

Lawrence Berkeley National Laboratory

Lawrence Berkeley National Laboratory

Title

Complete genome sequence of Kytococcus sedentarius type strain (strain 541T)

Permalink

<https://escholarship.org/uc/item/2d50t3rd>

Author

Sims, David

Publication Date

2008-08-15

Peer reviewed

Complete genome sequence of *Kytococcus sedentarius* type strain (541^T)

David Sims¹, Thomas Brettin^{1,2}, John C. Detter¹, Cliff Han¹, Alla Lapidus², Alex Copeland², Tijana Glavina Del Rio², Matt Nolan², Feng Chen¹, Susan Lucas², Hope Tice², Jan-Fang Cheng², David Bruce¹, Lynne Goodwin¹, Sam Pitluck², Galina Ovchinnikova², Amrita Pati², Natalia Ivanova², Konstantinos Mavromatis², Amy Chen³, Krishna Palaniappan³, Patrik D'haeseleer^{2,4}, Patrick Chain^{2,4}, Jim Bristow², Jonathan A. Eisen^{2,5}, Victor Markowitz³, Philip Hugenholtz², Susanne Schneider⁶, Markus Göker⁶, Rüdiger Pukall⁶, Nikos C. Kyrpides², and Hans-Peter Klenk^{6*}

¹ Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico, USA

² DOE Joint Genome Institute, Walnut Creek, California, USA

³ Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California, USA

⁴ Lawrence Livermore National Laboratory, Livermore, California, USA

⁵ University of California Davis Genome Center, Davis, California, USA

⁶ DSMZ - German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany

*Corresponding author: [Hans-Peter Klenk](mailto:hpk@dkfz.de)

Keywords: mesophile, free-living, marine, aerobic, opportunistic pathogenic, *Dermacoccaceae*

Kytococcus sedentarius (ZoBell and Upham 1944) Stackebrandt *et al.* 1995 is the type strain of the species, and is of phylogenetic interest because of its location in the *Dermacoccaceae*, a poorly studied family within the actinobacterial suborder *Micrococccineae*. *K. sedentarius* is known for the production of oligoketide antibiotics as well as for its role as an opportunistic pathogen causing valve endocarditis, hemorrhagic pneumonia, and pitted keratolysis. It is strictly aerobic and can only grow when several amino acids are provided in the medium. The strain described in this report is a free-living, nonmotile, Gram-positive bacterium, originally isolated from a marine environment. Here we describe the features of this organism, together with the complete genome sequence, and annotation. This is the first complete genome sequence of a member of the family *Dermacoccaceae* and the 2,785,024 bp long single replicon genome with its 2639 protein-coding and 64 RNA genes is part of the *Genomic Encyclopedia of Bacteria and Archaea* project.

Introduction

Strain 541^T (DSM 20547 = ATCC 14392 = JCM 11482 = CCM 314 and other culture collections) is the type strain of the species *Kytococcus sedentarius*, which is the type species of the genus *Kytococcus* [1]. Strain 541^T was first described as *Micrococcus sedentarius* (ZoBell and Upham 1944) [2] and later emended as *K. sedentarius* in a taxonomic dissection of the genus *Micrococcus* [1]. The organism is of interest for its biotechnological potential as source of natural antibiotics (oligoke-

tides), for its role as an opportunistic pathogen, and for its position in the tree of life, where it represents the scarcely populated genus *Kytococcus* (2 species) within in the actinobacterial family *Dermacoccaceae* [1] (Figure 1). *K. sedentarius* 541^T was first isolated around 1944 from a marine environment [2], but strains of the species were also frequently isolated from human skin [3]. More recently, closely related strains were also isolated from culture-dependant environmental screenings

of a non-saline alkaline groundwater environment in Cabeco de Vide in southern Portugal [4], screening for pelagic bacteria in South Korea [5], tropical marine sediments from the intertidal zone off the coast of the Republic Palau [6], from the ciliate *Collinia* sp.), an endoparasite of euphausiids from the Gulf of California (unpublished literature, GenBank record EU090136), and in a culture-independent analysis of the microbial burden and diversity in commercial airline cabins [7]. Screening of environmental genomic samples and surveys reported at the NCBI BLAST server indicated no closely related phylotypes that can be linked to the species or genus. Here we present a summary classification and a set of features for *K. sedentarius* strain 541^T (Table 1), together with the description of the complete genomic sequencing and annotation.

