Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*

Thean-Hock Tang*, Jean-Pierre Bachellerie†‡, Timofey Rozhdestvensky*, Marie-Line Bortolin†, Harald Huber§, Mario Drungowski¶, Thorsten Elge‖, Jürgen Brosius*, and Alexander Hüttenhofer*‡

*Institute of Experimental Pathology, Von-Esmarch-Strasse 56, 48149 Münster, Germany; †Laboratoire de Biologie Moléculaire Eukaryote du Centre National de la Recherche Scientifique, Université Paul-Sabatier, 31062 Toulouse, France; §Lehrstuhl für Mikrobiologie, Universität Regensburg, Universitätsstrasse 31, 93053 Regensburg, Germany; ¶Max-Planck-Institut für Molekulare Genetik, Harnackstrasse 23, 14195 Berlin, Germany; and ‖RessourcenZentrum/PrimärDatenbank Deutsches Ressourcenzentrum für Genomforschung GmbH, Heubnerweg 6, 14059 Berlin, Germany

Edited by Norman R. Pace, University of Colorado, Boulder, CO, and approved April 9, 2002 (received for review January 28, 2002)

In a specialized cDNA library from the archaeon *Archaeoglobus fulgidus* we have identified a total of 86 different expressed RNA sequences potentially encoding previously uncharacterized small non-messenger RNA (snmRNA) species. Ten of these RNAs resemble eukaryotic small nucleolar RNAs, which guide rRNA 2′-*O*-methylations (C/D-box type) and pseudouridylations (H/ACA-box type). Thereby, we identified four candidates for H/ACA small RNAs in an archaeal species that are predicted to guide a total of six rRNA pseudouridylations. Furthermore, we have verified the presence of the six predicted pseudouridines experimentally. We demonstrate that 22 snmRNAs are transcribed from a family of short tandem repeats conserved in most archaeal genomes and shown previously to be potentially involved in replicon partitioning. In addition, four snmRNAs derived from the rRNA operon of *A. fulgidus* were identified and shown to be generated by a splicing/processing pathway of pre-rRNAs. The remaining 50 RNAs could not be assigned to a known class of snmRNAs because of the lack of known structure and/or sequence motifs. Regarding their location on the genome, only nine were located in intergenic regions, whereas 33 were complementary to an ORF, five were overlapping an ORF, and three were derived from the sense orientation within an ORF. Our study further supports the importance of snmRNAs in all three domains of life.

The majority of genes from any given genome are transcribed into messenger RNAs (mRNAs). Their ORFs serve as a major criterion for identification of protein-coding genes. Within any genome, however, a considerable number of genes are transcribed into RNAs that are devoid of ORFs and therefore are not translated into proteins (for reviews see refs. 1–5). They can be long, resemble mRNAs but lacking ORFs, or represent small non-mRNAs (snmRNAs). These RNAs exert a variety of biological functions in housekeeping, including translation or splicing, as well as participate in many different aspects of metabolism or developmental controls. Their functions often are mediated by defined secondary/tertiary structures and/or intermolecular base pairings and carried out in association with specific proteins, within ribonucleoprotein particles.

In contrast to mRNAs, computational identification of snmRNAs is hampered by a lack of ORFs. Instead, one can rely only on known conserved sequence/structural motifs. Hence, the computational approach is unlikely to yield *per se* the complete set of snmRNAs within a given organism. Recently, several previously uncharacterized snmRNAs were identified in *Escherichia coli* intergenic regions through combined computational and phylogenetic approaches (6–8). Searches limited to intergenic regions may be problematic, because a number of hypothetical ORFs might not actually encode a protein but an snmRNA. Moreover, snmRNAs encoded by antisense transcripts located within known protein genes or snmRNAs overlapping an ORF such as detected in Archaea (9) also would be missed in these approaches. We therefore followed our unbiased experimental approach by generating cDNA libraries encoding snmRNA candidates from different model organisms, for which we have coined the termed "experimental RNomics" (10, 11).

The goal of this study was to identify snmRNAs in the hyperthermophilic sulfur-metabolizing organism, *Archaeoglobus fulgidus*, which belongs to the lineage of Euryarchaeota, one of the two major archaeal kingdoms (12). Its genome of 2.17 megabases contains a total of 2,436 predicted ORFs and, thus far identified, 51 non-mRNAs (13), i.e., 16S, 23S, and 5S rRNAs, a 7S RNA, RNase P RNA, and 46 tRNAs. In addition, four members of the burgeoning class of box C/D small nucleolar RNAs (snoRNAs), the guides for RNA 2′-*O*-methylation, have been predicted recently with a biocomputational approach (14). Here we report identification of 86 candidates for previously uncharacterized snmRNAs in the archaeon *A. fulgidus*.

Materials and Methods

Generation of a cDNA Library Encoding snmRNAs from *A. fulgidus*. We prepared total RNA from *A. fulgidus* cells by using the TRIzol method (GIBCO/BRL) and fractioned 200 μg of total RNA on a denaturing 8% polyacrylamide gel [7 M urea, 1× TBE buffer (90 mM Tris/64.6 mM boric acid/2.5 mM EDTA, pH 8.3)]. RNAs in the size range of ≈50–500 nt were excised from the gel, passively eluted, and ethanol-precipitated. Five micrograms of this fraction were tailed with CTP by using poly(A) polymerase (15). RNA was reverse-transcribed into cDNA by using primer GIBCO1 and cloned into pSPORT 1 vector employing the GIBCO Superscript system (GIBCO/BRL). cDNAs were amplified by PCR with primers FSP and RSP. PCR products were spotted by robots in high-density arrays on filters (16), performed at the Resource Center of the German Human Genome Project (Berlin).

For additional methods see *Supporting Materials and Methods*, which is published as supporting information on the PNAS web site, www.pnas.org.

Results and Discussion

Construction and Analysis of a cDNA Library Encoding snmRNA Candidates from *A. fulgidus*. To identify candidates for snmRNAs in *A. fulgidus*, we constructed a cDNA library based on small RNAs (sRNAs) (from ≈50 to ≈500 nt). Many cDNA sequences could be assigned to genes encoding known large or small non-mRNAs, as reported previously for a cDNA library containing snmRNAs from mouse (10). In addition to 5S rRNA, RNase P and degradation

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: snmRNA, small non-mRNA; snoRNAs, small nucleolar RNAs; sRNA, small RNA.

