Complete genome sequence of *Thioalkalivibrio paradoxus* type strain ARh 1\(^T\), an obligately chemolithoautotrophic haloalkaliphilic sulfur-oxidizing bacterium isolated from a Kenyan soda lake

Tom Berben\(^1\), Dimitry Y. Sorokin\(^2,3\), Natalia Ivanova\(^4\), Amrita Pati\(^4\), Nikos Kyrpides\(^4,5\), Lynne A. Goodwin\(^4\), Tanja Woyke\(^4\) and Gerard Muyzer\(^1\)*

**Abstract**

*Thioalkalivibrio paradoxus* strain ARh 1\(^T\) is a chemolithoautotrophic, non-motile, Gram-negative bacterium belonging to the *Gammaproteobacteria* that was isolated from samples of haloalkaline soda lakes. It derives energy from the oxidation of reduced sulfur compounds and is notable for its ability to grow on thiocyanate as its sole source of electrons, sulfur and nitrogen. The full genome consists of 3,756,729 bp and comprises 3,500 protein-coding and 57 RNA-coding genes. This organism was sequenced as part of the community science program at the DOE Joint Genome Institute.

**Keywords:** Haloalkaliphilic, Soda lakes, Sulfur-oxidizing bacteria, Thiocyanate

**Introduction**

Soda lakes are characterized by a high and stable pH (>9) due to the presence of molar concentrations of soluble carbonates as the dominant anions and a moderate to high salinity [1]. They are found in arid zones in many parts of the world, for example, in the Kulunda Steppe in Russia, North-Eastern China, the Rift Valley in Africa and the arid regions of California and Nevada (e.g., Mono Lake, Big Soda Lake). Despite their (extremely) haloalkaline character, these environments harbor a rich microbial diversity that is responsible for driving highly active biogeochemical cycles [2], of which the sulfur cycle is the most active. Our current research focuses on a group of chemolithoautotrophic sulfur-oxidizing bacteria that belong to the genus *Thioalkalivibrio* in the class *Gammaproteobacteria*. These organisms are of interest because of their role in the oxidative part of the sulfur cycle in soda lakes [3] and their application in the sustainable removal of sulfur from wastewater and gas streams [4]. To better understand the success of this group of organisms, we have sequenced the genomes of a large number of *Thioalkalivibrio* isolates. Here we present the genome sequence of *T. paradoxus* ARh 1\(^T\) (= DSM 13531\(^T\) = JCM 11367\(^T\)).

**Organism information**

**Classification and features**

This obligate aerobic and haloalkaliphilic strain, which was isolated from a mixed sample of sediments from Kenyan soda lakes, is a non-motile coccoid rod forming intracellular sulfur as an obligate intermediate during oxidation of thiosulfate and thiocyanate (Fig. 1). It is an obligate chemolithoautotroph, capable of using a variety of reduced, inorganic sulfur compounds, including sulfide, thiosulfate and polysulfide, as electron donor for carbon fixation. It can also oxidize CS\(_2\) (carbon disulfide). Of special interest is its ability to grow with thiocyanate (NCS\(^–\)) as electron donor, with a relatively high...
growth rate of 0.08–0.1 h\(^{-1}\) in continuous culture, compared to 0.01–0.015 h\(^{-1}\) for growth on thiosulfate \[5\]. Phylogenetic analysis based on 16S rRNA sequences shows that \textit{T. paradoxus} is closely related to \textit{Thioalkalivibrio nitratireducens} ALEN \text{2}\(\text{T}\) (Fig. 2). An overview of basic features of the organism is provided in Table 1.

**Genome sequencing information**

**Genome project history**

In order to better understand the diversity within the genus \textit{Thioalkalivibrio}, as well as their biogeochemical role in soda lakes, a large number of isolates (approximately 70) was sequenced at the Joint Genome Institute. The full genome of the type strain of \textit{Thioalkalivibrio paradoxus} presented here contains 3.8 million basepairs. Sequencing was performed at the JGI under project number 401912 and the sequence data was subsequently released in Genbank on December 31, 2013. A project overview is provided in Table 2.

**Growth conditions and genomic DNA preparation**

A buffer using sodium carbonate and bicarbonate, with a total salt concentration of 0.6 M Na\(^+\), was used for cultivation of the organism; the energy source was thiosulfate (40 mM). After harvesting, the cells were stored at \(-80\) °C for further processing. Genomic DNA was extracted using a standard chloroform-phenol-isooamyl alcohol mixture, followed by ethanol precipitation. After vacuum drying, the pellet was dissolved in water and the quantity and quality of the DNA determined using the JGI-provided Mass Standard Kit.

