Plasticity of archaeal C/D box sRNA biogenesis

Vanessa Tripp[1,2], Roman Martin[1], Alvaro Orell[1], Omer S. Alkhnbashi[3], Rolf Backofen[3], Lennart Randau[1,2]*

[1] Max Planck Institute for Terrestrial Microbiology, Karl-von-Frisch Strasse 10, 35043 Marburg, Germany

[2] LOEWE Center for Synthetic Microbiology, SYNMIKRO, Karl-von-Frisch-Strasse 16, 35043 Marburg, Germany

[3] Bioinformatics group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, 79110 Freiburg, Germany

* To whom correspondence should be addressed. Tel: +49 6421 178 600; Fax: +49 6421 178 599; Email: lennart.randau@mpi-marburg.mpg.de

Keywords

C/D box sRNA, RNA biogenesis, archaea, RNA modification

Summary

Archaeal and eukaryotic organisms contain sets of C/D box s(no)RNAs with guide sequences that determine ribose 2'-*O*-methylation sites of target RNAs. The composition of these C/D box sRNA sets is highly variable between organisms and results in varying RNA modification patterns which are important for ribosomal RNA folding and stability. Little is known about the genomic organization of C/D box sRNA genes in archaea. Here, we aimed to obtain first insights into the biogenesis of these archaeal C/D box sRNAs and analyzed the genetic context of more than 300 archaeal sRNA genes. We found that the majority of these genes do not possess independent promoters but are rather located at positions that allow for co-transcription with neighboring genes and their start or stop codons were frequently incorporated into the conserved boxC and D motifs. The biogenesis of plasmid-encoded C/D box sRNA variants was analyzed *in vivo* in *Sulfolobus acidocaldarius*. It was found that C/D box sRNA maturation occurs independent of their genetic context and relies solely on the presence of intact RNA kink-turn structures. The observed plasticity of C/D box sRNA biogenesis is suggested to enable their accelerated evolution and, consequently, allow for adjustments of the RNA modification landscape.

Introduction

Post-transcriptional modifications are found in RNA molecules of all three domains of life. In eukaryotes and archaea, the most prevalent modifications are pseudouridylations and 2'-*O*-methylations at the ribose moiety of rRNA, tRNA and snRNA nucleotides (Omer *et al.*, 2000, Maden *et al.*, 1995, Kiss-Laszlo *et al.*, 1996, Tycowski *et al.*, 1998). These modifications are introduced by two different RNA-guided mechanisms. In both cases, RNAs with sequences that are complementary to the target RNAs serve as guide molecules in ribonucleoprotein complexes that exhibit modification activity (Omer et al., 2000, Maxwell & Fournier, 1995, Balakin *et al.*, 1996). The RNAs that guide the introduction of 2'-*O*-methylations are termed C/D box small nucleolar (sno)RNAs in eukaryotes and their archaeal homologues are termed C/D box sno-like RNAs (C/D box sRNA) (Omer et al., 2000, Kiss-Laszlo et al., 1996). These RNA molecules are characterized by two conserved sequence elements called boxC at the 5' terminus (consensus sequence RUGAUGA) and boxD at the 3' terminus (consensus sequence CUGA) which are often duplicated in the center resulting in boxC' and boxD' elements (Kiss-Laszlo *et al.*, 1998). Upon C/D box s(no)RNA folding, the boxC and boxD sequences base-pair, which generates a helix-internal loop-helix structure termed kink-turn (k-turn) (Watkins *et al.*, 2000, Klein *et al.*, 2001). Pairing between the boxC' and boxD' sequences results in a similar structure termed k-loop, as the terminal helix is replaced by a loop (Nolivos *et al.*, 2005). Both structures typically consist of two sheared GA base pairs, a mismatch pair and a Watson-Crick base pair (Klein et al., 2001, Nolivos et al., 2005). The k-turn and k-loop structures are bound and stabilized by the protein L7Ae (Zago *et al.*, 2005, Kuhn *et al.*, 2002) which results in the recruitment of Nop5 and fibrillarin and the formation of an active C/D box ribonucleoprotein complex (RNP) (Aittaleb *et al.*, 2003, Lapinaite *et al.*, 2013, Yip *et al.*, 2016). The C/D box sRNA sequences that are located between the boxC and boxD' motifs and between the boxC' and boxD motifs are the guide sequences that exhibit complementarity to the sequences of the target RNA and determine the sites for the 2'-*O*-methylation reaction (Kiss-Laszlo et al., 1996). Thus, C/D box sRNAs usually contain two guide sequences. The introduction of the modifications increases the stability of the RNA targets as it provides protection from ribonucleolytic cleavage. Additionally, rRNA folding may be influenced as several C/D box sRNAs exist that target nucleotides that are apart in the rRNA sequence but close in the rRNA secondary structure. Thus, C/D box s(no)RNAs have been suggested to act as RNA chaperones (Helm, 2006, Herschlag *et al.*, 1993, Watkins & Bohnsack, 2012, Gaspin *et al.*, 2000, Steitz & Tycowski, 1995, Polikanov *et al.*, 2015, Dennis *et al.*, 2015)

Archaeal and eukaryotic C/D box s(no)RNAs fulfil the same function but the genomic organization of their genes differs. In eukaryotes, the genomic organization of C/D box snoRNA genes and the transcription and maturation are highly variable. In yeast, as well as in plants, most snoRNA genes

are transcribed from independent RNA polymerase II (or, less frequently, RNA polymerase III) promoters as mono- or polycistronic transcripts. In plants, C/D box snoRNA genes exist almost exclusively as polycistronic clusters (Li *et al.*, 2005, Liang *et al.*, 2002, Brown *et al.*, 2003, Dieci *et al.*, 2009, Leader *et al.*, 1997). Additionally, dicistronic tRNA-C/D box snoRNA genes are reported for plants (Kruszka *et al.*, 2003, Barbezier *et al.*, 2009). In vertebrates, only few genes are transcribed from independent promoters. In these organisms, most of the genes are located within introns of protein-coding and non-protein-coding genes. Usually, individual C/D box snoRNA genes are found in introns, but polycistrons were also found to exist (Weber, 2006, Tycowski *et al.*, 2004, Leader *et al.*, 1994, Pelczar & Filipowicz, 1998). Endo- and exoribonucleases are involved in the maturation of polycistronic C/D box snoRNA transcripts or intron-encoded C/D box snoRNAs. In yeast, polycistronic snoRNAs are flanked by sequences that form short hairpin structures and that are recognized by the RNaseIII-like endoribonuclease Rnt1 (Chanfreau *et al.*, 1998b, Chanfreau *et al.*, 1998a). Further trimming of the 5' and 3' ends occurs by 5'-3' and 3'-5' exoribonucleases that are not specific for the maturation of snoRNAs (Rat1, Xrn1; exosome) (Allmang *et al.*, 1999, Petfalski *et al.*, 1998, Qu *et al.*, 1999). Intron-encoded C/D box snoRNAs are matured by two different pathways of which one is splicing-dependent and the other is splicing-independent (Brown *et al.*, 2008). Predominantly, C/D box snoRNAs are processed from debranched introns by exoribonuclease activities (Villa *et al.*, 1998, Kiss & Filipowicz, 1995). Additionally, mature C/D box snoRNAs are produced by the action of endoribonucleases that cleave in the intronic sequences up- and downstream of the C/D box snoRNA and further trimming occurs by exoribonucleases. The second pathway seems to be important for snoRNA processing from introns in plants and yeast (Leader *et al.*, 1999, Villa et al., 1998).

In archaea, C/D box sRNA genes are most often located in intergenic regions or overlap with the 5' or 3' end of flanking open reading frames (ORFs) (Randau, 2012, Gaspin et al., 2000, Weisel *et al.*, 2010, Dennis *et al.*, 2001). Similar to eukaryotes, few polycistronic C/D box sRNA transcripts are described. In *Pyrococcus furiosus*, two clusters with two C/D box sRNA genes can be observed and in *Sulfolobus solfataricus*, one cluster of two C/D box sRNA genes exists (Gaspin et al., 2000, Dennis et al., 2001). Additionally, dicistronic tRNA-C/D box sRNA transcripts are reported and C/D box sRNAs can be found in the intron of pre-tRNA^{Trp} in several euryarchaeal organisms (Singh *et al.*, 2004, Clouet d'Orval *et al.*, 2001). A peculiar feature of C/D box sRNA maturation is the occurrence of circular archaeal C/D box sRNAs that are present in addition to linear molecules (Danan *et al.*, 2012, Starostina *et al.*, 2004, Su *et al.*, 2013, Randau, 2012).

In this study, we analyzed the localization and genomic context of C/D box sRNA genes in six archaeal model organisms. Co-transcriptional arrangements were identified which eliminate the need for independent promoters. Many C/D box sRNAs were found to overlap with flanking ORFs and C/D box

sRNA sequences can create 5' or 3'-UTRs. The effects of these C/D box sRNA gene fusions were tested in a reporter assay. Additionally, C/D box sRNA variants were analyzed *in vivo* in *Sulfolobus acidocaldarius* and revealed that the maturation of C/D box sRNAs occurs independent of the genetic context and relies solely on the integrity of the internal k-turn/k-loop elements.

