Genome sequence of the *Thermotoga thermarum* type strain (LA3\(^T\)) from an African solfataric spring

Markus Göker\(^1\), Stefan Springer\(^1\), Carmen Scheuner\(^1\), Iain Anderson\(^2\), Ahmet Zeytun\(^2,3\), Matt Nolan\(^2\), Susan Lucas\(^2\), Hope Tice\(^2\), Tijana Glavina Del Rio\(^2\), Jan-Fang Cheng\(^2\), Cliff Han\(^2,3\), Roxanne Tapia\(^2,3\), Lynne A. Goodwin\(^2,3\), Sam Pitluck\(^2\), Konstantinos Liolios\(^2\), Konstantinos Mavromatis\(^2\), Ioanna Pagani\(^2\), Natalia Ivanova\(^2\), Natalia Mikhailova\(^2\), Amrita Patti\(^2\), Amy Chen\(^2\), Krishna Palaniappan\(^2\), Miriam Land\(^2,3\), Loren Hauser\(^2,3\), Yun-juan Chang\(^2,5\), Cynthia D. Jeffries\(^2,5\), Manfred Rohde\(^6\), John C. Detter\(^2,3\), Tanja Woyke\(^2\), James Bristow\(^2\), Jonathan A. Eisen\(^2,7\), Victor Markowitz\(^4\), Philip Hugenholtz\(^2,8\), Nikos C. Kyrpides\(^2,9\), Hans-Peter Klenk\(^1*,\) Alla Lapidus\(^10,11\)

\(^1\) Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, 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
\(^6\) HZI – Helmholtz Centre for Infection Research, Braunschweig, Germany
\(^7\) University of California Davis Genome Center, Davis, California, USA
\(^8\) Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Australia
\(^9\) Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia
\(^10\) Theodosius Dobzhansky Center for Genome Bionformatics, St. Petersburg State University, St. Petersburg, Russia
\(^11\) Algorithmic Biology Lab, St. Petersburg Academic University, St. Petersburg, Russia

*Corresponding authors: Hans-Peter Klenk

Keywords: anaerobic, motile, thermophilic, chemoorganotrophic, solfataric spring, outer sheath-like structure, *Thermotogaceae*, GEBA

*Thermotoga thermarum* Windberger et al. 1989 is a member to the genomically well characterized genus *Thermotoga* in the phylum ‘*Thermotoga*’. *T. thermarum* is of interest for its origin from a continental solfataric spring vs. predominantly marine oil reservoirs of other members of the genus. The genome of strain LA3T also provides fresh data for the phylogenomic positioning of the (hyper-)thermophilic bacteria. *T. thermarum* strain LA3\(^T\) is the fourth sequenced genome of a type strain from the genus *Thermotoga*, and the sixth in the family *Thermotogaceae* to be formally described in a publication. Phylogenetic analyses do not reveal significant discrepancies between the current classification of the group, 16S rRNA gene data and whole-genome sequences. Nevertheless, *T. thermarum* significantly differs from other *Thermotoga* species regarding its iron-sulfur cluster synthesis, as it contains only a minimal set of the necessary proteins. Here we describe the features of this organism, together with the complete genome sequence and annotation. The 2,039,943 bp long chromosome with its 2,015 protein-coding and 51 RNA genes is a part of the Genomic Encyclopedia of Bacteria and Archaea project.

Introduction

Strain LA3\(^T\) (= DSM 5069 = NBRC 107925) is the type strain of the species *Thermotoga thermarum* [1], one out of currently nine species in the genus *Thermotoga* [2]. The genus name was derived from the Greek word thermê, heat, and the Latin word toga, Roman outer garment; *Thermotoga*,...
the hot outer garment [3]; the species epithet was derived from the Latin word thermarum, of warm springs, of warm baths [1]. Strain LA3T was originally isolated from a hot continental solfataric spring in Lac Abbé, southwest of Asbalto, Djibouti [1]. Here we present a summary classification and a set of features for *T. thermarum* LA3T, together with the description of the genomic sequencing and annotation.

