Abstract

Novel Functions for Large Noncoding Nucleic Acids in Bacteria

Danielle Lee Widner

2021

Due to the compact nature of bacterial genomes, large, highly structured noncoding RNAs (ncRNAs) are rare, yet when present these large ncRNAs perform sophisticated biochemical functions, often as ribozymes. Even rarer are DNAs that adapt complex single-stranded structures, as natural opportunities for such structures to form and evolve are limited. Presented in this thesis is research on one large ncRNA and three highly structured, single-stranded DNAs (ssDNAs) that expands our knowledge of the functional capabilities of structured nucleic acids.

OLE (Ornate, Large, Extremophilic) RNA is among the largest and most widespread ncRNAs in bacteria without an ascribed biochemical function. The RNA and its protein partner OapA (OLE-associated protein A) are required in *Bacillus halodurans* for proper adaption to cold temperatures, short chain alcohols, and slightly elevated magnesium. Surprisingly, mutating a conserved DxxxD motif in OapA to AxxxA (named the PM1 strain) causes *B. halodurans* to become even more sensitive to these stresses but only when OLE RNA is also expressed. This phenotype allowed a genetic screen to reveal additional components to the OLE ribonucleoprotein (RNP) complex, leading to the discovery of YbzG (renamed OapB) as a second OLE RNA-interacting protein and the focus of Chapter 2. OapB is a small (11 kDa), previously uncharacterized protein with a putative RNA binding domain that binds to the P13 region of OLE RNA with subnanomolar affinity (K_D ~ 700 pM). This interaction

requires the presence of a GNRA tetraloop, though additional contacts are required, as a hairpin with a GNRA tetraloop is insufficient for complex formation. The regions of OLE RNA that OapA and OapB interact with do not overlap, suggesting that the three components can come together to form the OLE RNP complex. I collaborated in an effort led by Dr. Yang Yang to determine the crystal structure of OapB in complex with a subregion of OLE RNA resolved to 2.1 Å, confirming that OapB interacts with the P13 GNRA tetraloop and revealing additional contacts in the P13 and P12.2 stems of OLE RNA. In addition to being found in nearly all *ole*-containing organisms, OapB is found in 1,670 species that lack the *ole* gene. To answer the question of what target RNA sites might look like in organisms that lack *ole*, I generate a consensus model for OapB-RNA interactions through *in vitro* selection of mutagenized OLE RNA constructs. This model can be used by future researchers to uncover the function of OapB in species that lack *ole*.

In addition to examining the composition of the OLE RNP complex I performed RNA-seq to explore the effect of the OLE RNA on gene expression, highlighted in Chapter 3. These RNA-seq datasets of wild type (WT), Δ*ole-oapA*, and PM1 strains of *B. halodurans* grown under either standard or stressed conditions (24 °C, 3% EtOH w/v, or 5 mM $MgCl_2$) revealed that multiple metal ion transporters are differentially regulated between WT and Δ*ole-oapA* or PM1 strains. In addition to my transcriptomics data and the magnesium sensitivity phenotype, bioinformatic evidence also supports the idea that the OLE RNP complex is involved in magnesium homeostasis. The *B. halodurans oapA* gene is a homolog of the *Aeribacillus pallidus citMHS* gene, which encodes a magnesium/citrate symporter. If the OLE RNP complex is involved in magnesium homeostasis I hypothesize that one function of OLE RNA may be to act as a fine-tuned sensor for intracellular Mg^{2+},

switching from one conformation to another as Mg^{2+} concentrations move outside ideal intracellular ranges. RNA is known to undergo structural changes as magnesium concentrations increase, and in the case of magnesium riboswitches those structural changes represent biologically relevant on and off states. I also hypothesize that this conformational switching may regulate OapA. To test this hypothesis, I have conducted in-line probing experiments on OLE RNA at varying Mg^{2+} concentrations. While my initial results reveal that OLE RNA undergoes a conformation change at biologically relevant Mg^{2+} concentrations, significant additional work is needed to test whether or not this conformational change is indeed a real regulatory mechanism.

Chapter 4 moves away from OLE RNA and focuses on three extraordinarily structured ssDNAs, the HEARO (HNH Endonuclease-Associated RNA and ORF), IS605-*orfB*-I, and IS605-*orfB*-II motifs. Each of these motifs is associated with the IS605 superfamily of transposons, a class of mobile genetic elements that transpose via an obligatory ssDNA intermediate. Prior to my work HEARO was believed to function as an RNA. Through bioinformatics, I have established a connection between HEARO and IS605 transposons on both the protein and nucleic acid level, showing that the motif almost certainly functions as an ssDNA. Furthermore, I have determined the phylogenetic distribution and frequency of HEARO elements per genome. For all three motifs I have compared ssDNA structure at the 5' and 3' ends of the transposons and analyzed the domain architecture of the TnpB homologs in comparison to canonical IS605 TnpB. Combined, these motifs represent three of the most complex natural ssDNAs and provide powerful insight into the evolution of such motifs.

Novel Functions for Large Noncoding Nucleic Acids in Bacteria

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Danielle Lee Widner

Dissertation Director: Ronald R. Breaker

December 2021

Table of Contents

Acknowledgements

First, I would like to thank Dr. Ron Breaker for his invaluable advice and support throughout my time in graduate school. Ron is one of those scientists who understands the rules of biology so fundamentally that he has successfully predicted what should exist and what should be possible on basic principles alone. I learned from him how to think critically about why a system evolves the way it does and how to ask and answer the most pressing questions.

I would like to thank Dr. Joan Steitz and Dr. Matt Simon for serving as members of my thesis committee. Their advice and feedback have helped push my research forward on multiple occasions.

I would like to thank the members of the Breaker lab, past and present, for the endless helpful conversations, both about research and not. In particular, I would like to thank Kim Harris for being an excellent leader of the OLE team, for always treating everyone on the team as valued contributors, and for carefully listening and providing feedback to every OLE-related hypothesis. Thank you to Adam Roth and Narasimhan Sudarsan for sharing your years of experience in RNA biology and for your ability to always point me in the direction of the right paper. Thank you to the other members of the OLE team, Yang Yang, Seth Lyon, and Chrishan Fernando, as well as to Harini Sadeeshkumar, Neil White, Sarah Malkowski, Lukas Corey, Aya Narunsky, Glenn Garfield, Cynthia Megyola, Keith Corbino, and Megan Fris.

I have had extraordinary mentors throughout my life. My 7th grade science teacher Marybeth Marx was there for me at one of the most difficult times in my life. We had both lost someone important and she would stay after the school day had ended to talk with me

about life. I am grateful beyond words for the encouragement and support she offered to me then and has continued to offer me in the years since.

Thank you to Dr. Nate Bowling my undergraduate research advisor for seeing that I was capable of working independently and for pushing me to take ownership of my projects.

I would like to thank the friends who have encouraged, challenged, and supported me along the way and the friends who became “framily” for keeping me motivated by asking when I was going to be a doctor every time they saw me.

Thank you to my family for always asking about my research even when they barely knew what I am studying.

List of Figures

List of Tables

Chapter One

Introduction: Large Noncoding Nucleic Acids in Bacteria

Bioinformatics as a tool for the discovery of structured nucleic acids

Comparative sequence analysis is a powerful tool for discovery of structured ncRNAs in prokaryotes. Bacteria and Archaea have compact genomes with far less intergenic sequence space than Eukaryotes, making it possible to systematically search through intergenic regions (IGRs) for novel ncRNAs. The ncRNAs uncovered by this method vary greatly in size, distribution, and function. They can be members of established RNA classes such as riboswitches, T-box RNAs, attenuators, or self-cleaving ribozymes or represent previously unknown classes of ncRNAs. The same methods that reveal ncRNAs can also uncover structured DNA motifs, though these are rarer, and few structured DNAs have verified functions (Bikard et al. 2010, Weinberg et al. 2017). Several different approaches have successfully generated these motifs, as listed below:

1) *Linage and environmental strategies*- This approach involves pooling bacterial IGRs by phylogenetic linage or source in the case of environmental metagenomes. It is particularly useful in identification of narrowly distributed ncRNAs (Weinberg et al. 2017).
2) *Large IGRs*- Structured ncRNAs are overrepresented in large IGRs. To take advantage of this natural enrichment, IGRs exceeding 600 nucleotides are pooled and searched for homologous sequences.
3) *Clustering by downstream genes*- For certain ncRNAs, such as riboswitches, the identity of downstream genes may be conserved in a large portion of examples. Through identification of genes likely to be associated with ncRNAs this targeted search method can dramatically reduce the number of IGRs that must be analyzed before finding a ncRNA (Weinberg et al. 2017).

4) *Proximity to genetic elements of interest*- Like searching through 'clustering by downstream genes,' searching by 'proximity to genetic elements of interest' takes advantage of the tendency of certain structured ncRNAs to exist near specific genes and domains. However, with this method the specific location of the structured ncRNA in relation to the genetic element of interest is more flexible. Therefore, this method looks at all IGRs within 6 kilobases of a target genetic element. Searching for ncRNAs through proximity to genetic elements of interest has been particularly fruitful in revealing novel self-cleaving ribozymes (Li et al. 2015, Harris et al. 2015).

5) *GC-IGR*- The GC-IGR method takes advantage of two common characteristics of ncRNAs, that they typically have high guanosine and cytidine (GC) content and that they reside within large IGRs. When all IGRs from a bacterial genome are plotted by IGR length versus percentage GC content, IGRs containing known ncRNAs cluster together (Meyer et al. 2009, Stav et al. 2019). Analyzing only the IGRs within that cluster enriches for novel ncRNAs.

Once a potential motif is identified, it is refined with the comparative sequence algorithms CMfinder (Yao et al. 2006) and Infernal 1.1 (Nawrocki and Eddy. 2013). Refined motifs are evaluated based on conservation of primary and secondary structure, phylogenic distribution, and gene context. Methods for functional validation vary greatly depending on the predicted class of ncRNA (or ssDNA).

Large noncoding RNAs

In Eukaryotes long noncoding RNAs (lncRNAs) are involved primarily in regulation of genes, loci, or entire chromosomes (Statello et al. 2021). As gene regulation is less complex in Prokaryotes due to the compact nature of their genomes and the extreme rarity of cellular differentiation, the RNAs involved in such processes are different. The regulatory RNAs of bacteria, such as riboswitches, T-box RNAs, sRNAs, and attenuators, are typically less than 200 nucleotides in length (Zhang and Ferre-D'Amare. 2015, Hoe et al. 2013, Weinberg et al. 2017). Bacteria do however encode a handful of large, highly structured noncoding RNAs (ncRNAs), many of which perform essential roles (Harris and Breaker. 2018) These large ncRNAs include well-studied classes such as 23S and 16S rRNA, RNase P, tmRNA, and Group I and II introns. Other classes such as GOLLD, ROOL, raiA, MISL, HOLDH, IMES-1, and Bacteriodales-2 have little known about them, some having only been observed through bioinformatics (Weinberg et al. 2017). One class, T-Large RNA, is a variation on a Group II self-slicing intron that produces a circularly permuted product (A. Roth, Z. Weinberg, K. Vanderschuren, M. H. Murdock, E. Poiata, and R. R. Breaker, unpublished data). Although a biochemical function for T-Large RNA has been established, its biological role remains unknown. While large ncRNAs are rare in bacteria their breadth of functions is wide and they frequently represent novel role for RNA. Studying these ncRNAs will likely reveal sophisticated and varied biochemical mechanisms that will expand our knowledge of RNA's capabilities.

OLE RNA

Of the bacterial large ncRNAs with an unknown function, OLE (Ornate, Large, Extremophilic) RNA is the most widespread and among the largest. Nearly 800 examples

of OLE RNA have been identified in environmental metagenomes and a wide variety of extremophilic Firmicutes, including species that are thermophilic, halophilic, alkalophilic, anaerobic, and solvent tolerant (Puerta-Fernandez et al. 2006). OLE RNA contains an exceptionally high degree of primary and secondary structure conservation. It contains 96 nucleotides that are at least 97% conserved and 24 pairing elements that demonstrate covariation in its over 600 nucleotides of sequence, suggests that it performs a complex biological role.

The *ole* gene is found in a conserved operon containing genes for isoprenoid biosynthesis, DNA repair, coenzyme metabolism, rRNA transcription and maturation, and arginine biosynthesis (Harris et al. 2018). The most notable feature of this operon is that in species where the operon contains the *ole* gene, and only in those species, a second novel gene is also present (Puerta-Fernandez et al. 2006). The role of that gene, named *oapA* (OLE-associated protein A), was presumed to be tied to the function of OLE RNA, and indeed, OapA forms a membrane bound dimer that localizes OLE RNA to the cell membrane, an extremely unusual feature for a large ncRNA (Block et al. 2010). The only other large noncoding bacterial RNA to do so is SRP RNA, which recruits the ribosome to the cellular membrane for the translation of secretory and transmembrane proteins (Herskovits et al. 2000).

In an attempt to uncover why this large ncRNA localizes to the cell membrane, *ole*, *oapA*, or *ole-oapA* were knocked out of *Bacillus halodurans*. Broad phenotype testing revealed that these knockout strains become sensitive to cold, short chain alcohols, and slightly elevated magnesium (Mg^{2+}) (Wallace et al. 2012, Harris et al. 2019). Surprisingly, these phenotypes are made more severe by a mutation to an invariable DxxxD motif

(mutated to AxxxA) in OapA. This phenotype, called PM1 (protein mutant 1), only occurs if OLE RNA is also expressed and does not lead to any additional sensitivities (Harris et al. 2018). The heightened severity of the PM1 phenotype allowed a genetic screen identifying mutants that alleviated extreme sensitivity to cold (Harris et al. 2018). From this screen a second OLE RNA-interacting protein was discovered, OapB (OLE-associated protein B), which is discussed in Chapter 2.

While studies of OapB will be important for future experiments involving reconstitution of the OLE ribonucleoprotein (RNP) complex, no immediate insights into function were gained from these studies. To more directly address the question of function I turned to RNA-seq, analyzing transcripts from wild type (WT), Δ*ole-oapA,* and PM1 *B. halodurans* strains under standard and stressed growth conditions. The results, discussed in Chapter 3, showed that multiple metal ion transporters were differentially regulated between WT and strains with a disrupted OLE RNP complex (Δ*ole-oapA* and PM1). Two of the differentially regulated genes were Mg^{2+} importers, adding to the evidence that the OLE RNP complex might play a role in Mg^{2+} homeostasis. In addition to being one of the three observed sensitivities when *ole* and/or *oapA* are knocked out, OapA has been bioinformatically linked to Mg^{2+} transporters. Although OapA has no close homologs, it does contain a DUF21 domain and the protein that OapA shares the greatest homology with is CitMHS, a Mg^{2+} and citrate symporter (Harris et al. 2019). Other proteins containing the DUF21 domain and notable sequence similarity to OapA are also Mg^{2+} transporters (Akanuma et al. 2014, Armitano et al. 2016, Harris et al. 2019). In Chapter 3 I further explore the connection between the OLE RNP complex and Mg^{2+} homeostasis and

present preliminary work investigating a potential mechanism for regulation of intracellular Mg^{2+} concentration by OLE RNA.

Structured single-stranded DNAs in Bacteria

While it has long been understood that RNA molecules can form complex secondary and tertiary structures, the ability of single-stranded DNA (ssDNA) to fold into a variety of biologically important forms has received far less attention. Because DNA is typically found in the iconic double helical structure, opportunities for the evolution of sophisticated ssDNAs are limited. There are however several circumstances in which DNA exists in a single-stranded form for a prolonged time, such as the transposition of certain mobile elements, the replication of rolling-circle plasmids, and the genomes of ssDNA viruses. While these mobile elements, plasmids, and viruses are highly enriched for structured DNAs, there are also classes of structured DNAs found within genomes. Within the genome ssDNAs can be found as G-quadraplexes, *attC* sites, and msDNA. Described below are the most common ssDNA motifs, as well as a few rarer motifs that exhibit exceptional structural complexity.

G-quadruplexes:

Guanine quadruplexes are a non-B form DNA structure that requires four repeats of two or more guanine resides each. Under favorable conditions the four repeats use Hoogsteen base pairing to form a stacked quadruplex. G-quadruplexes are common in Eukaryotic and Prokaryotic organisms and can function at the RNA or DNA level.

Within Prokaryotes the most well studied system of DNA G-quadruplexes is in *Neisseria meningitidis* and *Neisseria gonorrhoeae*, where they are implicated in initiating antigenic variation in flanking coding sequence through recombination with silent loci throughout the genome (Sechman et al. 2005, Cahoon et al. 2009, Cahoon and Seifert. 2013, Prister et al. 2020). In this system, transcription of an sRNA allows the quadruplex to form by opening up the dsDNA (Prister et al. 2020). Formation of the G-quadruplex then initiates a nonreciprocal, homologous recombination process (Meyer et al. 1984, Haas and Meyer. 1986).

In genera other than *Neisseria*, G-quadruplexes have been shown to have important albeit understudied roles. In *Deinococcus radiodurans* G-quadruplex formation appears to be involved in upregulation of certain radioresistance genes (Beaume et al. 2013). In *Escherichia coli* G-quadruplexes may also influence gene expression, as shown by a study that introduced a GFP gene with a G-quadruplex into the promotor region of several different locations in the genome (Holder and Hartig. 2014). Position in the genome and whether the gene was leading or lagging strand affected expression differently when the G-quadruplex was present (Holder and Hartig. 2014). The apparent differences in function between species suggests that G-quadruplexes are capable of performing myriad roles that can be tailored to fit the needs of a specific species.

attC sites:

attC sites are essential features of integrons, a highly adaptable system composed of an array of interchangeable genes referred to as cassettes. Integrons allow bacteria to store, shuffle, acquire, and excise open reading frames (ORFs) on demand. The system

contains the integrase IntI, a tyrosine recombinase that excises and inserts cassettes, an *attI* site, where the new and shuffled cassettes are integrated into the genome, a promoter (P_C), and anywhere from one to over 200 gene cassettes (Escudero et al. 2015). In addition to containing a gene each cassette includes a 3' *attC* site that is recognized by IntI (MacDonald et al. 2006). The structure of the *attC* site is typically a cruciform but can include more complex variations, such as a Y-shaped motif consisting of three pairing elements (Stokes et al. 1997, Weinberg et al. 2017).

Integrons are common, occuring in ~17% of sequenced bacteria (Cambray et al. 2010) and found at very high frequencies in diverse environmental samples, with one report showing 4000-18000 unique cassettes per 0.3 g of soil (Ghaly et al. 2019). This is unsurprising given that pairwise competition assays have shown class 1 integrons have a low fitness-cost in *E. coli* (Lacotte et al. 2017).

As a component of the bacterial adaptive immune response, integrons are often the most rapidly evolving part of the genome, partially because genes closest to the promoter are the most strongly expressed (Collis and Hall. 1995). In line with this, the SOS response drastically upregulates integron recombination (Guerin et al. 2009), as does biofilm formation combined with the stringent response (Strugeon et al. 2016).

msDNA:

Originally discovered as a high copy number satellite DNA in *Myxococcus xanthus*, multi-copy single-stranded DNA (msDNA) was soon revealed to be a structurally unique hybrid of covalently linked ssDNA and RNA (Yee et al. 1984, Furuichi et al. 1987 a & b). Synthesis of msDNA requires reverse transcriptase (RT), msr (the RNA component), and

msd (the ssDNA component), which are encoded in a genetic cassette referred to as a retron. The *msr* and *msd* genes overlap in a manner that allows msd to be reverse transcribed from the 3' end of the msr transcript. This is initiated by RT covalently linking the RNA and ssDNA via 2'-5' phosphodiester linkage at the priming guanosine (Dhundale et al. 1987, Shimamoto et al. 1995). As the ssDNA is synthesized its complementary RNA is degraded.

In addition to the unusual 2'-5' phosphodiester linkage of msDNA, the molecule contains secondary structure in both the ssDNA and RNA components. Typically, the ssDNA portion of the molecule contains a single highly stable stem-loop, and the RNA portion contains two shorter stem-loops. Though there are some notable deviations from the norm, such as Retron-Eco4 (also known as Ec83), which produces a purely ssDNA product (Lampson et al. 1990), and Retron-Sen1 (Se72) where the 2'-5' linkage is not seen and the final product is a dsDNA (Rychlik et al. 2001, Pilousova et al. 2011). With only a small fraction of the hundreds of predicted retrons experimentally validated (Simon and Zimmerly. 2008, Toro et al. 2014, Zimmerly et al. 2015), it is likely that additional atypical configurations remain to be discovered.

The function of msDNA remains unclear, with current evidence suggesting it plays a role in environmental adaptation. In *E. coli* overexpression of msDNA is known to increase the mutation rate (Maas et al. 1994), but only when there is a mismatch in the ssDNA base-pairing element (Maas et al. 1996 & Mao et al. 1996). Given that levels of msDNA increase when *E. coli* faces starvation conditions, it might act as part of the SOS response (Maas et al. 1994). Other work has found that *E. coli* with knockouts of retron genes have decreased intestinal persistence in mice (Elfenbein et al. 2015). Similarly, in

Salmonella Typhimurium deletion of retrons led to differences in gene expression under anaerobic conditions and left the bacteria unable to colonize mouse intestines (Elfenbein et al. 2015). Others have noted that production of msDNA can repress certain carbon utilization genes (Jeong et al. 2004). These potential functions merit further investigation, with study in a wider number of bacteria required to understand if these regulatory networks are broadly distributed or species specific.

Rolling circle plasmids:

The two primary modes of replication in prokaryotic organisms are theta and rolling circle replication. Of these, only rolling circle replication has favored the evolution of multiple DNA structures to facilitate its own propagation. Rolling-circle plasmids contain both a double-stranded and single-stranded origin of replication. The double-stranded origin of replication consists of a short inverse palindrome, while the single-stranded origin is composed of at least two paired elements and exceeds 100 nucleotides in length (Khan et al. 1997).

ARRPOF:

In addition to origins of replication a small number of rolling-circle plasmids contain an ssDNA motif known as ARRPOF (Area Required for Replication in a Plasmid Of Fusobacteria) between the double-stranded and single-stranded origins of replication (Weinberg et al. 2017). Averaging over 200 nucleotides and with at least three pseudoknots, ARRPOF is one of the most complex ssDNA motifs documented (Weinberg

et al. 2017). However, to our knowledge no biochemical or genetic experiments on this ssDNA have yet been attempted.

The IS605 superfamily of transposons:

The IS605 superfamily of insertion sequences are the smallest known transposable elements. They transpose via an obligatory ssDNA intermediate in a method known as 'peel-and-paste.' Peel-and-paste transposition requires TnpA, an HUH family transposase that catalyzes excision, circularization, and reinsertion of the transposon (Ton-Hoang et al. 2005, Lee et al. 2006, Guynet et al. 2008, Hickman et al. 2010). TnpA recognizes the left (5') and right (3') ends of the insertion sequence through two palindromic sequences that form imperfect hairpins during DNA replication (Ronning et al. 2005, Ton-Hoang et al. 2005, Ton-Hoang et al. 2010). In most IS605 superfamily transposons, hairpins at the 5' and 3' ends of the element are the only ssDNA structures present. However, comparative sequence analysis has revealed three separate exceptionally structured motifs associated with IS605 superfamily transposons. These motifs, HEARO (HNH endonuclease-associated RNA and ORF), IS605-*orfB*-I, and IS605-*orfB*-II, are some of the most structurally complex ssDNAs ever documented. They vary in location, with HEARO found exclusively at the 5' end and IS605-*orfB*-I and IS605-*orfB*-II found at the 3' end of their respective transposons. They also have significant differences in composition of associated ORFs. Each of these motifs is discussed in detail in Chapter 4.

Prospects for Discovery of Additional ssDNA Motifs

As genomic and metagenomic databases grow, the potential to find rare nucleic acid structures increases. Although new ssDNA motifs could be difficult to distinguish from ncRNA motifs, this also lends them to being discovered through the same comparative genomics searches that produce novel ncRNA motifs. In fact, it is very likely that additional ssDNA motifs have already been discovered but have been mistaken for structured ncRNAs. This will be particularly challenging for rare motifs found in non-model microbes and environmental metagenomes, yet the massive increase in metagenomic data will also allow motifs that occur in a limited range of species to be detected. Beyond providing valuable information about basic biology, pathogenicity (G-quadruplexes), antibiotic resistance (*attC* sites), and evolution of invaluable systems (HEARO and Cas9, see Chapter 4), ssDNAs can be utilized for numerous practical applications. For example, retrons have been manipulated to create genomically encoded analog memory (Farzadfard et al. 2014) and have been used in tandem with Cas9 to increase the rate of homology-directed repair (Sharon et al. 2018, Simon et al. 2018). With the amount of basic science to be learned and the breadth of practical applications, it is past time that scientists start paying more attention to the ability of DNA to form complex single-stranded structures.

References

Akanuma, G., Kobayashi, A., Suzuki, S., Kawamura, F., Shiwa, Y., Watanabe, S., Yoshikawa, H., Hanai, R., and Ishizuka, M. (2014) Defect in the formation of 70S ribosomes caused by lack of ribosomal protein L34 can be suppressed by magnesium. *J Bacteriol* **196**, 3820-3830

Armitano, J., Redder, P., Guimaraes, V. A., and Linder, P. (2016) An Essential Factor for High Mg(2+) Tolerance of Staphylococcus aureus. *Front Microbiol* **7**, 1888

Beaume, N., Pathak, R., Yadav, V. K., Kota, S., Misra, H. S., Gautam, H. K., and Chowdhury, S. (2013) Genome-wide study predicts promoter-G4 DNA motifs regulate selective functions in bacteria: radioresistance of D. radiodurans involves G4 DNA-mediated regulation. *Nucleic Acids Res* **41**, 76-89

Bikard, D., Loot, C., Baharoglu, Z., and Mazel, D. (2010) Folded DNA in action: hairpin formation and biological functions in prokaryotes. *Microbiol Mol Biol Rev* **74**, 570-588

Block, K. F., Puerta-Fernandez, E., Wallace, J. G., and Breaker, R. R. (2011) Association of OLE RNA with bacterial membranes via an RNA-protein interaction. *Mol. Microbiol***. 79**, 21-34

Cahoon, L. A., and Seifert, H. S. (2009) An alternative DNA structure is necessary for pilin antigenic variation in Neisseria gonorrhoeae. *Science* **325**, 764-767

Cahoon, L. A., and Seifert, H. S. (2013) Transcription of a cis-acting, noncoding, small RNA is required for pilin antigenic variation in Neisseria gonorrhoeae. *PLoS Pathog* **9**, e1003074

Cambray, G., Guerout, A. M., and Mazel, D. (2010) Integrons. *Annu Rev Genet* **44**, 141-166

Collis, C. M., and Hall, R. M. (1995) Expression of antibiotic resistance genes in the integrated cassettes of integrons. *Antimicrob Agents Chemother* **39**, 155-162

Dhundale, A., Lampson, B., Furuichi, T., Inouye, M., and Inouye, S. (1987) Structure of msDNA from Myxococcus xanthus: evidence for a long, self-annealing RNA precursor for the covalently linked, branched RNA. *Cell* **51**, 1105-1112

Elfenbein, J. R., Knodler, L. A., Nakayasu, E. S., Ansong, C., Brewer, H. M., Bogomolnaya, L., Adams, L. G., McClelland, M., Adkins, J. N., and Andrews-Polymenis, H. L. (2015) Multicopy Single-Stranded DNA Directs Intestinal Colonization of Enteric Pathogens. *PLoS Genet* **11**, e1005472

Escudero, J. A., Loot, C., Nivina, A., and Mazel, D. (2015) The Integron: Adaptation On Demand. *Microbiol Spectr* **3**, MDNA3-0019-2014

Farzadfard, F., and Lu, T. K. (2014) Synthetic biology. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* **346**, 1256272

Furuichi, T., Dhundale, A., Inouye, M., and Inouye, S. (1987) Branched RNA covalently linked to the 5' end of a single-stranded DNA in Stigmatella aurantiaca: structure of msDNA. *Cell* **48**, 47-53

Furuichi, T., Inouye, S., and Inouye, M. (1987) Biosynthesis and structure of stable branched RNA covalently linked to the 5' end of multicopy single-stranded DNA of Stigmatella aurantiaca. *Cell* **48**, 55-62

Ghaly, T. M., Geoghegan, J. L., Alroy, J., and Gillings, M. R. (2019) High diversity and rapid spatial turnover of integron gene cassettes in soil. *Environ Microbiol* **21**, 1567-1574

Guerin, E., Cambray, G., Sanchez-Alberola, N., Campoy, S., Erill, I., Da Re, S., Gonzalez-Zorn, B., Barbe, J., Ploy, M. C., and Mazel, D. (2009) The SOS response controls integron recombination. *Science* **324**, 1034

Guynet, C., Hickman, A. B., Barabas, O., Dyda, F., Chandler, M., and Ton-Hoang, B. (2008) In vitro reconstitution of a single-stranded transposition mechanism of IS608. *Mol Cell* **29**, 302-312

Haas, R., and Meyer, T. F. (1986) The repertoire of silent pilus genes in Neisseria gonorrhoeae: evidence for gene conversion. *Cell* **44**, 107-115

Harris, K. A., Lunse, C. E., Li, S., Brewer, K. I., and Breaker, R. R. (2015) Biochemical analysis of pistol self-cleaving ribozymes. *RNA* **21**, 1852-1858

Harris, K. A., and Breaker, R. R. (2018) Large Noncoding RNAs in Bacteria. *Microbiol Spectr* **6**

Harris, K. A., Zhou, Z., Peters, M. L., Wilkins, S. G., and Breaker, R. R. (2018) A second RNA-binding protein is essential for ethanol tolerance provided by the bacterial OLE ribonucleoprotein complex. *Proc. Natl. Acad. Sci. USA* **115**, E6319-E6328

Harris, K.A., Odzer, N.B., and Breaker, R.R. (2019) Disruption of the OLE ribonucleoprotein complex causes magnesium toxicity in *Bacillus halodurans*. *Mol. Microbiol.* **112**, 1552-1563

Herskovits, A. A., Bochkareva, E. S., and Bibi, E. (2000) New prospects in studying the bacterial signal recognition particle pathway. *Mol Microbiol* **38**, 927-939