Data deposition: The sequences reported in this paper have been deposited in the DDBJ/EMBL/GenBank databases (accession nos. AJ430234–AJ430319).

‡To whom reprint requests may be addressed. E-mail: huttenh@uni-muenster.de or bachel@ibcg.biotoul.fr.

www.pnas.org/cgi/doi/10.1073/pnas.112047299

Table 1. Compilation of expressed RNA sequences from class I from an *A. fulgidus* cDNA library derived from RNAs sized 50–500 nt

ERNS	cDNA clones	cDNA, nt	Northern blot, nt	Abund.	Adjacent genes	Strand	Comments		Accession no.
Group 1: C/D box snoRNAs									
Target: rRNAs							Modification	Guide box	
Afu-122	10	46	50	+++	AF2365/AF2366	><>	Um15 in 16S Um849 in 16S	D′ (11 nt) D (11 nt)	AJ430234
Afu-14	15	51	50	+++	AF0798/AF0799	><<	Cm2604 in 23S Um140 in 23S Cm2559 in 23S	D′ (9 nt) D (8 nt) D (8 nt)	AJ430235
Afu-113	1	84	270	++	AF0272 (>)	>	Um2129 in 23S	D′ (9 nt)	AJ430236
Target: tRNAs									
Afu-64	4	56	54	+++	AF0207/AF0208	><<	Cm34 in tRNA Leu (CAA)	D (10 nt)	AJ430237
Target: unknown									
Afu-67	7	161	145, 165	+	AF0596/AF0567	>>>	—	—	AJ430238
Afu-180	2	75	95	++	AF2289/AF2290	><>	—	—	AJ430239
Group 2: H/ACA box snoRNAs									
Afu-4	77	204	230	+++	AF2373/AF2374	>>	Ψ1167 in 16S, Ψ2601 in 23S, Ψ1364 in 23S		AJ430240
Afu-46	1	55	60	++	AF0456/AF0457	>>>	Ψ2639 in 23S		AJ430241
Afu-52	1	38	90	+++	AF2203/AF2204	>>>	Ψ2878 in 23S		AJ430242
Afu-190	2	34	76	+++	AF0244/AF0245	>><	Ψ1004 in 16S		AJ430243
Group 3: RNAs derived from repeats									
Locus 1									
Afu-31	1	41	68, 136, 204, 272, 340, 408	++	AF0443/AF0444	><>			AJ430244
Afu-44	1	38	dto.	++	AF0443/AF0444	><>			AJ430245
Afu-79	3	106	dto.	++	AF0443/AF0444	><			AJ430246
Afu-114	1	47	dto.	++	AF0443 (>)	<			AJ430247
Afu-170	1	40	dto.	++	tRNA/AF0443	><>			AJ430248
Afu-192	1	39	dto.	++	tRNA/AF0443	><>			AJ403249
Afu-235	1	31	dto.	++	tRNA/AF0443	><>			AJ430250
Locus 2									
Afu-98	2	95	68, 136, 204, 272, 340, 408	++	AF0001/AF0002	<><			AJ430251
Afu-124	1	38	dto.	++	AF0001/AF0002	<><			AJ430252
Afu-269	1	33	dto.	++	AF0001/AF0002	<><			AJ430253
Afu-356	1	47	dto.	++	AF0001/AF0002	<><			AJ430254
Locus 3									
Afu-17	6	65	75, 150, 225, 300, 375, 450	++	AF1880 (>)	<			AJ430255
Afu-74	1	35	dto.	++	AF1882 (>)	<			AJ430256
Afu-94	2	30	dto.	++	AF1881 (>)	<			AJ403257
Afu-148	2	37	dto.	++	AF1881/AF1882	<><			AJ430258
Afu-156	1	50	dto.	++	AF1881/AF1882	<><			AJ430259
Afu-185	2	46	dto.	++	AF1880/AF1881	><<			AJ430260
Afu-257	1	23	dto.	++	AF1880/AF1881	><<			AJ430261
Afu-260	1	39	dto.	++	AF0503 (>)	<			AJ430262
Afu-267	1	51	dto.	++	AF1882 (>)	<			AJ430263
Afu-280	1	41	dto.	++	AF1879/AF1880	<><			AJ430264
Afu-291	1	28	dto.	++	AF1879/AF1880	<><			AJ430265
Group 4: Spliced RNAs from the rRNA operon									
Afu-7	150	150	155	+++	AF1988/tRNA	<<<	16S-D RNA, similar to C/D sRNA		AJ430266
Afu-54	7	87	–	–	tRNA/AF1988	<><	16S-U RNA		AJ430267
Afu-73	13	141	–	–	tRNA/AF1987	<<<	23S-D RNA		AJ430268
Afu-357	7	229	–	–	AF1987/tRNA	<><	23S-U RNA		AJ430269

ERNS, expressed RNA sequences (Afu-); cDNA clones, number of independent cDNA clones identified from each RNA species; cDNA, nt, length of cDNA encoding an snmRNA as assessed by sequencing; Northern blot, nt, length of RNA as assessed by Northern blot analysis; dto., same as above; Abund., relative abundance of snmRNAs as estimated by Northern blot analysis; Adjacent genes, accession numbers of genes flanking the *snmRNA* gene; Strand, transcription from the + or − strand. Thereby, the middle arrow represents the *snmRNA* gene, whereas flanking arrows indicate orientation of the adjacent genes; >, the plus strand of the genome annotation; <, the minus strand of the genome annotation; Function/Comments, comments to proposed function of snmRNA. In the case of class I/groups 1 and 2 snmRNAs, modification refers to predicted modified nucleotides within rRNAs or tRNAs. For C/D-box RNAs, the guide box refers to the length of the antisense element (indicated in nt), preceded by its location in the 5′ domain (D′) or 3′ domain (D) of the snoRNA; Accession no., accession number of snmRNA sequence in the DDBJ/EMBL/GenBank databases.

fragments from 16S and 23S rRNAs, we also detected a few tRNAs (see *Supporting Results and Discussion* and Fig. 3, which are published as supporting information on the PNAS web site, www.pnas.org.) However, this abundant class of snmRNAs was strongly underrepresented, amounting to only 1% of RNAs identified, probably because of their stable structure and presence of nucleotide modifications that may interfere with reverse transcription during library construction. In addition, a considerable portion of cDNA clones (20%), which contained inserts below 14 nt in length, i.e., too short to be analyzed by a BLASTN database search, were not considered further in this study.

