Genome sequence of the moderately thermophilic halophile *Flexistipes sinusarabici* strain (MAS10^T)

Alla Lapidus[1], Olga Chertkov[1,2], Matt Nolan[1], Susan Lucas[1], Nancy Hammon[1], Shweta Deshpande[1], Jan-Fang Cheng[1], Roxanne Tapia[1,2], Cliff Han[1,2], Lynne Goodwin[1,2], Sam Pitluck[1], Konstantinos Liolios[1], Ioanna Pagani[1], Natalia Ivanova[1], Marcel Huntemann[1], Konstantinos Mavromatis[1], Natalia Mikhailova[1], Amrita Pati[1], Amy Chen[3], Krishna Palaniappan[3], Miriam Land[1,4], Loren Hauser[1,4], Evelyne-Marie Brambilla[5], Manfred Rohde[6], Birte Abt[5], Stefan Spring[5], Markus Göker[5], James Bristow[1], Jonathan A. Eisen[1,7], Victor Markowitz[4], Philip Hugenholtz[1,8], Nikos C. Kyrpides[1], Hans-Peter Klenk[5]*, and Tanja Woyke[1]

[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, California, USA
[4] Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[5] DSMZ - German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany
[6] HZI – Helmholtz Centre for Infection Research, Braunschweig, Germany
[7] University of California Davis Genome Center, Davis, California, USA
[8] Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Australia

*Corresponding author: Hans-Peter Klenk

Flexistipes sinusarabici Fiala *et al.* 2000 is the type species of the genus *Flexistipes* in the family *Deferribacteraceae*. The species is of interest because of its isolated phylogenetic location in a genomically under-characterized region of the tree of life, and because of its origin from a multiply extreme environment; the Atlantis Deep brines of the Red Sea, where it had to struggle with high temperatures, high salinity, and a high concentrations of heavy metals. This is the fourth completed genome sequence to be published of a type strain of the family *Deferribacteraceae*. The 2,526,590 bp long genome with its 2,346 protein-coding and 53 RNA genes is a part of the *G*enomic *E*ncyclopedia of *B*acteria and *A*rchaea project.

Introduction

Strain MAS10^T (= DSM 4947 = ATCC 49648) is the type strain of *Flexistipes sinusarabici* [1,2] which is the type and only species of the genus *Flexistipes* [1,2]. The strain was first isolated from the Atlantis II Deep brines of the Red Sea [1], together with four related isolates. The generic name derives from the Latin words *flexus*, a bending, turning, winding, and *stipes,* a branch of tree, stick [1]. The species epithet is derived from the Latin words *sinus,* a curve or fold in land, a gulf, and *arabicus*, Arabic, describing the place of isolation [1]. Since

