Genome Biology

RESEARCH Open Access

Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes

Zasha Weinberg[1,2*], Joy X Wang[1], Jarrod Bogue[2,4], Jingying Yang[2], Keith Corbino[1], Ryan H Moy[2,5], Ronald R Breaker[1,2,3*]

Abstract

Background: Structured noncoding RNAs perform many functions that are essential for protein synthesis, RNA processing, and gene regulation. Structured RNAs can be detected by comparative genomics, in which homologous sequences are identified and inspected for mutations that conserve RNA secondary structure.

Results: By applying a comparative genomics-based approach to genome and metagenome sequences from bacteria and archaea, we identified 104 candidate structured RNAs and inferred putative functions for many of these. Twelve candidate metabolite-binding RNAs were identified, three of which were validated, including one reported herein that binds the coenzyme *S*-adenosylmethionine. Newly identified *cis*-regulatory RNAs are implicated in photosynthesis or nitrogen regulation in cyanobacteria, purine and one-carbon metabolism, stomach infection by *Helicobacter*, and many other physiological processes. A candidate riboswitch termed *crcB* is represented in both bacteria and archaea. Another RNA motif may control gene expression from 3'-untranslated regions of mRNAs, which is unusual for bacteria. Many noncoding RNAs that likely act in *trans* are also revealed, and several of the noncoding RNA candidates are found mostly or exclusively in metagenome DNA sequences.

Conclusions: This work greatly expands the variety of highly structured noncoding RNAs known to exist in bacteria and archaea and provides a starting point for biochemical and genetic studies needed to validate their biologic functions. Given the sustained rate of RNA discovery over several similar projects, we expect that far more structured RNAs remain to be discovered from bacterial and archaeal organisms.

Background

Ongoing efforts to identify and characterize various structured noncoding RNAs from bacteria are revealing the remarkable functions that structured RNAs can perform [1-3]. To detect novel RNA classes in bacteria and archaea, a variety of bioinformatics strategies have been used [4-12]. In our recent efforts to identify novel structured RNAs, we applied a scheme based on detecting RNA secondary structures upstream of homologous protein-coding genes [13,14]. However, this strategy is best suited to finding *cis*-regulatory RNAs, not noncoding RNAs. Also, some *cis*-regulatory RNAs such as c-di-GMP riboswitches [14,15] or *ydaO* motif RNAs [5] are not often found upstream of homologous genes [13].

We therefore implemented a search system that is independent of protein-coding genes. In brief, our system clusters intergenic regions (IGRs) [16] by using a BLAST-based method [17] and infers secondary structures by using CMfinder [18]. Then, as before [19,20], the identified structures are used in homology searches to find homologues that allow CMfinder to refine further its structural alignment. The resulting alignments are scored and then analyzed manually to identify the most promising candidates and to infer possible biologic roles.

This method was applied to all available bacterial and archaeal genome sequences, as well as metagenome (that is, environmental) sequences, and identified 104 candidate RNA motifs described in this report. Some additional RNAs will be reported later (unpublished data)

* Correspondence: zasha.weinberg@yale.edu; ronald.breaker@yale.edu
[1]Howard Hughes Medical Institute, Yale University, P.O. Box 208103, New Haven, CT 06520-8103, USA

that bind cyclic di-GMP or tetrahydrofolate, that represent diverse variants of hammerhead self-cleaving ribozymes, or that exhibit exceptional characteristics suggesting a novel or unusual biochemical function [21]. In this report, we provide biochemical evidence that members of one of the 104 RNA motifs bind *S*-adenosylhomocysteine (SAH) and *S*-adenosylmethionine (SAM) *in vitro*, and presumably regulate the downstream genes coding for SAM synthetase. The rest of this report provides predicted structures of selected motifs and hypotheses regarding their biologic roles. The remaining motifs, as well as additional information on the selected motifs, are presented in Additional File 1. Discussions about individual motifs are largely independent, but are grouped into common putative functional roles. A list of all 104 motifs is provided in Table 1 and Additional File 2. Multiple-sequence alignments of motifs, the organisms in which their representatives appear, and predicted flanking genes are available in printable format in Additional File 3, and alignments are provided in machine-readable format in Additional Files 4 and 5. Consensus diagrams for all motifs are depicted in Additional File 6. Selected motifs (Table 1) were submitted for inclusion in the Rfam Database version 10.1 [22].

Results and discussion

Identification and analysis of RNA structures

Promising RNA motifs predicted by our automated bioinformatics procedure were subsequently evaluated manually (see Materials and Methods). As previously reported [14], we identified promising motifs by seeking RNAs that exhibit both regions of conserved nucleotide sequence and evidence of secondary structure. Evidence for the latter characteristic involved the identification of nucleotide variation between representatives of a motif that conserves a given structure. For example, one form of covariation involves mutations to two nucleotides that preserve a Watson-Crick base pair. Assessment of covariation can be complicated, because, for example, spurious evidence of covariation is sometimes a consequence of sequence misalignments. Therefore, final covariation assessments were performed manually.

Cis-regulatory RNAs in bacteria are typically located in 5' UTRs. However, transcription start sites for most genes have not been experimentally established. Therefore, when a motif commonly resides upstream of coding regions, we usually assume that it resides in 5' UTRs and is a *cis*-regulatory RNA. Additional analysis of our system and our scheme for naming motifs is described in Additional File 1.

Riboswitch candidates

Riboswitches [1,2,23] are RNAs that sense metabolites and regulate gene expression in response to changes in metabolite concentrations. Typically, they form domains within 5' UTRs of mRNAs, and their ligand binding triggers a folding change that modulates expression of the downstream gene. Therefore, good riboswitch candidates are consistently located in potential 5' UTRs. Most known riboswitches require complex secondary and tertiary structures to form tight and highly selective binding pockets for metabolite ligands. Therefore, motifs that comprise the strongest riboswitch candidates have complex secondary structures and stretches of highly conserved nucleotide positions. Motifs were analyzed manually according to these criteria.

We identified a total of 12 RNA motifs that exhibited these characteristics. Here we report the validation of a new SAM/SAH-binding RNA class, and analysis of other riboswitch candidates. Experimental validation of cyclic di-GMP-II and tetrahydrofolate riboswitches will be reported elsewhere. Details describing additional experimental validation efforts and ligands tested with other riboswitch candidates are presented in Additional File 1.

SAM/SAH-binding RNA

The coenzyme SAM and its reaction by-product SAH are frequently targeted ligands for riboswitches. Three structurally unrelated superfamilies [24] of SAM-binding riboswitches [25] and one SAH-binding riboswitch class [26] have been validated previously. All discriminate against SAM or SAH by orders of magnitude, despite the fact that SAM differs from SAH only by a single methyl group and associated positive charge.

Our current search produced a motif, termed SAM/SAH (Figure 1a), that is found exclusively in the order Rhodobacterales of α-proteobacteria. The RNA motif is consistently found immediately upstream of *metK* genes, which encode SAM synthetase. Because known SAM-binding riboswitches are frequently upstream of *metK* genes [25], the element's gene association suggests that it may function as part of a novel SAM-sensing riboswitch class.

A SAM/SAH RNA from *Roseobacter* sp. SK209-2-6, called "SK209-52 RNA," was subjected to in-line probing [27] in the presence of various concentrations of SAM or SAH (Figure 1b,c). SK209-52 RNA appears to bind SAH with a dissociation constant (K_D) of ~4.3 μ*M* and SAM with a K_D of ~8.6 μ*M* (Figure 1d). Similar results were obtained with SAM/SAH RNA constructs from other species (data not shown). However, because SAM undergoes spontaneous demethylation, SAM samples contain at least some of the breakdown product SAH. Thus, apparent affinity for SAM could result from binding only of contaminating SAH [26]. However, binding assays based on equilibrium dialysis and molecular-recognition experiments indicate that SAM/SAH RNAs do bind SAM (Additional File 1).

Table 1 Motifs identified in this work

Motif	RNA?	*cis*-reg?	Switch?	Taxa	Rfam
6S-flavo	**Y**	N	N	Bacteroidetes	RF01685
aceE	?	y	?	γ-Proteobacteria	
Acido-1	**y**	n	n	Acidobacteria	RF01686
Acido-Lenti-1	**y**	n	n	Acidobacteria, Lentisphaerae	RF01687
Actino-pnp	**Y**	Y	N	Actinomycetales	RF01688
AdoCbl-variant	**Y**	Y	Y	Marine	RF01689
asd	**Y**	?	?	Lactobacillales	RF01732
atoC	**y**	y	?	δ-Proteobacteria	RF01733
Bacillaceae-1	**Y**	n	n	Bacillaceae	RF01690
Bacillus-plasmid	**y**	?	n	*Bacillus*	RF01691
Bacteroid-*trp*	**y**	y	n	Bacteroidetes	RF01692
Bacteroidales-1	**Y**	?	?	Bacteroidales	RF01693
Bacteroides-1	**y**	?	n	*Bacteroides*	RF01694
Bacteroides-2	?	n	n	*Bacteroides*	
Burkholderiales-1	?	?	n	Burkholderiales	
c4 antisense RNA	**Y**	N	N	Proteobacteria, phages	RF01695
c4-a1b1	**Y**	N	N	γ-Proteobacteria, phages	
Chlorobi-1	**Y**	n	n	Chlorobi	RF01696
Chlorobi-RRM	**y**	y	n	Chlorobi	RF01697
Chloroflexi-1	**y**	?	n	*Chloroflexus aggregans*	RF01698
Clostridiales-1	**y**	n	n	Clostridiales, human gut	RF01699
COG2252	?	y	n	Pseudomonadales	
Collinsella-1	**y**	n	n	Actinobacteria, human gut	RF01700
crcB	**Y**	Y	Y	Widespread, bacteria and archaea	RF01734
Cyano-1	**y**	n	n	Cyanobacteria, marine	RF01701
Cyano-2	**Y**	n	n	Cyanobacteria, marine	RF01702
Desulfotalea-1	?	n	n	Proteobacteria	
Dictyoglomi-1	**y**	?	?	Dictyoglomi	RF01703
Downstream-peptide	**Y**	y	y	Cyanobacteria, marine	RF01704
epsC	**Y**	y	y	Bacillales	RF01735
fixA	?	y	n	*Pseudomonas*	
Flavo-1	**y**	n	n	Bacteroidetes	RF01705
flg-Rhizobiales	**y**	y	n	Rhizobiales	RF01736
flpD	**y**	?	n	Euryarchaeota	RF01737
gabT	**Y**	y	?	*Pseudomonas*	RF01738
Gamma-*cis*-1	?	y	n	γ-Proteobacteria	
glnA	**Y**	Y	y	Cyanobacteria, marine	RF01739
GUCCY-hairpin	?	?	n	Bacteroidetes, Proteobacteria	
Gut-1	**Y**	n	n	Human gut only	RF01706
gyrA	**y**	y	n	*Pseudomonas*	RF01740
hopC	**y**	Y	?	*Helicobacter*	RF01741
icd	?	y	n	*Pseudomonas*	
JUMPstart	**y**	Y	?	γ-Proteobacteria	RF01707
L17 downstream element	**y**	y	n	Lactobacillales, *Listeria*	RF01708
lactis-plasmid	**y**	?	n	Lactobacillales	RF01742
Lacto-*int*	?	?	n	Lactobacillales, phages	

