GENOMICS AND COMPUTATIONAL BIOLOGY

Vol. 1, No. 1 (2015): e18

Research Article

Processed Small RNAs in Archaea and BHB Elements

Sarah J. Berkemer[1,2,4,7,*], Christian Höner zu Siederdissen[1,2,4], Fabian Amman[2], Axel Wintsche[3,4], Sebastian Will[1,4], Ivo L. Hofacker[2,5,6], Sonja J. Prohaska[3,4], Peter F. Stadler[1,2,4,6,7,8,9,*]

[1]Bioinformatics Group, Department of Computer Science, Univ. Leipzig, Germany
[2]Institute for Theoretical Chemistry, Univ. Vienna, Austria
[3]Computational EvoDevo Group, Department of Computer Science, Univ. Leipzig, Germany
[4]Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany
[5]Research Group Bioinformatics and Computational Biology, Faculty of Computer Science, Univ. Vienna, Austria
[6]Center for non-coding RNA in Technology and Health, Univ. Copenhagen, Denmark
[7]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
[8]Fraunhofer Institut f ür Zelltherapie und Immunologie, Leipzig, Germany
[9]Santa Fe Institute, Santa Fe, USA

*To whom correspondence should be addressed: bsarah@bioinf.uni-leipzig.de; studla@bioinf.uni-leipzig.de

Received 2015-02-14; Accepted 2015-07-29; Published 2015-09-18

ABSTRACT

Bulge-helix-bulge (BHB) elements guide the enzymatic splicing machinery that in Archaea excises introns from tRNAs, rRNAs from their primary precursor, and accounts for the assembly of piece-wise encoded tRNAs. This processing pathway renders the intronic sequences as circularized RNA species. Although archaeal transcriptomes harbor a large number of circular small RNAs, it remains unknown whether most or all of them are produced through BHB-dependent splicing. We therefore conduct a genome-wide survey of BHB elements of a phylogenetically diverse set of archaeal species and complement this approach by searching for BHB-like structures in the vicinity of circularized transcripts. We find that besides tRNA introns, the majority of box C/D snoRNAs is associated with BHB elements. Not all circularized sRNAs, however, can be explained by BHB elements, suggesting that there is at least one other mechanism of RNA circularization at work in Archaea. Pattern search methods were unable, however, to identify common sequence and/or secondary structure features that could be characteristic for such a mechanism.

KEYWORDS

Circular RNA; Archaea; splicing; structure-based

INTRODUCTION

The small non-coding RNAs (sRNAs) of Archaea are much less understood than their eubacterial counterparts. At least in part this is a consequence of the comparably much smaller number of fully sequenced genomes since the large evolutionary distances between them hamper homology-based annotation. Archaea share ribosomal RNAs (rRNAs), transfer RNAs (tRNAs) and the RNA components of RNAse P and the signal recognition particle (7SRNA) with the other two domains of life. Although a large number of other sRNAs has been described for individual species, only three RNA classes are unambiguously recognizable within this diversity: Archaea and Eubacteria share CRISPR-Cas adaptive immune systems [1], and two classes of guide RNAs direct chemical modifications of rRNAs and other non-coding RNAs in both Eukarya and Archaea [2].

Both the Archaeal box C/D and the box H/ACA ribonucleoproteins (RNPs) contain core proteins clearly homologous to those of the Eukaryotic snoRNPs (reviewed in [3, 4]). Hence they are often referred to as snoRNAs even though Archaea do not have a nucleolus. We follow here this (ab)use of nomenclature. Box C/D sRNPs have a dimeric architecture [5] throughout the Archaea. Like their eukaryotic counterparts they catalyze 2’-O-ribose methylation at target sites determined by the box C/D snoRNA [2]. These RNAs have a length of about 50-60 nucleotides and feature two functionally essential kink-turns associated with the C/D and C’/D’ sequence motifs that are characteristic for the RNA class [6]. The box H/ACA snoRNAs guide the formation of pseudouridine [2, 7]. Their canonical secondary structure consists of a single stem-loop structure. In contrast, eukaryotic snoRNAs usually consist of two such stem-loops, each of which addresses an individual target. Beyond the canonical forms, recently several pseudouridylation guide RNAs with divergent secondary structures have been reported [8].

Circularized forms of small RNAs are abundant in some Archaea [9]. Some box C/D snoRNAs seem to be present predominantly or possibly exclusively as circular RNAs [9–11]. This is also true for assorted other small RNA species, among them the 5S rRNA [9]. The biosynthesis of most of these circular RNAs remains unknown. Only in a few select cases it is understood as a consequence of enzymatic splicing.

Archaea completely lack a spliceosomal splicing machinery. Enzymatic splicing, however, is a rather common feature in archaeal RNA processing. Mechanistically, it is closely related to tRNA splicing in Eukarya. A specific endonuclease recognizes and cleaves the so-called bulge-helix-bulge (BHB) structure, Figure 1; a specific ligase then joins the exons and circularizes the intron [12–15]. The prime example of BHB splicing is the removal of tRNA introns. In contrast to Eukarya, archaeal tRNAs may have multiple introns [16, 17] and the same mechanism implements a form of trans-splicing that composes tRNAs from two or three independently encoded fragments [18–23]. The maturation of the ribosomal RNA operon proceeds with the help of two BHB elements that form over long distances and guide formation of circularized precursors of the 16S and 23S rRNA, respectively [24, 25]. At least one pre-m RNA (CBF5, the archaeal homolog of dyskerin) contains an intron with a BHB element in many crenarchaeal species [26]. Although many of the box C/D snoRNAs in several archaeal species appear as circularized RNA molecules in the cell [27], this is known to be the consequence of BHB-guided splicing in only a single case, namely the box C/D snoRNA processed from a long intron of the tRNA-Trp precursor in *pyrococcus* species [28].

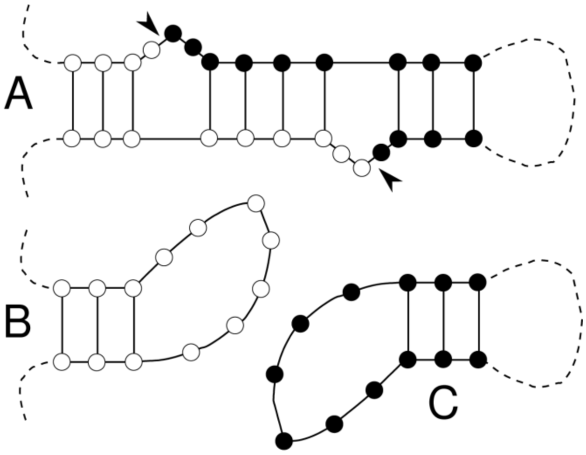

Figure 1: Schematic representation of BHB-guided enzymatic splicing. The substrate A is cleaved at the splice junctions (indicated by the arrows) located within two adjacent bulges separated by a short helix. Ligation yields the spliced product B and a spliced-out intron, a circularized RNA.

