RESEARCH PAPER

RNA Biology 10:7, 1211–1220; July 2013; © 2013 Landes Bioscience

RNA at 92 °C

The non-coding transcriptome of the hyperthermophilic archaeon *Pyrococcus abyssi*

Claire Toffano-Nioche,[1] Alban Ott,[1] Estelle Crozat,[1] An N. Nguyen,[1] Matthias Zytnicki,[2] Fabrice Leclerc,[1] Patrick Forterre,[1] Philippe Bouloc,[1] and Daniel Gautheret[1,*]

[1]Univ. Paris-Sud 11; CNRS; UMR8621; Institut de Génétique et Microbiologie; France; [2]URGI; INRA; Versailles; France

Keywords: transcriptome, hyperthermophile, archaea, non-coding RNA

The non-coding transcriptome of the hyperthermophilic archaeon *Pyrococcus abyssi* is investigated using the RNA-seq technology. A dedicated computational pipeline analyzes RNA-seq reads and prior genome annotation to identify small RNAs, untranslated regions of mRNAs, and cis-encoded antisense transcripts. Unlike other archaea, such as Sulfolobus and Halobacteriales, *P. abyssi* produces few leaderless mRNA transcripts. Antisense transcription is widespread (215 transcripts) and targets protein-coding genes that are less conserved than average genes. We identify at least three novel H/ACA-like guide RNAs among the newly characterized non-coding RNAs. Long 5' UTRs in mRNAs of ribosomal proteins and amino-acid biosynthesis genes strongly suggest the presence of cis-regulatory leaders in these mRNAs. We selected a high-interest subset of non-coding RNAs based on their strong promoters, high GC-content, phylogenetic conservation, or abundance. Some of the novel small RNAs and long 5' UTRs display high GC contents, suggesting unknown structural RNA functions. However, we were surprised to observe that most of the high-interest RNAs are AU-rich, which suggests an absence of stable secondary structure in the high-temperature environment of *P. abyssi*. Yet, these transcripts display other hallmarks of functionality, such as high expression or high conservation, which leads us to consider possible RNA functions that do not require extensive secondary structure.

Introduction

Widespread or "pervasive" transcription of genomic regions outside protein-coding genes is now well established in a wide range of eukaryotic species. The case for pervasive transcription in bacterial and archeal genomes is not as clear since these genomes are compact with short intergenic regions that are often part of transcribed operons. Yet, several recent studies used deep sequencing or tiling arrays to evaluate non-coding transcription in a variety of bacterial species and identified large amounts of small RNAs, antisense transcripts and UTR extensions of protein coding genes (reviewed in ref. 1). Although the term "pervasive transcription" is not often associated to non-eukaryotic organisms, it turns out that a large part of non-coding regions in bacteria are covered by regulatory RNAs or UTR extensions, and that many, if not all, bacterial coding genes produce antisense transcripts.

Screens for non-coding RNAs (ncRNAs) in archaea are not as developed as in the other domains of life. Early attempts of archeal Rnomics involved cDNA cloning[2] and computational screens possibly combined to northern blot or PCR validation.[3-6] Classes of RNAs identified by these archeal screens notably differed from bacterial RNAs and were dominated by modification guides H/ACA and C/D-box RNAs. These approaches are progressively superseded by high-throughput sequencing technologies enabling deep sequencing of total or size-selected RNA. In addition to new H/ACA and C/D box RNAs,[7-10] deep sequencing identified a number of new CRISPR (a defense system present in bacteria and archaea) RNA loci[7,8,10] as well as widespread antisense transcription of coding and non-coding loci.[7,8] Such "cis-encoded antisenses" are clearly a significant part of the non-coding transcriptome as they were visible already using low-throughput Rnomics.[11] Other recently identified classes of archeal RNAs include circular RNAs,[9] split RNA genes,[10] and a bacterial-like trans-acting small RNA.[12]

Among archaeal species sampled by Rnomics studies lie a number of hyperthermophiles (growth temperatures higher than 80°C): *Pyrococcus abyssi*, *Nanoarchaeum equitans*, *Sulfolobus solfataricus* (refs. 2–3, 5–6, and 10) and members of the Pyrobaculum genus.[7] Counter-intuitively, organisms living at high temperature do not necessarily have GC-rich genomes.[13-15] Indeed, topological constraints on circular DNA molecules may enforce double strand formation up to over 100°C independently of GC-content.[16] However, structured RNAs, such as tRNAs and rRNAs, tend to have very high GC-content in hyperthermophiles[13] possibly because they do not have such constraints. Computational biologists embraced this discrepancy to develop

*Correspondence to: Daniel Gautheret; Email: daniel.gautheret@u-psud.fr
Submitted: 05/05/13; Revised: 06/25/13; Accepted: 06/27/13
http://dx.doi.org/10.4161/rna.25567