Table 1: Motifs identified in this work *(Continued)*

Lacto-*rpoB*	**Y**	y	n	Lactobacillales	RF01709
Lacto-*usp*	**Y**	?	?	Lactobacillales	RF01710
Leu/phe leader	**Y**	Y	N	*Lactococcus lactis*	RF01743
livK	**y**	y	?	Pseudomonadales	RF01744
Lnt	**y**	y	?	Chlorobi	RF01711
manA	**Y**	Y	y	Marine, γ-Proteobacteria, cyanophage	RF01745
Methylobacterium-1	**Y**	n	n	*Methylobacterium*, marine	RF01712
Moco-II	**y**	Y	?	Proteobacteria	RF01713
mraW	**y**	y	?	Actinomycetales	RF01746
msiK	**Y**	Y	?	Actinobacteria	RF01747
Nitrosococcus-1	?	n	n	*Nitrosococcus*, Clostridia	
nuoG	**y**	y	?	Enterobacteriales (incl. *E. coli* K12)	RF01748
Ocean-V	**y**	n	n	Marine only	RF01714
Ocean-VI	?	?	?	Marine only	
pan	**Y**	Y	?	Chloroflexi, Firmicutes, δ-Proteobacteria	RF01749
Pedo-repair	**y**	?	n	*Pedobacter*	RF01715
pfl	**Y**	Y	Y	Several phyla	RF01750
pheA	?	y	n	Actinobacteria	
PhotoRC-I	**y**	y	n	Cyanobacteria, marine	RF01716
PhotoRC-II	**Y**	y	n	Marine, cyanophage	RF01717
Polynucleobacter-1	**y**	y	?	Burkholderiales, fresh water/estuary	RF01718
potC	**y**	y	?	Marine only	RF01751
psaA	**Y**	y	?	Cyanobacteria	RF01752
psbNH	**y**	y	n	Cyanobacteria, marine	RF01753
Pseudomon-1	**y**	n	n	Pseudomonadales	RF01719
Pseudomon-2	?	n	n	*Pseudomonas*	
Pseudomon-GGDEF	?	y	?	*Pseudomonas*	
Pseudomon-groES	**y**	y	?	*Pseudomonas*	RF01721
Pseudomon-Rho	**y**	Y	n	*Pseudomonas*	RF01720
Pyrobac-1	**y**	n	n	*Pyrobaculum*	RF01722
Pyrobac-HINT	?	y	n	*Pyrobaculum*	
radC	**Y**	y	?	Proteobacteria	RF01754
Rhizobiales-1	?	n	N	Rhizobiales	
Rhizobiales-2	**y**	?	n	Rhizobiales	RF01723
Rhodopirellula-1	?	y	?	Proteobacteria, Planctomycetes	
rmf	**Y**	y	?	Pseudomonadales	RF01755
rne-II	**Y**	y	N	Pseudomonadales	RF01756
SAM-Chlorobi	**y**	Y	?	Chlorobi	RF01724
SAM-I-IV-variant	**Y**	Y	Y	Several phyla, marine	RF01725
SAM-II long loops	**Y**	Y	Y	Bacteroidetes, marine	RF01726
SAM/SAH riboswitch	**Y**	Y	Y	Rhodobacterales	RF01727
sanguinis-hairpin	?	n	n	*Streptococcus*	
sbcD	**y**	?	n	Burkholderiales	RF01757
ScRE	?	y	n	*Streptococcus*	
Soil-1	?	n	n	Soil only	
Solibacter-1	?	n	n	*Solibacter usitatus*	
STAXI	**y**	?	n	Enterobacteriales	RF01728

Table 1: Motifs identified in this work *(Continued)*

sucA-II	y	y	?	Pseudomonadales	RF01758
sucC	Y	Y	?	γ-Proteobacteria	RF01759
Termite-*flg*	Y	y	n	Termite hind gut only	RF01729
Termite-*leu*	y	?	?	Termite hind gut only	RF01730
traJ-II	Y	Y	n	Proteobacteria, *Enterococcus faecium*	RF01760
Transposase-resistance	?	y	n	Several phyla	
TwoAYGGAY	y	n	n	Human gut, γ-Proteobacteria, Clostridiales	
wcaG	Y	y	y	Marine, cyanophage	RF01761
Whalefall-1	Y	n	n	Whalefall only	RF01762
yjdF	Y	Y	Y	Firmicutes	RF01764
ykkC-III	y	Y	y	Actinobacteria, δ-Proteobacteria	RF01763

Columns are as follows. "RNA?" : is this motif likely to represent a biological RNA? "Y" = certainly, "y" = probably, "?" = ambiguous, "n" = probably not, "N" = no. "*cis*-reg" : is the motif *cis*-regulatory? "switch?" : is the motif a riboswitch? Additional annotation and justification is in Additional File 2. "Taxa" : common taxon/taxa carrying this motif. Many motifs are discussed only in Additional file 1. "Rfam" : accession numbers of motifs that were submitted to the Rfam database for version 10.1. Note: consensus diagrams of some motifs were presented as supplementary data of a previous report [21] under simplified names: Acido-1 (previously ac-1), Dictyoglomi-1 (dct-1), Gut-1 (gt-1), *manA* (manA), Termite-*flg* (tf-1) and Whalefall-1 (wf-1).

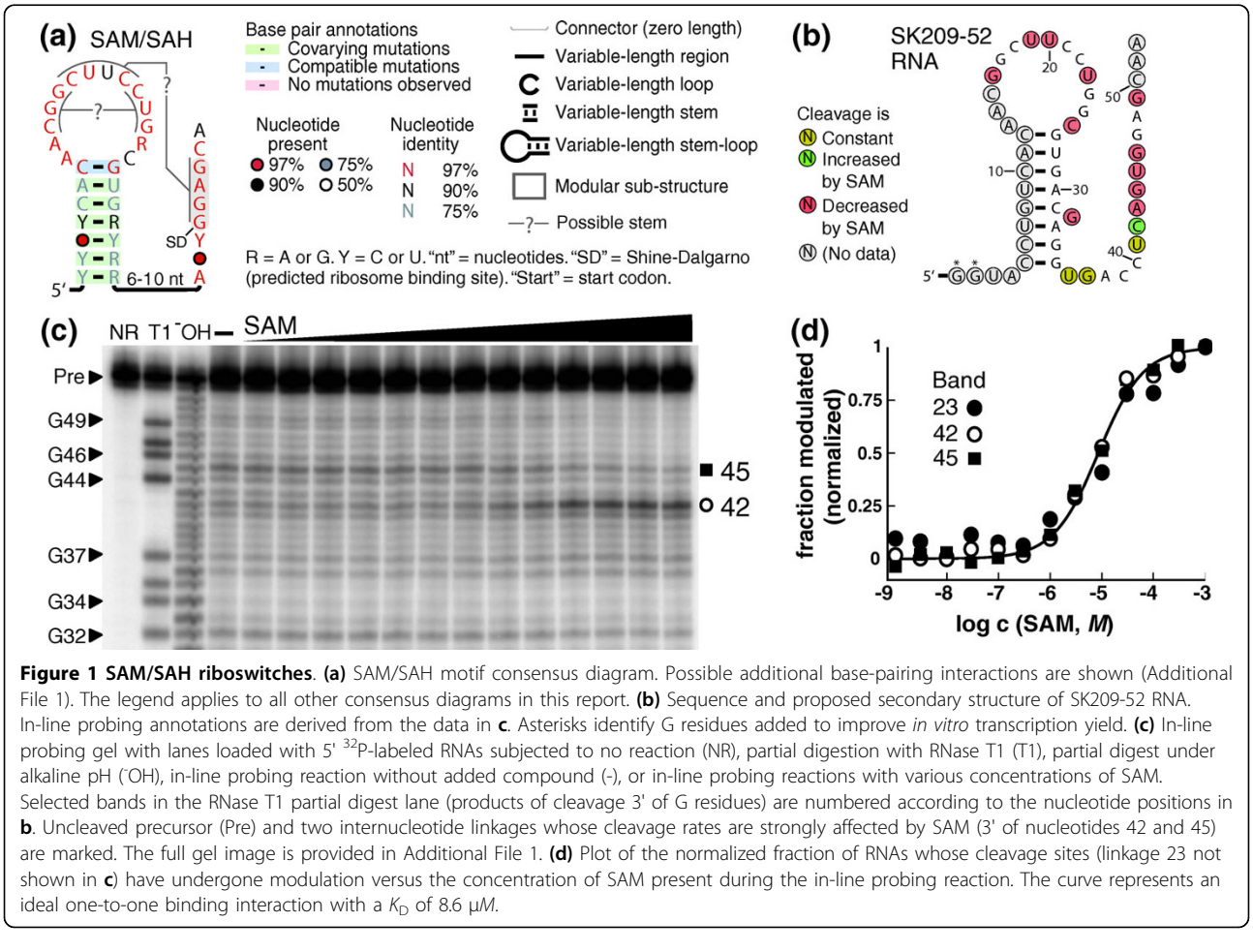

Figure 1 SAM/SAH riboswitches. **(a)** SAM/SAH motif consensus diagram. Possible additional base-pairing interactions are shown (Additional File 1). The legend applies to all other consensus diagrams in this report. **(b)** Sequence and proposed secondary structure of SK209-52 RNA. In-line probing annotations are derived from the data in **c**. Asterisks identify G residues added to improve *in vitro* transcription yield. **(c)** In-line probing gel with lanes loaded with 5' ^{32}P-labeled RNAs subjected to no reaction (NR), partial digestion with RNase T1 (T1), partial digest under alkaline pH (⁻OH), in-line probing reaction without added compound (-), or in-line probing reactions with various concentrations of SAM. Selected bands in the RNase T1 partial digest lane (products of cleavage 3' of G residues) are numbered according to the nucleotide positions in **b**. Uncleaved precursor (Pre) and two internucleotide linkages whose cleavage rates are strongly affected by SAM (3' of nucleotides 42 and 45) are marked. The full gel image is provided in Additional File 1. **(d)** Plot of the normalized fraction of RNAs whose cleavage sites (linkage 23 not shown in **c**) have undergone modulation versus the concentration of SAM present during the in-line probing reaction. The curve represents an ideal one-to-one binding interaction with a K_D of 8.6 μ*M*.

It is interesting to note that SAM/SAH aptamers, which are the smallest of the SAM and SAH aptamer classes, presumably cannot discriminate strongly against SAH. This lack of discrimination may mean that genes associated with this RNA are purposefully regulated by either SAM or SAH. However, SAM is more abundant in cells than is SAH [28]. This fact, coupled with the frequent association of the RNA motif with *metK* gene contexts of SAM/SAH RNAs, suggests that their biologic role is to function as part of a SAM-responsive riboswitch.

crcB motif

The *crcB* motif (Figure 2) is detected in a wide variety of phyla in bacteria and archaea. Thus, *crcB* RNAs join only one known riboswitch class (TPP) [29], and few other classes of RNAs, that are present in more than one domain of life. The *crcB* motif consistently resides in the potential 5' UTRs of genes, including those involved in DNA repair (*mutS*), K^+, or Cl^- transport, or genes encoding formate hydrogen lyase. In many cases, predicted transcription terminators overlap the