Expression Analysis of Candidates for snmRNAs by Northern Blotting. After comparative and genomic analysis of cDNA clones, we performed expression analysis of snmRNA candidates by Northern blotting (Tables 1 and 2; for examples from selected snmRNA candidates see *Supporting Results and Discussion* and Fig. 4, which are published as supporting information on the PNAS web site). In general, the sizes of snmRNAs were larger than the corresponding cDNAs. Our library construction leads to underrepresentation of the very 5′ ends of snmRNAs. Therefore, our sequences resemble mRNA-derived expressed sequence tags, and we designated them ERNS (expressed RNA sequences). We divided the snmRNA candidate species into two different classes based on the presence (class I) or absence (class II) of known structural motifs (Tables 1 and 2). Within the two classes, candidates were assigned to different groups based on further common features of their RNA structure and/or genomic location. In most cases, we failed to detect potential homologs of these snmRNA candidates in other sequenced archaeal genomes by BLASTN database searches, which may be because of their absence or the low conservation of primary

Table 2. Compilation of expressed RNA sequences from class II from an *A. fulgidus* cDNA library derived from RNAs sized 50–500 nt

ERNS	cDNA clones	cDNA, nt	Northern Blot, nt	Abund.	Adjacent genes	Strand	Comments		Accession no.
Group 1: snmRNAs in intergenic regions									
Afu-29	1	42	–	−	AF0808/AF0809	>>>			AJ430270
Afu-57	1	31	–	−	AF2312/AF2313	>>>			AJ430271
Afu-130	1	66	–	−	AF1505/AF1506	<<<			AJ430272
Afu-135	1	57	380	+	AF2370/AF2371	<>>			AJ430273
Afu-136	1	36	80	+	AF2220/AF2221	><<			AJ430274
Afu-242	1	44	340	+++	AF0230/AF0231	<<>			AJ430275
Afu-226	1	47	–	−	AF1882/tRNA	<><	similar to sRNA H/ACA		AJ430276
Afu-274	1	67	120	+	AF1052/AF1053	><<	similar to sRNA H/ACA		AJ430277
Afu-277	1	47	80	+	AF1318/AF1319	<<>	similar to sRNA H/ACA		AJ430278
Group 2: snmRNAs complementary to ORFs									
					Gene				
Afu-32	3	135	95	+++	AF2090-N (>)	<	compl. to 5′ end of ORF		AJ430279
Afu-97	1	177	100, 135	+	AF2247 (>)	<	dto.		AJ430280
Afu-142	1	71	–	−	AF1049 (<)	>	dto.		AJ430281
Afu-239	1	132	150	+	AF1987 (<)	>	similar to sRNA H/ACA		AJ430282
Afu-62	1	105	180, 210	++	AF1017 (<)	>	compl. to 3′ end of ORF		AJ430283
Afu-76	2	121	55, 355	+++	AF0790 (<)	>	dto.		AJ430284
Afu-115	1	43	–	−	AF2390 (<)	>	dto.		AJ430285
Afu-219	1	56	–	−	AF0227 (<)	>	dto.		AJ430286
Afu-304	1	70	48	+	AF0896 (>)	<	dto.		AJ430287
Afu-43	2	104	50	+++	AF0701 (>)	<	compl. to middle of ORF		AJ430288
Afu-47	1	73	–	−	AF1489 (<)	>	dto.		AJ430289
Afu-56	1	86	300	+	AF1277 (>)	<	dto.		AJ430290
Afu-86	1	45	–	−	AF1444 (<)	>	dto.		AJ430291
Afu-99	1	52	330	+++	AF0208 (<)	>	dto.		AJ430292
Afu-123	1	54	50	+++	AF0595 (>)	<	dto.		AJ430293
Afu-125	1	40	–	−	AF0592 (<)	>	dto.		AJ430294
Afu-139	1	148	–	−	AF0597 (<)	>	dto.		AJ430295
Afu-197	2	64	115	+	AF2236 (<)	>	dto.		AJ430296
Afu-202	1	62	180	+	AF0043 (>)	<	dto.		AJ430297
Afu-216	1	68	–	−	AF2197 (<)	>	dto.		AJ430298
Afu-220	1	67	–	−	AF1384 (>)	<	dto.		AJ430299
Afu-232	1	36	300	++	AF0072 (<)	>	dto.		AJ430300
Afu-264	2	128	–	−	AF1503 (<)	>	dto.		AJ430301
Afu-273	1	26	350	+	AF1820 (>)	<	dto.		AJ430302
Afu-275	1	74	200	+++	AF0037 (>)	<	dto.		AJ430303
Afu-278	1	38	300, 450	+	AF0335 (>)	<	dto.		AJ430304
Afu-288	1	34	380	+	AF0484 (>)	<	dto.		AJ430305
Afu-289	1	25	320	+++	AF0783 (>)	<	dto.		AJ430306
Afu-306	1	101	–	−	AF0553 (>)	<	dto.		AJ430307
Afu-309	1	39	–	−	AF0916 (>)	<	dto.		AJ430308
Afu-313	1	76	–	−	AF2236 (<)	>	dto.		AJ430309
Afu-339	1	45	400, 500, 600	++	AF0273 (>)	<	dto.		AJ430310
Afu-343	1	97	–	−	AF1392 (>)	<	dto.		AJ430311
Group 3: snmRNAs overlapping ORFs									
					Gene		overlap nt	ORF (nt)	
Afu-117	1	111	75	++	AF2404	>>	overlap 97 nt	290	AJ430312
Afu-160	1	105	150, 350	+	AF1849	>>	overlap 78 nt	1,869	AJ430313
Afu-191	1	105	230	+++	AF1544/AF1545	>>	overlap 20 nt	396	AJ430314
Afu-297	2	99	45, 310	++	tRNA	<<	overlap 3 nt	72	AJ430315
Afu-340	1	186	600	+	AF0937	>>	overlap 167 nt	1,272	AJ430316
Group 4: snmRNA transcribed from the sense orientation of ORFs									
					Gene			ORF (nt)	
Afu-158	2	89	145, 160	+	AF2230			267	AJ430317
Afu-225	1	49	600	+	AF0890			1,194	AJ430318
Afu-301	1	54	220, 250	+	AF1154			588	AJ430319

