***Supplementary Information***

**Challenges of Ligand Identification for the Second Wave of Orphan Riboswitch Candidates**

**Etienne B. Greenleea, Shira Stava, Ruben M. Atilhob, Kenneth I. Brewerb, Kimberly A. Harrisa, Sarah N. Malkowskic, Gayan Mirihana Arachchilaged, Kevin R. Perkinsa, Madeline E. Sherlockb, and Ronald R. Breaker a,b,d,\***

aDepartment of Molecular, Cellular and Developmental Biology, bDepartment of Molecular Biophysics and Biochemistry, cDepartment of Chemistry, dHoward Hughes Medical Institute, Yale University, P.O. Box 208103, New Haven, CT 06520-8103, USA.

\*To whom correspondence should be addressed.

Email: ronald.breaker@yale.edu

Twitter: @RonBreaker

Phone: 203 432-9389

**Summaries of weak riboswitch candidates**

The motifs listed below have some characteristics that suggest their biological functions might not involve the direct binding of small ligands. However, we have not completely ruled out the possibility that they function as riboswitches.

**The *sucC* motif.** A total of 72 unique representatives of the *sucC* motif were identified almost exclusively in species of *Pseudomonas*. The architecture of the motif (**Fig. S1A**) appears to include one well-supported hairpin (P1) and perhaps another simple hairpin (P2) that has only modest support from our comparative sequence analysis data. Conserved nucleotides are distributed among the base-paired regions and the flanking unpaired regions. Notably, P1 appears to be formed by a palindromic base-paired sub-structure that was not previously1 recognized. This architectural feature suggests that the RNA motif presents binding sites for a dimeric protein complex. However, a small and repetitive hairpin repeat is present in guanidine-II riboswitches,2 and yet this RNA selectively binds two guanidine ligands. Therefore, we are reluctant to rule out the possibility that this RNA motif directly binds a small molecule ligand.

The RNA motif primarily associates with the *sucC* gene (**Fig. S1B**), which encodes the beta subunit of succinyl-CoA synthetase. This is an intriguing gene association because many of the common riboswitch classes respond to RNA-derived coenzymes, and an example of a riboswitch that binds coenzyme A has yet to be experimentally validated.



**Fig. S1.** The consensus sequence and secondary structure model (A) and gene associations (B) for *sucC* motif RNAs.

**The *rmf* motif.** The *rmf* motif (**Fig. S2**) is comprised of four hairpins, with regions of highly conserved nucleotides located upstream, within the stems or loops, and downstream of these substructures. The proximity of these sequence and structural features to the RBS and start codon of the downstream ORF is consistent with gene regulatory function. Only 22 representatives of this motif were previously reported,1 but we now observe a total of 363 unique examples of the RNA motif in the Pseudodomonadaceae family and in several environmental samples. Although this motif is somewhat narrowly distributed, there is strong evidence for covariation supporting the consensus model.

The motif is exclusively found upstream of the *rmf* gene, which encodes ribosomal modulation factor (RMF). RMF is a small protein shown to convert 70S ribosomes to inactive homodimers during stationary phase.3 It has been shown that the expression of the *rmf* gene in *E. coli* is positively regulated by the level of guanosine tetraphosphate (ppGpp),4 a signaling molecule involved in the stringent response in bacteria. Therefore, it is intriguing to consider ppGpp as a potential ligand sensed by the *rmf* motif. However, many bacterial5 and fungal6 proteins participate in feedback regulation of their expression by binding to structured RNA elements in their own mRNAs. Therefore, the *rmf* motif might be recognized by the protein whose production it controls.



**Fig. S2.** The consensus sequence and secondary structure model for *rmf* motif RNAs.

**The Termite-*flg* motif.** Of the 67 known examples, members of the Termite-*flg* candidate riboswitch class are most often found in environmental sequences. Only two examples are the genome of a single sequenced organism, *Treponema primita* which is found in the gut of termites. The motif, as originally described,1 is comprised of one large and complex hairpin (P2) and another smaller hairpin (P3), and perhaps an additional small (P1) stem (**Fig. S3A**). Most of the conserved nucleotides are present in the loops of P2 and P3, the internal loop of P2, and within the putative initial base-paired region of P2. The RNA motif is commonly found in between tandemly-arranged *flgL* genes that encode flagellin proteins (**Fig. S3B**). A RBS sequence is frequently located several nucleotides downstream of the motif, followed by the start codon for the second *flgL* gene. Although sometimes flagellar protein genes are regulated by c-di-GMP riboswitches,7 the lack of other gene associations typical of this signaling molecule casts doubt on this possible ligand. Its genomic location and proximity to the RBS suggests a regulatory function, however we cannot conclude with confidence that this regulation occurs without the involvement of a regulatory protein or sRNA factor.



