Complete genome sequence of *Catenulispora acidiphila* type strain (ID 139908^{T})

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

[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, CA, 94720, USA
[4] Lawrence Livermore National Laboratory, Livermore, California 94550, USA
[5] Oak Ridge National Laboratory, Oak Ridge, TN, 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
acidophilic, free-living, vegetative and aerial mycelia, filamentous actinomycete, non-pathogenic, aerobic, *Catenulisporineae*

Abstract
Catenulispora acidiphila Busti *et al*. 2006 is the type species of the genus *Catenulispora*, and is of interest because of the rather isolated phylogenetic location of the genomically little studied suborder *Catenulisporineae* within the order *Actinomycetales*. *C. acidiphilia* is known for its acidophilic, aerobic lifestyle, but can also grow scantly under anaerobic conditions. Under regular conditions *C. acidiphilia* grows in long filaments of relatively short aerial hyphae with marked septation. It is a free living, non motile, Gram-positive bacterium isolated from a forest soil sample taken from a wooded area in Gerenzano, Italy. 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 actinobacterial family *Catenulisporaceae*, and the 10,467,782 bp long single replicon genome with its 9056 protein-coding and 69 RNA genes is a part of the ***G**enomic **E**ncyclopedia of **B**acteria and **A**rchaea* project.

Introduction
Catenulispora acidiphila strain ID 139908^{T} (DSM 44928 = NRRL B-24433 = JCM 14897) is the type species of the genus *Catenulispora* which is the type genus of family *Catenulisporaceae* as well as of the suborder *Catenulisporineae* [1]. The *Catenulisporacineae* are a rather small (six genera in two families) and young taxon [2, Fig. 1], for which so far no completed genome sequence has been reported. All four *Catenulispora* type strains were

isolated from paddy field or forest soil, prefer slightly acidic habitats, and form vegetative and aerial mycelia [1, 3-4].

Strains most probably belonging to the species *C. acidiphila* are also known from diversity studies performed on isolates collected from soils of various geographic origin: the 'Neo' strains from Italian and South American soils (Neo 1, 2, 6, 9, 15) as described by Busti *et al.* 2006 [5], several isolates from Ellinbank, Australia, (Ellin 5034, 5116, 5119) as described by Joseph *et al*. 2003 [6], and a Korean isolate D8-90T (AM690741), all of which share at least 99.3% 16S rRNA gene sequence identity with strain ID 139908^T. None of the samples sequenced in environmental genomic survey and screening programs surpassed 92% sequence similarity with strain ID 139908^T, indicating a lack of close links of these phylotypes to the species *C. acidiphila* or the genus *Catenulispora*.

Here we present a summary classification and a set of features for *C. acidiphila* ID 139908^T (Tab. 1), together with the description of the complete genomic sequencing and annotation.

Classification and features of organism

Fig. 1 shows the phylogenetic neighborhood of *C. acidiphila* strain ID 139908^T in a 16S rRNA based tree. All three 16S rRNA gene copies in the genome of strain D 139908^T are identical, and also match the previously published 16S rRNA sequence (AJ865857) generated from DSM 20547.

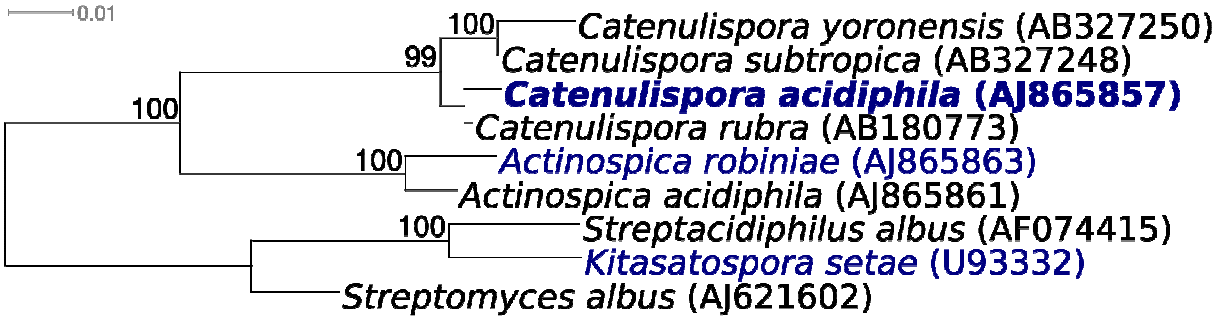
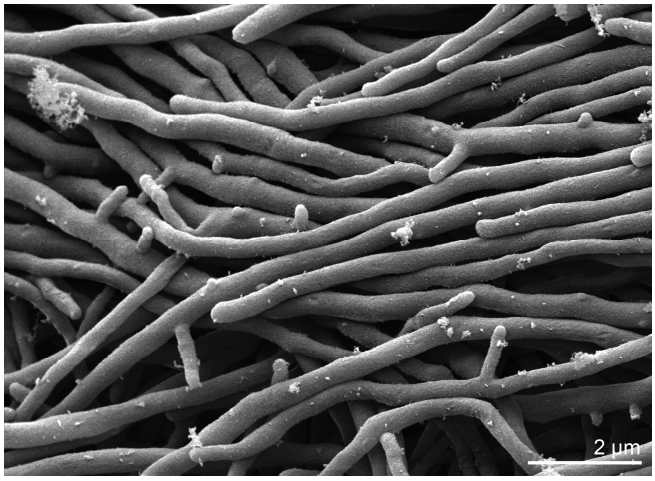


