Complete genome sequence of *Cellulophaga lytica* type strain (LIM-21^T)

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

[1] DOE Joint Genome Institute, Walnut Creek, California, USA
[2] DSMZ - German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany
[3] Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico, USA
[4] Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[5] Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[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

Keywords: aerobic, motile by gliding, Gram-negative, agarolytic, chemoorganotrophic, *Flavobacteriaceae,* GEBA

Cellulophaga lytica (Lewin 1969) Johansen *et al.* 1999 is the type species of the genus *Cellulophaga,* which belongs to the family *Flavobacteriaceae* within the phylum '*Bacteroidetes*' and was isolated from marine beach mud in Limon, Costa Rica. The species is of biotechnological interest because its members produce a wide range of extracellular enzymes capable of degrading proteins and polysaccharides. After the genome sequence of *Cellulophaga algicola* this is the second completed genome sequence of a member of the genus *Cellulophaga*. The 3,765,936 bp long genome with its 3,303 protein-coding and 55 RNA genes consists of one circular chromosome and is a part of the ***G**enomic **E**ncyclopedia of **B**acteria and **A**rchaea* project.

Introduction

Strain LIM-21^T (DSM 7489 = ATCC 23178 = JCM 8516) is the type strain of the species *Cellulophaga lytica*, which is the type species of the genus *Cellulophaga* [1]. The genus currently consists of five more validly named species [2]: *C. algicola* [3], *C. baltica, C. fucicola* [1], *C. pacifica* [4] and *C. tyrosinoxydans* [5]. The species was first described in 1969 by Lewin as '*Cytophaga lytica*' [6], and was subsequently transferred to the novel genus *Cellulophaga* as type strain *C. lytica* [1]. The genus name is derived from the Neo-Latin word 'cellulosum' meaning 'cellulose' and the latinized Greek word 'phagein' meaning 'to eat', yielding the Neo-Latin word 'Cellulophaga' meaning 'eater of cellulose' [2]. The species epithet is derived from the Neo-Latin word 'lytica' (loosening, dissolving) [2]. Here we present a summary classification and a set of features for *C. lytica* strain LIM-21^T, together with the description of the complete genomic sequencing and annotation.

Classification and features

A representative genomic 16S rRNA sequence of strain LIM-21^T was compared using NCBI BLAST 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 [7] and the relative frequencies, weighted by BLAST scores, of taxa and keywords (reduced to their stem [8]) were determined. The five most frequent genera were *Cellulophaga* (37.3%), *Flavobacterium* (8.5%), *Cytophaga* (6.3%), *Aquimarina* (5.8%) and *Arenibacter* (5.7%) (141 hits in total). Regarding the ten hits to sequences from members of the species, the average identity within HSPs was 99.0%, whereas the average coverage by HSPs was 93.3%. Regarding the eleven hits to sequences from other members of the genus, the average identity within HSPs was 94.0%, whereas the average coverage by HSPs was 93.1%. Among all other species, the one yielding the highest score was *Cytophaga lytica* (M62796), which corresponded to an identity of 99.2% and an HSP coverage of 96.9%. (Note that the Greengenes databases uses the INSDC (= EMBL/NCBI/DDBJ) annotation, which is not an authoritative source for nomenclature or classification). The highest-scoring environmental sequence was EU246790 ('Identification microorganism Libya untreated Mediterranean sea water feed reverse osmosis plant isolate RSW1-4RSW1-4 str. RSW1-4'), which showed an identity of 100.0% and an HSP coverage of 96.2%. The five most frequent keywords within the labels of environmental samples which yielded hits were 'sea' (5.6%), 'water' (4.8%), 'marin' (3.6%), 'sediment' (3.1%) and 'surfac' (2.8%) (109 hits in total). The single most frequent keyword within the labels of environmental samples which yielded hits of a higher score than the highest scoring species was 'feed, identif, libya, mediterranean, microorgan, osmosi, plant, revers, sea, untreat, water' (9.1%) (1 hit in total).

