bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Evolutionary insights on the structure and function of archaeal C/D box sRNAs as revealed by a comprehensive set from six *Pyrobaculum* species

Lauren M Lui[1,§], Andrew V. Uzilov[1], David L. Bernick[1], Andrea Corredor[1], Todd M. Lowe[1*], Patrick P. Dennis[2,§]

1 Department of Biomolecular Engineering, University of California, Santa Cruz, Santa Cruz, CA, 95064, USA

2 Whitman College, Walla Walla, WA, 99362, USA

§ Contributed equally to this manuscript

*To whom correspondence should be addressed. Tel: 1 831 459 1511; Email: lowe@soe.ucsc.edu

Email Addresses:

Lauren Lui:	lmlui@lbl.gov
Andrew Uzilov:	andrew.uzilov@gmail.com
David Bernick:	dbernick@soe.ucsc.edu
Andrea Corredor:	andrea.v.corredor@gmail.com
Todd Lowe:	lowe@soe.ucsc.edu
Patrick P Dennis:	dennispp@whitman.edu

Present Address:

Lauren Lui, Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, 94702, USA
Andrew Uzilov, Department of Genetics and Genomic Sciences, Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York NY 10029, USA

Keywords

C/D box sRNA, archaea, RNA-seq, comparative genomics, computational search model

ABSTRACT

Archaea use C/D box sno-like RNAs (sRNAs) to guide precise 2′-*O*-methyl modification of ribosomal and transfer RNAs. Although C/D box sRNAs are the most numerous archaeal small RNA class, most genomes have incomplete sRNA gene annotation because reliable, completely automated detection methods are unavailable. To study archaeal C/D box sRNA structure, function and evolution, we collected, predicted, and curated a comprehensive set of these sRNAs from six species within the crenarchaeal hyperthermophilic genus *Pyrobaculum*. We used high-throughput small RNA sequencing data, computational methods, and comparative genomics to curate 526 *Pyrobaculum* C/D box sRNAs and organized them into 110 families based on conservation of their guide sequences. Our comprehensive analysis provides a detailed narrative of C/D box sRNA evolutionary history, implications of target conservation on ribosome maturation, and sRNA impact on genome architecture. We find that in some cases the overlap of C/D box sRNAs with protein-encoding genes can abolish the ability of their guides to

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

target rRNA and tRNAs. Numerous duplications and rearrangements of sRNA genes are illustrated, and some of these appear to employ a mechanism similar to the retrotransposition of C/D box sRNAs genes in eukaryotes. Finally, we used these annotations to train a new computational search model.

INTRODUCTION

In eukaryotic cells, ribosome assembly occurs in the nucleolus, a specialized structure located within the nucleus. At this site, ribosomal RNA (rRNA) is transcribed, modified, processed, folded, and assembled along with ribosomal proteins into the large and small ribosomal subunits. The nucleolus also contains a large number of small (sno)RNAs that are implicated in the modification, folding, and maturation of rRNA (reviewed in (1)). These snoRNAs are incorporated into dynamic ribonucleoprotein (RNP) complexes that act as molecular machines along the ribosomal assembly line. Most snoRNAs contain guide sequences that base pair with rRNA, facilitating precise modification of ribonucleotides within the region of complementarity. The snoRNAs divide into two classes: C/D box snoRNAs, which guide 2′-*O*-methylation of ribose and H/ACA box snoRNAs, which guide the conversion of uridine to pseudouridine (2, 3). Although archaeal cells do not contain an organized nucleolar structure, they possess and utilize both C/D box and H/ACA box sno-like RNAs (sRNAs) in the modification of rRNA and assembly of ribosomal subunits (reviewed in (1, 4)).

Archaeal C/D box sRNAs are generally about 50 nucleotides (nts) in length and contain highly conserved C (RUGAUGA consensus) and D (CUGA consensus) box sequences at the 5′ and 3′ end of the molecule and less conserved versions (designated C′ and D′) near the center of the molecule (5). These RNAs fold into a balloon-like hairpin as a result of the formation of a kink-turn (K-turn) structural motif through the interaction of the C and D box sequences and a K-loop motif through the interaction of the D′ and C′ box sequences (Figure 1A). The K-turn and the K-loop are each recognized by the protein L7Ae (6). The binding of L7Ae stabilizes the RNA structure and allows two copies each of Nop56/58 (also called Nop5) and fibrillarin to bind, completing the assembly of the active ribonucleoprotein (RNP) complex (7). The fibrillarin protein is a *S*-adenosyl methonine dependent RNA methylase and is responsible for the catalytic activity of the RNP complex.

The two guide regions between the C and D′ boxes and between the C′ and D boxes are unstructured and each is available to base pair with an approximately 8–12nt long target sequence (Figure 1A). In addition to rRNA targets, a significant proportion of archaeal sRNAs have guide regions that are complementary to transfer RNA (tRNA) (7, 8). Methyl modification in the target RNA occurs at the nucleotide position that base pairs with the guide five nucleotides upstream from the first base of the D′ or D box sequence. This is known as the “N+five” rule and methylation targets are referred to as the D and D′ targets (2). Many C/D box sRNAs with canonical box features have guides that lack complementarity to rRNA and tRNA sequences; these are known as “orphan” guides and may target other RNAs but to date no conserved targets to alternative RNAs have been identified within the Archaea (1).

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The evolution of C/D box sRNAs affects ribosome function and the host genome. In all domains of life, ribose methylations help stabilize RNA structure and are most often found in functionally important regions of the ribosome, such as the peptidyl transferase center in domain V of the large ribosomal subunit (9). Although elimination of individual 2′-*O*-methylations by C/D box sRNA deletions appear to have little effect on the cell, global dysregulation of the methyltransferase fibrillarin has profound effects, possibly including cancer in humans (10, 11). In addition to their role in ribose methylation, the propagation of C/D box sRNAs genes may have a profound impact on the evolution and architecture of the genome. In mammals and nematodes, some C/D box sRNAs appear to duplicate via a retrotransposon-like mechanism (12–14). These duplications may lead to new functions of the sRNAs and like other transposons, impact genome evolution (15).

Although other studies have detected C/D box sRNA gene duplication and C/D box sRNA genes that overlap with protein-encoding genes (16–18), none have been comprehensive or carried out with the intent of understanding sRNA evolution and function within a genus. To understand better archaeal C/D box sRNA evolution and function, we have compiled the first complete or nearly complete set of 526 C/D box sRNAs present in six species of the hyperthermophilic genus *Pyrobaculum* using small RNA sequencing (RNAseq) data, comparative genomics, and improved computational detection. As a result, our set of C/D box sRNAs is more comprehensive than early studies in the *Sulfolobus* and *Pyrococcus* genera (16, 19, 20).

To detect the C/D box sRNA genes, we used small RNAseq data from five *Pyrobaculum* species for which genome sequences are available (18, 21): *P. aerophilum* (*Pae*), *P. arsentaticum* (*Par*), *P. calidifontis* (*Pca*), *P. islandicum* (*Pis*), and *P. oguniense* (*Pog*). In addition, the species *P. neutrophilus* comb. nov. (*Pne*; formerly *Thermoproteus neutrophilum* (22)) was used to supplement comparative genomics analyses. After identifying the *Pyrobaculum* C/D box sRNA set, we catalogued the sRNAs carefully into homologous families based on sequence similarity and methylation target prediction. In addition to this comprehensive annotation, we hand-curated computational predictions of the box sequences and the methylation targets to rRNAs and tRNAs in order to assist our analysis and study orphan guides. We used this extensive data set to examine (i) the variation in guide sequence within homologous sRNA families, (ii) the function of these sRNAs in modification of rRNA, and its assembly of ribosomes, (iii) the genomic context of sRNA genes, and (iv) the proliferation, mobility, plasticity and evolutionary divergence of sRNA genes as viewed within and between the six *Pyrobaculum* genomes. Finally, we created a new computational search model for archaeal C/D box sRNAs using these annotations.

MATERIALS AND METHODS

Computational prediction and organization of C/D box sRNA homolog families

To generate a complete or nearly complete set of sRNA gene predictions within the genus *Pyrobaculum*, we used computational covariance models and small RNA sequencing data of *Pae*, *Par*, *Pca*, *Pis*, and

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Pog (data reported in (18) except for *Pog*). The covariance model was created by using a hand-curated multiple structural alignment of the 65 *Pae* C/D box RNAs reported in the genome sequencing paper (23) as input to `cmbuild` from the Infernal v1.0 and v1.1 software packages (24, 25) with the hand-curated option `--rf` or `--hand` specified, respectively. The covariance model was calibrated with `cmcalibrate`. A final covariance model was built from a complete set of *Pae* sRNAs found from examining sequencing data and using comparative genomics with the other *Pyrobaculum* species (Supplementary Figure S1). RNA structure formed by the box features (Figure 1A), including the two G:A pairs of the K-turn and K-loop, was annotated in the alignment. The variable regions of archaeal C/D box sRNAs, the guides and variable loop, were annotated to be any base. We used `cmsearch` to scan the genomes and small RNA sequencing data using the glocal (`-g`) and no HMM filter (`--nohmm`) options. To improve specificity, candidates from genome scans that overlapped other annotated non-coding RNAs or overlapped Genbank RefSeq genes by more than 80% were discarded. One exception is if sRNAs were found by comparative genomics, they were still included in the final set even if they overlapped Genbank RefSeq genes.

To find additional orthologs of sRNAs genes within the *Pyrobaculum* genomes not found by the covariance model, the genomes were searched using sRNA sequences as queries to BLASTN (26). Genome Genbank/INSDC numbers are AE009441.1 (*Pae*), CP000660.1 (*Par*), CP000561.1 (*Pca*), CP000504.1 (*Pis*), CP001014.1 (*Pne*), and CP003316.1 (*Pog*). Top hits were manually curated, based on predicted promoters, conservation, and sequencing evidence. Families of C/D box sRNA homologs were created based on sequence similarity of guide sequences and predicted target sites of modification. The first 65 families (1-65) were assigned numbers based on the previously reported *Pae* sRNA numbering (18). Additional families containing newly identified sRNAs were numbered 100 to 141.