Figure 1. Phylogenetic tree of *C. acidiphila* ID 139908^T and all type strains of the genus *Catenulispora*, inferred from 1421 aligned characters [7, 8] of the 16S rRNA sequence under the maximum likelihood criterion [9]. The tree was rooted with the type strains of the genera within the *Streptomycetaceae* (*Streptomycineae*, *Actinomycetales*). Also included are the type strains from the sister family of *Catenulisporaceae*, *Actinospicaceae*. 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 [10] are printed in blue; published genomes in bold.

C. acidiphila strain ID 139908^T was described as Gram-positive, acidophilic, non-acid fast, non-motile, essentially aerobic bacterium forming both vegetative and aerial mycelia [1, Fig. 2]. Non-fragmentary vegetative mycelium and aerial hypha are straight to slighly flexuous and start to septate in chains of cylindrical arthrospores with a rugose surface when sporulation is induced [1]. Strain ID 139908^T grows on different agar media while producing brownish pigments and whitish aerial mass which turned to yellow/green with the age of bacteria [1]. The brownish pigments were not observed on tyrosine-supplemented Suter medium which indicated that they are not melanin-related [1]. The strain grows well in the presence of 3% (w/v) NaCl with a progressive reduction of pigmentation which started at 1% NaCl. Strain ID 139908^T grows better under aerobic conditions but is capable of reduced and non pigmented growth under microaerophilic and anaerobic conditions [1]. It is resistant to

lysozyme (at least 100µg/ml) [1] which was not reported for any of the strains of the genus *Catenulispora*. Optimum temperature for growth was 22-28°C and the pH for growth ranges from 4.3 to 6.8 with an optimum pH level 6.0 but scant growth was reported up to pH 7.5 [1]. The organism is capable to hydrolyse starch and casein, liquefy gelatine, and to utilize D-galactose, D-fructose, arabinose, xylose and gluconate but not glycerol, L-arabinose, D-mannitol, methyl-β-D-xylopyranoside, methyl-α-D-glucopyranoside, cellulose and sucrose [1].

Figure 2. Scanning electron micrograph of *C. acidiphila* strain ID 139908^T (M. Rohde, HZI Braunschweig)



Chemotaxonomy. Like the other *Catenulispora* strains [3, 4], the mureine of *C. acidiphila* strain ID 139908^T contains LL-diaminopimelic acid, glycine, glutamic acid and alanine [1] and can be assigned to type A3γ LL-Dpm–Gly (http://www.dsmz.de/species/murein.htm). Whole cell sugars contains large amounts of arabinose, together with xylose, ribose, rhamnose and glucose [1]. Strain ID 139908^T contains predominantly menaquinones with nine isoprene units: MK-9(H_6), -9(H_4), and MK -9(H_8) in a ratio of 4.5:2.8:1 [1], as also reported for other members of the genus [3, 4]. As in *C. rubra* [3] and in *C. subtopica* and *C. yoronensis* [4], the major cellular fatty acids are iso- (i-) and anteiso- (ai-) branched saturated acids: i-$C_{16:0}$ (47.1%) and ai-$C_{17:0}$ (12.7%), with smaller amounts of i-$C_{17:0}$ (5.7%), $C_{16:0}$ (5.6%), i-$C_{17:1}$ ω 9c (4.7%), i-$C_{15:0}$ (4.3%), i-$C_{16:1}$ (3.4%), $C_{16:1}$ω7c (3.2%), ai-$C_{17:1}$ ω 9c (2.8%), ai-$C_{15:0}$ (2.3%) [1]. Phosphatidylglycerol, diphosphatidylglycerol, phosphatidyl-inositol, phosphatidylinositol mannosides were identified as the dominating polar lipids together with two unknown phospholipids [1].

