RNA structures regulating ribosomal protein biosynthesis

Author: Kaila Deiorio-Haggar

Boston College

Graduate School of Arts and Sciences

Department of Biology

RNA STRUCTURES REGULATING RIBOSOMAL PROTEIN BIOSYNTHESIS

[a thesis]

by

KAILA DEIORIO-HAGGAR

Submitted in partial fulfillment of the requirements

for the degree of

Master of Science

August 2015

RNA STRUCTURES REGULATING RIBOSOMAL PROTEIN BIOSYNTHESIS

By: Kaila Deiorio-Haggar

Advisor: Dr. Michelle M. Meyer

Abstract

Most commonly known for being the blueprint for proteins, RNA also plays vital roles in gene regulation. Non-coding RNAs, functional RNA molecules that are not translated into proteins, are potential regulatory agents in bacteria. Ribosomal autogenous regulatory elements are short transcribed sequences between the promoter and a protein coding region that regulate expression of their associated gene(s), though they are not themselves translated. These sequences form RNA secondary structures that can regulate at either the transcriptional or translational level. These riboregulators have been well characterized in gram-negative bacteria such as *Escherichia coli*, but in gram-positive bacteria far less is known regarding how r-proteins are regulated.

My main goal has been to find riboregulators of r-protein synthesis in Bacilli and determine their consensus structures and phylogenetic distributions. I have utilized the RNA homology search program Infernal, coupled with our high-capacity genomic context visualization tool, to identify homologues of ribosomal-protein autogenous regulatory RNAs found in Bacilli. The alignments produced from this work determine consensus secondary structures and phylogenetic distribution of these regulator RNAs that provide new insight into the structure and function of these RNAs.

Chapter 1. Background and Significance

RNA, Ribosomes, and Ribosomal Proteins

In bacteria, ribosomal proteins are typically universally distributed and well conserved (Yutin et al., 2012). In addition, the over 50 genes encoding ribosomal proteins occur in long multi-gene operons whose structure is largely, but not completely, conserved (Coenye and Vandamme, 2006). Despite the universal nature of these proteins, their regulation does not appear to be conserved across eubacteria. A great deal of regulation occurs at the transcriptional level, and due to the abundance of rRNA and r-proteins in any given cell, additional regulation occurs in the level of translation initiation for r-proteins, mostly by ribosomal autogeous regulatory elements.

Ribosomal Autogenous Regulatory Elements

There are a wide variety of different bacterial regulatory RNAs ranging from riboswitches that require complex secondary- and tertiary- structure motifs for function (Serganov and Patel, 2012) to sRNAs that act predominantly through base-pairing interactions (Desnoyers et al., 2012; Storz et al., 2011). Some of the first regulatory RNAs to be described are those that autogenously regulate ribosomal protein biosynthesis in *Escherichia coli* (Nomura et al., 1980). These regulatory RNAs are non-coding RNAs (ncRNAs) that typically occur within 5'-untranslated or intergenic regions of transcripts encoding ribosomal proteins. When transcribed, these RNA sequences form unique sets of stem-loop structures. These secondary structures interact with a specific ribosomal protein binding partner to regulate an entire ribosomal protein operon (Zengel and Lindahl, 1994) (Figure 1). In many cases, the mRNA structure responsible for regulation is a mimic of the ribosomal RNA (rRNA) binding site for the same protein (Nomura et al., 1980).

The mechanism of gene regulation can be either transcriptional (Zengel and Lindahl, 1990) or translational (Guillier et al., 2002; Marzi et al., 2007; Schlax and Worhunsky, 2003), thus allowing these mRNA structures to act as a means of feedback inhibition, halting the production of unnecessary amounts of protein product. For example, *E. coli* uses both transcription termination via L4 binding for autogenous regulation (Zengel and Lindahl, 1990), as well as translational coupling via L20 for inhibition of its own operon (Guillier et al., 2002). These rRNA-binding structures were among some of the first RNA-based regulation methods studied, and because they were discovered before the genomic era, multiple sequence alignments are available for only a few of them. To address this lack, I have performed a comparative genomics study of ribosomal protein regulator RNAs in Bacilli.

Ribosomal Autogenous Regulatory Elements in *Escherichia coli*

The first ribosomal autogenous regulatory elements discovered were proposed and experimentally validated in *E. coli* during the 1980's and 90's. However, the first comparative genomics study of these regulators has only been recently performed (Fu et al., 2013). To date, over 10 ncRNAs regulating more than half of ribosomal protein genes have been described in *E. coli*. Most of these RNAs are narrowly distributed to Gammaproteobacteria (Fu et al., 2013), and their distribution pattern is consistent with vertical inheritance within Gammaproteobacteria (Gao et al., 2009). Of the *E. coli* ribosomal protein regulatory RNAs, only two (interacting with ribosomal proteins L1 and L10) (Fu et al., 2013; Iben and Draper, 2008) are widely present in Firmicutes (>85% of all sequenced Firmicutes). A third RNA structure (interacting with ribosomal protein S2) has been identified in >50% of sequenced Firmicutes and appears to be distributed equally among Bacilli and Clostridia (Fu et al., 2013; Meyer et al., 2009).

The mechanism of action and structure of many of the *E. coli* ribosomal regulators is now well understood (Aseev and Boni, 2011; Zengel and Lindahl, 1994). However, most of these regulatory elements are not annotated in standard genomic databases (Pruitt et al., 2003), and as of 2012, the RNA Families Database (Rfam) only contained alignments for 2 of the 12 known and studied ribosomal regulatory elements [Rfam: RF00140, RF00114]. Before the work performed by our laboratory (Fu et al., 2013), phylogenetic studies performed on these regulatory RNAs were not systematic and the knowledge gained was not preserved in databases used for genomic annotation (Allen et al., 1999; Aseev et al., 2009; Boni et al., 2001; Guillier et al., 2005; Iben and Draper, 2008). As a result, these important RNA molecules risk being overlooked as genomic annotation becomes more automated. Our recent work has combined the experimental data derived from *E. coli* sequences with homologous sequences identified through comparative genomic studies to compile alignments that describe RNAs (Fu et al., 2013). However, a number of RNA regulators are known in *E. coli* and not *Bacillus subtilis*.

Ribosomal Autogenous Regulatory Elements in *Bacillus subtilis*

Progress toward understanding the regulation of ribosomal protein biosynthesis in gram-positive bacteria, including model organisms such as *Bacillus subtilis*, has been much more limited than that of *E. coli*. While the gene order for most ribosomal proteins is conserved between *E. coli* and *B. subtilis* (Nikolaichik and Donachie, 2000), there are significant changes to the *spc*, alpha, and S10 *E. coli* operon structures (regulated by S8, S4, and L4, respectively). In *B. subtilis*, most of the genes contained within these operons are co-transcribed by a single transcript (Li et al., 1997), removing the need for individual regulation. Thus, not all r-protein regulatory elements present in *E. coli* will necessarily be present in *B. subtilis* as well.

RNA structures interacting with ribosomal proteins S4, S15, and L20 have been identified and experimentally validated in the Firmicute class Bacilli (Choonee et al., 2007; Grundy and Henkin, 1991; Scott and Williamson, 2001). While each of these ribosomal proteins is also a regulator in *E. coli*, the RNAs show little or no homology to the *E. coli* RNAs with the same function. Currently, Rfam alignments are only available for one of these RNAs, L20 [Rfam: RF00558] (Burge et al., 2013). While several additional putative RNA structures associated with ribosomal proteins have been identified in comparative genomic studies (Yao et al., 2007), none have been experimentally validated. I have updated and refined these existing putative RNA structures to better assess candidates for future experimental validation.

Main objective:

I identified a number of novel ribosomal regulatory elements within Bacilli, as well as curated existing alignments of such elements. Many ncRNA regulators of r-protein synthesis have been computationally proposed, but none have been experimentally verified. Nor has it been determined if such proposed RNAs are transcribed. I used existing alignments, and computationally confirmed the proposed structure, performed homology searches and examined the phylogenetic distribution and consensus sequence for each of these proposed RNAs.

Using the novel pipeline GAISR (Genomic Analysis for Illuminating Structured RNA, Figure 2), constructed in our lab, I identified four potential ribosomal regulatory elements that I believe interact with ribosomal proteins L25, L31, S1 and S16. Multiple sequence alignments were created for each of these RNAs. In addition there are multiple sequence alignments for proposed ribosomal regulatory elements believed to interact with ribosomal proteins S6, L13, L19, L21, and S10 (Yao et al., 2007). I used

computational methods to identify the elements most worthy of further study. My work will guide future experimental validation of these proposed elements.

I completed homology searches for nine existing multiple sequence alignments corresponding to *B. subtilis* ribosomal autogenous regulatory elements, created consensus secondary structure diagrams, and determined evolutionary diversity for each RNA. Evolutionary diversity will determine the prevalence of the given RNA structure. The consensus structure determines base-pairs and nucleotides of greatest conservation that have greater likelihood of being important for binding and can direct subsequent experimental studies. Additionally, transcriptomic data is a novel incorporation to the GAISR pipeline, and will be beneficial in both this and subsequent comparative genomics studies where experimental data for proposed RNA elements is lacking.

To gain understanding of the evolutionary diversity of ribosomal protein regulatory RNAs across a wider spectrum of bacteria, I identified several putative RNA structures associated with the S15 coding regions of Alphaproteobacteria, Actinobacteria, and Chlamydia during this process. During this process, I rediscovered the Gammaproteobacteria and Firmicute S15-associated regulators using the GAISR pipeline, indicating that our process can identify known regulators.

Figure 1: Overview of Ribosomal Autogenous Regulation. Ribosomal autogenous regulatory elements help to regulate expression of their associated gene, though they are not themselves translated. When transcribed, these RNA sequences form unique sets of stem-loop structures that are believed to be the key to the specificity in regulating gene expression, as these secondary structures interact with a specific ribosomal protein binding partner to regulate the entire ribosomal protein operon from which they originate.