**Figure S3.** Consensus sequence and secondary structure model (A) and gene associations (B) for the Termite-*flg* motif RNA.

**The *nuoG* motif.** This riboswitch candidate is found in several orders of the Enterobacteriaceae family and environmental sequences. A total of 38 unique representatives were identified. The motif, which remains largely unchanged from the previous prediction,1 often includes only a small P1 hairpin, which is well-supported by sequence covariation (**Fig. S4**). In 22 representatives, an additional imperfectly base-paired structure is present in the loop of P1. Immediately downstream of P1 is a predicted RBS and start codon. The gene downstream is exclusively *nuoG*, which in *E. coli* codes for the G subunit of the NADH-ubiquinone oxidoreductase (complex I).8 Complex I participates in generating the proton motive force that drives the oxidative phosphorylation process to synthesize ATP.

 Although it would be fascinating to identify a riboswitch that regulates expression of a gene that plays a central role in the oxidative phosphorylation process, there are reasons to favor a non-riboswitch function for the *nuoG* motif. For example, the simple architecture of the motif is atypical of the more complex RNA structures that serve as riboswitch aptamers. Instead of forming a small hairpin, the sequence could function as a palindromic DNA binding site for a protein dimer. However, the fact that some representatives have a large imperfectly base-paired insert into the loop of P1 might favor a role for this motif as an RNA transcript. Moreover, the narrow phylogenetic distribution and exclusive association with the gene for only one subunit of complex I would also be unusual for a metabolite-binding riboswitch. These characteristics cause us to rank this very low on the list of riboswitch candidates.



**Figure S4.** Consensus sequence and secondary structure model for the *nuoG* motif RNA.

**The *livK* motif.** The *livK* RNA motif was previously described1 as residing upstream of the *livK* gene. This gene product is a member of the Periplasmic Binding Protein Type 1 superfamily. However, the *livK* annotation is not strong and, for many representatives, this gene is annotated as another member of the same superfamily - *amiC*. This candidate riboswitch has 88 representatives found exclusively in the *Pseudomonas* genus and environmental samples. The motif is comprised of two stems with multiple instances of covarying mutations and numerous highly conserved nucleotides (**Fig. S5**). The motif is mostly found three times in each organism. The first representative is located upstream of the *amiC* gene, which is typically immediately followed by the *amiR* gene. These first two genes are transcriptional regulators that are known to regulate expression of the aliphatic amidase AmiE.9 The second representative resides after *amiR* and upstream of a gene that encodes a putative amidase enzyme. This gene does not appear to be the aliphatic amidase AmiE, but rather another unknown amidase. The third instance of the motif occurs on the opposite strand just next to the first described instance of the motif. This third motif is upstream of a gene that encodes for a predicted urea transporter.

For the known aliphatic amidase AmiE, AmiC functions as a negative regulator by binding the ligand acetamide, and AmiR positive regulates amidase by activing as an antiterminator.10 Urea can also directly inhibit this amidase.11 In contrast to the genetic context seen for the *livK* RNA motif, *amiE* is found upstream of *amiC* and *amiR*. Several protein regulatory factors are known to bind 5ˊ-leaders of mRNA transcripts they control, including their own mRNAs to achieve autoregulation. Thus, it seems most likely that the *livK* motif is a protein-binding site, but we have not ruled out the possibility that members of this RNA class function as riboswitches for a ligand such as acetamide or urea.



**Fig. S5.** Consensus sequence and secondary structure model for the *livK* motif RNA.

**The *Polynucleobacter*-1 motif.** Originally1 this motif was found in only environmental sequences, with one example present in *Polynucleobacter*. There are now 86 examples of this motif, with the majority still identified in environmental sequences and two additional examples in the *Limnohabitans* genus. The putative additional structure that is sometimes seen downstream is difficult to model, since many environmental sequences lack complete sequence information. Thus, it is genetically unclear what is located immediately downstream of this motif. There does not appear to be any conserved protein domains just after the motif, which is consistent with additional structure at the RNA level. When this additional sequence is available, a gene encoding an integrase is present. Therefore, this motif might represent a nucleic acid structure that is important for the function of a mobile genetic element, although we cannot yet rule out the possibility of riboswitch function.