Prediction of methylation targets

To identify the putative sites of 2′-*O*-methylation guided by *Pyrobaculum* C/D box sRNAs, we scanned mature rRNA and tRNA sequences for regions of complementarity to the D and D′ guides of the sRNAs. Mature rRNA sequences were obtained by a global alignment of the six *Pyrobaculum* rRNAs and removing introns based on conservation. Intron sites are indicated in Supplementary Figures S2 and S3. A uniform numbering system for sites of rRNA methylation was obtained by first constructing an alignment of the 16S and 23S rRNA sequences from the six species (Supplemental Figures S2 and S3). The predicted locations of modification were mapped on the alignment and assigned a position based on the *Pae* numbering. For a prediction to be considered credible, generally a minimum complementarity of nine continuous Watson-Crick base pairs centering at or near the “N+five” position was required. The criteria were relaxed in two specific instances. First, if the majority of members in an sRNA group met the prediction criteria, the prediction was extended to minority members that nearly met the criteria (for example, matches containing a mismatch or G:U base pair). Second, it has been noted that many sRNAs use their two guide regions to direct methylations to closely spaced nucleotides within the target RNA. Presumably this enhances target identification and creates greater stabilization of the guide target

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

interaction within the RNP complex. Consequently, when one guide exhibits strong complementarity to the target, the criterion for the second guide match is relaxed if (i) it is within 100 nts of the first complementarity, (ii) the weaker complementarity contains no more than one mismatch (iii) and the combined bit score for the two complementarities was 32 or higher (where a Watson-Crick base pair is 2, a G:U base pair is 1 and a mismatch is -2). Automation of this method is done by *findAntisense.py*. Description of the program and related files can be found at https://github.com/lmlui/findAntisense.

Northern Blot of polycistronic sRNAs

Northern blots were prepared as described in (18). The following DNA oligomers (Integrated DNA Technologies, Inc., Coralville, IA) were used as probes:
Pae sR21 sense (GCCAGTGTCCGAAAATTGACGAGCTCACCCTTTGC)

Pae sR21 antisense (GCAAAGGGTGAGCTCGTCAATTTTCGGACACTGGC).

Small RNA sequencing and read processing of *Pyrobaculum* species

Small RNA sequencing data for *Pae*, *Par*, *Pca*, and *Pis* is from a previous study from our lab (18). The libraries for these four species were sequenced on the Roche/454 GS FLX sequencer. Small RNA libraries for *Pog* were sequenced by the UC Davis Sequencing Facility on Illumina HiSeq 2000 to produce 2x75 nt paired-end sequencing reads. Sample preparation of small RNA libraries for *Pog* are described in (18, 27). Briefly, the small RNA size fraction was isolated by running total RNA in denaturing gel electrophoresis and extracting the region below tRNAs.

Reads with barcodes and linkers removed were mapped to genomes using BLAT(28). The resulting PSL file was processed to determine paired reads.

RESULTS

Most Pyrobaculum C/D box sRNAs homologous families have members in all six species

We identified 526 C/D box sRNA genes from six species of *Pyrobaculum* using evidence from (i) RNAseq data from *Pae*, *Par*, *Pca*, *Pis* and *Pog*, (ii) an improved computational covariance prediction model, and (iii) comparative genomics. Nearly all of the sRNAs from the five genomes with RNAseq data (436/442, 99%) are represented in the RNAseq libraries and have strong phylogenetic conservation within the *Pyrobaculum*. The new covariance model incorporates the K-turn and K-loop structural information. This computational approach also does not rely on using target detection to filter out false positives, which allowed us to detect sRNAs where both guides lack complementarity to rRNA or tRNA sequences (1, 29). The model predicts the sRNA box features; these were manually checked and adjusted when required.

The 526 sRNAs were organized into 110 different homologous families based on sequence conservation of their guide regions and predicted targets of methylation in tRNA and rRNA. The sequences, family organization, and genomic location of these sRNAs can be found at the Lowe Lab Archaeal snoRNA-like C/D box RNA Database (http://lowelab.ucsc.edu/snoRNAdb/) and (30). UCSC

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Archaeal Genome Browser tracks enabling the visualization of the small RNA sequencing data can also be found on the website. We found 26 additional sRNAs in this study (two in *Pae*, three in *Par*, four in *Pca*, three in *Pis*, four in *Pog*, and ten in *Pne*); the number of detected C/D box sRNA genes in individual species ranges between 84 and 92 (Figure 1B). Grouping the sRNAs into families allowed us to study more easily the evolutionary origins and relationships of C/D box sRNAs genes within the genus and to predict more accurately target sites of methylation within rRNA and tRNA.

Most of the homologous families are conserved, with 70 of the 110 families (64%) having representative sRNAs encoded in each of the six *Pyrobaculum* genomes. The remaining families (40) have representatives missing from one or more of the six genomes (Figure 1C). Eighteen of the families are unique with the representative present in only a single species. Each of these 18 sRNAs have small RNA sequencing reads and 15 have at least one predicted target to rRNA or tRNA. Within a family it is common for both guide regions to exhibit a high degree of sequence similarity indicative of a common ancestry. For example, of the 70 families that have a representative sRNAs form each of the six species, 62 exhibit a recognizable degree of sequence similarity (>70%) in both the D and the D′ guide regions among all members whereas the remaining eight families have a conserved sequence across all species in only one of the two guide regions (see alignment table at http://lowelab.ucsc.edu/snoRNAdb/ for numerous examples). Even when a particular guide region is conserved, it is frequently punctuated by nucleotide insertions or deletions or by nucleotide substitutions primarily at the 5′ or 3′ end of the guide that are less likely to impact the guide-target interaction. Even with the high degree of guide sequence similarity, not a single guide is perfectly conserved in any of the 70 families with representatives in all six species.

Mapping of predicted rRNA and tRNA methylation targets

To predict methylation targets in rRNA and tRNA, we used the “N+five” rule (2) (Figure 1A) and ranked hits based on extended complementarity between guide sequences and target RNAs (Supplementary Tables S1 and S2). Using the criteria described in the Methods Section we were able to predict targets for nearly 75% (767/1052) of the sRNA guides; 89% (468/526) of the sRNAs had predicted targets for one or both of the guides. Although 32 of the 110 sRNA families have targets in tRNAs, only about 17% (178/1052) of guides have tRNA targets; in many of these families, one guide targets a tRNA and the other guide has no predicted target.

Targets in ribosomal RNA. We mapped the positions of predicted methyl modification on the 16S and 23S rRNA secondary structure in order to visualize clustering patterns. We find that approximately 45% of sRNAs (235/526) use their D and D′ guides to target sites that are within 100 nts of each other in the primary rRNA sequence (Supplementary Figures S2-5, Supplementary Tables S1 and S2). We have suggested previously that this dual interaction at two closely positioned sites plays an important role in mediating the folding and stabilization of the nascent rRNAs and their assembly onto ribosomal subunits (17, 31). A computational study simulating C/D box sRNA chaperone function in rRNA folding also

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

suggests that double guide sRNAs may be especially important for proper long-range interactions in rRNAs (26). In addition to these dual guides, we also find that, in general, methylation sites cluster within functionally important regions, such as the peptidyl transferase center and helix 69 of 23S rRNA, whereas less important regions contain a lower density of modifications. Comparisons with positions of predicted modification in species outside of *Pyrobaculum* indicate that the precise sites of modification are, with a few notable exceptions, generally not conserved although the clustering pattern is conserved (17).

One-third of* Pyrobaculum *C/D box sRNAs target tRNAs. It has been shown previously that archaeal C/D box RNA can target modification to tRNAs as well as rRNAs (23). The tRNA methylation targets are at structurally conserved positions that are modified by tRNA methylases in other organisms. In our collection of 110 *Pyrobaculum* sRNA families, 32 are predicted to target modification to 23 different positions in tRNAs (Supplementary Tables S1 and S2).

The number of different tRNAs that can be targeted by a particular sRNA guide varies over a wide range and reflects the fact that some sequences in tRNAs are unique whereas others are shared among many different tRNA isoacceptors. The region surrounding position 34, the wobble base in the anticodon, is an example of a variable sequence. Guides from four different sRNAs target position C34 or U34, and each has only a single tRNA target (sR26:C34Trp, sR27:U34Gln, sR45:C34Val, sR46:U34Thr and sR51:C24Glu). Other guides have multiple tRNA targets; for example, the D guide of Pae sR64 exhibits complementarity to a conserved sequence in sixteen different tRNA families and directs modification to position G51 in the TΨC stem.

Long range interactions between sRNAs and rRNAs support the role of sRNAs as rRNA folding chaperones. Several sRNAs have D and D′ guides that have complementarities and predicted methylation targets that are more than 100 nts apart in the primary rRNA sequence but are close in the secondary structure. We suspect that these long-range interactions play an important role in the tertiary folding of rRNA during the assembly process. We find three instances of these interactions that are conserved among the six species in this study.

In 16S rRNA, the D guide of sR53 is complementary to the loop region of helix 18 and is predicted to methylate A509 (Supplementary Figure S6A). This entire stem loop 18 has been implicated in translational fidelity. The G507 (G530 in E coli 16S rRNA) is intimately associated with the interaction between the A site tRNA anticodon and the mRNA codon; site directed mutations at this position are lethal (27). Other mutations in this region affect translational fidelity and resistance to the antibiotic streptomycin. The D′ guide of sR53 has two separate complementarities to 16S sequences and is predicted to methylate at position C514 in the 3′ strand of helix 3 as well as at position C34 in the 5′ strand of helix 4. There are eight additional predicted sites of methyl modification in this region that are mediated by other *Pyrobaculum* sRNA families. It is unclear how these multiple sR53 guide interactions might occur within

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the nascent rRNA transcript and how they impact the folding and structural stability of the translational fidelity stem-loop.

Two other three-way interactions have been identified around the helix 26-27 junction in 23S rRNA (Supplementary Figure S6B) and in the core region around helix 28 that serves as the connection point for the four domains in 16S rRNA (Supplementary Figure S6C). In the first instance sR2 is predicted to use its D guide to modify position G655 and its D′ guide to modify both positions C667 and C781 in 23S rRNA. The second instance involving the core region in 16S includes helix 2, a complex pseudoknot that forms between the loop of helix 1 and the connector region between domain 2, and the core helix 28. The predicted modification of the sR56 D guide is at position U877, which is immediately 5′ to the helix 2 pseudoknot structure. The D′ guide is predicted to modify both positions G908 and G1337 in 16S rRNA. These interactions likely facilitate the complex folding events that arrange the four 16S domains around the central core helix 28.

Instances of mismatched base pairs at the "N+five" position. In several instances (two positions U109 and C1368 in 16S and five positions A608, G764, U912, C2045 and C2117 in 23S) we find a mismatch at the "N+five" position in the region of guide-target complementarity (Supplementary Tables S1 and S2). *In vivo* and *in vitro* studies have demonstrated that a Watson-Crick base pair at this position is essential for methylation of the target RNA (7, 28, 29). Nonetheless, a mismatch at the site of methylation within a conserved region of guide-target complementarity implies that the interaction may be beneficial but that the modification is either not needed or harmful to the function of the target RNA. Studies in yeast that use cross-linking to detect RNA-RNA interactions support this hypothesis (36, 37). Results from these studies indicate that *Saccharomyces cerevisiae* snoRNAs form interactions with rRNA that do not result in methylation and that these interactions may be involved in rRNA maturation.

