Conservation of Archaeal C/D Box sRNA-Guided RNA Modifications

Ruth Breuer, Jose-Vicente Gomes-Filho and Lennart Randau *

Prokaryotic RNA Biology, Philipps-Universität Marburg, Marburg, Germany

Post-transcriptional modifications fulfill many important roles during ribosomal RNA maturation in all three domains of life. Ribose 2'-O-methylations constitute the most abundant chemical rRNA modification and are, for example, involved in RNA folding and stabilization. In archaea, these modification sites are determined by variable sets of C/D box sRNAs that guide the activity of the rRNA 2'-O-methyltransferase fibrillarin. Each C/D box sRNA contains two guide sequences that can act in coordination to bridge rRNA sequences. Here, we will review the landscape of archaeal C/D box sRNA genes and their target sites. One focus is placed on the apparent accelerated evolution of guide sequences and the varied pairing of the two individual guides, which results in different rRNA modification patterns and RNA chaperone activities.

Keywords: C/D box, methylation, RNA modification, RNA folding, RNA structure

INTRODUCTION

The chemical modification of RNA has long been known to play a role in a wide variety of cellular processes in all three domains of life. The manifold modifications can be introduced co- or post-transcriptionally and concern all classes of RNA molecules. The most abundant RNA modification is the ribose-2'-O-methylation, which is commonly found on ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) and also present on small nuclear RNAs (snRNAs) in archaea and eukaryotes (Maden et al., 1995; Kiss-László et al., 1996; Tycowski et al., 1998; Omer et al., 2000; Vitali and Kiss, 2019). This modification fulfills many different functions: It can protect RNA from ribonucleolytic cleavage, stabilize single base pairs, exhibit a chaperone function and influence folding at high temperatures (Kawai et al., 1992; Herschlag et al., 1993; Williams et al., 2001; Helm, 2006). The latter function is especially important in thermophilic organisms, therefore, it is no surprise that thermophilic archaea exhibit a significantly larger number of 2'-O-methylations than mesophilic archaea (Noon et al., 1998; Omer et al., 2000; Su et al., 2013).

In bacteria, 2'-O-methylations are comparatively rare and introduced by site- or region-specific protein-only enzymes (Decatur and Fournier, 2002). In contrast, methylation of the ribose moiety is more commonly observed in archaea and eukaryotes, which both utilize an RNA-dependent mechanism involving so-called C/D box s(no)RNAs. Here, the methylation reaction is performed by a ribonucleoprotein (RNP) complex carrying a small nucleolar RNA (snoRNA), or its archaean homolog, the sno-like RNA (sRNA; Figure 1A; Lischwe et al., 1985; Ochs et al., 1985; Filipowicz and Kiss, 1993; Maxwell and Fournier, 1995; Gaspin et al., 2000; Omer et al., 2000).
C/D box sRNAs were found to be approximately 50–70 nt long in archaea and between 50 and 300 nt long in eukaryotes (Lui and Lowe, 2013). These RNA molecules are named for four conserved sequence elements: the box C and box C’ motifs with the consensus sequence RUGAUGA and the box D and box D’ motifs with the consensus sequence CUGA (Maxwell and Fournier, 1995; Kiss-Laszlo et al., 1998). During C/D box sRNA folding, the motifs C and D base-pair and form a helix-loop-helix structure termed kink-turn (k-turn). The k-turn is a short stem structure comprising non-canonical base pairs and carrying two sheared base pairs (AG and GA) at its top (Watkins et al., 2000; Klein et al., 2001). Base pairing between the motifs C’ and D’ results in a similar structure called the k-loop, which consists of a single stem closed by a terminal loop (Nolivos et al., 2005). The sequences located between the motifs C and D’, and C’ and D, are complementary to the target RNA sequences and therefore serve as guides for the identification of methylation sites (Figure 1A). The length of the archaeal guide sequences ranges from 10 to 12 nt. Target methylation occurs at the nucleotide complementary to the fifth nucleotide upstream of the box D/D’ motif (Kiss-Lásló et al., 1996; Tran et al., 2005). A recent study in Drosophila identifies the minimal functional eukaryotic C/D box snoRNA as a single-domain molecule with (i) a terminal stem with a consensus k-turn domain, (ii) one box C and one box D separated by a 14 nt long antisense element and (iii) a one-nucleotide spacer between box C and the antisense element (Deryusheva and Gall, 2019).