**Fig. S6.** Consensus sequence and secondary structure model for the *Polynucleobacter*-1 motif.

**Summaries of rejected motifs**

Additional RNA motifs published previously were further examined as possible riboswitch candidates that were dropped from further consideration. Specifically, the possible RNA motifs noted below were subjected to the same analyses as those in the main text, but were removed from the candidate riboswitch list for various reasons as noted for each motif. Note that some of these candidates were originally described as ‘ambiguous’,1 meaning that there was little or no good evidence in favor of riboswitch function.

**The *atoC* motif.** The structure of the *atoC* motif,1 based on the analysis of 64 representatives, is comprised of a simple small hairpin with no strictly conserved nucleotides. This hairpin is followed by an apparent terminator stem. The lack of sequence conservation is not consistent with the formation of a riboswitch aptamer.

**The *asd* motif.** The *asd* motif1 is not typically located in a position consistent with a cis-regulatory function. However, *asd* motif RNAs carry numerous conserved nucleotides and fold into a complex secondary structure, suggesting that it is a well-structured noncoding RNA.

**The Bacteriodales-1 motif.** TheBacteriodales-1 motif1 is commonly located upstream of ribosomal protein L20 genes, and therefore is most likely a binding site for L20 protein autoregulation.

**The Gamma-cis-1 motif.** This motif was originally published with 49 examples.1 At present, there are at least 4,704 examples that conform to a consensus model carrying conserved nucleotides primarily at the base of stem P1, and near a small internal loop in P3 (**Fig. S7A**). Multiple copies of the motif are commonly present within a given genome, and are associated with genes of unknown function (**Fig. S7B**). Often, two copies of the motif flank an ORF, with one copy on each side of the ORF. Additionally, the motif has an imperfect repeated sequence at the 5ˊ end and the 3ˊ end. This sequence closely matches the RYYYAAC consensus sequence of the inverse core site in *attC* sites.12 For these reasons, we conclude that this motif is likely a structured DNA element that acts as an *attC* site for an integron gene cassette.



**Figure S7.** Consensus sequence and secondary structure model (A) and gene associations (B) for the Gamma-cis-1motif. Nucleotides depicted are DNA given the likelihood that this motif functions as a structured DNA element.

**The Dictyoglomi-1 motif.** Most of the 29 representatives are from environmental sequences. The remaining sequences are phylogenetically diverse, and can be found in the Dictyoglomi, Actinobacteria, and Firmicutes phyla. When in Dictyoglomi, two distal versions of the motif occur that retain some of the distinct features of the consensus model (**Fig. S8**). The motif is mostly found upstream of a putative glycoside hydrolase and sometimes found upstream of a putative metal-dependent phosphohydrolase.

Some architectural features of the Dictyoglomi-1 motif suggest these RNAs represent a distant group of 6S RNAs13 which are natural aptamers for RNA polymerase that regulate transcription.14 Notably, there are two conserved C-G base-pairs that reside within one of the most conserved parts of the motif. This region resides at the base of an internal loop with an approximate length of 15 on the 5′ side. These features resemble the most important parts of known 6S RNAs. However, other regions appear to differ somewhat from known 6S RNAs, and therefore it is not certain that this RNA motif represents distal members of this general ncRNA class.



**Figure S8.** Updated consensus sequence and secondary structure model for Dictyoglomi-1 motif RNAs. Features similar to 6S RNAs include a large internal loop of conserved size, and conserved nucleotides at the base of this loop.

**The *hopC* motif.** The *hopC* motif1 appears to be a highly repetitive genetic element that is only narrowly distributed in bacteria.

**The JUMPstart motif.** The small consensus motif for JUMPstart is consistent with its proposed function as a binding site for the protein factor RfaH.1

**The Lnt motif.** Only 23 examples of the Lnt motif1 were identified. This motif is comprised of a simple small hairpin, and has mostly A and U residues conserved within a 3ˊ region that appears devoid of structure. This architecture as currently understood appears unlikely to function as a riboswitch aptamer.

**The *mraW* motif.** The *mraW* motif1 is a simple and highly repetitive motif that is likely to be part of a selfish genetic element.

**The methylobacterium-1 motif.** The methylobacterium-1 motif1 is mostly found in environmental sequences with the remainder from various species of α-proteobacteria. This RNA motif has a wide variety of genes immediately downstream, but is most frequently found near specific integrases and recombinases that have previously been associated with self-cleaving ribozymes.15 This genetic context suggests that the motif functions as part of a selfish genetic element.

