Riboswitch diversity and distribution

PHILLIP J. MCCOWN,[1,4] KEITH A. CORBINO,[2] SHIRA STAV,[1] MADELINE E. SHERLOCK,[3] and RONALD R. BREAKER[1,2,3]

[1]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06520-8103, USA
[2]Howard Hughes Medical Institute, Yale University, New Haven, Connecticut 06520-8103, USA
[3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520-8103, USA

ABSTRACT

Riboswitches are commonly used by bacteria to detect a variety of metabolites and ions to regulate gene expression. To date, nearly 40 different classes of riboswitches have been discovered, experimentally validated, and modeled at atomic resolution in complex with their cognate ligands. The research findings produced since the first riboswitch validation reports in 2002 reveal that these noncoding RNA domains exploit many different structural features to create binding pockets that are extremely selective for their target ligands. Some riboswitch classes are very common and are present in bacteria from nearly all lineages, whereas others are exceedingly rare and appear in only a few species whose DNA has been sequenced. Presented herein are the consensus sequences, structural models, and phylogenetic distributions for all validated riboswitch classes. Based on our findings, we predict that there are potentially many thousands of distinct bacterial riboswitch classes remaining to be discovered, but that the rarity of individual undiscovered classes will make it increasingly difficult to find additional examples of this RNA-based sensory and gene control mechanism.

Keywords: aptamer; coenzyme; ligand; noncoding RNA; RNA World

INTRODUCTION

Many bacteria use riboswitches to sense fundamental metabolites or ions and control the expression of genes whose protein products manage the cellular homeostasis of these ligands (Henkin 2008; Breaker 2012; Serganov and Nudler 2013). Much has been learned about the structures, functions, and distributions of riboswitches since the first experimental validation studies were reported in 2002 (Mironov et al. 2002; Nahvi et al. 2002; Winkler et al. 2002a,b). For example, many distinct RNA aptamer structures are exploited by riboswitches to selectively sense a surprising variety of ligands (Roth and Breaker 2009; Garst et al. 2011; Serganov and Patel 2012). These aptamers regulate gene expression by interacting with adjoining expression platforms that regulate gene expression primarily by controlling transcription termination or translation initiation (Barrick and Breaker 2007; Breaker 2012), although other mechanisms are also used by bacteria (e.g., Winkler et al. 2004; André et al. 2008; Lee et al. 2010; Hollands et al. 2012; Mellin et al. 2013).

Among the most widespread riboswitch classes are those that sense and respond to RNA-based coenzymes and second messengers. These observations strongly suggest that riboswitches might have emerged during the RNA World (Breaker 2009, 2012; Garst et al. 2011), when biological systems lacked proteins that otherwise would serve as biochemical sensors and switches. However, because no riboswitch can be unambiguously traced to the last common universal ancestor, these RNAs cannot be directly linked to the RNA World. For example, representatives of only two classes are often found in species outside the bacterial domain of life. Specifically, fluoride riboswitches (Baker et al. 2012) and thiamin pyrophosphate (TPP) riboswitches (Mironov et al. 2002; Winkler et al. 2002b) are sometimes present in species of archaea, whereas TPP riboswitches are very common among fungi (Sudarsan et al. 2003a; Cheah et al. 2007), algae (Croft et al. 2007), and plants (Kubodera et al. 2003; Sudarsan et al. 2003a; Wachter et al. 2007; Bocobza and Aharoni 2014).

In contrast, several riboswitch classes are exceedingly rare, and therefore could represent recent evolutionary inventions or perhaps ancient riboswitches that are being driven to the brink of extinction. One such rare riboswitch class is called $preQ_1$-III, which, as the name implies, is the third validated riboswitch class for the modified nucleobase called

[4]**Present address: Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN 46556, USA**

Corresponding author: ronald.breaker@yale.edu

Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.061234.117.

prequeuosine$_1$ (McCown et al. 2014). Only 86 examples of this riboswitch class were identified from metagenomic sequence data or from members of a single bacterial lineage. Even rarer are the variants of guanine riboswitches that sense 2′-deoxyguanosine (2′-dG). Only four examples of the 2′-dG-I riboswitch class have been found, all of which occur in a single organism, *Mesoplasma florum* (Kim et al. 2007). Similarly, only 12 members of the distinct 2′-dG-II class have been identified to date (Weinberg et al. 2017), and all of these are present only in metagenomic sequence data.

Not surprisingly, the order of riboswitch class discovery roughly corresponds to the number of representatives in each class, such that the most common riboswitches were discovered first. This has occurred because there is a higher probability of encountering a member of a common riboswitch class versus a member of a rare class, regardless of whether the search strategy involved a genetics-based or a bioinformatics-based search. Some very rare riboswitch classes were discovered only because their representatives are very close derivatives of a more common riboswitch class, such as adenine (Mandal and Breaker 2004) or the 2′-dG-I and -II riboswitches (Kim et al. 2007; Weinberg et al. 2017), all of which are variants of the vastly more common guanine riboswitch class initially discovered. Unfortunately, these close variants will remain hidden among the bioinformatics hits for a more prominent riboswitch class unless there are key distinctions, such as mutations to the ligand-binding core of the aptamer or gene associations that indicate a change in ligand specificity (Kellenberger et al. 2015; Nelson et al. 2015; Weinberg et al. 2017).

In recent years, the most productive strategy for the discovery of riboswitches has involved the use of computer algorithms to carry out comparative sequence analysis of the noncoding DNA portions of bacterial genomes (Ames and Breaker 2010). As noted above, these searches can uncover the existence of variants of a known riboswitch class by identifying RNAs that closely correspond to the consensus sequence and secondary structure model of the predominant class (e.g., Mandal and Breaker 2004; Kim et al. 2007; McCown et al. 2014). Moreover, bioinformatics algorithms can be used to find numerous new candidate riboswitches (e.g., Barrick et al. 2004; Weinberg et al. 2007, 2010) that must be subsequently validated by genetic, biochemical, and biophysical studies.

In the current study, we have comprehensively examined the known riboswitch classes using computational algorithms to create refined consensus sequence and structural models, identify rare variants, and establish the phylogenetic distributions of each class. Our results suggest that the most common riboswitch classes have largely been found, but that a strikingly large number of rarer classes remain to be discovered. Given their rarity and the inherent difficulties in searching for small noncoding RNA motifs, the vast majority of these undiscovered riboswitch classes will likely remain hidden.

RESULTS AND DISCUSSION

Strategy for the bioinformatics analysis of known riboswitch classes

Riboswitch aptamers are among the most highly conserved cellular components in biology, which reflects their need to form highly selective binding pockets for target ligands using only the four common nucleotides (Breaker 2011). In contrast, expression platforms can regulate gene expression via a variety of mechanisms and structures (Breaker 2012), and therefore are far less well conserved. Conserved aptamer features include nucleotide sequences (particularly at or near the ligand-binding core), secondary structures such as base-paired stems and pseudoknots, and motifs such as k-turns, E-loops, special tetraloops, and other common substructures (Serganov 2009; Garst et al. 2011). Distinctive patterns of conservation can be readily established by examining only a few representatives of a riboswitch aptamer class (e.g., Grundy and Henkin 1998; Gelfand et al. 1999; Barrick et al. 2004).

Once established, consensus sequence and structural models help to define each riboswitch class (Ames and Breaker 2010; Breaker 2011). The consensus model also can be used to direct bioinformatics algorithms to search for additional RNAs that closely correspond to the consensus. These algorithms can be made to rank candidate representative "hits" based on how well their sequences and predicted substructures conform to the existing consensus model. Outliers that are proven (or presumed) to function as riboswitches can then help inform how the consensus model for the aptamer should be updated to more accurately reflect the sequence and structural constraints on the riboswitch class. This bioinformatics strategy has been used previously to reveal the existence of structural variants of preQ$_1$-I, preQ$_1$-II, and preQ$_1$-III riboswitches (McCown et al. 2014) and to uncover variants of guanine riboswitches that exhibit altered ligand specificities (Mandal and Breaker 2004; Mandal et al. 2003; Kim et al. 2007). Most recently, we have used a bioinformatics search pipeline, guided by the known atomic-resolution structures of riboswitches, to identify additional rare variants whose ligand-binding specificities have been altered (Weinberg et al. 2017).

In the current study, we conducted a comprehensive analysis of the validated riboswitch classes to identify rare variants, to create a collection of updated consensus models, and to establish phylogenetic distributions. To identify additional riboswitch representatives, homology searches were performed using Infernal (Nawrocki and Eddy 2013) and RNAMotif (Macke et al. 2001). Alignment of riboswitch sequences was achieved by using RALEE (Griffiths-Jones 2005), and the creation of updated consensus models was achieved by using R2R (Weinberg and Breaker 2011).

Infernal is a program that searches sequence databases for new members of an RNA class by comparative sequence analysis (Nawrocki and Eddy 2013). Specifically, Infernal searches make use of a covariance model (or consensus model), which exploits the sequence of a putative RNA and its relationship to a given RNA secondary structure prediction model that is based on either existing bioinformatics or biochemical data. Infernal was designed to allow for the rapid annotation of structured RNAs in newly curated genomes. By adjusting the *E*-value threshold, the user can increase the number of false negatives or false positives. For example, an *E*-value threshold of one could be expected by random chance to yield a single false positive in the data set examined. By increasing the *E*-value threshold and permitting a higher number of false positives, careful examination of the results may yield new functional members of a riboswitch class that were missed initially due to variation of a putative conserved sequence or substructure represented in the previous consensus model (McCown et al. 2011). By iteratively searching for new representatives and adjusting the consensus model accordingly, distinct variants that were previously missed in searches based on an earlier model can be identified. We used this approach to expand the number of representatives for all validated riboswitch classes.

Similarly, we sought to identify more representatives by using RNAMotif (Macke et al. 2001), which makes use of a covariance model in a format like that used by Infernal. However, RNAMotif can include pseudoknots in the covariation model, which facilitated the preliminary analysis of sequence databases to find additional representatives of a riboswitch class. Revised consensus models were then developed for use by Infernal. Ultimately, Infernal generated extremely large sequence alignment files that were subjected to further computational and manual analyses. For example, RALEE was particularly useful for aligning large RNA regions with numerous representatives, which allowed the identification of conserved nucleotide sequences. R2R was used to compute values of conservation and covariation among nucleotides in each motif (Weinberg and Breaker 2011). In addition, R2R was used to create a preliminary graphic depiction of each motif, which was subsequently adjusted to prepare representations of consensus models.

