# Complete genome sequence of the thermophilic sulfate-reducing ocean bacterium *Thermodesulfatator indicus* type strain (CIR29812$^T$)

**Iain Anderson[1], Elizabeth Saunders[1,2], Alla Lapidus[1], Matt Nolan[1], Susan Lucas[1], Hope Tice[1], Tijana Glavina Del Rio[1], Jan-Fang Cheng[1], Cliff Han[1,2], Roxanne Tapia[1,2], Lynne A. Goodwin[1,2], Sam Pitluck[1], Konstantinos Liolios[1], Konstantinos Mavromatis[1], Ioanna Pagani[1], Natalia Ivanova[1], Natalia Mikhailova[1], Amrita Pati[1], Amy Chen[3], Krishna Palaniappan[3], Miriam Land[1,4], Loren Hauser[1,4], Cynthia D. Jeffries[1,4], Yun-juan Chang[1,4], Evelyne-Marie Brambilla[6], Manfred Rohde[5], Stefan Spring[6], Markus Göker[6], John C. Detter[1,2], Tanja Woyke[1], James Bristow[1], Jonathan A. Eisen[1,7], Victor Markowitz[3], Philip Hugenholtz[1,8], Nikos C. Kyrpides[1], Hans-Peter Klenk[6]***

**[1] DOE Joint Genome Institute, Walnut Creek, California, USA**
**[2] Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico, USA**
**[3] Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California, USA**
**[4] Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA**
**[5] HZI – Helmholtz Centre for Infection Research, Braunschweig, Germany**
**[6] Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, 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**

*__Corresponding author__: Hans-Peter Klenk

**Keywords**: strictly anaerobic, motile, Gram-negative, thermophilic, sulfate-reducing, chemolithoautotrophic, black smoker, *Thermodesulfobacteria*, *Thermodesulfobacteriaceae*, GEBA

*Thermodesulfatator indicus* Moussard *et al*. 2004 is a member of the Thermodesulfobacteriaceae, a family in the phylum Thermodesulfobacteria that is currently poorly characterized at the genome level. Members of this phylum are of interest because they represent a distinct, deep-branching, Gram-negative lineage. *T. indicus* is an anaerobic, thermophilic, chemolithoautotrophic sulfate reducer isolated from a deep-sea hydrothermal vent. Here we describe the features of this organism, together with the complete genome sequence, and annotation. The 2,322,224 bp long chromosome with its 2,233 protein-coding and 58 RNA genes is a part of the ***G**enomic **E**ncyclopedia of **B**acteria and **A**rchaea* project.

## Introduction

The genus *Thermodesulfatator* currently contains two species, both of which are anaerobic, thermophilic, chemolithoautotrophic sulfate reducers isolated from deep-sea hydrothermal vents [1,2]. Strain CIR29812$^T$ (= DSM 15286 = JCM 11887) is the type strain of the species Thermodesulfatator indicus [1]. The strain was isolated from a chimney fragment taken from a black smoker in the Kairai vent field, Central Indian Ridge [1]. The genus name was derived from a combination of the Greek term *thermos*, hot, and the Neo-Latin *desulfatator*, sulfate-reducer, meaning the thermophilic sulfate-reducer [1]; the species epithet was derived from the Latin adjective *indicus*, referring to the Indian Ocean, from where the strain was isolated [1]. The other species in this genus is T. atlanticus, which was isolated from the wall of a chimney at the

Rainbow vent field on the Mid-Atlantic Ridge [2]. The major difference between the two Thermodesulfatator species is that T. indicus is strictly chemolithoautotrophic, while T. atlanticus is able to utilize organic carbon sources [2]. Here we present a summary classification and a set of features for T. indicus CIR29812$^T$, together with the description of the genomic sequencing and annotation.

