Complete genome sequence of *Sanguibacter keddieii* type strain (ST-74^T)

Natalia Ivanova[1], Johannes Sikorski[2], David Sims[1], Thomas Brettin[1,3], John C. Detter[1,3], Cliff Han[1,3], Alla Lapidus[1], Alex Copeland[1], Tijana Glavina Del Rio[1], Matt Nolan[1], Feng Chen[1], Susan Lucas[1], Hope Tice[1], Jan-Fang Cheng[1], David Bruce[1,3], Lynne Goodwin[1,3], Sam Pitluck[1], Amrita Pati[1], Konstantinos Mavromatis[1], Amy Chen[4], Krishna Palaniappan[4], Patrik D'haeseleer[1,5], Patrick Chain[1,5], Jim Bristow[1], Jonathan A. Eisen[1,6], Victor Markowitz[4], Philip Hugenholtz[1], Markus Göker[2], Rüdiger Pukall[2], Hans-Peter Klenk[2], Nikos Kyrpides[1*]

[1] DOE Joint Genome Institute, Walnut Creek, California, USA
[2] DSMZ - German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany
[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, 94720, USA
[5] Lawrence Livermore National Laboratory, Livermore, California 94550, USA
[6] University of California Davis Genome Center, Davis, California, USA

Corresponding author: Nikos Kyrpides

Keywords
Blood isolate, aerobic, facultative anaerobic, *Sanguibacteraceae*, *Micrococcineae*

Abstract
Sanguibacter keddieii is the type species of the genus *Sanguibacter*, the only described genus within the family of *Sanguibacteraceae*. Phylogenetically, this family is located in the neighbourhood of the genus *Oerskovia* and the family *Cellulomonadaceae* within the actinobacterial suborder *Micrococcineae*. The strain described in this report was isolated from blood of apparently healthy cows. Here we describe the features of this organism, together with the complete genome sequence, and annotation. This is the first complete genome sequence of the family *Sanguibacteraceae*, and the 4,253,413 bp long single replicon genome with its 3735 protein-coding and 70 RNA genes is part of the ***G**enomic **E**ncyclopedia of **B**acteria and **A**rchaea* project.

Introduction
Strain ST-74^T (DSM 10542 = ATCC 51767 = JCM 11429 = NCIMB 703025 and other collections) is the type strain of the *Sanguibacter keddieii* species, and the type species of the genus *Sanguibacter* [1, Figure 1]. *S. keddieii* strain ST-74^T was isolated in 1995 by Fernandez-Garayzabal *et al*. from the blood of apparently healthy dairy cows in Spain [1] as the first member of the genus *Sanguibacter* and the family of *Sanguibacteraceae* [2]. On the basis of 16S rRNA sequence phylogeny, the small (six type strains) family *Sanguibacteraceae* is located in the neighbourhood to the genus *Oerskovia* [3] and the much larger micrococcineal families *Promicromonosporaceae* [2] and *Cellulomonadaceae* [2].

Like strain ST-74^T, two more type strains from the genus *Sanguibacter* (*S. suarezii* ST-26^T [1], and *S. inulinus* [4]) have been isolated from blood of cows. The type strains of the other

Sanguibacter species have been isolated from coastal sediment in the Eastern China Sea [5], from surface soil of a ginseng field in South Korea [6], from alpine subnival plants (DQ339590), and from a sea sand sample collected on the Weaver Peninsula on King George Island, Antarctica [7], which may suggest a global ecological versatility of this genus. Only two related but yet uncultivated phylotypes with more than 98.5% 16S rRNA sequence identity were reported from the gastrointestinal tract of pigs (AF371710), and from glacial meltwater at 6350 m on Mount Everest (EU584523), and no significant matches with any 16S rRNA sequences from environmental genomic samples and surveys are reported at the NCBI BLAST server (March 2009).