the time of its isolation in the late 1980s until now no closely related bacterium (16S rRNA identity >90%) was described. The resistance of the strain to moderate heat, high salt concentrations, and heavy metals [1] should make it an interesting target for extremophile biotechnology. Here we present a summary classification and a set of features for *F. sinusarabici* MAS10^T, together with the description of the complete genomic sequencing and annotation.
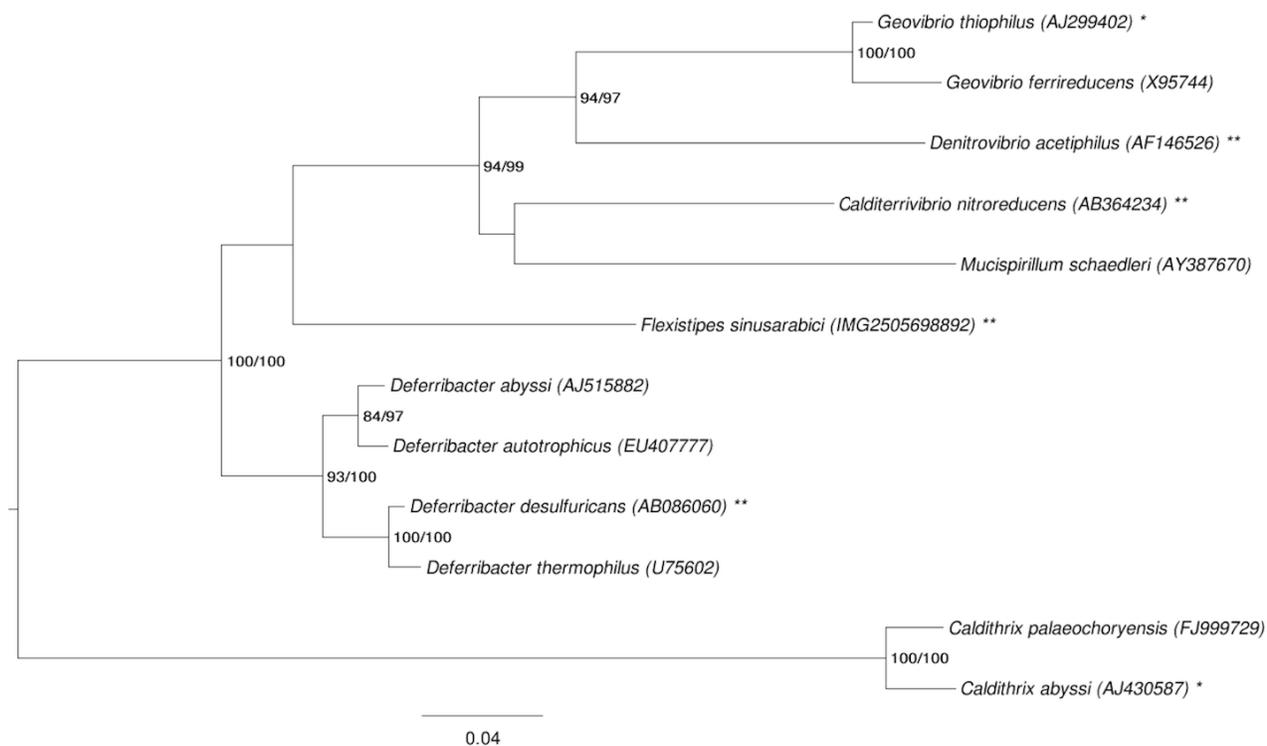
Figure 1. Phylogenetic tree highlighting the position of *F. sinusarabici* relative to the type strains of the other species within the phylum "*Deferribacteres*". The tree was inferred from 1,459 aligned characters [7,8] of the 16S rRNA gene sequence under the maximum likelihood (ML) criterion [9]. Rooting was done initially using the midpoint method [10] 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 250 ML bootstrap replicates [11] (left) and from 1,000 maximum parsimony bootstrap replicates [12] (right) if larger than 60%. Lineages with type strain genome sequencing projects registered in GOLD [13] are labeled with one asterisk, those also listed as 'Complete and Published' with two asterisks [14-16].
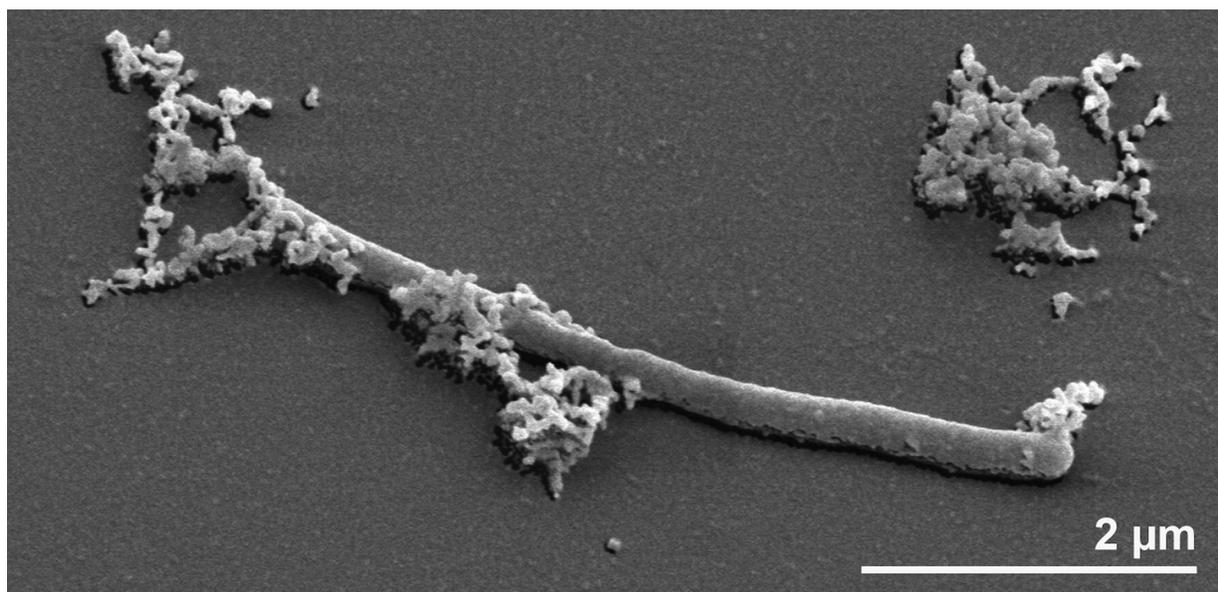

