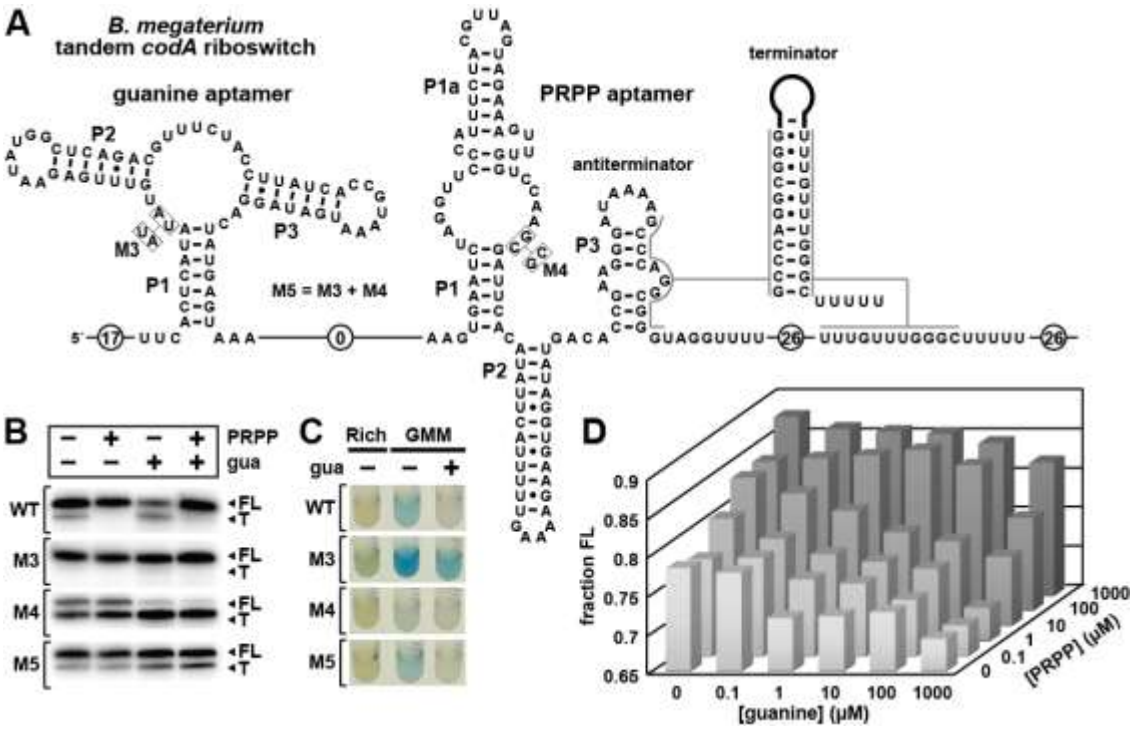

Figure 6-6. Guanine and PRPP competitively regulate transcript termination via a riboswitch integrating two aptamers. (A) Sequence and secondary structure of the tandem guanine-PRPP riboswitch derived from the *codA* gene of *B. megaterium*. An alternative RNA structure is depicted to include the terminator stem followed by a U-rich tract, which forms based on the relative abundance of each ligand. (B) PAGE analyses of single-round transcription termination assays of WT and various mutant *B. megaterium codA* tandem riboswitches in the presence or absence of the two ligands. FL and T denote full length and terminated transcripts, respectively. (C) Reporter gene expression of *B. subtilis* containing wild-type (WT) and mutant (M3 through M5) *B. megaterium* tandem riboswitch-*lacZ* reporter fusion constructs as described in panel A. Cells were grown in rich (lysogeny broth) medium or glucose minimal medium (GMM) containing either no (–) or 10 μM additional guanine (gua) and 50 μg mL^{-1} X-gal. **(**D) Plot of the fraction of full length *B. megaterium codA* tandem riboswitch transcripts contributing to the total number

of transcripts as a function of both guanine and PRPP concentrations. Data shown are the average of three experiments (see Figure 6-7). I performed the experiments shown in this figure.

