Complete genome sequence of *Archaeoglobus profundus* type strain (AV18^T)

Mathias von Jan[1], Alla Lapidus[2], Tijana Glavina Del Rio[2], Alex Copeland[2], Hope Tice[2], Jan-Fang Cheng[2], Susan Lucas[2], Feng Chen[2], Matt Nolan[2], Lynne Goodwin[2,3], Cliff Han[2,3], Sam Pitluck[2], Konstantinos Liolios[2], Natalia Ivanova[2], Konstantinos Mavromatis[2], Galina Ovchinnikova[2], Olga Chertkov[2], Amrita Pati[2], Amy Chen[4], Krishna Palaniappan[4], Miriam Land[2,5], Loren Hauser[2,5], Yun-Juan Chang[2,5], Cynthia D. Jeffries[2,5], Elizabeth Saunders[2], Thomas Brettin[2,3], John C. Detter[2,3], Patrick Chain[2,4], Konrad Eichinger[6], Harald Huber[6], Stefan Spring[1], Manfred Rohde[7], Markus Göker[1], Reinhard Wirth[6], Tanja Woyke[2], Jim Bristow[2], Jonathan A. Eisen[2,8], Victor Markowitz[4], Philip Hugenholtz[2], Nikos C Kyrpides[2], and Hans-Peter Klenk[1*]

[1] DSMZ - German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany
[2] DOE Joint Genome Institute, Walnut Creek, California, USA
[3] Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico, USA
[4] Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[5] Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[6] University of Regensburg, Microbiology – Archaeenzentrum, Regensburg, Germany
[7] HZI – Helmholtz Centre for Infection Research, Braunschweig, Germany
[8] University of California Davis Genome Center, Davis, California, USA
*Corresponding author: Hans-Peter Klenk

Archaeoglobus profundus (Burggraf *et al*. 1990) is a hyperthermophilic archaeon in the euryarchaeal class *Archaeoglobi*, which is currently represented by the single family *Archaeoglobaceae*, containing six validly named species and two strains ascribed to the genus 'Geoglobus' which is taxonomically challenged as the corresponding type species has no validly published name. All members were isolated from marine hydrothermal habitats and are obligate anaerobes. Here we describe the features of the organism, together with the complete genome sequence and annotation. This is the second completed genome sequence of a member of the class *Archaeoglobi*. The 1,563,423 bp genome with its 1,858 protein-coding and 52 RNA genes is a part of the *G*enomic *E*ncyclopedia of *B*acteria and *A*rchaea project.

Introduction

Strain AV18^T (= DSM 5631 = JCM 9629 = NBRC 100127) is the type strain of the species *Archaeoglobus profundus* [1,2]. It is the second of five species currently ascribed to the genus *Archaeoglobus*, of which the type species is *A. fulgidus*, described in 1988 [3]. Strains for all *Archaeoglobus* species were isolated from marine hydrothermal systems, yet *A. fulgidus* originates from a shallow marine hydrothermal system at Volcano, Italy [3] whereas *A. profundus* was isolated from a deep sea hot vent area (depth: 2000 m) at Guaymas, Mexico [1]. The genome sequence of the type strain from a third species of the *Archaeoglobaceae* – *Ferroglobus placidus* [4] – has been completed very recently (Feb 2010) at the Joint Genome Institute (CP001899). Here we present a summary classification and a set of features for *A. profundus* strain AV18^T, together with the description of the complete genomic sequencing and annotation.

Classification and features

Six species with validly published names and two strains ascribed to the not invalidly published genus 'Geoglobus' [5,6] are currently assigned to the *Archaeoglobi*, all of which were isolated from marine hydrothermal systems ranging from shallow water to deep sea habitats of 4,100 m depth. Five species thereof are accounted to the genus *Archaeoglobus*: *A. profundus*, *A. fulgidus*, *A. veneficus* [7], *A. infectus* [8] and *A. solfaticallidus* [9]. Publications about the taxonomy of the *Archaeoglobi* often mention another species of this genus ("*A. lithotrophicus*") isolated from deep oil reservoirs [10], but no formal species description has been published, therefore this ninth species is excluded from comparisons shown in this work.

Based on 16S rRNA gene sequences, the closest related type strain is *F. placidus* [4] with 96.5% sequence identity, while the other type strains of the genus *Archaeoglobus* share 91.9-95.0% sequence identity [11], with the non validly published '*Geoglobus*' strains inbetween (94.4%). The nearest related genera are *Pyrococci* and *Thermococci* with about 86% sequence identity. Searching the NCBI non-redundant nucleotide database with the 16S rRNA sequence of *A. profundus*, 73 sequences of at least 90% sequence identity were found. Fifty of these sequences belong to uncultured archaeal phylotypes from environmental samples, all others were identified as belonging to the *Archaeoglobaceae*. These samples originated from marine hydrothermal systems at the Mid-Atlantic Ridge [12,13] and AJ969472, the East Pacific Rise [14,15], Izu-Bonin Arc [16], and Southern Mariana Trough (AB293221, AB293225, AB293242, AB293237) in the Western Pacific Ocean, Iheya Basin (Okinawa Trough) in the East China Sea [17,18], the Gulf of California [19,20], a seafloor borehole at Juan de Fuca Ridge in the Pacific Ocean [21], from high temperature oil reservoirs [22], and from terrestrial hot springs in Europe [23], North America [24-26], East Asia (FJ638514, FJ638518-23 FJ638504, FJ638508) and Southeast Asia [27]. These numerous findings (as of January 2010) corroborate and extend the early assumption [1] that members of the *Archaeoglobaceae* may be widely distributed across hydrothermal habitats.

Figure 1 shows the phylogenetic neighborhood of *A. profundus* AV18[T] in a 16S rRNA based maximum likelihood [35] phylogenetic tree, which is in agreement with earlier inferences of the phylogeny of this taxon [5,6,8,9,31]. Remarkably, *A. profundus* clusters together with *F. placidus*, apart from the cluster containing the other three species of the genus *Archaeoglobus*, indicating polyphyly of the genus and therefore possibly the need for taxonomic emendation, as discussed previously [9].The sequence of the single 16S rRNA gene copy in the genome of *A. profundus* AV18[T] is identical with the previously published 16S rRNA gene sequence derived from DSM 5631 (AJ299219), which contained five ambiguous base calls.

Cells of *A. profundus* AV18[T] are reported as Gram stain-negative, highly irregular cocci, occurring singly or in pairs (Figure 2 and Table 1) [1]. They have dimensions of approximately 0.7-1.3 μm x 1.4-1.9 μm. The organism shows a blue-green fluorescence at 420 nm UV light, indicating the presence of coenzyme F_{420}, and contains a cell envelope composed of subunits covering the membrane, which is visible in thin sections [1]. Motility and flagella were not observed [1,43] in contrast to all other members of this genus, with the exception of *A. sulfaticallidus*, which was described very recently [9].

Growth of strain AV18[T] occurs between 65 and 90°C with an optimum at 82°C, at a pH ranging from 4.5 to 7.5 and a concentration of NaCl between 0.9 and 3.6% [1]. *A. profundus* is mixotrophic under strictly anaerobic conditions [1] with hydrogen as an essential energy source and sulfate, thiosulfate and sulfite as electron acceptors, producing H_2S [1].

All members of the genus *Archaeoglobus* can utilize hydrogen as electron donor, in addition, *A. fulgidus*, *A. veneficus* and *A. solfaticallidus* can use at least a subset of the organic compounds pyruvate, formate, acetate or lactate [9,43]. Electron acceptors are those of *A. profundus* (see above) except for *A. veneficus* and *A. infectus*) which are incapable of utilizing sulfate [9,43]. Carbon sources can be CO_2 (except for *A. profundus* and *A. infectus*) or organic compounds [8,9,43]. Due to differences mainly in metabolism, a new genus was introduced for *F. placidus* [4]: Unlike previously described *Archaeoglobales*, *F. placidus* is capable of growing by nitrogen reduction, and oxidation of ferrous iron or sulfide, but unable to reduce sulfate [4]. Besides, it is the only reported case of an archaeon which can anaerobically oxidize aromatic compounds, by reduction of

Fe(III) [37]. Other published species of this class are "*Geoglobus ahangari*" [5] and the recently reported "*G. acetivorans*" [6]. The genus "*Geoglobus*" again separates from the other *Archaeoglobaceae* by characteristic metabolic features: in cultivation experiments, the sole electron acceptor used by these species is Fe(III) and they are reported to be the first hyperthermophilic organisms exhibiting growth upon anaerobic oxidation of long chain fatty acids [5,6].
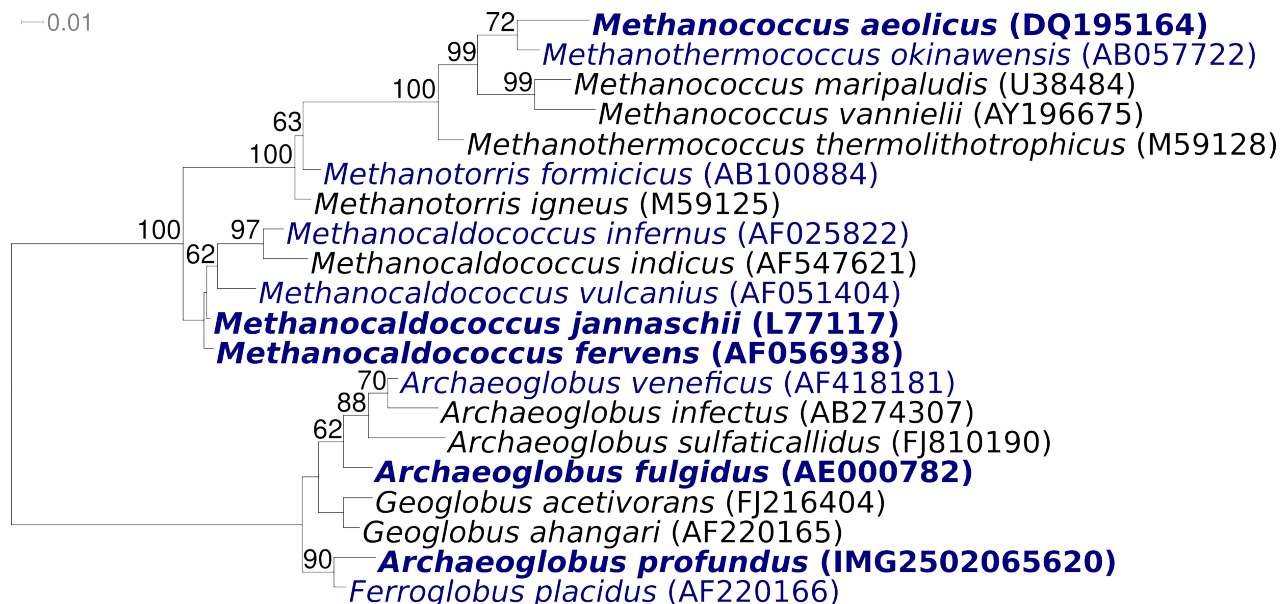