Classification and features

K. sedentarius cells are spherical/coccoid and occur predominantly in tetrads which can be arranged in cubical packets [1] (Figure 2). Cells are described as Gram-positive, nonmotile, non-

encapsulated, and not endospore-forming [1]. *K. sedentarius* 541^T is strictly aerobic and chemoor-ganotrophic, requires methionine and other amino acids for growth, and grows well in NaCl at concentrations up to 10% (w/v) [1].

K. sedentarius (strain NK0508) is capable of degrading diphenylarsenic acid [8], but not starch [1], and does not produce acids from most carbohydrates and alcohols [1]. Its optimal growth temperature is 28-36°C. Nitrate is reduced to nitrite by some *K. sedentarius* strains [1]. *K. sedentarius* is not only described as the source of the oligoketide antibiotics monensin A and B [9], but has also been associated with pitted keratolysis [10], opportunistic infections, and fatal hemorrhagic pneumonia [11].

Figure 1 shows the phylogenetic neighborhood of *K. sedentarius* strain 541^T in a 16S rRNA based tree. Analysis of the 16S rRNA gene copies in the genome of strain 541^T differed by one nucleotide from each other, and by up to two nucleotides from the previously published 16S rRNA sequence generated from DSM 20547 (X87755).

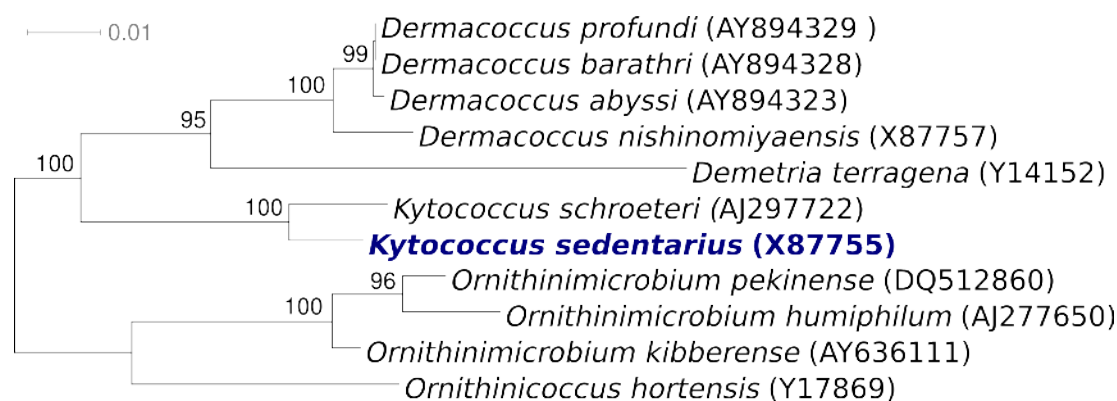


Figure 1. Phylogenetic tree of *K. sedentarius* strain 541^T with all type strains of the family *Dermacoccaceae*, inferred from 1,456 aligned 16S rRNA characters [12] under the maximum likelihood criterion [13, 14]. The tree was rooted with four members of the neighboring family *Intrasporangiaceae*. 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. Strains with a genome-sequencing project registered in GOLD [15] are printed in blue; published genomes in bold.

Chemotaxonomy

The murein of *K. sedentarius* strain 541^T contains L-Lys-Glu₂, a variation of cell wall type A4α [1]. Mycolic acids and teichonic acids were not reported [1]. Strain 541^T contains only completely unsaturated menaquinones with 8-11 isoprene

subunits (MK8 to MK11), with MK8 dominating [1]. The major cellular fatty acids are methyl-branched chain iso-C_{17:1} and anteiso-C_{17:0}, as well as the straight chain saturated C_{15:0} and C_{17:0} [1]. Phosphatidylglycerol, diphosphatidylglycerol, and

phosphatidylinositol were identified as dominating polar lipids [1]. Reported cytochromes include

aa₃, c₆₂₆, c₅₅₀, b₅₅₇, b₅₆₁, and b₅₆₄ [1].