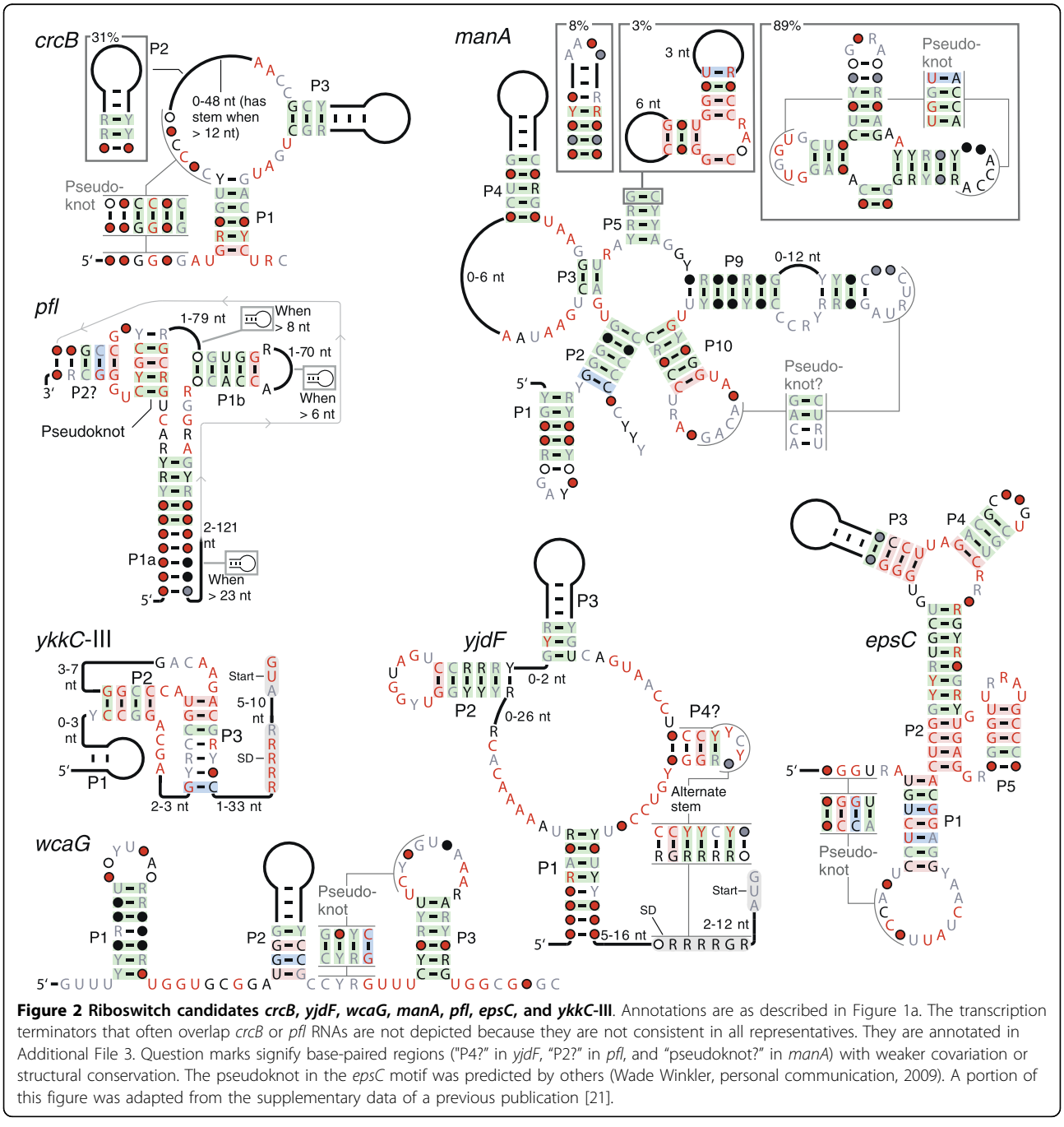

Figure 2 Riboswitch candidates *crcB*, *yjdF*, *wcaG*, *manA*, *pfl*, *epsC*, and *ykkC*-III. Annotations are as described in Figure 1a. The transcription terminators that often overlap *crcB* or *pfl* RNAs are not depicted because they are not consistent in all representatives. They are annotated in Additional File 3. Question marks signify base-paired regions ("P4?" in *yjdF*, "P2?" in *pfl*, and "pseudoknot?" in *manA*) with weaker covariation or structural conservation. The pseudoknot in the *epsC* motif was predicted by others (Wade Winkler, personal communication, 2009). A portion of this figure was adapted from the supplementary data of a previous publication [21].

conserved *crcB* motif. Therefore, if ligand binding of the putative riboswitch stabilizes the conserved structure predicted for these RNAs, higher ligand concentrations are expected to inhibit terminator stem formation and increase gene expression.

The *crcB* motif might regulate genes in response to stress conditions that can damage DNA and be mitigated by increased expression of other genes controlled by the RNAs (Additional File 1). If *crcB* RNAs are riboswitches, they presumably sense a metabolite present in organisms that is indicative of a common cellular condition in two domains of life.

pfl motif

The *pfl* motif (Figure 2) is found in four bacterial phyla. As with *crcB* RNAs, predicted transcription terminators overlap the 3' region of many *pfl* RNAs; thus, gene expression is likely increased in response to higher ligand concentrations. The genes most commonly associated with *pfl* RNAs are related to purine biosynthesis, or to synthesis of formyltetrahydrofolate (formyl-THF), which is used for purine biosynthesis. These genes include *purH*, *fhs*, *pfl*, *glyA*, and *folD*. PurH formylates AICAR by using formyl-THF as the donor. Formyl-THF can be synthesized by the product of *fhs* by using formate and THF as substrates. Formate, in turn, is produced in the reaction catalyzed by Pfl. The upregulation of Pfl to create formate for the synthesis of purines was observed previously [30]. Formyl-THF can also be produced from THF and serine by the combined action of GlyA and FolD. Thus, the five genes most commonly predicted to be regulated by *pfl* RNAs have a role in the synthesis of purines or formyl-THF. Most other genes apparently regulated by *pfl* RNAs (Additional File 3) encode enzymes that perform other steps in purine synthesis, or convert between THF or its 1-carbon adducts at least as a side effect (for example, *metH*) (Additional File 1).

yjdF motif

The *yjdF* motif (Figure 2) is found in many Firmicutes, including *Bacillus subtilis*. In most cases, it resides in potential 5' UTRs of homologues of the *yjdF* gene (Additional File 7), whose function is unknown. However, in *Streptococcus thermophilus*, a *yjdF* RNA motif is associated with an operon whose protein products synthesize nicotinamide adenine dinucleotide (NAD^+) (Additional File 3). Also, the *S. thermophilus yjdF* RNA lacks typical *yjdF* motif consensus features downstream of and including the P4 stem. Thus, if *yjdF* RNAs are riboswitch aptamers, the *S. thermophilus* RNAs might sense a distinct compound that structurally resembles the ligand bound by other *yjdF* RNAs. Or perhaps these RNAs have an alternate solution to form a similar binding site, as is observed with some SAM riboswitches [24].

manA and wcaG motifs

The *manA* and *wcaG* motifs (Figure 2) are found almost exclusively in marine metagenome sequences, but are each detected in T4-like phages that infect cyanobacteria (Additional File 3). Also, two *manA* RNAs are found in γ-proteobacteria. Remarkably, many phages of cyanobacteria have incorporated genes involved in metabolism, including exopolysaccharide production and photosynthesis [31-33], and some of these cyanophages carry *manA* or *wcaG* RNAs. RNA domains corresponding to the *manA* motif are commonly located in potential 5' UTRs of genes (Additional File 3) involved in mannose or fructose metabolism, nucleotide synthesis, *ibpA* chaperones, and photosynthetic genes. Distinctively, *wcaG* RNAs typically appear to regulate genes related to production of exopolysaccharides or genes that are induced by high-light conditions. Perhaps *manA* and *wcaG* RNAs are used by phages to modify their hosts' metabolism [33], although they may also be exploited by uninfected bacteria.

epsC motif

RNA domains corresponding to the *epsC* motif (Figure 2) are found in potential 5' UTRs of genes related to exopolysaccharide (EPS) synthesis, such as *epsC* [34], in *B. subtilis* and related species. Different species use different chemical subunits in their EPS [35], which acts in processes such as biofilm formation, capsule synthesis, and sporulation [35-37]. If *epsC* RNAs are riboswitches, they might sense an intermediate in EPS synthesis that is common to all bacteria containing *epsC* RNAs. Signalling molecules also regulate EPS synthesis in some bacteria [36,38], and are therefore also candidate riboswitch ligands.

The *epsC* motif was discovered independently by another group and named EAR (W. Winkler, personal communication, 2009). This candidate has been shown to exhibit transcription antitermination activity, likely by directly interacting with protein components of the transcription elongation complex (W. Winkler, personal communication, 2009), and therefore, this RNA motif may not also function as a metabolite-binding RNA. Intriguingly, the JUMPstart sequence motif [39] is found in the 5' UTRs of genes related to polysaccharide synthesis and also is associated with modification of transcriptional elongation [40-43]. We detected a conserved stem-loop structure among JUMPstart elements (Additional File 1).

ykkC-III motif

The previously identified *ykkC* [5] and mini-*ykkC* [14] motifs are associated with genes related to those associated with *ykkC*-III, but these RNAs have distinct conserved sequence and structural features. The new-found *ykkC*-III motif (Figure 2) is in potential 5' UTRs of *emrE* and *speB* genes. *emrE* is the most common gene family

associated with mini-*ykkC* and the second most common to be associated with *ykkC*, and *speB* is also associated with *ykkC* RNAs in many cases (Additional File 8). Although a perfectly conserved ACGA sequence in *ykkC*-III is similar to the less rigidly conserved ACGR terminal loops of mini-*ykkC* RNAs, the structural contexts are different (Additional File 1). All three RNA motifs have characteristics of gene-control elements that regulate similar genes, and perhaps respond to changing concentrations of the same metabolite. However, unlike mini-*ykkC*, whose small and repetitive hairpin architecture is suggestive of protein binding, both *ykkC* and *ykkC*-III exhibit more complex structural features that are suggestive of direct metabolite binding.

glnA and Downstream-peptide motifs

The *glnA* and Downstream peptide motifs carry similar sequence and structural features (Figure 3), although the genes they are associated with are very different. Many genes presumably regulated by *glnA* RNAs are clearly involved in nitrogen metabolism, and include nitrogen regulatory protein P_{II}, glutamine synthetase, glutamate synthase, and ammonium transporters. Another associated gene is PMT1479, which was the most repressed gene when *Prochlorococcus marinus* was starved for nitrogen [44]. Some *glnA* RNAs occur in tandem, which is an arrangement previously associated with more-digital gene regulation [45,46].

The Downstream-peptide motif is found in potential 5' UTRs of cyanobacterial ORFs whose products are typically 17 to 100 amino acids long and are predicted not to belong to a known protein family. We observe a pattern of synonymous mutations and insertions or deletions in multiples of three nucleotides (data not shown), supporting the prediction of a short conserved coding sequence. A previously predicted noncoding RNA called "yfr6" [47] is ~250 nucleotides in length and contains a short ORF. The 5' UTRs of these ORFs correspond to Downstream-peptide RNAs. Although only two full-length yfr6 RNAs were found, 634 Downstream-peptide RNAs were detected, suggesting that only the 5' UTR is conserved. Experiments on yfr6 showed that transcription starts ~20 nucleotides 5' to the proposed Downstream-peptide motif [47]. Also, a Downstream-peptide RNA resides in the potential 5' UTR of a gene that appears to be downregulated in response to nitrogen starvation [47]. A conserved amino acid sequence in predicted proteins associated with Downstream-peptide RNAs hints at a possible regulatory mechanism (Additional File 1). The proposed structural resemblance between *glnA* and Downstream-peptide RNAs suggests they may bind to chemically similar ligands, and previously conducted experiments suggest that both elements downregulate genes in response to nitrogen depletion.

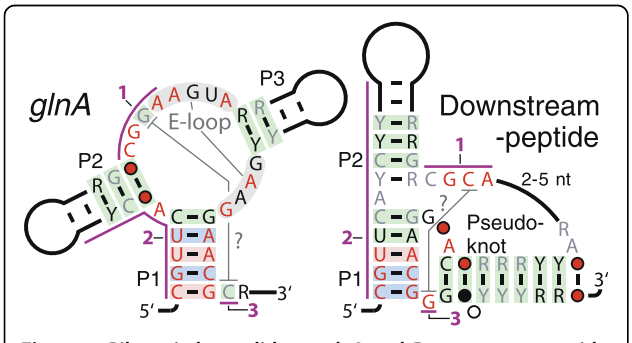

Figure 3 Riboswitch candidates *glnA* and Downstream-peptide. Annotations are as described in Figure 1a. Purple lines and numbers indicate conserved sequences or structures common to the two motifs.

Cyanobacterial photosystem regulatory motifs

psaA motif

Representatives of the *psaA* motif (Figure 4) occur in the potential 5' UTRs of Photosystem-I *psaAB* operons in certain cyanobacteria. The motif includes three hairpins that often include UNCG tetraloops [48]. Although the regulation of *psaAB* genes in species with *psaA* RNAs has not been studied, multiple *psa* genes in *Synechocystis* sp. PCC 6803 are regulated in response to light through DNA elements that are presumably transcription factor-binding sites [49]. Photosynthetic organisms upregulate photosystem-I (*psa*) genes under low-light conditions to maximize energy output, but must reduce their expression under sustained high-light conditions, to avoid damage from free radicals [50]. *psaA* RNAs could be involved in this regulation, although we have not found this RNA element upstream of *psa* genes other than *psaAB*.