**Features of the organism**

**16S rRNA gene analysis**

The single genomic 16S rRNA gene sequence of *T. thermarum* LA3T was compared with the Greengenes database [4] for determining the weighted relative frequencies of taxa and (truncated [5]) keywords as previously described [6,7]. The most frequently occurring genera were *Thermotoga* (53.9%), *Thermosipho* (29.1%), *Fervidobacterium* (11.0%), *Caldicellulosiruptor* (2.5%) and *Thermopallium* (1.4%) (130 hits in total). Regarding the two hits to sequences from members of the species, the average identity within HSPs was 100.0%, whereas the average coverage by HSPs was 95.7%. Regarding the 37 hits to sequences from other members of the genus, the average identity within HSPs was 92.1%, whereas the average coverage by HSPs was 98.4%. Among all other species, the one yielding the highest score was *Thermotoga hypogea* (U89768), which corresponded to an identity of 94.2% and an HSP coverage of 99.1%. (Note that the Greengenes database uses the INSDC (= EMBL/NCBI/DDBJ) annotation, which is not an authoritative source for nomenclature or classification.) The highest-scoring environmental sequence was DQ675048 (‘microbial production water -temperature petroleum reservoir clone QHO-B59’), which showed an identity of 99.0% and an HSP coverage of 82.0%. The most frequently occurring keywords within the labels of all environmental samples which yielded hits were ‘microbi’ (5.6%), ‘temperatur’ (3.2%), ‘spring’ (3.0%), ‘hot’ (2.6%) and ‘thermophil’ (2.3%) (117 hits in total). The most frequently occurring keywords within the labels of those environmental samples which yielded hits of a higher score than the highest scoring species were ‘microbi, petroleum, reservoir, temperatur’ (11.8%), ‘product, water’ (6.0%) and ‘aggregate-form, biodegrad, crude-oil-adh, fluid, niboli, oilfield, produc’ (5.8%) (2 hits in total). Some of these keywords fit well to the known ecology of *T. thermarum*.
comparing their newly isolated strain to the type strain of *K. olearia* (which might not yet have been publicly available when the study presented in [20] was conducted), published *T. shengliensis*, also in a novel genus. More recently, Nunoura et al. [21] added *K. arenicorallina* to the genus *Kosmotoga*. These authors also realized that *T. shengliensis* and *K. olearia* are more closely related to each other than *K. arenicorallina* to *K. olearia* and thus suggested to place *T. shengliensis* in *Kosmotoga* as *K. shengliensis* because *Kosmotoga* has priority over *Thermococcoides*.

Whereas the validation of *K. arenicorallina* was accepted by the International Journal of Systematic and Evolutionary Microbiology (IJSEM) [22], *K. shengliensis* was at first not accepted by the editors of IJSEM with reference to rule 31a [2] of the Bacteriological Code (Nunoura, pers. comm.). Probably the editors opined that a DNA-DNA hybridization experiment [23] between the type strains of *K. olearia* and *T. shengliensis* should be conducted to assess whether both represent a single or two distinct species. In the meantime, the name *K. shengliensis* has been validated, however. The advantages of this solution can be demonstrated by considering the number of conflicts between data and classification. With *Thermococcoides shengliensis* in use, the classification of the group caused one obvious problem, the paraphyly of *Kosmotoga* (Fig. 1), and one potential problem, that *K. shengliensis* and *T. shengliensis* might be conspecific. By accepting the proposal in [21] to assign *T. shengliensis* to the genus *Kosmotoga*, the first problem was solved and the second problem was not worsened.

*Th. thermarum* LA3T is Gram-negative-staining and rod-shaped, with a sheath that extends past the ends of the cell (Figure 2). Cells were reported to be 0.6 μm in width and 1.5-11 μm in length [1]. Flagella and motility were observed [1] (Table 1). Growth occurred between 55°C and 84°C with an optimum at 70°C [1]. The pH range for growth was 5.5-9.0 with 7.0 as the optimum [1]. The salinity range for growth was 0.2% to 0.55% NaCl with 0.35% as the optimum value [1]. Yeast extract was required for growth, and addition of glucose, maltose, or starch significantly increased cell yield [1]. H₂ and S⁰ both inhibited growth, and H₂S was not formed from S⁰ [1].
Figure 1 shows the phylogenetic neighborhood of *T. thermarum* in a 16S rRNA gene based tree. The sequence of the single 16S rRNA gene copy in the genome does not differ from the previously published 16S rRNA gene sequence (AB039769).

Figure 2. Scanning electron micrograph of *T. thermarum* LA3

Table 1. Classification and general features of *T. thermarum* LA3 according to the the MIGS recommendations [24].

| MIGS ID | Property                     | Term                                      | Evidence code |
|---------|------------------------------|-------------------------------------------|---------------|
|         | Current classification       |                                           |               |
|         | Domain                       | *Bacteria*                                | TAS [25]      |
|         | Phylum                       | *Thermotogae*                             | TAS [26,27]   |
|         | Class                        | *Thermotogae*                             | TAS [26,28]   |
|         | Order                        | *Thermotogales*                           | TAS [26,29]   |
|         | Family                       | *Thermotogaceae*                          | TAS [26,30]   |
|         | Genus                        | *Thermotoga*                              | TAS [3,31]    |
|         | Species                      | *Thermotoga thermarum*                    | TAS [1,32]    |
|         | Type strain                  | LA3                                        | TAS [1]       |
|         | Gram stain                   | negative                                   | TAS [1]       |
|         | Cell shape                   | rods with a ‘toga’ (a sheath-like structure) | TAS [1]       |
|         | Motility                     | motile                                     | TAS [1]       |
|         | Sporulation                  | not reported                               | TAS [1]       |
|         | Temperature range             | thermophile, 55-84°C                       | TAS [1]       |
|         | Optimum temperature          | 70°C                                       | TAS [1]       |
|         | Salinity                     | 0.2 - 0.6% NaCl (w/v), opt 0.35%           | TAS [31]      |
| MIGS-22 | Oxygen requirement           | anaerobe                                   | TAS [1]       |
| MIGS-6  | Habitat                      | low salinity hydrothermal well water       | TAS [1]       |
| MIGS-15 | Biotic relationship          | free living                                | TAS [1]       |
| MIGS-14 | Pathogenicity                | none                                       | NAS           |
| MIGS-23.1| Biosafety level             | 1                                          | TAS [33]      |
| MIGS-4  | Geographic location          | Lac Abbé, southwest of Asbalto, Djibouti   | TAS [1]       |
| MIGS-5  | Sample collection time       | 1989 or earlier                            | NAS           |
| MIGS-4.1| Latitude                     | 11.162                                     | NAS           |
| MIGS-4.2| Longitude                    | 41.781                                     | NAS           |
| MIGS-4.3| Depth                       | not reported                               | NAS           |
| MIGS-4.4| Altitude                    | 5 – 30 m                                   | TAS [1]       |

Evidence codes - 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). Evidence codes are from of the Gene Ontology project [34].
Chemotaxonomy
The analysis of complex lipids in strain LA3T showed that they were similar to those of T. maritima except that the less polar glycolipid was absent [1]. Analysis of core lipids showed that strain LA3T had one unidentified core lipid that was not present in T. maritima [1].

