frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 02 July 2012
doi: 10.3389/fmicb.2012.00231

Diversity of antisense and other non-coding RNAs in archaea revealed by comparative small RNA sequencing in four *Pyrobaculum* species

David L. Bernick[1], Patrick P. Dennis[2], Lauren M. Lui[1] and Todd M. Lowe[1]*

[1] *Department of Biomolecular Engineering, University of California, Santa Cruz, CA, USA*
[2] *Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA*

Edited by:
Frank T. Robb, University of California, USA

Reviewed by:
Mircea Podar, Oak Ridge National Laboratory, USA
Imke Schroeder, University of California Los Angeles, USA
Matthias Hess, Washington State University, USA
Lanming Chen, Shanghai Ocean University, China

***Correspondence:**
Todd M. Lowe, Department of Biomolecular Engineering, University of California, 1156 High Street, Santa Cruz, CA 95064, USA.
e-mail: lowe@soe.ucsc.edu

A great diversity of small, non-coding RNA (ncRNA) molecules with roles in gene regulation and RNA processing have been intensely studied in eukaryotic and bacterial model organisms, yet our knowledge of possible parallel roles for small RNAs (sRNA) in archaea is limited. We employed RNA-seq to identify novel sRNA across multiple species of the hyperthermophilic genus *Pyrobaculum*, known for unusual RNA gene characteristics. By comparing transcriptional data collected in parallel among four species, we were able to identify conserved RNA genes fitting into known and novel families. Among our findings, we highlight three novel *cis*-antisense sRNAs encoded opposite to key regulatory (ferric uptake regulator), metabolic (triose-phosphate isomerase), and core transcriptional apparatus genes (transcription factor B). We also found a large increase in the number of conserved C/D box sRNA genes over what had been previously recognized; many of these genes are encoded antisense to protein coding genes. The conserved opposition to orthologous genes across the *Pyrobaculum* genus suggests similarities to other *cis*-antisense regulatory systems. Furthermore, the genus-specific nature of these sRNAs indicates they are relatively recent, stable adaptations.

Keywords: antisense small RNA, archaea, transcriptome sequencing, comparative genomics, gene regulation, C/D box small RNA

INTRODUCTION

Archaeal species are known to encode a plethora of small RNA (sRNA) molecules. These sRNAs have a multitude of functions including suppression of messenger RNA (mRNA; Straub et al., 2009), targeting modifications to ribosomal (rRNA) or transfer RNA (tRNA; Omer et al., 2000; Bernick et al., 2012), specifying targets of the CRISPR immune defense system (Barrangou et al., 2007; Hale et al., 2008; Hale et al., 2009), *cis*-antisense regulation of transposase mRNA (Tang et al., 2002; Tang et al., 2005; Jager et al., 2009; Wurtzel et al., 2010), and encoding short proteins less than 30 amino acids in length (Jager et al., 2009).

Only a few previous studies have described sRNA genes in the phylum Crenarchaeota. In the *Sulfolobus* genus, C/D box and H/ACA-box guide sRNAs have been studied, including 18 guide sRNAs in *Sulfolobus acidocaldarius* (Omer et al., 2000), nine in *S. solfataricus* (Zago et al., 2005), and corresponding homologs detected computationally in *S. tokodaii* (Zago et al., 2005). These sRNAs form two distinct classes of guide RNAs: C/D box sRNAs which guide 2′-*O*-methylation of ribose, and H/ACA-box guide RNAs which direct isomerization of uridine to pseudouridine. Eukaryotes also share these two classes of guide RNAs with the same functions, but these homologs are dubbed small nucleolar RNAs (snoRNAs) because of their cellular localization. Recently, we employed high-throughput sequencing to identify ten conserved, novel families of H/ACA-like sRNA within the genus *Pyrobaculum* (Bernick et al., 2012).

Sulfolobus solfataricus has been further characterized using high-throughput sequencing (Wurtzel et al., 2010), revealing 18 CRISPR-associated sRNAs, 13 C/D box sRNAs, 28 *cis*-antisense encoded transposon-associated sRNAs, and 185 sRNA genes encoded antisense to other, non-transposon protein coding genes. It is unclear how many of the latter antisense transcripts are the result of transcriptional noise, overlapping but non-interacting gene products, or biologically relevant products of functional ncRNA genes. The diversity of sRNA genes is just beginning to be studied in depth in other members of the Crenarchaeota.

Genes that produce sRNA antisense to mRNA are known in all three domains of life and many of these sRNA have provided interesting examples of novel regulation. Within bacteria, antisense sRNAs are known and well-studied (Repoila et al., 2003; Aiba, 2007; Vogel, 2009). For example, utilization and uptake of iron in *Escherichia coli* is modulated by the sRNA *RyhB* that acts in concert with the ferric uptake regulator (Fur) protein (Masse et al., 2007). The sRNA is coded in *trans* to its regulatory targets, and the Sm-like protein Hfq is required for its function. In *Pseudomonas aeruginosa*, an analogous regulatory mechanism exists with the *PrrF* regulatory RNA (Wilderman et al., 2004).

In this study, we adapted techniques pioneered by researchers studying microRNA in eukaryotes (Lau et al., 2001; Henderson et al., 2006; Lu et al., 2006), to execute parallel high-throughput pyrosequencing of sRNAs across four *Pyrobaculum* species. This comparative transcriptomic approach enabled us to identify

novel conserved sRNA transcripts among four related hyperthermophiles (*Pyrobaculum aerophilum*, *P. arsenaticum*, *P. calidifontis*, and *P. islandicum*). We provide an overview of the distribution of sRNAs across species, and focus on two major classes: the highly abundant C/D box sRNAs, and sRNAs antisense to three biologically important protein coding genes. We augment our transcriptional analyses further with comparative genomics utilizing two additional *Pyrobaculum* species with sequenced genomes: *P. neutrophilum* (recently renamed from *Thermoproteus neutrophilus*) and *P. oguniense* (NCBI GenBank accession NC_016885.1).

MATERIALS AND METHODS

CULTURE CONDITIONS

Pyrobaculum aerophilum cells were grown anaerobically in media containing 0.5 g/L yeast extract, 1× DSM390 salts, 10 g/L NaCl, 1× DSM 141 trace elements, 0.5 mg/L $Fe(SO_4)_2(NH_4)_2$, pH 6.5, with 10 mM $NaNO_3$. *P. islandicum* and *P. arsenaticum* cells were grown anaerobically in media containing 10 g/L tryptone, 2 g/L yeast extract, 1× DSM390 salts, 1× DSM88 trace elements, and 20 mM $Na_2S_2O_3$. *P. calidifontis* cells were grown aerobically in 1 L flasks using 500 mL media containing 10 g/L tryptone, 2 g/L yeast extract, 1× DSM88 trace metals, 15 mM $Na_2S_2O_3$, pH 6.8, loosely capped with moderate shaking at 125 rpm. Anaerobic cultures were grown in 2 L flasks with 1 L media, prepared under nitrogen with resazurin as a redox indicator at 0.5 mg/L; 0.25 mM Na_2S was added as a reductant. All cultures were grown at 95°C to late log or stationary phase, monitored at OD_{600}.

The 10× DSM390 salts are comprised of (per liter ddH_2O) 1.3 g $(NH_4)_2SO_4$, 2.8 g KH_2PO_4, 2.5 g $MgSO_4{\cdot}7H2O$. The 100× DSM88 trace metal solution is comprised (per liter 0.12 N HCl), 0.9 mM $MnCl_2$, 4.7 mM $Na_2B_4O_7$, 76 μM $ZnSO_4$, 25 μM $CuCl_2$, 12.4 μM $NaMoO_4$, 18 μM $VOSO_4$, 6 μM $CoSO_4$. The 100× DSM141 trace metal solution is comprised of 7.85 mM Nitrolotriacetic acid, 12.2 mM $MgSO_4$, 2.96 mM $MnSO_4$, 17.1 mM NaCl, 0.36 mM $FeSO_4$, 0.63 mM $CoSO_4$, 0.68 mM $CaCl_2$, 0.63 mM $ZnSO_4$, 40 μM $CuSO_4$, 42 μM $KAl(SO_4)_2$, 0.16 mM H_3BO_3, 41 μM Na_2MoO_4, 0.1 mM $NiCl_2$, 1.14 μM Na_2SeO_3.

cDNA LIBRARY PREPARATION

Two preparations were constructed for each of *P. aerophilum*, *P. islandicum*, *P. arsenaticum*, and *P. calidifontis* cultures, yielding a total of eight cDNA libraries. The following protocol was used for each preparation.

Total RNA was extracted from exponential or stationary cultures; 100 μg of each preparation was loaded onto a 15% polyacrylamide gel, and size selected in the range 15–70 nt. The gel was post-stained with SYBR Gold and the tRNA band was used as the upper exclusion point. The lower exclusion point was set at 75% of the region between xylene cyanol (XC) and bromophenol blue (BP) loading dye bands (Ambion protocol). Samples were eluted, EtOH precipitated, and 3′ linker (5′-adenylated, 3′ ddC) was added as described by Lau et al., 2001; IDTDNA, Linker 1). A second gel purification was performed as above, excising the gel fragment above the XC dye band to remove excess 3′ linker. The recovered linked RNAs were reverse transcribed (RT) using Superscript III (Invitrogen) with a DNA primer complementary to Linker 1. Following RT, Exonuclease I (EXO1, Thermo) was added to the RT reaction mixture, and incubated for 30 min to remove excess primer. We utilized standard alkaline lysis treatment with NaOH-EDTA at 80°C for 15 min to remove any residual RNA, as well as to inactivate the reverse transcriptase and the EXO1 ssDNA nuclease. Neutralization and small fragment removal was performed with water-saturated G50 columns (Ambion NucAway). The recovered single stranded cDNA was dried to near completion using a Servo SpeedVac, followed by a second 5′-adenylated linker addition (IDTDNA – Linker 2) to the cDNA using T4 RNA ligase (Ambion).

A 2 μL volume of this reaction was amplified by PCR (20 μL reaction, 16 cycles). This was followed by a second amplification (20 μL reaction, 16 cycles) using 2 μL from the first amplification as template using Roche 454-specific hybrid adapters based on the method described by Hannon[1]. A four-base barcode was included in the 5′ hybrid primer. The final reaction was cleaned using the Zymo clean kit following the manufacturer's protocol.

SEQUENCING AND READ MAPPING

Sequencing was performed using a Roche/454 GS FLX sequencer, and the GS emPCR Kit II (Roche). Sequencing reads described in this work are provided online via the UCSC Archaeal Genome Browser[2] (Chan et al., 2012).