Our computer-assisted searches recorded more than 100,000 representatives of 38 validated riboswitch classes (Fig. 1). In most instances, the functions of newly found variants can be inferred because their genomic locations suggest they are regulating genes that are routinely associated with members of the parent riboswitch class. However, given that there are sometimes many hundreds or even thousands of members of a riboswitch class, experimental validation of one or more representatives is usually conducted only if the gene associations are very different or if the sequence or structural variation is substantial. Precedents exist for the discovery of rare riboswitch variants with novel ligand specificities that were difficult to distinguish from the predominant members of a larger riboswitch class. This problem was

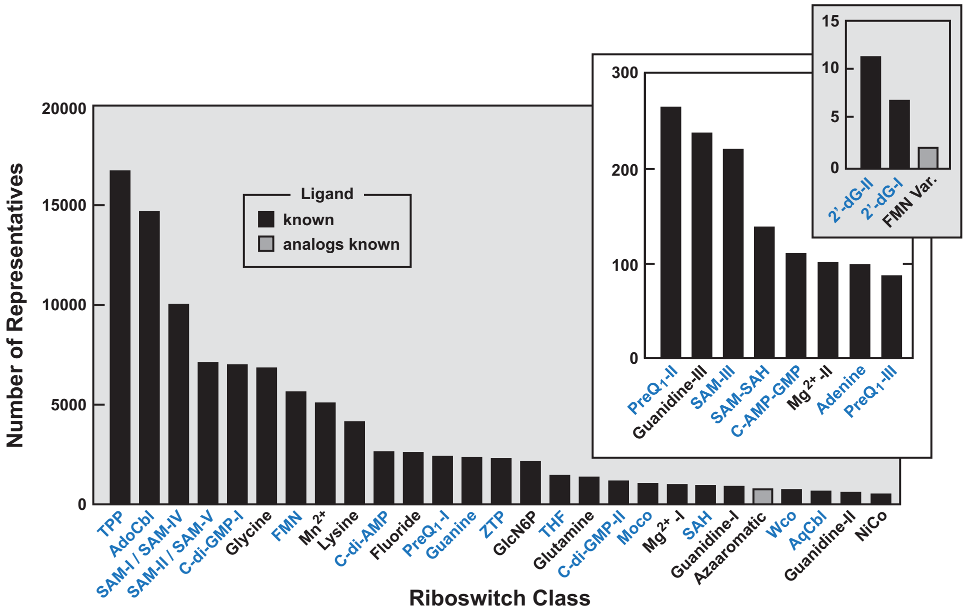


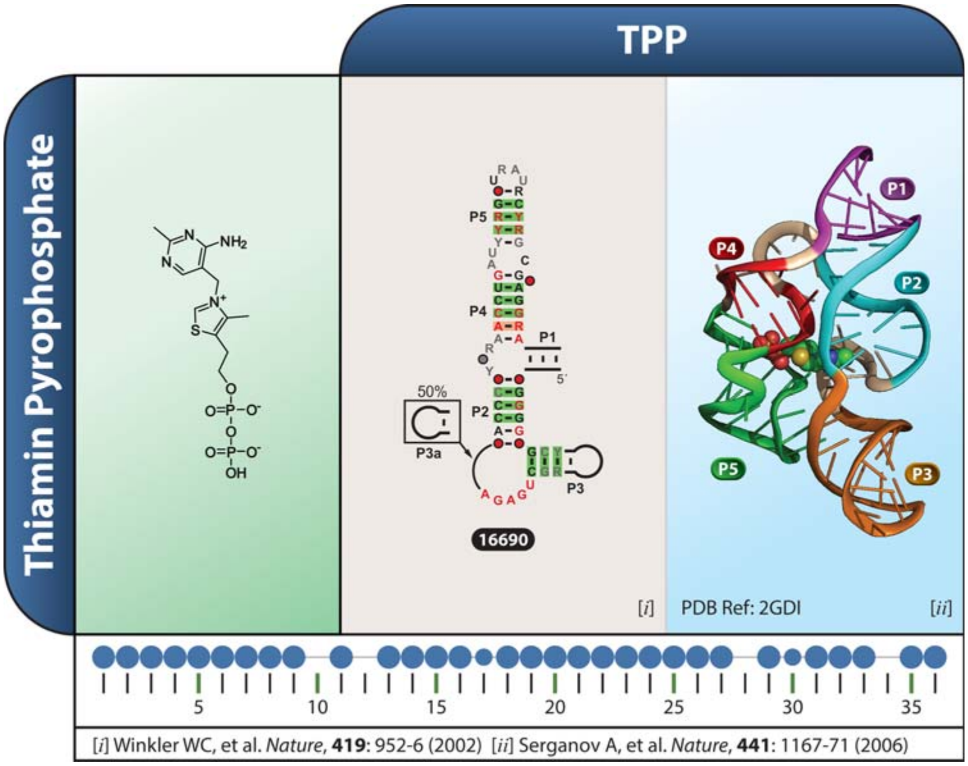
FIGURE 1. Rank order of riboswitch classes based on their abundance in genomic databases. Black and gray bars represent validated riboswitches with known and unknown natural ligands, respectively. Blue text identifies riboswitch classes whose ligands are derived from RNA nucleotides or their precursors. Each riboswitch class is named according to its ligand, wherein multiple structural classes for the same ligand are identified by Roman numerals. *Inset* plots include data for rarer riboswitch classes. (TPP) Thiamin pyrophosphate, (AdoCbl) adenosylcobalamin or coenzyme B_{12}, (SAM) *S*-adenosylmethionine, (C-di-GMP) cyclic-di-GMP, (FMN) flavin mononucleotide, (Mn^{2+}) divalent manganese, (C-di-AMP) cyclic-di-AMP, ($PreQ_1$) prequeuosine$_1$, (ZTP) 5-aminoimidazole-4-carboxamide ribonucleoside-5′-triphosphate, (GlcN6P) glucosamine-6-phosphate, (THF) tetrahydrofolate, (Moco) molybdenum cofactor, (Mg^{2+}) divalent magnesium, (SAH) *S*-adenosylhomocysteine, (Wco) tungsten cofactor, (AqCbl) aquacobalamin, (NiCo) divalent nickel and divalent cobalt, (c-AMP-GMP) cyclic AMP-GMP, (2′-dG) 2′-deoxyguanosine, (FMN-Var.) FMN riboswitch variant.

encountered for riboswitches such as adenine (Mandal and Breaker 2004) and 2′-dG-I (Kim et al. 2007) that are very similar in sequence and structure to guanine riboswitches. Similarly, riboswitches such as the c-AMP–GMP (Kellenberger et al. 2015; Nelson et al. 2015) and 2′-dG-II (Weinberg et al. 2017) classes remained hidden for many years after their parent classes had been discovered. Therefore, it is important to note that some bioinformatics hits assigned to a particular riboswitch class might actually represent unrecognized variants that have altered ligand specificity.

Graphical representations of riboswitch classes and their characteristics

On the completion of our bioinformatics analyses, we prepared a graphical representation of the distinguishing characteristics for each riboswitch class (see Supplemental File 1 for a complete collection). Each visual summary includes the following five characteristics for a given riboswitch class: (i) the name and chemical structure of the natural ligand, which also serves as the source for the name of each riboswitch class; (ii) the consensus sequence and secondary structure model for the aptamer; (iii) when available, an atomic-resolution model for the ligand-bound aptamer; (iv) the total number of riboswitch representatives currently predicted; (v) a depiction of the phylogenetic distribution of the riboswitch representatives among 36 bacterial divisions. A composite file including all riboswitch classes also can be printed as a 3′ × 4′ poster for display (Supplemental File 2).

An example of this graphical representation is provided for TPP riboswitches (Fig. 2), whose members are the most common among all known classes (Fig. 1). Several TPP riboswitch variants from species of bacteria and from eukaryotes have been demonstrated to selectively bind the coenzyme thiamin pyrophosphate (TPP) (e.g., see Mironov et al. 2002; Winkler et al. 2002b; Sudarsan et al. 2003a; Thore et al. 2006). However, members of this dominant riboswitch class can bind more weakly to thiamin monophosphate (TMP),



[i] Winkler WC, et al. *Nature*, **419**: 952-6 (2002) [ii] Serganov A, et al. *Nature*, **441**: 1167-71 (2006)

FIGURE 2. Distinguishing characteristics of TPP riboswitches. (*Left*) Chemical structure of thiamin pyrophosphate, the natural ligand for TPP riboswitches. (*Center*) Consensus sequence and secondary structure model for TPP riboswitch aptamers. Note that the model is based on all types, including those that carry mutations in the P4–P5 region and that cannot substantially discriminate against thiamin, thiamin monophosphate, and TPP. Red, black, and gray letters represent nucleotides conserved in 97%, 90%, and 75% of the representatives, respectively, wherein R is a purine and Y is a pyrimidine. Similarly, circles represent any nucleotide, wherein its presence is reflected by the same color-coding, with open circles representing a nucleotide that is present in at least 50% of the representatives. Green, blue, and red shading of base pairs reflects strong evidence for covariation, compatible mutations, or no covarying mutations, respectively. P1 through P5 represent base-paired substructures. Number highlighted in black is the total collection of representatives for this riboswitch class in all three domains of life. (*Right*) Ribbon diagram representing an atomic-resolution structure of a representative TPP riboswitch aptamer bound to its ligand (space-filling model). (*Bottom*) The distribution of riboswitches among 36 divisions (phyla or orders) of bacteria. Circle size represents the relative number of riboswitch representatives per nucleotide of sequenced DNA as annotated elsewhere (Fig. 5). See Supplemental File 1 for a complete collection of similar data and imagery for all riboswitch classes.

thiamin, and the TPP analogs amprolium, benfotiamine, and pyrithiamine (Winkler et al. 2002b; Sudarsan et al. 2005; Edwards and Ferré-D'Amaré 2006; Serganov et al. 2006). This highlights the challenges faced by biochemists seeking to validate the identity of the natural ligand for some riboswitches.

In some instances, cells contain substantial amounts of close metabolite derivatives, which can confound efforts to identify the precise ligand that triggers gene regulation. Indeed, in a separate study (A Roth, PJ McCown, K Zhang, J Lee, RR Breaker, in prep.), we identified a large collection of TPP riboswitch variants that carry mutations in the pyrophosphate-binding core of the typical TPP aptamer. These variants appear to have adapted to sense thiamin (vitamin B_1) with an affinity similar to that of its natural phosphorylated derivatives (TMP and TPP). These findings, and similar observations with other riboswitch classes (Mandal and Breaker 2004; Kim et al. 2007; Johnson et al. 2012; Kellenberger et al. 2015; Nelson et al. 2015; Weinberg et al. 2017), reveal that evolutionary forces can blur the lines between riboswitch classes through the acquisition of mutations within a ligand-binding pocket that can alter binding specificity.

Several atomic-resolution structure models have been published that depict how the various conserved nucleotides and secondary structure features collaborate to form a ligand-binding pocket for TPP (Edwards and Ferré-D'Amaré 2006; Serganov et al. 2006; Thore et al. 2006). Comparisons between the consensus model and the atomic-resolution model are mutually supportive, such that regions of the RNA structure that are not important for ligand recognition or for forming the binding pocket are highly variable in sequence and structure, whereas regions that are critical for forming the riboswitch aptamer are highly conserved.

Our bioinformatics analysis has revealed the presence of nearly 16,700 TPP riboswitch representatives in DNA sequence databases (RefSeq version 56 and additional environmental DNA sequence databases; see Materials and Methods). These RNAs are distributed among species from nearly all bacterial lineages. In eukaryotes, the ligand-binding domain conforms to the previously established consensus model for TPP aptamers (Sudarsan et al. 2003a), whereas representatives that carry mutations in the P4–P5 region that reduce molecular discrimination are present in many bacteria (A Roth, PJ McCown, K Zhang, J Lee, RR Breaker, in prep.). There are several possible explanations for the presence of these aptamer variants. Perhaps these organisms do not selectively bind TPP, but rather read out the total pool of thiamin, thiamine monophosphate, and TPP, similarly to the manner in which ZTP riboswitches recognize differently phosphorylated forms (Kim et al. 2015). Another possibility is that the concentration of TPP within an organism is substantially higher than those of the other natural analogs, and therefore the ability to discriminate against all possible ligands is not necessary. This is also presumed to be the case for some 2′-dG-I (Kim et al. 2007), SAM-SAH (Weinberg et al. 2010), and c-di-GMP (Nelson et al. 2015) riboswitches. However, sometimes the TPP riboswitch variants occur in the same genome, which suggests that riboswitches gain in regulatory sophistication by exploiting ligand specificity changes.

RNA-derived compounds are the most prevalent ligands for the common riboswitch classes

We have previously speculated that there perhaps are many thousands of riboswitch classes that remain to be discovered (Ames and Breaker 2010; Breaker 2012) (see also the Discussion below). However, because common riboswitches are more likely to be discovered than the very rare, we likely already have in hand a near-complete collection of the 25 most abundant classes. Thus, even with this small sampling, we can begin to speculate with a reasonable expectation for accuracy about the nature and origin of riboswitches and the ligands they sense.