Figure 2. Scanning electron micrograph of *F. sinusarabici* MAS10^T

Classification and features

A representative genomic 16S rRNA sequence of strain MAS10^T was compared using NCBI BLAST [3,4] 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 [5] and the relative frequencies of taxa and keywords (reduced to their stem [6]) were determined, weighted by BLAST scores. The most frequently occurring genera were *Acidithiobacillus* (60.0%), *Deferribacter* (26.8%), *Flexistipes* (8.2%), *Desulfuromonas* (2.2%) and *Calditerrivibrio* (1.8%) (80 hits in total). Regarding the single hit to sequences from members of the species, the average identity within HSPs was 98.0%, whereas the average coverage by HSPs was 96.9%. Among all other species, the one yielding the highest score was *Deferribacter abyssi* (AJ515881), which corresponded to an identity of 89.7% and an HSP coverage of 86.4%. (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 FR744611 ('succession potential reducers nitrate-treated facility determined temperature and nitrate availability production water Halfdan oil field clone PWB039'), which showed an identity of 96.7% and an HSP coverage of 93.1%. The most frequently occurring keywords within the labels of all environmental samples which yielded hits were 'microbi' (3.9%), 'acid' (3.4%), 'sediment' (3.3%), 'water' (3.0%) and 'oil' (2.4%) (170 hits in total). The most frequently occurring keyword within the labels of those environmental samples which yielded hits of a higher score than the highest scoring species was 'avail, determin, facil, field, halfdan, nitrat, nitrate-tr, oil, potenti, product, reduc, success, temperatur, water' (7.1%) (1 hit in total). While these keywords fit to the marine environment from which strain MAS10^T originated, they also point to sediments and oil fields which were so far not considered as habitats for *F. sinusarabici.*

Figure 1 shows the phylogenetic neighborhood of *F. sinusarabici* MAS10^T in a 16S rRNA based tree. The sequences of the two identical 16S rRNA gene copies in the genome differ by two nucleotides from the previously published 16S rRNA sequence M59231, which contains 25 ambiguous base calls.

Cells of strain MAS10^T are straight to bent rods, about 0.3 μm wide and 4–50 μm long (Figure 2) [1]. *F. sinusarabici* was described as non-motile [1].

Spore-formation was not observed [1]. MAS10^T cells stain Gram-negative, and growth is strictly anaerobic, with the best growth occurring within a temperature range of 45–50°C and a minimum doubling time of 8 ½ hours [1]. Optimal pH range for the strain is pH 6-8 [1]. Strain MAS10^T requires at least 3% NaCl for growth, but also grows at salt concentrations as high as 10% [1]. The organism prefers complex growth substrates such as yeast extract, meat extract, peptone and tryptone, while formate, lactate, citrate, malate, carbohydrate, amino acids and alcohols do not support cell growth [1]. Strain MAS10^T shows an unusual resistance against the transcription inhibitor rifampicin [1], which is however also commonly found among the spirochetes.

Chemotaxonomy

The chemotaxonomic data for MAS10^T is relatively sparse: No information on cell wall structure, quinones or polar lipids is available. The fatty acid composition is dominated by saturated unbranched acids: C_{18} (23.3%), C_{16} (15.1%), C_{17} (12.6%), with some branched acids *iso*-C_{14} (10.2%), *anteiso*-C_{15} (10.2%), *iso*-C_{16} (4.1%), *iso*-C_{15} (3.6%), and few unsaturated acids $C_{18:1}$ (9.9%), $C_{16:1}$ (2.8%), $C_{17:1}$ (2.5%) [1].

Genome sequencing and annotation
Genome project history

This organism was selected for sequencing on the basis of its phylogenetic position [28], and is part of the *Genomic Encyclopedia of Bacteria and Archaea* project [29]. The genome project is deposited in the Genome On Line Database [13] 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.

Growth conditions and DNA isolation

F. sinusarabici MAS10^T, DSM 4947, was grown anaerobically in DSMZ medium 524 (*Flexistipes* Medium) [30] at 47°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, but adding 10μl proteinase K for one hour extended lysis at 58°C. DNA is available through the DNA Bank Network [31].

Table 1. Classification and general features of *F. sinusarabici* MAS10[T] according to the MIGS recommendations [17] and the NamesforLife database [18].

