Published online 27 March 2007

Nucleic Acids Research, 2007, Vol. 35, No. 7 ***2283–2294***
doi:10.1093/nar/gkm057

Deciphering RNA structural diversity and systematic phylogeny from microbial metagenomes

Yanglong Zhu[1], Dileep K. Pulukkunat[2] and Yong Li[1,*]

[1]Department of Biochemistry and Molecular Biology, and Center for Genetics and Molecular Medicine, School of Medicine, University of Louisville, 319 Abraham Flexner Way, Louisville, KY, 40202, USA and [2]Ohio State Biochemistry Program, Department of Biochemistry, Ohio State University, Columbus, OH 43210, USA

Received August 15, 2006; Revised January 19, 2007; Accepted January 20, 2007

ABSTRACT

Metagenomics has been employed to systematically sequence, classify, analyze and manipulate the entire genetic material isolated from environmental samples. Finding genes within metagenomic sequences remains a formidable challenge, and noncoding RNA genes other than those encoding rRNA and tRNA are not well annotated in metagenomic projects. In this work, we identify, validate and analyze the genes coding for RNase P RNA (P RNA) from all published metagenomic projects. P RNA is the RNA subunit of a ubiquitous endoribonuclease RNase P that consists of one RNA subunit and one or more protein subunits. The bacterial P RNAs are classified into two types, Type A and Type B, based on the constituents of the structure involved in precursor tRNA binding. Archaeal P RNAs are classified into Type A and Type M, whereas the Type A is ancestral and close to Type A bacterial P RNA. Bacterial and some archaeal P RNAs are catalytically active without protein subunits, capable of cleaving precursor tRNA transcripts to produce their mature 5′-termini. We have found 328 distinctive P RNAs (320 bacterial and 8 archaeal) from all published metagenomics sequences, which led us to expand by 60% the total number of this catalytic RNA from prokaryotes. Surprisingly, all newly identified P RNAs from metagenomics sequences are Type A, i.e. neither Type B bacterial nor Type M archaeal P RNAs are found. We experimentally validate the authenticity of an archaeal P RNA from Sargasso Sea. One of the distinctive features of some new P RNAs is that the P2 stem has kinked nucleotides in its 5′ strand. We find that the single nucleotide J2/3 joint region linking the P2 and P3 stem that was used to distinguish a bacterial P RNA from an archaeal one is no longer applicable, i.e. some archaeal P RNAs have only one nucleotide in the J2/3 joint. We also discuss the phylogenetic analysis based on covariance model of P RNA that offers a few advantages over the one based on 16S rRNA.

INTRODUCTION

Metagenomics (also known as environmental genomics or community genomics) is the study of genomes of microbial organisms recovered directly from their natural environments (1,2). Through whole-genome shotgun sequencing of pooled DNA from environmental samples, metagenomics has been employed as a means of systematically investigating the nucleotide sequence, structure, regulation and function of genes. The main benefit of metagenomics is that it provides the capacity to effectively characterize the genetic diversity present in samples, bypassing the need for isolation and lab cultivation of individual species. Information from metagenomic data has the ability to enrich our knowledge of community metabolism, microbial diversity, environmental ecosystem and enzymology for industrial applications (3–5).

Finding and annotating genes is one of the fundamental goals in virtually all metagenomics projects, regardless of whether complete genome sequences can be assembled or not (2). Yet, it has become apparent that noncoding RNA genes, particularly the RNase P RNA gene of interest, are not well annotated in metagenomic projects. This gene encodes the RNA component (P RNA) of RNase P, a ubiquitous endoribonuclease responsible for cleavage of the 5′ leader of precursor tRNAs to generate the mature 5′-ends. The bacterial P RNA has been classified into two types: Type A and Type B. Type A is suggested to be the ancestral form and is the source from which Type B is derived (6). Significant differences in secondary structure are found between these two types; however, main elements prominent in the tertiary structure are conserved.

*To whom correspondence should be addressed. Tel: +1-502-852-7551; Fax: +1-502-852-6222; Email: yong.li@louisville.edu

Indeed, all P RNAs are composed of two domains: domain I is the Catalytic domain (C domain) and domain II is the Specificity domain (S domain) recognizing the TψC stem of the pre-tRNA substrate (7). Structural divergence exists in domain I, but the P4 pseudoknot region proposed to execute substrate binding and catalytic function is conserved (8). A striking difference between the two types lies in the folding of the pre-tRNA recognition responsible elements (in the S domain) (Figure 1). For Type A, stacked P13 and P14 helices are implicated, whereas for Type B the main element responsible is the P10.1 helix (6,9–11). Despite different helical packing in tertiary structures, the three strategic points that build the pre-tRNA recognition interface are in close proximity (10,11). Archaeal P RNAs are similar to bacterial ones in both primary and secondary structures, but they are not as catalytically proficient *in vitro* by themselves as bacterial ones (6,9,12).

Primary and secondary structural information of P RNA is widely employed to search P RNA genes in complete genomes (13,14). In this study, we use three structural models (bacterial Type A, bacterial Type B and Archaeal) to identify, validate and analyze the genes coding for P RNA in recently published metagenomes. Moreover, we use this RNA as a molecular marker to analyze community complexity and systematic phylogeny.

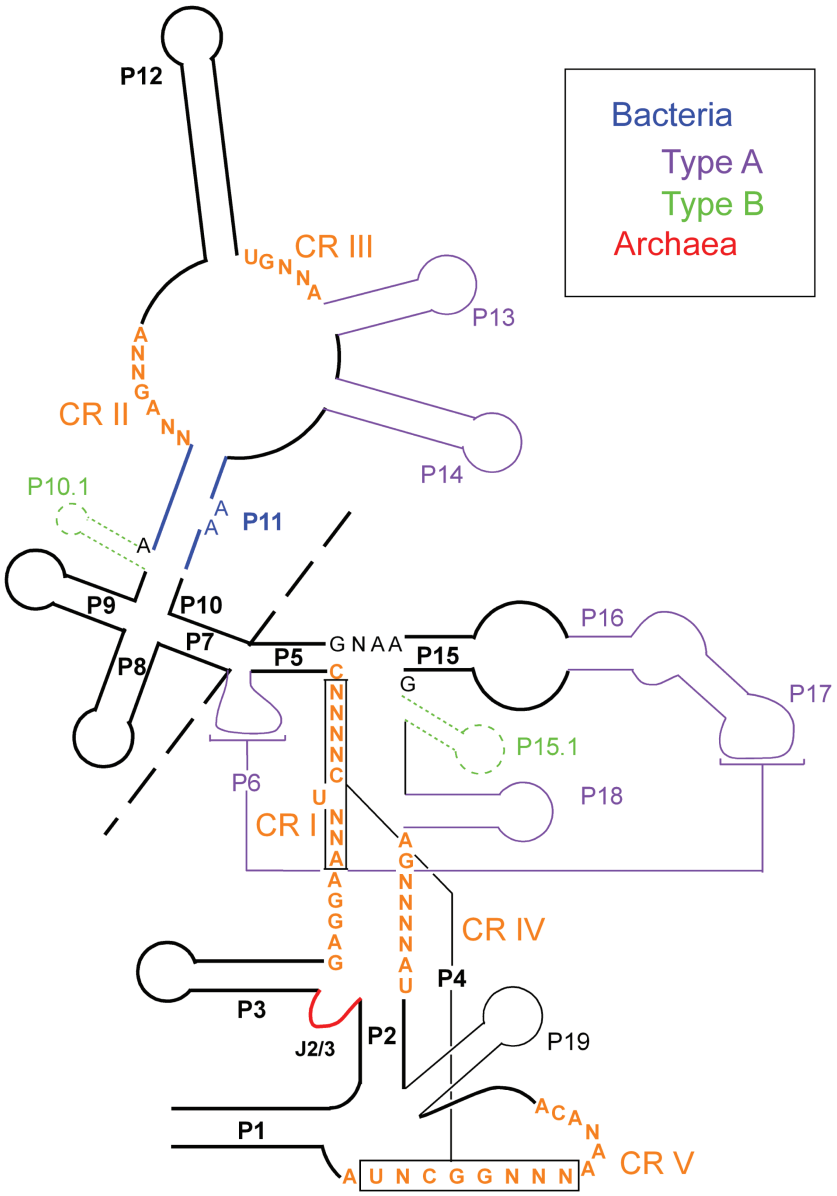


Figure 1. The P RNA structural model for bacteria and archaea. All P RNAs have five conserved regions (CR I–CR V). Bacterial and archaeal sequences are distinguished by the P11 stem. Type A and Type B bacterial P RNAs are distinctive based on the P10.1 and P15.1 stem (Type B; shown in dashed green lines). See text for details about the J2/3 region of archaeal P RNAs. The difference between Type A and Type M of archaeal P RNA is based on the P8 stem (Type M lacks P8, shown in dashed line). Catalytic domain (right bottom) and Specificity Domain (left top) are divided by dashed lines.