Most noteworthy is that an unusually large number of riboswitches sense ligands that are derived from RNA nucleotides or their precursors (Fig. 3). For example, five of the seven most common riboswitch classes selectively respond to the RNA-based coenzymes TPP (see above), B_{12} (adenosylcobalamin or AdoCbl) (Nahvi et al. 2002), SAM (sensed by two distinct classes in the top seven) (Epshtein et al. 2003; McDaniel et al. 2003; Winkler et al. 2003), and FMN (Fig. 1; Mironov et al. 2002; Winkler et al. 2002a). In total, more than a dozen classes or subclasses of riboswitches respond to nine nucleotide-like coenzymes. Intriguingly, these coenzymes are proposed to have been present during an evolutionary phase before proteins had emerged (Woese 1967;

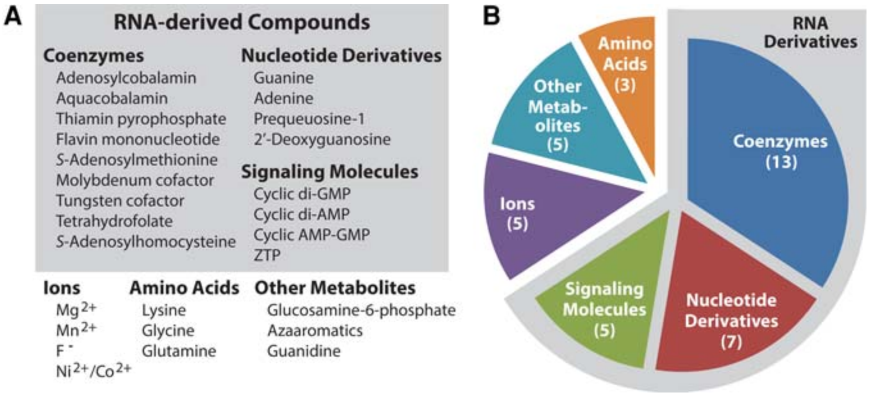


FIGURE 3. The ligands for known riboswitch classes. (*A*) List of the known riboswitch ligands grouped by ligand type. (*B*) Pie chart presenting the number of distinct riboswitch classes for the various ligand groups as presented in *A*. Note that the majority of riboswitches sense ligands derived from RNA monomers or their precursors.

White 1976; Benner et al. 1989). Perhaps riboswitch receptors for these coenzymes also have persisted through evolution, and represent modern versions of ancient RNA World regulatory RNAs (Nahvi et al. 2002; Breaker 2006, 2009). It has also been noted previously (Vitreschak et al. 2004) that the widespread nature and relative simplicity of riboswitches are consistent with their ancient origin.

Another striking trend is the abundance of RNA-based signaling molecules that are recognized by riboswitches. Five riboswitch classes are known that respond to c-di-GMP (two classes) (Sudarsan et al. 2008; Lee et al. 2010), c-di-AMP (Nelson et al. 2013), c-AMP-GMP (Kellenberger et al. 2015; Nelson et al. 2015), and ZTP (Kim et al. 2015). These observations again suggest that modern cells still carry traces of ancient RNA World signaling partnerships between riboswitches and small-molecule signals formed from ribonucleotide components (Breaker 2010; Nelson and Breaker 2017).

Perhaps in the future, additional molecular signals will be discovered by searching for the ligands of "orphan" riboswitches. Orphans exhibit characteristics typical of riboswitches, but their cognate ligands are not so readily elucidated. Notably, riboswitches for the signaling compounds c-di-AMP and c-AMP-GMP were discovered before these molecules were known to exist. Specifically, the *ydaO* orphan riboswitch class (Barrick et al. 2004) was reported several years before c-di-AMP was first discovered (Witte et al. 2008), followed thereafter by confirmation of c-di-AMP as the riboswitch ligand (Nelson et al. 2013). Similarly, c-di-GMP riboswitch variants (Sudarsan et al. 2008) were in hand, but remained unlinked to their natural ligand until after their ligand, c-AMP-GMP, had been discovered by other means (Davies et al. 2012). It seems very plausible that additional signaling molecules will be discovered by pairing certain orphan riboswitches to their natural ligands.

Five distinct riboswitch classes sense atomic ions

The vast majority of known riboswitch classes sense small-molecule metabolites. However, four different riboswitch classes have been discovered that sense divalent cations and another senses a monoanion. Moreover, given the importance of inorganic ions in biology, and given the inherent ability for polyanionic RNA polymers to interact with cationic ligands, we expect that the collection of riboswitch classes for metal ions will grow substantially.

Ion-binding riboswitches, particularly those that sense divalent metals, can pose challenges for those seeking to validate their functions. Most likely, all riboswitch classes will either form binding pockets for metal ions such as Mg^{2+}, or at least be influenced by nonspecific interactions with these cations via ionic interactions with the negative charges of the RNA phosphodiester backbone. To provide a convincing proof for inorganic-ion-responsive riboswitches, evidence for cation binding must be augmented with additional evidence for riboswitch function such as demonstrations of ligand-dependent gene control. Such strong lines of evidence have been reported for the five riboswitch classes described below.

Two classes of well-conserved aptamers bind Mg^{2+} and control genes involved in magnesium homeostasis. The most common of these is the Mg^{2+}-I riboswitch class (formerly called the *ykoK* motif), which uses an unusually large and elaborate aptamer domain (Barrick et al. 2004; Dann et al. 2007). Our current bioinformatics analyses revealed 971 representatives of Mg^{2+}-I riboswitches distributed among species from 12 disparate bacterial divisions (Supplemental File 1). Both the elaborate structure and the distribution of Mg^{2+}-I riboswitches are surprising, given that much simpler RNA structures that change their shape in response to Mg^{2+} binding should be common in sequence-space.

An important clue regarding the possible special utility of this riboswitch class is revealed by functional and biophysical studies. It has been demonstrated that Mg^{2+}-I riboswitches bind multiple Mg^{2+} ions in a highly cooperative fashion (Dann et al. 2007). Therefore, cells that carry this riboswitch class should be able to regulate gene expression more "digitally" (all on or all off) in response to very small changes in ligand concentration. Metabolite-binding riboswitches that exhibit similar cooperative ligand-binding characteristics have been reported for the ligands glycine (Mandal et al. 2004; Butler et al. 2011), THF (Trausch et al. 2011), c-di-AMP (Gao and Serganov 2014; Jones and Ferré-D'Amaré 2014; Ren and Patel 2014), and guanidine (Sherlock et al. 2017). For cells that carry Mg^{2+}-I riboswitches, a more digital genetic switch might be worth the extra cost of conserving and producing a more complex RNA architecture.

In contrast, a different Mg^{2+}-responsive riboswitch class (Cromie et al. 2006), herein called Mg^{2+}-II, is far simpler in architecture and requires far fewer conserved nucleotides. Surprisingly, our computational searches revealed only 101 representatives that are narrowly distributed in Proteobacteria. The vast majority of these riboswitches are associated with *mgtA* and *mgtE* genes, which code for Mg^{2+} transporters. It is not yet known whether members of this riboswitch class respond to the tight binding of only one Mg^{2+} ligand, or whether they can exhibit more complex functions. However, the structural characteristics suggest that members of this class might be functionally simpler than Mg^{2+}-I riboswitches.

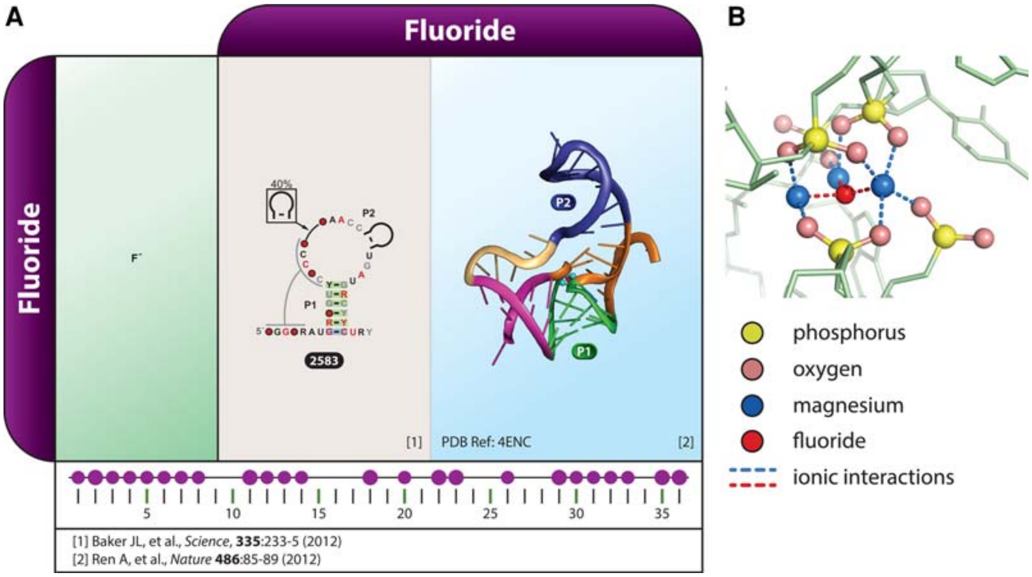
The selectivity of metal cation binding becomes important for riboswitches that respond to divalent metals that are far less abundant in cells compared to Mg^{2+}. At least one natural riboswitch class forms a highly specific and highly cooperative aptamer that responds fully only to Ni^{2+} and Co^{2+} ions. Members of the NiCo riboswitch class form an intricate binding pocket that selectively and cooperatively binds four Ni^{2+} or Co^{2+} ions, although Mn^{2+} also can be weakly bound noncooperatively (Furukawa et al. 2015). Ligand-binding selectivity is derived by the formation of both inner- and outer-sphere interactions largely involving N7 groups of guanine residues and 2′ oxygen atoms of ribose moieties. There is extensive sharing of nucleotides to form adjacent ligand-binding sites, which likely explains the source of cooperativity. Cooperative binding of Ni^{2+} or Co^{2+} might be important for cells to detect and respond to even small increases in the concentration of these ions, which can be toxic at even modest concentrations. Many of the genes associated with this riboswitch class are predicted to be heavy metal ion channels, and therefore this riboswitch class likely serves as a sensor of toxic concentrations of Ni^{2+} and/or Co^{2+}.

One of the longest-unresolved orphan riboswitch candidates was the *yybP* motif RNA (Barrick et al. 2004). After many years of uncertainty (Meyer et al. 2011), both genetic (Dambach et al. 2015) and structural (Price et al. 2015) data eventually confirmed the hypothesis (Waters et al. 2011) that members of this orphan riboswitch candidate naturally respond to Mn^{2+} ions. A total of 4383 representatives of this riboswitch class have been identified in species from 18 divisions of bacteria, indicating that this riboswitch and its natural ligand are broadly important for bacteria. Indeed, Mn^{2+} is very common in the environment and cells have had to evolve mechanisms to sense and respond to high concentrations of this divalent cation. The validation of *yybP* motif RNAs as Mn^{2+} riboswitches reveals a common sensor for this divalent cation and possible mechanisms by which cells overcome Mn^{2+} toxicity. However, it seems possible that rare variants of *yybP* motif RNAs could exist that have altered metal-binding specificities. If true, then members of this exceedingly widespread motif might help cells detect a greater diversity of ions on the periodic table.

Finally, one of the most remarkable riboswitch classes discovered to date selectively responds to fluoride. Initially, members of this class were reported as *crcB* orphan riboswitch candidates (Weinberg et al. 2010), but later were fortuitously found to function as fluoride-dependent gene control elements (Baker et al. 2012). Over 2500 fluoride riboswitches (Fig. 4A) were identified in 24 of the 36 divisions of bacteria. Moreover, members of this riboswitch class are also frequently present in species of archaea, suggesting that organisms have long exploited fluoride riboswitches to detect and respond to toxic levels of this anion. The most striking characteristic of members of this riboswitch class is that they can form a highly selective binding pocket for a negative point charge, despite carrying a negative phosphate group at every nucleotide in the RNA chain. An X-ray structure model of the binding site (Fig. 4B) wonderfully reveals how an RNA polyanion can form a tight binding pocket for fluoride (Ren et al. 2012). Notably, the RNA only makes use of aptamer phosphate groups to form a triangular Mg^{2+} cage, which selectively docks a single fluoride ion at its center. Perhaps other inorganic ions could be recognized by riboswitches that use similar strategies to form binding pockets for such unlikely ligands.