Table 1. Classification and general features of *C. acidiphila* strain ID 139908^T in accordance to the MIGS recommendations [11]

MIGS ID	Property	Term	Evidence code[a,b]
	Current classification	Domain *Bacteria*	
		Phylum *Actinobacteria*	TAS [12]
		Class *Actinobacteria*	TAS [13]
		Order *Actinomycetales*	TAS [13]
		Suborder *Catenulisporineae*	TAS [2]
		Family *Catenulisporaceae*	TAS [1]

		Genus *Catenulispora*	TAS [1]
		Species *Catenulispora acidipila*	TAS [1]
		Type strain ID 139908	TAS [1]
	Gram stain	positive	TAS [1]
	Cell shape	non-fragmentary vegetative mycelium	TAS [1]
	Motility	nonmotile	TAS [1]
	Sporulation	produces arthrospores when induced	TAS [1]
	Temperature range	mesophile, 11-37°C	TAS [1]
	Optimum temperature	22-28°C	TAS [1]
	Salinity	3% NaCl	TAS [1]
MIGS-22	Oxygen requirement	essentially aerobic; capable of reduced and non-pigmented growth under microaerophilic and anaerobic conditions	TAS [1]
	Carbon source	glucose, arabinose, xylose, manitol, fructose, glycerol	TAS [1]
	Energy source	starch	NAS
MIGS-6	Habitat	soil	TAS [1]
MIGS-15	Biotic relationship	unknown	
MIGS-14	Pathogenicity	none	NAS
	Biosafety level	1	TAS [14]
	Isolation	forest soil from wooden area	TAS [2]
MIGS-4	Geographic location	Gerenzano, Italy	TAS [2]
MIGS-5	Sample collection time	before 2006	TAS [1]
MIGS-4.1 MIGS-4.2	Latitude – Longitude	not reported	TAS [2]
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 [15]. 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 and annotation information

Genome project history

This organism was selected for sequencing on the basis of its 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 Genomes OnLine Database [10] and the complete genome sequence in GenBank NOT YET. Sequencing, finishing and annotation was 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	Two Sanger libraries - 8 kb pMCL200 and fosmid pcc1Fos
MIGS-29	Sequencing platforms	ABI3730
MIGS-31.2	Sequencing coverage	10x Sanger
MIGS-30	Assemblers	phrap
MIGS-32	Gene calling method	Prodigal
	INSDC / Genbank ID	not yet available
	Genbank Date of Release	not yet available
	GOLD ID	Gi02233
	NCBI project ID	21085
	Database: IMG-GEBA	2501533203
	Project relevance	Tree of Life, GEBA

Growth conditions and DNA isolation
C. acidiphila strain ID 139908^{T}, DSM 44928, was grown in DSMZ medium 987 (ISP2 Medium, http://www.dsmz.de) at 28°C. DNA was isolated from 0.5-1 g of cell paste using the JGI CTAB protocol; cells were lysed with 500 µl achromopeptidase, lysostaphin, and mutanolysin, each, during over night incubation on a shaker at 35°C.

Genome sequencing and assembly
The genome was sequenced using Sanger sequencing platform only. All general aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov, The Phred/Phrap/Consed software package (www.phrap.com) was used for sequence assembly and quality assessment. After the shotgun stage, reads were assembled with parallel phrap (High Performance Soft ware, LLC). Possible mis-assemblies were corrected with Dupfinisher [16] 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 (Roche Applied Science, Indianapolis, IN). A total of 2556 finishing reactions were produced to close gaps and to raise the quality of the finished sequence. The completed genome sequences of *C. acidiphila* contains 126,099 Sanger reads, achieving an average of 10x sequence coverage per base with an error rate less than 1 in 100,000.