MIGS ID	Property	Term	Evidence code
		Domain *Bacteria*	TAS [19]
		Phylum "*Deferribacteres*"	TAS [20,21]
		Class "*Deferribacteres*"	TAS [22,23]
	Current classification	Order *Deferribacterales*	TAS [22,24]
		Family *Deferribacteraceae*	TAS [22,25]
		Genus *Flexistipes*	TAS [1,2]
		Species *Flexistipes sinusarabici*	TAS [1,2]
		Type strain MAS10	TAS [1]
	Gram stain	negative	TAS [1]
	Cell shape	straight to acutely bent rods	TAS [1]
	Motility	non-motile	TAS [1]
	Sporulation	none	TAS [1]
	Temperature range	30–53°C, moderately thermophilic	TAS [1]
	Optimum temperature	45–50°C	TAS [1]
	Salinity	at least 3% NaCl, growths with up to 18% NaCl	TAS [1]
MIGS-22	Oxygen requirement	strictly anaerobic	TAS [1]
	Carbon source	complex organic components like yeast extract, meat extract, peptone, tryptone	TAS [1]
	Energy metabolism	heterotrophic	TAS [1]
MIGS-6	Habitat	marine, deep brine water	TAS [1]
MIGS-15	Biotic relationship	free-living	TAS [1]
MIGS-14	Pathogenicity	none	TAS [1]
	Biosafety level	1	TAS [26]
	Isolation	interface between upper brine layer and deep sea water	TAS [1]
MIGS-4	Geographic location	Atlantis II Deep brines, Red Sea	TAS [1]
MIGS-5	Sample collection time	1987 or before	NAS
MIGS-4.1	Latitude	21.37	TAS [1]
MIGS-4.2	Longitude	38.07	TAS [1]
MIGS-4.3	Depth	2,000 – 2,200 m	TAS [1]
MIGS-4.4	Altitude	-2,200 – 2,200 m	TAS [1]

Evidence codes - 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 [27].

Table 2. Genome sequencing project information

MIGS ID	Property	Term
MIGS-31	Finishing quality	Finished
MIGS-28	Libraries used	Four genomic libraries: one 454 pyrosequence standard library, two 454 PE libraries (3 kb, 15.5 kb insert size), one Illumina library
MIGS-29	Sequencing platforms	Illumina GAii, 454 GS FLX Titanium
MIGS-31.2	Sequencing coverage	162.0 × Illumina; 37.9 × pyrosequence
MIGS-30	Assemblers	Newbler version 2.3, Velvet version 0.7.63, phrap SPS-4.24
MIGS-32	Gene calling method	Prodigal 1.4, GenePRIMP
	INSDC ID	CP002858
	Genbank Date of Release	June 17, 2011
	GOLD ID	Gc01819
	NCBI project ID	45817
	Database: IMG-GEBA	2505679008
MIGS-13	Source material identifier	DSM 4947
	Project relevance	Tree of Life, GEBA

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 [32]. Pyrosequencing reads were assembled using the Newbler assembler (Roche). The initial Newbler assembly consisting of 175 contigs in two scaffolds was converted into a phrap [33] assembly by making fake reads from the consensus, to collect the read pairs in the 454 paired end library. Illumina GAii sequencing data (489.7 Mb) was assembled with Velvet [34] and the consensus sequences were shredded into 2.0 kb overlapped fake reads and assembled together with the 454 data. The 454 draft assembly was based on 170.4 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 [33] 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 [32], Dupfinisher [35], 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 605 additional reactions and 15 shatter libraries 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 [36]. 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 199.9 × coverage of the genome. The final assembly contained 248,918 pyrosequence and 395,536,860 Illumina reads.

Genome annotation

Genes were identified using Prodigal [37] as part of the Oak Ridge National Laboratory genome annotation pipeline, followed by a round of manual curation using the JGI GenePRIMP pipeline [38]. 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. Additional gene prediction analysis and functional annotation was performed within the Integrated Microbial Genomes - Expert Review (IMG-ER) platform [39].

Genome properties

The genome consists of a 2,526,590 bp long circular chromosome with a G+C content of 38.3% (Table 3 and Figure 3). Of the 2,399 genes predicted, 2,346 were protein-coding genes, and 53 RNAs; 85 pseudogenes were also identified. The majority of the protein-coding genes (75.2%) 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)	2,526,590	100.00%
DNA coding region (bp)	2,179,830	86.28%
DNA G+C content (bp)	967,539	38.29%
Number of replicons	1	
Extrachromosomal Elements	0	
Total genes	2,399	100.00%
RNA genes	53	2.21%
rRNA operons	2	
Protein-coding genes	2,346	97.79%
Pseudo genes	85	3.54%
Genes with function prediction	1,803	75.16%
Genes in paralog clusters	242	10.09%
Genes assigned to COGs	1,924	80.20%
Genes assigned Pfam domains	1,978	82.45%
Genes with signal peptides	366	15.26%
Genes with transmembrane helices	579	24.14%
CRISPR repeats	0	

Insight into the genome sequence
Comparative genomics

Lacking an available genome sequence of *Deferribacter abyssi*, the species yielding the highest score, the following comparative analyses were done with *D. desulfuricans* [14] (GenBank AP011529, AP011530) and *Calditerrivibrio nitroreducens* (GenBank CP002347, CP002348) [16], the phylogenetically closest organisms for which a genome sequence was available. The genomes of *F. sinusarabici*, *D. desulfuricans* and *C. nitroreducens* are similar in sizes (2.5 Mb, 2.5 Mb and 2.2 Mb, respectively) and have a similar, quite low G+C content (38%, 30% and 36%, respectively). Whereas *F. sinusarabici* has no plasmid, *D. desulfuricans* harbors a 5.9 kb plasmid; *C. nitroreducens* contains a 30.8 kb megaplasmid.