Figure 1. Phylogenetic tree highlighting the position of *A. profundus* AV18[T] relative to the other type strains within the family. The tree was inferred from 1,334 aligned characters [28,29] of the 16S rRNA gene sequence under the maximum likelihood criterion [30] and rooted in accordance with a current taxonomy [31]. The branches are scaled in terms of the expected number of substitutions per site. Numbers above branches are support values from 1,000 bootstrap replicates if larger than 60%. Lineages with type strain genome sequencing projects registered in GOLD [32] are shown in blue, published genomes in bold: *Methanococcus aeolicus* (CP000743), *Methanocaldococcus fervens* (CP001696), *Methanocaldococcus jannaschii* [33] and *A. fulgidus* [34], two of the very first organisms whose genome sequences have been revealed.
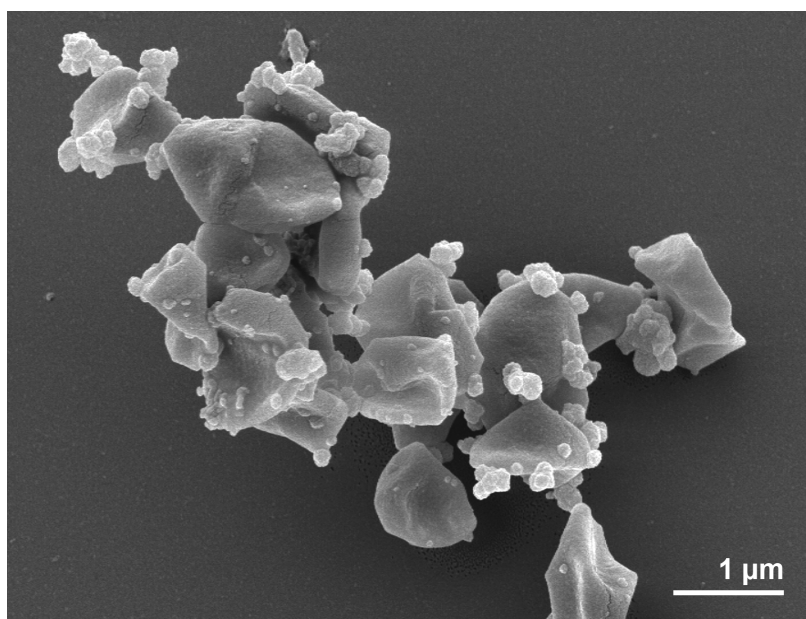

Figure 2. Scanning electron micrograph of cells of *A. profundus* strain AV18[T]

Table 1. Classification and general features of *A. profundus* strain AV18, according to the MIGS recommendations [36]

MIGS ID	Property	Term	Evidence code
		Domain *Archaea*	TAS [38]
		Phylum *Euryarchaeota*	TAS [39]
		Class *Archaeoglobi*	TAS [40]
		Order *Archaeoglobales*	TAS [41]
	Classification	Family *Archaeoglobaceae*	TAS [99]
		Genus *Archaeoglobus*	TAS [3]
		Species *Archaeoglobus profundus*	TAS [1]
		Type strain AV18	TAS [1]
	Gram stain	negative	TAS [1]
	Cell shape	coccoid, highly irregular	TAS [1]
	Motility	not motile	TAS [1]
	Sporulation	nonsporulating	NAS
	Temperature range	65-90°C	TAS [1]
	Optimum temperature	82°C	TAS [1]
	Salinity	>9-36 g/l (optimum 18 g/l)	TAS [1]
MIGS-22	Oxygen requirement	obligate anaerobic	TAS [1]
	Carbon source	acetate, pyruvate, lactate, yeast extract, meat extract, peptone, acetate containing crude oil	TAS [1]
	Energy source	H_2	TAS [1]
MIGS-6	Habitat	deep sea hydrothermal system	TAS [1]
MIGS-15	Biotic relationship	free-living	NAS
MIGS-14	Pathogenicity	none	TAS [100]
	Biosafety level	1	TAS [100]
	Isolation	cores of hot sediment	TAS [1]
MIGS-4	Geographic location	Guaymas, Mexico	TAS [1]
MIGS-5	Sample collection time	before or around 1989	NAS
MIGS-4.1	Latitude	14.84	
MIGS-4.2	Longitude	-17.23	NAS
MIGS-4.3	Depth	-2,000 m	TAS [1]
MIGS-4.4	Altitude	-2,000 m	TAS [1]

Evidence codes - IDA: Inferred from Direct Assay (first time in publication); TAS: Traceable Author Statement (i. e. a direct report exists in the literature); NAS: Non-traceable Author Statement (i. e. not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from the Gene Ontology project [42]. If the evidence code is IDA, then the property was directly observed for a living isolate by one of the authors or an expert mentioned in the acknowledgements.

Chemotaxonomy

In *A. profundus*, acyclic C_{40} tetraether, an unknown compound at an R_f in the range of cyclized glycerol-dialkyl-glycerol tetraethers, and a C_{20}:C_{20} diether constitute the membrane core lipids, whereas C_{20}:C_{25} diethers are absent, similar to *A. fulgidus* [1]. However, *A. profundus* differs from *A. fulgidus* in the composition of complex lipids, consisting of two phosphoglycolipids at R_f 0.10 and 0.13, and four glycolipids at R_f 0.40, 0.45, 0.60, 0.65, while the latter contains two phosphoglycolipids at R_f 0.10 and 0.215, one phospholipid at R_f 0.30 and one glycolipid at R_f 0.60 [1]. The cell envelope consists of an S-layer and is rifampicin and streptolydigin resistant [1].

Genome sequencing and annotation
Genome project history
This organism was selected for sequencing on the basis of its phylogenetic position [44], and is part of the **Genomic Encyclopedia of Bacteria and Archaea** project [45]. The genome project is deposited in the Genomes OnLine Database [32] and the complete genome sequence is available in GenBank. Sequencing, finishing and annotation were performed by the DOE Joint Genome Institute (JGI). A summary of the project information is shown in Table 2.

Table 2. Genome sequencing project information

MIGS ID	Property	Term
MIGS-31	Finishing quality	Finished
MIGS-28	Libraries used	Three 454 pyrosequence libraries, standard and two paired end (8 kb and 15kb insert sizes) and one Illumina library (300bp inset size)
MIGS-29	Sequencing platforms	454 Titanium, Illumina
MIGS-31.2	Sequencing coverage	136× 454 Titanium, 30× Illumina GAii
MIGS-30	Assemblers	Newbler, phrap
MIGS-32	Gene calling method	Prodigal, GenePRIMP
	INSDC ID	CP001857 (chromosome) CP001858 (plasmid)
	GenBank Date of Release	January 20, 2010
	GOLD ID	Gc01188
	NCBI project ID	32583
	Database: IMG-GEBA	2501939633
MIGS-13	Source material identifier	DSM 5631
	Project relevance	Tree of Life, GEBA

Growth conditions and DNA isolation
A. profundus AV18T, DSM 5631, was grown anaerobically in DSMZ medium 519 (*A. profundus* medium) [46] at 85°C. DNA was isolated from 1-1.5 g of cell paste using Masterpure Gram-positive DNA purification kit (Epicentre) with a modified protocol for cell lysis, st/DL according to Wu *et al*. [45].

Genome sequencing and assembly
The genome of strain AV18T was sequenced using a combination of 454 and Illumina sequencing platforms. All general aspects of library construction and sequencing can be found at http://www.jgi.doe.gov/. Pyrosequencing reads were assembled using the Newbler assembler version 2.0.0-PostRelease-10/28/2008 (Roche). Possible misassemblies were corrected with Dupfinisher [47] or transposon bombing of bridging clones (Epicentre Biotechnologies, Madison, WI). Gaps between contigs were closed by editing in Consed, by custom primer walk or PCR amplification. A total of 26 finishing reads were produced to close gaps, to resolve repetitive regions, and to raise the quality of the finished sequence. Illumina reads were used to improve the final consensus quality using an in-house developed tool (the Polisher, unpublished). The error rate of the completed genome sequence is less than 1 in 100,000. Pyrosequence provided 136× coverage of the genome and the final assembly contains 718,930 454-pyrosequence reads.

Genome annotation
Genes were identified using Prodigal [48] as part of the Oak Ridge National Laboratory genome annotation pipeline, followed by a round of manual curation using the JGI GenePRIMP pipeline [49]. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant database, UniProt, TIGR-Fam, Pfam, PRIAM, KEGG, COG, and InterPro databases. Additional gene prediction analysis and functional annotation was performed within the Integrated Microbial Genomes - Expert Review (IMG-ER) platform [50].

Genome properties

The 1,563,423 bp genome consists of a 1,560,622 bp chromosome and a 2,801 bp plasmid with an overall G+C content of 42.0% (Table 3 and Figure 3). Of the 1,909 genes predicted, 1,858 are protein-coding genes, and 52 RNAs; 35 pseudogenes were also identified. The majority of the protein-coding genes (60.0%) were assigned a putative function while the remaining ones were annotated as hypothetical proteins. The distribution of genes into COGs functional categories is presented in Table 4.

Insights from the genome sequence
Replicons

A. profundus AV18[T] is the second type strain of the *Archaeoglobi* with a fully sequenced genome to be published [34]. In contrast to *A. fulgidus*, the genome of AV18[T] has a small cryptic plasmid of 2,801 bp that contains four genes which appear to have no other function than the maintenance of this replicative unit. It displays a slightly lower G+C content (40%) than the rest of the genome and is negatively supercoiled, as demonstrated for pGS5 by López-García *et al.* [51]. The version of the plasmid presented here differs in three positions from the sequence of pGS5, resulting in one split gene.