Reads that included barcodes and sequencing linkers were selected from the raw sequencing data and used to identify reads from each of the eight pooled cDNA libraries. Reads were further consolidated, combining identical sequences with associated counts for viewing with the Archaeal Genome Browser. Reads were mapped to the appropriate genome [*P. aerophilum* (NC_003364.1); *P. arsenaticum* (NC_009376.1); *P. calidifontis* (NC_009073.1); *P. islandicum* (NC_008701.1); *P. oguniense* (NC_016885.1); *P. neutrophilum* (*T. neutrophilus*: NC_010525.1)] using BLAT (Kent, 2002), requiring a minimum of 90% identity (-minIdentity), a maximal gap of 3 (-maxIntron) and a minimum score (matches minus mismatches) of 16 (-minScore) using alignment parameters for this size range (-tileSize = 8 -stepSize = 4). Reads that mapped equally well to multiple positions in the genome were excluded from this study. The remaining, uniquely mapped reads were formatted and visualized as BED tracks within the UCSC Archaeal Genome Browser.

Of the 216,538 raw sequencing reads obtained, those that had readable barcodes and could be uniquely mapped to their respective genomes were: 39,294 in *P. calidifontis*, 30,827 in *P. aerophilum*, 31,206 in *P. arsenaticum*, and 42,951 in *P. islandicum*.

NORTHERN ANALYSIS

Northern blots were prepared using ULTRAhyb-Oligo (Ambion) following the manufacturer protocol[3] using Hybond-N+ (GE life sciences) membranes to transfer 10 μg/lane denatured total RNA (45 min, 50°C with glyoxyl loading buffer – Ambion). Size separation was conducted using 23 cm × 25 cm gels (1% agarose) in BPTE running buffer (30 mM bis-Tris, 10 mM PIPES, 1 mM EDTA, pH 6.5). The following DNA oligomers

[1] http://genoseq.ucla.edu/images/a/a9/SmallRNA.pdf
[2] http://archaea.ucsc.edu
[3] http://tools.invitrogen.com/content/sfs/manuals/8663MB.pdf

(Integrated DNA technologies) were used as probes: TFBi-iSense (CCTCCTCTGGAAAGCCCCTCAAGCTCCGA), TFBi-iAnti (TCGGAGCTTGAGGGGCTTTCCAGAGGAGG), PAEsR53 sense (GACCCCGATCGCCGAAAAATGACGAGTGGT).

COMPUTATIONAL PREDICTION OF ORTHOLOGOUS GENE CLUSTERS

Computational prediction of orthologous groups was established by computing reciprocal best BLASTP (Altschul et al., 1990; RBB) protein coding gene-pairs among pairs of four *Pyrobaculum* species. When at least three RBB gene-pairs select the same inter-species gene set (for example A pairs with B, B pairs with C, and C pairs with A), the cluster was considered an orthologous gene cluster.

COMPUTATIONAL PREDICTION OF C/D BOX sRNA HOMOLOG FAMILIES

C/D box sRNA homolog families were constructed from computational predictions with core C/D box features that were supported by transcripts from one or more of the four *Pyrobaculum* species (data from this study). Six *Pyrobaculum* genomes were searched for orthologs using these sRNA candidates as queries to BLASTN (Camacho et al., 2009). The highest scoring candidates were manually curated, then grouped into homologous C/D box sRNA families by multiple alignment.

RESULTS

SMALL RNA POPULATIONS

We prepared eight barcoded sequencing libraries using sRNA fractions (size range 16–70 nt) from anaerobic cultures of *P. aerophilum*, *P. arsenaticum*, *P. islandicum*, and an aerobic culture of *P. calidifontis*. These libraries were prepared using a 5′-independent ligation strategy (Pak and Fire, 2007) which preserves RNA strand orientation, captures both the 5′ and 3′ ends of the sRNA, and does not impose a bias for molecule selection based on 5′-phosphorylation state. Pyrosequencing, followed by selection of uniquely mapped sequence reads, allowed detection of reads associated with both known and novel genomic features (**Figure 1**), including:

(i) snoRNA-like guide RNAs, including known and novel C/D box sRNA and a new class of H/ACA-like sRNA (Bernick et al., 2012),
(ii) RNA sequences encoded *cis*-antisense (asRNA) to known protein coding genes,
(iii) RNA sequences derived from CRISPR arrays, thought to guide the CRISPR-mediated immune response,
(iv) unclassified novel sRNA, and
(v) degradation products of larger RNA including ribosomal RNA, messenger RNA and transfer RNA.

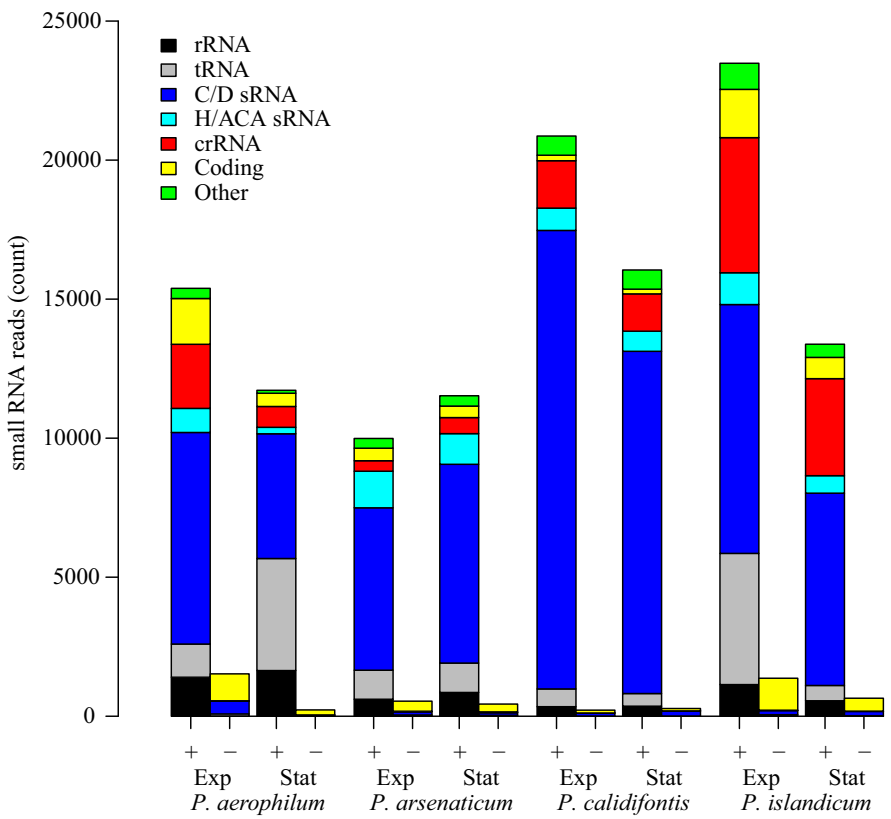

FIGURE 1 | Small RNA transcript abundance in four species of *Pyrobaculum*. Sense oriented reads (+) and antisense-oriented reads (−) shown in barplots for each species. Samples of each species were taken at both exponential (Exp) and stationary phases (Stat). RNA classifications were made based on mapping to genes coding for C/D box sRNA (C/D sRNA), H/ACA-like sRNA, CRISPR arrays (crRNA), fragments of coding regions (coding), ribosomal RNA (rRNA), and transfer RNA (tRNA).

Most antisense-oriented sequencing reads are associated with coding regions (**Figure 1**) in each of the species and growth phases examined. Antisense-oriented reads are frequently the result of convergent expression of a protein coding gene and a snoRNA-like guide RNA (**Tables A1–A4** in Appendix). We find, in some cases, that sequencing reads that appear to be antisense to snoRNA-like RNAs appear to be fragments of novel 3′ untranslated regions (3′ UTRs) of a convergently expressed protein coding region. These antisense-oriented sRNA reads are counted as antisense to the associated snoRNA-like sRNA. We made use of this transcriptional pattern to find novel C/D box sRNA and H/ACA-like sRNA; in these cases, highly abundant antisense reads to coding transcripts often proved to be a hallmark of novel C/D box and H/ACA-like sRNA (**Tables A2** and **A4** in Appendix). In a few remaining cases, we found novel *cis*-encoded antisense reads that were not derived from known classes of sRNA. We note that the proportion of reads belonging to each type of classified RNA is relatively stable across species and conditions (**Figure 1**), with the exception of two conditions in which tRNA fragments were enriched (*P. aerophilum* stationary phase, *P. islandicum* exponential phase). We are further investigating these differences, however the purpose and design of the sequencing portion of this study was aimed at qualitative discovery of novel sRNAs.

C/D BOX sRNA ACCOUNT FOR THE LARGEST FRACTION OF READS IN ALL SPECIES TESTED

In each of the eight small transcriptomes studied (four species sampled at exponential and stationary phase), C/D box sRNA accounted for the largest fraction of reads (**Figure 1**). A previous study (Fitz-Gibbon et al., 2002) has provided computational evidence for 65 C/D box sRNA candidates encoded in the genome of *P. aerophilum*. We now find an additional 23 C/D box sRNA candidates in that genome, representing a 35% increase in family size. By using transcriptional support from the four examined genomes (this study), combined with comparative genomic evidence that includes *P. oguniense* and *P. neutrophilum*, we find at least 74 C/D box sRNA in each *Pyrobaculum* spp. (**Table 1**). Of those genes, 70 appear to be conserved among all six genomes examined (**Figure 2**).

Table 1 | C/D box sRNA genes in each *Pyrobaculum* species based on transcriptional evidence or inferred by homology (*P. oguniense* and *P. neutrophilum*).

Species	C/D box sRNAs
P. aerophilum	88
P. arsenaticum	83
P. calidifontis	88
P. islandicum	84
P. oguniense	83
P. neutrophilum	74

All loci are manually curated.

CONVERGENTLY ORIENTED ncRNA ARE FREQUENTLY FOUND AT THE 3′ TERMINUS OF PROTEIN CODING GENES

It has been noted previously that in the genomes of *S. acidocaldarius* and *S. solfataricus*, C/D box sRNA genes occasionally exhibit antisense overlap to the 3′ end of protein encoding genes (Dennis et al., 2001). In the *Pyrobaculum* clade, we find numerous instances of a convergently oriented C/D box or H/ACA-like guide RNA gene that partially overlap, by a few nucleotides, the 3′ end of a protein-coding gene (**Tables A2** and **A4** in Appendix).

