

Permanent draft genome sequence of the gliding predator *Saprospira grandis* strain Sa g1 (= HR1)

Konstantinos Mavromatis¹, Olga Chertkov^{1,2}, Alla Lapidus¹, Matt Nolan¹, Susan Lucas¹, Hope Tice¹, Tijana Glavina Del Rio¹, Jan-Fang Cheng¹, Cliff Han^{1,2}, Roxanne Tapia^{1,2}, David Bruce^{1,2}, Lynne A. Goodwin^{1,2}, Sam Pitluck¹, Marcel Huntemann¹, Konstantinos Liolios¹, Ioanna Pagani¹, Natalia Ivanova¹, Natalia Mikhailova¹, Amrita Pati¹, Amy Chen³, Krishna Palaniappan³, Miriam Land^{1,4}, Evelyne-Marie Brambilla⁶, Manfred Rohde⁵, Stefan Spring⁶, Markus Göker⁶, John C. Detter^{1,2}, James Bristow¹, Jonathan A. Eisen^{1,7}, Victor Markowitz³, Philip Hugenholtz^{1,8}, Nikos C. Kyrpides¹, Hans-Peter Klenk^{6*}, and Tanja Woyke¹

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 authors: Hans-Peter Klenk (hpk@dsMZ.de)

Keywords: strictly aerobic, gliding, Gram-negative, mesophilic, chemoorganotrophic, marine, ixotrophy, *Saprospiraceae*, GEBA

Saprospira grandis Gross 1911 is a member of the *Saprospiraceae*, a family in the class 'Sphingobacteria' that remains poorly characterized at the genomic level. The species is known for preying on other marine bacteria via 'ixotrophy'. *S. grandis* strain Sa g1 was isolated from decaying crab carapace in France and was selected for genome sequencing because of its isolated location in the tree of life. Only one type strain genome has been published so far from the *Saprospiraceae*, while the sequence of strain Sa g1 represents the second genome to be published from a non-type strain of *S. grandis*. Here we describe the features of this organism, together with the complete genome sequence and annotation. The 4,495,250 bp long Improved-High-Quality draft of the genome with its 3,536 protein-coding and 62 RNA genes is a part of the *Genomic Encyclopedia of Bacteria and Archaea* project.

Introduction

Strain Sa g1 (= HR1 = DSM 2844 = ATCC 49590 = LMG 13157) belongs to the species *Saprospira grandis* [1,2] in the monospecific genus *Saprospira* [2,3]. The type strain of the species is Lewin WH^T (= ATCC 23119 = LMG 10407) [1,3] and is known for its predatory life style when capturing and preying on other bacteria via 'ixotrophy' [2]. Strain Sa g1 was isolated in 1975 from decaying crab carapace in Roscoff, France [4]. The genus name was derived from the Greek adjective *sapros*, meaning rotten/putrid, and the Latin *spira*, a coil/spiral, resulting in the Neo-Latin *Saprospira*, a spiral associated with decaying matter [5]; the species

epithet was derived from the Latin adjective *grandis*, large [5]. Life style and ecological role of members of the species was recently summarized by Saw *et al.* [6] when they reported the genome sequence of strain Lewin (isolated from La Jolla beach in San Diego; not to be confused with strain Lewin WH^T, the type strain of the species which was also isolated by Lewin, but from a rockpool near high water, Woods Hole). Strain Lewin was the first member of the genus *Saprospira* to be completely sequenced. Here we present a summary classification and a set of features for *S. grandis* Sa g1, together with the description of the genomic sequencing and annotation.