Genome sequencing and annotation

Genome project history
This organism was selected for sequencing on the basis of its phylogenetic position [35], and is part of the Genomic Encyclopedia of Bacteria and Archaea project [36,37]. The genome project is deposited in the Genomes On Line Database [14] and the complete genome sequence is deposited 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.

Growth conditions and DNA isolation
T. thermarum strain LA3T, DSM 5069, was grown anaerobically in DSMZ medium 498 (Thermotoga II medium) [38] at 80°C. DNA was isolated from 0.5-1 g of cell paste using MasterPure Gram-positive DNA purification kit (Epicentre MGP04100) following the standard protocol as recommended by the manufacturer with modification st/DL for cell lysis as described in Wu et al. 2009 [37]. DNA is available through the DNA Bank Network [39].

Genome sequencing and assembly
The genome was sequenced using a combination of Illumina and 454 sequencing platforms. All general aspects of library construction and sequencing can be found at the JGI website [40]. Pyrosequencing reads were assembled using the Newbler assembler (Roche). The initial Newbler assembly, consisting of one contig in one scaffold, was converted into a phrap [41] assembly by making fake reads from the consensus, to collect the read pairs in the 454 paired end library. Illumina GAii sequencing data (290.0 Mb) was assembled with Velvet [42] and the consensus sequences were shredded into 1.5 kb overlapped fake reads and assembled together with the 454 data. The 454 draft assembly was based on 14.0 Mb 454 draft data and all of the 454 paired end data. Newbler parameters are -consed-a 50-l 350-g -m -ml 20. The Phred/Phrap/Consed software package [41] was used for sequence assembly and quality assessment in the subsequent finishing process. After the shotgun stage, reads were assembled with parallel phrap (High Performance Software, LLC). Possible mis-assemblies were corrected with gapResolution [40], Dupfinisher [43], or sequencing cloned bridging PCR fragments with subcloning. Gaps between contigs were closed by editing in Consed, by PCR and by Bubble PCR primer walks (J.-F. Chang, unpublished). A total of 16 additional reactions were necessary to close gaps and to raise the quality of the finished sequence. Illumina reads were also used to correct potential base errors and increase consensus quality using a software Polisher developed at JGI [44]. The error rate of the completed genome sequence is less than 1 in 100,000. Together, the combination of the Illumina and 454 sequencing platforms provided 149.0 × coverage of the genome. The final assembly contained 414,118 pyrosequence and 1,166,274 Illumina reads.

Table 2. Genome sequencing project information

| MIGS ID | Property            | Term                                                                 |
|---------|---------------------|----------------------------------------------------------------------|
| MIGS-31 | Finishing quality   | Finished                                                             |
|         |                     | Three genomic libraries: one 454 pyrosequence standard librar, one 454 PE libra, one Illumina li- |
| MIGS-28 | Libraries used      | library, one 454 PE library (10 kb insert size), one Illumina li-
|         |                     | brary                                                               |
| MIGS-29 | Sequencing platforms| Illumina GAii, 454 GS FLX Titanium                                    |
| MIGS-31.2| Sequencing coverage| 142.2 × Illumina; 6.8 × pyrosequence                                |
| MIGS-30 | Assemblers          | Newbler version 2.3-PreRelease-10/20/2009, Velvet, phrap version SPS - 4.24 |
| MIGS-32 | Gene calling method | Prodigal                                                             |
|         | INSDC ID            | CP002351                                                             |
| MIGS-33 | GenBank Date of Release| November 21, 2011                                                  |
| MIGS-34 | GOLD ID             | Gc01826                                                              |
| MIGS-35 | NCBI project ID     | 41517                                                               |
| MIGS-36 | Database: IMG       | 25035080007                                                         |
| MIGS-37 | Source material identifier | DSM 5069                                                            |
Genome annotation
Genes were identified using Prodigal [45] as part of the Oak Ridge National Laboratory genome annotation pipeline, followed by a round of manual curation using the JGI GenePRIMP pipeline [46]. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) non-redundant database, UniProt, TIGRFam, Pfam, PRIAM, KEGG, COG, and InterPro databases. These data sources were combined to assert a product description for each predicted protein. Additional gene prediction analysis and functional annotation was performed within the Integrated Microbial Genomes - Expert Review (IMG-ER) platform [47].

Genome properties
The genome consists of one circular chromosome of 2,039,943 bp length with a 40.3% G+C content (Table 3 and Figure 3). Of the 2,066 genes predicted, 2,015 were protein-coding genes, and 51 RNAs; 69 pseudogenes were also identified. The majority of the protein-coding genes (74.3%) 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.