S. keddieii ST-74^T cells are facultatively anaerobic, Gram-positive, short, irregular shaped motile rods [1, Figure 2]. The colonies on tryptose soy agar (TSA, Difco) are circular, convex, with entire edges and yellow in color. Strain ST-74^T is Voges-Proskauer negative and does not reduce nitrate. Casein and gelatin are hydrolysed. Cellulose and Tween 80 are not hydrolysed. Acid is produced from a broad range of substrates: α-methyl-D-mannoside, α-methyl-D-glucoside, N-acetylglucosamine, amygdalin, rhamnose, D-rafinose, glycerol, L-arabinose, ribose, D-xylose, β-methyl-xyloside, galactose, glucose, fructose, D-mannose, rhamnose, arbutin, sorbitol, salicin, cellobiose, maltose, lactose, melibiose, sucrose, trehalose, raffinose, glycogen, β -gentibiose, turanose and lyxose [1]. The optimum growth temperature of strain ST-74^T is 25-30°C [1]; it grows at 35°C on agar [7] but not at 42°C [1].

Here we present a summary classification and a set of features for *S. keddieii* ST-74^T, (Table 1), together with the description of the complete genomic sequencing and annotation.

Classification and features

Figure 1 shows the phylogenetic neighborhood of *S. keddieii* strain ST-74^T in a 16S rRNA based tree. Analysis of the four 16S rRNA gene sequences in the genome of strain ST-74^T indicated that the genes differ by up to two nucleotides from each other, with two of the copies being identical with the previously published 16S rRNA sequence (X87755) generated from DSM 10542.

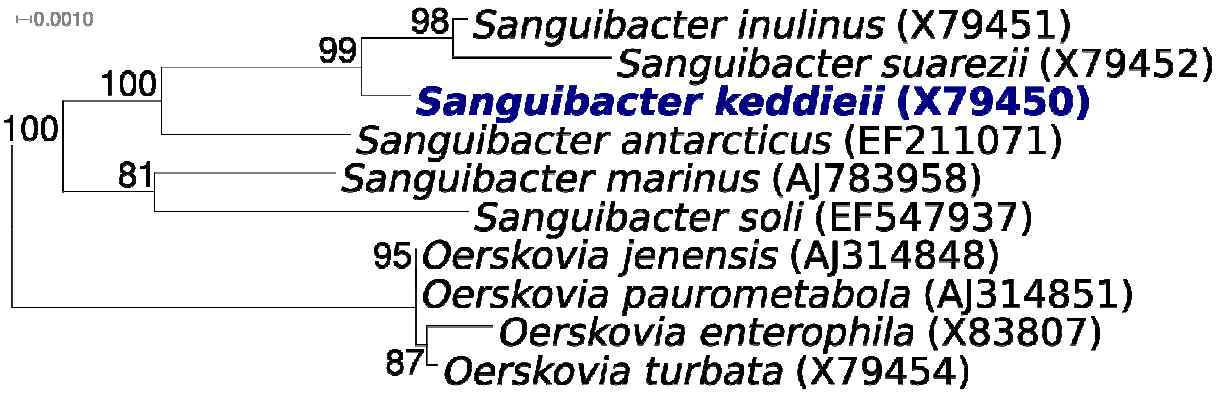
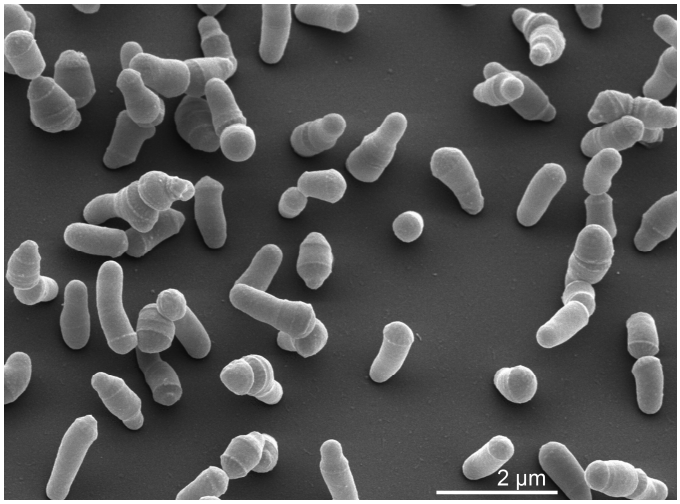