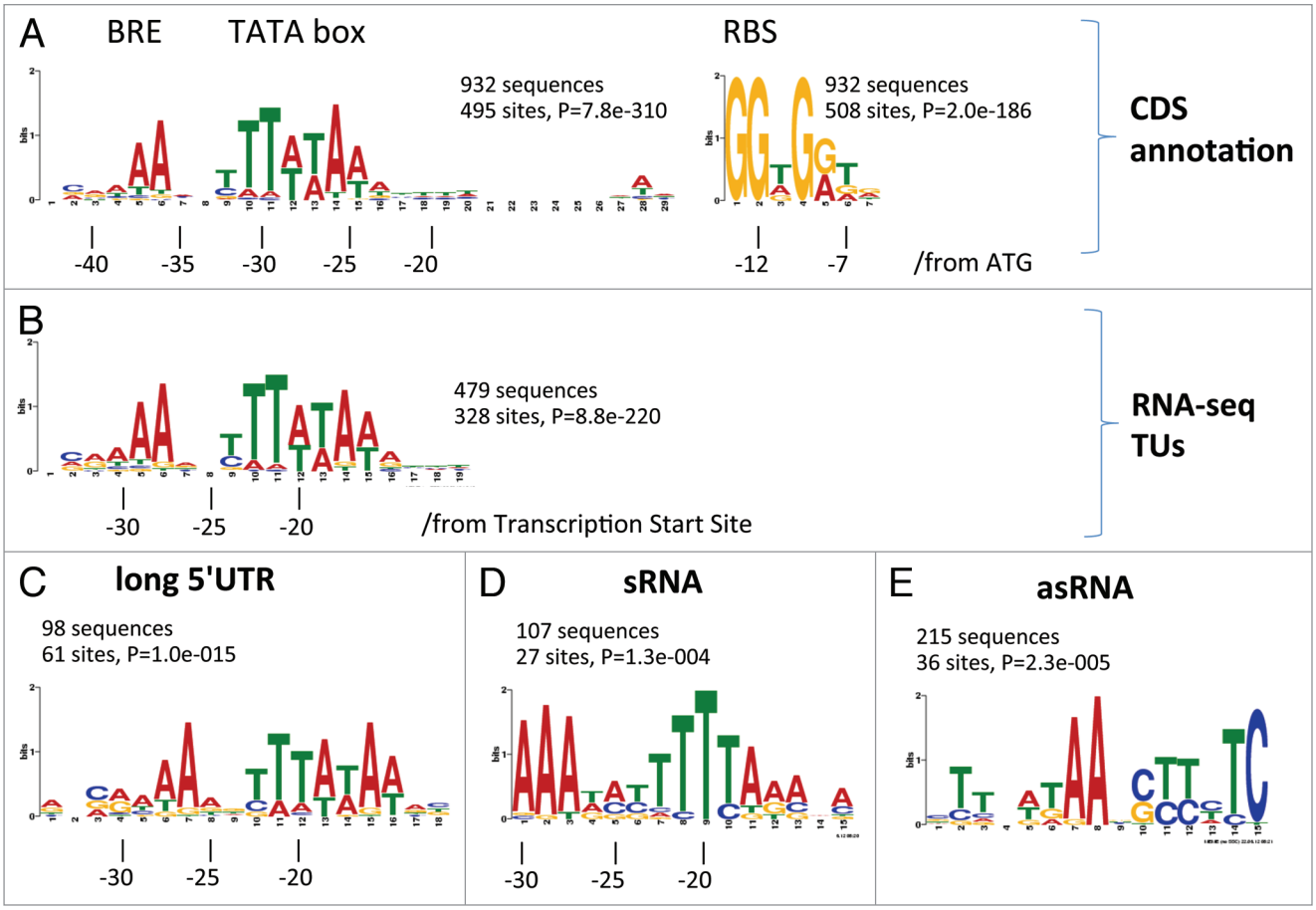

Figure 1. Consensus promoter motifs. Each frame shows the best ranking sequence motif identified by a MEME search[45] performed on the 50 nt upstream region of: (**A**) CDS annotations; (**B**) RNA-seq-based transcription units; (**C**) long 5' UTRs, (**D**) sRNAs, (**E**) asRNAs. Number of occurrences, MEME *P* value and number of sites are given for each motif. Motif coordinates are numbered from the first base of the RNA-seq transcription unit (**B–D**) or from the ATG start codon (**B**) and correspond to the dominant motif location. No dominant location was found for the asRNA motif (**E**).

ncRNA detection programs based on local GC-enrichment in otherwise AT-rich genomes.[2,4] These algorithms were successful in identifying a number of structured RNAs but the question of non-coding RNA genes in hyperthermophiles has been set aside since then. Currently, it is not known whether the multiple non-coding RNAs identified by deep sequencing adopt secondary structures like rRNA and tRNA do.

Here, we sequenced the complete transcriptome of *P. abyssi* grown at 92°C and reconstructed the set of coding and non-coding transcripts. Taking advantage of a higher coverage than in previous *P. abyssi* transcriptome screens, we confirmed a set of short or extended ORFs and estimated the number of novel non-coding elements in each major ncRNA category: independent transcripts, cis-encoded antisenses, and UTR extensions. We analyzed the properties of novel RNA candidates in each category and identified several new functional RNAs of the H/ACA box class, and possibly of the cis-regulatory RNA class. Finally, we analyzed the G:C content and abundance of new ncRNA elements and showed how they differ from ncRNAs identified in previous experimental and computational screens. This analysis revealed that a large fraction of the newly identified ncRNAs do not adopt stable secondary structures although they harbor other hallmarks of functional RNAs such as the presence of strong promoters, high expression levels, or phylogenetic conservation.

Results

RNA deep sequencing and transcript classification. We collected RNA from a *P. abyssi* culture grown at 92 °C sampled at successive growth stages up to stationary phase and submitted RNA to directional RNA-seq library preparation. Illumina sequencing produced 51 million single reads of length 40 nt of which 5.6 million mapped to unique, non-rRNA loci. The *P. abyssi* genome is very dense and comprises a large number of genes that overlap at their UTRs.[17] This context makes it difficult to distinguish new independent transcripts from the 5' and 3' extensions of previously annotated genes. We developed a protocol that clusters RNA-seq reads in a strand-specific fashion and detects overlaps with previous annotations in order to merge clusters into transcription units (TUs). Our basis for previous annotation was the NCBI annotation updated by that of Gao et al.[18] who identified 115 extra ORFs. TUs overlapping previous annotation were split into 5' UTR, 3' UTR, CDS, and operon

spacers. New TUs were analyzed for their coding potential and classified either as novel protein-coding genes (CDS) or independent ncRNAs, here called small RNAs (sRNA). Small RNAs overlapping TUs on the opposite strand were reclassified as cis-encoded antisense RNAs (asRNAs). 5' UTR were further classified into "long 5' UTRs" when longer than 50 nt.

Promoters and UTR regions. Consensus sequence motifs were extracted from the upstream region of annotated CDSs (**Fig. 1A**) and the region upstream of RNA-seq-based TUs (**Fig. 1B**). The Pyrococcus consensus promoter is composed of two boxes often referred to in archaea as BRE element and TATA box.[19] Here, the TATA box sequence is TTT(A/T)(T/A)AA, similar to that of *Methanococcus vannielii* (TTTATAATA),[20] and *Sulfolobus solfataricus* (TTTTAAA).[8] RNA-seq-based transcription start sites (TSS) are located in average 20 nt downstream of the 3' end of the TATA box (**Fig. 1B**). The fraction of transcription units featuring both a TATA and a BRE box (68%) is higher than the fraction of annotated CDS in this case (53%), indicating that our annotation of TUs improves previous CDS annotation.

Few leaderless transcripts are present in *P. abyssi*. Previous observations in *Sulfolobus*[8] and *Halobacteriales*[21] established a majority of leaderless transcripts in these species (> 69% in Sulfolobus, > 90% in halobacteriales), when the methanogen *Methanosarcina mazei* revealed a majority of long 5' UTR (up to 500 nt),[22] unveiling a variety of situations in Archaea. The median size of the *P. abyssi* 5' UTRs is 37 nt (**Fig. S1**). We predict 292 5' UTRs with a size under 20 nt (**Fig. S2A**), but most can be attributed to imperfect RNA-seq coverage rather than to a true lack of leader sequence. Indeed, an analysis of sequences upstream of the predicted TSS reveals an RBS motif in 180 of the 207 UTRs with a size of 6–16 nt, leaving only 27 transcripts that may lack a proper leader (**Fig. S2B**). Independently of the predicted UTR size, the distance between the 3' end of the TATA box and the RBS is constant (**Fig. S2B**) and suggests an actual size of these short 5' UTRs of about 20 nt.

The promoters of long 5' UTRs and independent non-coding RNAs (sRNAs) display consensus motifs similar to that of other promoters (**Fig. 1C and D**). However, the fraction of sRNAs with a major promoter (27/107 = 25%) is significantly lower than for protein coding genes (68%), suggesting a relatively lower accuracy of TSS definition in our sRNA annotation. Antisense RNAs (asRNAs) are not preceded by canonical promoters. Instead, a small proportion (36/215) are flanked by a motif (**Fig. 1E**), which does not correspond to a known regulatory sequence or its reverse complement. The fact the majority of asRNAs are not flanked by any visible promoter sequence suggests a large fraction of these RNAs result either from leaky transcription or processing of longer transcripts.

Our sequencing coverage was sufficient to identify 3' UTRs of size 10 nt or more for 446 genes. Considering all genes with a detectable 3' UTR, the median size of predicted 3' UTRs in Pyrococcus is 37 nt (**Fig. S1**). However, note that single-end RNA-seq protocols like the one we used here do not systematically sequence the 3' side of cDNA library fragments, thus actual 3' UTRs are probably longer.

Table 1. RNA abundance by class

RNA class	Number of elements	Total number of reads	Median RPKM
rRNA	5	32727463	14,818
tRNA	46	33414	264
Long 5' UTR	98	172014	185
sRNA	107	84797	63
Antisense RNA	215	13253	43
CDS	1,893	5261317	164

Overall portrait of non-coding elements. **Table 1** presents the amount of RNA-seq reads associated to each class of ncRNA element. The most populated classes of ncRNAs are asRNAs (215), followed by sRNAs (107) and long 5' UTRs (98). A substantial fraction of the sRNAs and long 5' UTRs (respectively, 37% and 45%) were already identified in previous studies and/or annotated as encoding ncRNA elements in databases. However, the vast majority of asRNAs (97%) were previously unreported (**Fig. 2A**). Non-coding RNAs are generally less conserved than coding genes at the nucleotide sequence level but more than intergenic region. Between 47–57% of the non-coding elements are specific to *P. abyssi*, compared with 28% for coding sequences and 76% for intergenic region (**Fig. 2B**).