## Classification and features

A representative genomic 16S rRNA sequence of *T. indicus* CIR29812$^T$ was compared using NCBI BLAST [3,4] under default settings (e.g., considering only the high-scoring segment pairs (HSPs) from the best 250 hits) with the most recent release of the Greengenes database [5] and the relative frequencies of taxa and keywords (reduced to their stem [6]) were determined, weighted by BLAST scores. The most frequently occurring genera were *Desulfovibrio* (22.5%), *Thermodesulfatator* (22.0%), *Thermodesulfobacterium* (16.9%), *Methylococcus* (10.9%) and *Thermodesulforhabdus* (5.7%) (38 hits in total). Regarding the two hits to sequences from members of the species, the average identity within HSPs was 99.9%, whereas the average coverage by HSPs was 95.8%. Among all other species, the one yielding the highest score was "*Geothermobacterium ferrireducens*" (AF411013), which corresponded to an identity of 90.1% and an HSP coverage of 64.7%. (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 AJ874315 ('continuous enrichment hydrothermal black chimney clone 850'), which showed an identity of 96.7% and an HSP coverage of 93.9%. The most frequently occurring keywords within the labels of all environmental samples which yielded hits were 'spring' (6.2%), 'microbi' (4.8%), 'hot' (4.2%), 'nation, park' (2.7%) and 'yellowston' (2.6%) (212 hits in total). These keywords fit reasonably well to the habitat of a thermophilic sulfate-reducer. Environmental samples which yielded hits of a higher score than the highest scoring species were not found.

Figure 1 shows the phylogenetic neighborhood of *T. indicus* in a 16S rRNA based tree. The sequences of the two 16S rRNA gene copies in the genome differ from each other by two nucleotides, and differ by up to four nucleotides from the previously published 16S rRNA sequence (AF393376).

*T. indicus* cells are Gram-negative rods with a length of 0.8-1.0 µm and a width of 0.4-0.5 µm [1]. An electron micrograph of *T. indicus* is shown in Figure 2. Cells are motile with a single polar flagellum and can be found separately or in groups of two or three cells [1]. The temperature range for growth is 55-80°C with an optimum at 70°C [1]. The salinity range is 10-35 g/L NaCl, with an optimum of 25 g/L NaCl [1]. The pH range is 6.0-6.7 with 6.25 as the optimum [1]. *T. indicus* is strictly anaerobic and strictly chemolithoautotrophic, growing with $H_2$ as electron donor, sulfate as electron acceptor, and $CO_2$ as the carbon source [1]. Some organic compounds stimulated growth [1]. Ammonium, nitrate, peptone and tryptone could serve as nitrogen sources [1].

### Chemotaxonomy

The major respiratory quinone found in *T. indicus* is menaquinone with seven isoprene subunits (MK-7) [1]. The major phospholipids are phosphatidylinositol and phosphatidyl-ethanolamine. Phosphatidylglycerol and three unidentified phospholipids are present in lesser amounts [1]. The major fatty acids are $C_{18:0}$ and $C_{18:1}$, and hydroxylated fatty acids are also present [1]. *T. indicus* was found to be sensitive to tetracycline, ampicillin, chloramphenicol, and rifampicin, and resistant to penicillin, kanamycin, and streptomycin [1].

## Genome sequencing and annotation

### Genome project history

This organism was selected for sequencing on the basis of its phylogenetic position [23], and is part of the ***G**enomic **E**ncyclopedia of **B**acteria and **A**rchaea* project [24]. The genome project is deposited in the Genomes On Line Database [13] 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.