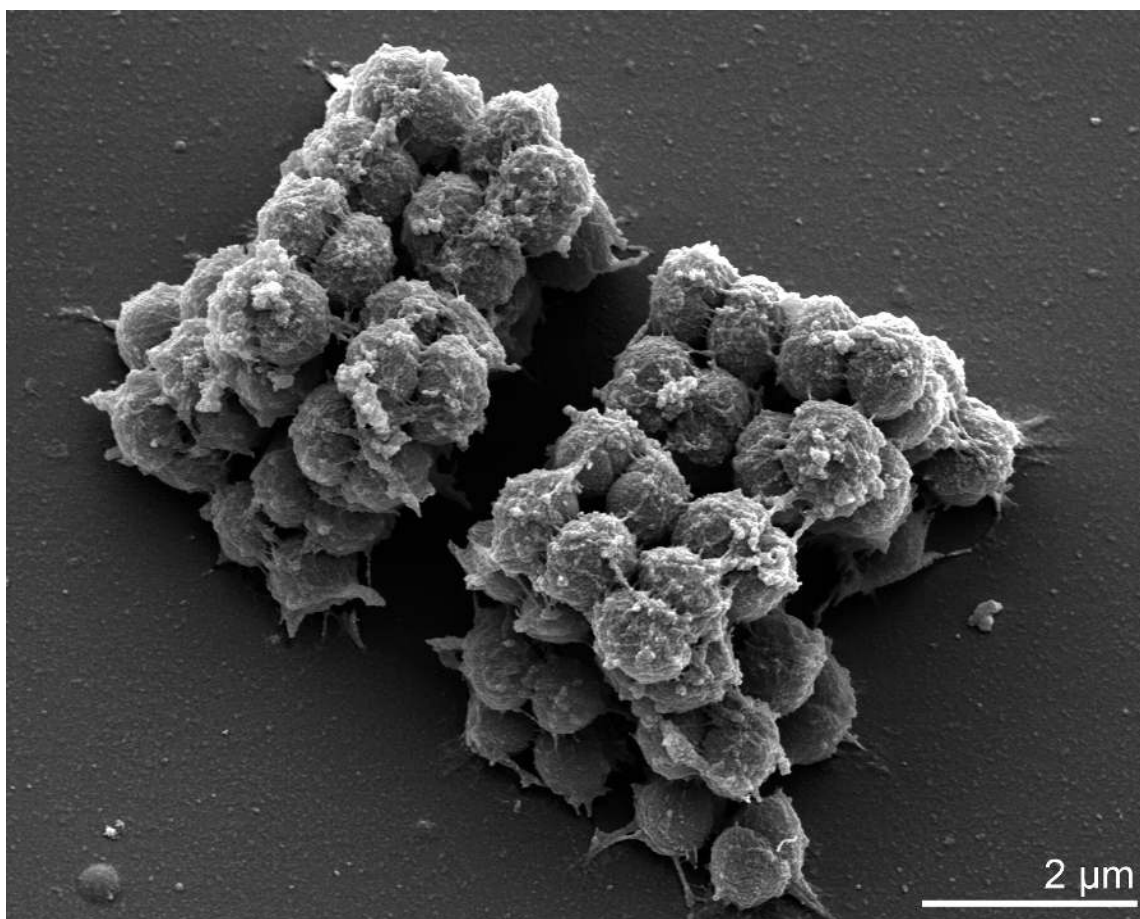


Figure 2. Scanning electron micrograph of *K. sedentarius* strain 541^T (Manfred Rohde, Helmholtz Centre for Infection Biology, Braunschweig)

Table 1. Classification and general features of *K. sedentarius* strain 541^T based on MIGS recommendations [16]

MIGS ID	Property	Term	Evidence code
		Domain <i>Bacteria</i>	
		Phylum <i>Actinobacteria</i>	
		Class <i>Actinobacteria</i>	TAS [17]
	Current classification	Order <i>Actinomycetales</i>	TAS [18]
		Suborder <i>Micrococcineae</i>	TAS [17]
		Family <i>Dermacoccaceae</i>	TAS [19]
		Genus <i>Kytococcus</i>	TAS [1]
		Species <i>Kytococcus sedentarius</i>	TAS [1]
		Type strain 541	
	Gram stain	positive	TAS [1]
	Cell shape	spherical, predominantly in tetrads	TAS [1]
	Motility	nonmotile	TAS [1]
	Sporulation	non-sporulating	TAS [1]
	Temperature range	mesophilic	TAS [1]
	Optimum temperature	28-36°C	TAS [1]
	Salinity	nonhalophilic, but growth in media up to 10% (w/v) NaCl	TAS [1]

Table 1. Classification and general features of *K. sedentarius* strain 541^T based on MIGS recommendations [16] (cont.)

