Introduction

Strain NCTC 11300 (\(=\) ATCC 33386 = NCTC 11300) is the type strain of the species Sebaldella termitidis [1]. The strain was first isolated from posterior intestinal content of Reticulitermes lucifugus (Mediterranean termites) by the French microbiologist Madeleine Sebald [1,2], and was initially classified as Bacteroides termitidis [3]. The unusually low G+C content, as well as biochemical features which did not correspond to those known for the other members of the genus Bacteroides [4], and the subsequently described novel 16S rRNA sequences [5] made the position of B. termitidis within the genus Bacteroides appear controversial, and guided Collins and Shah in 1986 to reclassify B. termitidis as the type strain of the novel genus Sebaldella [1]. Here we present a summary classification and a set of features for S. termitidis NCTC 11300, together with the description of the complete genomic sequencing and annotation.

Classification and features

NCTC 11300 represents an isolated species, with no other cultivated strain known in the literature belonging to the species. An uncultured clone with identical 16S rRNA sequence was identified in a mesophilic anaerobic digester that treats municipal wastewater sludge in Clos de Hilde, France [6], and another uncultured clone, PCD-1 (96.1% 16S rRNA sequence identity), was reported from the

---

Complete genome sequence of Sebaldella termitidis type strain (NCTC 11300)

Miranda Harmon-Smith, Laura Celia, Olga Chertkov, Alla Lapidus, Alex Copeland, Tijana Glavina Del Rio, Matt Nolan, Susan Lucas, Hope Tice, Jan-Fang Cheng, Cliff Han, John C. Detter, David Bruce, Lynne Goodwin, Sam Pitluck, Amrita Pati, Konstantinos Liolios, Natalia Ivanova, Konstantinos Mavromatis, Natalia Mikhailova, Amy Chen, Krishna Palaniappan, Miriam Land, Loren Hauser, Yun-Juan Chang, Cynthia D. Jeffries, Thomas Brettin, Markus Göker, Brian Beck, James Bistow, Jonathan A. Eisen, Victor Markowitz, Philip Hugenholtz, Nikos C. Kyrpides, Hans-Peter Klenk, and Feng Chen

1 DOE Joint Genome Institute, Walnut Creek, California, USA
2 ATCC- American Type Culture Collection, Manassas, Virginia, 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 DSMZ – German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany
7 University of California Davis Genome Center, Davis, California, USA

*Corresponding author: Hans-Peter Klenk

Keywords: anaerobic, mesophile, nonmotile, non-sporeforming, Gram-negative, termite intestine, ‘Fusobacteria’, ‘Leptotrichiaceae’, GEBA

Sebaldella termitidis (Sebald 1962) Collins and Shah 1986, is the only species in the genus Sebaldella within the fusobacterial family ‘Leptotrichiaceae’. The sole and type strain of the species was first isolated about 50 years ago from intestinal content of Mediterranean termites. The species is of interest for its very isolated phylogenetic position within the phylum Fusobacteria in the tree of life, with no other species sharing more than 90% 16S rRNA sequence similarity. The 4,486,650 bp long genome with its 4,210 protein-coding and 54 RNA genes is part of the Genomic Encyclopedia of Bacteria and Archaea project.

The Genomic Standards Consortium
Sebaldella termitidis type strain (NCTC 11300T)

digestive tract of the ground beetle *Poecilus chal- cites* [7]. The closest related type strains are those of the genus *Leptotrichia*, which share 85.9 to 89.96% 16S rRNA sequence similarity [8]. Neither environmental screenings nor metagenomic surveys provided any 16S rRNA sequence with significant sequence similarity to NCTC 11300T, indicating that members of the species *S. termitidis* and the genus *Sebaldella* are not very frequent in the environment (status February 2010).

