Mapping the RNA-Seq Trash Bin: Unusual Transcripts in Prokaryotic Transcriptome Sequencing Data

Gero Doose[a,1], Maria Alexis[a,1], Rebecca Kirsch[a], Sven Findeiß[d], David Langenberger[a,1], Rainer Machné[a,d], Mario Mörl[a], Steve Hoffmann[1], Peter F. Stadler[a,b,c,e,d,f]

[a] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
[b] *Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany*
[c] *Fraunhofer Institut für Zelltherapie und Immunologie – IZI Perlickstraße 1, D-04103 Leipzig, Germany*
[d] *Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*
[e] *Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark*
[f] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

Abstract

Prokaryotic transcripts constitute almost always uninterrupted intervals when mapped back to the genome. Split reads, i.e., RNA-seq reads consisting of parts that only map to discontiguous loci, are thus disregarded in most analysis pipelines. There are, however, some well-known exceptions, in particular tRNA splicing and circularized small RNAs in Archaea as well as self-splicing introns. Here, we reanalyze a series of published RNA-seq data sets, screening them specifically for non-contiguously mapping reads. We recover most of the known cases together with several novel archaeal ncRNAs associated with circularized products. In Eubacteria, only a handful of interesting candidates were obtained beyond a few previously described group I and group II introns. Most of the atypically mapping reads do not appear to correspond to well-defined, specifically processed products. Whether this diffuse background is, at least in part, an incidental by-product of prokaryotic RNA processing or whether it consists entirely of technical artefacts of reverse transcription or amplification remains unknown.

Keywords: RNA-seq, self-splicing introns, split tRNAs, circular sRNAs

1. Introduction

Common wisdom has it that prokaryotic transcripts correspond to intervals on the genomic DNA. In archaea, several exceptions to this simple rule are well known. As in eukaryotes, some of their tRNAs have introns that are spliced out by dedicated splicing endonucleases (1; 2). In contrast to Eukarya, enzymatically spliced introns can also be found in mRNAs (3) and in rRNAs (4). In some archaeal species, furthermore, tRNAs are composed of pieces that independently transcribed from different genomic locations (5; 2; 6; 7; 8).

Email addresses: gero@bioinf.uni-leipzig.de (Gero Doose), msalexis@stanford.edu (Maria Alexis), rebecca@bioinf.uni-leipzig.de (Rebecca Kirsch), sven@tbi.univie.ac.at (Sven Findeiß), david@bioinf.uni-leipzig.de (David Langenberger), raim@bioinf.uni-leipzig.de (Rainer Machné), moerl@uni-leipzig.de (Mario Mörl), steve@gmail.com (Steve Hoffmann), studla@bioinf.uni-leipzig.de (Peter F. Stadler)

Archaeal non-coding RNAs often are processed to yield a circular form. Large ORF-containing introns derived from rRNAs form stable RNA species in (9). Circular forms of both 23S and 16S rRNAs appear as processing intermediates during rRNA maturation (10). Circularized RNAs are produced from tRNA introns in (5), and a circularized box C/D snoRNA from *Pyrococcus furiosus* (11) turned out to be typical for box C/D snoRNAs, see also (8) for an example in *Nanoarchaeum equitans*. A recent study based on RNAse R treated RNA libraries systematically mapped circularized RNAs and showed that circularized RNAs are also abundant in *Sulfolobus solfataricus* and its relatives (12). In contrast to this rather complex situation in Archaea, Eubacterial transcriptomes are not known to harbour spliced transcripts with the exception of the hosts of self-splicing group I and group II introns (13; 14; 15).

It is well known, on the other hand, that reverse transcription can generate artefactual sequences that look

like splicing products, i.e., by leaving out stable RNA secondary structure features (16; 17; 18). Most analyses of prokaryotic RNA-seq data thus completely neglect sequencing reads that do not map as a single, uninterrupted interval. Of course, this strategy also hides any true splicing or circularization products. The purpose of this contribution is to systematically explore the content of the "trash bin" of RNA-seq analysis, aiming at the identification of atypically processed RNAs.