MATERIALS AND METHODS

Sequence sources

The metagenomic sequence source (Table 1) was from Acid Mine Biofilm (abbreviated as AM) (15), Sargasso Sea (SS) (16), Minnesota Soil (MS) (17), Whale Falls (W1, W2 and W3) (17), Deep Sea Sediment (DS) (18). In addition, we used other small-scale sequences from uncultured microbes (designated as 'Uncultured Others', UO) (19–25). The reference P RNAs were from RNase P database (http://www.mbio.ncsu.edu/RNaseP/) (26) and the Rfam database (http://rfam.wustl.edu/) (27).

Conserved sequence pattern search

We used a PERL script to search the metagenomics sequences, utilizing the pattern for the conserved regions CR I and CR V of P RNAs as identified previously (8,14). Since the conserved base patterns for Bacteria and Archaea are very similar, we combined them into a single pattern, (CR I) 5′GAGGAANNUCNNNNC3′ and (CR V) 5′ACANGANNNNGNNUC3′, and allow two mutations and a maximum of 4000 nt between these two conserved regions.

Analysis with INFERNAL

The INFERNAL program (version 0.7) was retrieved from Rfam (University of Washington at St. Louis) (13,27). INFERNAL is an implementation of covariance model (CM), which is a statistical model of RNA secondary structure and sequence consensus. P RNA seed sequence alignments (the models) were downloaded from the databases (University of Washington St. Louis http://rfam.wustl.edu/). Their accession numbers are RF000373 (archaeal), RF00010 (bacterial Type A) and RF00011 (bacterial Type B). A covariance model (CM) is built from each seed alignment. The INFERNAL command cmsearch was run on the preliminary candidate sequences from the conserved sequence pattern search using default parameters (cmsearch –W 500 CM SEQDB). The INFERNAL model match results are used as a guideline for finding conserved regions, hairpin and loop structures. In certain cases, the web-based utility Mfold version 3.2 (28) was used to predict the secondary structures that are subsequently compared with those derived from INFERNAL results.

Phylogeny reconstruction

We used Clustal W (version 1.83) and INFERNAL to generate multiple-sequence alignment from the P RNA sequences, respectively. The phylogenetic distance of the whole aligned sequences was calculated using the neighbor-joining (NJ) (Saitou and Nei 1987) methods as implemented in Clustal W with IUB weight matrix in the web service at http://align.genome.jp/. We then used drawgram from the PHYLIP software package with bootstrapping to assess the reproducibility of the reconstruction (http://evolution.genetics.washington.edu/phylip.html) (Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6; distributed by the author, Department of Genome Sciences, University of Washington, Seattle) to reconstruct the phylogenetic tree diagram.

Enzymatic assay of a P RNA from Sargasso Sea

The gene coding for a P RNA (AACY01084936) from Sargasso Sea (abbreviated as SS RNA) was chemically synthesized and placed downstream of a T7 promoter. The genes encoding RNase P protein subunits (Rpps) Pop5, Rpp30, Rpp29 and Rpp21 of *Methanocaldococcus jannaschii* (*Mja*) were cloned under the T7 promoter and transformed into *E. coli* strain BL-21 (DE3) expressing T7 RNA polymerase under the control of IPTG inducible *lac* promoter. The detailed description of the cloning approach and protein overexpression and purification will be published elsewhere. *In vitro* reconstitution of the SS RNA with *Mja* Rpps was achieved by pre-incubating 500 nM RNA with 5 μM *Mja* Rpps (Pop5, Rpp30, Rpp29 and Rpp21) in RC buffer [50 mM Tris-HCl (pH 7.5), 60 mM $MgCl_2$ and 1.0 M $(NH)_4OAc$] at 37°C for 10 min followed by 55°C for 10 min. RNase P assay was performed in RC buffer using 2 μM *E. coli* $ptRNA^{Tyr}$, a trace amount of which was labeled with α–[^{32}P]-GTP as the substrate at 55°C for 2 h. The reaction was stopped by adding 10 μl of quenching dye [10 M urea, 1 mM EDTA, 0.05% (w/v) xylene cyanol, 10% (v/v) phenol], and the products were separated in 8% (w/v) polyacrylamide/7 M urea gel and visualized by autoradiography.

Table 1. P RNAs identified in metagenomics projects

Metagenomic projects	Archaea	Bacterial Type A	Bacterial Type B	Total number of P RNA	Total nucleotides sequenced (M bps)	Reference
Acid Mine Biofilm (AM)	4 (3)[a]	1 (1)	0	5	75	(15)
Deep Sea Sediment (DS)	1	0	0	1	111	(18)
Minnesota Soil (MS)	0	15	0	15	100	(17)
Whale Falls (W1, W2 and W3)	0	17	0	17	25	(17)
Sargasso Sea (SS)	6	289 (2)	0	295	1045	(16)
Uncultured Others (UO)	1	0	0	1	0.6	(19–25)
Total	12	322	0	334	1911	
Known P RNA	50[b]	391[b]	79[b]	520[b]		(26,27)

[a]4(3) denotes that 4 P RNA found, yet 3 of them were reported in a previous study (14).
[b]There are 99/455/145 archaeal/bacteria Type A/bacterial Type B P RNA entries in Rfam and RNase P databases (26,27), yet some of them are repetitive sequences. This number represents the unique P RNA (not identical to any other entry).

RESULTS

Identification of P RNA genes based on covariance model

To search for the P RNA gene among metagenomes, we first applied the INFERNAL program directly to DNA sequences from various metagenomic projects (Table 1). However, we found it would be extremely time consuming with a standard workstation since INFERNAL cmsearch (covariance search) is computation intensive. It took a moderately powerful workstation (a PC with a 3.2 GHz Pentium 4 and 1 GB DDR-RAM) ~12 h to process a 1.5 M bp DNA sequence (the genome of *Aquifex aeolicus*). It is estimated that the time to screen the Sargasso Sea sequences (1045 M bp × 2) alone by cmsearch would be ~700 days on a typical workstation.

To reduce the size of the sequences to be input into cmsearch, a pre-screening was performed on the initial raw sequences using a PERL script (14). The PERL script was designed to search the conserved nucleotides of P RNAs, namely the Conserved Region CR I (GAGGAAN NUCNNNNC) and CR V (ACANAANNNNGNNUC). Two nucleotide mutations were allowed in the conserved sequences. The pre-screening yielded 2794 preliminary candidate sequences within about 1 h on the workstation we utilized. The set of the preliminary candidate sequences (the length range from ~200 to ~4200) was then subjected to INFERNAL analysis using covariance models built from the seed alignments of bacterial Type A, Type B and archaeal models from the Rfam database, respectively.