We measured RPKM (reads per kb per million) as an approximation of RNA abundance (**Table 1**). Since we used in our library preparation an exonuclease degrading 5'-monophosphorylated RNA species (Materials and Methods), we expect a depletion of rRNAs and tRNAs, as well as of any processed RNAs, including some sRNAs or asRNAs. Keeping this limitation in mind, the most abundant non-coding elements after rRNAs and tRNAs are long 5' UTRs (median: 185 RPKM), followed by sRNAs (63 RPKM) and asRNAs (43 RPKM). The relatively high abundance of long 5' UTRs only reflects that of coding transcripts in general, since the median RPKM in coding regions is 164 RPKM (**Table 1**). Analysis of the new transcription units revealed 26 new or extended protein coding sequences: nine in sRNAs and in 18 in long 5' UTRs (**Table S1**).

The coverage of intergenic regions by RNA-seq reads enabled us to combine a number of annotated genes into multicistronic transcription units (**Fig. S2**). Overall, 1,466 out of 1,944 annotated genes (75%) are part of multi-cistronic transcripts and are thus likely expressed as operons. Conversely, 444 out of 888 extended transcription units involve two or more CDSs or ncRNA genes. The longest polycistronic transcripts (28 CDS and 20 CDS) are ones encoding ribosomal proteins.

Widespread cis-antisense transcription includes known functional RNAs. Although we used a strict definition of antisense, which we require to be fully embedded in the opposite strand of a transcription unit, our asRNA list still includes four C/D box guide RNAs. Indeed, one of the longest and most expressed asRNAs (PabO115, size 277 nt: **Table S1**) matches a C/D box RNA (RFAM family RF01130) and sits right against the center of a five-gene operon (**Fig. S3**). Therefore, in the

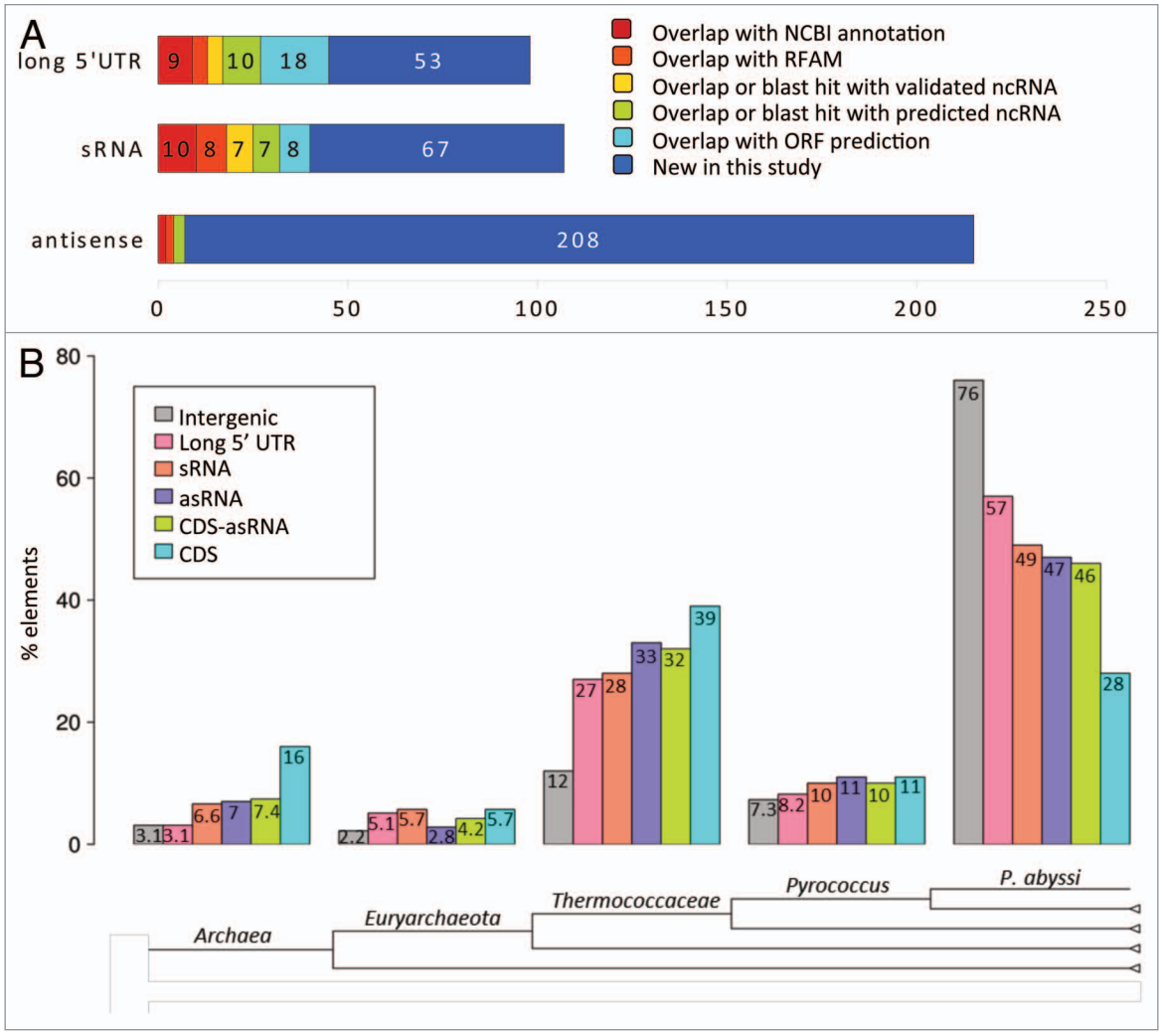

Figure 2. Characteristics of ncRNAs identified by RNA-seq. (**A**) Numbers of known RNAs vs. novel RNAs. Sources for known ncRNAs: NCBI features, RFAM,[23] and studies from Klein et al.[3] and Phok et al.[6] classes are ranked in order of decreasing confidence starting from experimentally validated RNAs (from left to right on the bar plot). When the known ncRNA is from *P. abyssi*, a minimum overlap of 25 nt with the known RNA and the new candidate is required to assign the candidate to a class. When the known ncRNA is from another Archaeal species, a BLASTN sequence conservation with a minimum BLASTN bit score of 42 is required to assign the candidate to a class. When a candidate appears in several classes, it is counted only in the class with highest confidence. (**B**) Conservation of ncRNA classes. At each taxonomic level, the histogram shows the fraction of elements conserved up to, and not deeper than, this taxonomic level (see Materials and Methods). Elements shown include the three ncRNA classes (107 sRNAs, 98 long 5' UTRs, 215 asRNAs), CDS, antisense-associated CDS (CDS-asRNA), and intergenic region. To avoid conservation bias due to size differences, conservation for CDS, CDS-asRNA, and intergenic region were obtained on randomly sampled fragments from CDS, CDS-asRNA, and intergenic regions with the same size distribution as ncRNA, asRNA, and ncRNA, respectively.

compact *P. abyssi* genome, functional trans-acting ncRNAs can be produced from fully antisense transcripts.

Conservation analysis does not support a predominance of trans-acting asRNAs. Indeed, such asRNAs would sustain a dual selection pressure, first due to their trans-acting role and second due to the coding sequence on the opposite strand. However, asRNAs are significantly less conserved than average protein coding genes (see "CDS," **Fig. 2B**), which argues against widespread trans-acting functions. The conservation level of asRNAs is similar to that of other ncRNAs (5' UTRs and sRNAs). This relatively low conservation also applies to other regions of genes harboring asRNA (see CDS-asRNA in **Fig. 2B**). Moreover, genes harboring asRNA are less expressed than average CDS (**Fig. S7**). Altogether, this shows that asRNAs generally occur in a class of genes that are less conserved and expressed than average protein-coding genes.