BHB elements are usually described as stringently defined secondary structure elements consisting of a 4 nt helix enclosed by two 3 nt bulges, Figure 1. Several different forms and variants have been reported in the literature however. BHB elements with one or two additional helices in intron and exon, denoted as hBHBh’, hBHB or BHBh’ [29], as well as BHB elements of type hBH or HBh’ with only a single bulge were discovered in [21]. Watanabe et al. [30] furthermore described relaxed BHB elements that feature mismatches within helices and bulges or helices with a difference respect to the standard length are also successfully processed [30, 31]. These relaxed BHB elements are recognized and cleaved by special forms of splicing endonucleases [30]. The BHB elements of tRNA introns and rRNAs are well-conserved across the *Crenarchaeota* phylum [29] and have been studied in detail [9, 27, 29, 32–36]. They fall into three classes originally defined in [29]. Two of them can be seen as relaxed versions of the structurally most complex group depicted in Figure 2 (top). These well-documented cases also exhibit several deviations from the standard BHB motif in Figure 1.

In this contribution[1] we explore two interrelated questions: (1) can BHB-element-dependent splicing explain all or at least most of the observed circularized or permuted small RNAs, and (2) to what extent can BHB elements be used in their own right as means of detecting novel small RNAs and/or likely sites of RNA processing. For the first question, we focus in particular on box C/D snoRNAs because the members of this abundant RNA class are often circularized. The second question is of practical relevance since in many cases RNA-seq data do not provide decisive information. (a) Circularized RNAs are depleted in most RNA-seq protocols unless specifically enriched e.g. by RNAse R treatment [9]. (b) Spliced tRNAs cannot be detected in cases where an unspliced paralog is present in the same genome [11]. (3) Circularized introns e.g. of tRNAs are often too short to be detectable by sequencing in their own right. For instance 5 out of 8 introns listed in [27] have a length of 26 nt or less.

METHODS

Known and Putative BHB Elements

The best known examples of BHB elements compiled in Figure 2 (top), which also include tRNA introns, already exhibit substantial deviations from the canonical structure including mismatches and elongations of stems. It is necessary, therefore to relax the definition of BHB elements to accommodate these cases. Without further biochemical evidence it is impossible to decide whether they are processed by the same enzyme(s) or not. The fact that the variability appears also among the extremely well-conserved tRNA introns however does suggest a common processing machinery.

BHB elements have been described extensively for tRNAs [29] and rRNAs. Following the workflow shown in Figure 3 we extended the alignments with BHB elements from ref. [29] with tRNA associated BHB elements from ref. [27]. Most archaeal tRNAs, including the spliced ones, but with the notable exception of split and permuted tRNAs are readily identified by *tRNAscan-SE* [38]. We therefore ran *tRNAscan-SE -A* for all genomes to complete tRNA data. A multiple sequence alignment, MSA1, was manually built from sequences with known BHB elements. It emphasizes the well-conserved secondary structure pattern, Figure 2 (top).

To investigate the box C/D snoRNAs we constructed a second alignment, MSA2, from 15 *Nanoarchaeum equitans* snoRNA sequences with unambiguous box C and box D motifs annotated in reference [39]; we constructed the alignment in a semi-automatic fashion. Since these short RNAs (length 45-70 nt) exist as circularized molecules, putative BHB elements similar to those of tRNA introns should be detectable overlapping both circularizing junctions if they are indeed processed in the same manner as tRNAs. We therefore included 15 nt flanking sequence, Figure 3 (bottom). In addition we

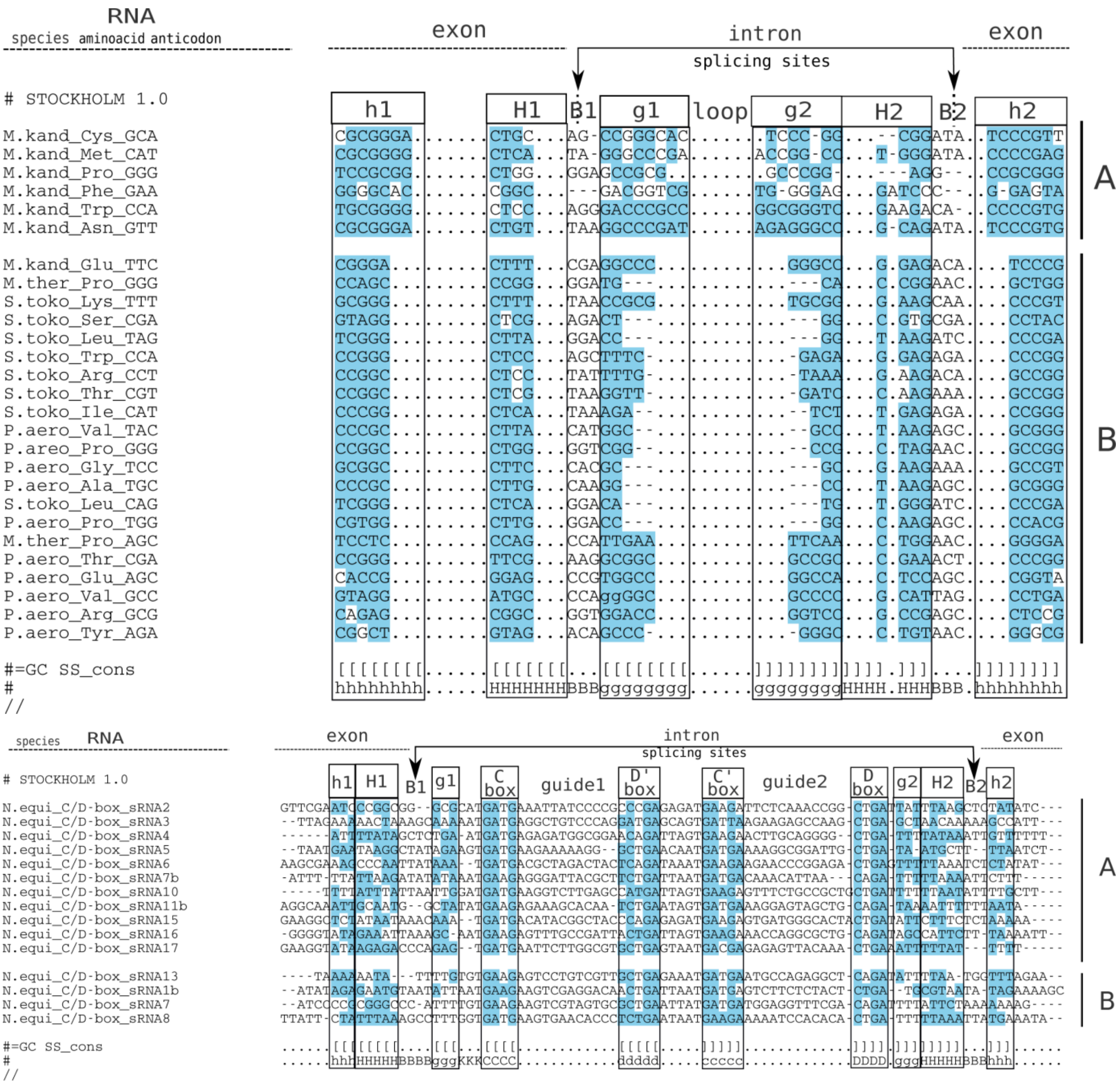

Figure 2: Consensus multiple alignment of genomic loci harboring BHB structural elements. MSA1 (top) is built from tRNAs with introns from [27] (block A) and [29] (block B). MSA2 below comprises box C/D snoRNAs from N. equitans [39]. The helical structure is different for BHB motifs of C/D box snoRNA motifs compared to tRNAs. Here, a part of the C- and D-boxes forms an extension of the g-stem. The two cleavage sites in the bulges (labeled B1 and B2 and marked by BBB at the bottom) are indicated by arrows. Base pairs are denoted by [,]. The three helical regions are highlighted and labelled H, h, and g, respectively.