This two-step search resulted in over 300 candidate sequences for manual analysis. We then identified the conserved regions CR II, CR III and CR IV using INFERNAL cmsearch results. Additional efforts including RNA folding with Mfold and manual sequence alignment for compensatory changes were used to determine the authenticity of these putative P RNA sequences. In the end, we found 334 sequences that met the criteria for P RNA as following: (1) five conserved regions, CR I to CR V (Figure 1) (8); (2) the structural folding for stem P1 through P19 (only 14 out of the 334 molecules do not have the upstream sequence of the P1 helix; these sequences all started from the beginning of contigs); (3) the conserved P11 structure: for bacterial sequences, a 5-bp stem is interrupted by 2 nt (mostly AA), while no interruption occurs for archaeal sequences; (4) the joint region for P2 and P3 of bacterial sequences is always a 'G' (rarely A), while that from archaea is 3–4 nt long (with a few exceptions, see below); (5) all the P RNA sequences do not overlap with any other gene annotated in the published metagenomics data, as all known P RNA genes. Combined with these criteria and INFERNAL model search scores, we determined that in these 334 P RNA genes, 322 are from bacteria and 12 from archaea. We designate these P RNAs bacterial or archaeal based on two criteria: one is the difference of their P11 stems (Figure 2) and another that none of bacterial P RNAs could be scored with the archaeal CM model and vice versa.

Six of these P RNAs (three archaeal: AADL01001488, AADL01001899 and AADL01001374; three bacterial: AADL01000212, AACY01000258 and AACY01029710; we use the contig accession number to refer the P RNA or its gene and the scaffold number is listed in Supplementary Table 1 if available) from Acid Mine Biofilm and Sargasso Sea projects were identified in a previous report (14). We found that from our newly identified P RNA, only one sequence, AADL01002043 (contig length 2576 bp) from Acid Mine Biofilm, is identical to a known P RNA (AADL01001488, contig length 68 363 bp; also from Acid Mine Biofilm; these two contig sequences only overlap at the P RNA regions). In the end, the total number of distinct P RNAs identified from metagenomics projects is 333, while 328 of them are novel sequences in this study (320 bacterial and 8 archaeal).

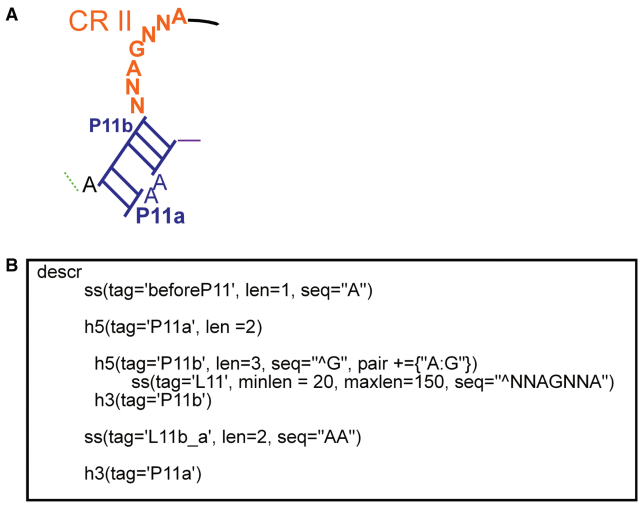


Figure 2. The bacterial and archaeal P RNAs are distinguished by their P11 stem. Only bacterial P11 stem is disrupted by AA dinucleotides. (**A**) The structure model for the bacterial P11 helix. (**B**) The RNAMotif descriptor to identify bacterial P11 and its frank region (A:G pair is frequently observed in the 3-bp segment).

Compared with trusted cutoffs previously reported (17.95 for bacterial Type A and 62.74 for archaeal P RNA), our lowest INFERNAL scores are 13.40 (AAFZ0109957) for bacterial Type A P RNAs and 14.11 (AACY01562772) for archaeal ones. Yet, these two sequences (AAFZ0109957 and AACY01562772) meet all our criteria for designating a P RNA and they are also predicted to be P RNA when aligned with all known P RNAs (Figure S1) and subjected to RNAz, a program for RNA structure conservation and thermodynamic stability (29). Moreover, we experimentally determined the authenticity of an archaeal P RNA (AACY01084936; see below), indicating that the reported cutoff values could be lowered.

No Type B bacterial P RNA from metagenomes

There are two types of bacterial RNase P based on sequence alignments of the RNA component: Type A (such as those from *Escherichia coli* and *T. thermophilus*) and Type B (such as that from *Bacillus subtilis*) (6). Type B RNase P RNA sequences, derived from the ancestral Type A form, are found exclusively in bacterial division of Firmicutes, most of which are Gram-positive strains, which include Bacilli, Clostridia and Mollicutes. Only a small fraction of Firmicutes possesses Type A P RNA (see below). It is quite surprising to note that among a

total of 334 P RNAs identified from all metagenomics sequences, there is not a single sequence of the Type B form. This conclusion was drawn upon the fact that no P RNA can be found through INFERNAL search (cmsearch) with the Type B P RNA Rfam model (RF00011, built upon *B. subtilis* and other P RNAs) and that no P RNAs possess the P10.1 stem (Figure 1), a structural signature for Type B P RNA (Supplementary Table S1 and Figure S1). It is also possible that the model RF00011 that we used to search Type B P RNA is biased against potential candidates that have lower similarity to known Type B P RNAs. We combined all three P RNA models, RF000373 (archaeal), RF00010 (bacterial Type A) and RF00011 (bacterial Type B), into a single P RNA model and conducted a similar INFERNAL search. While we retrieved every P RNA of bacterial Type A and archaeal sequences listed in Table S1 (with different scores, data not shown), we did not find any additional sequence scored by the INFERNAL program, indicating that the absence of Type B P RNA in metagenomics is unlikely a result of a biased model.

No Type M archaeal P RNA from metagenomes

P RNAs from Archaea are classified as two distinct structural types, Type A and Type M (30). Type A is the common and apparently ancestral structure class, and is strikingly similar to the ancestral Type A P RNAs of bacteria. The most significant structure feature of Type M P RNA is the lacking of P8 (30). Bacterial P8 stem is demonstrated to be involved in substrate TψC loop recognition (31). Type M P RNA also lacks L15, P16, P17 and P6, while it has an elongated and uninterrupted P10/P11 stem (30). We folded all 12 archaeal P RNAs (Table S1) according to the covariance model. All of them are of Type A with the presence of P8, P16, P17 and P6 stems and an interrupted P10/P11 stem (Figure 3 gives two examples: AACY01084936 and AB201308). We also performed an INFERNAL search with a CM model solely using known Type A or Type M archaeal sequences, and all 12 RNAs were scored with the Type A model but not the Type M model (data not shown).

An archaeal P RNA is active in the presence of archaeal RNase P proteins

To experimentally determine the authenticity of P RNA from metagenomes, we synthesized the gene encoding a P RNA (AACY01084936/CH004694 with an INFERNAL score of 22.95, Figure 3A) from Sargasso Sea (SS RNA). The RNA was prepared with *in vitro* T7 transcription and then subjected to precursor tRNA cleavage assay. The RNA alone was not active even with high concentration of magnesium and salts or with a bacterial RNase P protein (*E. coli* C5). However, when reconstituted with heterozygous RNase P proteins from an archaeon *Methanocaldococcus jannaschii* (*Mja*), the hybrid enzyme containing the metagenomic P RNA from Sargasso Sea and four *Mja* RNase P proteins was catalytically active to process a precursor tRNA into a mature tRNA and 5′-leader (Figure 3C). These results not only demonstrate the veracity of the search method to discover novel P RNA from metagenomes in this study, but also give us added confidence in the distinguishing of archaeal P RNAs from bacterial ones.