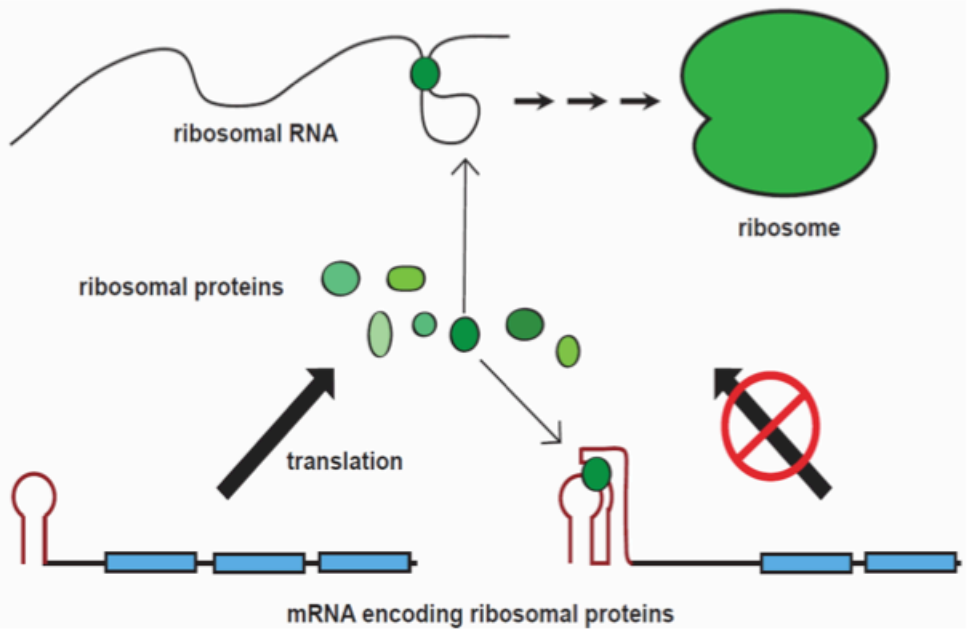
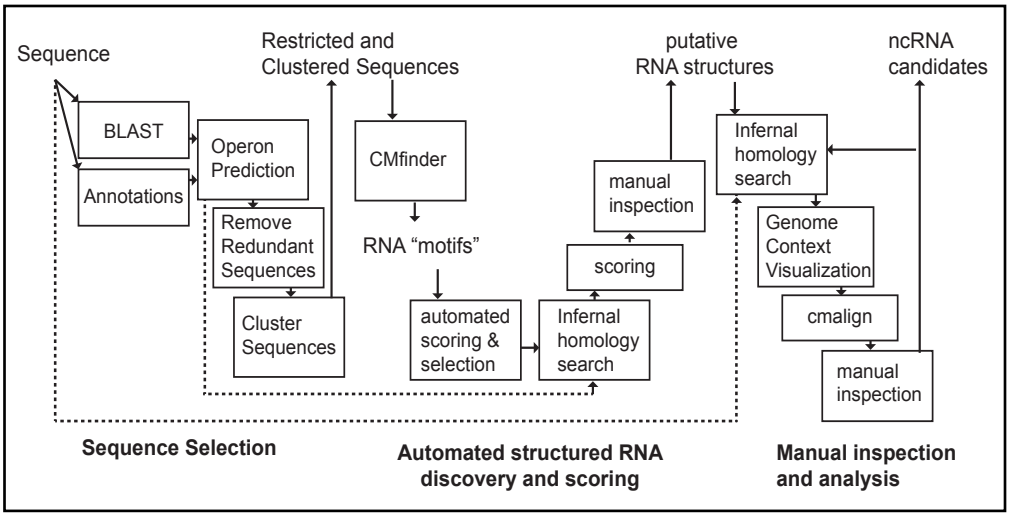


Figure 2: Overview of Computational Discovery Process (GAISR). Processes (CMfinder, scoring, etc.) are boxed and data types (sequences, putative ncRNAs, etc.) unboxed. Solid arrows show workflow and dashed lines show discontinuous information flow. GAISR has three elements: sequence selection, automated discovery, and manual analysis



Chapter 2. Comparative Genomic Analysis of Previously Experimentally Validated Bacillus Riboregulators

This chapter is excerpted heavily from: Deiorio-Haggar, et al. (2013).

In Bacilli, there are three experimentally validated ribosomal-protein autogenous regulatory RNAs that are not shared with *E. coli*. Each of these RNAs forms a unique secondary structure that interacts with a ribosomal protein encoded by a downstream gene, namely S4, S15, and L20. Only one of these RNAs, that interact with L20, is currently found in the RNA Families Database. I created, or modified existing, structural alignments for these three RNAs and used them to perform homology searches. I have determined that each structure exhibits a narrow phylogenetic distribution, mostly relegated to the Firmicute class Bacilli. This work, in conjunction with other similar work, demonstrates that there are most likely many non-homologous RNA regulatory elements regulating ribosomal protein biosynthesis that still await discovery and characterization in other bacterial species.

Initial multiple sequence alignments were obtained as described below for each RNA. The seed alignment for the L20-binding RNA was downloaded from the Rfam database (Rfam families: RF00558), the S4-binding RNA was obtained via Yao *et al.* (2007) supplementary material, and the S15-binding RNA was created manually via BLAST (completed genomes only) matches to the *G. stearothermophilus* sequence of interest and hand alignments of a few selected sequences as noted in the text. The alignments were all manually edited to remove sequences not compatible with published experimental data and to adjust base pairing. Any changes to base paring are discussed in the text for each RNA.

L20-interacting RNA

The RNA structure interacting with ribosomal protein L20 to regulate the *infC-L30-L20* operon was discovered independently by two studies at approximately the same time. One study experimentally analyzed the RNA (Choonee et al., 2007), while the other identified the RNA structure through a comparative genomics approach (Yao et al., 2007). In contrast to the L20-interacting RNA present in Gammaproteobacteria, the L20-interacting RNA in *B. subtilis* regulates transcription rather than translation, and does not directly precede the gene encoding its binding partner, *rplT* (Guillier et al., 2002). Instead, the protein-binding site is located at the start of the operon, preceding and regulating the genes *infC, rpmI,* and *rplT* (Choonee et al., 2007). The current Rfam alignment for this RNA [Rfam: RF00558] (generated by comparative genomics in Yao *et al*. (2007)), representing the L20-bound form, was used as the starting alignment for this study.

The L20-interacting RNA structure consists of three pairing elements (Figure 3), including a terminator stem (H3) that was experimentally confirmed using *in vitro* transcription termination assays (Choonee et al., 2007). A fourth pairing element directly preceding the terminator was noted in previous studies (Choonee et al., 2007). This fourth stem is present in >75% of species in the alignment, but shows no conservation of individual nucleotides. In addition, nuclease footprinting assays indicate it is unlikely to be involved in L20 binding (Figure 3) (Choonee et al., 2007).

Toe-printing and RNase probing assays have determined that L20 binds specifically at the junction of Helices 1 and 2, stabilizing Helix 2 in the process (Choonee et al., 2007). This region of the RNA bears striking resemblance to the L20 binding site on the 23S rRNA, and our studies show that this region is highly conserved, with few or no

mutations. The terminator stem does not appear to be involved in L20 binding (Choonee et al., 2007), and there is covariation throughout the stem, indicating that the secondary structure, rather than the sequence, is important for function. There is also a rigorously conserved pair of adenosines before the start of the first helix, but their effect on binding is unknown.

Of the three RNAs examined, the L20-binding RNA has the greatest penetration in sequenced Firmicutes (Figure 4). It is found in most Bacilli, and many Clostridia species. A few homologues of the L20-interacting RNA are also identified in Thermotogae and Actinobacteria. While Thermotogae are relatively closely related to Firmicutes (Wolf et al., 2004), the presence in Actinobacteria species suggests possible horizontal transfer events.

S15-interacting RNA

The mRNA structure interacting with ribosomal protein S15 regulates only *rpsO* (Scott and Williamson, 2005). Like its *E. coli* counterpart (Philippe et al., 1990), the RNA structure identified in Bacilli overlaps with the beginning of the coding region for *rpsO* (Scott and Williamson, 2005). The RNA was first identified in *B. stearothermophilus* (subsequently reclassified as *Geobacillus stearothermophilus*) (Nazina et al., 2001), and RNA-protein interaction was experimentally validated initially utilizing *in vitro* approaches. Regulatory activity of the RNA was subsequently demonstrated using *E. coli* as a surrogate organism (Scott and Williamson, 2005). There is no Rfam alignment for this RNA, nor has it been identified in previous comparative genomic works.

For this study, I manually constructed a starting alignment consisting of the 3-helix junction necessary for binding and regulation (Scott and Williamson, 2001; Scott and Williamson, 2005) using BLAST to identify the initial homologues. The starting

alignment contained sequences from several Geobacillus species, as well as a hand-aligned portion of the genomic region preceding *rpsO* in *B. subtilis* and *Caldicellulosiruptor bescii* DSM 6725 (a member of *Clostridia*). Utilizing this initial alignment I was able to identify the S15-binding RNA in most Bacilli species, and in both sequenced Negativiticutes species (Figure 4). However, its incidence in Clostridia is considerably lower, resulting in a lower overall frequency in Firmicutes. I also identified sequences in Deinococci and Fusobacteria (2 sequences in each phylum), suggesting potential horizontal transfer. However, it is difficult to make a definitive conclusion at this time due to the small number of putative homologues and lack of any experimental data to verify them.

Although alternative structures may exist, our final secondary structure is presumably stabilized by interactions with S15 (Figure 3) (Scott and Williamson, 2001). Consistent with deletion studies suggesting the length of Helix 1 may vary (Scott and Williamson, 2005), the sequence and length of Helix 1 in naturally occurring examples can range from 9 to 17 base pairs. The putative 'AUG' start codon within the loop of that helix (Figure 3, black box) is highly conserved, appearing in >97% of all sequences. Similarly to Helix 1, Helix 2 shows nucleotide sequence variability, but base pairing throughout the stem is largely maintained. While deletion studies have shown that the H2 helix may be reduced to 29 nucleotides and still retain functionality (Scott and Williamson, 2001), our alignment shows that the full-length stem is maintained in >90% of all sequences. A consecutive set of 'G-C' and 'G·U' pairs in H2, appearing as 'G-C' and 'R·Y' base pairs in Figure 2 due to sequence variability, were both expected to be highly conserved, as both base pairs are reported to be important for binding (Scott and Williamson, 2001). However, only the 'G-C' pair showed >90% conservation, while the 'G·U' pair ('R·Y' in Figure 1) exhibited much greater sequence variability, though base pairing was

maintained. Helix 3 includes a conserved 'GGAGG' that based on its location relative to the conserved 'AUG,' is likely part of the Shine-Dalgarno sequence.

S4-interacting RNA

The *B. subtilis* S4-interacting RNA regulates only the gene encoding S4, *rpsD*. In *E. coli*, the operon containing S4 also contains four additional ribosomal genes, and the S4 protein regulates the synthesis of all of them (Jinks-Robertson and Nomura, 1982). In *B. subtilis*, this gene cluster does not include *rpsD* (Jinks-Robertson and Nomura, 1982; Li et al., 1997). Rather *rpsD* is at a different location in the genome and is likely to be the only gene regulated by this RNA. *B. subtilis* S4 represses its own synthesis post-transcription initiation, but the mechanism of action remains unknown. The S4 protein is known to interact with 16S rRNA, and parts of the *B. subtilis* 5'-UTR (5'-untranslated region) have sequence and structural similarity to 16S (Grundy and Henkin, 1992). The S4-interacting RNA is found in only Firmicutes, Tenericutes, and Thermotogae, although within Firmicutes, the RNA does not appear in Clostridia or in the two Negativiticute genomes analyzed (Figure 4).