MIGS ID	Property	Term	Evidence code
MIGS-22	Oxygen requirement	mandatory aerobe	TAS [1]
	Carbon source	not reported	
	Energy source	unknown, not starch	NAS
MIGS-6	Habitat	marine	TAS [2]
MIGS-15	Biotic relationship	free-living	NAS
MIGS-14	Pathogenicity	in rare cases	TAS [10,11]
	Biosafety level	1	TAS [20]
	Isolation	slide submerged in sea water	TAS [2]
MIGS-4	Geographic location	probably San Diego	TAS [2]
MIGS-5	Sample collection time	about or before 1944	TAS [2]
MIGS-4.1	Latitude – Longitude	not reported	
MIGS-4.2			
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 [21]. If the evidence code is IDA, then the property was directly observed, for a live isolate by one of the authors, or another 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 Genome OnLine Database [15] and 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.

Table 2. Genome sequencing project information

MIGS ID	Property	Term
MIGS-31	Finishing quality	Finished
MIGS-28	Libraries used	Two genomic Sanger libraries: 8kb pMCL200 and fosmid pcc1Fos libraries.
MIGS-29	Sequencing platforms	ABI3730
MIGS-31.2	Sequencing coverage	17.3x Sanger
MIGS-30	Assemblers	phrap
MIGS-32	Gene calling method	Genemark 4.6b, tRNAScan-SE-1.23, infernal 0.81
	Genbank ID	ABUD00000000
	Genbank Date of Release	N/A
	NCBI project ID	21067
	GOLD ID	Gc01042
	Database: IMG-GEBA	2500901761
MIGS-13	Source material identifier	DSM 20547
	Project relevance	Tree of Life, GEBA

Growth conditions and DNA isolation

K. sedentarius strain 541^T, DSM20547, was grown in [DSMZ medium 92](#) (3% trypticase soy broth, 0.3% yeast extract) at 30°C. DNA was isolated

from 1-1.5 g of cell paste using Qiagen Genomic 500 DNA Kit (Qiagen, Hilden, Germany) with a modified protocol for cell lysis in first freezing for 20 min. (-70°C), then heating 5 min. (98°C), and

cooling 15 min to 37°C; adding 1.5 ml lysozyme (standard: 0.3 ml, only), 1.0 ml achromopeptidase, 0.12 ml lysostaphine, 0.12 ml mutanolysine, 1.5 ml proteinase K (standard: 0.5 ml, only), followed by overnight incubation at 35°C.

Genome sequencing and assembly

The genome was sequenced using a combination of 8 kb and fosmid DNA libraries. All general aspects of library construction and sequencing performed at the [JGI website](#). Draft assemblies were based on 60,742 total reads. The [Phred/Phrap/Consed](#) software package was used for sequence assembly and quality assessment [22-24]. After the shotgun stage, reads were assembled with parallel phrap (High Performance Software, LLC). Possible mis-assemblies were corrected with Dupfinisher [25] or transposon bombing of bridging clones (Epicentre Biotechnologies, Madison, WI). Gaps between contigs were closed by editing in Consed, custom priming, or PCR amplification (Roche Applied Science, Indianapolis, IN). A total of 1,255 additional reactions were necessary to close gaps and to raise the quality of the finished sequence. The completed genome sequence of *K. sedentarius* 541^T contains 61,582 reads. The error rate of the completed genome sequence is less than 1 in 100,000. Together all libraries provided > 17x coverage of the genome.

Genome annotation

Genes were identified using GeneMark [26] as part of the genome annotation pipeline in the [Integrated Microbial Genomes Expert Review](#) (IMG-ER) system [27], followed by a round of manual cura-

tion using JGI's [GenePRIMP](#) pipeline [28]. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) non-redundant database, UniProt, TIGR-Fam, Pfam, PRIAM, KEGG, COG, and InterPro databases. The tRNAscanSE tool [29] was used to find tRNA genes, whereas ribosomal RNAs were found by using the tool RNAmmer [30]. Other non-coding RNAs were identified by searching the genome for the Rfam profiles using INFERNAL (v0.81) [31]. Additional gene prediction analysis and manual functional annotation was performed within the [Integrated Microbial Genomes](#) (IMG) platform [32].