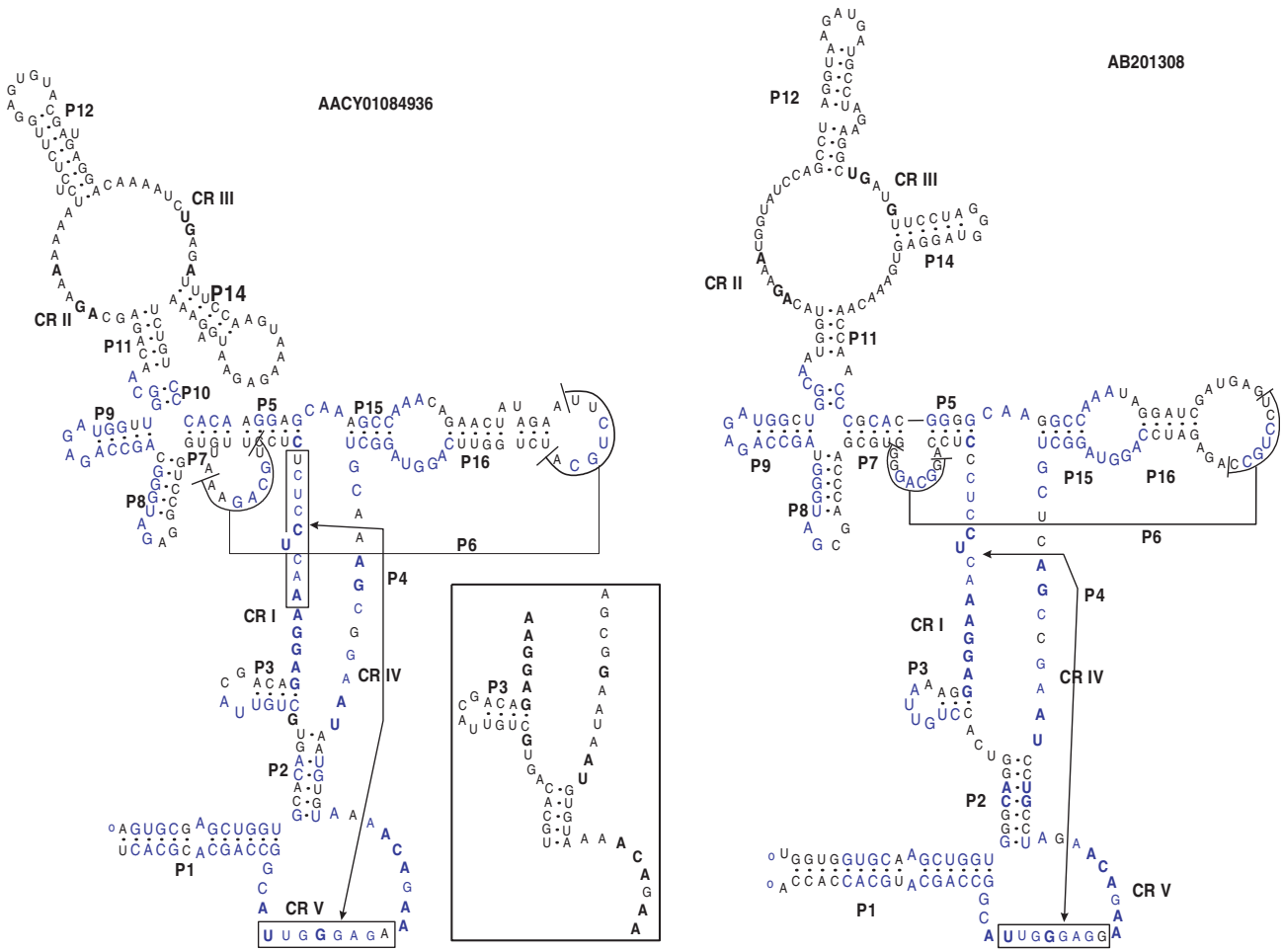
New structural features of Type A bacterial and archaeal P RNAs

Catalytic domain. The catalytic domain of P RNA consists of P1, P2, P3, P4, P5, P15 (P16, P17), P18 and P19 stems. We found 31 bacterial P RNAs with an unusual P2 stem (Table 2). These P2 stems have one kinked nucleotide (mostly U) or two kinked nucleotides as shown in Table 2: 29 with one and 2 with two kinked nucleotides in the middle of the P2 stem. All but one (AAFY01022437) of these sequences are from Sargasso Sea. The kinked nucleotides are located in the upstream strand of the P2 hairpin. Of those P RNAs with kinked P2 stems, only one is archaeal (AACY01562772). Thirteen of the kinked P2 stems are identical (Table 2). We notice that in most cases the kinked U residue can be positioned at either the fourth, or fifth nucleotide of the upstream P2 strand (Table 2). The P2 stem of P RNAs may contain one mismatched base pair, but kinked nucleotides have not been observed previously (14).

The joint between P2 and P3 (J2/3) helices has been used as a feature to distinguish bacterial and archaeal P RNA (14,32). The J2/3 has been characterized as a single 'G' (rarely A) in bacteria and 3–4 nt in archaea. The J2/3 with a single A was observed only in the P RNA from AACY01000258 and NC_002940 (*Haemophilus ducreyi* complete genome) previously (14). We found 17 J2/3 of the new P RNAs with a single nucleotide 'A' in the Sargasso Sea metagenomics sequences (AACY01014401, AACY01037205, AACY01046981, AACY01055431, AACY01076129, AACY01091819, AACY01371505, AACY01613792, AACY01799569, AACY01000258, AACY01001706, AACY01008126, AACY01207302, AACY01300066, AACY01650350 and AACY01681095).

Furthermore, we notice that some archaeal P RNAs possess bacterial type of J2/3 with a single G. Three archaeal P RNAs (all from Sargasso Sea, AACY01031212, AACY01084936 and AACY01513831) have a J2/3 joint with a single nucleotide G (Figure 3A), instead of 3–4 nt as in all known archaeal P RNAs. There is an alternative folding around J2/3 that would give 4 nt, yet it breaks two important criteria: first, the CR IV will be less conserved; secondly, the joint region that connects the P2 stem and CR V will be only 2 nt, instead of at least three as observed in all known P RNAs (14). The proposed folding for these RNAs has one mismatched base pair in its P2 stem, which is allowed in previously identified P RNAs (14). The existence of an archaeal P RNA with a single nucleotide J2/3 joint is further validated with the activity assay of the P RNA (AACY01084936) with RNase P proteins (see above).

Specificity domain. The P11 stem has been discovered to be a structural feature that distinguishes bacterial P RNAs from archaeal ones. The bacterial P11 helix consists of 5 bp, disrupted by 2 nt (mostly AA), and it is flanked by an A and the conserved NNAGNNA sequence (Figure 2A); the archaeal P11 is not disrupted by 2 nts. Using an



C

SS RNA	–	–	+	–	+	+
Mja Rpps	–	–	–	+	–	+
Eco Rpp	–	+	–	–	+	–
Eco RNA	–	+	–	–	–	–

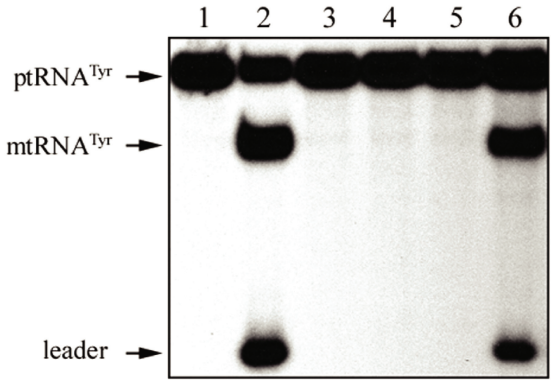


Figure 3. An archaeal P RNA identified from Sargasso Sea with a J2/3 of a single G. (**A**) The folding of this P RNA (AACY01084936) is drawn according to the INFERNAL alignment. In the inset box, an alternative folding of P2 stem and its flanking regions is provided. (**B**) The folding of AB201308. Nucleotides in blue are those identical ones in these two sequences. (**C**) Functional reconstitution of metagenomic P RNA (AACY01084936) with *Mja* RNase P proteins (Rpps). RNase P activity was reconstituted by mixing 500 nM RNA with 5 μM protein (*E. coli* RNase P protein C5, *Eco* Rpp, or *Mja* Rpps, lanes 5 and 6) and assayed at 55°C for 2 h using 2 μM *E. coli* ptRNA^{Tyr} as substrate. Lane 2 represents a control reaction *E. coli* RNase P. Lanes 3 and 4 represent control reactions in which the substrate was incubated with either the RNA alone or Mja Rpps, respectively.

