Complete genome sequence of the moderately thermophilic mineral-sulfide-oxidizing firmicute *Sulfobacillus acidophilus* type strain (NAL^T)

Iain Anderson[1], Olga Chertkov[1,2], Amy Chen[3], Elizabeth Saunders[1,2], Alla Lapidus[1], Matt Nolan[1], Susan Lucas[1], Nancy Hammon[1], Shweta Deshpande[1], Jan-Fang Cheng[1], Cliff Han[1,2], Roxanne Tapia[1,2], Lynne A. Goodwin[1,2], Sam Pitluck[1], Konstantinos Liolios[1], Ioanna Pagani[1], Natalia Ivanova[1], Natalia Mikhailova[1], Amrita Pati[1], Krishna Palaniappan[3], Miriam Land[1,4], Chongle Pan[1,4], Manfred Rohde[5], Rüdiger Pukall[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]*, and Konstantinos Mavromatis[1]

[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
aerobic, motile, Gram-positive, acidophilic, moderately thermophilic, sulfide- and iron-oxidizing, biomining, autotrophic, mixotrophic, soil, *insertis sedis, Clostridiales,* GEBA

Abstract
Sulfobacillus acidophilus Norris *et al.* 1996 is a member of the genus *Sulfobacillus* which comprises five species of the order *Clostridiales*. *Sulfobacillus* species are of interest for comparison to other sulfur and iron oxidizers and also have biomining applications. This is the first completed genome sequence of a type strain of the genus *Sulfobacillus*, and the second published genome of a member of the species *S. acidophilus*. The genome, which consists of one chromosome and one plasmid with a total size of 3,557,831 bp, harbors 3,626 protein-coding and 69 RNA genes, and is a part of the ***G**enomic **E**ncyclopedia of **B**acteria and **A**rchaea* project.

Introduction
The genus *Sulfobacillus* currently consists of five species [1], all of which are mildly thermophilic or thermotolerant acidophiles [2]. They grow mixotrophically by oxidizing ferrous iron, sulfur, and mineral sulfides in the presence of yeast extract or other organic compounds [3]. Some can also grow autotrophically [2,3]. The strains that have been tested are capable of anaerobic growth using Fe^{+3} as electron acceptor [2,4]. The genus *Sulfobacillus*, along with the genus *Thermoaerobacter*, has only tentatively been assigned to a family, "Clostridiales family XVII *incertae sedis"*. This group may form a deep branch within the phylum *Firmicutes* or may constitute a new phylum [5]. Strain NAL^T (= DSM

10332 = ATCC 700253) is the type strain of the species *Sulfobacillus acidophilus*. The genus name was derived from the Latin words 'sulfur' and 'bacillus' meaning 'small sulphur-oxidizing rod' [6]. The species epithet is derived from the Neo-Latin words 'acidum', acid, and 'philus', loving, meaning acid-loving [3]. The first genome from a member of the species *S. acidophilus*, strain TPY from a hydrothermal vent in the Pacific Ocean, was recently sequenced by Li *et al.* [7]. Here we present a summary classification and a set of features for *S. acidophilum* strain NAL^T, together with the description of the complete genomic sequencing and annotation.

Classification and features

A representative genomic 16S rRNA sequence of *S. acidophilus* NAL^T was compared using NCBI BLAST [8,9] 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 [10] and the relative frequencies of taxa and keywords (reduced to their stem [11]) were determined, weighted by BLAST scores. The most frequently occurring genera were *Sulfobacillus* (81.9%), *Thermaerobacter* (8.0%), *Laceyella* (2.8%), ′Gloeobacter′ (2.1%) and ′Synechococcus′ (2.0%) (76 hits in total). Regarding the six hits to sequences from members of the species, the average identity within HSPs was 98.9%, whereas the average coverage by HSPs was 97.2%. Regarding the 23 hits to sequences from other members of the genus, the average identity within HSPs was 93.1%, whereas the average coverage by HSPs was 81.2%. Among all other species, the one yielding the highest score was 'Sulfobacillus yellowstonensis' (AY007665), which corresponded to an identity of 99.4% and an HSP coverage of 97.0%. (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 HQ730681 ('Microbial Anaerobic Sediments Tinto River: Natural Acid and Heavy Metals Content extreme acid clone SN1 2009 12D'), which showed an identity of 94.5% and an HSP coverage of 99.0%. The most frequently occurring keywords within the labels of all environmental samples which yielded hits were 'acid' (4.8%), 'soil' (4.5%), 'hydrotherm' (3.7%), 'microbi' (3.7%) and 'mine' (3.0%) (172 hits in total). These keywords correspond well to the environment from which strain NAL^T was isolated. Environmental samples which yielded hits of a higher score than the highest scoring species were not found.