Overall, the antisense category is that with fewer previous annotation (**Fig. 2A**), which is unsurprising since previous screens mostly relied on sequence conservation or GC-content and systematically excluded coding sequences and their antisense strands. The number of actual functional RNAs among the 215 identified antisense elements remains to be determined. We selected 71 "high interest" candidate functional asRNAs based on their high abundance, large size, strong promoter, or high GC-content (noted PabO100 to PabO170 in **Table S1**). It should be noted that the only four asRNAs with a known function rank

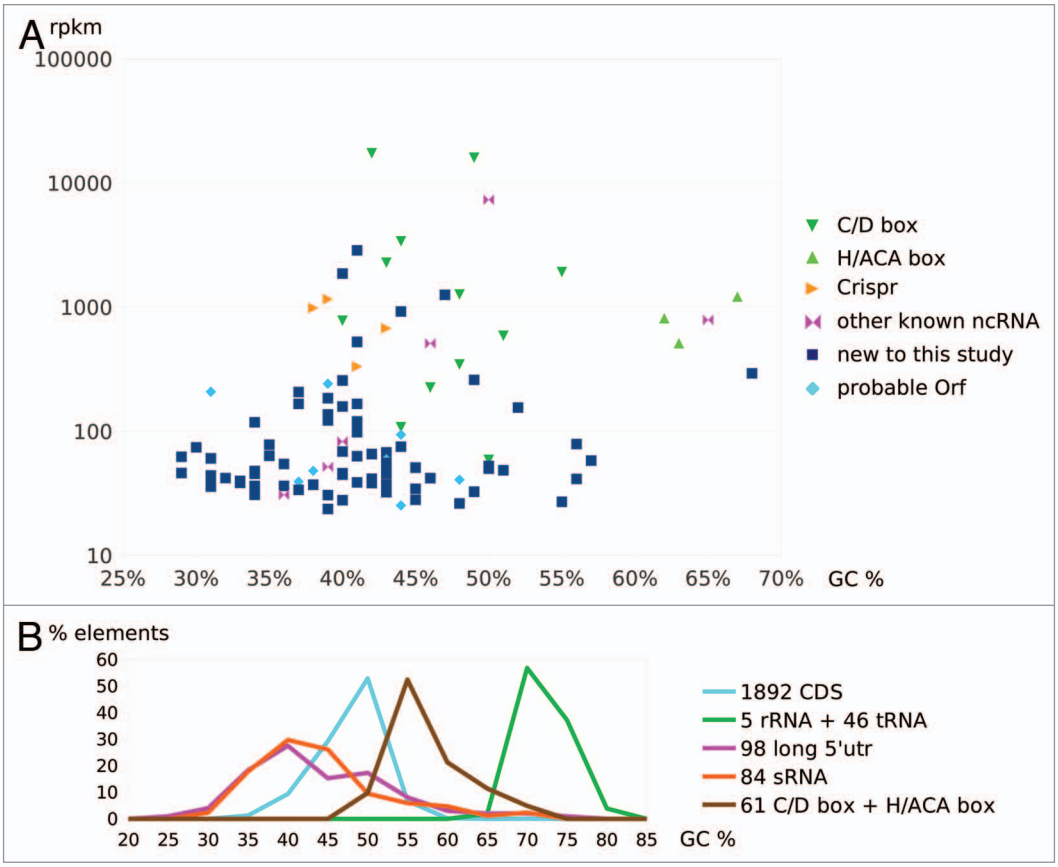

Figure 3. (**A**) Distribution of 107 sRNAs as a function of GC% and RPKM. Unknown RNAs are dark blue, known RNAs are colored as in legend. (**B**) GC-contents of transcript classes. Sequences were extracted following annotations (CDS, tRNA, rRNA, sno-like RNA) or RNA-seq analysis (sRNA and long 5' UTR).

among the top six asRNAs ranked by abundance. This suggests RPKM is a good criterion for selecting putative functional asRNAs.

New sRNA elements are mostly unstructured. Our annotation workflow identified 107 putative sRNAs (**Table S1**). We requalified nine of them as ORFs. Eighteen others were previously annotated in RFAM[23] or in prior publications as CRISPR, C/D box, or H/ACA box RNAs, while 14 were identified in previous screens[3,6] but had no assigned function.

Expectedly, sRNAs with high GC-contents, high conservation, or high expression were more likely to be reported in previous studies. Among 27 sRNAs meeting at least two out of these three criteria, only five are novel to this study (PabO1-5, **Table S1**). Of 67 sRNAs identified here for the first time, 47 can be considered as "high interest" based on any of the criteria: high expression, high conservation, strong promoter, or high-GC (noted PabO01 to PabO47, **Table S1**). Of note, 16 sRNAs have unusually long sizes of over 200 nt, which includes four CRISPR RNAs and nine high interest candidates that were not previously reported. Furthermore, two of the high interest sRNAs (PabO9, PabO10) are arranged in reciprocal antisense orientation (**Fig. S4**), reminiscent of a bacterial toxin/antitoxin system. These mutual antisense sRNAs were not classified as asRNAs by our annotation pipeline as they overlap only partially.

A high GC-content is a hallmark of structured RNA in hyperthermophiles.[13] Prior RNA detection screens based on GC-content mostly identified RNAs in the range 50–75% GC.[3] Here a threshold of 50% GC would retain only 15% of sRNAs. Nine of the novel RNAs are above this threshold and may thus constitute new instances of archaeal structured RNAs. Most of these new high-GC sRNAs are low-expression transcripts, which explains why they were not previously detected. **Figure 3A** shows a plot of abundances vs. GC-contents for all sRNAs, identified by their functional status. This confirms the general tendency for novel sRNAs to harbor low GC% and low abundance. Notable exceptions are PabO5 expressed at 293 RPKM and PabO17 at 155 RPKM. Three novel high-GC RNAs (PabO5, Pab043, PabO2) were computationally predicted by Phok et al.[6] but could not be confirmed experimentally.

If we compare the GC content distribution of new sRNAs with that of other transcript classes, the AU-richness of new sRNAs appears even more strikingly (**Fig. 3B**). Novel sRNAs peak at around 40–45% GC, against 70% for tRNAs and rRNAs, and 55% for C/D and H/ACA RNAs. Even coding regions have a GC content peak (50%) that is significantly higher than that of new sRNAs. This extreme AU-richness is a strong indication that most of the novel sRNAs identified in this study are generally unstructured. Importantly, however, AU-richness does not imply absence of function: for instance, CRISPR RNAs are generally

AU-rich (**Fig. 2A**). A strong indication of functionality in the *P. abyssi* AU-rich sRNAs is the fact that 41% of sRNAs identified by RNA-seq are conserved in thermococcales or deeper in the evolutionary tree (**Fig. 2B**), which is much higher than the 15% GC-rich fraction. Indeed, 16 AU-rich sRNAs (< 45% GC) are deeply conserved and 41 AU-rich sRNAs are considered "high quality" by the above criteria (**Table S1**).

New long 5' UTR elements suggest cis-regulated mRNAs. Our workflow identified 98 5' UTRs of 50 nt or more. Among these, 18 most likely include new ORFs that constitute extensions of previous annotated CDSs (**Table S1**). As for sRNAs, most of the "low hanging fruits" in the long 5' UTR category had been picked up by previous screens: among 26 long 5' UTRs meeting at least two criteria among high GC, high conservation, and high expression, only seven are novel to this study or were only known as computational predictions (PabO60–66). Thirteen of the long 5' UTRs overlap RFAM[23] entries for C/D or H/ACA guide RNAs. Location in 5' UTR is a frequent feature of archaeal guide RNAs, which are known to often overlap coding regions.[24-27] In such cases, it is difficult to infer from the RNA-seq data alone whether the ncRNA is transcribed independently or processed from the mRNA leader. Both situations have been reported.[8]

At the time we submit this manuscript, there is still no experimentally demonstrated riboswitch or other cis-regulatory RNA in archaea. Here, we can however pinpoint several strong candidates: the crcb RNA known as the fluoride riboswitch[28] is similar to one of our highly expressed long 5' UTRs. This ncRNA was predicted (and not confirmed) by the Klein et al.[3] and Phok et al.[6] screens and the presence of the riboswitch inferred in archaea by comparative genomics.[29] Here, we confirm its expression in the form of a 176 nt leader. Another long leader that was validated in previous screens[3,6] is proposed to be a pre-Q1 riboswitch by Phok et al.[6] (under the name RNA28) and confirmed in our analysis (**Table S1**). Besides previously proposed riboswitches and guide RNAs, we confirm the expression of Ssca, an archeal RNA of unknown function already identified by Klein[3] and Phok[6] and described as RFAM family RF00063.