Results

Promoter identification and genetic context of archaeal C/D box sRNA genes

In previous studies, C/D box sRNA genes did often not reveal independent promoters that are characterized by a TATA-box sequence in a distance of -26 +/- 3 bp upstream of the transcription start site (Starostina et al., 2004, Soppa, 1999, Randau, 2012). We recently described abundant C/D box sRNA production in six archaeal model organisms (*Methanococcus maripaludis*, *Nanoarchaeum equitans*, *S. acidocaldarius*, *Thermoproteus tenax*, *Methanopyrus kandleri*, *Ignicoccus hospitalis*) (Richter *et al.*, 2012, Randau, 2012, Dennis et al., 2015, Plagens *et al.*, 2014, Su et al., 2013). These data were utilized to analyze the genetic context of more than 300 identified C/D box sRNA genes and to search for promoters or conserved elements in their vicinity. The 50 nt upstream region of all C/D box sRNA genes was computationally screened for the presence of conserved motifs with a length of 6-15 nt. Three different conserved sequences were identified. These motifs all resemble TATA-box motifs and suggest the presence of promoters. However, only a minority of about 20 % of the C/D box sRNA genes were found to possess these elements within the analyzed 50 nt upstream region (Fig. 1; Table S2). Thus, the majority of C/D box sRNA genes do not contain promoters that would facilitate primary C/D box sRNA transcript production. Subsequently, the genetic context of the C/D box sRNA genes was analyzed to search for co-transcriptional arrangements. Strikingly, more than a quarter of the C/D box sRNA genes overlaps with the 3' end of flanking ORFs (Fig. 2A, B). These locations imply that extended transcripts exist and that the neighboring genes possess a 3'-UTR. The stop codons of the upstream genes are usually found within the C/D box sRNA sequences. In nearly half of the cases, the stop codons are situated within the boxC motif, but they can be also situated in the boxD', boxC' or boxD motifs (Fig. 2C). It should be noted, that the stop codon sequence “UGA” makes up most of the conserved boxC and boxD sequences and can evolve into a k-turn element with few nucleotide changes. In fewer cases, stop codons can be found within guide sequences, in the sequence between the boxC' and boxD' motif or in the sequence upstream of the boxC motif of the C/D box sRNAs. Some of these stop codon positions result in the occurrence of one or both guide regions within the open reading frame of the adjacent gene.

Only 7 % of the analyzed C/D box sRNA genes were found to overlap with the 5' end of flanking ORFs (Fig. 2B). The position of C/D box sRNA genes at the 5' end of flanking ORFs suggests that the transcripts of the adjacent genes possess a 5'-UTR. C/D box sRNAs that overlap with the 5' end of the flanking protein-coding regions contain the start codon of the coding region within their sequence. Most start codons are situated in the guide sequence or in the sequence downstream of the boxD motif (Fig. 2D). Additionally, the start codons can be found within all box motifs. The presence of this codon in the boxC, boxC' or boxD' motif, results in scenarios in which one or both guide sequences of the C/D box sRNAs are part of the neighboring coding region.

Another frequently observed organization of C/D box sRNA genes are clusters of two or three C/D box sRNA genes, suggesting their polycistronic transcription (Fig. 2B). This arrangement can be found multiple times in *S. acidocaldarius*, *M. kandleri* and *I. hospitalis.* In *T. tenax* and *I. hospitalis*, two postulated dicistronic tRNA-C/D box sRNA pairs exist. In *T. tenax*, the C/D box sRNA *TtesR134* gene is located directly downstream of a $tRNA^{Pro}$ gene and the tRNA 3' processing activity of RNaseZ would create a mature C/D box sRNA 5' terminus. In *I. hospitalis* the C/D box sRNA *IhosR11* gene is located downstream of a $tRNA^{Ser}$ gene, separated by a 14 nt spacer.

The majority of the remaining C/D box sRNA genes are located in intergenic regions and some of them possess their own promoter. In most cases, only few nucleotides separate the C/D box sRNA and the adjacent protein-coding gene which suggests the presence of UTRs. For example, 15 % of C/D box sRNA genes are located upstream or downstream of the neighboring protein-coding gene and the distance between both genes is less than 25 nt. In conclusion, the analysis of the genetic context of C/D box sRNA genes indicates that they often do not need to possess dedicated promoters to ensure their transcription. Instead, most C/D box sRNAs appear to be parts of longer co-transcriptional precursors that require processing to obtain individual mature C/D box sRNAs.

Identification of C/D box sRNA precursors

Next, we aimed to verify the presence of C/D box sRNA precursors. The investigated C/D box sRNAs were identified via RNA-Seq and careful analysis of the coverage plots revealed sequence reads that span the flanking C/D box sRNA gene regions. This observation is a first indication for the presence of longer precursor transcripts. Therefore, we employed Northern blot and RT-PCR to search for co-transcribed products. C/D box sRNAs in 5' and 3'-UTRs of flanking coding regions and dicistronic C/D box sRNA transcripts were identified for three respective example cases in *S. acidocaldarius* (Fig. 3). Northern blot analyses with probes against both individual RNAs of the putative C/D box sRNA Sac-sR125 and Sac-sR126 dicistron reveal two major bands. The size of the lower band corresponds to

the size of a single C/D box sRNA and the size of the upper band corresponds to the size of the C/D box sRNA dicistron. RT-PCR analyses verified the existence of dicistronic C/D box sRNAs.

The C/D box sRNA *Sac-sR121* gene is located directly upstream of *saci_1247* and the C/D box sRNA *Sac-sR24* gene is located directly downstream of *saci_0125*. RT-PCR analyses with one primer binding within the C/D box sRNA and the other one within the flanking mRNA portion revealed that shared transcripts exist. Thus, C/D box sRNAs can be located in UTRs. 3'-UTRs are common in archaea but 69 % of the *S. acidocaldarius* transcripts are leaderless, i.e. they lack a 5'-UTR (Brenneis *et al.*, 2007, Slupska *et al.*, 2001). However, several longer 5'-UTRs exist and approximately 6 % of the transcripts possess 5'-UTRs that are larger than 20 nucleotides (Wurtzel *et al.*, 2010).

Influence of C/D box sRNA sequences in the UTRs of adjacent genes

We investigated the effect of C/D box sRNA sequences in UTRs of fused genes. Therefore, plasmid-based reporter gene assays were established in *S. acidocaldarius* in which various C/D box sRNA genes were fused to the β-galactosidase gene (*lacS*) mimicking the genomic position of the naturally occurring neighboring gene. The transcription of the constructs was controlled by an inducible maltodextrin promoter and the β-galactosidase activity was measured following ONPG (o-nitrophenyl-β-D-galactopyranosid) hydrolysis.

Six *S. acidocaldarius* C/D box sRNA genes with a natural genomic localization upstream of protein-coding genes with 0 to 11 nt linkers were fused to the 5' end of lacS. Original linker sequences were maintained. A strong reduction of enzyme activity was observed in all cases in comparison to a lacS control without fused C/D box sRNA (Fig. 4A). For five of the six constructs, the enzyme activity was less than 20 % of the lacS control activity. Quantitative real-time PCR (qRT-PCR) analyses revealed that this loss of activity correlates with a reduced number of lacS fusion transcripts (Fig. S1).

In addition, four C/D box sRNA genes were investigated that are located at the 3' end of neighboring genes. The C/D box sRNA *Sac-sR24* gene is located directly downstream and the *Sac-sR106* gene is separated by 6 nts from a neighboring gene. Two C/D box sRNA genes, *Sac-sR2* and *Sac-sR101*, show a 9 nt overlap with protein-coding genes. In these cases, some nucleotides were exchanged at the terminus of lacS to accommodate the C/D box sRNA sequence. To exclude that these changes affect β-galactosidase activity, lacS controls without C/D box sRNA fusions were designed with identical nucleotide changes. Reporter assays revealed that these mutations have a minimal effect on activity. The presence of C/D box sRNAs in the 3'-UTRs was found to cause either neutral or negative effects on β-galactosidase activity. However, the reduction of activity was not as drastic as the one observed

for C/D box sRNA sequences in 5'-UTRs (Fig. 4B). Again, qRT-PCR analyses showed that this loss of activity correlates with reduced amounts of lacS fusion transcripts (Fig. S1).

In summary, these reporter assays suggest that C/D box sRNA gene sequences do not provide a benefit for their co-transcribed genes and 3'-terminal fusion appears to be less harmful. These observations are in agreement with the identified enrichment of C/D box sRNAs in 3'-UTRs and the predominance of fusions with genes whose annotation does not suggest essentiality.

Determination of C/D box sRNA maturation requirements

Archaeal C/D box sRNAs were found to be transcribed in operons or in UTRs of protein-coding genes. Therefore, we aimed to determine the sequence elements required for the maturation of individual, functional C/D box sRNAs. The 5' and 3' ends of C/D box sRNAs show a length heterogeneity of several nucleotides, suggesting exoribonucleolytic trimming (Dennis et al., 2015). We designed an assay that follows the production of a plasmid-encoded synthetic C/D box sRNA (based on Sac-sR121) *in vivo* in *S. acidocaldarius*. The C/D box sRNA *Sac-sR121* gene is located directly upstream of the 5'-methylthioadenosine phosphorylase gene *saci_1247*. It was cloned with 50 bp of the native upstream and downstream sequence behind an inducible maltodextrin promoter. The D' guide was exchanged against a unique sequence, resulting in the synthetic C/D box sRNA Sac-sR121*. This RNA can be detected via Northern blot analyses and distinguished from native C/D box sRNA Sac-sR121 transcripts (Fig. 5A). The assays allowed for the introduction of changes in the C/D box sRNA gene and in the up- and downstream regions to analyze their importance for maturation.

Total RNA was isolated from *S. acidocaldarius* cells harboring the C/D box sRNA Sac-sR121* *in vivo* constructs and a 16 nt long LNA probe was used to hybridize with the boxC sequence and the artificial guide sequence. These Northern Blot analyses revealed that a TATA-box in the native 50 bp upstream region is recognized and promotes transcription (Fig. 5B). Therefore, the TATA-box sequence was exchanged against a GC-containing sequence to facilitate utilization of an inducible promoter that generates a single transcription start site (Fig. 5A).