Origin of replication

Unlike in bacteria, the archaeal initiation of the replication fork can occur at more than one site (origin of replication, ORI) on the chromosome [59], which heuristics used for bacteria fail to locate. Likewise, the ORI in *A. profundus* could not be detected by the use of Ori-Finder [52], which is consistent with several attempts to discover the replication origin in *A. fulgidus* by such methods [53-57]. Well-conserved replication signature patterns are known from both *Crenarchaea* and *Euryarchaea* [59]. In the genome of *A. fulgidus*, two almost identical ORB elements of 22 bases length are located at 65.6% of the length of the genome, which is in agreement with the position of the (single) ORI of this organism, identified by experimental origin mapping [58].

Pattern searching in the non-coding regions of the genome sequence of AV18[T] revealed a situation very much comparable to that of *A. fulgidus*: Two identical, but inverted ORB elements (**T**TTCCA-CAGGAAA**T**AAAGGGG**T**) were identified between genes Arcpr_1540 and Arcpr_1543; with 1,264 bases of distance (containing two hypothetical proteins) between each other, differing in only two bases from either ORB element in *A. fulgidus*. This position marks the predicted origin of replication in *A. profundus*, which is likewise far away from the (single copy of) *cdc 6* (generally considered as marker gene for the ORI) Arcpr_0001, located at the very beginning of the chromosome sequence. The presence of further active chromosomal ORIs cannot be excluded, but the strong similarity to the situation in *A. fulgidus* suggests that the genome of AV18[T] also contains only one origin of replication.

Table 3. Genome Statistics

| Attribute	Value	% of Total
Genome size (bp)	1,563,423	100.00%
DNA coding region (bp)	1,474,996	94.34%
DNA G+C content (bp)	656,709	42.00%
Number of replicons	2	
Extrachromosomal elements	1	
Total genes	1,909	100.00%
RNA genes	52	2.67%
rRNA operons	1	
Protein-coding genes	1,858	97.33%
Pseudo genes	35	1.83%
Genes with function prediction	1,145	59.98%
Genes in paralog clusters	167	8.75%
Genes assigned to COGs	1,267	66.37%
Genes assigned Pfam domains	1,301	68.15%
Genes with signal peptides	141	7.39%
Genes with transmembrane helices	328	17.18%
CRISPR repeats	0	
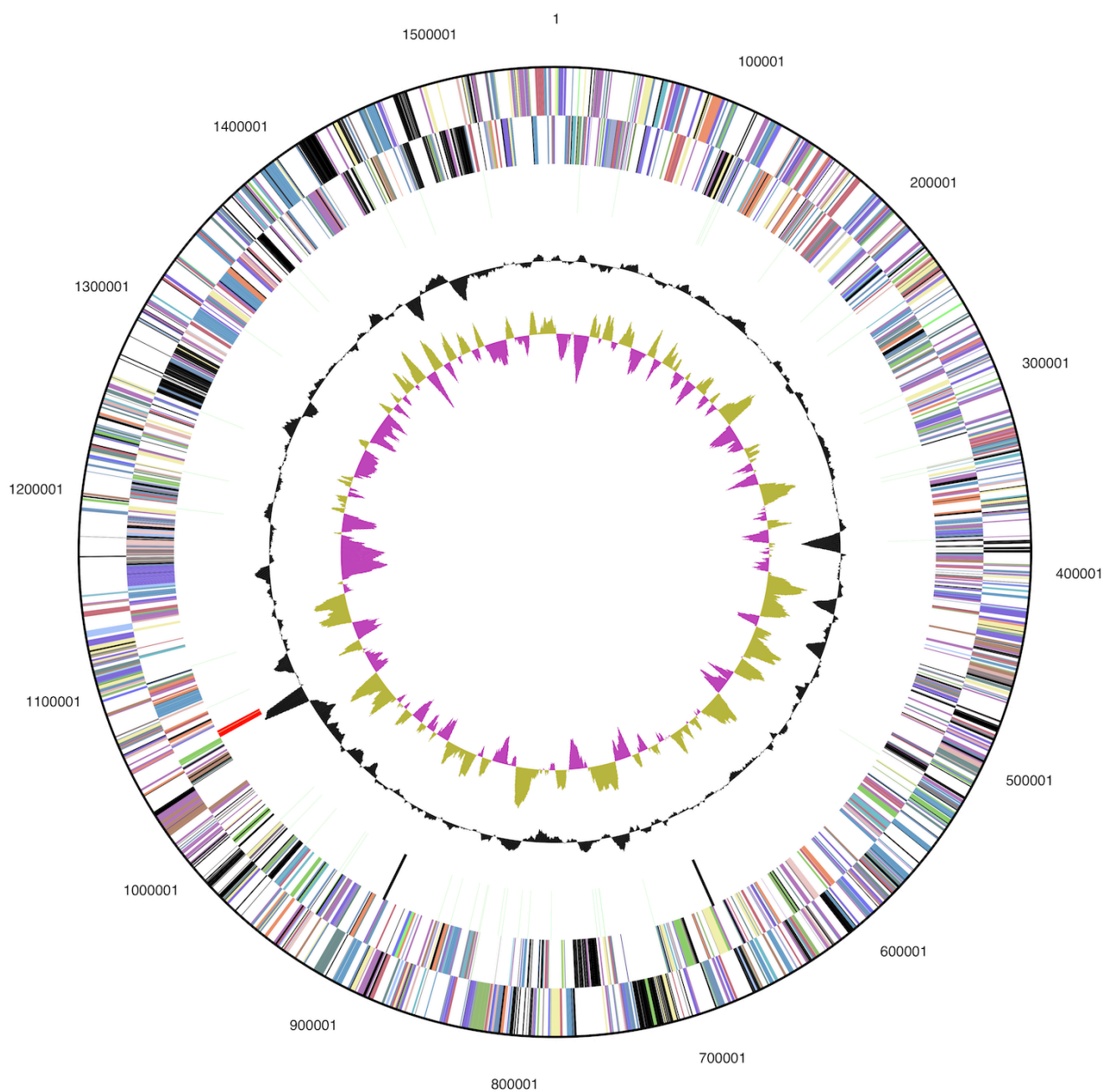
Figure 3. Graphical circular map of the genome (without the 2.8 kbp plasmid). From outside to the center: Genes on forward strand (color by COG categories), Genes on reverse strand (color by COG categories), RNA genes (tRNAs green, rRNAs red, other RNAs black), GC content, GC skew.

Shine-Dalgarno sequences

Before the start of the translation process, the recruitment of a ribosome to the mRNA is mediated by a species-specific DNA motif, the Shine-Dalgarno (SD) sequence [60], constituting the ribosome binding site (RBS) closely upstream of the coding region. In order to identify the SD consensus sequence in *A. profundus*, the Pattern Discovery Tool (oligo-analysis) of RSAT [61] was used for *de-novo* motif discovery within 50 bp regions upstream of all protein-coding genes in the genome of AV18T, with a background model estimated from its whole genome nucleotide sequence. The most frequently detected heptanucleotide was GGAGGTG, matching the complementary sequence one base shifted from the 3'-end of the 16S rRNA: TCTGCGGCTGGAT**CACCTCC**T-3' (bold: matching sequence) is obviously involved in ribosome recruitment.

Table 4. Number of genes associated with the general COG functional categories

Code	value	%age	Description
J	150	8.1	Translation, ribosomal structure and biogenesis
A	2	0.1	RNA processing and modification
K	67	3.6	Transcription
L	79	4.3	Replication, recombination and repair
B	4	0.2	Chromatin structure and dynamics
D	13	0.7	Cell cycle control, cell division, chromosome partitioning
Y	0	0.0	Nuclear structure
V	4	0.2	Defense mechanisms
T	46	2.5	Signal transduction mechanisms
M	40	2.2	Cell wall/membrane biogenesis
N	18	1.0	Cell motility
Z	0	0.0	Cytoskeleton
W	0	0.0	Extracellular structures
U	21	1.1	Intracellular trafficking and secretion
O	53	2.9	Posttranslational modification, protein turnover, chaperones
C	109	5.9	Energy production and conversion
G	41	2.2	Carbohydrate transport and metabolism
E	110	5.9	Amino acid transport and metabolism
F	47	2.5	Nucleotide transport and metabolism
H	85	4.6	Coenzyme transport and metabolism
I	24	1.3	Lipid transport and metabolism
P	57	3.1	Inorganic ion transport and metabolism
Q	7	0.4	Secondary metabolites biosynthesis, transport and catabolism
R	201	10.8	General function prediction only
S	169	9.1	Function unknown
-	642	34.6	Not in COGs

Using Prodoric Virtual Footprint software [62], the frequencies of heptanucleotides which are able to match (allowing one mismatch) the 3'-end of the 16S rRNA were determined. A significant drop was observed when the seven base window reached the base C at position eleven of the reverse complement 16S rRNA terminus (Table 5), indicating that interactions with the RBS are restricted to the ten most distal bases.

In total, the upstream regions of 950 genes match at least one of the four most frequently observed heptanucleotides, representing 51% of all protein-coding genes. Bakke *et al.* [63] recently evaluated three current genome annotation pipelines on the basis of the *Halorhabdus utahensis* genome [64] and recommended the integration of species-specific SD-motifs into the ORF-calling process of automated genome annotation pipelines, in order to determine the correct start codons of protein-coding genes. In several members of the *Archaea* (group A *sensu* Torarinsson *et al.* [65]), however, the benefits of this approach might be limited by the fact that single genes and first genes of operons are often leaderless (in *A. fulgidus*: 50%), thus containing no SD sequence [65]. Despite the expected abundance of leaderless transcripts, the percentage of genes preceded by SD sequences is significantly higher than the percentage observed in the genome of *H. utahensis* [64] based on the same annotation pipeline: Scanning 50 bp areas upstream of all *H. utahensis* genes with the most common heptanucleotide (allowing one mismatch) matched in only 8.6% of the respective areas of all genes, while the genome of strain AV18[T] reached 30.6%.