Several long 5' UTRs are located upstream of genes that are known to be cis-regulated in bacteria, such as ribosomal protein genes (rps15p, rpl15e, rpl21e, rps6e, rpl7ae), and amino acid metabolism pathway genes (aminotransferase, N-terminal protease, leu-, val-, and trp-aminoacyl tRNA synthetase). The presence of regulatory leader sequences upstream of ribosomal protein genes was suggested in *S. solfataricus*.[8] Here, we observe that Archaeal cis-regulatory leaders may also involve amino acid biosynthesis genes.

Few of the novel long 5' UTRs are GC-rich (**Fig. 3B**; **Table S1**), which argues against extensive secondary structure such as found for instance in the T-boxes of bacterial aminoacyl tRNA synthetases. Most of the high-GC leaders are already annotated as C/D or H/ACA box RNAs. Only five new leaders of unknown function have a GC content above 50%. One of them, PabO66, is 71% GC and is located upstream of an uncharacterized CDS.

Novel H/ACA RNAs. H/ACA RNAs are well-characterized small non-coding RNAs present both in eukaryotes and archaeas. They are involved in RNA-guided modifications where a specific U residue, often in rRNA, is converted into a pseudouridine, a common base modification in the ribosome. Other RNA targets like tRNAs also exhibit specific pseudo-uridylations, some of which can be directed by the same RNA guided machinery.[30] In Pyrococcus genomes, seven H/ACA sRNAs have been identified by an in silico approach and validated experimentally.[5] Those H/ACA RNAs are annotated in the Pyrococcus genomes and can also be retrieved from their RFAM identifier (Pab19, Pab21-snoR9: RF00065, Pab160, Pab35-HgcF: RF00058, Pab40-HgcG: RF00064, Pab91, Pab105-HgcE: RF00060). They were initially identified using a structure descriptor corresponding to a helix-loop-helix motif where an internal loop of variable length (from five to 11 residues on both 5' and 3' ends) connects two stems: a basal stem and an apical stem including at least seven and five base pairs, respectively. Another essential feature in the descriptor is the presence of a terminal loop including an embedded K-turn or K-loop sub-motif which is required to recruit one of the proteins (L7Ae) necessary for the H/ACA sRNP assembly.

A new family of H/ACA-related motifs was recently discovered in Pyrobaculum genomes.[7] These non-canonical H/ACA sRNAs differ from their canonical relatives found in *P. abyssi* in that the first stem is absent. However, the truncated motifs still maintain the second stem and a K-turn motif, two structural features which are sufficient to preserve the function of these H/ACA-like motifs as RNA guides.[7] The 5' and 3' ends of the H/ACA-like motifs remain unpaired and may be used as antisense elements for the RNA-RNA interaction with its target(s). Although the most abundant motifs are H/ACA like motifs in Pyrobaculum, both H/ACA and H/ACA-like motifs coexist and are functional. Nevertheless, the H/ACA motifs in Pyrobaculum are slightly different from those in Pyrococcus with a basal stem shortened by one or two base pairs (five or six base pairs instead of seven or eight) that may also include bulged residues.

We compared candidates ncRNAs from our RNA-seq analysis with the results of a genome wide search for novel H/ACA RNAs containing H/ACA or H/ACA-like motifs similar to those found in Pyrobaculum (see Materials and Methods). Four new H/ACA RNA candidates were identified (shown with their possible targets in **Table S2**). **Figure 4** presents the predicted secondary structure of the three most reliable candidates (PabO1, PabO48, and PabO78), based on the criteria mentioned above (high GC content, high conservation, high abundance) in a 2D structure representation including the H/ACA(like) motif and its possible targets. These findings suggest H/ACA-like RNAs are not specific to crenarchaea, and may also occur in euryarchaea.

Discussion

In hyperthermophile bacteria and archaea, structured RNA such as tRNA, rRNA, or modification guide RNA have seldom been found with a GC content below 50% (this study and refs. 13 and 14—some C/D box RNAs with very short hairpin structures may constitute rare exceptions). It is tempting to generalize this observation and propose that no low-GC transcript can ever fold into a stable secondary structure at extreme high temperature. Indeed,