Hickman, A. B., James, J. A., Barabas, O., Pasternak, C., Ton-Hoang, B., Chandler, M., Sommer, S., and Dyda, F. (2010) DNA recognition and the precleavage state during single-stranded DNA transposition in D. radiodurans. *EMBO J* **29**, 3840-3852

Hoe, C. H., Raabe, C. A., Rozhdestvensky, T. S., and Tang, T. H. (2013) Bacterial sRNAs: regulation in stress. *Int J Med Microbiol* **303**, 217-229

Holder, I. T., and Hartig, J. S. (2014) A matter of location: influence of G-quadruplexes on Escherichia coli gene expression. *Chem Biol* **21**, 1511-1521

Jeong, M. A., and Lim, D. (2004) A proteomic approach to study msDNA function in Escherichia coli. *J Microbiol* **42**, 200-204

Khan, S. A. (1997) Rolling-circle replication of bacterial plasmids. *Microbiol Mol Biol Rev* **61**, 442-455

Lacotte, Y., Ploy, M. C., and Raherison, S. (2017) Class 1 integrons are low-cost structures in Escherichia coli. *ISME J* **11**, 1535-1544

Lee, H. H., Yoon, J. Y., Kim, H. S., Kang, J. Y., Kim, K. H., Kim, D. J., Ha, J. Y., Mikami, B., Yoon, H. J., and Suh, S. W. (2006) Crystal structure of a metal ion-bound IS200 transposase. *J Biol Chem* **281**, 4261-4266

Li, S., Lunse, C. E., Harris, K. A., and Breaker, R. R. (2015) Biochemical analysis of hatchet self-cleaving ribozymes. *RNA* **21**, 1845-1851

Maas, W. K., Wang, C., Lima, T., Hach, A., and Lim, D. (1996) Multicopy single-stranded DNA of Escherichia coli enhances mutation and recombination frequencies by titrating MutS protein. *Mol Microbiol* **19**, 505-509

Maas, W. K., Wang, C., Lima, T., Zubay, G., and Lim, D. (1994) Multicopy single-stranded DNAs with mismatched base pairs are mutagenic in Escherichia coli. *Mol Microbiol* **14**, 437-441

MacDonald, D., Demarre, G., Bouvier, M., Mazel, D., and Gopaul, D. N. (2006) Structural basis for broad DNA-specificity in integron recombination. *Nature* **440**, 1157-1162

Mao, J. R., Inouye, S., and Inouye, M. (1996) Enhancement of frame-shift mutation by the overproduction of msDNA in Escherichia coli. *FEMS Microbiol Lett* **144**, 109-115

Meyer, M. M., Ames, T. D., Smith, D. P., Weinberg, Z., Schwalbach, M. S., Giovannoni, S. J., and Breaker, R. R. (2009) Identification of candidate structured RNAs in the marine organism 'Candidatus Pelagibacter ubique'. *BMC Genomics* **10**, 268

Meyer, T. F., Billyard, E., Haas, R., Storzbach, S., and So, M. (1984) Pilus genes of Neisseria gonorrheae: chromosomal organization and DNA sequence. *Proc Natl Acad Sci U S A* **81**, 6110-6114

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

Pilousova, L., and Rychlik, I. (2011) Retron Se72 utilizes a unique strategy of the self-priming initiation of reverse transcription. *Cell Mol Life Sci* **68**, 3607-3617

Prister, L. L., Ozer, E. A., Cahoon, L. A., and Seifert, H. S. (2019) Transcriptional initiation of a small RNA, not R-loop stability, dictates the frequency of pilin antigenic variation in Neisseria gonorrhoeae. *Mol Microbiol* **112**, 1219-1234

Prister, L. L., Yin, S., Cahoon, L. A., and Seifert, H. S. (2020) Altering the Neisseria gonorrhoeae pilE Guanine Quadruplex Loop Bases Affects Pilin Antigenic Variation. *Biochemistry* **59**, 1104-1112

Puerta-Fernandez, E., Barrick, J. E., Roth, A., and Breaker, R. R. (2006) Identification of a large noncoding RNA in extremophilic eubacteria. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19490-19495

Ronning, D. R., Guynet, C., Ton-Hoang, B., Perez, Z. N., Ghirlando, R., Chandler, M., and Dyda, F. (2005) Active site sharing and subterminal hairpin recognition in a new class of DNA transposases. *Mol Cell* **20**, 143-154

Rychlik, I., Sebkova, A., Gregorova, D., and Karpiskova, R. (2001) Low-molecular-weight plasmid of Salmonella enterica serovar Enteritidis codes for retron reverse transcriptase and influences phage resistance. *J Bacteriol* **183**, 2852-2858

Sechman, E. V., Rohrer, M. S., and Seifert, H. S. (2005) A genetic screen identifies genes and sites involved in pilin antigenic variation in Neisseria gonorrhoeae. *Mol Microbiol* **57**, 468-483

Sharon, E., Chen, S. A., Khosla, N. M., Smith, J. D., Pritchard, J. K., and Fraser, H. B. (2018) Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell* **175**, 544-557 e516

Shimamoto, T., Inouye, M., and Inouye, S. (1995) The formation of the 2',5'-phosphodiester linkage in the cDNA priming reaction by bacterial reverse transcriptase in a cell-free system. *J Biol Chem* **270**, 581-588

Simon, A. J., Morrow, B. R., and Ellington, A. D. (2018) Retroelement-Based Genome Editing and Evolution. *ACS Synth Biol* **7**, 2600-2611

Simon, D. M., and Zimmerly, S. (2008) A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res* **36**, 7219-7229

Statello, L., Guo, C. J., Chen, L. L., and Huarte, M. (2021) Gene regulation by long non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol* **22**, 96-118

Stav, S., Atilho, R. M., Mirihana Arachchilage, G., Nguyen, G., Higgs, G., and Breaker, R. R. (2019) Genome-wide discovery of structured noncoding RNAs in bacteria. *BMC Microbiol* **19**, 66

Stokes, H. W., O'Gorman, D. B., Recchia, G. D., Parsekhian, M., and Hall, R. M. (1997) Structure and function of 59-base element recombination sites associated with mobile gene cassettes. *Mol Microbiol* **26**, 731-745

Strugeon, E., Tilloy, V., Ploy, M. C., and Da Re, S. (2016) The Stringent Response Promotes Antibiotic Resistance Dissemination by Regulating Integron Integrase Expression in Biofilms. *mBio* **7**

Ton-Hoang, B., Guynet, C., Ronning, D. R., Cointin-Marty, B., Dyda, F., and Chandler, M. (2005) Transposition of ISHp608, member of an unusual family of bacterial insertion sequences. *EMBO J* **24**, 3325-3338

Ton-Hoang, B., Pasternak, C., Siguier, P., Guynet, C., Hickman, A. B., Dyda, F., Sommer, S., and Chandler, M. (2010) Single-stranded DNA transposition is coupled to host replication. *Cell* **142**, 398-408

Toro, N., and Nisa-Martinez, R. (2014) Comprehensive phylogenetic analysis of bacterial reverse transcriptases. *PLoS One* **9**, e114083

Wallace, J. G., Zhou, Z., and Breaker, R. R. (2012) OLE RNA protects extremophilic bacteria from alcohol toxicity. *Nucleic Acids Res.* **40**, 6898-6907

Weinberg, Z., Perreault, J., Meyer, M. M., and Breaker, R. R. (2009) Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. *Nature* **462**, 656-659

Weinberg, Z., Lünse, C. E.,Corbino, K. A., Ames, T. D., Nelson, J. W., Roth, A., Perkins, K. R., Sherlock, M. E., and Breaker, R. R. (2017) Detection of 224 candidate structured RNAs by comparative analysis of specific subsets of intergenic regions. *Nucleic Acids Res*. **45**, 10811-1082

Yao, Z., Weinberg, Z., and Ruzzo, W. L. (2006) CMfinder--a covariance model based RNA motif finding algorithm. *Bioinformatics* **22**, 445-452

Yee, T., Furuichi, T., Inouye, S., and Inouye, M. (1984) Multicopy single-stranded DNA isolated from a gram-negative bacterium, Myxococcus xanthus. *Cell* **38**, 203-209

Zhang, J., and Ferre-D'Amare, A. R. (2015) Structure and mechanism of the T-box riboswitches. *Wiley Interdiscip Rev RNA* **6**, 419-433

Zimmerly, S., and Wu, L. (2015) An Unexplored Diversity of Reverse Transcriptases in Bacteria. *Microbiol Spectr* **3**, MDNA3-0058-2014

Chapter Two

Structural basis of high-affinity association of OapB with OLE RNA

This chapter is adapted with permission from Widner DL, Harris KA, Corey L, Breaker RR. (2020). *Bacillus halodurans* OapB forms a high-affinity complex with the P13 region of OLE noncoding RNA. *J. Biol. Chem.* RA120. 012676. Portions are also adapted with permission from Yang Y, Harris KA, Widner DL, Breaker RR. (2021). Structural basis of high-affinity association of OapB with OLE RNA and insights into assembly of the OLE ribonucleoprotein complex. *Proc Natl Acad Sci U S A*
Author Contributions: DLW, KAH, YY, and RRB designed the experiments. DLW,

YY, and LC preformed experiments. DLW, YY, KAH, and RRB analyzed data. DLW,

KAH, YY, and RRB wrote the manuscripts.

Summary

Noncoding RNAs (ncRNAs) longer than 200 nucleotides are rare in bacteria, likely because bacterial genomes are under strong evolutionary pressures to maintain a small genome size. Of the long noncoding RNAs (ncRNAs) unique to bacteria, the OLE (ornate, large, extremophilic) RNA class is among the largest and most structurally complex. OLE RNAs form a ribonucleoprotein (RNP) complex by partnering with at least two proteins, OapA and OapB, that directly bind OLE RNA. The biochemical functions of the OLE RNP complex remain unknown but are required for proper adaptation to certain environmental stresses, such as cold temperatures, short chain alcohols, and high magnesium concentrations. Although the vast majority of bacteria with the *ole* gene also have the *oapB* gene, there are many whose genomes contain *oapB* but lack *ole*, suggesting that OapB has other RNA partners in some species that might exhibit similar structural features. In the current study, we used electrophoretic mobility shift assays (EMSAs) to examine the binding of OLE RNA fragments by OapB and found that OapB recognizes a small sub-region of OLE RNA, including stem P13, with a dissociation constant (K_D) of approximately 700 pM. Analyses with mutated RNA constructs, and the application of *in vitro* selection, revealed that strong binding of OLE RNA by OapB requires a stem containing a precisely located single-nucleotide bulge and a GNRA tetraloop. To provide the atomic details of OapB-OLE RNA interaction and to potentially reveal previously uncharacterized protein-RNA interfaces, Dr. Yang Yang determined the structure of OapB from *Bacillus halodurans* alone and in complex with an OLE RNA fragment at resolutions of 1.0 Å and 2.1 Å, respectively. The structure of OapB exhibits a K-shaped overall architecture wherein its conserved KOW motif and additional unique structural elements

of OapB form a bipartite RNA-binding surface that docks to the P13 hairpin and P12.2 helix of OLE RNA. These high-resolution structures elucidate the molecular contacts used by OapB to form a stable RNP complex and explain the high conservation of sequences and structural features at the OapB-OLE RNA binding interface. These findings provide new insights into the role of OapB in the assembly and biological function of OLE RNP complex and can guide the exploration of additional possible OLE RNA-binding interactions present in OapB.

Introduction

Noncoding RNAs (ncRNAs) greater than 200 nucleotides in length are rare in the bacterial domain of life (Harris et al. 2018, Cech and Steitz 2014), perhaps because most bacterial species are under extreme evolutionary pressure to maintain small, efficient genomes with little tolerance for producing long transcripts other than mRNAs. Intriguingly, the common classes of large and highly conserved bacterial ncRNAs reported to date mostly represent RNA-based enzymes such as ribosomal RNAs (Nissen et al. 2000), group I (Kruger et al. 1982) and group II (Peebles et al. 1986) self-splicing ribozymes, and RNase P endonucleases (Guerrier-Takada et al. 1983). It has been proposed that some ribozyme classes might be evolutionary descendants of catalytic polymers that were present in RNA World organisms, before the emergence of proteins (Benner et al. 1989). Thus, reports of the existence of unusually well-conserved and long ncRNAs in bacteria (Puerta-Fernandez et al. 2006, Weinberg et al. 2009, Weinberg et al. 2017) provide opportunities to study novel ribozyme classes with potentially ancient types of biochemical functions.

Previous studies reported the existence of OLE (ornate, large, extremophilic) RNAs, which represent a class of unusually long (~600 nucleotide) bacterial ncRNA transcripts with numerous highly conserved sequence and structural features (Puerta-Fernandez et al. 2006, Wallace et al. 2012, Harris et al. 2018). The presence of nucleotide sequences matching the OLE RNA consensus sequence and secondary structure model has been detected in the genomes of 296 Gram-positive bacterial species, and in several hundred additional metagenomic DNA sequences (Block et al. 2011). This distribution among various Firmicutes, particularly in extremophilic and anaerobic species of Bacilli and Clostridia, makes OLE RNA one of the most widespread large ncRNAs in bacteria whose biochemical function has yet to be determined.

Although the precise biological function of OLE RNA is unknown, various lines of evidence indicate that the RNA participates in the formation of an abundant ribonucleoprotein (RNP) complex. For example, the gene for OLE RNA (*ole*) is almost always located upstream of a gene (*oapA*) coding for a 21 kDa transmembrane protein of unknown function, termed OLE-associated protein A (OapA) (Puerta-Fernandez et al. 2006, Harris et al. 2018). The *ole* and *oapA* genes are mutually inclusive (each residing only in genomes that include the other), suggesting that they are functionally linked. Indeed, electrophoretic mobility shift assays (EMSAs) were used to demonstrate that OapA homodimers bind to OLE RNA by recognizing two short palindromic sequences, among other contacts with the RNA (Block et al. 2011).

Moreover, it was shown that the OLE-OapA RNP complex localizes to cell membranes (Block et al. 2011), presumably due to the predicted transmembrane helices of OapA (Block et al. 2011, Harris et al. 2018). In previous work, protocols were developed to

efficiently genetically alter the alkaliphile *Bacillus halodurans* and created knockouts of *ole* and *oapA* (Takami and Horikoshi. 1999, Wallace and Breaker. 2011), Genetic disruption of either the *ole* gene, the *oapA* gene, or both, yields strains that exhibit growth and survival defects when exposed to stresses such as cold (20°C) (Wallace and Breaker. 2011), short-chain alcohols such as ethanol (5% v/v) (Wallace et al. 2012), or slightly elevated Mg^{2+} concentrations (8 mM) (Harris et al. 2019). These observations, and other clues from its genetic context (Puerta-Fernandez et al. 2006), suggest that OLE RNA has a key function in maintaining cell membrane integrity and osmotic homeostasis.

Intriguingly, the stress phenotypes noted above are substantially worsened through mutation of a highly conserved DxxxD amino acid sequence in OapA (where x is less conserved) to the sequence AxxxA (referred to as OapA protein mutation 1 or 'PM1') (Harris et al. 2018). These phenotypes are only seen when OLE RNA is also present. These findings are notable for three reasons. First, the fact that PM1 causes more severe versions of the same phenotypes compared to an *oapA* or *ole* knockout (KO) suggests that the OLE RNP complex has one or more other constituents whose function is disrupted by the PM1 changes. Second, the loss of the severe phenotype in PM1 when OLE RNA is absent indicates that the RNA is essential for the stress-related functions of the RNP complex. Third, the severity of the stress phenotypes offers a powerful selective pressure to pursue genetic selections for mutations elsewhere in the genome that convert the severe phenotypes of a PM1 strain to the milder stress phenotypes experienced by the *ole* and/or *oapA* KO strains.

In one such genetic selection (Harris et al. 2018), six of ten isolated *B. halodurans* strains acquired genomic mutations in the gene *ybzG* to overcome the severe PM1

phenotypes. The YbzG protein had not been characterized previously but was known to carry a KOW motif (17), which is a putative RNA binding domain. Indeed, it has recently been demonstrated that the YbzG protein from *B. halodurans* selectively binds full-length OLE RNA, and it was thus renamed OLE-associated protein B (OapB) (Harris et al. 2018). Additionally, the H57Y mutant version of OapB, encountered in two independent PM1-resistant strains from the genetic selection, does not bind OLE RNA. These data demonstrate that OapB plays an essential role in forming the functional OLE RNP complex in *B. halodurans*.

Furthermore, OapB was found to bind a 160-nucleotide fragment of OLE RNA containing the base-paired regions P12 through P15, suggesting that the RNA binding target of OapB is limited to this OLE substructure (Harris et al. 2018). In the current study, I further examined the interaction of OLE RNA and OapB to establish the precise nucleotides and structural features required for forming the OLE-OapB RNP complex. Specifically, I employed EMSAs with truncated and mutated RNA constructs to localize the binding sites for OapB. In addition, I used an *in vitro* selection strategy to more precisely define the sequence and structural features of OLE RNA important for high-affinity binding by OapB.

Herein I also describe the crystal structures of OapB from *B. halodurans* at 1.0 Å resolution and OapB in complex with a subdomain of OLE RNA at 2.1 Å resolution, obtained through a collaboration led by Dr. Yang Yang. These high-resolution structures reveal the overall architecture of OapB alone and when docked to its OLE RNA binding target. In addition, the structures reveal the essential binding surfaces formed by OapB that enable its high-affinity association with OLE RNA and explain the importance of a

precisely located single-nucleotide bulge in stem P13 of OLE RNA that is required for its specific recognition by OapB. Structural comparisons of OapB with other KOW-motif-containing ribosomal proteins reveals a possible additional OLE RNA-binding site in OapB, which hints at how OapB might assist in the folding of the larger OLE RNP complex. These findings provide a more detailed understanding of the structures involved in this large bacterial ncRNA.

Results and Discussion

The *oapB* gene is found in most organisms that contain OLE RNA and in many Gram-positive bacteria that lack OLE RNA

Recently reported data (Harris et al. 2018) suggests that the partnership between OLE RNA and OapB (protein formerly called YbzG) is essential for the biological function of the OLE RNP complex in *B. halodurans*. Specifically, mutations to OapB were found to overcome the strong growth and survival deficiencies of the OapA PM1 strain as noted above. Likewise, deletion of the gene for OLE RNA similarly overcomes the PM1 phenotype. These results suggest that an intact OLE RNP complex is needed for cells to exhibit the strong PM1 phenotype, and that both OLE RNA and OapB are necessary to form a functional complex.

Unfortunately, Dr. Kim Harris has not yet been able to produce a viable genetic deletion of the *oapB* gene in *B. halodurans* to determine if the loss of OapB alone is phenotypically equivalent to strains lacking OapA or OLE RNA. The *oapB* gene is located between two essential genes: *map* (type 1 methionyl aminopeptidase) and *infA* (translation initiation factor IF-1) (Kobayashi et al. 2003, Koo et al. 2017). Replacement of the entire *oapB* gene

with a selectable marker gene might affect expression of the two flanking genes, causing propagation of the intended KO strain to fail due to polar effects, or other reasons unrelated to the function of OapB, as I suspect that *oapB* is a non-essential gene.

Regardless, there remain several questions about the natural role(s) of OapB. For example, do OLE RNAs from all species form a complex with OapB? Does OapB have a functional role beyond serving as a partner in the formation of OLE RNP complexes? To date, there are no additional validated functions for OapB proteins in any other species. Therefore, these uncertainties prompted a quantitative examination of the co-occurrences of the *ole* and *oapB* genes, which can yield clues relevant to these questions.

To evaluate the distributions of *ole* and *oapB* genes, a bioinformatic analysis approach was used both to expand the number of representatives and to identify instances of co-occurrences. Specifically, by using BLAST, a list was compiled of *oapB*-containing organisms and compared it to a previously generated list of *ole*-containing organisms (Harris et al. 2018). Representatives of the *oapB* gene are found exclusively in Firmicutes, with 1,952 examples in annotated genomes (**Fig. 1**, right). Importantly, over 85% of *oapB*-containing genomes lack *ole*. This indicates that, in species lacking OLE RNA, OapB has a separate biological role. Presumably, this role involves RNA binding because these additional OapB representatives are also generally very small (~100 amino acids), and only carry well-conserved KOW RNA-binding domains. However, *oapB* is known to be non-essential in *Bacillus subtilis* (Kobayashi et al. 2003, Koo et al. 2017). Additionally, a mutation to the *B. halodurans oapB* start codon resulted in a viable strain with no growth defects under normal laboratory culture conditions (Harris et al. 2018). Thus, OapB

function might be important in *B. subtilis* and other organisms lacking OLE RNA only under certain stress conditions.

Of the 296 fully sequenced bacterial genomes that were found to carry a representative *ole* gene, all but 14 species also harbor an *oapB* gene, corresponding to a greater than 95% co-occurrence (**Fig. 1**, left). Among the 14 species that lack *oapB,* no clear differences are observed in the structure or nucleotide conservation of OLE RNA in the P12 to P15 region (**Fig. 2**). Of the species that contain both *ole* and *oapB* it is unknown how many utilize OapB as a necessary factor in assembling a functional OLE RNP complex. Given the importance of OapB to the function of the *B. halodurans* OLE RNP complex and its broader importance to species lacking OLE RNA, we sought to further define the RNA binding site of this protein.

The primary OapB binding site includes the P13 region of the *B. halodurans* OLE RNA

Previous work demonstrated that OapB interacts with a 162-nucleotide construct spanning positions 449-608 of the full-length wild-type (WT) *B. halodurans* OLE RNA (Harris et al. 2018). A dissociation constant (K_D) of ~60 nM reported for the 449-608 OLE RNA construct provides a basis of comparison for the assessment of the function of various RNA constructs examined in the current study. Initially, I created a 100-nucleotide construct, called OLE-1 (**Fig. 3A**), which encompasses positions 481-580 of the WT OLE RNA. By generating and evaluating EMSA data using trace amounts of ^{32}P-labeled RNA, I determined that OLE-1 appears to be bound by two OapB proteins in a manner indicative of two separate binding events (**Fig. 3B**).

The initial shift in RNA electrophoretic mobility exhibits an OapB concentration response curve that is consistent with a stoichiometry of 1:1 and a K_D of ~800 pM (**Fig. 3C**). The second event causes a further reduction in mobility of the RNA, or a 'supershift', suggesting that a second OapB protein is binding to the existing RNA-OapB complex, albeit with a K_D that appears to be in the micromolar range. Under my assay conditions, I do not see evidence of cooperative binding between OLE-1 and two OapB proteins.

A prominent characteristic of OLE RNAs encompassed by the OLE-1 construct is the presence of a semi-repetitive architecture, primarily represented by the P13 and P14.1 stems, and the regions immediately flanking these stems (**Fig. 3A**). For example, both stems are formed by eight base-pairs that are interrupted by one or two unpaired nucleotides. In addition, both hairpins are closed by GNRA tetraloops. Loop sequences conforming to this consensus are commonly present in naturally occurring structured RNAs, where they can both stabilize base-pairing interactions of adjoining stems (Correll et al. 2003), and present unique tertiary folding opportunities by docking with tetraloop receptor structures (Takami and Horikoshi. 1999, Wallace and Breaker. 2011). Given the architectural similarity of P13 and P14.1, I hypothesized that an OapB molecule might individually bind to each of these repeated substructures.

I observed that replacement of the highly conserved GNRA tetraloop linking the left and right shoulders of base-paired region P14.1 with an alternative stable UNCG tetraloop (Kyrpides et al. 1996) (mutant construct M1, **Fig. 3A**) causes a loss of the second (weaker) OapB binding event, but has no effect on the first (stronger) binding interaction (**Fig. 3B**). The OapB binding curves and the K_D values for the WT and M1 constructs are essentially identical (**Fig. 3C**). In contrast, replacement of the GNRA tetraloop linking the shoulders

of P13 with a UNCG sequence (mutant construct M2) causes a loss of the first, but not the second OapB binding interaction. A construct wherein both GNRA loops were replaced (M3) exhibits no OapB binding, even at concentrations as high as 3.2 μM (**Fig. 3B**), again demonstrating the importance of the tetraloop sequences for formation of the OLE-OapB RNP complex.

To further establish the specific portions of OLE RNA required for the OapB binding events, I created RNA constructs OLE-2 and OLE-3 to examine each binding site in isolation. Construct OLE-2, encompassing P13 and flanking regions, exhibits robust single-event binding by OapB, again with a K_D value (~700 pM) that is essentially identical to that observed for the OLE-1 construct (**Fig. 4A**). In contrast, I did not observe binding of OapB to construct OLE-3, which encompasses P14.1 and its flanking regions (**Fig. 4B**). Given the small size of OLE-3 and its strong predicted secondary structure features, it is unlikely that the absence OapB binding is due to RNA misfolding.

A mutated version of the OLE-2 construct also was examined, wherein the P13 GNRA tetraloop was replaced with a UNCG tetraloop (OLE-4) (**Fig. 5A**). This alteration results in the complete loss of OapB binding, which matches that observed for the OLE-1 M2 mutant and further demonstrates the requirement of the P13 GNRA tetraloop. However, a construct called OLE-5 encompassing only the P13 stem plus a partial P14 stem does not serve as a binding site for OapB (**Fig. 5B**). These findings demonstrate that the P13 stem and GNRA tetraloop alone are insufficient for high-affinity OapB binding.

To observe if these effects hold true for the binding of full-length OLE RNA to OapB, a construct (FL-OLE M2) was made wherein the P13 GNRA tetraloop was replaced with a UNCG tetraloop. Comparison of the wild-type (FL-OLE) and mutated full-length OLE

RNA by EMSA (**Fig. 6**) reveals that disruption of the P13 GNRA tetraloop leads to a reduction in affinity to OapB, although the effects are less substantial than in truncated constructs due to the overall poorer affinity of the full-length RNA.

Taken together, these results localize the high-affinity OapB binding site to P13, but only when embedded in its larger natural structural context. Although the P14.1 region is similar in appearance to P13, the poor-affinity binding of OapB to construct OLE-1 (~10^{-3} weaker) and the lack of binding to OLE-3 casts doubt on the biological relevance of this second, weaker interaction with P14.1. In its native cellular context, the OLE RNP might involve the binding of two OapB proteins, but my current *in vitro* data supports only one strong RNA binding site with no evidence for cooperative interactions between two OapB molecules.

A single bulged nucleotide is essential for high-affinity binding of P13 by OapB

Upon confirming that the P13 region forms a higher-affinity OapB binding site compared to the P14.1 region, I further examined these two regions to determine what differences might explain our results. The preceding experiments highlight the critical importance of the GNRA tetraloops for both binding sites, although additional binding site discrimination must rely on other features of these two regions. The most obvious difference is that P13 contains a single bulged nucleotide in the stem (U497), whereas P14.1 contains a mismatched pair of nucleotides (C539 and A550) (**Fig. 3A**). To assess the importance of U497, constructs were made in which this nucleotide was either mutated or deleted. The mutation of U497 to any of the other three common nucleotides only weakens binding (**Fig. 7A**), whereas deletion of this bulged nucleotide completely eliminates

binding (**Fig. 7B**). Specifically, the K_D values range from 4 to 7 nM for constructs wherein the nucleotide identity at position 497 has been changed (**Fig. 7C**), compared to 700 pM for the WT OLE-2 construct (**Fig. 3C**).

These findings suggest that U497 might form one or more additional contacts with OapB, thereby resulting in a tighter RNA-protein complex. However, a far greater contributor to binding affinity is the presence of a bulge in the P13 helix. Perhaps the protein makes direct contact with the ribose and/or phosphate moieties of this bulged nucleotide, or perhaps the binding site of the protein prefers a distorted or bent RNA helix that results from the bulged nucleotide at this location.

In vitro selection reveals nucleotides critical for high-affinity OapB binding to OLE RNA

To generate a more detailed consensus sequence and structural model for the OapB binding site, I created a mutagenized RNA population that presents a single, high-affinity binding site based on the OLE-2 RNA construct (**Fig. 4A**). The mutagenized version of this construct, called OLE-6 (**Fig. 8A**), was formed by synthesizing a DNA population (**Table 1**) for *in vitro* transcription that carried a central 56-nucleotide region with 6% sequence degeneracy. This region was flanked by invariant 15-nucleotide primer-binding sites to enable reverse transcription and subsequent PCR reactions to amplify RNAs that were selected for their ability to be bound by OapB. The initial population (generation zero or 'G_0') included approximately 6 x 10^{11} molecules (1 pmol), and is expected to fully sample all possible sequences with 6 or fewer mutations relative to the WT sequence.

Starting from the G_0 population, I performed two rounds of *in vitro* selection, which were sufficient to produce a G_2 population that is bound by OapB with characteristics that are comparable to that observed for WT OLE-6 RNA (**Fig. 8B**). A minor band (d, **Fig. 8B**) appears to be dimers of OLE-6 that have mobility similar to that of the RNA-protein complex. However, the mobility of the dimer was sufficiently different from that of RNAs bound by OapB, to exclude RNA-only dimers from amplification in subsequent generations.

The G_2 population was then subjected to RNA sequencing, and the results were used to generate a consensus model for the OapB binding site (**Fig. 8C**). The overall secondary structure appears to be retained by the vast majority of the members of the G_2 population, wherein the most highly conserved nucleotides are present in the P13 stem-loop substructure. In particular, the GNRA tetraloop and the terminal base-pair of the P13 stem are conserved to a greater extent than any other region, with G501 of this tetraloop appearing in 100% of the sequence reads represented 500 times or more. Finally, although the U nucleotide identity at position 497 was not strictly conserved, the presence of the bulge at this location was retained. These results are consistent with our earlier data revealing that the bulge is essential for high-affinity binding by OapB, whereas the nucleotide identity is not (**Fig. 7**).

Bioinformatics search for RNA sequences that match the OapB binding site consensus

After obtaining the consensus model for the RNA binding site of *B. halodurans* OapB, I searched for matches to the model by employing comparative sequence analyses

(Kobayashi et al. 2003) of *B. halodurans* (**Table 2**) and *B. subtilis* (**Table 3**) genomes. All candidate OapB binding sites with an E value of less than or equal to 1 reside within protein coding genes. These candidates appear to be false positives, as coding regions are an uncommon location for protein-RNA interactions to occur. Furthermore, there is no apparent trend regarding gene functionality, and there is no overlap in the specific genes identified from the *B. halodurans* and *B. subtilis* genome analyses.