retrieved the known box C/D snoRNA sequences of *Sulfolobus solfataricus* [9, 40], *Sulfolobus acidocaldarius* [9, 41], *Nanoarchaeum equitans* [39], and *Methanopyrus kandleri* [27] together with 15 nt flanking sequence. For the construction of MSA2 we used *LocARNA* [42]. This tool simultaneously infers an alignment and a consensus secondary structure from a set of unaligned RNA sequences taking into account sequence similarity, structure similarity, and thermodynamics [43]. To include further knowledge, in our case C and D boxes as well as the kink turn motifs of the snoRNAs, we generated a constrained *LocARNA* alignment anchored at annotated columns of an initial manual alignment. We predicted for each sequence the thermodynamically most stable secondary structure that contains the consensus structure rather than modifying the structures to conform to the canonical BHB structure as much as possible. This allows us to also estimate and evaluate the structural variation.

Circular transcripts

Known circular sRNAs were compiled from *Sulfolobus solfataricus* [9, 40], *Sulfolobus acidocaldarius* [9, 41] , *Nanoarchaeum equitans* [39], and *Methanopyrus kandleri* [27]. Publicly available RNA-seq data of *S. acidocaldarius* [44, 45], *S. solfataricus* [40], *Nanoarchaeum equitans* and *Ignicoccus hospitalis* [39], and *M. kandleri* [27] were mapped and analyzed as outlined in [11]. In brief, read data sets for each species were pooled, the reads were quality-trimmed and then mapped to the corresponding genome using *segemehl* [46, 47] with the option --splits that forces reads that do not acceptably match as a single substring to be split and matched across splice sites. We required that a seed of at least 11 nt maps to each side of the split. The statistics of the mappings are summarized in Table 1. The *transrealign* tool, a component of the *segemehl* suite [47], was used to identify the circularizing splice junctions. We retained a circularization site if *transrealign* reported a backsplicing junction and a threshold of 5 reads covering the splice junction. In order to search for "intronic" sRNAs, we filtered the circularized candidates by length, requiring a minimum of 45 and maximum of 250 nt. The dataset consists of 5 archaeal species. Of these, 3 species belong to the phylum Crenarchaeota, 2 species are Euryarchaeota. A phylogenetic tree of the Archaea highlighting the position of the species investigated here can be found in the Supplementary Figure 1.

Species	mapped reads	split reads
S. acidocaldarius	23123836	108283
M. kandleri	15889885	128603
N. equitans	10728929	25721
I. hospitatlis	5567812	42757
S. solfataricus	8642437	74143

Table 1: Summary of mapped RNA-seq data sets. More details can be found in the Supplementary Table 16.

Evolutionary Conservation

In order to investigate the evolutionary conservation of some circular RNA candidates, we located potential homologs with *blast* in all publicly available archaeal genomes from the NCBI genbank. For each locus, we determined all corresponding sequences up to tolerant e-values of $E < 0.01$ for 20 query RNAs from *M. kandleri*, where no close relatives exist in the database. For 65 circular RNA candidates from *S. acidocaldarius* the much more conservative cut-off $E < 10^{-30}$ was used since the genomes of several closely related species are available. We constructed *ClustalW* alignments of detected homologs. These were evaluated with *RNAz* [48] to detect conserved RNA secondary structure.

Element-based search

A preliminary study [37] showed that covariance models (CMs) [49, 50] are much better suited to identify BHB elements than simpler but faster methods. We derived CMs from both MSA1 and MSA2 using *cmbuild*, a component of the *Infernal* suite (version 1.1) [51]. BHB elements are short (12-23 base pairs) compared to common non-coding RNA models in the Rfam database [52] and they enclose potentially large intronic sequences, see Figure 2. Both properties inherently limit the sensitivity matching the CMs against a genome.

While *Infernal* is, in principle, able to handle large local insertions, they come at a cost. For large models, this cost is absorbed by a sufficiently complex structure, thus still resulting in a very specific matcher. In our case, however, the BHB model alone does not provide enough information to offset the cost of the large intron. For MSA1 (Figure 2, top) we therefore modelled the intron site as an unspecific region in [37], and found that most structurally complex BHB elements were detectable with all pre-filters for *Infernal* disabled. Alternatively, one can try to include features of the intron into the CM. For MSA2, we therefore retained the sequences for the boxes C, D, C', and D' in the model. This results in a CM that specifically discovers box C/D snoRNAs that are surrounded by plausible BHB elements but does not generalize to other sRNA classes.

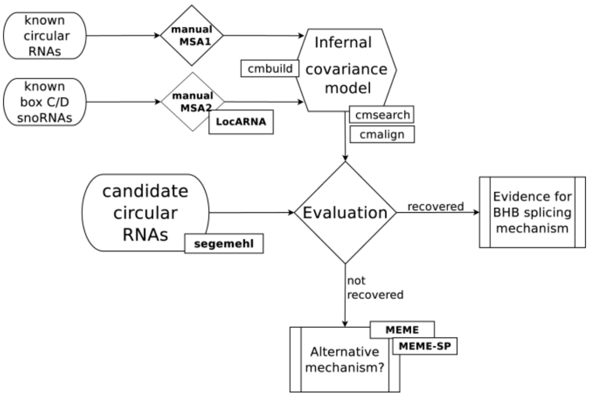

Figure 3: Workflow and summary of results of the genome-wide survey for BHB elements and circular or spliced RNA. From a cache of known circular RNA sequences as well as known box C/D snoRNAs we created two curated multiple-sequence alignments (MSA1 and MSA2) that serve as the basis for a stochastic Infernal covariance model. These are evaluated against RNA-seq data and putative BHB elements found in several of archaeal genomes. Known circularized RNA without BHB elements were further subjected to sequence and structure motif discovery using MEME and MEME-SP.

The CMs of MSA1 and MSA2 were used to search for additional BHB elements in the genomes of *M. kandleri*, *S. solfataricus*, *S. acidocaldarius*, *N. equitans*, and *I. hospitalis*. We used the global mode of *cmsearch* with MSA2 and the local mode with MSA1 because of the long intron. In addition to these whole-genome searches we investigated the known box C/D snoRNAs as well as the RNA candidates identified from the RNA-seq data, again using 15 nt flanking sequence. To this end we used cmalign, another component of the *Infernal* suite.