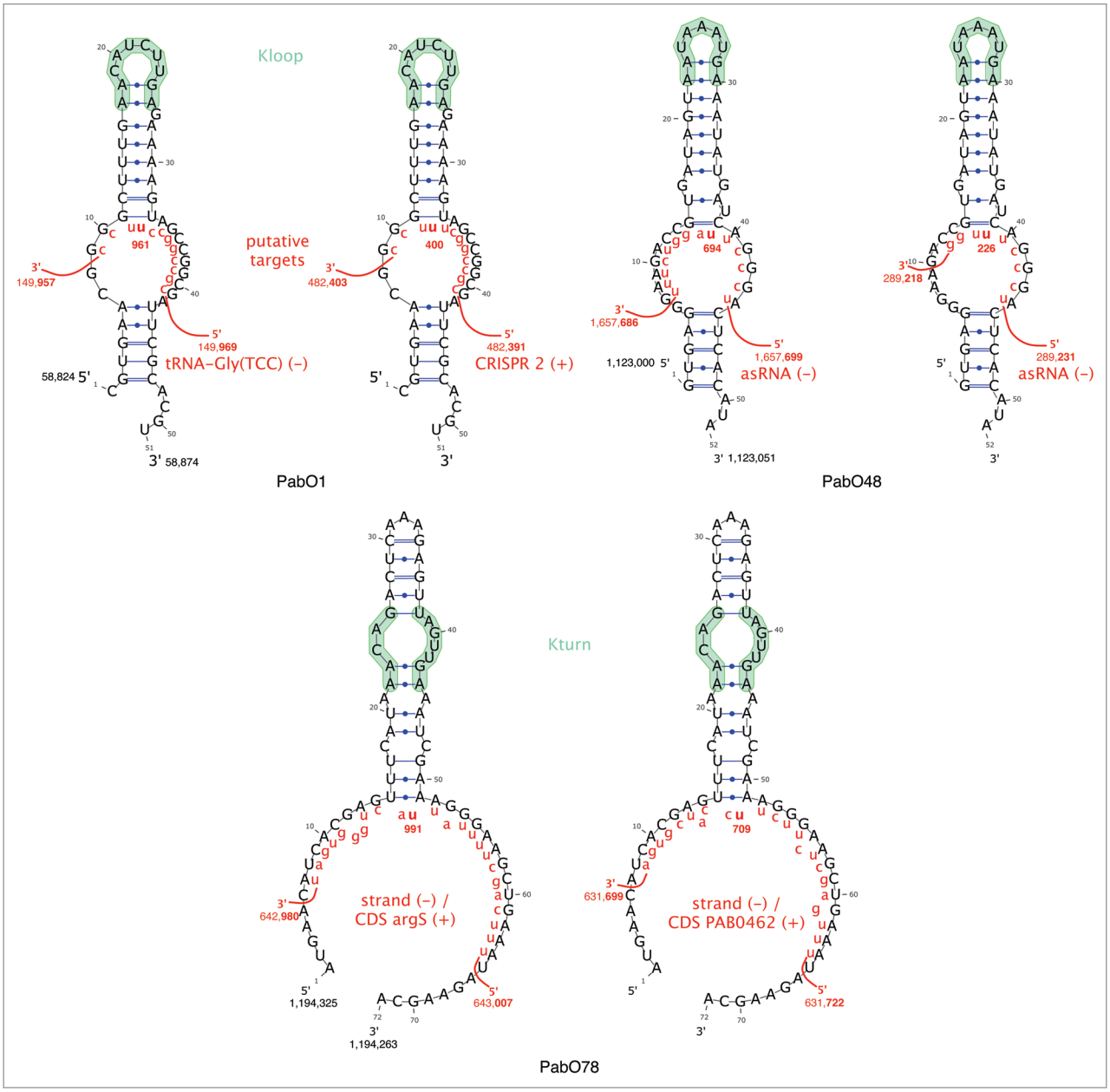

Figure 4. New H/ACA gene candidates (PabO1, PabO48, PabO78) corresponding to H/ACA or H/ACA-like motifs identified in RNA-seq ncRNAs. The 2D structure representations were generated using VARNA - version 3.9.[49] The guide RNA candidates include the annotation of the K-turn or K-loop motif (green). The RNA targets (red) are represented as paired to the guide RNA in the internal loop of the H/ACA motif or in the free ends of the H/ACA-like motif. The coding strand is indicated for all targets by a (+) or (–) symbol. Targets located on the non-coding strand with respect to the CDS, are noted "strand (–)." The CRISPR 2 (+) target follows the nomenclature adopted by Phok et al.[6] All targets shown are expressed at some degree in the genome.

factors other than a high GC content can contribute to RNA stabilization in hyperthermophiles. These include nucleotide modifications,[31] the presence of polyamines,[32] and high salt concentration, which was shown to stabilize DNA[33] and RNA.[34] Most likely, however, RNA molecules that require secondary structure to operate at high temperature should be, just like rRNAs and tRNAs, under selection for GC-rich helices in addition to any extra stabilizing factor. Consequently, we can reasonably assume an absence of secondary structure over most of the AU-rich sRNAs and mRNA leaders. However, an absence of secondary structure does not mean these RNAs are not functional. Indeed there are several strong indications of functionality in a large fraction of the AU-rich ncRNAs, including high abundance, the presence of strong promoters and, importantly, sequence conservation. Then what can these functions be? Small RNAs exert regulatory functions by targeting mRNAs and triggering RNA degradation or translation block. Cases have recently emerged showing such regulatory RNA activity exists in archaea.[12,35] Of course, RNA-RNA interactions would require at least a short GC-rich stretch for efficient targeting at high temperature, but this may not affect

the overall GC-content. Alternatively, AU-rich RNAs may act by recruiting proteins through sequence-based interactions or short local structures. Again, a short local GC-rich hairpin may form in an otherwise AU-rich RNA. This type of behavior would be reminiscent of the function of certain long non-coding RNAs (lncRNAs) in eukaryotes, which act mainly by recruiting proteins although they lack extensive sequence or structure signals.[36,37]

Pervasive transcription was first described in eukaryotic organisms and later extended to bacterial cells. It is now clear that archeal genomes also produce transcripts over most of their intergenic regions and antisense to a large number of genes. Klein et al.[3] observed that structured ncRNAs were rarer in the archaea *P. furiosus* and *M. janaschii* than in bacterial genomes. They hypothesized that the use of ncRNA could be selected against in high temperature environments, or that organisms with a reduced gene set such as many archaea did not require sophisticated RNA-based regulation. However, later transcriptome analysis of the reduced bacterial genome of *Mycoplasma pneumoniae* revealed over a hundred RNAs with potential regulatory functions, most of them antisense with respect to protein-coding genes.[38] Hence compact genomes may retain extensive RNA-based regulation. It is important though to dissociate RNA functionality from the existence of stable secondary structures. The emerging classes of eukaryotic regulatory RNAs such as endogenous siRNAs, lncRNAs, or XUTs[39] all seem to bind their RNA or protein targets without help from extensive secondary structure. Archaea have been invaluable models for studying various aspects of eukaryotic cell functions. It might turn out that RNA-based regulation is another key feature that may benefit from the archaeal model.

Materials and Methods

Strain growth. *Pyrococcus abyssi* GE5 was grown anaerobically in ASW-YT rich medium, containing artificial seawater, yeast extract and tryptone[40] at 92 °C. After two overnight transfers, cells were diluted to 2×10^6 cells/mL. After an initial growth of 3 h, samples (10 mL) were taken every 1.5 h until cells reach the stationary phase (11 h of growth). Samples were quickly cooled down in liquid nitrogen, spinned, and the pellets were frozen at −80°C.

RNA extraction and sequencing. Total RNAs were extracted using the TRIzol method (Invitrogen), following the manufacturer's instructions. RNA quality was monitored by agarose gel electrophoresis and concentration was measured using the NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific Inc.). A pool of equal amounts of each sample was checked for integrity by 2100 Bioanalyzer (Agilent Technologies Inc.) and treated to enrich in primary transcripts using a 5'-phosphate-dependent exonuclease (Terminator, Epicenter), following the manufacturer's instruction except for the amount of enzyme used in the reaction was 1 U per microgram of enriched RNAs and incubation was done at 30 °C for 1.5 h. Total RNA composition before and after treatment are shown in **Figure S8**. RNA samples were subsequently purified and concentrated using the RNA Clean-Up and Concentration Kit (Norgen Biotek Corp) before library preparation as described elsewhere in detail.[41] Briefly, the strand-specific RNA-seq template library was prepared starting from a pool of total samples (50 ng) following the directional mRNA-seq library preparation protocol provided by Illumina Inc. The library was sequenced (40 bp single-read) using an Illumina GA-IIx sequencer.

ncRNA classification into long 5' UTR, sRNA, asRNA. Oriented RNA-seq reads were mapped to the *P. abyssi* GE5 genome sequence using the Bowtie program.[42] Uniquely mapping reads were then processed using the DetR'prok ncRNA annotation pipeline,[43] available in the Galaxy Tool Shed (http://toolshed.g2.bx.psu.edu) under category "Sequence Analysis." Briefly, this workflow clusters mapped reads, compares clusters with previous annotation and generates extended annotations including 5' and 3' non-coding regions. Furthermore, extended annotations merge tandem protein coding, tRNA, and rRNA genes separated by intergenic regions shorter than 25 nt or fully covered by RNA-seq reads. The workflow then classifies non-coding fragments of extended annotations into long 5' UTRs, sRNAs, and asRNAs (**Fig. S5**). Antisense RNAs must be fully included in an extended annotation on the opposite strand. We used the *P. abyssi* genome annotation corrected by Gao et al.[18]

The ORF search was conducted using the EasyGene software[44] that uses a HMM model to score putative coding sequences based on codon statistics. The model is trained over a set of genes that is automatically extracted based on similarity with known genes.

Naming convention: we provide a name in the form PabOx (*P. abyssi* Orsay *x*) to any novel "high interest" RNA that was not experimentally confirmed in prior publications. We define as high interest any RNA meeting at least one criterion among the following. For sRNAs: high-GC, high conservation, high-RPKM, strong promoter; for long 5' UTRs: high-GC, high conservation, high-RPKM; for asRNA: high-GC, strong promoter, long size. We define high-GC as > 50% GC; high conservation as conserved in five species or more (see below definition of conservation), high-RPKM as more than twice the median RPKM value for this type of element and long size as over 200 nt.

Annotated ncRNAs and extended_annotations are available in **Supplemental Material** as GFF (General Feature Format) files providing locations, names, and qualitative information for each transcript (**Fig. S1**).

Promoter motif detection. We used the MEME web server[45] using default options except for "search given strand only" to detect motifs in the upstream sequences of CDS, transcription units, long 5' UTR, sRNA, and asRNA (**Fig. 1**). We extracted 50 nt sequences upstream of each annotation. Upstream sequences were excluded when another CDS was encountered before 50 nt. Sequences shorter than 10 nt were excluded from analysis.