Interestingly, archaean organisms harbor not only linear, but also circular C/D box sRNAs, though their role remains to be determined (Starostina et al., 2004; Danan et al., 2012; Randau, 2012; Su et al., 2013). The analysis of permuted RNA-seq reads allowed for the detection of circularization junctions of RNA molecules and revealed that C/D box sRNA termini can be fused. Inspection of these fusion sites indicated that the termini are not clearly defined, but can vary by few nucleotides for individual C/D box sRNA species. In addition, linear C/D box sRNAs are usually observed in parallel to circular variants. Notably, in Sulfolobus solfataricus, C/D box sRNAs occur predominantly in the linear form, whereas in Pyrococcus furiosus almost all C/D box sRNAs exist in both linear and circular forms with similar abundance (Starostina et al., 2004; Danan et al., 2012). Furthermore, archaean circular RNA molecules exist among tRNA introns and rRNA processing intermediates (Danan et al., 2012; Jüttner et al., 2020). Thermoproteus species were found to require circularization of signal recognition particle (SRP) RNAs to yield functional molecules (Plagens et al., 2015). In these cases, the 5’ and 3’ ends of the RNA molecule fold into close contact and form a bulge-helix-bulge motif which is recognized and cleaved by the tRNA splicing endonuclease and subsequently ligated by the tRNA ligase RtcB (Trotta et al., 1997; Englert et al., 2011; Popow et al., 2011). However, C/D box sRNA termini usually do not form canonical BHB motifs and the exact method of circularization remains unclear (Starostina et al., 2004). Since circular sRNA molecules have been nearly exclusively found in thermophiles thus far, it is suggested that the circularization provides stability at elevated growth temperatures (Starostina et al., 2004; Danan et al., 2012). Here, it is plausible that the close proximity of C/D box sRNA termini upon protein binding facilitates RNA ligation, representing a statistic event that is positively selected for due to the increased stability of the circularized products.

The C/D box sRNA is part of the C/D box RNP complex which contains three highly conserved proteins in archaea and four proteins in eukaryotes (Figure 1A). Upon adopting its secondary structure, the k-turn and k-loop of the C/D box sRNA are bound and stabilized by the RNA-binding protein L7Ae (Snu13/15.5 K in yeast/human; Kuhn et al., 2002; Omer et al., 2002; Gagnon et al., 2010). Binding of the C/D box sRNA by L7Ae depends on three essential features: (i) the terminal stem at the 5’ and 3’ ends of the C/D box sRNA,
which juxtaposes the boxes C and D motifs, (ii) two sheared GA base pairs formed by pairing of the box C and box D motifs, and (iii) the box C uridine which is part of the k-turn's internal loop (Kuhn et al., 2002). After binding of the C/D box sRNA by L7Ae, the assembly of the RNP is completed by binding of the proteins Nop5 (Nop56/Nop58 heterodimer in yeast and humans) and fibrillarin (Nop1/fibrillarin in yeast/human; Omer et al., 2002; Bortolin et al., 2003). The N-terminal and C-terminal domains of Nop5 interact with fibrillarin and the C/D box sRNA, respectively. Furthermore, the coiled-coil domain of Nop5 mediates Nop5-dimerization for optimal interaction with the C/D box sRNA (Aittaleb et al., 2003).

Fibrillarin exhibits a conserved S-adenosyl-methionine (SAM) binding motif and possesses methyl transfer activity. It was found that this activity is dependent on C/D box sRNA formation and could not be observed independent of the complex (Wang et al., 2001; Omer et al., 2002). For the 2'-O-methylation reaction, fibrillarin uses S-adenosyl-L-methionine as a methyl group donor and after depositing the methyl group at the 2'-OH moiety, the ribose preferably adopts an endo-conformation, thereby blocking sugar-edge interactions (Kawai et al., 1992; Auffinger and Westhof, 1997; Hansen et al., 2002; Motorin and Helm, 2010). Conversely, a fibrillarin-Nop5 heterodimer of Pyrococcus abyssi was recently found to perform in vitro 2'-O-methylation of rRNA independently of L7Ae and C/D box sRNAs (Tomkviene et al., 2017). In C/D box sRNPs containing fibrillarin, recent evidence shows that the guide RNA sequence determines the affinity of fibrillarin for the substrate and the extent of fibrillarin binding correlates with the efficiency of methylation (Graziadei et al., 2020).