Table 2. P2 stem with kinked nucleotides from bacterial P RNAs

Feature	Accession number	P2 Stem[a]	Note
One kinked nucleotide	AACY01001706		Alternatively U5 kinked
	AACY01023864		
	AACY01085511		
	AACY01100012		
	AACY01171914		
	AACY01177717		
	AACY01313347	5′**GGA*UU*GUC**3′	
	AACY01418924	3′**UCU-*A*CAG**5′	
	AACY01445082		
	AACY01568993		
	AACY01599112		
	AACY01634211		
	AACY01642849		
	AACY01756413		
	AACY01248068	5′**GGA*UU*AUC**3′ 3′**UCU-*A*UAG**5′	Alternatively U5 kinked
	AACY01254266		
	AACY01099407	5′**GGAUCGUC**3′ 3′**UCU-*G*CAG**5′	Alternatively 5 kinked, but less likely though
	AACY01044687	5′**GGAU*AU*UC**3′ 3′**UCUA*U*-AG**5′	Alternatively U7 kinked
	AACY01108785	5′**GGAUAAUC**3′ 3′**UCU**-UU**AG**5′	
	AACY01074654	5′AUAACUU3′ 3′UGUU-AG5′	
	AACY01189100	5′**GGAUCGUC**3′ 3′**UCUA-CAG**5′	
	AACY01207302	5′**GGA**CA**GU**U3′ 3′**UCU**-U**CAG**5′	
	AACY01283453	5′**GGA*UU*GU**U3′ 3′**UCU*A*-CA**A**5′	Alternatively U4 kinked
	AACY01300066	5′**GGAUAGUC**3′ 3′**UCU**-U**CAG**5’	
	AACY01313726	5′**GGAUAGUC**3′ 3′**UCUA*U*-AG**5′	Alternatively A5 kinked
	AACY01357080	5′**GGACAGUC**3′ 3′**UCU**-U**CAG**5’	
	AACY01518059	5′**GGACUGUC**3′ 3′**UCU-ACAG**5′	Alternatively U5 kinked
	AACY-1667100	5′**GGAUAGUC**3′ 3′**UCUA-CAG**5′	
	AAFY01022437	5′AC*GG*UCGCC3′ 3′UG*C*-AGUGG5′	Alternatively G3 kinked
Two kinked nucleotides	AACY01733681	5′**GGAUUGU**U3′ 3′CU**U-CAG**5′	
	AACY01749490	5′**GGA*UU*AGUC**3′ 3′**UCU*A*-CAG**5′	Alternatively U4 kinked

[a]Nucleotides in bold are conserved. The italicized ones can be paired alternatively.

RNAMotif (33) descriptor to search all of the new sequences, we found that the P11 structural feature used to distinguish bacterial and archaeal P RNA is still applicable. In other words, for the newly identified P RNAs, bacterial but not archaeal ones posses the 5-bp P11 stem interrupted by 2 nt (mostly AA) in the 3′ strand (Figure 2).

Community diversity

Using BLASTN against known P RNAs with these 333 metagenomics P RNAs as query sequences, we found that only nine sequences from Sargasso Sea have 97% or higher identity with known P RNA sequences (Table 3). All these sequences are highly similar to *Synechococcus* or *Prochlorococcus* species, which are two known dominating species in seawater.

We then use BLASTN to compare the 333 distinctive P RNA sequences with each other. We found that 243 of them have a sequence similarity of <97%, which is the species-level distinction criterion generally accepted for 16S rRNA phylogenetic analysis. With the deduction of nine of them, which are 97% or more identical to known P RNAs, we obtain 234 P RNA that may represent new species using 97% cutoff (200 of them are from Sargasso Sea). The 97% is frequently used as a similarity cutoff to determine community diversity (16). There are several genomes in the Sargasso Sea sequencing project that are near completion. We identified P RNA genes in some of these incomplete genomes (contig/scaffold accession numbers in parentheses): cf. *Burkholderia* SAR-1 2166002 (AACY01015267/CH004448), cf.

Table 3. Metagenomics P RNAs with over 97% identity with known ones

Contig/scaffold	Sequence similarity[a](%)	Known P RNA (Accession number/Annotation)
AACY01065403/CH009720	98	BX569689.1/*Synechococcus* sp. WH8102 genome
AACY01183524/CH034265	97	BX569689.1/*Synechococcus* sp. WH8102 genome
AACY01070379/CH011043	97	AJ272225.1/*Prochlorococcus marinus* str. TAK9803-2
AACY01043807/CH022746	97	AJ272225.1/*Prochlorococcus marinus* str. TAK9803-2
AACY01628203/CH172592	99	AJ272223.1 (*Prochlorococcus* sp. strain PCC9511)
AACY01010320/CH010342	97	AJ272224.1 (*Prochlorococcus* sp. rnpB gene, strain TATL2)
AACY01117529	97	AJ272219.1 (*Prochlorococcus* sp. rnpB gene, strain PAC1A)
AACY01637978	97	AJ272225.1 (*Prochlorococcus* sp. rnpB gene, strain TAK9803-2)

[a]Sequence similarity between the new metagenomics P RNA with contig/scaffold number and known P RNA.

γ-*Proteobacteria* SAR-1 2220231 (AACY01115054/CH004575), cf. *Spirochaetales* SAR-1 2222388 (AACY01011636/CH004627), cf. *Bacteria* SAR-1 2223244 (AACY01076953/CH004656), cf. γ-*Proteobacteria* SAR-1 2223787 (AACY01090586/CH004705) and cf. *Bacteria* SAR-1 2223849 (AACY01092346/CH004708).

Phylogenetic analysis

It has been reported that P RNAs from different taxa possess distinguishable structural features (6,9,32). Yet, the previous structural analyses of P RNA relying on manual alignment have been restricted to a limited number of sequences. Clustal W is one of the widely used alignment tools for phylogenetic analysis, yet it did not generate any credible alignment for PHYLIP phylogenetic analysis with P RNA even through extensive parameter manipulations (see Supplementary Files). We notice that the Clustal W alignment is not designed to recognize consensus secondary structures of distantly related sequences (with low primary sequence similarity). On the other hand, the INFERNAL program is based on covariance model, which contains a significant amount of secondary structure information. This prompted us to undertake a phylogenetic analysis using the covariance-model-based multiple sequence alignment. Compared with that from Clustal W, phylogenetic analysis with the INFERNAL automated alignment produces a much better classification as judged by the generally accepted phylogeny (Supplementary Figure S2). Over 95% of the sequences are clustered within phylogenetically related ones (Figure S2). In taxa of Euryarchaeota, Chlamydiae, Bacteroidetes, β-Proteobacteria and ε-Proteobacteria, all P RNA sequences are exclusively clustered in their own respective groups (Figure S2). All but one P RNA sequence from the taxa of Crenarchaeota, δ-Proteobacteria, Planctomycetes and α-Proteobacteria are clustered in their respective groups. For Firmicutes, γ-Proteobacteria and Spirochaetes, their P RNA sequences are also clustered correspondingly, but with a number of exceptions (Figure S2).

Comparable to previous diversity analysis of Sargasso Sea sequences using 16S rRNA or multiple-protein phylogenetic markers from Figure 6 and Supplementary Figure 4 of (16), our analysis based on P RNA (Figure S2 and Table 4) has revealed that species from α-Proteobacteria (55%) and γ-Proteobacteria (22%) dominate the Sargasso Sea, followed by Cyanobacteria. However, using 16S rRNA only, Venter *et al.* could not identify any archaeal species in the sequences (16). Euryarchaeota and Crenarchaeota certainly exist in Sargasso Sea, as judged by multiple-protein phylogenetic markers (16). Using the P RNA as a phylogenetic marker, there would be four Euryarchaeota and two Crenarchaeota in the Sargasso Sea sequencing project (Table 4). Particularly, our experimental data that the P RNA (AACY01084936) is of archaeal source further validates the existence of archaeal species in Sargasso Sea. P RNA-based diversity analysis is also sensitive enough to identify two Chloroflexi and three Fusobacteria (Table 4) that are unable to be detected by 16S rRNA-based analysis (16). Using P RNA-based phylogenetic analysis, we also significantly increased the number of P RNA that represent bacterial divisions of Bacterioides (16), Spirochaetes (22) and α-Proteobacteria (167) (Table 4).