ERNS, expressed RNA sequences (Afu-); cDNA clones, number of independent cDNA clones identified from each RNA species; cDNA, nt, length of cDNA encoding an snmRNA as assessed by sequencing; Northern blot, nt, length of RNA as assessed by Northern blot analysis; dto., same as above; Abund., relative abundance of snmRNA as estimated by Northern blot analysis; Adjacent genes, accession numbers of genes flanking the *snmRNA* gene; Strand, transcription from the + or − strand. Thereby, the middle arrow represents the *snmRNA* gene, whereas flanking arrows indicate orientation of the adjacent genes; >, the plus strand of the genome annotation; <, the minus strand of the genome annotation; ORF, length of the open reading frame (in nt) that overlaps with or contains the predicted snmRNA. Accession no., accession number of snmRNA sequence in the DDBJ/EMBL/GenBank databases.

structure, once more emphasizing the need of experimental RNomics in addition to biomathematical approaches.

Class I: Candidates for snmRNAs Exhibiting Known Sequence or Structure Motifs. Class I contains 36 candidates for snmRNAs assigned to four different groups (Table 1). Groups 1 and 2 correspond to candidates for archaeal members of C/D- or H/ACA-box snoRNAs, respectively. Group 3 shows snmRNAs transcribed from short, tandemly repeated sequences distributed in three separate clusters in the *A. fulgidus* genome. Finally, group 4 contains snmRNAs generated by a recently detected splicing event occurring at the bulge–helix–bulge motif of pre-rRNA processing stems.

Group 1: snmRNAs exhibiting C/D-box snoRNA motifs. Afu-14, Afu-64, Afu-113, and Afu-122 display all canonical features of archaeal methylation guide sRNAs (9, 14, and 17). Afu-14, Afu-64, and Afu-122 had been predicted in the *A. fulgidus* complete genome, termed sR1, sR4, and SR2, respectively (14). Their existence had not been tested experimentally. The fourth previously predicted *A. fulgidus* C/D sRNA, sR3, had been characterized experimentally as a methylation guide (18). It was not detected by us, probably because it corresponds to a pre-tRNA intron, presumably suggested to be active in cis before tRNA splicing and because tRNAs (or pre-tRNAs) are underrepresented in the cDNA library (see above). However, we identified a fourth, previously unpredicted C/D sRNA candidate, Afu-113: it harbors a 9-nt-long 5′ antisense element to 23S rRNA predicted to direct 2′-*O*-methylation of U2129.

Two additional members of this group, Afu-67 and Afu-180, significantly differ from previously reported archaeal sRNAs by the presence of nucleotide deviations from the box C and C′ consensus and by unusually long D′–C′ and C′–D intervals. Both RNAs lack an antisense element of at least 8 nt upstream of box D or box D′. Thus, a potential rRNA or tRNA target site is absent (the statistical

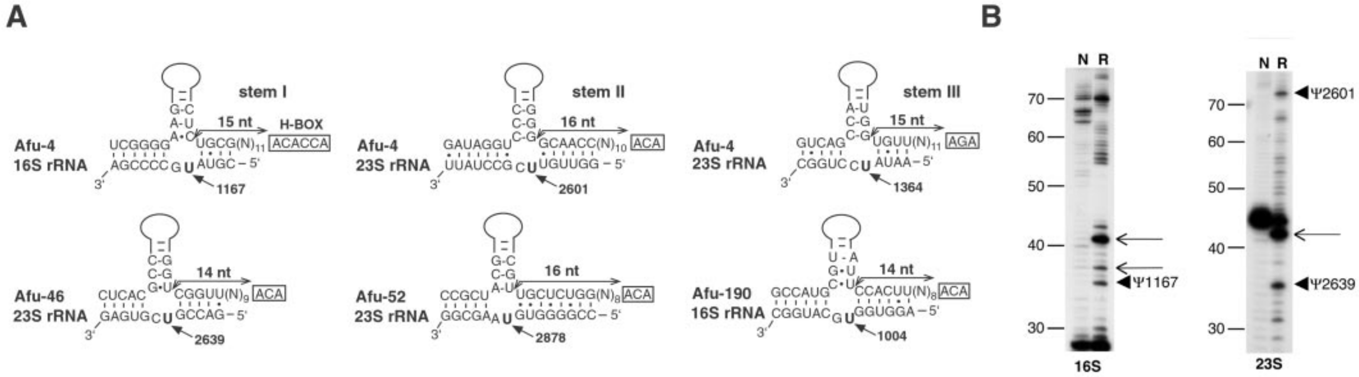


Fig. 1. The four pseudouridylation guide RNAs: Afu-4, Afu-46, Afu-52, and Afu-190 from *A. fulgidus*. (*A*) Potential base-pairing interaction with 16S or 23S rRNA involving each pseudouridylation pocket. The predicted site of pseudouridylation is denoted by an arrow, and its location within the cognate rRNA is indicated by numbering. Its distance from the ACA/AGA- (or H-) box (usually between 14 and 16 nt) is indicated also. The snmRNA sequence in a 5′-to-3′ orientation is shown in the upper strand, with the apical part of the long hairpin domain schematized by a solid line. (*B*) Verification of predicted pseudouridylation sites in *A. fulgidus* rRNA by primer extension. Total cellular RNA samples submitted or not to CMC modification (lanes R and N, respectively) were analyzed by primer extension using ^{32}P-labeled primers complementary to an appropriate segment of 16S or 23S rRNA. Predicted sites are denoted by arrowheads, additional pseudouridines thus far without a known cognate H/ACA guide are denoted by arrows, and cDNA sizes (in nt) are indicated in the margin.

significance of shorter, unperfect matches with rRNAs seems marginal). The function of these two sRNAs remains elusive.