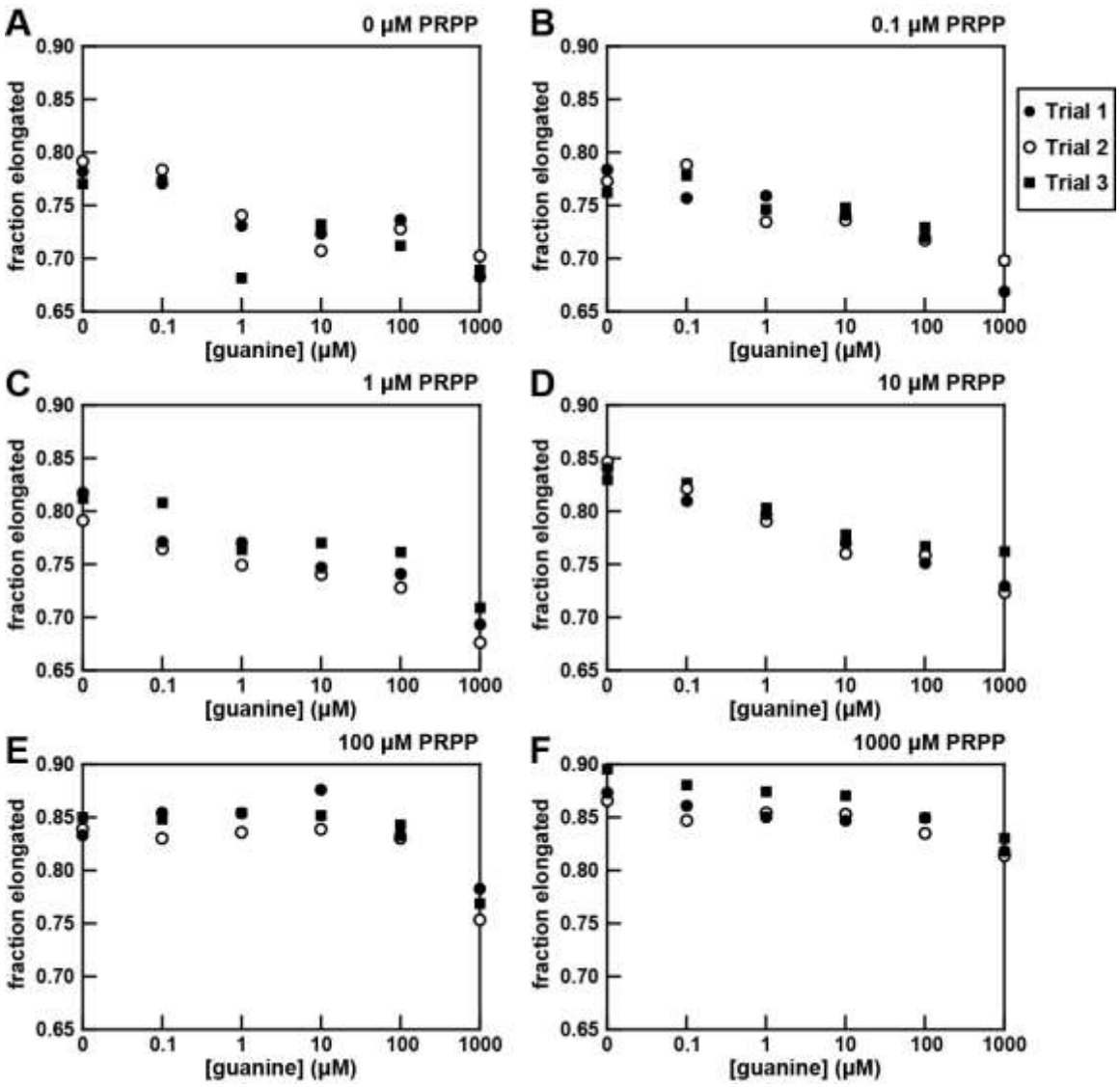

Figure 6-7. Comprehensive dataset for the in vitro transcription termination assays with the *B. megaterium codA* riboswitch construct. Data for three separate transcription termination 'trials' are plotted for six conditions tested as noted in panel A. Panels A through F include the PRPP concentrations used in the assays in the lower left of each plot. Details are as described in the legend to Figure 6-6. I performed the experiments shown in this figure.

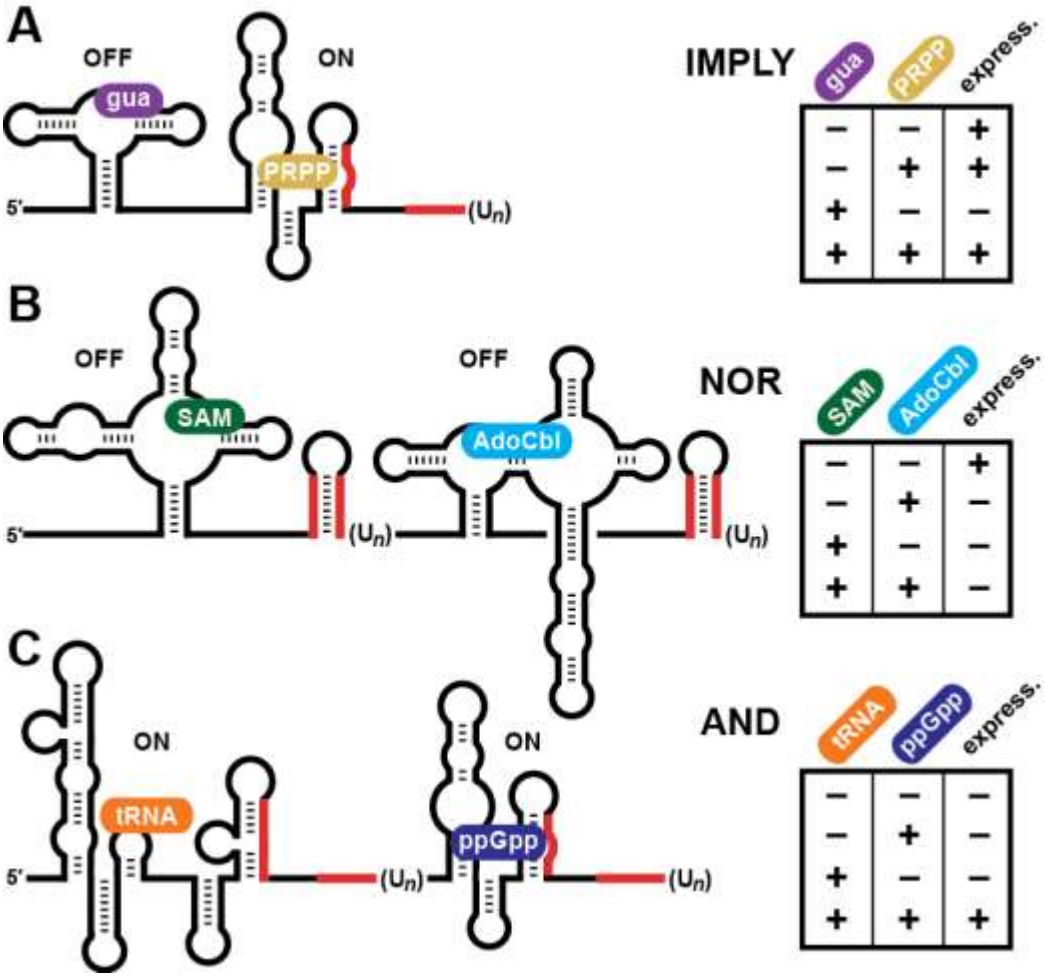

Figure 6-8. Natural examples of tandem riboswitches that approximate Boolean logic functions. (A) Tandem guanine and PRPP aptamers form an IMPLY logic gate to determine the state of downstream gene expression (express.). (B) Tandem *S*-adenosylmethionine (SAM) and adenosylcobalamin (AdoCbl) riboswitches form a NOR logic gate. (C) Tandem T box RNA and ppGpp riboswitches form an AND logic gate. The T box and ppGpp riboswitch elements can occur in either order.

TABLES

Table 6-1. Sequences of synthetic DNAs used in Chapter 6.