Figure 1 shows the phylogenetic neighborhood of *C. lytica* in a 16S rRNA based tree. The sequence of the four 16S rRNA gene copies in the genome differ from each other by up to four nucleotides, and differ by up to 15 nucleotides from the previously published 16S rRNA sequence (D12666), which contains 19 ambiguous base calls.

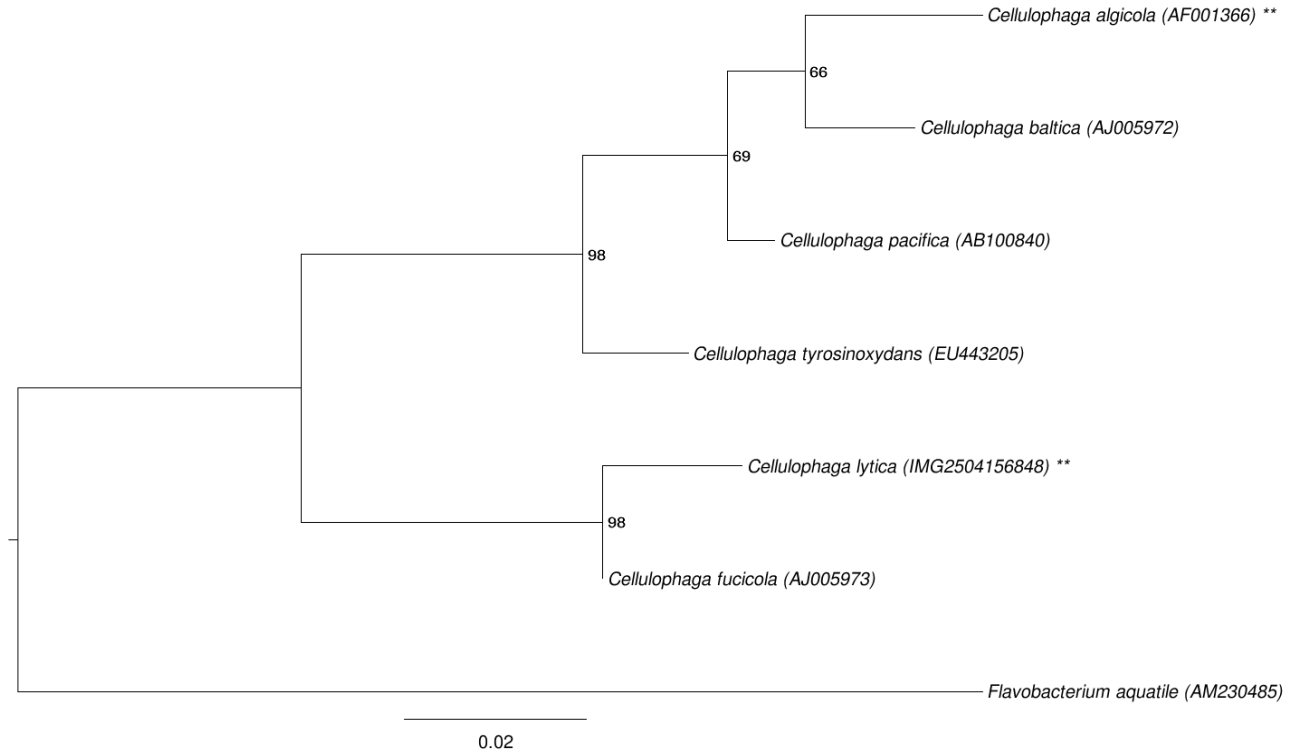

Figure 1. Phylogenetic tree highlighting the position of *C. lytica* relative to the other type strains within the genus. The tree was inferred from 1,458 aligned characters [9,10] of the 16S rRNA gene sequence under the maximum likelihood criterion [11] and rooted with the type strain of the type species of the family. The branches are scaled in terms of the expected number of substitutions per site. Numbers next to bifurcations are support values from 450 bootstrap replicates [12] if larger than 60%. Lineages with type strain genome sequencing projects tha are registered in GOLD [13] but remain unpublished are labeled with one asterisk, published genomes with two asterisks [14].