To find conserved, novel *cis*-encoded antisense RNA, we ranked conserved transcript abundance that overlapped orthologous protein coding genes. Among the top 34 predicted ortholog groups of genes with well-annotated function and conserved 3′ antisense transcription (**Table A2** in Appendix), 28 are convergent with C/D box sRNA and three are convergent with H/ACA-like sRNA. Among the top 19 predicted ortholog groups of unknown function with 3′ antisense transcription (**Table A4** in Appendix), 11 are convergent with C/D box sRNA, four are convergent H/ACA-like sRNA, and one is adjacent to a tRNA. Together, 87% of conserved, *cis*-antisense encoded sRNA are snoRNA-like guides, while only 2.6% are tRNA. In *P. aerophilum*, C/D box sRNA genes are nearly twice as abundant (88 compared to 46) as tRNA genes, but the sRNA genes are over 40-fold more likely to have a conserved overlap with the orthologous protein coding region. This may be an indication that these C/D box sRNA play a regulatory role with respect to the associated protein coding genes.

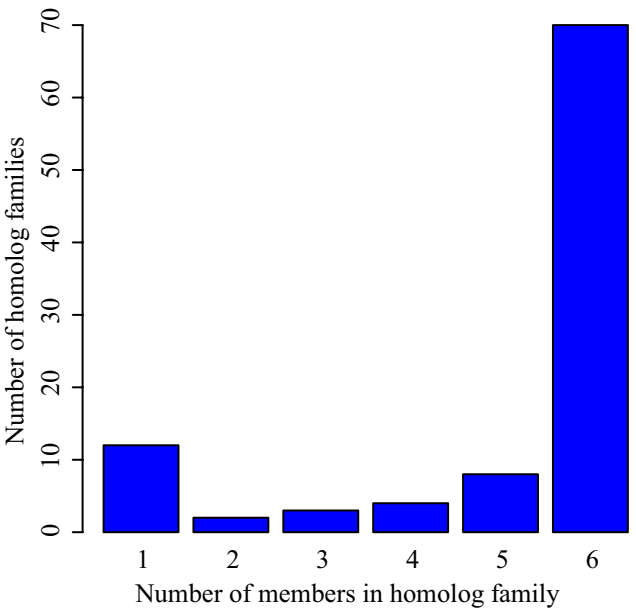

FIGURE 2 | Conservation of C/D box sRNA genes among six *Pyrobaculum* genomes. C/D box sRNA genes were organized by homolog family based on the location targeted by the encoded guide regions. These homolog families were then compared among the six studied species to verify conservation. While each individual species encodes more than 74 C/D box sRNA genes, 70 of those are conserved among each of the six studied *Pyrobaculum* spp. (group 6).

A notable example of a convergent ncRNA occurs at the 3′ terminus of the *electron transport flavoprotein* (*etf*) operon, where a C/D box sRNA, PAEsR53, overlaps the terminal gene (PAE0721 in *P. aerophilum*) in this four-gene operon. Like other operons within the *Pyrobaculum* genus, multiple promoters appear to drive expression of the *etf* operon (**Figure 3**). For this operon, an upstream promoter generates a 3400-nt-long full length *etfDH-ferredoxin-etfB-etfA* transcript. Two predicted internal promoters appear to generate respectively, the *ferredoxin-etfB-etfA* ∼2250 nt transcript, and the *etfA*-only 1040 nt transcript.

The *P. aerophilum* sRNA sequencing data revealed a strong abundance of sequences mapping to PAEsR53, as well as sequences of the same general size and location, mapping to the opposite strand (the UTR of the *etf* operon). Northern hybridization was performed to determine the origin of these "anti-PAEsR53" reads. **Figure 3** shows that these reads likely originate from the overlapping 3′ UTR of the *etf* operon, suggesting a possible interaction of the C/D box machinery with the *etf* mRNA. Predicted orthologs of this C/D box sRNA (PAEsR53) are syntenic with *etfA* in all *Pyrobaculum* species studied, overlapping the 3′ end of *etfA* orthologs by ∼12 bases. The overlap positions the D box guide sequence of PAEsR53 over the *etfA* stop codon in all *Pyrobaculum* species. If the guide RNA interacts through complementarity with the *etfA* mRNA, it could enable a 2′-*O*-methyl modification of the central "A" nucleotide within the conserved TAA stop codon in all four species.

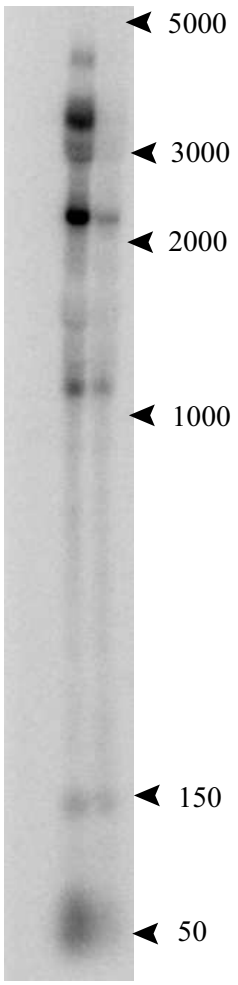

FIGURE 3 | Northern analysis of the 3′ UTR of the electron transport flavoprotein (*etf*) operon. *P. aerophilum* total RNA, exponential phase (left lane) and stationary phase (right lane). The probe was designed to anneal beyond the stop codon of the terminal gene in the *etf* operon, in the region of the convergently oriented PAEsR53 C/D box sRNA. Multiple bands at 3400, 2250, and 1040 nt are consistent with the *etf* operon and suboperon transcripts. The band near 50 nt, consistent with the RNA sequencing data, shows an apparent antisense transcript to PAEsR53 (sense relative to the 3′ UTR of the *etf* operon).

THE TRANSCRIPTION INITIATION FACTOR B GENES, *tfb*1 AND *tfb*2

The genomes of *Pyrobaculum* species contain a pair of paralogous genes that encode alternate forms of transcription initiation factor B (TFB). This factor is required for the initiation of basal level transcription at archaeal promoters (Santangelo et al., 2007).

In every sequenced *Pyrobaculum* species, TFB1 (PAE1645 and orthologs) contains a short N-terminal extension (22 amino acids in *P. aerophilum*) that is not present in the TFB2 proteins (PAE3329 and orthologs). Sequencing data reveals the presence of an abundant sRNA (*asR1*) encoded on the antisense strand that overlaps the 5′ end of *tfb1* (**Figure 4A**) in all four *Pyrobaculum* species examined (**Table A1** in Appendix). *Tfb1* also appears to have two promoters separated by 17–18 nt, such that the upstream promoter (P_u) is positioned to drive expression of full length *tfb1*, while the downstream promoter (P_d) generates transcripts that would lack a start codon near the start of the transcript.

In *P. aerophilum*, *asR1* sRNA is about 59 nt in length (**Table 2**; **Figure 4**), with a well-defined 5′ end that overlaps the extension region of the *tfb1* gene. The 3′ end of *asR1* is located just upstream of the *tfb1* translation initiation codon, precisely at the predicted start of transcription consistent with the P_u promoter. Importantly, there is an additional set of *asR1* sRNA reads of 41 nt in length, starting at the same 5′ position but terminating early, at the 5′ end of tfb1 transcripts consistent with the alternate P_d promoter. Mirroring the two variants of the antisense *asR1* transcript, deep sequencing revealed a large number of short sense strand sequencing reads, consistent with fragments representing the 5′ end of *tfb1* transcripts generated by P_u and P_d, spanning 50 and 32 nt in length respectively.

Northern analysis of total RNA from *P. aerophilum* confirmed the presence of a population of sense oriented transcripts of about 1000 nt in length, consistent with full length mRNA and another transcript population consistent with the sense oriented sRNAs described above (**Figure 5A**). When the antisense sRNA is probed, a population of short transcripts near 50 nt is detected (**Figure 5B**). The full length sense transcripts appear to be relatively constant in abundance across growth phase and culture conditions, consistent with data from a prior microarray study using the same RNA samples (Cozen et al., 2009). The correlated abundance of sense and antisense sRNA (**Figures 5C–E**) suggests that these sense::antisense pairs are associated, potentially as a double-stranded RNA. The elevated abundance of these pairs relative to

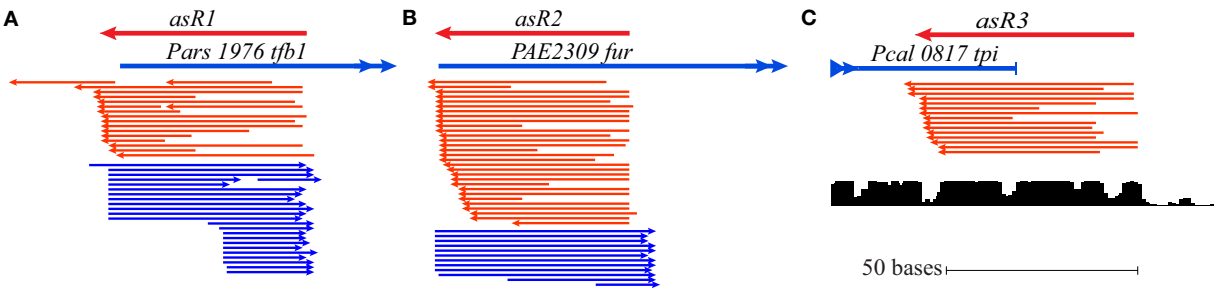

FIGURE 4 | *Cis*-antisense transcription occurring at three conserved loci in *Pyrobaculum* spp. The antisense sRNA genes (indicated as *asR1*, *asR2*, and *asR3*) are defined over the region of antisense-oriented sequencing reads (shown in red) with the associated gene (sense orientation) shown in blue. **(A)** Sequencing reads map to the 5′ terminus of transcription initiation factor B in both sense and antisense orientations (Pars_1976 is shown from *Pyrobaculum arsenaticum*). Sense oriented transcripts (blue) are present in two populations, consistent with the two conserved promoters, P_u and P_d. **(B)** The antisense sRNA (*asR2*) gene is defined over the region of antisense-oriented reads that map to the 5′ terminus of the ferric uptake regulator gene (*fur*; PAE2309 from *Pyrobaculum aerophilum* is shown). **(C)** Antisense-oriented reads map to the 3′ end of the triose-phosphate isomerase (*tpi*) locus (Pcal_0817 from *Pyrobaculum calidifontis*). Sequence conservation [lower graphic **(C)**] extends beyond the *tpi* stop codon (UCSC Archaeal Browser; Chan et al., 2012).