Pattern Discovery

To discover sequence and/or secondary structure elements that could be involved in circularization we extended all circular RNAs by 50nt at both circularization sites. In case of box C/D snoRNAs, 40nt upstream and downstream of the lateral C and D boxes, respectively, were used. We then analyzed these sequence intervals with several distinct pattern discovery tools.

We used *MEME* (version 4.8.1) [54] with parameters *-mod zoops -minw 4 -maxw 10* to identify sequence patterns. *MEME-SP* [55], an extension of *MEME*, was used to search for combined patterns of sequence and RNA secondary structure. *MEME-SP* uses the same expectation maximization framework as *MEME*, but learns from sequences that, in addition, are annotated with secondary structure profiles. These specify, for each position of each input sequence, the probability that the base is contained in a stem, an interior loop, or a bulge. This secondary structure information is computed for all locally stable suboptimal structures that can be formed, as hybridization structures form the sequence surrounding the circularization sites. The suboptimal hybridization structures are computed by *RNAduplex* [56] with parameter *-e 5* limiting the energy band of interest to the 5 kcal/mol above the most stable structure.

RESULTS

We systematically surveyed 5 archaeal genomes for circularized RNAs and BHB elements that might be responsible for the circularization using the workflow summarized in Figure 3. The *infernal*-based searches for putative BHB elements recovered most of the known BHB elements. We obtained a recall of 85% for the tRNAs with introns of *Methanopyrus kandleri* (6/7), *Sulfolobus solfataricus* (14/16), and *Sulfolobus acidocaldarius* (15/18). The well-described BHB elements flanking the 16S and 23S RNA are not recovered as expected, since these insertions are much too long for the CM-based approach (see Supplementary Table 2).

Most of the annotated box C/D snoRNAs feature BHB elements. A positive signal was recovered for 9 of the remaining 11 *N. equitans* [39] box C/D snoRNAs not already included in MSA2, for 112 of the 126 sequences from *M. kandleri*, for 7 of 20 loci in *S. solfataricus* [9] and 22 of the 24 box C/D snoRNAs reported for *S. acidocaldarius* [41], see Table 2.

Species	C/D	BHB	no BHB
M. kandleri	126	112 (80)	14 (8)
N. equitans	11	9 (6)	2 (1)
S. solfataricus	20	7 (3)	13 (5)
S. acidocaldarius	24	22 (16)	2 (2)

Table 2: Summary of BHB elements found in known box C/D snoRNA sequences. The numbers of sequences with a circularizing splice junction detected by *transrealign* is listed in parentheses. 15 additional sequences from N. equitans were used to build the covariance model, these are not included in the table. For full details we refer to the Supplementary Tables 5–16.

Restricting the CM-based search to the vicinity of the circularized sRNAs, 213 sequences from *M. kandleri*, 12 novel loci in *N. equitans*, 23 in *I. hospitalis*, 185 in *S. acidocaldarius* and 26 in *S. solfataricus* were identified as harboring possible BHB elements. Aligning them to the query CMs identified 30 novel putative box C/D snoRNAs as well as 25 sRNAs in *M. kandleri* fitting the tRNA-derived model. In *N. equitans* 3 new box C/D snoRNA sequences were found, while a single sequence matched the tRNA-derived CM. We identified 7 putative new box C/D RNAs in *I. hospitalis* and 3 loci fitting the tRNA-derived CM from 23 sequences with evidence for BHB elements. In *S. acidocaldarius* 185 sequences with putative BHB elements were found of which 12 fit the tRNA-derived model and 18 are candidates for new box C/D snoRNAs. No sequences were found for the tRNA-derived model in *S. solfataricus* but 8 of 26 sequences with BHB elements fit the box C/D snoRNA model. More detailed information can be found in the Supplementary Tables 17–21.

We detected several splice sites inside annotated coding regions. At least one example for introns in archaeal protein coding genes is well established: Homologs of the eukaryotic CBF5 protein in Archaea contain an intron spliced in a BHB-dependent manner [26, 31, 57]. The existence of additional cases has been proposed in [57]. Our analysis of the RNA-seq data with *transrealign* [47] recovers, as expected, the CBF5 intron of *S. solfataricus* and provides evidence also for other BHB-dependent introns in coding regions.

The sequences used to build MSA2 are box C/D snoRNAs of *N. equitans* verified by RNA-seq data in [39]. Due to the extremely small size of the *N. equitans* genes can be found in very close proximity of each other and most box C/D snoRNAs occur inside introns of tRNAs and also of mRNAs [39]. They are spliced and circularized afterwards to become functional. To our knowledge it is unknown whether spliced, unspliced, or both types of “host” mRNAs give raise to functional proteins in these cases. Figure 4 shows the secondary structure of box C/D snoRNA #5 of *N. equitans*, reported in [39] to form an intron within a putative rRNA methylase [39].

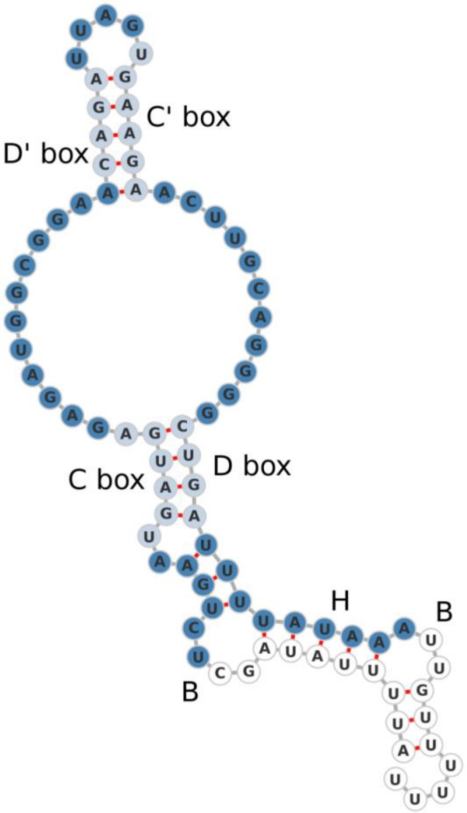

Figure 4: Secondary structure of the N. equitans box C/D sRNA #5 [39], drawn with *forna* [53]. White bases are located in the exon, blue bases belong to the intron. Light blue marks the four sequence boxes C, C’, D, and D’. As reported in [39], the sequence of sRNA #5 overlaps with the sequence of a putative rRNA methyltransferase gene. The two processing sites identified by the transition from white to dark blue coloring are clearly identifiable in RNA-seq data [11, 39].

In the *M. kandleri* RNA-seq data we identified 20 circularized products (besides the tRNA introns) of which 9 match candidate BHB elements. Of these, 6 are located within annotated protein coding genes. Conversely, of the 9 intergenic circular RNAs only three are associated with BHB elements. The remaining 11 circularization products derive from introns within ORFs. Of these, 7 preserve the reading frame, suggesting that enzymatic splicing may lead to functional isoforms (see Supplementary Table 2).