Riboswitches for amino acids and other metabolites

Despite the strong bias in favor of RNA World ligands, a few riboswitch classes sense and respond to metabolites that are not directly derived from RNA components. Lysine (Grundy et al. 2003; Sudarsan et al. 2003b), glycine (Mandal et al. 2004), and glutamine (Ames and Breaker 2011) are



[1] Baker JL, et al., *Science*, **335**:233-5 (2012)
[2] Ren A, et al., *Nature* **486**:85-89 (2012)

FIGURE 4. Fluoride riboswitches use their polyanionic RNA backbone to form a structure that selectively binds a fluoride anion. (*A*) Characteristics of fluoride riboswitches. Annotations are as described for Figure 2. (*B*) The atomic-resolution model for the ligand-binding site of a fluoride riboswitch. Fluoride (red) resides at the center of a Mg^{2+} triangle (blue), which is created by five negatively charged phosphates of the RNA aptamer backbone.

recognized by elaborate and relatively abundant riboswitch aptamers. Perhaps most puzzling is the fact that only three of the 20 common amino acids are directly sensed by riboswitches. Where are the riboswitch classes for the remaining 17 common amino acids?

One possible answer to this question is that, in Gram-positive bacteria, most amino acids are indirectly sensed by T box regulatory RNA elements (Green et al. 2010). T box RNA domains are present in the 5′ UTRs of many genes involved in amino acid synthesis and transport, as well as genes encoding aminoacyl-tRNA synthetases. Each T box RNA selectively base pairs with the anticodon loop of its cognate tRNA. The T box uses steric constraints to bind only tRNAs that lack an amino acid at the 3′ terminus. Binding of non-aminoacylated tRNAs activates gene expression, and thereby T box domains indirectly report on the amount of a specific amino acid present in the cell without ever directly binding the amino acid. Another possible answer is that attenuator systems that also indirectly sense amino acid deficiency are more common than had previously been known (S Stav, R Atilho, G Nguyen, G Mirihana-Arachichilage, RR Breaker, in prep.) Although T box regulatory elements and attenuator systems are widely used to sense amino acids in certain bacterial lineages, it seems very likely that novel riboswitch classes will be discovered that directly sense a greater variety of the common amino acids.

To date, only three other metabolites that are not directly derived from RNA components are known to be sensed by riboswitches. However, the riboswitch classes that sense these compounds all have noteworthy characteristics. The first such riboswitch detects the modified sugar compound glucosamine-6-phosphate (GlcN6P), which is selectively bound by RNA to regulate the expression of the *glmS* gene in many Gram-positive bacteria (Barrick et al. 2004; McCown et al. 2011). Importantly, members of this riboswitch class are called *glmS* ribozymes because they also function as self-cleaving RNAs (Winkler et al. 2004) that use the sugar ligand as a cofactor to promote RNA strand scission (Klein and Ferré-D'Amaré 2006; Cochrane et al. 2007; Ferré-D'Amaré 2010; Bingaman et al. 2017). GlcN6P-triggered RNA strand scission causes the associated mRNA to be rapidly degraded (Collins et al. 2007), thus repressing gene expression. This is the only metabolite-cofactor-dependent ribozyme discovered to date, although other riboswitch-triggered ribozymes have been reported (Lee et al. 2010).

Members of the second riboswitch class respond to a diverse collection of multi-ring molecules that can best be called "azaaromatic" ligands, and thus these RNAs are called azaaromatic riboswitches (Li et al. 2016). Within the genomic DNA database used for our study, 743 representatives of azaaromatic riboswitches were identified, which suggests that members of this class sense a natural compound that is of wide importance to bacteria. However, the riboswitch associates with only a single gene, called *yjdF*, whose protein product has an unknown function and therefore provides no clues regarding the riboswitch ligand. Moreover, the diversity of compounds that are bound by the RNA suggests that the RNA does not selectively bind a single natural ligand, but rather has adapted to sense a large diversity of azaaromatic or similar compounds. Indeed, when the *yjdF* gene is not associated with an azaaromatic riboswitch, it is commonly located immediately downstream from a gene for PadR proteins, whose functions are notable for their ability to broadly recognize planar multi-ring structures (Madoori et al. 2009). Perhaps azaaromatic riboswitches sense a broad range of such compounds as part of a toxic compound recognition and disposal system. If true, then azaaromatic riboswitches might be an early form of generalist ligand sensor, whereas PadR proteins might be more recent mimics of this versatile regulatory RNA.

The remaining non-RNA-derived ligand sensed by riboswitches is guanidine. Indeed, there are three distinct classes called guanidine-I, -II, and –III that selectively bind guanidine and regulate genes whose protein products overcome guanidine toxicity (Nelson et al. 2017; Sherlock and Breaker 2017; Sherlock et al. 2017). Each of these classes, originally called *ykkC* (Barrick et al. 2004), mini-*ykkC* (Weinberg et al. 2007), and *ykkC*-III (Weinberg et al. 2010) was considered an orphan riboswitch candidate, with the *ykkC* class resisting experimental validation for over a decade (Meyer et al. 2011). Guanidine-I riboswitches, for example, use a sophisticated RNA structure to form a guanidine-selective binding pocket that exploits all possible hydrogen-binding contacts and cation–π interactions with the ligand (Battaglia et al. 2017; Reiss et al. 2017). Guanidine-II riboswitches have perhaps the most striking functional characteristics given their exceedingly simple architecture. The two small and near-identical hairpins appear to form two guanidine binding pockets that function cooperatively (Sherlock et al. 2017). These subdomains represent the smallest natural ligand-binding domains known to date, suggesting that other tiny riboswitch aptamers might remain to be discovered.

Although the guanidyl moiety is present in fundamental metabolites such as guanine and arginine, free guanidine was not known to be a broadly important compound in bacteria until these riboswitches were experimentally validated. If riboswitch abundance and diversity are reasonable metrics for ligand relevance early in evolution, then free guanidine might have been a very important compound in the RNA World.

Orphan and other riboswitch candidates

In addition to the orphan riboswitches noted above whose ligands have been solved, other widespread orphan riboswitch candidates also have been experimentally validated. For example, representatives of the GEMM (Weinberg et al. 2007) and *pfl* (Weinberg et al. 2010) orphans have since been proven to respond, respectively, to c-di-GMP (Sudarsan et al. 2008; Kulshina et al. 2009; Smith et al. 2009) and ZTP (Jones and Ferré-D'Amaré 2015; Kim et al. 2015; Ren et al. 2015;

Trausch et al. 2015). A general theme is that the natural ligands for these long-standing riboswitch candidates remained mysterious because either the ligands, the biological processes regulated by the ligands, or both, were obscure or even entirely unknown. Given these outcomes, it is likely that other long-standing orphan riboswitch classes will be demonstrated to bind ligands that are broadly important for bacteria, but where the biochemical processes they regulate are generally underappreciated.

The ligands for nearly all of the first collection of orphan riboswitch classes (Barrick et al. 2004) have now been identified, although there are rarer variants of some of these RNAs that likely sense additional ligands and thus remain orphans (Nelson et al. 2017; Weinberg et al. 2017). We are unlikely to run out of challenging orphan riboswitch candidates anytime soon. Approximately 20 additional RNA motifs published six or more years ago exhibit characteristics strongly indicative of riboswitch function (E Greenlee, S Stav, K Perkins, ME Sherlock, KI Brewer, RR Breaker, in prep.). Furthermore, many additional orphan riboswitch candidates are present in a collection of over 200 novel ncRNA motifs (Z Weinberg, RR Breaker, in prep.).

For our current analysis, we have not included the characteristics of these orphan candidates because some might not prove to be riboswitches. Also, we chose not to further analyze four putative riboswitch classes previously proposed by others because their functions as novel riboswitches have not been convincingly validated. Specifically, a proposed riboswitch for aminoglycoside antibiotics (Jia et al. 2013) is most likely a previously studied DNA element exclusively associated with integrons (Roth and Breaker 2013). A proposed arginine riboswitch in the fungal species *Aspergillus nidulans* lacks demonstration of a saturable binding site or evidence for sequence and structural conservation (Borsuk et al. 2007). Similarly, a putative Mn^{2+} riboswitch lacks sufficient genetic or biochemical evidence of ligand-induced function (Shi et al. 2014). Finally, recently proposed novel riboswitches associated with homocysteine and homoserine metabolic genes (Leyn et al. 2014) are actually canonical members of the SAM-II riboswitch class (S Stav and RR Breaker, unpubl.).

For some variant riboswitches included here, analyses have not been detailed, or have provided only modest new information. For example, there are only a few different representatives of an FMN riboswitch variant originally called *CD3299* RNA (Blount et al. 2012; Blount 2013), which are primarily present in various strains of *Clostridium difficile* (Weinberg et al. 2017) and in select gut metagenomic samples that are nearly identical to *C. difficile* (RM Atilho, K Perkins, RR Breaker, in prep.). The CD3299 mRNA has been proposed to code for a riboflavin import protein (Gutiérrez-Preciado et al. 2015), although it is annotated as a putative multidrug transporter. The variant riboswitch RNAs are not associated with other known FMN metabolism genes, and they differ from the FMN riboswitch consensus sequence primarily at nucleotide positions involved in ligand binding. Although members of this FMN variant are rare, it is unlikely that the variants are simply defective FMN riboswitches that are being lost to evolution since many strains of *C. difficile* carry them. Recently, we have determined that these variants completely reject FMN or its most common natural derivatives riboflavin and FAD (RM Atilho, K Perkins, RR Breaker, in prep.). Rather, a representative was found to bind certain FMN degradation products or more distant chemical analogs that also trigger gene expression.

Indeed, it seems likely that many exceedingly rare riboswitch classes exist that have been derived recently through evolution, either by alterations to existing riboswitch representatives or by the emergence of novel motifs through natural structural exploration by mRNAs. In addition to the FMN riboswitch variant in *C. difficile*, other FMN riboswitch variants have been identified in a previous study (Pedrolli et al. 2012). One such representative appears to have evolved to functionally discriminate against the natural FMN analog called roseoflavin phosphate, a compound that exhibits antibacterial properties. Roseoflavin phosphate and its unphosphorylated precursor roseoflavin are known to bind tightly to several examples of FMN riboswitches (Lee et al. 2009; Ott et al. 2009; Serganov et al. 2009). This binding and subsequent suppression of genes coding for FMN biosynthesis might at least partly explain the source of the antibacterial activity of roseoflavin. However, organisms that produce roseoflavin as an antibacterial natural product must guard against self-poisoning. Variant FMN riboswitches that are not adversely affected even if roseoflavin binds could be one strategy whereby natural producers of roseoflavin avoid poisoning their own cells.

Despite the functional differences among some variant FMN riboswitches in recognizing FMN versus roseoflavin phosphate, in the current study we did not create distinct classes for these variants. Moreover, it is not yet clear how the variations in sequence and structure lead to differences in functional responses to the two ligands. Therefore, we could not computationally define and track the precise nucleotide changes that would enable the clear separation of FMN-like riboswitches into these two functional categories.