Table 5. Reverse complement of the 16S rRNA 3'-end

scanning heptanucleotide	abs. frequency	rel. frequency
AGGAGGT	555	29.9%
GGAGGTG	569	30.6%
GAGGTGA	565	30.4%
AGGTGAT	549	30.0%
GGTGATC	283	15.2%
GTGATCC	88	4.7%
TGATCCA	89	4.8%
GATCCAG	53	2.9%
ATCCAGC	58	3.1%
TCCAGCC	45	2.4%

Assessment of the frequencies of heptanucleotide sections with one acceptable mismatch in 50 bp regions upstream of all genes in AV18^T.

The heptanucleotide matching the very end of the 16S rRNA terminus is slightly less represented than the following shifted motifs, indicating that the final T of the 16S terminus might not be as essential for the RBS recognition as the preceding bases. This is consistent with recent insight into crystal structure and dynamics of the SD helix in an initiation-like 70S ribosome complex of *Thermus thermophilus*, showing base pairings of positions two to nine from the 3'-end of the 16S rRNA and the SD sequence of the mRNA, excluding an interaction with the very last base of the rRNA [66]. Transferring these results to the analysis of the SD sequence in strain AV18^T, the comparatively high observed frequency of motif AGGAGGT is likely due to the setting of the motif scan, which allows one mismatch. The same is true for the opposite side of the SD sequence, and the reason for the high frequency of motif AGGTGAT. Therefore, the predicted complete, species-specific consensus RBS motif of *A. profundus* is the 8-base pattern GGAGGTGA, which represents the functional sequence area of interaction in the initial contact between ribosome and mRNA in *A. profundus.*

tRNAs and Codon usage

By the use of tRNAscan-SE [67], a total of 48 tRNAs were identified and the coverage of all possible codons was assessed. Two codons are redundantly represented by tRNAs: AUC (two copies of Ile-tRNA gene) and AUG (four copies of Met-tRNA gene). None of the codons ending on U are present. Apart from these, AUA is the only codon that is not directly associated with a tRNA. The translation of this codon is strictly dependent on wobble modifications that are carried out by different modification systems in the three domains of life. Insight into the archaeal mechanism of AUG translation was gained very recently [68], involving the polyamine-conjugated modified base 2-agmatinylcytidine (agm^{2}C) at the wobble position of the corresponding tRNA, and the enzyme tRNA^{Ile}-agm^{2}C synthetase (TiaS), which catalyzes the agm^{2}C formation using agmatine and ATP. A candidate for this enzyme in *A. profundus* AV18^T is Arcpr_0572, identified by sequence similarity with the experimentally confirmed TiaS gene in *A. fulgidus* (AF2259). Arcpr_0572 displays the highest similarity to AF2004, one of three genes belonging to the same gene family in *A. fulgidus*. Therefore, a bidirectional best BLAST hit to the experimentally confirmed TiaS gene in *A. fulgidus* cannot be identified in *A. profundus*.

Redundant or missing representation of codons by tRNAs has apparently no effect on the frequency of codon usage (determined by program gp_cusage; data not shown), as both are used in some cases more frequently, in other cases less frequently than the corresponding alternative codon which is allocated exactly one tRNA. The

tRNAs for Trp, Tyr and one of the Met-tRNAs contain introns of 60, 17 and 26 bases length, respectively. Concerning the frequencies of the utilized start codons, 84.6% of the protein-coding genes start with AUG, while the frequency of this start codon in *A. fulgidus* is considerably lower (76.5%). The frequency of the alternative start codon GUG (10.4%) in *A. fulgidus* is almost twice as high (19.5%), reflecting the difference in GC-content (*A. fulgidus*: 48.6%), while UUG is rare in both (*A. profundus*: 4.4%, *A. fulgidus*: 3.2%). The correct prediction of start codons plays a decisive role in the ORF-calling process. In a comparison between three current genome annotation pipelines, 90% of the predicted genes shared the same stop codons, while only 48% thereof agreed in start codon prediction, resulting in different gene lengths [63]. The average gene length in the genome of AV18[T] is only 773 bp, while *A. fulgidus* genes are on average 815 bp long, a difference which – along with the different frequencies of alternative start codons – might also be caused by the different annotation pipelines used for both genomes [63].

Comparative genomics

The genome sequencing for the type strain of another species of the *Archaeoglobaceae*, *F. placidus* AEDII12DO[T], provided the opportunity for a genome-wide comparative analysis among three species of the *Archaeoglobaceae*. All of these analyses were performed using IMG online tools [69] with the default settings, unless stated otherwise. Metabolic pathways were reconstructed by the combination of online resources such as NCBI, KEGG [70],), BRENDA [71] and MetaCyc [72]. Orthology of genes was determined by bidirectional best BLAST [73] hits and the comparison of functional groups using EBI InterProScan [74]. Phylogenetic comparisons are restricted to validly named species only. This limitation excludes e.g. '*Nanoarchaeum equitans*', '*Cenarchaeum symbiosum*' and strains assigned to the category *Candidatus*.

The genome size of *A. profundus* AV18[T] (1.6 Mb) is significantly smaller than those of *A. fulgidus* (2.2 Mb, 2,468 protein-coding genes [34]) and *F. placidus* (2.2 Mb, 2,622 protein-coding genes). Figure 4 shows the numbers of shared genes in a Venn-diagram. *A. fulgidus* and *F. placidus* share a considerable number of genes that are not present in *A. profundus*. These genes are associated with a wide range of functions and pathways, some of which

will be discussed below in more detail. This fraction of genes includes the seven subunits of carbon monoxide dehydrogenase, two of the key enzymes for the β-oxidation of fatty acids, and genes belonging to the CRISPR/Cas system.

The genome of strain AV18[T] contains only a small percentage (8.7%) of paralogous genes, as compared to 12.8% in *F. placidus* and 17.1% *A. fulgidus* (http://img.jgi.doe.gov). Likewise, the percentage of genes with signal peptides in strain AV18[T] (7.4%) is considerably lower than those of *A. fulgidus* (10.8%) and *F. placidus* (10.1%.

DNA-polymerase genes

To date, four distinct DNA-dependent DNA-polymerase families are known. They are specifically distributed across the three domains of life, with the unrelated B and D family polymerases being present in *Archaea* [75]., The evolutionary divergence further discriminates *Crenarchaeota*, which have up to three family B monomeric DNA polymerases, and *Euryarchaeota*, which generally have one monomeric family B DNA polymerase and one heterodimeric family D DNA polymerase [76].

Three different family B DNA polymerases have been detected in Archaea [77-79], B3 being the single family B DNA-polymerase identified in the genome of *A. profundus* AV18[T]. The respective gene, Arcpr_0273 is also present in the genomes of *A. fulgidus* (in contrast to the current annotation, which assigned subtype B1 to this gene) and *F. placidus*. Each of the three *Archaeoglobi* contains also one copy of the euryarchaeal family D DNA polymerase, and *A. fulgidus* is unique by having a second family B DNA polymerase gene (AF0693), belonging to subtype B2.

RNA polymerase β subunit

The DNA-dependent RNA polymerase (RNAP) subunit B was previously reported as a suitable tool for phylogenetic reconstructions [80]. A split in the B subunit of the RNA polymerase – resulting in the fragments B' and B'' – has been reported for a subset of the euryarchaeal branch containing the methanogens and halophiles, based on the first five available archaeal sequences of this gene. This split has been described to be phylogenetically conserved and its use for supporting or refuting branching topologies has been suggested [80].
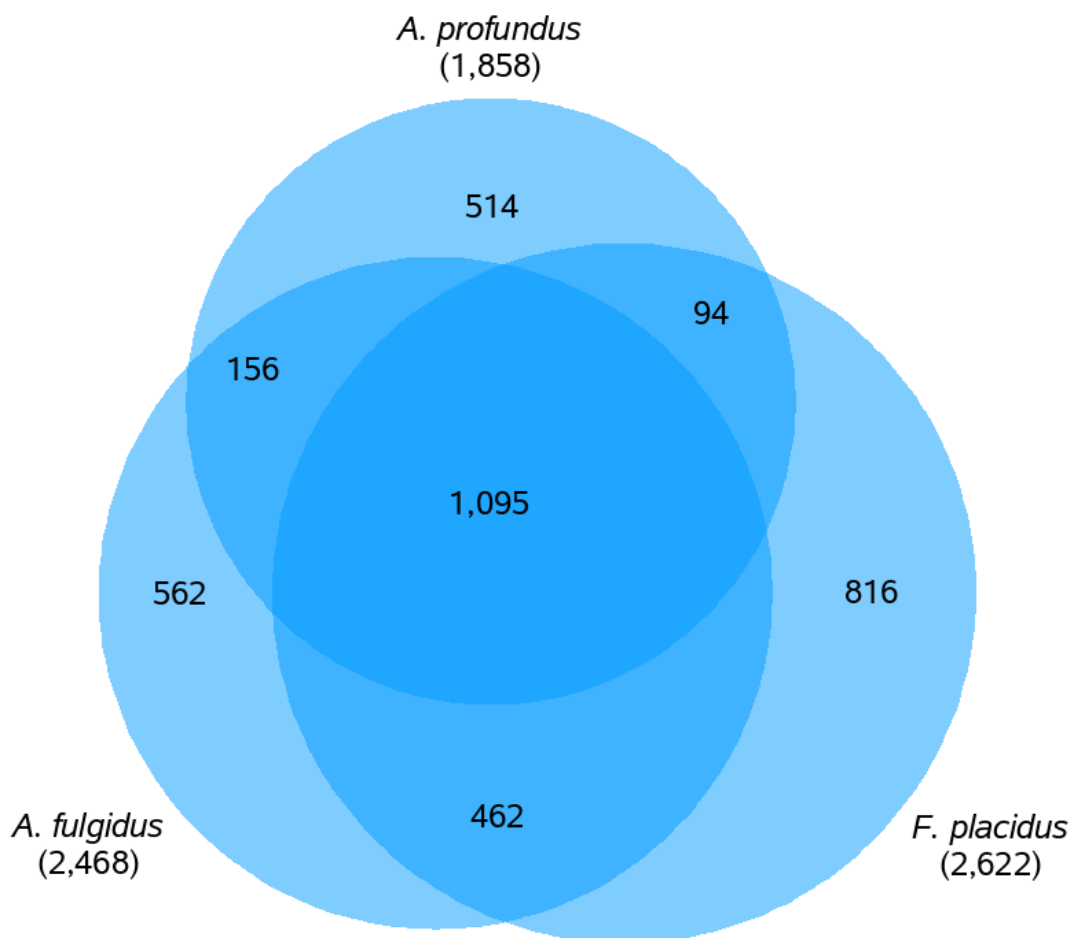
Figure 4. Venn-diagram depicting the intersections of protein sets (total numbers in parentheses) of the three completely sequenced *Archaeoglobi* genomes. All intersections concerning *A. profundus* are gene counts of AV18[T], the remaining intersection between *A. fulgidus* and *F. placidus* only, are gene counts in *A. fulgidus*. Due to variable copy numbers of several genes in the three species, the fragments do not add up to the total numbers of genes for *A. fulgidus* and *F. placidus*.

