Genome sequence of the soil bacterium *Saccharomonospora azurea* type strain (NA-128^T)

Hans-Peter Klenk[1*], Brittany Held[2], Susan Lucas[3], Alla Lapidus[3], Alex Copeland[3], Nancy Hammon[3], Sam Pitluck[3], Lynne A. Goodwin[2,3], Cliff Han[2,3], Roxanne Tapia[2,3], Evelyne-Marie Brambilla[1], Gabriele Pötter[1], Miriam Land[3,4], Natalia Ivanova[3], Manfred Rohde[5], Markus Göker[1], John C. Detter[2,3], Nikos C. Kyrpides[3], and Tanja Woyke[3]

[1] Leibniz Institute DSMZ – German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany
[2] Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico, USA
[3] DOE Joint Genome Institute, Walnut Creek, California, USA
[4] Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[5] HZI – Helmholtz Centre for Infection Research, Braunschweig, Germany

***Corresponding author**: Hans-Peter Klenk

Keywords: aerobic, chemoheterotrophic, Gram-positive, vegetative and aerial mycelia, spore-forming, non-motile, soil bacterium, *Pseudonocardiaceae*, CSP 2010

Saccharomonospora azurea Runmao *et al*. 1987 is a member of the genus *Saccharomonospora*, which is in the family *Pseudonocardiaceae* and thus far poorly characterized genomically. Members of the genus *Saccharomonospora* are of interest because they originate from diverse habitats, such as leaf litter, manure, compost, the surface of peat, and moist and over-heated grain, and may play a role in the primary degradation of plant material by attacking hemicellulose. Next to *S. viridis*, *S. azurea* is only the second member in the genus *Saccharomonospora* for which a completely sequenced type strain genome will be published. Here we describe the features of this organism, together with the complete genome sequence with project status 'Improved high quality draft', and the annotation. The 4,763,832 bp long chromosome with its 4,472 protein-coding and 58 RNA genes was sequenced as part of the DOE funded Community Sequencing Program (CSP) 2010 at the Joint Genome Institute (JGI).

Introduction

Strain NA-128^T (= DSM 44631 = ATCC 43670 = NBRC 14651) is the type strain of the species *Saccharomonospora azurea* [1], one of nine species currently in the genus *Saccharomonospora* [2]. The strain was originally isolated in the course of screening for new antibiotics from a soil sample collected near Guangyun City, Sichuan (China) [1]. The genus name *Saccharomonospora* was derived from the Greek words for *sakchâr*, sugar, *monos*, single or solitary, and *spora*, a seed or spore, meaning the sugar (-containing) single-spored (organism) [3]. The species epithet was derived from the Latin adjective *azurea*, azure, referring to the color of the areal mycelium [1]. Yoon *et al.* [4] showed in 1999 *via* DNA-DNA hybridization that '*S. caesia*' [5] (formerly known as '*Micropolyspora caesia*' [6]), which was not included on the Approved Lists [7], was a synonym of *S. azurea*. *S. azurea* and the other type strains of the genus *Saccharomonospora* were selected for genome sequencing in a DOE Community Sequencing Project (CSP 312) at Joint Genome Institute (JGI), because members of the genus (which originate from diverse habitats, such as leaf litter, manure, compost, surface of peat, moist and over-heated grain) might play a role in the primary degradation of plant material by attacking hemicellulose. This expectation was underpinned by the results of the analysis of the genome of *S. viridis* [8], one of the recently sequenced GEBA genomes [9]. The *S. viridis* genome, the only sequenced genome from the genus *Saccharomonospora* to date, contained an unusually large number (24) of genes for glycosyl hydrolases (GH) belonging to 14 GH fami-

lies, which were identified in the Carbon Active Enzyme Database [10]. Hydrolysis of cellulose and starch was also reported for other members of the genus (that are included in CSP 312), such as *S. marina* [11], *S. halophila* [12], *S. saliphila* [13], *S. paurometabolica* [14], and *S. xinjiangensis* [15]. Here we present a summary classification and a set of features for *S. azurea* AN-128^T, together with the description of the genomic sequencing and annotation.