Circular RNAs without BHB elements

Although the majority of tRNA introns and box C/D snoRNAs are processed with the help of BHB elements, there is a substantial number of circular RNA products that are not associated with recognizable BHB elements. So far, no RNA processing mechanism leading to circularized products in Archaea other than the BHB-element directed splicing has been described in any detail. The ligase Pab1020 has been reported to have RNA circularization activity [59], however. It likely uses double-stranded RNAs as substrate. To shed more light on the non-BHB circular RNAs, we searched for possible sequence and/or secondary structure elements that could be involved in circularization.

In order to gather additional evidence that some of the circularized RNA products might be functional sRNAs we tested some of them for evolutionarily conserved secondary structures. For *M. kandleri* we found homologs of 6 loci. Only a single one, however, showed strong evidence for structural conservation. This RNA, Figure 5, is not associated with a BHB element. For *S. acidocaldarius*, *RNAz* predicts RNA structure in 9 cases with *RNAz* class probabilities ≥ 0.5 (Supplementary Table 4).

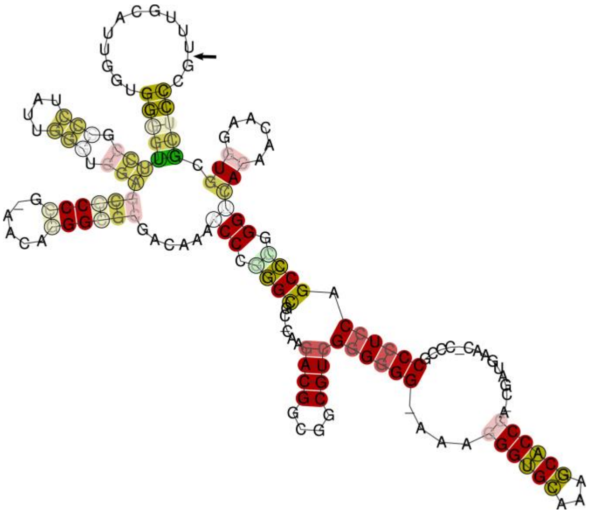

Figure 5: A well-conserved structured RNA in M. kandleri that is circularized but is not flanked by a BHB element. The structure drawing is produced by *RNAalifold* [58] in circular folding mode; the circularization site is indicated. The color coding represents the number of compensatory mutations (red=1, yellow=2, green=3) and conservation (saturation) of base pairs.

Using *MEME* credible sequence motifs were found only in the circular C/D-box snoRNA sequences published in [27]. As expected, the recovered patterns essentially correspond to the box C and box D sequences. No further conserved motifs could be found, especially when masking box C and box D sequences via prior, except short recurring stretches of guanines or cytosines scattered across the entire input sequence. The suboptimal local structures at the circularization sites we assayed with *MEME-SP*. However, no larger common patterns were detectable. Short stems mainly containing guanines and cytosines were found while adenines tended to occur in internal loops. Bulged positions did not feature significantly in any of the predicted motifs.

DISCUSSION

Many circular RNAs in Archaea are produced by BHB-element-dependent enzymatic splicing. In addition to tRNAs, rRNAs (including 5S rRNA), and 7S RNA, our analysis shows that this also pertains to a large subset of box C/D snoRNAs. The generic BHB model we derived from BHB elements reported in the literature does not conform to the strict, canonical BHB motif. In fact, a substantial fraction of the BHB elements previously described in the literature also deviate from the ideal model. Our choice of the covariance model has sufficient sensitivity to recognize most the candidates and at the same time has acceptable specificity, see Table 3. Models using the canonical-structure are effectively unable to discriminate between confirmed BHB-containing sequences and background due to the low information content of the BHB structure motif. This problem becomes even more severe if we disable the standard *Infernal* prior, and use only the canonical sequence-structure information for training. We conclude that a suitably generic model is necessary to detect BHB elements. The candidate elements can postprocessed by grouping them together with known types of BHB structures.

```
(A)
   g: ..auauuaaauaaauaaaAUGAuGAAGuuauugcgcgCuGAa
   g: ..<<<<<<<<----<<<---[[[[------------(((((-
   s: GAauacauAa.uuuuaA.AUGAUGAaguaggGCaaAuCUGAU
   s: ---<<<<<<<.---<<<.--[[[[------------(((((-
(B)
   g: UAAUGAuGaAuaaacgacacgCuGAUuuuuuuaauuuuau..
   g: ----)))))------------]]]]->>>>>>>>--->>>..
   s: UAAUGAUGaAagugCUgGAgaCUGA.UgcuUau.UAUuguua
   s: ----)))))------------]]]].>>>>>>>.--->>>--
```

model	g	Sdp	Szp
mean	19.7	8.3	-47.0
sd	19.8	16.5	20.8
#(bit > 0)	19/21	17/21	0/21
%	90	81	0

Table 3: Comparison between the generic BHB model (g) and two more strict models s_{dp} and s_{zp} that are closer to the canonical BHB element. The most strict model s_{zp} uses a near-zero prior to establish a model entirely based on canonical BHB motifs. The model called s_{dp} is based on canonical BHB motifs but uses the same prior which was used to build our general model. Each model was tested with 21 confirmed C/D box snoRNA sequences [39]. Only the generic model g can reliably find the BHB motif in all tested sequences. The table lists the mean Infernal bit score and the corresponding standard deviation for each model, as well as the number of hits resulting in a positive bit score #(bit > 0)

as well as the corresponding percentage. A table with scores for each sequence can be found in the Supplementary Table 4. The corresponding consensus sequences and structures, separated in left (A) and right (B) half, are (with C/D denoted [], C'/D' (), and BHB <>. The symbol '.' denotes gaps that are not part of the models but where

We identified 30 putative new box C/D snoRNAs flanked by BHB elements for *M. kandleri*, 18 for *S. acidocaldarius*, 8 for *S. solfataricus*, 7 for *I. hospitalis* and 3 for *N. equitans*. New RNAs spliced in a BHB-dependent manner and fitting the tRNA-derived model were also found for 4 of the 5 species (see also Table 4). On the other hand, a sizable number of unclassified sRNAs apparently are not associated with BHB elements. We could, however, not find common sequence or structure motifs that might be associated with a common alternative processing pathway.

Species	Splice sites	MSA1	MSA2
M. kandleri	213	25	30
N. equitans	12	1	3
I. hospitalis	23	3	7
S. acidocaldarius	185	12	18
S. solfataricus	26	0	8

Table 4: Summary of new BHB elements found with the MSA1 and MSA2 CMs among the circularized RNAs identified by transrealign from RNA-seq data. These loci are disjoint from the ones in Table 2. For full details we refer to the Supplementary Tables 17 – 21.

From a computational point of view, BHB elements are difficult to identify because they are formed in trans from distant components, neither of which features distinctive sequence patterns. We have considered two types of BHB elements bracketing a structure. In the tRNA case (MSA1), the BHB element brackets an intron to be spliced out. This presents an algorithmic obstacle given that covariance models [49] as implemented in *Infernal* assume that the RNA sequence to be matched to form a single region, with no interspersed elements. *Infernal* therefore scores the entire sequence, including any insertions relative to the consensus structure. We would prefer it to entirely exclude the intronic region from scoring because we have no a priori knowledge of the intronic region and hence are unable to model it. The nature of mutational events requires *Infernal* to be able to handle local insertions and deletions, which makes it possible to handle large insertions or deletions. This mechanism is, however, not able to cope with insertions that span several dozens to a few hundred nucleotides. Since insertions are necessarily awarded a small, but non-zero cost, they have an unavoidable detrimental effect on sensitivity. Due to the small size of the BHB structure itself, the MSA1 model is at the limit of what can be modelled successfully.