2. Methods

Sequence Data.. Publicly available RNA-seq data were downloaded from the short read archive for 4 Archea and 6 Eubacteria in the Electronic Supplement[1]. All these RNA-seq data were produced with non-strand-specific protocols. With the exception of the read data for *Escherichia coli* and *Salmonella enterica* all reads are single ended with length between 30 to 100 nts. According to requirements the raw reads where quality trimmed with `FASTX-Toolkit` and adapter clipped with `Cutadapt` (19).

Annotation.. Annotation sources are the GFF files for each analyzed species downloaded from the NCBI[2] and the Rfam[3] ftp servers, respectively. From the NCBI files all genes are extracted and the corresponding annotated elements, i.e., CDS, tRNA, and rRNA, are used. All genes that did not code for one of these elements are grouped into the separate class, "other". Since NCBI annotation files often miss non-coding RNAs (ncRNAs) and regulatory elements such as riboswitches, these were instead adopted from the Rfam gff files. A detailed summary of all used annotation items and their sources is provided in the Electronic Supplement.

Since the Rfam annotation did not feature well-known group I introns, we reasoned that either the Rfam seed alignment (RF00028) does not cover the diversity of bacterial group I introns, or the presence of open reading frames in these introns hampers infernal search. We therefore split the Rfam seed alignment as well as 14 alignments of group I subtypes (`http://www.rna.whu.edu.cn/gissd/`, (20)) into 27 overlapping blocks along the 5'→3' direction of the intron, constructed individual CMs, scanned the genomes for these sub-CMs and reconstructed potential group I introns by chaining adjacent hits in the correct order (non-overlapping 5' and 3' sub-CMs, and ≤5 kb distance between sub-CMs). The resulting 8 candidates, of which all but 2 have been described in literature (21; 22; 23), are listed in the Supplement.

For group II introns, we downloaded 35 intron sequences listed in the group II intron database (`http://webapps2.ucalgary.ca/~groupii/` on Feb 4th 2013, (24)) for different strains of the species considered here. Of these, 9 could be located in our reference genomes by `blastn`.

Read Mapping.. All reads were mapped with `segemehl`, version 0.1.4 (25; 26) with the split read option `-S`. Depending on the read length of RNA-seq data, the minimum fragment length *Z* and the minimum fragment score *U* we set to combinations from `-Z 20 -U 18` to (`-Z 14 -U 12`. These small values are motivated by the need to emphasize split reads. For all other parameters the default values were used. Reads that remained unmapped in the first pass were remapped with `Remapper`, a component of the `segemehl` suite.

Reads that were mapped with splits were assigned to one of three categories: "*normal*", same strand, same chromosom and insert between 15 nt and 200 kb, and matched fragments colinear with the genomic DNA; "*circular*", same strand, same chromosom and junction distance less than 200 kb with fragment order inverted relative to genomic DNA; "*(strand)switched*", same chromosom, junction distance less than 200 kb and fragments located on opposite strand. Splice sites determined by the read mapping were clustered with `haarz`, a component of the `segemehl` suite, to determine median split positions. The results of the mapping procedure are summarized in Tab. 1.

Overlaps between mapped reads and annotation data were computed with the help of `BEDTools` (27).

Analysis of ribosomal RNA loci.. To compare rRNA split read patterns across species we performed the following steps: (i) For each species operon structures have been defined based on the rRNA gene annotation. For bacteria 16S-23S-5S rRNA operons are used and 16S-23S rRNA operon in archaea. (ii) The sequences of the rRNA operons including 300 nt flanking sequence have been extracted from the corresponding genomes. In species with multiple copies of rRNA operons a clustalw alignment has been calculated and the consensus sequence extracted. Either the consensus sequence or the sequence of a unique encoded operon has been

[1] Machine readable data files and additional information can be found at `http://www.bioinf.uni-leipzig.de/supplements/12-002`.

[2] `ftp://ftp.ncbi.nih.gov/genomes/Bacteria/`

[3] `ftp://ftp.sanger.ac.uk/pub/databases/Rfam/11.0/genome.gff3.tar.gz`

Table 1: Summary of mapped reads.