Classification and features

A representative genomic 16S rRNA sequence of *S. azurea* NA-128^T was compared using NCBI BLAST [16,17] 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 [18] and the relative frequencies of taxa and keywords (reduced to their stem [19]) were determined, weighted by BLAST scores. The most frequently occurring genera were *Saccharomonospora* (47.9%), *Kocuria* (17.7%), *Corynebacterium* (9.4%), *Kibdelosporangium* (6.0%) and *Prauserella* (5.5%) (176 hits in total). Regarding the eight hits to sequences from members of the species, the average identity within HSPs was 99.5%, whereas the average coverage by HSPs was 99.8%. Regarding the 42 hits to sequences from other members of the genus, the average identity within HSPs was 97.0%, whereas the average coverage by HSPs was 98.3%. Among all other species, the one yielding the highest score was *Saccharomonospora xinjiangensis* (AJ306300), which corresponded to an identity of 98.9% and an HSP coverage of 100.1%. (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 FN667533 'stages composting process pilot scale municipal drum compost clone PS3734', which showed an identity of 100.0% and a HSP coverage of 97.9%. The most frequently occurring keywords within the labels of all environmental samples that produced hits were 'feedlot' (7.9%), 'top' (4.1%), 'beef, cattl, coli, escherichia, habitat, marc, neg, pen, primari, secondari, stec, surfac, synecolog' (3.9%), 'feedbunk' (2.3%) and 'compost' (1.7%) (74 hits in total). Environmental samples that yielded hits of a higher score than the highest scoring species were not found.

Figure 1 shows the phylogenetic neighborhood of *S. azurea* in a 16S rRNA based tree. The sequences of the three identical 16S rRNA gene copies in the genome do not differ from the previously published 16S rRNA sequence (Z38017).

Cells of *S. azurea* NA-128^T form an irregularly branched vegetative mycelium of 0.3 to 0.4 μm diameter (Figure 2) [1]. The monopodally branching aerial mycelium has a diameter of 0.3 to 0.6 µm [1]; the mature mycelium and the spores are azure to cyaneus when grown on Oatmeal agar (ISP3) or on Czapek sucrose agar [1]. Smooth, round spores are 0.8 to 1.0 mM long, mostly found on the aerial mycelium, but rarely on the substrate mycelium [1]. No distinct soluble pigment was detectable [1]. The growth range of strain NA-128^T spans from 24-40°C, with an optimum at 28-37°C [1]. Strain NA-128^T grows well in up to 7% NaCl containing medium, but is inhibited at 10% NaCl [1]. Physiological characteristics such as growth substrates, gelatin formation and peptonization of milk are described in detail by Runmao (1987) [1].

Chemotaxonomy

The cell wall of strain AN-128^T contains *meso*-diaminopimelic acid. Galactose and arabinose are present, indicating a type IV cell wall / type A whole cell sugar pattern [1]. The fatty acids spectrum is dominated by almost 80% hexadecanoic acids: *iso*-$C_{16:0}$ (27.0%), $C_{16:1\ cis-9}$ (17.0%), *iso*-$C_{16:0\ 2\text{-}OH}$ (14.0%), $C_{16:0}$ (palmitic acid, 13.0%), *iso*-$C_{16:1\ H}$ (7.0%), *anteiso*-$C_{16:0}$ (1.0%) [42]. There are no data available for polar lipids and quinines of this strain.

Genome sequencing and annotation

Genome project history

This organism was selected for sequencing as part of the DOE Joint Genome Institute Community Sequencing Program (CSP) 2010, CSP 312, "Whole genome type strain sequences of the genus *Saccharomonospora* – a taxonomically troubled genus with bioenergetic potential". The genome project is deposited in the Genomes On Line Database [26] 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.