First reports of the structure of the C/D box sRNA complex provided contradictory results for arrangement and number of associated proteins. However, it soon became clear that observed differences were caused by the type of C/D box sRNA that had been utilized in the in vitro experiments. While the usage of an artificial two-stranded RNA lacking the k-loop motif lead to the assembly of a monomeric complex consisting of one RNA and two copies of each protein, the usage of an in vitro transcribed natural C/D box sRNA sequence lead to the assembly of dimeric complex consisting of two RNAs and four copies of each protein (Figure 1B; Bleichert et al., 2009; Bleichert and Baserga, 2010; Xue et al., 2010; Lin et al., 2011; Bower-Phipps et al., 2012; Lapinaite et al., 2013). Accordingly, these results suggest that the nature of the RNA determines if mono- or diRNPs are assembled and influences these complexes' functional roles. These and other findings on the structural diversity of C/D box sRNPs are extensively reviewed by Yu et al. (2018).

C/D BOX sRNA TARGETS

Ribosomal Targets

A first study aiming to identify archaecal sRNAs employed co-immunoprecipitation with archaenal fibrillarin and Nop5 and identified 18 C/D box sRNAs in Sulfolobus acidocaldarius. Furthermore, methylations at the predicted target positions for six of these sRNAs were verified using deoxyribonucleotide triphosphate (dNTP) concentration-dependent primer extension assays (Omer et al., 2000). Subsequent experiments lead to the discovery of over 200 sRNAs across seven archaeal species, targeting mostly – though not exclusively – archaenal rRNAs. Here, it was also revealed that, in contrast to eukaryotes, most archaenal sRNAs possess two sequences able to guide methylation and that these double guides can target closely linked positions on the same RNA molecule (Omer et al., 2000). At the same time, another study reported the identification of a family of 46 archaenal sRNAs in the genomes of three species of the hyperthermophile Pyrococcus species Additionally, these sRNAs were experimentally verified in P. abyssi using Northern hybridization (Gaspin et al., 2000).

Shortly afterwards, another study used a combination of MALDI-MS and primer extension assays to locate conserved modification patterns in the A-loop region of the 23S rRNA in five archaenal and eubacterial species. The A-loop of the 23S rRNA (also known as helix 92), constitutes part of the peptidyl transferase loop in domain V of the 23S rRNA and its functional importance has been emphasized by several studies (Hansen et al., 2002). In fact, loss of the 2'-O-ribose methylation at position U2552 in the A-loop leads to decreased growth rate and reduced protein synthesis activity in Escherichia coli (Caldas et al., 2000). It was shown that despite variation in the exact positions of modifications in the helices 90–92, modifications in the A-loop are always present at positions equivalent to U2552 and/or G2553 in E. coli (Hansen et al., 2002). Projecting these previously identified modifications from E. coli onto their corresponding positions in the 2.4 Å X-ray crystal structure of the Haloarcula marismortui 50S ribosome subunit, all modifications were found to be clustered around the peptidyl transferase center (Ban et al., 2000; Hansen et al., 2002).