These findings highlight the mystery regarding the predominant biological function of OapB for each species in which this protein is carried. Perhaps the *B. halodurans* OapB has a single function – to serve as an essential RNA-binding partner in the OLE RNP complex. The features of its RNA binding site might be distinct for *B. halodurans*. If true, then the consensus model might not be useful for identifying biologically relevant binding sites in other species. Regardless, the majority of bacterial species that carry OapB do not have OLE RNAs (**Fig. 1**), and therefore the primary biological role for OapB in these species remains undiscovered.

Whereas the vast majority of species that carry OLE RNA also code for OapB, 14 species have been identified that appear to lack OapB (**Fig. 1**). Furthermore, the consensus model for the high-affinity RNA target site for *B. halodurans* is not strictly conserved across all OLE RNA representatives (Harris et al. 2013). These observations suggest that OapB is a non-essential partner in OLE RNP complexes from all species, and that the determinants of OapB binding might have co-evolved with the RNA if OapB is more widely essential for OLE RNP function. Overall, these observations suggest that OapB might function as a remarkably small RNA-binding protein that recognizes simple sequence and structural features to assist in the folding and functioning of certain RNA

transcripts. In *B. halodurans*, its primary function might be to assist in the formation of the OLE RNP complex, but additional functional roles cannot be ruled out.

The protein partners of OLE RNA interact with different regions of the polynucleotide

In the current study, I have specifically mapped two binding sites for OapB within OLE RNA. Both sites are located considerable distances 3′ relative to the specific binding sites of the previously established protein partner, OapA (Block et al. 2011). OapA presumably binds as a homodimer and relies on two YAGNCUR consensus sequence motifs located in the P2 and P4a regions of OLE RNA. This separation suggests that both OapA and OapB might bind full-length OLE RNA molecules simultaneously (**Fig. 9**). It is already known that shortened OLE RNA constructs that exclude P13 are tightly bound by OapA proteins (Block et al. 2011), whereas shortened constructs such as OLE-1 (**Fig. 3**) and OLE-2 (**Fig. 4**) that lack the two YAGNCUR consensus sequences are tightly bound by OapB. These observations are consistent with the hypothesis that both proteins can bind OLE RNA at the same time. Unfortunately, given the large shift in gel mobility upon the binding of OapA to OLE RNA (Block et al. 2011), and given the small size of OapB, I have not been able to demonstrate simultaneous binding using the EMSA methods described herein.

Importantly, the protein binding regions noted above encompass only small portions of OLE RNA. There are approximately 80 highly conserved nucleotides that reside outside of the key molecular recognition motifs important for binding of OapA and OapB (**Fig. 9**). Therefore, I expect that these other well-conserved sequence and structural features are recognized by other protein factors or, perhaps more likely, are critical for OLE RNA to

perform a complex biochemical function as exemplified by all other large, highly structured, and widespread ncRNAs in bacteria.

Determining the structures of OapB and the complex formed between OapB and OLE RNA

The initial full-length *B. halodurans* OapB construct carrying an N-terminal hexahistidine affinity tag failed to yield crystals after extensive screening of crystallization conditions. Secondary structure prediction suggests the N-terminal 5 to 10 residues of OapB are unstructured. Therefore, a new construct was designed encoding a near full-length OapB (spanning residue 5 to the C-terminus) fused to an N-terminal hexahistidine affinity tag. An HRV 3C protease cleavage site was inserted between the affinity tag and the OapB coding region to enable tag removal during protein purification. This preparation, done by Dr. Yang, yielded a structure of OapB, which was determined up to 1.0 Å resolution by single isomorphous replacement with anomalous scattering (SIRAS) using one native crystal and one derivative containing iodine (**Table 4**). The final model contains all the residues in the OapB construct used for crystallization (**Fig. 10 A and B**).

My biochemical characterization identified an OLE substructure comprising P12.2, P13, P14 and P15 regions as a minimized RNA element (hereinafter designated 'OLE min RNA') that retains specific and high-affinity binding to OapB. To obtain well-diffracting crystals of the OapB-OLE min RNA complex, Dr. Yang and I screened a series of RNA constructs with varied lengths of the P12.2 stem combined with different sequences of the tetraloop that caps the P14 stem. Dr. Yang and I reconstituted each OapB-OLE min RNA

complex *in vitro* using the truncated OapB construct described above and the resulting RNP complexes were purified through size-exclusion chromatography (**Fig. 11**).

The complex formed between OapB and a 60-nt OLE min RNA (**Fig. 10C**) was crystallized to yield two different forms (hereafter referred to as 'form I' and 'form II'), both with end-to-end stacking of P12.2 helices and diffracting X-rays to 2.1 Å (**Table 4**). Dr. Yang solved the structure by single-wavelength anomalous dispersion (SAD) using crystals soaked with cobalt (III) hexammine. To further improve the electron density map, the experimental phase obtained from the SAD dataset was combined with the phase information from a partial molecular replacement model using the OapB apo structure determined above as a search template. All nucleotides in the OLE min RNA used are observed with well-defined density in the refined maps (**Fig. 12**) and are built into the final models of the complex (**Fig. 10D**). Structures from the two crystal forms have the same overall architecture and highly similar OapB-OLE min RNA interfaces. Unless otherwise indicated, the structure descriptions and analyses provided below refer to the OapB-OLE min RNA complex structure determined from crystal form I.

Overview of the OapB structure alone and in complex with an OLE RNA fragment

The KOW motif encompasses ~27 amino acids and is most commonly found in several families of ribosomal proteins (Kyrpides et al. 1996). The N-terminal portion of OapB, which includes the KOW motif of the protein (amino acids 9-35), folds into a small β-barrel (SBB) domain (Youkharibache et al. 2019), followed by three α helices in the C-terminus (**Fig. 10B**). Specifically, the KOW motif spans β1, β2 and the loop between them (**Fig. 10A**). Upon binding to OLE RNA, OapB exhibits notable conformational

rearrangements in several loop regions, particularly in the β1-β2 loop (**Fig. 13A**). A series of basic residues at the OLE RNA-binding interface undergo major rotamer changes to better fit the RNA surface (**Fig. 13A**). Such induced fit mechanism is a common theme for many RNA-binding proteins when they form complexes with their cognate RNAs (Hainzl et al. 2005, Leulliot and Varani. 2001).

The OLE min RNA contains a near-perfectly base-paired four-way helical (4H) junction, in which four helical segments form two approximately parallel coaxial stacks (**Fig. 10D**). Stems P13 and P14 stack nearly coaxially, whereas stems P12.2 and P15 form a second coaxial stack. The two stacks are connected by a pair of antiparallel crossovers (**Figs. 10D and 14A**). Three out of the four inter-nucleotide contacts at the four-way junction are canonical Watson-Crick base-pairs. The exception involves two nucleotides at the upper end of stem P12.2, which form a water-mediated non-canonical C-U base-pair (**Fig. 14A**).

By comparison, a similar 4H junction is also present in the hairpin ribozyme (Rupert and Ferre-D'Amare. 2001), although its two coaxial stacks cross at a ~60° angle. Single-molecule FRET analyses of the simple 4H junction derived from the hairpin ribozyme have shown that, without the tertiary contacts in the full-length ribozyme, the 4H junction is highly polymorphic and dynamically samples different stacking conformers (Hohng et al. 2004, Walter et al. 1998). In the 4H junction formed by the OLE min RNA, the P13-closing GUGA tetraloop, which was shown to be the primary OapB-binding site, docks into the shallow minor groove of the adjoining P12.2 helix and forms a GNRA tetraloop-receptor-like interaction (Wu et al. 2012) (**Fig. 10D**). Within the GUGA tetraloop, nucleotide U502 interacts with the sugar edge of G577 (**Fig. 14D**), nucleotide G503 forms a type II A-minor-

like interaction (Doherty et al. 2001, Nissen et al. 2001) with U576 (**Fig. 14C**), and nucleotide A504 establishes a type I A-minor groove triple motif with the G489-C575 base-pair in stem P12.2 (**Fig. 14B**). These tertiary interactions likely lock the otherwise flexible 4H junction in the specific conformation observed in the crystal structure.

My mutagenesis analyses suggested that the GCGA tetraloop closing the stem P14.1 of OLE RNA constitutes a weak secondary OapB binding site. Given the pseudosymmetric nature of stems P13 and P14.1 relative to P14 (**Fig. 15A**), P14.1 likely also coaxially stacks with P14 and thus places the GCGA tetraloop at the distal end of the continuous helix formed by P13, P14 and P14.1, precluding any interactions between the two bound OapB molecules. This is consistent with my finding that binding of OapB to the two GNRA tetraloops in OLE RNA was not cooperative.

The molecular basis of OLE RNA recognition by OapB

The structure of OapB in complex with the OLE min RNA construct reveals extensive charge and shape complementarity at the protein-RNA binding interface, which buries a total of 680 $Å^2$ of otherwise solvent-accessible surface area upon their association (**Fig. 16A**). The major groove of hairpin P13 with its end-capped GUGA tetraloop sits on top of a highly basic hill-shaped protrusion composed of β4 and three inter-β strand loops in OapB (**Fig. 16B**). The GUGA tetraloop forms an extensive network of hydrogen bonds and salt bridges by contacting various basic amino acid residues mainly through backbone phosphates and 2'-hydroxyl groups (**Fig. 16B**). The lack of base-specific interactions, except for the two hydrogen bonds mediated by the first G501 in the tetraloop (**Fig. 16B**), indicates that the specificity of OapB binding to this region is dictated by the tertiary

conformation of the GNRA tetraloop, rather than exact sequence within this tetraloop family. Indeed, replacing the GUGA tetraloop of the P13 hairpin with another common stabilizing tetraloop based on the UNCG consensus (Antao et al. 1991), which is known to adopt a different conformation (Cheong et al. 1990, Heus and Pardi. 1991), abolishes OapB binding.

Previous genetic selection results revealed that mutations at either of two highly conserved residues in OapB, G19S and H57Y, mitigated the severity of PM1 phenotypes (Harris et al. 2018). These mutations disrupt the ability of OapB to bind OLE RNA, presumably preventing the proper assembly of the functional OLE RNP complex. Indeed, the OapB-OLE min RNA co-crystal structure provides a basis to understand how these OapB mutations disrupt OapB-OLE RNA binding.

The G19 residue of OapB resides at a strictly conserved position of the SBB domain (Kyrpides et al. 1996). This amino acid is located in the β1-β2 loop (**Figs. 10 A, B and 16B**), which projects into the major groove of the OLE RNA P13 hairpin (**Fig. 10D**). The absence of a side chain at this position of the wild-type protein allows for the close approach of OLE RNA to OapB, and the subsequent formation of two hydrogen bonds between the P13 GUGA tetraloop and the main chain nitrogen atoms of OapB G19 and R20 (**Fig. 16B and Fig. 15B**). A G19S mutation likely results in steric clashes that prevent proper contact between OapB and OLE RNA (**Fig. 15B**).

As H57 residue carries an imidazole side chain that is inserted between the U502 ribose at the bottom of the P13 GUGA tetraloop and the backbone of the P12.2 helix (**Fig. 15C**), the H57Y mutation is expected to cause a loss of two hydrogen bonds otherwise involving the side chain of histidine. In addition, the bulkier side chain of tyrosine is unable to fit

within the narrow gap originally occupied by histidine without causing substantial clashes with OLE RNA (**Fig. 15C**).

The close proximity of the P12.2 helix with the P13 hairpin tetraloop in the tertiary structure of OLE RNA creates a unified surface for OapB binding to these separate secondary-structure features (**Fig. 16 A and C**). OapB amino acids R35, N54, and N56, in addition to the above-mentioned H57 make multiple contacts with backbone phosphate groups of nucleotides in the P12.2 helix (**Fig. 16C**). These interactions likely play a role in OLE RNA folding by serving as part of a protein tether that stabilizes the close approach and orientation of the two coaxial stacks (the P12.2-P15 stack and the P13-P14 stack). These contacts appear to work synergistically with the GUGA tetraloop-P12.2 minor groove base triple tertiary interactions to direct the OLE RNA 4H junction to the observed conformation.

The presence of a single-nucleotide bulge in stem P13 (**Fig. 10C**) is a highly conserved feature of OLE RNA (Harris et al. 2018), which I found to be critical for OapB binding. In *B. halodurans* OLE RNA, the base of the bulged nucleotide U498 flips out to extrahelical space, distorting the P13 helix backbone and leading to an unusually short (less than 5 Å) phosphate-to-phosphate distance between U499 and G500 (**Fig. 16D**). The crowding of negative charges is neutralized by a pair of arginine residues (R20 and R49) from OapB that clamp the two phosphate groups from both sides (**Fig. 16D**). Notably, the bulged U498 adopts different conformations in the two crystal forms of OapB-OLE min RNA complex, but is not involved in any base-specific interactions with OapB in either form (**Fig. 16D and Fig. 15D**). This observation suggests this nucleotide is highly flexible and its identity is unlikely to contribute to the specificity of OLE RNA recognition by OapB.

Previous bioinformatics-derived secondary structure models for OLE RNA assigned the nucleotide equivalent to U497 as the bulged nucleotide (Puerta-Fernandez et al. 2006, Harris et al. 2018, Harris et al. 2019). My mutagenesis studies with a minimized OLE RNA showed that substituting U497 with other nucleotides weakens OapB binding by 6 to 10 fold, whereas removing this nucleotide almost completely abolishes binding. However, the *B. halodurans* OLE RNA sequence contains three consecutive uridines at positions 497 through 499, and therefore it was not initially clear which U nucleotide was bulged in the functional conformation. Specifically, deletion of U497 is equivalent to removing any of the three nucleotides at these positions, and therefore will be indistinguishable in their effects on OLE RNA structure and function. A deletion of a U nucleotide in this region is expected to eliminate the distorted RNA helical backbone by restoring contiguous base-pairing of P13, thereby displacing the U499 backbone phosphate and breaking its interactions with R20 and R49 of OapB. Mutating U497 to other nucleotides is expected to change its base-pairing potential with G507 and consequently shift the positions of the U498 and U499 backbone phosphate groups. However, the suboptimal RNA backbone conformation that results might be partially compensated by the flexibility of the flipped U498, resulting in a less deleterious impact on OapB binding affinity as is observed in our biochemical experiments. Therefore, it appears that the extrahelical U498 and G507-U497 wobble base-pair together place the two backbone phosphate groups for optimal interaction with the OapB arginine dyad.

Conclusions

The OapB protein from *B. halodurans* binds the P13 region of OLE RNA with an affinity in the mid-picomolar range. This strong interaction is required for cells to exhibit the most severe sensitivity phenotypes when exposed to cold, short-chain alcohols, or slightly elevated Mg^{2+} concentrations. Despite its small size (102 amino acids), the protein is remarkably selective for its RNA binding site. The consensus sequence for OapB binding includes a GNRA tetraloop on a stem closed with a G-C base-pair, and a bulged nucleotide (preferably uridine) located 3 base-pairs away and on the 5′ side of the tetraloop (**Fig. 8C**).

Although the RNA consensus model for OapB binding is relatively simple, and there are close matches for this architecture elsewhere in the RNAs of *B. halodurans* (**Table 2**), we do not have evidence that any of these additional binding site candidates are biologically relevant. It is apparent that additional regions of the high-affinity OapB binding site flanking the OLE RNA P13 stem are also required for complex formation (**Fig. 2B**). The nucleotide sequences for these additional molecular recognition determinants are not highly conserved. However, they might be sufficiently important for binding such that their absence precludes OapB from recognizing other cellular RNAs that closely mimic the P13 stem of *B. halodurans* OLE RNA. Regardless, OapB from most species must have other RNA targets, because the gene for this protein is present in more than 85% of sequenced bacterial species that do not contain the gene for OLE RNA. Additional experiments with species lacking OLE RNA will need to be conducted to determine the broader functions of OapB.

The high-resolution crystal structure of OapB with its cognate OLE RNA substructure reported here provides atomic detail of the architecture and assembly of an intriguing portion of the OLE RNP complex. The OLE RNA fragment bound by OapB includes the

stems P12.2, P13, P14 and P15, which together form an almost perfect 4H junction. This region contains RNA tertiary structural motifs that are recurrent in many other large ncRNAs (Natchiar et al. 2017), which highlights the importance of RNA tertiary structure to the biological and biochemical functions of OLE RNA.

Three subregions of OapB that are discontinuous in primary sequence come together to form a bipartite surface for association of the protein with OLE RNA. The binding interface is replete with hydrogen bonding and salt bridge interactions that tightly join the two molecules, which explains the strong binding affinity for the complex (K_D ~700 pM). Because most of these protein-RNA contacts are mediated by phosphate and ribose 2' hydroxyl groups of the RNA phosphodiester backbone, the high specificity of OLE RNA recognition by OapB is primarily dictated by the unique three-dimensional architecture of the OLE min RNA construct, as evident from the remarkable shape complementarity between the two partners at their binding interface (**Fig. 16A**). Additionally, the OapB-OLE min RNA complex structure explains the disruptive effects of mutations in OapB (Harris et al. 2018) and OLE RNA and provides a basis to understand why each of the three OapB mutations (G19S, G42V and H57Y) found in genetic selections (Harris et al. 2018) alleviate the strong dominant-negative phenotype observed with the PM1 strain.

In the OapB-OLE min RNA complex, the amount of OapB surface utilized for RNA binding is markedly lower compared to two structurally similar ribosomal proteins, RPL14 and RPL27. Detailed analyses of their RNA-binding patterns suggest that OapB α3 likely forms an additional RNA binding site and that this might be important for binding the full-length OLE RNA in the OLE RNP complex. Further analyses involving larger RNA

constructs will be needed to evaluate this hypothesis and to characterize the nature of this putative interaction.

Intriguingly, some of the most highly conserved nucleotide positions in this RNA construct are not involved in either binding OapB or in setting up the special 4H junction that is observed in the RNA-protein complex. Most notably, the conserved nucleotides in the loops of P14a and P15, and also those in the junction between P14 and P14a, would be located in the natural RNA far from the nucleotides bound by OapB. Therefore, the OLE RNP complex likely exploits these strongly conserved nucleotides to achieve other structural and functional objectives. It is likely that the *B. halodurans* OLE RNP complex requires OapB binding to assist in forming the local architecture necessary for these other objectives to be met, as well as helping to bring a distal part of OLE RNA near to these conserved nucleotides.

Because OapA and OapB interact only with small portions of OLE RNA it can be assumed that either OLE RNA is interacting with additional protein partners or that the other conserved substructures of this large ncRNA perform another function, possibly even acting as a ribozyme. Although increased understanding of the OLE-OapB complex does not readily provide additional insight into the function of the larger OLE RNP, it does lay the foundation for the eventual reconstitution of an active OLE RNP complex *in vitro*, allowing for the testing of specific hypotheses for biochemical function, and the further structural analysis of this unusual ncRNA.

Materials and Methods

Synthetic DNA oligonucleotides

All synthetic DNA oligonucleotides (**Table 1**) were purchased from Sigma Aldrich except for DLW137, which was purchased from the Keck Oligonucleotide Synthesis facility at Yale University.

Bioinformatics

OapB representatives were identified using NCBI BLAST with *B. halodurans* OapB (WP_010896340.1) as the query sequence. A list of organisms that contain *ole* was obtained from the OLE RNA alignment file created previously (Harris et al. 2018). The *oapB* gene was detected in each of these genomes using NCBI BLAST.

RNA constructs

Double-stranded DNA templates for *in vitro* RNA transcription were prepared by overlap extension reactions. Specifically, annealing incubations were conducted with 200 pmoles of each synthetic DNA strand in a mixture prepared by the addition of 10 μL 5x First Strand Buffer (Thermo Fisher Scientific), 1 μL of a 10 mM (each) dNTP mixture, and deionized H_2O in a volume of 44 μL, which was heated to 90°C for 1 min, then cooled to 23°C. To this mixture was added 5 μL of 0.1 M DTT and 1 μL (200 U) Superscript II Enzyme (Thermo Fisher Scientific). The reaction was incubated at 42°C for 1 hour followed by 70°C for 15 min. DNA templates that did not require overlap extension reactions were diluted to 4 μM and heated to 90°C for 1 min. All DNA constructs served as templates for *in vitro* transcription using T7 RNA polymerase. The resulting RNAs were purified by denaturing (7 M urea) 10% polyacrylamide gel electrophoresis (PAGE), dephosphorylated,

and 5′ ^{32}P-radiolabeled with γ^{32}P-ATP using methods described previously (Block et al. 2011).

Binding assays

The OapB stock used for binding assays was purified and stored as described previously (Harris et al. 2018). RNA binding assays with OapB were performed as described previously (Harris et al. 2018), but with two exceptions. First, OapB stock solutions were prepared in concentrations that were ten-fold higher than the final concentrations used in the binding assays, in half-log intervals. Second, bound complexes were separated from free 5′ ^{32}P-labeled OLE RNA constructs by using non-denaturing 10% PAGE at 20 W for 1-2 hours.

Selection for mutant OLE RNAs that bind OapB

A single-stranded synthetic DNA construct (**Table 1**, DLW137) containing a T7 RNA polymerase (T7 RNAP) promoter, 56 nucleotides prepared with a mutation rate of 6%, and two 15 nucleotide segments flanking the degenerate region to allow for amplification of the selected RNA representatives was ordered from the Keck Oligonucleotide Synthesis facility at Yale University. Double-stranded DNA templates for *in vitro* transcription (generated by primer-extension from DLW137 and DLW140) and the corresponding RNAs were produced as described above. Binding assays were prepared as previously described (Harris et al. 2018), with 100 μM OapB and 50 μM OLE-6 RNA, and were separated by nondenaturing 10% PAGE run for 3 h at 20 W. The band containing the OLE-OapB RNP complex was excised and subjected to crush-soak elution (200 mM NaCl, 10

mM Tris-HCl pH 7.5, 1 mM EDTA pH 8.0, 1 M urea) for 2 h. The eluted RNA was concentrated by precipitation with ethanol. Primers (**Table 1**, DLW145 and DLW167) were used to reverse transcribe the RNA. The resulting cDNAs were then amplified by PCR and used to generate a new pool of RNAs for the next round of selection. A total of two rounds of selection were performed and the resulting G_2 population was sequenced.

Construction of consensus model

The double-stranded G_2 DNA population was sequenced by the Yale Center for Genome Analysis, which yielded 10 million reads. Individual sequences were reconstructed from the pairwise reads, and identical reads were grouped and sorted by prevalence. Sequences with at least 500 reads (366 unique sequences) were compiled and used to generate the consensus RNA motif for OapB binding. Clustal Omega was used to align the sequences, followed by manual editing to refine the alignment (Sievers et al. 2011). The Stockholm file of aligned sequences was used to generate a consensus model for OapB binding with R2R (Weinberg and Breaker. 2011). To best display the data, R2R parameters were modified to report conservation of >99% (red), >95% (black), and >90% (gray), with all other parameters set at the default values.

Search for additional OapB interacting RNAs

The Stockholm file described above was modified to include only OLE RNA nucleotides 494-510. This trimmed file was run through *cmbuild* and *Infernal* algorithms to generate a calibrated covariance model (Nawrocki and Eddy. 2013). The calibrated covariance model

was run against the *B. halodurans* and *B. subtilis* genomes using *cmsearch* to generate a list of potential OapB-interacting sequences.

Protein expression and purification for crystallography.

An open reading frame encoding residue 5 to the C-terminus of *B. halodurans* OapB was fused to an N-terminal 6×His affinity tag and human rhinovirus (HRV) 3C protease cleavage sequence and cloned into the pET-11a expression vector (Millipore Sigma) between NdeI and BamHI restrictive enzyme sites. The corresponding OapB protein was overexpressed in *Escherichia coli* BL21 Star (DE3) cells (ThermoFisher Scientific) at 17 °C for 18 h.

Pelleted cells expressing recombinant OapB protein were resuspended in buffer A, which is composed of 50 mM Tris-HCl (pH 7.5), 500 mM NaCl, 20 mM imidazole (pH 7.5), 5% glycerol, and β-mercaptoethanol (β-ME). The mixture was homogenized using an M110EH Microfluidizer (Microfluidics International Corporation). Cell lysate was clarified by centrifugation using an SS-34 rotor (17,000 rpm for 1 h at 4 °C) and loaded onto a HisTrap HP affinity chromatography column (Cytiva Life Sciences) pre-equilibrated with buffer A. OapB protein was eluted from the HisTrap HP column over 20× column volumes (CV) with a linear gradient from buffer A to buffer B, the latter which is composed of 50 mM Tris-HCl (pH 7.5), 500 mM NaCl, 300 mM imidazole (pH 7.5), and 5% glycerol. Eluted protein from multiple fractions were pooled and supplemented with 2 mg of 6×His tagged HRV 3C protease and 1 mM (final concentration) β-ME. Protease cleavage and 6×His tag removal from OapB was performed at 4 °C for 24 h. The

protein sample was diluted with equal volume of buffer C, which is composed of 20 mM sodium phosphate (pH 7.0) and 1 mM β-ME, and the mixture was loaded onto a HiTrap Heparin HP column (Cytiva Life Sciences). After washing with 5× CV of buffer C, OapB protein was eluted over 10× CV with a linear gradient from buffer C to buffer D, the latter which is composed of 20 mM sodium phosphate (pH 7.0), 1 M NaCl, and 1 mM β-ME. The eluted OapB protein sample was supplemented with imidazole to a final concentration of 20 mM and passed through a HisTrap HP column again to remove residual 6×His tag, HRV 3C protease, and any uncleaved OapB protein. The flow through was collected, concentrated using an Amicon Ultra centrifugal filter with a 3 kDa cutoff, and further purified using a HiLoad 16/600 Superdex 200 pg size-exclusion chromatography (SEC) column in buffer E, composed of 20 mM Tris-HCl (pH 7.5), 500 mM NaCl, and 1 mM Tris(2-carboxyethyl) phosphine hydrochloride (TCEP-HCl).

***In vitro* transcription and OLE min RNA purification for crystallography.**

The DNA template for *in vitro* transcription of OLE min RNA was generated by annealing equimolar amounts of two DNA oligonucleotides with complementary sequences (5'-<u>TAA TAC GAC TCA CTA TA</u>G GCC AGT CTG GCG TTT GGT GAC AGC GCC AAG TTC TTC GGA ATT GGG AAA TCC TAC TGG CC-3' and 5'-**[G_m][G_m]**C CAG TAG GAT TTC CCA ATT CCG AAG AAC TTG GCG CTG TCA CCA AAC GCC AGA CTG GCC TAT AGT GAG TCG TAT TA-3'). The T7 RNA polymerase (RNAP) promoter sequence in the non-template DNA strand is underlined. Two guanosines denoted **[G_m]** on the 5' end of template DNA strand are 2'-O-methylated to reduce 3' end heterogeneity of the RNA transcript (Nawrocki and Eddy. 2013).

Transcription reactions (3 mL) were assembled using 1.2 nmol annealed DNA template, 300 µg T7 RNAP, 0.3 U inorganic pyrophosphatase (New England Biolabs) and 120 µL 25× RNAsecure RNase inactivation reagent (ThermoFisher Scientific) in 1× transcription buffer containing 80 mM HEPES-K (pH 7.5), 24 mM $MgCl_2$, 40 mM DTT, 2 mM spermidine, and 4 mM of each NTP. The resulting mixture was incubated at 37 °C for 2.5 h. After removing the pyrophosphate precipitate by centrifugation at 3,000 × g for 10 min, EDTA was added to the reaction mix to a final concentration of 50 mM. OLE min RNA was purified by extraction three times with acid phenol:chloroform (pH 4.5).The free nucleotides were subsequently removed using a Sephadex G-25 PD-10 desalting column (Cytiva Life Sciences) followed by size exclusion chromatography (SEC) using a HiLoad 16/600 Superdex 200 pg column in buffer F, containing 20 mM sodium cacodylate (pH 6.5) and 100 mM NaCl.

OapB-OLE min RNA complex assembly.

Purified OapB protein and OLE min RNA were mixed in a 2:1 molar ratio and incubated at room temperature for 1 h. OapB-OLE min RNA complex was separated from excessive OapB by SEC as described above using buffer G, containing 20 mM Tris-HCl (pH 7.5), 100 mM NaCl, and 1 mM TCEP. Fractions containing OapB-OLE min RNA were pooled and stored at 4 °C for later use.

Crystallization and data collection.

Crystallization of OapB was achieved using the sitting-drop vapor diffusion method at 19 °C in 50 mM Tris-HCl (pH 8.5), 100 mM KCl, 10 mM $MgCl_2$ and 30% polyethylene glycol (PEG) 400. Crystals were cryo-protected in buffer containing 50 mM Tris-HCl (pH 8.5), 100 mM KCl, 10 mM $MgCl_2$ and 36% PEG 400 before being flash frozen in liquid nitrogen. Iodine-derivative OapB crystals were prepared by soaking the crystals in buffer containing 50 mM Tris-HCl (pH 8.5), 100 mM KCl, 10 mM $MgCl_2$, 36% PEG 400 and 500 mM NaI at room temperature for 2 to 5 min. X-ray diffraction data were collected at 100 K at beamlines 24ID-E and 24ID-C of the Advanced Photon Source at Argonne National Laboratory. The datasets for the native and iodine-derivative crystals were initially collected at wavelengths of 0.9792 Å and 1.4586 Å, respectively. Because the diffraction limit of the OapB crystals clearly exceeds the maximum achievable resolutions at these wavelengths, one more dataset was collected using the iodine-derivative crystal at a wavelength of 0.8266 Å to push the resolution limit of the dataset to 1.0 Å.

OapB-OLE min RNA complex preparations were concentrated to about 400 µM and used in crystallization screening. OapB-OLE min RNA complex crystals were grown by the sitting-drop vapor diffusion method at 19 °C in either 100 mM HEPES (pH 7.0 to 7.5), 200 mM NH_4Cl and 39 to 42.5% 2-Methyl-2,4-pentanediol (MPD) (crystal form I) or 240 mM sodium malonate (pH 6.0 to 7.0) and 20 to 21% PEG 3,350 (crystal form II). Cobalt-derivative OapB-OLE min RNA complex crystals were prepared by soaking crystal form I in buffer containing 100 mM HEPES (pH 7.0), 200 mM NH_4Cl, 45% MPD and 2 mM cobalt (III) hexammine chloride ($[Co(NH_3)_6]Cl_3$) at 19 °C for 14 h. Data were collected at beamline 24ID-C of the Advanced Photon Source at wavelengths of 0.9792 Å and 1.6058

Å for native and cobalt-derivative crystals, respectively. All X-ray diffraction datasets were processed with the XDS package (Kao et al. 1999).