An estimate of the overall similarity between the three genomes was generated with the GGDC-Genome-to-Genome Distance Calculator [40,41]. This system calculates the distances by comparing the genomes to obtain HSPs (high-scoring segment pairs) and inferring distances from a set of formulas (1, HSP length / total length; 2, identities / HSP length; 3, identities / total length). Table 5 shows the results of the pairwise comparison between the three genomes.

The comparison of the *F. sinusarabici* and *D. desulfuricans* genomes revealed that 5.9% of the average of both genome lengths are covered with HSPs. The identity within these HSPs was 83.2%, whereas the identity over the whole genome was only 4.9%. Similar results were inferred for *F. sinusarabici* and *C. nitroreducens* (Table 5). The genomes of *D. desulfuricans* and *C. nitroreducens* show a significantly higher degree of similarity with 9.9% of the average of both genomes are covered with HSPs of 83.3% identity. The identity over the whole length of the genomes was 8.3%. These values corroborate the relationship between the three organisms as shown in the 16S rRNA-based phylogenetic tree in Figure 1, as there is no bootstrap support that *F. sinusarabici* is closer related to either *C. nitroreducens* or *D. desulfuricans*.
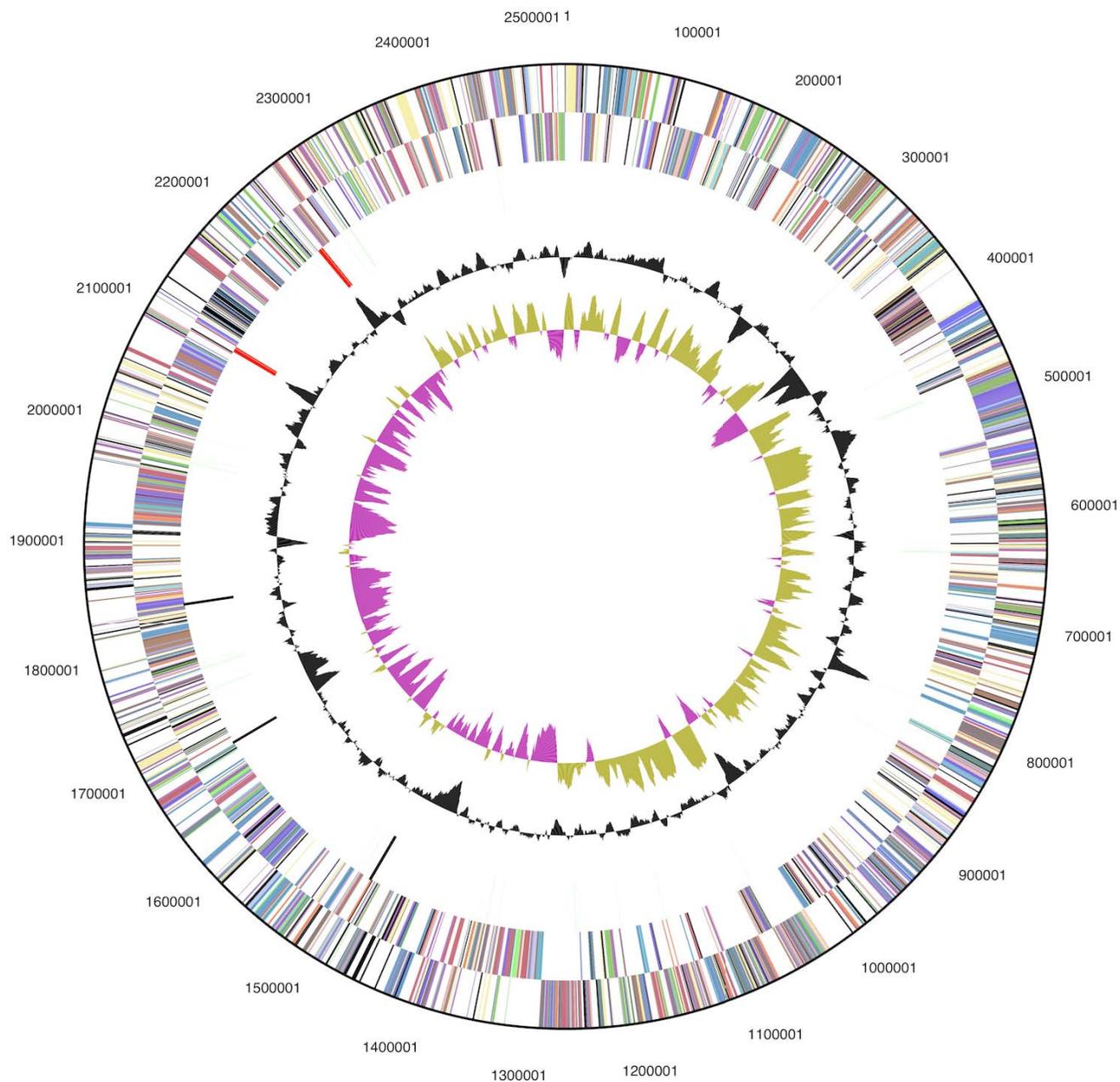
Figure 3. Graphical circular map of the chromosome. From outside to the center: Genes on forward strand (color by COG categories), Genes on reverse strand (color by COG categories), RNA genes (tRNAs green, rRNAs red, other RNAs black), GC content, GC skew.