Classification and features

A representative genomic 16S rRNA sequence of strain Sa g1 was compared using NCBI BLAST [7,8] 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 [9] and the relative frequencies of taxa and keywords (reduced to their stem [10]) were determined, weighted by BLAST scores. The most frequently occurring genera were *Saprospira* (82.0%), *Aureispira* (5.4%), "*Aureospira*" (4.8%), *Cytophaga* (3.9%) and *Lewinella* (3.8%) (16 hits in total). Regarding the three hits to sequences from members of the species, the average identity within HSPs was 99.4%, whereas the average coverage by HSPs was 98.6%. Among all other species, the one yielding the highest score was *Aureispira maritima* (AB278130), which corresponded to an identity of 87.3% and an HSP coverage of 98.0%. (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 FJ792500 ('Unexpectedly archaeal species shift between rare and dominant over thousand year time scales carbonate chimney Lost City Hydrothermal Field clone SGYF672'), which showed an identity of 99.2% and an HSP coverage of 100.3%. The most frequently occurring keywords within the labels of all environmental samples which yielded hits were 'lake' (3.8%), 'sludg' (2.9%), 'microbi' (2.8%), 'mat' (2.7%) and 'activ' (2.3%) (234 hits in total) and correspond to the already known habitats for strains of this species.

Figure 1 shows the phylogenetic neighborhood of *S. grandis* strain Sa g1 in a 16S rRNA based tree. The sequences of the four 16S rRNA gene copies in the genome differ from each other by up to one nucleotide, and differ by up to seven nucleotides from the previously published 16S rRNA sequence (M58795), which contains 52 ambiguous base calls.

