

Additional File 2 Summary of all RNA motifs supplementary for: Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea and their metagenomes

Zasha Weinberg, Joy X. Wang, Jarrod Bogue, Jingying Yang,
Keith Corbino, Ryan H. Moy, Ronald R. Breaker

March 22, 2010

The following table lists all RNA motifs predicted in this work, with annotation. For the “RNA?” column, the annotation codes are:

- Y/y/?/n/N: certainly, probably, ambiguous, probably not, definitely not. **Y** and **y** are capitalized to emphasize motifs that for which there is clear evidence of RNA structure.
- COV: strong covariation
- cov: modest covariation
- nc: modest non-canonicals
- seq: poor sequence conservation
- < : few seqs, so significance of covariation is unclear
- short: motif is short, so homology searches might not reveal sequences that break the model.
- V: Variable-length stems
- mod: modular stems
- tr : experimentally shown to be transcribed
- TFBS: predicted transcription-factor binding site implies that putative RNA is transcribed
- apt : experimental evidence of aptamer
- exp: other experimental evidence supporting RNA assignment

- hom(X): possible homology to RNA class X.
- fhom: presumable functional homolog to distinct RNA class in the same gene context
- tson: possibly transposon-associated, or associated with integrases
- loc: “local”-mode covariance model searches lead to reduced sequence conservation, perhaps simply because of unrelated hairpins, but this is cause for additional caution.
- rep: possibly repetitive (many representatives per genome)
- TT: predicted terminator associated with motif
- SD: overlap of Shine-Dalgarno implies a *cis*-regulatory role, which in turn implies RNA
- peptide: presence of short ORF with peptide corresponding to downstream amino acid synthesis genes implies RNA
- DNA?: some reason to suggest motif operates at the level of DNA.
- 4loop: common presence of terminal tetraloops (GNRA, UNCG, CUNG)
- Eloop: E-loop(s) is commonly present
- RBP: downstream gene encodes RNA-binding protein

For the “*cis*-reg” (does the motif represent a *cis*-regulatory RNA) column, annotation codes are:

- Y/y/?/n/N: (same as for “RNA?” column).
- one : always upstream of genes belonging to a single gene family. This arrangement is consistent with a *cis*-regulatory role for the motif, but if the organisms containing the motif representatives are closely related, the arrangement might simply reflect insufficient opportunities for the location of the motif to change via neutral rearrangements.
- > 1 : upstream of non-homologous genes. This phenomenon favors the hypothesis that it is functionally important for motif representatives to be upstream of protein-coding genes, because there were more events where motif representatives might have (under the null hypothesis) not been in a potential 5' UTR.
- alt : alternate hypothesis for function questions whether the motif is *cis*-regulatory
- over-TT : the motif commonly overlaps predicted transcription terminators, which would be consistent with a *cis*-regulatory function.
- over-SD : the motif commonly overlaps predicted Shine-Dalgarno sequences, which would be consistent with a *cis*-regulatory function.

The “switch?” column answers how likely it is that the motif represents a riboswitch. Evidence in favor of a *cis*-regulatory role (see above) also favors riboswitch function. Annotation codes are:

- Y/y/?/n/N: (same as for “RNA?” column)
- alt: alternate hypothesis that questions riboswitch function
- complex: motif structure is predicted to contain at least one multistem junction or pseudoknot.
- ss: secondary structure is not very complex
- phyla: motif displays high conservation of some nucleotide positions and is found in multiple phyla, suggesting a strong biochemical constraint consistent with riboswitch function.

- orders: motif present in multiple orders (of taxonomy) with some highly conserved nucleotide positions; see “phyla” annotation.

The “taxa/environments” column shows taxa or environmental samples in which motif representatives are commonly found.

The “#” column is the total number of motif representatives predicted.

The “Rfam acc.” column gives the Rfam Database accession for the motif, for motifs that are included in Rfam.