Species	input	mappable	unsplit	split	read class		
					normal	circular	switch
Eubacteria							
Bacillus cereus	15,498,220	15,264,233	15,250,993	13,240	3,853	1,631	6,324
Escherichia coli	52,515,346	44,429,568	44,115,280	314,288	8,573	20,544	39,587
Salmonella enterica	31,924,568	27,752,771	27,737,761	15,010	543	2,481	4,421
Pseudomonas PA14	78,141,620	65,573,260	65,300,316	272,944	12,271	8,706	12,372
Helicobacter pylori 26695	82,847,902	40,152,294	39,146,732	1,005,562	17,930	53,709	77,095
Synechocystis PCC6803	31,985,927	15,080,656	15,031,302	49,354	39,956	165	250
Archaea							
Nanoarchaeum equitans	17,253,447	11,173,688	11,096,897	35,034	7,393	12,860	1,932
Ignicoccus hospitalis	17,253,447	5,302,517	5,181,769	76,039	6,254	39,994	3,656
Pyrococcus furiosus	6,449,461	8,691,213	8,474,477	216,736	11,536	54,795	17,095
Sulfolobus solfataricus	17,356,356	11,965,214	11,921,178	44,036	3,681	6,893	11,623

used as reference operon. The only exceptions are *N. equitans* and *H. pylori*. In *N. equitans* the 16S and 23S rRNA are transcribed from separate loci, in *H. pylori* the 16S rRNA and an operon comprising the 23S and 5S rRNAs. The separate parts were concatenated with an intervening stretch of 160 Ns as reference sequence. (iii) For each species all rRNA gene overlapping reads have been remapped onto the reference operon. Hence all rRNA reads are projected to a single locus for each species. (iv) To compare split-read patterns between species, the species specific reference operons where aligned using `clustalw` (28) and the mapped read coordinates transferred onto the alignment.

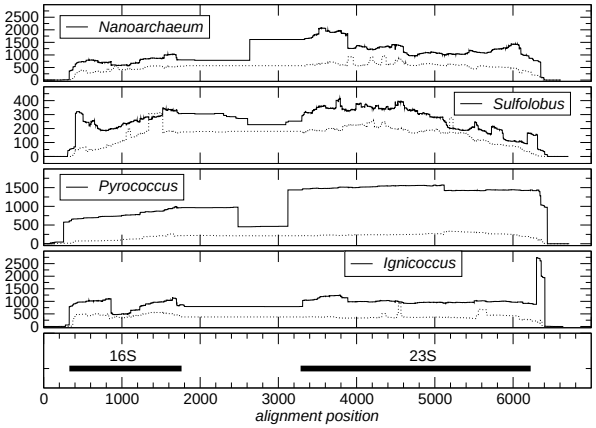

Figure 1: Density of circularized (thick line) and "spliced" reads (thin line) at the ribosomal rRNA loci. Coordinates refer to an multiple sequence alignment of the four Archaea species. For *N. equitans* the two separate RNAs genes are concatenated with 160 Ns are linker.

3. Results

3.1. Archaea

Several types of "atypical" RNAs are well known in Archaea. The most prominent form among them circularized RNAs. As expected we observe circularized precursor forms for both the 16S and 23S RNAs, see e.g. (10). In addition, large numbers of additional circularized products are observed, Fig. 1. The rRNA loci also feature substantial numbers of apparently spliced reads.

Most snoRNAs in Archaea also form circularized forms. Somewhat surprisingly, these are readily detectable from RNA-seq data even without prior treatment of the libraries to enrich circularized products as in the recent work of (12). The association of circularized products with small ncRNAs allows us to detect a number of novel ncRNA species in each of the four Archea, Tab. 2. The number of new candidates depends strongly on the species, presumably in response to the quality of the available genome annotation.

Enzymatic splicing to tRNAs is a well-known phenomenon in Archaea. It is typically invisible in RNA-seq data, however, because tRNAs are normally multi-copy genes and tRNAs with introns typically have nearly identical paralogs without an intron. In this situation, mature tRNAs are mapped to the intro-less locus even the molecule in reality was produced by splicing for the locus with intron. In 20 cases the intron is visible as a circularized by-product of splicing; a detailed table can be found in the Electronic Supplement.

In *Sulfolobus* an enzymatically spliced intron inter-

Table 2: Novel ncRNAs in Archaea. Since the RNA-seq data are not strand specific, the reading direction remains undetermined in most cases. Promoter or terminator elements annotated in the UCSC Archaea Browser identify a likely reading direction. Read support was added up for alternative junctions within a few nucleotides. In the Note column, 'mult.' designates the presence of multiple products, and ~ denotes loci adjacent to annotated ncRNAs.

