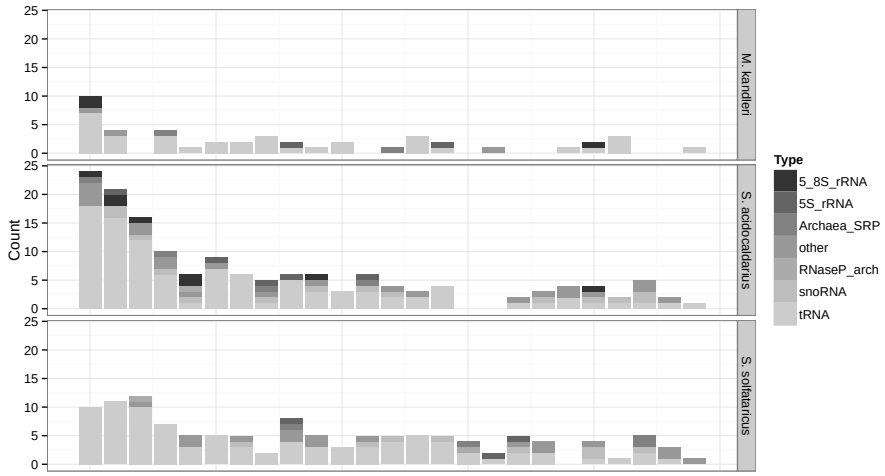
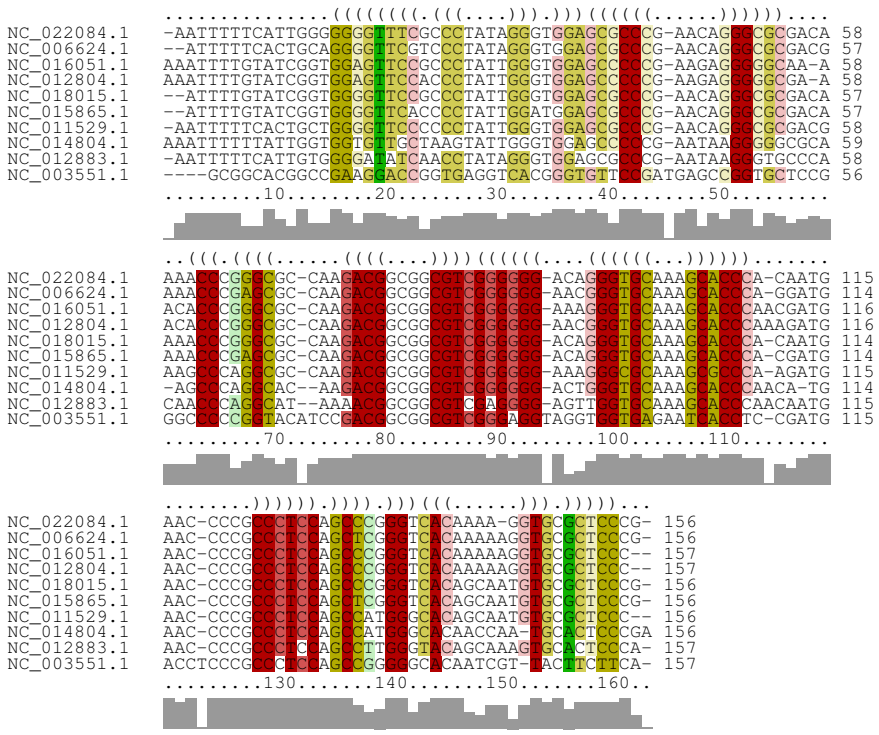
Comparative ncRNA Detection in Archaea

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

[1] Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria.
[2] Bioinformatics Group, Department of Computer Science University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany.
[3] Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany.
[4] Computational EvoDevo Group, Department of Computer Science, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany.
[5] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany.
[6] Fraunhofer Institut für Zelltherapie und Immunologie, Perlickstraße 1, D-04103 Leipzig, Germany.
[7] Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark.
[8] Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501.
E-mail:studla@bioinf.uni-leipzig.de



Supplementary Figure 1. For the top 2500 hits in each of the cm searches with our BHB model, the corresponding region (±250 nt) were searched for fitting to the description of a ncRNA from Rfam database [1]. The plot shows a histogram with bin size 100 for how many cm search are locally associated with a ncRNA, e.g. for *M. kandleri* top 100 ranked search hits, 7 are associated with tRNA, 1 with other (in this case a flpD motif), and two correspond to 5_8S_rRNA. The best hits are indeed enriched in association with ncRNA, whereas tRNA are predominant.