Northern blot experiments revealed a single band corresponding to the size of a mature C/D box sRNA. Inverse RT-PCR analyses and sequencing of the amplified products was used to verify maturation of C/D box Sac-sR121* and also revealed the existence of circular molecules. Circular C/D box sRNA versions were previously observed in several archaeal model organisms, but their role is not known (Table S1) (Danan et al., 2012, Starostina et al., 2004, Su et al., 2013, Randau, 2012). The sequencing result highlights that extra sequences upstream of the C/D box sRNA (created by the transcription start site of the inducible promoter) were removed. The 3' terminus might have been defined by termination at the C/D box sRNA 3' end or generated via the trimming of longer

transcripts. To test this option, the downstream region was exchanged against a sequence that is able to form a stable RNA hairpin. Only in this construct a longer transcript is detected in the Northern Blot analysis, which argues that precursor transcripts extend beyond the C/D box sRNA end and that the introduced hairpin protects precursors from exoribonucleolytic degradation. Additionally, minor amounts of mature C/D box sRNA transcripts were detected, which likely represent the instability of the unprotected ssRNA between the C/D box sRNA and the hairpin (Fig. 5B).

Next, we assayed if upstream and downstream sequences contain motifs that are required for efficient C/D box sRNA processing. Interestingly, completely randomized 50 bp upstream and downstream sequences still yielded a defined mature C/D box sRNA Sac-sR121* product (Fig. 5B). This observation is in agreement with the lack of conserved sequence elements in the vicinity of native C/D box sRNA genes. Therefore, we asked what is needed to generate the defined length of a functional C/D box sRNA and argue that any signal would need to be located within the boxC/C' or boxD/D' motifs or the k-turn/k-loop structures as these are the only sequences and structures that are conserved within all C/D box sRNAs. Mutations were introduced into these conserved sequences and the production of Sac-sR121* variants with altered k-turn/k-loop structures was assayed via Northern blot. It was found that all C/D box sRNA Sac-sR121* constructs harboring either single boxD mutations or mutations of the two GA base pairs involved in k-loop formation do not show any signal in the Northern blot analyses (Fig. 5C). The only mutations that were not detrimental disrupt the Watson-Crick base pair or the mismatch base pair involved in k-loop formation and show mature C/D box sRNA transcripts (Fig. 5C). The acceptance of variations in the boxC' and boxD' motifs that are involved in k-loop formation fits to the observed higher degree of conservation of boxC and boxD elements among the 61 *S. acidocaldarius* C/D box sRNAs. However, the nucleotides of the boxC' and boxD' motifs that form the two GA base pairs in the k-loop are also highly conserved (Dennis et al., 2015). Northern Blot signals were not observed for C/D box sRNA Sac-sR121* k-turn mutants and GA base pair mutants of the k-loop, which might indicate that k-turn and k-loop formation is crucial for C/D box sRNA stability. Mutations of the k-loop and k-turn abolish L7Ae binding which prevents the formation of the active ribonucleoprotein complex stabilizing the internal C/D box sRNA. It is possible that k-loop mutations are more likely to be tolerated due to the closer proximity of boxC' and boxD' motifs in the primary sequence.

In summary, only the presence of the conserved internal box sequences was found to be required for the generation of mature Sac-sR121* and potential endoribonuclease cleavage signals were not detected in surrounding sequences. This scenario allows for the observed highly variable genetic context of C/D box sRNA genes and enables them to “hijack” external promoters in co-transcriptional arrangements.

Discussion

C/D box sRNAs are highly abundant in thermophilic archaea and diverse sets of C/D box sRNA molecules guarantee extensive RNA modification patterns required for the folding and stabilization of RNA molecules. Detailed analyses of the genetic context of the C/D box sRNA genes revealed that most of them do not exhibit promoters for primary transcript production but are rather co-transcribed as parts of 5'-or 3'-UTRs of adjacent genes. Most archaeal transcripts are leaderless and lack a 5'-UTR (Slupska et al., 2001, Brenneis et al., 2007, Wurtzel et al., 2010, Cohen *et al.*, 2016). The effect of C/D box sRNAs within UTRs on the expression of the neighboring genes is not clear. Artificial 5' UTRs with random sequences were tested in a reporter assay in *Haloferax volcanii*. Translational efficiencies of the transcripts were similar or higher as the leaderless control (Brenneis et al., 2007). In contrast, L7Ae binding to k-turns in the 5' UTR was used as translational block in eukaryotes (Saito *et al.*, 2010). Similarly, our C/D box sRNA-reporter gene fusions revealed that C/D box sRNA genes that were fused to the 5' end of the reporter gene negatively influence expression and enzyme activity. Thus, these cases suggest a scenario of selfish C/D box sRNA genes that “hijack” a promoter which ensures their transcription at the expense of the downstream genes. Most archaeal C/D box sRNA genes were found to be fused to the 3' terminus of upstream genes. We hypothesized that these fusions could provide a stabilizing effect on the attached mRNA but reporter assays revealed a neutral effect at best. Thus, we do not have any indications that genes have a clear benefit from a fusion or an overlap with C/D box sRNA genes. In contrast, C/D box sRNA genes arise in genetic contexts that guarantee their production as parts of precursor transcripts that require further trimming. How did this C/D box sRNA localization evolve? It was suggested for eukaryotes that snoRNA genes evolve faster than protein-coding genes as their sequence is more variable. An open reading frame does not have to be maintained and the function of individual C/D box snoRNAs does not seem to be essential. The only sequences that are required for stability and functionality are the boxC/C' and boxD/D' motifs (Brown et al., 2003). Interestingly, our data show that many of the C/D box sRNA genes possess the sequence of the start or stop codon of an overlapped gene within their boxC/C' or boxD/D' motif. Thus, it is a very intriguing possibility that the start and stop codons of genes can accelerate the evolution of k-turn motifs within C/D box sRNA genes.

The diverse genetic contexts of archaeal C/D box sRNA genes imply that a universal maturation mechanism must exist which tolerates C/D box sRNA precursors that result from transcription initiated by adopted and often distant promoters. Our analyses of the production of a synthetic C/D box sRNA *in vivo* in *S. acidocaldarius* revealed that random upstream or downstream sequences are sufficient for efficient C/D box sRNA maturation. Consequently, sequence- or structure-specific endoribonucleases that recognize motifs outside of the C/D box sRNA sequence are not involved in

their maturation process. Recently, relaxed structure motifs for the tRNA splicing endonuclease were identified computationally in the vicinity of some C/D box sRNA genes (Berkemer *et al.*, 2015). Our results indicate that C/D box maturation is feasible without the presence of splicing endonuclease motifs. Instead, precursor-processing by unspecific exoribonucleases is plausible, which highlights that the archaeal C/D box sRNA processing pathway is similar to eukaryotic C/D box snoRNA biogenesis (Fig. 6). The maturation and final trimming of eukaryotic snoRNAs includes the activity of exoribonucleases (Allmang et al., 1999, Petfalski et al., 1998, Qu et al., 1999). However, the 5'-3' exoribonucleases that were identified e.g. in yeast cannot be identified in archaea (Hasenohrl *et al.*, 2011). Archaeal homologues of the bacterial RNaseJ family, exhibiting 5'-3' exoribonucleolytic activity, were identified, but their involvement in C/D box sRNA maturation has not been proven (Hasenohrl et al., 2011, Clouet-d'Orval *et al.*, 2010, Clouet-d'Orval *et al.*, 2015, Martens *et al.*, 2013). The processing of the 3' end of eukaryotic C/D box snoRNA involves exosome activity, which would also be plausible for archaeal C/D box sRNA maturation (Mitchell *et al.*, 1996, Mitchell *et al.*, 1997, Allmang et al., 1999).

Exoribonucleolytic trimming of C/D box sRNA precursors requires protection of the functional sequence from degradation. Our results underline that the integrity of the k-turn structure is crucial for C/D box sRNA maintenance in *S. acidocaldarius*. In contrast, the consensus sequence of the boxC' and boxD' motifs of the k-loop structure is more relaxed and only the two sheared GA base pairs were found to be crucial. This is in agreement with their previously observed importance for k-loop-mediated L7Ae binding (Nolivos et al., 2005). Elegant *in vitro* C/D box sRNP *in vitro* assembly approaches revealed that the C/D box sRNA internal loop is crucial for the formation of native dimeric complexes (Bleichert *et al.*, 2009, Bower-Phipps *et al.*, 2012, Lin *et al.*, 2011, Lapinaite et al., 2013, Yip et al., 2016).

The binding of L7Ae to the k-turn should be crucial for protecting the 5' and 3' ends of the RNAs. Circularization of C/D box sRNAs was described for several archaea and it was postulated that this represents an additional protection mechanism for 5' and 3' ends as archaeal C/D box sRNAs often lack the terminal stem that is typical for eukaryotic C/D box snoRNAs (Omer et al., 2000, Su et al., 2013, Randau, 2012, Gaspin et al., 2000, Starostina et al., 2004). It was shown for eukaryotic snoRNAs that the snoRNP proteins associate with the snoRNA before processing occurs and therefore constitute a processing barrier (Caffarelli *et al.*, 1996, Matera *et al.*, 2007). We hypothesize that a similar mechanism exists for archaeal C/D box sRNAs that prevents overtrimming of the termini (Fig. 6).

In conclusion, our results highlight the transcriptional plasticity of C/D box sRNA genes that is accompanied by a C/D box sRNA maturation pathway that is solely dependent on internal RNA

binding signals. These features allow for the accelerated evolution of C/D box sRNA genes which can take advantage of the incorporation of existing stop codons in their conserved box sequences. Consequently, the RNA modification landscape of an archaeal population can adapt fast to environmental challenges.

Experimental procedures

Strains and growth conditions

Sulfolobus acidocaldarius MW001 and strains containing the constructed plasmids were grown aerobically in Brock media (Brock *et al.*, 1972) with a pH of 3.5 at 75°C as described (Wagner *et al.*, 2012). The media were supplemented with 0.1 % (w/v) tryptone as well as 0.2 % (w/v) sucrose for normal growth or 0.2 % (w/v) dextrin for induction of the maltodextrin (mal) promoter. The growth of the cells was monitored by measurement of the optical density at 600 nm. Cell harvesting for total RNA preparations and ONPG-assays was performed at OD 0.4-0.6. For solid media the medium was supplemented with 1.2 % gelrite.

