Complete genome sequence of *Kangiella koreensis* type strain (SW-125[T])

Cliff Han[1,2], Johannes Sikorski[3], Alla Lapidus[1], Matt Nolan[1], Tijana Glavina Del Rio[1], Hope Tice[1], Jan-Fang Cheng[1], Susan Lucas[1], Feng Chen[1], Alex Copeland[1], Natalia Ivanova[1], Konstantinos Mavromatis[1], Galina Ovchinnikova[1], Amrita Pati[1], David Bruce[1,2], Lynne Goodwin[1,2], Sam Pitluck[1], Amy Chen[4], Krishna Palaniappan[4], Miriam Land[1,5], Loren Hauser[1,5], Yun-Juan Chang[1,5], Cynthia D. Jeffries[1,5], Patrick Chain[1,6], Elizabeth Saunders[1,2], Thomas Brettin[1,2], Markus Göker[3], Brian J. Tindall[3], Jim Bristow[1], Jonathan A. Eisen[1,7], Victor Markowitz[4], Philip Hugenholtz[1], Nikos C. Kyrpides[1], Hans-Peter Klenk[3*], and John C. Detter[1,2]

[1] DOE Joint Genome Institute, Walnut Creek, California, USA
[2] Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico, USA
[3] DSMZ - German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany
[4] Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[5] Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[6] Lawrence Livermore National Laboratory, Livermore, California, USA
[7] University of California Davis Genome Center, Davis, California, USA

*Corresponding author: Hans-Peter Klenk

Kangiella koreensis (Yoon *et al.* 2004) is the type species of the genus and is of phylogenetic interest because of the very isolated location of the genus *Kangiella* in the gammaproteobacterial order *Oceanospirillales*. *K. koreensis* SW-125[T] is a Gram-negative, non-motile, non-spore-forming bacterium isolated from tidal flat sediments at Daepo Beach, Yellow Sea, Korea. Here we describe the features of this organism, together with the complete genome sequence, and annotation. This is the first completed genome sequence from the genus *Kangiella* and only the fourth genome from the order *Oceanospirillales*. This 2,852,073 bp long single replicon genome with its 2647 protein-coding and 48 RNA genes is part of the **G**enomic **E**ncyclopedia of **B**acteria and **A**rchaea project.

Introduction

Strain SW-125[T] (= DSM 16069 = KCTC 12182 = JCM 12317) is the type strain of the species *Kangiella koreensis,* which is the type species of the tiny (two species containing) genus *Kangiella* [1]. This genus was only recently identified (2004) in the course of screening microorganisms from a tidal flat of the Yellow Sea in Korea. The genus is named *Kangiella* in order to honor Professor Kook Hee Kang, a Korean microbiologist, for his contribution to microbial research. The species name pertains to Korea, from where the strain was isolated [1]. Although many moderately halophilic or halotolerant bacteria have been isolated and characterized taxonomically from this habitat [1], literature on *Kangiella* is very limited. Presently,

the organism appears to be of interest solely for its position in the tree of life. Here we present a summary classification and a set of features for *K. koreensis* SW-125[T] together with the description of the complete genomic sequencing and annotation.

Classification and features

It is not evident from the taxonomic description of *K. koreensis* if any other strains beside SW-125[T] have been isolated from this species. Uncultured clones with high 16S rRNA gene sequence similarity to the sequence of strain SW-125[T] (AY520560) have been obtained from moderate saline crude oil contaminated soil in China (clone B109, 99%,

EU328030). The highest degree of similarity to sequences from environmental metagenomic libraries [2] was only 91% (As of June 2009).

Figure 1 shows the phylogenetic neighborhood of *K. koreensis* strain SW-125ᵀ in a 16S rRNA based tree. Analysis of the two identical 16S rRNA gene sequences in the genome of strain SW-125ᵀ differed