Supplementary Figure 2. Novel putative circRNA from *M. kandleri* (at 1500955-1501112). Alignment and consensus RNA secondary structure with homolog sequences in other archaea; the homologs were identified by blast search at e-value cut-off 0.01 as described in the main text. The figure furthermore reports the genome accession codes of the homolog sequences. The consensus structure and the output figure were generated using `RNAalifold` [2].

Supplementary Table 1. Comparison between tRNA introns accoring to tRNAscan results for *Methanopyrus kandleri*, *Sulfolobus solfataricus*, and *Sulfolobus acidocaldarius* and cm search results.

Species	tRNAscan			cm Search	
	tRNA Type	Intron Begin	Intron End	Rank	Bit Score
Methanopyrus kandleri	Trp	55,108	55,183	5.	21.5
Methanopyrus kandleri	Pro	1,499,308	1,499,322	124.	14.7
Methanopyrus kandleri	Pseudo	1,659,640	1,659,691	2.	25.3
Methanopyrus kandleri	Phe	1,639,150	1,639,119	not found	
Methanopyrus kandleri	Cys	1,062,337	1,062,317	79.	15.4
Methanopyrus kandleri	Asn	881,764	881,738	3.	25.2
Methanopyrus kandleri	Met	382,127	382,092	4.	22.2
Sulfolobus solfataricus	Asn	49,381	49,394	23.	12.5
Sulfolobus solfataricus	Met	466,263	466,279	209.	9.1
Sulfolobus solfataricus	Leu	637,204	637,218	not found	
Sulfolobus solfataricus	Leu	837,058	837,073	146.	10.0
Sulfolobus solfataricus	Ile	913,737	913,726	19.	12.6
Sulfolobus solfataricus	Pro	898,333	898,313	242.	8.9
Sulfolobus solfataricus	Thr	789,727	789,713	172.	9.4
Sulfolobus solfataricus	Tyr	642,512	642,500	not found	
Sulfolobus solfataricus	Ser	641,001	640,978	801.	6.7
Sulfolobus solfataricus	Arg	290,939	290,927	566.	7.3
Sulfolobus solfataricus	Arg	249,046	249,032	219.	9.0
Sulfolobus solfataricus	Thr	206,385	206,373	156.	9.7
Sulfolobus solfataricus	Met	184,841	184,817	13.	13.3
Sulfolobus solfataricus	Lys	138,407	138,386	1420.	5.7
Sulfolobus solfataricus	Lys	122,617	122,595	2.	17.3
Sulfolobus solfataricus	Trp	72,831	72,767	1321.	5.9
Sulfolobus acidocaldarius	Ser	512,669	512,693	62.	9.6
Sulfolobus acidocaldarius	Leu	512,819	512,833	5.	14.4
Sulfolobus acidocaldarius	Met	515,240	515,257	140.	8.3
Sulfolobus acidocaldarius	Lys	608,795	608,816	1.	16.7
Sulfolobus acidocaldarius	Pro	1,096,684	1,096,704	145.	8.3
Sulfolobus acidocaldarius	Met	1,166,860	1,166,879	10.	13.4
Sulfolobus acidocaldarius	Asn	2,181,266	2,181,254	129.	8.5
Sulfolobus acidocaldarius	Gly	2,160,121	2,160,107	988.	5.5
Sulfolobus acidocaldarius	Arg	1,241,011	1,240,995	4.	14.8
Sulfolobus acidocaldarius	Thr	1,188,440	1,188,425	73.	9.4
Sulfolobus acidocaldarius	Leu	716,510	716,493	12.	13.0
Sulfolobus acidocaldarius	Cys	610,584	610,569	not found	
Sulfolobus acidocaldarius	Lys	607,184	607,157	11.	13.0
Sulfolobus acidocaldarius	Thr	563,576	563,550	8.	13.9
Sulfolobus acidocaldarius	Phe	458,889	458,872	not found	
Sulfolobus acidocaldarius	Gly	458,680	458,666	not found	
Sulfolobus acidocaldarius	Arg	138,765	138,749	9.	13.7
Sulfolobus acidocaldarius	Trp	49,256	49,197	21.	11.4