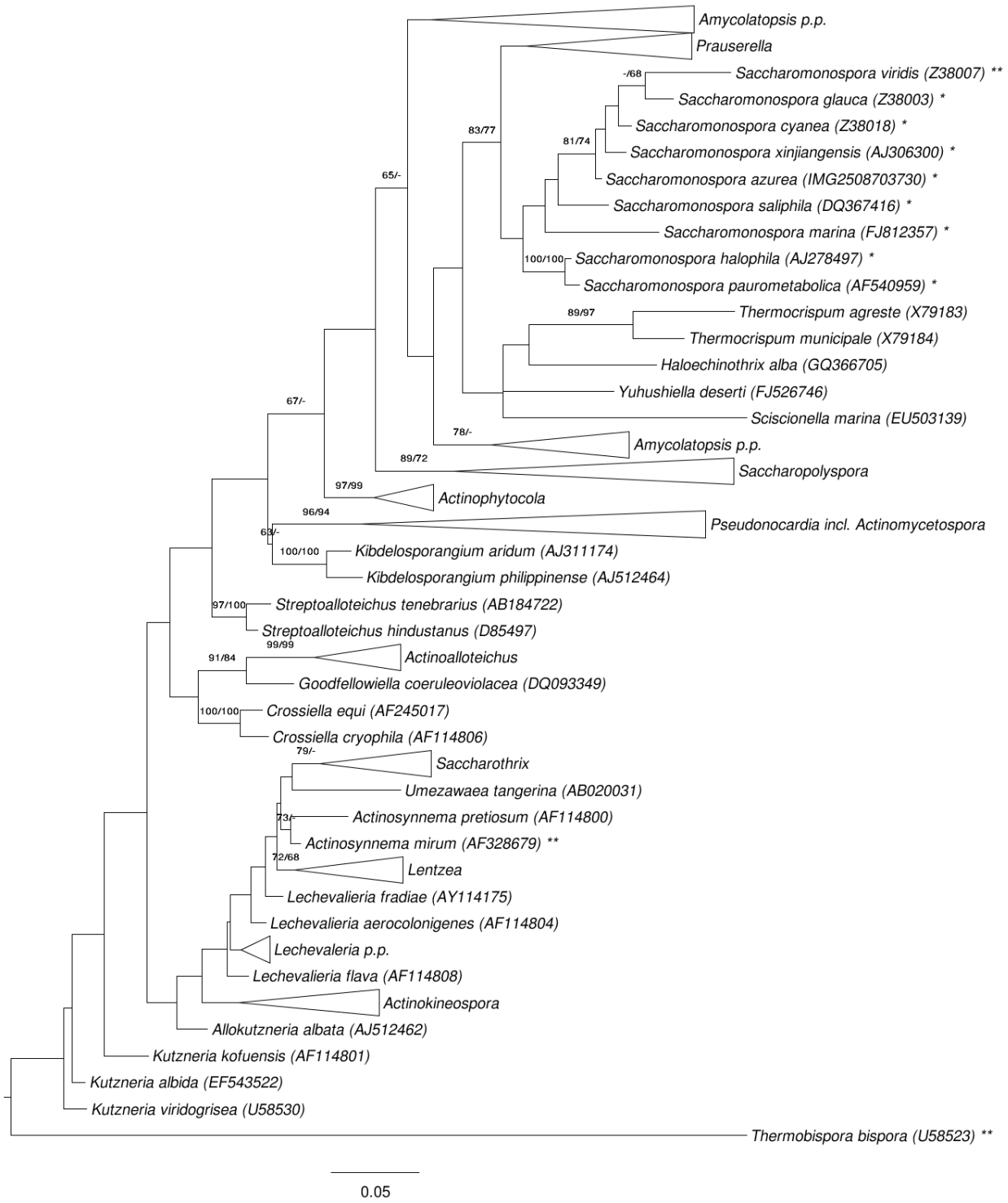

Figure 1. Phylogenetic tree highlighting the position of *S. azurea* relative to the type strains of the other species within the family *Pseudonocardiaceae*. The tree was inferred from 1,386 aligned characters [20,21] of the 16S rRNA gene sequence under the maximum likelihood (ML) criterion [22]. Rooting was done initially using the midpoint method [23] 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 550 ML bootstrap replicates [24] (left) and from 1,000 maximum parsimony bootstrap replicates [25] (right) if larger than 60%. Lineages with type strain genome sequencing projects registered in GOLD [26] are labeled with one asterisk, those also listed as 'Complete and Published' with two asterisks [8,27,28]. *Actinopolyspora iraqiensis* Ruan *et al*. 1994 was ignored in the tree. The species was proposed to be a later heterotypic synonym of *S. halophila* [29], although the name *A. iraqiensis* would have had priority over *S. halophila*. This taxonomic problem will soon be resolved with regard to the genomes of *A. iraqiensis* and *S. halophila*, which were both part of CSP 312.