Phylogenetic distributions of riboswitches

Interesting patterns for riboswitch distribution and prevalence emerged on examination of the 38 general riboswitch classes investigated in this study (Fig. 5). Whereas some riboswitch classes are broadly distributed among the various bacterial divisions, others are only narrowly distributed. Moreover, some phyla are particularly enriched for the known riboswitch classes. In particular, the phylum Firmicutes (including the classes Bacilli and Clostridia) have the most riboswitch representatives with 13,577 (Table 1). The phylum Proteobacteria ranks second with 10,751 riboswitch representatives. However, it is important to note that there is a potential for some bias in these distributions because sequenced

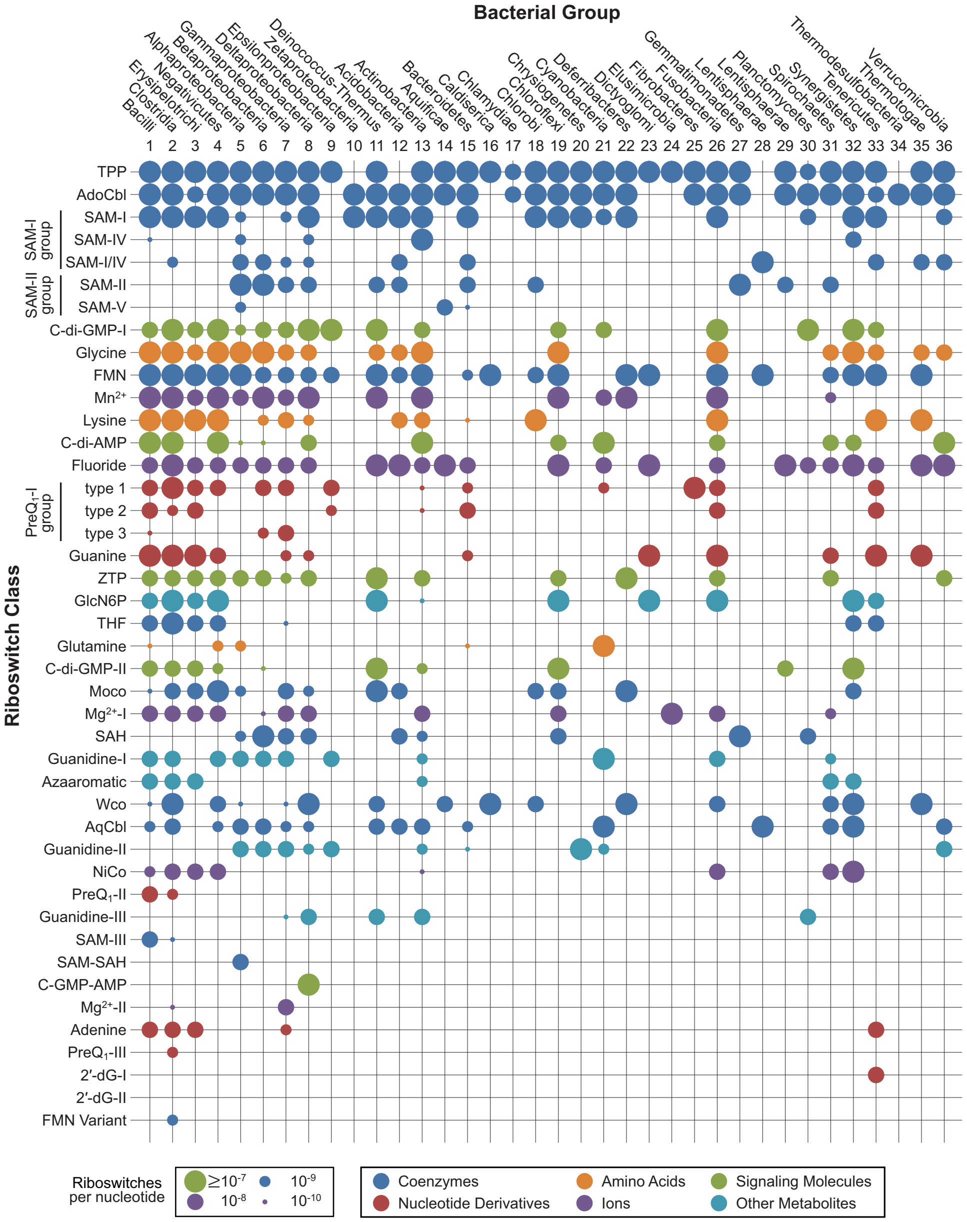


FIGURE 5. Grid depicting the presence and abundance of experimentally validated riboswitch classes. Abbreviations are as described for Figure 1 and data point sizes are as described for Figure 2. There are 38 general riboswitch classes, which are those that either bind a distinct ligand or use a distinct architecture to bind a ligand that is already known to be bound by another class. For example, guanine and adenine riboswitches are distinct or "general" classes because they bind distinct ligands, even though they exploit near-identical aptamer architectures. The type 1, 2, and 3 RNAs of the preQ$_1$-I riboswitch class use only modestly different aptamer sequences and structures to bind the ligand preQ$_1$, and thus constitute members of a single general class. Likewise, the RNAs called SAM-I (or S box), -IV, and -I/-IV are all close structural types of the general SAM-I class, and the RNAs originally called SAM-II and -V are types of the general SAM-II class. Notes: All known representatives of the 2′-dG-II riboswitch class are derived from metagenomic data, and there is insufficient DNA sequence information in these sequencing reads to confidently assign the hits to bacterial divisions. Caution should be used in interpreting rare instances of a riboswitch class in certain bacterial lineages, because some bioinformatics hits could be false positives.

genomes in the databases analyzed include an overrepresentation of species from Firmicutes and Proteobacteria (Sentausa and Fournier 2013). Regardless, organisms from these two bacterial divisions have a higher density (riboswitch representatives per nucleotide sequenced), suggesting that they indeed are enriched in known riboswitches (Fig. 5).

Riboswitches are rarely identified by computationally searching through bacteriophage genomes, and therefore we do not include phages in our graphical depictions. Previously, only two examples of c-di-GMP-I riboswitches were found in prophages of *C. difficile* (Sudarsan et al. 2008). Through our current bioinformatics search efforts,

TABLE 1. Distribution of all riboswitches per phylum

Lineage	Hits	Lineage (cont.)	Hits
Environmental	69345	Euryarchaeota	79
Firmicutes	13577	Planctomycetes	66
Bacilli	(7598)	Acidobacteria	58
Clostridia	(5221)	Verrucomicrobia	45
Negativicutes	(557)	Deferribacteres	30
Erysipelotrichi	(201)	Ascomycota	23
Proteobacteria	10,751	Chrysiogenetes	15
γ-proteobacteria	(4869)	Nitrospirae	14
α-proteobacteria	(2765)	Aquificae	14
β-proteobacteria	(2096)	Dictyoglomi	11
δ-proteobacteria	(914)	Chlamydiae	8
ε-proteobacteria	(105)	Fibrobacteres	7
ζ-proteobacteria	(2)	Gemmatimonadetes	7
Actinobacteria	4572	Crenarchaeota	6
Bacteroidetes	1256	Caldiserica	4
Fusobacteria	435	Thermodesulfobacteria	4
Cyanobacteria	357	Lentisphaerae	3
Deinococcus-Thermus	248	Elusimicrobia	2
Chloroflexi	229	Basidiomycota	2
Spirochaetes	197	Viridiplantae	2
Synergistetes	162	Class I Virus Family	2
Tenericutes	111	Korarchaeota	1
Chlorobi	92	Thaumarchaeota	1
Thermotogae	87		

Bacilli, Clostridia, Erysipelotrichi, and Negativicutes are members of the phylum Firmicutes. All the phylum names that have a Greek letter followed by the suffix -proteobacteria (e.g., Δ-proteobacteria) are members of the phylum Proteobacteria. Numbers of hits in parentheses are included in the totals for the relevant phyla.

we only identified two glutamine riboswitch aptamers encoded by two different cyanophage genomes. These findings strongly suggest that bacteriophages rarely use metabolite- or inorganic ion-binding riboswitches to control gene expression. However, we cannot rule out the possibility that bacteriophages exploit riboswitches more extensively. Perhaps they favor using classes that have yet to be discovered, or sense ligands that are far rarer than those that trigger the common riboswitch classes.

Similarly, our findings also reveal that species in the Archaeal domain of life sparingly use very few of the known riboswitch classes. The Archaeal phylum Euryarchaeota possess the most riboswitches with 79, while the phylum Crenarchaeota possesses six, and the phyla Korarchaeota and Thaumarchaeota each possess only one. Members of two riboswitch classes, TPP (Rodionov et al. 2002) and fluoride (Weinberg et al. 2010; Baker et al. 2012), are most common and were the only riboswitch classes previously known to be present in this domain of life (Sun et al. 2013). However, our study revealed that examples of FMN, Mg^{2+}-I, guanidine-II riboswitch classes are also occasionally present.

TPP riboswitches are also present in some eukaryotic species, primarily among fungi (72 examples), plants (23), and algae (four). Curiously, there are seven TPP riboswitches in a single protist species, *Perkinsus marinus*. Also, only one example was found in the metazoan *Hydra magnipapillata*, but the animal kingdom otherwise appears to be devoid of TPP riboswitches. Almost without exception, eukaryotes were found to be devoid of other riboswitch classes. A single representative of a SAM-I riboswitch has been identified in the protozoan *Trichomonas vaginalis* (RR Breaker, unpubl.). However, DNA sequences flanking this riboswitch appear to be of bacterial origin, which suggests that its presence is due to a more recent horizontal transfer of genetic information or possibly contamination during the DNA sequencing process. We found one example of a fluoride riboswitch in HSR1 RNA, which had been proposed to be a human ncRNA (Shamovsky et al. 2006). However, this RNA has proven to be a bacterial sequence that likely contaminated RNA samples isolated from human cells (Kim et al. 2010; Choi et al. 2015). We did not detect representatives of the other validated riboswitch classes from bacteria in any eukaryotic species, suggesting that these riboswitch classes might be only rarely (if at all) used by this domain of life.

It has been proposed that riboswitches might serve as targets for the development of novel antibacterial agents (Blount and Breaker 2006; Deigan and Ferré-D'Amaré 2011). Indeed, efforts to create novel compounds that bind riboswitches have yielded antibacterial efficacy (Kim et al. 2009; Mulhbacher et al. 2010; Blount et al. 2015; Howe et al. 2015), although some of these compounds might use broader mechanisms to inhibit bacterial growth (Kim et al. 2009; Kofoed et al. 2016). The distribution of riboswitches among various divisions of bacteria (Fig. 5) can be used to identify classes that might be most amenable to the development of broad-spectrum antibiotics. For example, TPP, AdoCbl, SAM-I, and FMN riboswitch classes might yield antibacterial compounds that target the widest number of pathogens. Presumably, drug binding to these riboswitches would be detrimental to bacteria because this should suppress the production and import of these essential coenzymes.

In contrast, most other classes are more narrowly distributed, and some of these classes would not serve as good drug targets for other reasons. However, some of the rarer riboswitch classes that regulate critical genes might be useful targets for compounds that serve as more focused or narrow-spectrum antibiotics. Given the importance of diverse microbiomes for human health, it might be best in some cases to use precision inhibitors of specific pathogenic species, rather than using a broad-spectrum antibiotic.

How many riboswitches remain to be discovered?

The rank order of riboswitch classes based on prevalence (Fig. 1) reveals that riboswitch classes with relatively few representatives predominate over classes with large numbers of representatives. These data, like many other natural phenomena, appear to follow a power law (Newman 2004) distribution, wherein one number is related to the fixed power of another. Specifically, the numbers Y and X are related by the equation $Y = mX^b$. It has been previously proposed that the abundance of representatives for riboswitch classes also have this characteristic (Ames and Breaker 2010), wherein Y is the number of representatives for a given riboswitch class, X is the rank order of the riboswitch class based on its abundance, m is the number of representatives for the most abundant riboswitch class, and b is the exponent (slope of the resulting line of the data when depicted on a log–log plot).