Supplementary Table 2. Comparison between circularized RNA according to RNA-seq read analysis and predicted BHB elements for *Methanopyrus kandleri*. The first two columns give the genomic position of the left and right circularizing bases. "Read Count" gives the number of reads supporting this particular circularization event. For each locus, which was reported to be associated with an BHB element, the Rank in the genomic screen and its bit score is provided. The last column describes the genomic neighborhood. If it is within an annotated gene, its locus tag is given. For loci in intergenic regions the distance to the upstream and downstream gene is given.

RNA-seq			cm Search		Genomic Surrounding
L. junc.	R. junc.	#Count	Rank	Bit Score	ncbi locus tag
69,921	69,985	2,065		–	MK0074 ⇐ 25nt 77nt⇒ MK0075
91,822	91,904	2,992	7,845.	8.1	⇐ MK0099 ⇒
205,318	205,387	12,182		–	MK0213 ⇐ 21nt 102nt⇒ MK0214
219,317	219,379	2,412		–	⇐ MK0233 ⇒
271,148	271,216	10,072	3,546.	9.8	⇐ MK0280 ⇒
361,063	361,125	1,809		–	⇐ MK0371 ⇒
384,798	384,945	634		–	MK0403 ⇐ 665nt 201nt⇒ MK0404
459,461	459,532	9,830	9,039.	7.7	⇐ MK0498 ⇒
519,288	519,358	22,145	15,966.	6.1	⇐ MK0556 ⇒
520,778	520,845	13,484		–	MK0557 ⇐6nt 269nt⇒ MK0558
755,163	755,226	304		–	⇐ MK0794 ⇒
790,521	790,585	1,161	19,971.	5.2	MK0830 ⇐ 6nt 142nt⇒ MK0831
993,172	993,238	50,802		–	⇐ MK1033 ⇒
1,104,263	1,104,330	6,223	17,300.	5.7	MK1128 ⇐ 573nt 260nt⇒ MK1129
1,238,104	1,238,178	9	4,630.	9.2	⇐ MK1253 ⇒
1,417,298	1,417,370	9,371		–	MK1390 ⇐ 36nt 224nt⇒ MK1391
1,417,378	1,417,443	5,658	14,942.	6.3	MK1390 ⇐ 116nt 144nt⇒ MK1391
1,444,847	1,444,927	2		–	⇐ MK1415 ⇒
1,500,955	1,501,112	2,648		–	MK1479 ⇐ 88nt 205nt⇒ MK1480
1,506,611	1,506,673	8,396	6,984.	8.4	⇐ MK1486 ⇒

Supplementary Table 3. Circulare RNA in *Sulfolobus solfataricus* [3] are tested for recovery in the cm screen using the consensus model in `glocal` mode. Additionally, the analysis was redone using the homology loci, if available, in *Sulfolobus acidocaldarius*. The homology search eas conducted with the `GotohScan` program [4]. The "Start" and "End" columns refer to position in the genomes NC_002754 and NC_007181, respectively.