**Figure 1.** Phylogenetic tree highlighting the position of *S. termitidis* NCTC 11300T relative to the other type strains within the family ‘Leptotrichiaceae’. The tree was inferred from 1,422 aligned characters [9,10] of the 16S rRNA gene sequence under the maximum likelihood criterion [11] and rooted in accordance with the current taxonomy. The branches are scaled in terms of the expected number of substitutions per site. Numbers above branches are support values from 1,000 bootstrap replicates if larger than 60%. Lineages with type strain genome sequencing projects registered in GOLD [12] are shown in blue, published genomes in bold, e.g. the recently published GEBA genomes from *Leptotrichia buccalis* [13], and *Streptobacillus moniliformis* [14].

Cells of strain NCTC 11300T are Gram-negative, obligately anaerobic, nonmotile, nonsporeforming rods of 0.3 to 0.5 x 2 to 12 μm with central swellings (Figure 2 and Table 1) [1]. Cells occur singly, in pairs, as well as in filaments [1]. Colonies on surface are transparent to opaque, circular measuring 1-2 mm in diameter, whereas colonies in deep agar are non pigmented and lenticular [1].

The major end products of the glucose metabolism by strain NCTC 11300T are acetic and lactic acids (with some formic acid) as opposed to succinic and acetic acids dominating in members of the genus *Bacteroides* [1]. Enzymes of the hexose-monophosphate-shunt are missing, while present in members of the genus *Bacteroides* [1,4]. A list of additional sugars and alcohols used or not-used for fermentation is provided by Collins and Shah [1].

**Figure 2.** Scanning electron micrograph of *S. termitidis* NCTC 11300T. (J. Carr, CDC, Atlanta, Georgia). More EM photos of the organism can be found at [http://phil.cdc.gov/phi](http://phil.cdc.gov/phi)
Table 1. Classification and general features of S. termitidis NCTC 11300 according to the MIGS recommendations [15]

<table>
<thead>
<tr>
<th>MIGS ID</th>
<th>Property</th>
<th>Term</th>
<th>Evidence code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Domain</td>
<td>Bacteria</td>
<td>TAS [16]</td>
</tr>
<tr>
<td></td>
<td>Phylum</td>
<td>Fusobacteria</td>
<td>TAS [17]</td>
</tr>
<tr>
<td>Current classification</td>
<td>Class ‘Fusobacteria’</td>
<td>TAS [17]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Order ‘Fusobacteriales’</td>
<td>TAS [17]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Family ‘Leptotrichiaceae’</td>
<td>TAS [18]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Genus Sebaldella</td>
<td>TAS [1,19]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Species Sebaldella termitidis</td>
<td>TAS [1,19]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type strain NCTC 11300</td>
<td>TAS [1]</td>
<td></td>
</tr>
<tr>
<td>Gram stain</td>
<td>Gram negative</td>
<td>TAS [1]</td>
<td></td>
</tr>
<tr>
<td>Cell shape</td>
<td>rod-shaped, with central swellings;</td>
<td>occur singly, in pairs and in filaments</td>
<td>TAS [1]</td>
</tr>
<tr>
<td>Motility</td>
<td>nonmotile</td>
<td>TAS [1]</td>
<td></td>
</tr>
<tr>
<td>Sporulation</td>
<td>nonsporulating</td>
<td>TAS [2]</td>
<td></td>
</tr>
<tr>
<td>Temperature range</td>
<td>mesophile</td>
<td>NAS</td>
<td></td>
</tr>
<tr>
<td>Optimum temperature</td>
<td>not determined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salinity</td>
<td>not reported</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIGS-22</td>
<td>Oxygen requirement</td>
<td>obligate anaerobic</td>
<td>TAS [1]</td>
</tr>
<tr>
<td>Carbon source</td>
<td>glucose and other sugars</td>
<td>TAS [1]</td>
<td></td>
</tr>
<tr>
<td>Energy source</td>
<td>fermentation of glucose and other sugars</td>
<td>TAS [1]</td>
<td></td>
</tr>
<tr>
<td>MIGS-6</td>
<td>Habitat</td>
<td>bacterial flora of termite gastrointestinal tract</td>
<td>TAS [1]</td>
</tr>
<tr>
<td>MIGS-15</td>
<td>Biotic relationship</td>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>MIGS-14</td>
<td>Pathogenicity</td>
<td>none reported</td>
<td>NAS</td>
</tr>
<tr>
<td>Biosafety level</td>
<td>2</td>
<td>TAS [20]</td>
<td></td>
</tr>
<tr>
<td>Isolation</td>
<td>posterior intestinal content of termites</td>
<td>TAS [2]</td>
<td></td>
</tr>
<tr>
<td>MIGS-4</td>
<td>Geographic location</td>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>MIGS-5</td>
<td>Sample collection time</td>
<td>1962 or before</td>
<td>TAS [1,2]</td>
</tr>
<tr>
<td>MIGS-4.1</td>
<td>Latitude</td>
<td>not reported</td>
<td></td>
</tr>
<tr>
<td>MIGS-4.2</td>
<td>Longitude</td>
<td>not reported</td>
<td></td>
</tr>
<tr>
<td>MIGS-4.3</td>
<td>Depth</td>
<td>not reported</td>
<td></td>
</tr>
<tr>
<td>MIGS-4.4</td>
<td>Altitude</td>
<td>not reported</td>
<td></td>
</tr>
</tbody>
</table>