The newly discovered kinked P2 stem may be an indicator for an unknown group of bacteria. Our phylogenetic analysis shows that all bacterial P RNAs with a kinked P2 stem are clustered into a closely related group at or below the class level. No P RNA genes from known organisms clustered in this group. This group is clustered within the division of α-Proteobacteria (Figure S2).

We also found that there are four archaeal P RNAs (AACY01031212, AACY01084936, AACY01513831 and AACY01562772) from Sargasso Sea clustered into an independent group close to Crenarchaeota. The first three sequences are archaeal P RNAs with a J2/3 joint of a single G, and they are 98% identical to each other. The fourth member of this group has the archaeal type of J2/3 (i.e. 3 nt) and a kinked P2 stem (Table 2); it is 86% identical to the first three. Together with its closest neighbor AB201308, the only P RNA found from Uncultured Others (UO) (22), this group of five parallels the Crenarchaeota phylum and may represent a novel class of archaea. We notice that AB201308 is structurally close to AACY01084936 with conservation at the P8, P9, P10, P2, P15 and CR IV regions (Figure 3B). The folding of AB201308 further supports the single nucleotide J2/3 of AACY01084936. The 17 bacterial P RNAs with a single nucleotide 'A' J2/3 joint from the Sargasso Sea metagenomics project, however, do not get clustered into a specific taxon; they are spread over Bacteroidetes, α-Proteobacteria, γ-Proteobacteria and Spirochaetes (Figure S2).

Table 4. Taxonomical distribution of metagenomics and known P RNA

		AM	DS	MS	W1	W2	W3	SS	UO	Total new per taxon	Total known from Rfam database
Archaea	Crenarchaeota							4	1	**5**	8
	Euryarchaeota	4	1					2		**7**	75
Actinobacteria								3		**3**	41
Bacterioides				4			1	11		**16**	9
Chlamydiae						1				**1**	30
Chloroflexi				1				2		**3**	2 + 1[a]
Cyanobacteria								18		**18**	27
Firmicutes								1 + 1[a]		**2**	115 + 6[a]
Fusobacteria								3		**3**[b]	2
Group 1	Deinococcus-Thermus			1						**1**	6
	Thermodesulfobacteria							1		**1**	1
Group 2	Chlorobi					1				**1**	3
	Thermotogae									**0**	4
	Verrucomicrobia			4[c]						**4**[c]	1
Nitrospirae		1								**1**	2
Planctomycetes				1				1		**2**	10
Proteobacteria	α			2		3	1	166		**172**	29
	β							1		**1**	11
	γ			1	1 + 1[a]	2		10 + 54[a]		**69**	47 + 5[a]
	δ			1	1	1				**3**	4
	ε				2					**2**	8
Spirochaetes						2		13 + 4[a]		**19**	3 + 6[a]
Thermodesulfobacteria										**1**	1
Total per metagenome		**5**	**1**	**15**	**5**	**10**	**2**	**295**	**1**	334	

[a]n1 + n2 represents the fact that the group is split in two branches on the phylogenetic tree (Figure S2). [b]These three sequences categorically assigned to Fusobacteria (Figure S2).
[c]These four sequences categorically assigned to the Group 2, yet it is not clear how close it is related to the Verrucomicrobia (Figure S2).

DISCUSSION

Gene finding with covariance model

Noncoding RNA genes produce functional RNA molecules rather than encoding proteins. The number of known RNA genes is expanding rapidly due to the deluge of genomic data and recent systematic efforts to detect RNA genes (34,35). The development of Rfam database and the companion INFERNAL program has greatly enhanced our ability to annotate noncoding RNA genes in completed genomes (13,27). Based on 'covariance models' (CMs—also called profile stochastic context-free grammars, or profile SCFG), the INFERNAL program recognizes the conservation of RNA secondary structure and aligns sequences accordingly (13,27). Yet, profile SCFG searches are computationally expensive, making it unrealistic for a laboratory with a standard workstation to annotate a noncoding RNA in metagenomes (with sequences over 1.5 G bases). In this study, we take advantage of the conserved regions of P RNA and preprocess the metagenomic sequences before applying the INFERNAL method to identify and analyze the gene coding for the RNA subunit (P RNA) of RNase P enzyme from metagenomes. Overall, we have identified a total of 328 new P RNAs from all published metagenomics projects. It is possible that some *bona fide* P RNA genes in these metagenomes were mistakenly filtered out in the first step of our search. To test if the addition of preprocessing step with 18 conserved nucleotides reduced the number of P RNA genes to be found, we compared the output with and without a nucleotide mutation in the conserved regions (CR I and CR V), yet we found only three P RNAs (AAFZ01016940, AAGA01017380 and AACY01700332) with one mutation. There were no P RNAs found with two mutations allowed in the conserved sequences. Therefore, it is unlikely that we would have found more P RNA genes in the metagenomes, should we have applied the INFERNAL program directly.

To illuminate the sensitivity and specificity of our search procedure, we apply the two-step method towards 171 microbial genomes published after January 1, 2005. We are able to find the gene encoding P RNA in every genome. The 171 sequences containing P RNA genes are provided in Supplementary Files. Moreover, the determination of a P RNA to be Type A (139 in total)/Type B (23) in bacteria or Type A (9) in archaea are also verified to be accurate based on further structural analysis (P13/14/P10.1 stems in bacteria and P8/L15/P16/P17/P6/P10/P11 stem in archaea, as discussed above).

Only Type A P RNAs from metagenomics sequences

No Type B bacterial P RNA was discovered within the 320 new bacterial P RNAs identified in metagenomic sequences, indicating that Firmicutes that include Bacilli and Mollicutes and mostly possess Type B bacterial P RNAs are less likely to live in these specific environments (acid mine, Sargasso Sea and a soil sample from Minnesota). Similarly, for archaeal P RNA, there is no Type M out of 12 sequences identified in this study. We cannot completely exclude the possibility of bias introduced by metagenomics sample collections.

New features of P RNA structure

A universally conserved kinked single uridine (U) residue (U69, *E. coli* P RNA numbering) in the P4 stem has been identified as a characteristic feature of P RNA, and such a structure may be directly involved in magnesium metal ion binding that is critical to catalysis (36). Yet, kinked nucleotide in the P2 stem had not been found in any P RNA previously. In the newly identified sequences, we found 31 P RNAs with a kinked nucleotide (the majority of them are Us) in their P2 stems. Some kinked U residues can be swapped between the fourth or fifth position of the upstream strand of the P2 stem. According to the crystal structure of *Thermotoga maritima* (37), the P2 stem is exposed to the solution; thus, we speculate that the kinked nucleotide may be involved in protein binding or metal ion binding for the active enzyme complex. The kinked upstream strand of P2 is located on the backside of the catalytic cavity; thus, with protein/metal ion binding, it may change the alignment and distance between the co-axial P2/P3 helix and the co-axial P1/P4/P5 helix, resulting in modulated catalysis.

Previously, two structural features had been used to distinguish a P RNA from bacteria or archaea (14). The first is the joint region J2/3 linking the P2 and P3 stems, that is always a single 'G' (rarely A) in bacteria, but 3–4 nt long, whereas in archaea (J2/3 of 4 nt only appears in *Halobacterium cutirubrum* and *Methanopyrus kandleri*) (14,26). The second is the P11 helix, flanked by an A and the conserved CR II (NNAGNNA) region, which is a 5-bp stem disrupted by 2 nt in bacterial P RNAs (including AACY01084936) but not in archaeal ones. In this study, we have identified three archaeal P RNAs that possess a J2/3 with a single G, so that the P11 stem is now the only structure that can be used to distinguish a bacterial P RNA from an archaeal one (Figure 2).