Crystal structure determination and refinement.

Heavy atom sites were identified using SHELXD (Kabsch. 2010) and structures were solved using AutoSol (Sheldrick. 2008) by SIRAS method for OapB and molecular replacement with single wavelength anomalous dispersion (MR-SAD) method (Terwilliger et al. 2009) for the OapB-OLE min RNA complex. The initial models were automatically built using the AutoBuild (Panjikar et al. 2009) module of the PHENIX software package (Terwilliger et al. 2008) and manually rebuilt in Coot (Adams et al. 2010). Heavy atoms were placed based on refined anomalous difference maps. The structure of the native OapB-OLE min RNA complex in crystal form II was determined by molecular replacement using the cobalt-derivative complex structure as a search template. The models for the OapB apo structure were refined using phenix.refine (Emsly et al. 2010) in PHENIX with anisotropic individual *B* factors for all non-hydrogen atoms. The models for the OapB-OLE min RNA complex were refined with Translation-Libration-Screw (TLS) parameters turned on. The final refined models were validated using MolProbity (Chou et al. 2013). Data collection and refinement statistics are summarized in ***SI Appendix*, Table S1**. All molecular representations were prepared using UCSF ChimeraX (Afonine et al. 2012).

Accession codes.

Atomic coordinates and associated structure factors have been deposited in the Protein Data Bank (PDB) with accession codes XXXX (native OapB apo structure), XXXX

(iodine-derivative OapB apo structure), XXXX (OapB-OLE min RNA complex, crystal form I) and XXXX (OapB-OLE min RNA complex, crystal form II).

Data availability – All data is available in the manuscript except for RNA sequencing results and alignment files used to generate Fig. 8C and alignment files used to generate Fig. 2. This data is available upon request from Ronald Breaker (see contact information above).

Author contributions – D.L.W., K.A.H., Y.Y., and R.R.B. devised the research plan, D.L.W., Y.Y., and L.C. conducted the biochemical assays, D.L.W. and K.A.H. conducted the bioinformatics analyses, and all authors evaluated the experimental data. D.L.W., Y.Y., K.A.H., and R.R.B. wrote the manuscripts and all authors provided edits and comments.

Acknowledgements – We thank Michelle Peters for purifying the OapB used for part of this study and other members of the Breaker laboratory for helpful discussion. K.A.H. was supported by NIH Grant F32GM116426. The project was supported by NIH Grant GM022778 (to R.R.B.). RNA research in the Breaker laboratory is also supported by the Howard Hughes Medical Institute.

Conflicts of interest – None of the authors declare any conflicts of interest related to the data presented in this article.

References

Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H., and Adams, P. D. (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr* **68**, 352-367

Antao, V. P., Lai, S. Y., and Tinoco, I., Jr. (1991) A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res* **19**, 5901-5905

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

Block, K. F., Puerta-Fernandez, E., Wallace, J. G., and Breaker, R. R. (2011) Association of OLE RNA with bacterial membranes via an RNA-protein interaction. *Mol. Microbiol***. 79**, 21-34

Cech, T. R., and Steitz, J. A. (2014) The noncoding RNA revolution – trashing old rules to forge new ones. *Cell* **157**, 77-94

Cheong, C., Varani, G., and Tinoco, I., Jr. (1990) Solution structure of an unusually stable RNA hairpin, 5'GGAC(UUCG)GUCC. *Nature* **346**, 680-682

Chou, F. C., Sripakdeevong, P., Dibrov, S. M., Hermann, T., and Das, R. (2013) Correcting pervasive errors in RNA crystallography through enumerative structure prediction. *Nat Methods* **10**, 74-76

Correll, C. C., and Swinger, K. (2003) Common and distinctive features of GNRA tetraloops based on a GUAA tetraloop structure at 1.4 Å resolution. *RNA* **9**, 355-363

Costa, M., and Michel, F. (1997) Rules for RNA recognition of GNRA tetraloops deduced by *in vitro* selection: comparison with *in vivo* evolution. *EMBO J.* 16, 3289-3302

Doherty, E. A., Batey, R. T., Masquida, B., and Doudna, J. A. (2001) A universal mode of helix packing in RNA. *Nat Struct Biol* **8**, 339-343

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

Fiore, J. L., and Nesbitt, D. J. (2013) An RNA folding motif: GNRA tetraloop-receptor interactions. *Q. Rev. Biophys.* **46**, 223-264

Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849–857

Hainzl, T., Huang, S., and Sauer-Eriksson, A. E. (2005) Structural insights into SRP RNA: an induced fit mechanism for SRP assembly. *RNA* **11**, 1043-1050

Harris, K. A., and Breaker, R. R. (2018) Large noncoding RNAs in bacteria. *Microbiol. Spectr.* **6**, RWR-0005-2017

Harris, K. A., Zhou, Z., Peters, M. L., Wilkins, S. G., and Breaker, R. R. (2018) A second RNA-binding protein is essential for ethanol tolerance provided by the bacterial OLE ribonucleoprotein complex. *Proc. Natl. Acad. Sci. USA* **115**, E6319-E6328

Harris, K.A., Odzer, N.B., and Breaker, R.R. (2019) Disruption of the OLE ribonucleoprotein complex causes magnesium toxicity in *Bacillus halodurans*. *Mol. Microbiol.* **112**, 1552-1563

Heus, H. A., and Pardi, A. (1991) Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science* **253**, 191-194

Hohng, S., Wilson, T. J., Tan, E., Clegg, R. M., Lilley, D. M., and Ha, T. (2004) Conformational flexibility of four-way junctions in RNA. *J Mol Biol* **336**, 69-79

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

Kao, C., Zheng, M., and Rudisser, S. (1999) A simple and efficient method to reduce nontemplated nucleotide addition at the 3 terminus of RNAs transcribed by T7 RNA polymerase. *RNA* **5**, 1268-1272

Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P., et al. (2003) Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. USA* **100**, 4678–4683

Koo, B. M., Kritikos, G., Farelli, J. D., Todor, H., Tong, K., Kimsey, H., Wapinski, I., Galardini, M., Cabal, A., Peters J. M., et al. (2017) Construction and analysis of two genome-scale deletion libraries for *Bacillus subtilis*. *Cell Syst.* **4**, 291-305

Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., and Cech, T. R. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell* **31**, 147–157

Kyrpides, N. C., Woese, C. R., and Ouzounis, C. A. (1996) KOW: a novel motif linking a bacterial transcription factor with ribosomal proteins. *Trends Biochem. Sci*. **21**, 425-426

Leulliot, N., and Varani, G. (2001) Current topics in RNA-protein recognition: control of specificity and biological function through induced fit and conformational capture. *Biochemistry* **40**, 7947-7956

Molinaro, M., and Tinoco, I., Jr. (1995) Use of ultra stable UNCG tetraloop hairpins to fold RNA structures: thermodynamic and spectroscopic applications. *Nucleic Acids Res.* **23**, 3056-3063

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

Nissen, P., Hansen, J., Ban, N., Moore, P. B., and Steitz, T. A. (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920-930

Nissen, P., Ippolito, J. A., Ban, N., Moore, P. B., and Steitz, T. A. (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc Natl Acad Sci U S A* **98**, 4899-4903

Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S., and Tucker, P. A. (2009) On the combination of molecular replacement and single-wavelength anomalous diffraction phasing for automated structure determination. *Acta Crystallogr D Biol Crystallogr* **65**, 1089-1097

Peebles, C. L., Perlman, P. S., Mecklenburg, K. L., Petrillo, M. L., Tabor, J. H., Jarrell, K. A., and Cheng, H. L. (1986) A self-splicing RNA excises an intron lariat. *Cell* **44**, 213-223

Puerta-Fernandez, E., Barrick, J. E., Roth, A., and Breaker, R. R. (2006) Identification of a large noncoding RNA in extremophilic eubacteria. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19490-19495

Rupert, P. B., and Ferre-D'Amare, A. R. (2001) Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. *Nature* **410**, 780-786

Sheldrick, G. M. (2008) A short history of SHELX. *Acta Crystallogr A* **64**, 112-122

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D. and Higgins, D. G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol*. **7**, 539

Takami, H., and Horikoshi, K. (1999) Reidentification of facultatively alkaliphilic *Bacillus* sp. C-125 to *Bacillus halodurans*. *Biosci. Biotechnol. Biochem.* **63**, 943-945

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L. W., Read, R. J., and Adams, P. D. (2008) Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr D Biol Crystallogr* **64**, 61-69

Terwilliger, T. C., Adams, P. D., Read, R. J., McCoy, A. J., Moriarty, N. W., Grosse-Kunstleve, R. W., Afonine, P. V., Zwart, P. H., and Hung, L. W. (2009) Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallogr D Biol Crystallogr* **65**, 582-601

Wallace, J. G., and Breaker, R. R. (2011) Improved genetic transformation methods for the model alkaliphile *Bacillus halodurans* C-125. *Lett. Appl. Microbiol.* **52**, 430-432

Wallace, J. G., Zhou, Z., and Breaker, R. R. (2012) OLE RNA protects extremophilic bacteria from alcohol toxicity. *Nucleic Acids Res.* **40**, 6898-6907

Walter, F., Murchie, A. I., and Lilley, D. M. (1998) Folding of the four-way RNA junction of the hairpin ribozyme. *Biochemistry* **37**, 17629-17636

Weinberg, Z., Perreault, J., Meyer, M. M., and Breaker, R. R. (2009) Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. *Nature* **462**, 656-659

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

Weinberg, Z., Lünse, C. E.,Corbino, K. A., Ames, T. D., Nelson, J. W., Roth, A., Perkins, K. R., Sherlock, M. E., and 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

Wu, L., Chai, D., Fraser, M. E., and Zimmerly, S. (2012) Structural variation and uniformity among tetraloop-receptor interactions and other loop-helix interactions in RNA crystal structures. *PLoS One* **7**, e49225

Youkharibache, P., Veretnik, S., Li, Q., Stanek, K. A., Mura, C., and Bourne, P. E. (2019) The Small beta-Barrel Domain: A Survey-Based Structural Analysis. *Structure* **27**, 6-26

Figures and Tables

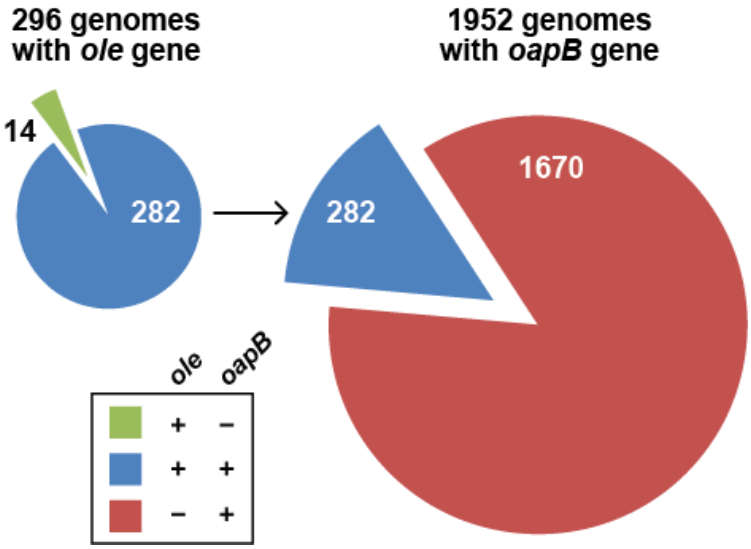

Figure 1. The *ole* and *oapB* genes frequently co-occur in bacterial genomes. Left: The gene for OLE RNA is present in the genomes of 296 bacterial species whose DNA has been completely sequenced. Of these, 282 species also carry a gene for OapB. Right: The gene for OapB is more widespread among bacteria than is the gene for OLE RNA.

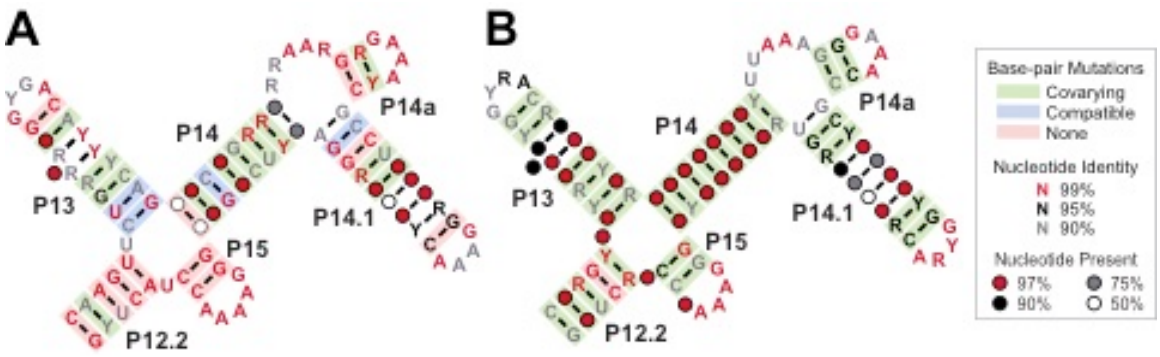

Figure 2. Consensus models for the OapB-interacting region of OLE RNA. *A*, The OLE RNA consensus model of residues 486-578 for the 14 OLE RNAs from organisms that do not contain the *oapB* gene. *B*, The OLE RNA consensus model of residues 486-578 generated from 795 genomic and metagenomic samples (13).

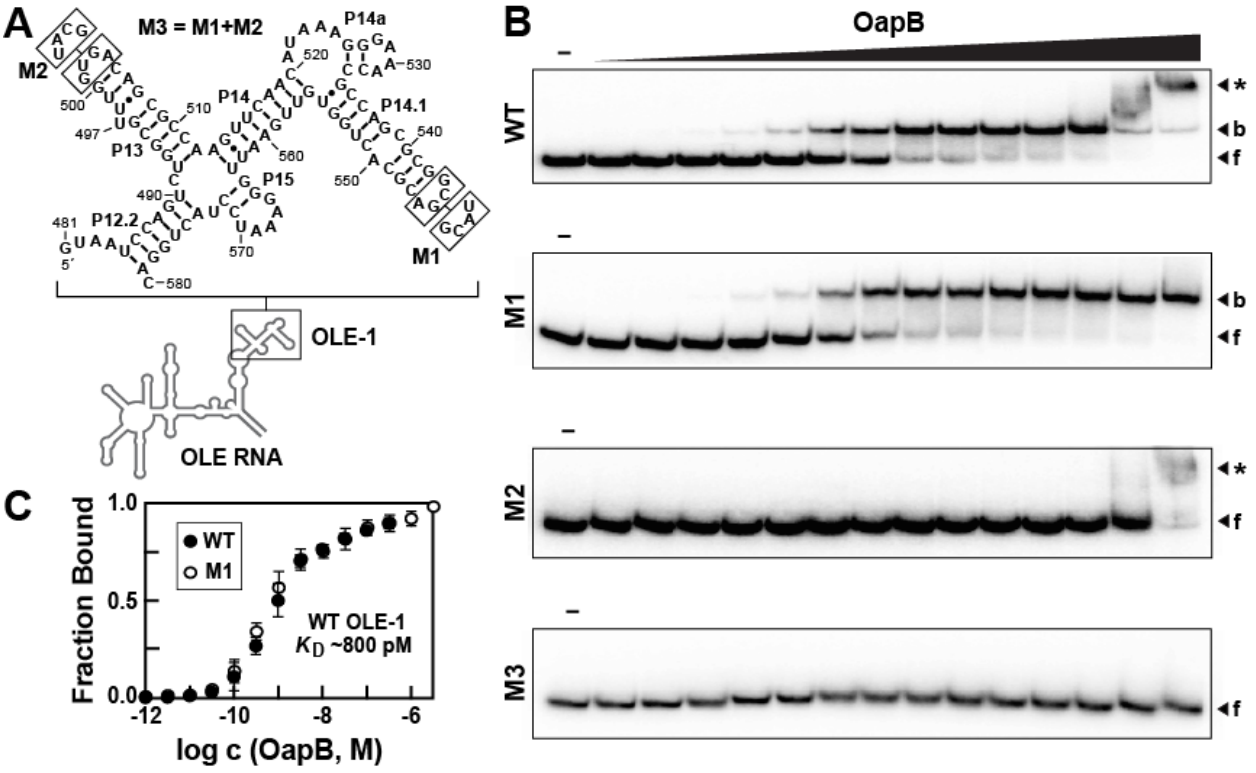

Figure 3. EMSA data for OapB binding to OLE-1, a 100-nucleotide fragment of *B. halodurans* OLE RNA. *A*, Sequence and secondary structure model for WT OLE-1 and various mutants. Boxed nucleotides are altered as depicted for mutant constructs M1 and M2. Construct M3 combines the mutations from M1 and M2. *B*, Representative EMSA autoradiograms derived by incubating trace amounts of 5′ ^{32}P-labeled RNA constructs as indicated either in the absence (–) of OapB, or in the presence of various concentrations of OapB ranging from 1 pM to 3.2 μM (in half-log increments). Bands corresponding to unbound or "free" RNA (f), bound RNA (b), and RNA undergoing a supershift (*) are annotated accordingly. *C*, Binding curves for WT and M1 OLE-1 RNA constructs with OapB derived from EMSA data as depicted in B. The plot for WT OLE-1 RNA excludes the highest two OapB concentrations examined to avoid the effects of the supershift. Error bars represent the standard error of the data based on at least three replicate experiments.

When absent, error bars are smaller than the symbols used to represent the value. I performed the experiments shown in this figure.

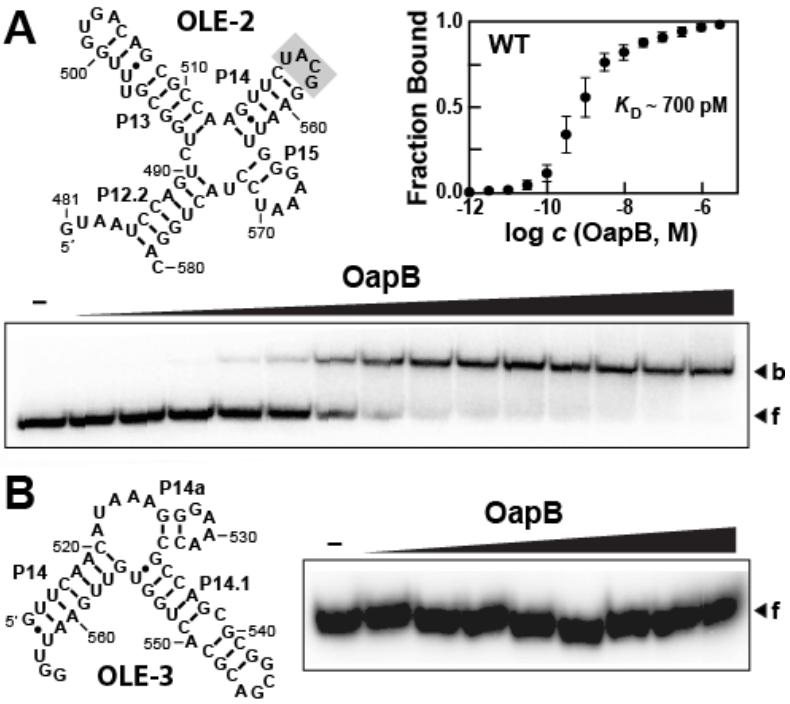

Figure 4. An OLE RNA fragment containing P13 and flanking regions is sufficient to form a complex with OapB. *A*, Characteristics of OapB binding to the OLE-2 RNA construct, which lacks the P14.1 region. Shaded nucleotides identify a UNCG tetraloop that was used in place of a portion of P14 and all of P14.1. OapB concentrations range from 1 pM to 1 μM (half-log increments). Additional annotations are as described in the legend to **Fig. 3**. *B*, OapB fails to bind the OLE-3 RNA construct, which lacks the P13 region. OapB concentrations range from 320 pM to 1 μM (half-log increments). Additional annotations are as described in the legend to **Fig. 3**. I performed the experiments shown in this figure.

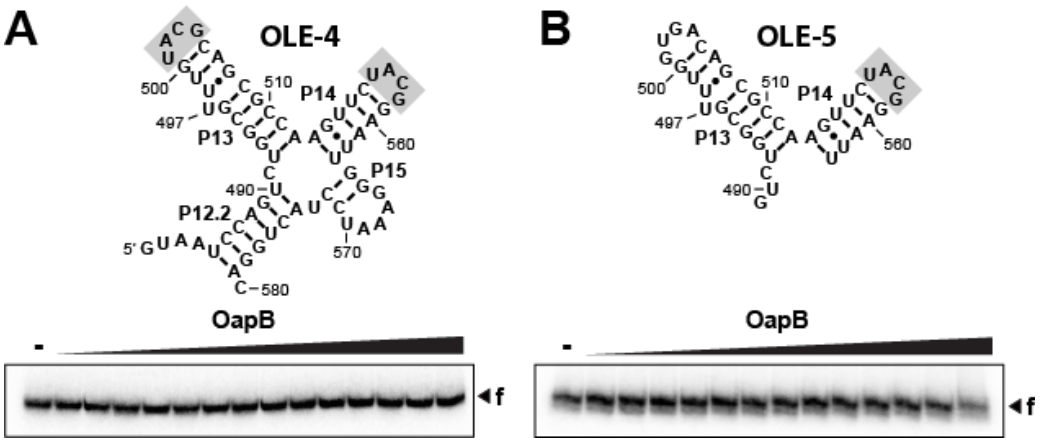

Figure 5. Modifications to the minimized RNA construct OLE-2 that eliminate OapB binding. *A*, Replacement of the GNRA tetraloop in OLE-2 with a UNCG tetraloop (OLE-4) abolishes binding to OapB. Shaded boxes designate non-natural nucleotide sequences used to create the OLE-4 construct. OapB concentrations in the EMSA reactions range from 1 pM to 3.2 μM (half-log increments). Additional annotations are as described in the legend to **Fig. 3**. *B*, A truncated OLE-2 construct encompassing only P13 and a partial P14 stem (OLE-5) does not serve as a binding site for OapB. OapB concentrations range from 1 pM to 1 μM (half-log increments). Additional annotations are as described in the legend to **Fig. 3**. I performed the experiments shown in this figure.

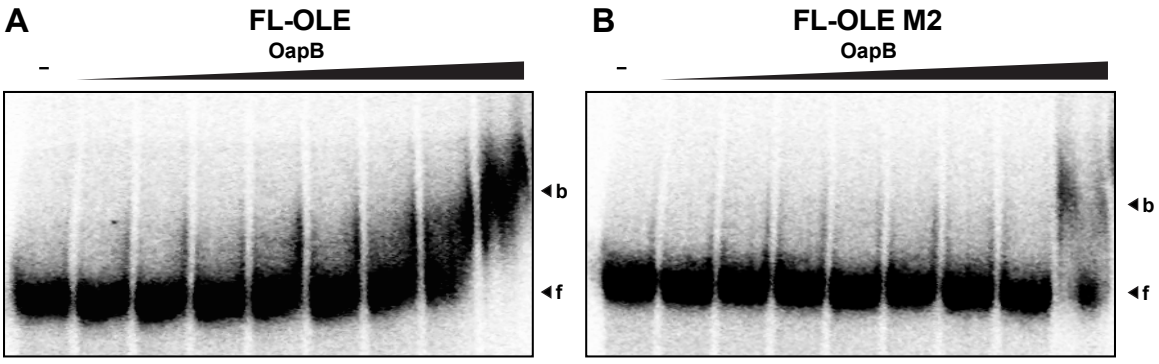

Figure 6. Binding of full-length wild-type and M2 OLE RNA to OapB. *A*, EMSA autoradiogram of OapB binding to full-length wild-type OLE RNA (FL-OLE). *B*, EMSA autoradiogram of OapB binding to full-length OLE RNA with the P13 GNRA tetraloop mutated to a UNCG tetraloop (FL-OLE M2). OapB concentrations range from 100 pM to 320 nM (half-log increments). Bands corresponding to unbound or "free" RNA (f) and bound RNA (b). I performed the experiments shown in this figure.

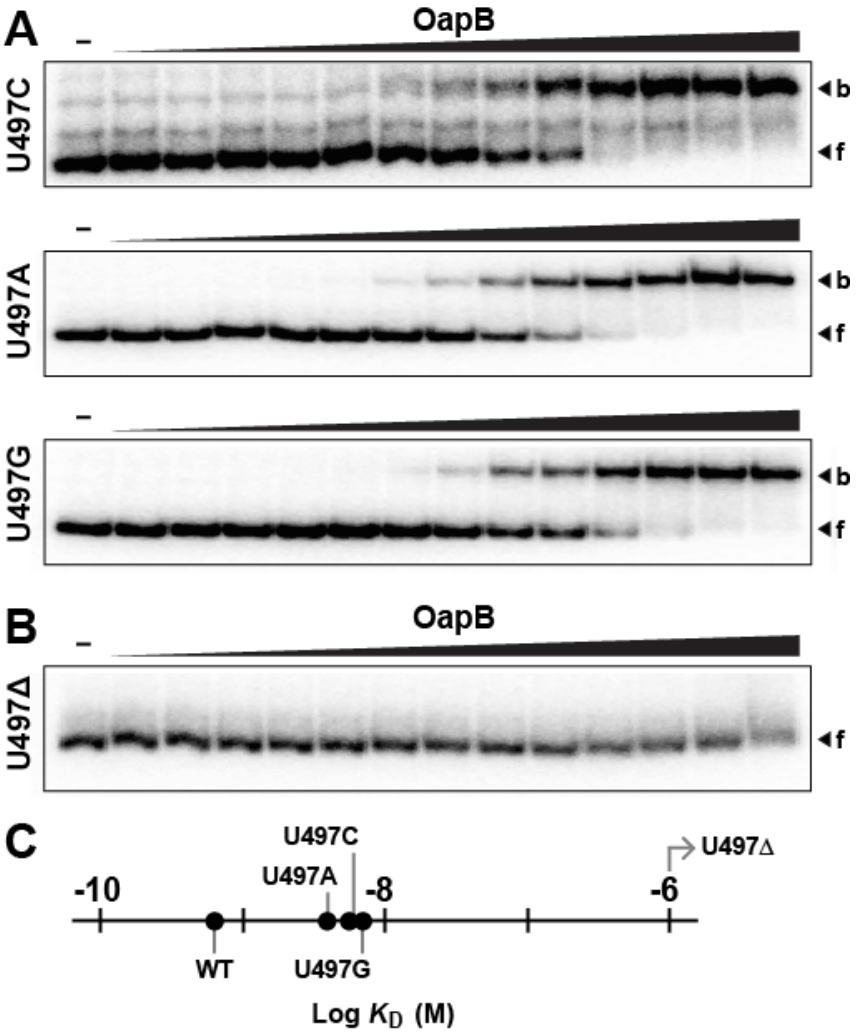

Figure 7. A bulged nucleotide in P13 is required for OapB binding. *A*, EMSA autoradiograms of OapB binding to OLE-2 constructs wherein U497 has been mutated to C, A or G, respectively. OapB concentrations range from 1 pM to 1 μM (half-log increments). Additional annotations are as described in the legend to **Fig. 3B**. *B*, EMSA autoradiogram of OapB binding to an OLE-2 construct wherein nucleotide 497 has been deleted (U497Δ). *C*, Plot of the K_D values for OapB binding to the WT OLE-2 construct, and for the various constructs with alterations at position 497. Note that the U497Δ construct has a K_D value that is poorer than 10^{-6} M. Data were derived from the EMSA assay data depicted in A and B. I performed the experiments shown in this figure.

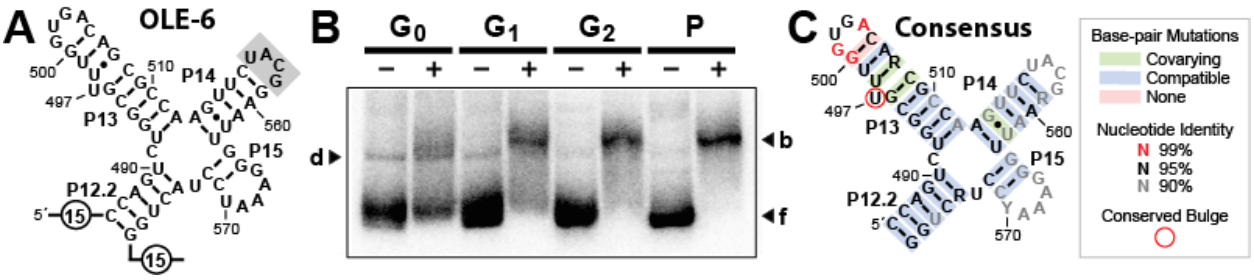

Figure 8. *In vitro* selection of mutant OLE-derived RNAs that are bound by OapB. *A*, Sequence and secondary structure model for the *in vitro* selection construct OLE-6, wherein the nucleotides depicted were mutagenized to a level of 6% degeneracy. Circled numbers identify the primer binding sites and the shaded nucleotides identify the UNCG tetraloop replacing the natural intervening sequences of OLE RNA. *B*, EMSA results for G_0 through G_2 RNA populations, and for the unmodified parental (P) OLE-6 construct. A band representing a putative protein-independent RNA dimer (d) is annotated. Additional annotations are as described in the legend to **Fig. 3**. *C*, Consensus model for the high-affinity OapB binding site based on the *B. halodurans* OLE RNP system. The consensus model was prepared by R2R using a covariation threshold of 10% (26). Red nucleotides are >99% conserved among the RNAs in the G_2 population with 500 or more representatives. I performed the experiments shown in this figure.