The starting alignment used here originated from the supplementary material of a comparative genomic screen of Firmicutes (Yao et al., 2007). However, the structure derived from comparative genomics did not match the experimental structure proposed by Grundy *et al.* (1992). In particular, the proposed structure incorporates sequence demonstrated to have no regulatory activity. For the work presented here, the structure was manually edited to match the structural prediction based on experimental data (Grundy and Henkin, 1992).

This structure contains two hairpins with variable bulges and stems (Figure 3). I found the hairpin branching from the first helix to be especially variable, both in sequence and

in presence. The 'GUAA' bulge (Figure 3, gray box), remains conserved, and is proposed to interact with the S4 protein (Grundy and Henkin, 1991; Grundy and Henkin, 1992), as it has sequence identity to a similar bulge on the 16S rRNA (Stern et al., 1986; Vartikar and Draper, 1989). Also, in most sequences aligned, the Shine-Dalgarno sequence follows relatively closely downstream of the second variable helix of the RNA (Figure 3). I evaluated the presence of two pseudoknots proposed in the original description of this RNA structure (Grundy and Henkin, 1991). Mutations to these putative pseudoknots did not affect regulation (Grundy and Henkin, 1992), and our alignment shows no support for the existence of pseudoknots in this structure.

I also performed homology searches with the original alignment derived from comparative genomics by Yao and coworkers (Yao et al., 2007), as this structure has the potential to be an alternative non-interacting conformation for the RNA. The final alignments of the two possible structures share significant taxonomic overlap, 172 out of 177 species, indicating that both structures are possible in most species. The majority of the alternative structure is quite similar to the S4-binding structure. H1 is largely intact in the alternative structure, with a conserved 'GUAA' bulge near the top of the stem, as well as similarly conserved loops. There are some minor changes to the base pairing of H1, causing the protein binding site ('GUAA') and the top loop to be smaller than their correlates in the S4-binding structure. More dramatically, the alternative structure lacks the H2 stem, and instead has an elongated H1 stem that partially overlaps the existing H2. In order to elongate the H1 stem, the alternative structure includes a 5' extension of 25-30 nucleotides. Although this 5' extension corresponds exactly with the transcriptional start site (Grundy and Henkin, 1990), deletion of this region was shown to have no impact on S4 protein binding (Grundy and Henkin, 1992).

Discussion

Scientists have been aware of autogenous regulators of ribosomal protein synthesis since 1980 (Nomura et al., 1980). While there is a good understanding of the RNA structures that regulate ribosomal protein regulation in *E. coli*, our knowledge of these elements outside of *E. coli* is sorely lacking. While three of the *E. coli* RNA regulators are widely distributed, the majority are not. In Bacilli, experimentally validated RNA structures interacting with S4, S15, and L20 are known to regulate ribosomal proteins. These structures show no homology to RNAs interacting with homologous proteins from *E. coli*.

For this work, I created alignments for two of the three RNAs unique to Bacilli and assessed homologues in the context of previous experimental results. I have found that the three regulatory elements examined – interacting with ribosomal proteins S4, S15, and L20 – have a narrow evolutionary distribution, even so narrow as to exclude the Firmicute class Clostridia from the S4 distribution (Figure 4). Based on various experimental analyses it is apparent that there are multiple evolutionarily distinct regulatory RNAs responding to the same ribosomal protein in different bacterial phyla (Choonee et al., 2007; Grundy and Henkin, 1992; Guillier et al., 2002; Scott and Williamson, 2001; Vartikar and Draper, 1989). Furthermore, most of the characterized RNA structures responsible for regulating ribosomal protein biosynthesis appear to be narrowly distributed (Fu et al., 2013), and comparative genomic studies have discovered a number of putative RNA structures associated with ribosomal proteins in Firmicutes and other bacterial species that have yet to be verified (Naville and Gautheret, 2010; Yao et al., 2007). The combination of these observations suggests strongly that there are many distinct RNA structures responsible for ribosomal protein regulation that have yet to be identified and experimentally characterized. In the future, identifying and

validating non-homologous regulatory RNAs that likely have the same function in non-model organisms will lead to a more comprehensive understanding of each RNA-protein interaction and to elucidation of the evolutionary trajectories for these regulatory RNAs.

Figure 3: Consensus sequence and secondary structures of Bacilli ribosomal regulatory elements. Start codons (AUG) are depicted inside a black box when occurring within the RNA structure. Gray boxes indicate areas of high conservation and possible binding. Dotted boxes surround areas proposed to be important for binding. Co-varying base pairs are shaded red or green only when Watson-Crick base pairing is in >95% of the aligned sequences. Helix numbering is consistent with previously published data for each RNA. (Deiorio-Haggar et al. (2013))

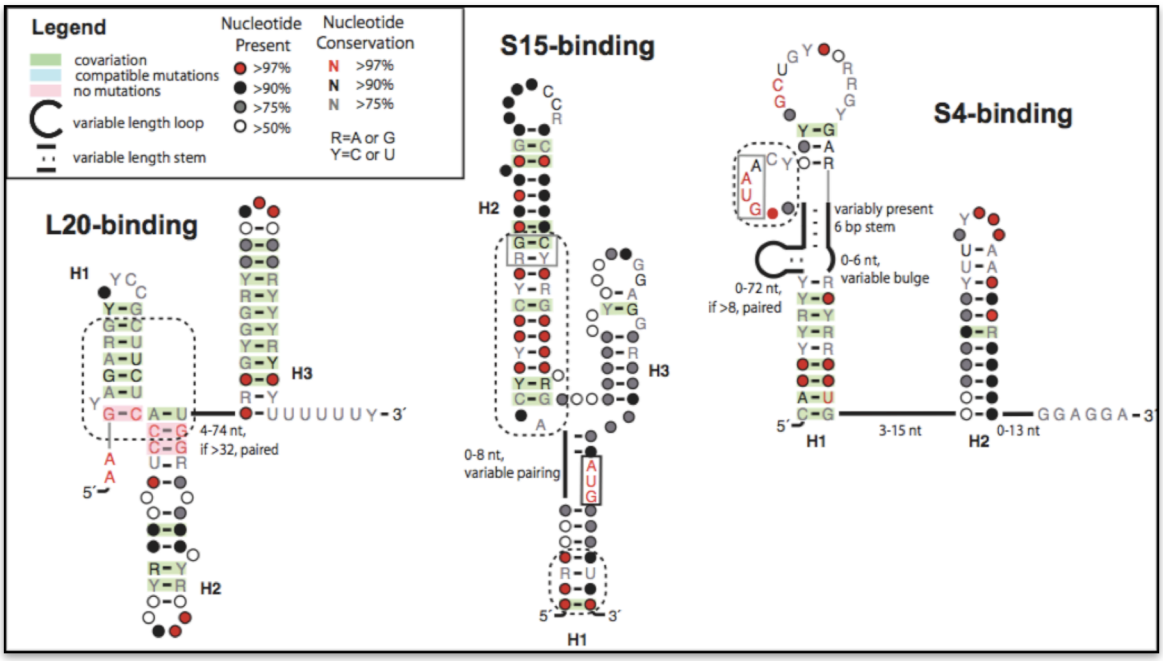
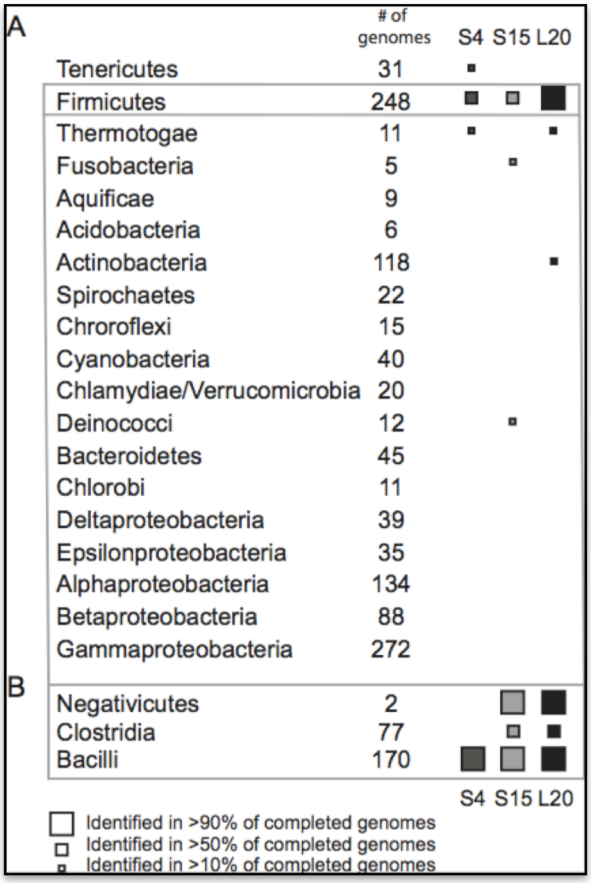


Figure 4: Phylogenetic distribution of Bacilli autogenous ribosomal regulators
A. Distribution of autogenous regulators of ribosomal protein synthesis in eubacterial phyla. Phyla are arranged based on close relation to one another.
B. Distribution of autogenous regulators of ribosomal protein synthesis for classes within the phylum Firmicutes. (Deiorio-Haggar et al. (2013))



Chapter 3. Comparative Genomic Analysis of Proposed Bacillus Riboregulators

I curated a number of existing alignments for proposed Bacillus regulatory elements. The RNAs discussed here have been computationally proposed, but were never experimentally verified in any way. I used existing alignments (Yao, et al. 2007) to computationally confirm the proposed structures of these RNAs and perform homology searches, in order to guide future experimentation of these proposed elements.

***rpsF*-Preceding RNA**

This section excerpted heavily from Fu, et al. (2013).

A computational search for structured RNAs in Firmicute genomes identified an RNA motif preceding *rpsF* (Yao et al. 2007). Additional putative RNA structures preceding this gene have been identified by similar searches in other bacterial phyla (Weinberg et al. 2007; Weinberg et al. 2010). Examination of the motifs obtained from these searches revealed a common structural element that includes a stem-loop with a bulge containing a pair of conserved cytosines. I combined the shared regions of these motifs into a common alignment and subsequently used this alignment to identify additional homologues in completed microbial genomes.