The motifs are as follows:

motif	RNA?	<i>cis-reg?</i>	switch?	Taxa / environments	#	Rfam acc.
6S-flavo	Y COV hom(6S)	N	N	Bacteroidetes	92	RF01685
<i>aceE</i>	? cov nc	y	?	γ -Proteobacteria	20	
Acido-1	y COV <	n	n	Acidobacteria	4	RF01686
Acido-Lenti-1	y COV nc	n	n	Acidobacteria, Lentisphaerae	89	RF01687
Actino-pnp	Y cov fhom	Y	N	Actinomycetales	100	RF01688
AdoCbl-variant	Y COV V mod hom(AdoCbl)	Y >1	Y complex	Marine	144	RF01689
<i>asd</i>	Y COV V mod TT	?	? complex	Lactobacillales	44	RF01732
<i>atoC</i>	y cov TT	y	?	δ -Proteobacteria	8	RF01733
Bacillaceae-1	Y COV nc	n	n	Bacillaceae	149	RF01690
<i>Bacillus</i> -plasmid	y cov	?	n	<i>Bacillus</i>	32	RF01691
Bacteroid- <i>trp</i>	y peptide	y	n alt	Bacteroidetes	47	RF01692
Bacteroidales-1	Y COV 4loop TFBS TT	?	?	Bacteroidales	47	RF01693
<i>Bacteroides</i> -1	y cov nc 4loop	?	n	<i>Bacteroides</i>	40	RF01694
<i>Bacteroides</i> -2	? COV nc	n	n	<i>Bacteroides</i>	66	
Burkholderiales-1	? cov nc loc rep	?	n	Burkholderiales	88	
c4 antisense RNA	Y COV V 4loop exp hom(c4)	N	N	Proteobacteria, phages	593	RF01695
c4-alb1	Y COV tr	N	N	γ -Proteobacteria, phages	87	
Chlorobi-1	Y COV TT	n	n	Chlorobi	15	RF01696

motif	RNA?	cis-reg?	switch?	Taxa / environments	#	Rfam acc.
Chlorobi-RRM	y cov RBP	y	n	Chlorobi	20	RF01697
Chloroflexi-1	y COV nc rep <	?	n	<i>Chloroflexus aggregans</i>	3	RF01698
Clostridiales-1	y COV nc 4loop V loc	n	n	Clostridiales, human gut	206	RF01699
COG2252	? cov nc V RBP	y	n	Pseudomonadales	11	
<i>Collinsella</i> -1	y COV nc loc	n	n	Actinobacteria, human gut	6	RF01700
<i>crcB</i>	Y COV V	Y >1 over-TT	Y phyla complex	Widespread, bacteria and archaea	449	RF01734
Cyano-1	y cov	n	n	Cyanobacteria, marine	276	RF01701
Cyano-2	Y COV nc V	n	n	Cyanobacteria, marine	57	RF01702
Desulfotalea-1	? cov nc	n	n	Proteobacteria	229	
Dictyoglomi-1	y cov Eloop	?	?	Dictyoglomi	4	RF01703
Downstream-peptide	Y COV V 4loop tr	y	y orders complex	Cyanobacteria, marine	634	RF01704
<i>epsC</i>	Y COV	y >1	y complex	Bacillales	35	RF01735
<i>fixA</i>	? cov nc	y	n	<i>Pseudomonas</i>	10	
Flavo-1	y COV V nc	n	n	Bacteroidetes	239	RF01705
<i>flg</i> -Rhizobiales	y cov 4loop	y	n	Rhizobiales	14	RF01736
<i>flpD</i>	y cov	?	n	Euryarchaeota	8	RF01737
<i>gabT</i>	Y COV nc mod SD	y SD	?	<i>Pseudomonas</i>	21	RF01738
Gamma-cis-1	? cov nc tson loc	y	n	γ -Proteobacteria	50	
<i>glnA</i>	Y COV V Eloop	Y >1	y orders complex	Cyanobacteria, marine	1039	RF01739
GUCCY-hairpin	? cov short	?	n	Bacteroidetes, Proteobacteria	651	
Gut-1	Y	n	n	Human	43	RF01706