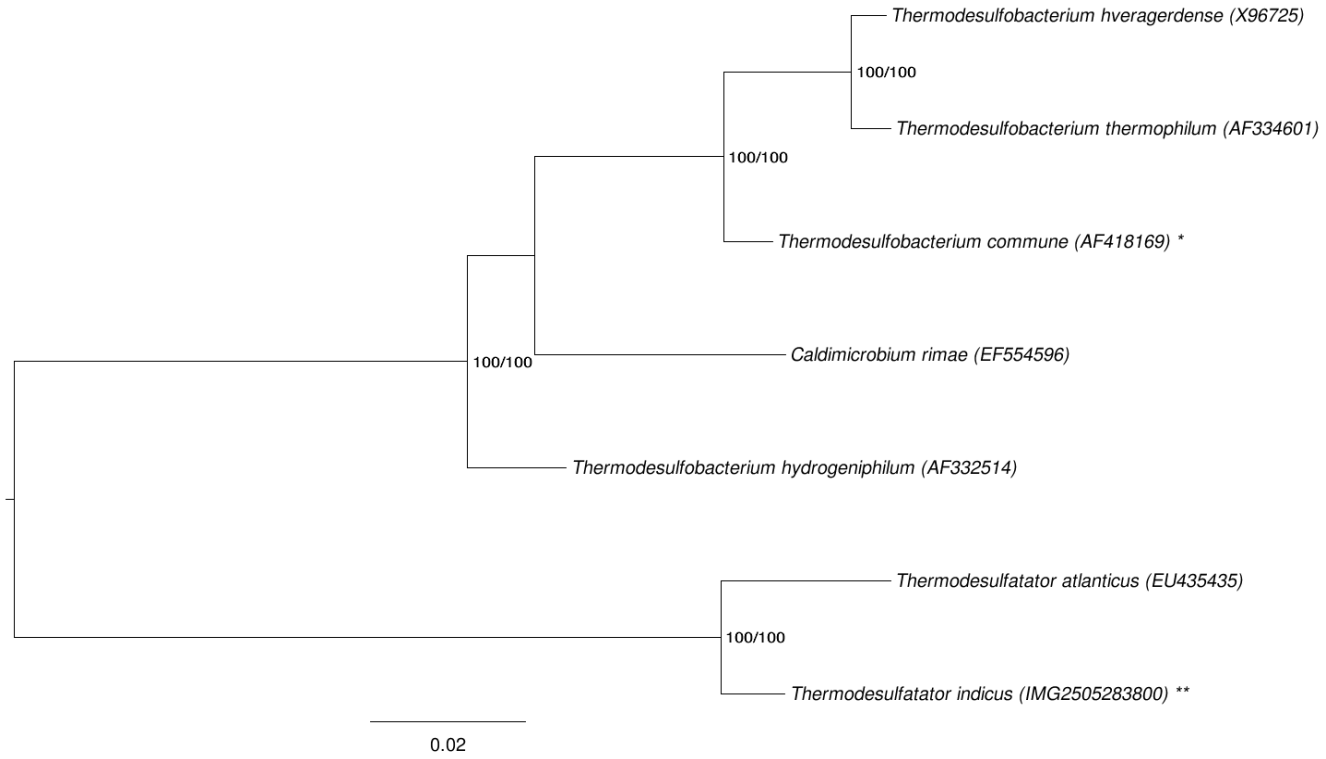


**Figure 1.** Phylogenetic tree highlighting the position of *T. indicus* relative to the type strains of the other species within the phylum *Thermodesulfobacteria*. The tree was inferred from 1,475 aligned characters [7,8] of the 16S rRNA gene sequence under the maximum likelihood (ML) criterion [9]. Rooting was done initially using the midpoint method [10] and then checked for its agreement with the current classification (Table 1). The branches are scaled in terms of the expected number of substitutions per site. Numbers adjacent to the branches are support values from 1,000 ML bootstrap replicates [11] (left) and from 1,000 maximum-parsimony bootstrap replicates [12] (right) if larger than 60%. Lineages with type strain genome sequencing projects registered in GOLD [13] are labeled with one asterisk, those also listed as 'Complete and Published' with two asterisks.