The list of sRNAs with both BRE and TATA box motifs (**Table S1**) was produced as the combination of two searches: (1) the results of a MEME search from the 50 nt upstream sequences of all sRNAs and (2) the presence of a BRE-TATA box motif in the sRNA upstream region (50 nt) given by a FIMO[46] search, performed over the whole genome using the BRE-TATA box motif defined from 50 nt upstream regions of expressed genes (default option with *P* value threshold set to 10e-3).

Conservation analysis. All identified ncRNA sequences were submitted to a BLASTN (version 2.2.15 with parameters: -W 7 -r 2 -q -3 -G 5 -E 2 -e 10) search on a database of 76 archaeal genomes, selected based both on their "complete" status given by the Genomes OnLine Database (http://www.genomesonline.org) in March 2012, and their availability in the NCBI genome repository. An empirical filter to remove BLASTN alignments with a bit score below 42 (corresponding to an expected value of 0.06) was applied and all retained hits were inspected to assess the extent of conservation for each query. Inspection was based on the taxonomy lineage offered by the ORGANISM record of the genome GenBank file, where "Archaea" is the deepest level and species strain the uppermost level. Each BLASTN hit corresponds to a species and, thus, a lineage. For each BLASTN hit, the uppermost common level of the two lineages was retained. The conservation level of a sequence element (**Fig. 2B**; **Table S1**) was then given by the uppermost common level among all lineages defined by the BLASTN hits of this element. For example, an RNA is found in two species: *Pyrococcus abyssi* and *Methanosarcina acetivorans*. The taxonomic description for *P. furiosus* is "Archaea; Euryarchaeota; Thermococci; Thermococcales; Thermococcaceae; Pyrococcus; abyssi" and that for *M. acetivorans* is "Archaea; Euryarchaeota; Methanomicrobia; Methanosarcinales; Methanosarcinaceae; Methanosarcina; acetivorans." Then the uppermost common level between these two species is "Euryarchaeota."

H/ACA-like motif search. A standard descriptor-based search was performed using RNAMotif[47] to identify canonical or non-canonical H/ACA motifs with structural features similar to those found in Pyrobaculum[7] except that we also allowed the possible substitution of the K-turn motif by a K-loop. A series of filters were then used to post-process the results, which initially included around 3300 hits: around 100 H/ACA or H/ACA-like motifs with a K-turn, ~3100 for H/ACA or H/ACA-like motifs with a K-loop. The first filter was expression-based, retaining only hits in expressed regions from the RNA-seq data, thus eliminating 33% and 87% of the H/ACA(like) motifs with a K-turn and a K-loop, respectively. The second filter was sequence and structure conservation-based, keeping hits conserved in other archaea where the H-ACA(like) motif is also preserved. The third filter was function-based, looking for the presence of potential RNA targets (using the YASS program[48]) that might be chemically modified by the RNA guide machinery. The following base pairing constraints were imposed between guide and target: a minimum of seven "consecutive" base pairs (spanning the unpaired UN dinucleotide corresponding to the pseudouridylable position) for the canonical H/ACA motifs where the first stem is still present, but slightly truncated with respect to the typical H/ACA motifs in *Pyrococcus abyssi* (e.g., Pae sR201 and sR202 in *Pyrobaculum aerophilum*, see Bernick et al.[7]). In the case of H/ACA-like motifs, a minimum of 11 consecutive base pairs was required. In both cases, the constraints are consistent with the minimum number of base pairs observed in H/ACA or H/ACA-like motifs when paired to their RNA target(s).[5,7] The basic workflow of the full search is summarized in **Figure S6**.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

RNA library preparation and sequencing were performed at the CNRS Imagif platform. The bioinformatics pipeline was developed with help from the eBio bioinformatics platform. Work was funded in part by Agence Nationale pour la Recherche (grant ANR-2010-BLAN-1602-01) "Duplex-omics," and by a Marie Curie research fellowship to EC.

Supplemental Material

Supplemental material may be found here:
www.landesbioscience.com/journals/rnabiology/article/25567

References

1. Lasa I, Toledo-Arana A, Gingeras TR. An effort to make sense of antisense transcription in bacteria. RNA Biol 2012; 9:1039-44; PMID:22858676; http://dx.doi.org/10.4161/rna.21167
2. Tang TH, Bachellerie JP, Rozhdestvensky T, Bortolin ML, Huber H, Drungowski M, et al. Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. Proc Natl Acad Sci USA 2002; 99:7536-41; PMID:12032318; http://dx.doi.org/10.1073/pnas.112047299
3. Klein RJ, Misulovin Z, Eddy SR. Noncoding RNA genes identified in AT-rich hyperthermophiles. Proc Natl Acad Sci USA 2002; 99:7542-7; PMID:12032319; http://dx.doi.org/10.1073/pnas.112063799
4. Schattner P. Searching for RNA genes using base-composition statistics. Nucleic Acids Res 2002; 30:2076-82; PMID:11972348; http://dx.doi.org/10.1093/nar/30.9.2076
5. Muller S, Leclerc F, Behm-Ansmant I, Fourmann JB, Charpentier B, Branlant C. Combined in silico and experimental identification of the Pyrococcus abyssi H/ACA sRNAs and their target sites in ribosomal RNAs. Nucleic Acids Res 2008; 36:2459-75; PMID:18304947; http://dx.doi.org/10.1093/nar/gkn077
6. Phok K, Moisan A, Rinaldi D, Brucato N, Carpousis AJ, Gaspin C, et al. Identification of CRISPR and riboswitch related RNAs among novel noncoding RNAs of the euryarchaeon Pyrococcus abyssi. BMC Genomics 2011; 12:312; PMID:21668986; http://dx.doi.org/10.1186/1471-2164-12-312
7. Bernick DL, Dennis PP, Höchsmann M, Lowe TM. Discovery of Pyrobaculum small RNA families with atypical pseudouridine guide RNA features. RNA 2012; 18:402-11; PMID:22282340; http://dx.doi.org/10.1261/rna.031385.111
8. Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R. A single-base resolution map of an archaeal transcriptome. Genome Res 2010; 20:133-41; PMID:19884261; http://dx.doi.org/10.1101/gr.100396.109
9. Danan M, Schwartz S, Edelheit S, Sorek R. Transcriptome-wide discovery of circular RNAs in Archaea. Nucleic Acids Res 2012; 40:3131-42; PMID:22140119; http://dx.doi.org/10.1093/nar/gkr1009
10. Randau L. RNA processing in the minimal organism Nanoarchaeum equitans. Genome Biol 2012; 13:R63; PMID:22809431; http://dx.doi.org/10.1186/gb-2012-13-7-r63
11. Straub J, Brenneis M, Jellen-Ritter A, Heyer R, Soppa J, Marchfelder A. Small RNAs in haloarchaea: identification, differential expression and biological function. RNA Biol 2009; 6:281-92; PMID:19333006; http://dx.doi.org/10.4161/rna.6.3.8357
12. Jäger D, Pernitzsch SR, Richter AS, Backofen R, Sharma CM, Schmitz RA. An archaeal sRNA targeting cis- and trans-encoded mRNAs via two distinct domains. Nucleic Acids Res 2012; 40:10964-79; PMID:22965121; http://dx.doi.org/10.1093/nar/gks847
13. Galtier N, Lobry JR. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol 1997; 44:632-6; PMID:9169555; http://dx.doi.org/10.1007/PL00006186
14. Grogan DW. Hyperthermophiles and the problem of DNA instability. Mol Microbiol 1998; 28:1043-9; PMID:9680196; http://dx.doi.org/10.1046/j.1365-2958.1998.00853.x
15. Hurst LD, Merchant AR. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc Biol Sci 2001; 268:493-7; PMID:11296861; http://dx.doi.org/10.1098/rspb.2000.1397