In contrast, the box C/D snoRNAs are bracketed by BHB elements from the outside, i.e., the region of interest is located between the two parts of the BHB element. As box C/D snoRNAs share a common secondary structure and several sequence patterns, the MSA2 CM is a high quality model. The downside is that it is limited to known RNA classes delimited by a BHB element.

The consensus structures of both the tRNA and the box C/D snoRNA model deviate from the canonical BHB structure. There are multiple reasons for this. First, as pointed out in the Methods, non-canonical BHB-like structures are known. In addition, even canonical BHB motifs exhibit variations in sequence and structure. When combined into a single model, or two to incorporate C/D structure, this leads to a consensus that is structurally relaxed compared to the canonical BHB structure. We have, in effect, trained a less specific, but more sensitive model.

While it might be of interest to have more explicit, directly interpretable models of the structural constraints, such an approach is limited by two factors: (i) we only have a moderate training set of known BHB elements and (ii) software such as *Infernal* builds inherently statistical models. To investigate the effect of more stringent structural constraints, we have constructed models from using subsets of the training data with less structural variation. For these restrictive models, we observe a drastic decrease in sensitivity and we lose the ability to detect a large fraction of known BHB elements. We have to conclude, therefore, that the structural variability captured in our generic covariance model is much closer to biology reality than the strict canonical structure shown in Figure 1.

The survey presented here serves as a starting point for a more detailed investigation into the realm of archaeal sRNAs and their processing. It also poses several questions for future research. It highlights the need for the methods to efficiently search for non-contiguous patterns. This requires a more general approach to building matchers than provided by CMs. Modifications seem feasible allowing for "exclusions" with size of a few hundred nucleotides. Recent work on formal methods [60–62] holds promise to tackle this challenge with reasonable effort.

The paradigm certainly breaks down when BHB elements bring together independently transcribed parts of tRNAs from vastly distant genomic loci. Theoretically, one could try to dissect the genome into some 104 tiles about 100 nt in length and to consider all 108 combinations of pairs of tiles, resulting in an artificial genome slightly larger than the human genome (3 109nt), i.e., just within computational reach for the small genomes of the Archaea. Special cases, such as split tRNAs [18, 63], can of course be handled with much less effort. The high conservation of tRNA sequences makes it possible to identify the genomic loci (tiles) that contain tRNA parts first, and restrict the combination of tiles to those candidates only. The *SPLITS* tool [64] is based upon this idea. However, for unknown genes, or those that evolve rapidly and thus lack highly conserved regions, this approach is not applicable.

CONCLUSION

In this work we combined high throughput sequencing data and statistical models for homology search to identify new candidates for BHB-dependent splicing in Archaea. In itself, neither approach is sufficient. While RNA-seq can identify candidates for BHB-dependent splicing, such high throughput data also produce large number of false positives, in particular when inefficiently processed loci are included in the candidate set. On the other hand, their small size, poor definition at sequence level, relatively high level of structural variation, and their inherently non-local nature makes BHB elements an extremely difficult target for homology search – even when the sophisticated machinery of covariance models is employed. The combination of RNA-seq based detection of circularized RNAs with the evaluation of the sequence/structure patterns around the circularization site, however, allows us discriminate between BHB-related splicing and other loci that are presumably processed by other mechanisms. Although our data strongly suggest that a second splicing mechanism exists, the present data do not reveal specific hints such as sequence or structure motifs shared by a (subset of) splicing candidates that appear unrelated to BHB elements. The identification of an alternative splicing mechanism will likely be facilitated by the growing number and increasing sequencing depth of transcriptome studies in Archaea.

Non-local elements such as BHB structures require more specialized computational machinery. The existence of intronic elements alone makes it more complicated to design structural models that are specific enough to recognize BHB-like sequences in genomes. Once trans-splicing comes into play (as in the case of split tRNA genes) more complicated statistical models would be required. Even if sufficiently specific models could be devised, the computational requirements are likely to be prohibitive, however. The genomes investigated here are so diverse that, with the exception of tRNAs, it has in general not been possible to establish the homology of BHB elements. Comparative genomics approaches thus are not particularly helpful to narrow down candidate BHB elements. This may change, however, when transcriptome data become available with a much denser taxon sampling.

ACKNOWLEDGEMENTS

This work was funded, in part, by the Austrian FWF, project “SFB F43 RNA regulation of the transcriptome”, the German ministry of science (0316165C as part of the e:Bio initiative), a PhD stipend from the European Social Fund (2000729457) to (AW), and a DAAD travel fellowship to (SJB). We thank Udo Bläsi and Sonja-Verena Albers for sharing RNA-seq data. We acknowledge support from the German Research Foundation (DFG) and Universität Leipzig within the program of Open Access Publishing.

AUTHOR CONTRIBUTIONS

SJB and CHzS designed the structural alignments, and performed the genome-wide scans. SJB, FA, and ILH analyzed the deepSeq data. SW optimized the structural alignments. AW and SJP performed the search for non-BHB patterns. PFS guided the study. SJB, CHzS, and PFS wrote the paper. All authors read and approved the final manuscript.

CONFLICT OF INTEREST DECLARATION

The authors declare no conflict of interest.

SUPPLEMENTARY DATA

High resolution files of the main figures, together with the supplementary items listed below, are available for download at Genomics and Computational Biology online.

Supplement. This file includes a figure showing taxonomical information about the archaeal species used in this contribution, a figure depicting the alignment of a novel circular ncRNA, as well as tables summarizing information of genomic loci that show evidence for a flanking BHB element.