I identified over 1300 sequences matching this RNA motif across many bacterial phyla, including Proteobacteria, Actinobacteria, Cyanobacteria, Spirochaetes, and Firmicutes. A consensus representation of this alignment is in Figure 5A. The double cytosine in the bulge is strikingly conserved, and the H1 stem sequesters a potential Shine-Dalgarno sequence in many examples, suggesting possible regulatory activity. H2 is variable in

length (13-107 nucleotides in length, typically paired), but the region proximal to the bulge is fairly conserved. Apart from this shared motif, the organization of the operon is not well conserved. The RNA invariably precedes *rpsF*, but the subsequent genomic context can vary. In Cyanobacteria, *rpsF* occurs in isolation in the genome. In most Deltaproteobacteria and some Gammaproteobacteria *rpsF* and *rpsR* occur together with no intervening gene. However, in Betaproteobactera and the remaining Gammaproteobacteria they are separated by *priB,* and in Firmicutes and Actinobacteria they are separated by *ssbA*. Furthermore, in *E. coli* the operon contains *rplI* (encoding L9) (Isono et al. 1978), which occurs in isolation at a different locus in Firmicute genomes (Akanuma et al. 2012). This riboregulator shows the widest phylogenetic distribution of all Bacillus-discovered riboregulators I studied, and is found in nearly all phyla of bacteria (Figure 6).

rplM-Preceding RNA

This RNA structure is located in the 5' untranslated region of the Bacilli operon encoding L13 (*rplM*) and S9 (*rpsI*). The L13 protein binds to the central portion of the 23S RNA in conjunction with L19 (see below) (Schmidt et al. 1981; Oβwald et al. 1990), while the S9 protein binds the 3' major domain of the 16S RNA (Powers et al. 1988).

The *rplM*-preceding regulatory RNA consists of a single stem-loop structure with a variable bluge. It is followed by a conserved 'GGAG' that is thought to be part of the Shine-Dalgarno, as it is followed by an 'RUG' that is most likely the start of the L13 coding region (Figure 5B). L13 has an exceedingly narrow phylogenetic distribution, relegated only to Firmicutes, and only found in Bacilli, to the exclusion of Clostridia (Figure 6).

***rplS*-Preceding RNA**

The Bacillus *rplS*-preceding structure is proposed to regulate only L19 (*rplS*), as this gene is expressed in a separate operon (Yao et al. 2007). In *E. coli*, L19 is the last of four genes in its operon, and though the gene order is conserved in some Firmicutes, the distance between *rplS* and the stop codon of the gene directly upstream is large enough to suggest a separate transcript is made for L19 (Yao et al. 2007). In *E. coli*, L21 has been shown to bind to the 23S RNA, together with L13, with cross-links to helix 25 (Oβwald et al. 1990).

The Firmicute *rplS*-preceding structure consists of three helices, the last of which contains a conserved 'AGGAG' that is most likely part of the Shine-Dalgarno sequence, as it is followed by a conserved 'AUG' that represents the start of the L19 coding region (Figure 5C). The transcription start site for this operon in *B. subtilis* has been confirmed using 5'-RACE (unpublished work performed by the Meyer Lab). L19 is only found in Firmicutes, and while most of the L19 homologues found were in Bacilli, some examples were found in both Negativiticutes and Clostridia (Figure 6).

***rplU*-Preceding RNA**

The Firmicute L21-interacting RNA binds *rplU* (L21), and regulates the L21-ysxB-L27 operon, while in *E. coli*, there is no intervening *ysxB* gene. *YsxB* is uncharacterized and its function is currently unknown. L21 interacts with the 23S RNA complex in *E. coli* (Wower et al. 1981), and may have the same function in Bacillus.

Transcription of this RNA element has been confirmed using 5'-RACE, and the presence of this element in the 5'-UTR of the operon transcript has been validated via RT-PCR (unpublished work performed by the Meyer Lab). The structure contains a conserved 'AGGAGG' that is most likely part of the Shine-Dalgarno sequence, given its position

upstream of the 'AUG' start site for L21 (Figure 5D). The *rplU*-preceding RNA is a more widely distributed regulator, as it can be found in all three classes of Firmicutes and its fairly close relatives, Tenericutes, Thermotogae, and Fusobacteria, suggesting early evolution (Figure 6).

***rpsJ*-Preceding RNA**

This RNA is thought to regulate the S10 operon, which includes 29 genes. This long operon has two potential promoter sequences, ~200 and ~140 bases upstream of the S10 gene. It has been proposed that the these two promoters working in concert may function in the uninterrupted transcription of the large operon (Li et al. 1997). In *E. coli*, different sections of this large gene cluster are independently transcribed by S10 and L4. Therefore, it is unclear which r-protein is the most obvious potential regulator for this RNA (Li et al. 1997).

This structure consists of two hairpins and contains a conserved 'GGAG' that is most likely part of the Shine-Dalgarno sequence, as it is followed by a highly conserved 'AUG,' also in the stem-loop structure (Figure 5E). S10 seems to have executed a possible horizontal transfer event, as this RNA can be found in both Firmicutes, mainly Bacilli, and Gammaproteobacteria, a fairly distant bacterial phylum (Figure 6).

Discussion

Of the putative ribosomal regulators studied here, S6 has the most obvious potential for experimental analysis, due to its high degree of prevalence in nearly all bacterial phyla as well as its high degree of conservation and covariation. Experiments subsequent to my analysis have shown that this RNA is a functional regulator (Fu et al, 2014). Following that, I would say that L19 and L21 show great promise due to the placements of the Shine-Dalgarno and protein start codon within the regulatory structure. These

placements indicate a mechanism of repression that can be tested with directed mutagenesis to either lessen or heighten the stem's ability to base-pair to itself. L13 is an interesting specimen as it is found in Bacilli, to the exclusion of Clostridia. It would be interesting to computationally determine if there is another L13 RNA regulator in Clostridia, and then compare the two. I am undecided as to the experimental tractability of S10. Although it has good nucleotide conservation and base pair covariation, its presence in both Firmicutes and Gammaproteobacteria is intriguing and deserves further computational study before extensive experimental analysis.

Figure 5. Consensus sequence and secondary structures of putative Bacilli ribosomal regulatory elements. A: *rpsF*-preceding structure; B: *rplM*-preceding structure; C: *rplS*-preceding structure; D: *rplU*-preceding structure; E: *rpsJ*-preceding structure. In each figure applicable, the Shine-Dalgarno is boxed in gray, and the protein start codon is boxed in black.

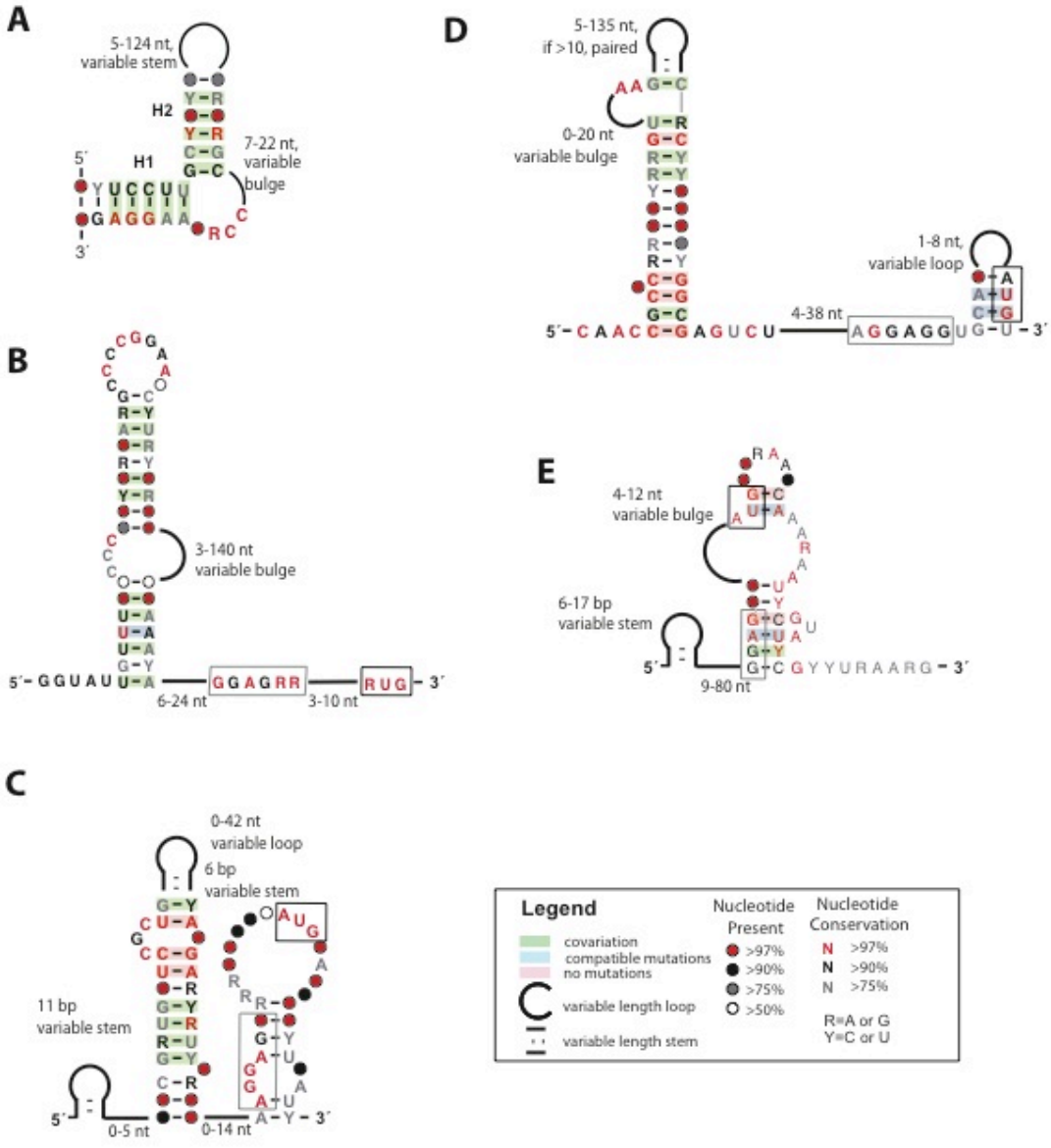
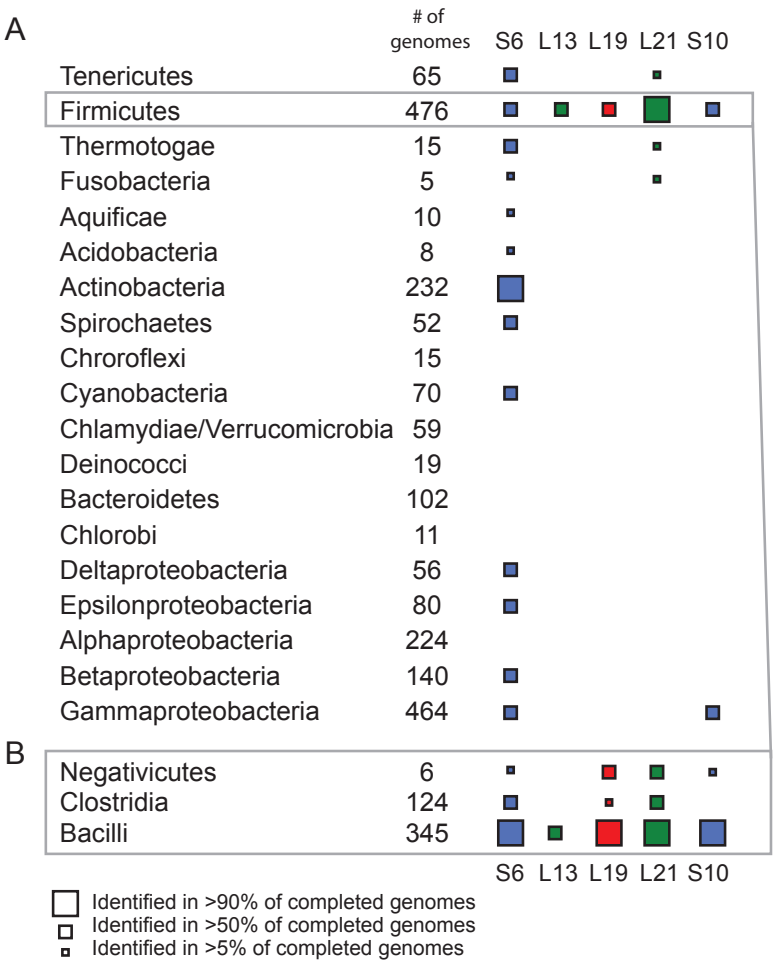