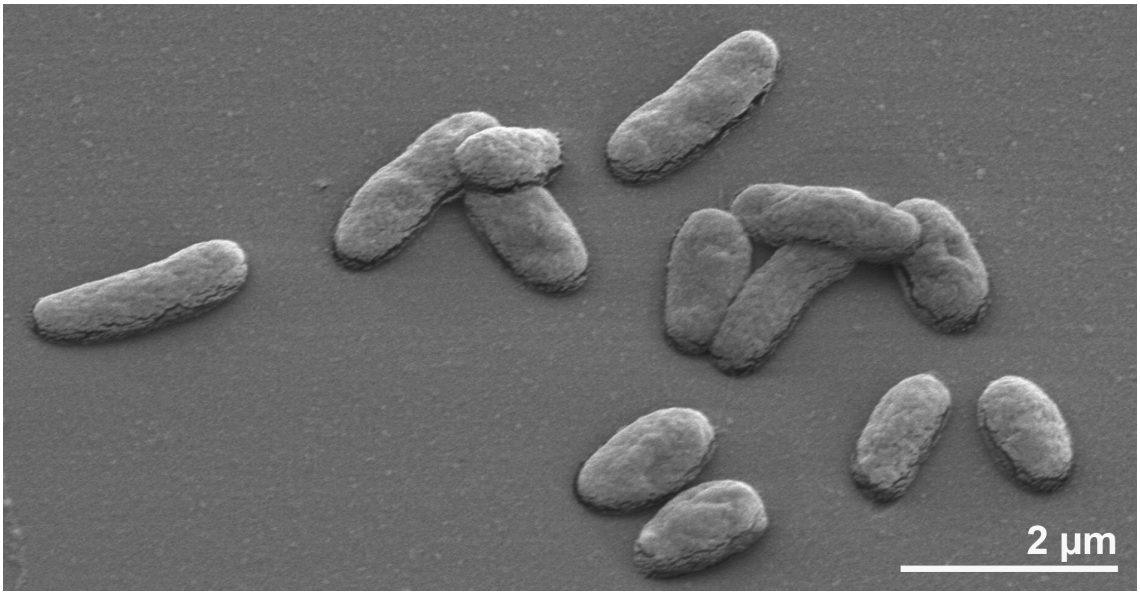


**Figure 2.** Scanning electron micrograph of T. indicus CIR29812$^T$

**Table 1.** Classification and general features of *T. indicus* CIR29812, according to the MIGS recommendations [14].

| MIGS ID | Property | Term | Evidence code |
|---|---|---|---|
| | Current classification | Domain *Bacteria* | TAS [15] |
| | | Phylum *Thermodesulfobacteria* | TAS [16] |
| | | Class *Thermodesulfobacteria* | TAS [17,18] |
| | | Order *Thermodesulfobacteriales* | TAS [17,19] |
| | | Family *Thermodesulfobacteriaceae* | TAS [17,20] |
| | | Genus *Thermodesulfatator* | TAS [1] |
| | | Species *Thermodesulfatator indicus* | TAS [1] |
| | | Type-strain CIR29812 | TAS [1] |
| | Gram stain | negative | TAS [1] |
| | Cell shape | small rods | TAS [1] |
| | Motility | motile *via* single polar flagellum | TAS [1] |
| | Sporulation | non-sporulating | TAS [1] |
| | Temperature range | thermophile, 55-80°C | TAS [1] |
| | Optimum temperature | 70°C | TAS [1] |
| | Salinity | 10-35 g NaCl per liter, optimum at 25 g | TAS [1] |
| MIGS-22 | Oxygen requirement | strictly anaerobic | TAS [1] |
| | Carbon source | $CO_2$ | TAS [1] |
| | Energy metabolism | chemolithoautotrophic | TAS [1] |
| MIGS-6 | Habitat | deep-sea hydrothermal vent field | TAS [1] |
| MIGS-15 | Biotic relationship | free living | TAS [1] |
| MIGS-14 | Pathogenicity | none | NAS |
| | Biosafety level | 1 | TAS [21] |
| MIGS-23.1 | Isolation | chimney fragment from black smoker | TAS [1] |
| MIGS-4 | Geographic location | Kairai vent field, Central Indian Ridge | TAS [1] |
| MIGS-5 | Sample collection time | April 2001 | TAS [1] |
| MIGS-4.1 | Latitude | -25.317 | TAS [1] |
| MIGS-4.2 | Longitude | 70.033 | TAS [1] |
| MIGS-4.3 | Depth | 2,420 m | TAS [1] |
| MIGS-4.4 | Altitude | -2,420 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. 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 [22].

**Table 2.** Genome sequencing project information

| MIGS ID | Property | Term |
|---|---|---|
| MIGS-31 | Finishing quality | Finished |
| MIGS-28 | Libraries used | Four genomic libraries: one 454 pyrosequence standard library, two 454 PE libraries (7 and 11 kb insert sizes), one Illumina library |
| MIGS-29 | Sequencing platforms | Illumina GAii, 454 GS FLX Titanium |
| MIGS-31.2 | Sequencing coverage | 183.8 × Illumina; 126.8 × pyrosequence |
| MIGS-30 | Assemblers | Newbler version 2.3-PreRelease-6-30-2009-gcc-3.4.6, Velvet version 1.0.13, phrap |
| MIGS-32 | Gene calling method | Prodigal |
| | INSDC ID | CP002683 |
| | GenBank Date of Release | November 21, 2011 |
| | GOLD ID | Gc01827 |
| | NCBI project ID | 40057 |
| | Database: IMG-GEBA | 2505119042 |
| MIGS-13 | Source material identifier | DSM15286 |
| | Project relevance | Tree of Life, GEBA, Bioenergy |

## Growth conditions and DNA isolation

*T. indicus* strain CIR29812$^T$, DSM 15286, was grown anaerobically in DSMZ medium 383 (*Desulfobacterium* medium) [25] at 70°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/LALM for cell lysis as described in Wu *et al*. 2009 [24]. DNA is available through the DNA Bank Network [26].