The cells of *C. lytica* are slender flexible rods, cylindrical with blunt ends. Their lengths and widths range from 1.5-10 and 0.3-0.4 µm, respectively (Figure 2 and Table 1) [25]. *C. lytica* is motile by gliding [25]. Colonies have a bright yellow color caused by zeaxanthin as the main pigment; flexirubin-type pigments are not formed [24,28]. *C. lytica* requires Na^+ and grows at NaCl concentrations up to 8% [3,5], in the presence of 10% NaCl no growth was observed [4]. The temperature range for growth is between 4°C [4] and 40°C [25], with an optimum between 22-30°C [25].

C. lytica is aerobic and chemoorganotrophic [24]. The organism can degrade agar, alginate, gelatin and starch [24,25], but not casein, cellulose (filter paper), chitin, alginic acid, elastin or fibrinogen [1,25]. There are conflicting observations describing the ability of *C. lytica* to degrade carboxymethylcellulose (CMC). Whereas most authors [3,5,24,25] describe the hydrolysis of CMC, Nedashkovskaya *et al*. 2004 [4] did not observe its degradation by *C. lytica.* Nitrate reduction and denitrification are negative [25]. *C. lytica* is catalase [24] and oxidase positive [25]. Acid is formed oxidatively from cellobiose, galactose, glucose, lactose, maltose and xylose [4]. *C. lytica* is sensitive to oleandomycin, lincomycin and shows resistance to benzylpenicillin, carbencillin, gentamicin, kanamycin, neomycin, ampicillin, streptomycin and tetracycline [4].

Chemotaxonomy

The fatty acid profiles of four *C. lytica* strains were analyzed by Bowman in 2000 [3]. The predominant cellular acids of these four analyzed *C. lytica* strains were branched-chain saturated and unsaturated fatty acids and straight-chain saturated and monounsaturated fatty acids, namely i-$C_{15:0}$ (18.9%), i-$C_{15:1}\omega10c$ (10.3%), i-$C_{17:1}\omega7c$ (5.1%), $C_{15:0}$ (9.3%), $C_{16:1}\omega7c$ (9.0%), i-$C_{15:0}$ 3-OH (6.2%), i-$C_{16:0}$ 3-OH (5.2%) and i-$C_{17:0}$ 3-OH (20.8%) [3]. The isoprenoid quinones of *C. lytica* were not determined, but for *C. pacifica* the presence of MK-6 as the major lipoquinone was described [4]. Polar lipids have not been studied.