ABBREVIATIONS

RNP: ribonucleoprotein

RNA: ribonucleic acid

BHB: bulge-helix-bulge

MSA: multiple sequence alignment

REFERENCES

1. Barrangou R. **CRISPR-Cas systems and RNA-guided interference**. Wiley Interdiscip Rev RNA. 2013;4:267–278. doi:10.1002/wrna.1159.
2. Dennis PP, Omer A. **Small non-coding RNAs in Archaea**. Curr Op Microbiol. 2005;8:685–694. doi:10.1016/j.mib.2005.10.013.
3. Lafontaine DLJ, Tollervey D. **Birth of the snoRNPs: the evolution of the modification-guide snoRNAs**. Trends Biochem Sci. 1998;23:383–388. doi:10.1016/S0968-0004(98)01260-2.
4. Yip WS, Vincent NG, Baserga SJ. **Ribonucleoproteins in archaeal pre-rRNA processing and modification**. Archaea. 2013;2013:1–15. doi:10.1155/2013/614735.
5. Bower-Phipps KR, Taylor DW, Wang HW, Baserga SJ. **The box C/D sRNP dimeric architecture is conserved across domain Archaea. RNA**. 2012;18:1527–1540. doi:10.1261/rna.033134.112.
6. Omer AD, Zago M, Chang A, Dennis PP. **Probing the structure and function of an archaeal C/D-box methylation guide sRNA**. RNA. 2006;12:1708–1720. doi:10.1261/rna.31506.
7. Blaby IK, Majumder M, Chatterjee K, Jana S, Grosjean H, de Cr´ecy-Lagard V, et al. **Pseudouridine formation in archaeal RNAs: The case of Haloferax volcanii**. RNA. 2011;17:1367–1380. doi:10.1261/rna.2712811.
8. Bernick DL, Dennis PP, H¨ochsmann M, Lowe TM. **Discovery of Pyrobaculum small RNA families with atypical pseudouridine guide RNA features**. RNA. 2012;18:402–411. doi:10.1261/rna.031385.111.
9. Danan M, Schwartz S, Edelheit S, Sorek R. **Transcriptome-wide discovery of circular RNAs in archaea**. Nucleic Acids Res. 2012;40:3131–3142. doi:10.1093/nar/gkr1009.
10. Starostina NG, Marshburn S, Johnson LS, Eddy SR, Terns RM, Terns MP. **Circular box C/D RNAs in Pyrococcus furiosus**. Proc Natl Acad Sci USA. 2004;101:14097–14101.
11. Doose G, Alexis M, Kirsch R, Findeiß S, Langenberger D, Machn´e R, et al. **Mapping the RNA-seq Trash Bin: Unusual Transcripts in Prokaryotic Transcriptome Sequencing Data.** RNA Biology. 2013;10:1204–1210. doi:10.4161/rna.24972.
12. Thompson LD, Daniels CJ. **A tRNA (Trp) intron endonuclease from Halobacterium volcanii. Unique substrate recognition properties**. J Biol Chem. 1988;263:17951–17959.
13. Kjems J, Garrett RA. **Novel splicing mechanism for the ribosomal RNA intron in the archaebacterium**

Desulfurococcus mobilis. Cell. 1988;54:693–703. doi:10.1016/S0092-8674(88)80014-X.

14. Salgia SR, Singh SK, Gurha P, Gupta R. **Two reactions of Haloferax volcanii RNA splicing enzymes: Joining of exons and circularization of introns.** RNA. 2003;9:319–330. doi:10.1261/rna.2118203.

15. Heinemann IU, S¨ oll D, Randau L. **Transfer RNA processing in archaea: unusual pathways and enzymes**. FEBS Lett. 2010;584:303–309. doi:10.1016/j.febslet.2009.10.067.

16. Sugahara J, Yachie N, Arakawa K, Tomita M. **In silico screening of archaeal tRNA-encoding genes having multiple introns with bulge-helix-bulge splicing motifs.** RNA. 2007;13:671–681. doi:10.1261/rna.309507.

17. Yoshihisa T. **Handling tRNA introns, archaeal way and eukaryotic way.** Front Genet. 2014;5:213. doi:10.3389/fgene.2014.00213.

18. Randau L, M¨unch R, Hohn MJ, Jahn D, S¨ oll D. **Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5'-and 3'-halves**. Nature. 2005;433:537–541.

19. Randau L, S¨ oll D. **Transfer RNA genes in pieces.** EMBO Rep. 2008;9:623–628. doi:10.1038/embor.2008.101.

20. Fujishima K, Sugahara J, Kikuta K, Hirano R, Sato A, Tomita M, et al. **Tri-split tRNA is a transfer RNA made from 3 transcripts that provides insight into the evolution of fragmented tRNAs in archaea.** Proc Natl Acad Sci USA. 2009;106:2683–2687. doi:10.1073/pnas.0808246106.

21. Sugahara J, Fujishima K, Morita K, Tomita M, Kanai A. **Disrupted tRNA gene diversity and possible evolutionary scenarios.** J Mol Evol. 2009;69:497–504. doi:10.1007/s00239-009-9294-6.

22. Richter H, Mohr S, Randau L. **C/D box sRNA, CRISPR RNA and tRNA processing in an archaeon with a minimal fragmented genome.** Biochem Soc Trans. 2013;41:411–415. doi:10.1042/BST20120276.

23. Soma A. **Circularly permuted tRNA genes: their expression and implications for their physiological relevance and development.** Front Genet. 2014;5:63. doi:10.3389/fgene.2014.00063.

24. Tang TH, Rozhdestvensky TS, d'Orval BC, Bortolin ML, Huber H, Charpentier B, et al. **RNomics in Archaea reveals a further link between splicing of archaeal introns and rRNA processing.** Nucleic Acids Res. 2002;30:921–930. doi:10.1093/nar/30.4.921.

25. Yip WS, Vincent NG, Baserga SJ. **Ribonucleoproteins in Archaeal Pre-rRNA Processing and Modification.** Archaea. 2013;2013:614735. doi:10.1155/2013/614735.

26. Yokobori S, Itoh T, Yoshinari S, Nomura N, Sako Y, Yamagishi A, et al. **Gain and loss of an intron in a protein-coding gene in Archaea: the case of an archaeal RNA pseudouridine synthase gene.** BMC Evol Biol. 2009;9:198. doi:10.1186/1471-2148-9-198.

27. Su AAH, Tripp V, Randau L. **RNA-Seq analyses reveal the order of tRNA processing events and the maturation of C/D box and CRISPR RNAs in the hyperthermophile Methanopyrus kandleri.** Nucleic Acids Res. 2013;41:6250–6258. doi:10.1093/nar/gkt317.

28. Clouet d'Orval B, Bortolin ML, Gaspin C, Bachellerie JP. **Box C/D RNA guides for the ribose methylation of archaeal tRNAs: the tRNATrp intron guides the formation of two ribose-methylated nucleosides in the mature tRNATrp.** Nucleic Acids Res. 2001;29:4518–4529. doi:10.1093/nar/29.22.4518.

29. Marck C, Grosjean H. **Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea: evolutionary implications.** RNA. 2003;9:1516–1531. doi:10.1261/rna.5132503.

30. Watanabe Yi, Yoshinari S. **Intron and RNA splicing in Archaea.** Viva Orig. 2013;41:12–13.

31. Yoshinari S, Itoh T, Hallam SJ, DeLong EF, Yokobori Si, Yamagishi A, et al. **Archaeal pre-mRNA splicing: a connection to hetero-oligomeric splicing endonuclease.** Biochem Biophys Res Commun. 2006;346:1024–1032. doi:10.1016/j.bbrc.2006.06.011.

32. Tocchini-Valentini GD, Fruscoloni P, Tocchini-Valentini GP. **Coevolution of tRNA intron motifs and tRNA endonuclease architecture in Archaea.** Proc Natl Acad Sci USA. 2005;102:15418–15422. doi:10.1073/pnas.0506750102.

33. Ghosh Z, Chakrabarti J, Mallick B, Das S, Sahoo S, Sethi HS. **tRNA-isoleucine-tryptophan composite gene.** Biochem Biophys Res Commun. 2006;339:37–40. doi:10.1016/j.bbrc.2005.10.183.