Figure 3. Graphical map of the chromosome. From outside to the center: Genes on forward strand (colored by COG categories), Genes on reverse strand (colored by COG categories), RNA genes (tRNAs green, rRNAs red, other RNAs black), GC content (black), GC skew (purple/olive).
Table 3. Genome Statistics

| Attribute                                      | Value   | % of Total |
|------------------------------------------------|---------|------------|
| Genome size (bp)                               | 2,039,943 | 100.00%    |
| DNA coding region (bp)                         | 1,859,937 | 91.18%     |
| DNA G+C content (bp)                           | 822,588  | 40.32%     |
| Number of replicons                            | 1       |            |
| Extrachromosomal elements                      | 0       |            |
| Total genes                                    | 2,066   | 100.00%    |
| RNA genes                                      | 51      | 2.47%      |
| rRNA operons                                   | 1       |            |
| tRNA genes                                     | 46      | 2.23%      |
| Protein-coding genes                           | 2,015   | 97.53%     |
| Pseudo genes                                   | 69      | 3.34%      |
| Genes with function prediction (proteins)      | 1,535   | 74.30%     |
| Genes in paralog clusters                      | 912     | 44.14%     |
| Genes assigned to COGs                         | 1,719   | 83.20%     |
| Genes assigned Pfam domains                    | 1,704   | 82.48%     |
| Genes with signal peptides                     | 327     | 15.83%     |
| Genes with transmembrane helices               | 549     | 26.57%     |
| CRISPR repeats                                 | 7       |            |

Insights into the genome sequence

Because a number of complete genome sequences of type strains from the phylum has already been published, we conducted a phylogenomic analysis using the bioinformatics pipeline established in [48] and further modified as described in [18,49]. The resulting supermatrix comprised 1,889 genes and 582,906 characters before, 1,168 genes and 360,527 characters after cleaning with MARE. The selected model was PROTGAMMALGF; the resulting tree had a log likelihood of -3,783,776.37 and is shown in Figure 3. The best maximum-parsimony tree found had a length of 404,859 steps (not counting uninformative characters) and was topologically identical. The gene-content matrix comprised 3,267 characters and yielded best trees with a log likelihood of -13,904.74 and a parsimony score of 2,243, respectively. Bootstrapping support values from all four applied methods are shown in Figure 4 if larger then 60%. The phylogenomic trees disagree with the 16S rRNA tree (Fig. 1) in some respects. For instance, *Thermosipho* appears as a sister group of *Fervidobacterium*. Hence we assessed whether the 16S rRNA alignment described above, if reduced to the strains used in the phylogenomic analysis, is in significant conflict with the phylogenomic topology. Using the kind of constraint analysis mentioned above, search under the maximum-likelihood criterion yielded a best tree with a score of -5,425.82 if the search was unconstrained but a tree with a log likelihood of -5,436.37 under the constraint; this was not significantly worse in the SH test as implemented in RAxML (α = 0.05). Under maximum parsimony, the globally best trees had a score of 512, whereas the best constrained tree was 529 steps in length; this was significantly worse according to KH test implemented in PAUP* (p = 0.0148).

Currently there is neither evidence for a significant discrepancy between 16S rRNA and whole-genome data, nor a significant disagreement between 16S rRNA and the classification after *Thermococcoides shengiensis* was placed in *Kosmotoga* as *K. shengiensis* (see above). Nevertheless, as usual [36] the phylogenomic trees are much better resolved than the 16S rRNA phylogenies, and the *Kosmotoga-Thermococcoides* question could also be addressed in greater detail if the genomes of the type strains were available, as digital replacements of DNA-DNA hybridization have been implemented [23]. The classification of the group thus can only benefit from additional genome-sequenced type strains.
### Table 4. Number of genes associated with the general COG functional categories

| Code | Value | Percentage | Description                                                   |
|------|-------|------------|---------------------------------------------------------------|
| J    | 138   | 7.2%       | Translation, ribosomal structure and biogenesis              |
| A    | 0     | 0.0%       | RNA processing and modification                               |
| K    | 85    | 4.5%       | Transcription                                                |
| L    | 108   | 5.7%       | Replication, recombination and repair                         |
| B    | 2     | 0.1%       | Chromatin structure and dynamics                               |
| D    | 22    | 1.2%       | Cell cycle control, cell division, chromosome partitioning   |
| Y    | 0     | 0.0%       | Nuclear structure                                            |
| V    | 26    | 1.4%       | Defense mechanisms                                           |
| T    | 79    | 4.1%       | Signal transduction mechanisms                                |
| M    | 79    | 4.1%       | Cell wall/membrane biogenesis                                 |
| N    | 68    | 3.6%       | Cell motility                                                |
| Z    | 0     | 0.0%       | Cytoskeleton                                                 |
| W    | 0     | 0.0%       | Extracellular structures                                     |
| U    | 43    | 2.3%       | Intracellular trafficking and secretion, and vesicular transport |
| O    | 58    | 3.0%       | Posttranslational modification, protein turnover, chaperones  |
| C    | 128   | 6.7%       | Energy production and conversion                              |
| G    | 211   | 11.0%      | Carbohydrate transport and metabolism                        |
| E    | 201   | 10.5%      | Amino acid transport and metabolism                           |
| F    | 60    | 3.1%       | Nucleotide transport and metabolism                           |
| H    | 77    | 4.0%       | Coenzyme transport and metabolism                             |
| I    | 35    | 1.8%       | Lipid transport and metabolism                                |
| P    | 99    | 5.2%       | Inorganic ion transport and metabolism                        |
| Q    | 20    | 1.1%       | Secondary metabolites biosynthesis, transport and catabolism  |
| R    | 238   | 12.5%      | General function prediction only                              |
| S    | 134   | 7.0%       | Function unknown                                             |
| -    | 347   | 16.8%      | Not in COGs                                                  |