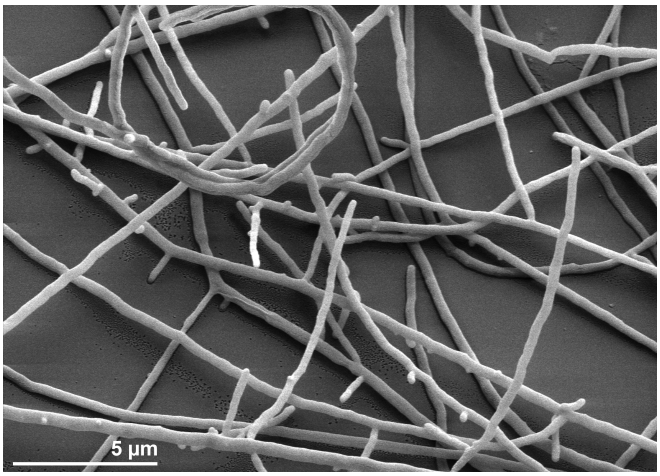

Figure 2. Scanning electron micrograph of *S. azurea* AN-128^T

Table 1. Classification and general features of *S. azurea* AN-128 ^{T} according to the MIGS recommendations [30].

MIGS ID	Property	Term	Evidence code
	Current classification	Domain *Bacteria*	TAS [31]
		Phylum *Actinobacteria*	TAS [32]
		Class *Actinobacteria*	TAS [33]
		Subclass *Actinobacteridae*	TAS [33,34]
		Order *Actinomycetales*	TAS [7,33-35]
		Suborder *Pseudonocardineae*	TAS [33,34,36]
		Family *Pseudonocardiaceae*	TAS [33,34,36-38]
		Genus *Saccharomonospora*	TAS [7,39]
		Species *Saccharomonospora azurea*	TAS [1]
		Type-strain AN-128	TAS [1]
	Gram stain	positive	NAS
	Cell shape	variable	NAS
	Motility	non-motile	NAS
	Sporulation	single spores with smooth surface, mainly on aerial mycelium	TAS [1]
	Temperature range	mesophile, 24–40°C	TAS [1]
	Optimum temperature	28–37°C	TAS [1]
	Salinity	grows in up to 7% NaCl; 10% is inhibitory	TAS [1]
MIGS-22	Oxygen requirement	aerobic	TAS [1]
	Carbon source	Mono, di- and trisaccharides	TAS [1]
	Energy metabolism	chemoheterotrophic	NAS
MIGS-6	Habitat	soil	TAS [1]
MIGS-15	Biotic relationship	free living	NAS
MIGS-14	Pathogenicity	none	NAS
	Biosafety level	1	TAS [40]
MIGS-23.1	Isolation	soil	TAS [1]
MIGS-4	Geographic location	Guangyuan City, Sichuan (China)	TAS [1]
MIGS-5	Sample collection time	1986 or before	NAS
MIGS-4.1	Latitude	32.45	NAS
MIGS-4.2	Longitude	105.84	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 [41]. 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.