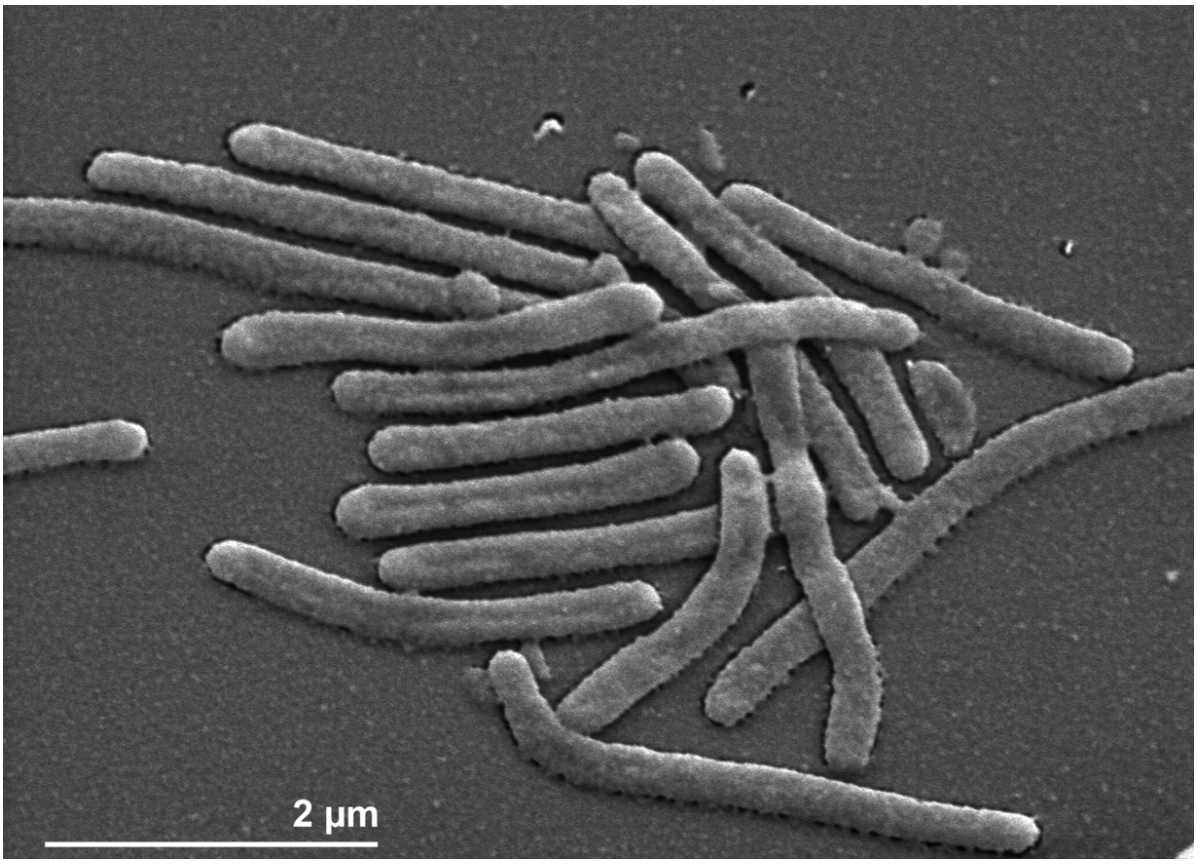

Figure 2. Scanning electron micrograph of *C. lytica* LIM-21^T

Table 1. Classification and general features of *C. lytica* LIM-21^T according to the MIGS recommendations [15].

MIGS ID	Property	Term	Evidence code
	Current classification	Domain *Bacteria*	TAS [16]
		Phylum '*Bacteroidetes*'	TAS [17]
		Class *Flavobacteria*	TAS [18]
		Order '*Flavobacteriales*'	TAS [19]
		Family *Flavobacteriaceae*	TAS [18,20-23]
		Genus *Cellulophaga*	TAS [1]
		Species *Cellulophaga lytica*	TAS [1]
		Type strain LIM-21	TAS [1]
	Gram stain	negative	TAS [24]
	Cell shape	rod-shaped	TAS [24]
	Motility	motile by gliding	TAS [24]
	Sporulation	none	TAS [24]
	Temperature range	4-40°C	TAS [4,25]
	Optimum temperature	22-30°C	TAS [25]
	Salinity	up to 8% NaCl	TAS [3,5]
MIGS-22	Oxygen requirement	aerobic	TAS [24]
	Carbon source	carbohydrates	TAS [24]
	Energy metabolism	chemoheterotroph	TAS [24]
MIGS-6	Habitat	mud	TAS [24]
MIGS-15	Biotic relationship	free-living	NAS
MIGS-14	Pathogenicity	none	NAS
	Biosafety level	1	TAS [26]
	Isolation	beach mud	TAS [24]
MIGS-4	Geographic location	Limon, Costa Rica	TAS [24]
MIGS-5	Sample collection time	1969	NAS
MIGS-4.1	Latitude	10.1	NAS
MIGS-4.2	Longitude	-83.5	NAS
MIGS-4.3	Depth	not reported	NAS
MIGS-4.4	Altitude	not reported	NAS

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 of the Gene Ontology project [27]. If the evidence code is IDA, then the property was directly observed by one of the authors or an 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 [29], and is part of the ***G**enomic **E**ncyclopedia of **B**acteria and **A**rchaea* project [30]. The genome project is deposited in the Genomes 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.