	Sulfolobus solfataricus				Sulfolobus acidocaldarius			
	RNA-seq		cm Search		RNA-seq		cm Search	
Name	Start	End	Rank	Bit Score	Start	End	Rank	Bit Score
5S rRNA/SSOr02	77,945	78,067	863.	6.5	1,293,914	1,294,035	–	
16S rRNA/SSOr03	871,658	873,216	–		1,108,641	1,107,094	–	
23S rRNA/SSOr04	873,334	876,429	–		1,106,947	1,103,875	–	
tRNA-Trp/SSOt04	72,767	72,831	1,321.	5.9	49,197	49,262	548.	6.4
tRNA-Lys/SSOt07	138,386	138,407	1,420.	5.7	607,138	607,204	11.	13.0
tRNA-Met/SSOt11	184,817	184,841	13.	13.3	–	–	–	
tRNA-Pro/SSOt42	898,313	898,333	242.	8.9	1,096,702	1,096,684	145.	8.3
tRNA-Ser/SSOt33	640,978	641,001	1,094.	6.2	512,691	512,669	62.	9.6
C/D box sR106	285,707	285,760	–		2,179,509	2,179,560	–	
C/D box Sso-180	362,308	362,369	–		669,556	669,612	–	
C/D box sR133	442,392	442,417	–		–	–	–	
C/D box sR102	563,241	563,296	–		1,388,934	1,388,984	–	
C/D box Sso-sR8	647,783	647,833	601.	7.1	1,885,917	1,885,967	–	
C/D box Sso-sR4	666,143	666,186	1,779.	5.3	–	–	–	
C/D box Sso-sR10	794,186	794,240	–		1,152,443	1,152,394	–	
C/D box Sso-207	816,021	816,075	–		–	–	–	
C/D box SSOs02	829,352	829,405	–		1,117,732	1,117,685	–	
C/D box Sso-sR12	2,189,397	2,189,456	–		–	–	–	
C/D box sR105	2,237,915	2,237,962	–		217,040	217,087	–	
H/ACA box sR109	59,5510	595,579	308.	8.3	458,983	459,052	351.	7.0
ncRNA	442,786	442,854	–		–	–	–	
ncRNA	722,538	722,578	3005.	4.4	417,691	417729	–	
Sso-117	1,576,633	1,576,671	–		–	–	–	
Sso-109	1,927,228	1,927,258	–		–	–	–	
7S rRNA/SSOr01	49,977	50023	839.	6.6	72,370	72,326	763.	5.9
Sso-214	105,148	105,181	–		–	–	–	
RNase P	224,732	224,765	–		586,242	586,211	–	
Sso-83	581,818	581,860	–		–	–	–	
ncRNA	1,275,500	127,5567	–		–	–	–	
SSO0393	343,138	343,264	–		650,082	650,206	–	
Intergenic region	871,573	871657	1,107.	6.2	1,108,730	1,108,647	–	
Intergenic region	873,215	873331	–		1,107,091	1,106,972	–	
SSO0389	335,563	335,635	–		–	–	–	
SSO0845	725,923	726,085	–		–	–	–	
SSO2359	2,154,297	2,154,322	–		782,915	782,892	–	
SSO2619	2,385,872	2,385,901	–		738,654	738,681	–	
SSO2642	2,404,146	2,404,146	–		2,114,638	2,114,694	–	

Supplementary Table 4. circRNA candidates of *M. kandleri* and *S. acidolarius* with putatively conserved stable secondary structures as predicted by `RNAz`. As described in the main text, circRNA candidates were identified by mapping RNA-Seq data, homologs were located in all archaeal genomes, potential homologs were aligned, and subsequently evaluated by `RNAz`. The table lists the candidates that are predicted as putative structural RNAs together with the number of homologous sequences in the locus alignment and the assigned RNAz class probability.

circRNA candidate locus			Number of	RNAz class
Accession	Start	Stop	Seqs	probability
NC_003551.1	1500955	1501112	10	0.9992
NC_007181.1	1214595	1214683	4	0.680306
NC_007181.1	1254692	1254799	4	0.938321
NC_007181.1	1107137	1107281	39	0.998730
NC_007181.1	1803656	1803770	3	0.722219
NC_007181.1	183648	183733	5	0.703805
NC_007181.1	1995955	1996059	3	0.619096
NC_007181.1	553923	554080	4	0.799874
NC_007181.1	753148	753230	4	0.982281
NC_007181.1	766362	766509	4	0.646306
NC_007181.1	773268	773364	4	0.644060
NC_007181.1	86425	86509	3	0.554138

References

1. Gardner, P.P., Daub, J., Tate, J., Moore, B.L., Osuch, I.H., Griffiths-Jones, S., Finn, R.D., Nawrocki, E.P., Kolbe, D.L., Eddy, S.R., et al.: Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Research **39**(suppl 1) (2011) D141–D145
2. Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R., Stadler, P.F.: RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics **9** (2008) 474
3. Danan, M., Schwartz, S., Edelheit, S., Sorek, R.: Transcriptome-wide discovery of circular RNAs in archaea. Nucleic Acids Res. **40** (2012) 3131–3142
4. Hertel, J., de Jong, D., Marz, M., Rose, D., Tafer, H., Tanzer, A., Schierwater, B., Stadler, P.F.: Non-coding RNA annotation of the genome of *Trichoplax adhaerens*. Nucleic acids research **37**(5) (2009) 1602–1615