HHS Public Access

Author manuscript

Microbiol Spectr. Author manuscript; available in PMC 2018 July 12.

Published in final edited form as:

Microbiol Spectr. 2018 July ; 6(4): . doi:10.1128/microbiolspec.RWR-0005-2017.

Large noncoding RNAs in bacteria

Kimberly A. Harris[1,2] and **Ronald R. Breaker**[1,2,3,*]

[1]Howard Hughes Medical Institute, Yale University, Box 208103, New Haven, CT 06520-8103, USA

[2]Department of Molecular, Cellular and Developmental Biology, Yale University, Box 208103, New Haven, CT 06520-8103, USA

[3]Department of Molecular Biophysics and Biochemistry, Yale University, Box 208103, New Haven, CT 06520-8103, USA

Abstract

Bacterial noncoding RNA (ncRNA) classes longer than 200 nucleotides are rare, but are responsible for performing some of the most fundamental tasks in living cells. RNAs such as 16S and 23S ribosomal RNA, group I and group II introns, RNase P ribozymes, tmRNAs, and coenzyme B_{12} riboswitches, are diverse in structure and accomplish biochemical functions that rival the activities of proteins. Over the last decade, a number of new classes of large ncRNAs have been uncovered in bacteria. A total of 21 classes with no established functions have been identified through the use of bioinformatics search strategies. Based on precedents for bacterial large ncRNAs performing sophisticated functions, it seems likely that some of these structured ncRNAs also will prove to carry out complex functions. Thus, determining their roles will provide a better understanding of fundamental biological processes. A few studies have produced data that provide clues to the purposes of some of these recently-found classes, but the true functions of most classes remain mysterious.

Diversity of Large ncRNA Functions

Although bacteria harbor far fewer long ncRNAs than eukaryotes, the known classes of large structured ncRNAs in bacteria perform essential roles in the core processes of information transfer, metabolism, and physiological adaptation (1). For example, many classes are central to genetic information processing: ribosomal RNAs act as ribozymes (2) to translate mRNAs, RNase P ribozymes process precursor tRNAs (3, 4), tmRNAs rescue stalled ribosomes (5, 6), and riboswitches bind ions and metabolites to regulate gene expression (7–10). Furthermore, most of the large structured ncRNA classes whose functions are known operate as ribozymes that perform essential chemical reactions such as peptide bond formation (2), RNA splicing (11, 12) and RNA cleavage (3). Two of these ribozyme classes, namely group I and group II introns, are sometimes components of selfish genetic elements that both splice mRNAs and mobilize to various regions in DNA genomes (13, 14). Of course many self-splicing ribozymes also carry protein coding regions, located either in their

*Correspondence: Dr. Ronald R. Breaker, Tel: (203) 432-9389, Fax: (203) 432-0753, ronald.breaker@yale.edu.

exon flanks or inserted into non-critical portions of their ribozyme structure. However, these coding regions are usually incidental to the main functions performed by the ncRNA's structure. Collectively, these ncRNAs have an enormous influence on both genetic and cellular processes, which suggests the intriguing possibility that newly-found large ncRNA classes may also serve fundamental roles in biology.

Given what we currently know about bacterial ncRNAs, these additional classes can be expected to possess functions ranging from catalytic activity to gene regulation, but also may have biological and biochemical functions that have yet to be observed for RNA. It is not practical at this time to put boundaries around the possible set of functions for these ncRNAs. However, it seems highly unlikely that large ncRNAs that have extensive and non-repetitive conserved sequences, and that have complexly-folded structures, will prove to perform only simple biochemical tasks like base-pairing to another RNA or serving as a passive binding site for a protein factor.

Notably, where are the ribozymes that promote chemical transformations typical of metabolic enzymes? To date, ribosomes are the only natural ribozymes known to perform chemistry other than phosphoester transfer or hydrolysis. If a newly-found ribozyme class performs a critical task, but is a legacy biocatalyst from the RNA World (15) that has persisted in deeply-branching bacterial lineages, the RNA might have a role that has been replaced by proteins in eukaryotic organisms. Or, it is also possible that the ncRNA serves a purpose for which an RNA molecule is well-suited, and, therefore, the ncRNA has emerged more recently in evolution.

To date, seven classes of large structured ncRNAs (consistently greater than 200 nucleotides) with established biochemical functions are known to exist in bacteria (Figure 1). These RNA classes are defined by their distinct biochemical functions and/or distinct consensus sequence and structural models. The signal recognition particle (SRP) RNA is excluded here because many representatives are shorter than 200 nucleotides (16). Three additional putative classes of large ncRNAs, called OLE (17), GOLLD (18), and HEARO (18), have been investigated to some degree, but their biochemical roles are not well understood. Another 18 possible large ncRNA classes were recently discovered by using bioinformatics and await experimental validation (19). If all of these candidates indeed represent additional classes of large ncRNAs, establishing their functions would increase the number of large ncRNA activities in bacteria by four fold.

Based on the considerable size and structural sophistication of these ncRNA candidates, we believe that significant opportunities exist to discover entirely new biological and biochemical roles for RNA in modern cells. Importantly, all structured bacterial ncRNAs larger than 350 nucleotides with known functions are catalytic RNAs or function in catalytic complexes (Figure 1). Moreover, by using the number of multi-stem junctions and pseudoknots as a metric for structural complexity, we also note that ncRNAs with the most sophisticated structures tend to be ribozymes. These observations support the hypothesis that some of the additional mysterious large ncRNAs with sizes and structural complexities similar to those of known ribozymes might possess hidden catalytic abilities.

Discovery of Novel ncRNAs

Computational search strategies have been very productive for revealing hundreds of structured ncRNA classes in bacteria. For example, comparative sequence analysis algorithms have been used to identify multiple novel classes of riboswitches (19–23) and ribozymes (24, 25), as well as exceptionally large and structurally complex ncRNAs (17, 18, A. Roth, Z. Weinberg, K. Vanderschuren, M. H. Murdock, E. Poiata, and R. R. Breaker, unpublished data). Each of the ncRNA classes discussed below was uncovered computationally through such phylogenetic analyses. This general method identifies nucleotide positions that exhibit strong sequence conservation and secondary-structure features that are supported by nucleotide covariation indicative of Watson-Crick base-pairing.

Bacterial genomic and metagenomic DNA sequences are particularly amenable for searches directed toward the discovery of structured RNA molecules with functions other than coding for proteins. First, the abundance of bacterial genomic sequence data provides a deep dataset for conducting searches for novel ncRNA classes by using comparative sequence analysis. Second, structured RNAs such as ribozymes and riboswitches have never been observed to reside entirely within the coding regions of messenger RNAs. Noncoding nucleotides comprise only a small fraction of the total nucleotides within most bacterial genomes. Thus, undiscovered classes of structured RNAs are enriched in these noncoding regions, which reduces the demands placed on computer algorithms that use comparative sequence analysis as a search mechanism. Third, the expansive evolutionary separations between many diverse bacterial species allow researchers to gain confidence in novel RNA classes that remain exceptionally well conserved.