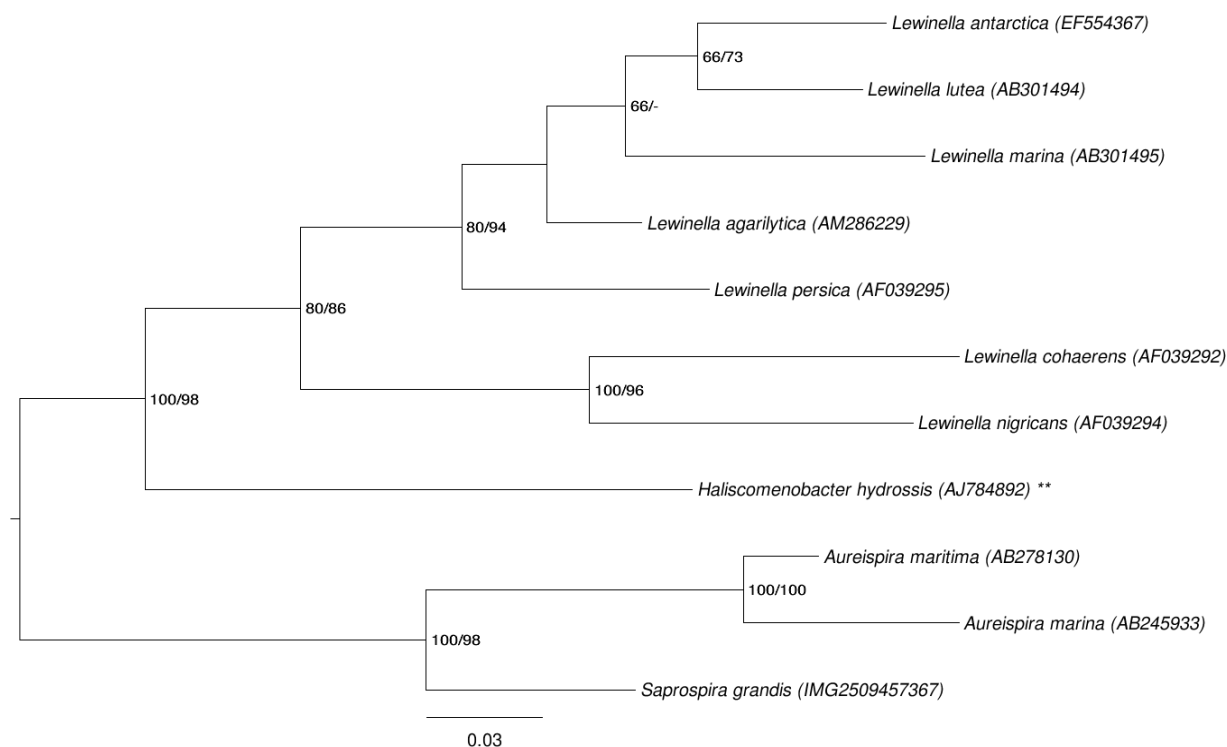


Figure 1. Phylogenetic tree highlighting the position of *S. grandis* relative to the type strains of the other species within the family Saprospiraceae. The tree was inferred from 1,413 aligned characters [11,12] of the 16S rRNA gene sequence under the maximum likelihood (ML) criterion [13]. Rooting was done initially using the mid-point method [14] 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 250 ML bootstrap replicates [15] (left) and from 1,000 maximum parsimony bootstrap replicates [16] (right) if larger than 60%. Lineages with type strain genome sequencing projects registered in GOLD [17] are labeled with one asterisk, those also listed as 'Complete and Published' with two asterisks [18].

General features of *S. grandis* were summarized previously by Saw *et al.* in the short genome report of strain Lewin [6], and are therefore not repeated here. Individual features of strain Sa g1 are largely unknown due to a lack of relevant publications, as

are chemotaxonomical data. A description of the isolation and some morphological features of strain Sa g1 are reported by Reichenbach [4]. Figure 2 shows an electron micrograph of the *S. grandis* Sa g1 cells.

Table 1. Classification and general features of *S. grandis* Sa g1 according to the MIGS recommendations [19] and the Names for Life database [2].

MIGS ID	Property	Term	Evidence code
		Domain <i>Bacteria</i>	TAS [20]
		Phylum <i>Bacteroidetes</i>	TAS [21,22]
		Class ' <i>Sphingobacteria</i> '	TAS [21,23]
	Current classification	Order <i>Sphingobacteriales</i>	TAS [21,24]
		Family <i>Saprospiraceae</i>	TAS [21,25]
		Genus <i>Saprospira</i>	TAS [3,26,27]
		Species <i>Saprospira grandis</i>	TAS [3,26]
		Strain Sa g1	TAS [4]
	Gram stain	negative	TAS [28,29]
	Cell shape	helical filaments	TAS [28,29]
	Motility	<i>via</i> gliding	TAS [28,29]
	Sporulation	non-sporulating	NAS
	Temperature range	mesophile, 6-47°C	TAS [28,29]
	Optimum temperature	25-30°C	TAS [4,28,29]
	Salinity	seawater	TAS [28,29]
MIGS-22	Oxygen requirement	strictly aerobe	TAS [28,29]
	Carbon source	peptides, proteins	TAS [28,29]
	Energy metabolism	chemoorganotroph	TAS [28,29]
MIGS-6	Habitat	marine littoral zone	TAS [28]
MIGS-15	Biotic relationship	free living	TAS [28]
MIGS-14	Pathogenicity	not reported	
	Biosafety level	1	TAS [30]
MIGS-23.1	Isolation	decaying crab carapace	TAS [4]
MIGS-4	Geographic location	Roscoff, France	TAS [4]
MIGS-5	Sample collection time	September 1975	TAS [4]
MIGS-4.1	Latitude	48.70	NAS
MIGS-4.2	Longitude	-3.97	
MIGS-4.3	Depth	not reported	
MIGS-4.4	Altitude	not reported	

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 [31]. 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.