Figure 1 shows the phylogenetic neighborhood of *S. acidophilus* NAL^T in a 16S rRNA based tree. The sequences of the five 16S rRNA gene copies in the genome differ from each other by up to eight nucleotides, and differ by up to four nucleotides from the previously published 16S rRNA sequence (AB089842), which contains two ambiguous base calls.

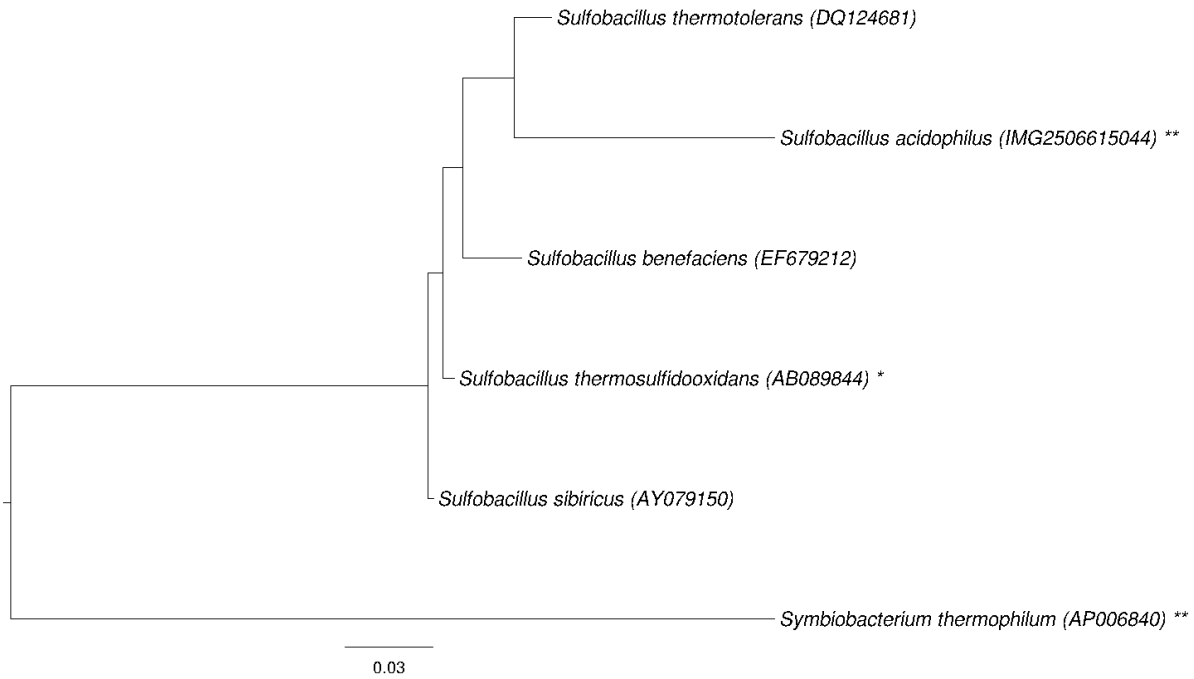

Figure 1. Phylogenetic tree highlighting the position of *S. acidophilus* relative to the type strains of the other species within the genus *Sulfobacillus*. The tree was inferred from 1,422 aligned characters [12,13] of the 16S rRNA gene sequence under the maximum likelihood (ML) criterion [14]. The comparatively closely related genus *Symbiobacterium* [51] was included for rooting the tree. The branches are scaled in terms of the expected number of substitutions per site. Numbers adjacent to the branches, if any, are support values from 1,000 ML bootstrap replicates [16] (left) and from 1,000 maximum-parsimony bootstrap replicates [17] (right) if larger than 60% (i.e., there were none). Lineages with type strain genome sequencing projects registered in GOLD [18] are labeled with one asterisk, those also listed as 'Complete and Published' with two asterisks [15].