PhotoRC-I, PhotoRC-II, and psbNH motifs

Two distinct RNA structures (Figure 4) are associated with genes belonging to the photosynthetic reaction center family of proteins that are probably *psbA*. PhotoRC-I RNAs are present in known cyanobacteria and in marine environmental samples, whereas PhotoRC-II RNAs are detected only in marine samples and a cyanophage. These motifs and *psbNH* are further described in Additional File 1.

Other motifs

L17 downstream element

The L17 downstream element (Additional File 6) is located downstream (within the potential 3' UTRs) of genes that encode ribosomal protein L17. In many cases, no annotated genes are located immediately downstream of the element. Although the motif might actually be transcribed in the opposite orientation, the structure as shown is more stable because it carries

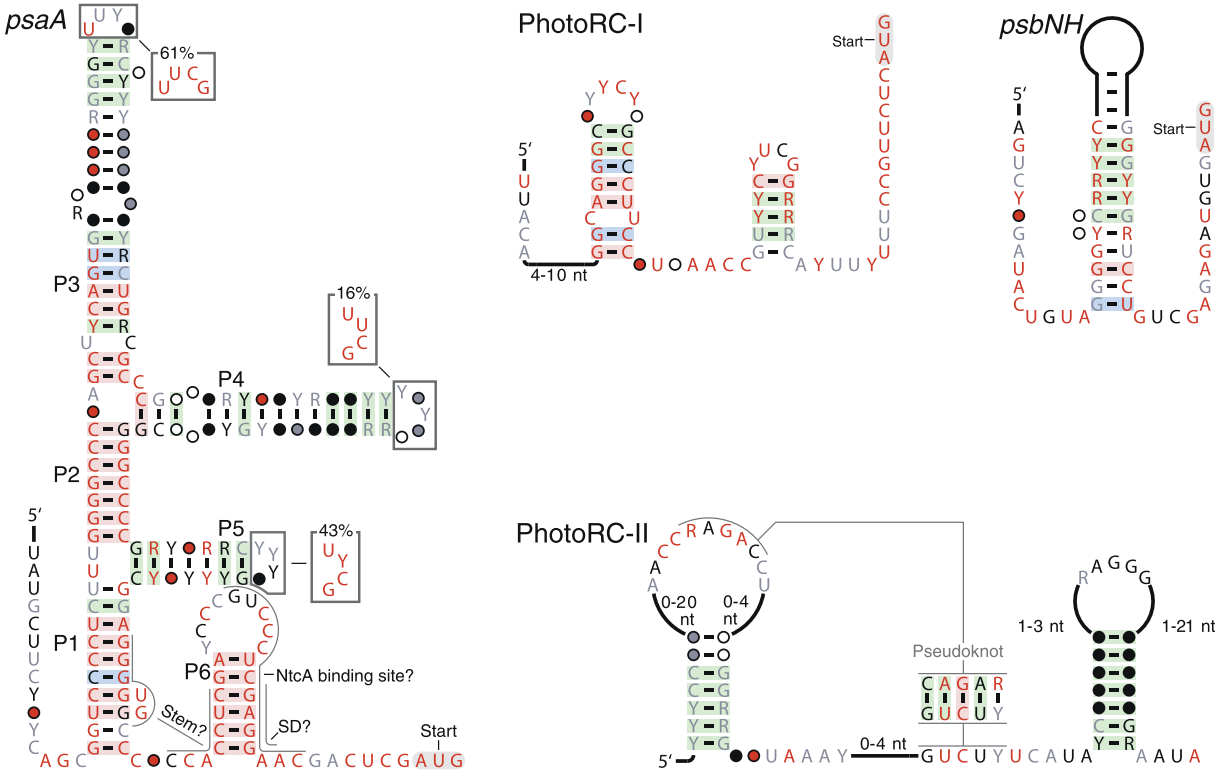

Figure 4 Cyanobacterial motifs related to photosynthesis. Annotations are as described in Figure 1a.

many G-U base pairs and GNRA tetraloops [48]. These structures would be far less stable in the corresponding RNA transcribed from the complementary DNA template. RNA molecules overlapping an L17 downstream element were recently detected by microarrays and designated SR79100 [51]. The expression of ribosomal proteins is frequently regulated by a feedback mechanism in which the protein binds an RNA structure in the 5' UTR of its mRNA, called a ribosomal leader [52]. We did not detect obvious similarity between the L17 downstream element and rRNA, although this situation is typical of ribosomal leaders [53]. Thus, the L17 downstream element could function in the 3' UTR and be part of a feedback-regulation system for L17 production. Regulation of a gene by a structured RNA domain located in the 3' UTR is highly unusual in bacteria. However, precedents include an element in a ribosomal protein operon that regulates both upstream and downstream genes [54], and regulation of upstream genes is observed in a phage [55] and proposed in *Listeria* [56].

***hopC* motif**

The *hopC* motif (Additional File 6) is found in *Helicobacter* species in the potential 5' UTRs of *hopC/alpA* gene and co-transcribed *hopB/alpB* genes. Previous studies established that expression of the *hopCB* operon is increased in response to low pH [57]. The experimentally determined 5' UTRs of the *hopCB* operon mRNA in *H. pylori* 60190 [57] contains a predicted *hopC* motif RNA. HopCB is needed for optimal binding to human epithelial cells [58] and is presumably involved in infection of the human stomach.

***msiK* motif**

The *msiK* motif is always found in the potential 5' UTRs of *msiK* genes [59,60], which encode the ATPase subunit for ABC-type transporters of at least two complex sugars [61], and probably many more [62]. The motif comprises an 11-nucleotide bulge within a long hairpin. The 3' side of the basal pairing region includes a predicted ribosome binding site, which may be part of the regulatory mechanism. Existing data indicate that *msiK* genes are not regulated in response to changing levels of glucose [59,61], so perhaps the RNA participates in a feedback-inhibition loop by binding MsiK proteins (Additional File 1).

***pan* motif**

The *pan* motif (Additional File 6) is found in three phyla and is present in the genetically tractable organism *B. subtilis*. Each *pan* RNA consists of a stem interrupted by two highly conserved bulged A residues. Most *pan* RNAs occur in tandem, and their simple structure and dimeric arrangement is suggestive of a dimeric protein-binding motif. The RNAs are located upstream of

operons containing *panB*, *panC*, or aspartate decarboxylase genes, which are involved in synthesizing pantothenate (vitamin B_5).

rmf motif

The *rmf* motif is found in the potential 5' UTRs of *rmf* genes in *Pseudomonas* species. These genes encode ribosome-modulation factor, which acts in the stringent response to depletion of nutrients and other stressors [63]. Because Rmf interacts with rRNA, the protein Rmf might bind to the 5' UTR of its mRNA. Alternately, because the RNA is relatively far from the *rmf* start codon, *rmf* RNAs might be noncoding RNAs that are expressed separate from the adjacent coding region.

SAM-Chlorobi motif

The SAM-Chlorobi motif is found in the potential 5' UTRs of operons containing all predicted *metK* and *ahcY* genes within the phylum Chlorobi. As noted earlier, *metK* encodes SAM synthetase, and in most other organisms, *metK* homologues are controlled by changing SAM concentrations that are detected by SAM-responsive riboswitches. In contrast, *ahcY* encodes *S*-adenosylhomocysteine (SAH) hydrolase, and this gene is known to be controlled by SAH-responsive riboswitches in some organisms [26]. Sequences conforming to a strong promoter sequences [64,65] imply that SAM-Chlorobi RNAs are transcribed (Additional File 1). However, preliminary analysis of several SAM-Chlorobi RNA constructs by using in-line probing did not reveal binding to SAM or SAH (Additional File 1).

STAXI motif

The Ssbp, Topoisomerase, Antirestriction, XerDC Integrase (STAXI) motif is composed mainly of a pseudoknot structure repeated at least two and usually three times (Figure 5). Tandem STAXI motifs are frequently near to genes that encode proteins that bind or manipulate DNA, including single-stranded DNA-binding proteins (Ssbp), integrases and topoisomerases, or antirestriction proteins. Also, they are occasionally located near c4 antisense RNAs [66] (Additional File 1). Because genes proximal to STAXI representatives encode DNA-manipulation proteins, it is possible that the STAXI motif represents a single-stranded DNA that adopts a local structure when duplex DNA is separated, as occurs during DNA replication, repair, or when bound by some proteins. However, the UUCG tetraloops that frequently occur within the STAXI motif repeats are known to stabilize RNA, whereas the corresponding TTCG are not particularly stabilizing for DNA structures [67]. This suggests that the motif is more likely to serve its function as an RNA structure.

Noncoding RNAs

Several motifs that are most likely expressed as noncoding RNAs unaffiliated with mRNAs also were identified (Figure 5, Table 1). Gut-1 and whalefall-1 RNAs are found only in environmental sequences, and Bacteroides-2 is found in only one sequenced organism (Additional File 1). Thus, bacteria from multiple environmental samples express noncoding RNAs that are not represented in any cultivated organisms whose genomes have been sequenced [68,21]. Similarly, Acido-1 and Dictyoglomi-1 RNAs are found in phyla in which few genome sequences are available. Further observations regarding all noncoding RNA candidates can be found in Additional File 1.

Expansion of representatives of previously characterized structured RNAs

Existing homology search methods for RNAs frequently fail to detect representatives of known RNA classes whose sequences have diverged extensively. However, our computational pipeline occasionally reveals examples of such RNAs. Details regarding RNA representatives that expand the collection of 6S RNAs, AdoCbl riboswitches, SAM-II riboswitches, and SAM-I/SAM-IV riboswitches are provided in Additional File 1. The RNAs that expand the collection of the superfamily of SAM-I [69] and SAM-IV [24] riboswitches (Additional File 6) are typically found in metagenome sequences. These variant SAM-I/SAM-IV riboswitches share many of the structural features of both families (Additional File 6), but lack an internal loop in the P2 stem, which is present in SAM-I/SAM-IV riboswitches (Additional File 1).

Conclusions

Numerous structured RNA candidates have been identified in the genomic and metagenomic DNA sequence data from bacteria and archaea. The predicted RNAs exhibit a great diversity of conserved sequences and structural features, and their genomic locations are indicative of a wide variety of mechanisms of action (for example, *cis* vs. *trans*) and putative biologic roles. Our findings suggest that the bacterial and archaeal domains of life will continue to be a rich source of novel structured RNAs.

Although some of the RNAs identified perform the same function as previously validated RNA classes (for example, 6S-Flavo RNA, SAM/SAH riboswitches), the vast majority of the predicted RNA motifs are likely to perform novel functions. Given that many of these RNAs are specific to certain lineages or uncultivated environmental samples, technologies that more rapidly make available DNA sequence information from additional lineages of bacteria and archaea are likely to accelerate the discovery of more classes of structured RNAs. This discovery rate might also be increased by improvements in computational analysis methods. These findings should yield a diverse collection of structured noncoding RNAs that will reveal a more complete understanding of the roles that RNAs perform in microbial cells.