Here, the validity of this observation was reassessed, based on a larger number of available archaeal RNAP subunit B genes (n=77) from all of the currently available fully sequenced genomes. For organisms exhibiting the above mentioned split, the corresponding amino acid sequences of the B' and B'' component were joined and a phylogenetic tree was inferred (Figure 5), showing clusters that are largely consistent with the 16S rRNA tree topology [31]. The topology of this tree suggests a polyphyletic origin of the split in the B subunit, however, the best tree under the constraint of monophyly is not significantly worse (α=0.01) than the tree shown [30]. Therefore, this tree is not significantly in conflict with the assumption of a unique origin of the split into the B' and B'' components of RNAP. Further mapping of the species exhibiting the conserved split against the 16S rRNA

phylogeny confirmed the suggestion that this split is the result of a singular event which had taken place in the evolution of the *Euryarchaea* [80]. The lowest branching family containing this conserved split are *Archaeoglobaceae* represented by *A. profundus*, *A. fulgidus* and *F. placidus* (genes: Arcpr_0976/7, AF1886/7, Ferp_0762/3). Likewise, all taxa which diverged later from the main branch, i.e. *Methanococci*, *Methanobacteria*, *Methanomicrobia*, *Halobacteria* and possibly *Methanopyrus kandleri* (the basal position of the latter in the 16S rRNA-based phylogenetic tree is disputed [82]), contain this split without exception. Taxa which diverged earlier (*Thermococci*, *Thermoplasmata* and all *Crenarchaeota*) have the unfragmented version of the B subunit, equally without exception among validly named organisms.

─── 0.1

```
                              90┌ Halomicrobium mukohataei arg-2
                            93└ Haloarcula marismortui ATCC 43049
                                  Halorhabdus utahensis AX-2
                          90    Natronomonas pharaonis Gabara
                        100    Haloterrigena turkmenica 4k
                                Halogeometricum borinquense PR3
                    71   100└ Halorubrum lacusprofundi ACAM 34
                           100 Methanoculleus marisnigri JR1.1
                        68    Methanoculleus marisnigri JR1.2
                       100    Methanospirillum hungatei JF-1
                99             Methanocorpusculum labreanum Z
                       100    Methanohalophilus mahii SLP (Salt Lake Paterek)
                       100    Methanococcoides burtonii DSM 6242
        100              92    Methanosarcina acetivorans C2A
                               Methanosaeta thermophila PT
                      99    Archaeoglobus fulgidus VC-16
                      100   Archaeoglobus profundus AV18
                            Ferroglobus placidus AEDII12DO
    93                   100┌ Thermoplasma volcanium GSS1
                       100    Thermoplasma acidophilum AMRC-C165
                               Picrophilus torridus DSM 9790
                   94┌ Thermococcus gammatolerans EJ3
                   └ Thermococcus kodakarensis KOD1
              100   99┌ Pyrococcus furiosus Vc 1
                     └ Pyrococcus horikoshii JA-1
                        Thermococcus sibiricus MM 739
                    100┌ Methanocaldococcus fervens AG86
              100     └ Methanocaldococcus jannaschii JAL-1
                   100┌ Methanococcus vannielii ATCC 35089
                     └ Methanococcus aeolicus Nankai-3
                  79┌ Methanobacterium thermautotrophicumm Delta H
              100  └ Methanobrevibacter smithii PS
          100        Methanosphaera stadtmanae MCB-3
                     Methanothermus fervidus ATCC 43054
                      Methanopyrus kandleri AV19
                 100┌ Thermosphaera aggregans M11TL
            100    └ Desulfurococcus kamchatkensis 1221n
                     Staphylothermus marinus F1
               81┌ Aeropyrum pernix K1
        80    77└ Ignicoccus hospitalis KIN4/I
                    Hyperthermus butylicus ATCC_700455
        100         Ignisphaera aggregans AQ1.S1
                 100┌ Sulfolobus tokodaii 7
            100    └ Sulfolobus acidocaldarius ATCC 33909
                    Metallosphaera sedula TH2
                 61┌ Thermoproteus neutrophilus V24Sta
                   └ Pyrobaculum islandicum GEO3
                    Pyrobaculum aerophilum IM2
                100 Pyrobaculum arsenaticum PZ6*
            100    Pyrobaculum calidifontis VA1
                     Caldivirga maquilingensis IC-167
```

Figure 5. Phylogenetic tree of archaeal type strains with fully sequenced genomes, inferred using the maximum likelihood criterion [30], based on an alignment of the RNA polymerase B subunit sequence and rooted with the node which separates *Cren-* and *Euryarchaeota*. The alignment was inferred by Muscle [81] software, using the PROTCATLGF substitution model. Bootstrapping was performed using RAxML [30] and values above 60% mark the corresponding nodes. Species containing a conserved split in the RNA polymerase B subunit gene are displayed in bold.

CRISPRs

Clusters of Regularly Interspaced Short Palindromic Repeats (CRISPRs) represent a recently discovered prokaryotic defense system against viral attacks [83,84]. Although frequently observed in members of the *Archaea* (~90%), *A. profundus* completely lacks any CRISPRs. In contrast, the genome of *A. fulgidus* contains three large CRISPR spacer/repeat arrays, consisting of 44 to 60 repeats of lengths between 30 and 37 bases per repeat [34]. *Ferroglobus* contains twelve CRISPR arrays of variable repeat lengths and copy numbers (JGI, unpublished).

Motility and chemotaxis genes

A widespread phenomenon among *Archaea* and *Bacteria* is their ability to sense environmental conditions by the chemotaxis system and actively move towards more favorable locations by the activity of the flagellum. The archaeal flagella are non-homologous to those of *Bacteria*, and their components are encoded by one or two well-conserved gene clusters (*fla* clusters) [85], which have been subject to extensive phylogenetic studies [86]. *A. profundus* is reported to be non-motile [1,43], showing no flagellation, in contrast to most *Archaeoglobi*, including *A. fulgidus* [3] and *F. placidus* [4]. Unexpectedly, the genome sequence revealed the presence of a complete *fla* gene cluster (Arcpr_1384 – Arcpr_1391) and the preflagellin peptidase *FlaK* gene (Arcpr_0277), [85]. The situation in *A. fulgidus* (AF_1048– AF_1055, *flaK*-gene: AF_0936) and *F. placidus* (Ferp_1456–Ferp_1463, *flaK*-gene: Ferp_0061) is virtually identical in content, order and orientation of genes of the *fla* cluster, therefore the different phenotypes are unexpected. However, a conflict between presence of the flagella genes and the phenotypically observed lack of motility is not unique for *A. profundus*, but has also been reported for *Methanosarcina* species [86]. Also the reverse, even more surprising case – observed motility, but lacking homologues of the genes coding for flagellum components – has been reported for *Pyrobaculum aerophilum* and *M. kandleri* [86]. Some of our electron micrograph images (data not shown) displayed structures which might be flagella on few *A. profundus* cells, mainly observed in larger cell clots. This indicates that *A. profundus* might be flagellated under certain conditions, not necessarily for motility reasons, but also functions such as cell-cell adhesion to form cell aggregates (as reported for *Methanosarcinales*) are thinkable.

In any case, the possibility of artifacts (e.g. the presence of fragments from damaged cells) causing the observed structures on our electron micrographs cannot be excluded.

Unlike the flagellum genes, the archaeal chemotaxis system is homologous to the one in bacteria (for a review see [87]). Using the IMG Phylogenetic Profiler, the genomes of *A. profundus*, *A. fulgidus* and *F. placidus* revealed the same genetic components for a chemotaxis system (AF1034, AF1037–AF1042, AF1044; Arcpr_1371–Arcpr_1376, Arcpr_1378, Arcpr_1379; Ferp_1072–Ferp_1377, Ferp_1379, Ferp_1990), with the only exception that *A. fulgidus* displays two copies of the methyl-accepting chemotaxis protein, while the others only have one. This observation again supports the hypothesis that *A. profundus* might be motile under certain conditions, otherwise not only its flagellum-genes, but also the genetic components for chemotaxis would remain unused. However, the archaeal system of motility and chemotaxis is not yet fully unraveled. Especially the proteins constituting the flagellar motor and the link between chemotactic signal transduction and the motility apparatus [88]. The lack of undescribed essential components for this complex cannot be ruled out for *A. profundus*, which might be the reason for the observed immobility.

β-oxidation of long-chain fatty acids

The ability of the *Archaeoglobi* to anaerobically oxidize long-chain fatty acids has been discussed controversially: although a β-oxidation system in *A. fulgidus* was predicted from the genome sequence [34], followed by reports of growth on crude and olive oil [89], "*G. ahangari*" was later reported to be the first hyperthermophile with this capacity [5]. Very recently, *A. fulgidus* VC-16 was demonstrated to be capable of growth on a wide range of fatty acids and alkenes as sole source of energy, using thiosulfate or sulfate as the electron acceptor [90]. Likewise, the genome of *F. placidus* contains at least the four key enzymes for β-oxidation, suggesting the presence of this pathway

In the first description of *A. profundus*, minor growth on acetate containing crude oil was observed [1]. With the here reported complete genome sequence, it becomes clear that this organism is unique within its sequenced relatives in lacking two of the four key enzymes for β-

oxidation: 3-hydroxyacetyl-CoA-dehydrogenase (EC:1.1.1.35) and enoyl-CoA-hydratase (EC:4.2.1.17). Therefore it can now be posited that the reported growth on crude oil was most likely due to the contained traces of acetate, as the organism lacks essential components required for the oxidation of long-chain fatty acids *via* β-oxidation.