## 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 [27]. Pyrosequencing reads were assembled using the Newbler assembler (Roche). The initial Newbler assembly consisting of 49 contigs in one scaffold was converted into a phrap [28] assembly by making fake reads from the consensus, to collect the read pairs in the 454 paired end library. Illumina GAii sequencing data (427.0 Mb) was assembled with Velvet [29] 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 298.3 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 [28] 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 (C. Han, unpublished), Dupfinisher [30], 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 95 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 (A. Lapidus, unpublished). 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 310.6 × coverage of the genome. The final assembly contained 759,221 pyrosequence and 11,861,111 Illumina reads.

### Genome annotation

Genes were identified using Prodigal [31] as part of the Oak Ridge National Laboratory genome annotation pipeline, followed by a round of manual curation using the JGI GenePRIMP pipeline [32]. 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. Non-coding genes and miscellaneous features were predicted using tRNAscan-SE [33], RNAMMer [34], Rfam [35], TMHMM [36], and SignalP [37].

## Genome properties

The genome consists of a 2,322,224 bp long circular chromosome with a 42.4% G+C content (Table 3 and Figure 3). Of the 2,291 genes predicted, 2,233 were protein-coding genes, and 58 RNAs; 38 pseudogenes were also identified. The majority of the protein-coding genes (73.2%) 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.

**Table 3.** Genome Statistics

| Attribute | Value | % of Total[a] |
|---|---|---|
| Genome size (bp) | 2,322,224 | 100.00% |
| DNA coding region (bp) | 2,101,503 | 90.50% |
| DNA G+C content (bp) | 985,214 | 42.43% |
| Number of replicons | 1 | |
| Extrachromosomal elements | 0 | |
| Total genes | 2,291 | |
| RNA genes | 58 | |
| rRNA operons | 2 | |
| tRNA genes | 49 | |
| Protein-coding genes | 2,233 | 100.00% |
| Pseudo genes | 38 | 1.70% |
| Genes with function prediction (proteins) | 1,678 | 75.15% |
| Genes in paralog clusters | 959 | 42.95% |
| Genes assigned to COGs | 1,845 | 82.62% |
| Genes assigned Pfam domains | 917 | 41.07% |
| Genes with signal peptides | 351 | 15.72% |
| Genes with transmembrane helices | 499 | 22.35% |
| CRISPR repeats | 3 | |

a) The total is based on either the size of the genome in base pairs or the total number of protein coding genes in the annotated genome.