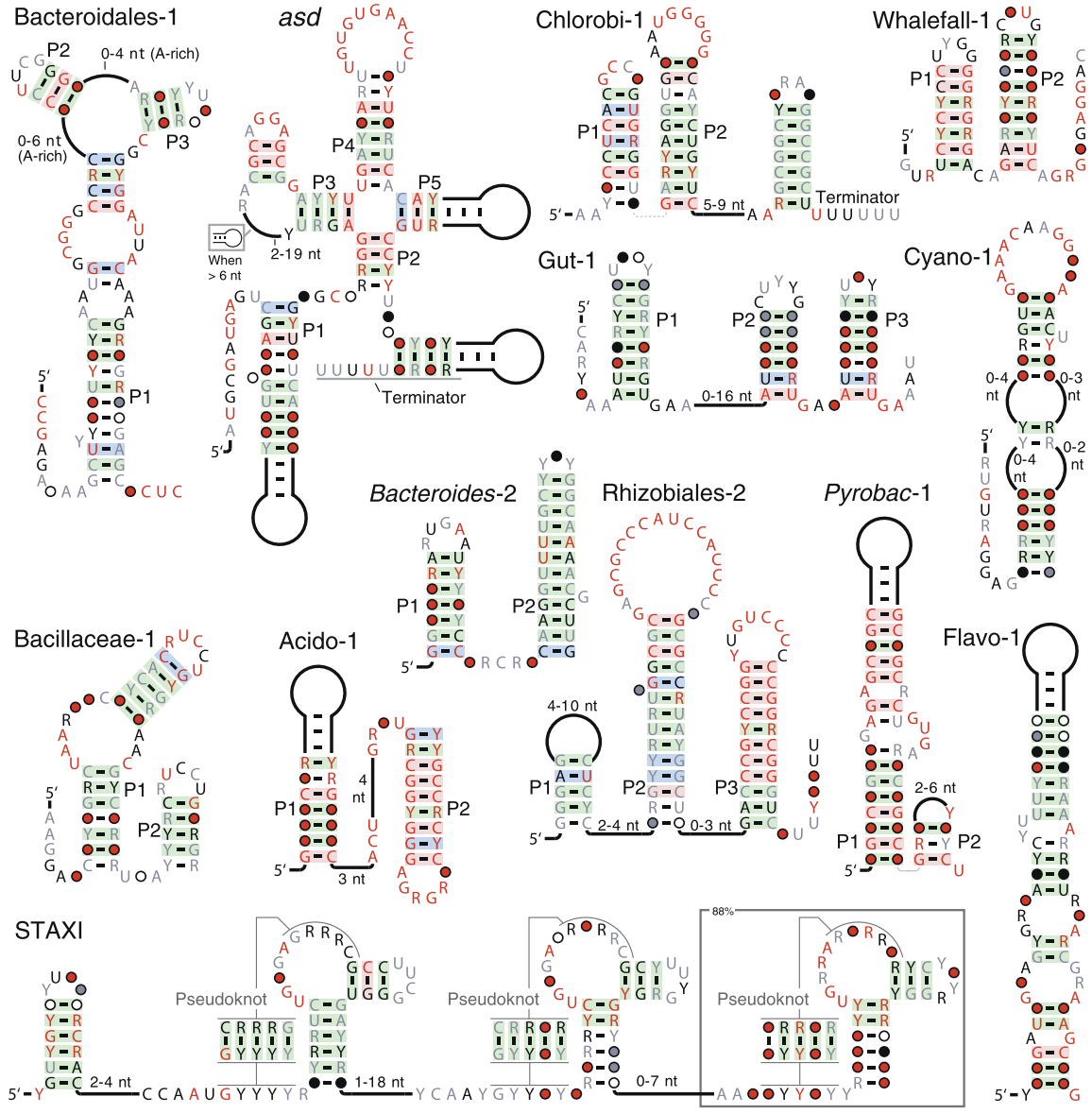

Figure 5 Examples of other candidate RNAs. Annotations are as described in Figure 1a. The Bacteroidales-1 motif has more conserved nucleotides than depicted (Additional File 6). A portion of this figure was adapted from the supplementary data of a previous publication [21].

Materials and methods

DNA sequence sources and gene annotations

The microbial subsets of RefSeq [70] version 25 or 32 (Additional file 9) were searched, along with metagenome sequences from acid mine drainage [71], soil and whale fall [72], human gut [73,74], mouse gut [75], gutless sea worms [76], sludge [77], Global Ocean Survey scaffolds [78,79], other marine sequences [80], and termite hindgut [81]. Locations and identities of protein-coding genes were derived from RefSeq or IMG/M [82] annotations, or from "predicted proteins" [83] in Global Ocean Survey sequences. However, genes in some sequences [74,80,81] were predicted by using MetaGene (dated Oct. 12, 2006) with default parameters [84]. Conserved protein domains were annotated by using the Conserved Domain Database version 2.08 [85].

Annotations for tRNAs and rRNAs were derived from the sources noted earlier, or were predicted by using tRNAscan-SE [86] run in bacterial mode. To detect additional rRNAs, annotated rRNAs whose descriptions read "ribosomal RNA" or "#S rRNA" (# represents any number) were used in WU-BLAST queries with command-line flags -hspsepQmax = 4000 -E 1e-20 -W 8 [13]. Other RNAs were detected with Rfam [22] and

WU-BLAST, as described previously [13]. We also used published alignments of riboswitches [87] as queries with RaveNnA global-mode searches [19,20], selecting hits manually based primarily on E-values.

Automated motif identification

To reduce false positives in sequence comparisons, the pipeline was run separately on related taxa or metagenome sources (Additional File 9). For each run, InterGenic Regions (IGRs) of at least 30 nucleotides were extracted between protein-coding, tRNA and rRNA genes.

To generate clusters, an early version of a recently described algorithm was used [16]. Specifically, IGRs were compared by using nucleotide NCBI BLAST [17] version 2.2.17 and parameters -W 7 -G 2 -E 2 -q -2 -m 8. Self-matches were ignored. BLAST scores below a parameter S (see later) were considered insignificant and were ignored. Each BLAST match defines two "nodes," corresponding to the matching sequences. Nodes that overlap by at least five nucleotides are merged, along with their BLAST homologies. A cluster consists of all nodes that have direct or indirect (transitive) BLAST matches. Closely related sequences that span multiple distinct elements in an entire IGR can lead to spurious node merges. Therefore, homologies with BLAST scores >100 are ignored.

If a node's length in nucleotides is L, and L < 500, then the node is extended on either side by (500-L)/2 nucleotides, but is constrained to remain within the original IGR. CMfinder can easily tolerate nodes of 500 nucleotides. When L > 1,000, nodes are shrunk by (L - 1,000)/2 nucleotides around the center. The L > 1,000 case is extremely rare. Only clusters with at least three members were reported.

For each pipeline run, we tried a range of values for the parameter S = 35, 40, ..., 85, and determined how many known RNAs were detected with each value. Based on these data, a set of S values was selected manually, and the union of clusters arising from each S was used as input to CMfinder [18]. CMfinder was used to predict motifs exactly as before [13]. Automated homology searches were then performed as described [13], except that covariance model scores used the null3 model [88]. Motifs were scored by using a previously established method [13], and by using tools comprising Pfold [89] to infer a phylogenetic tree, and then running pscore [90]. We also automatically eliminated motifs that had no covarying base-pair positions, that had an average G+C content <24%, that had representatives whose nucleotide coordinates overlapped the reverse-complements of other representatives on average by ≥30% of their nucleotides, or that had fewer than six positions that were ≥97% conserved (when sequences were weighted with the GSC algorithm). Source code is provided (Additional File 10).

Manual analysis of motifs

The manual analysis of each candidate RNA motif proceeded essentially as described previously [14]. For motifs that were likely to be *cis*-regulatory, we routinely searched for articles referencing the locus tags of apparently regulated genes, by using Google Scholar [91]. We also used mutual information analysis [87] to predict additional base-pairing interactions. Motifs less likely to represent structured RNAs were rejected by using previously established criteria [14]. In motif consensus diagrams, covariation and levels of conservation were calculated using earlier protocols [14], but ≤10% noncanonic pairs were tolerated in alignment columns that correspond to conserved base pairs. RNAs were drawn with R2R (Z.W., R.R.B., unpublished software) and Adobe Illustrator.

Assessing the novelty of motifs

To determine whether the predicted RNA structures were reported previously, we searched the Rfam database [22], and various articles not yet incorporated into Rfam that performed detailed analysis or experiments on new-found candidate RNAs [10,47,92-110]. Although some raw predictions of a previous report [9] overlap some of our RNA motifs (Additional File 11), these raw predictions have never been subjected to detailed evaluation. Additionally, extensive Google searches [111] for genes associated with *crcB* RNAs revealed that one of the 358 raw predictions of conserved elements on the RibEx web server [112] overlaps several of the *crcB* RNAs we found. This conserved element was called RLE0038 and was not previously subjected to detailed evaluation. We have not determined whether other coinciding predictions are present on this web server because its data are not available in a machine-readable format.

In-line probing experiments

RNA constructs were prepared by *in vitro* RNA transcription by using T7 RNA polymerase and the appropriate DNA templates that were created by overlap extension of synthetic DNA oligonucleotides by using SuperScript II reverse transcriptase (Invitrogen), as instructed by the manufacturer. RNA transcripts were purified by using denaturing (8 *M* urea) polyacrylamide gel electrophoresis (PAGE). RNAs were eluted from the gel, dephosphorylated by using alkaline phosphatase, and 5' radiolabeled with [γ-^{32}P] by using methods reported previously [26]. 5' ^{32}P-labeled fragments resulting from in-line probing reactions were subjected to denaturing PAGE, and were imaged and analyzed as previously described [26].

Equilibrium dialysis experiments

Equilibrium dialysis experiments were conducted in a Dispo-Equilibrium Biodialyzer (The Nest Group, Inc., Southboro, MA, USA), which comprises two chambers

(A and B) separated by a 5,000-kDa MW cut-off membrane. Chamber A was loaded with 20 μl solution of 500 n*M* ^{3}H-SAM, and Chamber B was loaded with 20 μ*M* specified RNA in a buffer containing 50 m*M* MOPS (pH 7.2 at 20°C), 20 m*M* $MgCl_2$, and 500 m*M* KCl. The chambers were equilibrated at 25°C for 10 h before a 3-μl aliquot was removed from each chamber. Radioactivity of the aliquots was measured with a liquid scintillation counter. Each experiment was repeated 3 times, and average B/A values and standard deviations were calculated.

Additional file 1: Supplementary results and discussion. Additional analysis of motifs, including those not discussed in the manuscript, and in-line probing experiments on riboswitch candidates.

Additional file 2: Summary and evaluation of all motifs. Table 1, with summary of supporting evidence, and numbers of representatives of each motif.

Additional file 3: Taxa of motif representatives, genes flanking representatives and annotated multiple-sequence alignments. For each motif, this file shows the taxa of each motif representative, depicts genes flanking these representatives and describes conserved domains that the genes encode. Also, a multiple-sequence alignment is provided for each motif, and includes secondary structure and other annotations.

Additional file 4: Raw text alignment files, including annotation. Raw alignments of RNAs, including annotations (for example, predicted transcription terminators, flanking sequences) in "Stockholm" text format. The alignment format and appropriate viewing programs are discussed on Wikipedia [113]. The Stockholm files can be retrieved from the .tar.gz archive file by using programs such as WinZip (Windows), StuffIt Expander (Mac), or tar/gzip (UNIX).

Additional file 5: Raw text alignment files, just the motifs. Raw alignments of RNA motifs with minimal annotation and no flanking sequences, in "Stockholm" text format. The Stockholm files can be retrieved from the .tar.gz archive file by using programs such as WinZip (Windows), StuffIt Expander (Mac), or tar/gzip (UNIX).

Additional file 6: Consensus diagrams of all motifs. Consensus diagrams depicting all motifs in high resolution.

Additional file 7: Alignment of YjdF proteins. Multiple-sequence alignment of proteins predicted to be homologous to YjdF of *Bacillus subtilis*.

Additional file 8: Genes associated with *ykkC*, mini-*ykkC* and *ykkC*-III RNAs. The frequencies with which various gene families are associated with *ykkC*, mini-*ykkC* or *ykkC*-III RNAs are listed.

Additional file 9: Partitioning of genomes and metagenomes. Describes how genomes and metagenomes were divided into pipeline runs.

Additional file 10: Source code implemented as part of this project. Source code files and a README.pdf file are provided to assist in detailed understanding of the methods. The files can be retrieved from the .tar.gz archive file, as described for Additional file 4.

Additional file 11: Overlap with previous raw predictions. Overlaps of our RNA motifs with raw predictions of a prior study [9]. Tab-delimited text file.