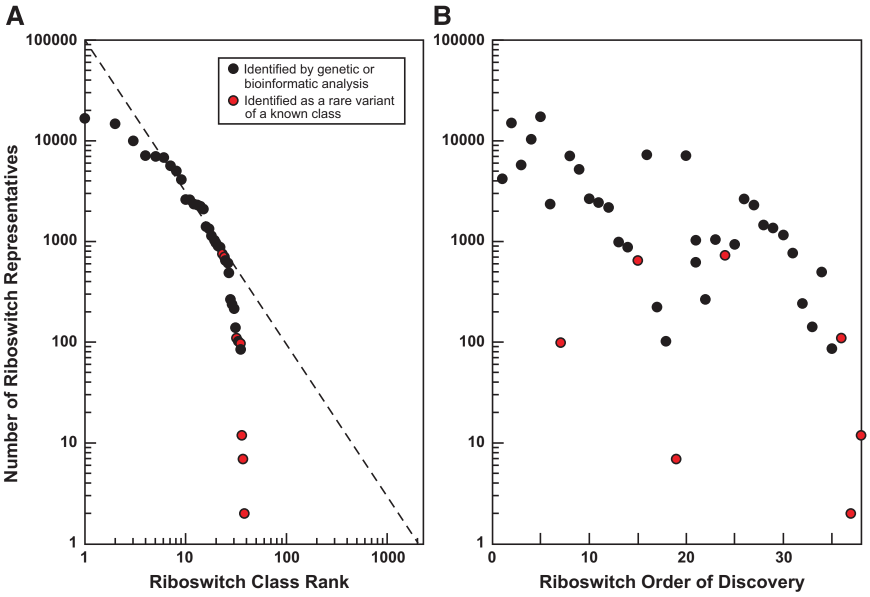


FIGURE 6. Evidence that riboswitch classes conform to a power law distribution. (*A*) A log–log plot depicting the validated riboswitch classes ranked (*x*-axis, X) in order of their abundance (*y*-axis, Y). The dashed line represents an ideal power law distribution from the equation $Y = mX^b$ where m and b were set to 100,000 and −1.6, respectively, to yield a line that best tracks the data. (*B*) Plot of the abundance of each riboswitch class on a log scale versus the order of discovery. Red points represent classes discovered as rare variants of more common classes.

Such a distribution in riboswitch abundance seems reasonable because the evolutionary forces that influence riboswitch representative numbers should play out similarly at all scales. These forces, which determine whether a riboswitch class is abundant or rare, include such things as the date of evolutionary emergence of a riboswitch class, its utility to the host cell, its size and information content, its evolutionary malleability, and a variety of other characteristics. Through evolution, some riboswitches will become unusually abundant, particularly if they emerged very early in evolution and serve a fundamental sensory and regulatory role that persists in nearly all modern species. In contrast, some riboswitch classes will be exceptionally rare, primarily due to their more recent evolutionary appearance, the lack of need for their functions in most modern species, and/or their displacement by competing genetic factors such as proteins. These forces will likely have been present throughout the eons of evolution, yielding many riboswitch classes with abundances spanning many orders of magnitude.

Data that conform to a simple power law relationship should yield a straight line on a log–log plot. At first glance, a log–log plot of the riboswitch class rank order versus the number of representatives for each class identified in the DNA sequence databases used in this study appears to be a poor fit to a power law equation (Fig. 6A). However, there are two simple effects that can explain the data points varying from the expected distribution. First, the most abundant riboswitch classes encounter a unique limitation that the vast majority of riboswitch classes do not encounter. The power law simulation used to establish the line indicates that the most common riboswitch class should have ∼100,000 representatives, and the second-most common class should have ∼33,000. However, the most common riboswitch class in bacteria, which senses TPP, has only 16,701 members, and the second-most common class, which senses AdoCbl, has only 14,336. To attain the predicted values, there would need to be nearly sixfold and 2.3-fold more TPP and AdoCbl riboswitches, respectively, in each average host cell than are observed. These numbers are entirely unnecessary for cells to manage the metabolic processes needed to maintain the availability of these coenzymes. As a result, the abundances of the most common riboswitch classes are expected to fall short of those predicted by the power law. Such deviations are sometimes observed for the distributions of other natural phenomena as well (Newman 2004).

Second, the abundances of the rarer riboswitch classes fall precipitously below those predicted by the simulated power law distribution. However, this effect is easily explained by the certainty that we have not discovered all riboswitch classes in the data set analyzed in this study. Specifically, rarer riboswitches are more difficult to discover by whatever method has been used to discover riboswitch classes to date. For example, rare riboswitches are less likely to be uncovered by comparative sequence analysis approaches because multiple representatives are necessary to make convincing consensus models and establish genome contexts—two important requirements for defining strong riboswitch candidates. Rarity also reduces the probability that a geneticist will encounter a riboswitch-mediated gene control process through mutational screens or other genetic analyses. Thus, we have a more complete collection of the common riboswitch classes, and a progressively incomplete collection of the ever-rarer classes. We expect that, as rarer riboswitch classes are discovered, the new data points on riboswitch class abundance will more closely follow the power law projection.

A plot of the order of discovery for novel riboswitch classes versus abundance reveals that this expected correlation indeed exists (Fig. 6B). Therefore, we do not expect to see the discovery of additional riboswitches discovered that rank among the top 25 classes, at least within the current DNA sequence data sets examined. More importantly, if the power law distribution as depicted roughly holds, then there are likely to be hundreds of riboswitch classes that remain undiscovered among the sequence databases we have analyzed, and many thousands of classes among the many bacterial species on the planet that have yet to be studied (Stewart 2012). Of course, the vast majority of these undiscovered riboswitch classes will be exceedingly rare, but their discovery will be progressively aided by searching through the increasing amounts of novel DNA sequences being made available. Each new riboswitch find has the potential to teach us more about how RNA folds to selectively bind its target ligand, and much more about how cells regulate the diverse biological processes they use.

Also, by using the power law relationship, we can estimate the total number of individual riboswitch representatives remaining to be discovered. Surprisingly, ~85% of all the intergenic regions (IGRs) predicted by the power law relationship to carry riboswitches have already been discovered (left portion of Fig. 6A), and they mostly carry riboswitches from the common classes (Fig. 7). This means that only 15% of the total number of IGRs predicted to carry riboswitches includes several hundred novel riboswitch classes remaining to be discovered (right portion of Fig. 6A). Given the trends seen with past discoveries and the diminishing numbers of IGRs carrying novel riboswitch classes, novel riboswitch classes should become progressively more difficult to discover.

Concluding remarks

The current collection of riboswitch classes represents a structurally and functionally rich array of ligand-binding gene control elements built entirely of RNA. However, the diversity of riboswitch classes that exist in modern cells is likely to be far greater. If the current distribution trend (Fig. 6A) holds, there could be more than 1000 riboswitch classes remaining undiscovered in the current genomic sequence databases. Unfortunately, since most of these projected classes are expected to be exceedingly rare, they will likely remain undiscovered for many years to come. Fortunately, a variety of methods exist that could be used to identify additional classes, particularly those that are more numerously represented. For example, targeted analysis of IGRs for genes whose regulation is of interest can yield evidence for novel riboswitch classes (Cromie et al. 2006; Fuchs et al. 2006). Also, bioinformatics approaches that are designed to identify structured noncoding RNAs will likely remain a productive route to discovering even quite rare riboswitches (e.g., Meyer et al. 2009; Weinberg et al. 2010; Z Weinberg, RR Breaker, in prep.; S Stav, R Atilho, G Nguyen, G Mirihana-Arachichilage, RR Breaker, in prep.).

This brings up an important question—is it worth the necessarily increasing effort to find rare riboswitch classes? So far, almost every new riboswitch class has revealed interesting structural and/or functional features that expand our understanding of the capabilities of RNA as a medium for forming molecular sensors and switches. Furthermore, the physical association between each riboswitch and the protein coding region whose expression it regulates exposes a link between its ligand and the likely functions of various transporters and enzymes. This effect has been made abundantly clear with the reports of riboswitches for fluoride (Baker et al. 2012) and guanidine (Nelson et al. 2017), which revealed the existence of widespread mitigation systems for these toxic ligands, including specialized transporters. Given that there are likely a large number of riboswitches yet to be discovered, there are also many more opportunities to explore the functional diversity of RNA and gain insight into biological pathways and processes that are currently underappreciated or are entirely unknown.

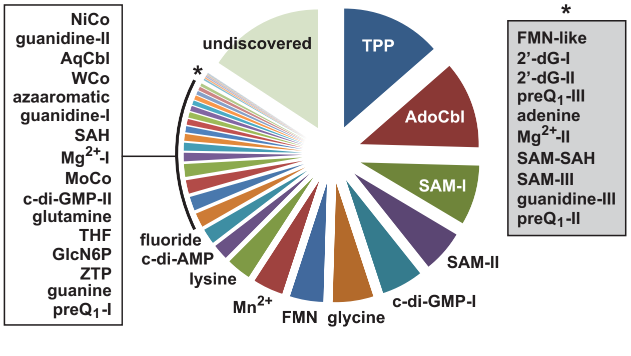


FIGURE 7. Distribution of IGRs carrying riboswitches. Abbreviations are as described for Figure 1, wherein the boxed riboswitch names are listed from *top* to *bottom* as the least common to most common, respectively. The size of the plot segment labeled "undiscovered" was established by summing the numbers of riboswitch classes remaining to be discovered as presented in Figure 6A and as predicted by the power law distribution wherein $m = 100{,}000$ and $b = -1.6$.

MATERIALS AND METHODS

RNA bioinformatics analyses

Bioinformatics updates of the consensus models and phylogenetic distributions for all riboswitch classes began with homology searches as previously described (McCown et al. 2011, 2014; Ruff et al. 2016). Homology searches were conducted with INFERNAL 1.1 using the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) Database release 56 (Pruitt et al. 2009; O'Leary et al. 2016) augmented with multiple environmental DNA sequence data sets (Nawrocki and Eddy 2013; Nelson et al. 2013; McCown et al. 2014).

To aggressively search for distal variants belonging to the known riboswitch classes, an *E*-value for possible representatives of as high as 10,000 were considered. However, the total number of representatives reported for each class was derived by filtering out likely false positives as described below. To eliminate false positives from the candidate list, we removed hits that (i) were found within protein coding regions, (ii) were found within previously defined ncRNA motifs, (iii) were not able to adopt the major secondary-structure features of the class, (iv) contained regions of predominantly homopolymer character, or (v) were low-ranking hits that lacked a reasonable genomic context (e.g., wrong orientation for riboswitch function or unreasonable gene association).

Candidates that included only a portion of the riboswitch sequence, but consisting of at least 50% sequence or structure identity with the consensus model, were accepted for the following riboswitch classes: TPP, AdoCbl/AqCbl, SAM-I, SAM-III, SAM-IV, SAM-I/IV, SAH, c-di-GMP-I, c-di-GMP-II, c-di-AMP, c-AMP-GMP, preQ$_1$-II, preQ$_1$-III, guanine, adenine, 2′-dG-I, 2′-dG-II, lysine, glycine, glutamine, Mg^{2+}-I, Mg^{2+}-II, ZMP, F$^-$, *glmS*, FMN, THF, NiCo, Moco/Wco, azaaromatic, Mn^{2+}, guanidine-I, and guanidine-III. Truncated sequences were not included as representatives of the following riboswitch classes due to their small sizes and simple structures: SAM-II, SAM-V, SAM-SAH, preQ$_1$-I, and guanidine-II. All consensus sequence and structure diagrams were generated with the R2R program (Weinberg and Breaker 2011).

Phylogenetic distributions

The representatives for each riboswitch class were mapped onto a phylogenetic tree using phyla based on the Interactive Tree of Life (Letunic and Bork 2011) and on the NIH Taxonomy website (http://www.ncbi.nlm.nih.gov/taxonomy).

Structural models

Atomic-resolution structure models were generated with PyMOL (Schrödinger, LLC) using PDB files from the RCSB protein databank (www.rcsb.org/pdb) (Berman et al. 2000). PNG graphic files were exported and slightly modified for appearance before inserting into the graphics program Adobe Illustrator.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank all members of the Breaker laboratory for helpful discussions, Z. Weinberg for advice on computational searches including the data on eukaryotic TPP riboswitches, and N. Carriero and R. Bjornson for assistance with the Yale Life Sciences High Performance Computing Center funded by the National Institutes of Health (RR19895). We also thank M. Tortora (Fairfield University) for assistance in figure generation. P.J.M. was supported in part by the National Institutes of Health grant T32GM007499. M. E.S. was supported by a National Institutes of Health Cellular and Molecular Biology Training Grant (T32FM007223). This work was also supported by a National Institutes of Health grant (GM022778) to R.R.B. and by the Howard Hughes Medical Institute.