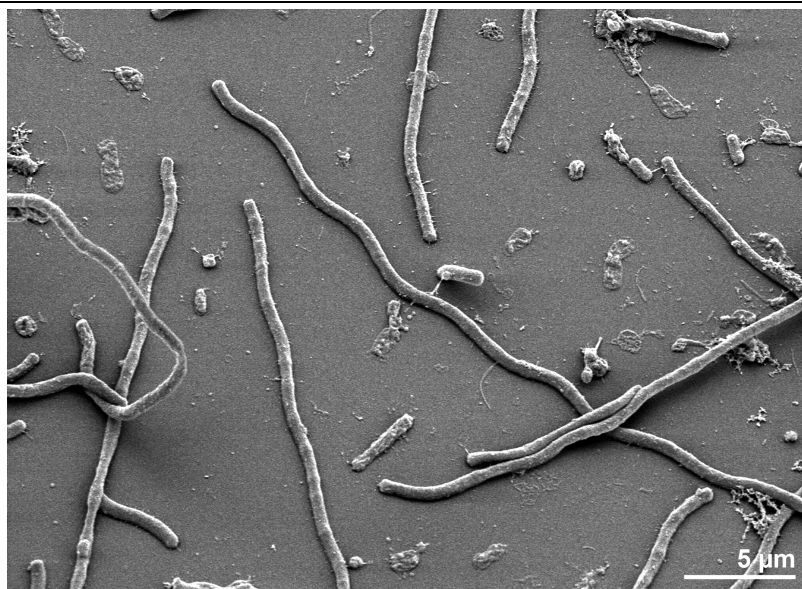


Figure 2. Scanning electron micrograph of *S. grandis* Sa g1

Genome sequencing and annotation

Genome project history

This organism was selected for sequencing on the basis of its phylogenetic position [32], and is part of the *Genomic Encyclopedia of Bacteria and Archaea* project [33]. The genome project is deposited in the Genomes On Line Database [17] and the complete genome sequence is deposited in GenBank. Sequencing, finishing and annotation were performed by the DOE Joint Genome Institute (JGI). A summary of the project information is shown in Table 2.

Growth conditions and DNA isolation

S. grandis strain Sa g1, DSM 2844, was grown in DSMZ medium 172 (*Cytophaga* (marine) medium) [34] at 28°C. DNA was isolated from 0.5-1 g of cell paste using Jetflex Genomic DNA Purification kit (GENOMED 600100) following the standard protocol as recommended by the manufacturer without modification. DNA will be available on request through the DNA Bank Network [46].

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 [35]. Pyrosequencing reads were assembled using the Newbler assembler (Roche). The initial Newbler assembly consisting of 551 contigs in six scaffolds was converted into a phrap [36] assembly by making fake reads from the

consensus, to collect the read pairs in the 454 paired end library. Illumina GAii sequencing data (3,575.7 Mb) was assembled with Velvet [37] 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 72.8 Mb of 454 paired end data. Newbler parameters are -consed -a 50 -l 350 -g -m -ml 20. The Phred/Phrap/Consed software package [36] 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 [35], Dupfinisher [38], 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 45 additional reactions were necessary to close gaps and to raise the quality of the final contigs. Illumina reads were also used to correct potential base errors and increase consensus quality using a software Polisher developed at JGI [39]. The final assembly consists of 84 contigs in 5 scaffolds. Together, the combination of the Illumina and 454 sequencing platforms provided $777.1 \times$ coverage of the genome. The final assembly contained 235,183 pyrosequence and 45,502,670 Illumina reads.

Table 2. Genome sequencing project information

MIGS ID	Property	Term
MIGS-31	Finishing quality	Improved-High-Quality Draft
MIGS-28	Libraries used	Two genomic libraries: one 454 PE library (9 kb insert size), one Illumina library
MIGS-29	Sequencing platforms	Illumina GAii, 454 GS FLX Titanium
MIGS-31.2	Sequencing coverage	768.5 × Illumina; 8.6 × pyrosequence
MIGS-30	Assemblers	Newbler version 2.3, Velvet version 1.0.13, phrap version 1.080812
MIGS-32	Gene calling method	Prodigal 1.4, GenePRIM
	INSDC ID	requested February 9, 2012
	GenBank Date of Release	Pending
	GOLD ID	Gi03955
	NCBI project ID	61003
	Database: IMG-GEBA	2509276035
MIGS-13	Source material identifier	DSM 2844
	Project relevance	Tree of Life, GEBA

Genome annotation

Genes were identified using Prodigal [40] as part of the Oak Ridge National Laboratory genome annotation pipeline, followed by a round of manual curation using the JGI GenePRIMP pipeline [41]. 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. Additional gene prediction analysis and functional annotation was performed within the Integrated Microbial Genomes - Expert Review (IMG-ER) platform [42].