Name	Sequence	Annotation
MES18	TAATACGACTCACTATAGGTTCACTCATATATGTTTGAGAATATGGCTCAG	Forward PCR primer for transcription of the WT *B. megaterium* 208 *codA* RNA from genomic DNA
MES216	CAATGAGCAAAAACCTACCCGCC	Reverse PCR primer for transcription of the WT *B. megaterium* 208 *codA* RNA from genomic DNA
MES131	TAATACGACTCACTATAGGAAAGTGTATCTAGGGATCCGGAATCAATTCAAGGACCGAGCGGTACAGGTACAACATGTACTACACC	Forward template for transcription of the WT *F. ignava* 106 *purC* RNA
MES132	GGTACGCGAAACCTATCCTCCTGGGCTTTTCTCCCTCTGGTGTAGTACATGTTGTACC	Reverse template for transcription of the WT *F. ignava* 106 *purC* RNA
MES48	GCGCACATGAGAATTCCAGCGACGCTGTTGATCCTTTTAAATAAGTCTGATAAAATGTGAACTAATTGAGCGATCCTTTTTCAAAACTTGAGGGAAGTGCAGCTTAGGGTTCCGC	IDT G-block containing the wild-type *H. modesticaldum purB* riboswitch and controlled by the *B. subtilis lysC* promoter, for cloning into pDG1661 as

	GTGCAACGGCCTGGACCGAGCGC TGCAAAAAGGGGTCAACCGATG ATGTTTGCCCTTTTTACACCGGA GGGAGAAAAGCCCAGGCGGATA GGTTCCTTGCATTGGCAAGGCGG GAATCTATCTGCCCGGGCTTTGC TTTTTGCCAAAGGATCCCCAGCT GCGC	well as serving as a template for the dsDNA template for *in vitro* transcription termination
MES49	GCGCACATGAGAATTCCAGCGAC GCTGTTGATCCTTTTAAATAAGT CTGATAAAATGTGAACTAATTGA GCGATCCTTTTTCAAAACTTGAG GGAAGTGCAGCTTAGGGTTCCGC GTGCAACGGCCTGGACCGAGCGC TGCAAAAAGGGGTCAACCGATG ATGTTTGCCCTTTTTACACCGGA GGGAGAAAAAAACAGGCGGATA GGTTCCTTGCATTGGCAAGGCGG GAATCTAAATAAACGGGCTTTGC TTTTTGCCAAAGGATCCCCAGCT GCGC	IDT G-block containing the terminator mutation (M1) *H. modesticaldum purB* riboswitch and controlled by the *B. subtilis lysC* promoter, for cloning into pDG1661 as well as serving as a template for the dsDNA template for *in vitro* transcription termination
MES50	GCGCACATGAGAATTCCAGCGAC GCTGTTGATCCTTTTAAATAAGT	IDT G-block containing the mutation of a conserved

	CTGATAAAATGTGAACTAATTGAGCGATCCTTTTTCAAAACTTGAGGGAAGTGCAGCTTAGGGTTCCGCGTGCAACGGCCTGGACCGACGGCTGCAAAAAGGGGTCAACCGATGATGTTTGCCCTTTTTACACCGGAGGGAGAAAAGCCCAGGCGGATAGGTTCCTTGCATTGGCAAGGCGGGAATCTATCTGCCCGGGCTTTGCTTTTTGCCAAAGGATCCCCAGCTGCGC	nucleotide (M2) *H. modesticaldum purB* riboswitch and controlled by the *B. subtilis lysC* promoter, for cloning into pDG1661 as well as serving as a template for the dsDNA template for *in vitro* transcription termination.
MES51	ACATGAGAATTCCAGCGACGC	Forward primer for amplification of the *H. modesticaldum purB* reporter construct
MES52	AGCTGGGGATCCTTTGGC	Reverse primer for amplification of the *H. modesticaldum purB* reporter construct
MES202	CCAGCGACGCTGTTGATCCTTTTAAATAAGTCTGATAAAATGTGAACTAATTGAGAGATAGTTTTTGAAAAATTGAGGGAAGTGCAGCTTAGGG	Forward primer for the construction of all single-round transcription termination templates from *H. modesticaldum*

MES203	GCGCACATGAGAATTCCAGCGACGCTGCGCAGCTGGGGATCCTTTGGC	Reverse primer for the construction of all single-round transcription termination templates from *H. modesticaldum*
MES103	TACGACGAATTCCAAAAATAATGTTGATCCTTTTAAATAAGTCTGATAAAATGTGAACTAAGAAAAAGTTAAATACGGTT	Forward primer for the construction of *B. megaterium codA* reporter construct as well as single-round transcription termination templates
MES104	GCTCAGGATCCAAATAGACACATAAGAAAA	Reverse primer for the construction of *B. megaterium codA* reporter construct as well as single-round transcription termination templates
MES42	GCTAAATACGGTTCACTCATAATTGTTTGAGAATATGGCTCAG	Forward primer for construction of the *B. megaterium codA* reporter as well as single-round transcription termination template containing M3 via QuikChange mutagenesis
MES43	CTGAGCCATATTCTCAAACAATTATGAGTGAACCGTATTTAGC	Reverse primer for construction of the *B. megaterium codA*

		reporter as well as single-round transcription termination template containing M3 via QuikChange mutagenesis
MES44	CTACGTTAGTAGAAAGTTTGGTCCAACGGATTCACATATTCATTTTTGAAAAG	Forward primer for construction of the *B. megaterium codA* reporter as well as single-round transcription termination template containing M4 (and M5 using M3 as the source) via QuikChange mutagenesis
MES45	CTTTTCAAAAATGAATATGTGAATCCGTTGGACCAAACTTTCTACTAACGTAG	Reverse primer for construction of the *B. megaterium codA* reporter as well as single-round transcription termination template containing M4 (and M5 using M3 as the source) via QuikChange mutagenesis

CHAPTER SEVEN

Discovery and possible functions of the *ykkC* subtype 2c and 2d riboswitch candidates

INTRODUCTION

Approximately 30% of *ykkC* motif RNAs represent distinct variants from the guanidine-I riboswitches and these remaining candidate riboswitches were initially all grouped into *ykkC* subtype 2.[1] I experimentally determined that RNAs from this subtype do not respond to guanidine, and they could be sorted by their nucleotide identity at the position directly prior to the closing of the P1 stem. This nucleotide is an A in guanidine-I riboswitches, and typically a C or a U for subtype 2 RNAs.

Upon further investigation of the RNA sequences, the phylogeny of organisms containing these RNAs, and the genes associated with each instance of the RNA, I began to suspect that *ykkC* subtype 2 contained multiple riboswitch classes. With assistance from my colleague Shira Stav, I sorted each RNA sequence by its downstream gene association to create consensus models for RNAs within each gene association group (using a cutoff of 10 sequences as a minimum for a gene association to be considered). In some cases, I observed additional conservation upstream of the P1 stem consisting mostly of purine nucleotides, as well as downstream of the P3 helix, which have a general 'RRRYYYY' consensus. These features are absent from the consensus model for guanidine-I riboswitches and are also absent from certain subtype 2 variants. In total, I sorted these sequences into four candidate riboswitch classes called subtypes 2a through 2d.