by two nucleotides from the previously published 16S rRNA sequence generated from DSM 16069 (AY520560). The slight difference between the genome data and the reported 16S rRNA gene sequence is most likely due to sequencing errors in the previously reported sequence data.
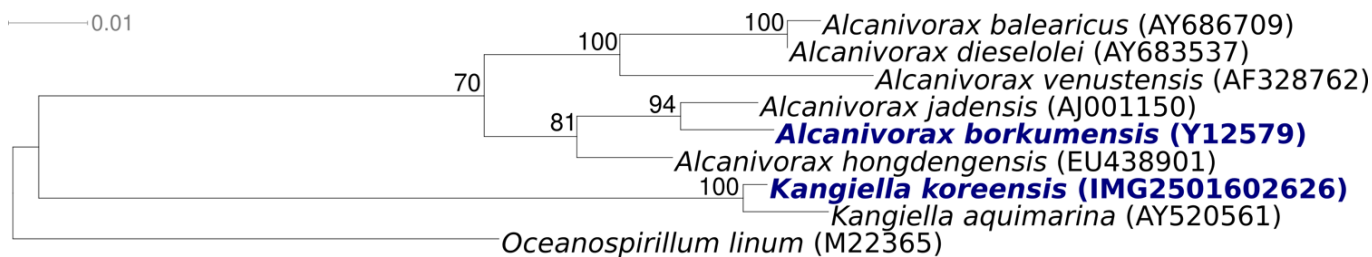
Figure 1. Phylogenetic tree highlighting the position of *K. koreensis* SW-125T relative to the other type strains in the phylogenetic neighborhood.. The tree was inferred from 1,476 aligned characters [3,4] of the 16S rRNA gene sequence under the maximum likelihood criterion [5], and rooted with the type strain of the order Oceanospirillales. 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%. Strains with a genome sequencing project registered in GOLD [6] are printed in blue; published genomes in bold.

K. koreensis cells are rods of 0.2-0.5 × 1.5-4.5 μm in size (Table 1 and Figure 2). The colonies are smooth, raised, circular to irregular, light yellowish-brown in color and 2.0–3.0 mm in diameter after seven days incubation at 30°C on marine agar 2216 (MA) (Difco) [1]. The following physiological features are from Yoon *et al.* [1]. The growth conditions have been explored in quite detail. The growth at various temperatures was determined after incubation for at least 15 days on marine agar 2216 (Difco). The optimal growth temperature was at 30-37°C, with a minimum temperature of 4°C and a maximum temperature of 43°C [1]. The conditions of growth in dependence of pH were determined in marine broth 2216 (Difco). The optimal pH is 7.0 – 8.0. Growth is still possible at pH 5.5, but not at pH 5.0 [1]. Growth at various NaCl concentrations (1–15 %) was investigated in MB or trypticase soy broth (TSB, Difco). The optimal growth occurs in the presence of 2-3% NaCl (MB), growth still occurs in the presence of 12% NaCl (MB), but not without NaCl (TSB) or in the presence of more than 13% NaCl (MB) [1]. Growth under anaerobic conditions occurs on MA supplemented with nitrate. Strain SW-125ᵀ hydrolyses casein, tyrosine, Tween 20, Tween 40 and Tween 60, but not hypoxanthine or

xanthine [1]. Furthermore, H_2S is not produced, and nitrate is not reduced under aerobic conditions but to nitrogen gas under anaerobic conditions [1]. Acid is not produced from the following sugars: adonitol, L-arabinose, D-cellobiose, D-fructose, D-galactose, D-glucose, lactose, maltose, D-mannitol, D-mannose, D-melezitose, melibiose, *myo*-inositol, D-raffinose, L-rhamnose, D-ribose, D-sorbitol, sucrose, D-trehalose or D-xylose [1]. Unfortunately, a list of carbon sources from which acid is produced is not delivered [1]. When assayed with the API ZYM system, alkaline phosphatase, esterase (C4), esterase lipase (C8), leucine arylamidase, valine arylamidase, trypsin and naphthol-AS-BI-phosphohydrolase are present, but lipase (C14), cystine arylamidase, α-chymotrypsin, acid phosphatase, α-galactosidase, β-galactosidase, β-glucuronidase, α-glucosidase, β-glucosidase, N-acetyl-β-glucosaminidase, α-mannosidase and α-fucosidase are absent [1]. Strain SW-125ᵀ was found to be susceptible to polymyxin (50 U), streptomycin (50 μg), penicillin (20 U), chloramphenicol (50 μg), ampicillin (10 μg), cephalothin (30 μg) and erythromycin (15 μg), and to be resistant to novobiocin (5 μg) and tetracycline (30 μg) [1].
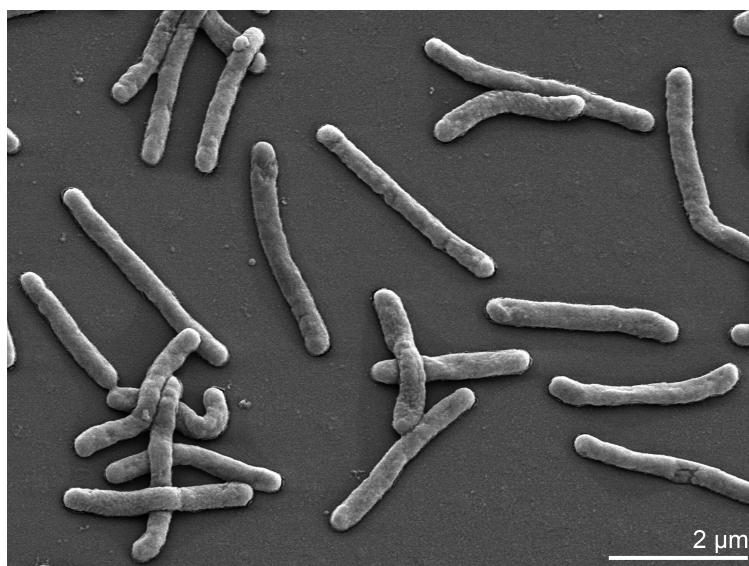
Figure 2. Scanning electron micrograph of *K. koreensis* SW-125[T] (Manfred Rohde, Helmholz Centre for Infection Research, Braunschweig).