Escherichia coli DH5α and ER1821 bearing the plasmid pM.EsaBC4I were used for cloning and methylation of plasmid DNA. The strains were grown in Lysogeny broth medium supplemented with the appropriate antibiotics at 37°C and 200 rpm.

Plasmid construction and transformation

To obtain strains overexpressing C/D box sRNA gene variants and C/D box sRNA-lacS fusions, the plasmid pSVA1431, containing a maltose inducible promoter, was utilized. This plasmid is described as pCmalLacS in Berkner *et al.* 2010 (Berkner *et al.*, 2010) but the distance between the TATA box and the transcription start site is 2 bp shorter in pSVA1431 (Sonja-Verena Albers, Albert-Ludwigs-University Freiburg, personal communication). All constructs were cloned behind the mal promoter via the NcoI and EagI restriction sites (Table S1). For C/D box sRNA fusions to the 5' end of lacS, the NcoI restriction site was exchanged to a KasI restriction site in the plasmid to prevent an artificial translation start by the 'ATG' from the NcoI restriction site. Methylated plasmids were electroporated into competent *S. acidocaldarius* MW001 cells using the Gene Pulser ® Electroporation System (BioRad) with the input parameters 1500 V, 600 Ω, 25 µF (Wagner et al., 2012).

Prediction of transcription initiation signals

The 50 nt upstream sequences of 343 C/D box sRNA genes from six archaea were analyzed for conserved sequences and structures. The C/D box sRNAs were identified in RNA-Seq analyses in previous studies (Table S2) (Richter et al., 2012, Randau, 2012, Plagens et al., 2014, Su et al., 2013, Dennis et al., 2015). To check for sequence conservation, sequences were clustered into related families based on a similarity matrix using Markov clustering with reasonable cutoff (80 %) (Enright *et al.*, 2002). Afterwards, multiple sequence alignments were generated using MAFFT (Katoh *et al.*, 2002) and regions with at least 85 % sequence conservation and a length of 6 to 15 nucleotides were extracted. Finally, sequence logos were generated using WebLogo (Crooks *et al.*, 2004). For the identification of conserved structure motifs hierarchical cluster trees were generated using RNAclust which reflected sequence and structure similarity which was given by LocARNA (Will *et al.*, 2007, Will *et al.*, 2012). For each node in the cluster tree consensus structures were searched.

RT-PCR analyses

Reverse transcription PCR (RT-PCR) was used to analyze dicistronic C/D box sRNA and C/D box sRNA/mRNA transcripts. Total RNA was isolated using the TRIzol reagent (Life Technologies) and treated with DNaseI (Thermo Fisher Scientific) following the manufacturer's instructions. 450 ng of DNaseI-treated RNA was reverse transcribed using gene specific primers and Superscript III reverse transcriptase (Life Technologies) using the manufacturer's instruction. A negative control was performed without the addition of reverse transcriptase to check for DNA contamination. PCR amplification was performed with Taq polymerase and gene-specific primers binding in the C/D box sRNA gene and the neighboring gene or within two neighboring C/D box sRNA genes, respectively (Table S1).

Inverse RT-PCR was performed to determine the sequences of the synthetic C/D box sRNAs as a fraction of archaeal C/D box sRNAs always exists in a circular variant (Su et al., 2013, Danan et al., 2012, Starostina et al., 2004). Total RNA was isolated using the TRIzol reagent (Life Technologies) and 1 µg of the total RNA preparation was used for the reverse transcription reaction with gene specific primers (Table S1) and Superscript III reverse transcriptase (Life Technologies) according to the manufacturer's instruction. PCR amplification was performed with Taq polymerase and outward-facing C/D box sRNA gene-specific primers. The amplification products were cloned with the Topo ® TA cloning kit (Thermo Fisher Scientific) and sequenced.

Northern blot analyses

Total RNA was isolated using the TRIzol reagent (Life Technologies) according to the manufacturer's instruction. Electrophoresis of 10 µg of total RNA was performed in 8 % denaturing polyacrylamide

gels. A semi dry electrophoretic transfer system was used to transfer the RNA onto a positively charged nylon membrane. The RNA was immobilized by UV crosslinking to the membrane. Pre-hybridization was performed in DIG Easy Hyb buffer (Roche) for 30 min at 42°C and subsequently hybridization was achieved over night at 42°C with radiolabeled probes complementary to the designed guide region and native C/D box sRNAs, respectively (Table S1). The membrane was washed two times (2xSSC, 0.1 % SDS and 1xSSC, 0.1 % SDS) and the detection of radioactivity was carried out by phosphorimaging.

β-galactosidase assays with ONPG

The β-galactosidase activity was determined as described (Wagner et al., 2012). Briefly, 2 ml of the *S. acidocaldarius* cultures were resuspended in Z-buffer (10 mM KCl, 1 mM $MgSO_4$, 60 mM Na_2HPO_4, 40 mM NaH_2PO_4 [pH 7]) containing 1 mM PMSF and 0.5 % Triton X-100 to an OD_{600nm} of 3.2. The β-galactosidase enzymatic activity was measured at 42°C in a 200 µl reaction which consisted of 170 µl Z-buffer, 20 µl cell suspension and 10 µl ONPG solution (12 mg ONPG in 1 ml Z-buffer). The ONPG hydrolysis was measured at 410 nm for 3 h in 5 min intervals in a microplate reader. Additionally, the protein concentration was determined using the Bio-Rad protein assay based on the Bradford protein quantitation method (Bradford, 1976) according to the manufacturer's instructions and a calibration curve created with bovine serum albumine (BSA) dilutions.

For the quantification of the β-galactosidase activity the following equation was used (Miller, 1972, Wagner *et al.*, 2014):

$$Miller = \frac{60{,}000*\left(A_{410(t2\text{-}t1)}\text{- autolysis } A_{410(t2\text{-}t1)}\right)*7}{time\ (s) * volume\ of\ the\ sample\ (ml) * protein\ concentration\ (mg/ml)}$$

Quantitative RT-PCR

Total RNA of was isolated using the TRIzol reagent (Life Technologies) and treated with DNaseI (Thermo Fisher Scientific) following the manufacturer's instructions. 1 µg of DNaseI-treated RNA was reverse transcribed using random hexamer primers and Superscript III reverse transcriptase (Life Technologies) using the manufacturer's instruction. Real-time qPCR was carried out in a Bio-Rad CFX Connect Real-Time System and each reaction contained 10 µl iQ SYBR Green Supermix (Bio-Rad), 100 nM of two specific primers (see Table S1) and cDNA equivalent to 10 ng total RNA in a total volume of 20 µl. The amplification protocol includes an initial denaturation step at 95°C for 3 min, 40 cycles of 95°C for 15 s, 60°C for 20 s, 72°C for 20 s and a final denaturation step at 95°C for 15 s. A melting curve analysis demonstrated the absence of primer-dimer formation. The assays were performed

with biological duplicates, technical duplicates, a no-template control and no-RT controls. The expression of *secY* (*saci0574*) was used for normalization (Reimann *et al.*, 2013). Quantification cycles (C_qs) were determined with Bio-Rad CFX manager software and mean relative gene expression ratios including standard deviations were calculated using the $2(-\Delta\Delta C_t)$ method (Livak & Schmittgen, 2001).

Acknowledgements

We thank Sonja-Verena Albers for advice and kindly providing plasmid pSVA1431. Funding was provided by the Deutsche Forschungsgemeinschaft (DFG), project RA 2169/3-1.

Author contribution

VT, RM and AO performed the analyses of the genetic context and the *in vivo* experiments, designed experiments and wrote the manuscript. OSA and RB performed computational analyses for the identification of promoters. LR designed experiments and wrote the manuscript.

References

Aittaleb, M., R. Rashid, Q. Chen, J.R. Palmer, C.J. Daniels & H. Li, (2003) Structure and function of archaeal box C/D sRNP core proteins. *Nature structural biology* **10**: 256-263.

Allmang, C., J. Kufel, G. Chanfreau, P. Mitchell, E. Petfalski & D. Tollervey, (1999) Functions of the exosome in rRNA, snoRNA and snRNA synthesis. *The EMBO journal* **18**: 5399-5410.

Balakin, A.G., L. Smith & M.J. Fournier, (1996) The RNA world of the nucleolus: two major families of small RNAs defined by different box elements with related functions. *Cell* **86**: 823-834.

Barbezier, N., G. Canino, J. Rodor, E. Jobet, J. Saez-Vasquez, A. Marchfelder & M. Echeverria, (2009) Processing of a dicistronic tRNA-snoRNA precursor: combined analysis *in vitro* and *in vivo* reveals alternate pathways and coupling to assembly of snoRNP. *Plant physiology* **150**: 1598-1610.

Berkemer, S.J., C. Höner zu Siederdissen, F. Amman, A. Wintsche, S. Will, I. Hofacker, S.J. Prohaska & P.F. Stadler, (2015) Processed Small RNAs in Archaea and BHB Elements. *Genomics and Computational Biology* **1**: e18.

Berkner, S., A. Wlodkowski, S.V. Albers & G. Lipps, (2010) Inducible and constitutive promoters for genetic systems in *Sulfolobus acidocaldarius*. *Extremophiles : life under extreme conditions* **14**: 249-259.

Bleichert, F., K.T. Gagnon, B.A. Brown, 2nd, E.S. Maxwell, A.E. Leschziner, V.M. Unger & S.J. Baserga, (2009) A dimeric structure for archaeal box C/D small ribonucleoproteins. *Science* **325**: 1384-1387.

Bower-Phipps, K.R., D.W. Taylor, H.W. Wang & S.J. Baserga, (2012) The box C/D sRNP dimeric architecture is conserved across domain Archaea. *RNA* **18**: 1527-1540.

Bradford, M.M., (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry* **72**: 248-254.