At a cursory glance, all of these subtypes appear to be extremely similar because they share many of the same highly-conserved nucleotides. I was able to use the crystal structures[2,3] of representatives of the guanidine-I riboswitch class (subtype 1) to evaluate the importance and possible role of many of these nucleotides in each variant subtype. For subtypes 2a, 2b, and 2c, the additional conservation preceding the P1 stem and following

the P3 stem seemed likely to play a role in ligand recognition, but because these nucleotides are not conserved in guanidine-I riboswitches, I did not initially have a good understanding of their positioning relative to the rest of the aptamer. Recently, structural data has been collected that indicates some of these nucleotides have a role in phosphate sensing and form additional base pairs, termed the P0 stem (A. Knappenberger, C. Reiss, S. Strobel, personal communication).

After sorting these representatives by their gene associations, I noticed that sequences associated with branched-chain amino acid metabolism appeared highly similar to those associated with glutamate synthases and *natA* transporters. These were collectively termed *ykkC* subtype 2a, which I then experimentally verified as representatives of the ppGpp riboswitch class (Chapter 5). The RNAs of subtype 2b, which control expression of *de novo* purine biosynthesis genes, are also extremely similar to the sequences of ppGpp riboswitches. I subsequently experimentally validated (Chapter 6) that this riboswitch class responds to a highly similar ligand, phosphoribosyl pyrophosphate (PRPP).

The other two genes commonly controlled by *ykkC* subtype 2 RNAs that were included in these initial analyses encode nucleoside diphosphate linked to X (NUDIX) hydrolases, and a transporter domain of unknown function (DUF) associated with phosphonate utilization (phn_DUF6). The consensus models of RNAs with these gene associations, termed *ykkC* subtype 2c (Figure 7-1A) and 2d (Figure 7-1B), respectively, each have more prominent changes at nucleotide positions that are known to be involved in ligand recognition compared to the extremely subtle difference between ppGpp and PRPP riboswitches. Representative RNAs of these two classes do not exhibit structural

modulation in response to guanidine, PRPP or ppGpp (data not shown) and both still await ligand identification and validation.

SUBTYPE 2C RIBOSWITCH CANDIDATES

Structural features of *ykkC* subtype 2c RNAs. The consensus sequence and secondary structure model for the NUDIX hydrolase-controlling subtype 2c RNAs has the most similarity to those for PRPP and ppGpp riboswitches. Another similarity between subtypes 2a, 2b, and 2c is that all are found exclusively in bacteria of the phylum Firmicutes. In terms of structural similarities, the newly-identified P0 stem found in ppGpp and PRPP riboswitches is likely to form in subtype 2c RNAs, although the conservation immediately upstream of the P1 stem differs when comparing subtype 2c RNAs to ppGpp or PRPP riboswitches. The nucleotide position directly before the P1 stem, which is typically conserved as a G for ppGpp and PRPP riboswitches, is either a U or a G for subtype 2c RNAs. At the position between directly prior to the closing of the P1 stem, which was the nucleotide I initially identified that separates subtype 1 from subtype 2, the *ykkC* subtype 2c RNA has either a C or G. This is the only *ykkC* subtype for which a G at that position is commonly observed. Another position of interest is the conserved 'GCA' linker between the P2 and P3 stems. The conserved G in this region is commonly an A in some subtypes, and not conserved in others. Subtype 2c RNAs also have a considerable amount of conservation in the internal loop of the P3 stem, including the G which forms a long-range base pair observed for guanidine-I and PRPP riboswitches, but not ppGpp riboswitches. With only 46 unique representatives, some of this high degree of conservation might not be as meaningful as it is for the other subtypes which have many

more examples. However, even with few sequences the stems still exhibit a high level of covariation and non-conserved loops are completely variable, indicating that most nucleotides that are conserved are likely important for ligand recognition.

Genes associated with *ykkC* subtype 2c RNAs. NUDIX hydrolases are a broad family of proteins that cleave the phosphodiester bond between the α and β phosphates of various nucleotide substrates.[4,5] The founding member of this class of proteins is MutT, an enzyme involved in DNA damage repair which primarily acts on 8-oxo-2′-deoxyguanosine di- and triphosphate as its substrates.[5,6] The damaged DNA molecule is cleaved by the enzyme to produce the monophosphate version of the nucleotide. Because nucleoside monophosphates cannot be used as a substrate for DNA polymerase, cleavage by the NUDIX hydrolase removes these damaged nucleotides from the cellular nucleotide pool. Similar NUDIX hydrolase family proteins act on the di-and tri-phosphate versions of other types of damaged DNA and RNA bases, including 8-oxo-adenosine, 2-hydroxy-adenosine, and 5-hydroxy-cytosine.[5,7] NUDIX hydrolase family proteins act on other nucleotide-containing compounds including coenzymes or nucleotide sugars such as NADH, TPP, Ap_4A, and ADP-glucose.[5,7,8] Some NUDIX hydrolases display broad substrate specificity and act on a range of nucleoside di- and triphosphates. In general, NUDIX hydrolases are considered to be "housekeeping" enzymes that maintain a balance between various nucleotide pools.[5,9]

At first, I only grouped together the *ykkC* RNA examples upstream of genes encoding NUDIX hydrolases (42 examples in the RefSeq 76 database) as *ykkC* subtype 2c. Recall that I did not initially evaluate representatives upstream of genes that are only rarely (<10 examples) associated with a *ykkC* RNA because of the difficulty of creating a meaningful

consensus based on so few sequences. Upon analysis at a later time, I found four examples of *ykkC* subtype 2 RNAs upstream of a gene encoding a Haloacid dehalogenase-like (HAD) hydrolase, another protein superfamily that cleaves phosphoester bonds of various substrates including nucleotides.[10-12] I assigned these four examples to subtype 2c in part because of the similarities in downstream gene functions. However, the observation that these four HAD-like hydrolase associated sequences perfectly match the consensus sequence of those examples upstream of NUDIX hydrolase genes provided me with even more compelling evidence that these two gene associations belong to the same candidate riboswitch class.