Metabolic network analysis

The metabolic Pathway/Genome Database (PGDB) was computationally generated using Pathway Tools software version 12.5 [33] and MetaCyc version 12.5 [34], 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 [35].

Genome properties

The genome is 2,785,024 bp long and comprises one main circular chromosome with a 71.6% GC content (Table 3 and Figure 3). Of the 2,703 genes predicted, 2,639 were protein-coding genes, 64 encoded RNAs. Eighty-four pseudogenes were also identified. In addition, 72.1% of the genes were assigned with a putative function while the remaining ones were annotated as hypothetical proteins.

Table 3. Genome Statistics

Attribute	Value	% of Total
Genome size (bp)	2,785,024	
DNA Coding region (bp)	2,558,989	91.88%
DNA G+C content (bp)	1,994,844	71.63%
Number of replicons	1	
Extrachromosomal elements	0	
Total genes	2703	100.00%
RNA genes	64	2.37%
rRNA operons	2	
Protein-coding genes	2639	97.63%
Pseudo genes	84	3.11%
Genes with function prediction	1948	72.07%
Genes in paralog clusters	288	10.65%
Genes assigned to COGs	1851	68.48%
Genes assigned Pfam domains	1908	70.59%
Genes with signal peptides	539	19.94%
Genes with transmembrane helices	595	22.01%
CRISPR repeats	0	0

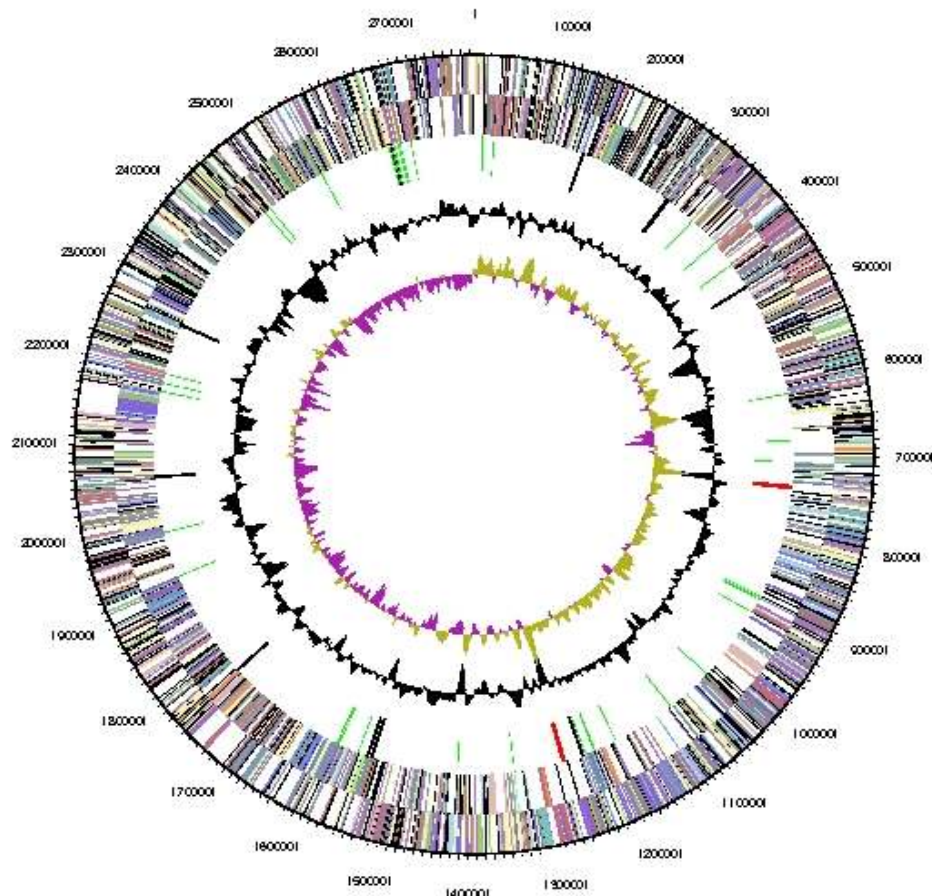


Figure 3. Graphical circular map of the genome. From outside to the center: Genes on forward strand (color by COG categories), Genes on reverse strand (color by COG categories), RNA genes (tRNAs green, rRNAs red, other RNAs black), GC content, GC skew.