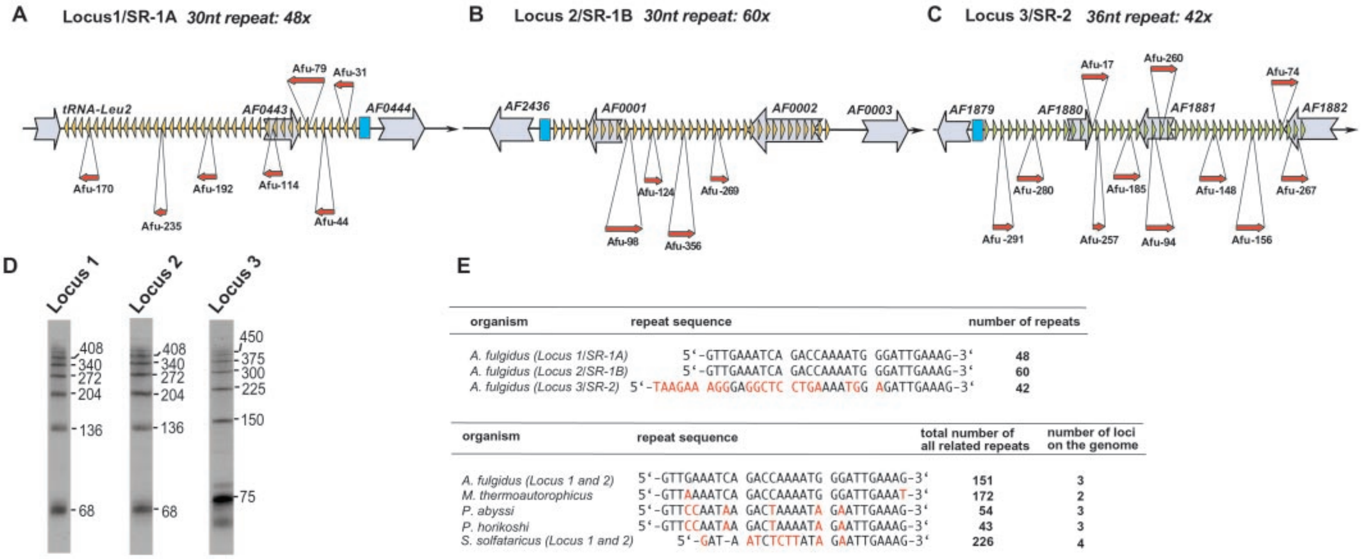
Group 2: snmRNAs exhibiting H/ACA-box motifs. Based on hallmark sequence and structural features, Afu-4, Afu-46, Afu-52, and Afu-190 are likely to represent archaeal counterparts for the abundant class of H/ACA-box snoRNAs, which guide pseudouridylation of ribosomal or spliceosomal RNAs in eukaryotic cells (19–22). The few archaeal rRNAs analyzed thus far contain only a very small number of pseudouridines, similar to bacteria but in marked contrast to Eukarya (23, 24). However, integral protein components of box H/ACA small nucleolar ribonucleoprotein particles, Gar1p, Nhp2p, and Cbf5p, the common pseudouridine synthase, are present in archaeal genomes, raising the possibility that pseudouridine formation in archaeal rRNAs involves counterparts of eukaryotic box H/ACA small nucleolar ribonucleoprotein particles (25).

Afu-4 features a highly stable secondary structure organized in three long stems, each containing a typical pseudouridylation pocket immediately followed by a downstream, single-stranded H- or ACA-box motif (for structure see Fig. 5, which is published as supporting information on the PNAS web site). Although the AGA-box downstream from Afu-4 stem III deviates from the canonical motif, the presence of AGA has been observed in yeast and *Trypanosoma* H/ACA snoRNAs (20, 26). Sequences in the large internal loop of each stem can form canonical bipartite guide duplexes of 9–13 bp around a target uridine in 16S rRNA or 23S rRNA (Fig. 1*A*). Remarkably, the target uridine is always separated from the box motif by the typical distance, 15–16 nt (20). Based on such indirect but statistically highly significant evidence, Afu-4 is predicted to guide pseudouridylation of three uridines, U1167 in 16S and U2601 and U1364 in 23S rRNA (Fig. 1*A*). One of these uridines, U2601, is universally pseudouridylated in Eukarya. Interestingly, this latter nucleotide is not pseudouridinylated in the crenarchaeote *Sulfolobus acidocaldarius* (23, 24). A homolog of Afu-4 was discovered computationally in the archaeon *Pyrococcus furiosus* in the lab of Sean Eddy (R. J. Kline, Z. Misulovin, and S. Eddy, personal communication).

Afu-46, Afu-52, and Afu-190 also represent likely rRNA pseudouridylation guides. They feature single, long, stable stems positioned immediately upstream from an ACA motif (for structures see *Supporting Results and Discussion* and Fig. 3). They resemble *Trypanosoma* H/ACA-box snoRNAs, which also consist of a single hairpin (26), and depart from the two-domain structures widespread in yeast, animal, and plant snoRNAs (ref. 20; see ref. 27 for review). In each case, the long stem contains an internal loop with all the features of a pseudouridylation pocket. The respective internal loops are able to form a 10–13-bp bipartite RNA duplex around a target uridine in 16S or 23S rRNA (Fig. 1*A*). Also, the targeted uridine is separated from the ACA- (or H-) box by 14–16 nt. Afu-46, Afu-52, and Afu-190 therefore are predicted to guide pseudouridylation of U2639 and U2878 in 23S rRNA and U1004 in 16S rRNA, respectively (Fig. 1*A*). Presence of pseudouridines at the expected rRNA sites was confirmed by primer extension performed on *N*-cyclohexyl-*N*′-D-(4-methyl-morpholinium)ethylcarbdiimide *p*-tosylate-treated total cellular RNA. As predicted, the six rRNA pseudouridines were detected unambiguously (Fig. 1*B* and data not shown). All four snmRNAs from this group are relatively rich in GC as compared with the *A. fulgidus* genome (48.6%), especially Afu-4 (60%).