motif	RNA?	cis-reg?	switch?	Taxa / environments	#	Rfam acc.
	COV nc V			gut only		
<i>gyrA</i>	y	y >1	n	<i>Pseudomonas</i>	24	RF01740
<i>hopC</i>	COV nc y	SD Y	?	<i>Helicobacter</i>	4	RF01741
	cov tr					
<i>icd</i>	< ? cov	y	n	<i>Pseudomonas</i>	20	
JUMPstart	nc y	Y	?	γ -Proteobacteria	243	RF01707
	COV nc					
L17 downstream element	tr hom(JUMPstart) y	y (3')	n alt	Lactobacillales, <i>Listeria</i>	82	RF01708
<i>lactis</i> -plasmid	cov V y	?	n	Lactobacillales	14	RF01742
Lacto- <i>int</i>	cov ? COV	?	n	Lactobacillales,	54	
Lacto- <i>rpoB</i>	NC tson Y	y	n	phages Lactobacillales	60	RF01709
Lacto- <i>usp</i>	COV Y	? alt	?	Lactobacillales	6	RF01710
Leu/phe leader	COV hom(6S)? Y	Y	N	<i>Lactococcus</i> <i>lactis</i>	9	RF01743
	COV V					
<i>livK</i>	peptide hom(leu leader) y	y	?	Pseudomonadales	8	RF01744
Lnt	cov y	y	?	Chlorobi	10	RF01711
<i>manA</i>	cov Y	Y >1	y phyla complex	Marine, γ -Proteobacteria, cyanophage	195	RF01745
<i>Methylobacterium</i> -1	COV V mod Y	n	n	<i>Methylobacterium</i> , marine	36	RF01712
Moco-II	COV nc y	Y TT	?	Proteobacteria	8	RF01713
	cov TT					
<i>mraW</i>	< y	y	?	Actinomycetales	60	RF01746
<i>msiK</i>	COV Y	Y SD	? orders	Actinobacteria	136	RF01747
	COV V					
<i>Nitrosococcus</i> -1	SD ? cov	n	n	<i>Nitrosococcus</i> , Clostridia	14	
<i>nuoG</i>	nc 4loop y	y	?	Enterobacteriales (incl. <i>E. coli</i> K12)	86	RF01748
Ocean-V	COV V SHORT y	n	n	Marine only	3	RF01714
Ocean-VI	cov < ? cov 4loop	?	?	Marine only	308	