Table 2 | Terminal cis-antisense encoded sRNA in *Pyrobaculum* species.

sRNA		Len	Sequence
asR1	Pae	59	5′-AACTCGGAGCTTGAGGGGCTTTCCAGAGGAGGGGGGATTTGAGACCGACATAGCGTGTT
	Par	79	5′-TATGCGGAGCTTTAGGGGCTTGCCGGAAGAAGGTAGGCTTGTACTCGACATAGCGTGTTTATAAGCTTTTCTAGCGTAT
	Pca	33	5′-..TACGGAATTTTAAGGGCTTGCCGGGCGGGGTAG
	Pis	63	5′-.ATACGTAGCTTAAGGGGTTTCCCAGAAGACGTCGGACTTGACGACGACATAGCGAGTTTATAA
asR2	Pae	60	5′-GGGAGTCACTCTGTACCCCCTCTCCTTCAACGCTTGTACTAACTGGGCTGACTCCATCGT
	Pca	54	5′-......GACGCGGTATCCCTTCTCCTTTAGCGTGGCGACGAGCTGTGCCGTCTCCATAAT
asR3	Pae	65	5′-ACCCCCGAATTGGGGGCAAATGAGCGGCGGACACTTAAGGCGGCCCCGCCGCGAGCGGTTTCGCC
	Par	58	5′-.CCCCCGGA.CGGGGGCGAATGAGCGGCGGGCACCTGTGGCGGCTCCGCCGCGACTACT
	Pca	63	5′-.ACCCCGGA.TGGGGGCGGATGAGCGGCAGACACCTAAGGCGGCCCTGCCGCGACCAAGGGCTT
	Pis	59	5′-GACCCCCTGCTGGGGGCATATGAGCGGCGGGCACCTAAGGCGGCTCCGCCGCGACTGTA

Position of start codon (on coding strand) shown underlined for asR1 and asR2 (CAT). Position of stop codon (on coding strand) underlined (CTA, TTA) for asR3. Pae (P. aerophilum); Par (P. arsenaticum); Pca (P. calidifontis); Pis (P. islandicum); len (length of sRNA approximated from sequence read population).

the mRNA (**Figure 5A**) suggests that the sRNA pairs are stabilized within a dsRNA complex. The role of *asR1* with respect to *tfb1* transcripts is unclear, though the modulation of sRNA (both sense and antisense) while *tfb1* mRNA remains at constant and low abundance is reminiscent of negative feedback control.

The presence of complementary sense and antisense transcripts has been observed in a previous RNA sequencing study (Tang et al., 2005). Those authors suggested that the presence of an antisense transcript might enhance the stability of the mRNA target. As exemplified with *tfb1*, the presence of *cis*-antisense transcripts in our data are often accompanied by the presence of complementary sense strand fragments of similar size. This observation suggests that formation of a dsRNA duplex between the antisense sRNA and the 5′ region of the mRNA target may trigger destabilization of the mRNA; or alternatively, that base pairing between the antisense sRNA and the 5′ end of the nascent mRNA early in elongation may trigger premature transcription termination. For either mechanism, the result appears to be a constant level of *tfb1* mRNA under a variety of different culture conditions and growth phases.

THE FERRIC UPTAKE REGULATOR GENE (*fur*)

In a number of bacteria, the ferric uptake regulator FUR, is a transcriptional regulator of genes encoding proteins involved in iron homeostasis and protection from the toxic effects of iron under aerobic conditions. Some bacteria also encode a FUR-associated sRNA, for example *ryhB*; its synthesis is negatively regulated by FUR. The *ryhB* sRNA functions as a negative regulator of genes whose transcription is indirectly activated by FUR. The mechanism of *ryhB* sRNA negative regulation involves base pairing followed by selective degradation of the targeted mRNA (Andrews et al., 2003).

A homolog of the *fur* gene is conserved in the genomes of all known *Pyrobaculum* species. Embedded in each of the associated genes and located about 75 nt downstream from the 5′ start codon is an antisense-oriented, promoter-like sequence. In the two studied facultative aerobes (*P. aerophilum* and *P. calidifontis*), we detected a novel 54 nt-long *cis*-antisense transcript (**Table A1** in Appendix), designated as *asR2*, with precise transcription initiation consistent with the noted antisense promoter-like sequence. The 3′ end of the *asR2* transcript (**Table 2**; **Figure 4B**) transcript

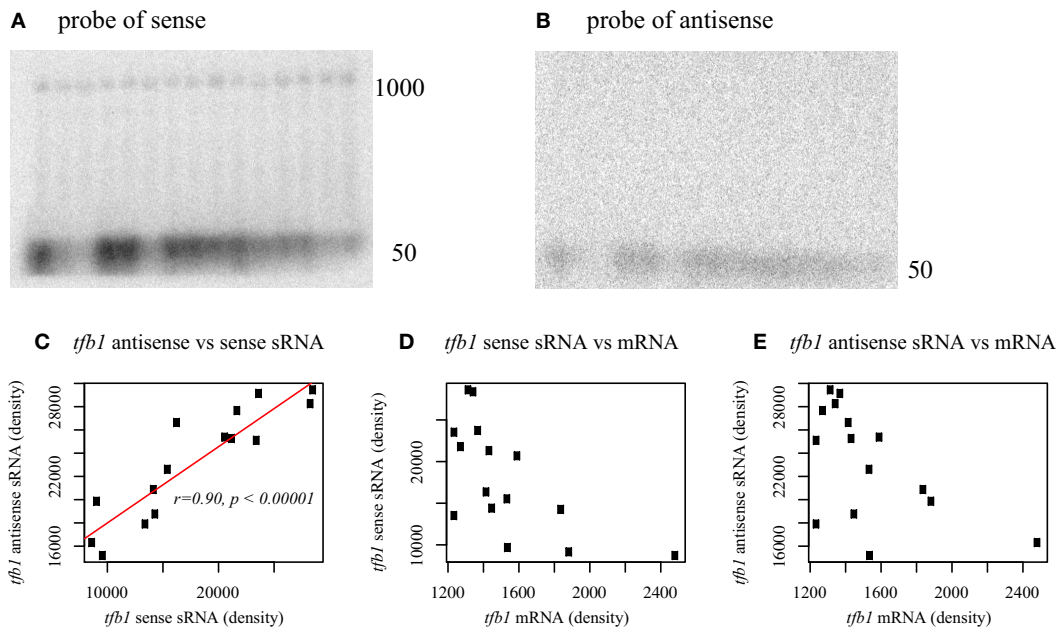

FIGURE 5 | Northern blot analysis of tfb1 from *Pyrobaculum aerophilum* using probes to the sense strand of *tfb* (A) and antisense strand (B). Sense transcripts of *tfb1* occur at both full length (∼1000 nt) and at 50 nt. Antisense transcripts occur at ∼50 nt. Lanes 1–15 (upper panels, left to right); total RNA across five respiratory growth conditions in three time series. Lanes 1–3 stationary phase, Lanes 4–6 growth with O_2, 7–9 growth with NO_3, 10–12 growth with As(V), 13–15 Fe(III). Each set of three lanes extracted from a time series for all five respiratory conditions at $T = (2.5, 4.5, 7.5\,h)$ with indicated terminal electron acceptor. Sense and antisense sRNA transcript abundance, inferred from band density, is positively correlated across growth conditions **(C)**, while no significant correlation is found between full length *tfb1* mRNA and either sRNA population **(D,E)**. Full length *tfb1* transcripts [1000 nt] remain nearly constant under all conditions tested **(A)**. Band density established using imagej (http://rsb.info.nih.gov/ij/).

terminates just upstream of the *fur* translation start codon. Both the *asR2* transcript and a complementary RNA fragment apparently derived from the 5′ end of *fur* mRNA, were present at high levels in anaerobically grown *P. aerophilum* and at modest levels in *P. calidifontis*. In the strict anaerobes (*P. islandicum*, *P. arsenaticum*), it appears that sequencing depth was insufficient to resolve any antisense-sense pairs under the limited set of growth conditions; however, we note that the predicted promoter for *asR2* in the facultative aerobes is equally well-conserved across all *Pyrobaculum* species.

THE TRIOSE-PHOSPHATE ISOMERASE (*tpi*) GENE

The *tpi* gene encodes triose-phosphate isomerase, an enzyme that is central to the modified Embden–Meyerhoff glycolytic pathway in *Pyrobaculum* species (Reher et al., 2007). We detected a 65-nt-long antisense transcript *asR3* (**Table A2** in Appendix) that overlaps the 3′ end of the *tpi* gene (**Figure 4C**) in all four of the species examined. Upon further examination of the 3′ terminal portion of *tpi*, we also detected a conserved sequence and associated secondary structure that is present in all sequenced *Pyrobaculum* spp. (**Figure 6**), which we term the tpi-element. In *P. aerophilum*, *P. islandicum*, and *P. calidifontis*, the tpi-element includes the stop codon of *tpi*, while the entire element is encoded immediately downstream of the *tpi* stop codon in the remaining *Pyrobaculum* spp.

A dsRNA formed by an interaction of *asR3* with the tpi-element could potentially compete against the mRNA intramolecular structure, and thus modulate function of the highly conserved tpi-element. Alternatively, *asR3* might itself be the active element of the pair, and in that case, presence of free *tpi* transcript might act as a repressor of *asR3*. In this model, *asR3* may have other *trans* targets in the genome and play a more general role in coordination of glycolysis in *Pyrobaculum* species.

DISCUSSION

Comparative transcriptomics has revealed compelling, conserved cases of novel *cis*-encoded transcripts that are antisense to core protein coding genes involved in transcription initiation and metabolism. We have considered these most obviously as potential regulators of their opposite strand partners, but they might also have broader regulatory roles.