The distribution of genes into COGs functional categories is presented in Table 4, and a cellular overview diagram is presented in Figure 4, fol-

lowed by a summary of metabolic network statistics shown in Table 5.

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

Code	Value	%	Description
J	151	5.7	Translation
A	1	0.0	RNA processing and modification
K	143	5.4	Transcription
L	160	6.1	Replication, recombination and repair
B	2	0.1	Chromatin structure and dynamics
D	22	0.8	Cell cycle control, mitosis and meiosis
Y	0	0.0	Nuclear structure
V	56	2.1	Defense mechanisms
T	73	2.8	Signal transduction mechanisms
M	111	4.2	Cell wall/membrane biogenesis

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

Code	Value	%	Description
N	2	0.1	Cell motility
Z	1	0.0	Cytoskeleton
W	0	0.0	Extracellular structures
U	27	1.0	Intracellular trafficking and secretion
O	64	2.4	Posttranslational modification, protein turnover, chaperones
C	99	3.8	Energy production and conversion
G	116	4.4	Carbohydrate transport and metabolism
E	185	7.0	Amino acid transport and metabolism
F	75	2.8	Nucleotide transport and metabolism
H	101	3.8	Coenzyme transport and metabolism
I	86	3.3	Lipid transport and metabolism
P	117	4.4	Inorganic ion transport and metabolism
Q	46	1.7	Secondary metabolites biosynthesis, transport and catabolism
R	229	8.7	General function prediction only
S	160	6.1	Function unknown
-	788	29.9	Not in COGs

Table 5. Metabolic Network Statistics

Attribute	Value
Total genes	2703
Enzymes	531
Enzymatic reactions	922
Metabolic pathways	185
Metabolites	662