Brenneis, M., O. Hering, C. Lange & J. Soppa, (2007) Experimental characterization of Cis-acting elements important for translation and transcription in halophilic archaea. *PLoS genetics* **3**: e229.

Brock, T.D., K.M. Brock, R.T. Belly & R.L. Weiss, (1972) Sulfolobus: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature. *Archiv fur Mikrobiologie* **84**: 54-68.

Brown, J.W., M. Echeverria & L.H. Qu, (2003) Plant snoRNAs: functional evolution and new modes of gene expression. *Trends in plant science* **8**: 42-49.

Brown, J.W., D.F. Marshall & M. Echeverria, (2008) Intronic noncoding RNAs and splicing. *Trends in plant science* **13**: 335-342.

Caffarelli, E., A. Fatica, S. Prislei, E. De Gregorio, P. Fragapane & I. Bozzoni, (1996) Processing of the intron-encoded U16 and U18 snoRNAs: the conserved C and D boxes control both the processing reaction and the stability of the mature snoRNA. *The EMBO journal* **15**: 1121-1131.

Chanfreau, G., P. Legrain & A. Jacquier, (1998a) Yeast RNase III as a key processing enzyme in small nucleolar RNAs metabolism. *Journal of molecular biology* **284**: 975-988.

Chanfreau, G., G. Rotondo, P. Legrain & A. Jacquier, (1998b) Processing of a dicistronic small nucleolar RNA precursor by the RNA endonuclease Rnt1. *The EMBO journal* **17**: 3726-3737.

Clouet-d'Orval, B., D.K. Phung, P.S. Langendijk-Genevaux & Y. Quentin, (2015) Universal RNA-degrading enzymes in Archaea: Prevalence, activities and functions of beta-CASP ribonucleases. *Biochimie* **118**: 278-285.

Clouet-d'Orval, B., D. Rinaldi, Y. Quentin & A.J. Carpousis, (2010) Euryarchaeal beta-CASP proteins with homology to bacterial RNase J Have 5'- to 3'-exoribonuclease activity. *The Journal of biological chemistry* **285**: 17574-17583.

Clouet d'Orval, B., M.L. Bortolin, C. Gaspin & J.P. Bachellerie, (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 research* **29**: 4518-4529.

Cohen, O., S. Doron, O. Wurtzel, D. Dar, S. Edelheit, I. Karunker, E. Mick & R. Sorek, (2016) Comparative transcriptomics across the prokaryotic tree of life. *Nucleic acids research*. **44**(W1):W46-53

Crooks, G.E., G. Hon, J.M. Chandonia & S.E. Brenner, (2004) WebLogo: a sequence logo generator. *Genome research* **14**: 1188-1190.

Danan, M., S. Schwartz, S. Edelheit & R. Sorek, (2012) Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic acids research* **40**: 3131-3142.

Dennis, P.P., A. Omer & T. Lowe, (2001) A guided tour: small RNA function in Archaea. *Molecular microbiology* **40**: 509-519.

Dennis, P.P., V. Tripp, L. Lui, T. Lowe & L. Randau, (2015) C/D box sRNA-guided 2'-O-methylation patterns of archaeal rRNA molecules. *BMC genomics* **16**: 632.

Dieci, G., M. Preti & B. Montanini, (2009) Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics* **94**: 83-88.

Enright, A.J., S. Van Dongen & C.A. Ouzounis, (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic acids research* **30**: 1575-1584.

Gaspin, C., J. Cavaille, G. Erauso & J.P. Bachellerie, (2000) Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the *Pyrococcus* genomes. *Journal of molecular biology* **297**: 895-906.

Hasenohrl, D., R. Konrat & U. Blasi, (2011) Identification of an RNase J ortholog in *Sulfolobus solfataricus*: implications for 5'-to-3' directional decay and 5'-end protection of mRNA in Crenarchaeota. *RNA* **17**: 99-107.

Helm, M., (2006) Post-transcriptional nucleotide modification and alternative folding of RNA. *Nucleic acids research* **34**: 721-733.

Herschlag, D., F. Eckstein & T.R. Cech, (1993) The importance of being ribose at the cleavage site in the *Tetrahymena* ribozyme reaction. *Biochemistry* **32**: 8312-8321.

Katoh, K., K. Misawa, K. Kuma & T. Miyata, (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* **30**: 3059-3066.

Kiss-Laszlo, Z., Y. Henry, J.P. Bachellerie, M. Caizergues-Ferrer & T. Kiss, (1996) Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell* **85**: 1077-1088.

Kiss-Laszlo, Z., Y. Henry & T. Kiss, (1998) Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. *The EMBO journal* **17**: 797-807.

Kiss, T. & W. Filipowicz, (1995) Exonucleolytic processing of small nucleolar RNAs from pre-mRNA introns. *Genes & development* **9**: 1411-1424.

Klein, D.J., T.M. Schmeing, P.B. Moore & T.A. Steitz, (2001) The kink-turn: a new RNA secondary structure motif. *The EMBO journal* **20**: 4214-4221.

Kruszka, K., F. Barneche, R. Guyot, J. Ailhas, I. Meneau, S. Schiffer, A. Marchfelder & M. Echeverria, (2003) Plant dicistronic tRNA-snoRNA genes: a new mode of expression of the small nucleolar RNAs processed by RNase Z. *The EMBO journal* **22**: 621-632.

Kuhn, J.F., E.J. Tran & E.S. Maxwell, (2002) Archaeal ribosomal protein L7 is a functional homolog of the eukaryotic 15.5kD/Snu13p snoRNP core protein. *Nucleic acids research* **30**: 931-941.

Lapinaite, A., B. Simon, L. Skjaerven, M. Rakwalska-Bange, F. Gabel & T. Carlomagno, (2013) The structure of the box C/D enzyme reveals regulation of RNA methylation. *Nature* **502**: 519-523.

Leader, D.J., G.P. Clark, J. Watters, A.F. Beven, P.J. Shaw & J.W. Brown, (1997) Clusters of multiple different small nucleolar RNA genes in plants are expressed as and processed from polycistronic pre-snoRNAs. *The EMBO journal* **16**: 5742-5751.

Leader, D.J., G.P. Clark, J. Watters, A.F. Beven, P.J. Shaw & J.W. Brown, (1999) Splicing-independent processing of plant box C/D and box H/ACA small nucleolar RNAs. *Plant Molecular Biology* **39**: 1091-1100.

Leader, D.J., J.F. Sanders, R. Waugh, P. Shaw & J.W. Brown, (1994) Molecular characterisation of plant U14 small nucleolar RNA genes: closely linked genes are transcribed as polycistronic U14 transcripts. *Nucleic acids research* **22**: 5196-5203.

Li, S.G., H. Zhou, Y.P. Luo, P. Zhang & L.H. Qu, (2005) Identification and functional analysis of 20 Box H/ACA small nucleolar RNAs (snoRNAs) from *Schizosaccharomyces pombe*. *The Journal of biological chemistry* **280**: 16446-16455.

Liang, D., H. Zhou, P. Zhang, Y.Q. Chen, X. Chen, C.L. Chen & L.H. Qu, (2002) A novel gene organization: intronic snoRNA gene clusters from *Oryza sativa*. *Nucleic acids research* **30**: 3262-3272.

Lin, J., S. Lai, R. Jia, A. Xu, L. Zhang, J. Lu & K. Ye, (2011) Structural basis for site-specific ribose methylation by box C/D RNA protein complexes. *Nature* **469**: 559-563.

Livak, K.J. & T.D. Schmittgen, (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**: 402-408.

Maden, B.E., M.E. Corbett, P.A. Heeney, K. Pugh & P.M. Ajuh, (1995) Classical and novel approaches to the detection and localization of the numerous modified nucleotides in eukaryotic ribosomal RNA. *Biochimie* **77**: 22-29.

Martens, B., F. Amman, S. Manoharadas, L. Zeichen, A. Orell, S.V. Albers, I. Hofacker & U. Blasi, (2013) Alterations of the transcriptome of *Sulfolobus acidocaldarius* by exoribonuclease aCPSF2. *PLoS One* **8**: e76569.

Matera, A.G., R.M. Terns & M.P. Terns, (2007) Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nature reviews. Molecular cell biology* **8**: 209-220.

Maxwell, E.S. & M.J. Fournier, (1995) The small nucleolar RNAs. *Annual Review of Biochemistry* **64**: 897-934.

Miller, J.H., (1972) Experiments in Molecular Genetics. *New York: Cold Spring Harbor Laboratory*.

Mitchell, P., E. Petfalski, A. Shevchenko, M. Mann & D. Tollervey, (1997) The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'-->5' exoribonucleases. *Cell* **91**: 457-466.

Mitchell, P., E. Petfalski & D. Tollervey, (1996) The 3' end of yeast 5.8S rRNA is generated by an exonuclease processing mechanism. *Genes & development* **10**: 502-513.

Nolivos, S., A.J. Carpousis & B. Clouet-d'Orval, (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 research* **33**: 6507-6514.

Omer, A.D., T.M. Lowe, A.G. Russell, H. Ebhardt, S.R. Eddy & P.P. Dennis, (2000) Homologs of small nucleolar RNAs in Archaea. *Science* **288**: 517-522.

Pelczar, P. & W. Filipowicz, (1998) The host gene for intronic U17 small nucleolar RNAs in mammals has no protein-coding potential and is a member of the 5'-terminal oligopyrimidine gene family. *Molecular and cellular biology* **18**: 4509-4518.

Petfalski, E., T. Dandekar, Y. Henry & D. Tollervey, (1998) Processing of the precursors to small nucleolar RNAs and rRNAs requires common components. *Molecular and cellular biology* **18**: 1181-1189.

Plagens, A., V. Tripp, M. Daume, K. Sharma, A. Klingl, A. Hrle, E. Conti, H. Urlaub & L. Randau, (2014) *In vitro* assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex. *Nucleic acids research* **42**: 5125-5138.