Acknowledgements
We thank Nick Carriero and Rob Bjornson for assisting our use of the Yale Life Sciences High Performance Computing Center (NIH grant RR19895-02), Paul Gardner for sharing a list of recently published RNA discovery articless, and Adam Roth, Narisiman Sudarsan, Michelle Meyer, Jonathan Perreault, Jeff Barrick, Zizhen Yao, Elizabeth Tseng, Larry Ruzzo, and Breaker lab members for helpful comments. R.R.B. is a Howard Hughes Medical Institute Investigator.

Author details
[1]Howard Hughes Medical Institute, Yale University, P.O. Box 208103, New Haven, CT 06520-8103, USA. [2]Department of Molecular, Cellular and Developmental Biology, Yale University, P.O. Box 208103, New Haven, CT 06520-8103, USA. [3]Department of Molecular Biophysics and Biochemistry, Yale University, P.O. Box 208103, New Haven, CT 06520-8103, USA. [4]Current address: Department of Biology, University of Rochester, Rochester, NY 14627, USA. [5]Current address: School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

Authors' contributions
ZW and RRB conceived of the study, ZW prepared bioinformatics scripts, and RRB supervised the study. ZW, JY, KC, RM, and JXW analyzed motif predictions to infer conserved RNA structures. JXW, ZW, and JB tested riboswitch candidates by using in-line probing. JXW and JB conducted SAM/SAH experiments. ZW and RRB wrote the manuscript, with assistance from all authors.

Received: 18 November 2009 Revised: 18 January 2010
Accepted: 15 March 2010 Published: 15 March 2010

References

1. Roth A, Breaker RR: **The structural and functional diversity of metabolite-binding riboswitches.** *Annu Rev Biochem* 2009, **78**:305-335.
2. Waters LS, Storz G: **Regulatory RNAs in bacteria.** *Cell* 2009, **136**:615-628.
3. Narberhaus F, Vogel J: **Regulatory RNAs in prokaryotes: here, there and everywhere.** *Mol Microbiol* 2009, **74**:261-269.
4. Rivas E, Eddy SR: **Noncoding RNA gene detection using comparative sequence analysis.** *BMC Bioinformatics* 2001, **2**:8.
5. Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I, Wickiser JK, Breaker RR: **New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control.** *Proc Natl Acad Sci USA* 2004, **101**:6421-6426.
6. Corbino KA, Barrick JE, Lim J, Welz R, Tucker BJ, Puskarz I, Mandal M, Rudnick ND, Breaker RR: **Evidence for a second class of *S*-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria.** *Genome Biol* 2005, **6**:R70.
7. Meyer IM: **A practical guide to the art of RNA gene prediction.** *Brief Bioinform* 2007, **8**:396-414.
8. Meyer MM, Ames TD, Smith DP, Weinberg Z, Schwalbach MS, Giovannoni SJ, Breaker RR: **Identification of candidate structured RNAs in the marine organism *Candidatus* 'Pelagibacter ubique'.** *BMC Genomics* 2009, **10**:268.
9. Livny J, Teonadi H, Livny M, Waldor MK: **High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs.** *PLoS One* 2008, **3**:e3197.
10. Marchais A, Naville M, Bohn C, Bouloc P, Gautheret D: **Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles.** *Genome Res* 2009, **19**:1084-1092.
11. Klein RJ, Misulovin Z, Eddy SR: **Noncoding RNA genes identified in AT-rich hyperthermophiles.** *Proc Natl Acad Sci USA* 2002, **99**:7542-7547.
12. Schattner P: **Searching for RNA genes using base-composition statistics.** *Nucleic Acids Res* 2002, **30**:2076-2082.
13. Yao Z, Barrick J, Weinberg Z, Neph S, Breaker R, Tompa M, Ruzzo WL: **A computational pipeline for high-throughput discovery of *cis* -regulatory noncoding RNA in prokaryotes.** *PLoS Comput Biol* 2007, **3**:e126.
14. Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, Wang JX, Lee ER, Block KF, Sudarsan N, Neph S, Tompa M, Ruzzo WL, Breaker RR: **Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline.** *Nucleic Acids Res* 2007, **35**:4809-4819.
15. 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**:411-413.
16. Tseng HH, Weinberg Z, Gore J, Breaker RR, Ruzzo WL: **Finding non-coding RNAs through genome-scale clustering.** *J Bioinform Comput Biol* 2009, **7**:373-388.
17. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.

18. Yao Z, Weinberg Z, Ruzzo WL: **CMfinder–a covariance model based RNA motif finding algorithm.** *Bioinformatics* 2006, **22**:445-452.
19. Weinberg Z, Ruzzo WL: **Sequence-based heuristics for faster annotation of non-coding RNA families.** *Bioinformatics* 2006, **22**:35-39.
20. Eddy SR, Durbin R: **RNA sequence analysis using covariance models.** *Nucleic Acids Res* 1994, **22**:2079-2088.
21. Weinberg Z, Perreault J, Meyer MM, Breaker RR: **Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis.** *Nature* 2009, **462**:656-659.
22. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A: **Rfam: updates to the RNA families database.** *Nucleic Acids Res* 2009, **37**: D136-140.
23. Montange RK, Batey RT: **Riboswitches: emerging themes in RNA structure and function.** *Annu Rev Biophys* 2008, **37**:117-133.
24. Weinberg Z, Regulski EE, Hammond MC, Barrick JE, Yao Z, Ruzzo WL, Breaker RR: **The aptamer core of SAM-IV riboswitches mimics the ligand-binding site of SAM-I riboswitches.** *RNA* 2008, **14**:822-828.
25. Wang JX, Breaker RR: **Riboswitches that sense *S*-adenosylmethionine and *S*-adenosylhomocysteine.** *Biochem Cell Biol* 2008, **86**:157-168.
26. Wang JX, Lee ER, Morales DR, Lim J, Breaker RR: **Riboswitches that sense *S*-adenosylhomocysteine and activate genes involved in coenzyme recycling.** *Mol Cell* 2008, **29**:691-702.
27. Soukup GA, Breaker RR: **Relationship between internucleotide linkage geometry and the stability of RNA.** *RNA* 1999, **5**:1308-1325.
28. Ueland PM: **Pharmacological and biochemical aspects of *S* -adenosylhomocysteine and *S*-adenosylhomocysteine hydrolase.** *Pharmacol Rev* 1982, **34**:223-253.
29. Sudarsan N, Barrick JE, Breaker RR: **Metabolite-binding RNA domains are present in the genes of eukaryotes.** *RNA* 2003, **9**:644-647.
30. Derzelle S, Bolotin A, Mistou MY, Rul F: **Proteome analysis of *Streptococcus thermophilus* grown in milk reveals pyruvate formate-lyase as the major upregulated protein.** *Appl Environ Microbiol* 2005, **71**:8597-8605.
31. Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW: **Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations.** *PLoS Biol* 2005, **3**:e144.
32. Rohwer F, Thurber RV: **Viruses manipulate the marine environment.** *Nature* 2009, **459**:207-212.
33. Lindell D, Jaffe JD, Coleman ML, Futschik ME, Axmann IM, Rector T, Kettler G, Sullivan MB, Steen R, Hess WR, Church GM, Chisholm SW: **Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution.** *Nature* 2007, **449**:83-86.
34. Lemon KP, Earl AM, Vlamakis HC, Aguilar C, Kolter R: **Biofilm development with an emphasis on *Bacillus subtilis*.** *Curr Top Microbiol Immunol* 2008, **322**:1-16.
35. Leoff C, Saile E, Sue D, Wilkins P, Quinn CP, Carlson RW, Kannenberg EL: **Cell wall carbohydrate compositions of strains from the *Bacillus cereus* group of species correlate with phylogenetic relatedness.** *J Bacteriol* 2008, **190**:112-121.
36. Nakhamchik A, Wilde C, Rowe-Magnus DA: **Cyclic-di-GMP regulates extracellular polysaccharide production, biofilm formation, and rugose colony development by *Vibrio vulnificus*.** *Appl Environ Microbiol* 2008, **74**:4199-4209.
37. Torres-Cabassa A, Gottesman S, Frederick RD, Dolph PJ, Coplin DL: **Control of extracellular polysaccharide synthesis in *Erwinia stewartii* and *Escherichia coli* K-12: a common regulatory function.** *J Bacteriol* 1987, **169**:4525-4531.
38. Liang W, Silva AJ, Benitez JA: **The cyclic AMP receptor protein modulates colonial morphology in *Vibrio cholerae*.** *Appl Environ Microbiol* 2007, **73**:7482-7487.
39. Hobbs M, Reeves PR: **The JUMPstart sequence: a 39 bp element common to several polysaccharide gene clusters.** *Mol Microbiol* 1994, **12**:855-856.
40. Marolda CL, Valvano MA: **Promoter region of the *Escherichia coli* O7-specific lipopolysaccharide gene cluster: structural and functional characterization of an upstream untranslated mRNA sequence.** *J Bacteriol* 1998, **180**:3070-3079.
41. Nieto JM, Bailey MJ, Hughes C, Koronakis V: **Suppression of transcription polarity in the *Escherichia coli* haemolysin operon by a short upstream element shared by polysaccharide and DNA transfer determinants.** *Mol Microbiol* 1996, **19**:705-713.
42. Leeds JA, Welch RA: **Enhancing transcription through the *Escherichia coli* hemolysin operon, *hlyCABD*: RfaH and upstream JUMPStart DNA sequences function together via a postinitiation mechanism.** *J Bacteriol* 1997, **179**:3519-3527.
43. Wang L, Jensen S, Hallman R, Reeves PR: **Expression of the O antigen gene cluster is regulated by RfaH through the JUMPstart sequence.** *FEMS Microbiol Lett* 1998, **165**:201-206.
44. Tolonen AC, Aach J, Lindell D, Johnson ZI, Rector T, Steen R, Church GM, Chisholm SW: **Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability.** *Mol Syst Biol* 2006, **2**:53.
45. Mandal M, Lee M, Barrick JE, Weinberg Z, Emilsson GM, Ruzzo WL, Breaker RR: **A glycine-dependent riboswitch that uses cooperative binding to control gene expression.** *Science* 2004, **306**:275-279.
46. Welz R, Breaker RR: **Ligand binding and gene control characteristics of tandem riboswitches in *Bacillus anthracis*.** *RNA* 2007, **13**:573-582.
47. Axmann IM, Kensche P, Vogel J, Kohl S, Herzel H, Hess WR: **Identification of cyanobacterial non-coding RNAs by comparative genome analysis.** *Genome Biol* 2005, **6**:R73.
48. Pace NR, Thomas BC, Woese CR: **Probing RNA structure, function, and history by comparative analysis.** *The RNA World* Cold Spring Harbor, New York: Cold Spring Harbor Laboratory PressGesteland RF, Cech TR, Atkins JF , 2 1999, 113-141.
49. Muramatsu M, Hihara Y: **Coordinated high-light response of genes encoding subunits of Photosystem I is achieved by AT-rich upstream sequences in the cyanobacterium *Synechocystis* sp. strain PCC 6803.** *J Bacteriol* 2007, **189**:2750-2758.
50. Muramatsu M, Hihara Y: **Characterization of high-light-responsive promoters of the *psaAB* genes in *Synechocystis* sp. PCC 6803.** *Plant Cell Physiol* 2006, **47**:878-890.
51. Perez N, Trevino J, Liu Z, Ho SC, Babitzke P, Sumby P: **A genome-wide analysis of small regulatory RNAs in the human pathogen group A *Streptococcus*.** *PLoS One* 2009, **4**:e7668.
52. Zengel JM, Lindahl L: **Diverse mechanisms for regulating ribosomal protein synthesis in *Escherichia coli*.** *Prog Nucleic Acid Res Mol Biol* 1994, **47**:331-370.
53. Batey RT: **Structures of regulatory elements in mRNAs.** *Curr Opin Struct Biol* 2006, **16**:299-306.
54. Mattheakis L, Vu L, Sor F, Nomura M: **Retroregulation of the synthesis of ribosomal proteins L14 and L24 by feedback repressor S8 in *Escherichia coli*.** *Proc Natl Acad Sci USA* 1989, **86**:448-452.
55. Guarneros G, Montanez C, Hernandez T, Court D: **Posttranscriptional control of bacteriophage lambda gene expression from a site distal to the gene.** *Proc Natl Acad Sci USA* 1982, **79**:238-242.
56. Toledo-Arana A, Dussurget O, Nikitas G, Sesto N, Guet-Revillet H, Balestrino D, Loh E, Gripenland J, Tiensuu T, Vaitkevicius K, Barthelemy M, Vergassola M, Nahori MA, Soubigou G, Regnault B, Coppee JY, Lecuit M, Johansson J, Cossart P: **The *Listeria* transcriptional landscape from saprophytism to virulence.** *Nature* 2009, **459**:950-956.
57. McGowan CC, Necheva AS, Forsyth MH, Cover TL, Blaser MJ: **Promoter analysis of *Helicobacter pylori* genes with enhanced expression at low pH.** *Mol Microbiol* 2003, **48**:1225-1239.
58. Odenbreit S, Faller G, Haas R: **Role of the AlpAB proteins and lipopolysaccharide in adhesion of *Helicobacter pylori* to human gastric tissue.** *Int J Med Microbiol* 2002, **292**:247-256.
59. Hurtubise Y, Shareck F, Kluepfel D, Morosoli R: **A cellulase/xylanase-negative mutant of *Streptomyces lividans* 1326 defective in cellobiose and xylobiose uptake is mutated in a gene encoding a protein homologous to ATP-binding proteins.** *Mol Microbiol* 1995, **17**:367-377.
60. Parche S, Amon J, Jankovic I, Rezzonico E, Beleut M, Barutcu H, Schendel I, Eddy MP, Burkovski A, Arigoni F, Titgemeyer F: **Sugar transport systems of *Bifidobacterium longum* NCC2705.** *J Mol Microbiol Biotechnol* 2007, **12**:9-19.
61. Schlösser A, Kampers T, Schrempf H: **The *Streptomyces* ATP-binding component MsiK assists in cellobiose and maltose transport.** *J Bacteriol* 1997, **179**:2092-2095.
62. Bertram R, Schlicht M, Mahr K, Nothaft H, Saier MH Jr, Titgemeyer F: **In silico and transcriptional analysis of carbohydrate uptake systems of *Streptomyces coelicolor* A3(2).** *J Bacteriol* 2004, **186**:1362-1373.
63. Niven GW, El-Sharoud WM: **Ribosome modulation factor.** *Bacterial Physiology: A Molecular Approach* Berlin: Springer-VerlagEl-Sharoud WM 2008, 293-311.