We observe two types of mismatches at the "N+five" position in the context of sRNA families: (i) only one member has a mismatched base pair and (ii) the mismatch is conserved in multiple members. For example, the sR09 family has five members and the D and D′ guides are highly conserved. However, the *Par* D guide contains an A-to-U nucleotide substitution at the critical "N+five" position of the D guide, changing the guide-target interaction to U:U at this position (Figure 2A). The sR09 family has dual guides and we suspect that these guide-target interactions play an important role in the localized folding of the 23S rRNA (Figure 2C).

An example of a conserved mismatch occurs in the sR33 family. This family has members in all six *Pyrobaculum* species and the D and D′ guides target closely spaced position in 23S rRNA (Figure 2B,D). The D′ guide of all members is predicted to be incapable of methylation at C2045 located in helix 69 because of a C:U mismatch at the "N+five" position. The proper folding of helix 69 is critical because of its interaction with the anticodon stems of A site and P site tRNAs during protein synthesis (30). We consider the D′ guide-target interaction credible because of its strong conservation and its close proximity to the D guide interaction. In these instances of mismatch at the "N+five" position we predict that methylation does

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

not occur, but the sRNAs act as chaperones for productive and efficient folding of the rRNA during the ribosome assembly process (1, 4).

Guides with no predicted targets in tRNA or rRNA (orphan guides). In the 526 different *Pyrobaculum* sRNAs (representing 1052 guides) that we have identified, there are 285 guides (27%) that show no significant complementarity to either rRNA or tRNA sequences (see Supplementary Tables S1 and S2). We searched for mRNA and other non-coding RNA targets for these orphan guides, but no significant and conserved complementarities were observed. In an analysis of orphan guides conserved in the six *Pyrobaculum* species, none of these guide families had conserved mRNA targets.

In some instances some orphan guides appear to be the result of nucleotide substitutions. For example, in the sR30 family, the D guide of all six members is predicted to target C2724 in 23S rRNA, whereas the D′ guide is predicted to target C2708 in only four of the members (Supplementary Table S1 and http://lowelab.ucsc.edu/snoRNAdb/). The two sRNAs containing the disrupted D′ guides occur in *Pog* and *Par*, a sub-lineage within the *Pyrobaculum* genus.

Guide divergence also appears to occur via genomic arrangements or sRNA duplication that result in an overlap between an sRNA gene and a protein-encoding gene. Of the sRNAs that overlap the 5′- or 3′-end of a protein-encoding gene in the sense orientation, 88% and 61% of the respective overlapping guides do not have predicted targets in rRNA or tRNA (see later sections for more discussion).

The origin of other orphan guides is less clear. There are a few instances where guide families are conserved but only one of the members has targets. For example, both *Pae* sR64 and *Pae* sR101 have tRNA targets, but the five other homologs in their families are dual orphan guides (Supplementary Tables S1 and S2). These guide families are relatively well conserved with only point mutations and it is unclear if these are instances of gain- or loss-of-targeting function.

There are only three families with six members (sR43, sR50 and sR108) where both guides are orphan guides. In the sR43 family both guides are highly conserved among members and would be expected to recognize the same target sequence whereas in the other two families, the guide sequences are only moderately conserved and would likely not all recognize the same target sequence. The lack of conservation for the sR50 family may be partially explained by overlap with the promoter of a nearby protein-encoding gene.

Proliferation, mobility, plasticity and evolutionary divergence of C/D box sRNA genes within the *Pyrobaculum* genus

Grouping the *Pyrobaculum* C/D box sRNAs into homologous families has greatly facilitated our understanding of target conservation and the origins and evolution of these sRNAs. Here we describe sequence similarity of guide sequences between different homologous sRNA families.

Composite and transposed sRNAs. Of the 18 sRNA single-member families, five have a guide that shares some resemblance to a guide in a different sRNA family. These are designated as either

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

transposed or composite sRNAs (Table 1). Transposed sRNAs share one guide with another defining family, but typically the guide has been transposed from D to D′ or visa versa, from D′ to D position, compared to the defining family. Composite sRNAs have D and D′ guides that each match one of the guides in two different families. For example, *Pca* sR12/45 has a D guide that is similar to the D guide of the sR12 family and a D′ guide that is similar to the D′ guide of the sR45 family (Figure 3C). We suggest that the genes encoding these composite and transposed sRNAs are generated by genomic rearrangements between different sRNAs or sRNA genes.

Duplication of sRNAs. Duplication of a full-length sRNA gene can also occur as evidenced by the highly similar *Pae* sR113a and 113b (Figure 3A). None of the other *Pyrobaculum* species in this study have members in this sRNA family. The 5′-flanking regions in front of the two genes are unrelated. In contrast, the 3′-flanking regions are identical for 14 bp with sequence similarity extending a further 30 bp. The sequence 3′ to the sR113a gene encodes sR08 on the opposite strand, while the sequence 3′ to the sR113b gene contains what appears to be the remnant of the sRNA gene that has been obliterated by the presence of ORF PAE3005. The sR113a and sR08 genes are convergently transcribed and separated by a 1 bp intergenic space.

Other examples of duplication are of *Pog* sR46 and *Pae* sR62. The sR46 family has members in all six species. *Pog* sR46a is a duplication of *Pog* sR46 and has a nearly identical D′ guide and a D guide with three nt substitutions. The other apparent duplication involves *Pae* sR62 (Figure 3B). This gene is located at position 2104084-2104133 on the chromosome. A highly similar sequence presumably representing an sR62 pseudogene remnant occurs at position 2101402-2101488.

Super families of sRNAs. The tracking of sequence similarity between guides from different families can reveal ancient origins and evolutionary relationships between the homologous sRNA gene families. We have uncovered evidence suggesting that the sR45, 12, 56 and 57 families share a complex evolutionary history (Figure 3C). The sR56 and sR57 families appear to represent an ancient duplication appearing early within the *Pyrobaculum* lineage. Only a homolog of sR56 appears in the closely related species, *Thermoproteus tenax* (*Tte*). Each family has representatives in all six *Pyrobaculum* species and one of the families (sR57) is a circular permutation of the other (sR56). The D and D′ guides of sR56 target modifications to 16S U877 and G908 respectively and the D and D′ guides of sR57 target modifications to 16S G906 and A879 respectively. The shared core sequence between the D guide of sR57 and the D′ guide of sR56 is UUCACC and the shared core sequence between the D guide of sR56 and the D′ guide of sR57 is UCCUUUA. These cores sequences are offset by two nucleotides due to indels within the respective guides and this accounts for the two nt shift in target specificities. The two aberrant (transposed) members of the sR57 family (*Pae* sR57a and *Pca* sR57b) are circular permutations of each other and share only the single guide UC-CC-CUU (dashes indicate indels) with the D guide of the core

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

sR57 family. Archaeal sRNAs are known to circularize (27, 39–41) and the sR57 may be an example of circularization and re-insertion into the genome.

The sR12 family is also implicated in this complex interconnection of families. It has a D′ guide that exhibits sequence similarity to the D′ guide of the sR56 family (CU-UC-CCUC). Indels in the sR12 D′ guide changes the target specificity to position 23S G1221. As mentioned above, the D guide in the sR12 family is shared with the D guide of the composite sR12/45. The second D′ guide of sR12/45 is derived from the sR45 family; this guide is predicted to target methylation to position C34 in the anticodon loop of $tRNA^{Val}$.

The relationships between these four related families illustrate several important aspects of sRNA gene evolutions including: (i) gene duplication; (ii) target migration (resulting from insertion/deletion) or divergence (resulting from nucleotide substitution) that alters or abolishes guide-target interactions; (iii) rearrangements, including guide replacement and/or circular permutation.

MITE-like elements resembling sRNAs. Many of the families with only one sRNA member occur in *Pca* (Figure 1C). This species exhibits modular duplications and rearrangements between and within sRNA families as evidenced by the transposed and composite sRNAs described above (see Table 1). A careful analysis has also revealed the presence of a MITE-like element present in at least 15 copies within the *Pca* genome (Figure 4, Supplementary Table S3). MITEs are miniature inverted-repeat transposon elements that are characterized by a combination of terminal inverted-repeats and internal sequences too short to encode proteins. These elements are Class II transposons that occur in plants and other archaea (35).

These elements in *Pca* have characteristics of both sRNAs and MITEs. Each copy contains C box and D box sequences and highly degenerate internal D′ and C′ sequences. The guides located between the C and D′ and C′ and D boxes exhibit only modest sequence similarity across the 15 copies. Highly conserved imperfect inverted-repeat sequences flank the C and D boxes (Figure 4). The elements have a large average distance (322bp) from the nearest protein-encoding gene compared to other sRNAs (22bp). The presence of these MITE-like elements in regions of the genome where there are no other genomic features suggest that they are located in regions of genomic instability, which may be hotspots for insertion by mobile elements or by these MITE-like elements themselves.

Five of the element copies were classified as C/D box sRNAs (sR131, sR133, sR137, sR139, sR141) and contain moderately degenerate internal box sequences. The genomic location of the other ten copies of this element that were not considered to be sRNAs are listed in Supplementary Table S3. For sR141, an amazing 13.6% of the uniquely mapped RNAseq reads in *Pca* are generated from this single locus. Of the other *Pca* sRNAs (non-MITE-like), the highest percentage of uniquely mapped reads to an sRNA was 4% and on average 0.5% of total unique reads mapped to each sRNA. The other copies have expression levels similar to other sRNAs. The MITE-like sRNAs also tend to have a higher percentage of antisense reads compared to other sRNAs. On average, 39% of reads from a MITE-like sRNA locus are

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

antisense, whereas on average 9.3% of reads from other *Pca* sRNAs are antisense. We suggest that this element may play a role in the generation, mobilization and proliferation of C/D box sRNAs or their modular components. We have not observed these MITE-like elements in the other five species although they may well exist in lower copy numbers and with less sequence conservation.

Association of C/D box sRNA and tRNA genes.

Most archaeal C/D box sRNAs are independently transcribed, but in a few cases C/D box sRNA genes are known to be polycistronic (1). Transcription of archaeal C/D box sRNAs genes with protein-encoding genes has been reported in *Sulfolobus* and the *Pyrococcus* genera (16, 19); in *Nanoarchaeum equitans*, a few instances of di-cistronic C/D box sRNA-tRNA transcripts have also been reported (36). In the *Pyrobaculum*, we find conserved instances of C/D box RNAs co-transcribed with other sRNAs, a tRNA, and protein-encoding genes.