Table 2. Genome sequencing project information

MIGS ID	Property	Term
MIGS-31	Finishing quality	Improved high quality draft
MIGS-28	Libraries used	Three genomic libraries: one 454 pyrosequence standard library, one 454 PE library (12 kb insert size), one Illumina library
MIGS-29	Sequencing platforms	Illumina GAii, 454 GS FLX Titanium
MIGS-31.2	Sequencing coverage	1,025.0 × Illumina; 8.6 × pyrosequence
MIGS-30	Assemblers	Newbler version 2.3, Velvet version 1.0.13, phrap version SPS - 4.24
MIGS-32	Gene calling method	Prodigal
	INSDC ID	AGIU00000000, CM001466
	GenBank Date of Release	March 6, 2012
	GOLD ID	Gi07579
	NCBI project ID	62037
	Database: IMG	2508501044
MIGS-13	Source material identifier	DSM 44631
	Project relevance	Bioenergy and phylogenetic diversity

Growth conditions and DNA isolation

Strain NA-128^T, DSM 44631, was grown in DSMZ medium 83 (Czapek Peptone Medium) [43] 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 with the following modifications: extended cell lysis time (60 min.) with additional 30µl Achromopeptidase, Lysostaphin, Mutanolysin; proteinase K was applied in 6-fold the supplier recommended amount for 60 min. at 58°C. The purity, quality and size of the bulk gDNA preparation were assessed by JGI according to DOE-JGI guidelines. DNA is available through the DNA Bank Network [44].

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 [45]. Pyrosequencing reads were assembled using the Newbler assembler (Roche). The initial Newbler assembly consisting of 215 contigs in one scaffold was converted into a phrap [46] assembly by making fake reads from the consensus, to collect the read pairs in the 454 paired end library. Illumina GAii sequencing data (5,162.6 Mb) was assembled with Velvet [47] 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 80.3 Mb 454 draft data and all of the 454 paired end data. Newbler parameters are -consed -a 50 -l 350 -g -m -ml 20. The Phred/Phrap/Consed software package [46] 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 [45], Dupfinisher [48], 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 158 additional reactions were necessary to close gaps and to raise the quality of the finished sequence. Illumina reads were also used to correct potential base errors and increase consensus quality using a software Polisher developed at JGI [49].

The error rate of the completed genome sequence is less than 1 in 100,000. Together, the combination of the Illumina and 454 sequencing platforms provided 1,033.6 × coverage of the genome. The final assembly contained 345,324 pyrosequence and 64,928,268 Illumina reads.

Genome annotation

Genes were identified using Prodigal [50] as part of the Oak Ridge National Laboratory genome annotation pipeline, followed by a round of manual curation using the JGI GenePRIMP pipeline [51]. 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 [52].

Genome properties

The genome consists of a 4,763,852 bp long chromosome with a 70.1% G+C content (Table 3 and Figure 3). Of the 4,530 genes predicted, 4,472 were protein-coding genes, and 58 RNAs; 96 pseudogenes were also identified. The majority of the protein-coding genes (73.8%) 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)	4,763,852	100.00%
DNA coding region (bp)	4,287,642	90.00%
DNA G+C content (bp)	3,331,901	70.08%
Number of replicons	1	
Extrachromosomal elements	0	
Total genes	4,530	100.00%
RNA genes	58	1.28%
rRNA operons	3	
tRNA genes	47	1.04%
Protein-coding genes	4,472	98.72%
Pseudo genes	96	2.12%
Genes with function prediction (proteins)	3,342	73.77%
Genes in paralog clusters	2,354	51.96%
Genes assigned to COGs	3,312	73.11%
Genes assigned Pfam domains	3,450	76.16%
Genes with signal peptides	1,332	29.40%
Genes with transmembrane helices	1,070	23.62%
CRISPR repeats	0	