64. Bayley DP, Rocha ER, Smith CJ: **Analysis of *cepA* and other *Bacteroides fragilis* genes reveals a unique promoter structure.** *FEMS Microbiol Lett* 2000, **193**:149-154.
65. Chen S, Bagdasarian M, Kaufman MG, Walker ED: **Characterization of strong promoters from an environmental *Flavobacterium hibernum* strain by using a green fluorescent protein-based reporter system.** *Appl Environ Microbiol* 2007, **73**:1089-1100.
66. Citron M, Schuster H: **The c4 repressors of bacteriophages P1 and P7 are antisense RNAs.** *Cell* 1990, **62**:591-598.
67. Antao VP, Tinoco I Jr: **Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops.** *Nucleic Acids Res* 1992, **20**:819-824.
68. Shi Y, Tyson GW, DeLong EF: **Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column.** *Nature* 2009, **459**:266-269.
69. Winkler WC, Nahvi A, Sudarsan N, Barrick JE, Breaker RR: **An mRNA structure that controls gene expression by binding *S* -adenosylmethionine.** *Nat Struct Biol* 2003, **10**:701-707.
70. Pruitt K, Tatusova T, Maglott D: **NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, **33**:D501-D504.
71. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment.** *Nature* 2004, **428**:37-43.
72. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM: **Comparative metagenomics of microbial communities.** *Science* 2005, **308**:554-557.
73. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome.** *Science* 2006, **312**:1355-1359.
74. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, Takami H, Morita H, Sharma VK, Srivastava TP, Taylor TD, Noguchi H, Mori H, Ogura Y, Ehrlich DS, Itoh K, Takagi T, Sakaki Y, Hayashi T, Hattori M: **Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes.** *DNA Res* 2007, **14**:169-181.
75. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI: **An obesity-associated gut microbiome with increased capacity for energy harvest.** *Nature* 2006, **444**:1027-1031.
76. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ, Szeto E, Kyrpides NC, Mussmann M, Amann R, Bergin C, Ruehland C, Rubin EM, Dubilier N: **Symbiosis insights through metagenomic analysis of a microbial consortium.** *Nature* 2006, **443**:950-955.
77. Garcia Martin H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, Yeates C, He S, Salamov AA, Szeto E, Dalin E, Putnam NH, Shapiro HJ, Pangilinan JL, Rigoutsos I, Kyrpides NC, Blackall LL, McMahon KD, Hugenholtz P: **Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities.** *Nat Biotechnol* 2006, **24**:1263-1269.
78. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcon LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, *et al*: **The Sorcerer II Global Ocean sampling expedition: northwest Atlantic through eastern tropical Pacific.** *PLoS Biol* 2007, **5**:e77.
79. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304**:66-74.
80. Konstantinidis KT, Braff J, Karl DM, DeLong EF: **Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre.** *Appl Environ Microbiol* 2009, **75**:5345-5355.
81. Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, Sorek R, Tringe SG, Podar M, Martin HG, Kunin V, Dalevi D, Madejska J, Kirton E, Platt D, Szeto E, Salamov A, Barry K, Mikhailova N, Kyrpides NC, Matson EG, Ottesen EA, Zhang X, Hernandez M, Murillo C, Acosta LG, *et al*: **Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite.** *Nature* 2007, **450**:560-565.
82. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IM, Grechkin Y, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Hugenholtz P, Kyrpides NC: **IMG/M: a data management and analysis system for metagenomes.** *Nucleic Acids Res* 2008, **36**:D534-D538.
83. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia JM, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, *et al*: **The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families.** *PLoS Biol* 2007, **5**:e16.
84. Noguchi H, Park J, Takagi T: **MetaGene: prokaryotic gene finding from environmental genome shotgun sequences.** *Nucleic Acids Res* 2006, **34**:5623-5630.
85. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D, Bryant SH: **CDD: a Conserved Domain Database for protein classification.** *Nucleic Acids Res* 2005, **33**:D192-D196.
86. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25**:955-964.
87. Barrick JE, Breaker RR: **The distributions, mechanisms, and structures of metabolite-binding riboswitches.** *Genome Biol* 2007, **8**:R239.
88. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009, **25**:1335-1337.
89. Knudsen B, Hein J: **Pfold: RNA secondary structure prediction using stochastic context-free grammars.** *Nucleic Acids Res* 2003, **31**:3423-3428.
90. Yao Z: **Genome scale search of noncoding RNAs: bacteria to vertebrates.** Seattle, WA: University of Washington; Dissertation 2008.
91. **Google Scholar.** [http://scholar.google.com].
92. Liu JM, Livny J, Lawrence MS, Kimball MD, Waldor MK, Camilli A: **Experimental discovery of sRNAs in *Vibrio cholerae* by direct cloning, 5S/tRNA depletion and parallel sequencing.** *Nucleic Acids Res* 2009, **37**:e46.
93. Livny J, Brencic A, Lory S, Waldor MK: **Identification of 17 *Pseudomonas aeruginosa* sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2.** *Nucleic Acids Res* 2006, **34**:3484-3493.
94. Sonnleitner E, Sorger-Domenigg T, Madej MJ, Findeiss S, Hackermuller J, Huttenhofer A, Stadler PF, Blasi U, Moll I: **Detection of small RNAs in *Pseudomonas aeruginosa* by RNomics and structure-based bioinformatic tools.** *Microbiology* 2008, **154**:3175-3187.
95. Gonzalez N, Heeb S, Valverde C, Kay E, Reimmann C, Junier T, Haas D: **Genome-wide search reveals a novel GacA-regulated small RNA in *Pseudomonas* species.** *BMC Genomics* 2008, **9**:167.
96. Steglich C, Futschik ME, Lindell D, Voss B, Chisholm SW, Hess WR: **The challenge of regulation in a minimal photoautotroph: non-coding RNAs in *Prochlorococcus*.** *PLoS Genet* 2008, **4**:e1000173.
97. Ulve VM, Sevin EW, Cheron A, Barloy-Hubler F: **Identification of chromosomal alpha-proteobacterial small RNAs by comparative genome analysis and detection in *Sinorhizobium meliloti* strain 1021.** *BMC Genomics* 2007, **8**:467.
98. Valverde C, Livny J, Schluter JP, Reinkensmeier J, Becker A, Parisi G: **Prediction of *Sinorhizobium meliloti* sRNA genes and experimental detection in strain 2011.** *BMC Genomics* 2008, **9**:416.
99. del Val C, Rivas E, Torres-Quesada O, Toro N, Jimenez-Zurdo JI: **Identification of differentially expressed small non-coding RNAs in the legume endosymbiont *Sinorhizobium meliloti* by comparative genomics.** *Mol Microbiol* 2007, **66**:1080-1091.
100. Saito S, Kakeshita H, Nakamura K: **Novel small RNA-encoding genes in the intergenic regions of *Bacillus subtilis*.** *Gene* 2009, **428**:2-8.
101. Padalon-Brauch G, Hershberg R, Elgrably-Weiss M, Baruch K, Rosenshine I, Margalit H, Altuvia S: **Small RNAs encoded within genetic islands of *Salmonella typhimurium* show host-induced expression and role in virulence.** *Nucleic Acids Res* 2008, **36**:1913-1927.
102. Pichon C, Felden B: **Small RNA genes expressed from *Staphylococcus aureus* genomic and pathogenicity islands with specific expression among pathogenic strains.** *Proc Natl Acad Sci USA* 2005, **102**:14249-14254.