Genome annotation
Genes were identified using Prodigal [17] as part of the Oak Ridge National Laboratory genome annotation pipeline, followed by a round of manual curation using JGI's GenePRIMP pipeline (http://geneprimp.jgi-psf.org) [18]. 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 (IMG-ER) platform (http://img.jgi.doe.gov/er) [19].

Genome properties

The genome is 10,467,782 bp long and comprises one circular chromosome with a 69.8% GC content (Tab. 3). Of the 9125 genes predicted, 9056 were protein coding genes, and 69 RNAs; 142 pseudogenes were also identified. 68.2% of the genes were assigned with a putative

function while the remaining are annotated as hypothetical proteins. The distribution of genes into GOGs functional categories is presented in Table 4.

Table 3. Genome Statistics

Attribute	Value	% of Total
Genome size (bp)	10,467,782	
DNA Coding region (bp)	9,386,056	89.67%
DNA G+C content (bp)	7,303,066	69.77%
Number of replicons	1	
Extrachromosomal elements	0	
Total genes	9125	100.00%
RNA genes	69	0.76%
rRNA operons	3	
Protein-coding genes	9056	99.24%
Pseudo genes	142	1.56%
Genes with function prediction	6219	68.15%
Genes in paralog clusters	2379	26.07%
Genes assigned to COGs	5805	63.62%
Genes assigned Pfam domains	6202	67.97%
Genes with signal peptides	2279	24.98%
Genes with transmembrane helices	2231	24.45%
CRISPR repeats	4	

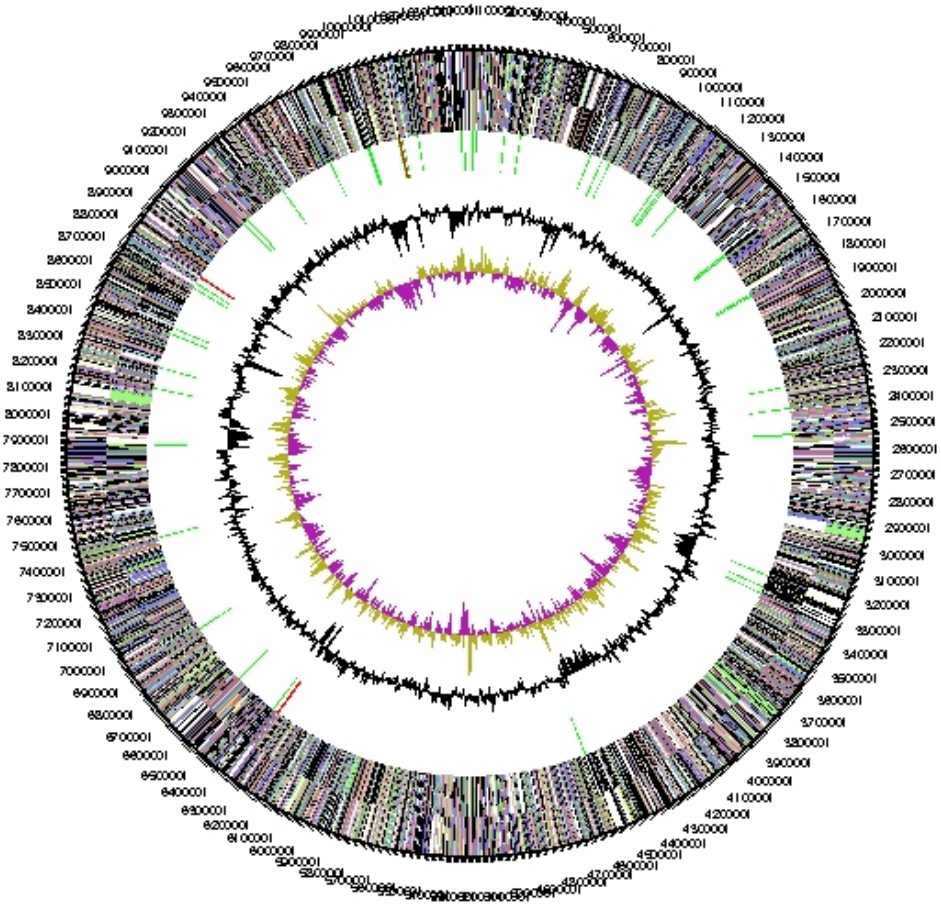


Figure 3. Graphical circular map of the genome. From outside to the center: Genes on forward strand (color by COG categories), Genes on reverse strand (color by COG categories), RNA genes (tRNAs green, sRNAs red, other RNAs black), GC content, GC skew.