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 of the Gene Ontology project [21]. If the evidence code is IDA, then the property was directly observed for a live isolate by one of the authors or an expert mentioned in the acknowledgements.

Chemotaxonomy
The cell wall structure of strain NCTC 11300T has not yet been reported. Nonhydroxylated and 3-hydroxylated fatty acids were present [1]. The major long chain fatty acids are saturated and monounsaturated straight chain acids: C\textsubscript{16:0} (37%) and C\textsubscript{18:1} (41%), with methyl branched acids being absent [1], as opposed to straight-chain saturated, anteiso- and iso-methyl branched-chain acids in members of the genus Bacteroides, which are missing the monounsaturated acids [1]. Menaquinones were not detected, as opposed to members of the genus Bacteroides [1].

Genome sequencing and annotation
Genome project history
This organism was selected for sequencing on the basis of its phylogenetic position, and is part of the Genomic Encyclopedia of Bacteria and Archaea project [22]. The genome project is deposited in the Genome OnLine Database [12] 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.

http://standardsigenomics.org
**Table 2. Genome sequencing project information**

<table>
<thead>
<tr>
<th>MIGS-31</th>
<th>Finishing quality</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIGS-28</td>
<td>Libraries used</td>
<td>One genomic 8kb pMCL200 library, one 454 pyrosequence library and one Illumina library</td>
</tr>
<tr>
<td>MIGS-29</td>
<td>Sequencing platforms</td>
<td>Sanger, 454 Titanium, Illumina</td>
</tr>
<tr>
<td>MIGS-31.2</td>
<td>Sequencing coverage</td>
<td>9.2× Sanger; 30.3× 454 Titanium</td>
</tr>
<tr>
<td>MIGS-30</td>
<td>Assemblers</td>
<td>Newbler, phrap</td>
</tr>
<tr>
<td>MIGS-32</td>
<td>Gene calling method</td>
<td>Prodigal, GenePRIMP</td>
</tr>
<tr>
<td>INSDC ID</td>
<td></td>
<td>CP001739 (chromosome), CP001740, CP001741 (plasmids)</td>
</tr>
<tr>
<td>Genbank Date of Release</td>
<td>November 19, 2009</td>
<td></td>
</tr>
<tr>
<td>GOLD ID</td>
<td>Gc01144</td>
<td></td>
</tr>
<tr>
<td>NCBI project ID</td>
<td>29539</td>
<td></td>
</tr>
<tr>
<td>Database: IMG-GEBA</td>
<td>2501846314</td>
<td></td>
</tr>
<tr>
<td>MIGS-13</td>
<td>Source material identifier</td>
<td>ATCC 33386</td>
</tr>
<tr>
<td>Project relevance</td>
<td>Tree of Life, GEBA</td>
<td></td>
</tr>
</tbody>
</table>

**Growth conditions and DNA isolation**

*S. termitidis* NCTC 11300T, ATCC 33386™, was grown anaerobically in ATCC medium 1490 (Modified chopped meat medium) [23] at 37°C. DNA was isolated from cell paste using a basic CTAB extraction and then quality controlled according to JGI guidelines.