Our analysis indicates that the transcriptional relationships between sRNA and tRNA genes within the *Pyrobaculum* genus are extremely fluid. We identified a novel archaeal transcript in *Pae*, *Pis*, and *Pca* that contains three C/D box sRNAs (sR101, sR21, and sR100). These three genes are polycistronic based on genomic proximity, northern hybridization, and overlapping RNA-Seq reads (Figure 5A and B). In *Pne*, *Par*, and *Pog*, there is no homolog of sR100, but sR21 and sR101 are still syntenic. In *Par* and *Pog* the two genes are approximately 180 nts from each other and appear to be expressed from separate promoters. In *Pis* and *Pne* the sR34 and sR40 genes are also co-transcribed (based on RNA-seq reads) and separated by 10, and -4 nts respectively. In *Pne* the D box of sR34 is located within the C box of sR40 (four nt overlap); it is unclear how this overlap affects the maturation of the two sRNAs. In *Par*, *Pog*, and *Pca* the genes are separated by 16, 16, and 78 nts respectively and are convergently transcribed whereas in *Pae* the two genes are separated by more than 2000 nts.

Plant species and the archaeon *Nanoarchaeum equitans* have C/D box sRNA genes that are reported to be co-transcribed with tRNAs (37). In the *Pyrobaculum* genus we find one case of a C/D box sRNA that is likely co-transcribed with elongator $tRNA^{Met}$. In *Pae*, *Pis*, and *Pne*, the sR44 gene is positioned 8 bp or less from the 3′-end of $tRNA^{Met}$ gene (Figure 5C). In *Par* and *Pog*, the sR44 and $tRNA^{Met}$ genes share the same synteny, but the genes are separated by about 100 nts and their expression appears to be driven from separate promoters. In *Pca*, the sR44 gene is approximately 13 Kbps downstream of the $tRNA^{Met}$ gene. In *Tte*, there is no homolog of sR44 and none of its orthologs of $tRNA^{Met}$ are linked to C/D box sRNA genes.

These two examples demonstrate fluidity of C/D box sRNAs genes within the *Pyrobaculum* genus. None of the four sRNAs discussed (sR21, sR100, sR101, and sR44) have homologs in *Tte* (Figure 5). Within the polycistronic example the sR100 was lost from the transcription unit in the *Par*/*Pog*/*Pne* lineage and in *Pog* and *Par* the remaining sR100 and sR101 genes developed individual promoters. Similarly, the sR44 gene appears to have become linked to the $tRNA^{Met}$ gene in the ancestor of *Pae*, *Pis*, *Pne*, *Pog*,

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

and *Par* lineage; *Pca* is an out group to these species and does not have the same sRNA-tRNA linkage (Figure 5C). Separate promoters for the two genes occur in the *Pog*/*Par* sub-lineage.

Impact of overlapping of C/D box sRNA genes and protein-encoding genes

In our previous study (18), we noted that *Pyrobaculum* C/D box sRNAs genes are over 40-fold more likely than tRNA genes to have conserved overlap with orthologous protein-encoding genes. Other studies have also noted the 3′-antisense overlap of C/D box sRNAs with protein-encoding genes (16) We looked more closely at this relationship since overlap could impact the function of both gene types. In addition, antisense interactions suggest the possibility that C/D box sRNAs might guide modification of mRNAs or be involved in antisense regulation.

In our set of 526 *Pyrobaculum* C/D box sRNAs, 97 exhibit either partial or complete overlap with protein-encoding genes (Figure 6 and Supplementary Table S4). For this analysis, we considered only overlaps that extend either into the D′ guide region (eight nts or more beyond the 5′-end of the sRNA gene) or into the D guide region (five nts or more beyond the 3′-end of the sRNA gene) since shorter overlaps ending in the D box or C box were not expected to impact target specificity. We note however that there were 23 instances of where sRNA C box of the sRNA gene overlaps the 3' end of a protein encoding ORF in the sense orientation and provides the in frame translation termination codons for the ORF (C box RUGAUGA). We classified the more extensive overlaps into five categories (Figure 6A-E and Supplementary Table S4). Instances of overlap with the 5′-end of an mRNA were checked manually to confirm that the start codon of the mRNA was called correctly; start codons were adjusted based on protein sequence conservation.

The first and largest category involves sRNA genes that overlap a protein-encoding gene and was divided into three subcategories: (i) overlap at the 5′-end of the protein ORF in the sense orientation; (ii) overlap at the 3′-end of the protein ORF in the sense orientation and (iii) overlap at the 3′-end of the protein ORF in the antisense orientation (Figure 6A). There were no C/D box sRNA genes that overlapped the 5′-end of a protein-encoding gene in the antisense orientation. In the first subcategory only two of the 17 overlapping guides (12%) were predicted to have methylation targets in rRNA or tRNA. We suspect that in many of these instances, the sRNAs are co-transcribed with the mRNA based on promoter analysis. The translation initiation codons for the respective ORFs are located either in the C′ box or in the D guide region of the sRNA sequence. A recent study by Tripp *et al*. has reached similar conclusion based on an analysis of 300 sRNAs from six divergent species of archaea (44).

The second and third subcategories with overlapping guides in the sRNAs at the 3′-end of the protein-encoding gene had in comparison numerous predicted targets (40 of 47 for antisense sRNAs guides and 7 of 18 for sense sRNA guides; see Supplementary Table S4). This disparity suggests that the translational initiation site or N-terminal amino acid sequence of the protein is important in protein structure function and that the 5′-end of a gene cannot easily be usurped for sRNA guide function. In

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

contrast the 3′-end of protein-encoding genes appears more flexible and accommodates in many instances both C-terminal amino acid sequence encoding and sRNA guide function.

The high proportion of sRNA located near or overlapping the 3′-end of protein-encoding genes may suggest that they play a role in gene regulation and possibly mRNA stability. Sense strand sRNAs that are co-transcribed with mRNA need to be excised and rescued from decaying mRNA transcripts. The sRNAs that are antisense could participate in antisense regulation through the formation of an RNA/RNA duplex or trigger methylation of the mRNA through a more limited guide target interaction. We also note in our RNAseq reads that many sRNA genes generate both sense strand and antisense strand transcripts. In other archaea, small antisense RNAs have been shown to regulate gene expression by binding to 3′-UTRs (reviewed in (45)). A role for these antisense sRNA transcripts has not been defined.

The second category represents sRNAs that are contained completely within protein-encoding genes (Figure 6B and E). Nine of the ten of these are in the antisense category and all have at least one guide that has a target in rRNA or tRNA. These internal sRNAs are located near the 3′-end of the protein-encoding gene, again suggesting that this region is flexible and can accommodate both amino acid coding and guide function without detriment.

The third category contains a single sRNA gene (*Pne* sR42) that is antisense and spans the four nt intergenic space between two sense strand and co-transcribed protein-encoding genes (Figure 6C). In other *Pyrobaculum* species there is a longer intergenic space and the sR42 members overlap only the 3′-end of the upstream NAD dependent deacetylase gene. In the final two categories, four members of the sR127 family are located at the convergence of two protein-encoding genes. In *Pne* and *Pis* the sR127 gene spans the intergenic space between the protein encoding genes (Figure 6D). In *Par* and *Pog* the Uridine phosphorylase gene contains a 3′-extension not found in *Pne* and *Pis* that extends through the entire sR127 gene and into the convergently transcribed hypothetical protein-encoding gene (Figure 6E).

In summary, overlap of an sRNA gene at the 5′ end of an ORF in the sense orientation is detrimental to the targeting function of the overlapping guide, but overlap on the 3′ end of an ORF in either sense or antisense orientation is less so. Some families such as sR05, sR118, and sR3 have conserved overlap (Supplementary Table S4). However, there are many instances where only a subset of the sRNAs in a family have conserved overlap, indicating that the position of sRNAs in relation to ORFs is fluid. Some orphan guides may be a result of loss-of-function by overlap with an ORF, rather than the result of developing targets other than tRNA or rRNA.

New computational model based on curated *Pyrobaculum* C/D box sRNAs

One of the most important aspects of our detailed curation of 526 archaeal sRNAs was the ability to generate a gold-standard set for training a computational model. After training the model and obtaining score distributions of true positives and random sequences (false positives), we determined a high confidence threshold of 17 bits, and a moderate confidence of 13 bits for archaeal C/D box sRNA predictions (Supplementary Figure S7).

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

In our previous study where we used small RNAseq data from four species of *Pyrobaculum*, we identified several unannotated transcripts that were likely to be conserved C/D box sRNAs with box features (specifically the K-turn motif formed by box base pairing) divergent from the canonical C/D box model (18). Therefore, we developed a more useful covariance model that accommodates orphan guides and incorporates box sequences, K-turn and K-loop structure, and length of spacers (guides and variable loop) (Figure 1A). The length of the spacers in the model is based on the longest observed length in the training set. The final model was trained on a structural alignment of all *Pae* sRNAs and created with the Infernal v1.1 software package (25). We are confident that the training set does not contain false members since all the member C/D box RNAs are either conserved within the *Pyrobaculum* genus or have confirming small RNA sequencing data.

Analysis of predictions generated from this model indicates (i) that removing predictions that have >80% overlap with mRNAs and overlap with known non-coding RNAs greatly reduces the number of false positives and (ii) that across different species, thresholds can be used with our model to evaluate *de novo* predictions (Supplementary Figure S7). We used Infernal v1.0 (24) to pick up between 0-5 more predictions per species since it is slightly more sensitive than Infernal v1.1 (25) (Supplementary Table S5). It is possible to increase the sensitivity of Infernal v1.1 using the `--max` option, but the number of candidates to evaluate manually becomes prohibitive. Infernal 1.0 and 1.1 produce different subsets of candidates; we used both for the final prediction set and manually curated the sRNAs that were predicted by only one or the other.

To test how well this model works on divergent archaea, we used it to search three species in the euryarchaeal genus *Pyrococcus*: *P. abyssi* (Pab), *P.furiosus* (Pfu), and *P. horikoshii* (Pho). The genus contains many sRNA gene predictions and has been a model for studing C/D box sRNA structure and function (19, 20). We were surprised to find eight new C/D box sRNA among these species (two in *Pab*, six in *Pfu*, and one in *Pho*, Supplementary Table S5). These predictions are conserved with other *Pyrococcus* sRNAs or have small RNA sequencing evidence (Supplementary Figure S8). The model has better specificity in the *Pyrococcus* than in the *Pyrobaculum* (Supplementary Figure S7). The reason for the better specificity is that the *Pyrococcus* C/D box sRNAs have much less variation in their box sequences and more canonical K-turns compared to the *Pyrobaculum* (Supplementary Figure S9). The larger variation in the *Pyrobaculum* may make this model efficacious for scanning other Archaea.