Figure 6. Phylogenetic distribution of putative Bacilli autogenous ribosomal regulators

A. Distribution of autogenous regulators of ribosomal protein synthesis in eubacterial phyla. Phyla are arranged based on close relation to one another.

B. Distribution of autogenous regulators of ribosomal protein synthesis for classes within the phylum Firmicutes.



Chapter 4. Novel Bacillus Ribosomal Regulatory RNAs

I identified several novel ribosomal regulatory elements within Bacilli. Using the pipeline GAISR, I identified RNAs which directly precede the ribosomal genes *rplY, rpmE, rpsA,* and *rpsP*. I created multiple sequence alignments, performed homology searches, and identified which proposed regulators would be most and least worthy of experimental validation.

rplY-Preceding RNA

The Bacillus *rplY*-preceding structure regulates only *rplY,* the gene encoding L25. In *E. coli*, *rplY* undergoes translational autogenous regulation, though it is unknown if this mechanism is shared in Bacillus (Aseev et al. 2015). The Bacillus L25 protein consists of two domains, an N-terminal domain that shares homology to the *E. coli* L25 RNA and binds the E-loop of the 5S ribosomal subunit, and a far more variable C-terminal end that has homology to the CTC (catabolite controlled) protein family (Schmalisch et al. 2002). Thus *B. subtilis* L25 shares homology with other general stress and CTC proteins both within and without Bacillus (Gongadze et al. 2008). Because the Bacillus L25-CTC ribosomal protein is only expressed under low-nitrogen stress conditions, it has not been as extensively studied in this organism as it has in other model organisms such as *E. coli* and *Thermus thermophilus*, in which it is constitutively expressed (Schmalisch et al. 2002; Aseev et al. 2015).

The RNA structure I identified shows extensive nucleotide conservation with base pair covariation and showing some pairs with no mutation. The two helix structure is followed by a conserved ‘GGUGGA’ that is thought to be part of the Shine-Dalgarno, as it is followed by a conserved ‘AUG’ believed to be the start of the L25 coding region (Figure 7A). The *rplY*-preceding RNA has one of the most narrow distributions of all RNAs I

have studied, being found in only Staphylococcus bacteria (Figure 8).

***rpmE*-Preceding RNA**

In *E. coli* the L31 protein exists in both the full-length (7.9 kDa) and the fragmented (7.0 kDa) forms (Arnold and Reilly, 1999), and both forms have been shown to have a transient connection to the 50S RNA, due to differing disulfide bond configurations (Eistetter et al. 1999). In *B. subtilis*, however, there is evidence of two L31-protein paralogues, *rpmE* and *ytiA*, the differing expression of which has been shown to be zinc-dependent (Akanuma et al. 2006). *B. subtilis* expresses the RpmE protein in exponential growth phase, while *ytiA* was found to be expressed during stationary phase (Nanamiya et al. 2008). The ribosomal regulatory element discussed here is proposed to regulate the *rpmE* form of the L31 gene.

The closing stem and helix 2 show covariation, and there is a conserved 'AAGGA' before the closing stem that is most likely part of the Shine-Dalgarno Sequence, as the following 'AUG' is thought to represent the coding region of *rpmE* (Figure 7B). L31 has a narrow distribution, found only in Firmicutes, and within that, in nearly all Bacilli to the exclusion of Clostridia (Figure 8).

***rpsA*-Preceding RNA**

The interactions of r-protein S1 with the ribosome are still unclear (Sengupta et al. 2011), though it is known that its presence is required for the translation of its own transcript as well as others (Boni et al. 2001; Sørensen et al. 1998). In *E. coli*, a non-coding RNA element is found in the 5'-UTR of *rpsA* that regulates the translation of the S1 protein (Boni et al. 2000) and interacts with the S1 protein and 30S ribosome *in vitro* (Boni et al. 2001). A putative RNA structure precedes *rpsA* in many species of both Cyanobacteria (Weinberg et al. 2007) and Gammaproteobacteria (Fu et al. 2013). These

RNAs bear no resemblance to the RNA in *B. subtilis*.

The Firmicute S1 binding structure consists of three helices, with little to no mutation. These helices are followed by a conserved 'AGGAGG' that is most likely part of the Shine-Dalgarno sequence, as it is followed by a conserved 'AUG' that represents the start of the S1 coding region (Figure 7C). S1 has the fewest available sequences in its alignment, and as such has one of the narrowest distributions. Only found in Firmicutes, S1 is only found in the Bacillales class within Bacilli (Figure 8).

***rpsP*-Preceding RNA**

Unfortunately, not all putative elements show promise computationally as potential regulators, as was the case with the putative S16-regulatory element. Of the three proposed structures tested, there was not enough genetic variation or sequence diversity to either choose one proposed structure over another, nor to convince me that this regulatory element was worthy of further study.

Discussion

The *rpmE*-preceding RNA shows the greatest amount of potential of these three regulators, due to its comparatively wider distribution and the amount of covariation found in its consensus diagram. Also, the placement of the Shine-Dalgarno before the closing stem of the structure indicates a possible entrapment mechanism being used (Marzi et al, 2007). The *rplY*- and *rpsA*-preceding RNAs show good covariation and sequence similarity, but their low prevalence and relatively few homologues suggest that there may be additional *rplY* and *rpsA* regulators in other Firmicutes. These other regulators would be most interesting to study in tandem with the ones discussed here, so I would hesitate to experiment on these RNAs at this time without first doing further computational study.

Figure 7. Consensus sequence and secondary structures of novel Bacilli ribosomal regulatory elements. A: *rplY*-preceding structure; B: *rpmE*-preceding structure; C: *rpsA*-preceding structure. In each figure, the Shine-Dalgarno is boxed in gray, and the protein start codon is boxed in black.

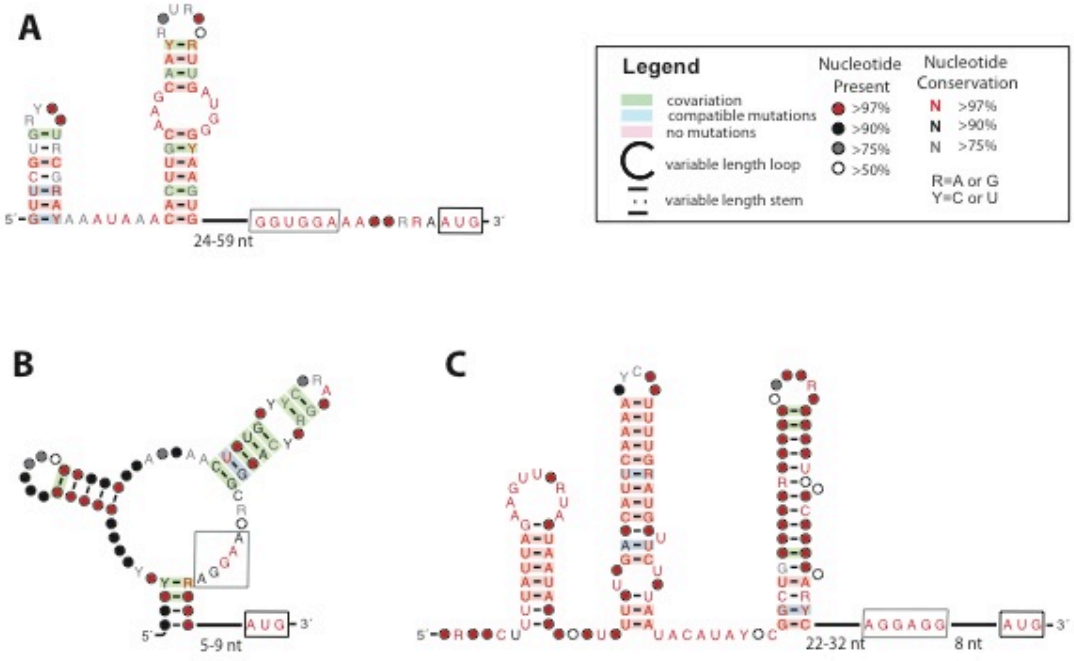
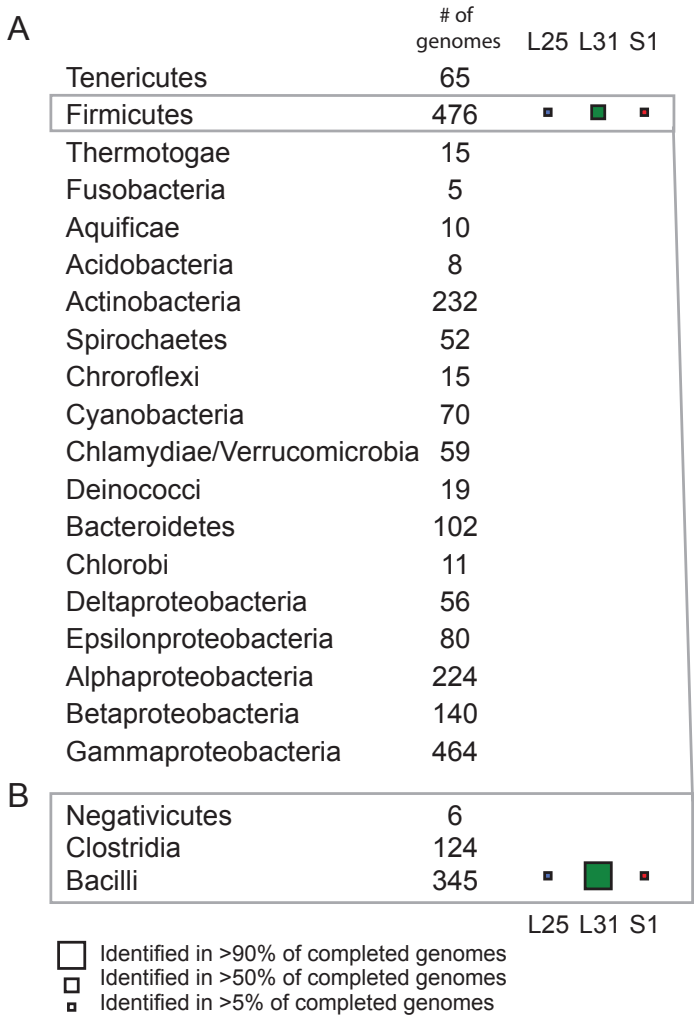