The advent of RNA-seq has enabled researchers to efficiently identify C/D box sRNAs among any organism's total RNA pool. Subsequently, their guide sequences can be used to computationally predict their potential RNA targets on the basis of their hybridization potential. Analyses of RNA-seq coverage revealed large numbers of abundantly transcribed small RNAs with readily identifiable C and D box sequences. These postulated C/D box sRNAs were, for example, described for model archaea of different archaenal phyla: Nanoarchaeum equitans (26 C/D box sRNAs), Ignicoccus hospitalis (128 C/D box sRNAs), Methanococcus maripaludis (7 C/D box sRNAs), Methanopyrus kandleri (127 C/D box sRNAs), Pyrobaculum caldifontis (88 C/D box sRNAs), S. acidocaldarius (61 C/D box sRNAs) and Thermoproteus tenax (52 C/D box sRNAs). Using the guide sequences of these C/D box sRNAs, 719 potential 2'-O-methylation sites in the archaenal 23S and 16S rRNA sequences were identified and hinted at common targets and rRNA regions (Figure 2). This dataset revealed some shared methylation targets but did not reveal a single position to be uniformly present in all seven species. Instead, it became clear that methylation targets cluster in hotspot regions of the RNA molecules. Among all investigated species, these methylation hotspots have been detected in the functionally important and evolutionary conserved regions of the ribosome (Liang et al., 2009; Dennis et al., 2015; Lui et al., 2018).
The archaeal consensus 23S rRNA structure exhibits six domains surrounding a central core. Conserved methylation hotspots are identified in domain II helices 35 and 35a, domain IV helices 61 and 68–71, and domain V helices 90–93 (Dennis et al., 2015). These regions correspond to the ancient core of the ribosome where domain V lies at the center of the large ribosomal subunit. One major cluster of hotspots surrounds the catalytic peptidyl transferase center located in domain V, where peptide bond formation and peptide release occurs (Figure 2; Petrov et al., 2013; Dennis et al., 2015; Lui et al., 2018). Another predicted cluster lies in domain IV where helices 68, 69, and 71 form part of the interface between the large and the small ribosomal subunits (Cate et al., 1999; Dennis et al., 2015; Lui et al., 2018).

The archaeal consensus 16S rRNA structure consists of four domains connected by a central core which is located close to the functional decoding center of the small ribosomal subunit (Wimberly et al., 2000). Here, conserved methylation hotspots are predicted in helices 3, 18, and 27 (Dennis et al., 2015). In fact, helix 18 is the core of the decoding center and responsible for monitoring the codon-anticodon pairings (Ogle et al., 2001). Due to their location, the methylated nucleotides likely contribute to stabilizing the decoding center as well as the association of the four domains. Furthermore, predicted methylation clusters are especially dense in regions which are not protected by RNA-binding proteins. It has therefore been proposed that the modifications help stabilize the structure of these exposed regions and in turn support subunit interactions (Dennis et al., 2015). These findings are corroborated by a study across six Pyrobaculum species, revealing that most rRNA-targeting C/D box sRNAs are dual guides targeting sites within 100 nt of each other (Lui et al., 2018).

**Non-ribosomal Targets**

The initial prediction and experimental verification of archaeal rRNA targets of C/D box sRNAs revealed additional antisense elements matching tRNAs (Gaspin et al., 2000; Omer et al., 2000). Further investigation revealed the presence of four C/D box sRNAs targeting the first position of the anticodon of the tRNAs tRNA-Leu (CAA), tRNA-Leu (UAA), tRNA-Met and tRNA-Trp in three Pyrococcus species. One of these sRNAs, termed sR50, corresponds to the intron of its predicted target, the pre-tRNA-Trp and was shown to guide the methylation of the pre-tRNA-Trp nucleotides Cm34 and Um39 in in vitro experiments in Haloferax volcanii (Omer et al., 2000; D’Orval et al., 2001). This proposed cis-acting mechanism was later shown to be a trans-acting or intramolecular mechanism by a study which also revealed the sequential pattern of C/D box sRNA-guided methylation (Bortolin et al., 2003; Singh et al., 2004). Recently, an in-depth analysis of C/D box sRNA families in Pyrobaculum species computationally predicted tRNA targets for 16% of the identified guide sequences (as opposed to 56% rRNA targets; Lui et al., 2018). Unsurprisingly, the tRNA methylation targets correspond to
structurally conserved regions, however, in contrast to rRNA methylation via double-guide sRNAs, tRNA methylation is mediated by sRNAs where one guide targets the tRNA while the other guide has no target. Interestingly, a tRNA-targeting sRNA can mediate methylation of either a single or many tRNAs, depending on whether its guide targets a unique sequence like the region surrounding the wobble base in the anticodon (position 34) or a conserved sequence shared across different tRNA families (Lui et al., 2018). Conversely, a few tRNAs in S. acidocaldarius and P. furiosus contain several methylated or predicted methylated nucleotides which have not yet been linked to corresponding sRNAs or stand-alone specific methyltransferases (Wolff et al., 2020).