These features of bacterial genomes permit the use of computational search strategies to uncover the most common ncRNA classes. However, such RNAs that are exceedingly rare cannot easily be identified by existing comparative sequence analysis algorithms. These searches fail when they do not have sufficiently distinct representatives of a ncRNA class for comparison. Fortunately, sequencing technologies, sequence databases, and computational resources continue to grow and improve. Therefore, the future of ncRNA discovery via computational searching is promising, and offers the opportunity to discover ever rarer ncRNAs that exist only in small biological niches.

Large ncRNAs with Unknown Functions

Among the most common functions for the highly-structured ncRNA classes, which excludes less structured classes such as bacterial sRNAs (26), are RNA self-cleavage (24, 25, 27) and riboswitch-mediated ligand binding and gene control (28). Most representatives of these natural self-cleaving ribozyme and riboswitch classes are shorter than 200 nucleotides. Thus the discovery of a novel structured ncRNA class whose representatives are consistently longer than 200 nucleotides should encourage researchers to consider possible biochemical functions other than RNA self-cleavage or riboswitch regulation. Below are brief descriptions of large bacterial RNAs whose functions have yet to be established.

OLE RNA

The OLE (Ornate, Large, Extremophilic) RNA class was first described in 2006 based on the discovery of 15 representatives from bacterial genomes (17). Currently, there are 657 unique representatives known that reside exclusively in the genomes of extremophilic species and environmental metagenomes (K. A. Harris, Z. Zhou, M. L. Peters, S. G. Wilkins, and R. R. Breaker, submitted for publication). Strikingly, this RNA is found in a wide range of species in Firmicutes, wherein about half of its ~600 nucleotides are conserved with covariation supporting a complex secondary structure comprised of several multistem junctions (Figure 2) (17, 29, 30). The intricate network of bulges and loops and the positioning of conserved nucleotides suggest that this RNA forms a complex tertiary structure that is critical for its function.

Because OLE RNAs are so widespread in anaerobic extremophiles, it is tempting to speculate that they may have a role in protecting these species from the extreme environments in which they thrive. Transcriptome analysis in *Bacillus halodurans* revealed that *ole*, the gene for OLE RNA, is one of the most highly expressed. Excluding rRNA and tRNA transcripts, the OLE RNA is the sixteenth most abundant transcript under normal growth conditions. OLE RNA abundance further increases when cells are exposed to short-chain alcohols, including ethanol, which is produced during anaerobic growth (30). Again excluding RNAs responsible for translation, OLE RNA becomes the fifth most common in cells grown in the presence of 5% ethanol. OLE RNA is surpassed in abundance only by SRP RNA, tmRNA, and two mRNAs. Under these conditions, OLE RNA transcripts are processed and remain relatively stable, with a half-life of approximately three hours (30).

The *ole* gene is commonly embedded in a large operon that contains genes involved in isoprenoid biosynthesis, DNA repair, coenzyme metabolism, and transcription regulation (17). Directly downstream of *ole* is a gene of unknown function that encodes the OLE-Associated Protein (OAP). The position of the tandem-arranged *ole* and *oap* genes immediately downstream of the *ispA* gene and immediately upstream of the *dxs* gene is highly conserved in almost all of the bacterial genomes carrying OLE RNA (K. A. Harris, Z. Zhou, M. L. Peters, S. G. Wilkins, and R. R. Breaker, submitted for publication). This suggests that the role of OLE RNA might be related to cell membrane biochemistry, given that the IspA protein (geranyltranstransferase) and the Dxs protein (1-deoxy-D-xylulose-5-phosphate synthase) are key enzymes in the isoprenoid biosynthesis pathway (31).

OAP is a 21-kDa, predicted four-helix transmembrane protein that specifically binds OLE RNA *in vitro* (29). The complex has an apparent 2:1 OAP:OLE RNA stoichiometry, suggesting that the protein might function as a dimer. Because OAP is predicted to be a transmembrane protein and binds OLE RNA, the ability of OLE RNA to localize to cell membranes was examined by using fluorescence *in situ* hybridization (FISH) microscopy. Indeed, OLE RNA localizes to the cell membrane, but only in the presence of OAP (29). This finding again suggests that the function of OLE RNA might be related to the biochemistry of membranes, or perhaps cell walls.

Knockouts of both *ole* and *oap* show they are not essential in *B. halodurans* (30) when grown under normal conditions. However, growth of *B. halodurans* strains lacking the *ole*

and/or *oap* genes is reduced five-fold compared to wild-type cells in the presence of 5% ethanol. Alcohols can cause a range of stresses to bacterial cells, such as increased membrane permeability that allows ions and small molecules to enter the cytoplasm (32). Therefore, cells have ethanol-induced responses to upregulate specific pathways and make changes to protein and lipid composition of the cell membrane (33–35). Because OLE RNA and OAP localize to cell membranes (29), the ribonucleoprotein (RNP) may have a role in a response mechanism to this stress, such as stabilizing or producing membrane components to fortify against leakage. Furthermore, *B. halodurans* cells lacking *ole* and/or *oap* genes are less tolerant of growth in cold temperatures (30) and other growth conditions (K. A. Harris and R. R. Breaker, unpublished data). Still, it is not known how OLE RNAs help cells adapt to these stresses.

OLE RNA is currently the most prevalent structured ncRNA class larger than 500 nucleotides whose function is unknown. This distinction alone makes it a particularly attractive target for further analysis. Moreover, OLE RNAs are among the most complex and well-conserved ncRNAs known to exist in bacteria (Figure 1). Nearly all RNAs that are similar in size and structural complexity to OLE RNAs whose functions are already known (e.g. RNase P, group I and II self-splicing RNAs) function as ribozymes with biologically important activities. Therefore, establishing the biochemical function of OLE RNAs will likely reveal the action of a new ribozyme, or meaningful knowledge about a fundamental aspect of the cells that carry this molecule.

GOLLD RNA

With an average of more than 800 nucleotides, the GOLLD (Giant, Ornate, Lake- and Lactobacillales-Derived) RNA is the third-largest bacterial ncRNA discovered to date, behind only 16S and 23S rRNA (18). A common arrangement originally reported for GOLLD RNAs includes numerous RNA substructures that are indicative of the formation of an exceedingly complex tertiary structure (Figure 3). A total of 391 representatives have been identified in Lactobacillales and Actinomycetales orders, and among environmental DNA sequences. The motif consists of distinct 5′- and 3′-domains. The 3′ half is highly conserved, contains most of the long-range interactions, and is present in all GOLLD RNAs. The 5′ half appears to diverge into variant structures where some substructures are absent or substituted (18). This type of domain variation is not uncommon for complex structured RNAs (36).

GOLLD RNAs are commonly encoded by bacteriophages, frequently located adjacent to tRNA genes (18). However, genes for GOLLD RNAs are sometimes present in bacterial genomes and unaffiliated with bacteriophages. This suggests that GOLLD RNAs have a biochemical function that is beneficial to both bacterial cells and the viruses that infect them or, alternatively, are a type of selfish genetic element.