Nitrate reduction

Currently, *F. placidus* is the only validly named member of the *Archaeoglobi* which has been shown to be able to use nitrate as electron acceptor. A cluster of genes encoding a putative nitrate reductase has yet been identified in *A. fulgidus* (AF0173–AF0176) and discussed in the literature [34,91], again resulting in a conflict between genetic equipment and observed metabolic features, as a biochemical evidence for nitrate reduction is still missing in *A. fulgidus*. Homologues of these genes are also present in *A. profundus*, though distributed in two separate locations in the genome (Arcpr0672, Arcpr0674, Arcpr1727, Arcpr1728) and in *F. placidus* (Ferp_0121-Ferp_0124). The latter contains another nitrate reductase gene cluster (Ferp_0311–Ferp_0314, additional gamma subunit: Ferp_1088), which might be the reason for the observed nitrate respiration in culture conditions, while specificity and activity of the more widely distributed hypothetical nitrate reductase gene cluster remains subject to further experiments.

Sulfate reduction

The reduction of sulfurous compounds is the central electron accepting pathway in the metabolism of *A. profundus*. The genetic equipment for the catalysis of the corresponding reactions is largely equivalent to the one previously described for *Desulfovibrio* species and postulated for *Desulfohalobium retbaense* [92]. The respective genes of *A. profundus* have been determined by sequence comparisons and identification of the corresponding functional groups. A notable difference to the mechanism of sulfate-reduction in *Desulfovibrio* species is the absence of a periplasmic cytochrome buffer composed of cytochrome c_3.

Thus, genes encoding a molybdopterin oxidoreductase MOP complex – as described for *Desulfovibrio desulfuricans G20* [93] – have not been identified in the genome of *A. profundus*. The MOP complex is thought to transfer electrons to menaquinone by interacting with periplasmic reduced cytochrome c_3. The regeneration of the reduced menaquinone pool is most likely performed by a set of F_{420}-nonreducing hydrogenase family proteins (genes: Arcpr_1002, Arcpr_1005 and Arcpr_1006) which transfer electrons originating from the oxidation of hydrogen – *via* a co-localized gene (Arcpr_1004) encoding a membrane associated cytochrome b – to oxidized menaquinone molecules in the membrane. Another option for the reduction of the menaquinone pool is given by a $F_{420}H_2$:quinone oxidoreductase complex, utilizing electrons supplied by $F_{420}H_2$. This reduced electron carrier originates from the pathway of reverse methanogenesis, which is a typical feature of the *Archaeoglobi*. The $F_{420}H_2$:quinone oxidoreductase complex has been studied in *A. fulgidus* [94,95] and a similar gene cluster exists in *A. profundus* (Arcpr_1575–Arcpr_1584). One of three additional proteins which have been found in the purified complex of *A. fulgidus* [94] has also been identified in *A. profundus* (Arcpr_0247) by reciprocal BLAST search.

The quinone-interacting membrane-bound oxidoreductase (QMO)-complex (Arcpr_0661–Arcpr_0663) transfers electrons *via* the heterodimeric AprAB complex (Arcpr_1261, Arcpr_1262) from the reduced menaquinone pool in the membrane to activated sulfate (APS, adenosine-5'-phosphosulfate), forming sulfite. Likewise, the membrane-associated DsrMKJOP (Arcpr_1727–Arcpr_1731) complex transfers electrons from the same source to the dissimilatory sulfite reductase (Arcpr_0139-Arcpr_0141), catalyzing the reduction from sulfite to sulfide. Both processes are used to generate a membrane potential with the major purpose of ATP production.

Carbon monoxide dehydrogenase

The enzymatic equipment used for reverse methanogenesis in *A. fulgidus* is equivalent to the the "Eastern branch" of the Wood-Ljungdahl pathway, which is also present in acetogenic organisms [96]. This pathway consists of two branches, each reducing a CO_2 into a methyl- and a carbonyl-moiety, respectively, which are joined forming acetyl-CoA. This metabolic capacity is not present in *A. profundus*, due to a blocked "Western branch" (acetyl-CoA decarbonylase/synthase is absent), a fact which has been discovered already in 1995 [97]. The consequence for *A. profundus* is its inability to grow autotrophically [43,97]. In both *A. fulgidus* and *F. placidus*, all genes for the complete set of seven different subunits of the acetyl-CoA decarbonylase/synthase are present and both can

grow autotrophically, like all other described *Archaeoglobaceae*, except *A. profundus* and *A. infectus*. However, *A. profundus* might be able to use the presence of the Eastern branch of the Wood-Ljungdahl pathway for a certain amount of CO_2-fixation, as the intermediate 5,10-methylene-tetrahydro-methanopterine can be branched off to other pathways, e. g. by formaldehyde-activating enzyme (Arcpr_1052) into formaldehyde, or by glycine hydroxymethyltransferase (Arcpr_0687, Arcpr_1587) to the glycine, serine and threonine metabolism.

Besides providing comprehensive insight into the genetic equipment, the completely sequenced genome of *A. profundus* revealed instances in which the presence of certain genes suggests capabilities which were not observed in laboratory cultivation, such as flagellation or chemotaxis. Reasons for this might be paralogous genes, *e. g.* having altered, yet unidentified substrate specificity, defect genes, pseudogenes or genes which are permanently transcriptionally deactivated, as reported for hydrogenase genes in *Methanosarcina acetivorans* [98]. Alternatively, the biochemic capacities might only be exhibited under specific unknown environmental conditions, which are yet to be reproduced in laboratory experiments.

Acknowledgements

References

1. Burggraf S, Jannasch HW, Nicolaus B, Stetter KO. *Archaeoglobus profundus* sp. nov., represents a new species within the sulfate-reducing archaebacteria. *Syst Appl Microbiol* 1990; **13**:24-28.

2. Validation List no. 34. Validation of the publication of new names and new combinations previously effectively published outside the IJSB. *Int J Syst Bacteriol* 1990; **40**:320-321. doi:10.1099/00207713-40-3-320

3. Stetter KO. *Archaeoglobus fulgidus* gen. nov., sp. nov.: a new taxon of extremely thermophilic archaebacteria. *Syst Appl Microbiol* 1988; **10**:172-173.

4. Hafenbradl D, Keller M, Dirmeier R, Rachel R, Roßnagel P, Burggraf S, Huber H, Stetter KO. *Ferroglobus placidus* gen. nov., sp. nov., a novel hyperthermophilic archaeum that oxidizes Fe^{2+} at neutral pH under anoxic conditions. *Arch Microbiol* 1996; **166**:308-314. PubMed doi:10.1007/s002030050388

5. Kashefi K, Tor JM, Holmes DE, Gaw Van Praagh CV, Reysenbach AL, Lovley DE. *Geoglobus ahangari* gen. nov., sp. nov., a novel hyperthermophilic archaeon capable of oxidizing organic acids and growing autotrophically on hydrogen with Fe(III) serving as sole electron acceptor. *Int J Syst Evol Microbiol* 2002; **52**:719-728. PubMed doi:10.1099/ijs.0.01953-0

6. Slobodkina GB, Kolganova TV, Querellou J, Bonch-Osmolovskaya EA, Slobodkin AI. *Geoglobus acetivorans* sp. nov., an iron(III)-reducing archaeon from a deep-sea hydrothermal vent. *Int J Evol Microbiol* 2009; **59**:2880-2883. doi:10.1099/ijs.0.011080-0

7. Huber H, Jannasch H, Rachel R, Fuchs T, Stetter KO. *Archaeoglobus veneficus* sp. nov., a novel facultative chemolithoautotrophic hyperthermophilic sulfide reducer, isolated from abyssal black smokers. *Syst Appl Microbiol* 1997; **20**:374-380.

8. Mori K, Maruyama A, Urabe T, Suzuki K, Hanada S. *Archaeoglobus infectus* sp. nov., a novel thermophilic chemolithoheterotrophic archaeon isolated from a deep-sea rock collected at Suiyo Seamount, Izu-Bonin Arc, western Pacific Ocean. *Int J Syst Evol Microbiol* 2008; **58**:810-816. PubMed doi:10.1099/ijs.0.65422-0

9. Steinsbu BO, Thorseth IH, Nakagawa S, Inagaki F, Lever MA, Engelen B, Øvreås L, Pedersen RB. *Archaeoglobus sulfaticallidus* sp. nov., a novel thermophilic and facultatively lithoautotrophic sulfate-reducer isolated from black rust exposed to hot ridge flank crustal fluids. *Int J Syst Evol Microbiol* 2010; PubMed doi:10.1099/ijs.0.016105-0

10. Stetter KO, Huber R, Blöchl E, Kurr M, Eden RD, Fielder M, Cash H, Vance I. Hyperthermophilic archaea are thriving in deep North Sea and Alaskan oil reservoirs. *Nature* 1993; **365**:743-745. doi:10.1038/365743a0

11. Chun J, Lee JH, Jung Y, Kim M, Kim S, Kim BK, Lim YW. EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. *Int J Syst Evol Microbiol* 2007; **57**:2259-2261. PubMed doi:10.1099/ijs.0.64915-0

12. Reysenbach AL, Longnecker K, Kirshtein J. Novel bacterial and archaeal lineages from an in situ growth chamber deployed at a mid-atlantic ridge hydrothermal vent. *Appl Environ Microbiol* 2000; **66**:3798-3806. PubMed doi:10.1128/AEM.66.9.3798-3806.2000

13. Voordeckers JW, Do MH, Hügler M, Ko V, Sievert SM, Vetriani C. Culture dependent and independent ATP citrate lyase genes: a comparison of microbial communities from different black smoker chimneys on the Mid-Atlantic Ridge. *Extremophiles* 2008; **12**:627-640. PubMed doi:10.1007/s00792-008-0167-5

14. Nercessian O, Reysenbach AL, Prieur D, Jeanthon C. Archaeal diversity associated with *in situ* samplers deployed in hydrothermal vents on the East Pacific Rise (13°N). *Environ Microbiol* 2003; **5**:492-502. PubMed doi:10.1046/j.1462-2920.2003.00437.x