The *T. thermarum* genome has numerous differences from the other *Thermotoga* genomes, particularly with regard to cofactor metabolism. Some of these differences are shared with *T. lettingae*, which is more closely related to *T. thermarum* than the other *Thermotoga* species with sequenced genomes (Figs. 1 and 4). There appears to be a significant difference in Fe-S cluster synthesis between *T. thermarum* and the other *Thermotoga* species. Fe-S cluster synthesis requires at the minimum a cysteine desulfurase to produce sulfur and a scaffold protein for Fe-S clus-
ter assembly (reviewed in [50]). There are three Fe-S cluster biosynthesis pathways in bacteria: Nif, Isc, and Suf [51]. *T. maritima* uses the Suf system. It has an operon with sufCBDSU genes and another operon with a second copy of sufCB [51]. The SufS protein is a cysteine desulfurase. In *Bacillus subtilis*, which has a similar set of Suf proteins as *T. maritima*, the SufU protein has been shown to be a scaffold protein [52]. In *Escherichia coli*, which lacks the SufU protein, SufB is a scaffold protein, and SufC and SufD are required for iron acquisition [53]. In *E. coli* the Suf genes are expressed under iron starvation conditions [51]. *T. maritima*, therefore, may have two scaffold proteins, SufU and SufB. *T. thermarum* has a cluster of four genes (Theth_0902-0905) including two cysteine desulfurases and two proteins similar to SufU, but the SufBCD proteins are not present in the *T. thermarum* genome. Thus *T. thermarum* appears to encode a minimal set of Fe-S cluster synthesis proteins. It is possible that in *Thermotogales* and *Firmicutes* SufU is used as the scaffold protein if iron is plentiful, while SufBCD is required under low-iron conditions. *T. thermarum* may have access to more iron in its environment than other *Thermotoga* species. Interestingly, adjacent to the Fe-S cluster biosynthesis genes in *T. thermarum* is a transporter for which the closest characterized homolog is ZupT from *E. coli*, which transports iron and other divalent metals [54]. *T. lettingae* has similar Fe-S cluster synthesis genes as *T. thermarum* but also encodes the sufCB genes.

All of the *Thermotoga* species lack uroporphyrinogen synthesis and most of vitamin B12 synthesis, and the only enzyme of B12 metabolism common to all *Thermotoga* genomes is the adenosylcobalamin from cobalamin. However, *T. thermarum* contains several genes clustered together (Theth_1729-1737) involved in the later steps of cobalamin synthesis, suggesting that it can utilize precursors of cobalamin that the other *Thermotoga* species cannot utilize. Most of these genes are also found in *T. lettingae*. *T. thermarum* and *T. lettingae* are the only *Thermotoga* species to have genes for tungsten-dependent aldehyde:ferredoxin oxidoreductases (Theth_0853, Theth_1019). Theth_0853 has 68% amino acid identity to the formaldehyde:ferredoxin oxidoreductase of *Pyrococcus furiosus*, suggesting it was recently acquired. These enzymes use a bis-molybdenum cofactor with tungsten in place of molybdenum [55]. *T. thermarum* and *T. lettingae* are also the only *Thermotoga* species to have genes for tungsten transport (Theth_0538-540) and molybdopterin biosynthesis (Theth_0439-440, Theth_0535-536, Theth_1749). However, genes for molybdopterin synthase (moaD, moaE) could not be identified, suggesting they may have alternative genes for this step of the pathway. *T. thermarum* also has molybdenum cofactor guanyllyltransferase (Theth_0112) for production of molybdopterin guanine dinucleotide. Adjacent to this enzyme are a formate dehydrogenase accessory protein, a formate transporter pseudogene, and a molybdopterin dinucleotide-dependent formate dehydrogenase pseudogene. There are no other genes in *T. thermarum* with the molybdopterin dinucleotide binding domain (pfam01568) suggesting that molybdopterin dinucleotide synthesis is no longer necessary.