Genome properties

The Improved-High-Quality draft assembly of the genome consists of 84 contigs in four scaffolds representing the chromosome (4,422,561 bp, 11,045 bp, 2,786 bp and 2,223 bp length, respectively) and one 56,635 bp plasmid scaffold, with an overall 46.1% G+C content (Table 3 and Figure 3). Of the 3,598 genes predicted, 3,536 were protein-coding genes, and 62 RNAs; 70 pseudogenes were also identified. The majority of the protein-coding genes (57.4%) 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
Genome size (bp)	4,495,250	100.00%
DNA coding region (bp)	3,693,336	82.16%
DNA G+C content (bp)	2,067,067	46.06%
Number of scaffolds	5*	
Extrachromosomal elements	1	
Total genes	3,598	100.00%
RNA genes	62	1.72%
rRNA operons	3**	
tRNA genes	48	1.33%
Protein-coding genes	3,536	98.28%
Pseudo genes	70	1.95%
Genes with function prediction (proteins)	2,064	57.37%
Genes in paralog clusters	1,575	43.77%
Genes assigned to COGs	2,064	57.37%
Genes assigned Pfam domains	2,072	57.59%
Genes with signal peptides	1,109	30.82%
Genes with transmembrane helices	687	19.09%
CRISPR repeats	5	

* four scaffolds for the chromosome and one for a plasmid

** only two rRNA operons appear to be complete; the third copy appears to be split into two incomplete fractions due to unresolved assembly problems.

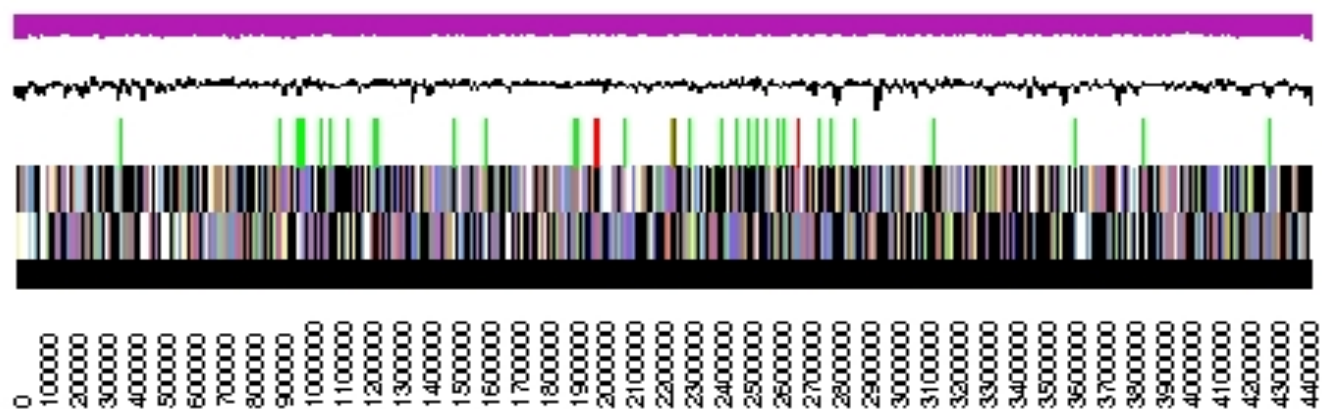


Figure 3. Graphical map of the largest scaffold, SapgrDRAFT_Contig123.4, which represents >99.6% of the chromosome. The smaller contigs of the chromosome and the plasmid are not shown, but accessible through the img/er pages on the JGI web pages [35,42]. From bottom to top: Genes on forward strand (colored 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	Description
J	152	6.8	Translation, ribosomal structure and biogenesis
A	0	0.0	RNA processing and modification
K	118	5.3	Transcription
L	167	7.5	Replication, recombination and repair
B	1	0.0	Chromatin structure and dynamics
D	34	1.5	Cell cycle control, cell division, chromosome partitioning
Y	0	0.0	Nuclear structure
V	53	2.4	Defense mechanisms
T	98	4.4	Signal transduction mechanisms
M	217	9.7	Cell wall/membrane biogenesis
N	25	1.1	Cell motility
Z	1	0.0	Cytoskeleton
W	0	0.0	Extracellular structures
U	55	2.5	Intracellular trafficking and secretion, and vesicular transport
O	121	5.4	Posttranslational modification, protein turnover, chaperones
C	108	4.9	Energy production and conversion
G	51	2.3	Carbohydrate transport and metabolism
E	125	5.6	Amino acid transport and metabolism
F	57	2.6	Nucleotide transport and metabolism
H	101	4.5	Coenzyme transport and metabolism
I	88	4.0	Lipid transport and metabolism
P	91	4.1	Inorganic ion transport and metabolism
Q	35	1.6	Secondary metabolites biosynthesis, transport and catabolism
R	316	14.2	General function prediction only
S	215	9.7	Function unknown
-	1,534	42.6	Not in COGs