Information on the composition of the peptidoglycan composition is unavailable. The predominant respiratory lipoquinone of *K. koreensis* SW-125[T] is the ubiquinone Q-8 (comprising approximately 84-88%) [1]. The fatty acids comprise iso-$C_{11:0}$ (5.6%), iso-$C_{13:0}$ (0.4%), iso-$C_{15:1}$ F (1.2%), iso-$C_{15:0}$ (57.6%), iso-$C_{16:0}$ (0.7%), $C_{16:0}$ (1.1%), iso-$C_{17:0}$ (7.2%), iso-$C_{17:1}$ ω 9c (8.6%), iso-$C_{11:0}$ 3-OH (10.5%), iso-$C_{17:1}$ ω9c (8.6%), iso-$C_{11:0}$ 3-OH(10.5%), iso-$C_{15:0}$ 3-OH (0.9%), iso-$C_{17:0}$ 3-OH (1.0%) and summed feature 1 (iso-$C_{15:1}$ and/or $C_{13:0}$ 3-OH) (3.2%) [1]. The predominance of iso-branched chain fatty acids indicates that the initial step in fatty acid synthesis is determined by an enzyme with a high degree of specificity for branched chain precursors (rather than acetate). The polar lipids of neither members of this species nor members of this genus have been investigated.

Table 1. Classification and general features of *K. koreensis* SW-125[T] according to the MIGS recommendations [7].

MIGS ID	Property	Term	Evidence code
		Domain *Bacteria*	TAS [8]
		Phylum *Proteobacteria*	TAS [9]
		Class *Gammaproteobacteria*	TAS [10,11]
	Current classification	Order *Oceanospirillales*	TAS [12,11
		Family *Incertae sedis*	NAS
		Genus *Kangiella*	TAS [1]
		Species *Kangiella koreensis*	TAS [1]
		Type strain SW-125	
	Gram stain	negative	TAS [1]
	Cell shape	rods, 0.2-0.5 × 1.5-4.5 μm	TAS [1]
	Motility	nonmotile	TAS [1]
	Sporulation	non-sporulating	TAS [1]
	Temperature range	4-43°C	TAS [1]
	Optimum temperature	30-37°C	TAS [1]

Table 1. Classification and general features of *K. koreensis* SW-125^T according to the MIGS recommendations (cont.) [7].

MIGS ID	Property	Term	Evidence code
	Salinity	requires 2-3% (w/v) NaCl, growth at 12% but not 13% NaCl	TAS [1]
MIGS-22	Oxygen requirement	aerobic and anaerobic growth	TAS [1]
	Carbon source	no specific information available	
	Energy source	peptone	TAS [1]
MIGS-6	Habitat	tidal flats	TAS [1]
MIGS-15	Biotic relationship	free living	NAS
MIGS-14	Pathogenicity	unknown	
	Biosafety level	1	TAS [13]
	Isolation	tidal flat sediment	TAS [1]
MIGS-4	Geographic location	Daepo Beach, Yellow Sea, Korea	TAS [1]
MIGS-5	Sample collection time	2004 or before	TAS [1]
MIGS-4.1 MIGS-4.2	Latitude, Longitude	33.245, 126.409	NAS
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 [14]. If the evidence code is IDA, then the property was observed for a living isolate by one of the authors or an expert mentioned in the acknowledgements.