Coordinates				Note
Pyrococcus furiosus				
128135	128190	?	8	
258945	259007	?	915	
505270	505323	?	4	
505760	505814	+	1	
860511	860567	?	3	
Sulfolobus solfataricus				
434665	434719	-	2	
1275505	1275576	?	71	3 variants
Nanoarchaeum equitans				
432130	432227	+	159	5' of 16S
396865	396957	+	95	3' of 23S
339418	339570	?	53	mult.
248142	248285	?	1	
Ignicoccus hospitalis				
28125	28202	?	883	
54013	54076	-	9	
62481	62544	?	3	
62543	62607	?	7	~ previous
69658	69725	?	411	
74304	74365	?	4	
507227	507289	?	112	
576736	576811	?	828	
598273	598363	?	41	
599309	599358	?	2	
617433	617521	+	2017	~ Iho-sR86
720628	720706	?	18	
734264	734345	+	20	3' of 23S
824008	824070	+	8	~ Iho-sR109
1000660	1000778	+	6	~ Iho-sR131
1000717	1000778	+	3	~ Iho-sR131
1066825	1066891	?	500	
1266699	1266795	?	461	mult.

rupts the coding sequence of the cbf5 gene (29). This case is readily detectable in the form multiple splitreads of the normal type. A second well-supported candidate is located close to the annotated translation start site of the putative protein SSO1586. With a length of 144 nt it preserves the reading frame. Since the entire sequence of the putative protein is conserved it might encode a functional isoform.

An interesting case of trans-splicing are the split tRNAs reported in *Nanoarchaeum* (30; 6; 8). Some of them are not directly observable as split reads, however. This is the case e.g. for tRNA-Lys and tRNA-Gln, which have nearly identical paralogs that attract the mature tRNA reads to the unspliced loci irrespective of their true origin. The tRNA-Met and tRNA-Glu are visible at least with a few reads, while tRNA-His is invisible. This is explained by that high conservation of tRNA genes and the fact that RNA-seq data used here comprise a mixture of *N. equitans* and *I. hospitalis*. The tRNA-His sequence in *I. hospitalis* thus captures the transspliced tRNA-His reads from *N. equitans*. No other transsplicing events supported by a larger number of reads was observed in the data sets analyzed here.

Table 3: Splice junction overlaps with crisper and rRNA.

Species	**all**	**crisper**	**rRNA**
Eubacteria			
Bacillus cereus	11,808	0	7,955
Escherichia coli	68,704	6	37,616
Salmonella enterica	7,445	0	6,050
Pseudomonas PA14	33,349	528	16,819
Helicobacter pylori 26695	148,734	0	114,388
Synechocystis PCC6803	40,371	7	786
Archaea			
Nanoarchaeum equitans	22,185	0	16,607
Ignicoccus hospitalis	49,904	0	12,615
Pyrococcus furiosus	83,426	23,544	23,502
Sulfolobus solfataricus	22,197	150	13,781

Given the high expression levels of rRNAs, it is not surprising that a large fraction (ranging from 25% in *Ignicoccus* to 75% in *Nanoarchaeum*) of split reads maps to the rRNA loci, see Tab. 3. The number of spliced reads nevertheless is systematically smaller than the number of reads crossing a circularization point, see Fig. 1 above.

Surprisingly, about a quarter of the split read data for *Pyrococcus* maps to the CISPR loci. It is tempting to speculate that inclusion of an organisms own sequences in CRISPRs is akin to an autoimmune reaction. Without further validation, however, we cannot rule out that artifacts in reverse transcription or amplification are responsible for these "trans-spliced" reads.