Approximately 15% of the examples carry tRNAs embedded within a variable region of the motif. The significance of these overlapping arrangements is unclear. For example, the biochemical function of GOLLD RNAs might somehow relate to the processing or activity of tRNAs. Alternatively, GOLLD RNAs might have a function that is completely independent from tRNAs, and they are only occasionally co-expressed as an efficient way to

produce large amounts of specific ncRNAs that are separated by post-transcriptional processing. As observed in the initial report (17), *Lactobacillus* bacteriophages and prophages often have large noncoding regions surrounding their tRNA genes. Bacteriophages also exhibit high rates of host-parasite recombination and are capable of both horizontal and vertical genetic transfer (37, 38). Consequently, there is a possibility that GOLLD RNAs are not relevant to the bacteriophage life cycle.

Experiments performed in *Lactobacillus brevis*, which harbors the *golld* gene in a prophage, revealed that when the prophage is induced, GOLLD RNA expression levels correlated with bacteriophage particle production. Mapping of the 5′- and 3′-termini of these RNAs demonstrated that the entire predicted structure of GOLLD RNA is produced (18). Given the wide distribution of *golld* among a number of bacteriophages, it is plausible that GOLLD RNA has a useful function for phage reproduction. However, initial analysis of a bacteriophage carrying a *golld* knockout resulted in mutant phages that reproduced without evidence of a replication defect (39). Therefore, the biochemical function of GOLLD RNA does not seem to be essential for the replication of phages, at least in certain bacterial hosts.

HEARO RNA

The proposed HEARO (HNH Endonuclease-Associated RNA and ORF) RNAs are highly structured molecules of ~350 nucleotides surrounding an embedded open reading frame (ORF) (18). The motif does not have many highly-conserved nucleotides, but exhibits plentiful evidence for covariation to support the formation of many base-paired regions. HEARO representatives are located in species from ten different bacterial phyla, predominantly Firmicutes, Proteobacteria, Cyanobacteria, and Actinobacteria. One example in the archaean *Methanosarcina mazei* has previously been reported (18). However, in some bacterial species, dozens of *hearo* genes are present. This pattern of *hearo* distribution is strongly indicative of a function as a selfish genetic element. Such mobile elements can operate as RNA or DNA (40). It is not yet clear that HEARO functions as a structured ncRNA, or whether its function is manifested as a structured single-stranded DNA (ssDNA) element.

In most instances, the motif contains an embedded ORF that encodes a putative HNH endonuclease. the presence of this protein-coding region indicates that the motif is at least occasionally transcribed. Indeed, expression of the HEARO RNA was detected in one bacterium (18). HNH endonucleases are a family of homing endonucleases, which are commonly embedded within group I and group II introns and are involved in the transfer of these elements (41). Close relatives of the HEARO ORF include the ORF associated with IS605 selfish genetic elements. These are known to exploit small structured DNA motifs as part of their replicative cycle (42). Each HEARO representative is much larger and complex than these IS element ssDNA motifs, and so it is unclear if they are related. Experiments to seek self-splicing activity of HEARO RNA transcripts have yielded no positive results (18). Further research is needed to determine if the RNA is processed, interacts with any proteins, and ultimately is functional as an RNA polymer beyond coding for an endonuclease.

T-Large ribozymes are permuted group II introns

In rare instances, computational approaches to discover structured ncRNAs reveal distant variants of known ncRNA classes (e.g. 43–46). A large RNA called “T-Large” was found (A. Roth, Z. Weinberg, K. Vanderschuren, M. H. Murdock, E. Poiata, and R. R. Breaker, unpublished data) that resembles group II self-splicing introns, but is circularly permuted. This RNA does promote phosphoester transfer reactions like those required for normal RNA splicing, and even exploits the same splice sites. However, these splice sites are present in the precursor RNA in reverse order, such that the ribozyme yields a circularized rather than linear exon product and yields a branched T-like product rather than a lariat intron. T-Large representatives are fairly widespread in Proteobacteria, Cyanobacteria, and Deinococcus-Thermus phyla, and are common in environmental DNA samples. Unfortunately, the biological role of the unusual RNA splicing reaction products produced by T-Large ribozymes has yet to be established.

Other newly-discovered large ncRNAs

With the exception of T-Large, the ncRNA classes noted above have unknown biochemical functions. All the RNAs have unclear biological roles, and we are even uncertain whether the complex structure of the HEARO motif is relevant as an RNA molecule or as a single-stranded DNA. Although we have not studied each of these ncRNA candidates continuously for the last decade since their discoveries (17, 18), the fact that their biochemical and biological functions remain mysterious is both intriguing and cautionary. Perhaps new functional features of RNA will be revealed and new biological processes exposed upon establishing the roles these RNAs serve. However, it is inherently challenging to study the biochemical functions of an RNA whose connection to a specific biological process remains obscure.

Intriguingly, there is reason to believe that additional large bacterial ncRNA classes will continue to be revealed as new genomes are sequenced and new bioinformatics methods are developed. Recently, at least 16 candidate large ncRNA classes have been reported as part of a collection of 224 structured RNA motifs (19), which add to the list of candidates reported previously (18). The most noteworthy classes are discussed below, whereas the remaining motifs are listed in Table 1. Full descriptions of all these ncRNAs can be found in the previous publications describing their initial identification (17–19).

(i) IMES-1 RNA—Several IMES (Identified in Marine Environmental Sequences) ncRNA classes that are abundantly expressed were uncovered by comparative sequence analysis (18). With an average length of ~220 nucleotides, the IMES-1 class (also reported elsewhere as Groups 3, 4, 11, and 19 psRNA) (47), is the longest of the four IMES RNA classes identified. Its secondary structure contains one pseudoknot, one multistem junction, and features a large number of highly conserved nucleotides.

Metatranscriptome data from cells isolated from the Pacific Ocean revealed the presence of IMES-1 RNA with a five-fold higher abundance than that measured for 5S rRNA (18, 47, 48). It is unclear why these RNAs are so highly expressed, but RNAs with this high level of expression are extremely unusual. Unfortunately, the over 400 examples of IMES-1 have

been located only in environmental metagenomes, making it challenging to study this ncRNA and establish its biological importance.

(ii) ROOL RNA—The ROOL (Rumen-Originating, Ornate, Large) RNAs were originally identified (19) among metagenomic DNA sequences isolated from cow rumen. A total of 397 distinct examples have been identified predominantly in the Lactobacillales and Clostridiales orders. The predicted secondary structure of the ROOL motif (19) makes it one of the most complex ncRNAs (Figure 1). However, it has fewer highly-conserved nucleotides than most other structured ncRNAs of similar size.

ROOL RNAs share some contextual similarities with GOLLD RNAs, such as frequent proximity to tRNA genes, association with bacteriophages and prophages, and occurrence in species of Lactobacillales. ROOL RNAs however do not have any recognized sequence or structural similarities to GOLLD RNAs, other than the fact that both RNA classes exhibit intricately-folded structures with several pseudoknots. GOLLD and ROOL RNAs are the most complexly structured bacterial large ncRNAs of unknown function (Figure 1). It is possible that they have similar biological functions, but certainly their distinct structures merit separate classifications.