Speculations on a ligand for the *ykkC* subtype 2c candidate riboswitch class. Based on the ligand identities for the similar subtypes 2a and 2b, as well as the known substrates for NUDIX and HAD-like hydrolases,[4-12] it is extremely likely that the ligand for subtype 2c is another nucleotide derivative. Analysis of the expression platform for subtype 2c RNAs revealed that this riboswitch candidate should turn on expression of the downstream gene based on predicted interactions with a downstream terminator stem, as is the case for guanidine-I,[1] ppGpp, and PRPP riboswitches. Therefore, a high concentration of this mystery nucleotide ligand turns on expression of a gene encoding a hydrolase protein. This hydrolase protein then breaks down the mystery nucleotide ligand into its monophosphate form. Based on this logic, I find it unlikely that the ligand will be a coenzyme or nucleotide signaling molecule. It seems improbable that the only genes associated with an aptamer for a coenzyme or signaling molecule would be those with the function to break down that compound. Signaling molecules have a much wider network of genes that they regulate[13,14]

and coenzymes typically shut off their own biosynthesis pathways when they reach sufficient levels instead of breaking down those molecules already synthesized.[15,16]

Instead, it seems more likely that the mystery nucleotide ligand is a compound that always needs to be degraded when it accumulates to high levels, such as a damaged nucleotide. Any nucleotide with undesirable modifications to its nucleobase induced by oxidative or other types of damage will not function properly in a DNA or RNA polymer. As such, these should immediately be degraded and removed from the available nucleotide pool, which matches the logic and proposed function of this candidate riboswitch class. Initial in-line probing experiments indicate that multiple representatives of *ykkC* subtype 2c RNA display structural modulation at positions within the aptamer known to form the binding pocket for other *ykkC* subtypes in response to nucleoside diphosphates (data not shown). These interactions appear to be specific, in that mutations to conserved nucleotides in the aptamer abolish binding. These RNAs exhibit the highest, albeit still modest, affinity for 2′-deoxyadenosine diphosphate. I expect this riboswitch candidate to display a better affinity for its true natural ligand, which could be a modified and/or damaged DNA nucleotide. However, this variant riboswitch candidate could generally sense the deoxynucleoside diphosphate pool, although if this is the case then the biological rationale is currently unclear.

SUBTYPE 2D RIBOSWITCH CANDIDATES

Structural features of *ykkC* subtype 2d RNAs. The consensus sequence and secondary structure model for the phn_DUF6-controlling subtype 2d RNAs has the most similarity to the guanidine-I riboswitch class. Unlike subtypes 2a through 2c,

representatives of *ykkC* subtype 2d are never found in Firmicute organisms and are instead exclusive to Proteobacteria and Actinobacteria. The conserved aptamer structure for *ykkC* subtype 2d RNAs is confined to the P1-P3 stems observed for guanidine-I riboswitches and there is no evidence for the formation of a P0 stem. Overall, a few alterations in the conservation or nucleotide identity at positions within the internal loops of the P1 and P3 stem distinguish guanidine-I riboswitches from *ykkC* subtype 2d RNAs.

Unlike all other classes of *ykkC* riboswitches, the subtype 2d riboswitch candidate appears to regulate at the level of translation instead of transcription. The aptamer sequence is extremely close to the start codon of the adjacent open reading frame (ORF) with no evidence or space for an intrinsic transcription terminator sequence. Based on the location of the aptamer, ligand binding should allow access of the ribosome to its binding site in a manner similar to guanidine-II and guanidine-III riboswitches.[17,18]

Genes associated with *ykkC* subtype 2d RNAs. The phn_DUF6[19] transporter proteins fall into the same drug/metabolite transporter (DMT) superfamily[20] as the small-multidrug resistance (SMR) transporters controlled by guanidine-I riboswitches, which were recently demonstrated to function as guanidine transporters.[1] The DUF6 domain was renamed EamA and transporters in this family act on a huge diversity of substrates.[21] In some cases, the guanidine transporters controlled by guanidine-I riboswitches are annotated as EamA, demonstrating the shared homology of these protein families. Unfortunately, the wide variety of possible substrates for this family of transporters makes it difficult to identify a single riboswitch ligand that would regulate their expression.

I later found another gene that is also associated with *ykkC* subtype 2d RNAs. In some rare instances, *ykkC* subtype 2 representatives are found upstream of a gene encoding a

transporter with homology to the TauE/SafE family of proteins.[22,23] I assigned these *ykkC* RNA representatives to subtype 2d because they match the consensus sequence as well as the phylogenetic distribution of that subtype, similar to the assignment of additional subtype 2c examples as described above. This second family of transporters associated with subtype 2d RNAs used to be known as DUF81 and were renamed TauE or SafE, which export sulfite and sulfoacetate.[22,23] Although there is still some uncertainty about the substrates for these transporters, the range of substrates for this family of transporters is much narrower than that of the EamA superfamily.

Speculations on a ligand for the *ykkC* subtype 2d candidate riboswitch class. Based on the nature of the genes associated with subtype 2d RNAs and the prediction that the expression of these genes should increase when the ligand is abundant in the cell, it is plausible that the ligand is toxic to cells. In a manner akin to guanidine and fluoride riboswitches and the transporters they each regulate, a high concentration of this mystery ligand would activate *ykkC* subtype 2d riboswitches in order to facilitate this compound's export. Based on the similarities in the consensus model between subtype 2d RNAs and guanidine-I aptamers as well as potential substrates for the TauE/SafE transporters, this mystery ligand is likely a relatively small, charged, toxic molecule. Some molecules that might fit this description are phosphite, sulfite, nitrite and their derivatives. Ongoing experiments are exploring these and other compounds as possible ligands for the *ykkC* subtype 2d candidate riboswitch class.

CONCLUSIONS

Overall, a variety of factors had to be overcome in order to finally solve the *ykkC* orphan riboswitch, although it still has partial lingering orphan status until the ligands for subtypes 2c and 2d have been validated. My finding that the *ykkC* RNA motif represents multiple riboswitch classes that sense such radically different ligand molecules was surprising and completely unique to all other known riboswitch classes. While variant aptamers had previously been known, the structures and identities of the ligands for the variant class had always been highly similar to the parent class.[24-30] To allow for recognition of such different compounds, the nucleotide changes in the binding pocket between the different *ykkC* subtypes are drastic. For example, only one out of the six nucleotides that directly interacts with the guanidine ligand in the guanidine-I RNAs[2,3] remains highly conserved in ppGpp or PRPP riboswitches. However, the changes in nucleotide identity and conservation that enable the formation of these unique binding pockets were initially undetectable among dozens of other highly-conserved nucleotides.

A huge proportion of the *ykkC* aptamer is conserved compared to its total size in comparison to other riboswitch class aptamers. It is now clear that the majority of these highly-conserved nucleotides facilitate the A-minor interactions between a portion of the P1 stem and the loop of the P3 stem.[2,3] A few additional nucleotides are conserved to form an S-turn in all subtypes except ppGpp riboswitches. The nucleotides conserved in the P0 stem of ppGpp and PRPP riboswitches are not conserved in guanidine-I riboswitches, which constitute the vast majority of *ykkC* motif RNAs, and therefore could not be identified bioinformatically until the sequences were manually separated and evaluated. Once the roles of these highly-conserved nucleotides were understood, the nucleotide

positions that vary between the subtypes became more prominent and obvious. The overall *ykkC* motif architecture appears to be an excellent basis for creating an RNA with the ability to specifically recognize a wide range of small molecules.