Table 2. Genome sequencing project information

MIGS ID	Property	Term
MIGS-31	Finishing quality	Finished
MIGS-28	Libraries used	Three genomic libraries: one 454 pyrosequence standard library, one 454 PE library (8 kb insert size), one Illumina library
MIGS-29	Sequencing platforms	Illumina GAii, 454 GS FLX Titanium
MIGS-31.2	Sequencing coverage	1,605.2 × (Illumina); 22.9 × (pyrosequence)
MIGS-30	Assemblers	Newbler version 2.5-internal-10Apr08, Velvet version 0.7.63, phrap version SPS-4.24
MIGS-32	Gene calling method	Prodigal 1.4, GenePRIMP
	INSDC ID	CP002534
	Genbank Date of Release	February 28, 2011
	GOLD ID	Gc01668
	NCBI project ID	50743
	Database: IMG-GEBA	2504136007
MIGS-13	Source material identifier	DSM 7489
	Project relevance	Tree of Life, GEBA

Growth conditions and DNA isolation

C. lytica LIM-21^T, DSM 7489, was grown in DSMZ medium 514 (BACTO marine broth) [31] at 28°C. DNA was isolated from 0.5-1 g of cell paste using MasterPure Gram-positive DNA purification kit (Epicentre MGP04100) following the standard protocol as recommended by the manufacturer with modification st/DL for cell lysis as described in Wu *et al*. [30]. DNA is available through the DNA Bank Network [32].

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 [33]. Pyrosequencing reads were assembled using the Newbler assembler version 2.5-internal-10Apr08 (Roche). The initial Newbler assembly consisting of 28 contigs in one scaffold was converted into a phrap version SPS - 4.24 [34] assembly by making fake reads from the consensus, to collect the read pairs in the 454 paired end library. Illumina GAii sequencing data (3,907 Mb) was assembled with Velvet [35] 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 156.1 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 [34] 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 [33], Dupfinisher [36], or sequencing cloned bridging PCR fragments with subcloning or transposon bombing (Epicentre Biotechnologies, Madison, WI). Gaps between contigs were closed by editing in Consed, by PCR and by Bubble PCR primer walks (J.-F.Chang, unpublished). A total of 238 additional reactions and two 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 [37]. 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 1,628.1 × coverage of the genome. The final assembly contained 282,018 pyrosequence and 78,832,334 Illumina reads.

Genome annotation

Genes were identified using Prodigal [38] as part of the Oak Ridge National Laboratory genome annotation pipeline, followed by a round of manual curation using the JGI GenePRIMP pipeline [39]. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant 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 [40].

Genome properties

The genome consists of a 3,765,936 bp long chromosome with a G+C content of 32.1% (Figure 3 and Table 3). Of the 3,358 genes predicted, 3,303 were protein-coding genes, and 55 RNAs; 19 pseudogenes were also identified. The majority of the protein-coding genes (65.5%) were assigned with 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.