Insights into the genome sequence

Comparison with the genome sequence of *S. grandis* strain Lewin

The two complete copies of the 16S rRNA gene in the Sa g1 genome show 99.5% sequence identity with those of strain Lewin [6], but only 98.0% sequence identity with the respective sequence from the yet to be genome-sequenced type strain Lewin WH^T (ATCC 23119, M58795) [43]; this discrepancy is due to the huge number of ambiguous base calls in M58795, and is relativized by 99.4% identical bases within the HSPs. Given the different habitats of the two sequenced strains it is interesting to compare some basic genome features and their membership to the same species.

The second largest scaffold in the Improved-High-Quality draft assembly of the Sa g1 genome (SapgrDRAFT_Contig162.5) has a size of 56,635 bp, which is comparable to the size of plasmid SGRA01 in *S. grandis* strain Lewin, 54,948 bp (CP002832) [6]. A BLAST search against the NCBI nr database revealed a full length colinearity with about 94% sequence identity between SapgrDRAFT_Contig162.5 and the plasmid of *S. grandis* strain Lewin. Further comparison of the two sequences with the GGDC-Genome-to-Genome Distance Calculator [44,45] revealed distances of only 0.0704 (formula 1) to 0.1342 (formula 3), corresponding to 82.7 to 88.3% DDH values. SapgrDRAFT_3602 encodes a protein involved in initiation of plasmid replication, RepB, while the largest fraction of (13) genes encoded on this scaffold belong to COG function category 'nucleotide transport and metabolism' (similar to SGRA01); therefore suggesting that SapgrDRAFT_Contig162.5 is a plasmid whose sequence was not circularized during the genome assembly.

The largest scaffold in the draft assembly of Sa g1 (SapgrDRAFT_Contig123.4) has a size of 4,422,561 bp, which is comparable to the size of the *S. grandis* strain Lewin chromosome, 4,345,237 bp. The overall genome statistics (see Table 3) of the two strains is similar in some features, such as G+C content (46.1% strain Sa g1 vs. 46.4% strain Lewin), total number of genes (3,598 vs. 4,311), genes with function predictions (2,064 vs. 2,173), three rRNA operons (both), but

deviates more in others, such as genes in paralogue clusters (1,575 vs. 215), genes with signal peptides (1,109 vs. 589), and genes with transmembrane helices (687 vs. 778), which may reflect the differences in the gene calling and annotation process (strain Lewin is not yet featured in IMG [42]). As for the number of genes associated with the general COG functional categories (see Table 4) there are categories with very similar content, such as transcription (118 genes, both), translation (152 vs. 160), defense mechanisms (53 vs. 52), cell motility (25 vs. 26), lipid transport (88 vs. 90) and cell wall/membrane biogenesis (217 vs. 206), while other categories deviate more significantly, such as replication (167 vs. 186), cell cycle control (34 vs. 20), intracellular trafficking (55 vs. 44), energy production and conversion (108 vs. 123), and secondary metabolites biosynthesis and catabolism (35 vs. 52), which again might be partially attributed to different procedures in the annotation processes.

The sequences of SapgrDRAFT_Contig123.4 and the chromosome of strain Lewin (CP002831), which represent roughly 99% of the respective genomes, were also compared with the GGDC-Genome-to-Genome Distance Calculator [44,45]. The inferred distances from formulas 1 and 3 were 0.1139 and 0.1741, respectively, corresponding to 83.1% and 77.9% DDH values, respectively, estimated via regression-based predictions. These values indicate that both strains belong to the same species, *S. grandis*.