Figure 1. Phylogenetic tree of *S. keddieii* strain ST-74^T with all type strains of the family *Sanguibacteraceae*, inferred from 1468 aligned characters [8] of the 16S rRNA sequence under the maximum likelihood criterion [9, 10]. The tree was rooted with the type strains from the neighbour genus *Oerskovia*. The branches are scaled in terms of the expected number of substitutions per site. Numbers above branches are support values from 1000

bootstrap replicates if larger than 60%. Strains with a genome sequencing project registered in GOLD [11] are printed in blue; published genomes in bold.

Figure 2. Scanning electron micrograph of *S. keddieii* ST-74 ^{T} (M. Rohde, Helmholtz Centre for Infection Biology, Braunschweig)



Only little is known about the chemotaxonomy of strain ST-74^{T}. Saturated straight chain and branched chain are the major cellular fatty acids. In strain ST-74^{T} the straight chain fatty acids 16:0 (53.3%), 18:0 (10.1%), 14:0 (5.8%) predominate lower amounts of branched chain anteiso-15:0 (11.4%) and iso-16:0 (5.4%) fatty acids. This is in contrast to other species in the genus *Sanguibacter* and in the neighbouring taxa *Oerskovia* and *Cellulomonadaceae*, where branched chain fatty acids dominate [12]. Only traces of unsaturated acids, anteiso-15:1 (1.6%), and no mycolic acids were detected [1], as in the neighbouring taxa *Oerskovia* and *Cellulomonadaceae*. The murein of *S. keddieii* contains L-Lys-Ser-D-Glu, variation A4α [1], strikingly different from members of the genus *Oerskovia* and in members of the family *Cellulomonadaceae* [1]. Menaquinones are the sole respiratory lipoquinones present, with a partially saturated menaquinone containing nine-isoprene subunits MK-9(H_4) dominating [1]. The location of the points of unsaturation are in the 2nd and 3rd isoprene units, adjacent to the napthoquinone nucleus (MK-9 (II, III-H_4), in *O. turbata*. The phospholipid composition has not been reported, but phosphatidylglycerol, diphosphatidylglycerol, phosphatidylinositol, together with phosphoglycolipids have been reported in members of the neighbouring taxa *Oerskovia* and *Cellulomonadaceae* [12].

Table 1. Classification and general features of *S. keddieii* ST-74 ^{T} in accordance to the MIGS recommendations [13]

MIGS ID	Property	Term	Evidence code[a,b]
	Current classification	Domain *Bacteria*	TAS [2]
		Phylum *Actinobacteria*	TAS [2]
		Class *Actinobacteria*	TAS [14]
		Order *Actinomycetales*	TAS [1]
		Family *Sanguibacteraceae*	TAS [1]
		Genus *Sanguibacter*	
		Species *Sanguibacter keddieii*	
		Type strain ST-74	
	Gram stain	positive	TAS [1]
	Cell shape	short, irregular rods	TAS [1]
	Motility	motile	TAS [1]
	Sporulation	not reported	
	Temperature range	mesophilic	TAS [1]
	Optimum temperature	25-30°C	TAS [1]
	Salinity	not reported	
MIGS-22	Oxygen requirement	primarily aerobe; facultatively anaerobic; no nitrate reduction	TAS [1]
	Carbon source	broad variety of sugars	TAS [1]
	Energy source	carbohydrates	NAS
MIGS-6	Habitat	animal blood	TAS [1]
MIGS-15	Biotic relationship	free living	NAS
MIGS-14	Pathogenicity	none	NAS
	Biosafety level	2	TAS [15]
	Isolation	blood of apparently healthy cow	TAS [1]
MIGS-4	Geographic location		
MIGS-5	Sample collection time	before 1995	TAS [1]
MIGS-4.1 MIGS-4.2	Latitude – Longitude	not reported	
MIGS-4.3	Depth	not reported	
MIGS-4.4	Altitude	not reported	

a) 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 http://www.geneontology.org/GO.evidence.shtml of the Gene Ontology project [16]. If the evidence code is IDA, then the property should have been directly observed, for the purpose of this specific publication, for a live isolate by one of the authors, or an expert or reputable institution mentioned in the acknowledgements.