Group 3: snmRNAs derived from short tandem repeats. A total of 22 cDNA clones map to three different *A. fulgidus* genomic loci within clusters of a peculiar class of short tandem repeats, termed SRSRs for short regularly spaced repeats (Fig. 2 *A*–*C*; for Northern blot see Fig. 2*D*). A similar family of repeats, apparently devoid of any protein-coding potential, is present in most completely sequenced archaeal genomes (28). The tandemly repeated short sequence elements (24–37 nt depending on the organism) are always organized in a few large clusters in which 10–100 identical copies are interspersed by divergent spacers of constant size (31–51 nt depending on the organism). Sequences of the repeat units but not the spacers are conserved among Archaea including the crenarchaeon *Sulfolobus solfataricus* (Fig. 2*E*). In *A. fulgidus*, this repeat family is distributed among three genomic loci. Locus 1 and locus 2, designated as SR-1A and SR-1B, respectively (13), contain 48 and 60 copies, respectively, of a 30-nt repeat (Fig. 2 *A*, *B*, and *E*) interspersed with 38-nt-long unique spacers. The third locus, SR-2 (Fig. 2*C*), contains 42 copies of a 36-nt repeat (related in sequence to the 30-nt SR-1 repeat, Fig. 2*E*) interspersed with 41-nt-long unique spacers. The detection of snmRNA candidates mapping to all three loci (Table 1) provides evidence that short, regularly spaced repeats are transcribed in all cases from the same DNA strand. Transcription of tandemly repeated genetic elements in prokaryotes has not been reported before (for a review see ref. 28).

Our findings are corroborated by Northern analysis with probes against the respective repeat motifs. We observed ladder-like band patterns (Fig. 2*D*), with multiples of ≈68 nt (locus 1 and 2 repeats) or 75 nt (locus 3 repeat). These increments almost precisely correspond to the interval between successive repeats in each



organism	repeat sequence	number of repeats
A. fulgidus (Locus 1/SR-1A)	5‘-GTTGAAATCA GACCAAAATG GGATTGAAAG-3‘	48
A. fulgidus (Locus 2/SR-1B)	5‘-GTTGAAATCA GACCAAAATG GGATTGAAAG-3‘	60
A. fulgidus (Locus 3/SR-2)	5‘-TAAGAA AGGGAGGCTC CTGAAAATGG AGATTGAAAG-3‘	42

organism	repeat sequence	total number of all related repeats	number of loci on the genome
A. fulgidus (Locus 1 and 2)	5‘-GTTGAAATCA GACCAAAATG GGATTGAAAG-3‘	151	3
M. thermoautorophicus	5‘-GTTAAAATCA GACCAAAATG GGATTGAAAT-3‘	172	2
P. abyssi	5‘-GTTCCAATAA GACTAAAATA GAATTGAAAG-3‘	54	3
P. horikoshi	5‘-GTTCCAATAA GACTAAAATA GAATTGAAAG-3‘	43	3
S. solfataricus (Locus 1 and 2)	5‘-GAT-A ATCTCTTATA GAATTGAAAG-3‘	226	4

Fig. 2. Location (*A–C*), Northern blot analysis (*D*), and sequence (*E*) of repeats from *A. fulgidus* and other archaeal species. (*A–C*) Location of repeats on locus 1, 2, and 3 of the *A. fulgidus* genome (not drawn to scale) in respect to annotated ORFs or the tRNA-Leu-2 gene. Repeats are shown by yellow or green triangles, which indicate the direction of transcription. The relative position of cDNA clones representing snmRNAs is indicated by red arrows, protein or tRNA genes are shown by light gray arrows, and putative promotor sequences for the three repeat loci are indicated by blue boxes. (*D*) Northern blot analysis of repeated sequences. For each locus, an oligonucleotide derived from the respective repeat motif was used as a probe. (*E*) Sequence alignment of repeats from three loci in *A. fulgidus* (*Upper*) compared with short regularly spaced repeats of four different archaeal species (*Lower*). Deviations from the *A. fulgidus* repeat sequence are indicated in red. The total number of all related repeats as well as the number of loci to which repeats are mapped are indicated on the right.

cluster (68 nt for locus 1 and 2 and 77 nt for locus 3). We observed a similar ladder-like pattern in a Northern blot from *S. solfataricus* by using a corresponding probe (data not shown). These patterns suggest that in all Archaea examined the clustered repeats are transcribed in the form of long precursor(s), subsequently processed into monomers or multimers of the repeat motif. Accordingly, clone Afu-79 spans two consecutive repeat units. Moreover, except for one clone, Afu-98, the 3′ end of the 22 cDNA sequences derived from the three *A. fulgidus* repeat clusters locates in the middle of the repeat motif.

A. fulgidus repeats in loci 1 and 2 are preceded by a highly similar region (285 nt, 90% identity), which also is related to the sequence upstream from locus 3. This finding could reflect the importance of these regions for promoting cluster transcription (Fig. 2 *A–C*, and data not shown). Each of the three *A. fulgidus* repeat clusters spans one or two hypothetical ORFs (Fig. 2*A–C*), However, none of these ORFs exhibit sequence similarity to ORFs in any archaeal genome. Furthermore, with the exception of a hypothetical protein assigned as AF1880 in locus 3 (Fig. 2*C*), these hypothetical mRNAs would be transcribed from the opposite DNA strand.

Clusters of this family of repeats are present also in the genomes of the halophile *Haloferax volcanii* and its megaplasmid (29). Remarkably, the introduction of extra copies of the repeats within a plasmid alters its distribution among the daughter cells, suggesting that the tandem repeats are involved in replicon partitioning (29). Our results raise the possibility that their role in this control might be mediated by transcription of the repeat clusters.