Polikanov, Y.S., S.V. Melnikov, D. Soll & T.A. Steitz, (2015) Structural insights into the role of rRNA modifications in protein synthesis and ribosome assembly. *Nature structural & molecular biology* **22**: 342-344.

Qu, L.H., A. Henras, Y.J. Lu, H. Zhou, W.X. Zhou, Y.Q. Zhu, J. Zhao, Y. Henry, M. Caizergues-Ferrer & J.P. Bachellerie, (1999) Seven novel methylation guide small nucleolar RNAs are processed

from a common polycistronic transcript by Rat1p and RNase III in yeast. *Molecular and cellular biology* **19**: 1144-1158.

Randau, L., (2012) RNA processing in the minimal organism *Nanoarchaeum equitans*. *Genome biology* **13**: R63.

Reimann, J., D. Esser, A. Orell, F. Amman, T.K. Pham, J. Noirel, A.C. Lindas, R. Bernander, P.C. Wright, B. Siebers & S.V. Albers, (2013) Archaeal signal transduction: impact of protein phosphatase deletions on cell size, motility, and energy metabolism in *Sulfolobus acidocaldarius*. *Molecular & cellular proteomics* **12**: 3908-3923.

Richter, H., J. Zoephel, J. Schermuly, D. Maticzka, R. Backofen & L. Randau, (2012) Characterization of CRISPR RNA processing in *Clostridium thermocellum* and *Methanococcus maripaludis*. *Nucleic acids research* **40**: 9887-9896.

Saito, H., T. Kobayashi, T. Hara, Y. Fujita, K. Hayashi, R. Furushima & T. Inoue, (2010) Synthetic translational regulation by an L7Ae-kink-turn RNP switch. *Nature chemical biology* **6**: 71-78.

Singh, S.K., P. Gurha, E.J. Tran, E.S. Maxwell & R. Gupta, (2004) Sequential 2'-O-methylation of archaeal pre-tRNATrp nucleotides is guided by the intron-encoded but trans-acting box C/D ribonucleoprotein of pre-tRNA. *The Journal of biological chemistry* **279**: 47661-47671.

Slupska, M.M., A.G. King, S. Fitz-Gibbon, J. Besemer, M. Borodovsky & J.H. Miller, (2001) Leaderless transcripts of the crenarchaeal hyperthermophile *Pyrobaculum aerophilum*. *Journal of molecular biology* **309**: 347-360.

Soppa, J., (1999) Transcription initiation in Archaea: facts, factors and future aspects. *Molecular microbiology* **31**: 1295-1305.

Starostina, N.G., S. Marshburn, L.S. Johnson, S.R. Eddy, R.M. Terns & M.P. Terns, (2004) Circular box C/D RNAs in *Pyrococcus furiosus*. *Proc Natl Acad Sci U S A* **101**: 14097-14101.

Steitz, J.A. & K.T. Tycowski, (1995) Small RNA chaperones for ribosome biogenesis. *Science* **270**: 1626-1627.

Su, A.A., V. Tripp & L. Randau, (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 research* **41**: 6250-6258.

Tycowski, K.T., A. Aab & J.A. Steitz, (2004) Guide RNAs with 5' caps and novel box C/D snoRNA-like domains for modification of snRNAs in metazoa. *Current biology : CB* **14**: 1985-1995.

Tycowski, K.T., Z.H. You, P.J. Graham & J.A. Steitz, (1998) Modification of U6 spliceosomal RNA is guided by other small RNAs. *Molecular cell* **2**: 629-638.

Villa, T., F. Ceradini, C. Presutti & I. Bozzoni, (1998) Processing of the intron-encoded U18 small nucleolar RNA in the yeast *Saccharomyces cerevisiae* relies on both exo- and endonucleolytic activities. *Molecular and cellular biology* **18**: 3376-3383.

Wagner, M., M. van Wolferen, A. Wagner, K. Lassak, B.H. Meyer, J. Reimann & S.V. Albers, (2012) Versatile Genetic Tool Box for the Crenarchaeote *Sulfolobus acidocaldarius*. *Frontiers in microbiology* **3**: 214.

Wagner, M., A. Wagner, X. Ma, J.C. Kort, A. Ghosh, B. Rauch, B. Siebers & S.V. Albers, (2014) Investigation of the malE promoter and MalR, a positive regulator of the maltose regulon, for an improved expression system in *Sulfolobus acidocaldarius*. *Applied and environmental microbiology* **80**: 1072-1081.

Watkins, N.J. & M.T. Bohnsack, (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.

Watkins, N.J., V. Segault, B. Charpentier, S. Nottrott, P. Fabrizio, A. Bachi, M. Wilm, M. Rosbash, C. Branlant & R. Luhrmann, (2000) A common core RNP structure shared between the small nucleoar box C/D RNPs and the spliceosomal U4 snRNP. *Cell* **103**: 457-466.

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

Weisel, J., S. Wagner & G. Klug, (2010) The Nop5-L7A-fibrillarin RNP complex and a novel box C/D containing sRNA of *Halobacterium salinarum* NRC-1. *Biochemical and biophysical research communications* **394**: 542-547.

Will, S., T. Joshi, I.L. Hofacker, P.F. Stadler & R. Backofen, (2012) LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *Rna* **18**: 900-914.

Will, S., K. Reiche, I.L. Hofacker, P.F. Stadler & R. Backofen, (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS computational biology* **3**: e65.

Wurtzel, O., R. Sapra, F. Chen, Y. Zhu, B.A. Simmons & R. Sorek, (2010) A single-base resolution map of an archaeal transcriptome. *Genome research* **20**: 133-141.

Yip, W.S., H. Shigematsu, D.W. Taylor & S.J. Baserga, (2016) Box C/D sRNA stem ends act as stabilizing anchors for box C/D di-sRNPs. *Nucleic acids research*.

Zago, M.A., P.P. Dennis & A.D. Omer, (2005) The expanding world of small RNAs in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Molecular microbiology* **55**: 1812-1828.

Figure legends

Figure 1: Conserved sequence motifs within the genomic 50 nt upstream region of archaeal C/D box sRNA genes. (A) The genomic upstream regions of 343 archaeal C/D box sRNA genes were analyzed for conserved sequences and three sequence motifs were identified that are represented with sequence logos (Crooks et al., 2004). (B) The total number of analyzed C/D box sRNA genes in the individual organisms (Mma: *Methanococcus maripaludis*; Neq: *Nanoarchaeum equitans*; Sac: *Sulfolobus acidocaldarius*; Tte: *Thermoproteus tenax*, Mka: *Methanoyrus kandleri*; Iho: *Ignioccus hospitalis*), the number of genes with the conserved motifs and motifs identified at standard promoter position (-26 +/- 3 bp upstream of the transcription start site) are displayed. (C) The distribution of the three identified promoter motifs upstream of C/D box sRNA genes in the six analyzed organisms is shown. For some C/D box sRNA gene upstream regions more than one of the indicated motifs could be identified.

Figure 2: Genetic context of archaeal C/D box sRNA genes and positions of start and stop codons within C/D box sRNAs that overlap with the 5' or 3' end of flanking ORFs. (A) The schematic C/D box sRNA gene structure is shown. (B) The genetic context of C/D box sRNA genes from *Methanococcus maripaludis*, *Sulfolobus acidocaldarius*, *Thermoproteus tenax*, *Methanopyrus kandleri* and I*gnioccus hospitalis* was analyzed (for details see Table S2). C/D box sRNA genes that overlap with a flanking ORF suggest the presence of transcripts with UTRs. C/D box sRNA genes that are located upstream or downstream of the neighboring gene, separated by only few nt may also exist as shared transcripts. (C) Positions of the stop codons of coding regions with overlapping C/D box sRNA genes. The stop codon UGA is used more often than UAA and UAG. (D) Positions of the start codons of coding regions with overlapping C/D box sRNA genes.

Figure 3: Verification of dicistronic C/D box sRNA transcripts and C/D box sRNAs in 5' and 3'-UTRs of flanking coding regions via Northern blot and RT-PCR analyses. (A) Northern blot analyses were performed with 10 µg extracted total RNA from *S. acidocaldarius* MW001 and 5'-labeled probes complementary to C/D box sRNA Sac-sR125 and Sac-sR126, respectively. Both analyses reveal two major bands that correspond to the size of a mature C/D box sRNA (~60/65 nt) and the C/D box sRNA dicistron (~125 nt). (B) RT-PCR analyses of the C/D box sRNA tandem Sac-sR125/126 reveal the existence of a dicistronic transcript. In the control reaction (C) the reverse transcriptase was not added during the reaction. A size standard (M) was used to determine the size of the resulting products (2-Log DNA ladder, New England Biolabs). (C, D) RT-PCR analyses to test for the existence of C/D box sRNA Sac-sR121 in the 5'-UTR of Saci_1247 and for of C/D box sRNA Sac-sR24 in the 3'-UTR of Saci_0125 reveal products of the expected size and confirm the existence of precursor transcripts.

Figure 4: Relative β-galactosidase activity of C/D box sRNA-lacS fusions. (A) C/D box sRNA genes that are located at the 5' end of protein-coding genes were fused to *lacS* relative to their genomic localization. The enzyme activity of the β-galactosidase in these fusion constructs was measured and referred to the control lacS without fused C/D box sRNA. A reduction of enzyme activity can be observed for all C/D box sRNA-lacS fusions. (B) C/D box sRNA genes that are located at the 3' end of protein-coding genes were fused to *lacS* relative to their genomic localization. The enzyme activity of the β-galactosidase in these fusion constructs was measured and referred to the control lacS without fused C/D box sRNA. The C/D box sRNA fusions have either a negative or neutral effect on the β-galactosidase activity. Error bars indicate the standard deviation of three biological replicates.