The sequence of the three smaller scaffolds (SapgrDRAFT_Contig118.2 with 11,045 bp length, SapgrDRAFT_Contig106.1 with 2,786 bp and SapgrDRAFT_Contig119.3 with 2,223 bp) were compared against the NCBI nr database and revealed significant similarities only with the chromosome of strain Lewin.

Acknowledgements

We would like to gratefully acknowledge the help of Maren Schröder (DSMZ) for growing *S. grandis* 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-AC05-000R22725, as well as German Research Foundation (DFG) INST 599/1-2.

References

- Dawyndt P, Vancanneyt M, De Meyer H, Swings J. Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources. *IEEE Trans Knowl Data Eng* 2005; **17**:1111-1126. <http://dx.doi.org/10.1109/TKDE.2005.131>
- Garrity G. NamesforLife. BrowserTool takes expertise out of the database and puts it right in the browser. *Microbiol Today* 2010; **37**:9.
- Skerman VBD, McGowan V, Sneath PHA, eds. Approved Lists of Bacterial Names. *Int J Syst Bacteriol* 1980; **30**:225-420. <http://dx.doi.org/10.1099/00207713-30-1-225>
- Reichenbach H. *Saprospira grandis* (Leucotrichales) – Wachstum und Bewegung. Film E2424 des Instituts Wiss Film, Göttingen, *Publ Wiss Film, Sekt Biol* 1980; **13**:26/E2424.
- Euzéby JP. List of Bacterial Names with Stranding in Nomenclature: a folder on the internet. *Int J Syst Bacteriol* 1997; **47**:590-592. <http://dx.doi.org/10.1099/00207713-47-2-590>
- Saw JHW, Yuryew A, Kanbe M, Hou S, Young AG, Aizawa SI, Alam M. Complete genome sequencing and analysis of *Saprospira grandis* str. Lewin, a predatory marine bacterium. *Stand Genomic Sci* 2012; **6**:84-93.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **215**:403-410. [PubMed](http://pubmed.ncbi.nlm.nih.gov/2322822/)
- Korf I, Yandell M, Bedell J. BLAST, O'Reilly, Sebastopol, 2003.
- 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://pubmed.ncbi.nlm.nih.gov/16907960/) <http://dx.doi.org/10.1128/AEM.03006-05>
- Porter MF. An algorithm for suffix stripping. Program: *electronic library and information systems* 1980; **14**:130-137.
- Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. *Bioinformatics* 2002; **18**:452-464. [PubMed](http://pubmed.ncbi.nlm.nih.gov/12111111/) <http://dx.doi.org/10.1093/bioinformatics/18.3.452>
- Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000; **17**:540-552. [PubMed](http://pubmed.ncbi.nlm.nih.gov/11111111/) <http://dx.doi.org/10.1093/oxfordjournals.molbev.a026334>
- Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol* 2008; **57**:758-771. [PubMed](http://pubmed.ncbi.nlm.nih.gov/18111111/) <http://dx.doi.org/10.1080/10635150802429642>
- 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>
- 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
- Swofford DL. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.0 b10. Sinauer Associates, Sunderland, 2002.
- Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 2012; **40**:D571-D579. [PubMed](http://pubmed.ncbi.nlm.nih.gov/22111111/) <http://dx.doi.org/10.1093/nar/gkr1100>
- Daligault H, Lapidus A, Zyetun A, Nolan M, Lucas S, Glavina Del Rio T, Tice H, Cheng JF, Tapia R, Han C, et al. Complete genome sequence of *Halicomenobacter hydrossis* type strain (O^T).