Group 4: snmRNAs derived from ribosomal pre-rRNA sequences. Four different snmRNA species, designated 16S-D, 16S-U, 23S-D, and 23S-U RNA, representing processing products of pre-16S or pre-23S rRNAs, can be assigned to this group (Table 1). All four RNAs are generated by cleavage and subsequent ligation of pre-rRNA spacers at the BHB motifs flanking pre-16S and pre-23S processing intermediates in the rRNA primary transcript, thus revealing an additional link between rRNA processing and RNA splicing in Archaea (30).

Interestingly, 16S-D RNA, by far the most abundant RNA present in our cDNA library, contains structural motifs typical of methylation guide sRNAs. The 16S-D RNA binds to the L7Ae protein, a core component of archaeal C/D-box ribonucleoprotein particles, supporting the notion that it might have an important, but still-unknown role in pre-rRNA biogenesis or even might target RNA molecules other than rRNA (30).

Class II: Candidates for snmRNAs with Unknown Sequence or Structure Motifs. Class II contains the remaining 50 candidates for snmRNAs not assigned to class I, i.e., devoid of known sequence or structure motifs (Table 2). These snmRNAs were grouped according to the location of their coding region on the *A. fulgidus* genome. We distinguished between candidates located in intergenic regions (group 1), transcribed in the antisense orientation to an ORF (group 2), overlapping an ORF (group 3), or encoded within an ORF in the sense orientation (group 4).

Group 1: snmRNAs located in intergenic regions. In general, candidate clones located in intergenic regions should represent the most likely previously uncharacterized functional snmRNAs. Indeed, more than 30 candidates for snmRNAs could be identified in *E. coli* intergenic regions by using different computational and/or experimental approaches (6–8). We have detected nine candidates for snmRNAs located in the intergenic regions of the *A. fulgidus* genome. Three snmRNAs, Afu-226, Afu-274, and Afu-277, are remotely reminiscent of eukaryotic H/ACA snoRNAs because of the presence of sizeable stems and an ACA motif in a single-stranded 3′ tail. Among them, Afu-274 seems to be the best candidate with its two major stems separated by a hinge exhibiting a potential H-box. Although its 3′ stem could well accommodate a pseudouridylation pocket, the best potential bipartite guide sequence (for U1349 in 16S rRNA) in this pocket is only 8 nt long, however, and the internal loop is unusually large for a typical H/ACA snoRNA. Because we could not unambiguously assign them as bona fide H/ACA RNAs, Afu-226, Afu-274, and Afu-277 were annotated as class II/group 1 RNAs.

Group 2: snmRNAs transcribed in the antisense orientation to an ORF. From the group of snmRNAs that are transcribed in the antisense orientation of an ORF encoding a known or hypothetical protein we identified 33 snmRNA candidates (Table 2). In *E. coli*, small antisense RNAs complementary to the translational initiation sites of specific mRNAs have been implicated in the regulation of translation. Such activity may be mediated via masking of the Shine–Dalgarno sequence and the AUG start codon (for reviews see refs. 1 and 31). We therefore assigned the 33 previously uncharacterized antisense RNAs to three different subgroups on the basis of the location of complementarity to the presumed respective mRNA targets. Four snmRNAs (Afu-32, Afu-97, Afu-142, and Afu-239) are complementary to the putative translation initiation region and might be involved in regulation of translation in analogy to *E. coli* snmRNAs. The potential target of Afu-142, gsp-E4, resembles a type II secretion protein, whereas the three other candidate antisense RNAs in this subset correspond to hypothetical proteins. The five snmRNAs complementary to the translation termination site and the remaining snmRNAs complementary to regions located within an ORF still might regulate translation by as-yet-unknown mechanisms.

Group 3: snmRNAs overlapping ORFs. From this group, five independent candidates for snmRNAs could be identified that overlapped ORFs. Expression of all these candidates was confirmed by Northern blot analysis, a prerequisite to be assigned to this group. At the same time, the ORF of the respective overlapping gene was required to be considerably larger than the size of the snmRNA, thereby excluding identification of false positives derived from degradation of mRNAs (Table 2). Coexpression of the potential snmRNA genes and their respective overlapping protein genes would require either processing of a fraction of a presumptive common transcript into the snmRNA or separate transcription of the snmRNA and mRNA from two independent promoters.

Group 4: snmRNAs transcribed from the sense orientation of ORFs. For this last group, it remains to be demonstrated that its RNA members actually represent defined previously uncharacterized snmRNAs rather than degradation products of known mRNAs. Based on the detection of a Northern hybridization signal with a size considerably smaller than that of the predicted ORF, three potential snmRNAs candidates, Afu-158, Afu-225, and Afu-301, of a total of 182 RNA species derived from the coding region of mRNAs can reasonably be assigned to this group. Any of them, however, might still represent a (stable) degradation product of an mRNA. Conversely, it is noteworthy that the three snmRNAs all map within genes for hypothetical proteins, opening the possibility that the proposed ORFs have been annotated wrongly and encode a snmRNA instead.

Expressed Sequence Tags for Hypothetical Proteins from the *A. fulgidus* Genome. In addition to the 86 candidates for snmRNAs described above, we have identified a total of 182 cDNA clones derived from ORFs encoding hypothetical or known proteins (76 and 106 cDNA clones, respectively). These RNAs might merely represent degradation products of mRNAs as mentioned above. As for the 76 RNA species from hypothetical ORFs, however, we cannot rule out the possibility that some of those ORFs were annotated incorrectly and encode a stable snmRNA instead of a protein. In any case, the 182 sequences should be treated as useful information, because they represent expressed sequence tags of potential protein genes transcribed from the genome of *A. fulgidus*. Sequences of these cDNA clones are therefore presented under *Supporting Results and Discussion* and Table 3.

Conclusions

Identification of 86 candidates for previously uncharacterized snmRNAs in the archaeon *A. fulgidus* illustrates the power of our experimental RNomics approach in the search for previously uncharacterized snmRNA species in model organisms. The key finding of this work is the discovery of four H/ACA-type RNAs in Archaea, guiding formation of six pseudouridines in 16S or 23S rRNAs. In addition, we show that nested families of sRNAs are generated from short tandem repeats (termed short regularly spaced repeat elements), which are common to Archaea but previously have not been shown to be transcribed. This study sets the stage for the functional analysis of all snmRNAs identified in our screen. The experimental RNomics approach can and will be applied to different model organisms from bacteria to man and undoubtedly will result in the identification of previously uncharacterized classes of functionally important snmRNAs.