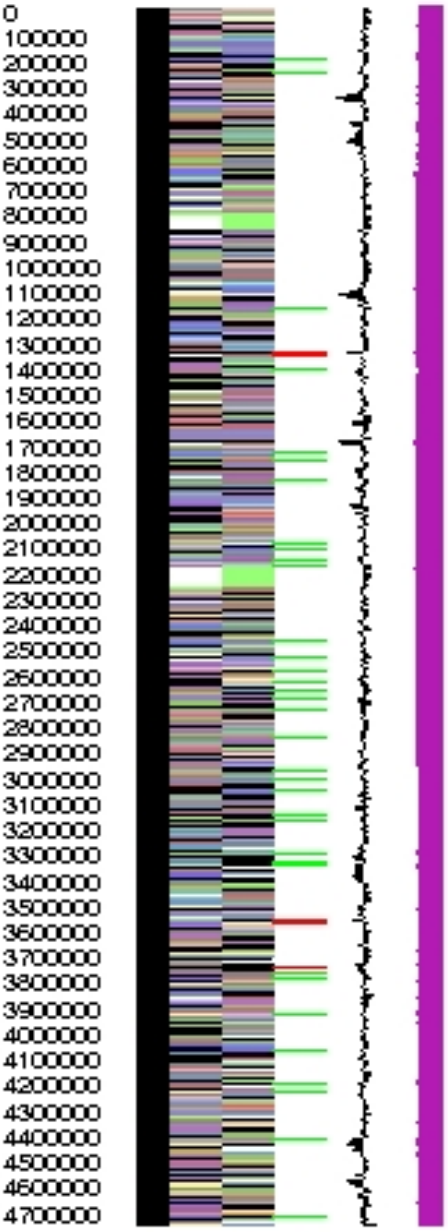

Figure 3. Graphical map of the chromosome. From left to the right: Genes on forward strand (color by COG categories), Genes on reverse strand (color by COG categories), RNA genes (tRNAs green, rRNAs red, other RNAs black), GC content, GC skew.

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

Code	value	%age	Description
J	171	4.6	Translation, ribosomal structure and biogenesis
A	1	0.0	RNA processing and modification
K	394	10.6	Transcription
L	175	4.7	Replication, recombination and repair
B	2	0.1	Chromatin structure and dynamics
D	35	0.9	Cell cycle control, cell division, chromosome partitioning
Y	0	0.0	Nuclear structure
V	58	1.6	Defense mechanisms
T	190	5.1	Signal transduction mechanisms
M	156	4.2	Cell wall/membrane biogenesis
N	6	0.2	Cell motility
Z	0	0.0	Cytoskeleton
W	0	0.0	Extracellular structures
U	36	1.0	Intracellular trafficking and secretion, and vesicular transport
O	134	3.6	Posttranslational modification, protein turnover, chaperones
C	245	6.6	Energy production and conversion
G	259	7.0	Carbohydrate transport and metabolism
E	313	8.4	Amino acid transport and metabolism
F	91	2.4	Nucleotide transport and metabolism
H	194	5.2	Coenzyme transport and metabolism
I	179	4.8	Lipid transport and metabolism
P	176	4.7	Inorganic ion transport and metabolism
Q	152	4.1	Secondary metabolites biosynthesis, transport and catabolism
R	478	12.8	General function prediction only
S	282	7.6	Function unknown
-	1,218	26.9	Not in COGs

Acknowledgements

The work conducted by the US Department of Energy Joint Genome Institute was supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

1. Runmao H. *Saccharomonospora azurea* sp. nov., a new species from soil. *Int J Syst Bacteriol* 1987; **37**:60-61. http://dx.doi.org/10.1099/00207713-37-1-60
2. Garrity G. NamesforLife. BrowserTool takes expertise out of the database and puts it right in the browser. *Microbiol Today* 2010; **37**:9.
3. Euzéby JP. List of Bacterial Names with Standing in Nomenclature: a folder available on the internet. *Int J Syst Bacteriol* 1997; **47**:590. PubMed http://dx.doi.org/10.1099/00207713-47-2-590
4. Yoon JH, Kim SB, Lee ST, Park YH. DNA-DNA relatedness between *Saccharomonospora* species: '*Saccharomonospora caesia*' as a synonym of *Saccharomonospora azurea*. *Int J Syst Bacteriol* 1999; **49**:671-673. http://dx.doi.org/10.1099/00207713-49-2-671
5. Kurup VP. Taxonomic study of some members of *Micropolyspora* and *Saccharomonospora*. *Microbiologica* 1981; **4**:249-259.
6. Kalakoutskii LV. A new species of the genus *Micropolyspora – Micropolyspora caesia* n. sp. [English translation of Microbiologiya]. *Microbiology* 1964; **33**:765-768.
7. Skerman VBD, McGowan V, Sneath PHA. Approved Lists of Bacterial Names. *Int J Syst Bacteriol* 1980; **30**:225-420. http://dx.doi.org/10.1099/00207713-30-1-225
8. Pati A, Sikorski J, Nolan M, Lapidus A, Copeland A, Glavina Del Rio T, Lucas S, Chen F, Tice H, Pitluck S, *et al*. Complete genome sequence of *Saccharomonospora viridis* type strain (P101 ^{T}). *Stand Genomic Sci* 2009; **1**:141-149. PubMed http://dx.doi.org/10.4056/sigs.20263
9. 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