103. Swiercz JP, Hindra , Bobek J, Haiser HJ, Di Berardo C, Tjaden B, Elliot MA: **Small non-coding RNAs in *Streptomyces coelicolor*.** *Nucleic Acids Res* 2008, **36**:7240-7251.
104. Rasmussen S, Nielsen HB, Jarmer H: **The transcriptionally active regions in the genome of *Bacillus subtilis*.** *Mol Microbiol* 2009, **73**:1043-1057.
105. Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, Assefa SA, He M, Croucher NJ, Pickard DJ, Maskell DJ, Parkhill J, Choudhary J, Thomson NR, Dougan G: **A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*.** *PLoS Genet* 2009, **5**:e1000569.
106. Tezuka T, Hara H, Ohnishi Y, Horinouchi S: **Identification and gene disruption of small noncoding RNAs in *Streptomyces griseus*.** *J Bacteriol* 2009, **191**:4896-4904.
107. Yoder-Himes DR, Chain PS, Zhu Y, Wurtzel O, Rubin EM, Tiedje JM, Sorek R: **Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing.** *Proc Natl Acad Sci USA* 2009, **106**:3976-3981.
108. Geissmann T, Chevalier C, Cros MJ, Boisset S, Fechter P, Noirot C, Schrenzel J, Francois P, Vandenesch F, Gaspin C, Romby P: **A search for small noncoding RNAs in *Staphylococcus aureus* reveals a conserved sequence motif for regulation.** *Nucleic Acids Res* 2009, **37**:7239-7257.
109. Arnvig KB, Young DB: **Identification of small RNAs in *Mycobacterium tuberculosis*.** *Mol Microbiol* 2009, **73**:397-408.
110. Georg J, Voss B, Scholz I, Mitschke J, Wilde A, Hess WR: **Evidence for a major role of antisense RNAs in cyanobacterial gene regulation.** *Mol Syst Biol* 2009, **5**:305.
111. **Google.** [http://www.google.com].
112. Abreu-Goodger C, Merino E: **RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements.** *Nucleic Acids Res* 2005, **33**:W690-692.
113. **Stockholm format.** [http://en.wikipedia.org/wiki/Stockholm_format].
114. Fuchs RT, Grundy FJ, Henkin TM: **The S(MK) box is a new SAM-binding RNA for translational regulation of SAM synthetase.** *Nat Struct Mol Biol* 2006, **13**:226-233.
115. Poiata E, Meyer MM, Ames TD, Breaker RR: **A variant riboswitch aptamer class for *S* -adenosylmethionine common in marine bacteria.** *RNA* 2009, **15**:2046-2056.
116. Platt MD, Schurr MJ, Sauer K, Vazquez G, Kukavica-Ibrulj I, Potvin E, Levesque RC, Fedynak A, Brinkman FS, Schurr J, Hwang SH, Lau GW, Limbach PA, Rowe JJ, Lieberman MA, Barraud N, Webb J, Kjelleberg S, Hunt DF, Hassett DJ: **Proteomic, microarray, and signature-tagged mutagenesis analyses of anaerobic *Pseudomonas aeruginosa* at pH 6.5, likely representing chronic, late-stage cystic fibrosis airway conditions.** *J Bacteriol* 2008, **190**:2739-2758.
117. Sriramulu DD, Nimtz M, Romling U: **Proteome analysis reveals adaptation of *Pseudomonas aeruginosa* to the cystic fibrosis lung environment.** *Proteomics* 2005, **5**:3712-3721.
118. Jarrige AC, Mathy N, Portier C: **PNPase autocontrols its expression by degrading a double-stranded structure in the pnp mRNA leader.** *EMBO J* 2001, **20**:6845-6855.
119. Cardineau GA, Curtiss R: **Nucleotide sequence of the *asd* gene of *Streptococcus mutans:* identification of the promoter region and evidence for attenuator-like sequences preceding the structural gene.** *J Biol Chem* 1987, **262**:3344-3353.
120. Hendriksen WT, Bootsma HJ, Estevao S, Hoogenboezem T, de Jong A, de Groot R, Kuipers OP, Hermans PW: **CodY of *Streptococcus pneumoniae* : link between nutritional gene regulation and colonization.** *J Bacteriol* 2008, **190**:590-601.
121. Kim K, Meyer RJ: **Copy-number of broad host-range plasmid R1162 is regulated by a small RNA.** *Nucleic Acids Res* 1986, **14**:8027-8046.
122. Vitreschak AG, Lyubetskaya EV, Shirshin MA, Gelfand MS, Lyubetsky VA: **Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis.** *FEMS Microbiol Lett* 2004, **234**:357-370.
123. Eddy SR: **A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure.** *BMC Bioinformatics* 2002, **3**:18.
124. Leaphart AB, Thompson DK, Huang K, Alm E, Wan XF, Arkin A, Brown SD, Wu L, Yan T, Liu X, Wickham GS, Zhou J: **Transcriptome profiling of *Shewanella oneidensis* gene expression following exposure to acidic and alkaline pH.** *J Bacteriol* 2006, **188**:1633-1642.
125. Storz G, Zheng M: **Oxidative stress.** *Bacterial Stress Responses* Washington, DC: ASM PressStorz G, Hengge-Aronis R 2000, 47-59.
126. Lee JC: **Structural studies of ribosomal RNA based on cross-analysis of comparative models and three-dimensional crystal structures.** Austin, Texas: University of Texas; Dissertation 2003.
127. Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, Delong EF: **Microbial community gene expression in ocean surface waters.** *Proc Natl Acad Sci USA* 2008, **105**:3805-3810.
128. Forchhammer K: **Global carbon/nitrogen control by PII signal transduction in cyanobacteria: from signals to targets.** *FEMS Microbiol Rev* 2004, **28**:319-333.
129. Walt A, Kahn ML: **The *fixA* and *fixB* genes are necessary for anaerobic carnitine reduction in *Escherichia coli*.** *J Bacteriol* 2002, **184**:4044-4047.
130. Chou HT, Kwon DH, Hegazy M, Lu CD: **Transcriptome analysis of agmatine and putrescine catabolism in *Pseudomonas aeruginosa* PAO1.** *J Bacteriol* 2008, **190**:1966-1975.
131. Espinosa-Urgel M, Ramos JL: **Expression of a *Pseudomonas putida* aminotransferase involved in lysine catabolism is induced in the rhizosphere.** *Appl Environ Microbiol* 2001, **67**:5219-5224.
132. Ochsner UA, Wilderman PJ, Vasil AI, Vasil ML: **GeneChip expression analysis of the iron starvation response in *Pseudomonas aeruginosa* : identification of novel pyoverdine biosynthesis genes.** *Mol Microbiol* 2002, **45**:1277-1287.
133. Yamanishi Y, Mihara H, Osaki M, Muramatsu H, Esaki N, Sato T, Hizukuri Y, Goto S, Kanehisa M: **Prediction of missing enzyme genes in a bacterial metabolic network: reconstruction of the lysine-degradation pathway of *Pseudomonas aeruginosa*.** *FEBS J* 2007, **274**:2262-2273.
134. Vencato M, Tian F, Alfano JR, Buell CR, Cartinhour S, DeClerck GA, Guttman DS, Stavrinides J, Joardar V, Lindeberg M, Bronstein PA, Mansfield JW, Myers CR, Collmer A, Schneider DJ: **Bioinformatics-enabled identification of the HrpL regulon and type III secretion system effector proteins of *Pseudomonas syringae* pv. phaseolicola 1448A.** *Mol Plant Microbe Interact* 2006, **19**:1193-1206.
135. Bonomo RA, Szabo D: **Mechanisms of multidrug resistance in *Acinetobacter* species and *Pseudomonas aeruginosa*.** *Clin Infect Dis* 2006, **43(Suppl 2)**:S49-S56.
136. Duan K, Liu CQ, Supple S, Dunn NW: **Involvement of antisense RNA in replication control of the lactococcal plasmid pND324.** *FEMS Microbiol Lett* 1998, **164**:419-426.
137. Kok J: **Inducible gene expression and environmentally regulated genes in lactic acid bacteria.** *Antonie Van Leeuwenhoek* 1996, **70**:129-145.
138. Regulski EE, Moy RH, Weinberg Z, Barrick JE, Yao Z, Ruzzo WL, Breaker RR: **A widespread riboswitch candidate that controls bacterial genes involved in molybdenum cofactor and tungsten cofactor metabolism.** *Mol Microbiol* 2008, **68**:918-932.
139. Wijayarathna CD, Wachi M, Nagai K: **Isolation of *ftsI* and *murE* genes involved in peptidoglycan synthesis from *Corynebacterium glutamicum*.** *Appl Microbiol Biotechnol* 2001, **55**:466-470.
140. Panagiotidis CH, Boos W, Shuman HA: **The ATP-binding cassette subunit of the maltose transporter MalK antagonizes MalT, the activator of the *Escherichia coli* mal regulon.** *Mol Microbiol* 1998, **30**:535-546.
141. Ravcheev DA, Gelfand MS, Mironov AA, Rakhmaninova AB: **[Purine regulon of gamma-proteobacteria: a detailed description.].** *Genetika* 2002, **38**:1203-1214.
142. Bochner BR, Ames BN: **ZTP (5-amino 4-imidazole carboxamide riboside 5'-triphosphate): a proposed alarmone for 10-formyl-tetrahydrofolate deficiency.** *Cell* 1982, **29**:929-937.
143. Rohlman CE, Matthews RG: **Role of purine biosynthetic intermediates in response to folate stress in *Escherichia coli*.** *J Bacteriol* 1990, **172**:7200-7210.
144. Weng M, Nagy PL, Zalkin H: **Identification of the *Bacillus subtilis pur* operon repressor.** *Proc Natl Acad Sci USA* 1995, **92**:7455-7459.
145. Su Z, Mao F, Dam P, Wu H, Olman V, Paulsen IT, Palenik B, Xu Y: **Computational inference and experimental validation of the nitrogen assimilation regulatory network in cyanobacterium *Synechococcus* sp. WH 8102.** *Nucleic Acids Res* 2006, **34**:1050-1065.
146. Fujita M, Amemura A, Aramaki H: **Transcription of the *groESL* operon in *Pseudomonas aeruginosa* PAO1.** *FEMS Microbiol Lett* 1998, **163**:237-242.
147. Seraphin B: **The HIT protein family: a new family of proteins present in prokaryotes, yeast and mammals.** *DNA Seq* 1992, **3**:177-179.

148. Lombardo MJ, Rosenberg SM: ***radC102* of *Escherichia coli* is an allele of *recG*.** *J Bacteriol* 2000, **182**:6287-6291.
149. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A: **Pfam: clans, web tools and services.** *Nucleic Acids Res* 2006, **34**:D247-D251.
150. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R: **NCBI GEO: archive for high-throughput functional genomic data.** *Nucleic Acids Res* 2009, **37**:D885-D890.
151. Nalca Y, Jansch L, Bredenbruch F, Geffers R, Buer J, Haussler S: **Quorum-sensing antagonistic activities of azithromycin in *Pseudomonas aeruginosa* PAO1: a global approach.** *Antimicrob Agents Chemother* 2006, **50**:1680-1688.
152. Chugani S, Greenberg EP: **The influence of human respiratory epithelia on *Pseudomonas aeruginosa* gene expression.** *Microb Pathog* 2007, **42**:29-35.
153. Diwa A, Bricker AL, Jain C, Belasco JG: **An evolutionarily conserved RNA stem-loop functions as a sensor that directs feedback regulation of RNase E gene expression.** *Genes Dev* 2000, **14**:1249-1260.
154. Gupta RS: **The phylogeny and signature sequences characteristics of Fibrobacteres, Chlorobi, and Bacteroidetes.** *Crit Rev Microbiol* 2004, **30**:123-143.
155. Montange RK, Batey RT: **Structure of the *S* -adenosylmethionine riboswitch regulatory mRNA element.** *Nature* 2006, **441**:1172-1175.
156. Connelly JC, Leach DR: **The *sbcC* and *sbcD* genes of *Escherichia coli* encode a nuclease involved in palindrome inviability and genetic recombination.** *Genes Cells* 1996, **1**:285-291.
157. Arthur DC, Ghetu AF, Gubbins MJ, Edwards RA, Frost LS, Glover JN: **FinO is an RNA chaperone that facilitates sense-antisense RNA interactions.** *EMBO J* 2003, **22**:6346-6355.
158. Passalacqua KD, Varadarajan A, Ondov BD, Okou DT, Zwick ME, Bergman NH: **Structure and complexity of a bacterial transcriptome.** *J Bacteriol* 2009, **191**:3203-3211.
159. 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**:774-784.
160. Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL, Breaker RR: **Genetic control by a metabolite binding mRNA.** *Chem Biol* 2002, **9**:1043.
161. Nahvi A, Barrick JE, Breaker RR: **Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes.** *Nucleic Acids Res* 2004, **32**:143-150.
162. Fox KA, Ramesh A, Stearns JE, Bourgogne A, Reyes-Jara A, Winkler WC, Garsin DA: **Multiple posttranscriptional regulatory mechanisms partner to control ethanolamine utilization in *Enterococcus faecalis*.** *Proc Natl Acad Sci USA* 2009, **106**:4435-4440.
163. Regulski EE, Breaker RR: **In-line probing analysis of riboswitches.** *Methods Mol Biol* 2008, **419**:53-67.
164. Johansen LE, Nygaard P, Lassen C, Agerso Y, Saxild HH: **Definition of a second *Bacillus subtilis* pur regulon comprising the *pur* and *xpt-pbuX* operons plus *pbuG*, *nupG* (*yxjA*), and *pbuE* (*ydhL*).** *J Bacteriol* 2003, **185**:5200-5209.

doi:10.1186/gb-2010-11-3-r31
Cite this article as: Weinberg *et al.*: Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biology* 2010 **11**:R31.