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. The genome project is deposited in the Genomes OnLine Database [12] and the complete genome sequence 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
K. koreensis SW-125^T, DSM 16069, was grown in DSMZ medium 514 (BACTO Marine Broth) [15] at 28°C. DNA was isolated from 0.5-1 g of cell paste using Qiagen Genomic 500 DNA Kit (Qiagen, Hilden, Germany) following the manufacturer's protocol, but with a modification 'L' for cell lysis, as described in Wu *et al.* [16].

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 performed at the JGI can be found at the JGI website (http://www.jgi.doe.gov/). 454 Pyrosequencing reads were assembled using the Newbler assembler version 1.1.02.15 (Roche). Large Newbler contigs were broken into 3,167 overlapping fragments of 1,000 bp and entered into the 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 or transposon bombing of bridging clones [17]. Gaps between contigs were closed by editing in Consed, custom primer walk or PCR amplification. 329 Sanger finishing reads were produced to close gaps, to resolve repetitive regions, and to raise the quality of the finished sequence. The final assembly consists of 24,350 Sanger and 478,372 pyrosequence (454) reads. Together all sequence types provided 49.6x coverage of the genome. The error rate of the completed genome sequence is less than 1 in 100,000.

Table 2. Genome sequencing project information

| MIGS ID	Property	Term
MIGS-31	Finishing quality	Finished
MIGS-28	Libraries used	Two genomic libraries: one 8 kb pMCL200 Sanger library and one 454 pyrosequence standard library
MIGS-29	Sequencing platforms	ABI3730, 454 GS FLX
MIGS-31.2	Sequencing coverage	8.6x Sanger; 41× pyrosequence
MIGS-30	Assemblers	Newbler version 1.1.02.15, phrap
MIGS-32	Gene calling method	Prodigal
	INSDC ID	CP001707
	INSCD date of release	August 28, 2009
	GOLD ID	Gc01097
	INSDC project ID	29443
	Database: IMG-GEBA	2501533215
MIGS-13	Source material identifier	DSM 16069
	Project relevance	Tree of Life, GEBA

Genome annotation

Genes were identified using Prodigal [18] as part of the Oak Ridge National Laboratory genome annotation pipeline, followed by a round of manual curation using the JGI GenePRIMP pipeline (http://geneprimp.jgi-psf.org/) [19]. 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 functional annotation was performed within the Integrated Microbial Genomes Expert Review platform (http://img.jgi.doe.gov/er) [20].

Genome properties

The genome is 2,852,073 bp long and comprises one main circular chromosome with a 43.7% GC content. (Table 3 and Figure 3). Of the 2,695 genes predicted, 2,647 were protein coding genes, and 48 RNAs; 14 pseudogenes were also identified. The majority of the protein-coding genes (71.7%) were assigned 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