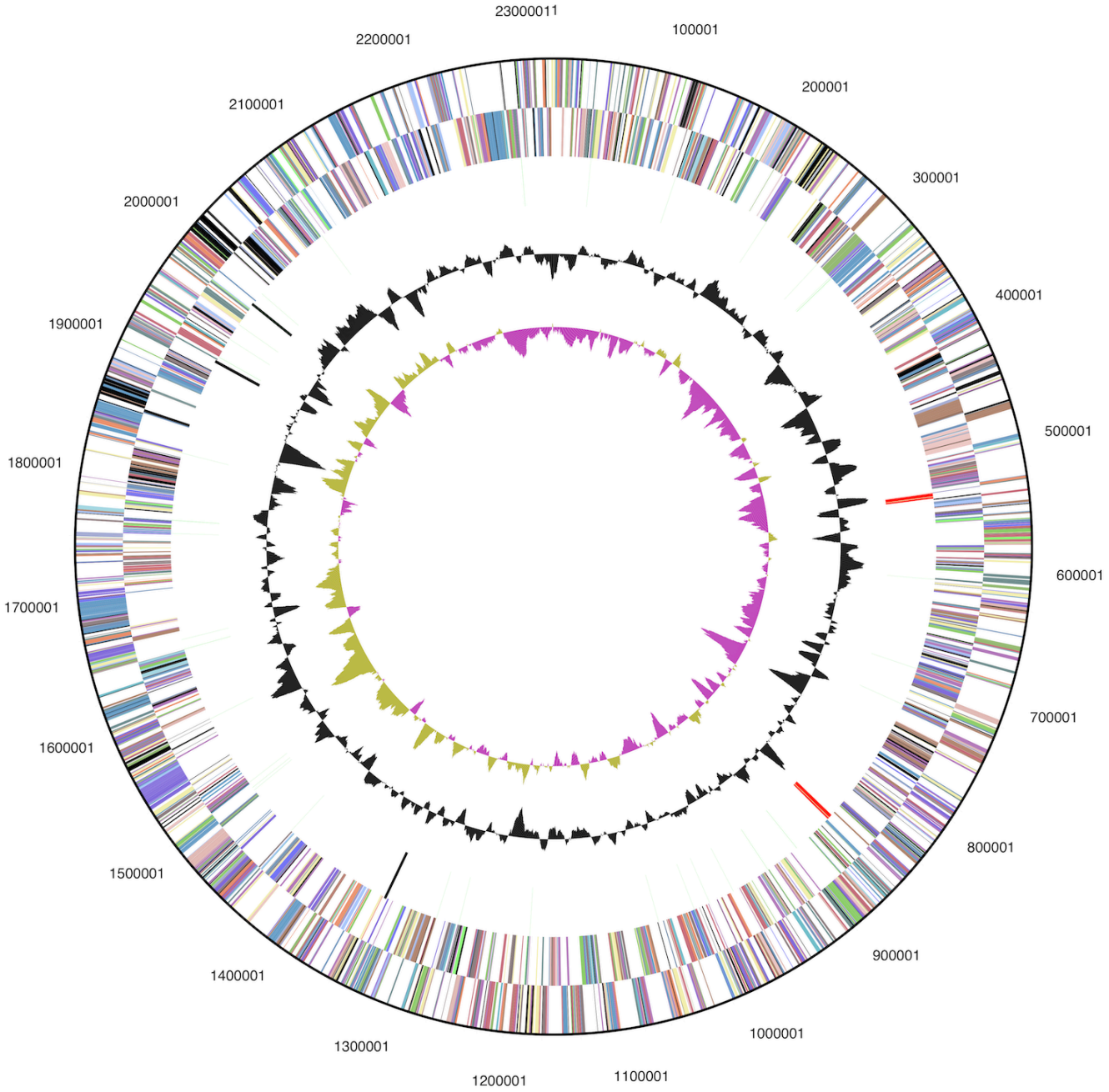


**Figure 3.** Graphical map of the chromosome. From outside to the center: Genes on forward strand (colored by COG categories), Genes on reverse strand (colored by COG categories), RNA genes (tRNAs green, rRNAs red, other RNAs black), GC content, GC skew.

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

| Code | value | %age[a] | Description |
|---|---|---|---|
| J | 155 | 6.9 | Translation, ribosomal structure and biogenesis |
| A | 2 | 0.1 | RNA processing and modification |
| K | 72 | 3.2 | Transcription |
| L | 144 | 6.4 | Replication, recombination and repair |
| B | 2 | 0.1 | Chromatin structure and dynamics |
| D | 35 | 1.6 | Cell cycle control, cell division, chromosome partitioning |
| Y | 0 | 0.0 | Nuclear structure |
| V | 17 | 0.8 | Defense mechanisms |
| T | 114 | 5.1 | Signal transduction mechanisms |
| M | 129 | 5.8 | Cell wall/membrane biogenesis |
| N | 84 | 3.8 | Cell motility |
| Z | 0 | 0.0 | Cytoskeleton |
| W | 0 | 0.0 | Extracellular structures |
| U | 83 | 3.7 | Intracellular trafficking and secretion, and vesicular transport |
| O | 88 | 3.9 | Posttranslational modification, protein turnover, chaperones |
| C | 151 | 6.8 | Energy production and conversion |
| G | 67 | 3.0 | Carbohydrate transport and metabolism |
| E | 166 | 7.4 | Amino acid transport and metabolism |
| F | 58 | 2.6 | Nucleotide transport and metabolism |
| H | 123 | 5.5 | Coenzyme transport and metabolism |
| I | 39 | 1.7 | Lipid transport and metabolism |
| P | 82 | 3.7 | Inorganic ion transport and metabolism |
| Q | 19 | 0.9 | Secondary metabolites biosynthesis, transport and catabolism |
| R | 225 | 10.1 | General function prediction only |
| S | 152 | 6.8 | Function unknown |
| - | 388 | 17.4 | Not in COGs |

a) The percentage is based on the total number of protein coding genes in the annotated genome.