(iii) *raiA* motif RNA—Discovered in Firmicutes and Actinobacteria, the *raiA* RNA class has 1347 representatives that exhibit a moderately complex secondary structure containing two pseudoknots and two multistem junctions (19). This RNA motif is named for its frequent occurrence in the 5′ untranslated region (UTR) of *raiA* genes. This gene encodes the RaiA protein, which binds ribosomes to halt translation during cell stress (49). The *raiA* motif RNA potentially functions as a *cis*-regulatory ncRNA. However, these RNAs are occasionally more than 600 base-pairs upstream of the neighboring protein-coding region, and are sometimes closer to the 3′ end of the gene located immediately upstream, commonly *comFC*. Bacterial *cis*-regulatory RNAs rarely reside in the 3′ UTRs of the genes they control. Therefore, it seems unlikely that *raiA* motif RNAs control expression of ComFC proteins, which are involved in genetic competence for DNA uptake (50).

The *raiA* motif RNA is also found upstream of genes encoding periplasmic binding proteins (PBPs) that transport a variety of substrates. Again, it seems possible that *raiA* motif RNAs function as riboswitches or another type of *cis*-regulatory domain, but ligand candidates that trigger changes in gene expression are not immediately apparent.

(iv) Additional candidates—A major challenge when initially evaluating candidate structured ncRNAs is to build confidence in the hypothesis that they even represent RNAs. For example, there are three candidates that are striking with regard to their size and structural complexity, but it is possible that they actually function as ssDNAs. The ARRPOF (Area Required for Replication in a Plasmid of *Fusobacterium*) motif, and the GEBRO (GC-Enriched, Between Replication Origins) motif (19) are two of the most complexly-structured ncRNA candidates of unknown function (Figure 1). They each form multiple pseudoknots and carry a large number of conserved nucleotides. These complex nucleic acid structures presumably aid in plasmid replication or perform regulatory roles associated with plasmids. Similarly, the PAGEV (Plasmid-Associated gamma-Proteobacteria Especially Vibrionales)

motif also frequently is present in plasmids (19). It is not yet certain what polynucleotide form these structures use to carry out their biological functions.

A fourth ncRNA candidate that might actually function as a ssDNA, called IS605-*orfB*-I, appears to form one multistem junction and one pseudoknot (19). Representative of this class reside 3′ of genes encoding a transposase of the IS605 OrfB family. The IS605 family of transposases use ssDNA as a transposition intermediate (42, 51). Thus, it is possible that the IS605-*orfB*-I motif is functional as ssDNA, but employs a secondary structure that is much larger and more complex than analogous elements in other IS605 representatives.

An additional ten large ncRNA motif candidates are summarized in Table 1. These ncRNAs may be new large ncRNA classes, but it is not certain that they form structures sophisticated enough to perform challenging biochemical functions. The Bacteroidales-2, HOLDH (Human Oral, Large, Distant to HINT), MISL (Mostly Independently Structured, Large), and RT-2 (Reverse Transcriptase 2) motifs possess some distinct secondary structure features. However, we do not currently have any additional clues to speculate on their functions beyond what has been stated previously (19). Clostridiales-3, EGFOA-assoc-1, *ilvB*-OMG, *lysM*-Actino, RT-7, and throat-1 are potential new large ncRNAs, but they do not appear to form complex structures (18).

Experimental Validation of Novel ncRNA Functions

Due to the tremendous outputs of bioinformatics search pipelines, there is a growing number of interesting large ncRNA candidates to study. However, the challenge of assigning functions to these ncRNA classes remains a difficult barrier to surpass. In past decades, novel ncRNAs were sometimes discovered by researchers who were studying a particular biological or biochemical process. As a result, they had strong clues regarding the possible function of the RNA, as was the case for RNase P (3), group I (11) and group II (12) introns, and ribosomes (2). The discovery of a novel ncRNA class by bioinformatics sometimes provides fewer clues regarding its function, and so additional experimental tactics are needed to define its biological and biochemical activities.

Below we briefly describe some of these possible experimental approaches, which have been useful for those elucidating the functions of other biomolecules, such as proteins or eukaryotic long ncRNAs (lncRNAs). This is hardly an exhaustive list, and surely each ncRNA class will need tailored experiments for in-depth analysis. As the field moves forward, new approaches will undoubtedly be developed to characterize these challenging ncRNAs.

Bacterial genetics

Bacteria tend to cluster genes into operons that code for proteins involved in a single biochemical or physiological process, such as a metabolic pathway or a stress response. Therefore, it might be possible to infer the function of a ncRNA that frequently clusters with genes for a given biological pathway. The information derived from the genomic location of a ncRNA has been particularly useful for determining the functions of *cis*-acting ncRNA regulatory elements, especially if the gene association of the ncRNA gene is highly

conserved. For example, the location of a ncRNA sequence upstream of a metabolite synthase gene is a strong indication that the RNA functions as a riboswitch that responds to the metabolite made by the enzyme encoded immediately downstream.

If the large ncRNA of interest resides in a genetically-tractable, culturable organism, genetics-based methods can be very powerful. Expression levels of the ncRNA can be determined by transcriptomics analysis under specific growth conditions (52). Gene deletions and overexpression constructs can be used to determine if the ncRNA is essential, or if there are particular phenotypes associated with the ncRNA (30, 53). Additionally, plasmids can be transformed into a knockout strain to express mutated or truncated versions of the ncRNA. If deleterious ncRNA mutations are identified, genetic screens may be used to identify gene mutations that rescue the phenotypes of defective cells or binding partners (54). RNA-seq of the knockout strains or cells grown under stress conditions may provide insight into genes upregulated to compensate for the loss of the ncRNA.

Biochemistry and chemical biology

Experiments that probe direct binding interactions of the ncRNA with small molecules, metabolites, or ions can be useful in determining if ncRNAs have a ligand or cofactor (55, 56). Modified nucleotides are critical for the structure and function of some of the known noncoding RNAs, such as rRNAs and tRNAs. Unfortunately, it is not yet known if any of the large ncRNAs described herein (Table 1) carry such modification. If a candidate ncRNA is naturally modified, this might hinder the biochemical analysis of RNAs made by in vitro transcription using only the four standard nucleotides. There are a myriad of RNA pull-down and co-purification methods paired with mass spectrometry or sequencing to identify proteins, DNAs, or RNAs that interact with the ncRNA of interest (57, 58). Techniques such as gradient profiling by sequencing (Grad-seq), which captures RNAs based on their biochemical profiles and protein interactions, can provide valuable clues to implicate the ncRNA in particular biochemical pathways if the RNA binds proteins of known functions (59). Cells also can be treated with antibiotics or chemical inhibitors that target specific biosynthetic pathways to provide additional phenotypic insights. For example, genetic knockout cells lacking the ncRNA might exhibit unusual growth characteristics when exposed to unusual nutrient sources, or otherwise sub-lethal doses of antibiotics or other toxic agents (53).