Several computational studies also identified numerous "orphan guides," which are defined as snoRNA guide sequences that do not show complementary to any known RNA target (Hüttenhofer et al., 2001; Yang et al., 2006; Lui et al., 2018). Looking for target RNA sequences outside of rRNAs or tRNAs has relieved many guide sequences of their "orphan status," thereby contributing to the expanding functional diversity of snoRNAs (see also Discussion section and reviewed in Falaleeva et al., 2017; Bergeron et al., 2020). At least some orphan guides can be considered as a consequence of accumulated mutations in the guide sequence, eventually leading to the evolution of guides for novel methylation targets (Dennis et al., 2001; Lui et al., 2018).

C/D Box sRNAs and Their Role as RNA Chaperones

Shortly after their discovery, it was suggested that C/D box sRNAs might function as RNA chaperones (Steitz and Tycowski, 1995). This theory gained further support by a subsequent computer simulation of long-range rRNA interactions at elevated temperatures in the presence of double-guided C/D box sRNAs (Schoemaker and Guityaev, 2006). Indeed, there exist several C/D box sRNAs whose guide sequences target positions at a considerable distance from each other. For example, N. equitans exhibits two instances where the predicted targets are distant in sequence but close in the secondary structure: the guides of sR17 target nucleotides situated on opposing strands of helix 28, the defining helix of domain III of the 16S rRNA and the guides of sR15 target nucleotides on opposing strands of helix 30 which defines a large subsection of domain III (Dennis et al., 2015). These findings lead to the conclusion that the C/D box sRNAs act as chaperones by bringing distant sRNA sequences together and facilitate their annealing, thereby assisting in ribosome subunit assembly (Gaspin et al., 2000; Dennis et al., 2015).

C/D BOX sRNA GENE AND GUIDE EVOLUTION

Genomic Context Variability

In yeast, most snoRNA genes are transcribed from independent RNA polymerase II or III promoters as mono- or polycistronic transcripts (Li et al., 2005; Dieci et al., 2009). In plants, the C/D box sRNA genes exist almost exclusively as polycistronic clusters, or, to a lesser extent, as dicistronic tRNA-C/D box snoRNAs (Leader et al., 1997; Kruszka et al., 2003; Barbezier et al., 2009). In vertebrates, independent promoters are rare and snoRNA genes are usually located in introns of protein-coding or non-protein-coding genes (Pelczar and Filipowicz, 1998; Weber, 2006) and only few of them exist as polycistrons (Leader et al., 1994; Tycowski et al., 2004). The polycistronic transcripts or intron-located C/D box snoRNAs are processed and matured by endo- and exoribonucleolytic activities (Caffarelli et al., 1994; Kiss and Filipowicz, 1995; Villa et al., 1998).

In archaea, analysis of the genomic context of C/D box sRNA genes revealed a variable organization of promoter and processing elements. A 2017 study on the genomic context of C/D box sRNAs in six archaeal model organisms concluded that only a minority of archaeal C/D box sRNAs are transcribed from independent promoters as only 20% of all investigated genes exhibited a conserved, TATA box-like motif in their 50 nt upstream region (Tripp et al., 2017). Instead, the majority of C/D box sRNA genes overlap with either the 5′ or the 3′ end of a neighboring open reading frame (ORF; Gaspin et al., 2000; Dennis et al., 2001; Randau, 2012; Tripp et al., 2017; Lui et al., 2018). Twenty-five percent of genes show overlap with a 3′ end and carry the stop codon of the upstream gene within their sequence. Though the stop codon can be found in any of the four conserved motifs, in almost 50% of cases it was located in the C box motif (Tripp et al., 2017). Notably, the stop codon "UGA" is found in most of the C box and D box motifs and only a few nucleotide changes separate it from evolving into a k-turn element. Therefore, it was proposed that the start or stop codons of overlapping genes are responsible for the accelerated evolution of k-turn motifs in C/D box sRNA genes (Tripp et al., 2017). A 5′ overlap was identified for 7% of the investigated C/D box sRNA genes. Similarly, the start codon of an overlapping downstream coding region was found to be located within different parts of C/D box sRNA sequences, most commonly however, in the guide sequence or downstream of the D box motif (Tripp et al., 2017). Analyses of the impact of these overlaps on the transcription rate of a reporter gene revealed neutral or only slightly negative effects for 3′ overlapping C/D box sRNAs genes. However, a 5′ overlap caused a significant reduction in transcription of the downstream gene, which is in agreement with the rare presence of this gene arrangement in nature (Tripp et al., 2017). In some cases, this scenario might result in the formation of pseudogene sequences downstream of C/D box sRNA genes.