10. Carbon Active Enzyme Database. http://www.cazy.org
11. Liu Z, Li Y, Zheng L, Huang YJ, Li WJ. *Saccharomonospora marina* sp. nov., isolated from an ocean sediment of the East China Sea. *Int J Syst Evol Microbiol* 2010; **60**:1854-1857. PubMed http://dx.doi.org/10.1099/ijs.0.017038-0
12. Al-Zarban SS, Al-Musaallam AA, Abbas I, Stackebrandt E, Kroppenstedt RM. *Saccharomonospora halophila* sp. nov., a novel halophilic actinomycete isolated from marsh soil in Kuwait. *Int J Syst Evol Microbiol* 2002; **52**:555-558. PubMed
13. Syed DG, Tang SK, Cai M, Zhi XY, Agasar D, Lee JC, Kim CJ, Jiang CL, Xu CL, Li WJ. *Saccharomonospora saliphila* sp. nov., a halophilic actinomycete from an Indian soil. *Int J Syst Evol Microbiol* 2008; **58**:570-573. PubMed http://dx.doi.org/10.1099/ijs.0.65449-0
14. Li WJ, Tang SK, Stackebrandt E, Kroppenstedt RM, Schumann P, Xu LH, Jiang CL. *Saccharomonospora paurometabolica* sp. nov., a moderately halophilic actinomycete isolated from soil in China. *Int J Syst Evol Microbiol* 2003; **53**:1591-1594. PubMed http://dx.doi.org/10.1099/ijs.0.02633-0
15. Jin X, Xu LH, Mao PH, Hseu TH, Jiang CL. Description of *Saccharomonospora xinjiangensis* sp. nov. based on chemical and molecular classification. *Int J Syst Bacteriol* 1998; **48**:1095-1099. PubMed http://dx.doi.org/10.1099/00207713-48-4-1095
16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **215**:403-410. PubMed
17. Korf I, Yandell M, Bedell J. BLAST, O'Reilly, Sebastopol, 2003.
18. 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://dx.doi.org/10.1128/AEM.03006-05
19. Porter MF. An algorithm for suffix stripping. Program: *electronic library and information systems* 1980; **14**:130-137.
20. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. *Bioinformatics* 2002; **18**:452-464. PubMed http://dx.doi.org/10.1093/bioinformatics/18.3.452
21. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000; **17**:540-552. PubMed http://dx.doi.org/10.1093/oxfordjournals.molbev.a026334
22. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol* 2008; **57**:758-771. PubMed http://dx.doi.org/10.1080/10635150802429642
23. 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
24. 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
25. Swofford DL. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.0 b10. Sinauer Associates, Sunderland, 2002.
26. Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Kyrpides NC. The genomes on line database (GOLD) in 2009: Status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 2010; **38**:D346-D354. PubMed http://dx.doi.org/10.1093/nar/gkp848
27. Land M, Lapidus A, Mayilraj S, Chen R, Copeland A, Glavina Del Rio T, Nolan M, Lucas S, Tice H, Cheng JF, *et al*. Complete genome sequence of *Actinosynnema mirum* type strain (101 ^{T}). *Stand Genomic Sci* 2009; **1**:46-53. PubMed http://dx.doi.org/10.4056/sigs.21137
28. Liolios K, Sikorski J, Jando M, Lapidus A, Copeland A, Glavina Del Rio T, Nolan M, Lucas S, Tice H, Cheng JF, *et al*. Complete genome sequence of *Thermobispora bispora* type strain (R51 ^{T}). *Stand Genomic Sci* 2010; **2**:318-326. PubMed http://dx.doi.org/10.4056/sigs.962171
29. Tang SK, Wang Y, Klenk HP, Shi R, Lou K, Zhang YJ, Chen C, Ruan JS, Li WJ. *Actinopolyspora alba* sp. nov. and Actinopolyspora erythraea sp. nov., isolated from a salt field, and reclassification of *Actinopolyspora iraqiensis* Ruan *et al*. 1994 as a heterotypic synonym of *Saccharomonospora halophila*. *Int J Syst Evol Microbiol* 2011; **61**:1693-1698. PubMed http://dx.doi.org/10.1099/ijs.0.022319-0
30. 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
31. 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