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

Code	COG counts and percentage of protein-coding genes		Description
	Genome		
	value	% of total	
J	182	2.0	Translation, ribosomal structure and biogenesis
A	2	0.0	RNA processing and modification
K	607	6.7	Transcription
L	173	1.9	Replication, recombination and repair
B	2	0.0	Chromatin structure and dynamics
D	34	0.4	Cell cycle control, mitosis and meiosis
Y	0	0.0	Nuclear structure
V	96	1.1	Defense mechanisms

T	389	4.3	Signal transduction mechanisms
M	210	2.3	Cell wall/membrane biogenesis
N	45	0.5	Cell motility
Z	1	0.0	Cytoskeleton
W	0	0.0	Extracellular structures
U	46	0.5	Intracellular trafficking and secretion
O	149	1.6	Posttranslational modification, protein turnover, chaperones
C	306	3.4	Energy production and conversion
G	441	4.9	Carbohydrate transport and metabolism
E	425	4.7	Amino acid transport and metabolism
F	108	1.2	Nucleotide transport and metabolism
H	223	2.5	Coenzyme transport and metabolism
I	226	2.5	Lipid transport and metabolism
P	241	2.7	Inorganic ion transport and metabolism
Q	265	2.9	Secondary metabolites biosynthesis, transport and catabolism
R	670	7.4	General function prediction only
S	328	3.6	Function unknown
-	3251	35.9	Not in COGs

Acknowledgements

We gratefully acknowledge the help of Marlen Jando for growing *C. acidiphila* cultures and Susanne Schneider for DNA extraction and quality analysis (both at DSMZ). This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396, as well as German Research Foundation (DFG) INST 599/1-1.

References

1. Busti E, Cavaletti L, Monciardini P, Schumann P, Rohde M, Sosio M, Donadio S. *Catenulispora acidiphila* gen. nov., sp. nov., a novel, mycelium-forming actinomycete, and proposal of *Catenulisporaceae* fam. nov. *Int J Syst Evol Bacteriol* 2006, **56**:1741-6.

2. Cavaletti L, Monciardini P, Schumann P, Rohde M, Bamonte R *et al*. *Actinospica robiniae* gen. nov., sp. nov. and *Actinospica acidiphila* sp. nov.: proposal for *Actinospicaceae* fam. nov. and *Catenulisporinae* subord. nov. in the order *Actinomycetales*. *Int J Syst Evol Microbiol* 2006, **56:**1747–53.

3. Tamura T, Ishida Y, Sakane T. Suzuki K. (2007). *Catenulispora rubra* sp. nov., an acidophilic actinomycete isolated from forest soil. *Int J Syst Evol Microbiol* 2007, 57**:**2272–4.

4. Busti E, Monciardini P, Cavaletti L, Bamonte R, Lazzarini A, Sosio M, Donadio S. Antibiotic-producing ability by representatives of a newly discovered lineage of actinomycetes. *Microbiology* 2006, **152:**675-83.

5. Tamura T., Ishida Y**.,** Otoguro M, Suzuki, K. *Catenulispora subtropica* sp. nov. and *Catenulispora* ***yoron****ensis* sp. nov. *Int J Syst Evol Microbiol* 2008, **58:**1552-5.

6. Joseph SJ, Hugenholtz P, Sangwan P, Osborne CA, Janssen PH. Laboratory cultivation of widespread and previously uncultured soil bacteria. *Appl Environ Microbiol* 2003, **69:**7210-5.

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

8. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000, **17**:540-52.

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

10. 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.

11. 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. *Nat Biotechnol* 2008, **26**:541-7.

12. Cavalier-Smith T. The neomuran origin of actinobacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int J Syst Bacteriol* 2002, **52:**7-76.

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

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

15. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat Genet* 2000, **25**:25-9.

16. Sims D, Brettin T, Detter JC, Han C, Lapidus A, *et al*. Complete genome of *Kytococcus sedentarius* type strain (strain 541^{T}). *SIGS*, 2009 reviewed.

17. http://compbio.ornl.gov/prodigal/

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

19. Markowitz VM, Mavromatis K, Ivanova NN, Chen I-MA, Kyrpides NC. Expert Review of Functional Annotations for Microbial Genomes. *Bioinformatics* 2009, in press.