3.2. Eubacteria

In contrast to Archaea, split reads are expected to be rare in Eubacteria. In fact, the only well-understood source are self-splicing introns. In the 6 genomes considered here, 8 group I and 9 group II introns could be

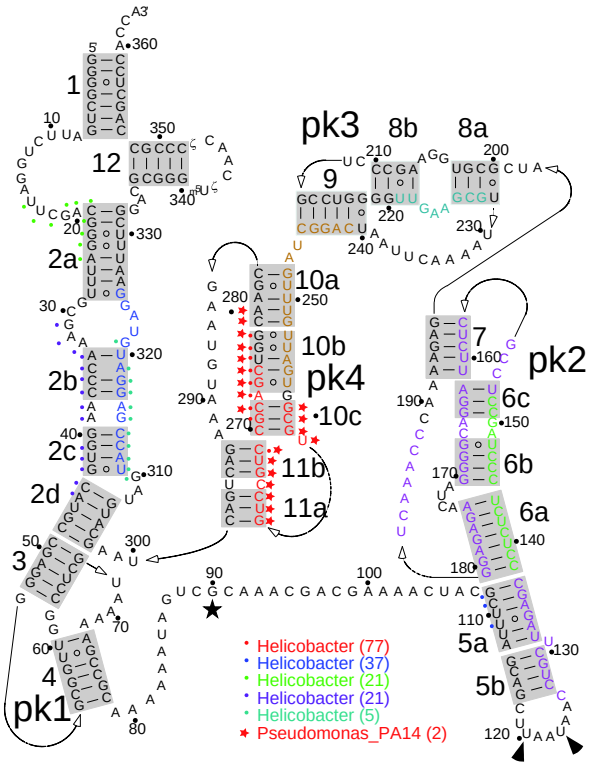

Figure 2: Mapping of "introns" observed in multiple reads from *Helicobacter* and *Pseudomonas* to the tmRNA structure (*E. coli* tmRNA model from (32)) shows that the excisions are concentrated in the pseudoknotted regions.

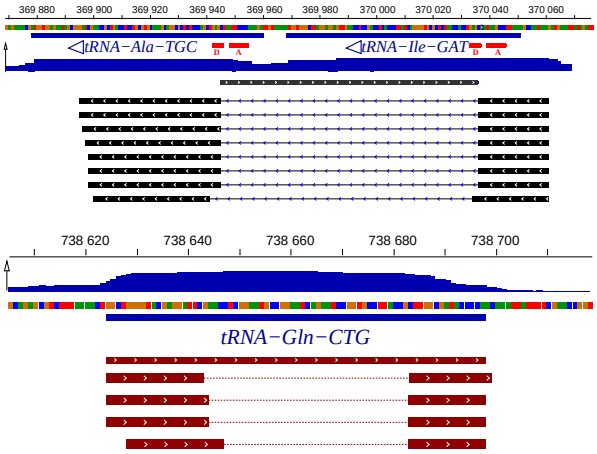

Figure 3: Unusual eubacterial reads associated with tRNAs. **Above:** Spliced fusion of two adjacent tRNAs in *H. pylori*. Red marks indicate the 5'-side of the acceptor stem and D-stem, resp. The apparent intron extends roughly from the end of the acceptor stem of tRNA-Ile to the begin of the D-stem of tRNA-Ala. The coverage suggests that the two adjacent tRNAs are produced from a signle primary transcript. **Below:** A circularized tRNA in *Salmonella*.

tentatively annotated computationally. Not all of them are visible in the RNA-seq data in the form of split-reads. Only a group I intron in the initator-tRNA of *Synechocystis* (21) and a group I intron in the recA gene of *B. cereus* (31; 23) are well represented in our data. All of the detectable group II introns are located in *B. cereus* (22; 23). Only the B.c.I3 intron located within the DNA polymerase III subunit α is supported by many split reads. The two plasmid-borne introns designated B.c.I4 and B.c.I5 are visible as a single split read each. More details on the self-splicing introns can be found in the Electronic Supplement.

Surprisingly, our mapping data also show a large number of split and circularized reads that cannot be explained by known splicing mechanisms. As in Archaea, a large fraction of the split reads again maps to the rRNA operons, Tab. 3. With the exception of *Synechocystis*, rRNA accounts for the dominating part of the unusual RNAs. We have not been able to isolate candidates for well-defined stable processing products, however.

Beyond the self-spicing introns and the rRNA loci only a moderate number of "splice sites" is supported by multiple, non-identical reads. Among the most peculiar examples are tmRNAs with missing subsequences, Fig. 2, which appear in several species. Although the excisions appear to be concentrated in the highly structured, pseudoknotted regions, only some of them are easily explained as "RTfact" resulting from the RT reading through the base of stem and thus omitting the entire structural domain enclosed by the stem.