Genome sequencing

Genome project history

This organism was selected for sequencing on the basis of each phylogenetic position, and is part of the ***G**enomic **E**ncyclopedia of **B**acteria and **A**rchaea* project. The genome project is deposited in the Genome OnLine Database [11] and the complete genome sequence in

GenBank NOT YET. 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	Three genomic libraries: two Sanger libraries - 8 kb pMCL200 and fosmid pcc1Fos - and one 454 pyrosequence standard library
MIGS-29	Sequencing platforms	ABI3730, 454 GS FLX
MIGS-31.2	Sequencing coverage	10.4 x Sanger; 20 x pyrosequence
MIGS-30	Assemblers	Newbler version 1.1.02.15, phrap
MIGS-32	Gene calling method	Genemark 4.6b, tRNAScan-SE-1.23, infernal 0.81
	INSDC / Genbank ID	N/A
	Genbank Date of Release	N/A
	GOLD ID	Gi02151
	Database: IMG-GEBA	2500901759
	Project relevance	Tree of Life, GEBA

Growth conditions and DNA isolation

S. keddieii ST-74^T, DSM10542, was grown in DSMZ medium 92 (3% trypticase soy broth, 0.3% yeast extract; see http://www.dsmz.de/microorganisms/media_list.php) at 30°C. DNA was isolated from 1-1.5 g of cell paste using Qiagen Genomic 500 DNA Kit (Qiagen, Hilden, Germany) following the manufacturer's protocol, but with with an extended (one hour) incubation at 37°C for cell lysis with lysozyme and proteinase K.

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 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 4,746 overlapping fragments of 1000bp 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 [17] 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. 2,397 Sanger finishing reads were produced to close gaps, to resolve repetitive regions, and to raise the quality of the finished sequence. The error rate of the completed genome sequence is less than 1 in 100,000. Together all sequence types provided 30.4 x coverage of the genome.

Genome annotation

Genes were identified using GeneMark [18] as part of the genome annotation pipeline in the Integrated Microbial Genomes Expert Review (IMG-ER) system (http://img.jgi.doe.ogv/er)

[19], followed by a round of manual curation using JGI's GenePRIMP pipeline (http://geneprimp.jgi-psf.org) [20]. 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. The tRNAScanSE tool [21] was used to find tRNA genes, whereas ribosomal RNAs were found by using the tool RNAmmer [22]. Other non coding RNAs were identified by searching the genome for the Rfam profiles using INFERNAL (v0.81) [23]. Additional gene prediction analysis and manual functional annotation was performed within the Integrated Microbial Genomes (IMG) platform (http://img.jgi.doe.gov/) [24].

Metabolic network analysis

The metabolic Pathway/Genome Database (PGDB) was computationally generated using Pathway Tools software version 12.5 [25] and MetaCyc version 12.5 [26], based on annotated EC numbers and a customized enzyme name mapping file. It has undergone no subsequent manual curation and may contain errors, similar to a Tier 3 BioCyc PGDB [27].

Genome properties

The genome is 4,253,413 bp long and comprises one main circular chromosome with a 71.9% GC content (Table 3 and Figure 3). Of the 3805 genes predicted, 3735 were protein coding genes, and 70 RNAs. 25 pseudogenes were also identified. 74.4% of the genes were assigned with a putative function while the remaining are annotated as hypothetical proteins. The properties and the statistics of the genome are summarized in Table 3. The distribution of genes into COGs functional categories is presented in Table 4, and a cellular overview diagram is presented in Figure 4, followed by a summary of metabolic network statistics shown in Table 5.