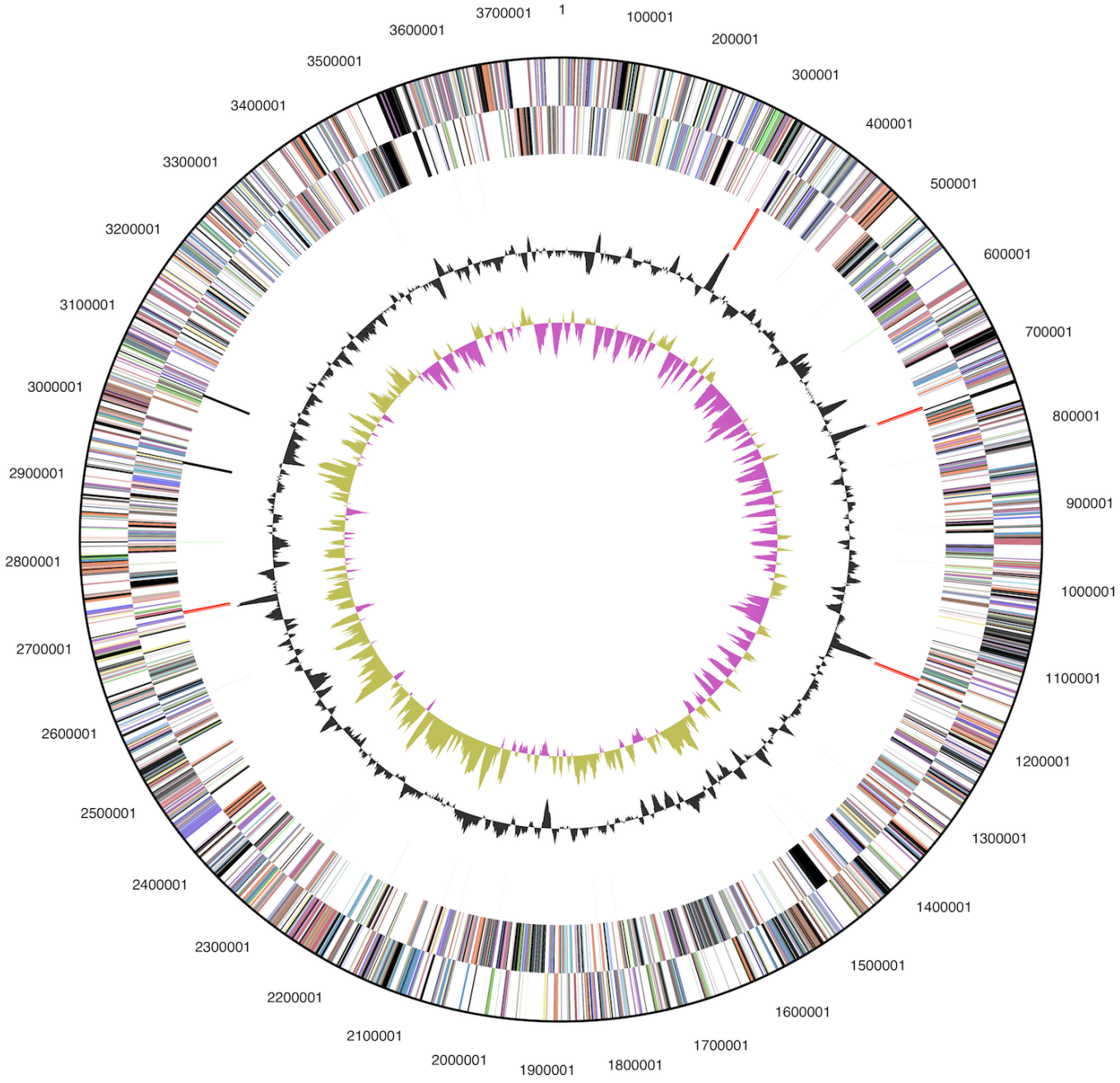

Figure 3. Graphical circular map of the chromosome. From outside to the center: Genes on forward strand (color by COG categories), Genes on reverse strand (color by COG categories), RNA genes (tRNAs green, rRNAs red, other RNAs black), GC content, GC skew.

Table 3. Genome Statistics

Attribute	Value	% of Total
Genome size (bp)	3,765,936	100.00%
DNA coding region (bp)	3,443,047	91.43%
DNA G+C content (bp)	1,209,276	32.11%
Number of replicons	1	
Extrachromosomal elements	0	
Total genes	3,358	100.00%
RNA genes	55	1.64%
rRNA operons	4	
Protein-coding genes	3,303	98.36%
Pseudo genes	19	0.57%
Genes with function prediction	2,200	65.52%
Genes in paralog clusters	344	10.24%
Genes assigned to COGs	2,098	62.48%
Genes assigned Pfam domains	2,346	69.86%
Genes with signal peptides	1,005	29.93%
Genes with transmembrane helices	794	23.65%
CRISPR repeats	0	