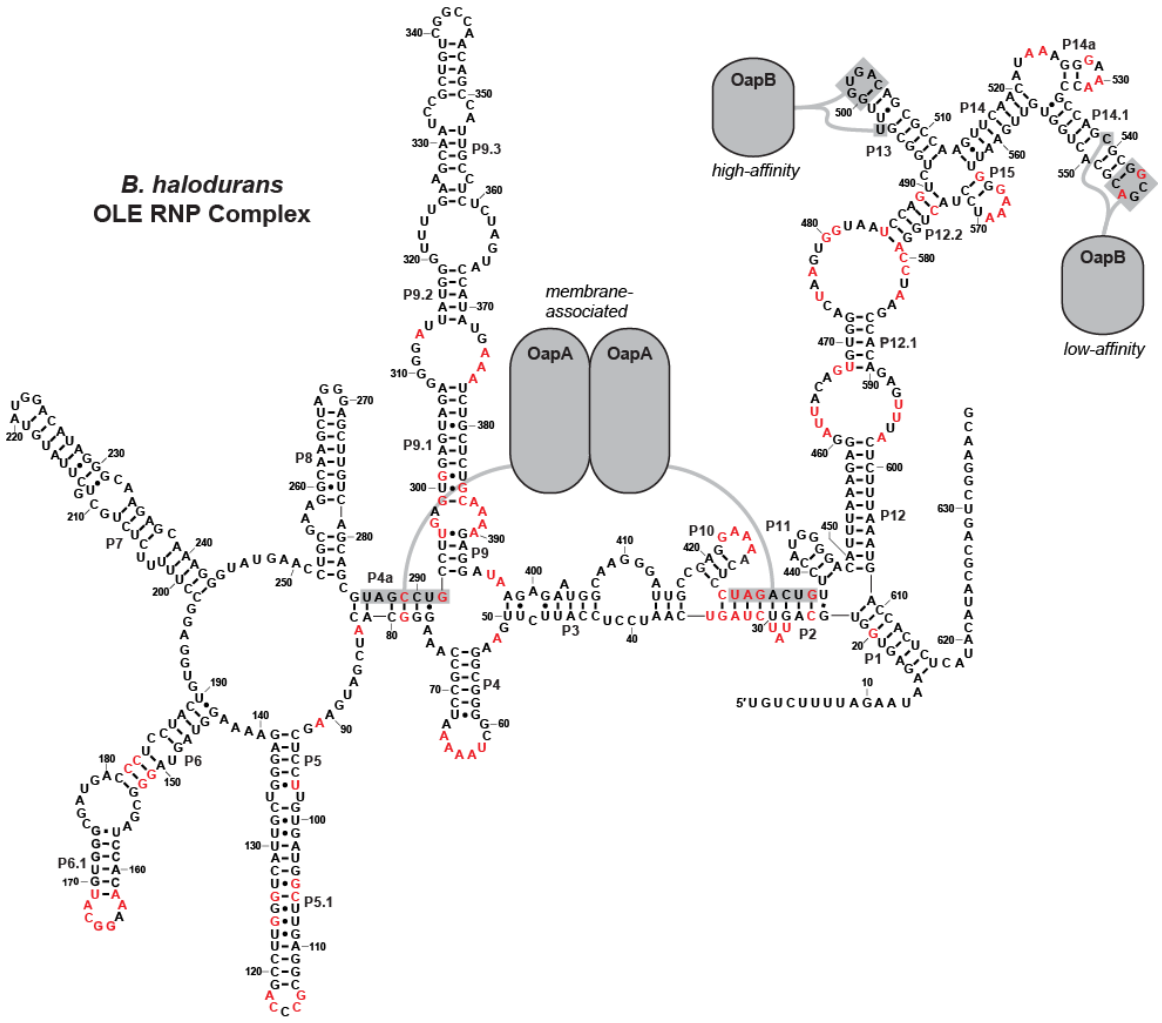

Figure 9. Model for OLE RNP complex.

A model for the *B. halodurans* OLE RNP complex showing where OapA and OapB interact. Gray shaded nucleotides are required for protein binding. Red residues are greater than 97% conserved across all OLE RNA examples (13).

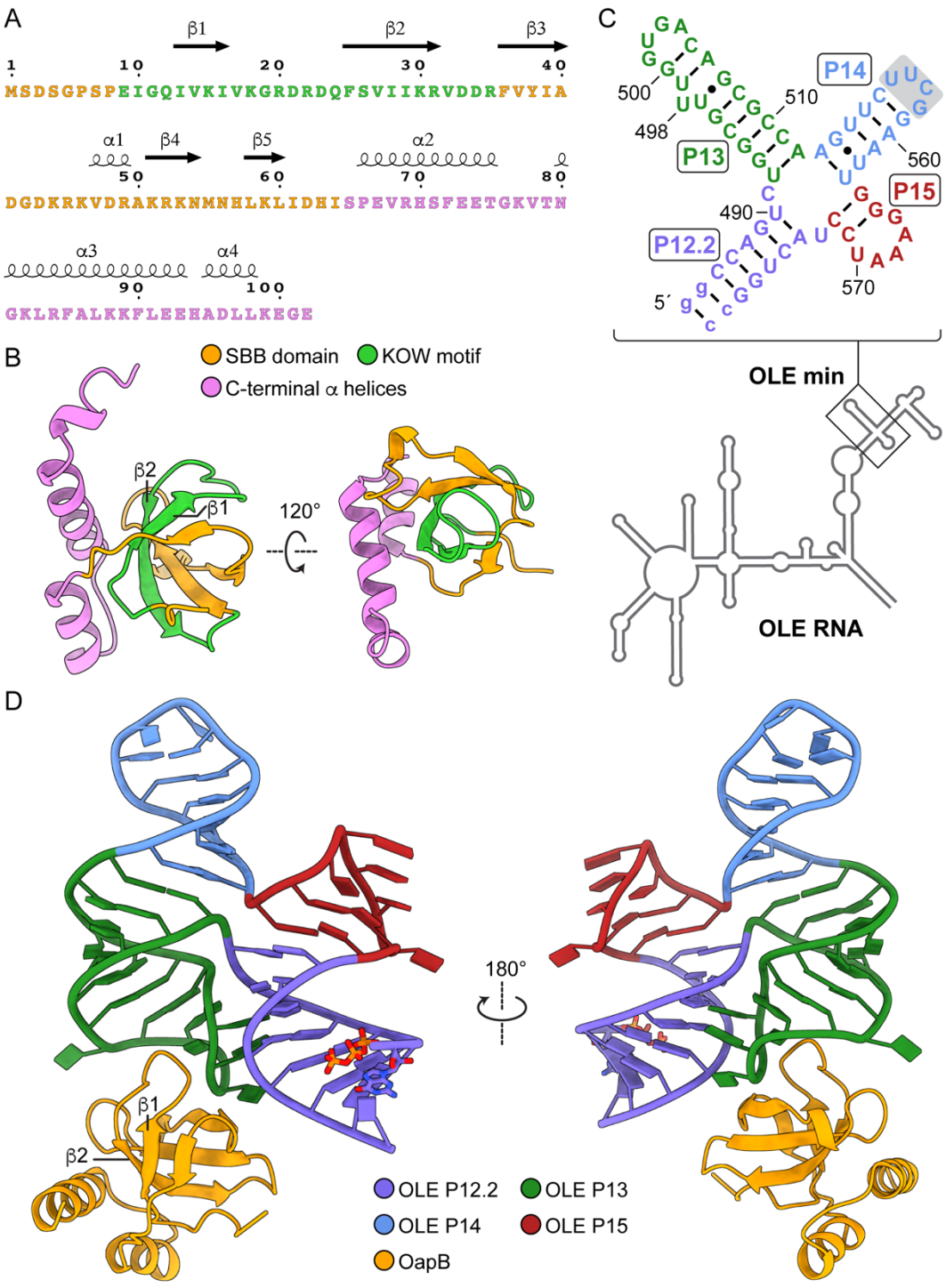

Fig. 10. Constructs and structures of OapB and the OapB-OLE min RNA complex. *A*, Sequence and secondary structure and *B*, crystal structure of *B. halodurans* OapB. The N-terminal SBB domain is colored orange, the C-terminal helices are colored violet, and

the conserved 27-residue KOW motif is colored green. *C*, Sequence and secondary structure model of the OLE min RNA used in crystallization. Stem P14 is capped with a non-native UUCG tetraloop (shaded box). Stem P12.2 is made more stable with two non-native G-C base pairs (lowercase letters). *D*, Front and back view of OapB-OLE min RNA complex crystal structure. Dr. Yang performed the experiments and analysis shown in this figure.

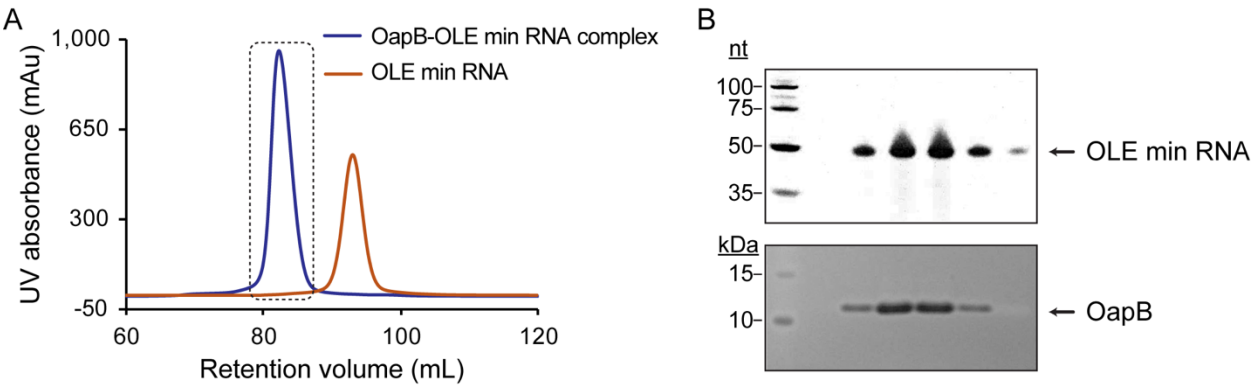

Figure 11. OapB-OLE min RNA complex reconstitution. *A*, Size-exclusion chromatography (SEC) profiles of OLE min RNA alone (orange) and OapB-OLE min RNA complex (blue). Due to lack of tryptophan residues in OapB, the peak corresponding to excessive OapB is not visible. *B*, The SEC purification fractions in the dashed line box were resolved on denaturing polyacrylamide gels and stained with SYBR Gold for RNA or stained with Coomassie blue for protein. A denatured double-stranded DNA marker was used in the gel stained with SYBR Gold. Dr. Yang performed the experiments shown in this figure.

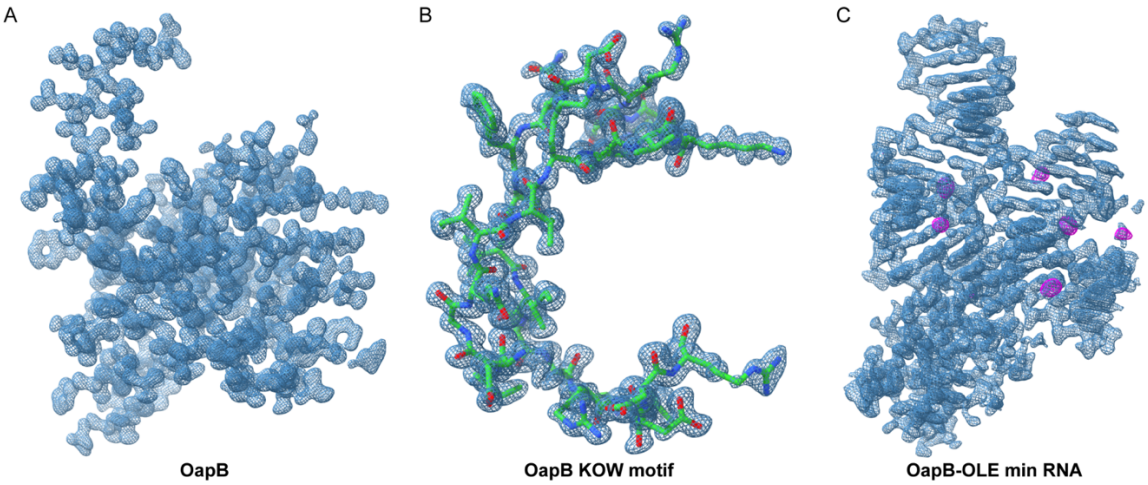

Figure 12. Electron density maps for OapB and OapB-OLE min RNA complex structures. *A*, Final refined $2F_o$-F_c map of the 1.0 Å OapB apo structure contoured at 1.0 σ. *B*, Structural model of the KOW motif in OapB superimposed with the refined $2F_o$-F_c map contoured at 1.0 σ. *C*, Final refined $2F_o$-F_c map (blue mesh) of the 2.1 Å OapB-OLE min RNA complex structure contoured at 1.0 σ. The anomalous difference map (magenta mesh) contoured at 5.0 σ shows six cobalt (III) hexammine-binding sites in OLE min RNA. Dr. Yang performed the analysis shown in this figure.

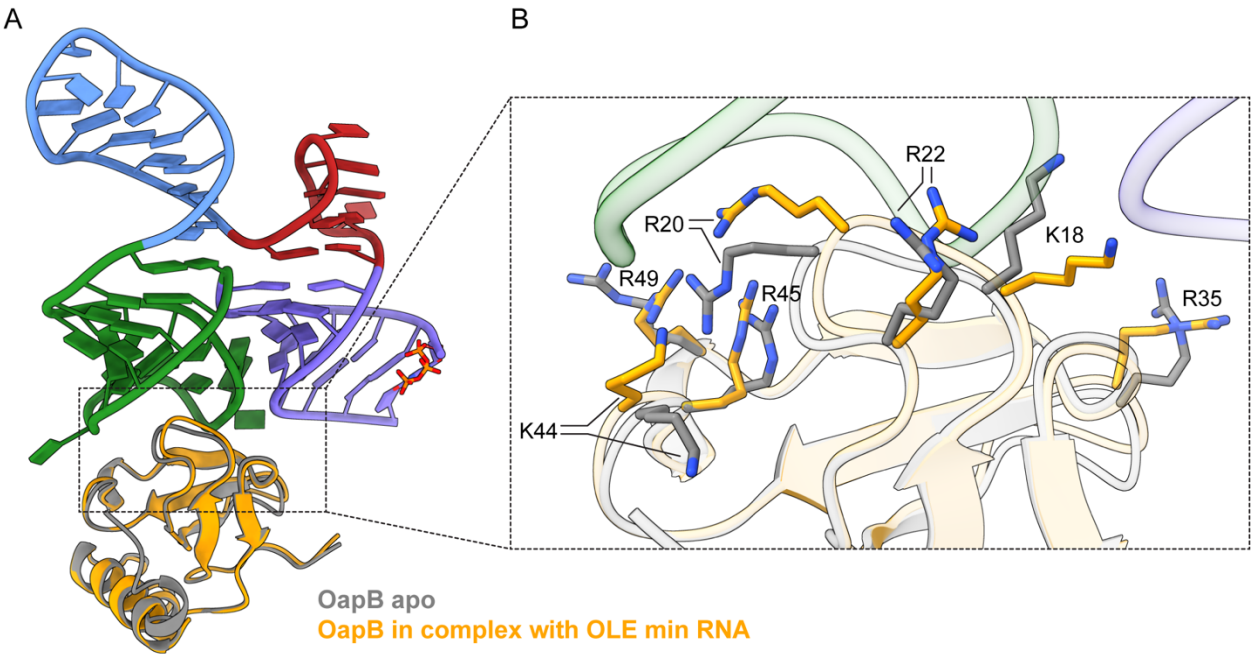

Figure 13. Conformational changes of OapB upon OLE RNA binding. *A*, Superimposition of OapB apo and OapB-OLE min RNA complex structures reveals notable conformational changes in several loops regions of OapB. *B*, Expanded view of the conformational changes of OapB residues at the OLE RNA-binding interface. Yang Yang performed the analysis shown in this figure. Dr. Yang performed the analysis shown in this figure.

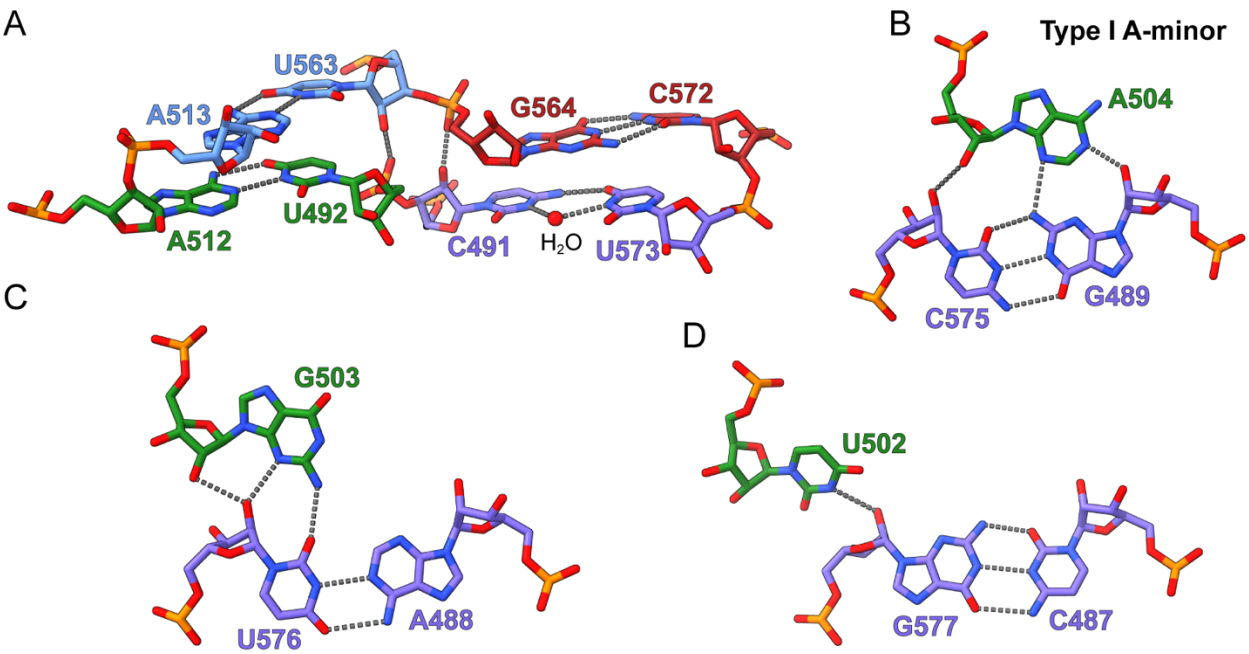

Fig. 14. Tertiary interactions formed by OLE min RNA. *A*, Base-pairs observed at the 4H junction of OLE min RNA. Nucleotides from stems P12.2, P13, P14 and P15 are colored in purple, green, blue and red, respectively. The water molecule mediating the non-canonical C491-U573 base-pair in stem P12.2 is shown as a red sphere. Hydrogen bonds are shown as gray dashed lines. *B*–*D*, Minor groove base triple interactions formed between P12.2 helix and *B*, A504, *C*, G503 and *D*, U502 in the P13 hairpin tetraloop. Dr. Yang performed the analysis shown in this figure.

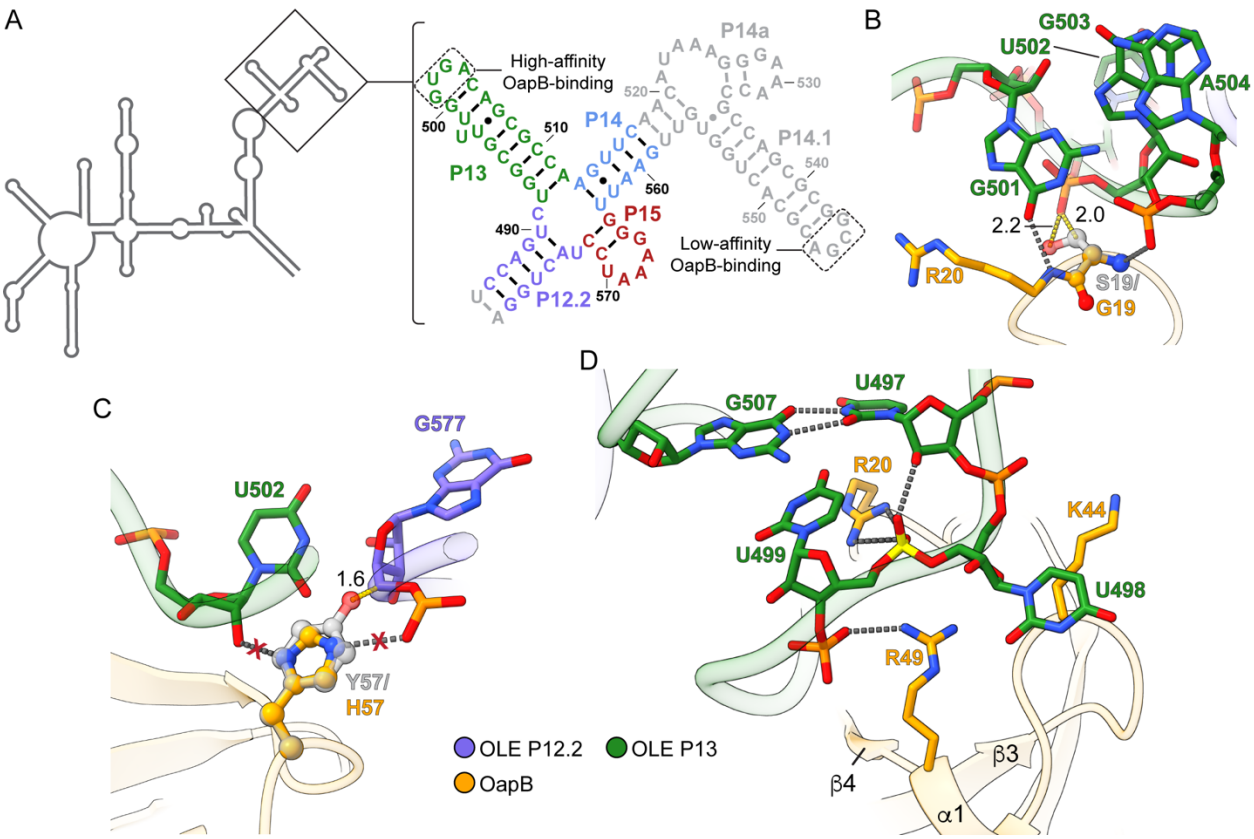

Figure 15. Molecular details of the OapB-OLE RNA interactions. *A*, Sequence and secondary structure models of the OLE RNA substructure with two OapB binding sites. Nucleotides not present in the OLE min RNA used for crystallization are shown as gray. *B*, Modeling of the G19S mutation in the OapB-OLE min RNA complex structure. G19 (orange) and S19 (transparent gray) are shown in ball-and-stick representation. Hydrogen bonds are shown as gray dashed lines. Potential steric clashes upon mutation are indicated by yellow dashed lines with distances indicated (in Å). *C*, Modeling of the H57Y mutation in the OapB-OLE min RNA complex structure. H57 (orange) and Y57 (transparent gray) are shown in ball-and-stick representation. Hydrogen bonds that would be lost due to H57Y mutation are labeled with red crosses. The potential steric clash upon mutation is indicated by yellow dashed lines with the distance indicated (in Å). *D*, Detailed interactions around

the bulged U498 region in crystal form II of the OapB-OLE min RNA complex. The phosphate group connecting U498 and U499 is highlighted in yellow. Dr. Yang performed the analysis shown in this figure.

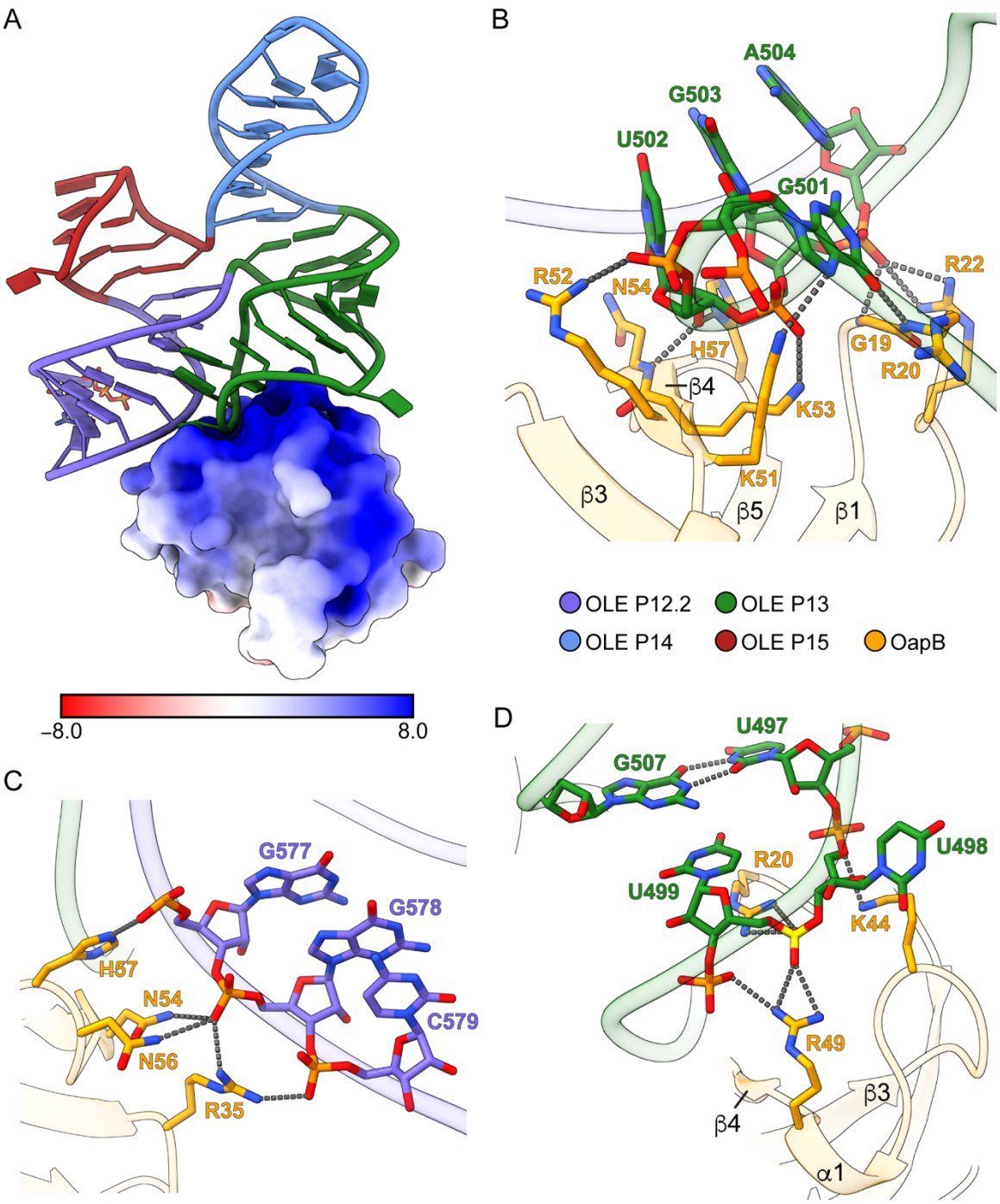

Fig. 16. Molecular basis of OLE RNA recognition by OapB. *A*, Electrostatic surface representation of OapB in complex with OLE min RNA. The unit of electrostatic potential is kT/e^{-}. *B*, Close-up view of the OapB-OLE RNA interactions at the P13 GNRA tetraloop

region. Hydrogen bonds are shown as gray dashed lines. *C*, Close-up view of the interactions between OapB and the helix backbone of P12.2. *D*, Close-up view of the OapB-OLE RNA interactions near the bulged U498 nucleotide. The phosphate group connecting U498 and U499 is highlighted in yellow. Dr. Yang performed the analysis shown in this figure.

Table 1. Synthetic DNA oligonucleotides. Yellow shading denotes nucleotide changes from the WT or parental constructs.