Cells of *S. acidophilus* NAL^T are rods 3.0-5.0 µm in length and 0.5-0.8 µm in width (Figure 2) [3]. Cells are Gram-positive and form spherical endospores [3]. Flagella were not observed [3]. Strain NAL^T was found to grow between 28°C and 62°C with an optimum at 48°C [19]. The upper and lower temperatures for growth were not determined but were predicted to be 10°C and 62°C [19]. The pH range for growth was 1.6-2.3 with an optimum at 1.8 [19]. Three strains of *S. acidophilus* have been found to be facultative anaerobes able to use Fe^{+3} as an electron acceptor under anaerobic conditions [4]; but strain NAL^T was not tested in this study. Strain NAL^T can grow autotrophically or mixotrophically by oxidizing Fe^{+2}, sulfur, or mineral sulfides or heterotrophically on yeast extract [3]. *S. acidophilus* and other sulfobacilli have potential applications in biomining. Strain NAL^T increased the leaching of numerous mineral sulfides [19]; however its sensitivity to low concentrations of metals may limit its usefulness in biomining [19]. *S. acidophilus* NAL^T was chosen for sequencing based on its phylogenetic and physiological properties and is part of the Genomic Encyclopedia of *Bacteria* and *Archaea* (GEBA) project [20].

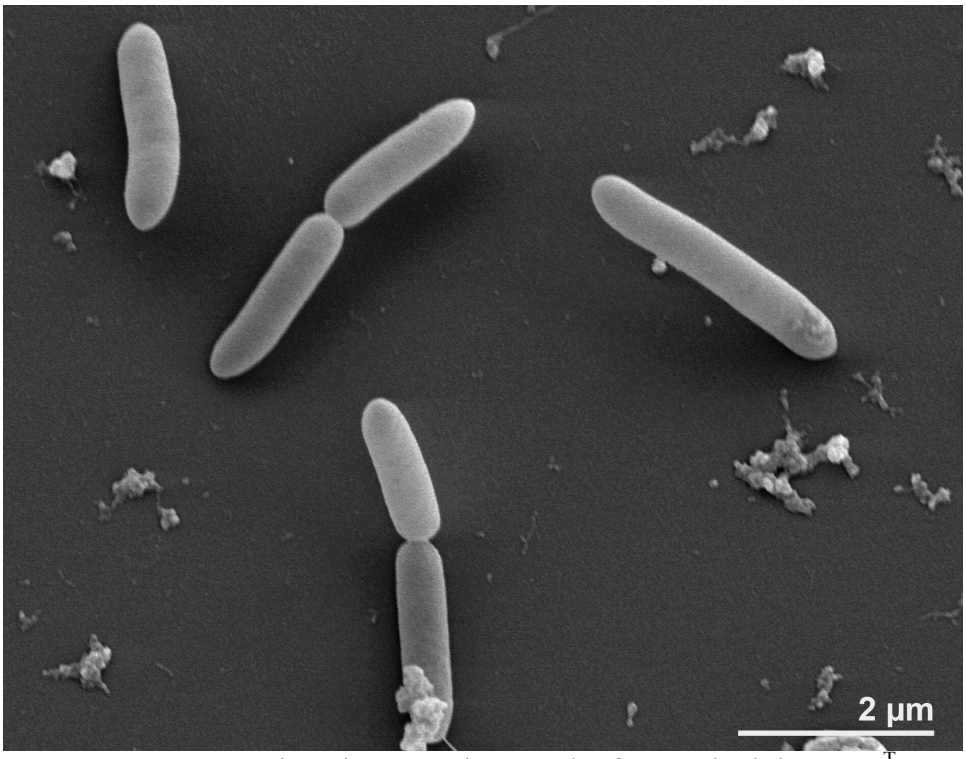

Figure 2. Scanning electron micrograph of *S. acidophilus* NAL^T

Chemotaxonomy

No data is available with respect to chemotaxonomy of *S. acidophilus* strain NAL^T.

Table 1. Classification and general features of *S. acidophilus* NAL^T in accordance with the MIGS recommendations [21] and the NamesforLife database.