*T. thermarum* has fewer glycosyl hydrolases than the other *Thermotoga* species [56], but it has genes for transport and utilization of oligogalacturonides that are not present in the others. *T. thermarum* has an ABC transporter (Theth_0394-0396) similar to the oligogalacturonide ABC transporter from *Erwinia chrysanthemi* [57], while none of the other *Thermotoga* genomes contains genes similar to any of the known oligogalacturonide transporters. Close to the transporter is the kduI gene (Theth_0398) involved in oligogalacturonide degradation, which is also not found in other *Thermotoga* species. The transporter genes and kduI gene have 60-70% amino acid identity to genes from *Thermoanaerobacter*, suggesting recent acquisition from *Clostridia*. Other genes found only in *T. thermarum* and *T. lettingae* include enzymes for histidine degradation (Theth_0380, Theth_1683, Theth_0980) and serine degradation (Theth_1895-1896). *Thermotoga* species generally grow on a variety of carbohydrates, but the presence of these pathways suggests amino acids may be a carbon and energy source for some species. All *Thermotoga* species have genes for the Rnf complex, which couples an ion gradient to the transfer of electrons between NADH and ferredoxin. In addition *T. thermarum* and *T. lettingae* have genes for the NqrBCDEFG subunits of a sodium-translocating NADH:quinone dehydrogenase (Theth_1137-1141). They lack the NqrA subunit, which contains the quinone binding site [58], so the other participant in the reaction (besides NADH) is unknown.
Figure 4. Phylogenetic tree inferred from completely sequenced genomes of the ‘Thermotogae’ type strains. The tree was inferred from 360,527 aligned amino acid characters under the maximum likelihood (ML) criterion and rooted using midpoint rooting [11]. The branches are scaled in terms of the expected number of substitutions per site. Numbers above the branches are bootstrapping support values (if larger than 60%) from (i) maximum-likelihood supermatrix analysis; (ii) maximum-parsimony supermatrix analysis; (iii) maximum-likelihood gene-content analysis; (iv) maximum-parsimony gene-content analysis.

Acknowledgements
We would like to gratefully acknowledge the help of Maren Schröder for growing T. thermarum cultures and Evelyne-Marie Brambilla for DNA extraction and quality control (both of the DSMZ) as well as the valuable hints provided by T. Nunoura and J.P. Euzéby. This work was performed under the auspices of the US Department of Energy Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396, UT-Battelle and Oak Ridge National Laboratory under contract DE-AC05-00OR22725, as well as German Research Foundation (DFG) INST 599/1-2.

References
1. Windberger E, Huber R, Trincone A, Fricke H, Stetter KO. Thermotoga thermarum sp. nov. and Thermotoga neapolitana occurring in African continental solfataric springs. Arch Microbiol 1989; 151:506-512; http://dx.doi.org/10.1007/BF00454866
2. Euzéby JP. List of bacterial names with standing in nomenclature: a folder available on the Internet. Int J Syst Bacteriol 1997; 47:590-592; PubMed http://dx.doi.org/10.1099/00207713-47-2-590
3. Huber R, Langworthy TA, König H, Thomm M, Woese CR, Slytr UB, Stetter KO. Thermotoga maritima sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90°C. Arch Microbiol 1986; 144:324-333; http://dx.doi.org/10.1007/BF00409880
4. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compat-
ible with ARB. *Appl Environ Microbiol* 2006; 72:5069-5072; [PubMed](http://dx.doi.org/10.1128/AEM.03006-05)

5. Porter MF. An algorithm for suffix stripping. Program: electronic library and information systems 1980; 14:130-137.

6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; 215:403-410. [PubMed](http://dx.doi.org/10.1016/S0022-2836(05)80360-2)

7. Göker M, Cleland D, Saunders E, Lapidus A, Nolan M, Lucas S, Hammon N, Deshpande S, Cheng JF, Tapia R, et al. Complete genome sequence of *Isosphaera pallida* type strain (IS1B)³. *Stand Genomic Sci* 2011; 4:63-71. [PubMed](http://dx.doi.org/10.4056/sigs.1533840)

8. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. *Bioinformatics* 2002; 18:452-464; [PubMed](http://dx.doi.org/10.1093/bioinformatics/18.3.452)

9. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000; 17:540-552; [PubMed](http://dx.doi.org/10.1093/oxfordjournals.molbev.a026334)

10. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol* 2008; 57:758-771; [PubMed](http://dx.doi.org/10.1080/10635150802429642)

11. Hess PN, De Moraes Russo CA. An empirical test of the midpoint rooting method. *Biol J Linn Soc Lond* 2007; 92:669-674; [PubMed](http://dx.doi.org/10.1111/j.1095-8312.2007.00864.x)

12. Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A. How many bootstrap replicates are necessary? *Lect Notes Comput Sci* 2009; [5541]:184-200; [PubMed](http://dx.doi.org/10.1007/978-3-642-02008-7_13)

13. Swoford DL. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods*), Version 4.0 b10. Sinauer Associates, Sunderland, 2002.

14. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 2012; 40:D571-D579; [PubMed](http://dx.doi.org/10.1093/nar/gkr1100)

15. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, et al. Evidence for lateral gene transfer between *Archaea* and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 1999; 399:323-329; [PubMed](http://dx.doi.org/10.1038/20601)

16. Zhaxybayeva O, Swithers KS, Lapierre P, Fournier GP, Bickhart DM, DeBoy RT, Nelson KE, Nesbo CL, Doolittle WF, Gogarten JP, Noll KM. On the chimeric nature, thermophilic origin, and phylogenetic placement of the *Thermotogales*. *Proc Natl Acad Sci USA* 2009; 106:5865-5870; [PubMed](http://dx.doi.org/10.1073/pnas.0901260106)