Cleavage of tRNAs as a response to stress, first discovered as response to phage infection in *E. coli* (33; 34), is a general phenomenon in all domains of life, see e.g. (35; 36; 37). At least in some cases, tRNA cleavage seems to have evolved into an interal regulation mechanism (38). Fragments of tRNAs, furthermore, may act as regulatory ncRNAs in both Eukarya (39; 40) and Archaea (41). Healing of the cleaved tRNAs is likewise a frequently observed phenomenon, see e.g. (42; 43). The ligases involved in tRNA splicing in Eukarya (44) and Archaea (45) utilize the 2',3'-cyclic phosphates generated by endonucleolytic cleavage. Members of the same protein family have also been found in Eubacteria, see (46) for a recent review of RNA ligases The *E. coli* ligase RtcB, a component of the RNA repair operon reseals tRNAs cleaved in the anticodon loop (43). It has been shown to be capable of catalyze tRNAs splicing in yeast (47). It is not unreasonable to assume, therefore, that unexpected tRNA-derived RNAs, including "trans-splicing" products ap-

pear as by-products of the the tRNA cleavage/repair pathways, and hence are present in the cell. In *Helicobacter*, for example, we find what lookes like a spliced common precursor of two adjacent tRNAs, see Figure 3. In *Salmonella*, the matured tRNA-Gln-CTG is associated with circularized reads.

4. Discussion

The preparation of RNA-seq libraries contains a reverse transcription step that may account for many of the observed non-canonical splicing events. Such RT artifacts have investigated in detail e.g. in (16; 17; 18). While we cannot rule out in most cases that the observed reads are such "RTfacts", there are plausibe alternative mechanism that could produce atypical transcript structures.

On other hand, the data contain a large number of true positive examples for both Archaea and Eubacteria in which splicing or circularization has been demonstrated in independent experiments. Hence clearly not all of the observed split reads are technical artefacts. In some cases, the molecular mechanisms the lead to the "spliced" RNAs is well known. This is the case for the self-splicing introns and for the processing of tRNAs (48) and rRNAs (10) in Archaea. The splicing endonuclease processing in Archaea has a broad range of target and is known to be involved also the transsplicing of tRNAs from indepedently encoded fragments as well as in the splicing of mRNAs. Homologous enzymes are present also in diverse eubacterial species, where they form a tRNA cleavage/repair pathway (briefly reviewed in the previous section). Thus there appears to be an ancient RNA repair system present all domains of life, which could account for many or even most of the spliced and circularized RNAs observed here.

In *E. coli*, the stress-induced toxin MazF cleaves certain single-stranded mRNA at or closely upstream of the start codon and removes a 43 nt fragment comprises the anti-Shine-Dalgarno from the 3' terminus of the 16S rRNA (49). Ribosomes with the truncated 16S rRNA specifically translate leaderless mRNAs, presumably as a stress response (50). The abundance of leaderless transcripts also in other proteobacteria (51; 52) might similar mechanisms are more wide-spread. In conjunction with a variety of RNA ligases (46), they might account for at least a part of the atypical sequences observed here.

Apparent splice junctions that are supported by multiple read counts, thus, are at least good candidates atypically processed RNAs that deserve further attention. In Archaea, the combination of atypical reads and a local, (nearly) isolated peak of coverage provides at least a very strong indication of for processed ncRNAs. In all four Archaea considered here, additional candidates, Tab. 2, could be identified.

On the other hand, several well-described cases of atypical transcripts, such as the trans-spliced tRNAs in *Nanoarchaeum* were observed only in a very small number of reads. This can be explained only in part by the presence of unspliced paralogs that attract the processed reads to the contiguous locus in the mapping procedure because an unspliced alignment is always preferred over a spliced one. Low expression, or support by only a small number of splice junctions, thus does not necessarily imply that an atypical transcript is a technical artefact or, even if present in the cell, devoid of biological function.

Acknowledgements

This work was supported in part by the German Research Foundation (STA 850/7-2, under the auspicies of SPP-1258 "Sensory and Regulatory RNAs in Prokaryotes"), more funding and other acknowledgments!