MIGS ID	Property	Term	Evidence code
	Current classification	Domain *Bacteria*	TAS [23]
		Phylum "*Firmicutes*"	TAS [24-26]
		Class *Clostridia*	TAS [27,28]
		Order *Clostridiales*	TAS [29,30]
		Family "XVII *incertae sedis*"	TAS [5]*
		Genus *Sulfobacillus*	TAS [31,32]
		Species *Sulfobacillus acidophilus*	TAS [3,33]
		Type strain NAL	TAS [3]
	Gram stain	positive	TAS [3]
	Cell shape	rods	TAS [3]
	Motility	non-motile	NAS
	Sporulation	spherical endospores	TAS [3]
	Temperature range	not reported	
	Optimum temperature	48°C	TAS [19]
	Salinity	not reported	
MIGS-22	Oxygen requirement	facultative anaerobe	TAS [4]
	Carbon source	CO_2, organic compounds	TAS [3]
	Energy metabolism	autotrophic, mixotrophic, heterotrophic	TAS [3]
MIGS-6	Habitat	acidic sulfidic and sulfurous sites	TAS [19]
MIGS-15	Biotic relationship	free-living	TAS [3]
MIGS-14	Pathogenicity	none	NAS
	Biosafety level	1	TAS [34]
	Isolation	coal spoil heap	TAS [3]

MIGS-4	Geographic location	Alvecote, North Warwickshire, UK	TAS [3]
MIGS-5	Sample collection time	1988	TAS [3]
MIGS-4.1	Latitude	52.638	TAS [3]
MIGS-4.2	Longitude	-1.641	TAS [3]
MIGS-4.3	Depth	not reported	
MIGS-4.4	Altitude	not reported	

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). These evidence codes are from of the Gene Ontology project [35].
* not *Alicyclobacillaceae* [32] as described in NamesforLife [22]; see Figure 2 and [51].

Genome sequencing and annotation

Genome project history
This organism was selected for sequencing on the basis of its phylogenetic position [36], and is part of the ***G**enomic **E**ncyclopedia of **B**acteria and **A**rchaea* project [20]. The genome project is deposited in the Genomes OnLine Database [18] 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.

Table 2. Genome sequencing project information

MIGS ID	Property	Term
MIGS-31	Finishing quality	Finished
MIGS-28	Libraries used	Four genomic libraries: two 454 pyrosequence standard library, one 454 PE library (6 kb and 10 kb insert size), one Illumina library
MIGS-29	Sequencing platforms	Illumina GAii, 454 GS FLX Titanium
MIGS-31.2	Sequencing coverage	168.4 x Illumina; 51.2 x pyrosequence
MIGS-30	Assemblers	Newbler version 2.3-PreRelease-6/30/2009, Velvet 1.0.13, phrap version SPS - 4.24
MIGS-32	Gene calling method	Prodigal 1.4, GenePRIMP
	INSDC ID	CP003179 (chromosome) CP003180 (plasmid, unnamed)
	Genbank Date of Release	December 14, 2011
	GOLD ID	Gc02053
	NCBI project ID	40777
	Database: IMG-GEBA	2506520015
MIGS-13	Source material identifier	DSM 10332
	Project relevance	Tree of Life, GEBA

Growth conditions and DNA isolation
S. acidophilus strain NAL^T, DSM 10332, was grown in DSMZ medium 709 (*Acidomicrobium* medium) [37] at 45°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 [20]. DNA is available through the DNA Bank Network [38].

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 [39]. Pyrosequencing reads were assembled using the Newbler assembler (Roche). The initial Newbler assembly consisting of 104 contigs in tree scaffolds was converted into a phrap [40] assembly by making fake reads from the consensus, to collect the read pairs in the 454 paired end library. Illumina GAii sequencing data (599.7 Mb) were assembled with Velvet [41] 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 143.7 Mb of 454 draft data and all of the 454 paired-end data. Newbler parameters were -consed -a 50 -l 350 -g -m -ml 20. The Phred/Phrap/Consed software package [40] 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 [42], 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 640 additional reactions and eight shatter libraries 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 the software Polisher developed at JGI [43]. 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 219.6 × coverage of the genome. The final assembly contained 612,059 pyrosequence and 16,626,072 Illumina reads.

Genome annotation

Genes were identified using Prodigal [44] as part of the Oak Ridge National Laboratory genome annotation pipeline, followed by a round of manual curation using the JGI GenePRIMP pipeline [45]. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant database, UniProt, TIGR-Fam, Pfam, PRIAM, KEGG, COG, and InterPro databases. Additional gene prediction analysis and functional annotation was performed within the Integrated Microbial Genomes - Expert Review (IMG-ER) platform [46].