**The Ocean-VI motif.** Upon further analysis, these sequences of the original Ocean-VImotif1 overlaps an ORF coding for a YhaM superfamily protein. Therefore, the presumed motif is likely to reflect the conservation of the coding region and not a noncoding RNA structure.

**The *pan* motif.** The *pan* motif1 is comprised of two repetitive elongates hairpins with symmetrical bulged A nucleotides. This simple but repetitive structure is indicative of the binding of protein dimers.

**The *Pseudomon*-*groES* motif.** The *Pseudomon*-*groES* motif1 is formed by four modest-length hairpins with only a few highly conserved nucleotides mostly residing within base-paired regions. Of the 364 distinct representatives, all are found immediately upstream of the *groES* gene in a cis-regulatory position. However, its architecture appears more suited to respond to a protein or sRNA signal, rather than function as a metabolite-binding riboswitch.

**The *radC* motif.** The *radC* motif1 is commonly associated with genes encoding DNA-binding proteins, and therefore might represent a structured DNA motif.

**The Rhodopirellula-1 motif.** The Rhodopirellula-1 motif1 is likely to be similar in function to an *attC* DNA endonuclease cleavage site.

**The Termite-*leu* motif.** The Termite-leu motif1 is formed by two relatively long hairpins with no highly conserved nucleotides. The absence of conserved nucleotides strongly suggests that the motif does not function as a riboswitch aptamer.

**The Whalefall-1 motif.** The Whalefall-1 motif1 is a very simple hairpin structure and yields many possible representatives whose genomic locations are not consistent with a regulatory function. Also, the genes nearby do not have a consistent functional theme, unlike most riboswitch classes.

**References**

1. Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, et al. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. Genome Biol 2010; 11: R31.
2. Sherlock ME, Malkowski SN, Breaker RR. Biochemical validation of a second guanidine riboswitch class in bacteria. Biochemistry 2017; 56: 352-8.
3. Yamagishi M, Matsushima H, Wada A, Sakagami M, Fujita N, Ishihama A. Regulation of the *Escherichia coli* *rmf* gene encoding the ribosome modulation factor: growth phase- and growth rate-dependent control. EMBO J 1993; 12: 625-30.
4. Izutsu K, Wada A, Wada C. Expression of ribosome modulation factor (RMF) in *Escherichia coli* requires ppGpp. Genes Cells 2001; 6:665-76.
5. Meyer MM. The role of mRNA structure in bacterial translational regulation. WIREs RNA 2017; 8:e1370.
6. Li S, Breaker RR. Identification of 15 candidate structured noncoding RNA motifs in fungi by comparative genomics. 2017; BMC Genomics (*in press*).
7. Sudarsan N, Lee ER, Weinberg Z, Moy RH, Kim JN, Link KH, Breaker RR. Riboswitches in eubacteria sense the second messenger cyclic-di-GMP. Science 2008; 321: 411-3.
8. Falk-Krzesinski HJ, Wolfe AJ. Genetic analysis of the *nuo* locus, which encodes the proton-translocating NADH dehydrogenase in *Escherichia coli*. J Bacteriol 1998; 180: 1174-84.
9. Wilson SA, Wachira SJ, Drew RE, Jones D, Pearl LH. Antitermination of amidase expression in *Pseudomonas aeruginosa* is controlled by a novel cytoplasmic amide-binding protein. EMBO J 1993; 12: 3637-42.
10. O'hara BP, Norman RA, Wan PT, et al. Crystal structure and induction mechanism of AmiC-AmiR: a ligand-regulated transcription antitermination complex. EMBO J 1999; 18: 5175-86.
11. Tata R, Marsh P, Brown PR. Arg-188 and Trp-144 are implicated in the binding of urea and acetamide to the active site of the amidase from *Pseudomonas aeruginosa*. Biochim Biophys Acta 1994; 1205: 139-45.
12. Stokes HW, O'gorman DB, Recchia GD, Parsekhian M, Hall RM. Structure and function of 59-base element recombination sites associated with mobile gene cassettes. Mol Microbiol 1997; 26: 731-45.
13. Barrick JE, Sudarsan N, Weinberg Z, Ruzzo WL, Breaker RR. 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. RNA 2005; 11: 774-84.
14. Cavanagh AT, Wassarman KM. 6S RNA, a global regulator of transcription in *Escherichia coli*, *Bacillus subtilis*, and beyond. Annu Rev Microbiol 2014; 68: 45-60.
15. Weinberg Z, Kim PB, Chen TH, et al. New classes of self-cleaving ribozymes revealed by comparative genomics analysis. Nat Chem Biol 2015; 11: 606-10.