The fraction of shared genes in the three genomes is shown in a Venn diagram (Figure 4). The numbers of pairwise shared genes were calculated with the phylogenetic profiler function of the IMG/ER platform [33]. The homologous genes within the genomes were detected with a maximum E-value of 10^{-5} and a minimum identity of 30%. Roughly 61% of all genes in the genomes (1,400 genes) are shared by all three genomes, with about equal numbers of genes (224 and 246) shared on a pairwise basis by *F. sinusarabici* and *D. desulfuricans* or by *D. desulfuricans* and *C. nitroreducens*, respectively, and to the exclusion of the third genome. Within the 567 unique genes of *F. sinusarabici* that have no detectable homologs in the genomes of *D. desulfuricans* and *C. nitroreducens* (under the sequence similarity thresholds used for the comparison) the 86 genes (3.7% based on the whole gene number) encoding transposases appear to be noteworthy.

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

Code	value	%age	Description
J	145	7.0	Translation, ribosomal structure and biogenesis
A	1	0.1	RNA processing and modification
K	84	4.0	Transcription
L	205	9.8	Replication, recombination and repair
B	2	0.1	Chromatin structure and dynamics
D	21	1.0	Cell cycle control, cell division, chromosome partitioning
Y	0	0.0	Nuclear structure
V	26	1.3	Defense mechanisms
T	115	5.5	Signal transduction mechanisms
M	135	6.5	Cell wall/membrane/envelope biogenesis
N	36	1.7	Cell motility
Z	0	0.0	Cytoskeleton
W	0	0.0	Extracellular structures
U	62	3.0	Intracellular trafficking, secretion, and vesicular transport
O	81	3.9	Posttranslational modification, protein turnover, chaperones
C	174	8.3	Energy production and conversion
G	66	3.2	Carbohydrate transport and metabolism
E	198	9.5	Amino acid transport and metabolism
F	52	2.5	Nucleotide transport and metabolism
H	115	5.5	Coenzyme transport and metabolism
I	60	2.9	Lipid transport and metabolism
P	86	4.1	Inorganic ion transport and metabolism
Q	32	1.5	Secondary metabolites biosynthesis, transport and catabolism
R	246	11.8	General function prediction only
S	145	7.0	Function unknown
-	475	19.8	Not in COGs

Table 5. Pairwise comparison of *F. sinusarabici*, *D. desulfuricans* and *C. nitroreducens* using the GGDC-Calculator.

		1, HSP length /total length [%]	2, identities /HSP length [%]	3, identities /total length [%]
F. sinusarabici	*D. desulfuricans*	5.9	83.2	4.9
F. sinusarabici	*C. nitroreducens*	5.1	83.3	4.3
D. desulfuricans	*C. nitroreducens*	9.9	83.3	8.3