Attribute	Value	% of Total
Genome size (bp)	2,852,073	100.00%
DNA Coding region (bp)	2,585,246	90.64%
DNA G+C content (bp)	1,245,988	43.69%
Number of replicons	1	
Extrachromosomal elements	0	
Total genes	2,695	100.00%
RNA genes	48	1.78%
rRNA operons	2	
Protein-coding genes	2,647	98.22%
Pseudo genes	14	0.52%
Genes with function prediction	1,932	71.69%
Genes in paralog clusters	163	6.05%
Genes assigned to COGs	2,034	75.47%
Genes assigned Pfam domains	1,995	74.03%
Genes with signal peptides	691	25.64%
Genes with transmembrane helices	727	26.98%
CRISPR repeats	0	
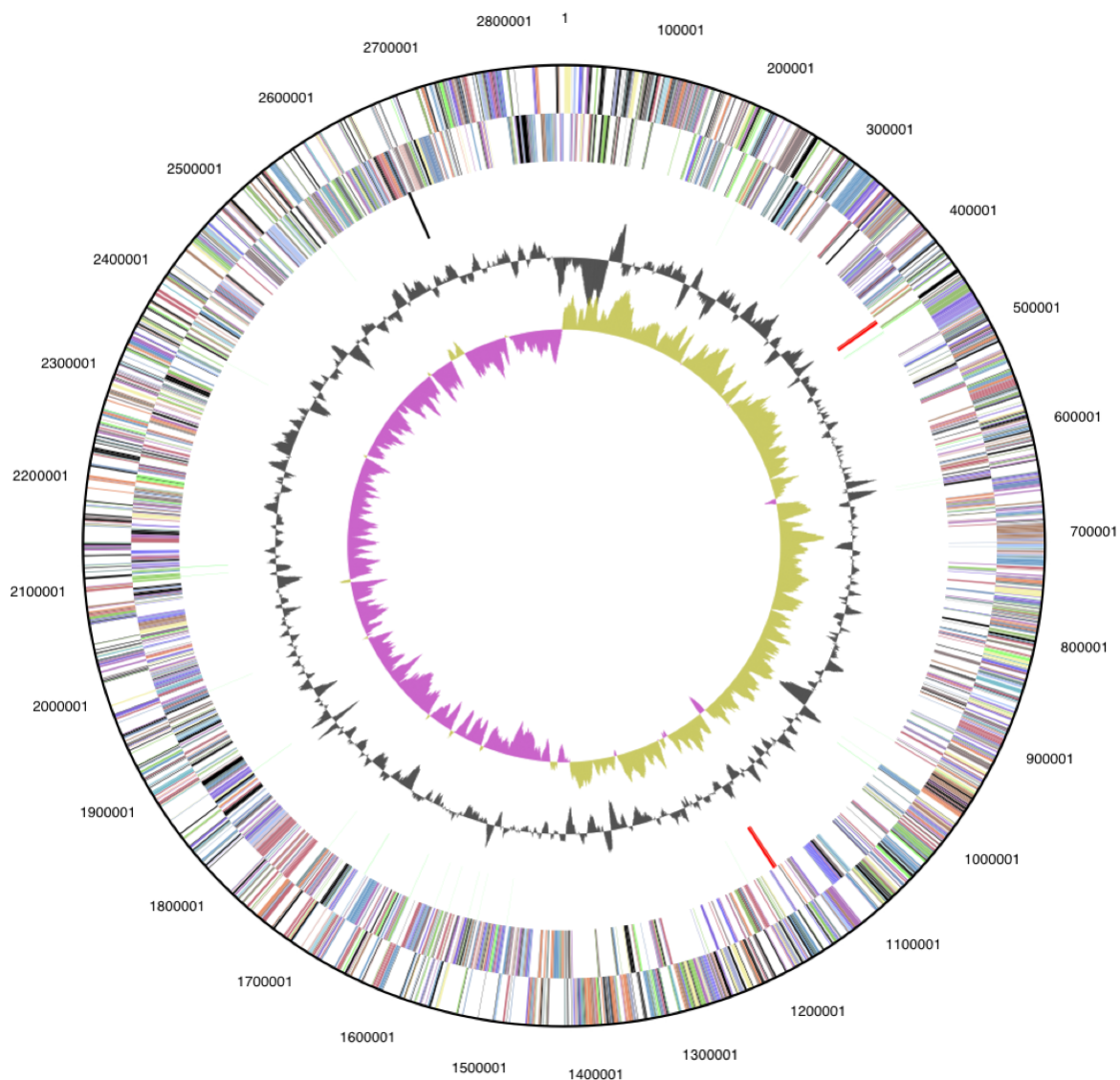
Figure 3. Graphical circular map of the genome. From outside to the center: Genes on forward strand (color by COG categories), Genes on reverse strand (color by COG categories), RNA genes (tRNAs green, rRNAs red, other RNAs black), GC content, GC skew.

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

Code	Value	% age	Description
J	170	6.4	Translation, ribosomal structure and biogenesis
A	1	0.1	RNA processing and modification
K	129	4.9	Transcription
L	106	4.0	Replication, recombination and repair
B	0	0.0	Chromatin structure and dynamics
D	29	1.1	Cell cycle control, mitosis and meiosis
Y	0	0.0	Nuclear structure
V	30	1.1	Defense mechanisms
T	134	5.1	Signal transduction mechanisms
M	139	5.3	Cell wall/membrane biogenesis