**Genome sequencing and assembly**

The genome was sequenced using a combination of Sanger and 454 sequencing platforms. All general aspects of library construction and sequencing can be found at [http://www.jgi.doe.gov/](http://www.jgi.doe.gov/). 454 Pyrosequencing reads were assembled using the Newbler assembler version 1.02.15 (Roche). Large Newbler contigs were broken into 4,966 overlapping fragments of 1,000 bp and entered into assembly as pseudo-reads. The sequences were assigned quality scores based on Newbler consensus q-scores with modifications to account for overlap redundancy and to adjust inflated q-scores. A hybrid 454/Sanger assembly was made using the parallel phrap assembler (High Performance Software, LLC). Possible mis-assemblies were corrected with Dupfinisher [24] or transposon bombing of bridging clones (Epicentre Biotechnologies, Madison, WI). Gaps between contigs were closed by editing in Consed, custom primer walk or PCR amplification. A total of 796 Sanger finishing reads were produced to close gaps, to resolve repetitive regions, and to raise the quality of the finished sequence. Illumina reads were used to improve the final consensus quality using an in-house developed tool (the Polisher, unpublished). The error rate of the completed genome sequence is less than 1 in 100,000. Together all sequence types provided 39.5× coverage of the genome. The final assembly contains 45,934 Sanger and 760,187 pyrosequence reads.

**Genome annotation**

Genes were identified using Prodigal [25] as part of the Oak Ridge National Laboratory genome annotation pipeline, followed by a round of manual curation using the JGI GenePRIMP pipeline [26]. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant database, UniProt, TIGRFam, Pfam, PRIAM, KEGG, COG, and InterPro databases. Additional gene prediction analysis and manual functional annotation was performed within the Integrated Microbial Genomes Expert Review (IMG-ER) platform [27].

**Genome properties**

The genome consists of a 4,418,842 bp long chromosome, and two plasmids with 54,160 bp and 13,648 bp length, respectively, with a 33.4% GC content (Table 3 and Figure 3). Of the 4,264 genes predicted, 4,210 were protein-coding genes, and 54 RNAs; 59 pseudogenes were identified. The majority of the protein-coding genes (60.4%) were assigned with a putative function while those remaining were annotated as hypothetical proteins. The distribution of genes into COGs functional categories is presented in Table 4.
Table 3. Genome Statistics

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome size (bp)</td>
<td>4,486,650</td>
<td>100.00%</td>
</tr>
<tr>
<td>DNA coding region (bp)</td>
<td>3,918,335</td>
<td>87.33%</td>
</tr>
<tr>
<td>DNA G+C content (bp)</td>
<td>1,497,450</td>
<td>33.38%</td>
</tr>
<tr>
<td>Number of replicons</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Extrachromosomal elements</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Total genes</td>
<td>4,264</td>
<td>100.00%</td>
</tr>
<tr>
<td>RNA genes</td>
<td>54</td>
<td>1.27%</td>
</tr>
<tr>
<td>rRNA operons</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Protein-coding genes</td>
<td>4,210</td>
<td>98.73%</td>
</tr>
<tr>
<td>Pseudogenes</td>
<td>59</td>
<td>1.38%</td>
</tr>
<tr>
<td>Genes with function prediction</td>
<td>2,576</td>
<td>60.41%</td>
</tr>
<tr>
<td>Genes in paralog clusters</td>
<td>1,253</td>
<td>29.39%</td>
</tr>
<tr>
<td>Genes assigned to COGs</td>
<td>2,299</td>
<td>60.95%</td>
</tr>
<tr>
<td>Genes assigned Pfam domains</td>
<td>2,787</td>
<td>65.36%</td>
</tr>
<tr>
<td>Genes with signal peptides</td>
<td>801</td>
<td>18.79%</td>
</tr>
<tr>
<td>Genes with transmembrane helices</td>
<td>901</td>
<td>21.13%</td>
</tr>
<tr>
<td>CRISPR repeats</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Graphical circular maps of the chromosome and the two plasmids. From outside to the center: Genes on forward strand (color by COG categories), Genes on reverse strand (color by COG categories), RNA genes (tRNAs green, rRNAs red, other RNAs black), GC content, GC skew.
Table 4. Number of genes associated with the general COG functional categories