Table 3. Genome Statistics

Attribute	Value	% of Total
Genome size (bp)	4,253,413	100.00%
DNA Coding region (bp)	3,872,139	91.04%
DNA G+C content (bp)	3,057,630	71.89%
Number of replicons	1	
Extrachromosomal elements	0	
Total genes	3805	100.00%
RNA genes	70	2.37%
rRNA operons	4	
Protein-coding genes	3735	98.16%
Pseudo genes	25	0.66%
Genes with function prediction	2832	74.43%
Genes in paralog clusters	501	13.17%
Genes assigned to COGs	2706	71.12%
Genes assigned Pfam domains	2785	73.19%
Genes with signal peptides	1126	29.59%
Genes with transmembrane helices	937	24.63%
CRISPR repeats	0	

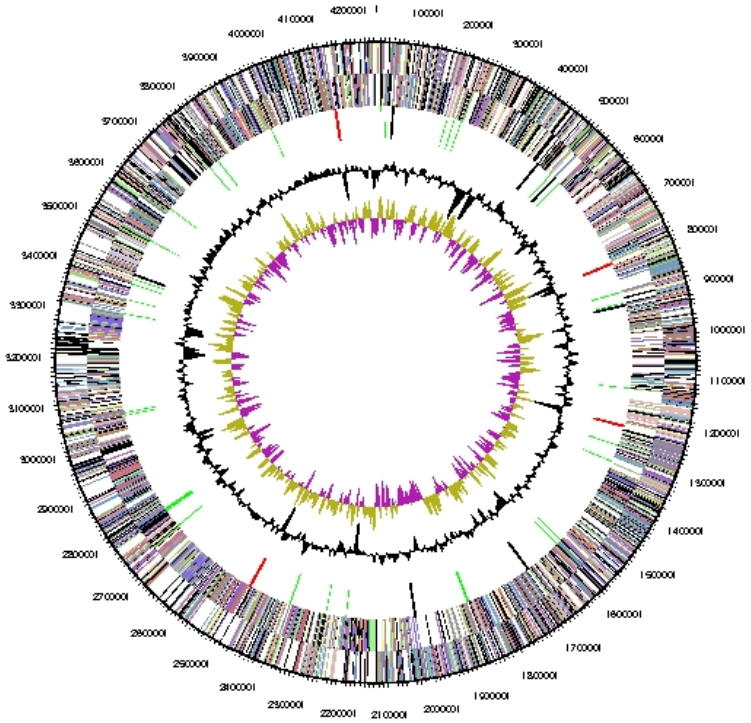


Figure 3. Graphical circular map of the genome. From outside to the center: Genes on forward strand (color by COG categories), Genes on reverse strand (color by COG categories), RNA genes (tRNAs green, sRNAs red, other RNAs black), GC content, GC skew

Table 4. Number of genes associated with the 21 general COG functional categories

Code	COG counts and percentage of protein-coding genes		Description
	Genome		
J	166	5.0	Translation
A	1	0.0	RNA processing and modification
K	317	10.0	Transcription
L	120	4.0	Replication, recombination and repair
B	1	0.0	Chromatin structure and dynamics
D	25	1.0	Cell cycle control, mitosis and meiosis
Y	0	0.0	Nuclear structure
V	69	2.0	Defense mechanisms
T	173	6.0	Signal transduction mechanisms
M	134	4.0	Cell wall/membrane biogenesis
N	55	2.0	Cell motility
Z	3	0.0	Cytoskeleton
W	0	0.0	Extracellular structures
U	41	1.0	Intracellular trafficking and secretion
O	84	3.0	Posttranslational modification, protein turnover, chaperones
C	174	6.0	Energy production and conversion