Approximately 3-11% of known C/D box sRNAs in the *Pyrobaculum* and *Pyrococcus* are not predicted with this covariance model. Even with a bit score threshold as low as -50, these sRNAs are not predicted. We examined the false negatives and found that in most cases one or more box features were unusual, often resulting in non-canonical K-motifs. For example, the sR42 family has members in all six *Pyrobaculum* species, but *Pae* sR42 has an unusual D′ (GCAA) and C′ (AUGGCGU) box motifs. Not only do these box sequences diverge from the consensus (CUGA for D box and RUGAUGA for C box), they do not create kink-turns with the canonical GA/AG base pairs. To capture these unusual C/D box sRNAs, another model may be needed.

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

DISCUSSION

We used RNAseq data, computational methods and comparative genomics to identify a likely comprehensive set of 526 C/D box sRNAs from six species within the genus *Pyrobaculum*. We organized these sRNAs into 110 homologous families based on sequence similarity of their D and D′ guides and mapped their predicted methylation sites in 16S and 23S rRNA. With this set of families and predicted targets, we were able to explore known and hypothetical functions of C/D box sRNAs, study their impact on the genomic organization and architecture, and visualize many aspects of their evolutionary origins and diversification.

The combination of the *Pyrobaculum* C/D box sRNA catalogue and our extensive map of the corresponding methylation sites provides evolutionary perspective on the canonical functions of archaeal C/D box sRNAs. Our analysis indicates that slightly less than two-thirds of the predicted targets are conserved among the six species. This is much higher than in a panarchaeal study of archaea where only one target was conserved among all seven of the species (each from different orders) (17), but even at the genus level it appears that methylation sites are only moderately conserved. In addition, there are several instances of conserved *Pyrobaculum* sRNA families where members can have slightly different targets because of insertions or deletions within one of their guide. For example, the D′ guide of sR127 targets methylation to 23S position 2609 in *Pca*, to 2610 in *Pis* and *Pne*, and 2612 in *Pae*, *Par* and *Pog* (see Supplementary Figure S3). In these and other instances the region of interaction within rRNA is conserved (and likely to be important at least within the genus) but the particular site of methylation is not.

There are intriguing instances of dual guide target long range interactions, but these are also rarely conserved past the genus level. We imagine that these long range interactions help arrange localized regions into the more complex tertiary 16S or 23S rRNA structures. The variation in methylation sites is reflective of the sequence diversity of the guides and reinforces the hypothesis that the aggregate of methylations in certain regions of the rRNA is generally more important than particular sites of modification (9).

The genomes of hyperthermophilic archaea contain large numbers of C/D box sRNAs genes that are presumably necessary to assist in the assembly and function of ribosomes at high temperatures. Where the genes for these sRNAs originate, how they were propagated within thermophilic genomes and how their guide functions were tuned to meet the needs of ribosome assembly and function has until now remained unclear. We have identified within the *Pyrobaculum* genus, instances of C/D box sRNA gene expansion and diversification resulting from a number of processes: (i) gene duplications, (ii) gene rearrangements, (iii) guide replacement and guide translocations, (iv) transposon-like mechanisms, and (v) guide divergence caused by nucleotide substitutions, and insertions and deletions.

The extensive overlap of C/D box sRNAs genes with protein-encoding genes raises new questions about their sequence constraints, excision from mRNA transcripts and role in the regulation of mRNA stability and translation. In numerous instances the sequence of sRNA guides is usurped by the coding constraints of the mRNA sequence, particularly in the region at the 5’ end of the gene encoding the N-

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

terminal amino acid sequence of the protein. Sense strand sRNAs are interesting for two reasons. First, their maturation requires precise excision from the mRNA transcript. We suggest that if translation is restricted, the sRNA partially or completely assembles into a complex with L7Ae, and possible Nop 56/58 and fibrillarin within the mRNA. The RNP complex likely protects the sRNA sequence from the nucleases that degrade the unprotected parts of the mRNA transcript and allows the precise excision of the sRNA from the mRNA transcript. Tripp *et al.* have recently made a similar suggestion and using artificial constructions have provided some experimental evidence to support this idea (38). Second, the C and D box sequences within the mRNA have the potential to form a K-turn structural motif in the presence of available L7Ae protein. This complex is known to auto regulate mRNA expression in both natural and synthetic constructions (40–42). Antisense sRNAs may function as antisense regulators or may use their guide sequences to carry out site-specific methylation of the mRNA and influence the structure, function or stability of the mRNA.

This comprehensive effort to identify the complete set of C/D box sRNA genes from six species within the hyperthermophilic genus *Pyrobaculum* has provided unique and valuable insights into (i) their structure and function, (ii) their role in ribosome subunit biogenesis, (iii) their evolutionary origin, propagation and divergence and (iv) their role in influencing protein gene expression and shaping overall genome architecture.

FUNDING

This work was supported by the National Science Foundation [EF-082277055]; US National Institutes of Health (NIH) bioinformatics training grant [1 T32 GM070386-01 to A.U. and L.L.]; University of California Bio-technology Research and Education Program [Graduate Research and Education in Adaptive Bio-Technology (GREAT) Training Program to D.B.]; Google [Anita Borg Scholarship to L.L.]; and Achievement Rewards for College Scientists Foundation [scholarship to L.L.]. Funding for open access charge: National Institutes of Health.

REFERENCES

1. Lui,L. and Lowe,T. (2013) Small nucleolar RNAs and RNA-guided post-transcriptional modification. *Essays Biochem.*, **54**, 53–77.
2. Kiss-László,Z., Henry,Y., Bachellerie,J.-P., Caizergues-Ferrer,M. and Kiss,T. (1996) Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell*, **85**, 1077–88.
3. Ganot,P., Bortolin,M.-L. and Kiss,T. (1997) Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell*, **89**, 799–809.
4. Watkins,N.J. and Bohnsack,M.T. (2012) The box C/D and H/ACA snoRNPs: key players in the modification, processing and the dynamic folding of ribosomal RNA. *Wiley Interdiscip. Rev. RNA*, **3**, 397–414.
5. Balakin,A.G., Smith,L. and Fournier,M.J. (1996) The RNA world of the nucleolus: two major families of small RNAs defined by different box elements with related functions. *Cell*, **86**, 823–34.
6. Nolivos,S., Carpousis,A.J. and Clouet-d'Orval,B. (2005) The K-loop, a general feature of the *Pyrococcus* C/D guide RNAs, is an RNA structural motif related to the K-turn. *Nucleic Acids Res.*, **33**, 6507–14.

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

7. Omer,A.D., Ziesche,S., Ebhardt,H. and Dennis,P.P. (2002) *In vitro* reconstitution and activity of a C/D box methylation guide ribonucleoprotein complex. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 5289.

8. Clouet d'Orval,B., Bortolin,M.-L., Gaspin,C. and Bachellerie,J.-P. (2001) Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNATrp intron guides the formation of two ribose-methylated nucleosides in the mature tRNATrp. *Nucleic Acids Res.*, **29**, 4518–29.

9. Decatur,W.A. and Fournier,M.J. (2002) rRNA modifications and ribosome function. *Trends Biochem. Sci.*, **27**, 344–51.

10. Tollervey,D., Lehtonen,H., Jansen,R., Kern,H. and Hurt,E.C. (1993) Temperature-sensitive mutations demonstrate roles for yeast fibrillarin in pre-rRNA processing, pre-rRNA methylation, and ribosome assembly. *Cell*, **72**, 443–457.

11. Marcel,V., Ghayad,S., Belin,S., Therizols,G., Morel,A.P., Solano-Gonz??lez,E., Vendrell,J., Hacot,S., Mertani,H., Albaret,M., *et al.* (2013) P53 Acts as a Safeguard of Translational Control by Regulating Fibrillarin and rRNA Methylation in Cancer. *Cancer Cell*, **24**, 318–330.

12. Zemann,A., op de Bekke,A., Kiefmann,M., Brosius,J. and Schmitz,J. (2006) Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Res.*, **34**, 2676–2685.

13. Weber,M.J. (2006) Mammalian small nucleolar RNAs are mobile genetic elements. *PLoS Genet.*, **2**, e205.

14. Schmitz,J., Zemann,A., Churakov,G., Kuhl,H., Grützner,F., Reinhardt,R. and Brosius,J. (2008) Retroposed SNOfall--a mammalian-wide comparison of platypus snoRNAs. *Genome Res.*, **18**, 1005–10.

15. Deragon,J.M. and Capy,P. (2000) Impact of transposable elements on the human genome. *Ann. Med.*, **32**, 264–73.

16. Dennis,P.P., Omer,A. and Lowe,T. (2001) A guided tour: small RNA function in Archaea. *Mol. Microbiol.*, **40**, 509–519.

17. Dennis,P.P., Tripp,V., Lui,L., Lowe,T. and Randau,L. (2015) C/D box sRNA-guided 2′-O-methylation patterns of archaeal rRNA molecules. *BMC Genomics*, **16**, 632.

18. Bernick,D.L., Dennis,P.P., Lui,L.M. and Lowe,T.M. (2012) Diversity of Antisense and Other Non-Coding RNAs in Archaea Revealed by Comparative Small RNA Sequencing in Four *Pyrobaculum* Species. *Front. Microbiol.*, **3**, 231.

19. Gaspin,C., Cavaillé,J., Erauso,G. and Bachellerie,J.-P. (2000) Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the Pyrococcus genomes. *J. Mol. Biol.*, **297**, 895–906.

20. Omer,A.D., Lowe,T.M., Russell,G., Ebhardt,H., Eddy,S. and Dennis,P. (2000) Homologs of small nucleolar RNAs in archaea. *Science (80-.).*, **288**, 517–22.

21. Bernick,D.L., Dennis,P.P., Höchsmann,M. and Lowe,T.M. (2012) Discovery of Pyrobaculum small RNA families with atypical pseudouridine guide RNA features. *RNA*, **18**, 402–11.

22. Chan,P.P., Cozen,A.E. and Lowe,T.M. (2013) Reclassification of *Thermoproteus neutrophilus* Stetter and Zillig 1989 as *Pyrobaculum neutrophilum* comb. nov. based on phylogenetic analysis. *Int. J. Syst. Evol. Microbiol.*, **63**, 751–754.

23. Fitz-Gibbon,S.T., Ladner,H., Kim,U.-J., Stetter,K.O., Simon,M.I. and Miller,J.H. (2002) Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 984–9.

24. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–7.

25. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

26. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

27. Uzilov,A. V (2013) Novel applications of high-throughput RNA sequencing: mapping RNA structure and discovering circular RNAs. http://escholarship.org/uc/item/5284w609.

28. Kent,W.J. (2002) BLAT---The BLAST-Like Alignment Tool. *Genome Res.*, **12**, 656–664.

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

29. Lowe,T. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science (80-.).*, **283**, 1168–71.
30. Lui,L.M. (2015) Evolution of structure and function of Kink-turn containing RNAs in the Domain Archaea. http://escholarship.org/uc/item/8007d3p7.
31. Ziesche,S.M., Omer,A.D. and Dennis,P.P. (2004) RNA-guided nucleotide modification of ribosomal and non-ribosomal RNAs in Archaea. *Mol. Microbiol.*, **54**, 980–93.
32. Schoemaker,R.J.W. and Gultyaev,A.P. (2015) Computer simulation of chaperone effects of Archaeal C/D box sRNA binding on rRNA folding. *Nucleic Acids Res.*, **34**, 2015–2026.
33. Powers,T. and Noller,H.F. (1990) Dominant lethal mutations in a conserved loop in 16S rRNA. *Proc. Natl. Acad. Sci. U. S. A.*, **87**, 1042–1046.
34. Appel,C.D. and Maxwell,E.S. (2007) Structural features of the guide:target RNA duplex required for archaeal box C/D sRNA-guided nucleotide 2'-*O*-methylation. *RNA*, **13**, 899–911.
35. Cavaillé,J., Nicoloso,M. and Bachellerie,J.-P. (1996) Targeted ribose methylation of RNA *in vivo* directed by tailored antisense RNA guides. *Nature*, **383**, 732–735.
36. Martin,R., Hackert,P., Ruprecht,M., Simm,S., Brüning,L., Mirus,O., Sloan,K.E., Kudla,G., Schleiff,E. and Bohnsack,M.T. (2014) A pre-ribosomal RNA interaction network involving snoRNAs and the Rok1 helicase. *RNA*, **20**, 1173–82.
37. Kudla,G., Granneman,S., Hahn,D., Beggs,J.D. and Tollervey,D. (2011) Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 10010–5.
38. Stark,H., Rodnina,M. V, Wieden,H.-J., Zemlin,F., Wintermeyer,W. and van Heel,M. (2002) Ribosome interactions of aminoacyl-tRNA and elongation factor Tu in the codon-recognition complex. *Nat. Struct. Biol.*, **9**, 849–854.
39. Su,A., Tripp,V. and Randau,L. (2013) RNA-Seq analyses reveal the order of tRNA processing events and the maturation of C/D box and CRISPR RNAs in the hyperthermophile Methanopyrus kandleri. *Nucleic Acids Res.*, **41**, 6250–8.
40. Danan,M., Schwartz,S., Edelheit,S. and Sorek,R. (2012) Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Res.*, **40**, 3131–42.
41. Starostina,N.G., Marshburn,S., Johnson,L.S., Eddy,S.R., Terns,R.M. and Terns,M.P. (2004) Circular box C/D RNAs in Pyrococcus furiosus. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 14097–101.
42. Filée,J., Siguier,P. and Chandler,M. (2007) Insertion sequence diversity in archaea. *Microbiol. Mol. Biol. Rev.*, **71**, 121–157.
43. Randau,L. (2012) RNA processing in the minimal organism Nanoarchaeum equitans. *Genome Biol.*, **13**, R63.
44. Tripp,V., Martin,R., Orell,A., Alkhnbashi,O.S. and Backofen,R. (2016) Plasticity of archaeal C / D box sRNA biogenesis. *Mol. Microbiol.*, 10.1111/mmi.13549.
45. Babski,J., Maier,L.-K., Heyer,R., Jaschinski,K., Prasse,D., Jäger,D., Randau,L., Schmitz,R.A., Marchfelder,A. and Soppa,J. (2014) Small regulatory RNAs in Archaea. *RNA Biol.*, **11**, 1–10.
46. Mao,H., White,S.A. and Williamson,J.R. (1999) A novel loop-loop recognition motif in the yeast ribosomal protein L30 autoregulatory RNA complex. *Nat. Struct. Biol.*, **6**, 1139–1147.
47. Cléry,A., Bourguignon-Igel,V., Allmang,C., Krol,A. and Branlant,C. (2007) An improved definition of the RNA-binding specificity of SECIS-binding protein 2, an essential component of the selenocysteine incorporation machinery. *Nucleic Acids Res.*, **35**, 1868–84.
48. Saito,H., Kobayashi,T., Hara,T., Fujita,Y., Hayashi,K., Furushima,R. and Inoue,T. (2010) Synthetic translational regulation by an L7Ae-kink-turn RNP switch. *Nat. Chem. Biol.*, **6**, 71–8.
49. Bernick,D.L., Karplus,K., Lui,L.M., Coker,J.K.C., Murphy,J.N., Chan,P.P., Cozen,A.E. and Lowe,T.M. (2012) Complete genome sequence of *Pyrobaculum oguniense*. *Stand. Genomic Sci.*, **6**, 336–45.
50. Nawrocki,E.P. (2009) Structural RNA Homology Search and Alignment using Covariance Models. http://openscholarship.wustl.edu/etd/256.

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Tables and Figures

Table 1: Composite and transposed C/D box sRNAs. Two unusual types of sRNAs (composite and transposed) were identified. Composite sRNAs have D guide that shows sequence similarity to a guide in one sRNA family and D′ guides that show sequence similarity to a guide in a second sRNA families. These are given both family numbers separated by a forward slash (/). Transposed sRNAs have either a D guide that is shared with the D′ guide of the defining family or visa versa, a D′ guide that is shared with the D guide of the defining family. Transposed sRNAs are identified with the number of the defining family followed by a lower-case a or b. The *Pae* sR57b is considered as a transposed sRNA since the D′ guide normally associated with the sR57 is not present (to view these sRNA sequences, see http://lowelab.ucsc.edu/snoRNAdb/).

Composite C/D box sRNA	D′ guide	D guide
Pca sR12/45	Shared with D′ guide of sR45 family	Shared with D guide of sR12 family
Pca sR103/109	Shared with D′ guide of sR103 family	Shared with D guide of sR109 family
Transposed C/D box sRNA(s)	**D′ guide**	**D guide**
Pca sR26a	Shared with D guide of sR26 family	Not shared
Pog 46a	Shared with D′ guide of sR46 family	Not shared
Pca sR57a	Shared with D guide of sR57 family	Not shared
Pae sR57b	Not shared	Shared with D guide of sR57 family
Par and *Pog* sR13a	Shared with D guide of sR13 family	Not shared

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

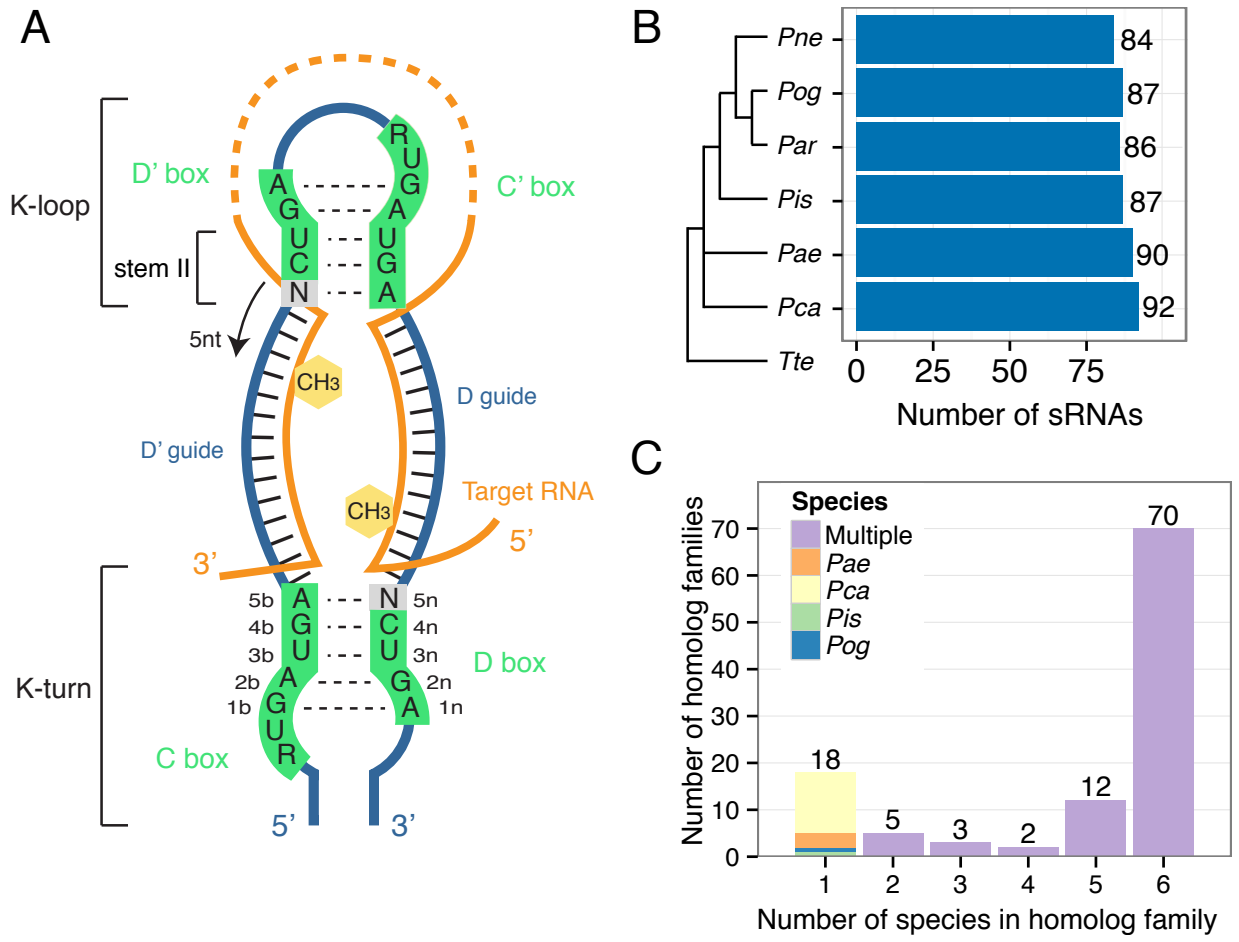

Figure 1: **Organization of *Pyrobaculum* C/D box RNAs into 110 homologous families** (A) The typical structure of an archaeal C/D box sRNA is depicted. The structure contains two K-motifs, the K-turn formed by the interaction between the C and D box sequences and the K-loop formed by the interaction between the D′ and C′ motifs (black dashed lines). The two guide regions located respectively between the D′ and C boxes and between the D and C′ boxes (green), base pair with the target RNA (orange) and methylation (yellow hexagon) occurs in the target nucleotide that base pairs with the guide five nts upstream from the start of the D′ or D box sequence. This is the "N+five" rule. (B) The number of identified sRNAs in each of the six species of *Pyrobaculum* is indicated with species tree as determined by 16S rRNA alignment (49). *Thermoproteus tenax* (*Tte*) is included as an outgroup. (C) C/D box sRNAs were organized into 110 homologous families based on sequence similarity of the guides and predicted targets in rRNA and tRNAs. C/D box sRNA numbers indicate to which family each belongs. Thus, *Pae* sR01, *Par* sR01, etc. belong to the sR01 family. C/D box sRNAs were first grouped into families using the original annotation numbering in *Pae* (1-65) (23). All other C/D box sRNAs were grouped into families starting at number 100. The majority of sRNAs fall into families with representatives on all six species.