15. Moussard H, Moreira D, Cambon-Bonavita MA, López-García P, Jeanthon C. Uncultured *Archaea* in a hydrothermal microbial assemblage: phylogenetic diversity and characterization of a genome fragment from a euryarchaeote. *FEMS Microbiol Ecol* 2006; **57**:452-469. PubMed doi:10.1111/j.1574-6941.2006.00128.x

16. Higashi Y, Sunamura M, Kitamura K, Nakamura K, Kurusu Y, Ishibashi J, Urabe T, Maruyama A. Microbial diversity in hydrothermal environments of Suiyo Seamount, Izu-Bonin Arc, using a catheter-type in situ growth chamber. *FEMS Microbiol Ecol* 2004; **47**:327-336. PubMed doi:10.1016/S0168-6496(04)00004-2

17. Nakagawa S, Takai K, Inagaki F, Chiba H, Ishibashi J, Kataoka S, Hirayama H, Nunoura T, Horikoshi K, Sako Y. Variability in microbial community and venting chemistry in a sediment-hosted back-arc hydrothermal system: Impacts of subseafloor phase-separation. *FEMS Microbiol Ecol* 2005; **54**:141-155. PubMed doi:10.1016/j.femsec.2005.03.007

18. Takai K, Horikoshi K. Genetic diversity of archaea in deep-sea hydrothermal vent environments. *Genetics* 1999; **152**:1285-1297. PubMed

19. Teske A, Hinrichs KU, Edgcomb V, de Vera Gomez A, Kysela D, Sylva SP, Sogin ML, Jannasch HW. Microbial diversity of hydrothermal sediments in the Guaymas Basin: evidence for anaerobic methanotrophic communities. *Appl Environ Microbiol* 2002; **68**:1994-2007. PubMed doi:10.1128/AEM.68.4.1994-2007.2002

20. Pagé A, Tivey MK, Stakes DS, Reysenbach AL. Temporal and spacial archaeal colonization of hydrothermal vent deposits. *Environ Microbiol* 2008; **10**:874-884. PubMed doi:10.1111/j.1462-2920.2007.01505.x

21. Nakagawa S, Inagaki F, Suzuki Y, Steinsbu BO, Lever MA, Takai K, Engelen B, Sako Y, Wheat CG, Horikoshi K and Integrated Ocean Drilling Program Expedition 301 Scientists. Microbial community in black rust exposed to hot ridge flank crustal fluids. *Appl Environ Microbiol* 2006; **72**:6789-6799. PubMed doi:10.1128/AEM.01238-06

22. Gittel A, Sørensen KB, Skovhus TL, Ingvorsen K, Schramm A. Prokaryotic community structure and sulfate reducer activity in water from high-temperature oil reservoirs with and without nitrate treatment. *Appl Environ Microbiol* 2009; **75**:7086-7096. PubMed doi:10.1128/AEM.01123-09

23. Kormas KA, Tamaki H, Hanada S, Kamagata Y. Apparent richness and community composition of *Bacteria* and *Archaea* in geothermal springs. *Aquat Microb Ecol* 2009; **57**:113-122. doi:10.3354/ame01333

24. Spear JR, Walker JJ, McCollom TM, Pace NR. Hydrogen and bioenergetics in the Yellowstone geothermal ecosystem. *Proc Natl Acad Sci USA* 2005; **102**:2555-2560. doi:10.1073/pnas.0409574102

25. Costa KC, Navarro JB, Shock EL, Zhang CL, Soukup D, Hedlund BP. Microbiology and geochemistry of great boiling and mud hot springs in the United States Great Basin. *Extremophiles* 2009; **13**:447-459. PubMed doi:10.1007/s00792-009-0230-x

26. Vick TJ, Dodsworth JA, Costa KC, Shock EL, Hedlund BP. Microbiology and geochemistry of Little Hot Creek, a hot spring environment in the Long Valley Caldera. *Geobiology* 2010; **8**:140-154. PubMed doi:10.1111/j.1472-4669.2009.00228.x

27. Kanokratana P, Chanapan S, Pootanakit K, Eurwi-laichtr L. Diversity and abundance of *Bacteria* and *Archaea* in the Bor Khuleng Hot Spring in Thailand. *J Basic Microbiol* 2004; **44**:430-444. doi:10.1002/jobm.200410388

28. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. *Bioinformatics* 2002; **18**:452-464. PubMed doi:10.1093/bioinformatics/18.3.452

29. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000; **17**:540-552. PubMed

30. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web-servers. *Syst Biol* 2008; **57**:758-771. PubMed doi:10.1080/10635150802429642

31. Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer KH, Ludwig W, Glöckner FO, Rosselló-Móra R. The all-species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst Appl Microbiol* 2007; **31**:241-250. doi:10.1016/j.syapm.2008.07.001

32. Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 2010; **38**:D346-D354. PubMed doi:10.1093/nar/gkp848

33. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, *et al*. Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii. Science* 1996; **273**:1058-1073. PubMed doi:10.1126/science.273.5278.1058

34. Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, *et al*. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus. Nature* 1997; **390**:364-370. PubMed doi:10.1038/37052

35. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 1981; **17**:368-376. PubMed doi:10.1007/BF01734359

36. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Tompson N, Allen MJ, Angiuoli SV, *et al*. The minimum information about a genome sequence" (MIGS) specification. *Nat*

Biotechnol 2008; **26**:541-547. PubMed doi:10.1038/nbt1360

37. Tor JM, Lovley DR. Anaerobic degradation of aromatic compounds coupled to Fe(III) reduction by *Ferroglobus placidus. Environ Microbiol* 2001; **3**:281-287. PubMed doi:10.1046/j.1462-2920.2001.00192.x

38. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains *Archaea, Bacteria*, and *Eucarya. Proc Natl Acad Sci USA* 1990; **87**:4576-4579. PubMed doi:10.1073/pnas.87.12.4576

39. Garrity GM, Holt JG. Phylum AII. *Euryarchaeota* phy. nov. *In:* Bergey's Manual of Systematic Bacteriology, vol. 1. 2nd ed. Edited by: Garrity GM, Boone DR, Castenholz RW. Springer, New York; 2001; pp 211-355.

40. Garrity GM, Holt JG. Class VI. *Archaeoglobi* class. nov. *In:* Bergey's Manual of Systematic Bacteriology, vol. 1. The Archaea and the deeply branching and phototrophic Bacteria. Springer, New York; 2001; pp 349.

41. Huber H, Stetter KO. Order I. *Archaeoglobales* ord. nov. *In*: Bergey's Manual of Systematic Bacteriology, second edition vol. The *Archaea* and the deeply branching and phototrophic Bacteria. Springer, New York; 2001; pp 349.

42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; **25**:25-29. PubMed doi:10.1038/75556

43. Hartzell P, Reed DW. The genus *Archaeoglobus*. In: Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E. (eds.), *Prokaryotes*, Third Edition, Springer, New York 2006; **3**:82-100.

44. Klenk HP, Göker M. En route to a genome-based classification of *Archaea* and *Bacteria*? *Syst Appl Microbiol* 2010; **33**:175-182. PubMed doi:10.1016/j.syapm.2010.03.003

45. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova N, Kunin V, Goodwin L, Wu M, Tindall BJ, *et al*. A phylogeny-driven genomic encyclopedia of *Bacteria* and *Archaea. Nature* 2009; **462**:1056-1060. PubMed doi:10.1038/nature08656

46. List of growth media used at DSMZ: http://www.dsmz.de/microorganisms/media_list.php

47. Sims D, Brettin T, Detter JC, Han C, Lapidus A, Copeland A, Glavina Del Rio T, Nolan M, Chen F, Lucas S, *et al*. Complete genome sequence of *Kytococcus sedentarius* type strain (541[T]). *Stand Genomic Sci* 2009; **1**:12-20. doi:10.4056/sigs.761

48. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; **11**:119. PubMed doi:10.1186/1471-2105-11-119

49. Pati A, Ivanova N, Mikhailova N, Ovchinikova G, Hooper SD, Lykidis A, Kyrpides NC. GenePRIMP: A Gene Prediction Improvement Pipeline for microbial genomes. *Nat Methods* 2010; **7**:455-457. PubMed

50. Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC. IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* 2009; **25**:2271-2278. PubMed doi:10.1093/bioinformatics/btp393

51. López-García P, Forterre P, van der Oost J, Erauso G. Plasmid pGS5 from the hyperthermophilic archaeon *Archaeoglobus profundus* is negatively supercoiled. *J Bacteriol* 2000; **182**:4998-5000. PubMed doi:10.1128/JB.182.17.4998-5000.2000

52. Gao F, Zhang CT. Ori-Finder: a web-based system for finding oriCs in unannotated bacterial genomes. *BMC Bioinformatics* 2008; **9**:79. PubMed doi:10.1186/1471-2105-9-79

53. Grigoriev A. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res* 1998; **26**:2286-2290. PubMed doi:10.1093/nar/26.10.2286

54. Mrázek J, Karlin S. Strand compositional asymmetry in bacterial and large viral genomes. *Proc Natl Acad Sci USA* 1998; **95**:3720-3725. PubMed doi:10.1073/pnas.95.7.3720

55. Salzberg SL, Salzberg AJ, Kerlavage AR, Tomb JF. Skewed oligomers and origins of replication. *Gene* 1998; **217**:57-67. PubMed doi:10.1016/S0378-1119(98)00374-6

56. Lopez P, Philippe H, Myllykallio H, Forterre P. Identification of putative chromosomal origins of replication in archaea. *Mol Microbiol* 1999; **32**:883-886. PubMed doi:10.1046/j.1365-2958.1999.01370.x

57. Rocha EPC, Danchin A, Viari A. Universal replication biases in bacteria. *Mol Microbiol* 1999; **32**:11-16. PubMed doi:10.1046/j.1365-2958.1999.01334.x

58. Maisnier-Patin S, Malandrin L, Birkeland NK, Bernander R. Chromosome replication patterns in the hyperthermophilic euryarchaea *Archaeoglobus fulgidus* and *Methanocaldococcus (Methanococcus) jannaschii*. *Mol Microbiol* 2002; **45**:1443-1450. PubMed doi:10.1046/j.1365-2958.2002.03111.x

59. Robinson NP, Dionne I, Lundgren M, Marsh VL, Bernander R, Bell SD. Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*. *Cell* 2004; **116**:25-38. PubMed doi:10.1016/S0092-8674(03)01034-1