G	354	12.0	Carbohydrate transport and metabolism
E	237	8.0	Amino acid transport and metabolism
F	77	3.0	Nucleotide transport and metabolism
H	119	4.0.	Coenzyme transport and metabolism
I	80	3.0	Lipid transport and metabolism
P	199	7.0	Inorganic ion transport and metabolism
Q	43	1.0	Secondary metabolites biosynthesis, transport and catabolism
R	362	12.0	General function prediction only
S	213	7.0	Function unknown
-	1029	27.5	Not in COGs

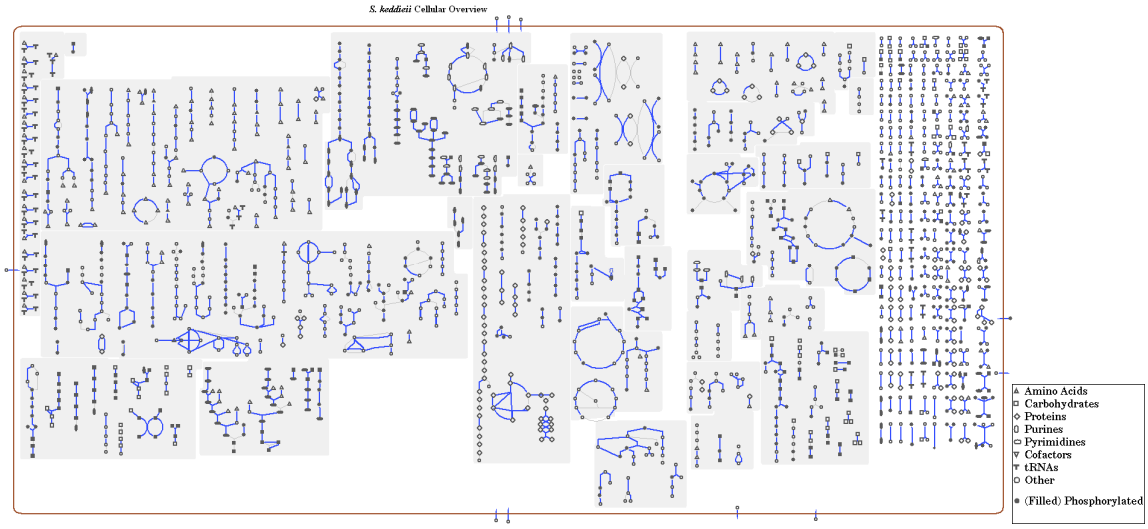


Figure 4. Schematic cellular overview diagram of all pathways of the *S. keddieii* ST-74^T metabolism. Nodes represent metabolites, with shape indicating class of metabolite (see key to right). Lines represent reactions.

Table 5. Metabolic Network Statistics

Attribute	Value
Total genes	3805
Enzymes	714
Enzymatic reactions	935
Metabolic pathways	205
Metabolites	676

Acknowledgements

We would like to gratefully acknowledge the help of Katja Steenblock for growing *S. keddieii* ST-74^T cultures, and Brian J. Tindall for chemotaxonomic advice (both 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 as well as German Research Foundation (DFG) INST 599/1-1.

References

1. Fernández-Garayzábal JF, Dominguez L, Pascual C, Jones D, Collins MD. Phenotypic and phylogenetic characterization of some unknown coryneform bacteria isolated from bovine blood and milk: description of *Sanguibacter* gen.nov. *Lett Appl Microbiol* 1995, **20**:69-75.

2. Stackebrandt E, Rainey FA, Ward-Rainey NL. Proposal for a new hierarchic classification system, *Actinobacteria* classis nov. *Intl J Syst Bacterial* 1997, **47**:479-91.

3. Prauser H, Lechevalier MP, Lechevalier HA. Description of *Oerskovia* gen. nov. to harbor Oerskovia's motile *Nocardia*. *Appl Microbiol* 1970, **19**:534.

4. Pascual C, Collins MD, Grimont PAD, Dominguez L, Fernandes-Garayzabal JF. *Sanguibacter inulinus* sp. nov. *Int J Syst Bacteriol* 1996, **46**:811-3.