We found that 28 of the top 34 cases of conserved 3′ antisense expression among orthologous *Pyrobaculum* proteins of known function coincide with convergent C/D box guide RNAs. This finding suggests that guide directed 2′-*O*-methylation of the mRNA in the region or downstream of the stop codon might be an unrecognized component of mRNA metabolism and gene regulation. It has been shown that pseudouridine modification of a stop codon can suppress termination of translation (Karijolich and Yu, 2011), but there are currently no studies of the possible implications of 2′-*O*-methyl modification on mRNA translation or stability. Alternatively, the presence of abundant mRNA fragments at the 3′ end may indicate that a sense-antisense interaction between the C/D box sRNA and mRNA terminus results in truncation of the mRNA

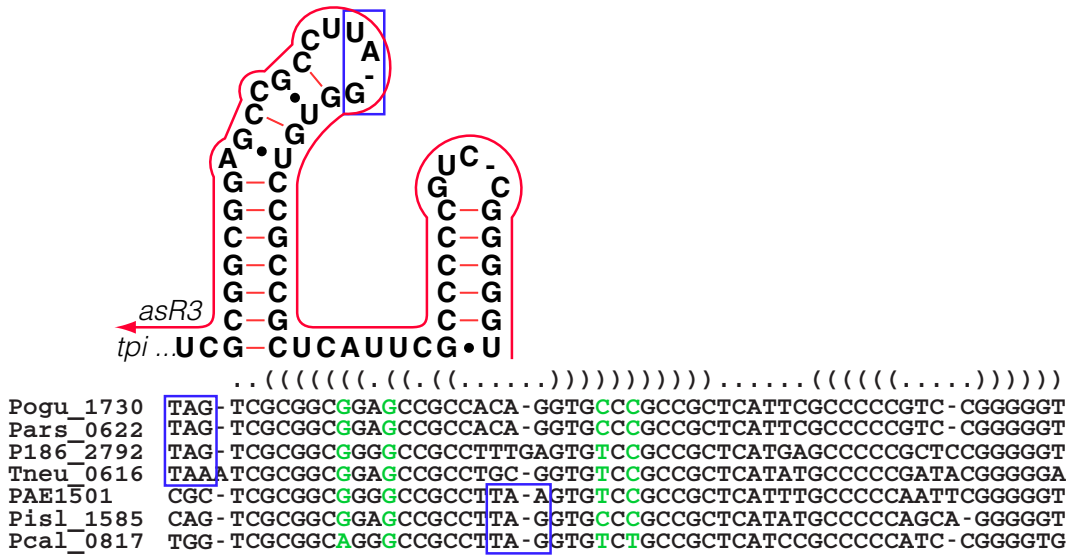

FIGURE 6 | The *Pyrobaculum* tpi-element and the associated antisense element, *asR3*. In all sequenced *Pyrobaculum* species, a highly conserved primary sequence forms a predicted secondary structure element (upper panel) at the 3′ end of the triose-phosphate isomerase gene (*tpi*). The depicted secondary structure contains the stop codon (blue box) in *P. aerophilum* (PAE1501), *P. islandicum* (Pisl_1585), and *P. calidifontis* (Pcal_0817). In *P. oguniense* (Pogu_1730), *P. arsenaticum* (Pars_0622), *Pyrobaculum* sp.1860 (P186_2792), and *P. neutrophilum* (Tneu_0616), the stop codon is immediately upstream of the tpi-element. *asR3* (red line) is encoded on the opposite strand, and has potential to compete/interfere with the tpi-element secondary structure. The genomic alignment of the 3′ portion of *tpi* and 3′ UTR is shown with the consensus secondary structure (nested parentheses). Base paired columns with one or more substitutions that maintain secondary structure are highlighted (green).

by an unknown mechanism, leading to mRNA destabilization and degradation.

The coordinated regulatory program implemented by Fur and its regulatory sRNA *ryhB* in some bacteria, provides a mechanism that yields both repression of some genes and activation (de-repression) of others. This program balances the needs of iron storage and utilization while protecting from iron-induced toxicity under oxic conditions. These dual roles are mediated by the inverse expression patterns of Fur and *ryhB*. Fur negatively regulates *ryhB*, which negatively regulates downstream genes. This circuit allows Fur to derepress (activate) those downstream genes. In published studies, active transcription in one direction can negatively regulate expression of the *cis*-encoded antisense partner (Lapidot and Pilpel, 2006), thus creating exclusive access to the shared genomic region. Likewise in *Pyrobaculum*, the *cis*-antisense transcription observed may yield the same type of inverse expression pattern. In this view, if the *cis*-antisense gene product is capable of repressing transcription or translation of targets in *trans*, then positive expression of *Pyrobaculum fur*, *tfb1*, or *tpi* may act through their corresponding antisense partner to activate (derepress) additional members of the associated regulon. Identification and verification of targets in *trans* is difficult in species that are not amenable to genetic manipulation such as *Pyrobaculum*, although future studies will explore computationally predicted targets.

The presence of *asR1*, a *cis*-encoded antisense RNA found within *tfb1* but not *tfb2* is of special interest when we consider these possible models of action for the *cis*-encoded antisense RNA. A specific TFB and TATA binding protein (TBP) pair in the archaeal halophile *Halobacterium* sp. NRC-1 has been shown to activate transcription under heat shock conditions (Coker and DasSarma, 2007). The observations that there are two instances of *tfb* in all *Pyrobaculum* genomes, and that only one harbors an antisense gene, suggest that *tfb1* might be essential only under particular conditions and/or initiate transcription for a subset of *Pyrobaculum* genes. Under this model, *tfb1* transcription might be held at low levels by the presence of *asR1* and possibly a dsRNA-binding complex. Under the alternative view, the *cis*-encoded *asR1* might facilitate activation of a *trans*-encoded regulon via de-repression. In the former view, the mechanism(s) that regulate sRNA transcription, stability, and mRNA interaction are central, while in the latter model, the sRNA is a downstream effector molecule of the independently regulated top-strand mRNA partner. In either case, resolving the molecular details of the sRNA's interaction with *tfb1* are needed to better understand this potential high-level mechanism for broad gene regulation in *Pyrobaculum*.

The tpi-element and its associated antisense partner, *asR3*, may provide a novel regulatory circuit acting from the 3′ UTR of *tpi*. The structure of the tpi-element (**Figures 4C** and **6**) contains the stop codon in some species while in other species the conserved structure is just downstream of the *tpi* stop codon. Possibilities for the function of the tpi-element include early transcription termination or translation termination. In either case, the tpi-element could be acting as a novel 3′ UTR riboswitch by binding a small molecule, or alternatively may be involved in a protein-binding event. Just as described above, the *cis*-antisense element *asR3*, encoded opposite the tpi-element, may act as a repressor of tpi-element function, or may have a role in *trans* with other genes in the *tpi* regulon.

In this study, we have described 74 or more expressed C/D box sRNA in each of four transcriptomes, most of which are conserved among multiple *Pyrobaculum* species. We have shown evidence that an unexpectedly large number of these sRNA overlap protein coding genes. Three novel sRNAs *asR1*, *asR2*, and *asR3* overlap genes involved in core transcription, iron regulation and core metabolism. Sequencing data have revealed the presence of sRNA originating from both strands, and these transcripts can be supported by promoter analysis, and verified by northern analyses. By contrast, less than 1% of transcripts mapped to CRISPR arrays show any evidence of dual strand transcripts (**Figure 1**). We suggest that the presence of dual-stranded transcript reads is an indication of an interaction of an sRNA with a convergently oriented mRNA, potentially mediated by one or more unknown dsRNA-binding complexes.

Future RNA-seq studies employing deeper sequencing technologies, alternative growth conditions, and other archaeal species will likely uncover many more cases of candidate regulatory antisense RNA. This work suggests multiple new research directions and will require complementary methodologies to better understand the complexity of sRNA function in Archaea. Given the conserved patterns of *cis*-antisense RNA transcripts now apparent, we anticipate rapid progress from follow-up studies that will demonstrate new modes of gene regulation homologous or analogous to those found in bacteria and eukaryotes.

ACKNOWLEDGMENTS

We are grateful to members of the Joint Genome Institute for making 454 sequencing possible (P. Richardson and J. Bristow for providing resources, and E. Lindquist and N. Zvenigorodsky for sample preparation and analysis). We thank Aaron Cozen for his generous procedural guidance and for the use of RNA blots used in the study. This work was supported by National Science Foundation Grant EF-082277055 (Todd M. Lowe and David L. Bernick); the Graduate Research and Education in Adaptive Bio-Technology (GREAT) Training Program sponsored by the University of California Bio-technology Research and Education Program (David L. Bernick); and by the National Science Foundation while Patrick P. Dennis was working at the Foundation. The opinions, findings, and conclusion expressed in this publications are ours and do not necessarily reflect the views of the National Science Foundation.

AUTHOR CONTRIBUTIONS

David L. Bernick designed and performed the experimental and computational analyses, and wrote the manuscript. Lauren M. Lui analyzed the C/D box sRNA sequencing data. Patrick P. Dennis provided assistance with the manuscript, collaborative review, and structure determination of C/D box sRNA. Todd M. Lowe provided scientific direction, contributed to interpretation of results, and edited the manuscript.

REFERENCES

Aiba, H. (2007). Mechanism of RNA silencing by Hfq-binding small RNAs. *Curr. Opin. Microbiol.* 10, 134–139.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Andrews, S. C., Robinson, A. K., and Rodriguez-Quinones, F. (2003). Bacterial iron homeostasis. *FEMS Microbiol. Rev.* 27, 215–237.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–1712.

Bernick, D. L., Dennis, P. P., Hochsmann, M., and Lowe, T. M. (2012). Discovery of *Pyrobaculum* small RNA families with atypical pseudouridine guide RNA features. *RNA* 18, 402–411.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421. doi:10.1186/1471-2105-10-421

Chan, P. P., Holmes, A. D., Smith, A. M., Tran, D., and Lowe, T. M. (2012). The UCSC Archaeal Genome Browser: 2012 update. *Nucleic Acids Res.* 40, D646–D652.

Coker, J. A., and DasSarma, S. (2007). Genetic and transcriptomic analysis of transcription factor genes in the model halophilic archaeon: coordinate action of TbpD and TfbA. *BMC Genet.* 8, 61. doi:10.1186/1471-2156-8-61

Cozen, A. E., Weirauch, M. T., Pollard, K. S., Bernick, D. L., Stuart, J. M., and Lowe, T. M. (2009). Transcriptional map of respiratory versatility in the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. *J. Bacteriol.* 191, 782–794.

Dennis, P. P., Omer, A., and Lowe, T. (2001). A guided tour: small RNA function in archaea. *Mol. Microbiol.* 40, 509–519.

Fitz-Gibbon, S. T., Ladner, H., Kim, U. J., Stetter, K. O., Simon, M. I., and Miller, J. H. (2002). Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. *Proc. Natl. Acad. Sci. U.S.A.* 99, 984–989.

Hale, C., Kleppe, K., Terns, R. M., and Terns, M. P. (2008). Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. *RNA* 14, 2572–2579.