<table>
<thead>
<tr>
<th>Code</th>
<th>Value</th>
<th>%age</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>152</td>
<td>3.6</td>
<td>Translation, ribosomal structure and biogenesis</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0.0</td>
<td>RNA processing and modification</td>
</tr>
<tr>
<td>K</td>
<td>265</td>
<td>6.3</td>
<td>Transcription</td>
</tr>
<tr>
<td>L</td>
<td>130</td>
<td>3.1</td>
<td>Replication, recombination and repair</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0.0</td>
<td>Chromatin structure and dynamics</td>
</tr>
<tr>
<td>D</td>
<td>22</td>
<td>0.5</td>
<td>Cell cycle control, cell division, chromosome partitioning</td>
</tr>
<tr>
<td>Y</td>
<td>0</td>
<td>0.0</td>
<td>Nuclear structure</td>
</tr>
<tr>
<td>V</td>
<td>47</td>
<td>1.1</td>
<td>Defense mechanisms</td>
</tr>
<tr>
<td>T</td>
<td>96</td>
<td>2.3</td>
<td>Signal transduction mechanisms</td>
</tr>
<tr>
<td>M</td>
<td>155</td>
<td>3.7</td>
<td>Cell wall/membrane biogenesis</td>
</tr>
<tr>
<td>N</td>
<td>17</td>
<td>0.4</td>
<td>Cell motility</td>
</tr>
<tr>
<td>Z</td>
<td>0</td>
<td>0.0</td>
<td>Cytoskeleton</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
<td>0.0</td>
<td>Extracellular structures</td>
</tr>
<tr>
<td>U</td>
<td>41</td>
<td>1.0</td>
<td>Intracellular trafficking, secretion and vesicular transport</td>
</tr>
<tr>
<td>O</td>
<td>71</td>
<td>1.7</td>
<td>Posttranslational modification, protein turnover, chaperones</td>
</tr>
<tr>
<td>C</td>
<td>128</td>
<td>3.0</td>
<td>Energy production and conversion</td>
</tr>
<tr>
<td>G</td>
<td>468</td>
<td>11.1</td>
<td>Carbohydrate transport and metabolism</td>
</tr>
<tr>
<td>E</td>
<td>219</td>
<td>5.2</td>
<td>Amino acid transport and metabolism</td>
</tr>
<tr>
<td>F</td>
<td>93</td>
<td>2.2</td>
<td>Nucleotide transport and metabolism</td>
</tr>
<tr>
<td>H</td>
<td>106</td>
<td>2.5</td>
<td>Coenzyme transport and metabolism</td>
</tr>
<tr>
<td>I</td>
<td>59</td>
<td>1.4</td>
<td>Lipid transport and metabolism</td>
</tr>
<tr>
<td>P</td>
<td>105</td>
<td>2.5</td>
<td>Inorganic ion transport and metabolism</td>
</tr>
<tr>
<td>Q</td>
<td>32</td>
<td>0.8</td>
<td>Secondary metabolites biosynthesis, transport and catabolism</td>
</tr>
<tr>
<td>R</td>
<td>403</td>
<td>9.6</td>
<td>General function prediction only</td>
</tr>
<tr>
<td>S</td>
<td>241</td>
<td>5.7</td>
<td>Function unknown</td>
</tr>
<tr>
<td>-</td>
<td>1,665</td>
<td>39.5</td>
<td>Not in COGs</td>
</tr>
</tbody>
</table>

Acknowledgements

We would like to gratefully acknowledge the help of Janice Carr (Centers of Disease Control, Atlanta, Georgia) for providing the EM photo of *S. thermitidis* NCTC 11300^T_. 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, Los Alamos National Laboratory under contract No. DE-AC02-06NA25396, and Oak Ridge National Laboratory under contract DE-AC05-00OR22725

References


225 Standards in Genomic Sciences


Sebaldella termitidis type strain (NCTC 11300T)