Genome properties

The genome consists of one circular chromosome of 3,472,898 bp and one circular plasmid of 84,933 bp length with an overall G+C content of 56.8% (Table 3 and Figures 3 and 4). Based on coverage of 454 paired ends, the plasmid may be inserted into the chromosome in about half of the population. Of the 3,695 genes predicted, 3,626 are protein-coding genes, and 69 are RNAs; 155 pseudogenes were also identified. The majority of the protein-coding genes (68.3%) were assigned with 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)	3,557,831	100.00%
DNA coding region (bp)	3,106,298	87.31%
DNA G+C content (bp)	2,019,235	56.75%

Number of replicons	2	
Extrachromosomal elements	1	
Total genes	3,695	
RNA genes	69	
rRNA operons	5	
Protein-coding genes	3,626	100.00%
Pseudo genes	155	4.27%
Genes with function prediction	2,475	68.26%
Genes in paralog clusters	1,896	52.29%
Genes assigned to COGs	2,740	75.57%
Genes assigned Pfam domains	413	11.39%
Genes with signal peptides	652	17.98%
Genes with transmembrane helices	910	25.10%
CRISPR repeats	2	

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.

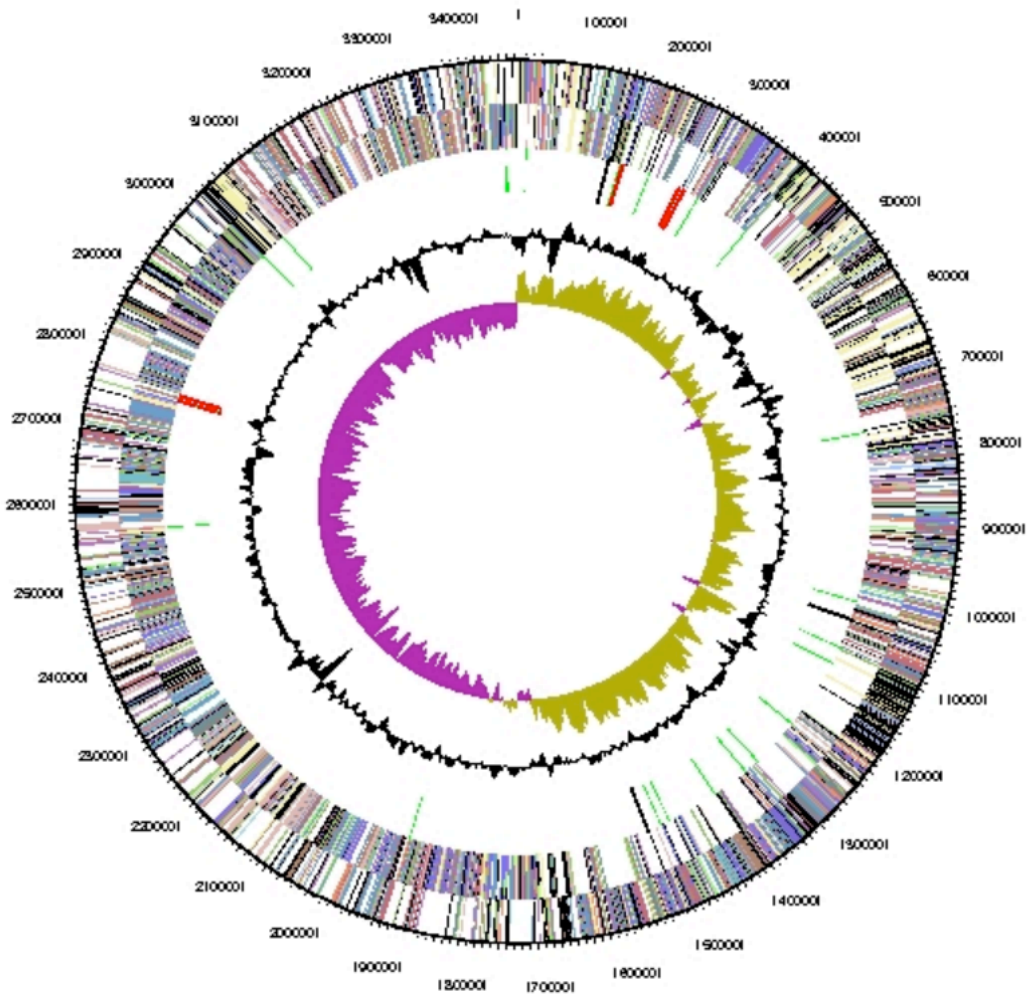

Figure 3. Graphical map of the chromosome. From outside to the center: Genes on forward strand (colored by COG categories), Genes on reverse strand (colored by COG categories), RNA genes (tRNAs green, rRNAs red, other RNAs black), GC content, GC skew.