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., Terns, R. M., and Terns, M. P. (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* 139, 945–956.

Henderson, I. R., Zhang, X., Lu, C., Johnson, L., Meyers, B. C., Green, P. J., and Jacobsen, S. E. (2006). Dissecting *Arabidopsis thaliana* DICER function in small RNA processing, gene silencing and DNA methylation patterning. *Nat. Genet.* 38, 721–725.

Jager, D., Sharma, C. M., Thomsen, J., Ehlers, C., Vogel, J., and Schmitz, R. A. (2009). Deep sequencing analysis of the Methanosarcina mazei Go1 transcriptome in response to nitrogen availability. *Proc. Natl. Acad. Sci. U.S.A.* 106, 21878–21882.

Karijolich, J., and Yu, Y. T. (2011). Converting nonsense codons into sense codons by targeted pseudouridylation. *Nature* 474, 395–398.

Kent, W. J. (2002). BLAT – the BLAST-like alignment tool. *Genome Res.* 12, 656–664.

Lapidot, M., and Pilpel, Y. (2006). Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. *EMBO Rep.* 7, 1216–1222.

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294, 858–862.

Lu, C., Kulkarni, K., Souret, F. F., Muthuvalliappan, R., Tej, S. S., Poethig, R. S., Henderson, I. R., Jacobsen, S. E., Wang, W., Green, P. J., and Meyers, B. C. (2006). MicroRNAs and other small RNAs enriched in the *Arabidopsis* RNA-dependent RNA polymerase-2 mutant. *Genome Res.* 16, 1276–1288.

Masse, E., Salvail, H., Desnoyers, G., and Arguin, M. (2007). Small RNAs controlling iron metabolism. *Curr. Opin. Microbiol.* 10, 140–145.

Omer, A. D., Lowe, T. M., Russell, A. G., Ebhardt, H., Eddy, S. R., and Dennis, P. P. (2000). Homologs of small nucleolar RNAs in Archaea. *Science* 288, 517–522.

Pak, J., and Fire, A. (2007). Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. *Science* 315, 241–244.

Reher, M., Gebhard, S., and Schonheit, P. (2007). Glyceraldehyde-3-phosphate ferredoxin oxidoreductase (GAPOR) and nonphosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPN), key enzymes of the respective modified Embden-Meyerhof pathways in the hyperthermophilic crenarchaeota *Pyrobaculum aerophilum* and Aeropyrum pernix. *FEMS Microbiol. Lett.* 273, 196–205.

Repoila, F., Majdalani, N., and Gottesman, S. (2003). Small non-coding RNAs, co-ordinators of adaptation processes in *Escherichia coli*: the RpoS paradigm. *Mol. Microbiol.* 48, 855–861.

Santangelo, T. J., Cubonova, L., James, C. L., and Reeve, J. N. (2007). TFB1 or TFB2 is sufficient for *Thermococcus kodakaraensis* viability and for basal transcription in vitro. *J. Mol. Biol.* 367, 344–357.

Straub, J., Brenneis, M., Jellen-Ritter, A., Heyer, R., Soppa, J., and Marchfelder, A. (2009). Small RNAs in haloarchaea: identification, differential expression and biological function. *RNA Biol.* 6, 281–292.

Tang, T. H., Bachellerie, J. P., Rozhdestvensky, T., Bortolin, M. L., Huber, H., Drungowski, M., Elge, T., Brosius, J., and Hutenhofer, A. (2002). Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7536–7541.

Tang, T. H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R., Bachellerie, J. P., and Huttenhofer, A. (2005). Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol. Microbiol.* 55, 469–481.

Vogel, J. (2009). A rough guide to the non-coding RNA world of *Salmonella*. *Mol. Microbiol.* 71, 1–11.

Wilderman, P. J., Sowa, N. A., Fitzgerald, D. J., Fitzgerald, P. C., Gottesman, S., Ochsner, U. A., and Vasil, M. L. (2004). Identification of tandem duplicate regulatory small RNAs in *Pseudomonas aeruginosa* involved in iron homeostasis. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9792–9797.

Wurtzel, O., Sapra, R., Chen, F., Zhu, Y., Simmons, B. A., and Sorek, R. (2010). A single-base resolution map of an archaeal transcriptome. *Genome Res.* 20, 133–141.

Zago, M. A., Dennis, P. P., and Omer, A. D. (2005). The expanding world of small RNAs in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Mol. Microbiol.* 55, 1812–1828.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 April 2012; accepted: 06 June 2012; published online: 02 July 2012.

Citation: Bernick DL, Dennis PP, Lui LM and Lowe TM (2012) Diversity of antisense and other non-coding RNAs in archaea revealed by comparative small RNA sequencing in four Pyrobaculum species. Front. Microbio. ***3****:231. doi: 10.3389/fmicb.2012.00231*

This article was submitted to Frontiers in Evolutionary and Genomic Microbiology, a specialty of Frontiers in Microbiology.

APPENDIX

Table A1 | Orthologous genes with 5′ sequencing reads. Orthologous groups are shown in each row where the locus tag number (e.g., 1645 for gene PAE1645) is followed by counts of (antisense, sense) reads. Groups are ranked by the total number of reads found within groupings formed by the number of species in a group with antisense sequencing reads. Read counts are accumulated by considering the largest region covered by at least one read in an overlapping region along a given strand, and assigning the read count to that region. Footnoted gene IDs have associated snoRNA-like sRNA (C/D box or H/ACA-like) – a, antisense oriented; s, sense oriented.

Product	*P. aerophilum*	*P. arsenaticum*	*P. calidifontis*	*P. islandicum*
Transcription initiation factor IIB	1645 (225,409)	1976 (12,16)	0584 (1,1)	1667 (8,39)
DNA-cytosine methyltransferase	1659 (4,0)	1839 (1,0)	0576	1675 (2,0)
Rhomboid family protein	1099 (3,0)	0267 (1,0)	0686	1249 (2,0)
Ferric uptake regulator, Fur family	2309 (40,11)	1526	1653 (1,0)	1023
30S ribosomal protein S12P	0670	2326 (2,0)	2096	0698 (7,1)
Cobalamin adenosyltransferase	1715	0782	0623[a] (86,4)	1701 (0,3)
Thiol:disulfide interchange protein	3152	1672	1794	0523 (32,0)
30S ribosomal protein S11P	3179 (15,2)	1654 (0,2)	1813	0540
NAD-dependent deacetylase	3500	1959[a] (12,3)	1963	0793
NADH dehydrogenase subunit A	3520 (9,0)	1954 (0,1)	1983	0847
30S ribosomal protein S3P	1779	0769	0553	1729 (9,0)
Translation initiation factor IF–1A	1072 (7,0)	0278	0681	1256
Putative transcriptional regulator, GntR family	2315 (0,10)	1532 (4,2)	1659 (0,2)	1028
Valyl-tRNA synthetase	2297 (4,0)	1497 (0,1)	1649	1019 (0,1)
Putative signal-transduction protein with CBS domains	2961 (4,0)	1332 (0,1)	1143	0364
Major facilitator superfamily MFS_1	1550[s] (3,5)	0660[s] (0,2)	0530[s] (0,2)	
Elongation factor 1, beta/beta′/delta chain	0695 (3,1)	2345	2114	0684
Egghead-like protein	0042 (3,2)	1076	2043 (0,1)	0056
V-type ATP synthase subunit B	1146	0237 (3,0)	0698	1264
Conserved protein (possible ATP binding)	0793 (3,11)	0044	2138	1084
Putative transcriptional regulator, ModE family	0813 (2,0)	0057	0023	1100
50S ribosomal protein L18e	0672	2328 (2,0)	2098	0696 (0,1)
ABC transporter related	1393 (2,0)	0445	1879	1525
Peptidase M50	1702	2238 (2,1)	0618	1696
Cation diffusion facilitator family transporter	0568 (2,0)	2239	1215	0125 (0,2)
paREP10	1480	0613 (2,0)	0811 (0,1)	1575
Exosome complex RNA-binding protein Rrp42	2206 (2,1)	1938	0932	0835
Inner-membrane translocator	3412 (2,0)	1174	1046	0977
NADH-ubiquinone oxidoreductase subunit		2274	2047	0329 (1,0)
CopG domain protein DNA-binding domain protein	2357 (1,0)	1561	1689	0622
Inner-membrane translocator	3348 (1,0)	1760	0444	0590
Amino acid-binding ACT domain protein	2296	1510	1648	1018 (1,1)
Hydrogen sulfite reductase	2596 (1,0)	1213	1457	
DNA-directed RNA polymerase subunit P	2258 (0,1)	1825	1624 (1,0)	0899
DNA polymerase, beta domain protein	1893	0821 (1,0)	1502	
Phosphate ABC transporter, inner membrane subunit PstC	1396	0443	1881	1527 (1,0)
Nicotinamide-nucleotide adenylyltransferase	1438	0405 (1,0)	0794	1561
Peptidase S8 and S53, subtilisin, kexin, sedolisin	1983	2056		1805 (1,0)
Glu/Leu/Phe/Val dehydrogenase,C terminal	3438	1871	1031	0980 (1,0)
D-isomer specific 2-hydroxyacid dehydrogenase, NAD-binding	3320 (1,1)	1736	1741	0566
Electron transfer flavoprotein, alpha subunit	0721	2372 (1,0)	2132	0645 (0,1)
Sua5/YciO/YrdC/YwlC family protein	2978 (1,0)	1345	1129	0378 (0,1)
AAA ATPase	3527	1626 (1,0)	1978	0145
MazG nucleotide pyrophosphohydrolase	1159 (1,0)	0222	0722	
30S ribosomal protein S8P	2098	2009 (0,2)	0176	1865 (1,0)
Transcriptional regulator, XRE family	0783	0037 (1,0)	2145	1076

(Continued)