A significant fraction of C/D box sRNA genes was found to occur in clusters of two or three genes indicating polycistronic transcription. Several dicistronic transcripts, including examples of tRNA- C/D box sRNA fusions, were identified (Tripp et al., 2017). In these cases, different C/D box sRNAs were, for example, found to be located directly downstream of genes coding for tRNASeR in I. hospitalis, tRNAPro in T. tenax, and tRNAVal in N. equitans. Consequently, tRNA 3′ maturation is suggested to generate the 5′ terminus of the respective C/D box sRNAs. In some cases, C/D box
sRNAs were also found to be located within tRNA introns and shown to mediate tRNA methylation in cis (D’Orrval et al., 2001). The majority of the remaining genes are located in intergenic regions and some of them exhibit the aforementioned, conserved motifs, indicating the presence of an independent promoter. Others are located up- or downstream of neighboring protein-coding genes at a distance of less than 25 nt. Consequently, most C/D box sRNA genes do not require independent promoters, as they are part of longer precursor transcripts that are subsequently processed into mature C/D box sRNAs (Tripp et al., 2017).

Mutational analyses of S. acidocaldarius upstream and downstream regions of a C/D box sRNA gene revealed that these surrounding sequences can be changed without affecting C/D box sRNA maturation. Instead, the presence of the conserved, internal box motifs responsible for forming the k-turn and k-loop structures was found to be essential (Tripp et al., 2017). These observations suggest that the insertion of a C/D box sRNA gene into a transcriptionally active genome context is sufficient to obtain mature C/D box sRNAs. In this model, C/D box sRNPs formation would result in the protection of the C/D box sRNA body via protein-RNA contacts while the exposed RNA termini would gradually be processed by cellular nucleases and/or chemical RNA degradation at elevated temperatures. In addition, interactions with RNA ligases would then yield fractions of circularized C/D box sRNAs without accessible RNA termini.

### Identification of Guide Sequences

The conserved box C and box D sequences of C/D box sRNAs and their evenly spaced arrangement into two k-turns for L7Ae binding allow for the computational prediction of C/D box sRNA genes among archaeal sequences. One of the first programs that used these features to scan genomes for snoRNAs and their putative methylation targets in rRNA is snoScan (Lowe and Eddy, 1999). This program applies probabilistic models of snoRNAs and initially identified 22 novel guides in Saccharomycys cerevisiae. Although the initial development of snoScan focused on the prediction of rRNA methylation sites, different RNA sequences can also be used for target prediction. Another tool that utilizes probabilistic models is snoSeeker (Yang et al., 2006). This program searches for box C and box D elements, terminal stem pairing and, optionally, target sequences, enabling prediction of both guide and orphan sRNAs. The algorithm SnoReport utilizes support vector machines (SVM) and RNA secondary structure prediction to identify C/D box sRNA sequences (Hertel et al., 2008; de Araujo Oliveira et al., 2016). A more recent version, SnoReport 2.0, takes advantage of features of known C/D box sRNAs detected in invertebrates to improve its SVM during the training phase. In addition, a k-turn test, in which the predicted sRNAs must present G.A dinucleotides in box C and D, at least one uridine for the U-U pair and a Watson-Crick base pair between the sixth nt of the C box and the first nt of the D box significantly reduced the number of false positives (de Araujo Oliveira et al., 2016). Additionally, snoStrip (Bartschat et al., 2014) is a comprehensive pipeline that applies the following steps: first, a sequence-based homology search is performed using BLASTn (Altschul et al., 1990) and further complemented with the generation of covariance models. Next, the detection of characteristic C and D box motifs is performed through temporary alignments using MUSCLE (Edgar, 2004). If the location of a box motif agrees in all alignments, the position is annotated as a candidate box sequence. After defining conserved sequence elements, a secondary structure analysis is employed to ensure that only correctly folded C/D box sRNAs are further analyzed. Finally, prediction of the putative targets is achieved using Plexy, a tool that calculates the optimal thermodynamic interactions of a C/D box sRNA with candidate targets (Kehr et al., 2011). As the repertoire of sequenced C/D box sRNAs increases, several databases have been created to categorize these molecules, including the Plant snoRNA database (Brown et al., 2001). For archaea, two databases can be highlighted: Rfam and snoNADb (Lowe and Eddy, 1999; Omer et al., 2000; Kalvari et al., 2021). The Rfam database uses a generalized search based on covariance models to annotate a wide diversity of non-coding RNAs, including C/D box sRNAs, that are conserved in three or more species. The database snoNADb compiles homologs of C/D box sRNAs that were predicted for crenarchaeal and euryarchaeal species, while also providing information about their putative targets.