Name	Description	Sequence
KH199	OLE-1, forward	TAATACGACTCACTATAGGTAATCCAGTCTGG CGTTTGGTGACAGCGCCAAGTTCAACATAAAG GGAAACCG
KH200	OLE-1, reverse	GTCCAGTAGGATTTCCCAATTCAACACCAGTG CGTCGCCGCGCTGGCGGTTTCCCTTTATGTTGA ACTTGGCG
DLW070	M1 OLE-1, forward	TAATACGACTCACTATAGGTAATCCAGTCTGG CGTTTGGTGACAGCGCCAAGTTCAACATAAAG GGAA
DLW071	M1 OLE-1, reverse	GTCCAGTAGGATTTCCCAATTCAACACCAGTG CGCGTACGCGCTGGCGGTTTCCCTTTATGTTGA ACT
DLW072	OLE-3, forward	TAATACGACTCACTATAGTTCAACATAAAGGG AAACCGCCAGC
DLW073	OLE-3, reverse	ATTCAACACCAGTGCGTCGCCGCGCTGGCGGT TTCCCTTTATG
DLW074	OLE-2, forward	TAATACGACTCACTATAGGTAATCCAGTCTGG CGTTTGGTGACAGCGCCAA
DLW075	OLE-2, reverse	GTCCAGTAGGATTTCCCAATTCCGTAGAACTTG GCGCTGTCACCAAACGC
DLW076	OLE 513-566 construct forward	TAATACGACTCACTATAGTTCAACATAAAGGG AAACCGCCAGCGCGGCGACGCACTGGTGTTGA
DLW077	OLE 511-566 construct reverse	TCAACACCAGTGCGTCGCCGCGCTGGCGGTTT CCCTTTATGTTGAACTATAGTGAGTCGTATTA
DLW098 (RNA)	OLE P13 construct	GUCUGGCGUUUGGUGACAGCGCCAAG
DLW099	OLE-4, forward	TAATACGACTCACTATAGGTAATCCAGTCTGG CGTTTGTACGCAGCGCCAA
DLW100	OLE-4, reverse	GTCCAGTAGGATTTCCCAATTCGCTAGAACTTG GCGCTGCGTACAAACGC
DLW111	M3 OLE-1, Forward	TAATACGACTCACTATAGGTAATCCAGTCTGG CGTTTGTACGCAGCGCCAAGTTCAACATAAAG GGAA
DLW112	M3 OLE-1, Reverse	GTCCAGTAGGATTTCCCAATTCAACACCAGTG CGCGTACGCGCTGGCGGTTTCCCTTTATGTTGA ACT
DLW113	M2 OLE-1, Forward	TAATACGACTCACTATAGGTAATCCAGTCTGG CGTTTGTACGCAGCGCCAAGTTCAACATAAAG GGAA
DLW114	M2 OLE-1, Reverse	GTCCAGTAGGATTTCCCAATTCAACACCAGTG CGTCGCCGCGCTGGCGGTTTCCCTTTATGTTGA ACT

DLW115 (RNA)	OLE-5	GUCUGGCGUUUGGUGACAGCGCCAAGUUCUA CGGAAUU
DLW137	T7 RNAP Promoter+15 fixed nts+mutagenized region+15 fixed residues	TAATACGACTCACTATAGGGCTATGGACTGAA TCCAGTCTGGCGTTTGGTGACAGCGCCAAGTTC TACGGAATTGGGAAATCCTACTGGACCTAGTT CATACGG
DLW140	T7 RNAP Promoter +5 nts for extending mutagenesis library	TAATACGACTCACTATAGGGCTA
DLW145	For DLW137 RT, reverse	CCGTATGAACTAGGT
DLW167	For DLW137 RT, forward, T7 promoter + 15 bases	TAATACGACTCACTATAGGGCTATGGACTGAA T
DLW168	OLE-2, U490C and U490C + A574G forward	TAATACGACTCACTATAGGTAATCCAGCCTGG CGTTTGGTGACAGCGCCAA
DLW170	OLE-2, U490C + A574G reverse	GTCCAGCAGGATTTCCCAATTCCGTAGAACTTG GCGCTGTCACCAAACGC
DLW171	OLE-2, U497C forward	TAATACGACTCACTATAGGTAATCCAGTCTGG CGCTTGGTGACAGCGCCAA
DLW172	OLE-2, U497C reverse	GTCCAGTAGGATTTCCCAATTCCGTAGAACTTG GCGCTGTCACCAAGCGC
DLW173	OLE-2, Δ497 forward	TAATACGACTCACTATAGGTAATCCAGTCTGG CGTTGGTGACAGCGCCAA
DLW174	OLE-2, Δ497 reverse	GTCCAGTAGGATTTCCCAATTCCGTAGAACTTG GCGCTGTCACCAACGCC
DLW175	OLE-2, U498C forward	TAATACGACTCACTATAGGTAATCCAGTCTGG CGTCTGGTGACAGCGCCAA
DLW176	OLE-2, U498C reverse	GTCCAGTAGGATTTCCCAATTCCGTAGAACTTG GCGCTGTCACCAGACGC
DLW177	OLE-2, G500A forward	TAATACGACTCACTATAGGTAATCCAGTCTGG CGTTTAGTGACAGCGCCAA
DLW178	OLE-2, G500A reverse	GTCCAGTAGGATTTCCCAATTCCGTAGAACTTG GCGCTGTCACTAAACGC
DLW181	OLE-2, U497G forward	TAATACGACTCACTATAGGTAATCCAGTCTGG CGGTTGGTGACAGCGCCAA
DLW182	OLE-2, U497G reverse	GTCCAGTAGGATTTCCCAATTCCGTAGAACTTG GCGCTGTCACCACACGC
DLW183	OLE-2, G500A + C505U forward	TAATACGACTCACTATAGGTAATCCAGTCTGG CGTTTAGTGATAGCGCCAA
DLW184	OLE-2, G500A + C505U reverse	GTCCAGTAGGATTTCCCAATTCCGTAGAACTTG GCGCTATCACTAAACGC
DLW215	OLE-2, U497A forward	TAATACGACTCACTATAGGTAATCCAGTCTGG CGATTGGTGACAGCGCCAA

DLW216	OLE-2, U497A reverse	GTCCAGTAGGATTTCCCAATTCCGTAGAACTTG GCGCTGTCACCATACGC

Table 2. Computational search results for additional OapB-interacting RNAs in *B. halodurans*.

Hit Number	Start-End	Strand	E value	Gene
1	2908483-2908467	-	9.0×10^{-8}	*ole*
2	3562330-3562314	-	9.2×10^{-5}	AYT26_RS17310 (Na+ /H+ antiporters NhaC family protein)
3	618968-618946	-	0.023	AYT26_RS03175 (GNAT family N-acetyltransferase)
4	985772-985757	-	0.17	AYT26_RS04795 (6-phospho-beta-glucosidase)
5	508763-508785	+	0.18	AYT26_RS02615 (helix-turn-helix domain-containing protein)
6	2882577-2882593	+	0.33	*xylA* (xylose isomerase)
7	2679483-2679463	-	0.35	*sigE* (RNA polymerase sporulation sigma factor SigE)
8	273046-273010	-	0.5	*urtB* (urea ABC transporter permease subunit UrtB)
9	3621304-3621289	-	0.52	AYT26_RS17615 (hypothetical protein)
10	764127-764111	-	0.58	*uxaC* (glucuronate isomerase)
11	2130805-2130785	-	0.86	AYT26_RS10300 (cellulase family glycosylhydrolase)
12	2605020-2605004	-	0.95	*smc* (chromosome segregation protein SMC)

Table 3. Computational search results for additional OapB-interacting RNAs in *B. subtilis* subsp. subtilis str. 168.

Hit Number	Start-End	Strand	E value	Gene
1	2546679-2526693	+	0.016	*gcvPB* (glycine decarboxylase (subunit 2)(glycine cleavage system protein P))
2	3184232-3184258	+	0.043	*gbsB* (choline dehydrogenase)
3	3433691-3433675	-	0.11	*cysJ* (assimilatory sulfite reductase (flavoprotein alpha-subunit))
4	1460057-1460041	-	0.21	*ptsI* (phosphotransferase system (PTS) enzyme I)
5	934132-934105	-	0.27	*yfhO* (conserved membrane protein)
6	1071126-1071142	+	0.39	*prsA* (molecular chaperone lipoprotein)
7	3960227-3960242	+	0.52	*licC* (phosphotransferase system (PTS) lichenan-specific enzyme IIC component)
8	969585-969598	+	0.55	*yhbB* (conserved hypothetical protein)
9	1946584-1946600	+	0.59	*yngA* (putative conserved membrane protein possibly involved in arabinogalactan metabolism)
10	1541126-1541169	+	0.7	*nprE* (extracellular neutral metalloprotease)
11	681656-681676	+	0.83	*ydjP* (putative aminoacrylate hydrolase)
12	884096-884076	-	0.87	*acoR* (transcriptional regulator (AcrR-acetoin))

Table 4. X-ray crystallography data collection, phasing and refinement statistics

	OapB apo native (PDB XXXX)	OapB apo I-derivative (PDB XXXX)	OapB-OLE min RNA complex (Crystal form I) (PDB XXXX)	OapB-OLE min RNA complex (Crystal form II) (PDB XXXX)
Data collection				
Space group	*P* 1 2_1 1	*P* 1 2_1 1	*R* 3 2	*P* 6_5
Cell dimensions				
a, *b*, *c* (Å)	47.918, 34.909, 49.811	47.904, 34.912, 49.637	98.817, 98.817, 198.897	71.99, 71.99, 223.905
α, β, γ (°)	90, 90.24, 90	90, 90.151, 90	90, 90, 120	90, 90, 120
Resolution (Å)	47.92–1.20 (1.245–1.20)*	49.64–1.00 (1.036–1.0)	41.83–2.098 (2.173–2.098)	54.47–2.1 (2.175–2.1)
R_{sym} or R_{merge}	0.034 (0.156)	0.062 (1.24)	0.046 (1.22)	0.044 (0.860)
I / σI	17.8 (5.1)	10.8 (1.1)	29.1 (1.7)	16.4 (1.2)
Completeness (%)	98.3 (92.9)	96.4 (92.1)	99.7 (99.1)	99.2 (99.2)
Redundancy	3.3 (2.8)	4.2 (4.1)	9.6 (9.3)	3.5 (3.6)
Refinement				
Resolution (Å)	47.92–1.20 (1.245–1.20)	49.64–1.00 (1.036–1.0)	41.83–2.098 (2.173–2.098)	54.47–2.1 (2.175–2.1)
No. reflections	50646 (4788)	85600 (8113)	22151 (2149)	37854 (3791)
R_{work} / R_{free}	0.1212/0.1521	0.1488/0.1628	0.1658/0.1997	0.1717/0.2041
No. atoms				
Macromolecules	1761	1788	2007	3969
Ligand/ion	2	17	97	64
Water	314	239	240	404
B-factors				
Macromolecules	14.26	13.26	72.18	70.19
Ligand/ion	14.16	20.48	89.04	84.33
Water	28.7	25.75	74.05	68.42
R.m.s deviations				
Bond lengths (Å)	0.010	0.011	0.002	0.006
Bond angles (°)	1.14	1.12	0.51	1.05
Validation				
MolProbity score	1.19	1.46	1.22	1.39
Clashscore	2.23	4.38	2.24	3.21
Poor rotamers (%)	0.00	0.00	0.00	0.00
Ramachandran plot				
Favored (%)	96.88	96.35	96.59	95.93
Allowed (%)	3.12	3.65	3.41	4.07
Disallowed (%)	0.00	0.00	0.00	0.00

*Values in parentheses are for highest-resolution shell.

Chapter Three

Disruption of the OLE RNP complex causes differential regulation of metal ion transporters

Author contributions: D.L.W. and R.R.B. designed experiments. D.L.W. and C.F. (Chrishan Fernando) conducted experiments. D.L.W., A.N. (Aya Narunsky), and S.L. (Seth Lyon) analyzed data. D.L.W. and A.N. wrote the chapter.

Summary

OLE RNA is atypical for a bacterial RNA in that it is large, highly structured, and localizes to the cellular membrane through an interaction with a protein partner. This protein partner, OapA, possesses some sequence similarity to known magnesium (Mg^{2+}) transporters, leading to the discovery that *Bacillus halodurans* becomes sensitive to Mg^{2+} when the *ole* and/or *oapA* genes are knocked out. To understand how the OLE ribonucleoprotein (RNP) complex affects *B. halodurans* gene expression total RNA was isolated from strains with wild type and disrupted OLE RNP complexes. From this it was found that under standard growth conditions (37 °C, LB pH 10.0) strains with a disrupted OLE RNP complex downregulate two different Mg^{2+} importers. Whereas under stressed growth conditions (24 °C, 3% EtOH, or 5 mM $MgCl_2$) strains with a disrupted OLE RNP complex downregulate genes for predicted manganese (Mn^{2+}) importers and upregulated a Mn^{2+} riboswitch controlled *terC* class Mn^{2+} exporter. Additionally, several genes involved in glutamate and proline synthesis were upregulated in strains with a disrupted OLE RNP complex. With the RNA-seq data strengthening the evidence for OLE RNA playing a role in Mg^{2+} homeostasis, I hypothesized that OLE RNA might respond to intracellular Mg^{2+} by undergoing a structural rearrangement that regulates OapA function. In-line probing results confirmed that OLE RNA does indeed dramatically restructure at physiologically relevant Mg^{2+} concentrations, that this restructuring occurs similarly in OLE RNAs from different

species, and that it can be disrupted by a single point mutation. However, as large RNAs universally utilize cations for proper folding, this finding is only the first step to understanding how the OLE RNP complex regulates Mg^{2+} homeostasis.

Introduction

As discussed in chapter 2, OLE RNA has characteristics that suggest it almost certainly preforms a sophisticated biological function. It is large, highly structured, and contains many conserved residues, a rarity for bacterial noncoding RNAs (ncRNAs). When OLE RNA and/or its protein partner OapA (OLE-asscoiated protein A) are knocked out *Bacillus halodurans* exhibits a stress profile that is distinct from that of any known ncRNA. Knockouts of *ole*, *oapA*, or *ole-oapA* become sensitive to cold, short chain alcohols, and slightly elevated magnesium (Mg^{2+}) (Wallace et al. 2012, Harris et al. 2019). While sensitivity to cold and short chain alcohols are relatively common phenotypes due to the complexity of the response they elicit, sensitivity to slightly elevated Mg^{2+} is extremely uncommon (Armitano et al. 2016).

OLE RNA is also unusual in that it localizes to the cell membrane via an interaction with OapA (Block et al. 2011). OapA is a transmembrane protein with no close homologs (Block et al. 2011). However, in *Aeribacillus pallidus* the *oapA* gene is annotated as *citMHS* (Harris et al. 2019). In other species CitMHS is a Mg^{2+}/citrate symporter (Boorsma et al. 1996, Wakeman et al. 2014). Additionally, OapA contains a DUF21 domain that is found in various transporter proteins (El-Gebali et al. 2019). Two other, somewhat closely related DUF21 domain containing proteins have been shown to function as Mg^{2+} exporters (Akanuma et al. 2014, Armitano et al. 2016, Harris et al. 2019).

In this chapter I discuss additional evidence that the OLE ribonucleoprotein (RNP) complex functions in regulating Mg^{2+} homeostasis and provide the first evidence that the complex might also play a role in manganese (Mn^{2+}) homeostasis. With strong support for a connection between the OLE RNP complex and regulation of intracellular Mg^{2+} I began testing the specific hypothesis that OLE RNA might directly respond to increasing Mg^{2+} concentrations by altering its conformation. This chapter includes preliminary evidence that OLE RNA undergoes a conformational change as Mg^{2+} concentrations increase beyond the ideal physiological range.

Results and Discussion

RNA-seq reveals differential expression of multiple metal ion transporters in *B. halodurans* strains with disrupted OLE RNP complexes

Previous RNA-seq data showed that OLE RNA is abundant in *B. halodurans* and that levels of the RNA increase 5.4 fold under ethanol stress (Wallace et al. 2012). In this study I expand significantly on that work, looking for genes that are differentially regulated between wild type (WT), Δ*ole-oapA*, and PM1 strains of *B. halodurans* under standard (LB pH 10.0 at 37 °C) or one of three stressed growth conditions (24 °C, 3% EtOH v/v, or 5 mM $MgCl_2$).

Under standard growth conditions very few genes are substantially downregulated in strains with disrupted OLE RNP complexes (Δ*ole-oapA* and PM1) compared to the WT *B. halodurans*. In fact, the only gene to be substantially downregulated in both Δ*ole-oapA* and PM1 strains is a gene for an MgtE class magnesium importer (**Fig. 1 A**). A second gene for another MgtE class magnesium importer is also consistently downregulated in

both Δ*ole-oapA* and PM1 strains compared to WT, though it should be noted that this second gene has extremely low expression (<10 TPKM). Both these genes are regulated by at least one M-box RNA, a class of Mg^{2+} riboswitch that turns off gene expression when intracellular Mg^{2+} levels are elevated (Dann et al. 2007, Ramesh and Winkler 2010), suggesting that intracellular Mg^{2+} is high in Δ*ole-oapA* and PM1 strains. In addition to regulation at the transcriptional level, MgtE transporters lock into a closed conformation that prevents transport when intracellular Mg^{2+} concentrations are high (Hattori et al. 2007, Hattori et al. 2009). Multiple layers of control are typical for Mg^{2+} homeostasis in Firmicutes (Groisman et al. 2013, Chandrangsu et al. 2017, Tracshel et al. 2019).

Under stressed conditions (24 °C, 3% EtOH, or 5 mM $MgCl_2$) a larger number of genes were differentially regulated. While these genes varied greatly from one stressed condition to another, one set of related genes was consistently differentially regulated in both Δ*ole-oapA* and PM1 strains, Mn^{2+} importers and exporters (**Fig. 2 B**). Three genes that encode punitive Mn^{2+} importers or components of Mn^{2+} importers were consistently downregulated in Δ*ole-oapA* and PM1 strains compared to WT transcript levels. Two of the genes (AYT26_RS07215 and AYT26_RS07220) are adjacent to one another in the genome and likely regulated together.

In additional to the putative Mn^{2+} importers, a TerC-class Mn^{2+} exporter was differentially regulated between WT and strains with disrupted OLE RNP complexes (**Fig. 2 B**). The *terC* transcript is consistently upregulated in Δ*ole-oapA* and PM1 strains, except for one of the three replicates for the Δ*ole-oapA* strain under ethanol stress, which I believe to be a statistical outlier. The *B. halodurans terC* gene is controlled by a Mn^{2+} responsive riboswitch that terminates transcription when intracellular Mn^{2+} levels are low (Dambach

et al. 2015, Price et al. 2015). Interestingly, under standard growth condition none of these genes are differentially regulated (**Fig. 2 B**). These data suggest that under stressed conditions intracellular Mn^{2+} is elevated in Δ*ole-oapA* and PM1 strains. Whereas a previous connection had been observed between the OLE RNP complex and Mg^{2+}, this is the first evidence that the complex might also be involved in Mn^{2+} homeostasis.

Genes involved in glutamate and proline synthesis are frequently upregulated in *B. halodurans* strains with disrupted OLE RNP complexes

Several genes involved in glutamate and proline synthesis are upregulated in Δ*ole-oapA* and PM1 strains. Genes for glutamine-hydrolyzing carbamoyl-phosphate synthase small subunit (AYT26_RS12810), glutamine--fructose-6-phosphate transaminase (isomerizing) (*GFPT1*, AYT26_RS01550), glutamate 5-kinase (AYT26_RS07785), and glutamate-5-semialdehyde dehydrogenase (AYT26_RS07780) are upregulated in Δ*ole-oapA* and/or PM1 in at least two conditions (**Fig. 2**). Both glutamine-hydrolyzing carbamoyl-phosphate and GFPT1 are involved in catalysis of glutamine to glutamate (Piette et al. 1984, Teplyakov et al. 1999). From that point, glutamate 5-kinase and glutamate-5-semialdehyde dehydrogenase each catalyze a step in converting glutamate to proline (Majumdar et al. 2016). Other glutamate-related genes such as NADP-specific glutamate dehydrogenase (AYT26_RS10685), alpha-L-glutamate ligase (AYT26_RS04430), and glutamate synthase subunit beta (AYT26_RS08890) are sometimes differentially regulated, but in a less consistent manner.

This suggests that Δ*ole-oapA* and PM1 strains are producing a large amount of glutamate that is subsequently converted into proline. Overproduction of proline, a

common compatible solute, is a direct response to high osmolarity in closely related *Bacillus subtilis* (Brill et al. 2011). Why disruption of the OLE RNP complex causes increased osmolarity is unclear at this time, though I speculate that a disruption in divalent cation homeostasis might led to high osmolarity in *B. halodurans* through a yet unknown mechanism.

OLE RNA undergoes a structural rearrangement at biologically relevant Mg^{2+} concentrations

With numerous links between the OLE RNP complex and Mg^{2+} I began to hypothesize possible functions of OLE RNA in Mg^{2+} homeostasis. Given that structured ncRNAs have previously been observed responding to intracellular Mg^{2+} concentrations by switching from an 'on-state' to an 'off-state' (Cromie et al. 2006, Dann et al. 2007) I asked if OLE RNA could potentially do the same. If the OLE RNP complex functions in a manner similar to MpfA proteins, DUF21 domain containing Mg^{2+} exporters with some sequence similarity to OapA (Akanuma et al. 2014, Armitano et al. 2016, Harris et al. 2019), then it might be possible to predict the function of OLE RNA. MpfA contains three domains, the transmembrane DUF21 domain, a CBS domain, and a CorC/HlyC domain. Both the CBS and CorC/HlyC domains are responsible for regulation of Mg^{2+} transport by the DUF21 domain (Akanuma et al. 2014). While the exact regulatory mechanism for MpfA has not been established, in MgtE, another Mg^{2+} transporter, the CBS domain binds ATP, helping to coordinate Mg^{2+} binding and switch the transporter from an open to a closed conformation (Tomita et al. 2017). I hypothesized that OapA might function as a

Mg^{2+} transporter and that OLE RNA might take the place of the CBS and CorC/HlyC domains in regulating OapA.

To test this hypothesis, I first explored how OLE RNA responds to increasing Mg^{2+} concentrations. In-line probing of *B. halodurans* OLE RNA shows that the RNA undergoes a structural rearrangement near physiological Mg^{2+} concentrations (**Fig. 3**). Interestingly, two of the sites with the most pronounced modulation, C24 and C82, are within the base-paired regions that bind OapA (Block et al. 2011) (**Fig. 4**). One of those nucleotides, C24 shows a particularly unusual modulation pattern in which the band increases then decreases in intensity. This pattern suggests that the residue goes from a highly structured state at low Mg^{2+} to a relatively unstructured state around physiological Mg^{2+} concentrations, then returns to a highly structured state at hyper-physiological Mg^{2+} concentrations. This structural rearrangement is not specific to *B. halodurans* OLE RNA. *Fictibacillus gelatini* OLE RNA modulates similarly with increasing Mg^{2+} concentrations (**Fig. 5**).

A point mutation prevents structural rearrangement of OLE RNA at biologically relevant Mg^{2+} concentrations

Structural rearrangement of large ncRNAs is common under increasing Mg^{2+} (Draper 2004, Lambert and Draper. 2007, Dann et al. 2007), so in order to differentiate OLE RNA from other systems Chrishan Fernando and I sought to find a single point mutation that prevented modulation at biologically relevant Mg^{2+} concentrations. Before testing multiple point mutations, I verified that a truncated OLE RNA (containing nucleotides 30-433) no longer modulates (**Fig. 6**). With the $OLE_{30\text{-}433}$ construct modulation is disrupted at all positions except one, shown with an asterisk. I speculate that the position

that retains modulation does so because of local rather than global restructuring of the RNA.

Chrishan Fernando and I generated and tested several point mutants of OLE RNA to search for variants that prevented structural rearrangement at biological Mg^{2+} concentrations. Specific mutants were chosen using criteria such as targeting highly conserved nucleotides, positions potentially capable of long-range interactions (Wallace et al. 2012), and positions that modulate under in-line probing conditions. Six point mutations tested (C56A, G71U, C85A, G114U, U462A, and G566U) did not affect modulation of OLE RNA (data not shown). Only one point mutant, C36A, prevented the restructuring of OLE RNA under physiological Mg^{2+} concentrations (**Fig. 7**). While this result is encouraging and supports the hypothesis that OLE RNA might function as a fine-tuned sensor for intracellular Mg^{2+} concentrations, I stress that significant additional data is required to understand how OLE RNA and OapA regulate Mg^{2+} homeostasis.

Conclusions

Multiple lines of evidence suggest that the OLE RNP complex plays a critical role in Mg^{2+} homeostasis, though a specific mechanism of action remains elusive. All available data is consistent with the possibility that OapA is a magnesium exporter and that OLE RNA could function as an exceptionally sophisticated regulatory RNA. This hypothesis represents a potential novel function for a large ncRNA that is not completely without precedent. Some Mg^{2+} transporters are known to directly respond to intracellular Mg^{2+} concentrations by shifting from one conformation to another (Hattori et al. 2007, Hattori et al. 2009), as are ncRNAs (Dann et al. 2007). What is not clear from the available data is

why, if this hypothesis is correct, OLE RNA is as large and highly conserved as it is. This could be explained with one of two possibilities. First, the OLE RNP complex might have additional functions beyond Mg^{2+} homeostasis. Second, regulation of a transporter by a Mg^{2+}-sensing RNA might require more structural features and conservation than is typical for other roles ncRNAs fulfill. Regardless, additional experiments are required to show whether the structural change OLE RNA undergoes at physiological Mg^{2+} concentrations is biologically relevant.

This work represents the first evidence that the OLE RNP complex may also play a role in Mn^{2+} homeostasis. Though it is also possible that the differential expression of Mn^{2+} transporters is a consequence of disruptions to Mg^{2+} homeostasis, as it has recently been shown that loss and dysregulation of Mg^{2+} transporters in *B. subtilis* can protect cells from Mn^{2+} and Co^{2+} intoxication (Pi et al. 2020). Discovering why Mg^{2+} transporters are differentially regulated under standard growth conditions and Mn^{2+} transporters become differentially regulated when stress is introduced could have broader implications in understanding how bacteria cope with metal ion stress.

Materials and Methods

Isolation of Total RNA

The Δ*ole-oapA* and PM1 strains of *B. halodurans* used in this study were previously described in (Wallace et al. 2012) and (Harris et al. 2018) respectively. Cells were diluted to OD_{600} = 0.01 in LB pH 10.0 from an overnight culture and grown at 37 °C. For standard growth samples, the cells were grown to OD_{600} = 0.35-0.40 and harvested. For stressed conditions, the cells were grown to OD_{600} = 0.2 then exposed to the appropriate stress for

3 hours before harvest (24 °C, 3% EtOH v/v, or 5 mM $MgCl_2$). Cells were lysed by freeze-thawing in 100 µl TE buffer (10 mM Tris-HCl, 1 mM EDTA pH 8.0) + 3 mg/ml lysozyme. Total RNA was isolated using a Qiagen RNeasy Mini Kit (Cat. No. / ID: 74104). DNA was removed using the TURBO DNA-*free*™ kit (Thermo Fisher Scientific, Catalog number: AM1907). rRNA was removed from up to 10 µg of sample with the RiboMinus™ Transcriptome Isolation kit, Bacteria (Thermo Fisher Scientific, Catalog number: K155004). PCR for *ole* DNA was performed on relevant samples to ensure that all DNA had been digested. Purified RNA samples were then sent to the Yale Center for Genomic Analysis for deep sequencing. Each sample was sequenced to a depth of ~10 million reads.

Analysis of RNA-seq data

Data preparations. We conducted # experiments, each with a varying number of samples (6-#) under different stress conditions and extracted the RNA transcripts for each experiment as described above. We mapped the reads of the NGS sequencing data to the *Bacillus halodurans* C-125 reference genome (genome assembly GCF_000011145.1) using the HISAT2 alignment program (Kim et al. 2019). We then used the Picard command-line tools (http://broadinstitute.github.io/picard/) to filter reads of low quality, and the SAMtools (Li et al. 2009) package to sort the alignment and prepare the reads for processing.

Calculating the TPKM counts of the data. Using the Bedtools package, we calculated the coverage of the identified sequences in the reference genome (Quinlan et al. 2010). We assembled those reads into full-length transcripts with the StringTie assembler (Pertea et

al. 2015) and used DESeq2 (Love et al. 2014) to analyze the differentially expressed genes between the various samples. Finally, to evaluate the change in transcripts between the samples we calculate the transcripts per kilobase million (TPKM) for each of the genes. TPKM normalizes the number of transcripts to the length of the respective gene and the sequencing depth.

Generation of heatmaps. Heatmaps were made using the average TPM value per gene across biological replicates as the input data. Each TPM value was transformed by $\log_2$ and then the pheatmap package in R was used to scale each dataset and plot heatmaps (Gu et al. 2016).

RNA oligonucleotide preparation and in-line probing

Colony PCR was performed on *B. halodurans* as previously described (Block et al. 2011). RNA products were made by adding 10 µl of PCR product to a 100 µl *in vitro* transcription reaction and incubating at 37 °C for 2-4 hours. 5' ^{32}P-radiolabeling, purification, and in-line probing conditions were as previously described (Regulski and Breaker. 2008) with the exceptions that $MgCl_2$ was removed from the in-line probing buffer and the reaction was run for 24 hours.

Generation of point mutations in OLE RNA

Point mutations to OLE RNA were generated through site-directed mutagenesis of the WT pHCMC05::*ole-oapA* plasmid (Wallace et al. 2012) with the QuikChange II XL kit (Agilent, Part number: 200521).