Table A1 | Continued

Product	*P. aerophilum*	*P. arsenaticum*	*P. calidifontis*	*P. islandicum*
Acyl-CoA dehydrogenase domain protein	2070	2103	0199	1853 (1,0)
2-dehydropantoate 2-reductase	3409 (1,0)	2003	0383	1363
FHA domain containing protein	0816	0060 (1,0)	0026	1103
PaREP1 domain containing protein	3235		0464 (0,3)	1514 (1,0)
30S ribosomal protein S19e	3043 (1,1)	1790	0988	0440 (0,39)
CutA1 divalent ion tolerance protein	2325	1539	1667	1044 (1,0)
Nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase	2075 (1,1)	2019	0203	1857
Inner-membrane translocator	2083 (1,0)	1504	0826	0317 (0,2)
30S ribosomal protein S7P	0733 (1,1)	0001	0006	0655 (0,3)
Ribosomal protein L11	3104 (1,0)	1602	1832	0464
Metallophosphoesterase	3211	1639 (1,0)	0239	1924
Acetolactate synthase, large subunit, biosynthetic type	3300	1724 (1,1)	1753	0554 (0,2)
NAD+ synthetase	1219 (0,1)	0310 (1,0)	0793	1302
30S ribosomal protein S3Ae	3472 (1,0)	1852	1182	0771 (0,1)
Band7 protein	0750	0015	2166 (1,0)	1055
TGS domain protein	1649	1844	0581	1670 (1,0)
MoaD family protein	0727 (0,1)	2368	2136 (1,0)	0649
Putative circadian clock protein, KaiC	0729 (1,1)	2366 (0,2)	0010	0651
Tryptophanyl-tRNA synthetase	3091 (1,0)	1612	1822	0454
Aldehyde ferredoxin oxidoreductase	0622 (1,4)	2285	2057	0738
Inner-membrane translocator	3350	1761	0445 (0,1)	0591 (1,0)
Tyrosyl-tRNA synthetase	0630	2290 (1,0)	2062	0733
NADH-quinone oxidoreductase, B subunit	2928	1001	1957	0336 (1,0)
Prephenate dehydratase	0893[s] (0,51)	0111	0075	1150 (1,0)

Table A2 | Orthologous genes with 3′ sequencing reads. Orthologous groups, read counts, and footnotes displayed are as described in **Table A1**.

Product	*P. aerophilum*	*P. arsenaticum*	*P. calidifontis*	*P. islandicum*
Electron transfer flavoprotein, alpha subunit	0721[a] (381,54)	2372[a] (258,14)	2132[a] (2145,21)	0645[a] (22,6)
DNA-directed RNA polymerase, M/15 kDa subunit	3480[a] (153,0)	1847 (1,0)	1177[a] (94,0)	0776[a] (31,0)
NAD-dependent deacetylase	3500[a] (83,2)	1959[a] (6,0)	1963 (1,0)	0793[a] (26,1)
SMC domain protein	2280[a] (2,7)	1811[a] (9,0)	1637[a] (7,1)	0884[a] (10,0)
Triosephosphate isomerase	1501 (2,1)	0622 (1,0)	0817 (13,0)	1585 (1,0)
Metallophosphoesterase	2243[a] (287,0)	1913	0956[a] (26,0)	0802[a] (63,0)
Resolvase, N-terminal domain	3513[a] (250,0)	1963[a] (66,0)	1967	0797[a] (27,0)
Succinate dehydrogenase subunit D	0719[a] (94,10)	2361[a] (129,0)	2130	0668[a] (46,0)
HhH-GPD family protein	0880[a] (23,2)	0101[a] (10,3)	0066	1140[a] (235,0)
Elongation factor EF-2	0332[a] (183,0)	2139	0213[a] (15,0)	1957[a] (3,0)
Twin-arginine translocation protein, TatA/E family subunit	1546[a] (32,16)	0666	0534[a] (53,1)	1615[a] (7,2)
Aldo/keto reductase	2929[a] (66,14)	1002[a] (1233,1)	0966	
Putative agmatinase	2260	1823	1626[a] (13,2)	0897[a] (611,2)
MazG nucleotide pyrophosphohydrolase	1159 (133,1)	0222 (352,2)	0722	
Ferric uptake regulator, Fur family	2309[a] (128,0)	1526[a] (141,1)	1653	1023
Seryl-tRNA synthetase	3158[a] (50,0)	1667	1802	0528[a] (39,36)
Uridylate kinase	3159	1665	1804 (20,0)	0530 (39,0)
Purine and other phosphorylases, family 1	1476	0610	0814[a] (24,0)	1572 (23,0)
Isoleucyl-trna synthetase	1617	1993[a] (5,0)	0601[a] (2,1)	1650
Transcriptional regulator, Fis family	3027[s] (4,10)	1779[a] (3,0)	0999[s] (0,5)	0429[s] (0,47)
GCN5-related N-acetyltransferase	3246	1807	1556[a] (4,0)	0488 (2,0)
Putative circadian clock protein, kaic	0729	2366[a] (5,0)	0010	0651 (1,0)
Conserved protein (RNA polymerase related?)	1975	2051 (1,0)	1587	1800 (1,0)
Lysine exporter protein (LYSE/YGGA)	2077	2018	0708[a] (5260,0)	1858
Translation initiation factor IF-2 subunit gamma	0064	1171	0242[a] (162,0)	0078
Alpha-l-glutamate ligases, rimk family	1818	0723	0506[a] (116,0)	1747
Oxidoreductase, molybdopterin binding	0389	0833	1263	1366[a] (84,0)
Ribosomal protein L25/L23	1972	2048	1585[a] (66,8)	1798
Proliferating-cell nuclear antigen-like protein	0720	2362	2131[a] (36,0)	0667
3-dehydroquinate synthase	1685	1827	0566[a] (25,0)	1689
DNA polymerase, beta domain protein region	1153	1067	0856 (23,0)	
Haloacid dehalogenase domain protein hydrolase	1785	0739	0554[a] (20,0)	1734
Mn2+-dependent serine/threonine protein kinase	2192	1948	0924	0825 (12,0)
Radical SAM domain protein	2153 (0,1)	0818	1068	0189[a] (9,0)
DNA polymerase I	2180	0798	1087	0816 (6,0)
Ribonuclease HII	1216	0312	0780	1305 (5,0)
Bifunctional GMP synthase/glutamine amidotransferase protein	3369	1772	1723	0600 (4,0)
Band 7 protein	0750	0015	2166	1055 (4,0)
Alpha-l-glutamate ligases, RimK family	0645 (4,0)	2302	2074	0721
Ribonucleoside-diphosphate reductase, adenosylcobalamin-dependent	3155 (4,0)	1670	1797	0525
Thermosome	3273 (0,3)	1704 (0,1)	1771	0501 (3,5)
Metallophosphoesterase	1087 (3,0)	0270	1512	1254
Peptidase M24	2025	2086	1010	1836 (3,0)
Pyruvate/ketoisovalerate oxidoreductase, gamma subunit	3279	1708	1767	0497 (3,0)
DNA polymerase, beta domain protein region	0045	1137 (3,0)	0385	
Creatininase	2983	1350	1124	0383 (2,0)
Acetyl-CoA acetyltransferase	1220	0309	0781	1301 (2,2)
Indole-3-glycerol-phosphate synthase	0570 (2,0)	2240	1213	0124 (0,2)
Regulatory protein, ArsR	0731	2364	0008 (2,0)	0653
PaREP1 domain containing protein	0002	1095	1373[a] (2,0)	
Putative signal-transduction protein with CBS domains	3588		1394	0254 (2,0)

(Continued)

Table A2 | Continued

Product	*P. aerophilum*	*P. arsenaticum*	*P. calidifontis*	*P. islandicum*
Exosome complex exonuclease Rrp41	2207 (2,1)	1937	0933	0836
ABC transporter related	3413 (2,0)	1175	1045	0976
Uroporphyrinogen III synthase HEM4	0589	2250	1712	0116 (2,0)
Potassium transport membrane protein, conjectural	2422 (2,0)	1446	0314	0883
Undecaprenyl diphosphate synthase	2942 (2,0)	1319	1157	0348
Nucleotidyl transferase	0837	0080	0043	1119 (2,0)
Carbon starvation protein CstA	1423 (2,0)	0894	0860	
Carboxypeptidase Taq	0885 (1,2)	0104	0069 (0,1)	1143
Leucyl-tRNA synthetase	1107	0260	0691	1246 (1,0)
HEPN domain protein	1894	0820 (1,0)	1501	
DNA-directed RNA polymerase subunit E, RpoE2	3563 (1,0)	2230	1991	0921
5-carboxymethyl-2-hydroxymuconate Δ-isomerase	2688	0535	1503 (1,0)	
ATPase	1789	0736		1446 (1,0)
Oligosaccharyl transferase, STT3	3030 (1,0)	1781	0997	0431
paREP7	0906		0492	0185 (1,0)
Haloacid dehalogenase domain protein hydrolase	2017^{s} (0,12)	2080^{s} (0,15)	1016^{s} (1,4)	1830
Egghead-like protein	0042	1076	2043	0056 (1,0)
Putative transcriptional regulator, CopG family	1443 (0,2)	0399	0796	1563 (1,1)
Asparaginyl-tRNA synthetase	2973 (1,0)	1342	1133	0375
Succinate dehydrogenase iron-sulfur subunit	0717	2359	2128	0670 (1,0)
Peptidase T2, asparaginase 2	3083 (1,0)	1892	0970	0908
Radical SAM domain protein	0596	2255	1716	0113 (1,0)
30S ribosomal protein S25e	2188 (0,1)	0790 (0,1)	1079	0808 (1,0)
Ribosomal-protein-alanine acetyltransferase	2246 (1,0)		0958	1001
PilT protein domain protein	3561 (1,0)	1614	1989	0923
Peptidase S8 and S53, subtilisin, kexin, sedolisin	0712	2355	2124 (1,0)	0674
Nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase	2075^{s} (0,49)	2019	0203	1857^{s} (1,8)
Beta-lactamase domain protein	2160	0810	1074	0803 (1,0)
Xanthine dehydrogenase accessory factor	2669	0253	1324 (1,0)	
ABC transporter related	3269	1702 (1,0)	1774	0503
tRNA CCA-pyrophosphorylase	3325	1740	1737	0570 (1,0)
Starch synthase	3429	1878 (1,0)	1038	0968
Dual specificity protein phosphatase	1536	0675 (1,0)	0541	1603
Putative endoribonuclease L-PSP	3003^{s} (1,62)	1258^{s} (0,210)	1096^{s} (0,81)	0414^{s} (0,65)
Sulfite reductase, dissimilatory-type beta subunit	2597	1212 (1,0)	1456	
Methyltransferase small	0261	2199	0236	0747 (1,0)
Putative transcriptional regulator, AsnC family	1507 (1,0)	0627 (0,1)	0822	1590
Methyltransferase type 11	1165	0216 (1,0)	1364	1338
Serine/threonine protein kinase	0815	0059 (1,0)	0025	1102
Transcriptional regulator, PadR-like family	0013	1087 (1,0)		0038
Inner-membrane translocator	3350 (1,0)	1761	0445	0591
Geranylgeranyl reductase	2989	1355	1119	0388 (1,0)
Extracellular solute-binding protein, family 5	2391	1494	0422	0602 (1,0)
2-methylcitrate synthase/citrate synthase II	1689 (1,0)	2234	0563	1692
30S ribosomal protein S6e	1505 (0,1)	0626 (1,0)	0821	1589

Table A3 | Hypothetical genes with 5′ sequencing reads. Orthologous groups, read counts, and footnotes displayed are as described in **Table A1**.