Figure 8. Phylogenetic distribution of novel Bacilli autogenous ribosomal regulators

A. Distribution of autogenous regulators of ribosomal protein synthesis in eubacterial phyla. Phyla are arranged based on close relation to one another.

B. Distribution of autogenous regulators of ribosomal protein synthesis for classes within the phylum Firmicutes.



Chapter 5. S15 Riboregulators Discovered Outside Bacillus

This chapter excerpted heavily from Slinger, et al. (2014).

The RNAs that interact with ribosomal protein S15 are representative of what is likely a common phenomenon. To date, three different RNA structures that interact with ribosomal protein S15 have been identified in *E. coli*, *Thermus thermophilus*, and *Geobacillus stearothermophilus* (formerly *Bacillus stearothermophilus*) (Figure 9). Each RNA structure appears to have distinct binding determinants, and they bear little resemblance to the rRNA binding-site for S15 (Mathy et al. 2004; Scott and Williamson, 2005). Yet, each allows negative regulation of *rpsO*, the gene encoding S15. I used comparative genomics to identify several putative RNA structures associated with the S15 coding region across diverse microbial genomes. These structures are diverse and many also encompass potential regulatory regions including ribosome-binding sites.

rpsO-Preceding Structure Discovery

To streamline the identification of putative RNA structures associated with the coding region for ribosomal protein S15 (*rpsO*), I used GAISR to examine the genomic region corresponding to the 5'-untranslated region of the gene encoding S15, *rpsO,* in fully sequenced bacterial genomes. From the initial search I identified 52 RNA motifs, originating from 16 initial phylogenetic sequence clusters.

From these initial RNA motifs, I identified five promising RNA structures based on motif complexity and a combination of conserved and diverse nucleotide arrangements. Among these structures were the two known RNAs that allow regulation of *rpsO* in Firmicutes and Gammaproteobacteria (Fu et al. 2013; Deiorio-Haggar et al. 2013). Of note, the RNA structure reported for *Thermus thermophilus* was not identified by my

search, suggesting that more RNAs may be present that were not uncovered here. There are several potential reasons for this result, including biases in sequence coverage (there were only 19 sequences derived from Deinococcus/Thermus available for analysis), and my use of a single RNA discovery tool for identification of RNA structures may limit the ability to identify some putative RNA structures. In addition, no tool for RNA *de novo* discovery is designed to identify potential pseudoknotted structures, yet these are very common in biologically functional RNAs (Staple et al. 2005). Pseudoknotted structures are typically identified manually during the curation process.

Alignments corresponding to the three promising novel structures were curated and additional examples identified using Infernal homology searches. In addition, the phylogenetic distribution of each putative ncRNA was examined and consensus diagrams of the three final candidate ncRNAs are shown in Figure 10.

S15 riboregulators identified are diverse in sequence and secondary structure

My first RNA was identified in mainly in the Rhizobiales, Rhodobacteriales, and Rhodospirillales orders of Alphaproteobacteria. However, a few examples were identified in species classified as Caulobacterales, Richettsiales, Sphingomonadales. The putative RNA structure consists of two pairing elements, H1 and H2, and in 50% of examples there is a long-linker region between the two pairing elements (up to 400 nt), that is typically base-paired, although the patterning of this base-pairing does not appear to be well-conserved. The most highly conserved portion of the putative RNA is the H1 helix. This helix shows extensive evidence of covariation and the loop region is highly conserved, suggesting that it is important for protein binding. The H2 helix is less well-conserved, but typically encompasses a putative ribosome binding site in the 3' portion.

The original alignment included a non-conserved H0 pairing element, but subsequent experimental analysis indicated that the transcription start site being midway through the loop prevented formation of this helix (Figure 10D) (Supplementary Data from Slinger et al. 2014).

My second RNA was identified mainly in the Actinomycetales order of Actinobacteria. The putative RNA structure contains a core pseudoknotted structure that bears some resemblance to the RNA structure originating from *E. coli* (Figure 9A), and there are weakly scoring homologs that appear in various Gammaproteobacterial (e.g. Pseudomonas) lacking the known *E. coli* S15 regulator (Fu et al. 2013). However, the closing pseudoknot occurs prior to any potential regulatory sequences suggesting that the "entrapment" mechanism proposed for the *E. coli* RNA is not likely to play a role here (Serganov et al. 2002; Philippe et al. 1993). Like the RNA described above, a ribosome-binding site is apparent in the 3' portion of the terminal helix suggesting a potential translational regulatory mechanism. Subsequent experimental analysis confirmed the transcription start site upstream of the start of the structural motif (Figure 10E) (Supplementary Data from Slinger et al. 2014).

My third RNA originates from Chlamydia, and is the one I have the least confidence in, mainly due to the limited sequence diversity available for analysis. However, there is a very strongly conserved hairpin overlapping start of *rpsO* in approximately 30 sequenced strains of Chlamyida and a second potential short pairing element displaying some covariation and compatible mutations. Notably, very few regulatory RNAs have been identified in Chlamydia. Only examples of the TPP and cobalamin riboswitches have been identified in Chlamydia (Gardner et al. 2011), and in these cases there appear to only be isolated sequences rather than elements that are conserved in many genomes. Pre-existing analysis of transcript start sites in *Chlamydia trachomatis* indicates that the

transcript start site is just upstream of the first predicted hairpin (Albrecht et al. 2010). Furthering the lack of confidence in this RNA, experimental analysis of the transcription of this RNA showed the transcription start site to be partway through the first helix, preventing formation of the majority of the structural motif (Figure 10F) (Supplementary Data from Slinger et al. 2014).

Discussion

Despite their shared function, none of the *rpsO*-preceding RNA structures show any obvious sequence or structural similarity (Figure 9). While this collection of RNA regulators highlights RNA structural diversity, examination of their phylogenetic distributions indicates that most bacterial phyla have no previously described S15 regulation (Fu et al. 2013; Deiorio-Haggar et al. 2013). This S15 work shows that nature may invent many unique ways to solve a single biological problem. Due to the diversity of RNAs found for S15 regulation, I expect novel structures in diverse phyla to be found as more genomes are sequenced and computational search sensitivity increases.

Figure 9. Ribosomal protein S15-interacting RNA leader sequences and structures originating from different phyla of bacteria.
A: *Escherichia coli*, B: *Thermus thermophilus*, and C: *Geobacillus stearothermophilus*. In each structure the *rpsO* start codon is boxed, and a bar is placed over the ribosome binding site. (Slinger et al. 2014).