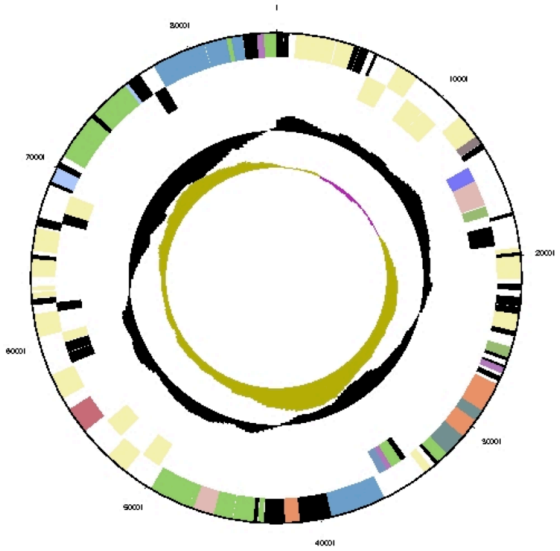

Figure 4. Graphical map of the plasmid. From outside to the center: Genes on forward strand (colored by COG categories), Genes on reverse strand (colored by COG categories), RNA genes (tRNAs green, rRNAs red, other RNAs black), GC content, GC skew.

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

Code	value	%age[a]	Description
J	149	4.1	Translation, ribosomal structure and biogenesis
A	0	0.0	RNA processing and modification
K	188	5.2	Transcription
L	269	7.4	Replication, recombination and repair
B	1	0.0	Chromatin structure and dynamics
D	26	0.7	Cell cycle control, cell division, chromosome partitioning
Y	0	0.0	Nuclear structure
V	34	0.9	Defense mechanisms
T	111	3.1	Signal transduction mechanisms
M	149	4.1	Cell wall/membrane/envelope biogenesis
N	47	1.3	Cell motility
Z	0	0.0	Cytoskeleton
W	0	0.0	Extracellular structures
U	62	1.7	Intracellular trafficking, secretion, and vesicular transport
O	129	3.6	Posttranslational modification, protein turnover, chaperones
C	244	6.7	Energy production and conversion
G	215	5.9	Carbohydrate transport and metabolism
E	257	7.1	Amino acid transport and metabolism
F	89	2.5	Nucleotide transport and metabolism
H	153	4.2	Coenzyme transport and metabolism
I	130	3.6	Lipid transport and metabolism

P	121	3.3	Inorganic ion transport and metabolism
Q	81	2.2	Secondary metabolites biosynthesis, transport and catabolism
R	326	9.0	General function prediction only
S	239	6.6	Function unknown
-	886	24.4	Not in COGs

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

Insights into the genome sequence

Comparative genomics

While the sequencing of the genome described in this paper was underway, Li *et al.* from the Third Institute of Oceanography, Xiamen, China published the complete genome sequence of strain TPY [7]. The two genomes differ in size by less than 7,000 bp. We here take the opportunity to compare the completed genome sequences from these two stains, NAL^T and TPY, both belonging to *S. acidophilus*. While the biological material for the here described genome from the type stain, NAL^T, is publicly available from the DSMZ open collection for postgenomic analyses, no source of the biological material (MIGS-13 criterion, see Table 2) of strain TPY was provided in [7].

To estimate the overall similarity between the genomes of strains NAL^T and TPY (Genbank accession number: CP002901), the GGDC-Genome-to-Genome Distance Calculator [47,48] was used. The system calculates the distances by comparing the genomes to obtain HSPs (high-scoring segment pairs) and interfering distances from the set of formulas (1, HSP length / total length; 2, identities / HSP length; 3, identities / total length). The comparison of the genomes of strains NAL^T and TPY revealed that 99.65% of the average of both genome lengths are covered with HSPs. The identity within these HSPs was 99.01%, whereas the identity over the whole genome (counting regions not covered by HSPs as non-identical) was 98.67%. The inferred digital DNA-DNA hybridization values for the two strains are 96.47% (formula 1 in [48], 86.08% (formula 2) and 97.05% (formula 3), respectively. These results clearly demonstrate that according to the whole genome sequences of strains NAL^T and TPY the similarity is very high, supporting the membership of both strains in the same species.