Even though these tools and different strategies are available for C/D box sRNA prediction, this class of RNA is still underrepresented in archaeal annotations (Gardner et al., 2010) and only a combination of RNA-seq analyses, comparative genomics and computational methods allow for complete C/D box sRNA identification (Lui et al., 2018). Since most prediction algorithms were developed using eukaryotic C/D box sRNAs as training sets, it is hypothesized that features which are exclusive to archaeal C/D box sRNAs are absent, therefore impacting the overall efficiency and reliability of the predictions. Here, the lower degree of conservation of C’ and D’ boxes in eukaryotes in comparison to archaea represents one clear difference (Yang et al., 2020). The recent increase in the availability of archaeal transcriptome datasets is an asset to expand the repertoire of hand-curated C/D box sRNAs. Utilizing experimentally validated datasets, tools that are based on pre-generated covariance models (e.g. INFERNAL - cmsearch) can take advantage of the conserved C’ and D’ motifs to drastically increase the number of predicted C/D box sRNAs in Archaea (Lui et al., 2018). The reduction of the stringency of the search parameters for C/D box sRNA genes results in increasing amounts of false positive sequences resembling C/D box sRNA genes. These hits can be viewed as sequence space with increased probability of evolving novel C/D box sRNA elements and might impact the dynamics of guide sequence generation.

### DISCUSSION

C/D box sRNAs were early found to possess other functions besides their established role in the 2'-O-methylation of rRNA and tRNA (Dennis et al., 2001). C/D box snoRNAs of yeast and
eukaryotes (especially humans) have been shown to be involved in diverse functions including rRNA processing, RNA base acetylation, regulation of mRNA 3' processing, and alternative pre-mRNA splicing (Kass et al., 1990; Falaleeva et al., 2016; Huang et al., 2017; Sharma et al., 2017b). Recently, it was shown that C/D box snoRNA-guided methylation of mRNA regulates protein expression and enzyme activity (Elliott et al., 2019). Additionally, it was revealed that many snoRNAs are processed into shorter forms such as miRNA (called sno-miRNA) and efficiently exert gene regulatory functions (Brameier et al., 2011). It was also recently discovered that snoRNAs retained in longer RNAs can interact with non-canonical proteins and act as a decoy, thereby hindering their activity (Bergeron et al., 2020). In fact, the influence of C/D box snoRNAs in the human metabolism is very significant: the C/D box snoRNA U60 is involved in intracellular cholesterol trafficking and regulation of cholesterol homeostasis (Brandis et al., 2013). Lack of expression of the C/D box snoRNA cluster SNORD116 causes Prader-Willi-Syndrome, a neurobehavioral disorder manifesting itself in hyperphagia and leading to morbid obesity (Ding et al., 2008; Duker et al., 2010). Furthermore, the snoRNA U50 was found to be deleted in several common cancers, with a particularly strong association in breast cancer and prostate cancer (Dong et al., 2008, 2009; Siprashvili et al., 2015). With an evident link between snoRNAs and human cancer and other systemic diseases being established, a strong resurgence of eukaryotic snoRNA research has been noted. New findings in this area continually expand our knowledge of diverse snoRNA functions and have most recently been reviewed by Deogahria and Majumder (2019), Liang et al. (2019) and Bratković et al. (2020).