## Acknowledgements

We would like to gratefully acknowledge the help of Maren Schröder (DSMZ) for growing *T. indicus* cultures. 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. Moussard H, L'Haridon S, Tindall BJ, Banta A, Schumann P, Stackebrandt E, Reysenbach AL, Jeanthon C. *Thermodesulfatator indicus* gen. nov., sp. nov., a novel thermophilic chemolithoautotrophic sulfate-reducing bacterium isolated from the Central Indian Ridge. *Int J Syst Evol Microbiol* 2004; **54**:227-233. PubMed http://dx.doi.org/10.1099/ijs.0.02669-0

2. Alain K, Postec A, Grinsard E, Lesongeur F, Prieur D, Godfroy A. *Thermodesulfatator atlanticus* sp. nov., a thermophilic, chemolithoautotrophic, sulfate-reducing bacterium isolated from a Mid-Atlantic Ridge hydrothermal vent. *Int J Syst Evol Microbiol* 2010; **60**:33-38. PubMed http://dx.doi.org/10.1099/ijs.0.009449-0

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

4. Korf I, Yandell M, Bedell J. BLAST, O'Reilly, Sebastopol, 2003.

5. 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 compatible with ARB. *Appl Environ Microbiol* 2006; **72**:5069-5072. PubMed http://dx.doi.org/10.1128/AEM.03006-05

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

7. 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

8. 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

9. 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

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

11. 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. http://dx.doi.org/10.1007/978-3-642-02008-7_13

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

13. Liolios K, Chen IM, Mavromatis K, Tavernarakis N, 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 http://dx.doi.org/10.1093/nar/gkp848

14. 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

15. 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 http://dx.doi.org/10.1073/pnas.87.12.4576

16. Garrity GM, Holt JG. Phylum BIII. *Thermodesulfobacteria* phy. nov. *In:* Garrity GM, Boone DR, Castenholz RW (*eds*), Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 1, Springer, New York, 2001, p. 389.

17. List Editor. Validation List no. 85. Validation of 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

18. Hatchikian EC, Ollivier B, Garcia JL. Class I. *Thermodesulfobacteria* class. nov. *In*: Garrity GM, Boone DR, Castenholz RW (*eds*), Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 1, Springer, New York, 2001, p. 389.

19. Hatchikian EC, Ollivier B, Garcia JL. Order I. *Thermodesulfobacteriales* ord. nov. *In*: Garrity GM, Boone DR, Castenholz RW (*eds*), Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 1, Springer, New York, 2001, p. 389.

20. Hatchikian EC, Ollivier B, Garcia JL. Family I. *Thermodesulfobacteriaceae* fam. nov. *In:* Garrity GM, Boone DR, Castenholz RW (*eds*), Bergey's Manual of Systematic Bacteriology, Second edition, Volume 1, Springer, New York, 2001, p. 390.

21. BAuA. 2010, Classification of bacteria and archaea in risk groups. http://www.baua.de TRBA 466, p. 235.

22. 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

23. 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

24. 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

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

26. 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

27. The DOE Joint Genome Institute. http://www.jgi.doe.gov

28. Phrap and Phred for Windows. MacOS, Linux, and Unix. http://www.phrap.com

29. 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

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

31. 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 http://dx.doi.org/10.1186/1471-2105-11-119

32. 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

33. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997; **25**:955-964. PubMed

34. Lagesen K, Hallin PF, Rødland E, Stærfeldt HH, Rognes T, Ussery DW. RNammer: consistent annotation of rRNA genes in genomic sequences. *Nucleic Acids Res* 2007; **35**:3100-3108. PubMed http://dx.doi.org/10.1093/nar/gkm160

35. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. *Nucleic Acids Res* 2003; **31**:439-441. PubMed http://dx.doi.org/10.1093/nar/gkg006

36. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol* 2001; **305**:567-580. PubMed http://dx.doi.org/10.1006/jmbi.2000.4315

37. Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 2004; **340**:783-795. PubMed http://dx.doi.org/10.1016/j.jmb.2004.05.028