Structural biology and biophysics

High-resolution structure models, such as those generated by x-ray crystallography analyses, are important for ascertaining the molecular details of ncRNAs. Unfortunately, with large ncRNAs that likely interact with proteins, it has not been practical to make attempts to crystallize a potentially floppy, long ncRNA that might lack critical binding partners. Perhaps cryo-electron microscopy (cryo-EM) methods employed with ncRNAs gently removed from their cellular environments could yield useful data regarding the fine structures of ncRNAs and their biochemical partners. Otherwise, careful structural analysis might need to await the outcomes of other experiments seeking to establish fundamental details of the functions of candidate large ncRNAs.

Unlike eukaryotic lncRNAs that can be visualized in cells by FISH microscopy, bacterial cells are significantly smaller, making it difficult to precisely determine the cellular localization of RNAs in bacteria. However, with high-resolution microscopes and mathematical modeling, subcellular localization of RNAs can be determined with techniques such as FISH or live-cell MS2–GFP (bacteriophage MS2 coat protein fused to green fluorescent protein) (60). Such methods can be particularly informative if the ncRNA is naturally abundant and localizes to a prominent feature in a bacterial cell, such as the cell membrane (29).

Prospects for Large ncRNA Discoveries and Impacts

Most of the large ncRNAs discussed herein have functions that remain elusive. Because of their striking complexity, sequence conservation, and differences from known ncRNA classes, their biochemical and biological roles are likely to be novel. The key challenge in this area will be to decipher the precise functions of these ncRNAs. New large ncRNA classes appear to present more difficulties for those seeking to establish functions than most previously discovered ribozymes and ncRNAs, which were typically discovered more serendipitously. The first ribozyme to be experimentally validated, a self-splicing group I intron, was uncovered in an mRNA that was known to be spliced (11). The RNase P RNA was known to be a component of a pre-tRNA-processing ribonucleoprotein complex before ribozyme activity was demonstrated (3). By contrast, large bacterial ncRNAs such as OLE RNA, GOLLD RNA, and a variety of others do not yet exhibit an obvious link to a known biochemical process.

Indeed, the biological functions of some ncRNA classes, such as tmRNAs (6) and 6S RNAs (61), took decades from their initial discovery date to elucidate. It seems likely that this lag time between discovery and functional validation might also occur for many of these newly-found ncRNA classes, which are all less widespread and perhaps have less prominent functions than tmRNAs and 6S RNAs. For instance, OLE RNAs were first reported twelve years ago (17), and GOLLD RNAs (18) nine years ago. Some progress in elucidating the structural and functional characteristics have so far only been reported for OLE RNAs (29, 30).

The accelerating rate of genome sequencing and the ever-expanding amounts of metagenomic data ensure that more novel ncRNAs will continue to be found. Despite the challenging tasks of experimentally validating their functions, each new ncRNA offers the possibility of identifying processes that have never been observed in biology previously. These and future discoveries will continue to alter the landscape of known RNA functions. If the known large ncRNAs, many of which have revolutionized our understanding of biology, are any indication of the impact of these classes of unknown function, then it will be well worth the effort.

Acknowledgments

We thank Adam Roth, Danielle Widner, and other members of the Breaker Laboratory for helpful discussions. Noncoding RNA research in the Breaker Laboratory is supported by the National Institutes of Health grant

F32GM116426 to K.A.H. and P01GM022778 to R.R.B. R.R.B is also an Investigator with the Howard Hughes Medical Institute.

References

1. Cech TR, Steitz JA. The noncoding RNA revolution-trashing old rules to forge new ones. Cell. 2014; 157:77–94. [PubMed: 24679528]
2. Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. The Structural Basis of Ribosome Activity in Peptide Bond Synthesis. Science. 2000; 289:920–930. [PubMed: 10937990]
3. Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. The RNA moiety of ribonuclease P is the catalytic subunit. Cell. 1986; 35:849–857.
4. Ellis JC, Brown JW. The RNase P family. RNA Biol. 2014; 6:362–369.
5. Keiler KC, Waller PR, Sauer RT. Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. Science. 1996; 271:990–993. [PubMed: 8584937]
6. Janssen BD, Hayes CS. The tmRNA ribosome-rescue system. Adv Protein Chem Struct Biol. 2012; 86:151–191. [PubMed: 22243584]
7. Mandal M, Breaker RR. Gene regulation by riboswitches. Nat Rev Mol Cell Biol. 2004; 5:451–463. [PubMed: 15173824]
8. Roth A, Breaker RR. The structural and functional diversity of metabolite-binding riboswitches. Annu Rev Biochem. 2009; 78:305–334. [PubMed: 19298181]
9. Serganov A, Nudler E. A decade of riboswitches. Cell. 2013; 152:17–24. [PubMed: 23332744]
10. Süss B. [Riboswitch chapter citation]
11. Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. Cell. 1982; 31:147–157. [PubMed: 6297745]
12. Peebles CL, Perlman PS, Mecklenburg KL, Petrillo ML, Tabor JH, Jarrell KA, Cheng HL. A self-splicing RNA excises an intron lariat. Cell. 1986; 44:213–223. [PubMed: 3510741]
13. Hausner G, Hafez M, Edgell DR. Bacterial group I introns: mobile RNA catalysts. Mob DNA. 2014; 5:8. [PubMed: 24612670]
14. Tori N, Jiménex-Zurdo JI, Garcia-Rodgiguez MG. Bacterial group II introns: not just splicing. FEMS Microbiol Rev. 2007; 31:342–358. [PubMed: 17374133]
15. Benner SA, Ellington AD, Tauer A. Modern metabolism as a palimpsest of the RNA world. Proc Natl Acad Sci USA. 1989; 86:7054–7058. [PubMed: 2476811]
16. Rosenblad MA, Larsen N, Samuelsson T, Zwieb C. Kinship in the SRP RNA family. RNA Biol. 2009; 6:508–516. [PubMed: 19838050]
17. Puerta-Fernandez E, Barrick JE, Roth A, Breaker RR. Identification of a large noncoding RNA in extremophilic eubacteria. Proc Natl Acad Sci U S A. 2006; 103:19490–19495. [PubMed: 17164334]
18. Weinberg Z, Perreault J, Meyer MM, Breaker RR. Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. Nature. 2009; 462:656–659. [PubMed: 19956260]
19. Weinberg Z, Lünse CE, Corbino KA, Ames TD, Nelson JW, Roth A, Perkins KR, Sherlock ME, Breaker RR. Detection of 224 candidate structured RNAs by comparative analysis of specific subsets of intergenic regions. Nucleic Acids Res. 2017; doi: 10.1093/nar/gkx699
20. 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. [PubMed: 15096624]
21. 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. [PubMed: 16086852]
22. 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. [PubMed: 17621584]

23. Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, Breaker RR. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. Genome Biol. 2010; 11:R31. [PubMed: 20230605]
24. Roth A, Weinberg Z, Chen AG, Kim PB, Ames TD, Breaker RR. A widespread self-cleaving ribozyme class is revealed by bioinformatics. Nat Chem Biol. 2014; 10:56–60. [PubMed: 24240507]
25. Weinberg Z, Kim PB, Chen TH, Li S, Harris KA, Lünse CE, Breaker RR. New classes of self-cleaving ribozymes revealed by comparative genomics analysis. Nat Chem Biol. 2015; 11:606–610. [PubMed: 26167874]
26. Nitzan M, Rehani R, Margalit H. Integration of bacterial small RNAs in regulatory networks. Annu Rev Biophys. 2017; 46:131–148. [PubMed: 28532217]
27. Jimenez RM, Polanco JA, Lupták A. Chemistry and biology of self-cleaving ribozymes. Trends Biochem Sci. 2015; 40:648–661. [PubMed: 26481500]
28. McCown PJ, Corbino KA, Stav S, Sherlock ME, Breaker RR. Riboswitch diversity and distribution. RNA. 2017; 23:995–1011. [PubMed: 28396576]
29. Block KF, Puerta-Fernandez E, Wallace JG, Breaker RR. Association of OLE RNA with bacterial membranes via an RNA-protein interaction. Mol Microbiol. 2011; 79:21–34. [PubMed: 21166891]
30. Wallace JG, Zhou Z, Breaker RR. OLE RNA protects extremophilic bacteria from alcohol toxicity. Nucleic Acids Res. 2012; 40:6898–6907. [PubMed: 22561371]
31. Julsing MK, Rijpkema M, Woerdenbag HJ, Quax WJ, Kayser O. Functional analysis of genes involved in the biosynthesis of isoprene in *Bacillus subtilis*. Appl Microbiol Biotechnol. 2007; 75:1377–1384. [PubMed: 17458547]
32. Ingram LO. Ethanol tolerance in bacteria. Crit Rev Biotechnol. 1990; 9:305–319. [PubMed: 2178781]
33. Huffer S, Clark ME, Ning JC, Blanch HW, Clark DS. Role of alcohols in growth, lipid composition, and membrane fluidity of yeasts, bacteria, and archaea. Appl Environ Microbiol. 2011; 77:6400–6408. [PubMed: 21784917]
34. Yang S, Giannone RJ, Dice L, Yang ZK, Engle NL, Tschaplinski TJ, Hettich RL, Brown SD. *Clostridium thermocellum* ATCC27405 transcriptomic, metabolomic and proteomic profiles after ethanol stress. BMC Genomics. 2012; 13:336. [PubMed: 22823947]
35. Williams TI, Combs JC, Lynn BC, Strobel HJ. Proteomic profile changes in membranes of ethanol-tolerant *Clostridium thermocellum*. Appl Microbiol Biotechnol. 2007; 74:422–432. [PubMed: 17124583]
36. Michel F, Westhof E. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J Mol Biol. 1990; 216:585–610. [PubMed: 2258934]
37. Mazodier P, Davies J. Gene transfer between distantly related bacteria. Annu Rev Genet. 1991; 25:147–171. [PubMed: 1812805]
38. Chibani-Chennoufi S, Bruttin A, Dillmann ML, Brussow H. Phage-host interaction: an ecological perspective. J Bacteriol. 2004; 186:3677–3686. [PubMed: 15175280]
39. Chen AG. PhD thesis Yale University; New Haven, CT: 2014 Functional investigation of ribozymes and ribozyme candidates in viruses, bacteria and eukaryotes.
40. Curcio MJ, Derbyshire KM. The outs and ins of transposition: from mu to kangaroo. Nat Rev Mol Cell Biol. 2003; 4:865–77. [PubMed: 14682279]
41. Stoddard BL. Homing endonuclease structure and function. Q Rev Biophys. 2005; 38:49–95. [PubMed: 16336743]
42. He S, Corneloup A, Guynet C, Lavatine L, Caumont-Sarcos A, Siguier P, Marty B, Dyda F, Chandler M, Ton Hoang B. The IS200/IS605 family and “peel and paste” single-strand transposition mechanism. Microbiol Spectr. 2015; 3 doi:10.1128.
43. Webb CH, Riccitelli NJ, Ruminski DJ, Lupták A. Widespread occurrence of self-cleaving ribozymes. Science. 2009; 326:953. [PubMed: 19965505]
44. Perreault J, Weinberg Z, Roth A, Popescu O, Chartrand P, Ferbeyre G, Breaker RR. Identification of hammerhead ribozymes in all domains of life reveals novel structural variations. PLoS Comput Biol. 2011; 7:e1002031. [PubMed: 21573207]

45. McCown PJ, Liang JJ, Weinberg Z, Breaker RR. Structural, functional, and taxonomic diversity of three preQ$_1$ riboswitch classes. Chem Biol. 2014; 21:880–889. [PubMed: 25036777]
46. Weinberg Z, Nelson JW, Lünse CE, Sherlock ME, Breaker RR. Bioinformatic analysis of riboswitch structures uncovers variant classes with altered ligand specificity. Proc Natl Acad Sci USA. 2017; 114:E2077–E2085. [PubMed: 28265071]
47. Shi Y, Tyson GW, DeLong EF. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. Nature. 2009; 459:266–269. [PubMed: 19444216]
48. 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 U S A. 2008; 105:3805–3810. [PubMed: 18316740]
49. Agafonov DE, Kolb VA, Spirin AS. Ribosome-associated protein that inhibits translation at the aminoacyl-tRNA binding stage. EMBO Rep. 2001; 2:399–402. [PubMed: 11375931]
50. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. CDD: NCBI's conserved domain database. Nucleic Acids Res. 2015; 43:D222–226. [PubMed: 25414356]
51. Barabas O, Ronning DR, Guynet C, Hickman AB, Ton-Hoang B, Chandler M, Dyda F. Mechanism of IS200/IS605 family DNA transposases: activation and transposon-directed target site selection. Cell. 2008; 132:208–220. [PubMed: 18243097]
52. Creecy JP, Conway T. Quantitative bacterial transcriptomics with RNA-seq. Curr Opin Microbiol. 2015; 23:133–140. [PubMed: 25483350]
53. Bochner BR. Global phenotypic characterization of bacteria. FEMS Microbiol Rev. 2009; 33:191–205. [PubMed: 19054113]
54. Shuman HA, Silhavy TJ. The art and design of genetic screens: *Escherichia coli*. Nat Rev Genet. 2003; 4:419–431. [PubMed: 12776212]
55. Regulski EE, Breaker R. In-line probing of riboswitches. Methods Mol Biol. 2008; 419:53–67. [PubMed: 18369975]
56. Rice GM, Busan S, Karabiber F, Favorov OV, Weeks KM. SHAPE analysis of small RNAs and riboswitches. Methods Enzymol. 2014; 549:165–187. [PubMed: 25432749]
57. McHugh C, Russell P, Guttman M. Methods for comprehensive experimental identification of RNA-protein interactions. Genome Biol. 2014; 15:203. [PubMed: 24467948]
58. Simon MD. Insight into lncRNA biology using hybridization capture analyses. Biochim Biophys Acta. 2016; 1859:121–127. [PubMed: 26381323]
59. Smirnov A, Forstner KU, Holmqvist E, Otto A, Gunster R, Becher D, Reinhardt R, Vogel J. Grad-seq guides the discovery of ProQ as a major small RNA-binding protein. Proc Natl Acad Sci USA. 2016; 113:11591–11596. [PubMed: 27671629]
60. Montero Llopis P, Jackson AF, Sliusarenko O, Surovtsev I, Heinritz J, Emonet T, Jacobs-Wagner C. Spatial organization of the flow of genetic information in bacteria. Nature. 2010; 466:77–81. [PubMed: 20562858]
61. Wassarman KM, Storz G. 6S RNA regulates *E. coli* RNA polymerase activity. Cell. 2000; 101:613–623. [PubMed: 10892648]