- Stand Genomic Sci* 2011; **4**:352-360. [PubMed](#)
<http://dx.doi.org/10.4056/sigs.1964579>
19. 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>
20. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms. Proposal for the domains *Archaea* and *Bacteria*. *Proc Natl Acad Sci USA* 1990; **87**:4576-4579. [PubMed](#)
<http://dx.doi.org/10.1073/pnas.87.12.4576>
21. Validation List No. 143. *Int J Syst Evol Microbiol* 2012; **62**:1-4.
<http://dx.doi.org/10.1099/ij.s.0.039487-0>
22. Krieg NR, Ludwig W, Euzéby J, Whitman WB. Phylum XIV. Bacteroidetes phyl. nov. In: Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL, Ludwig W, Whitman WB (eds), *Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 4*, Springer, New York, 2011, p. 25.
23. Kämpfer P. Class III. Sphingobacteriia class. nov. In: Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL, Ludwig W, Whitman WB (eds), *Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 4*, Springer, New York, 2011, p. 330.
24. Kämpfer P. Order I. Sphingobacteriales ord. nov. In: Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL, Ludwig W, Whitman WB (eds), *Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 4*, Springer, New York, 2011.
25. Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL, Ludwig W, Whitman WB. Family III. *Saprosiraceae* fam. nov. In: Krieg NR, Staley JT, Brown DR, Hedlund BP, Paster BJ, Ward NL, Ludwig W, Whitman WB (eds), *Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 4*, Springer, New York, 2010, p. 358.
26. Gross J. Über freilebende Spironemaceen. *Mitteilungen aus der Zoologischen Station zu Neapel* 1911; **20**:188-203.
27. Lewin RA, Leadbetter ER. Genus V. *Saprosira* Gross 1911, 190; Lewin 1962, 560 emend. mut. char. In: Buchanan RE, Gibbons NE (eds), *Bergey's Manual of Determinative Bacteriology, Eighth Edition*, The Williams and Wilkins Co., Baltimore, 1974, p. 109-111.
28. Lewin RA. Growth and nutrition of *Saprosira grandis* Gross (*Flexibacterales*). *Can J Microbiol* 1972; **18**:361-365. [PubMed](#)
<http://dx.doi.org/10.1139/m72-055>
29. Reichenbach H. The Genus *Saprosira*. In: Dworkin M, Falkow, S, Rosenberg, E, Schleifer, KH, Stackebrandt, E (eds). *The Prokaryotes*. 3rd ed. Volume 3. New York, NY: Springer; 2006. p 591-601.
30. BAuA. 2010, Classification of *Bacteria* and *Archaea* in risk groups. <http://www.baua.de TRBA 466>, p. 198.
31. 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>
32. 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>
33. 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>
34. List of growth media used at DSMZ:
<http://www.dsmz.de/catalogues/catalogue-microorganisms/culture-technology/list-of-media-for-microorganisms.html>.
35. The DOE Joint Genome Institute.
www.jgi.doe.gov
36. Phrap and Phred for Windows, MacOS, Linux, and Unix. www.phrap.com
37. 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>
38. 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.
39. 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.

40. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal Prokaryotic Dynamic Programming Genefinding Algorithm. *BMC Bioinformatics* 2010; **11**:119. [PubMed](#) <http://dx.doi.org/10.1186/1471-2105-11-119>
41. Pati A, Ivanova N, Mikhailova N, Ovchinnikova 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>
42. Markowitz VM, Ivanova NN, Chen IMA, Chu K, Kyrpides NC. IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* 2009; **25**:2271-2278. [PubMed](#) <http://dx.doi.org/10.1093/bioinformatics/btp393>
43. Gherna R, Woese CR. A partial phylogenetic analysis of the 'flavobacter-bacteroides' phylum: basis for taxonomic restructuring. *Syst Appl Microbiol* 1992; **15**:513-521. [PubMed](#) [http://dx.doi.org/10.1016/S0723-2020\(11\)80110-4](http://dx.doi.org/10.1016/S0723-2020(11)80110-4)
44. Auch AF, von Jan M, Klenk HP, Göker M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand Genomic Sci* 2010; **2**:117-134. [PubMed](#) <http://dx.doi.org/10.4056/sigs.531120>
45. 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. [PubMed](#) <http://dx.doi.org/10.4056/sigs.541628>
46. 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. [PubMed](#) <http://dx.doi.org/10.1089/bio.2010.0029>