The comparison of the number of genes belonging to the different COG categories revealed few differences between the genomes of strains NAL^T and TPY. Strain NAL^T has 2,740 genes with COGs assigned, while strain TPY has 2700. We analyzed the differences in COG assignment between the two strains and found that in almost all cases they could be explained by differences in the gene calls or pseudogene assignment, i.e. in one genome two parts of a pseudogene were called as two separate genes, while in the other genome they were combined into one pseudogene. The only clear case of a difference in gene content between the two strains is the presence of a transposable element consisting of two genes (Sulac_1668, Sulac_1669) disrupting a subunit of a potassium transporter (Sulac_1667) in strain NAL^T. There were also cases where a gene in one strain was split into two genes in the other strain. For example, Sulac_2178 corresponds to TPY_1983 and TPY1984, and Sulac_0347 corresponds to TPY_0381 and TPY_0382. In both cases the differences are due to a single base indel.

A dot plot showed that there are large blocks of synteny between the two genomes with some rearrangements (data not shown). The genes found on the plasmid in strain NAL^T are found in two regions of the chromosome in strain TPY. Sulac_3528-3555 correspond to TPY_0524-0552, while Sulac_3556-3626 correspond to TPY_2310-2244. This suggests that in strain TPY the plasmid was inserted into the chromosome and then split into two pieces.
We analyzed CRISPR repeats with the CRISPR Recognition Tool [49] and found major differences between the two strains. They both have two regions of CRISPR repeats, but the strain TPY repeat regions have 8 and 9 repeats while the strain NAL^T repeat regions have 27 and 43 repeats. All of the spacers in the TPY repeat regions are found in NAL^T, but NAL^T has many additional spacers. This agrees with previous results suggesting that CRISPRs evolve quickly, and differences can be found in closely related strains [50].

Acknowledgements

We would like to gratefully acknowledge the help of Gabriele Gehrich-Schröter for growing *S. acidophilus* cultures and Susanne Schneider for DNA extraction (both at DSMZ). This work was performed under the auspices of the US Department of Energy's 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. Euzéby JP. List of bacterial names with standing in nomenclature: A folder available on the Internet. *Int J Syst Bacteriol* 1997; **47**:590-592.
2. Johnson DB, Joulian C, d'Hugues P, Hallberg KB. *Sulfobacillus benefaciens* sp. nov., an acidophilic facultative anaerobic Firmicute isolated from mineral bioleaching operations. *Extremophiles* 2008; **12**:789-798.
3. Norris PR, Clark DA, Owen JP, Waterhouse S. Characteristics of *Sulfobacillus acidophilus* sp. nov. and other moderately thermophilic mineral-sulphide-oxidizing bacteria. *Microbiology* 1996; **142**:775-783.
4. Bridge TAM, Johnson DB. Reduction of soluble iron and reductive dissolution of ferric iron-containing minerals by moderately thermophilic iron-oxidizing bacteria. *Appl Environ Microbiol* 1998; **64**:2181-2186.
5. Ludwig W, Schleifer KH, Whitman WB. Revised road map to the phylum *Firmicutes*. *In:* Bergey's Manual of Systematic Bacteriology, 2nd ed., vol. 3. De Vos P, Garrity G, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer KH, Whitman WB eds. 2009. Springer-Verlag, New York. pp. 1-13.
6. Golovacheva RS, Karavaiko GI. *Sulfobacillus*, a new genus of thermophilic sporulating bacteria. *Mikrobiologiia* 1978; **47**:815-822.
7. Li B, Chen Y, Liu Q, Hu S, Chen X. Complete genome analysis of *Sulfobacillus acidophilus* strain TPY, isolated from a hydrothermal vent in the Pacific Ocean. *J Bacteriol* 2011; **193**:5555.
8. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. *J Mol Biol* 1990; **215**:403-410.
9. Korf I, Yandell M, Bedell J. BLAST, O'Reilly, Sebastopol, 2003.
10. 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 Env Microbiol* 2006; **72**:5069-5072.
11. Porter MF. An algorithm for suffix stripping. *Program: electronic library and information systems* 1980; **14**:130-137.
12. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. *Bioinformatics* 2002; **18**:452-464.
13. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000; **17**:540-552.
14. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web-servers. *Syst Biol* 2008; **57**:758-771.
15. Ueda K, Yamashita A, Ishikawa J, Shimada M, Watsuji TO, Morimura K, Ikeda H, Hattori M, Beppu T. Genome sequence of *Symbiobacterium thermophilum*, an uncultivatable bacterium that depends on microbial commensalism. *Nucleic Acids Res* 2004; **32**:4937-4944.
16. Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A. How many bootstrap replicates are necessary? *Lect Notes in Comput Sci* 2009; **5541**:184-200.
17. Swofford DL. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.0 b10. Sinauer Associates, Sunderland, 2002.
18. 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.
19. Watling HR, Perrot FA, Shiers DW. Comparison of selected characteristics of *Sulfobacillus* species and review of their occurrence in acidic and bioleaching environments. *Hydrometallurgy* 2008; **93**:57-65.
20. 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.
21. 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.
22. Garrity G. NamesforLife. BrowserTool takes expertise out of the database and puts it right in the browser. *Microbiol Today* 2010; **37**:9.
23. 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.
24. Gibbons NE, Murray RGE. Proposals Concerning the Higher Taxa of Bacteria. *Int J Syst Bacteriol* 1978; **28**:1-6.
25. Garrity GM, Holt JG. The Road Map to the Manual. *In:* Garrity GM, Boone DR, Castenholz RW (*eds*), Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 1, Springer, New York, 2001, p. 119-169.
26. Murray RGE. The Higher Taxa, or, a Place for Everything...?. *In:* Holt JG (*ed*), Bergey's Manual of Systematic Bacteriology, First Edition, Volume 1, The Williams and Wilkins Co., Baltimore, 1984, p. 31-34.
27. List Editor. List of new names and new combinations previously effectively, but not validly, published. List no. 132. *Int J Syst Evol Microbiol* 2010; **60**:469-472.
28. Rainey F A. Class II. Clostridia class nov. *In:* De Vos P, Garrity G, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer KH, Whitman WB (*eds*), Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 3, Springer-Verlag, New York, 2009, p. 736.