Product	*P. aerophilum*	*P. arsenaticum*	*P. calidifontis*	*P. islandicum*
Hypothetical protein	3282 (3,3)	1710	1765	0495 (11,6)
Hypothetical protein	0301	2159[a] (673,0)	0225	2002
Hypothetical protein	3499	1958	1962	0792[a] (26,1)
Hypothetical protein	0432 (6,0)		0474	0140
Hypothetical protein	1798	0175 (4,0)	0509	1742
Protein of unknown function DUF107	0749	0014	2167	1056 (4,0)
Hypothetical protein	1503 (3,0)	0624	0819	1587
Hypothetical protein	2934	1279 (3,0)	1165	0340 (0,1)
Hypothetical protein	1517	0632 (3,0)	0877	
Hypothetical protein	3546	1625 (3,0)	1977	0933
Hypothetical protein	0433 (3,6)		0479	0139
Hypothetical protein	3051[s] (0,284)	1797	0981[s] (0,152)	0447[s] (3,129)
Hypothetical protein	0838 (2,2)	0081	0044	1120
Hypothetical protein	1710 (0,3)	0785 (2,3)	0620	1698
Hypothetical protein	0728	2367 (2,0)	2137	0650
Hypothetical protein	1147	0229	0706	1284 (2,0)
Hypothetical protein	1522	0636	0874	1594 (2,0)
Hypothetical protein	2941 (2,0)	1318	1158	0347
Hypothetical protein	2338	1549	1677	0634 (2,0)
Hypothetical protein	2822	0279	1187	1388 (2,1)
Hypothetical protein	1943	2025 (2,0)	1297	
Hypothetical protein	2416 (2,0)	1479	0319	0879 (0,1)
Hypothetical protein	0746 (2,0)	0012	2168	1006
Hypothetical protein	1069 (0,13)	0281	0680	1257 (2,0)
Hypothetical protein	3081 (2,0)	1891	0969	0907 (0,1)
Hypothetical protein	0800	0050	0016 (0,1)	1092 (2,4)
Protein of unknown function DUF77	1158	0223 (2,0)	0711	1327
Hypothetical protein	3135 (0,2)	1683	1783	0512 (2,0)
Hypothetical protein	1683	1828 (1,0)	0567 (0,1)	1688
Hypothetical protein	0789	0040 (1,0)	2142	1079
Protein of unknown function DUF72	2078 (1,0)	2017	0205	1859
Hypothetical protein	1641 (0,3)	1979 (1,1)	0587	1664 (0,1)
Protein of unknown function DUF54	2213	1931	0938	0842 (1,0)
Protein of unknown function DUF437	0638 (1,0)	2296	2068	0727
Hypothetical protein	3550	1622	1974	0930 (1,0)
Hypothetical protein	3467	1857	1188	0993 (1,0)
Hypothetical protein	1318	0471		1365 (1,1)
Hypothetical protein	3556 (1,0)	1619	1969	0927
Hypothetical protein	2598	1211 (1,0)	1455	
Hypothetical protein	1643 (1,0)	1977	0585	1666
Hypothetical protein	2824	0884 (1,0)	0867	1387
Hypothetical protein	2322	1537	1664	1037 (1,0)
Hypothetical protein	2177	0799	1088	0817 (1,0)
Hypothetical protein	1449 (0,1)	0601 (0,1)	0800 (1,0)	1567 (0,1)
Protein of unknown function DUF52	0818	0062 (1,0)	0028	1105
Hypothetical protein	1173	0212	0743	1320 (1,0)
Hypothetical protein	3004[s] (1,62)	1259	1095	0415[s] (0,65)
Protein of unknown function DUF72	3079 (0,1)	1889 (1,3)	0967	0905
Hypothetical protein	2268 (1,0)	1820	1629	0894
Hypothetical protein	3324	1739	1738 (0,1)	0569 (1,0)

(Continued)

Table A3 | Continued

Product	*P. aerophilum*	*P. arsenaticum*	*P. calidifontis*	*P. islandicum*
Protein of unknown function UPF0027	0998	0172	0141 (1,0)	1219
Hypothetical protein	2210	1934 (1,0)	0941	0839
Hypothetical protein	1797	0174 (1,1)	0508	1741
Hypothetical protein	0718	2360	2129	0669 (1,0)
Hypothetical protein	1448	0600	0799	1566 (1,0)
Hypothetical protein	1613	2156	0603	1648 (1,10)
Hypothetical protein	1018	0184	0656 (0,1)	1203 (1,0)
Hypothetical protein	2429	1441	0309	1892 (1,0)
Hypothetical protein	3148	1674	1792	0521 (1,0)
Hypothetical protein	1676	1833		1682 (1,0)
Hypothetical protein	2403 (1,0)	1470	0327	0871 (0,3)

Table A4 | Hypothetical genes with 3′ sequencing reads. Orthologous groups, read counts, and footnotes displayed are as described in **Table A1**.

Product	*P. aerophilum*	*P. arsenaticum*	*P. calidifontis*	*P. islandicum*
Protein of unknown function DUF6, transmembrane	1545[s] (16,32)	0667[s] (2,20)	0535[s] (1,53)	1614[s] (2,7)
Hypothetical protein	1519[a] (550,1)	0634[a] (197,0)	0875[a] (202,0)	1596
Hypothetical protein	0577[a] (6,0)	2243[a] (18,7)	1195[a] (2,0)	0121
Hypothetical protein	1836 (0,2)	0710 (1,0)	0502 (5,0)	1752 (2,0)
Hypothetical protein	3249 (4,0)	1805	1855	0485 (2148,0)
Hypothetical protein	1234 (349,0)	0295 (1,0)	1511	0244
Protein of unknown function DUF1614	2020	2082 (15,0)	1014	1832 (58,0)
Hypothetical protein	1687[a] (48,4)	2232[a] (17,3)	0565	1690
Hypothetical protein	3138[a] (15,23)	1680[a] (27,2)	1786	0515
Hypothetical protein	0889	0108	0073 (4,0)	1147[a] (28,3)
Protein of unknown function DUF192	2955[a] (15,1)	1329	1147[a] (10,4)	0358
Hypothetical protein	3550	1622	1974 (1,0)	0930[s] (2,44)
Hypothetical protein	3005[a] (177,0)	1260	1094	0416
Hypothetical protein	3630	1030	2015	0003 (74,0)
Hypothetical protein	3245[a] (63,1)	1808		0487
Hypothetical protein	2069	2102	0198[a] (44,0)	1852
Hypothetical protein	3468	1856	1186	0994 (31,0)
Hypothetical protein	0730	2365 (18,0)	0009	0652
Hypothetical protein	3497	1956	1958 (12,6)	0790
Hypothetical protein	0936	0136 (8,0)	0105	1172
Hypothetical protein	3135 (0,1)	1683	1783[s] (8,6)	0512[s] (0,116)
Hypothetical protein	3156	1669	1798	0526 (6,0)
Hypothetical protein	0748	0013[s] (6,6)		1058[s] (0,1982)
Protein of unknown function DUF62	3627	2209[s] (0,5)	2013	0002 (6,0)
Hypothetical protein	1177	0209	0746[a] (5,16)	1317
Hypothetical protein	3189	1645	1819	0545 (4,0)
Hypothetical protein	3295 (4,0)	1719	1758	0550
Hypothetical protein	2549	0845 (3,0)	1351	0257
Hypothetical protein	1173 (3,0)	0212	0743	1320
Protein of unknown function UPF0027	0998	0172 (3,0)	0141	1219
Hypothetical protein	2504 (3,0)	1412		1927
Hypothetical protein	2326	1541	1669	1042 (3,0)
Hypothetical protein	1549[s] (3,5)	0661[s] (0,2)	0531[s] (0,5)	1618[s] (0,4)
Hypothetical protein		0611[s] (2,15)	0813[s] (0,24)	1573 (0,23)
Protein of unknown function DUF64	0371		1533	1367 (2,0)
Hypothetical protein	2285 (2,0)	1583	1640	1010
Hypothetical protein	1307	0473		1351[s] (2,14)
Hypothetical protein	1816	0724 (2,0)		0422 (0,1)
Hypothetical protein	3251 (2,0)	1803	1853	0483
Hypothetical protein	1497	0619	0805	1582 (2,0)
Hypothetical protein	1895	0678 (2,0)	0646	
Protein of unknown function DUF224, cysteine-rich region domain protein	1762	0754	0545	1721 (2,0)
Hypothetical protein	3568	2225	1996 (2,0)	0916
Hypothetical protein	1998	2066	1030	1817 (2,0)
Protein of unknown function DUF115	2328	1542	1670	1041 (2,0)
Hypothetical protein	2337	1548	1676	0635 (2,0)
Protein of unknown function DUF100	0944 (1,0)	0141	0111	1181
Protein of unknown function DUF340, membrane	1479	0612	0812	1574 (1,0)
Hypothetical protein	3304 (1,0)	1727	1750	0557
Hypothetical protein	0882 (1,0)	0102	0067	1141
Hypothetical protein	1130	0243 (1,0)	0697	1241

(Continued)

Table A4 | Continued

Product	*P. aerophilum*	*P. arsenaticum*	*P. calidifontis*	*P. islandicum*
Hypothetical protein	1449	0601	0800	1567 (1,0)
Hypothetical protein	2190	1946	0927	0827 (1,0)
Hypothetical protein	0927 (1,0)	0131	0083	1161
Hypothetical protein	0708	2353	2122	0676 (1,0)
Hypothetical protein	2606	1232 (1,0)	1431	
Hypothetical protein	2187	0791 (1,0)	1080	0809 (0,1)
Hypothetical protein	0239	1512 (1,0)	0690	
Protein of unknown function DUF1028	3380	1006 (1,0)	0160	
Hypothetical protein	2154 (1,0)	0817	1069	0942
Hypothetical protein	3161	1666	1803	0529 (1,0)
Hypothetical protein	2058	2311	0401	1850 (1,0)
Hypothetical protein	0840	0083 (1,0)	0046	1122