Figure 5: Cloning scheme and Northern blot analysis of the maintenance of synthetic C/D box sRNA Sac-sR121* constructs in *S. acidocaldarius*. (A) Investigated constructs contain C/D box sRNA Sac-sR121 with a synthetic D' guide and 50 bp of the native upstream and downstream sequence of the C/D box sRNA gene. The constructs were cloned into the plasmid pSVA1431 containing an inducible maltodextrin (mal) promoter or the native promoter. In the absence of the native promoter, the TATA box in the 50 bp upstream region was exchanged as specified. (B) Northern blot analyses revealed the existence of a TATA box in the 50 bp upstream sequence. Random sequences within the 50 bp upstream and downstream regions of C/D box sRNA genes did not abolish C/D box sRNA maturation. Unspecific binding of the probe to *S. acidocaldarius* MW001 RNAs is tested in lane 1. (C) The indicated k-turn and k-loop mutants were analyzed for their impact on C/D box sRNA maturation and stability. K-turn mutations were not tolerated, but k-loop mutations outside the conserved GA bp were allowed.

Figure 6: Archaeal C/D box sRNA biogenesis model. C/D box sRNAs are frequently co-transcribed as part of 3' or 5'-UTRs. The start and stop codon sequences of the flanking coding regions can be incorporated within the conserved boxC and boxD elements. C/D box sRNA maturation occurs independent of the genetic context and is proposed to utilize sequence-independent exoribonucleases whose primary role is mRNA ribolysis. An intact k-turn structure is crucial for C/D box sRNA stability and it is suggested that L7Ae binding to this structure or complete C/D box sRNP assembly protects the mature C/D box sRNA from exoribonuclease degradation.

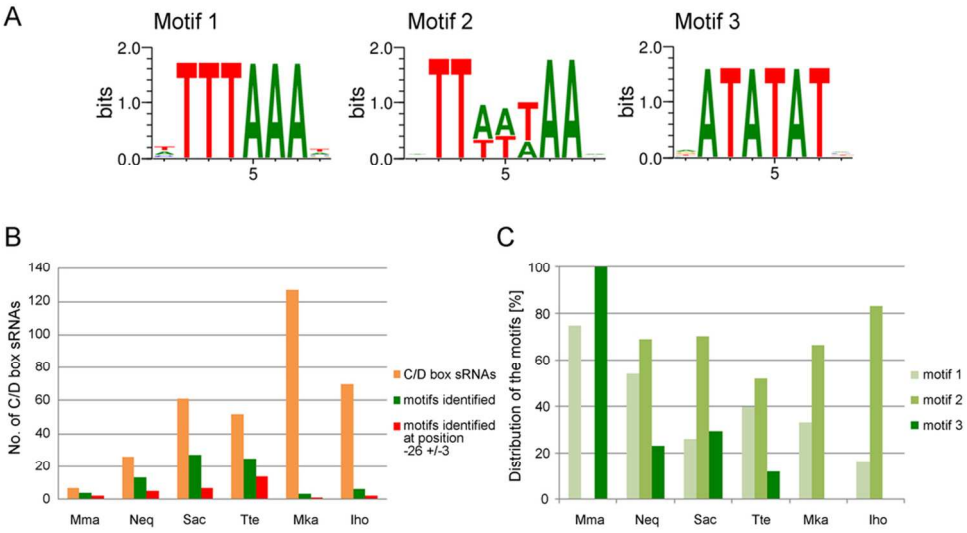


Figure 1: Conserved sequence motifs within the genomic 50 nt upstream region of archaeal C/D box sRNA genes. (A) The genomic upstream regions of 343 archaeal C/D box sRNA genes were analyzed for conserved sequences and three sequence motifs were identified that are represented with sequence logos (Crooks et al., 2004). (B) The total number of analyzed C/D box sRNA genes in the individual organisms (Mma: Methanococcus maripaludis; Neq: Nanoarchaeum equitans; Sac: Sulfolobus acidocaldarius; Tte: Thermoproteus tenax, Mka: Methanoyrus kandleri; Iho: Ignioccus hospitalis), the number of genes with the conserved motifs and motifs identified at standard promoter position (-26 +/- 3 bp upstream of the transcription start site) are displayed. (C) The distribution of the three identified promoter motifs upstream of C/D box sRNA genes in the six analyzed organisms is shown. For some C/D box sRNA gene upstream regions more than one of the indicated motifs could be identified.

Figure 1
94x56mm (300 x 300 DPI)

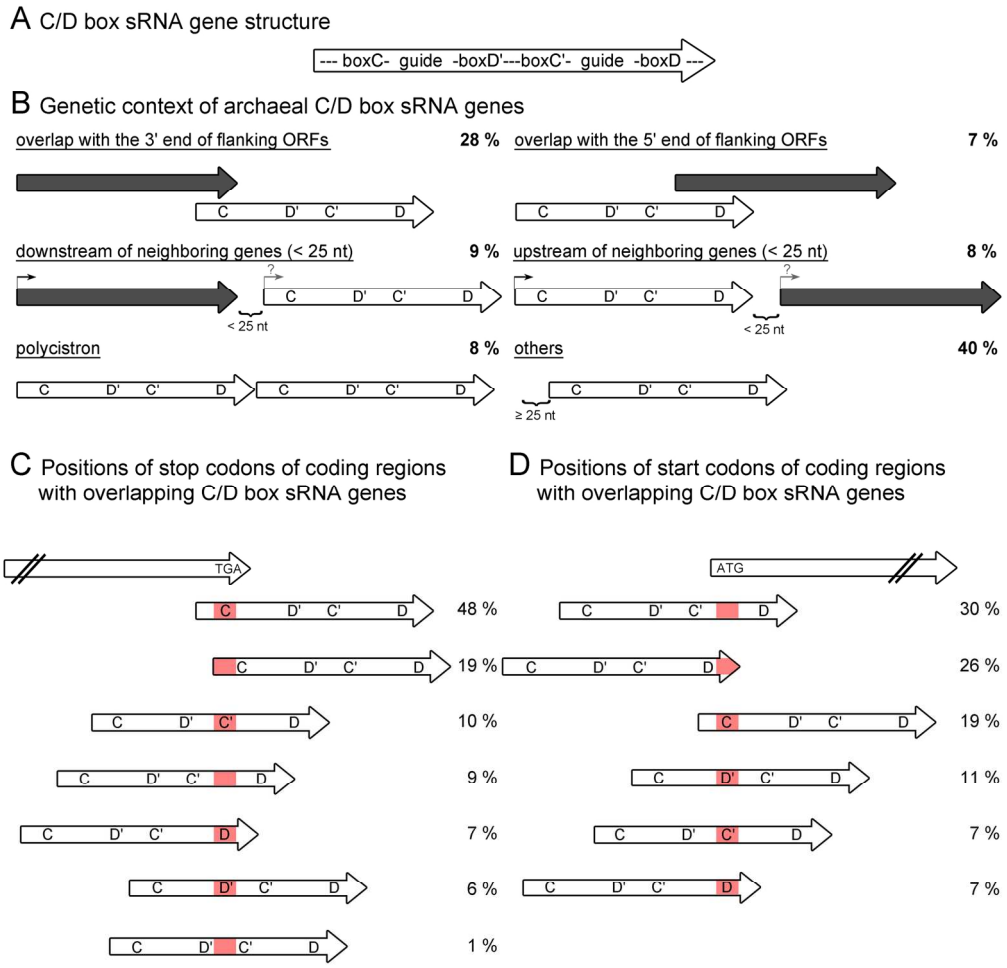


Figure 2: Genetic context of archaeal C/D box sRNA genes and positions of start and stop codons within C/D box sRNAs that overlap with the 5' or 3' end of flanking ORFs. (A) The schematic C/D box sRNA gene structure is shown. (B) The genetic context of C/D box sRNA genes from Methanococcus maripaludis, Sulfolobus acidocaldarius, Thermoproteus tenax, Methanopyrus kandleri and Ignioccus hospitalis was analyzed (for details see Table S2). C/D box sRNA genes that overlap with a flanking ORF suggest the presence of transcripts with UTRs. C/D box sRNA genes that are located upstream or downstream of the neighboring gene, separated by only few nt may also exist as shared transcripts. (C) Positions of the stop codons of coding regions with overlapping C/D box sRNA genes. The stop codon UGA is used more often than UAA and UAG. (D) Positions of the start codons of coding regions with overlapping C/D box sRNA genes.

Figure 2
158x156mm (300 x 300 DPI)

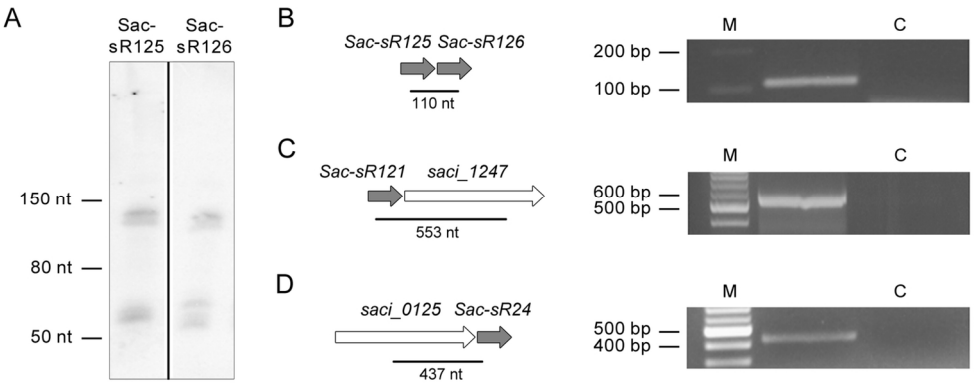


Figure 3: Verification of dicistronic C/D box sRNA transcripts and C/D box sRNAs in 5' and 3'-UTRs of flanking coding regions via Northern blot and RT-PCR analyses. (A) Northern blot analyses were performed with 10 µg extracted total RNA from S. acidocaldarius MW001 and 5'-labeled probes complementary to C/D box sRNA Sac-sR125 and Sac-sR126, respectively. Both analyses reveal two major bands that correspond to the size of a mature C/D box sRNA (~60/65 nt) and the C/D box sRNA dicistron (~125 nt). (B) RT-PCR analyses of the C/D box sRNA tandem Sac-sR125/126 reveal the existence of a dicistronic transcript. In the control reaction (C) the reverse transcriptase was not added during the reaction. A size standard (M) was used to determine the size of the resulting products (2-Log DNA ladder, New England Biolabs). (C, D) RT-PCR analyses to test for the existence of C/D box sRNA Sac-sR121 in the 5'-UTR of Saci_1247 and for of C/D box sRNA Sac-sR24 in the 3'-UTR of Saci_0125 reveal products of the expected size and confirm the existence of precursor transcripts.