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

Code	Value	% age	Description
N	40	1.5	Cell motility
Z	1	0.0	Cytoskeleton
W	0	0.0	Extracellular structures
U	81	3.1	Intracellular trafficking and secretion
O	130	4.9	Posttranslational modification, protein turnover, chaperones
C	141	5.3	Energy production and conversion
G	41	1.5	Carbohydrate transport and metabolism
E	197	7.4	Amino acid transport and metabolism
F	54	2.0	Nucleotide transport and metabolism
H	118	4.5	Coenzyme transport and metabolism
I	84	3.2	Lipid transport and metabolism
P	113	4.3	Inorganic ion transport and metabolism
Q	53	5.3	Secondary metabolites biosynthesis, transport and catabolism
R	235	8.9	General function prediction only
S	223	8.4	Function unknown
-	613	23.2	Not in COGs

Acknowledgements

References

1. Yoon JH, Oh TK, Park YH. *Kangiella koreensis* gen. nov., sp. nov. and *Kangiella aquimarina* sp. nov., isolated from a tidal flat of the Yellow Sea in Korea. *Int J Syst Evol Microbiol* 2004; **54**:1829-1835. PubMed PubMed doi:10.1099/ijs.0.63156-0

2. Venter JC, Remington K, Heidelberg J, Halpern A, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, *et al*. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 2004; **304**:66-74. PubMed doi:10.1126/science.1093857

3. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. *Bioinformatics* 2002; **18**:452-464. PubMed doi:10.1093/bioinformatics/18.3.452

4. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000; **17**:540-552. PubMed

5. Stamatakis A, Hoover P, Rougemont J. A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Syst Biol* 2008; **57**:758-771. doi:10.1080/10635150802429642 PubMed doi:10.1080/10635150802429642

6. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 2008; **36**:D475-D479. PubMed PubMed doi:10.1093/nar/gkm884

7. 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 doi:10.1038/nbt1360

8. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl*

Acad Sci USA 1990; **87**: 4576-4579. PubMed doi:10.1073/pnas.87.12.4576

9. Woese CR, Stackebrandt E, Macke TJ, Fox GE. A phylogenetic definition of the major eubacterial taxa. *Syst Appl Microbiol* 1985; **6**: 143-151. PubMed

10. Garrity GM, Bell JA, Lilburn T. Class III. *Gammaproteobacteria* class. nov. *In*: DJ Brenner, NR Krieg, JT Staley, Garrity GM (eds), Bergey's Manual of Systematic Bacteriology, second edition, vol. 2 (The *Proteobacteria*), part B (The *Gammaproteobacteria*), Springer, New York, 2005, p. 1.

11. List Editor Validation of publication of new names and new combinations previously effectively published outside the IJSEM. List no. 106. *Int J Syst Evol Microbiol* 2005; **55**: 2235-2238. doi:10.1099/ijs.0.64108-0

12. Garrity GM, Bell JA, Lilburn T. Order VIII. *Oceanospirillales* ord. nov. *In*: DJ Brenner, NR Krieg, JT Staley, Garrity GM (eds), Bergey's Manual of Systematic Bacteriology, second edition, vol. 2 (The *Proteobacteria*), part B (The *Gammaproteobacteria*), Springer, New York, 2005, p. 270.

13. Anonymous. Biological Agents: Technical rules for biological agents www.baua.de TRBA 466.

14. 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 PubMed doi:10.1038/75556

15. List of growth media used at DSMZ: http://www.dsmz.de/microorganisms/ media_list.php

16. Wu M, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova N, Kunin V, Goodwin L, Wu M, Tindall BJ, *et al.*. A phylogeny-driven genomic encyclopedia of Bacteria and Archaea. (In press)

17. Sims D, Brettin T, Detter JC, Han C, Lapidus A, Copeland A, Glavina Del Rio T, Nolan M, Chen F, Lucas S, *et al*. Complete genome of *Kytococcus sedentarius* type strain (541[T]). *Stand Genomic Sci* 2009; **1**:12-20. doi:10.4056/sigs.761

18. Anonymous. Prodigal Prokaryotic Dynamic Programming Genefinding Algorithm. Oak Ridge National Laboratory and University of Tennessee 2009 http://compbio.ornl.gov/prodigal/

19. Pati A, Ivanova N, Mikhailova N, Ovchinikova G, Hooper SD, Lykidis A, Kyrpides NC. GenePRIMP: A Gene Prediction Improvement Pipeline for microbial genomes. (Submitted)

20. Markowitz VM. Mavromatis K, Ivanova NN, Chen I-MA, Chu K, Kyrpides NC. Expert Review of Functional Annotations for Microbial Genomes. *Bioinformatics* 2009; **25**:2271-2278 PubMed doi:10.1093/bioinformatics/btp393