29. Skerman VBD, McGowan V, Sneath PHA. Approved Lists of Bacterial Names. *Int J Syst Bacteriol* 1980; **30**:225-420.
30. Prévot AR. *In:* Hauderoy P, Ehringer G, Guillot G, Magrou. J., Prévot AR, Rosset D, Urbain A (*eds*), Dictionnaire des Bactéries Pathogènes, Second Edition, Masson et Cie, Paris, 1953, p. 1-692.
31. List Editor. Validation List no. 36. Validation of the publication of new names and new combinations previously effectively published outside the IJSB. *Int J Syst Bacteriol* 1991; **41**:178-179.
32. da Costa M S, Rainey F A. Family II. *Alicyclobacillaceae* fam. nov. *In:* De Vos P, Garrity G, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer KH, Whitman WB (eds), Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 3, Springer-Verlag, New York, 2009, p. 229.
33. List Editor. Validation List no. 59. Validation of the publication of new names and new combinations previously effectively published outside the *IJSB. Int J Syst Bacteriol* 1996; **46**:1189-1190.
34. BAuA. Classification of bacteria and archaea in risk groups. TRBA 466. p. 227. Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, Germany. 2010.
35. 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. *Nat Genet* 2000; **25**:25-29.
36. Klenk H, Göker M. En route to a genome-based classification of *Archaea* and *Bacteria*? *Syst Appl Microbiol* 2010; **33**:175-182.
37. List of growth media used at DSMZ: http//www.dsmz.de/catalogues/catalogue-microorganisms/culture-technology/list-of-media-for-microorganisms.html.
38. 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.
39. JGI website. http://www.jgi.doe.gov/.
40. The Phred/Phrap/Consed software package. http://www.phrap.com.
41. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008; **18**:821-829.
42. 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.
43. 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.
44. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; **11**:119.
45. Pati A, Ivanova NN, Mikhailova N, Ovchinnikova G, Hooper SD, Lykidis A, Kyrpides NC. GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nat Methods* 2010; **7**:455-457.
46. Markowitz VM, Ivanova NN, Chen I-MA, Chu K, Kyrpides NC. IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* 2009; **25**:2271-2278.
47. 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.

48. Auch AF, Klenk HP, Göker M. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Stand Genomic Sci* 2010; **2**:142-148.
49. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 2007; **8**:209.
50. Tyson GW, Banfield JF. Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol* 2008; **10**:200-207.
51. Yarza P, Ludwig W, Euzéby J, Amann R, Schleifer KH, Glöckner FO, Rosselló-Móra R. Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. *Syst Appl Microbiol* 2010; **33**:291-299.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.