**Genome sequencing and assembly**

The draft genome of \textit{Thioalkalivibrio paradoxus} ARh \text{1}\(\text{T}\) was generated at the DOE Joint Genome Institute (JGI) using Illumina data \[6\]. For this genome, we constructed and sequenced an Illumina short-insert paired-end library with an average insert size of 270 bp which generated 18,589,770 reads and an Illumina long-insert paired-end library with an average insert size of 7,058.67 \(\pm\) 3247.54 bp which generated 20,051,794 reads totaling 5,796 Mbp of Illumina data (unpublished, Feng Chen). All general aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov/. The initial draft assembly contained 83 contigs in 11 scaffolds. The initial draft data was assembled with ALLPATHS \[7\], version 39750, and the consensus was computationally shredded into 10 Kbp overlapping fake reads (shreds). The Illumina draft data was also assembled with Velvet, version 1.1.05 \[8\], and the consensus sequences were computationally shredded into 1.5 Kbp overlapping fake reads (shreds). The Illumina draft data was assembled again with Velvet using the shreds from the first Velvet assembly to guide the next assembly. The consensus from the second Velvet assembly was shredded into 1.5 Kbp overlapping fake reads. The fake reads from the ALLPATHS assembly and both Velvet assemblies and a subset of the Illumina CLIP paired-end reads were assembled using parallel phrap, version 4.24 (High Performance Software, LLC). Possible mis-assemblies were corrected with manual editing in Consed \[9–11\]. Gap closure was accomplished using repeat resolution software (Wei Gu, unpublished), and sequencing of bridging PCR fragments with Sanger and/or PacBio (unpublished, Cliff Han) technologies. A total of 50 additional sequencing reactions were completed to close gaps and to raise the quality of the final sequence. The size of the genome is 3.8 Mb and the final assembly is based on 5,796 Mbp of Illumina draft data, which provides an average 1,486X coverage of the genome.

**Genome annotation**

The assembled sequence was annotated using the JGI prokaryotic annotation pipeline \[12\] and was further reviewed.
using the Integrated Microbial Genomes - Expert Review (IMG-ER) platform [13]. Genes were identified using Prodigal [14], followed by manual curation using GenePRIMP [15]. Predicted CDSs were translated and used to search the NCBI non-redundant, UniProt, TIGRFam, Pfam, KEGG, COG and InterPro databases. The tRNAscanSE tool [16] was used to detect tRNA genes and ribosomal RNA genes were detected using models constructed from SILVA [17]. Other RNA genes were predicted using Rfam profiles in Infernal [18]. CRISPR elements were detected using CRT [19] and PILER-CR [20]. Further annotation was performed using the Integrated Microbial Genomics (IMG) platform [21].

### Genome properties

The finished genome with a G+C percentage of 66.06 % comprises a single chromosome of approximately 3.8 Mb (Fig. 3). There are 3557 genes of which 3,500 are protein-coding genes (a summary of genome properties is shown in Table 3).

| MIGS ID | Property | Term |
|---------|----------|------|
| MIGS 28 | Libraries used | Illumina |
| MIGS 29 | Sequencing platforms | Illumina HiSeq 2000 |
| MIGS 31.2 | Fold coverage | 1,486X |
| MIGS 30 | Assemblers | Velvet [8], ALLPATHS R39750 [7] |
| MIGS 32 | Gene calling method | Prodigal [14], GenePRIMP [15] |
| MIGS 13 | Source Material Identifier | DSM 13531 |
| MIGS 14 | Project relevance | Biotechnology |

### Table 1 Classification and general features of *Thioalkalivibrio paradoxus* ARh 1 T [24]

| MIGS ID | Property | Term |
|---------|----------|------|
| MIGS ID | Property | Term |
| MIGS 31 | Finishing quality | Finished |
| MIGS-28 | Libraries used | Illumina |
| MIGS 29 | Sequencing platforms | Illumina HiSeq 2000 |
| MIGS 31.2 | Fold coverage | 1,486X |
| MIGS 30 | Assemblers | Velvet [8], ALLPATHS R39750 [7] |
| MIGS 32 | Gene calling method | Prodigal [14], GenePRIMP [15] |
| MIGS 13 | Source Material Identifier | DSM 13531 |
| MIGS 14 | Project relevance | Biotechnology |

### Table 2 Project information

| MIGS ID | Property | Term |
|---------|----------|------|
| MIGS 31 | Finishing quality | Finished |
| MIGS-28 | Libraries used | Illumina |
| MIGS 29 | Sequencing platforms | Illumina HiSeq 2000 |
| MIGS 31.2 | Fold coverage | 1,486X |
| MIGS 30 | Assemblers | Velvet [8], ALLPATHS R39750 [7] |
| MIGS 32 | Gene calling method | Prodigal [14], GenePRIMP [15] |
| MIGS 13 | Source Material Identifier | DSM 13531 |
| MIGS 14 | Project relevance | Biotechnology |