60. Shine J, Dalgarno L. The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: Complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci USA* 1974; **71**:1342-1346. PubMed doi:10.1073/pnas.71.4.1342

61. van Helden J, André B, Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 1998; **281**:827-842. PubMed doi:10.1006/jmbi.1998.1947

62. Münch R, Hiller K, Barg H, Heldt D, Linz S, Wingender E, Jahn D. PRODORIC: prokaryotic database of gene regulation. *Nucleic Acids Res* 2003; **31**:266-269. PubMed doi:10.1093/nar/gkg037

63. Bakke P, Carney N, DeLoache W, Gearing M, Ingvorsen K, Lotz M, McNair J, Penumetcha P, Simpson S, Voss L, *et al*. Evaluation of three automated genome annotations for *Halorhabdus utahensis*. *PLoS ONE* 2009; **4**:e6291. PubMed doi:10.1371/journal.pone.0006291

64. Anderson I, Tindall BJ, Pomrenke H, Göker M, Lapidus A, Nolan M, Copeland A, Glavina Del Rio T, Chen F, Tice H, *et al*. Complete genome sequence of *Halorhabdus utahensis* type strain (AX-2[T]). *Stand Genomic Sci* 2009; **1**:218-225. doi:10.4056/sigs.31864

65. Torarinsson E, Klenk HP, Garrett RA. Divergent transcriptional and translational signals in archaea. *Environ Microbiol* 2005; **7**:47-54. PubMed doi:10.1111/j.1462-2920.2004.00674.x

66. Korostelev A, Trakhanov S, Asahata H, Laurberg M, Lancaster L, Noller HF. Interactions and dynamics of the Shine Dalgarno helix in the 70S ribosome. *Proc Natl Acad Sci USA* 2007; **104**:16840-16843. PubMed doi:10.1073/pnas.0707850104

67. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in ge-

nomic sequence. *Nucleic Acids Res* 1997; **25**:955-964. PubMed doi:10.1093/nar/25.5.955

68. Ikeuchi Y, Kimura S, Numata T, Nakamura D, Yokogawa T, Ogata T, Wada T, Suzuki T, Suzuki T. Agmatine-conjugated cytidine in a tRNA anticodon is essential for AUA decoding in archaea. *Nat Chem Biol* 2010; **6**:277-282. PubMed doi:10.1038/nchembio.323

69. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Anderson I, Lykidis A, Mavromatis K, *et al*. The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Res* 2010; **38**(Database issue):D382-D390. PubMed doi:10.1093/nar/gkp887

70. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000; **28**:27-30. PubMed doi:10.1093/nar/28.1.27

71. Chang A, Scheer M, Grote A, Schomburg I, Schomburg D. BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res* 2009; **37**(Database issue):D588-D592. PubMed doi:10.1093/nar/gkn820

72. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 2010; **38**(Database issue):D473-D479. PubMed doi:10.1093/nar/gkp875

73. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **215**:403-410. PubMed

74. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, *et al*. InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009; **37**(Database issue):D211-D215. PubMed doi:10.1093/nar/gkn785

75. Tahirov TH, Makarova KS, Rogozin IB, Pavlov YI, Koonin EV. Evolution of DNA polymerases: an inactivated polymerase-exonuclease module in Pol ε and a chimeric origin of eukaryotic polymerases from two classes of archaeal ancestors. *Biol Direct* 2009; **4**:11. PubMed doi:10.1186/1745-6150-4-11

76. Castrec B, Rouillon C, Henneke G, Flament D, Querellou J, Raffin JP. Binding to PCNA in euryarchaeal DNA replication requires two PIP motifs for DNA polymerase D and one PIP motif for DNA polymerase B. *J Mol Biol* 2009; **394**:209-218. PubMed doi:10.1016/j.jmb.2009.09.044

77. Pisani FM, De Martino C, Rossi MA. DNA polymerase from the archaeon *Sulfolobus solfataricus* shows sequence similarity to family B DNA polymerases. *Nucleic Acids Res* 1992; **20**:2711-2716. PubMed doi:10.1093/nar/20.11.2711

78. Prangishvili DA. DNA-dependent DNA polymerases of the thermoacidophilic archaebacterium *Sulfolobus acidocaldarius*. *Mol Biol USSR* 1986; **20**:477-488.

79. Edgell DR, Klenk HP, Doolittle WF. Gene duplications in evolution of archaeal family B DNA polymerases. *J Bacteriol* 1997; **179**:2632-2640. PubMed

80. Klenk HP, Zillig W. DNA-dependent RNA polymerase subunit B as a tool for phylogenetic reconstructions: branching topology of the archaeal domain. *J Mol Evol* 1994; **38**:420-432. PubMed doi:10.1007/BF00163158

81. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; **32**:1792-1797. PubMed doi:10.1093/nar/gkh340

82. Brochier C, Forterre P, Gribaldo S. Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the *Methanopyrus kandleri* paradox. *Genome Biol* 2004; **5**:R17. PubMed doi:10.1186/gb-2004-5-3-r17

83. Jansen R, Embden JD, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* 2002; **43**:1565-1575. PubMed doi:10.1046/j.1365-2958.2002.02839.x

84. Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. *Science* 2010; **327**:167-170. PubMed doi:10.1126/science.1179555

85. Ng SY, Chaban B, Jarrell KF. Archaeal flagella, bacterial flagella and type IV pili: a comparison of genes and posttranslational modifications. *J Mol Microbiol Biotechnol* 2006; **11**:167-191. PubMed doi:10.1159/000094053

86. Desmond E, Brochier-Armanet C, Gribaldo S. Phylogenomics of the archaeal flagellum: rare horizontal gene transfer in a unique motility structure. *BMC Evol Biol* 2007; **7**:106. PubMed doi:10.1186/1471-2148-7-106

87. Szurmant H, Ordal GW. Diversity in chemotaxis mechanisms among the bacteria and archaea.

Microbiol Mol Biol Rev 2004; **68**:301-319. PubMed doi:10.1128/MMBR.68.2.301-319.2004

88. Schlesner M, Miller A, Streif S, Staudinger WF, Müller J, Scheffer B, Siedler F, Oesterhelt D. Identification of *Archaea*-specific chemotaxis proteins which interact with the flagellar apparatus. *BMC Microbiol* 2009; **9**:56. PubMed doi:10.1186/1471-2180-9-56

89. Stetter KO, Huber R. The role of hyperthermophilic prokaryotes in oil fields. In: Bell CR, Brylinsky M, Johnson-Green P (eds.), *Microbial Biosystems: New Frontiers: Proceedings of the 8th International Symposium on Microbial Ecology*, Atlantic Canada Society for Microbial Ecology, Halifax, Canada 2000, pp. 369-375.

90. Khelifi N, Grossi V, Hamdi M, Dolla A, Tholozan JL, Ollivier B, Hirschler-Réa A. Anaerobic oxidation of fatty acids by the hypothermophilic sulfate-reducing archaeon *Archaeoglobus fulgidus. Appl Environ Microbiol* 2010; **76**:3057-3060. PubMed doi:10.1128/AEM.02810-09

91. Richardson DJ, Berks BC, Russell DA, Spiro S, Taylor CJ. Functional, biochemical and genetic diversity of prokaryotic nitrate reductases. *Cell Mol Life Sci* 2001; **58**:165-178. PubMed doi:10.1007/PL00000845

92. Spring S, Nolan M, Lapidus A, Del Rio TG, Copeland A, Tice H, Cheng JF, Lucas S, Land M, Chen F, *et al*. Complete genome sequence of *Desulfohalobium retbaense* type strain (HR$_{100}^{T}$). *Stand Genomic Sci* 2010; **2**:38-48. doi:10.4056/sigs.581048

93. Li X, Luo Q, Wofford NQ, Keller KL, McInerney MJ, Wall JD, Krumholz LR. A molybdopterin oxidoreductase is involved in H$_2$ oxidation in *Desulfovibrio desulfuricans* G20. *J Bacteriol* 2009; **191**:2675-2682. PubMed doi:10.1128/JB.01814-08

94. Kunow J, Linder D, Stetter KO, Thauer RK. F$_{420}$H$_2$:quinone oxidoreductase from *Archaeoglobus fulgidus*. Characterization of a membrane-bound multisubunit complex containing FAD and iron-sulfur clusters. *Eur J Biochem* 1994; **223**:503-511. PubMed doi:10.1111/j.1432-1033.1994.tb19019.x

95. Brüggemann H, Falinski F, Deppenmeier U. Structure of the F$_{420}$H$_2$:quinone oxidoreductase of *Archaeoglobus fulgidus*. Identification and overproduction of the F$_{420}$H$_2$:oxidizing subunit. *Eur J Biochem* 2000; **267**:5810-5814. PubMed doi:10.1046/j.1432-1327.2000.01657.x

96. Pierce E, Xie G, Barabote RD, Saunders E, Han CS, Detter JC, Richardson P, Brettin TS, Das A, Ljungdahl LG, Ragsdale SW. The complete genome sequence of *Moorella thermoacetica* (f. *Clostridium thermoaceticum*). *Environ Microbiol* 2008; **10**:2550-2573. PubMed doi:10.1111/j.1462-2920.2008.01679.x

97. Vorholt J, Kunow J, Stetter KO, Thauer RK. Enzymes and coenzymes of the carbon monoxide dehydrogenase pathway for autotrophic CO$_2$ fixation in *Archaeoglobus fulgidus* and the lack of carbon monoxide dehydrogenase in the heterotrophic *A. profundus. Arch Microbiol* 1995; **163**:112-118. doi:10.1007/s002030050179

98. Guss AM, Kulkarni G, Metcalf WW. Differences in hydrogenase gene expression between *Methanosarcina acetivorans* and *Methanosarcina barkeri. J Bacteriol* 2009; **191**:2826-2833. PubMed doi:10.1128/JB.00563-08

99. Huber H, Stetter KO. Family I. *Archaeoglobaceae* fam. nov. *In*: Bergey's Manual of Systematic Bacteriology, second edition vol. The *Archaea* and the deeply branching and phototrophic *Bacteria*. Springer, New York; 2001; pp 349.

100. Classification of *Bacteria* and *Archaea* in risk groups. http://www.baua.de TRBA 466