A remarkable difference between the compared organisms is their motility. Whereas *F. sinusarabici* is described to be non-motile, *D. desulfuricans* is motile by twitching [14] and *C. nitroreducens* is also described to be motile [16]. The mechanism of twitching motility is still unknown but it is thought that moving across surfaces is caused by extension and retraction of type IV pili. A set of genes that is responsible for twitching motility was identified in several organisms; in *Pseudomonas aeruginosa* a gene cluster involved in pilus biosynthesis and twitching motility was characterized, the gene products of this gene cluster show a high degree of sequence similarity to the chemotaxis (che) proteins of enterics and the gliding bacterium *Myxococcus xanthus* [42]. A closer look into the genome sequences of *F. sinusarabici, D. desulfuricans* and *C. nitroreducens* revealed the presence of different gene sets coding for chemotaxis proteins. In contrast to *D. desulfuricans* and *C. nitroreducens*, *F. sinusarabici* lacks four *che* genes (*che*B, *che*R, *che*V, *che*W). In *P. aeruginosa* a mutation in the *pil*I gene, a homolog to *che*W, lead to a blocking of pilus production [42]. It can be assumed that the missing *che*W gene in *F. sinusarabici* might be responsible for the non-motility of the cells, despite the rather large number of 36 genes annotated in the cell motility category of table 4.
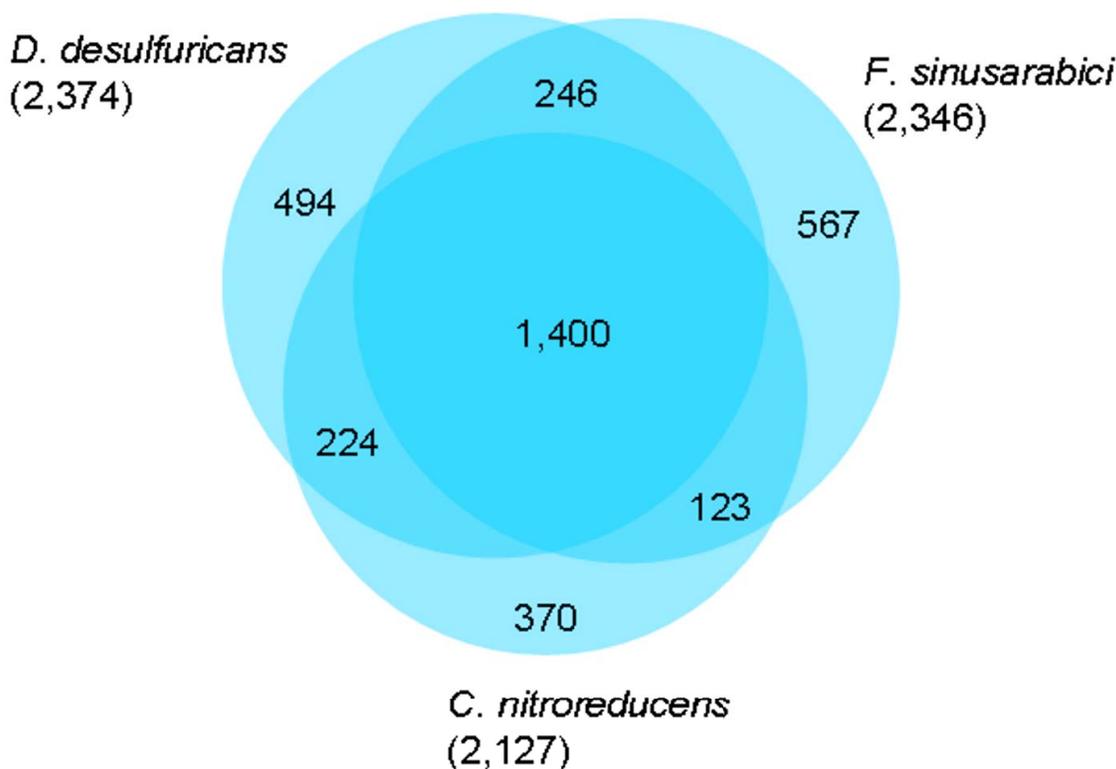
Figure 4. Venn diagram depicting the intersections of protein sets (total number of derived protein sequences in parentheses) of *F. sinusarabici, D. desulfuricans* and *C. nitroreducens*.

References

1. Fiala G, Woese CR, Langworthy TA, Stetter KO. *Flexistipes sinusarabici* a novel genus and species of eubacteria occurring in the Atlantis II Deep brines of the Red Sea. *Arch Microbiol* 1990; **154**:120-126. doi:10.1007/BF00423320

2. Validation List No. 75. *Int J Syst Evol Microbiol* 2000; **50**:1415-1417. PubMed doi:10.1099/00207713-50-4-1415

3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Bascic local alignment search tool. *J Mol Biol* 1990; **215**:403-410. PubMed

4. Korf I, Yandell M, Bedell J. BLAST, O'Reilly, Sebastopol, 2003.

5. 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 doi:10.1128/AEM.03006-05

6. Porter MF. An algorithm for suffix stripping. Program: electronic library and information systems 1980; **14**:130-137.

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

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

9. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol* 2008; **57**:758-771. PubMed doi:10.1080/10635150802429642

10. Hess PN, De Moraes Russo CA. An empirical test of the midpoint rooting method. *Biol J Linn Soc Lond* 2007; **92**:669-674. doi:10.1111/j.1095-8312.2007.00864.x

11. 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. doi:10.1007/978-3-642-02008-7_13

12. Swofford DL. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.0 b10. Sinauer Associates, Sunderland, 2002.

13. Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, 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 doi:10.1093/nar/gkp848

14. Takaki Y, Shimamura S, Nakagawa S, Fukuhara Y, Horikawa H, Ankai A, Harada T, Hosoyama A, Oguchi A, Fukui S, *et al.* Bacterial lifestyle in a deep-sea hydrothermal vent chimney revealed by the genome sequence of the thermophilic bacterium *Deferribacter desulfuricans* SSM1. *DNA Res* 2010; **17**:123-137. PubMed doi:10.1093/dnares/dsq005