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

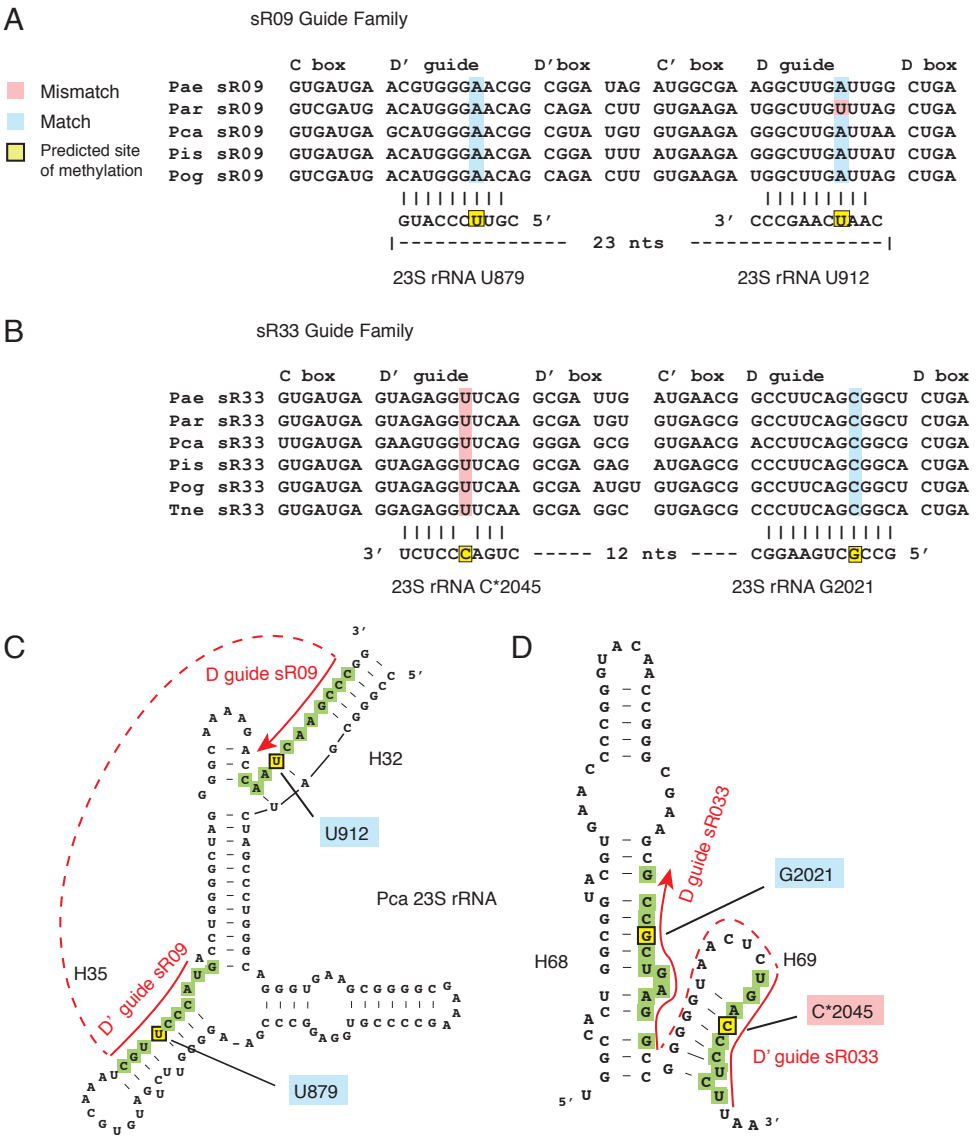

Figure 2: **Conserved Instances of mismatch at the site of modification ("N+5" position) indicate that in some cases the target interaction, rather than the modification, is important.** Sequence alignments of sRNAs in the sR09 (A) and sR33 (B) families are presented and their guide complementarities to 23S rRNA are indicated below. The critical "N+five" nucleotide in the guide regions is highlighted in blue when there is a Watson-Crick base pair between the guide and target and in rose when there is a mismatch base pair. Mismatched base pairs are also indicated by an asterisk in the rRNA position. The yellow highlight represents the "N+five" position in the rRNA target. (C) The secondary structure of helix 32-35 in 23S rRNA is depicted showing the complementarity of the sR09 D guide near position U912 and the D′ guide near position U879. Based on *in vivo* and *in vitro* studies demonstrating a requirement for a canonical base pair at "N+five" position (7, 34, 35), we predict that the *Par* sR09 base pairs with 23S rRNA at both positions, but that the D guide interaction does not result in modification of U912. Secondary structure generated by SSU-ALIGN package (50). (D) The secondary structure of helix 68-69 region in 23S rRNA is depicted showing the complementarity of the sR33 D guide near position G2021 and the D′ guide near position C2045. All six sR33 members have a mismatch at position 2045 at the "N+five" position in the guide-target interaction.

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

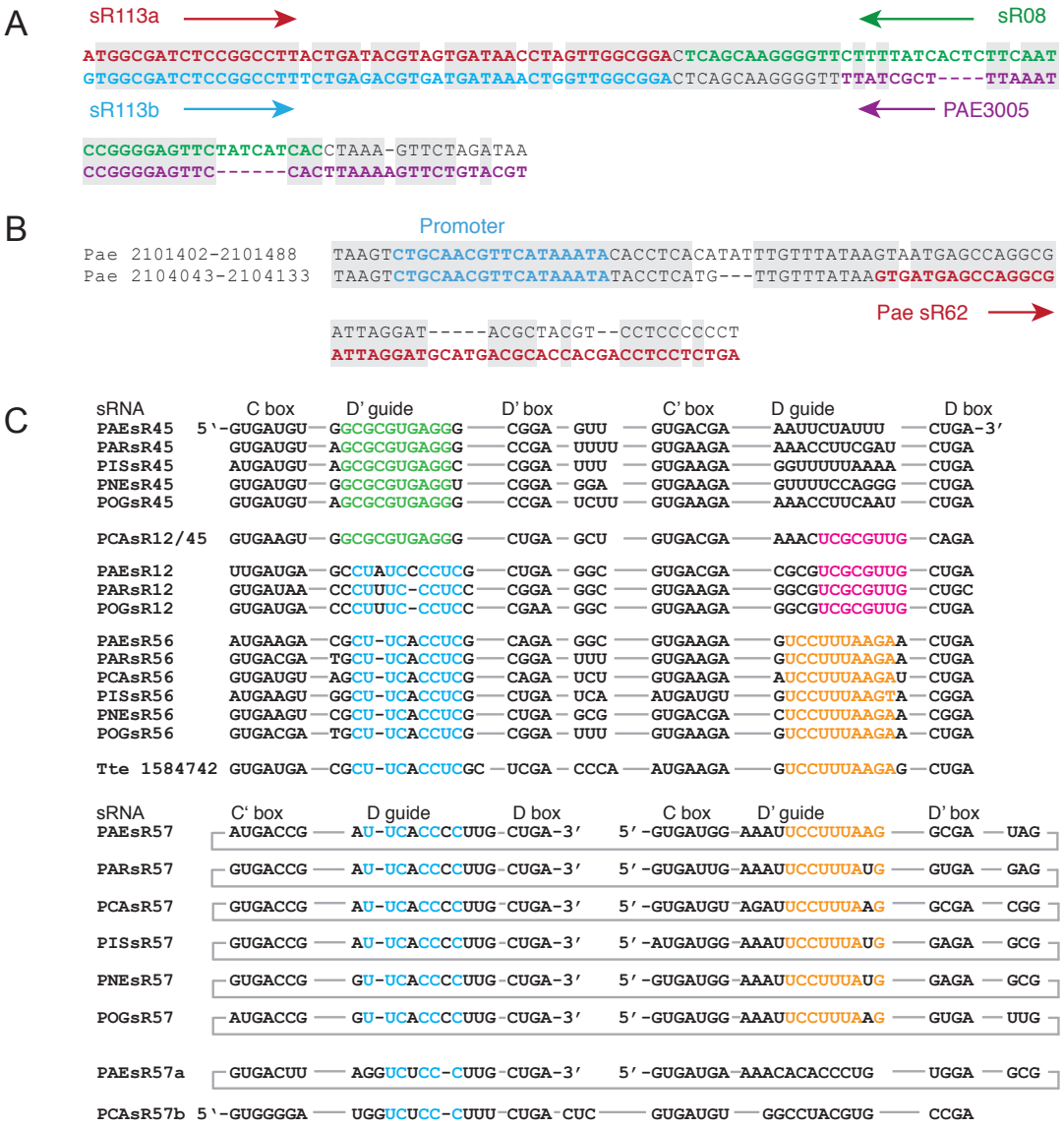

Figure 3: **Proliferation and evolutionary interconnections between sRNA genes.** (A) Duplication of the sR113 gene. The sequence similarity between the two nearly identical *Pae* sR113a (red) and sR113b (blue) genes and their 3′ flanking regions is illustrated (grey highlight). The 3′ flanking regions of the sR113a gene contains the sR08 gene (green) that is transcribed on the opposite strand and is separated from sR113a by a single nucleotide. The 3′ flanking region of the sR113b gene contains a remnant of an sR08 like gene that is partially buried in the PAE3005 ORF that is separated from the sR113b gene by 14 nucleotides. There is no sequence similarity between the 5′ flanking regions of the sR113a and 113B genes. (B) The *Pae* chromosome contains imperfect duplicate sequences that are separated by 2Mbps. Both copies retain a promoter-like sequence (blue) that is likely used to drive expression of the *Pae* sR62 gene (red). The second sequence contains an apparent remnant of the sR62 gene as suggested by the hyphenated regions of sequence identity (grey highlight). (C) Interconnected guide sequence similarity between different families of sRNAs. The colored sequences (green, magenta, blue and orange) indicate different sequence similarities in the guide regions of sRNAs of the interconnected sRNA families sR45, sR12/45, sR56, and sR57. The sRNA in the outgroup species *Thermproteus tenax* (*Tte*; chromosome start 1584742) is related to the sR57 family.