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

Code	value	%age	Description
J	154	6.7	Translation, ribosomal structure and biogenesis
A	0	0.0	RNA processing and modification
K	181	7.9	Transcription
L	107	4.7	Replication, recombination and repair
B	1	0.0	Chromatin structure and dynamics
D	18	0.8	Cell cycle control, cell division, chromosome partitioning
Y	0	0.0	Nuclear structure
V	41	1.8	Defense mechanisms
T	125	5.5	Signal transduction mechanisms
M	193	8.4	Cell wall/membrane/envelope biogenesis
N	4	0.2	Cell motility
Z	0	0.0	Cytoskeleton
W	0	0.0	Extracellular structures
U	30	1.3	Intracellular trafficking, secretion, and vesicular transport
O	99	4.3	Posttranslational modification, protein turnover, chaperones
C	122	5.3	Energy production and conversion
G	122	5.3	Carbohydrate transport and metabolism
E	194	8.5	Amino acid transport and metabolism
F	59	2.6	Nucleotide transport and metabolism
H	121	5.3	Coenzyme transport and metabolism
I	79	3.5	Lipid transport and metabolism
P	159	6.9	Inorganic ion transport and metabolism
Q	29	1.3	Secondary metabolites biosynthesis, transport and catabolism
R	276	12.1	General function prediction only
S	177	7.7	Function unknown
-	1,260	37.5	Not in COGs

Insights from the genome sequence

A closer look at the genome sequence of strain LIM-21^T revealed a set of genes which might be responsible for the yellow-orange color of *C. lytica* cells by encoding enzymes that are involved in the synthesis of carotenoids. Carotenoids are produced by the action of geranylgeranyl pyrophosphate synthase (Celly_1682), phytoene synthase (Celly_0459), phytoene desaturase (Celly_0458), lycopene cyclase (Celly_0462) and carotene hydroxylase (Celly_0461). Geranylgeranyl pyrophosphate synthases start the biosynthesis of carotenoids by combining farnesyl pyrophosphate with C_5 isoprenoid units to C_{20}-molecules, geranylgeranyl pyrophosphate. The phytoene synthase catalyzes the condensation of two geranylgeranyl pyrophosphate molecules followed by the removal of diphosphate and a proton shift leading to the formation of phytoene. Sequential desaturation steps are conducted by the phytoene desaturase followed by cyclization of the ends of the molecules catalyzed by the lycopene cyclase [41]. This above mentioned set of genes was also found in the genome of *C. algicola* [14].

Strain LIM 21^T produces a wide range of extracellular enzymes degrading proteins and polysaccharides. *C. lytica*, like the other members of the genus *Cellulophaga*, cannot hydrolyze filter paper or cellulose in its crystalline form, though they can hydrolyze the soluble cellulose derivative carboxymethylcellulose (CMC). The genome sequence of strain LIM 21^T revealed the presence of three cellulases (Celly_0269, Celly_0304, Celly_0965), probably responsible for the hydrolysis of CMC. In addition two β-glucosidases (Celly_3249, Celly_1282) were identified in the genome, catalyzing the breakdown of the glycosidic β-1,4 bond between two glucose molecules in cellobiose. The deduced amino acid sequence of Celly_0304 shows 90% identity to the deduced sequence of the *C. algicola* cellulase coding gene Celly_0025. The identity of the deduced amino acid sequences of the cellulase encoding genes Celly_0269 and Celly_2753 is 65%. The neighborhoods of these two *C. lytica* cellulase genes have a similar structure like the respective genome regions in *C. algicola*, with orthologs belonging to different COG categories.

The LIM 21^T genome contains 15 genes coding for sulfatases, which are located in close proximity to glycoside hydrolase genes suggesting that sulfated polysaccharides may be used as substrates. α-L-fucoidan could be a substrate, as three α-L-fucosidases (Celly_0440, Celly_0442, Celly_0449) are located in close proximity to nine sulfatases (Celly_0432, Celly_0425, Celly_0426, Celly_0436, Celly_0431, Celly_0433, Celly_0435, Celly_0438, Celly_0444). Sakai and colleagues report the existence of intracellular α-L-fucosidases and sulfatases, which enable '*Fucophilus fucoidanolyticus*' to degrade fucoidan [42].

The above mentioned sulfatases and fucosidases containing region of *C. lytica* is similar to the recently described region of *C. algicola* with five α-L-fucosidases and three sulfatases [14]. This fucoidan degrading ability could be also shared by *Coraliomargarita akajimensis*, as the annotation of the genome sequence revealed the existence of 49 sulfatases and 12 α-L-fucosidases [43].