References

Akanuma, G., Kobayashi, A., Suzuki, S., Kawamura, F., Shiwa, Y., Watanabe, S., Yoshikawa, H., Hanai, R., and Ishizuka, M. (2014) Defect in the formation of 70S ribosomes caused by lack of ribosomal protein L34 can be suppressed by magnesium. *J Bacteriol* **196**, 3820-3830

Armitano, J., Redder, P., Guimaraes, V. A., and Linder, P. (2016) An Essential Factor for High Mg(2+) Tolerance of Staphylococcus aureus. *Front Microbiol* **7**, 1888

Block, K. F., Puerta-Fernandez, E., Wallace, J. G., and Breaker, R. R. (2011) Association of OLE RNA with bacterial membranes via an RNA-protein interaction. *Mol. Microbiol***. 79**, 21-34

Boorsma, A., van der Rest, M. E., Lolkema, J. S., and Konings, W. N. (1996) Secondary transporters for citrate and the Mg(2+)-citrate complex in Bacillus subtilis are homologous proteins. *J Bacteriol* **178**, 6216-6222

Brill, J., Hoffmann, T., Bleisteiner, M., and Bremer, E. (2011) Osmotically controlled synthesis of the compatible solute proline is critical for cellular defense of Bacillus subtilis against high osmolarity. *J Bacteriol* **193**, 5335-5346

Chandrangsu, P., Rensing, C., and Helmann, J. D. (2017) Metal homeostasis and resistance in bacteria. *Nat Rev Microbiol* **15**, 338-350

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

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

Dann, C. E., 3rd, 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

Draper, D. E. (2004) A guide to ions and RNA structure. *RNA* **10**, 335-343

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. (2019) The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427-D432

Groisman, E. A., Hollands, K., Kriner, M. A., Lee, E. J., Park, S. Y., and Pontes, M. H. (2013) Bacterial Mg2+ homeostasis, transport, and virulence. *Annu Rev Genet* **47**, 625-646

Gu, Z., Eils, R., and Schlesner, M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849

Harris, K. A., and Breaker, R. R. (2018) Large noncoding RNAs in bacteria. *Microbiol. Spectr.* **6**, RWR-0005-2017

Harris, K. A., Zhou, Z., Peters, M. L., Wilkins, S. G., and Breaker, R. R. (2018) A second RNA-binding protein is essential for ethanol tolerance provided by the bacterial OLE ribonucleoprotein complex. *Proc. Natl. Acad. Sci. USA* **115**, E6319-E6328

Harris, K.A., Odzer, N.B., and Breaker, R.R. (2019) Disruption of the OLE ribonucleoprotein complex causes magnesium toxicity in *Bacillus halodurans*. *Mol. Microbiol.* **112**, 1552-1563

Hattori, M., Tanaka, Y., Fukai, S., Ishitani, R., and Nureki, O. (2007) Crystal structure of the MgtE Mg2+ transporter. *Nature* **448**, 1072-1075

Hattori, M., Iwase, N., Furuya, N., Tanaka, Y., Tsukazaki, T., Ishitani, R., Maguire, M. E., Ito, K., Maturana, A., and Nureki, O. (2009) Mg(2+)-dependent gating of bacterial MgtE channel underlies Mg(2+) homeostasis. *EMBO J* **28**, 3602-3612

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915

Lambert, D., and Draper, D. E. (2007) Effects of osmolytes on RNA secondary and tertiary structure stabilities and RNA-Mg2+ interactions. *J Mol Biol* **370**, 993-1005

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079

Love, M. I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550

Majumdar, R., Barchi, B., Turlapati, S. A., Gagne, M., Minocha, R., Long, S., and Minocha, S. C. (2016) Glutamate, Ornithine, Arginine, Proline, and Polyamine Metabolic Interactions: The Pathway Is Regulated at the Post-Transcriptional Level. *Front Plant Sci* **7**, 78

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290-295

Pi, H., Wendel, B. M., and Helmann, J. D. (2020) Dysregulation of Magnesium Transport Protects Bacillus subtilis against Manganese and Cobalt Intoxication. *J Bacteriol* **202**

Piette, J., Nyunoya, H., Lusty, C. J., Cunin, R., Weyens, G., Crabeel, M., Charlier, D., Glansdorff, N., and Pierard, A. (1984) DNA sequence of the carA gene and the control region of carAB: tandem promoters, respectively controlled by arginine and the pyrimidines, regulate the synthesis of carbamoyl-phosphate synthetase in Escherichia coli K-12. *Proc Natl Acad Sci U S A* **81**, 4134-4138

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

Puerta-Fernandez, E., Barrick, J. E., Roth, A., and Breaker, R. R. (2006) Identification of a large noncoding RNA in extremophilic eubacteria. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19490-19495

Quinlan, A. R., and Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842

Ramesh, A., and Winkler, W. C. (2010) Magnesium-sensing riboswitches in bacteria. *RNA Biol* **7**, 77-83

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

Teplyakov, A., Obmolova, G., Badet-Denisot, M. A., and Badet, B. (1999) The mechanism of sugar phosphate isomerization by glucosamine 6-phosphate synthase. *Protein Sci* **8**, 596-602

Tomita, A., Zhang, M., Jin, F., Zhuang, W., Takeda, H., Maruyama, T., Osawa, M., Hashimoto, K. I., Kawasaki, H., Ito, K., Dohmae, N., Ishitani, R., Shimada, I., Yan, Z., Hattori, M., and Nureki, O. (2017) ATP-dependent modulation of MgtE in Mg(2+) homeostasis. *Nat Commun* **8**, 148

Trachsel, E., Redder, P., Linder, P., and Armitano, J. (2019) Genetic screens reveal novel major and minor players in magnesium homeostasis of Staphylococcus aureus. *PLoS Genet* **15**, e1008336

Wakeman, C. A., Goodson, J. R., Zacharia, V. M., and Winkler, W. C. (2014) Assessment of the requirements for magnesium transporters in Bacillus subtilis. *J Bacteriol* **196**, 1206-1214

Wallace, J. G., Zhou, Z., and Breaker, R. R. (2012) OLE RNA protects extremophilic bacteria from alcohol toxicity. *Nucleic Acids Res.* **40**, 6898-6907

Weinberg, Z., Perreault, J., Meyer, M. M., and Breaker, R. R. (2009) Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. *Nature* **462**, 656-659

Figures and Tables

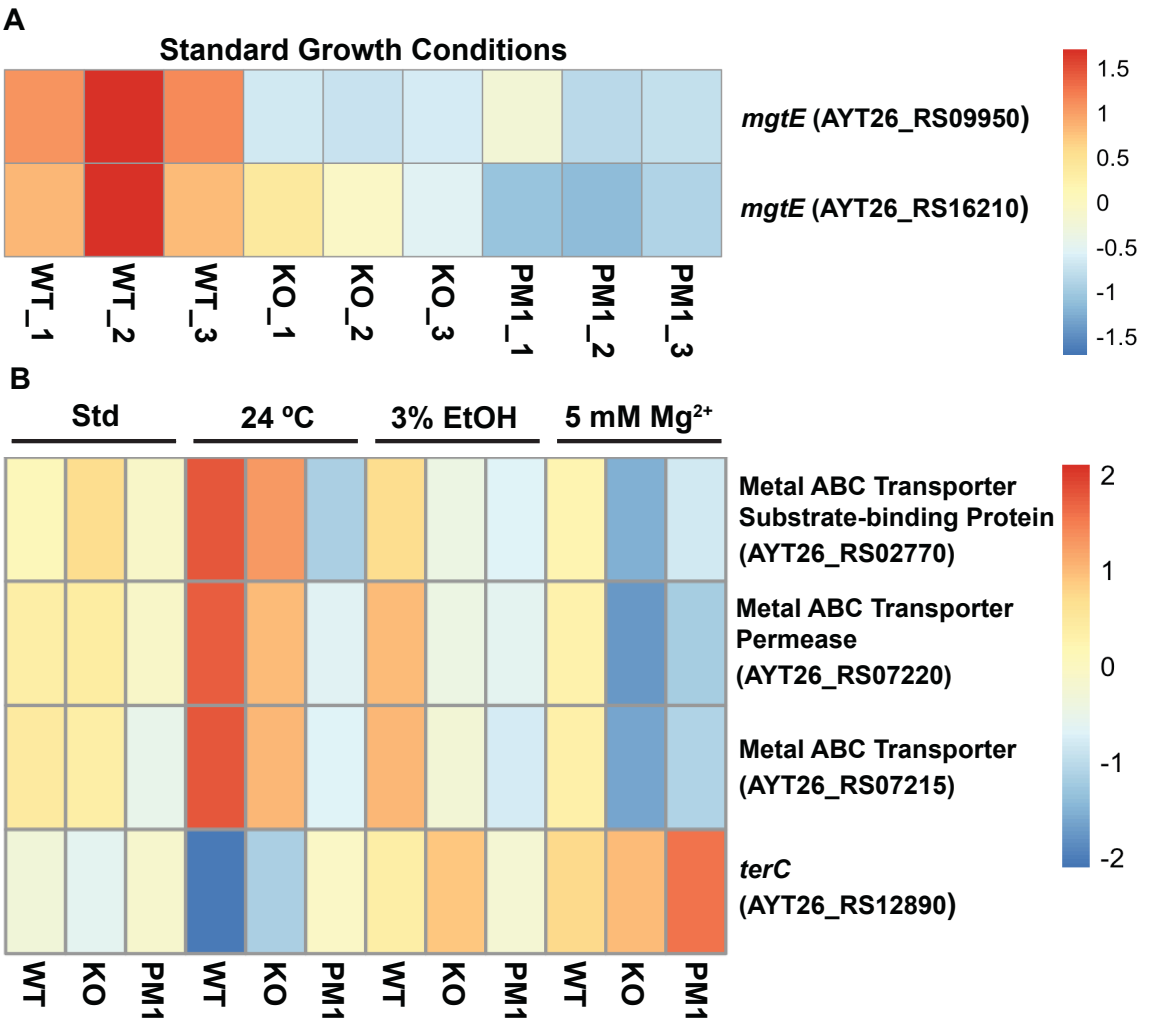

Figure 1. Differential expression of metal ion transporters in wild type (WT), Δ*ole-oapA* (KO), and PM1 strains of *Bacillus halodurans*. A) Expression of two putative *mgtE* genes (AYT26_RS09950 and AYT26_RS16210) under standard growth conditions (37 °C, LB pH 10.0). Each strain includes three replicates. B) Expression of three putative Mn^{2+} importers (AYT26_RS 02770, AYT26_RS07215, and AYT26_RS07220) and one putative *terC* class Mn^{2+} exporter (AYT26_RS12890) under standard growth conditions (Std) or stressed with either cold (24 °C), ethanol (3% v/v), or $MgCl_2$ (5 mM). Each grid section represents the average of three replicates. All like samples fall within one standard

deviation of one another except one of the three replicates KO strain *terC* under ethanol stress, which was expressed at significantly higher levels in that sample. I performed the experiments, and Aya Narunsky, Seth Lyon, and I performed the bioinformatic analysis for this figure.

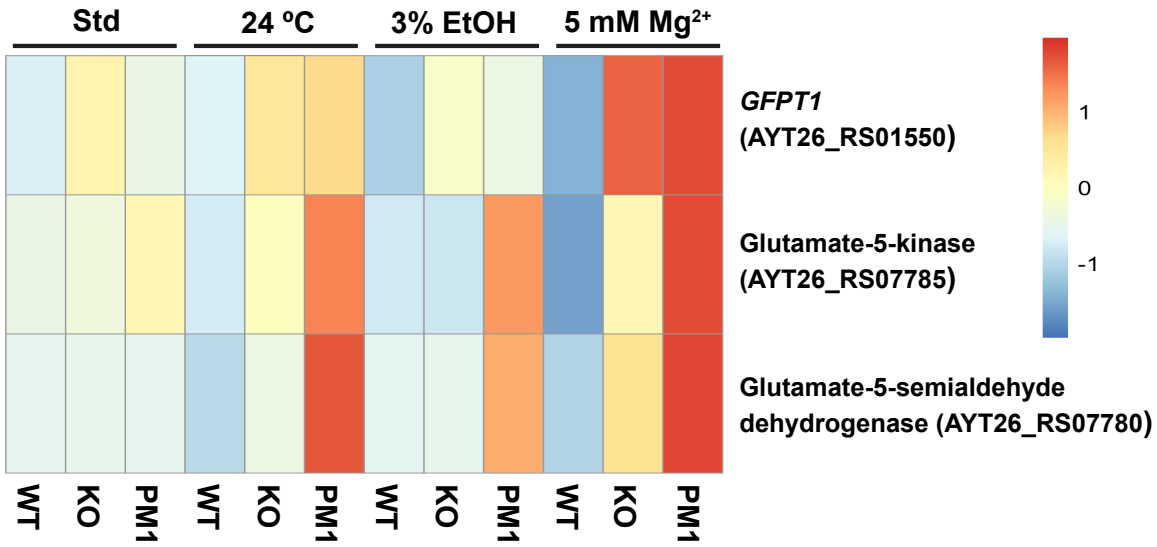

Figure 2. **Differential expression of glutamate and proline synthesis genes in wild type (WT), Δ*ole-oapA* (KO), and PM1 strains of *B. halodurans.*** Expression of three genes involved in the conversion of glutamine to glutamate and proline under standard growth conditions (Std) or stressed with either cold (24 °C), ethanol (3% EtOH v/v), or slightly elevated magnesium (5 mM $MgCl_2$). Each grid section represents the average of three replicates. I performed the experiments, and Aya Narunsky, Seth Lyon, and I performed the bioinformatic analysis for this figure.

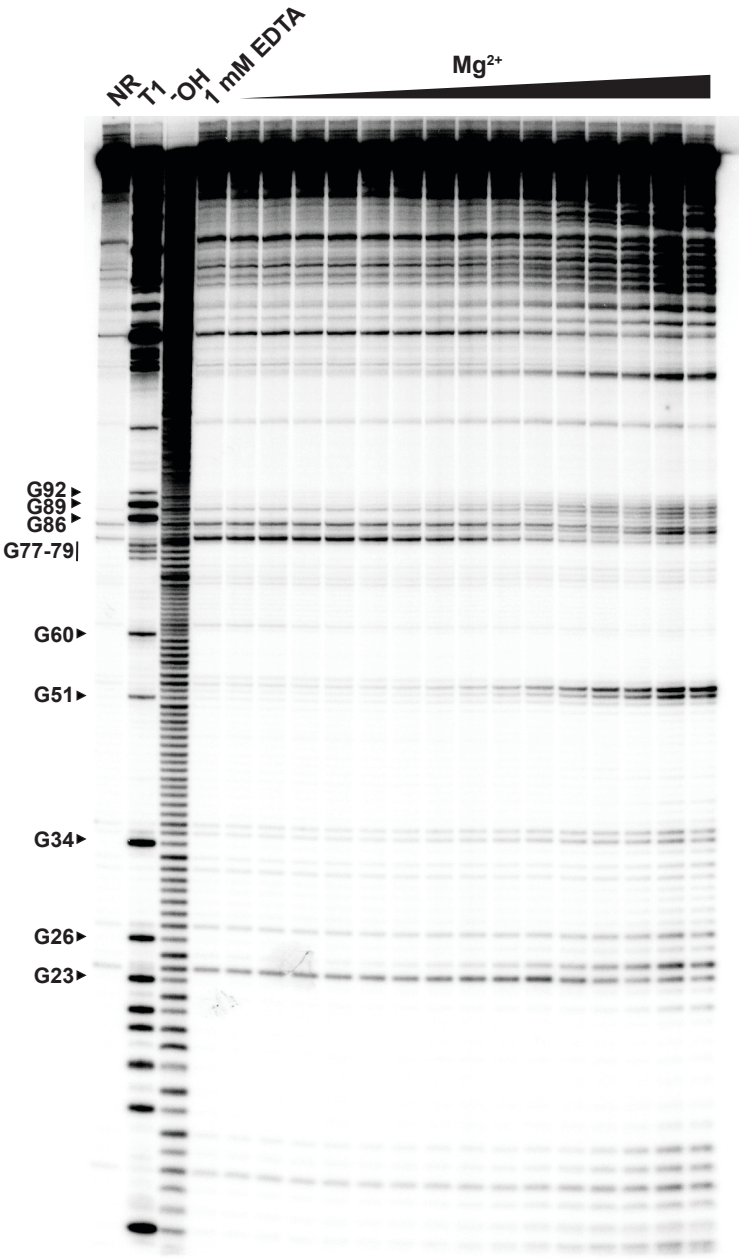

Figure 3. In-line probing of *B. halodurans* OLE RNA at increasing Mg^{2+} concentrations. PAGE autoradiograph of 5' ^{32}P-labeled OLE RNA subjected to in-line probing conditions with 1 mM EDTA or $MgCl_2$ (10 μM to 32 mM in quarter-log increments). NR, T1, and $^{-}$OH lanes were subjected to no reaction, digestion with RNase T1 (cleavage after G residues), and incomplete digestion under alkaline conditions (cleaves after each nucleotide), respectively. The RNA could be mapped up to nucleotide G92, after

which point the individual bands from the alkaline digest became too compressed to confidently distinguish individual nucleotides. Increased scission at higher Mg^{2+} concentrations is a feature of in-line probing and not necessarily an indication of structural change (Regulski and Breaker. 2008). I performed the experiment for this figure.

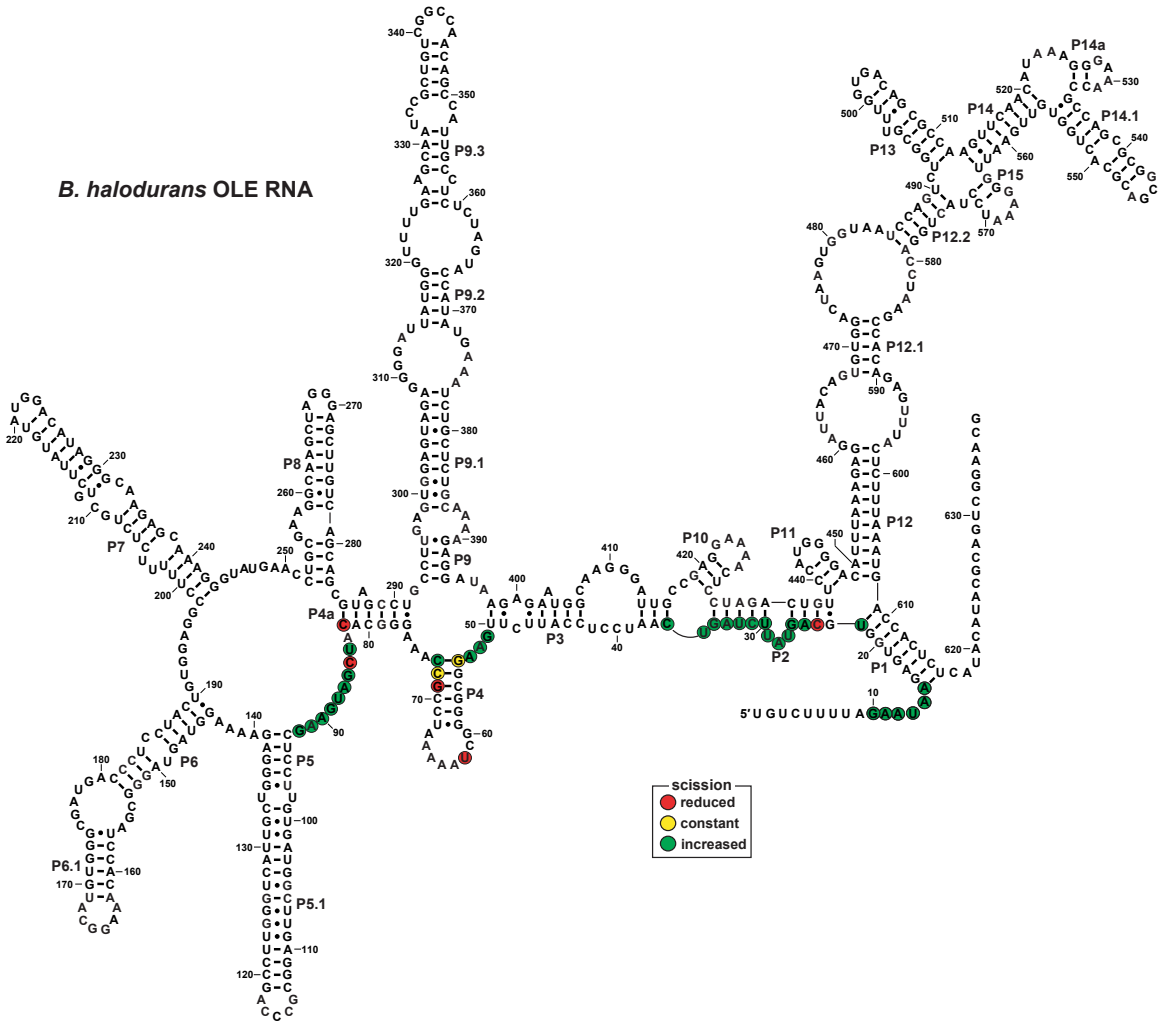

Figure 4. A primary and secondary structure model of *B. halodurans* OLE RNA highlighting nucleotides that modulate with increasing Mg^{2+}. The nucleotide sequence of the construct used in **Fig. 3** denoting increases and decreases in scission. Nucleotides beyond G92 could not be accurately mapped.

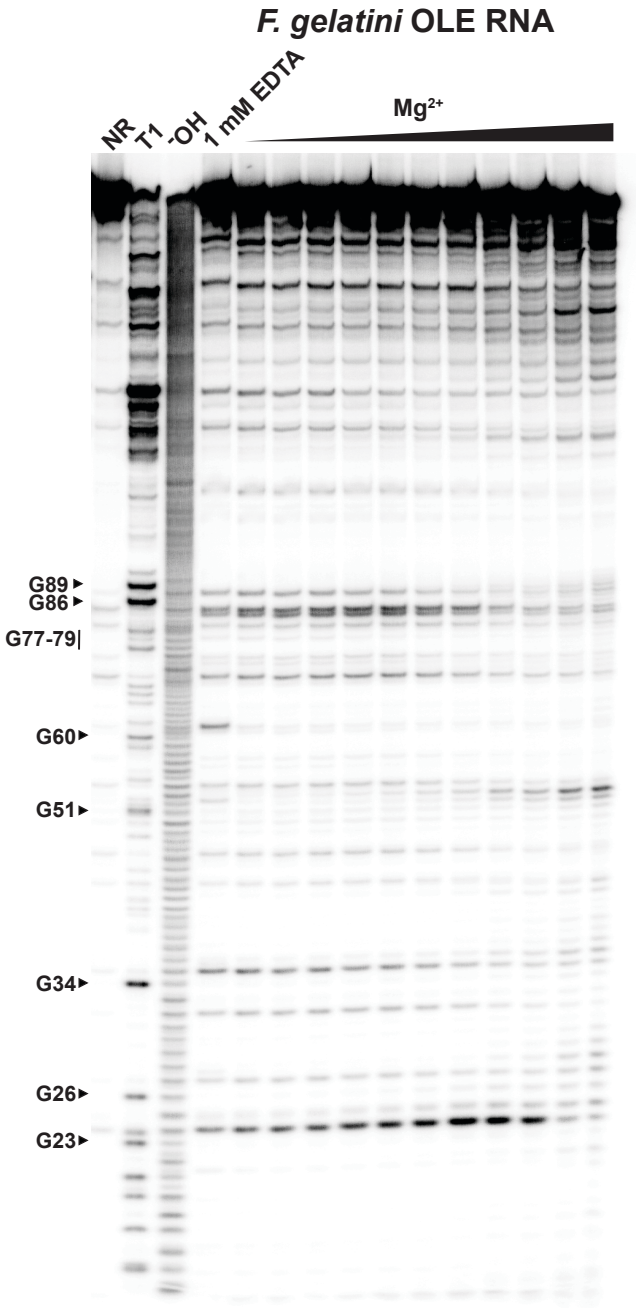

Figure 5. In-line probing of *Fictobacillus gelatini* OLE RNA with increasing Mg^{2+}. PAGE autoradiograph of 5' ^{32}P-labeled OLE RNA subjected to in-line probing conditions with 1 mM EDTA or $MgCl_2$ (100 μM to 32 mM in quarter-log increments). All annotations are as described in **Fig. 3**. Chrishan Fernando performed the experiment for this figure.

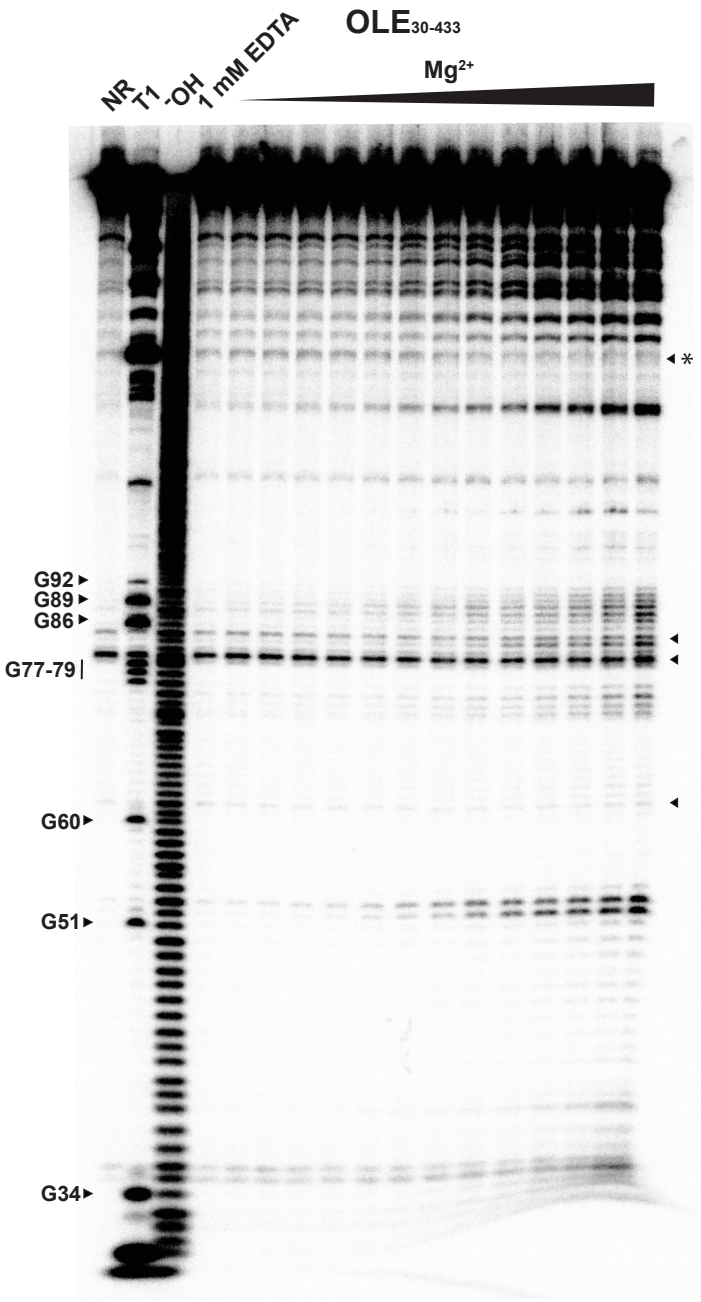

Figure 6. In-line probing of a truncated OLE RNA with increasing Mg^{2+}. PAGE autoradiograph of 5' ^{32}P-labeled OLE RNA_{30-433} subjected to in-line probing conditions with 1 mM EDTA or $MgCl_2$ (100 μM to 100 mM in quarter-log increments). All annotations are as described in **Fig. 3**. Arrows on the righthand side of the gel show bands

that modulate in full-length OLE RNA. The asterisk denotes a band that continues to modulate in $OLE_{30\text{-}433}$. I performed the experiment for this figure.

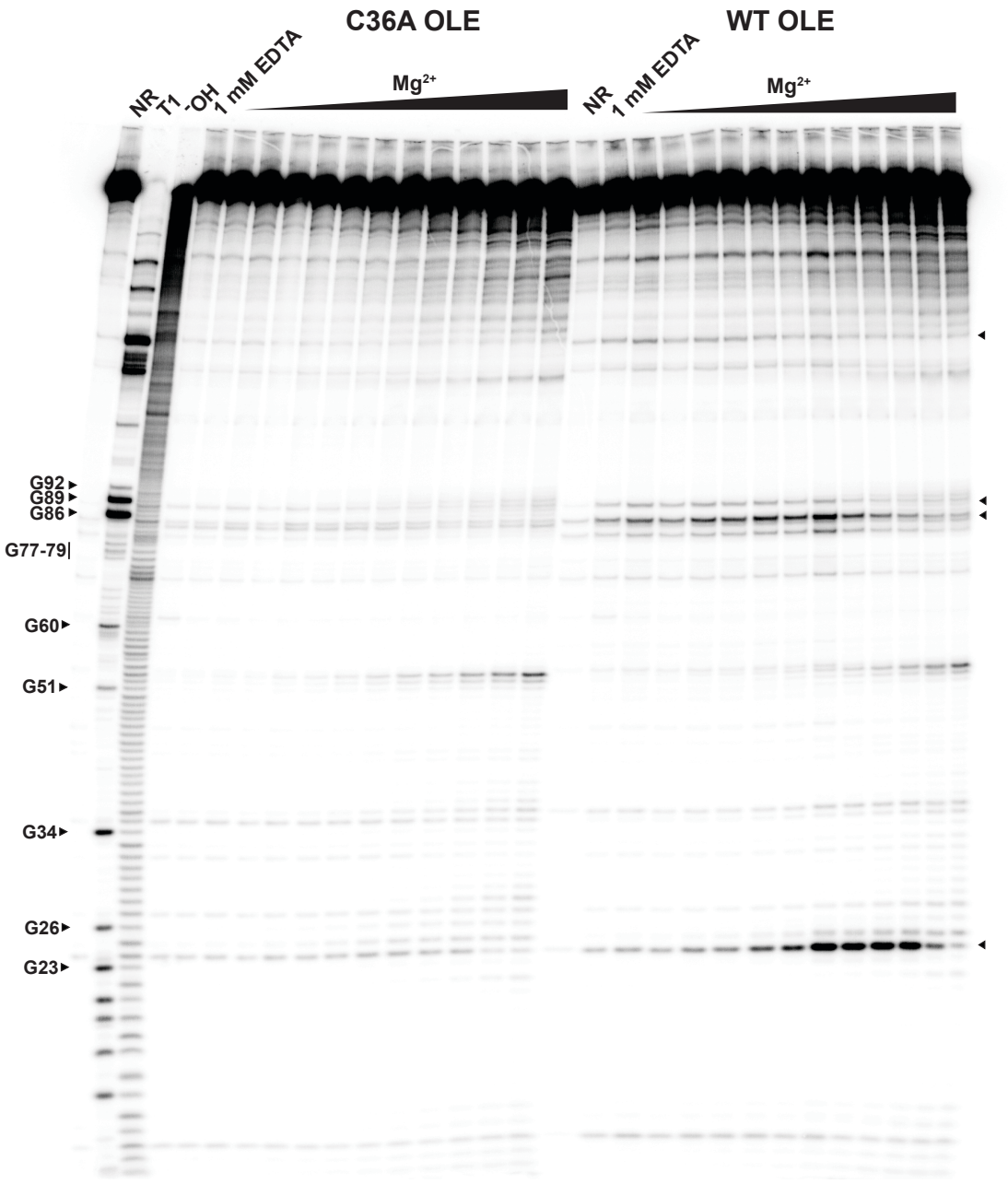

Figure 7. In-line probing of C36A and wild type (WT) OLE RNA with increasing Mg^{2+}. PAGE autoradiograph of 5' ^{32}P-labeled C36A and WT OLE RNA subjected to in-line probing conditions with 1 mM EDTA or $MgCl_2$ (56 μM to 32 mM in quarter-log increments). All annotations are as described in **Fig. 3**. Arrows on the righthand side

indicate where modulation differs between C36A and WT OLE RNA. Chrishan Fernando performed the experiment for this figure.

Chapter Four

Remarkably structured single-stranded DNA motifs within the IS605 transposon superfamily

Author Contributions: D.L.W. performed data analysis. D.L.W. wrote the chapter.

Summary

Most structured single-stranded DNA (ssDNA) motifs are compact and simplistic, yet in circumstances where DNA exists outside the double helix for prolonged periods more complex ssDNA structures can evolve. Recently three ssDNA motifs associated with the IS605 transposon superfamily, HEARO (HNH Endonuclease Associated RNA and ORF), IS605-*orfB*-I, and IS605-*orfB*-II, were found to be among the most structurally complex ssDNA motifs documented. Although members of the IS605 superfamily are known to utilize DNA hairpins in transposition, the extent of secondary structure in these new motifs suggests that they serve additional biochemical functions. Of these motifs HEARO is the most widespread and structurally complex. In this work I analyze the distribution of HEARO and its frequency per host genome. I compare its ssDNA structure to other members of the IS605 superfamily, including those containing IS605-*orfB*-I and IS605-*orfB*-II motifs and analyze the domain architecture of its associated endonuclease in comparison to closely related proteins. For the IS605-*orfB*-I and IS605-*orfB*-II motifs I quantify co-occurrence with the two IS605-associated proteins, TnpA and TnpB. From that analysis I show that IS605-*orfB*-I motifs are typically associated with non-autonomous transposons, while IS605-*orfB*-II motifs are predominately associated with autonomous transposons. Thru careful study of three convergent motifs, this work offers new insight into the evolution of structurally complex ssDNAs.