Author contributions: P.J.M. and R.R.B. designed the study, and P. J.M., M.E.S., and S.S. conducted the computational analyses. P.J.M., K.A.C., and R.R.B. created the final graphics. R.R.B. wrote the manuscript, and all authors contributed to editing.

Received February 23, 2017; accepted April 4, 2017.

REFERENCES

Ames TD, Breaker RR. 2010. Bacterial riboswitch discovery and analysis. In *The chemical biology of nucleic acids* (ed. Mayer G). Wiley, Chichester, UK.

Ames TD, Breaker RR. 2011. Bacterial aptamers that selectively bind glutamine. *RNA Biol* **8:** 82–89.

André G, Even S, Putzer H, Burguière P, Croux C, Danchin A, Martin-Verstraete I, Soutourina O. 2008. S-box and T-box riboswitches and antisense RNA control a sulfur metabolic operon of *Clostridium acetobutylicum*. *Nucleic Acids Res* **36:** 5955–5969.

Baker JL, Sudarsan N, Weinberg Z, Roth A, Stockbridge RB, Breaker RR. 2012. Widespread genetic switches and toxicity resistance proteins for fluoride. *Science* **335:** 233–235.

Barrick JE, Breaker RR. 2007. The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biol* **8:** R239.

Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, et al. 2004. New riboswitch motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc Natl Acad Sci* **101:** 6421–6426.

Battaglia RA, Price IR, Ke A. 2017. Structural basis for guanidine sensing by the ykkC family of riboswitches. *RNA* **23:** 578–585.

Benner SA, Ellington AD, Tauer A. 1989. Modern metabolism as a palimpsest of the RNA world. *Proc Natl Acad Sci* **86:** 7054–7058.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* **28:** 235–242.

Bingaman JL, Zhang S, Stevens DR, Yennawar NH, Hammes-Schiffer S, Bevilacqua PC. 2017. The GlcN6P cofactor plays multiple catalytic roles in the *glmS* ribozyme. *Nat Chem Biol* **13:** 439–445.

Blount KF. 2013. *Methods for treating or inhibiting infection by* Clostridium difficile. U.S. patent appl. no. 13/576,989.

Blount KF, Breaker RR. 2006. Riboswitches as antibacterial drug targets. *Nat Biotechnol* **24:** 1558–1564.

Blount KF, Coish PDG, Dixon BR, Myung J, Osterman D, Wickens P, Avola S, Baboulas N, Bello A, Berman J, et al. 2012. *Flavin derivatives*. U.S. patent appl. no. 13/381,809.

Blount KF, Megyola C, Plummer M, Osterman D, O'Connell T, Aristoff P, Quinn C, Chrusciel RA, Poel TJ, Schostarez HJ, et al. 2015. Novel riboswitch-binding flavin analog that protects mice against Clostridium difficile infection without inhibiting *cecal flora*. *Antimicrob Agents Chemother* **59:** 5736–5746.

Bocobza SE, Aharoni A. 2014. Small molecules that interact with RNA: riboswitch-based gene control and its involvement in metabolic regulation in plants and algae. *Plant J* **79:** 693–703.

Borsuk P, Przykorska A, Blachnio K, Koper M, Pawlowicz JM, Pekala M, Weglenski P. 2007. L-arginine influences the structure and function of arginase mRNA in *Aspergillus nidulans*. *Biol Chem* **388:** 135–144.

Breaker RR. 2006. Riboswitches and the RNA World. In *The RNA World*, 3rd ed. (ed. Gesteland RF, Cech TR, Atkins JF), pp. 89–107. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Breaker RR. 2009. Riboswitches: from ancient gene-control systems to modern drug targets. *Future Microbiol* **4:** 771–773.

Breaker RR. 2010. RNA second messengers and riboswitches: relics from the RNA World? *Microbe* **5:** 13–20.

Breaker RR. 2011. Prospects for riboswitch discovery and analysis. *Mol Cell* **43:** 867–879.

Breaker RR. 2012. Riboswitches and the RNA World. *Cold Spring Harb Perspect Biol* **4:** a003566.

Butler EB, Xiong Y, Wang J, Strobel SA. 2011. Structural basis of cooperative ligand binding by the glycine riboswitch. *Chem Biol* **25:** 293–298.

Cheah MT, Wachter A, Sudarsan N, Breaker RR. 2007. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature* **447:** 497–500.

Choi D, Oh HJ, Goh CJ, Lee K, Hahn Y. 2015. Heat shock RNA1, known as a eukaryoic temperature-sensing noncoding RNA, is of bacterial origin. *J Microbiol Biotechnol* **25:** 1234–1240.

Cochrane JC, Lipchock SV, Strobel SA. 2007. Structural investigation of the GlmS ribozyme bound to its catalytic cofactor. *Chem Biol* **14:** 97–105.

Collins JA, Irnov I, Baker S, Winkler WC. 2007. Mechanism of mRNA destabilization by the *glmS* ribozyme. *Genes Dev* **21:** 3356–3368.

Croft MT, Moulin M, Webb ME, Smith AG. 2007. Thiamine biosynthesis in algae is regulated by riboswitches. *Proc Natl Acad Sci* **104:** 20770–20775.

Cromie MJ, Shi Y, Latifi T, Groisman EA. 2006. An RNA sensor for intracellular Mg^{2+}. *Cell* **125:** 71–84.

Dambach M, Sandoval M, Updegrove TB, Anantharaman V, Aravind L, Waters LS, Storz G. 2015. The ubiquitous *yybP-ykoY* riboswitch is a manganese-responsive regulatory element. *Mol Cell* **57:** 1099–1109.

Dann CE III, Wakeman CA, Sieling CL, Baker SC, Irnov I, Winkler WC. 2007. Structure and mechanism of a metal-sensing regulatory RNA. *Cell* **130:** 878–892.

Davies BW, Bogard RW, Young TS, Mekalanos JJ. 2012. Coordinated regulation of accessory genetic elements produces cyclic di-nucleotides for *V. cholerae* virulence. *Cell* **149:** 358–370.

Deigan KE, Ferré-D'Amaré AR. 2011. Riboswitches: discovery of drugs that target bacteria gene-regulatory RNAs. *Acc Chem Res* **44:** 1329–1338.

Edwards TE, Ferré-D'Amaré AR. 2006. Crystal structures of the *thi*-box riboswitch bound to thiamine pyrophosphate analogs reveal adaptive RNA-small molecule recognition. *Structure* **14:** 1459–1468.

Epshtein V, Mironov AS, Nudler E. 2003. The riboswitch-mediated control of sulfur metabolism in bacteria. *Proc Natl Acad Sci USA* **100:** 5052–5056.

Ferré-D'Amaré AR. 2010. The *glmS* ribozyme: use of a small molecule coenzyme by a gene-regulatory RNA. *Q Rev Biophys* **43:** 423–447.

Fuchs RT, Grundy FJ, Henkin TM. 2006. The S_{MK} box is a new SAM-binding RNA for translational regulation of SAM synthase. *Nat Struct Mol Biol* **13:** 226–233.

Furukawa K, Ramesh A, Zhou Z, Weinberg Z, Vallery T, Winkler WC, Breaker RR. 2015. Bacterial riboswitches cooperatively bind Ni^{2+} or Co^{2+} ions and control expression of heavy metal transporters. *Mol Cell* **57:** 1088–1098.

Gao A, Serganov A. 2014. Structural insights into recognition of c-di-AMP by the *ydaO* riboswitch. *Nat Chem Biol* **10:** 787–792.

Garst AD, Edwards AL, Batey RT. 2011. Riboswitches: structures and mechanisms. *Cold Spring Harb Perspect Biol* **3:** a003533.

Gelfand MS, Mironov AA, Jomantas J, Kozlov YI, Perumov DA. 1999. A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. *Trends Genet* **15:** 439–444.

Green NJ, Grundy FJ, Henkin TM. 2010. The T box mechanism: tRNA as a regulatory molecule. *FEBS Lett* **584:** 318–324.

Griffiths-Jones S. 2005. RALEE—RNA ALignment editor in Emacs. *Bioinformatics* **21:** 257–259.

Grundy FJ, Henkin TM. 1998. The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria. *Mol Microbiol* **30:** 737–749.

Grundy FJ, Lehman SC, Henkin TM. 2003. The L box regulon: lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. *Proc Natl Acad Sci* **100:** 12057–12062.

Gutiérrez-Preciado A, Torres AG, Merino E, Bonomi HR, Goldbaum FA, Garcia-Angulo VA. 2015. Extensive identification of bacterial riboflavin transporters and their distribution across bacterial species. *PLoS One* **10:** e0126124.

Henkin TM. 2008. Riboswitch RNAs: using RNA to sense cellular metabolism. *Genes Dev* **22:** 3383–3390.

Hollands K, Proshkin S, Sklyarova S, Epshtein V, Mironov A, Nudler E, Groisman EA. 2012. Riboswitch control of Rho-dependent transcription termination. *Proc Natl Acad Sci* **109:** 5376–5381.

Howe JA, Wang H, Fischmann TO, Balibar CJ, Xiao L, Galgoci AM, Malinverni JC, Mayhood T, Villafania A, Nahvi A, et al. 2015. Selective small-molecule inhibition of an RNA structural element. *Nature* **526:** 672–677.

Jia X, Zhang J, Sun W, He W, Jiang H, Chen D, Murchie AI. 2013. Riboswitch control of aminoglycoside antibiotic resistance. *Cell* **152:** 68–81.

Johnson JE Jr, Reyes FE, Polaski JT, Batey RT. 2012. B_{12} cofactors directly stabilize an mRNA regulatory switch. *Nature* **492:** 133–137.

Jones CP, Ferré-D'Amaré AR. 2014. Crystal structure of a c-di-AMP riboswitch reveals an internally pseudo-dimeric RNA. *EMBO J* **33:** 2692–2703.

Jones CP, Ferré-D'Amaré AR. 2015. Recognition of the bacterial alarmone ZMP through long-distance association of two RNA subdomains. *Nat Struct Mol Biol* **22:** 679–685.

Kellenberger KA, Wilson SC, Hickey SF, Gonzalez TL, Su Y, Hallberg ZF, Brewer TF, Iavarone AT, Carlson HK, Hsieh YF, et al. 2015. GEMM-I riboswitches from Geobacter sense the bacterial second messenger cyclic AMP-GMP. *Proc Natl Acad Sci* **112:** 5383–5388.

Kim JN, Roth A, Breaker RR. 2007. Guanine riboswitch variants from *Mesoplasma florum* selectively recognize 2′-deoxyguanosine. *Proc Natl Acad Sci* **104:** 16092–16097.

Kim JN, Blount KF, Puskarz I, Lim J, Link KH, Breaker RR. 2009. Design and antimicrobial action of purine analogues that bind Guanine riboswitches. *ACS Chem Biol* **4:** 915–927.

Kim DS, Lee Y, Hahn Y. 2010. Evidence for bacterial origin of heat shock RNA-1. *RNA* **16:** 274–279.

Kim PB, Nelson JW, Breaker RR. 2015. An ancient riboswitch class in bacteria regulates purine biosynthesis and one-carbon metabolism. *Mol Cell* **57:** 317–328.

Klein DJ, Ferré-D'Amaré AR. 2006. Structural basis of *glmS* ribozyme activation by glucosamine-6-phosphate. *Science* **313:** 1752–1756.

Kofoed EM, Yan D, Katakam AK, Reichelt M, Lin B, Kim J, Park S, Date SV, Monk IR, Xu M, et al. 2016. De novo guanine biosynthesis but not riboswitch-regulated purine salvage pathway is required for *Staphylococcuss aureus* infection in vivo. *J Bacteriol* **198:** 2001–2015.