Using RiboMethSeq to analyze the 2'-O-methylation patterns on eukaryotic rRNAs, it was shown that a knockdown of the methyltransferase fibrillarin (FBL) in HeLa-cells leads to a site-specific decrease of methylation levels. Affected sites were identified in conserved and/or functionally important regions of the ribosome, like its “core,” close to the A- and P-sites, the intersubunit bridges and the peptide exit tunnel, while 2'-O-Me sites close to the peptidyl transferase center were not subject to variation in methylation levels upon FBL knockdown (Erales et al., 2017). Another study from the same year mapping 2'-O-methylation sites vulnerable to fibrillarin depletion on human rRNAs, also investigated the C/D box snoRNAs whose guide sequences target these “vulnerable” methylation sites. However, these studies did not find a direct correlation between the sites with a variable methylation level and abundance of the snoRNAs which target them (Sharma et al., 2017a). More recently, RiboMethSeq was adapted to map 2'-O-methylation sites on rRNAs in human breast cancer samples (Marcel et al., 2020). Here, the identified methylation sites were divided into two classes: one class encompassing a larger group of rRNA 2'-O-methylation sites with a low inter-patient variability, termed “stable” sites, and a second class encompassing a smaller group of rRNA 2'-O-methylation sites with a high inter-patient variability in methylation levels, termed “variable” sites. These stable sites were found to be located in the decoding center, the peptidyl transferase center and the polypeptide exit tunnel, while the variable sites were located in layers 1 or 2 nt away from these functional regions. Furthermore, it is suggested that the 2'-O-methylation levels at the variable RNA sites are associated with breast cancer subtype and tumor grade, indicating that not only tumor size but also the pattern of RNA 2'-O-methylation influences factors like tumor aggressiveness and patient survival (Marcel et al., 2020).

Additional functional roles of archaeal C/D box sRNAs are likely also to be discovered, which is supported by the existence of many “orphan” C/D box sRNA guides in archaea without easily detectable complementary methylation targets. Guide sequences of C/D box sRNAs define the methylation landscape of their hybridization targets. As these targets are mostly highly conserved rRNA molecules, it is initially surprising to see that C/D box sRNA do not exhibit a similar degree of conservation. As described in Ribosomal Targets section, ubiquitous methylation of functionally and structurally important rRNA regions can be achieved by different sets of C/D box sRNAs with varied guide sequences and guide sequence pairs. This dynamic evolution of guides has been analyzed in detail in six Pyrobaculum species containing 526 different C/D box sRNAs that were organized into 110 homologous families (Lui et al., 2018). At the genus level, less than two-thirds of the predicted targets were found to be conserved among the six Pyrobaculum species and guide sequences exhibited short insertions, deletions or substitutions. In the Pyrobaculum species dataset, 28% of guides showed no significant complementarity to potential RNA targets (Lui et al., 2018). As C/D box sRNA genes often overlap with adjacent genes that provide promoter elements and processing signals, it is also possible that the overlapping sequence results in the creation of an orphan guide sequence that is paired with a second guide that provides methylation benefits for the cell. Therefore, the presence of orphan guide sequences can partly be considered to be a consequence of the plasticity of the genomic context of C/D box sRNA genes. Here, it remains to be understood why C/D box sRNAs exhibit dynamic scenarios of polycistronic transcriptional units with different mRNA and tRNA partners. In mammalian cells, snoRNAs have been found in retroposable elements and it was proposed that retroposition followed by genetic drift would be able to increase snoRNA diversity and change their modification landscape (Weber, 2006). Mobile features of archaean C/D box sRNA genes remain to be discovered.

**AUTHOR CONTRIBUTIONS**

RB, J-VG-F, and LR conceptualized and wrote the manuscript. All authors contributed to the article and approved the submitted version.

**FUNDING**

Funding was provided by the Deutsche Forschungsgemeinschaft (DFG), project RA 2169/4-1.