Due to the nature of the method used for sorting these sequences, many genes only rarely associated with a *ykkC* RNA were not assigned to a specific subtype. This allows for the possibility of even more classes of *ykkC* riboswitch variants exist that were too rare to be detected. Subtypes 2c and 2d likely each constitute a single riboswitch class, but the possibility remains for there to be additional variant aptamers within those subtypes. There could also be additional variant classes within subtype 1, which were all assumed to bind guanidine and thus were not thoroughly evaluated for variant aptamers. Perhaps there are even more surprises hidden within this incredible RNA scaffold that might only be revealed by more genomic sequence data and thorough future analysis.

Moving forward, there are a variety of lessons to be learned from solving *ykkC* RNAs that may or may not apply to the future validation efforts for other orphan riboswitch candidates. The first is to always question whether a gene annotation is accurate. This was the case for F^- transporters regulated by F^- riboswitches, which had been annotated as Cl^- transporters.[31] However, the incorrect annotation of the substrates for two entirely separate classes of proteins under the regulatory network of *ykkC* riboswitches severely hindered the validation effort. It certainly would not have taken over a decade to identify guanidine as the ligand for a riboswitch class that regulates genes encoding guanidine carboxylases and guanidine transporters, as opposed to urea carboxylases and multi-drug resistance transporters, respectively.

Of course, the knowledge of these gene functions was gained only after guanidine was identified as the riboswitch ligand through an unbiased approach. Aside from established biosynthesis pathways and some other protein systems that are particularly well characterized, it is extremely difficult to accurately determine which annotations for protein functions and substrates are accurate, especially for transporters and certain broad enzyme families. It will likely continue to be the case that riboswitch validation will help determine gene functions or specificities whereas determination of a protein function might only rarely contribute to orphan riboswitch validation.

A second lesson is to be aware of possible variant aptamers and test ligand hypotheses on RNAs associated with the specific genes that led to that particular ligand hypothesis. Both ppGpp and PRPP were tested for binding to *ykkC* motif RNAs, but always using the *Bacillus subtilis* representative RNA sequence, which falls into the category of guanidine-I riboswitches. Perhaps the *ykkC* motif is completely unique in its broad range of aptamers and their corresponding ligands, but many unsolved riboswitch candidates are upstream of a variety of genes with varying functions that could encompass separate riboswitch classes.

Another hindering factor was fortuitously avoided in validating the ppGpp and PRPP riboswitch classes in that the previous consensus model for all *ykkC* motif RNAs did not account for the entire aptamer structure for these subtypes. My decision to look for additional conservation upstream and downstream of what had been the entire aptamer for guanidine-I riboswitches allowed for identification of additional nucleotides vital to ligand binding. If the representative ppGpp and PRPP riboswitch RNA constructs had been designed to be the same length as those for guanidine-I riboswitches, ligand binding would not have been observed.

The initial consensus models for c-di-GMP[32] and SAM-VI[33] riboswitches were also missing important structural features and in both of these cases, the RNA constructs tested by in-line probing were also fortunately long enough to encompass the entire aptamer.[13,33] Defining the exact beginning and end points for an RNA motif can be challenging based solely on bioinformatic data, especially for riboswitch candidates with few representatives. As such, there is a distinct possibility that multiple orphan riboswitches have been tested with the correct cognate ligand, but binding was not observed because a portion of the aptamer was missing. While it might be prudent to test the entire noncoding region harboring a riboswitch candidate, the introduction of too many additional nucleotides, which could contain competing structures, increases the chances of misfolding *in vitro*. Although no single foolproof strategy exists for *in vitro* examination of orphan riboswitch candidates, the analysis of multiple constructs with varying amounts of additional flanking sequence might help eliminate this issue.

While the efforts to solve both the *ykkC*[1] and *yjdF*[34] riboswitch candidates both benefited from unbiased screening approaches using *in vivo* reporter systems, these are some of the only riboswitch classes out of the ~40 validated classes this approach could have worked for. Many of the molecules sensed by riboswitches are internally produced signaling molecules or metabolites. For a screening approach to work, the ligand must be able to enter the cells either by diffusion or through a dedicated transporter, and the correct ligand has to be included in the screen in the first place. Creating a comprehensive library of all possible metabolites, ions, and toxins would take a tremendous amount of effort compared to some more feasible, albeit lower throughput, alternatives.

Another strategy often suggested for ligand discovery involves attempting to pull out the ligand from a cellular extract using the aptamer RNA, then ascertaining the molecular identity of the ligand through mass spectrometry analysis. First, this approach relies on finding cellular conditions in which the ligand is present in cells, which is commonly the rate-limiting step in ligand identification. Even if these conditions are found and the ligand identity is still unclear with the knowledge of conditions that the cells have been subjected to for the ligand to be produced, successful isolation of the ligand compound through a pull down followed by assignment of the compound through mass spectrometry analysis would be a daunting experimental task.

Unfortunately, there is no "one size fits all" strategy for solving orphan riboswitch candidates because each class of these RNAs presents unique challenges. Some challenges are easy to immediately identify, such as the association with genes of unknown functions or technical challenges of performing experiments with the ligand of interest. Other challenges can only be realized in retrospect, such as missing portions of the aptamer sequence or the association with gene that are mis-annotated. Progress in solving orphan riboswitch candidates is slow and does not appear to be getting any faster, even with the knowledge gained from previous successes. Despite the frustrating nature of the experimental pursuit of unsolved riboswitch candidates, the reward in terms of the scientific knowledge revealed by each solved riboswitch class merits our continued work in this field.

In the case of the *ykkC* motif, the discoveries made through identifying its ligands have certainly warranted the years spent investigating these RNAs. The first biological receptors specific for guanidine were identified, including three riboswitch classes and two protein

families. These findings indicate the widespread presence of guanidine in bacteria when this compound had not previously been considered to be biologically relevant. The source of guanidine in cells as well as the characterization of pathways to diminish its toxic effects will likely be the subject of many future studies. The validation of most *ykkC*-I RNAs as guanidine riboswitches allowed for immediate validation of the mini-*ykkC* and *ykkC*-III motif RNAs as guanidine riboswitch classes II and III. Additional remaining orphan riboswitch candidates also have sufficient overlap in gene associations to suggest that they might sense the same ligand and pursuing these motifs might again lead to solving multiple riboswitch classes simultaneously.