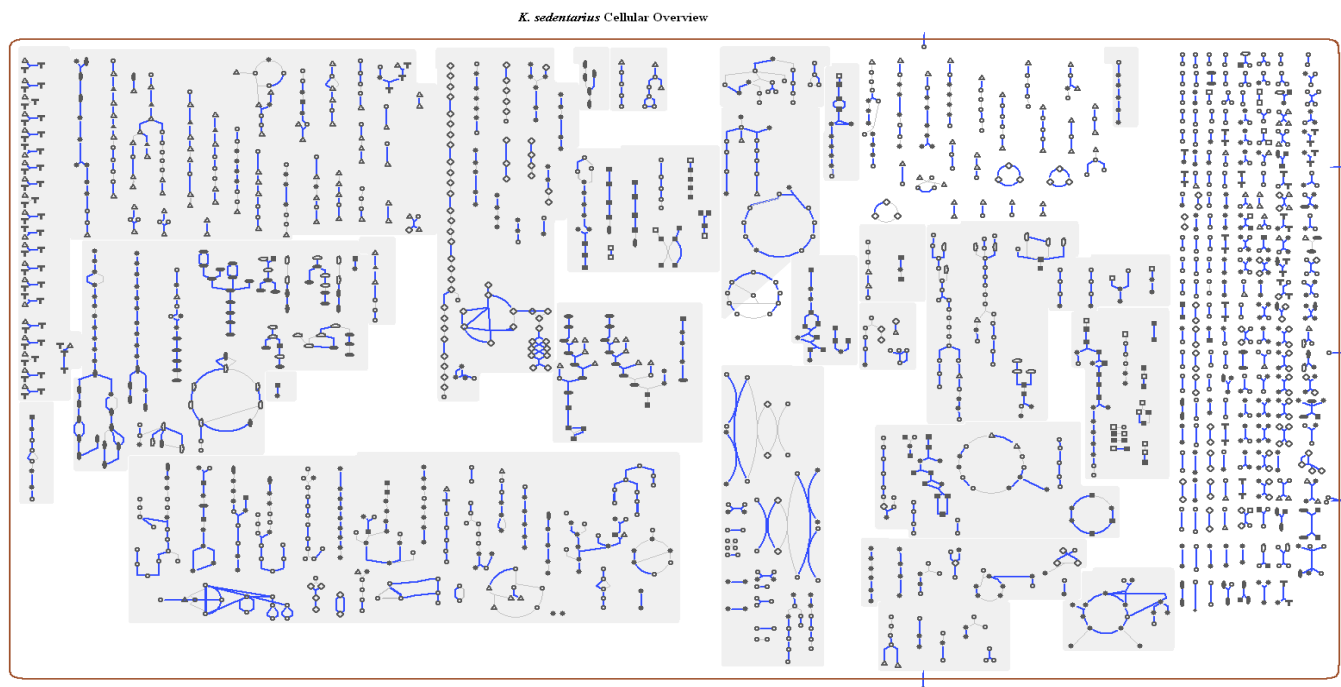


Figure 4. Schematic cellular overview of all pathways of the *K. sedentarius* strain 541^T metabolism. Nodes represent metabolites, with shape indicating class of metabolite. Lines represent reactions.

Acknowledgements

We would like to gratefully acknowledge the help of Katja Steenblock (DSMZ) for growing *K. sedentarius* 541^T 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, as well as German Research Foundation (DFG) INST 599/1-1.

References

- Stackebrandt E, Koch C, Gvozdiak O, Schumann P. Taxonomic dissection of the genus *Micrococcus*: *Kocuria* gen. nov., *Nesterenkonia* gen. nov., *Kytococcus* gen. nov., *Dermaococcus* gen. nov., and *Micrococcus* Cohn 1872 gen. emend. *Int J Syst Bacteriol* 1995; **45**:682-692 [P-MID:7547287](#) [doi:10.1099/00207713-45-4-682](#)
- ZoBell CE, Upham HC. A list of marine bacteria including descriptions of sixty new species. *Bull Scripps Inst Oceanogr Calif* 1944; **5**:239-292
- Kloss WE, Tornabebe TG, Schleifer KH. Isolation and characterization of *Micrococci* from human skin, including two species: *Micrococcus lylae* and *Micrococcus kristinae* *Int J Syst Bacteriol* 1974; **24**:79-101 [doi:10.1099/00207713-24-1-79](#)
- Tiago I, Chung AP, Verissimo A. Bacterial diversity in a nonsaline alkaline environment: heterotrophic aerobic populations. *Appl Environ Microbiol* 2004; **70**:7378-7387 [P-MID:15574939](#) [doi:10.1128/AEM.70.12.7378-7387.2004](#)
- Bhattarai HD, Lee YK, Cho KH, Lee HK, Shin HE. The study of antagonistic interactions among pelagic bacteria: a promising way to coin environmental friendly antifouling compounds. *Hydrobiologica* 2006; **568**:417-423 [doi:10.1007/s10750-006-0220-2](#)
- Gontang EA, Fenical W, Jensen PR. Phylogenetic diversity of gram-positive bacteria cultured from marine sediments. *Appl Environ Microbiol* 2007; **73**:3272-3279 [P-MID:17400789](#) [doi:10.1128/AEM.02811-06](#)
- Osman S, La Duc MT, Dekas A, Newcombe D, Venkateswaran K. Microbial burden and diversity of commercial airline cabin air during short and long durations of travel. *ISME J* 2008; **2**:482-497 [PMID:18256704](#) [doi:10.1038/ismej.2008.11](#) [P-MID:18256704](#) [doi:10.1038/ismej.2008.11](#)
- Nakamiya K, Nakayama T, Ito H, Edmonds JS, Shibata Y, Morita M. Degradation of arylarsenic compounds by microorganisms. *FEMS Microbiol Lett* 2007; **274**:184-191 [P-MID:17697081](#) [doi:10.1111/j.1574-6968.2007.00835.x](#)
- Pospisil S, Benada O, Kofronova O, Petricek M, Janda L, Havlicek V. *Kytococcus sedentarius* (formerly *Micrococcus sedentarius*) and *Dermaococcus nishinomiyaensis* (formerly *Micrococcus nishinomiyaensis*) produce monensins, typical *Streptomyces cinnamonensis* metabolites. *Can J Microbiol* 1998; **44**:1007-1014 [P-MID:9933918](#) [doi:10.1139/cjm-44-10-1007](#)
- Longshaw CM, Wright JD, Farrell AM, Holland KT. *Kytococcus sedentarius*, the organism associated with pitted keratolysis, produces two keratin-degrading enzymes. *J Appl Microbiol* 2002; **93**:810-817 [P-MID:12392527](#) [doi:10.1046/j.1365-2672.2002.01742.x](#)
- Levenga H, Donnelly P, Blijlevens N, Verweij P, Shirango H, de Pauw B. Fatal hemorrhagic pneumonia caused by infection due to *Kytococcus sedentarius*--a pathogen or passenger? *Ann Hematol* 2004; **83**:447-451 [P-MID:14689234](#) [doi:10.1007/s00277-003-0831-x](#)
- Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. *Bioinformatics* 2002; **18**:452-464 [P-MID:11934745](#) [doi:10.1093/bioinformatics/18.3.452](#)
- Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 1981; **17**:368-376 [P-MID:7288891](#) [doi:10.1007/BF01734359](#)
- Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol* 2008; **57**:758-761 [P-MID:18853362](#) [doi:10.1080/10635150802429642](#)
- Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic

- onomic projects and their associated metadata. *Nucleic Acids Res* 2008; **36**:D475-479 [P-MID:17981842](#) [doi:10.1093/nar/gkm884](#)
16. 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 [P-MID:18464787](#) [doi:10.1038/nbt1360](#)
 17. Stackebrandt E, Rainey F, Ward-Rainey N. Proposal for a new hierarchic classification system, *Actinobacteria* classis nov. *Int J Syst Bacteriol* 1997; **47**:479-491 [doi:10.1099/00207713-47-2-479](#)
 18. Buchanan RE. Studies in the Nomenclature and Classification of the *Bacteria*: VIII. The Subgroups and Genera of the *Actinomycetales*. *J Bacteriol* 1918; **3**:403-406 [PMID:16558803](#)
 19. 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 *Micrococccineae*. *Int J Syst Evol Microbiol* 2000; **50**:1279-1285 [P-MID:10843073](#)
 20. Anonymous. Biological Agents: Technical rules for biological agents. <www.baua.de>.
 21. 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 [P-MID:10802651](#) [doi:10.1038/75556](#)
 22. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 1998; **8**:186-192 [P-MID:9521922](#) [doi:10.1101/gr.8.3.186](#)
 23. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 1998; **8**:175-185 [PMID:9521921](#) [doi:10.1101/gr.8.3.175](#)
 24. Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. *Genome Res* 1998; **8**:195-202 [PMID:9521923](#) [doi:10.1101/gr.8.3.195](#)
 25. Han CS, Chain P. Finishing repeat regions automatically with Dupfinisher CSREA Press. In: Arabnia AR, Valafar H, editors. Proceedings of the 2006 international conference on bioinformatics & computational biology; 2006 June 26-29. CSREA Press. p 141-146.
 26. 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**:2607-2618 [P-MID:11410670](#) [doi:10.1093/nar/29.12.2607](#)
 27. Markowitz V, Mavromatis K, Ivanova N, Chen I-M, Chu K, Palaniappan K, Szeto E, Anderson I, Lykidis A, Kyrpides N. Expert Review of Functional Annotations for Microbial Genomes. (Submitted 2009).
 28. Pati A. GenePRIMP: A Gene Prediction Improvement Pipeline for microbial genomes (In preparation 2009).
 29. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997; **25**:955-962 [P-MID:9023104](#) [doi:10.1093/nar/25.5.955](#)
 30. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 2007; **35**:3100-3108 [P-MID:17452365](#) [doi:10.1093/nar/gkm160](#)
 31. 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-D124 [P-MID:15608160](#) [doi:10.1093/nar/gki081](#)
 32. Markowitz VM, Szeto E, Palaniappan K, Grechkin Y, Chu K, Chen IM, Dubchak I, Anderson I, Lykidis A, Mavromatis K, *et al.* The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res* 2008; **36**:D528-D533 [P-MID:17933782](#) [doi:10.1093/nar/gkm846](#)
 33. Karp P, Paley S, Romero P. The pathway tools software. *Bioinformatics* 2002; **18** Suppl1:S225-232 [PMID:12169551](#)
 34. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, *et al.* The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* 2008; **36**:D623-631 [PMID:17965431](#) [doi:10.1093/nar/gkm900](#)
 35. 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**:D475-479 [P-MID:17981842](#) [doi:10.1093/nar/gkm884](#)

33:6083-6089 [P-](#)
[MID:16246909](#) [doi:10.1093/nar/gki892](#)