An archaeal P RNA from Sargasso Sea is active in the presence of archaeal RNase P proteins

One of the P RNAs identified from Sargasso Sea metagenomic project (AACY01084936) is catalytically active only when reconstituted with Rpps from *Methanocaldococcus jannaschii* (*Mja*). The fact that this RNA is not active on its own or with bacterial RNase P RNA protein, but is active with archaeal Rpps implies that it is a genuine P RNA and it is from an archaeon. The success of such a reconstitution also validates the presence of an archaeal P RNA with a single G J2/3 region, which was identified as a bacteria-only feature previously (14). We also reconstituted this P RNA with Rpps from *Methanothermobacter thermoautotrophicus* (*Mth*) and those from *Pyrococcus furiosus* (*Pfu*), yet the resulting enzyme complexes were not catalytic in both cases (data not shown). Interestingly, *Mja* P RNA is Type M, while P RNAs from *Mth* and *Pfu* are Type A (26). However, cross-Type reconstitution occurs, i.e. Type A bacterial P RNA can reconstitute with a RNase P protein from a bacterium possessing a Type B P RNA, and vice versa (38). The metagenomic P RNA (AACY01084936) is firmly classified as a Type A archaeal P RNA based on its structural features: the presence of P18, L15/P16/P17/P6 and interrupted P10/P11 stems (Figure 3A).

Community complexity of Sargasso Sea

Various techniques have been used to estimate species sample size in ecological systems for community diversity studies. Two methods were used to analyze the Sargasso Sea microbial community including: (1) molecular phylogeny with small subunit ribosomal RNA (SSU rRNA) as the taxon identifier (39), and (2) multiple protein-coding genes and 16S rRNA as the taxon identifier (16,40,41). With these two approaches, the Sargasso Sea microbe community is estimated to have at least 1633 ribotypes (39) to some 1800 phylotypes (16). Among these 1800 species, Venter *et al.* claimed to have identified 148 previously unknown species/phylotypes using a 97% cutoff. The P RNA gene is essential, and each organism has one copy of the gene [very rarely predicted to have two copies, yet experimentally untested (42)], so P RNA-based phylogeny may offer some advantages over SSU rRNA (usually with multiple gene copies). We have not found any two P RNAs coexisting in a single annotated scaffold sequence in this study (Supplementary Table S1), further underscoring the singleton feature of P RNA genes. We take our P RNA-based results to a phylogenetic analysis to identify species and resolve community complexity. Using a 97% cutoff that eliminates 95 P RNA sequences, we identified 200 new P RNA/species (from a total of 295289 bacterial and 6 archaeal) from the Sargasso Sea microbial community, which is 35% more than the previously claimed 148 unknown species using 16S rRNA (16). Since previously published data (16) did not provide standard GenBank accession numbers for rRNA sequences used for molecular phylogeny, we were unable to correlate the 16S rRNA and P RNA into a respective contig/scaffold sequence. It is possible that the larger number of known rRNA sequences, reported over 10 000 (43) compared with only hundreds of P RNA (26,27), reduces the number of newly identified rRNA sequences. It is also possible that higher variability of P RNA sequences than that of 16S rRNA leads more 'species' to be identified, which would further implicate the advantage of using P RNA genes as phylogenetic markers.

Construction of the P RNA-based phylogeny

The secondary structure of P RNA was progressively resolved by extensive phylogenetic comparisons (44–47), which is in sound agreement with available crystal structures (10,11). The success of applying phylogenetic comparison to P RNA secondary structure determination indicates that its secondary structure may be of phylogenetic importance. However, earlier studies have not provided a comprehensive analysis across the phylogeny. Cho and colleagues (48) showed that a Clustal W alignment of some 20 Gram-positive bacterial P RNAs formed group relationships similar to ones found in 16S rRNA-based phylogeny. Subsequent studies at genus level demonstrate that P RNA alone provides a better resolution than 16S rRNA in *Prochlorococcus* (49). Another example is that the 16S rRNAs of

Haemobartonella canis and *Mycoplasma haemofelis* were nearly identical (homology of 99.3–99.7%); in contrast, RNase P RNAs have a lower degree (94.3–95.5%) of sequence homology (50). Our attempt to use P RNA secondary structure to perform system phylogeny drew our attention to the INFERNAL program, which is designed for aligning multiple highly conserved or remotely related primary sequences, as long as they have a well-conserved secondary structure. We constructed a covariance model for the prokaryotes, and generated a single alignment encompassing all metagenomics P RNAs (334) and previously known (699) prokaryotic ones to produce a phylogenetic tree (Figures S1 and S2). Based on the clustering of the P RNA sequences from known species, we found the P RNA-based clustering analysis provided rather satisfactory phylogeny at the levels of phylum and below (Figure S2).

Summary

The development of noncoding RNA gene finding methods has enhanced our ability to understand the molecular diversity and community complexity in microbial populations that cannot be cultivated individually. Our analysis of metagenomic P RNA in this study provides the following insights into RNA molecules from metagenomic sequencing projects. First, using a covariance model, we have identified 328 new P RNAs of Type A from metagenomics projects and expanded by 60% the total number of this catalytic RNA from prokaryotes. Second, only Type A P RNA is found in metagenomic sequences, indicating that, in the defined environments where the metagenomic DNA samples were collected, species may preferably carry the ancestral form of RNA molecules. Third, we disclose new structural features of P RNA, namely the kinked P2 stem and the novel archaeal J2/3 joint region. Finally, our experimental approach to reconstitute a metagenomic P RNA with RNase P proteins further endorses the veracity of the computational methods and genuineness of the discovered P RNAs.

ACKNOWLEDGEMENTS

This project is supported by a pilot study grant for Systems Biology from the Center for Genetics and Molecular Medicine, University of Louisville. Y.L. is supported by a Beginning-Grant-In-Aid grant from the American Heart Association. The computational resource in this project is supported by Grant 2 P20 RR-16481 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH). D.K.P is supported by American Heart Association Pre-Doctoral Fellowship (0515218B). D.K.P. gratefully acknowledges research support and funding from a National Institutes of Health grant (R01 GM067807) to Mark P. Foster and Venkat Gopalan (OSU). Y.L. would like to dedicate this work to his late mentor Dr Ying-lai Wang (1907–2001). Y.L. is grateful to Drs Sidney Altman, Venkat Gopalan and Marlene Steffen for their critical reading of the manuscript. Funding to pay the Open Access publication charge was provided by the University of Louisville.

Conflict of interest statement. None declared.