5. Huang Y, Dai X, He L, Wang Y-N, Wang B-J, Liu Z, Liu S-J. *Sanguibacter marinus* sp. nov., isolated from coastal sediment. *Int J Syst Evol Microbiol* 2005, **55**:1755-8.

6. Kim MK, Pulla RK, Kim S-Y, Yi T-H, Soung N-K, Yang D-C. *Sanguibacter soli* sp. nov., isolated from soil of a ginseng field. *Int J Syst Evol Microbiol* 2008, **58**:538-41

7. Hong SG, Lee YK, Yim JH, Chun J, Lee HK. *Sanguibacter antarcticus* sp. nov., isolated from Antarctic sea sand. *Int J Syst Evol Microbiol* 2008, **58**:50-2.

8. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. *Bioinformatics* 2002, **18**:452-64.

9. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 1981, **17**:368-76.

10. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web-servers. *Syst Biol* 2008, **57**:758-71.

11. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC. The Genomes OnLine Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata *Nucleic Acids Res* 2008, **36**, D475-9.

12. Minnikin DE, Collins MD, Goodfellow M. Fatty acid and polar lipid composition in the classification of *Cellulomonas*, *Oerskovia* ad related taxa. *J Appl Bacteriol* 1979, **47**:87-9.

13. Field D, Garrity G, Gray T, Morrison N, Selengut J, et al. Towards a richer description of our complete collection of genomes and metagenomes: the “Minimum Information about a Genome Sequence” (MIGS) specification. *Nature Biotechnology* 2008, **26**:541-7.

14. Stackebrandt E, Schumann P. Description of *Bogoriellaceae* fam. nov., *Dermacoccaceae* fam. nov., *Rarobacteraceae* fam. nov. and *Sanguibacteraceae* fam. nov. and emendation of some families of the suborder *Micrococcineae*. *Intl J Syst Evol Micobiol* 2000, **50**: 1279-85.

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

16. 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 PMID:10802651 doi:10.1038/75556

17. Han CS, Chain P. Finishing repeat regions automatically with Dupfinisher. *In*: Proceeding of the 2006 international conference on bioinformatics & computational biology. Hamid R. Arabnia & Homayoun Valafar (eds), CSREA Press. June 26-29, 2006:141-6.

18. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 2001, **29**:2607-18.

19. Markowitz VM, Mavromatis K, Ivanova NN, Chen I-MA, Chu, K., and Kyrpides NC. Expert Review of Functional Annotations for Microbial Genomes. *Bioinformatics* 2009; in press. PMID: 19561336 doi:10.1093/bioinformatics/btp393

20. Pati A., et al. GenePRIMP: A Gene Prediction Improvement Pipeline for microbial genomes. *in preparation* 2009

21. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997, **25**:955-64. PMID:9023104 doi:10.1093/nar/25.5.955

22. Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 2007, **35**: 3100-3108. PMID:17452365 doi:10.1093/nar/gkm160

23. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* 2005, **33**:D121-4. PMID:15608160 doi:10.1093/nar/gki081

24. Markowitz, VM., Szeto, E., Palaniappan, K., Grechkin, Y., Chu, K., Chen, I-MA., Dubchak, I., Anderson, I., Lykidis, A., Mavromatis, K., Ivanova, NN., and Kyrpides, N.C. The Integrated Microbial Genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res* 2008, **36:**D528-33. PMID:17933782 doi:10.1093/nar/gkm846

25. Karp PD, Paley SM, Romero P. The Pathway Tools Software. *Bioinformatics* 2000, **18**:S225-32.

26. Karp P, Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley SM, Rhee SY, Shearer A, Tissier C, Walk TC, Zhang P. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of pathway/Genome Databases. *Nucleic Acids Res* 2008, **36**:D623-31. PMID:17965431 doi:10.1093/nar/gkm900

27. Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 2005, **33**:6083-9. PMID:16246909 doi:10.1093/nar/gki892