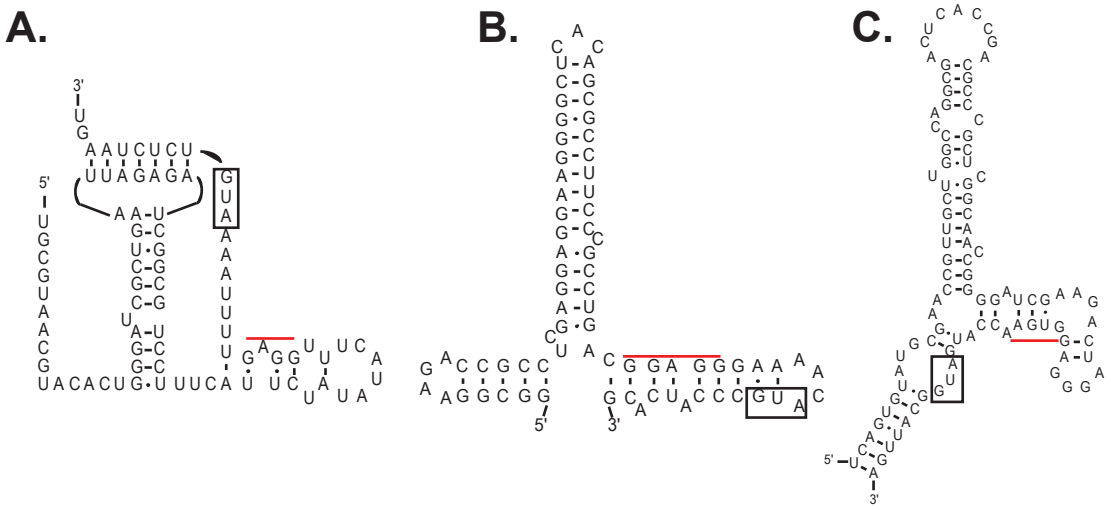
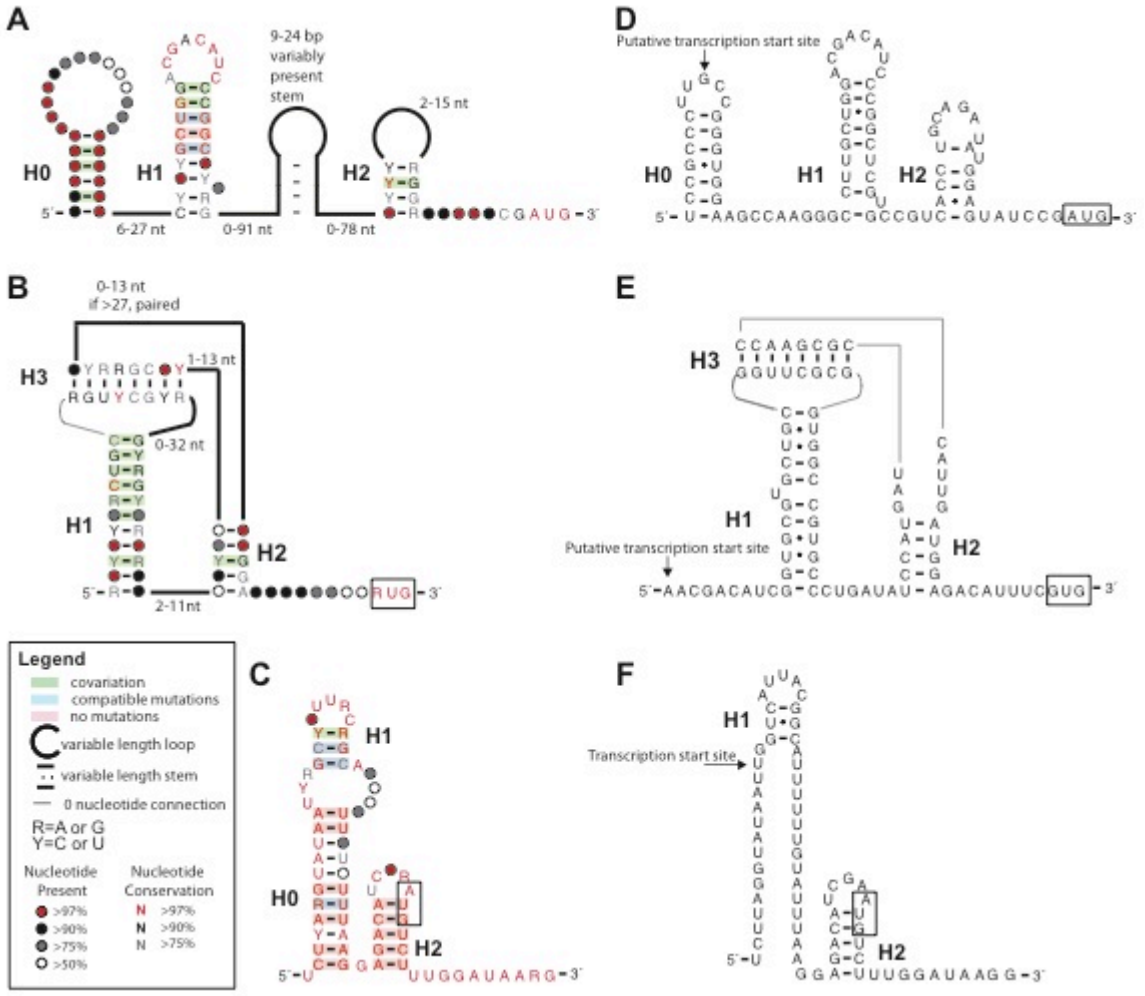


Figure 10. Consensus diagrams of putative RNA structures
A: putative RNA consensus structure originating from Alphaproteobacteria; B: putative RNA consensus structure originating from Actinobacteria; C: putative RNA consensus originating from Chlamydia; D: proposed Alphaproteobacteria RNA structure, showing putative transcription start site; E: proposed Actinobacteria RNA structure, showing putative transcription start site; and F: proposed Chlamydia RNA structure, showing putative transcription start site.
(Slinger et al. 2014).



Conclusions

My work on RNA structures regulating ribosomal biosynthesis has prompted experimental analysis into RNA regulatory elements that have produced publications on both the S15- and S6-interacting structures (Slinger et al, 2014; Fu et al, 2013). Additional experimental work is currently being done on many of the other putative RNA regulators I have discovered and studied. Not including the S15 and S6 regulators that have already been verified, I find the *rpmE*- and *rplS*-preceding structures to be the most biologically relevant.

This work shows that there are significant gaps in our knowledge of r-proteins and their regulation. Computational analysis can help to direct a more efficient flow of experimental analysis in this area, though it is not without its flaws. It is only when computational analysis is performed in conjunction with experimental work that true progress can be made.

Materials and Methods:

Computational Discovery of Ribosomal Autogenous Regulatory Elements

Our lab has created a computational pipeline, GAISR (Genomic Analysis for Illuminating Structured RNA) for *de novo* ncRNA discovery and candidate match refinement. GAISR is based on a pipeline published by Walter Ruzzo and coworkers (Yao et al., 2007) and used by Ronald Breaker and others successfully to identify ncRNA regulatory elements in bacteria (Meyer et al., 2009; Weinberg et al., 2007; Weinberg et al., 2010). GAISR utilizes several pre-existing tools, including CMfinder, a *de novo* ncRNA discovery tool (Yao et al., 2007), and Infernal 1.0, an RNA homology search tool (Nawrocki et al., 2009). GAISR also streamlines sequence selection as well as processing of candidate RNA matches (Figure 2).

Sequence selection: To generate multiple sequence alignments, BLAST matches to the target sequence outside of the genus of interest are first collected and aligned. GAISR then uses a custom operon database created from genomic data to find candidate matches within 5'-untranslated region (5'-UTR) and intergenic regions of genes of choice (Dam et al., 2007; Mao et al., 2009). All candidate matches are filtered to remove redundant sequences (>90% sequence identity for >70% sequence length) and clustered by taxon.

Automated discovery: GAISR uses CMfinder, a *de novo* ncRNA discovery tool (Yao et al., 2007), which takes the multiple sequence alignment and identifies structured RNA motifs within each sequence (Meyer et al., 2009; Weinberg et al., 2007; Weinberg et al., 2010). Similar motifs in each sequence are aligned, and additional matches to the RNA motif are found via Infernal 1.0 homology search (Nawrocki et al., 2009). RNA motif

alignments are then scored (Yao et al., 2007) and manually inspected to finalize the putative RNA structures.

RNA analysis: Once initial alignments are thus obtained, Infernal 1.0 is used to build and calibrate covariance models of the alignment in order to perform the homology search. Our searches are performed against a custom sequence database. This custom database contains eubacterial genomic regions surrounding ribosomal proteins from all complete eubacterial genomes in refseq46 (Pruitt et al., 2003). This database is ~57MB, which is a decrease of ~100x compared to the entire refseq46-microbial database, which causes a direct increase in the speed and sensitivity of our search process. A single search with the final alignments is typically performed against the entire refseq46-microbial database to verify that no sequences were omitted. After each homology search, new homologs are screened by genomic context using our HTML5-based genomic context visualization tool GenomeChart (Miller et al., 2013). Homologs are then manually screened and adjusted in order to better their fit to the existing alignment and to maintain consistency with experimental data, if applicable.

The search process is typically repeated 2-3 times in order to expand sequence diversity. Pseudoknotted or alternative structures are screened in additional searches. Percentages of bacteria in each phyla containing each RNA are calculated based on the number of completed genomes within refseq46. Consensus secondary structure diagrams are created from the alignments using GCS-weighting in R2R (Weinberg and Breaker, 2011).

Computational Refinement of Matches

In order to refine the multitude of matches resulting from GAISR's homologous search, our lab has created SCCS (Sequence Conservation, Context, and Size) as an optional

addition to the pipeline. As Infernal finds homologues, it can include copious amounts of undesirable bad matches. SCCS filters those potential matches automatically based on hit size (removing matches that are too small to be real) and context (removing matches that are not followed by the same gene/operon) (J. Anthony manuscript in preparation).

Manual Curation of Alignments

In order to create a final aligned product, each alignment must be edited or curated. Given the manual nature of this process, it is subjected to the whims of the curator, though there are standards to which one must conform. Each match is assessed based on overall length, stem and loop presence and quality, and presence of any and all highly conserved bases. A single stem-loop structure may be variably present without being a detriment to the remainder of the alignment. However, one must be sure that an alignment with multiple stem-loops is not skipping a single motif unnecessarily. Often a stem-loop motif of a given match will only be partially aligned, leaving it up to the curator to manually shift bases left and right in order for the match to better fit the overall alignment. In this way, loops and bulges are created and either maintained or removed, via insertion of a gap column or manual lengthening of the stem. Base pair and single nucleotide conservation is another way to edit an alignment. The most major highly conserved regions we look out for are the Shine-Dalgarno and the supposed start codon of the regulator RNA, as there is little sequence diversity across those regions in all bacteria. At times these highly conserved bases are a part of a stem-loop structure, and the lack of them can be cause for immediate removal of any given homology match from the alignment.

References

Akanuma, G, H Nanamiya, Y Natori, K Yano, S Suzuki, S Omata, M Ishizuka, Y Sekine and F Kawamura 2012. Inactivation of Ribosomal Protein Genes in Bacillus subtilis Reveals Importance of Each Ribosomal Protein for Cell Proliferation and Cell Differentiation. *Journal of Bacteriology* **194**: 6282-6291.

Akanuma, Genki, et al. "Liberation of zinc-containing L31 (RpmE) from ribosomes by its paralogous gene product, YtiA, in Bacillus subtilis." Journal of bacteriology 188.7 (2006): 2715-2720.

Arnold, Randy J., and James P. Reilly. "Observation of Escherichia coli Ribosomal Proteins and Their Post-translational Modifications by MALDI-MS."

Aseev, L. V., and I. V. Boni. "Extraribosomal functions of bacterial ribosomal proteins." Molecular Biology 45.5 (2011): 739-750.

Aseev, L. V., et al. "Conservation of regulatory elements controlling the expression of the rpsB-tsf operon in γ-proteobacteria." Molecular biology 43.1 (2009): 101-107.

Aseev, Leonid V., Natalia S. Bylinkina, and Irina V. Boni. "Regulation of the rplY gene encoding 5S rRNA binding protein L25 in Escherichia coli and related bacteria." rna 21.5 (2015): 851-861.

Boni, Irina V., et al. "Non-canonical mechanism for translational control in bacteria: synthesis of ribosomal protein S1." The EMBO journal 20.15 (2001): 4222-4232.

Boni, Irina V., Valentina S. Artamonova, and Marc Dreyfus. "The Last RNA-Binding Repeat of the Escherichia coliRibosomal Protein S1 Is Specifically Involved in Autogenous Control." Journal of bacteriology 182.20 (2000): 5872-5879.

Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, et al. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res 2013; 41:D226-32.

Choonee N, Even S, Zig L, Putzer H. Ribosomal protein L20 controls expression of the *Bacillus subtilis infC* operon via a transcription attenuation mechanism. Nucleic Acids Res 2007; 35:1578-88.