REFERENCES

1. Handelsman,J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, **68**, 669–685.
2. Chen,K. and Pachter,L. (2005) Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput. Biol.*, **1**, 106–112.
3. Tringe,S.G., von Mering,C., Kobayashi,A., Salamov,A.A., Chen,K., Chang,H.W., Podar,M., Short,J.M., Mathur,E.J. *et al.* (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554–557.
4. Schloss,P.D. and Handelsman,J. (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol.*, **6**, 229.
5. Galperin,M.Y. (2004) Metagenomics: from acid mine to shining sea. *Environ. Microbiol.*, **6**, 543–545.
6. Haas,E.S., Banta,A.B., Harris,J.K., Pace,N.R. and Brown,J.W. (1996) Structure and evolution of ribonuclease P RNA in Gram-positive bacteria. *Nucleic Acids Res.*, **24**, 4775–4782.
7. Loria,A. and Pan,T. (1997) Recognition of the T stem-loop of a pre-tRNA substrate by the ribozyme from Bacillus subtilis ribonuclease P. *Biochemistry*, **36**, 6317–6325.
8. Chen,J.L. and Pace,N.R. (1997) Identification of the universally conserved core of ribonuclease P RNA [letter]. *RNA*, **3**, 557–560.
9. Haas,E.S. and Brown,J.W. (1998) Evolutionary variation in bacterial RNase P RNAs. *Nucleic Acids Res.*, **26**, 4093–4099.
10. Krasilnikov,A.S., Xiao,Y., Pan,T. and Mondragon,A. (2004) Basis for structural diversity in homologous RNAs. *Science*, **306**, 104–107.
11. Krasilnikov,A.S., Yang,X., Pan,T. and Mondragon,A. (2003) Crystal structure of the specificity domain of ribonuclease P. *Nature*, **421**, 760–764.
12. Pannucci,J.A., Haas,E.S., Hall,T.A., Harris,J.K. and Brown,J.W. (1999) RNase P RNAs from some Archaea are catalytically active. *Proc. Natl. Acad. Sci. USA*, **96**, 7803–7808.
13. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
14. Li,Y. and Altman,S. (2004) In search of RNase P RNA from microbial genomes. *RNA*, **10**, 1533–1540.
15. Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S. *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.
16. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
17. Tringe,S.G. and Rubin,E.M. (2005) Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.*, **6**, 805–814.
18. Hallam,S.J., Putnam,N., Preston,C.M., Detter,J.C., Rokhsar,D., Richardson,P.M. and DeLong,E.F. (2004) Reverse methanogenesis: testing the hypothesis with environmental genomics. *Science*, **305**, 1457–1462.
19. Lopez-Garcia,P., Brochier,C., Moreira,D. and Rodriguez-Valera,F. (2004) Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers. *Environ. Microbiol.*, **6**, 19–34.
20. Meyerdierks,A., Kube,M., Lombardot,T., Knittel,K., Bauer,M., Glockner,F.O., Reinhardt,R. and Amann,R. (2005) Insights into the genomes of archaea mediating the anaerobic oxidation of methane. *Environ. Microbiol.*, **7**, 1937–1951.
21. Moreira,D., Rodriguez-Valera,F. and Lopez-Garcia,P. (2004) Analysis of a genome fragment of a deep-sea uncultivated Group II euryarchaeote containing 16S rDNA, a spectinomycin-like operon and several energy metabolism genes. *Environ. Microbiol.*, **6**, 959–969.

22. Nunoura,T., Hirayama,H., Takami,H., Oida,H., Nishi,S., Shimamura,S., Suzuki,Y., Inagaki,F., Takai,K. *et al.* (2005) Genetic and functional properties of uncultivated thermophilic crenarchaeotes from a subsurface gold mine as revealed by analysis of genome fragments. *Environ. Microbiol.*, **7**, 1967–1984.
23. Piel,J., Hui,D., Wen,G., Butzke,D., Platzer,M., Fusetani,N. and Matsunaga,S. (2004) Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge Theonella swinhoei. *Proc. Natl. Acad. Sci. USA*, **101**, 16222–16227.
24. Quaiser,A., Ochsenreiter,T., Klenk,H.P., Kletzin,A., Treusch,A.H., Meurer,G., Eck,J., Sensen,C.W. and Schleper,C. (2002) First insight into the genome of an uncultivated crenarchaeote from soil. *Environ. Microbiol.*, **4**, 603–611.
25. Treusch,A.H., Kletzin,A., Raddatz,G., Ochsenreiter,T., Quaiser,A., Meurer,G., Schuster,S.C. and Schleper,C. (2004) Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea. *Environ. Microbiol.*, **6**, 970–980.
26. Brown,J.W. (1999) The Ribonuclease P Database. *Nucleic Acids Res.*, **27**, 314.
27. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–124.
28. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
29. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA*, **102**, 2454–2459.
30. Harris,J.K., Haas,E.S., Williams,D., Frank,D.N. and Brown,J.W. (2001) New insight into RNase P RNA structure from comparative analysis of the archaeal RNA. *RNA*, **7**, 220–232.
31. Harris,M.E., Nolan,J.M., Malhotra,A., Brown,J.W., Harvey,S.C. and Pace,N.R. (1994) Use of photoaffinity crosslinking and molecular modeling to analyze the global architecture of ribonuclease P RNA. *Embo. J.*, **13**, 3953–3963.
32. Haas,E.S., Armbruster,D.W., Vucson,B.M., Daniels,C.J. and Brown,J.W. (1996) Comparative analysis of ribonuclease P RNA structure in Archaea. *Nucleic Acids Res.*, **24**, 1252–1259.
33. Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.
34. Eddy,S.R. (1999) Noncoding RNA genes. *Curr. Opin. Genet. Dev.*, **9**, 695–699.
35. Eddy,S.R. (2002) Computational genomics of noncoding RNA genes. *Cell*, **109**, 137–140.
36. Frank,D.N. and Pace,N.R. (1997) In vitro selection for altered divalent metal specificity in the RNase P RNA. *Proc. Natl. Acad. Sci. USA*, **94**, 14355–14360.
37. Torres-Larios,A., Swinger,K.K., Krasilnikov,A.S., Pan,T. and Mondragon,A. (2005) Crystal structure of the RNA component of bacterial ribonuclease P. *Nature*, **437**, 584–587.
38. Buck,A.H., Dalby,A.B., Poole,A.W., Kazantsev,A.V. and Pace,N.R. (2005) Protein activation of a ribozyme: the role of bacterial RNase P protein. *Embo. J.*, **24**, 3360–3368.
39. Acinas,S.G., Klepac-Ceraj,V., Hunt,D.E., Pharino,C., Ceraj,I., Distel,D.L. and Polz,M.F. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*, **430**, 551–554.
40. Carlson,C.A., Giovannoni,S.J., Hansell,D.A., Goldberg,S.J., Parsons,R., Otero,M.P., Vergin,K. and Wheeler,B.R. (2002) Effect of nutrient amendments on bacterioplankton production, community structure, and DOC utilization in the northwestern Sargasso Sea. *Aquat. Microb. Ecol.*, **30**, 19–36.
41. Ciccarelli,F.D., Doerks,T., von Mering,C., Creevey,C.J., Snel,B. and Bork,P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283–1287.
42. Piccinelli,P., Rosenblad,M.A. and Samuelsson,T. (2005) Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes. *Nucleic Acids Res.*, **33**, 4485–4495.
43. Wuyts,J., Perriere,G. and Van de Peer,Y. (2004) The European ribosomal RNA database. *Nucleic Acids Res.*, **32**, D101–D103.
44. Gutell,R.R., Power,A., Hertz,G.Z., Putz,E.J. and Stormo,G.D. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.*, **20**, 5785–5795.
45. Haas,E.S., Morse,D.P., Brown,J.W., Schmidt,F.J. and Pace,N.R. (1991) Long-range structure in ribonuclease P RNA. *Science*, **254**, 853–856.
46. Pace,N.R., Smith,D.K., Olsen,G.J. and James,B.D. (1989) Phylogenetic comparative analysis and the secondary structure of ribonuclease P RNA-a review. *Gene*, **82**, 65–75.
47. Westhof,E. and Altman,S. (1994) Three-dimensional working model of M1 RNA, the catalytic RNA subunit of ribonuclease P from Escherichia coli. *Proc. Natl. Acad. Sci. USA*, **91**, 5133–5137.
48. Cho,M., Yoon,J.H., Kim,S.B. and Park,Y.H. (1998) Application of the ribonuclease P (RNase P) RNA gene sequence for phylogenetic analysis of the genus Saccharomonospora. *Int. J. Syst. Bacteriol.*, **48 Pt 4**, 1223–1230.
49. Schon,A., Fingerhut,C. and Hess,W.R. (2002) Conserved and variable domains within divergent rnase P RNA gene sequences of Prochlorococcus strains. *Int. J. Syst. Evol. Microbiol.*, **52**, 1383–1389.
50. Birkenheuer,A.J., Breitschwerdt,E.B., Alleman,A.R. and Pitulle,C. (2002) Differentiation of Haemobartonella canis and Mycoplasma haemofelis on the basis of comparative analysis of gene sequences. *Am. J. Vet. Res.*, **63**, 1385–1388.