We dedicate this paper to Harry Noller and Carl Woese who kindled and sustained our interest in RNA. This work was supported by the German Human Genome Project through the Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie Grant 01KW9966 (to J.B. and A.H.), an Interdiziplinäres Zentrum für Klinische Forschung grant (Teilprojekt IKF G6, Münster to A.H.), laboratory funds from the Centre National de la Recherche Scientifique and Université Paul Sabatier, Toulouse, and by a grant from the Ministère de l'Education Nationale, de la Recherche et de la Technologie (Program de Recherche Fondamentale en Microbiologie et Maladies Infectieuses et Parasitaires, 2001–2002, to J.P.B.).

1. Wassarman, K. M., Zhang, A. & Storz, G. (1999) *Trends Microbiol.* **7,** 37–45.
2. Mattick, J. S. (2001) *EMBO Rep.* **2,** 986–991.
3. Erdmann, V. A., Barciszewska, M. Z., Hochberg, A., de Groot, N. & Barciszewski, J. (2001) *Cell Mol. Life Sci.* **58,** 960–977.
4. Erdmann, V. A., Barciszewska, M. Z., Szymanski, M., Hochberg, A., de Groot, N. & Barciszewski, J. (2001) *Nucleic Acids Res.* **29,** 189–193.
5. Eddy, S. R. (2001) *Nat. Rev. Genet.* **2,** 919–929.
6. Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E. G., Margalit, H. & Altuvia, S. (2001) *Curr. Biol.* **11,** 941–950.
7. Rivas, E., Klein, R. J., Jones, T. A. & Eddy, S. R. (2001) *Curr. Biol.* **11,** 1369–1373.
8. Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G. & Gottesman, S. (2001) *Genes Dev.* **15,** 1637–1651.
9. Gaspin, C., Cavaille, J., Erauso, G. & Bachellerie, J. P. (2000) *J. Mol. Biol.* **297,** 895–906.
10. Hüttenhofer, A., Kiefmann, M., Meier-Ewert, S., O'Brien, J., Lehrach, H., Bachellerie, J.-P. & Brosius, J. (2001) *EMBO J.* **20,** 2943–2953.
11. Filipowicz, W. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 14035–14037.
12. Woese, C. R., Achenbach, L., Rouviere, P. & Mandelco, L. (1991) *Syst. Appl. Microbiol.* **14,** 364–371.
13. Klenk, H. P., Clayton, R. A., Tomb, J. F., White, O., Nelson, K. E., Ketchum, K. A., Dodson, R. J., Gwinn, M., Hickey, E. K., Peterson, J. D., *et al.* (1997) *Nature* (*London*) **390,** 364–370.
14. Omer, A. D., Lowe, T. M., Russell, A. G., Ebhardt, H., Eddy, S. R. & Dennis, P. P. (2000) *Science* **288,** 517–522.
15. DeChiara, T. M. & Brosius, J. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 2624–2628.
16. Schmitt, A. O., Herwig, R., Meier-Ewert, S. & Lehrach, H. (1999) in *PCR Applications: Protocols for Functional Genomics*, eds. Innis, M. A., Gelfand, D. H. & Sninsky, J. J. (Academic, San Diego), pp. 457–472.
17. Dennis, P. P., Omer, A. & Lowe, T. (2001) *Mol. Microbiol.* **40,** 509–519.
18. Clouet d'Orval, B. C., Bortolin, M. L., Gaspin, C. & Bachellerie, J. P. (2001) *Nucleic Acids Res.* **29,** 4518–4529.
19. Smith, C. M. & Steitz, J. A. (1997) *Cell* **89,** 669–672.
20. Ganot, P., Bortolin, M. L. & Kiss, T. (1997) *Cell* **89,** 799–809.
21. Bachellerie, J. P., Cavaille, J. & Qu, L. H. (2000) in *Ribosome: Structure, Function, Antibiotics, and Cellular Interactions*, eds. Garrett, R., Douthwaite, S., Liljas, A., Matheson, A., Moore, P. B. & Noller, H. (Am. Soc. Microbiol., Washington, DC), pp. 191–203.
22. Jady, B. E. & Kiss, T. (2001) *EMBO J.* **20,** 541–551.
23. Kowalak, J. A., Bruenger, E., Crain, P. F. & McCloskey, J. A. (2000) *J. Biol. Chem.* **275,** 24484–24489.
24. Ofengand, J. & Rudd, K. (2000) in *Ribosome: Structure, Function, Antibiotics, and Cellular Interaction*, eds. Garrett, R., Douthwaite, S., Liljas, A., Matheson, A., Moore, P. B. & Noller, H. (Am. Soc. Microbiol., Washington, DC), pp. 175–190.
25. Watanabe, Y. & Gray, M. W. (2000) *Nucleic Acids Res.* **28,** 2342–2352.
26. Liang, X. H., Liu, L. & Michaeli, S. (2001) *J. Biol. Chem.* **276,** 40313–40318.
27. Kiss, T. (2001) *EMBO J.* **20,** 3617–3622.
28. Mojica, F. J., Diez-Villasenor, C., Soria, E. & Juez, G. (2000) *Mol. Microbiol.* **36,** 244–246.
29. Mojica, F. J., Ferrer, C., Juez, G. & Rodriguez-Valera, F. (1995) *Mol. Microbiol.* **17,** 85–93.
30. Tang, T.-H., Rozhdestvensky, T., Clouet d'Orval, B., Huber, H., Charpentier, B., Branlant, C., Bachellerie, J.-P., Brosius, J. & Hüttenhofer, A. (2002) *Nucleic Acids Res.* **30,** 921–930.
31. Wagner, E. G. H., Altuvia, S. & Romby, P. (2001) in *Homology Effects: Advances in Genetics*, ed. Wu, T. (Academic, San Diego), in press.