32. Garrity GM, Holt JG. The Road Map to the Manual. *In:* Garrity GM, Boone DR, Castenholz RW (*eds*), Bergey's Manual of Systematic Bacteriology, Second Edition, Volume 1, Springer, New York, 2001, p. 119-169.

33. Stackebrandt E, Rainey FA, Ward-Rainey NL. Proposal for a new hierarchic classification system, *Actinobacteria* classis nov. *Int J Syst Bacteriol* 1997; **47**:479-491. http://dx.doi.org/10.1099/00207713-47-2-479

34. Zhi XY, Li WJ, Stackebrandt E. An update of the structure and 16S rRNA gene sequence-based definition of higher ranks of the class *Actinobacteria*, with the proposal of two new suborders and four new families and emended descriptions of the existing higher taxa. *Int J Syst Evol Microbiol* 2009; **59**:589-608. PubMed http://dx.doi.org/10.1099/ijs.0.65780-0

35. Buchanan RE. Studies in the nomenclature and classification of bacteria. II. The primary subdivisions of the *Schizomycetes*. *J Bacteriol* 1917; **2**:155-164. PubMed

36. Labeda DP, Goodfellow M, Chun J, Zhi X-Y, Li W-J. Reassessment of the systematics of the suborder *Pseudonocardineae*: transfer of the genera within the family *Actinosynnemataceae* Labeda and Kroppenstedt 2000 emend. Zhi et al. 2009 into an emended family *Pseudonocardiaceae* Embley *et al*. 1989 emend. Zhi *et al*. 2009. *Int J Syst Evol Microbiol* 2011; **61**:1259-1264. PubMed http://dx.doi.org/10.1099/ijs.0.024984-0

37. List Editor. Validation List no. 29. Validation of the publication of new names and new combinations previously effectively published outside the IJSB. *Int J Syst Bacteriol* 1989; **39**:205-206. http://dx.doi.org/10.1099/00207713-39-2-205

38. Embley MT, Smida J, Stackebrandt E. The phylogeny of mycolate-less wall chemotype IV Actinomycetes and description of *Pseudonocardiaceae* fam. nov. *Syst Appl Microbiol* 1988; **11**:44-52. http://dx.doi.org/10.1016/S0723-2020(88)80047-X

39. Nonomura H, Ohara Y. Distribution of actinomycetes in soil. X. New genus and species of monosporic actinomycetes in soil. *J Ferment Technol* 1971; **49**:895-903.

40. BAuA. 2010, Classification of *Bacteria* and *Archaea* in risk groups. http://www.baua.de TRBA 466, p. 194.

41. 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

42. Wink JM. Compendium of Actinobacteria. http://www.dsmz.de/microorganisms/wink_pdf/DSM44631.pdf.

43. List of growth media used at DSMZ: http://www.dsmz.de/catalogues/catalogue-microorganisms/culture-technology/list-of-media-for-microorganisms.html.

44. 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. http://dx.doi.org/10.1089/bio.2010.0029

45. The DOE Joint Genome Institute. http://www.jgi.doe.gov

46. Phrap and Phred for Windows. MacOS, Linux, and Unix. http://www.phrap.com

47. 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

48. 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.

49. 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.

50. 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

51. Pati A, Ivanova N, Mikhailova N, Ovchinikova 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

52. 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