References

[1] Marck C, Grosjean H. Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea: evolutionary implications. RNA 9, 2003:1516–1531.

[2] 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 13, 2007:671–681.

[3] Yoshinari S, et al. Archaeal pre-mRNA splicing: a connection to hetero-oligomeric splicing endonuclease. Biochem Biophys Res Commun 346, 2006:1024–1032.

[4] Tocchini-Valentini GD, Fruscoloni P, Tocchini-Valentini GP. Evolution of introns in the archaeal world. Proc Natl Acad Sci USA 108, 2011:4782–4787.

[5] Salgia SR, Singh SJ, Gurha P, Gupta R. Two reactions of *Haloferax volcanii* RNA splicing enzymes: Joining of exons and circularization of introns. RNA 9, 2003:319–330.

[6] Randau L, Söll D. Transfer RNA genes in pieces. EMBO Rep 9, 2008:623–628.

[7] Fujishima K, 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 106, 2009:2683–2687.

[8] Randau L. RNA processing in the minimal organism *Nanoarchaeum equitans*. Genome Biol 13, 2012:R63.

[9] Dalgaard JZ, Garrett RA. Protein-coding introns from the 23S rRNA-encoding gene form stable circles in the hyperthermophilic archaeon *Pyrobaculum organotrophum*. Gene 121, 1992:103–110.

[10] Tang TH, et al. RNomics in Archaea reveals a further link between splicing of archaeal introns and rRNA processing. Nucleic Acids Res 30, 2002:921–930.

[11] 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 101, 2004:14097–14101.

[12] Danan M, Schwartz S, Edelheit S, Sorek R. Transcriptome-wide discovery of circular RNAs in Archaea. Nucleic Acids Res 40, 2012:3131–3142.

[13] Cech TR. Self-splicing of group i introns. Annual Review of Biochemistry 59, 1990:543–568. doi:10.1146/annurev.bi.59.070190.002551. PMID: 2197983.

[14] Nielsen H, Johansen SD. Group I introns: Moving in new directions. RNA Biology 6, 2009:375–383.

[15] Edgell DR, Chalamcharla VR, Belfort M. Learning to live together: mutualism between self-splicing introns and their hosts. BMC Biol 9, 2011:22.

[16] Cocquet J, Chong A, Zhang G, Veitia RA. Reverse transcriptase template switching and false alternative transcripts. Genomics 88, 2006:127–131.

[17] Roy SW, Irimia M. When good transcripts go bad: artifactual RT-PCR 'splicing' and genome analysis. Bioessays 30, 2008:601–605.

[18] Houseley J, Tollervey D. Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. PLoS ONE 5, 2010:e12271.

[19] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 17, 2011.

[20] Zhou Y, et al. GISSD: Group I intron sequence and structure database. Nucleic Acids Res 36, 2008:D31–37.

[21] Biniszkiewicz D, Cesnaviciene E, Shub DA. Self-splicing group I intron in cyanobacterial initiator methionine tRNA: evidence for lateral transfer of introns in bacteria. EMBO J 13, 1994:4629–4635.

[22] Tourasse NJ, Stabell FB, Reiter L, Kolstø AB. Unusual group II introns in bacteria of the *Bacillus cereus* group. J Bacteriol 187, 2005:5437–5451.

[23] Tourasse NJ, Kolsto AB. Survey of group I and group II introns in 29 sequenced genomes of the *Bacillus cereus* group: insights into their spread and evolution. Nucleic Acids Res 36, 2008:4529–4548.

[24] Candales MA, et al. Database for bacterial group II introns. Nucleic Acids Res 40, 2012:D187–D190.

[25] Hoffmann S, et al. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comp Biol 5, 2009:e1000502.

[26] Hoffmann S, et al. A multi-split mapping algorithm for splicing, trans-splicing, and fusion detection in single-end reads 2012. Submitted.

[27] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 2010:841–842.

[28] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22, 1994:4673–4680.

[29] Yokobori S, 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 9, 2009:198.

[30] Randau L, Munch R, Hohn MJ, Jahn D, Söll D. *Nanoarchaeum equitans* creates functional tRNAs from separate genes for their 5'- and 3'-halves. Nature 433, 2005:537–541.