The identification of subtype 2 *ykkC* RNAs as well as the validation of the ligands for some of those variant riboswitch classes have provided a new perspective into the adaptability of a single RNA scaffold. A riboswitch class was finally found for ppGpp, a common nucleotide signaling molecule that had long been postulated[30,35-37] as a riboswitch ligand. A riboswitch class for PRPP, another fundamental RNA derivative, was found to regulate RNA monomer biosynthesis, which has strong RNA World implications. The ppGpp and PRPP riboswitch classes are immensely intriguing in that their consensus models are nearly identical and yet each demonstrates specificity for its cognate ligand. Many examples of these *ykkC* RNAs also conspire with other structured RNAs to accomplish complex regulatory systems that mimic two-input logic gates[38] and create new types of logic function. The unsolved *ykkC* subtype 2c and 2d RNAs have the potential to reveal even more new and exciting biology in the future. Orphan riboswitch validation is difficult, but the impact of the findings resulting from the pursuit of the corresponding ligand for each orphan will continue to justify the effort.

REFERENCES

(1) Nelson, J.W., Atilho, R.M., Sherlock, M.E., Stockbridge, R.B., Breaker, R.R. (2017) Metabolism of free guanidine in bacteria is regulated by a widespread riboswitch class. *Mol Cell* 65, 220-230.

(2) Reiss, C.W., Xiong, Y., Strobel, S.A. (2017) Structural basis for ligand binding to the guanidine-I riboswitch. *Structure 25*, 195-202.

(3) Battaglia, R.A., Price, I.R., Ke, A. (2017) Structural basis for guanidine sensing by the *ykkC* family of riboswitches. *RNA 23*, 578-585.

(4) McLennan, A.G. (2006) The Nudix hydrolase superfamily. *Cell Mol Life Sci 63*, 123-143.

(5) Bessman, M.J., Frick, D.N., O'Handley, S.F. (1996) The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed, "housecleaning" enzymes. *J Biol Chem 271*, 25059-25062.

(6) Takagi, Y., Setoyama, D., Ito, R., Kamiya, H., Yamagata, Y., Sekiguchi, M. (2012) Human MTH3 (NUDT18) protein hydrolyzes oxidized forms of guanosine and deoxyguanosine diphosphates: comparison with MTH1 and MTH2. *J Biol Chem 287*, 21541-21549.

(7) Nguyen, V.N., Park, A., Xu, A., Srouji, J.R., Brenner, S.E., Kirsch, J.F. (2016) Substrate specificity characterization for eight putative nudix hydrolases. Evaluation of criteria for substrate identification within the Nudix family. *Proteins 84*, 1810-1822.

(8) Muñoz, F.J., Baroja-Fernández, E., Morán-Zorzano, M.T., Alonso-Casajús, N., Pozueta-Romero, J. (2006) Cloning, expression and characterization of a Nudix hydrolase that catalyzes the hydrolytic breakdown of ADP-glucose linked to starch biosynthesis in *Arabidopsis thaliana*. *Plant Cell Physiol 47*, 926-934.

(9) Fisher, D.I., Cartwright, J.L., Harashima, H., Kamiya, H., McLennan, A.G. (2004) Characterization of a nudix hydrolase from Deinococcus radiodurans with a marked specificity for (deoxy)ribonucleoside 5'-diphosphates. *BMC Biochem 5*, 7.

(10) Kuznetsova, E., Proudfoot, M., Gonzalez, C.F., Brown, G., Omelchenko, M.V., Borozan, I., Carmel, L., Wolf, Y.I., Mori, H., Savchenko, A.V., Arrowsmith, C.H., Koonin, E.V., Edwards, A.M., Yakunin, A.F. (2006) Genome-wide analysis of substrate specificities of the *Escherichia coli* haloacid dehalogenase-like phosphatase family. *J Biol Chem 281*, 36149-36161.

(11) Lu, Z., Dunaway-Mariano, D., Allen, K.N. (2005) HAD superfamily phosphotransferase substrate diversification: structure and function analysis of HAD subclass IIB sugar phosphatase BT4131. *Biochemistry 44*, 8684-8696.

(12) Caparrós-Martín, J.A., McCarthy-Suárez, I., Culiáñez-Macià, F.A. (2013) HAD hydrolase function unveiled by substrate screening: enzymatic characterization of *Arabidopsis thaliana* subclass I phosphosugar phosphatase AtSgpp. *Planta 237*, 943–954.

(13) Sudarsan, N., Lee, E.R., Weinberg, Z., Moy, R.H., Kim, J.N., Link, K.H., Breaker, R.R. (2008) Riboswitches in eubacteia sense the second messenger cyclic di-GMP. *Science 321*, 411-413.

(14) Nelson, J.W., Sudarsan, N., Furukawa, K., Weinberg, Z., Wang, J.X., Breaker, R.R. (2013) Riboswitches in eubacteria that sense the second messenger c-di-AMP. *Nat Chem Biol 9*, 834-839.

(15) Winkler ,W.C., Nahvi, A., Breaker, R.R. (2002) Thiamine derivatives that bind messenger RNAs directly to regulate bacterial gene expression. *Nature 419*, 952-956.

(16) Winkler, W.C., Nahvi, A., Sudarsan, N., Barrick, J.E., Breaker, R.R. (2003) An mRNA structure that controls gene expression by binding *S*-adenosylmethionine. *Nat Struct Biol 10*, 701-707.

(17) Sherlock, M.E., Malkowski, S.N., Breaker, R.R. (2017) Biochemical validation of a second guanidine riboswitch class in bacteria. *Biochemistry 56*, 352-358.

(18) Sherlock, M.E., Breaker, R.R. (2017) Biochemical validation of a third guanidine riboswitch class in bacteria. *Biochemistry 56*, 359–363.

(19) Villarreal-Chiu, J.F., Quinn, J.P., McGrath, J.W. (2012) The genes and enzymes of phosphonate metabolism by bacteria, and their distribution in the marine environment. *Front Microbiol 3*, 19.

(20) Jack, D.L., Yang, N.M., Saier, M.H. (2001) The drug/metabolite transporter superfamily. *Eur J Biochem 268*, 3620-3639.

(21) Västermark, Å., Almén, M.S., Simmen, M.W., Fredriksson, R., Schiöth, H.B. (2011) Functional specialization in nucleotide sugar transporters occurred through differentiation of the gene cluster EamA (DUF6) before the radiation of *Viridiplantae*. *BMC Evol Biol 11*, 123.

(22) Weinitschke, S., Denger, K., Cook, A.M., Smits, T.H. (2007) The DUF81 protein TauE in *Cupriavidus necator* H16, a sulfite exporter in the metabolism of C2 sulfonates. Microbiology 153, 3055-3060.