15. Kiss H, Lang E, Lapidus A, Copeland A, Nolan M, Glavina Del Rio T, Chen F, Lucas S, Tice H, Cheng JF, *et al.* Complete genome sequence of *Denitrovibrio acetiphilus* tye strain (N2460^T). *Stand Genomic Sci* 2010; **2**:270-279. PubMed doi:10.4056/sigs.892105

16. Pitluck S, Sikorski J, Zeytun A, Lapidus A, Nolan M, Lucas S, Hammon N, Deshpande S, Cheng JF, Tapia R, *et al.* Complete genome sequence of *Calditerrivibrio nitroreducens* type strain (Yu37-1^T). *Stand Genomic Sci* 2011; **4**:54-62. PubMed doi:10.4056/sigs.1523807

17. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, *et al.* The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 2008; **26**:541-547. PubMed doi:10.1038/nbt1360

18. Garrity G. NamesforLife. BrowserTool takes expertise out of the database and puts it right in the browser. *Microbiol Today* 2010; **37**:9.

19. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains *Archaea, Bacteria*, and *Eucarya*. *Proc Natl Acad Sci USA* 1990; **87**:4576-4579. PubMed doi:10.1073/pnas.87.12.4576

20. Garrity GM, Holt JG. Phylum BIX. *Deferribacteres* phy. nov. In: Garrity GM, Boone DR, Castenholz RW (eds), *Bergey's Manual of Systematic Bacteriology*, Second Edition, Volume 1, Springer, New York, 2001, p. 465.

21. Jumas-Bilak E, Roudière L, Marchandin H. Description of 'Synergistetes' phyl. nov. and emended description of the phylum 'Deferribacteres' and of the family *Syntrophomonadaceae*, phylum 'Firmicutes'. *Int J Syst Evol Microbiol* 2009; **59**:1028-1035. PubMed doi:10.1099/ijs.0.006718-0

22. List Editor. Validation List no. 85. Validation of publication of new names and new combinations previously effectively published outside the IJSEM. *Int J Syst Evol Microbiol* 2002; **52**:685-690. PubMed doi:10.1099/ijs.0.02358-0

23. Huber H, Stetter KO. Class I. *Deferribacteres* class. nov. In: Garrity GM, Boone DR, Castenholz RW (eds), *Bergey's Manual of Systematic Bacteriology*, Second Edition, Volume 1, Springer, New York, 2001, p. 465.

24. Huber H, Stetter KO. Order I. *Deferribacterales* ord. nov. *In:* Garrity GM, Boone DR, Castenholz RW (*eds*), *Bergey's Manual of Systematic Bacteriology*, Second Edition, Volume 1, Springer, New York, 2001, p. 465.

25. Huber H, Stetter KO. Family I. *Deferribacteraceae* fam. nov. *In:* Garrity GM, Boone DR, Castenholz RW (*eds*), *Bergey's Manual of Systematic Bacteriology*, Second Edition, Volume 1, Springer, New York, 2001, p. 465-466.

26. BAuA. Classification of bacteria and archaea in risk groups. *TRBA* 2005; **466**:190.

27. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*. Gene Ontology: tool for the unification of biology. *Nat Genet* 2000; **25**:25-29. PubMed doi:10.1038/75556

28. Klenk HP, Göker M. En route to a genome-based classification of *Archaea* and *Bacteria*? *Syst Appl Microbiol* 2010; **33**:175-182. PubMed doi:10.1016/j.syapm.2010.03.003

29. 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 doi:10.1038/nature08656

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

31. 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. *Biopreservation and Biobanking* 2011; **9**:51-55. doi:10.1089/bio.2010.0029

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

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

34. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008; **18**:821-829. PubMed doi:10.1101/gr.074492.107

35. 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.

36. 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.

37. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; **11**:119. PubMed doi:10.1186/1471-2105-11-119

38. Pati A, Ivanova NN, Mikhailova N, Ovchinnikova G, Hooper SD, Lykidis A, Kyrpides NC. GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nat Methods* 2010; **7**:455-457. PubMed doi:10.1038/nmeth.1457

39. 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 doi:10.1093/bioinformatics/btp393

40. Auch AF, von Jan M, Klenk HP, Göker M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand Genomic Sci* 2010; **2**:117-134. PubMed doi:10.4056/sigs.531120

41. Auch AF, Klenk HP, Göker M. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Stand Genomic Sci* 2010; **2**:142-148. PubMed doi:10.4056/sigs.541628

42. Darzins A. Characterization of a *Pseudomonas aeruginosa* gene cluster involved in pilus biosynthesis and twitching motility: sequence similarity to the chemotaxis proteins of enterics and the gliding bacterium *Myxococcus xanthus*. *Mol Microbiol* 1994; **11**:137-153. PubMed doi:10.1111/j.1365-2958.1994.tb00296.x