16. Marguet E, Forterre P. DNA stability at temperatures typical for hyperthermophiles. Nucleic Acids Res 1994; 22:1681-6; PMID:8202372; http://dx.doi.org/10.1093/nar/22.9.1681
17. Cohen GN, Barbe V, Flament D, Galperin M, Heilig R, Lecompte O, et al. An integrated analysis of the genome of the hyperthermophilic archaeon Pyrococcus abyssi. Mol Microbiol 2003; 47:1495-512; PMID:12622808; http://dx.doi.org/10.1046/j.1365-2958.2003.03381.x
18. Gao J, Wang J. Re-annotation of two hyperthermophilic archaea *Pyrococcus abyssi* GE5 and *Pyrococcus furiosus* DSM 3638. Curr Microbiol 2012; 64:118-29; PMID:22057919; http://dx.doi.org/10.1007/s00284-011-0035-x
19. Soppa J. Normalized nucleotide frequencies allow the definition of archaeal promoter elements for different archaeal groups and reveal base-specific TFB contacts upstream of the TATA box. Mol Microbiol 1999; 31:1589-92; PMID:10200975; http://dx.doi.org/10.1046/j.1365-2958.1999.01274.x
20. Thomm M, Wich G, Brown JW, Frey G, Sherf BA, Beckler GS. An archaebacterial promoter sequence assigned by RNA polymerase binding experiments. Can J Microbiol 1989; 35:30-5; PMID:2497942; http://dx.doi.org/10.1139/m89-005
21. Brenneis M, Hering O, Lange C, Soppa J. Experimental characterization of Cis-acting elements important for translation and transcription in halophilic archaea. PLoS Genet 2007; 3:e229; PMID:18159946; http://dx.doi.org/10.1371/journal.pgen.0030229
22. Jäger D, Sharma CM, Thomsen J, Ehlers C, Vogel J, Schmitz RA. Deep sequencing analysis of the Methanosarcina mazei Gö1 transcriptome in response to nitrogen availability. Proc Natl Acad Sci USA 2009; 106:21878-82; PMID:19996181; http://dx.doi.org/10.1073/pnas.0909051106
23. 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(Database issue):D141-5; PMID:21062808; http://dx.doi.org/10.1093/nar/gkq1129
24. Gaspin C, Cavaillé J, Erauso G, Bachellerie JP. Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the Pyrococcus genomes. J Mol Biol 2000; 297:895-906; PMID:10736225; http://dx.doi.org/10.1006/jmbi.2000.3593
25. Dennis PP, Omer A, Lowe T. A guided tour: small RNA function in Archaea. Mol Microbiol 2001; 40:509-19; PMID:11359559; http://dx.doi.org/10.1046/j.1365-2958.2001.02381.x
26. Zago MA, Dennis PP, Omer AD. The expanding world of small RNAs in the hyperthermophilic archaeon Sulfolobus solfataricus. Mol Microbiol 2005; 55:1812-28; PMID:15752202; http://dx.doi.org/10.1111/j.1365-2958.2005.04505.x
27. Muller S, Charpentier B, Branlant C, Leclerc F. A dedicated computational approach for the identification of archaeal H/ACA sRNAs. Methods Enzymol 2007; 425:355-87; PMID:17673091; http://dx.doi.org/10.1016/S0076-6879(07)25015-3
28. Baker JL, Sudarsan N, Weinberg Z, Roth A, Stockbridge RB, Breaker RR. Widespread genetic switches and toxicity resistance proteins for fluoride. Science 2012; 335:233-5; PMID:22194412; http://dx.doi.org/10.1126/science.1215063
29. Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, et al. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. Genome Biol 2010; 11:R31; PMID:20230605; http://dx.doi.org/10.1186/gb-2010-11-3-r31
30. Muller S, Urban A, Hecker A, Leclerc F, Branlant C, Motorin Y. Deficiency of the tRNATyr:Psi 35-synthase aPus7 in Archaea of the Sulfolobales order might be rescued by the H/ACA sRNA-guided machinery. Nucleic Acids Res 2009; 37:1308-22; PMID:19139072; http://dx.doi.org/10.1093/nar/gkn1037
31. Kowalak JA, Dalluge JJ, McCloskey JA, Stetter KO. The role of posttranscriptional modification in stabilization of transfer RNA from hyperthermophiles. Biochemistry 1994; 33:7869-76; PMID:7516708; http://dx.doi.org/10.1021/bi00191a014
32. Imanaka T. Molecular bases of thermophily in hyperthermophiles. Proc Jpn Acad Ser B Phys Biol Sci 2011; 87:587-602; PMID:22075760; http://dx.doi.org/10.2183/pjab.87.587
33. Marguet E, Forterre P. Protection of DNA by salts against thermodegradation at temperatures typical for hyperthermophiles. Extremophiles 1998; 2:115-22; PMID:9672686; http://dx.doi.org/10.1007/s007920050050
34. Hethke C, Bergerat A, Hausner W, Forterre P, Thomm M. Cell-free transcription at 95 degrees: thermostability of transcriptional components and DNA topology requirements of Pyrococcus transcription. Genetics 1999; 152:1325-33; PMID:10430563
35. Prasse D, Ehlers C, Backofen R, Schmitz RA. Regulatory RNAs in archaea: first target identification in Methanoarchaea. Biochem Soc Trans 2013; 41:344-9; PMID:23356309; http://dx.doi.org/10.1042/BST20120280
36. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 2011; 25:1915-27; PMID:21890647; http://dx.doi.org/10.1101/gad.17446611
37. Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 2011; 477:295-300; PMID:21874018; http://dx.doi.org/10.1038/nature10398
38. Güell M, van Noort V, Yus E, Chen WH, Leigh-Bell J, Michalodimitrakis K, et al. Transcriptome complexity in a genome-reduced bacterium. Science 2009; 326:1268-71; PMID:19965477; http://dx.doi.org/10.1126/science.1176951
39. Tisseur M, Kwapisz M, Morillon A. Pervasive transcription - Lessons from yeast. Biochimie 2011; 93:1889-96; PMID:21771634; http://dx.doi.org/10.1016/j.biochi.2011.07.001
40. Sato T, Fukui T, Atomi H, Imanaka T. Targeted gene disruption by homologous recombination in the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1. J Bacteriol 2003; 185:210-20; PMID:12486058; http://dx.doi.org/10.1128/JB.185.1.210-220.2003
41. Toffano-Nioche C, Nguyen AN, Kuchly C, Ott A, Gautheret D, Bouloc P, et al. Transcriptomic profiling of the oyster pathogen Vibrio splendidus opens a window on the evolutionary dynamics of the small RNA repertoire in the Vibrio genus. RNA 2012; 18:2201-19; PMID:23097430; http://dx.doi.org/10.1261/rna.033324.112
42. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009; 10:R25; PMID:19261174; http://dx.doi.org/10.1186/gb-2009-10-3-r25
43. Toffano-Nioche C, Luo Y, Kuchly C, Wallon C, Steinbach D, Zytnicki M, et al. Detection of noncoding RNA in Bacteria and Archaea using the DETR'PROK Galaxy pipeline. Methods 2013; 13:00204-1; PMID:23806640
44. Larsen TS, Krogh A. EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics 2003; 4:21; PMID:12783628; http://dx.doi.org/10.1186/1471-2105-4-21
45. Bailey TL, Elkan C. "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California, 1994.
46. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics 2011; 27:1017-8; PMID:21330290; http://dx.doi.org/10.1093/bioinformatics/btr064
47. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res 2001; 29:4724-35; PMID:11713323; http://dx.doi.org/10.1093/nar/29.22.4724
48. Noé L, Kucherov G. YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Res 2005; 33(Web Server issue):W540-3; PMID:15980530; http://dx.doi.org/10.1093/nar/gki478
49. Darty K., Denise A, Ponty Y. VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics 2009; 25:1974-5; PMID:19398448; http://dx.doi.org/10.1093/bioinformatics/btp250