(23) Krejcík, Z., Denger, K., Weinitschke, S., Hollemeyer, K., Paces, V., Cook, A.M., Smits, T.H. (2008) Sulfoacetate released during the assimilation of taurine-nitrogen by *Neptuniibacter caesariensis*: purification of sulfoacetaldehyde dehydrogenase. *Arch Microbiol 190*, 159-68.

(24) Kim, J.N., Roth, A., Breaker, R.R. (2007) Guanine riboswitch variants from *Mesoplasma florum* selectively recognize 2'-deoxyguanosine. *Proc Natl Acad Sci USA 104*, 16092-16097.

(25) Weinberg, Z., Nelson, J.W., Lünse, C.E., Sherlock, M.E., Breaker, R.R. (2017) Bioinformatic analysis of riboswitch structures uncovers variant classes with altered ligand specificity. *Proc Natl Acad Sci USA 114*, 2077-2085.

(26) Regulski EE, Moy RH, Weinberg Z, Barrick, J.E., Yao, Z., Ruzzo, W.L., Breaker, R.R. (2008) A widespread riboswitch candidate that controls bacterial genes involved in molybdenum cofactor and tungsten cofactor metabolism. *Mol Microbiol 68*, 918-32.

(27) Nahvi, A., Sudarsan, N., Ebert, M.S., Zou, X., Brown, K.L., Breaker, R.R. (2002) Genetic control by a metabolite binding mRNA. *Chem Biol 9*, 1043-1049.

(28) Nahvi, A., Barrick, J.E., Breaker, R.R. (2004) Coenzyme B_{12} riboswitches are widespread genetic control elements in prokaryotes. *Nucleic Acid Res 32*, 143-150.

(29) Johnson, J.E., Reyes, F.E., Polaski, J.T., Batey, R.T. (2012) B_{12} cofactors directly stabilize an mRNA regulatory switch. *Nature 492*, 133-137.

(30) Nelson, J.W., Sudarsan, N., Phillip, G.E., Stav, S., Lünse, C.E., McCown, P.J., Breaker, R.R. (2015) Control of bacterial exoelectrogenesis by c-AMP-GMP. *Proc Natl Acad Sci USA 112*, 5389-5394.

(31) Baker, J.L., Sudarsan, N., Weinberg, Z., Roth, A., Stockbridge, R.B., Breaker, R.R. (2012) Widespread genetic switches and toxicity resistance proteins for fluoride. *Science 335*, 233-235.

(32) Weinberg, Z., Wang, J.X., Bogue, J., Yang, J., Corbino, K., Moy, R.H., Breaker, R.R. (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol 11*, 31.

(33) Arachchilage, G.M., Sherlock, M.E., Weinberg, Z., Breaker, R.R. (2017) SAM-VI RNAs selectively bind *S*-adenosylmethionine and exhibit similarities to SAM-III riboswitches. *RNA Biol* (in press).

(34) Li, S., Hwang, X.Y., Stav, S., Breaker, R.R. (2016) The yjdF riboswitch candidate regulates gene expression by binding diverse azaaromatic compounds. *RNA 22*, 530-541.

(35) Nelson, J.W., Breaker, R.R. (2017) The lost language of the RNA World. *Sci Signal 483*, eaam8812.

(36) Breaker, R.R. (2010) RNA second messengers and riboswitches: relics from the RNA World? *Microbe 5*, 13-20.

(37) Kim, P.B., Nelson, J.W., Breaker, R.R. (2015) An ancient riboswitch class in bacteria regulates purine biosynthesis and one-carbon metabolism. *Mol Cell 57*, 317-328.

(38) Sudarsan, N., Hammond, M.C., Block, K.F., Welz, R., Barrick, J.E., Roth, A., Breaker, R.R. (2006) Tandem riboswitch architectures exhibit complex gene control functions. *Science 13,* 300-304.

FIGURES

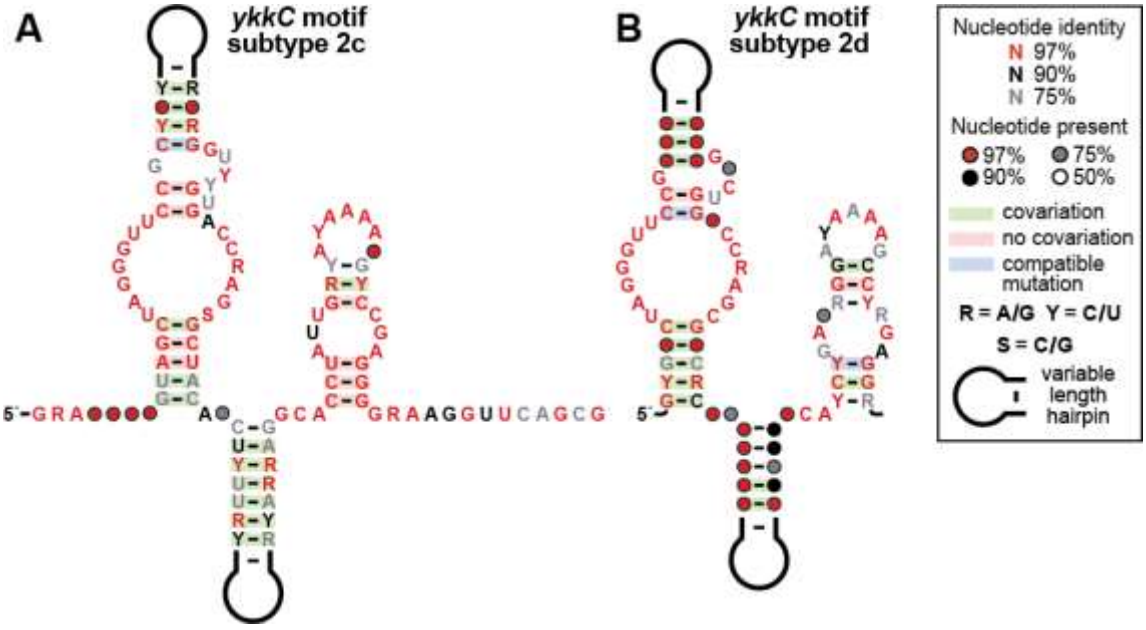

Figure 7-1. Additional unsolved variant *ykkC* motif RNAs. (A) Consensus sequence and secondary structure model for *ykkC* subtype 2c candidate riboswitches derived from 46 unique representatives. (B) Consensus sequence and secondary structure model for *ykkC* subtype 2d candidate riboswitches derived from 144 unique representatives.