17. Swithers KS, Dipippo JL, Bruce DC, Detter C, Tapia R, Han S, Goodwin LA, Han J, Woyke T, Pitluck S, et al. Genome Sequence of *Kosmotoga olearia* Strain TBF 19.5.1, a thermophilic bacterium with a wide growth temperature range, isolated from the Troll B oil platform in the North Sea. *J Bacteriol* 2011; 193:5566-5567; [PubMed](http://dx.doi.org/10.1128/JB.05828-11)

18. Abt B, Han C, Scheuner C, Lu M, Lapidus A, Nolan M, Lucas S, Hammon N, Deshpande S, Cheng JF, et al. Complete genome sequence of the termite hindgut bacterium *Spirochaeta coccosoides* type strain (SPN1)³, reclassification in the genus *Sphaerochaeta* as *Sphaerochaeta coccosoides* comb. nov. and emendations of the family *Spirochaetaceae* and the genus *Sphaerochaeta*. *Stand Genomic Sci* 2012; 6:194-209; [PubMed](http://dx.doi.org/10.4056/sigs.2796069)

19. DiPippo JL, Nesbo CL, Dahle H, Doolittle WF, Birklund NK, Noll KM. *Kosmotoga olearia* gen. nov., sp. nov., a thermophilic, anaerobic heterotroph isolated from an oil production fluid. *Int J Syst Evol Microbiol* 2009; 59:2991-3000; [PubMed](http://dx.doi.org/10.1099/ijs.0.008045-0)

20. Feng Y, Cheng L, Zhang X, Li X, Deng Y. Zhang Hui. *Thermococcoides shengliensis* gen. nov., sp. nov., a new member of the order *Thermotogales* isolated from oil-production fluid. *Int J Syst Evol Microbiol* 2010; 60:932-937; [PubMed](http://dx.doi.org/10.1099/ijs.0.013912-0)

21. Nunoura T, Hirai M, Imachi H, Miyazaki M, Makita H, Hirayama H, Fukushima Y, Yamamoto H, Takai K. *Kosmotoga arenicorallina* sp. nov. a thermophilic and obligately anaerobic heterotroph isolated from a shallow hydrothermal system occurring within a coral reef, southern part of the Yaeyama Archipelago, Japan, reclassification of *Thermococcoides shengliensis* as *Kosmotoga shengliensis* comb. nov., and emended descrip-
Thermotoga thermarum

...tion of the genus Kosmotoga. Arch Microbiol 2010; 192:811-819.; PubMed http://dx.doi.org/10.1007/s00203-010-0611-7

22. Euzéby JP. Validation List: List of new names and new combinations previously effectively, but not validly, published. Int J Syst Evol Microbiol 2011; 61:1-3.; PubMed http://dx.doi.org/10.1099/ijs.0.030445-0

23. Auch AF, von Jan M, Klenk HP, Göker M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci 2010; 2:117-134.; PubMed http://dx.doi.org/10.4056/sigs.531120

24. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 2008; 26:541-547.; PubMed http://dx.doi.org/10.1038/nbt1360

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

26. List Editor. Validation List no. 85. Validation of the publication of new names and new combinations previously effectively published outside the IJSEM. Int J Syst Evol Microbiol 2002; 52:685-690.; PubMed http://dx.doi.org/10.1099/ijs.0.02358-0

27. Reysenbach AL. Phylum B II. Thermotogae phyl. nov. In: Boone DR, Castenholz RW, Garrity GM (eds): Bergey's Manual of Systematic Bacteriology, second edition, vol. 1 (The Archaea and the deeply branching and phototrophic Bacteria), Springer-Verlag, New York, 2001, p. 369-387.

28. Reysenbach AL. Class I. Thermotogae class nov. In: Boone DR, Castenholz RW, Garrity GM (eds): Bergey's Manual of Systematic Bacteriology, second edition, vol. 1 (The Archaea and the deeply branching and phototrophic Bacteria), Springer-Verlag, New York, 2001, p. 369-370.

29. Reysenbach AL. Order I. Thermotogales ord. nov. In: Boone DR, Castenholz RW, Garrity GM (eds): Bergey's Manual of Systematic Bacteriology, second edition, vol. 1 (The Archaea and the deeply branching and phototrophic Bacteria), Springer-Verlag, New York, 2001, p. 369-370.

30. Reysenbach AL. Family I. Thermotogaceae fam. nov. In: Boone DR, Castenholz RW, Garrity GM (eds): Bergey's Manual of Systematic Bacteriology, second edition, vol. 1 (The Archaea and the deeply branching and phototrophic Bacteria), Springer-Verlag, New York, 2001, p. 370.