Kubodera T, Watanabe M, Yoshiuchi K, Yamashita N, Nishimura A, Nakai S, Gomi K, Hanamoto H. 2003. Thiamine-regulated gene expression of *Aspergilus oryzae thiA* requires splicing of the intron containing a riboswitch-like domain in the 5′ UTP. *FEBS Lett* **555:** 516–520.

Kulshina N, Baird NJ, Ferré-D'Amaré AR. 2009. Recognition of the bacterial second messenger cyclic diguanylate by its cognate riboswitch. *Nat Struct Mol Biol* **16:** 1212–1217.

Lee ER, Blount KF, Breaker RR. 2009. Roseoflavin is a natural antibacterial compound that binds to FMn riboswitches and regulates gene expression. *RNA Biol* **6:** 187–194.

Lee ER, Baker JL, Weinberg Z, Sudarsan N, Breaker RR. 2010. An allosteric self-splicing ribozyme triggered by a bacterial second messenger. *Science* **329:** 845–848.

Letunic I, Bork P. 2011. Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39:** W475–W478.

Leyn SA, Suvorova IA, Kholina TD, Sherstneva SS, Novichkov PS, Gelfand MS, Rodionov DA. 2014. Comparative genomics of transcriptional regulation of methionine metabolism in proteobacteria. *PLoS One* **9:** e113714.

Li S, Hwang XY, Breaker RR. 2016. The *yjdF* riboswitch candidate regulates gene expression by binding diverse azaaromatic compounds. *RNA* **22:** 530–541.

Macke T, Ecker D, Gutell R, Gautheret D, Case DA, Sampath R. 2001. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res* **29:** 4724–4735.

Madoori PK, Aqustiandari H, Driessen AJ, Thunnissen AM. 2009. Structure of the transcriptional regulator LmrR and its mechanism of multidrug recognition. *EMBO J* **28:** 156–166.

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

Mandal M, Boese B, Barrick JE, Winkler WC, Breaker RR. 2003. Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell* **113:** 577–586.

Mandal M, Lee M, Barrick JE, Weinberg Z, Emilsson GM, Ruzzo WL, Breaker RR. 2004. A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science* **306:** 275–279.

McCown PJ, Roth A, Breaker RR. 2011. An expanded collection and refined consensus model of *glmS* ribozymes. *RNA* **17:** 728–736.

McCown PJ, Liang JJ, Weinberg Z, Breaker RR. 2014. Structural, functional, and taxonomic diversity of three preQ$_1$ riboswitches. *Chem Biol* **21:** 880–889.

McDaniel BAM, Grundy FJ, Artsimovitch I, Henkin TM. 2003. Transcription termination control of the S box system: direct measurement of *S*-adenosylmethionine by the leader RNA. *Proc Natl Acad Sci* **100:** 3083–3088.

Mellin JR, Tiensuu T, Bécavin C, Gouin E, Johansson J, Cossart P. 2013. A riboswitch-regulated antisense RNA in *Listeria monocytogenes*. *Proc Natl Acad Sci* **110:** 13132–13137.

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

Meyer MM, Hammond MC, Salinas Y, Roth A, Sudarsan N, Breaker RR. 2011. Challenges of ligand identification for riboswitch candidates. *RNA Biol* **8:** 5–10.

Mironov AS, Gusarov I, Rafikov R, Lopez LE, Shatalin K, Kreneva RA, Perumov DA, Nudler E. 2002. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **111:** 747–756.

Mulhbacher J, Brouillette E, Allard M, Fortier LC, Malouin F, Lafontaine DA. 2010. Novel riboswitch ligand analogs as selective inhibitors of guanine-related metabolic pathways. *PLoS Pathog* **6:** e1000865.

Nahvi AS, Sudarsan N, Ebert MS, Zou X, Brown KL, Breaker RR. 2002. Genetic control by a metabolite binding mRNA. *Chem Biol* **9:** 1043–1049.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29:** 2933–2935.

Nelson JW, Breaker RR. 2017. The lost language of the RNA World. *Sci Signal* (in press).

Nelson JW, Sudarsan N, Furukawa K, Weinberg Z, Wang JX, Breaker RR. 2013. Riboswitches in eubacteria that sense the second messenger c-di-AMP. *Nat Chem Biol* **9:** 834–839.

Nelson JW, Sudarsan N, Phillip GE, Stav S, Lünse CE, McCown PJ, Breaker RR. 2015. Control of bacterial exoelectrogenesis by c-AMP-GMP. *Proc Natl Acad Sci* **112:** 5389–5394.

Nelson JW, Atilho RM, Sherlock ME, Stockbridge RB, Breaker RR. 2017. Metabolism of free guanidine in bacteria is regulated by a widespread riboswitch class. *Mol Cell* **65:** 220–230.

Newman MEJ. 2004. Power laws, Pareto distributions and Zipf's law. *Contemp Phys* **46:** 323–351.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44:** D733–D745.

Ott E, Stolz J, Lehmann M, Mack M. 2009. The RFN riboswitch of *Bacillus subtilis* is a target for the antibiotic roseoflavin produced by *Streptomyces davawensis*. *RNA Biol* **6:** 276–280.

Pedrolli DB, Matern A, Wang J, Ester M, Siedler K, Breaker R, Mack M. 2012. A highly specialized flavin mononucleotide riboswitch responds differently to similar ligands and confers roseoflavin resistance to *Streptomyces davawensis*. *Nucleic Acids Res* **40:** 8662–8673.

Price IR, Gaballa A, Ding F, Helmann JD, Ke A. 2015. Mn^{2+}-sensing mechanisms of *yybP-ykoY* orphan riboswitches. *Mol Cell* **57:** 1110–1123.

Pruitt KD, Tatusova T, Klimke W, Maglott DR. 2009. NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Res* **37:** D32–D36.

Reiss CW, Xiong Y, Strobel SA. 2017. Structural basis for ligand binding to the guanidine-I riboswitch. *Structure* **25:** 195–202.

Ren A, Patel DJ. 2014. c-di-AMP binds the *ydaO* riboswitch in two pseudo-symmetry–related pockets. *Nat Chem Biol* **10:** 780–786.

Ren A, Rajashankar KR, Patel DJ. 2012. Fluoride ion encapsulation by Mg^{2+} ions and phosphates in a fluoride riboswitch. *Nature* **486:** 85–89.

Ren A, Rajashankar KR, Patel DJ. 2015. Global RNA fold and molecular recognition for a *pfl* riboswitch bound to ZMP, a master regulator of one-carbon metabolism. *Structure* **23:** 1375–1381.

Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. 2002. Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. *J Biol Chem* **277:** 48939–48959.

Roth A, Breaker RR. 2009. The structural and functional diversity of metabolite-binding riboswitches. *Annu Rev Biochem* **78:** 305–334.

Roth A, Breaker RR. 2013. Integron *attI1* sites, not riboswitches, associate with antibiotic resistance genes. *Cell* **153:** 1417–1418.

Ruff KM, Muhammad A, McCown PJ, Breaker RR, Strobel SA. 2016. Singlet glycine riboswitches bind ligand as well as tandem riboswitches. *RNA* **22:** 1728–1738.

Sentausa E, Fournier PE. 2013. Advantages and limitations of genomics in prokaryotic taxonomy. *Clin Microbiol Infect* **19:** 790–795.

Serganov A. 2009. The long and short of riboswitches. *Curr Opin Struct Biol* **19:** 251–259.

Serganov A, Nudler E. 2013. A decade of riboswitches. *Cell* **152:** 17–24.

Serganov A, Patel DJ. 2012. Metabolite recognition principles and molecular mechanisms underlying riboswitch function. *Annu Rev Biophys* **41:** 343–370.

Serganov A, Polonskaia A, Phan AT, Breaker RR, Patel DJ. 2006. Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature* **441:** 1167–1171.

Serganov A, Huang L, Patel DJ. 2009. Coenzyme recognition and gene regulation by a flavin mononucleotide riboswitch. *Nature* **458:** 233–237.

Shamovsky I, Ivannikov M, Kandel ES, Gershon D, Nudler E. 2006. RNA-mediated response to heat shock in mammalian cells. *Nature* **440:** 556–560.

Sherlock ME, Breaker RR. 2017. Biochemical validation of a third guanidine riboswitch class in bacteria. *Biochemistry* **56:** 359–363.

Sherlock ME, Malkowski SN, Breaker RR. 2017. Biochemical validation of a second guanidine riboswitch class in bacteria. *Biochemistry* **56:** 352–358.

Shi Y, Zhao G, Kong W. 2014. Genetic analysis of riboswitch-mediated transcriptional regulation responding to Mn^{2+} in *Salmonella*. *J Biol Chem* **289:** 11353–11366.

Smith KD, Lipchock SV, Ames TD, Wang J, Breaker RR, Strobel SA. 2009. Structural basis of ligand binding by a c-di-GMP riboswitch. *Nat Struct Mol Biol* **16:** 1218–1223.

Stewart EJ. 2012. Growing unculturable bacteria. *J Bacteriol* **194:** 4151–4160.

Sudarsan N, Barrick JE, Breaker RR. 2003a. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA* **9:** 644–647.

Sudarsan N, Wickiser JK, Nakamura S, Ebert MS, Breaker RR. 2003b. An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes Dev* **17:** 2688–2697.

Sudarsan N, Cohen-Chalamish S, Nakamura S, Emilsson GM, Breaker RR. 2005. Thiamine pyrophosphate riboswitches are targets for the antimicrobial compound pyrithiamine. *Chem Biol* **12:** 1325–1335.

Sudarsan N, Lee ER, Weinberg Z, Moy RH, Kim JN, Link KH, Breaker RR. 2008. Riboswitches in eubacteria sense the second messenger cyclic di-GMP. *Science* **321:** 411–413.

Sun EI, Leyn SA, Kazanov MD, Saier MH Jr, Novichkov PS, Rodionov DA. 2013. Comparative genomics of metabolic capacities of regulons controlled by *cis*-regulatory RNA motifs in bacteria. *BMC Genomics* **14:** 597–615.

Thore S, Leibundgut M, Ban N. 2006. Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. *Science* **312:** 1208–1211.

Trausch JJ, Ceres P, Reyes FE, Batey RT. 2011. The structure of a tetrahydrofolate-sensing riboswitch reveals two ligand binding sites in a single aptamer. *Structure* **19:** 1413–1423.

Trausch JJ, Marcano-Velázquez JG, Matyjasik MM, Batey RT. 2015. Metal ion-mediated nucleobase recognition by the ZTP riboswitch. *Chem Biol* **22:** 829–837.

Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS. 2004. Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet* **20:** 44–50.

Wachter A, Tunc-Ozdemir M, Grove BC, Green PJ, Shintani DK, Breaker RR. 2007. Riboswitch control of gene expression in plants by splicing and alternative 3′ end processing. *Plant Cell* **19:** 3437–3450.

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

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

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

Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, Breaker RR. 2010. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol* **11:** R31.

Weinberg Z, Nelson JW, Lünse CE, Sherlock ME, Breaker RR. 2017. Bioinformatic analysis of riboswitch structures uncovers variant classes with altered ligand specificity. *Proc Natl Acad Sci* **114:** E2077–E2085.

White HB III. 1976. Coenzymes as fossils of an earlier metabolic state. *J Mol Evol* **7:** 101–104.

Winkler WC, Cohen-Chalamish S, Breaker RR. 2002a. An mRNA structure that controls gene expression by binding FMN. *Proc Natl Acad Sci* **99:** 15908–15913.

Winkler WC, Nahvi A, Breaker RR. 2002b. Thiamine derivatives that bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419:** 952–956.

Winkler WC, Navi A, Sudarsan N, Barrick JE, Breaker RR. 2003. An mRNA structure that controls gene expression by binding *S*-adenosylmethionine. *Nat Struct Biol* **10:** 701–707.

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

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

Woese C. 1967. *The genetic code: the molecular basis for genetic expression*. Harper and Row, New York.