### Table 3 Genome statistics

| Attribute | Value | % of Total |
|-----------|-------|------------|
| Genome size (bp) | 3,756,729 | 100 |
| DNA coding (bp) | 3,305,445 | 87.99 |
| DNA G+C (bp) | 2,500,004 | 66.55 |
| Total genes | 3,557 | 100 |
| Protein coding genes | 3,500 | 98.40 |
| RNA genes | 57 | 1.60 |
| Pseudo genes | 124 | 3.49 |
| Genes in internal clusters | 176 | 3.46 |
| Genes with function prediction | 2,739 | 77.00 |
| Genes assigned to COGs | 2,317 | 65.14 |
| Genes with Pfam domains | 2,835 | 79.70 |
| Genes with signal peptides | 271 | 7.62 |
| Genes with transmembrane helices | 841 | 23.64 |
| CRISPR repeats | 8 | 2.36 |

*Evidence codes - IDA: Inferred from Direct Assay; 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 [32]
Fig. 2 Phylogenetic tree, based on 16S rRNA sequences, of *Thioalkalivibrio* and various members of the Ectothiorhodospiraceae family. ARB [22] was used for tree construction and MEGA6 [23] for the bootstrap analysis. Alphaproteobacteria were used as the outgroup and pruned from the finished tree.

Fig. 3 Genome map of *Thioalkalivibrio paradoxus* ARh 1. From outer to inner ring: genes on the forward strand; genes on the reverse strand; RNA genes (tRNA: green; rRNA: red; other: black); GC content and GC skew.
in Table 3). Approximately two-thirds of the protein coding genes could be assigned to a COG functional category (Table 4).

**Conclusions**

The availability of high-quality genomic sequences of the type strains of *Thioalkalivibrio*, the dominant genus of sulfur-oxidizing bacteria in soda lakes, is an invaluable tool for gaining a more complete understanding of the biogeochemistry of these extreme environments. Additionally, this information may provide new insights into the exact mechanisms of adaptation these bacteria have evolved to not only survive, but thrive in this habitat. Finally, the genome may contain clues that will help improve the existing biotechnological applications of this organism in bioremediation.

**Competing interests**

The authors declare that they have no competing interests.

**Authors’ contributions**

TB drafted this manuscript, with GM and DS providing critical review and discussion. DS was responsible for cultivation and DNA extraction. Sequencing and annotation were performed at the JGI by NI, AP, NK, LAG and TW. All authors approve of the final version.

**Acknowledgements**

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Tom Berben and Gerard Muyzer are supported by ERC Advanced Grant PARASOL (No. 322551). Dimitry Sorokin is supported by RFFR Grant 13-04-00049.

**Author details**

1. Microbial Systems Ecology, Department of Aquatic Microbiology, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, The Netherlands. 2. Winogradsky Institute of Microbiology, RAS, Moscow, Russia. 3. Department of Biotechnology, Delft University of Technology, Delft, The Netherlands. 4. Joint Genome Institute, Walnut Creek, California, USA. 5. Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia.

**Received:** 7 October 2015 **Accepted:** 9 November 2015 **Published online:** 19 November 2015

**References**

1. Kempe S, Kazmierczak J. Soda Lakes. In: Reitner J, Thiel V, editors. Encyclopedia of Geobiology. Netherlands: Springer; 2011.

2. Sorokin DY, Berben T, Melton ED, Overman L, Vavourakis CD, Muyzer G. Microbial diversity and biogeochemical cycling in soda lakes. Extremophiles. 2014;18:791–809.

3. Sorokin DY, Banciu H, Robertson LA, Kuenen JG, Muyzer G. Halophilic and halokalkaliphilic sulfur-oxidizing bacteria from hypersaline habitats and soda lakes. In: Rosenberg E, editor. The Prokaryotes - Prokaryotic Physiology and Biochemistry. Berlin-Heidelberg: Springer; 2013. p. 530–51.