34. Tocchini-Valentini GD, Fruscoloni P, Tocchini-Valentini GP. **Processing of multiple-intron-containing pretRNA.** Proc Natl Acad Sci USA. 2009;106(48):20246–20251. doi:10.1073/pnas.0911658106.

35. Chan PP, Cozen AE, Lowe TM. **Discovery of permuted and recently split transfer RNAs in Archaea.** Genome Biol. 2011;12(4):R38.

36. Yamazaki S, Yoshinari S, Kita K, Watanabe Y, Kawarabayasi Y. **Identification of an entire set of tRNA molecules and characterization of cleavage sites of the intron-containing tRNA precursors in acidothermophilic crenarchaeon Sulfolobus tokodaii strain7**. Gene. 2011;489:103–110. doi:10.1016/j.gene.2011.08.003.

37. Höner zu Siederdissen C, Berkemer S, Amman F, Wintsche A, Will S, Prohaska SJ, et al. **Comparative Detection of Processed Small RNAs in Archaea**. In: IWBBIO 2014. Univ. Granada; 2014. p. 286–297. http://iwbbio.ugr.es/2014/papers/IWBBIO_2014_paper_33.pdf.

38. Lowe TM, Eddy SR. **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** Nucleic Acids Res. 1997;25:955–964. doi:10.1093/nar/25.5.0955.

39. Randau L. **RNA processing in the minimal organism Nanoarchaeum equitans**. Genome Biol. 2012;13:R63. doi:10.1186/gb-2012-13-7-r63.

40. Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R. **A single-base resolution map of an archaeal transcriptome**. Genome Res. 2010;20:133–141. doi:10.1101/gr.100396.109.

41. Omer AD, Lowe TM, Russell AG, Ebhardt H, Eddy SR, Dennis PP. **Homologs of small nucleolar RNAs in Archaea.** Science. 2000;288:517–522. doi:10.1126/science.288.5465.517.

42. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. **Inferring Non-Coding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering**. PLoS Comput Biol. 2007;3:e65. doi:10.1371/journal.pcbi.0030065.

43. Sankoff D. **Simultaneous solution of the RNA folding, alignment, and proto-sequence problems**. SIAM J Appl Math. 1985;45:810–825. doi:10.1137/0145048.

44. Märtens B, Amman F, Manoharadas S, Zeichen L, Orell A, Albers SV, et al. **Alterations of the Transcriptome of Sulfolobus acidocaldarius by Exoribonuclease aCPSF2.** PLoS one. 2013;8:e76569. doi:10.1371/journal.pone.0076569.

45. Reimann J, Esser D, Orell A, Amman F, Pham TK, Noirel J, et al. **Archaeal Signal Transduction: Impact of Protein Phosphatase Deletions on Cell Size, Motility, and Energy Metabolism in Sulfolobus acidocaldarius**. Mol Cell Proteomics. 2013;12:3908–3923. doi:10.1074/mcp.M113.027375.

46. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, et al. **Fast mapping of short sequences with mismatches, insertions and deletions using index structures.** PLoS Comp Biol. 2009;5:e1000502. doi:10.1371/journal.pcbi.1000502.

47. Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, et al. **A multi-split mapping algorithm for circular RNA, splicing, trans-splicing, and fusion detection.** Genome Biol. 2014;15:R34. doi:10.1186/gb-2014-15-2-r34.

48. Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF. **RNAz 2.0: improved noncoding RNA detection.** Pac Symp Biocomput. 2010;15:69–79.

49. Eddy SR, Durbin R. **RNA sequence analysis using covariance models.** Nucleic Acids Res. 1994;22:2079–2088. doi:10.1093/nar/22.11.2079.

50. Nawrocki EP, Kolbe DL, Eddy SR. **Infernal 1.0: inference of RNA alignments.** Bioinformatics. 2009;25:1335–1337. doi:10.1093/bioinformatics/btp157.

51. Nawrocki EP, Eddy SR. **Infernal 1.1: 100-fold faster RNA homology searches.** Bioinformatics. 2013;29:2933–2935. doi:10.1093/bioinformatics/btt509.

52. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, et al. **Rfam: Wikipedia, clans and the “decimal” release.** Nucleic Acids Res. 2011;39:D141–D145. doi:10.1093/nar/gkq1129.

53. Kerpedjiev P, Hammer S, Hofacker IL. **forna (force-directed RNA): Simple and Effective Online RNA Secondary Structure Diagrams.** Bioinformatics. 2015; doi:10.1093/bioinformatics/btv37.

54. Bailey TL, Elkan C. **Fitting a mixture model by expectation maximization to discover motifs in bipolymers.** Proc Int Conf Intell Syst Mol Biol. 1994;2:28–36.

55. Wintsche A, Stadler PF, Prohaska SJ. **MEME-SP: de novo prediction of short RNA motifs**; 2014. In preparation.

56. Lorenz R, Bernhart SH, H¨oner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. **ViennaRNA Package 2.0.** Alg Mol Biol. 2011;6:26. doi:10.1186/1748-7188-6-26.

57. Watanabe Yi, Yokobori Si, Inaba T, Yamagishi A, Oshima T, Kawarabayasi Y, et al. **Introns in protein-coding genes in Archaea.** FEBS letters. 2002;510:27–30. doi:10.1016/S0014-5793(01)03219-7.

58. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. **RNAalifold: improved consensus structure prediction for RNA alignments.** BMC Bioinformatics. 2008;9:474. doi:10.1186/1471-2105-9-474.

59. Brooks MA, Meslet-Cladi´ere L, Graille M, Kuhn J, Blondeau K, Myllykallio H, et al. **The structure of an archaeal homodimeric ligase which has RNA circularization activity.** Protein Sci. 2008;17:1336–1345. doi:10.1110/ps.035493.108.

60. Giegerich R, H¨oner zu Siederdissen C. **Semantics and Ambiguity of Stochastic RNA Family Models.** IEEE/ACM Trans Comp Biol Bioinf. 2011;8:499–516. doi:10.1109/TCBB.2010.12.

61. Höner zu Siederdissen C, Hofacker IL, Stadler PF. **How to Multiply Dynamic Programming Algorithms.** In: Brazilian Symposium on Bioinformatics (BSB 2013). vol. 8213 of Lect. Notes Bioinf. Heidelberg: Springer; 2013. p. 82–93. doi:10.1007/978-3-319-02624-4.

62. Höner zu Siederdissen C, Hofacker IL, Stadler PF. **Product Grammars for Alignment and Folding.** IEEE/ACM Trans Comp Biol Bioinf. 2014;12:507–519. doi:10.1109/TCBB.2014.2326155.

63. Randau L, Pearson M, Söll D. **The complete set of tRNA species in Nanoarchaeum equitans**. FEBS letters. 2005;579:2945–2947. doi:10.1016/j.febslet.2005.04.051.

64. Sugahara J, Yachie N, Sekine Y, Soma A, Matsui M, Tomita M, et al. **SPLITS: a new program for predicting split and intron-containing tRNA genes at the genome level.** In Silico Biol. 2006;6:411–418.

ENDNOTES

[1] This contribution is an extended and updated version of a paper [37] accepted for the IWBBIO 2014 conference in Granada.