Figure 3
130x53mm (300 x 300 DPI)

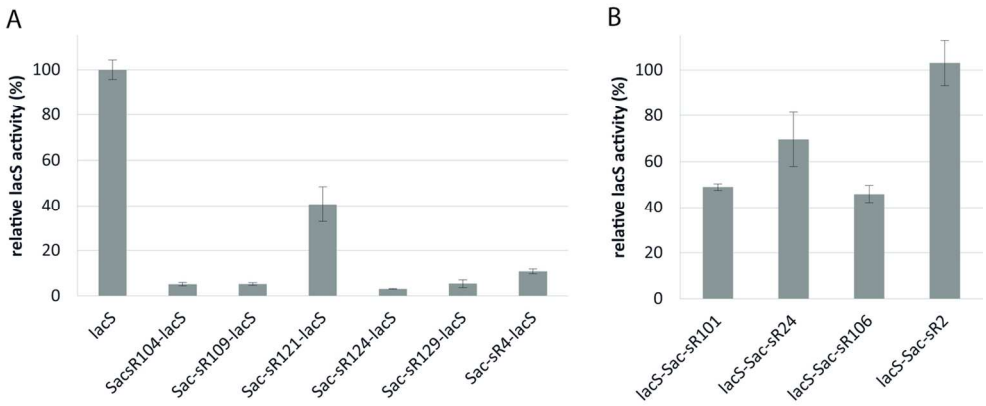


Figure 4: Relative β-galactosidase activity of C/D box sRNA-lacS fusions. (A) C/D box sRNA genes that are located at the 5' end of protein-coding genes were fused to lacS relative to their genomic localization. The enzyme activity of the β-galactosidase in these fusion constructs was measured and referred to the control lacS without fused C/D box sRNA. A reduction of enzyme activity can be observed for all C/D box sRNA-lacS fusions. (B) C/D box sRNA genes that are located at the 3' end of protein-coding genes were fused to lacS relative to their genomic localization. The enzyme activity of the β-galactosidase in these fusion constructs was measured and referred to the control lacS without fused C/D box sRNA. The C/D box sRNA fusions have either a negative or neutral effect on the β-galactosidase activity. Error bars indicate the standard deviation of three biological replicates.

Figure 4

131x53mm (300 x 300 DPI)

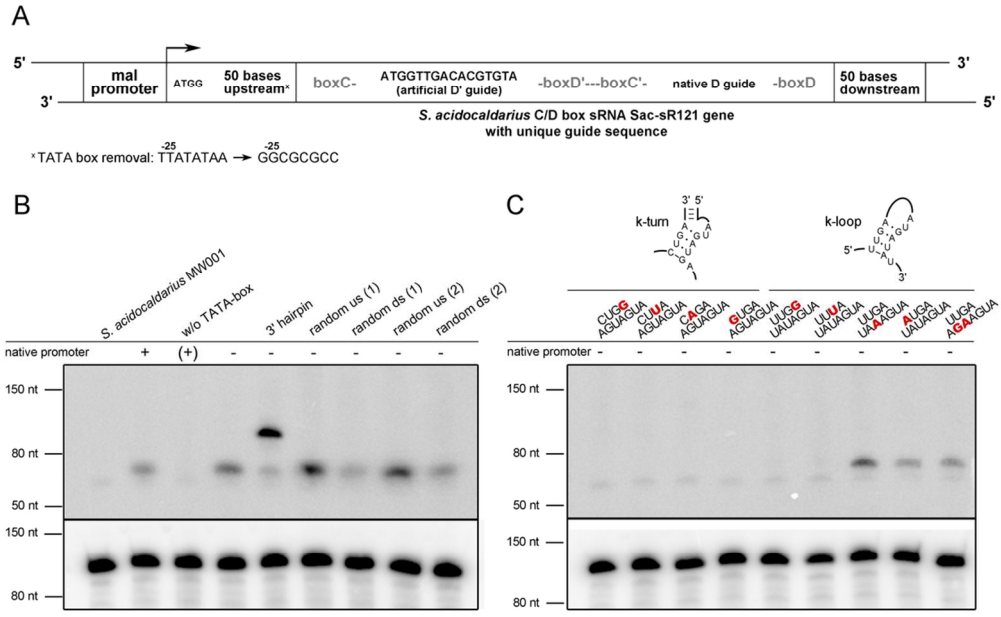


Figure 5: Cloning scheme and Northern blot analysis of the maintenance of synthetic C/D box sRNA Sac-sR121* constructs in S. acidocaldarius. (A) Investigated constructs contain C/D box sRNA Sac-sR121 with a synthetic D' guide and 50 bp of the native upstream and downstream sequence of the C/D box sRNA gene. The constructs were cloned into the plasmid pSVA1431 containing an inducible maltodextrin (mal) promoter or the native promoter. In the absence of the native promoter, the TATA box in the 50 bp upstream region was exchanged as specified. (B) Northern blot analyses revealed the existence of a TATA box in the 50 bp upstream sequence. Random sequences within the 50 bp upstream and downstream regions of C/D box sRNA genes did not abolish C/D box sRNA maturation. Unspecific binding of the probe to S. acidocaldarius MW001 RNAs is tested in lane 1. (C) The indicated k-turn and k-loop mutants were analyzed for their impact on C/D box sRNA maturation and stability. K-turn mutations were not tolerated, but k-loop mutations outside the conserved GA bp were allowed.

Figure 5
104x68mm (300 x 300 DPI)

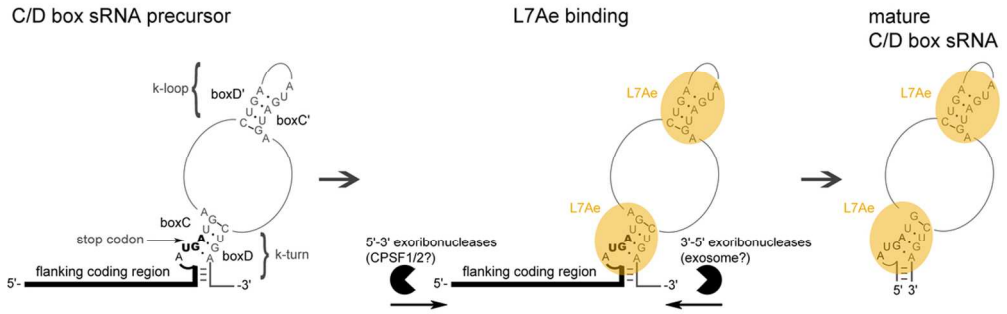
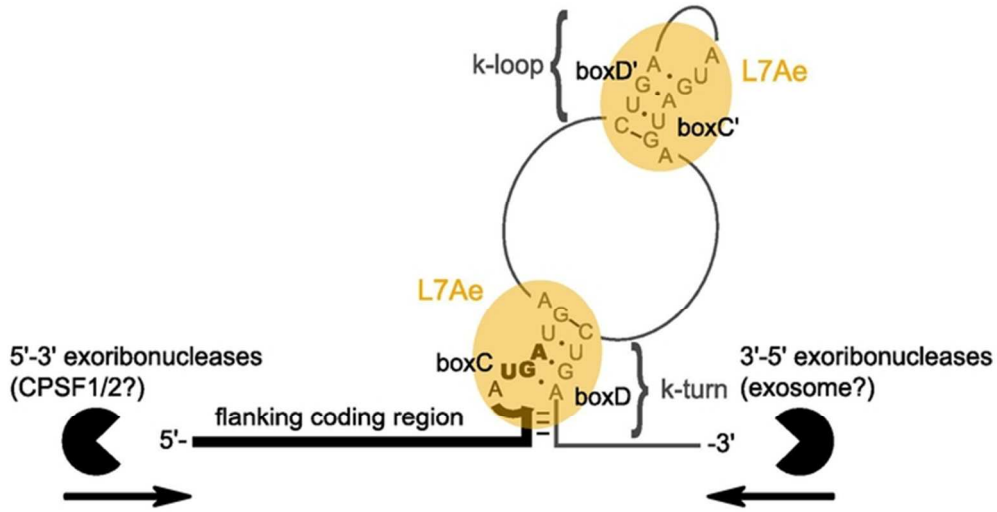


Figure 6: Archaeal C/D box sRNA biogenesis model. C/D box sRNAs are frequently co-transcribed as part of 3' or 5'-UTRs. The start and stop codon sequences of the flanking coding regions can be incorporated within the conserved boxC and boxD elements. C/D box sRNA maturation occurs independent of the genetic context and is proposed to utilize sequence-independent exoribonucleases whose primary role is mRNA ribolysis. An intact k-turn structure is crucial for C/D box sRNA stability and it is suggested that L7Ae binding to this structure or complete C/D box sRNP assembly protects the mature C/D box sRNA from exoribonuclease degradation.

Figure 6
107x36mm (300 x 300 DPI)



C/D box sRNAs constitute one of the most abundant RNA families in Archaea and are used to guide methylation of target RNA molecules. We analyzed the biogenesis of this RNA family and observed its independence of the genetic context. This plasticity of C/D box sRNA biogenesis is suggested to enable their accelerated evolution and allow for adjustments of the RNA modification landscape.

Graphical Abstract

59x44mm (300 x 300 DPI)