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

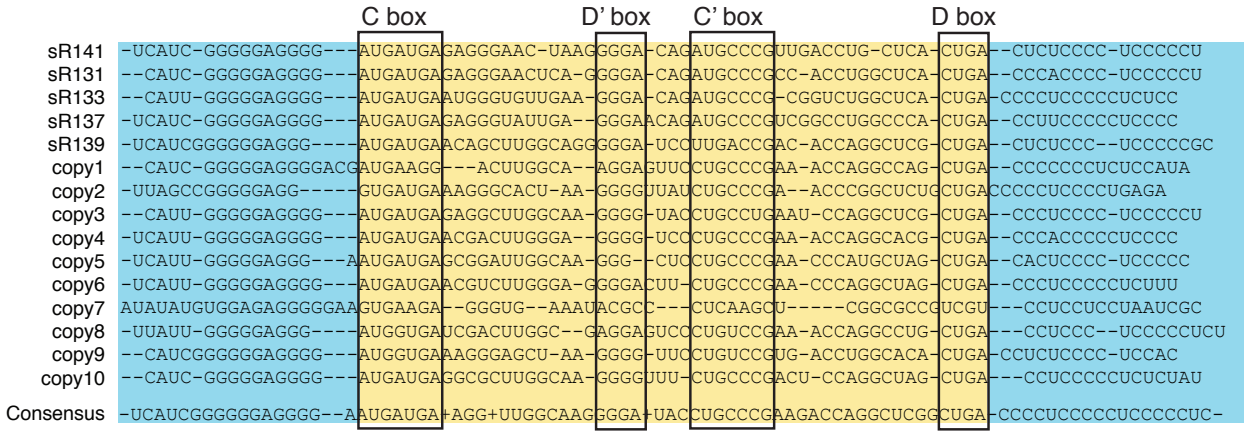

Figure 4: **MITE-like element in the *Pca* genome.** The chromosome of *Pca* contains fifteen copies of a MITE-like sRNA element. The sequences are aligned to illustrate the high degree of conserved sequence similarity in the 5′ and 3′ flanking inverted repeat sequences (blue highlight). The sRNA-like sequences (yellow highlight) contain canonical C and D boxes but generally degenerate D′ and C′ boxes (boxed). The conservation between the D and D′ guide sequences in the 15 elements is moderate with a consensus sequence at the bottom. Five of these elements were cataloged as authentic sRNAs.

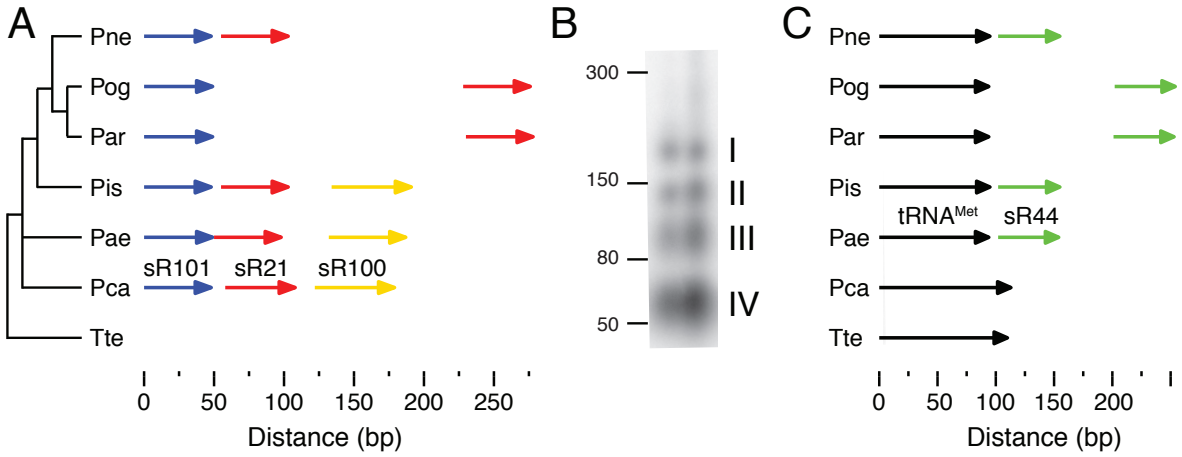

Figure 5: **Genomic context of sRNA genes.** (A) Genomic organization of the sR101 (blue), sR21 (red) and sR100 (yellow) genes in the six species of *Pyrobaculum*. The 16S rRNA phylogenetic tree with *Thermoproteus tenax* (*Tte*) as the outgroup, is illustrated on the left; the sRNA gene locations above a bp distance scale is illustrated to the right for the *Pyrobaculum* species. There is no representative of the sR100 gene in *Pne*, *Pog*, *Par* and in *Par* and *Pog* there is an approximately 200 nt insertion between the sR101 and sR21 genes. (B) Northern hybridization using RNA extracted from *Pae* cells with probes to *Pae* sR21. The position of molecular size markers is indicated in nts on the left and the identity of the four detectable transcripts is indicated on the right. (I) full length polycistronic sR101+sR21+sR100; (II) sR101+sR21; (III) sR21+sR100; (IV) sR21. (C) Linkage of tRNA and sRNA genes. In *Pae*, *Pis*, and *Pne* the sR44 genes (grey arrows) are located eight nts or less from the 3′ end of a $tRNA^{Met}$ gene (green arrows). In *Pog* and *Par* the distance between the $tRNA^{Met}$ and sR44 gene is increased to about 100 nts. In *Pca*, sR44 is located approximately 13 Kbps downstream of the $tRNA^{Met}$ gene. There is no representative of the sR44 family in *Tte*.

bioRxiv preprint doi: https://doi.org/10.1101/121921; this version posted March 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

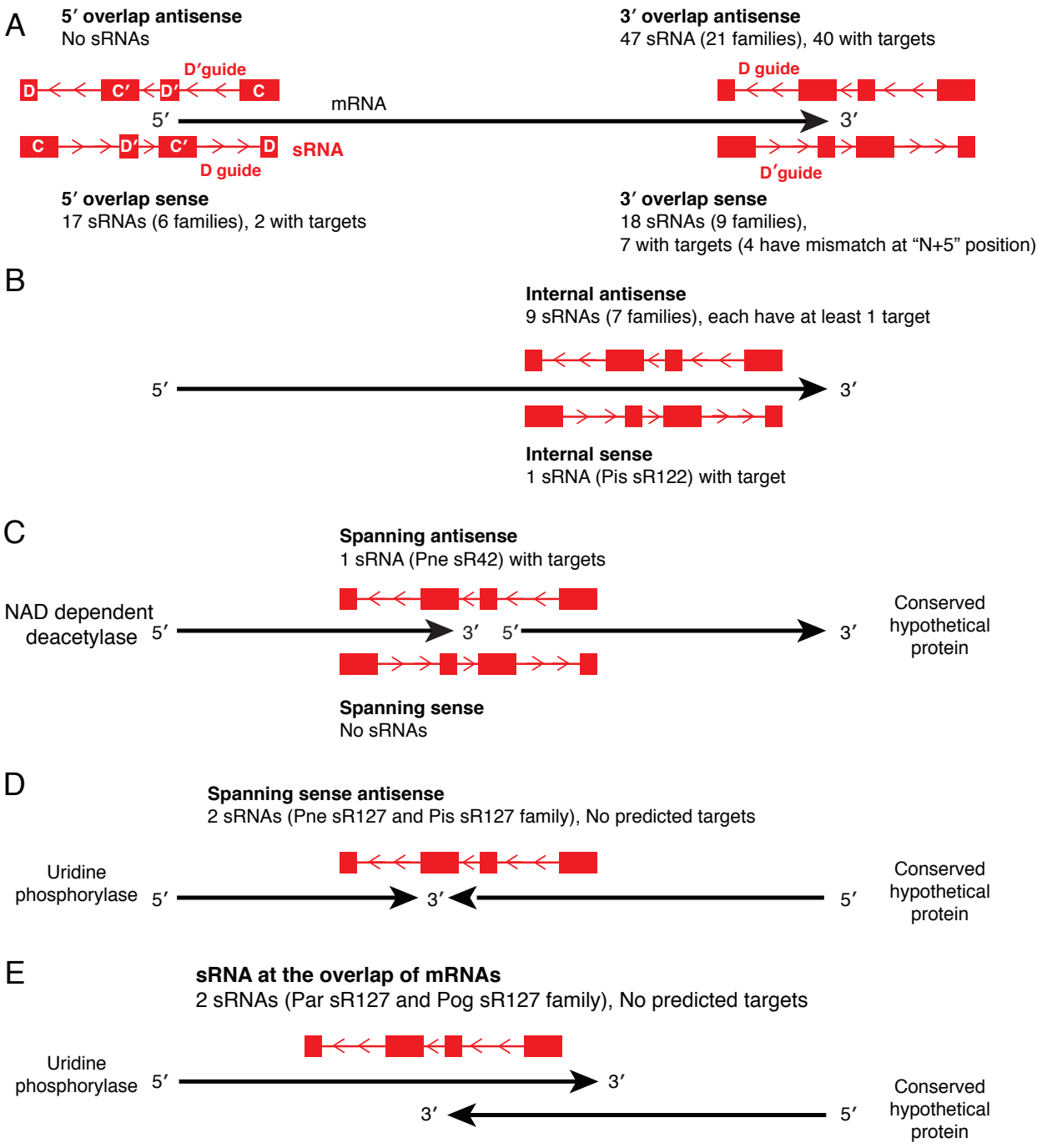

Figure 6: **The overlap between sRNA genes and protein coding genes.** The overlap between sRNA genes and protein-coding genes is divided into five categories (A-E). The protein genes are shown as black arrows with the 5′ and 3′ polarity indicated. Overlapping sRNA genes are shown in red with polarity indicated by the internal arrows; the C, D′, C′ and D box sequences are indicated as shown in the top left sRNA. The number of sRNA genes, the number of families that they represent and the number that have predicted targets is indicated for each type of overlap. Details relating to these sRNAs are given in the text and in the Supplementary Table S4.