[31] Ko M, Choi H, Park C. Group I self-splicing intron in the recA gene of *Bacillus anthracis*. J Bacteriol 184, 2002:3917–3922.

[32] Zwieb C, Gorodkin J, B K, Burks J, Wower J. tmRDB (tmRNA database). Nucleic Acids Res 31, 2003:446–447.

[33] David M, Borasio GD, Kaufmann G. Bacteriophage T4-induced anticodon-loop nuclease detected in a host strain restrictive to RNA ligase mutants. Proc Natl Acad Sci USA 79, 1982:7097–7101.

[34] Amitsur M, Levitz R, Kaufmann G. Bacteriophage T4 anticodon nuclease, polynucleotide kinase and RNA ligase reprocess the host lysine tRNA. EMBO J 6, 1987:2499–2503.

[35] Saikia M, et al. Genome-wide identification and quantitative analysis of cleaved tRNA fragments induced by cellular stress. J Biol Chem 287, 2012:42708–42725.

[36] Thompson DM, Lu C, Green PJ, Parker R. tRNA cleavage is a conserved response to oxidative stress in eukaryotes. RNA 14, 2008:2095–2103.

[37] Thompson DM, Parker R. Stressing out over tRNA cleavage. Cell 138, 2009:215–219.

[38] Jöchl C, et al. Small ncRNA transcriptome analysis from *Aspergillus fumigatus* suggests a novel mechanism for regulation of protein-synthesis. Nucleic Acids Res 36, 2008:2677–2689.

[39] Li Y, et al. Stress-induced tRNA-derived RNAs: a novel class of small RNAs in the primitive eukaryote *Giardia lamblia*. Nucleic Acids Res 36, 2008:6048–6055.

[40] Lee YS, Shibata Y, Malhotra A, Dutta A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). Genes Dev 23, 2009:2639–2649.

[41] Gebetsberger J, Zywicki M, Künzi A, Polacek N. tRNA-derived fragments target the ribosome and function as regulatory non-coding RNA in *Haloferax volcanii*. Archaea 2012, 2012:260909.

[42] Keppetipola N, Nandakumar J, Shuman S. Reprogramming the tRNA-splicing activity of a bacterial RNA repair enzyme. Nucleic Acids Res 35, 2007:3624–3630.

[43] Tanaka N, Shuman S. RtcB is the RNA ligase component of an *Escherichia coli* RNA repair operon. J Biol Chem 286, 2011:7727–7731.

[44] Konarska M, Filipowicz W, Gross HJ. RNA ligation via 2'-phosphomonoester, 3'5'-phosphodiester linkage: Requirement of 2',3'-cyclic phosphate termini and involvement of a 5'-hydroxyl polynucleotide kinase. Proc Natl Acad Sci USA 79, 1982:1474–1478.

[45] Englert M, Sheppard K, Aslanian A, Yates III JR, Söll D. Archaeal 3'-phosphate RNA splicing ligase characterization identifies the missing component in tRNA maturation. Proc Natl Acad Sci USA 108, 2011:1290–1295.

[46] Popow J, Schleiffer A, Martinez J. Diversity and roles of (t)RNA ligases. Cell Mol Life Sci 69, 2012:2657–2670.

[47] Tanaka N, Meineke B, Shuman S. RtcB, a novel RNA ligase, can catalyze tRNA splicing and HAC1 mRNA splicing *in vivo*. J Biol Chem 286, 2011:30253–30257.

[48] Heinemann IU, Söll D, Randau L. Transfer RNA processing in archaea: unusual pathways and enzymes. FEBS Lett 584, 2010:303–309.

[49] Vesper O, et al. Selective translation of leaderless mRNAs by specialized ribosomes generated by MazF in *Escherichia coli*. Cell 147, 2011.

[50] Sci TB. Selective translation during stress in *Escherichia coli*. Moll, I and Engelberg-Kulka, H 37, 2012:493–498.

[51] Sharma CM, et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 464, 2010:250–255.

[52] Schmidtke C, et al. Genome-wide transcriptome analysis of the plant pathogen *Xanthomonas* identifies sRNAs with putative virulence functions. Nucleic Acids Res 40, 2012:2020–2031.