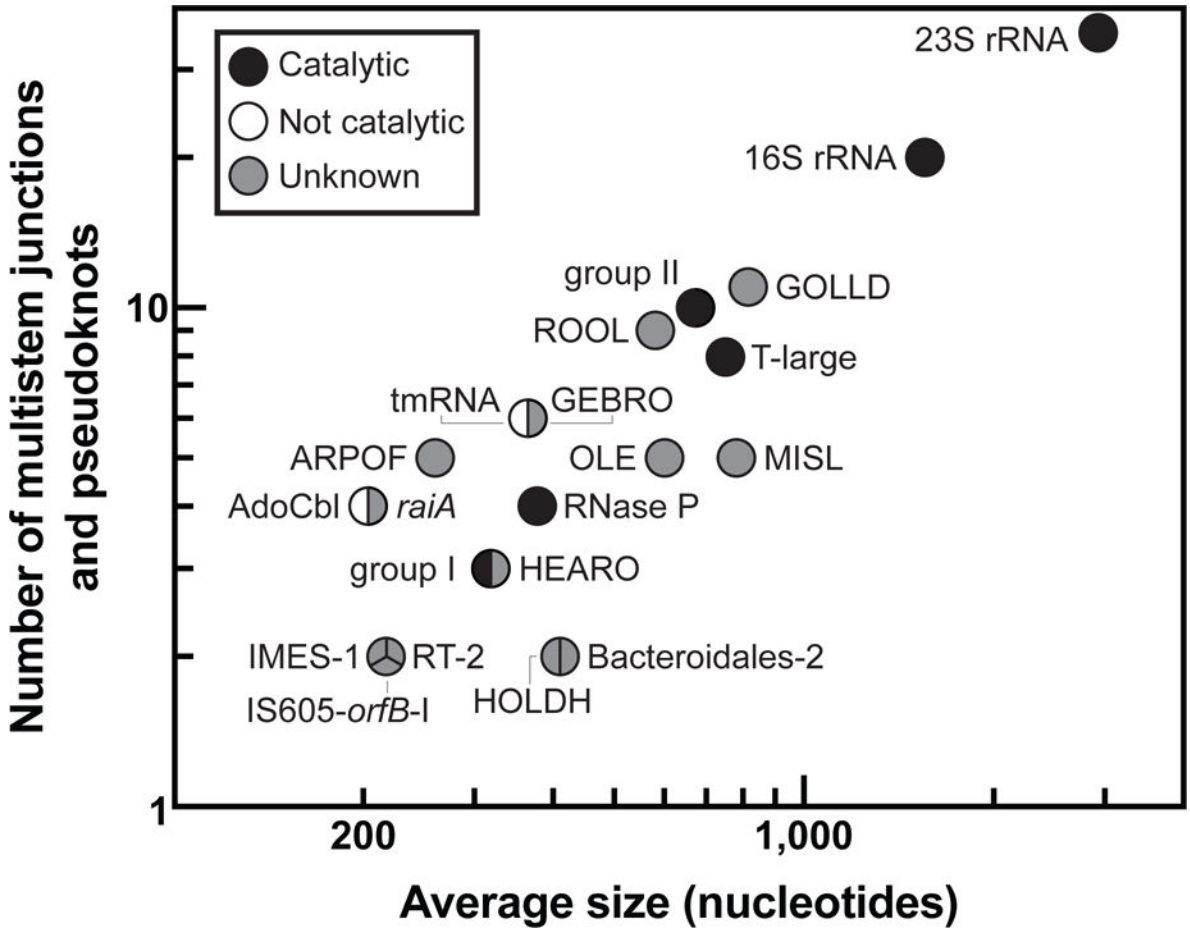

Figure 1. Size and structural complexity of large and highly-structured ncRNAs in bacteria

Structural complexity is represented by the number of multistem junctions and pseudoknots present in the predicted secondary structure models as described previously (17). Overlapping points representing different ncRNAs are depicted with split circles. Narrowly distributed ncRNAs and ncRNAs with fewer than two multistem junctions and pseudoknots were omitted. For example, noncoding RNAs such as large sRNAs and CRISPR RNAs are commonly longer than 200 nucleotides, but have repetitive and simple hairpin secondary structures that are bound by proteins. Although, 23S rRNA forms the active site for the peptidyl transferase reaction catalyzed by ribosomes, 16S rRNA functions in complex with the catalytic RNA component and is classified accordingly.