**Table 4 Number of genes associated with the 25 general COG functional categories**

| Code | Value | Percent | Description |
|------|-------|---------|-------------|
| J    | 204   | 7.95    | Translation, ribosomal structure and biogenesis |
| A    | 2     | 0.08    | RNA processing and modification |
| K    | 103   | 4.02    | Transcription |
| L    | 96    | 3.74    | Replication, recombination and repair |
| B    | 1     | 0.04    | Chromatin structure and dynamics |
| D    | 32    | 1.25    | Cell cycle control, Cell division, chromosome partitioning |
| V    | 116   | 4.52    | Defense mechanisms |
| T    | 119   | 4.64    | Signal transduction mechanisms |
| M    | 201   | 7.84    | Cell wall/membrane biogenesis |
| N    | 34    | 1.33    | Cell motility |
| U    | 51    | 1.99    | Intracellular trafficking and secretion |
| O    | 158   | 6.16    | Posttranslational modification, protein turnover, chaperones |
| C    | 228   | 8.89    | Energy production and conversion |
| G    | 91    | 3.55    | Carbohydrate transport and metabolism |
| E    | 162   | 6.32    | Amino acid transport and metabolism |
| F    | 61    | 2.38    | Nucleotide transport and metabolism |
| H    | 150   | 5.85    | Coenzyme transport and metabolism |
| I    | 95    | 3.70    | Lipid transport and metabolism |
| P    | 178   | 6.94    | Inorganic ion transport and metabolism |
| Q    | 36    | 1.40    | Secondary metabolites biosynthesis, transport and catabolism |
| R    | 237   | 9.24    | General function prediction only |
| S    | 146   | 5.69    | Function unknown |
| -    | 1,240 | 34.86   | Not in COGs |

The total is based on the total number of protein coding genes in the genome.
4. Sorokin DY, van den Bosch PLF, Abbas B, Janssen AJH, Muyzer G. Microbiological analysis of the population of extremely halalkaliphilic sulfur-oxidizing bacteria dominating in lab-scale sulfide-removing bioreactors. Appl Microbiol Biotechnol. 2008;80:965–75.

5. Sorokin DY, Tourova TP, Lysenko AM, Mityushina LL, Kuenen JG. Thioalkalivibrio thiocyanatoxidans sp. nov. and Thioalkalivibrio porosudis sp. nov., novel alkaliphilic, obligately autotrophic, sulfur-oxidizing bacteria capable of growth on thiocyanate, from soda lakes. Int J Syst Evol Microbiol. 2002;52:657–64.

6. Bennett S. Solexa Ltd. Pharmacogenomics. 2004;5:433–8.

7. Gnerre S, MacCallum I, Przybylski D, Ribeiro F, Burton J, Walker B, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 2011;108:1513–8.

8. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9.

9. Ewing B, Hiller L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 1998;8:175–85.

10. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998;8:186–94.

11. Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. Genome Res. 1998;8:195–202.

12. Mavromatis K, Ivanova NN, Chen IM, Szeto E, Markowitz VM, Kyrpides NC. The DOE-JGI standard operating procedure for the annotations of microbial genomes. Stand Genomic Sci. 2009;1:63–7.

13. Markowitz VM, Ivanova NN, Chen IM, Chu K, Kyrpides NC. IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics. 2009;25:2271–8.

14. Hyatt D, Chen G, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119–30.

15. Pati A, Ivanova NN, Mikhailova N, Ovchinnikova G, Hooper SD, Lykidis A, et al. GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. Nat Methods. 2010;7:455–2.

16. Lowe TM, Eddy SR. TRNAScan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nuc Acids Res. 1997;25:955–64.

17. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig WG, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nuc Acids Res. 2007;35:7188–96.

18. Navrotsky EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009;25:1335–7.

19. Bland C, Ramsey TL, Sabre F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics. 2007;8:209.

20. Edgar RC. PILETER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics. 2007;8:18.

21. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Pillay M, et al. IMG 4 version of the integrated microbial genomes comparative analysis system. Nucleic Acids Res. 2014;42:D560–7.

22. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, et al. ARB: a software environment for sequence data. Nucl Acids Res. 2004;32:1363–71.

23. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol. 2013;30:2725–9.

24. Field D, Gantzi G, Gray T, Morrison N, Selengut J, Sterk P, et al. Towards a richer description of our complete collection of genomes and metagenomes “Minimum Information about a Genome Sequence ” (MIGS) specification. Nat Biotechnol. 2008;26:541–7.

25. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 1990;87:4576–9.

26. Garrity GM, Bell JA, Lilburn T, Egan P. Reassignment of the Genus Ectothiorhodospira from the Class Chromatiales to a new Class, Ectothiorhodospiraceae fam. nov., and emended description of the Chromatiales Bavendamm 1924. Int J Syst Evol Microbiol. 1984;34:338–9.

27. Garrity GM, Bell JA, Lilburn T, Egan P. Reassignment of the Genus Ectothiorhodospira from the Class Chromatiales to a new Class, Ectothiorhodospiraceae fam. nov., and emended description of the Chromatiales Bavendamm 1924. Int J Syst Evol Microbiol. 1984;34:338–9.

28. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig WG, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nuc Acids Res. 2007;35:7188–96.

29. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009;25:1335–7.

30. Bland C, Ramsey TL, Sabre F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics. 2007;8:209.

31. Edgar RC. PILETER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics. 2007;8:18.

32. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.