Coenye T, Vandamme P. Organisation of the S10, spc and alpha ribosomal protein gene clusters in prokaryotic genomes. FEMS Microbiol Lett 2005; 242:117-26.

Deiorio-Haggar K, Anthony J, Meyer MM: RNA structures regulating ribosomal protein biosynthesis in bacilli. *RNA Biology* 2013, 10.

Desnoyers G, Bouchard MP, Masse E. New insights into small RNA-dependent translational regulation in prokaryotes. Trends Genet 2012.

Eistetter, Anthony J., et al. "Characterization of Escherichia coli 50S ribosomal protein L31." FEMS microbiology letters 180.2 (1999): 345-349.

Fu Y, Deiorio-Haggar K, Anthony J, Meyer MM: Most RNAs regulating ribosomal protein biosynthesis in Escherichia coli are narrowly distributed to

Gammaproteobacteria. *Nucleic Acids Research* 2013, 41:3491–3503.

Gao, Beile, Ritu Mohan, and Radhey S. Gupta. "Phylogenomics and protein signatures elucidating the evolutionary relationships among the Gammaproteobacteria." International journal of systematic and evolutionary microbiology 59.2 (2009): 234-247.

Gongadze, G. M., et al. "Bacterial 5S rRNA-binding proteins of the CTC family." Biochemistry (Moscow) 73.13 (2008): 1405-1417.

Grundy FJ, Henkin TM. Characterization of the *Bacillus subtilis rpsD* regulatory target site. J Bacteriol 1992; 174:6763-70.

Grundy FJ, Henkin TM. Cloning and analysis of the *Bacillus subtilis rpsD* gene, encoding ribosomal protein S4. J Bacteriol 1990; 172:6372-9.

Grundy FJ, Henkin TM. The *rpsD* gene, encoding ribosomal protein S4, is autogenously regulated in *Bacillus subtilis*. J Bacteriol 1991; 173:4595-602.

Guillier M, Allemand F, Raibaud S, Dardel F, Springer M, Chiaruttini C. Translational feedback regulation of the gene for L35 in *Escherichia coli* requires binding of ribosomal protein L20 to two sites in its leader mRNA: a possible case of ribosomal RNA-messenger RNA molecular mimicry. RNA 2002; 8:878-89.

Iben JR, Draper DE. Specific interactions of the L10(L12)4 ribosomal protein complex with mRNA, rRNA, and L11. Biochemistry 2008; 47:2721-31.

Isono, K and M Kitakawa 1978. Cluster of ribosomal protein genes in Escherichia coli containing genes for proteins S6, S18, and L9. *Proc. Natl. Acad. Sci. USA* **75**: 6163-6167.

Jinks-Robertson S, Nomura M. Ribosomal protein S4 acts in trans as a translational repressor to regulate expression of the alpha operon in *Escherichia coli*. J Bacteriol 1982; 151:193-202.

Li X, Lindahl L, Sha Y, Zengel JM. Analysis of the *Bacillus subtilis* S10 ribosomal protein gene cluster identifies two promoters that may be responsible for transcription of the entire 15-kilobase S10-*spc*-alpha cluster. J Bacteriol 1997; 179:7046-54.

Marzi S, Myasnikov AG, Serganov A, Ehresmann C, Romby P, Yusupov M, et al. Structured mRNAs regulate translation initiation by binding to the platform of the ribosome. Cell 2007; 130:1019-31.

Mathy N, Pellegrini O, Serganov A, Patel DJ, Ehresmann C, Portier C: Specific recognition of rpsO mRNA and 16S rRNA by Escherichia coli ribosomal protein S15 relies on both mimicry and site differentiation. *Molecular Microbiology* 2004, 52:661–675.

Meyer MM, Ames TD, Smith DP, Weinberg Z, Schwalbach MS, Giovannoni SJ, et al. Identification of candidate structured RNAs in the marine organism 'Candidatus Pelagibacter ubique'. BMC Genomics 2009; 10:268.

Nanamiya, Hideaki, et al. "Zinc is a key factor in controlling alternation of two types of L31 protein in the Bacillus subtilis ribosome." Molecular microbiology 52.1 (2004): 273-

283.

Naville M, Gautheret D. Premature terminator analysis sheds light on a hidden world of bacterial transcriptional attenuation. Genome Biol 2010; 11:R97.

Nazina TN, Tourova TP, Poltaraus AB, Novikova EV, Grigoryan AA, Ivanova AE, et al. Taxonomic study of aerobic thermophilic bacilli: descriptions of *Geobacillus subterraneus gen. nov., sp. nov.* and *Geobacillus uzenensis sp. nov*. from petroleum reservoirs and transfer of *Bacillus stearothermophilus, Bacillus thermocatenulatus, Bacillus thermoleovorans, Bacillus kaustophilus, Bacillus thermodenitrificans* to *Geobacillus* as the new combinations *G. stearothermophilus, G. thermocatenulatus, G. thermoleovorans, G. kaustophilus, G. thermodenitrificans* Int J Syst Evol Microbiol 2001; 51:433-46.

Nikolaichik, Yevgeny A., and William D. Donachie. "Conservation of gene order amongst cell wall and cell division genes in Eubacteria, and ribosomal genes in Eubacteria and Eukaryotic organelles." Genetica 108.1 (2000): 1-7.

Nomura M, Yates JL, Dean D, Post LE. Feedback regulation of ribosomal protein gene expression in *Escherichia coli*: structural homology of ribosomal RNA and ribosomal protein MRNA. Proc Natl Acad Sci U S A 1980; 77:7084-8.

OBwald, Monika, Barbara Greuer, and Richard Brimacombe. "Localization of a series of RNA-protein cross-link sites in the 23S and 5S ribosomal RNA from Escherichia coli, induced by treatment of 50S subunits with three different bifunctional reagents." Nucleic acids research 18.23 (1990): 6755-6760.

Philippe C, Portier C, Mougel M, Grunberg-Manago M, Ebel JP, Ehresmann B, et al. Target site of *Escherichia coli* ribosomal protein S15 on its messenger RNA. Conformation and interaction with the protein. J Mol Biol 1990; 211:415-26.

Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. "NCBI Reference Sequence project: update and current status." Nucleic acids research 31.1 (2003): 34-37.

Schlax PJ, Xavier KA, Gluick TC, Draper DE. Translational repression of the *Escherichia coli* alpha operon mRNA: importance of an mRNA conformational switch and a ternary entrapment complex. J Biol Chem 2001; 276:38494-501.

Schmalisch, Matthias, Ines Langbein, and Jörg Stülke. "The general stress protein Ctc of Bacillus subtilis is a ribosomal protein." Journal of molecular microbiology and biotechnology 4.5 (2002): 495-501.

Schmidt, Francis J., et al. "The binding site for ribosomal protein L11 within 23 S ribosomal RNA of Escherichia coli." Journal of Biological Chemistry 256.23 (1981): 12301-12305.

Scott LG, Williamson JR. Interaction of the *Bacillus stearothermophilus* ribosomal protein S15 with its 5'-translational operator mRNA. J Mol Biol 2001; 314:413-22.

Scott LG, Williamson JR. The binding interface between *Bacillus stearothermophilus* ribosomal protein S15 and its 5'-translational operator mRNA. J Mol Biol 2005; 351:280-90.

Sengupta, Jayati, Rajendra K. Agrawal, and Joachim Frank. "Visualization of protein S1 within the 30S ribosomal subunit and its interaction with messenger RNA." Proceedings of the National Academy of Sciences 98.21 (2001): 11991-11996.

Serganov A, Patel DJ. Molecular recognition and function of riboswitches. Curr Opin Struct Biol 2012; 22:279-86.

Serganov A, Polonskaia A, Ehresmann B, Ehresmann C, Patel DJ. Ribosomal protein S15 represses its own translation via adaptation of an rRNA-like fold within its mRNA. EMBO J 2003; 22:1898-908.

Slinger, B. L., Deiorio-Haggar, K., Anthony, J. S., Gilligan, M. M., & Meyer, M. M. (2014). Discovery and validation of novel and distinct RNA regulators for ribosomal protein S15 in diverse bacterial phyla. BMC genomics, 15(1), 657.

Sørensen, Michael A., Jens Fricke, and Steen Pedersen. "Ribosomal protein S1 is required for translation of most, if not all, natural mRNAs in Escherichia coli in vivo." Journal of molecular biology 280.4 (1998): 561-569.

Staple DW, Butcher SE: Pseudoknots: RNA structures with diverse functions. *PLoS Biol* 2005, 3:e213.

Stern S, Wilson RC, Noller HF. Localization of the binding site for protein S4 on 16 S ribosomal RNA by chemical and enzymatic probing and primer extension. J Mol Biol 1986; 192:101-10.

Storz G, Vogel J, Wassarman KM. Regulation by small RNAs in bacteria: expanding frontiers. Mol Cell 2011; 43:880-91.

Vartikar JV, Draper DE. S4-16 S ribosomal RNA complex. Binding constant measurements and specific recognition of a 460-nucleotide region. J Mol Biol 1989; 209:221-34.

Weinberg, Z, JE Barrick, Z Yao, A Roth, JN Kim, J Gore, JX Wang, ER Lee, KF Block, N Sudarsan, et al. 2007. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Research* **35**: 4809-4819.

Weinberg, Z, JX Wang, J Bogue, J Yang, K Corbino, RH Moy and RR Breaker 2010. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biology* **11**: 1-17.

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

Wolf M, Muller T, Dandekar T, Pollack JD. Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. Int J Syst Evol Microbiol 2004; 54:871-5.

Wower, Iwona, et al. "The use of 2-iminothiolane as an RNA-protein cross-linking agent in Escherichia coli ribosomes, and the localisation on 23S RNA of sites cross-linked to proteins L4, L6, L21, L23, L27 and L29." Nucleic acids research 9.17 (1981): 4285-4302.

Yao Z, Barrick J, Weinberg Z, Neph S, Breaker R, Tompa M, et al. A computational pipeline for high- throughput discovery of cis-regulatory noncoding RNA in prokaryotes. PLoS Comput Biol 2007; 3:e126.

Yutin N, Puigbo P, Koonin EV, Wolf YI. Phylogenomics of prokaryotic ribosomal proteins. PLoS One 2012; 7:e36972.

Zengel JM, Lindahl L. Diverse mechanisms for regulating ribosomal protein synthesis in *Escherichia coli*. Prog Nucleic Acid Res Mol Biol 1994; 47:331-70.

Zengel JM, Lindahl L. Ribosomal protein L4 stimulates *in vitro* termination of transcription at a NusA-dependent terminator in the S10 operon leader. Proc Natl Acad Sci U S A 1990; 87:2675-9.