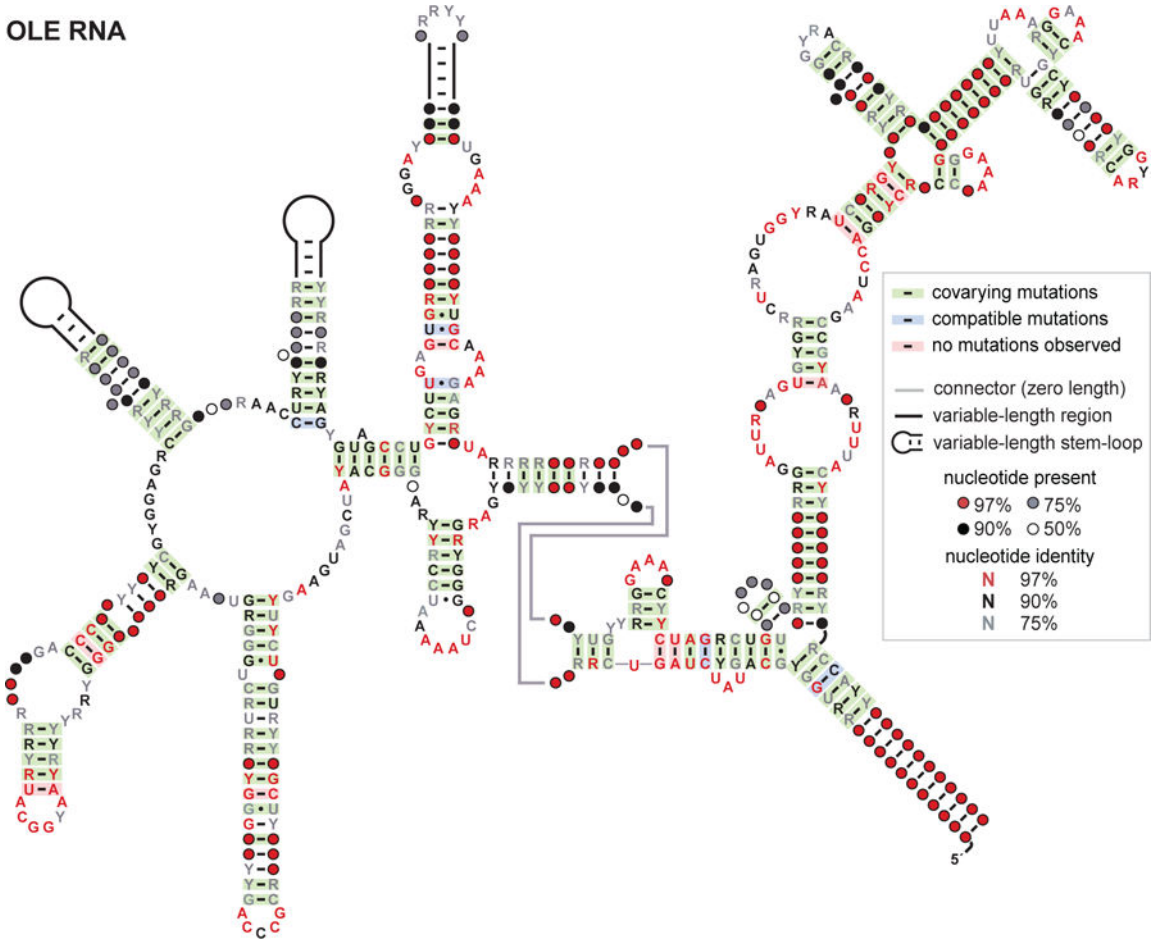

Figure 2. Consensus sequence and secondary structure model for OLE RNAs

This model is based on the alignment of 657 unique representatives from genomic sequences from RefSeq version 63 and metagenomic sequences as described in (24). R and Y represent purine and pyrimidine nucleotides, respectively.

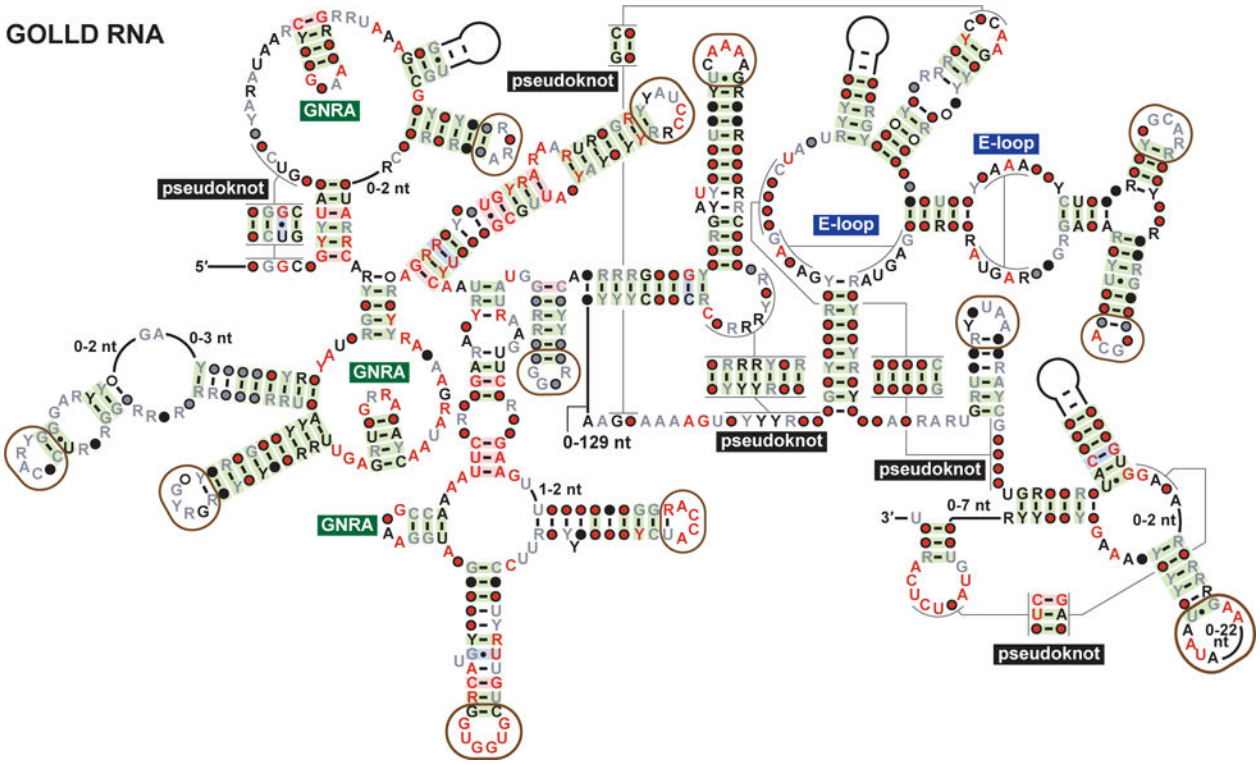

Figure 3. Consensus sequence and secondary structure model for GOLLD RNAs

This model is based on the alignment of sequences identified in (17). Notable predicted substructures include two E-loops, three GNRA tetraloops, and five pseudoknots. Of the 20 hairpin loops, five form pseudoknots or represent the GNRA tetraloops. A total of 12 of the remaining 15 hairpin loops carry highly-conserved nucleotides, suggesting that they might be involved in forming RNA tertiary contacts that are important for the function of GOLLD RNA. Other annotations are as described for Figure 2.

Table 1
Large ncRNAs in bacteria with unpublished functions

The initial reports of these RNA motifs were published elsewhere (15–17).

RNA	Avg. Size	#	Nucleotide Conservation[a]	Taxa
ARRPOF	260	78	35%	Fusobacteria
Bacteroidales-2	419	355	56%	Bacteroidales
Clostridiales-3	252	559	40%	Clostridiales
EGFOA-assoc-1	251	23	64%	environmental
GEBRO	349	66	49%	*Streptococcus* (Firmicutes)
GOLLD	829	391	34%	Firmicutes, Actinobacteria
HEARO	350	3283	29%	Firmicutes, Proteobacteria, Cyanobacteria, Actinobacteria
HOLDH	401	22	52%	environmental
ilvB-OMG	209	39	70%	OMG group (γ-Proteobacteria)
IMES-1	217	491	59%	marine environmental
IS605-*orfB*-I	213	444	37%	*Enterococcus* (Firmicutes)
lysM-Actino	211	359	54%	Actinomycetales
MISL	782	55	33%	Verrucomicrobia
OLE	596	657	40%	Firmicutes
PAGEV	223	123	33%	γ-Proteobacteria
raiA	211	1347	43%	Actinobacteria, Firmicutes
ROOL	581	397	17%	Firmicutes, Fusobacteria, Tenericutes
RT-2	214	482	21%	Coriobacteriales (Actinobacteria), Clostridiales (Firmicutes)
RT-7	201	202	34%	Bacteroidales
T-Large	765	291	33%	Proteobacteria, Cyanobacteria, and Deinococcus–Thermus
throat-1	294	63	71%	throat and tongue metagenomes

[#]Number of examples, with data derived from the microbial dataset of RefSeq version 63.

[a]Nucleotide conservation is computed as a percentage of the average size in nucleotides divided by the total number of nucleotides conserved in 75% or more of the representatives for each motif.