motif	RNA?	cis-reg?	switch?	Taxa / environments	#	Rfam acc.
<i>pan</i>	Y COV nc V TT	Y TT	?	Chloroflexi, Firmicutes, δ -Proteobacteria	121	RF01749
<i>Pedo-repair</i>	y COV <	?	n	<i>Pedobacter</i>	6	RF01715
<i>pfl</i>	Y COV V mod TT	Y over-TT	Y phyla complex	Several phyla	234	RF01750
<i>pheA</i>	? cov nc	y	n	Actinobacteria	29	
PhotoRC-I	y cov 4loop	y	n	Cyanobacteria, marine	66	RF01716
PhotoRC-II	Y COV	y	n	Marine, cyanophage	467	RF01717
<i>Polynucleobacter-1</i>	y COV nc	y >1	?	Burkholderiales, fresh water/estuary	16	RF01718
<i>potC</i>	y cov	y >1	? ss	Marine only	136	RF01751
<i>psaA</i>	Y COV V nc 4loop	y V	? orders? complex	Cyanobacteria	31	RF01752
<i>psbNH</i>	y cov nc	y	n	Cyanobacteria, marine	39	RF01753
<i>Pseudomon-1</i>	y COV TT	n	n	Pseudomonadales	23	RF01719
<i>Pseudomon-2</i>	? cov NC	n	n	<i>Pseudomonas</i>	16	
<i>Pseudomon-GGDEF</i>	? cov nc	y	?	<i>Pseudomonas</i>	7	
<i>Pseudomon-groES</i>	y COV 4loop tr	y	?	<i>Pseudomonas</i>	20	RF01721
<i>Pseudomon-Rho</i>	y cov nc	Y	n	<i>Pseudomonas</i>	22	RF01720
<i>Pyrobac-1</i>	y COV V	n	n	<i>Pyrobaculum</i>	8	RF01722
<i>Pyrobac-HINT</i>	? cov <	y	n	<i>Pyrobaculum</i>	4	
<i>radC</i>	Y COV V 4loop	y	? complex	Proteobacteria	167	RF01754
Rhizobiales-1	? cov nc loc rep	n	N	Rhizobiales	795	
Rhizobiales-2	y cov	?	n	Rhizobiales	16	RF01723
Rhodopirellula-1	? cov nc rep	y	?	Proteobacteria, Planctomycetes	153	
<i>rmf</i>	Y COV	y	?	Pseudomonadales	22	RF01755
<i>rne-II</i>	Y COV V mod fhom	y	N	Pseudomonadales	21	RF01756

motif	RNA?	cis-reg?	switch?	Taxa / environments	#	Rfam acc.
SAM-Chlorobi	y cov TFBS	Y	?	Chlorobi	10	RF01724
SAM-I-IV-variant	Y COV V 4loop mod hom(SAM-I/IV)	Y	Y phyla complex	Several phyla, marine	473	RF01725
SAM-II long loops	Y COV hom(SAM-II)	Y >1	Y phyla complex	Bacteroidetes, marine	162	RF01726
SAM/SAH riboswitch	Y COV	Y over-SD?	Y apt	Rhodobacterales	64	RF01727
<i>sanguinis</i> -hairpin	? cov nc rep	n	n	<i>Streptococcus</i>	12	
<i>sbcD</i>	y cov V TT <	?	n	Burkholderiales	8	RF01757
ScRE	? cov nc	y	n	<i>Streptococcus</i>	59	
Soil-1	? cov nc 4loop	n	n	Soil only	15	
<i>Solibacter</i> -1	? cov nc rep	n	n	<i>Solibacter</i> <i>usitatus</i>	34	
STAXI	y COV V 4loop DNA?	?	n	Enterobacteriales	77	RF01728
<i>sucA</i> -II	y cov V	y	?	Pseudomonadales	22	RF01758
<i>sucC</i>	Y COV V 4loop	Y	?	γ -Proteobacteria	25	RF01759
Termite- <i>flg</i>	Y COV nc	y	n	Termite hind gut only	13	RF01729
Termite- <i>leu</i>	y COV V	?	?	Termite hind gut only	20	RF01730
<i>traJ</i> -II	Y COV V nc fhom	Y	n	Proteobacteria, <i>Enterococcus</i> <i>faecium</i>	71	RF01760
Transposase-resistance	? V COV nc tson loc	y	n	Several phyla	218	
TwoAYGGAY	y COV nc loc	n	n	Human gut, γ -Proteobacteria, Clostridiales	216	
<i>wcaG</i>	Y COV	y >1	y complex	Marine, cyanophage	120	RF01761
Whalefall-1	Y COV nc 4loop	n	n	Whalefall only	16	RF01762
<i>yjdF</i>	Y COV V SD 4loop	Y over-SD	Y orders complex	Firmicutes	110	RF01764
<i>ykkC</i> -III	y	Y >1	y phyla	Actinobacteria,	44	RF01763

motif	RNA?	<i>cis</i> -reg?	switch?	Taxa / environments	#	Rfam acc.
	COV _{nc}		complex	δ -Proteobacteria		