31. List Editor. Validation List no. 22. Validation of the publication of new names and new combinations previously effectively published outside the IJSB. Int J Syst Bacteriol 1986; 36:573-576.; http://dx.doi.org/10.1099/00207713-36-4-573

32. List Editor. Validation List no. 41. Validation of the publication of new names and new combinations previously effectively published outside the IJSB. Int J Syst Bacteriol 1992; 42:327-328.; http://dx.doi.org/10.1099/00207713-42-2-327

33. BAuA. 2010, Classification of bacteria and archaea in risk groups. http://www.baua.de TRBA 466, p. 237.

34. 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 http://dx.doi.org/10.1038/75556

35. Göker M, Klenk HP. Phylogeny-driven target selection for large-scale genome-sequencing (and other) projects. Stand Genomic Sci 2013; 8:360-374. PubMed http://dx.doi.org/10.4056/sigs.3446951

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

37. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, et al. A phylogeny-driven Genomic Encyclopaedia of Bacteria and Archaea. Nature 2009; 462:1056-1060.; PubMed http://dx.doi.org/10.1038/nature08656

38. List of growth media used at DSMZ: http://www.dsmz.de/catalogues/catalogue-microorganisms/culture-technology/list-of-media-for-microorganisms.html

39. Gemeinholzer B, Dröge G, Zetzsche H, Haszprunar G, Klenk HP, Güntsch A, Berendsohn WG, Wägele JW. The DNA Bank Network: the start from a German initiative Biopreserv Biobank 2011; 9:51-55. http://dx.doi.org/10.1089/bio.2010.0029

40. The DOE Joint Genome Institute. www.jgi.doe.gov

41. Phrap and Phred for Windows. MacOS, Linux, and Unix. www.phrap.com
42. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008; 18:821-829; PubMed http://dx.doi.org/10.1101/gr.074492.107

43. Han C, Chain P. Finishing repeat regions automatically with Dupfinisher. In: Proceedings of the 2006 international conference on bioinformatics & computational biology. Arabnia HR, Valafar H (eds), CSREA Press. June 26-29, 2006: 141-146.

44. Lapidus A, LaButti K, Foster B, Lowry S, Trong S, Goltsman E. POLISHER: An effective tool for using ultra short reads in microbial genome assembly and finishing. AGBT, Marco Island, FL, 2008.

45. Hyatt D, Chen GL, Locascio PF, Land ML, Lamer FW, Hauser LJ. Prodigal Prokaryotic Dynamic Programming Genefinding Algorithm. BMC Bioinformatics 2010; 11:119; PubMed http://dx.doi.org/10.1186/1471-2105-11-119

46. 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 http://dx.doi.org/10.1038/nmeth.1457

47. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Pati A, Huntemann M, et al. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res 2012; 40:D115-D122. PubMed http://dx.doi.org/10.1093/nar/gkr1044

48. Anderson I, Scheuner C, Göker M, Mavromatis K, Hooper SD, Porat I, Klenk HP, Ivanova N, Kyrpides N. Novel Insights into the Diversity of Catabolic Metabolism from Ten Haloarchaeal Genomes. PLoS ONE 2011; 6:e20237; PubMed http://dx.doi.org/10.1371/journal.pone.0020237

49. Göker M, Scheuner C, Klenk HP, Stielow JB, Menzel W. Codivergence of mycoviruses with their hosts. PLoS ONE 2011; 6:e22252; PubMed http://dx.doi.org/10.1371/journal.pone.0022252

50. Py B, Barras F. Building Fe-S proteins: bacterial strategies. Nat Rev Microbiol 2010; 8:436-446; PubMed http://dx.doi.org/10.1038/nrmicro2356

51. Johnson DC, Dean DR, Smith AD, Johnson MK. Structure, function, and formation of biological iron-sulfur clusters. Annu Rev Biochem 2005; 74:247-281; PubMed http://dx.doi.org/10.1146/annurev.biochem.74.082803.133518

52. Albrecht AG, Netz DJ, Miethke M, Pierik AJ, Burghaus O, Peuckert F, Lill R, Marahiel MA. SufU is an essential iron-sulfur cluster scaffold protein in Bacillus subtilis. J Bacteriol 2010; 192:1643-1651; PubMed http://dx.doi.org/10.1128/JB.01536-09

53. Saini A, Mapolelo DT, Chahal HK, Johnson MK, Outten FW, Suf D, Suf C. ATPase activity are required for iron acquisition during in vivo Fe-S cluster formation on SufB. Biochemistry 2010; 49:9402-9412; PubMed http://dx.doi.org/10.1021/bi1011546

54. Grass G, Franke S, Taudte N, Nies DH, Kucharski LM, Maguire ME, Rensing C. The metal permease ZupT from Escherichia coli is a transporter with a broad substrate spectrum. J Bacteriol 2005; 187:1604-1611; PubMed http://dx.doi.org/10.1128/JB.187.5.1604-1611.2005

55. Chan MK, Mukund S, Kletzin A, Adams MWW, Rees DC. Structure of a hyperthermophilic tungstopterin enzyme, aldehyde ferredoxin oxidoreductase. Science 1995; 267:1463-1469; PubMed http://dx.doi.org/10.1126/science.7878465

56. Carbohydrate-Active Enzymes Database. www.cazy.org.

57. Hugouvieux-Cotte-Pattat N, Reverchon S, Two transporters, TogT and TogMNAB, are responsible for oligogalacturonide uptake in Erwinia chrysanthemi 3937. Mol Microbiol 2001; 41:1125-1132; PubMed http://dx.doi.org/10.1046/j.1365-2958.2001.02565.x

58. Casutt MS, Nedielkov R, Wendelspiss S, Vossler S, Gerken U, Murai M, Miyoshi H, Möller HM, Steuber J. Localization of ubiquinone-8 in the Na+-pumping NADH:quinone oxidoreductase from Vibrio cholerae. J Biol Chem 2011; 286:40075-40082; PubMed http://dx.doi.org/10.1074/jbc.M111.224980