Introduction

While structured RNAs are common in biology, DNA rarely forms structures other than the iconic double helix and consequently structured single-stranded DNAs (ssDNAs) are rare in biology. However, there are a handful of situations in which DNA naturally enters a single-stranded state. These situations are most common in mobile genetic elements (MGEs). MGEs such as ssDNA viruses, rolling circle plasmids, integrons, and some transposons exist as ssDNA in at least one stage of their replicative cycles (Bikard et al. 2010). Often the most complex ssDNA structures are found in ssDNA phages where the pressure and opportunity to evolve such structures is highest (Lambert et al. 1986, Das et al. 2010, Bikard et al. 2010). Additionally, non-mobile, structured DNA motifs can occur in the form of G-quarduplexes and msDNA (Bartas et al. 2019, Simon et al. 2019).

In this chapter I will discuss three of the most highly structured naturally occurring ssDNAs, HEARO (HNH Endonuclease Associated RNA and ORF), IS605-*orfB*-I, and IS605-*orfB*-II. When first published in 2009 HEARO was predicted to function as an RNA in part because no single-stranded DNA (ssDNA) of comparable complexity had previously been reported (Weinberg et al. 2009) However, I discovered that the HNH endonuclease associated with the HEARO motif is a homolog of the IS605 TnpB protein. Unlike HEARO, the two smaller and more narrowly distributed IS605-associated motifs, IS605-*orfB*-I and IS605-*orfB*-II, were immediately recognized to function as ssDNA motifs through a strong co-occurrence with *tnpB* genes associated with IS605 insertion sequences (Weinberg et al. 2017). The transposon family containing IS605, IS200, IS1341, and IS607 insertion sequences is notable for its use of ssDNA intermediates in a form of transposition known as peel-and-paste. Here I will refer to peel-and-paste transposons collectively as the IS605 superfamily.

Two proteins are associated with the IS605 superfamily, TnpA and TnpB. The transposase, TnpA, is part of the HUH endonuclease family, a class of proteins adept at cleaving and rejoining ssDNA (Chandler et al. 2013). TnpA forms a homodimer that recognizes imperfect ssDNA hairpins at the 5' and 3' ends of the transposon, with one monomer binding to each hairpin and exciting the ssDNA through nucleophilic attack by a tyrosine residue (Ronning et al. 2005, Ton-Hoang et al. 2005, Lee et al. 2006, Guynet et al. 2008, Hickman et al. 2010). This forms a 5' covalent bond with the ssDNA molecule and induces a *trans* to *cis* conformational shift in the tyrosine-containing α-helix (He et al. 2013). This conformational shift positions TnpA to circularize the ssDNA by transferring the 5' phosphate to the 3'-OH (He et al. 2013). TnpA transports the circularized transposon to a new insertion location that matches a specific tetra- or pentanucleotide target sequence and integrates the ssDNA into the genome.

The role of TnpB in this process remains a mystery. In fact, TnpB is not required for transposition (Kersulyte et al. 2002, Pasternak et al. 2010) and has been shown to inhibit transposition of ISDra2 in *Deinococcus radiodurans* (Pasternak et al. 2013). Therefore, IS605 superfamily transposons that contain only TnpB, often referred to as IS1341, are predicted to be non-autonomous MGEs that rely on TnpA proteins encoded elsewhere in the genome for mobility. From this it can be predicted that HEARO transposons, which typically contain either TnpB or no proteins-coding genes at all, likely function as non-autonomous MGEs that rely on outside transposases. This is consistent with the finding that HEARO-associated TnpB is enriched in genomes containing Y1 serine and tyrosine transposases (Kapitonov et al. 2016).

In this chapter I expand on what is known about HEARO, IS605-*orfB*-I, and IS605-*orfB*-II through bioinformatic analysis of the three motifs. For HEARO, I look at the distribution of HEARO elements by kingdom and phyla and the distribution of HEARO elements per genome. For IS605-*orfB*-I and IS605-*orfB*-II, I look at the frequency with which the motifs associate with *tnpA* and *tnpB* genes, showing that both are sometimes but not always part of autonomous transposons. Finally, all three motifs are compared to other IS605 elements in regard to nucleic acid structure and TnpB domain composition.

Results and Discussion

Distribution of HEARO

HEARO is widespread in prokaryotes, residing in 426 species across both the bacterial and archaeal kingdoms. Within bacteria, HEARO is present in eleven phyla, both Gram-negative and Gram-positive, with particularly high representation in Firmicutes and Cyanobacteria (**Fig. 1**). HEARO is found in 20 archaeal species and strains, all within the phyla Euryarchaeota and the class Methanomicrobia. The majority of these archaeal examples are in various strains of *Methanosarcina*. The archaeal HEARO elements are most closely related to elements found in Cyanobacteria and Firmicutes, suggesting a crossover event from one of those phyla.

Multiple copies of HEARO frequently occur per genome

Different classes of transposons occur at different frequencies within host genomes. Transposons that carry beneficial genes such as antibiotic resistance cassettes generally exist only as a single copy within the host, whereas purely selfish transposable elements

often occur multiple times per genome. The limiting factors for selfish transposon propagation are host defense systems (such as CRISPR) and the fitness cost inflicted on the host. A transposable element that kills its host by over propagating also kills itself.

IS605 superfamily transposons are selfish genetic elements, conferring no known benefits to their hosts. For this reason, I sought to systematically explore the fitness cost of HEARO transposons by quantifying the number of copies per host genome. My analysis showed a high level of variability in copy frequency. Even though about half of all HEARO-containing species have only one copy, in some strains the transposon has propagated to extremely high numbers, with one species containing 143 HEARO transposons (**Fig. 2**). With >75% of species containing ≤4 copies and >90% of species containing ≤11 copies, it is clear that although host defenses and the fitness cost keep copy number low in most species, HEARO transposons efficiently propagate within genomes despite being non-autonomous MGEs.

Distribution of TnpA and TnpB proteins encoded in HEARO, IS605-*orfB*-I, and IS605-*orfB*-II

HEARO transposons typically encode only TnpB, but a *tnpA* gene is found near the element in rare instances and is common in genomes that contain HEARO (Kapitonov et al. 2016). In the rare instances where TnpA is present, the element is presumably transformed from a non-autonomous to an autonomous transposon. HEARO is also frequently found at genetic loci where no *tnpA* or *tnpB* genes are found (Weinberg et al. 2009). The presence of HEARO without accompanying protein coding genes raises interesting and yet unaddressed questions about the potential for ultra-simplistic non-

autonomous MGEs. If ORF-free HEARO elements can propagate throughout a genome, could miniature ORF-free IS605 variants also exist? Due to the simplicity of such motifs, which would presumably contain only two imperfect palindromes and a tetra- or pentanucleotide target sequence, it would be incredibly difficult to accurately detect such variants bioinformatically.

I was curious to know if ORF-free transposons were common for IS605-*orfB*-I and IS605-*orfB*-II and what ORFs were associated with the two motifs. It was immediately apparent that unlike HEARO, IS605-*orfB*-I and IS605-*orfB*-II are associated with either TnpA, TnpB, or both in the vast majority of cases. Cases without either ORF were not examined closely because of their rarity and because some ORF-free transposons would be expected simply as a byproduct of inactive transposon decay. Among the elements with ORFs differences observed between the two motifs, the most common array for IS605-*orfB*-I was for only the *tnpB* gene to be present, though in slightly over a third of instances *tnpA* was also present (**Table 1**). For IS605-*orfB*-II this trend was reversed, with most elements containing both *tnpA* and *tnpB* and a small fraction (~13%) containing only *tnpB* (**Table 1**). That suggests that most IS605-*orfB*-I elements are part of non-autonomous transposons, and most IS605-*orfB*-II elements are part of autonomous ones.

Comparison of ssDNA Structure and Gene Content between HEARO, IS605-*orfB*-I, IS605-*orfB*-II, and other IS605 variants

While there is some variation from one class to another, the canonical IS605 transposon contains two imperfect hairpins at the 5’ and 3’ ends of the insertion sequence. These hairpins contain one bulged nucleotide on the righthand pairing element that allows

TnpA to distinguish between the top and bottom strands (Barabas et al. 2008). In some instances, such as IS608 from *Helicobacter pylori*, a second hairpin is present at the 5' end (He et al. 2013) (**Fig. 3**). By contrast, HEARO averages over 300 nucleotides in length and contains three multistem junctions and one pseudoknot (a long-distance base pairing interaction) (Weinberg et al. 2009) (**Fig. 3**). Though less complex than HEARO, each of the IS605-*orfB* motifs contains one pseudoknot and at least five additional stem regions (Weinberg et al. 2017) (**Fig. 3**). Whereas HEARO contains a sophisticated structure at the 5' end of the transposon, IS605-*orfB*-I and IS605-*orfB*-II have highly structured 3' termini (**Fig. 3**). Why structure at one end versus the other might be beneficial is unclear at this time. Additional bioinformatics analyses are required to address if HEARO, IS605-*orfB*-I, and IS605-*orfB*-II are unique or if there is a continuum of structures ranging from simple imperfect hairpins to numerous 5' and 3' variations.

Comparison of Protein Domains between HEARO, IS605-*orfB*-I, IS605-*orfB*-II, other IS605 variants, and Cas9

In addition to homology with IS605 TnpB proteins, the HEARO-associated ORF is a progenitor of Cas9 (Kapitonov et al. 2016). A study searching for Cas9 homologs revealed that the HEARO-associated TnpB (renamed IscB) shares multiple domains with the protein (Kapitonov et al. 2016). Both IscB and Cas9 have an arginine rich helix, an HNH endonuclease domain, and three RuvC domains (**Fig. 4**). In Cas9 the arginine bridging helix interacts with target DNA and the HNH and RuvC domains each nick one strand of target DNA (Jinek et al. 2014, Nishimasu et al. 2014). Cas9 also contains domains that are not present in IscB, REC domains that bind the guide RNA and a PAM-interacting

(PI) domain that confers PAM specificity (Jinek et al. 2014, Nishimasu et al. 2014). Of the TnpB homologs IscB is the most direct progenitor of Cas9, as they share an HNH endonuclease domain (Kapitonov et al. 2016) (**Fig. 4**). The only endonuclease domains in other TnpB homologs are those of RuvC, meaning that while canonical IS605 TnpB can nick a single strand of DNA, IscB can make a double-stranded break. Why the HEARO-associated IscB and not other TnpB homologs would require the capacity to make a double-stranded break is unclear, but will hopefully be elucidated once the function of TnpB is understood. Additionally, IscB has lost the helix-turn-helix (HTH) and zinc finger (CCCC) domains present in IS605 TnpB. The TnpB homologs associated with IS605-*orfB*-I and IS605-*orfB*-II are more closely related to canonical TnpB than to IscB, though neither homolog contains an N-terminal HTH domain and only the TnpB associated with IS605-*orfB*-II possesses a CCCC domain (**Fig. 4**). Once again, without a prescribed function for TnpB, understanding the significance of these differences is difficult.

Conclusions

Natural selection does not preserve complex nucleic acid structures except when they serve an important purpose, even if only for the propagation of a selfish element. In the case of HEARO, IS605-*orfB*-I, and IS605-*orfB*-II, this function unknown. One unprecedented hypothesis considered was that HEARO and perhaps also IS605-*orfB*-I and IS605-*orfB*-II could possess a catalytic function. Although numerous examples of catalytic DNAs have been evolved *in vitro*, no naturally occurring deoxyribozymes have been discovered (Silverman. 2016). If naturally occurring deoxyribozymes do exist they would presumably reside in the DNA equivalents of common ribozyme locations (Weinberg et

al. 2019), such as ssDNA transposons and phages. This, in addition to the fact that HEARO is as large and structurally complex as a Group I self-splicing intron (Harris and Breaker. 2018), made it worth asking if, though unprecedented, HEARO and possibly IS605-*orfB*-I and IS605-*orfB*-II might function as deoxyribozymes. However, initial experiments to determine if HEARO could take the place of TnpA and excise the ssDNA transposon from the genome have produced only negative findings (unpublished results).

Regardless, the size and complexity of HEARO, IS605-*orfB*-I, and IS605-*orfB*-II suggests that they perform a sophisticated biochemical function. Determining that function might offer insight into the function of TnpB. Though, since they reside at different ends of the transposon it is not even clear whether all three motifs perform the same function. Farther research, particularly involving biochemical and genetic approaches, is needed before any conclusions can be drawn on the function of these complex and unusual structures. The discovery of these motifs expands the small field of structured ssDNAs and raises questions about how widespread such elements are within the IS605 superfamily.

The connection between IscB and Cas9 raises new questions. Is the link between Cas9 and a protein associated with one of nature's most complex ssDNA structures coincidence, or was IscB uniquely positioned to evolve into a precision DNA targeting system because of its coevolution with HEARO? Until the function of HEARO is determined such questions remain impossible to answer, giving all the more reason to study these unusual and fascinating ssDNAs.

Methods

Databases

Additional examples of the HEARO, IS605-*orfB*-I, and IS605-*orfB*-II motifs were found with the comparative sequence algorithms CMfinder (Yao et al. 2006) and Infernal 1.1 (Nawrocki and Eddy. 2013) as described in Weinberg et al. 2017. The databases consisted of Refseq version 76 for HEARO and Refseq version 80 for IS605-*orfB*-I and IS605-*orfB*-II as well as a previously described collection of microbial environmental sequences (Weinberg et al. 2017).

Phylogenetic distribution was determined by pulling taxonomical data for 10,780 HEARO motifs and sorting by phyla. Examples from environmental databases were removed. Frequency of HEARO elements per genome was determined using that taxonomical data.

Determination of co-occurrence with TnpA and TnpB

Co-occurance of IS605-*orfB*-I and IS605-*orfB*-II with TnpA and TnpB was determined by individually examining the genetic context of the two genes upstream of each motif in NCBI Sequence Viewer. ORFs without clear annotations were run on blastp to determine if they are homologs of TnpA or TnpB. Environmental reads in which the entire transposon was not present or where completeness was ambiguous were removed from the analysis.

References

Barabas, O., Ronning, D. R., Guynet, C., Hickman, A. B., Ton-Hoang, B., Chandler, M., and Dyda, F. (2008) Mechanism of IS200/IS605 family DNA transposases: activation and transposon-directed target site selection. *Cell* **132**, 208-220

Bartas, M., Cutova, M., Brazda, V., Kaura, P., Stastny, J., Kolomaznik, J., Coufal, J., Goswami, P., Cerven, J., and Pecinka, P. (2019) The Presence and Localization of G-Quadruplex Forming Sequences in the Domain of Bacteria. *Molecules* **24**

Bikard, D., Loot, C., Baharoglu, Z., and Mazel, D. (2010) Folded DNA in action: hairpin formation and biological functions in prokaryotes. *Microbiol Mol Biol Rev* **74**, 570-588

Chandler, M., de la Cruz, F., Dyda, F., Hickman, A. B., Moncalian, G., and Ton-Hoang, B. (2013) Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nat Rev Microbiol* **11**, 525-538

Das, B., Bischerour, J., Val, M. E., and Barre, F. X. (2010) Molecular keys of the tropism of integration of the cholera toxin phage. *Proc Natl Acad Sci U S A* **107**, 4377-4382

Guynet, C., Hickman, A. B., Barabas, O., Dyda, F., Chandler, M., and Ton-Hoang, B. (2008) In vitro reconstitution of a single-stranded transposition mechanism of IS608. *Mol Cell* **29**, 302-312

Harris, K. A., and Breaker, R. R. (2018) Large Noncoding RNAs in Bacteria. *Microbiol Spectr* **6**

He, S., Guynet, C., Siguier, P., Hickman, A. B., Dyda, F., Chandler, M., and Ton-Hoang, B. (2013) IS200/IS605 family single-strand transposition: mechanism of IS608 strand transfer. *Nucleic Acids Res* **41**, 3302-3313

Hickman, A. B., James, J. A., Barabas, O., Pasternak, C., Ton-Hoang, B., Chandler, M., Sommer, S., and Dyda, F. (2010) DNA recognition and the precleavage state during single-stranded DNA transposition in D. radiodurans. *EMBO J* **29**, 3840-3852

Jinek, M., Jiang, F., Taylor, D. W., Sternberg, S. H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., Kaplan, M., Iavarone, A. T., Charpentier, E., Nogales, E., and Doudna, J. A. (2014) Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997

Kapitonov, V. V., Makarova, K. S., and Koonin, E. V. (2015) ISC, a Novel Group of Bacterial and Archaeal DNA Transposons That Encode Cas9 Homologs. *J Bacteriol* **198**, 797-807

Kersulyte, D., Velapatino, B., Dailide, G., Mukhopadhyay, A. K., Ito, Y., Cahuayme, L., Parkinson, A. J., Gilman, R. H., and Berg, D. E. (2002) Transposable element ISHp608 of Helicobacter pylori: nonrandom geographic distribution, functional organization, and insertion specificity. *J Bacteriol* **184**, 992-1002

Lambert, P. F., Waring, D. A., Wells, R. D., and Reznikoff, W. S. (1986) DNA requirements at the bacteriophage G4 origin of complementary-strand DNA synthesis. *J Virol* **58**, 450-458

Lee, H. H., Yoon, J. Y., Kim, H. S., Kang, J. Y., Kim, K. H., Kim, D. J., Ha, J. Y., Mikami, B., Yoon, H. J., and Suh, S. W. (2006) Crystal structure of a metal ion-bound IS200 transposase. *J Biol Chem* **281**, 4261-4266

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

Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935-949

Pasternak, C., Ton-Hoang, B., Coste, G., Bailone, A., Chandler, M., and Sommer, S. (2010) Irradiation-induced Deinococcus radiodurans genome fragmentation triggers transposition of a single resident insertion sequence. *PLoS Genet* **6**, e1000799

Pasternak, C., Dulermo, R., Ton-Hoang, B., Debuchy, R., Siguier, P., Coste, G., Chandler, M., and Sommer, S. (2013) ISDra2 transposition in Deinococcus radiodurans is downregulated by TnpB. *Mol Microbiol* **88**, 443-455

Ronning, D. R., Guynet, C., Ton-Hoang, B., Perez, Z. N., Ghirlando, R., Chandler, M., and Dyda, F. (2005) Active site sharing and subterminal hairpin recognition in a new class of DNA transposases. *Mol Cell* **20**, 143-154

Silverman, S. K. (2016) Catalytic DNA: Scope, Applications, and Biochemistry of Deoxyribozymes. *Trends Biochem Sci* **41**, 595-609

Simon, A. J., Ellington, A. D., and Finkelstein, I. J. (2019) Retrons and their applications in genome engineering. *Nucleic Acids Res* **47**, 11007-11019

Ton-Hoang, B., Guynet, C., Ronning, D. R., Cointin-Marty, B., Dyda, F., and Chandler, M. (2005) Transposition of ISHp608, member of an unusual family of bacterial insertion sequences. *EMBO J* **24**, 3325-3338

Weinberg, C. E., Weinberg, Z., and Hammann, C. (2019) Novel ribozymes: discovery, catalytic mechanisms, and the quest to understand biological function. *Nucleic Acids Res* **47**, 9480-9494

Weinberg, Z., Perreault, J., Meyer, M. M., and Breaker, R. R. (2009) Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. *Nature* **462**, 656-659

Weinberg, Z., Lünse, C. E.,Corbino, K. A., Ames, T. D., Nelson, J. W., Roth, A., Perkins, K. R., Sherlock, M. E., and 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

Yao, Z., Weinberg, Z., and Ruzzo, W. L. (2006) CMfinder--a covariance model based RNA motif finding algorithm. *Bioinformatics* **22**, 445-452

Figures and Tables

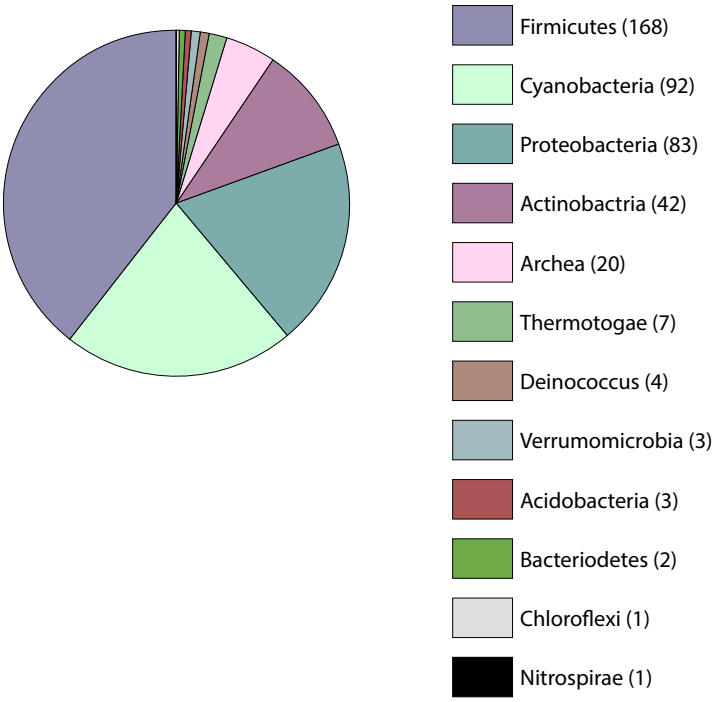

Figure 1. Distribution of HEARO in Bacteria and Archaea. A chart showing the number of species containing at least one HEARO element across bacterial phyla and within Archaea. All archaeal species are within the phyla Euryarchaeota.

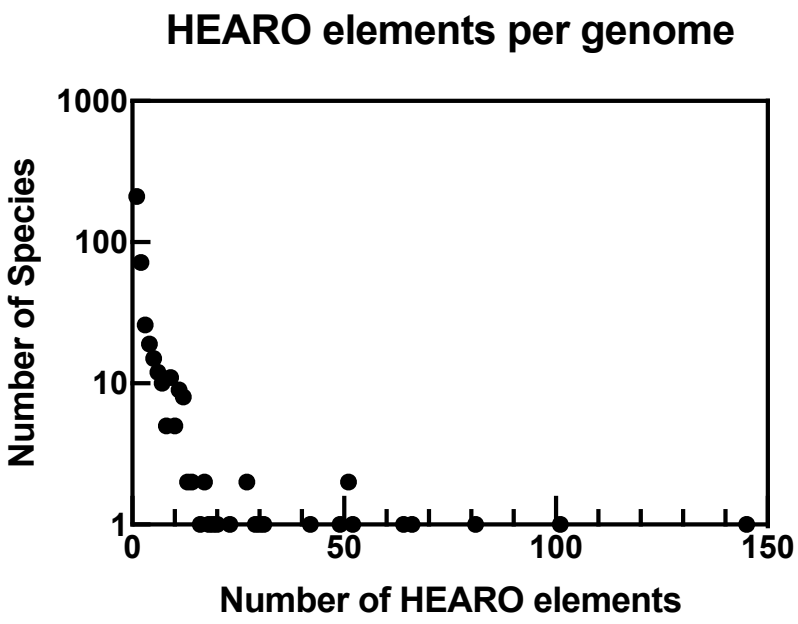

Figure 2. Frequency of HEARO elements per host species. A graph plotting the number of HEARO elements in a species by the number of species that contain that number of HEARO elements.

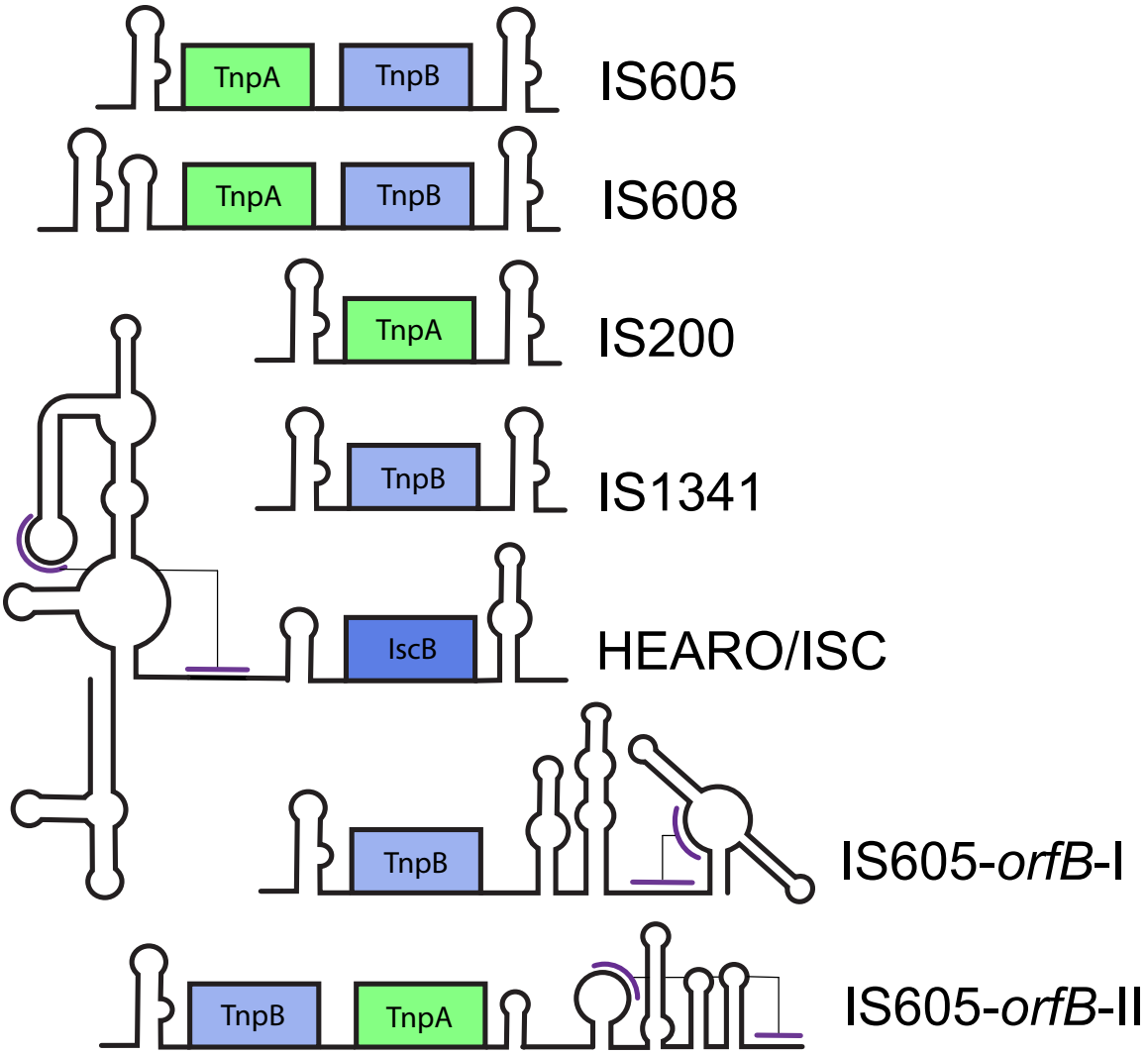

Figure 3. Comparison of the most common ssDNA structure and gene arrangements for IS605 superfamily members. The most common arrangements of several IS605 variants including ssDNA structure and arrangement of TnpA and TnpB genes. HEARO transposons have also been classified as ISC (Kapitonov et al. 2016). Long-range pseudoknot interactions are shown in purple. Elements shown with identical 5' or 3' structures in model can have minor variations in secondary structure between specific examples.

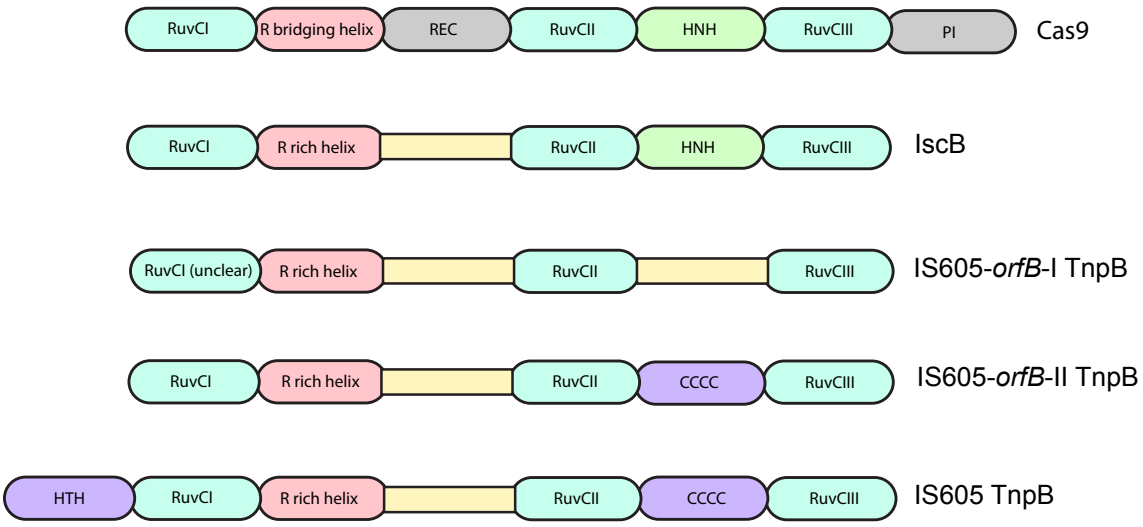

Figure 4. Comparison of domain architecture of Cas9 and various TnpB proteins. A diagram showing the domain architecture of Cas9, IscB, and the TnpB homologs associated with IS605-*orfB*-I and IS605-*orfB*-II, and IS605. Abbreviations are as follows: RuvC endonuclease domains (RuvCI, RuvCII, and RuvCIII), arginine-rich bridging helix (R bridging helix), recognition domains (REC), HNH endonuclease domain (HNH), PAM-interacting domain (PI), helix-turn-helix domain (HTH), and zinc finger domain (CCCC).

Table 1. Co-occurrence of IS605-*orfB*-I and IS605-*orfB*-II with TnpA and TnpB

	TnpA(-)/TnpB(+)	TnpA(+)/TnpB(-)	TnpA(+)/TnpB(+)
IS605-*orfB*-I	63	1	35
IS605-*orfB*-II	6	0	41