Comparative genomics

The genomes of the two recently sequenced *Cellulophaga* type strains differ significantly in their size, *C. lytica* having 3.8 Mb and *C. algicola* 4.9 Mb and their number of pseudogenes, 19 (0.6%) and 122 (2.8%), respectively. Liu *et al*., 2004 have shown that pseudogenes in prokaryotes are not uncommon; the analysis of 64 genomes, including archaea, pathogen and nonpathogen bacteria, revealed an occurrence of pseudogenes of at least 1-5% of all gene-like sequences, with some genomes containing considerably higher amounts [44].

To estimate the overall similarity between the genomes of *C. lytica* and *C. algicola* the GGDC-Genome-to-Genome Distance Calculator [45,46] was used. The system calculates the distances by comparing the genomes to obtain HSPs (high-scoring segment pairs) and interfering distances from the set of formulas (1, HSP length / total length; 2, identities / HSP length; 3, identities / total length). The comparison of the genomes of *C. lytica* with *C. algicola* revealed that 25% of the average of both genome lengths are covered with HSPs. The identity within these HSPs was 82%, whereas the identity over the whole genome was only 20%. These results demonstrate that according to the whole genomes of *C. lytica* and *C. algicola* the similarity is not very high, although the comparison of 16S rRNA gene sequences showed only 7.7% differences.

In order to compare the *C. lytica* and *C. algicola* genomes, correlation values (Pearson coefficient) according to the similarity on the level of COG cat-

egory, pfam, enzymes and TIGRfam were calculated. The highest correlation value (0.98) was reached on the level of pfam data; the correlation values on the basis of COG, enzyme and TIGRfam data were 0.88, 0.92 and 0.93, respectively. As a correlation value of 1 indicates the highest correlation, we can find a quite high correlation between the genomes of *C. lytica* and *C. algicola* at least considering the pfam data [40].

The comparison of the number of genes belonging to the different COG categories revealed no large differences in the genomes of *C. lytica* and *C. algicola*. A slightly higher fraction of genes belonging to the categories transcription (*C. lytica* 8.63%, *C. algicola* 6.85%), translation (*C. lytica* 7.34%, *C. algicola* 6.30%), amino acid metabolism (*C. lytica* 9.25%, *C. algicola* 8.19%), inorganic ion transport and metabolism (*C. lytica* 7.58%, *C. algicola* 6.85%) and posttranslational modification (*C. lytica* 4.72%, *C. algicola* 3.90%) were identified in *C. lytica*. The part of genes belonging to the following COG categories was slightly smaller in *C. lytica* than in *C. algicola*: carbohydrate metabolism (*C. lytica* 5.82%, *C. algicola* 6.77%), defense mechanisms (*C. lytica* 1.95%, *C. algicola* 2.48%), secondary metabolites biosynthesis (*C. lytica* 1.38%, *C. algicola* 2.05%).

The synteny dot plot in Figure 4 shows a nucleotide based comparison of the genomes of *C. lytica* and *C. algicola*. Only in some parts of the genome a relatively high degree of similarity becomes visible. There exists a fragmented collinearity between the two genomes.

The Venn-diagram (Figure 5) shows the number of shared genes. *C. lytica* and *C. algicola* share a great number of genes (592 genes) that are not present in the genome of *Flavobacterium johnsoniae* [47]. This fraction of genes includes genes coding for enzymes that are responsible for the degradation of polysaccharides, for example fucoidan and agar. While 15 sulfatases and three α-L-fucosidases were identified in the genome of *C. lytica,* and 22 sulfatases and five α-L-fucosidases were identified in the genome of *C. algicola*, only four sulfatase genes and no α-L-fucosidase genes were identified in the genome of *F. johnsoniae*. In addition, three agarases were i*ntified in the genomes of *C. lytica* and *C. algicola*, each, whereas the genome of *F. johnsoniae* contains no agarase gene. *F. johnsoniae* is a chitin hydrolyzing organism; the genes involved in the utilization of chitin were described by McBride *et al.* (2009) [47]. *C. lytica* [1,25] and *C. algicola* [3] are non-chitinolytic, and there were no homologs to the chitin utilizing loci of *F. johnsoniae* identified in their genomes. To the group of genes that are shared by all three genomes belong the genes that code for enzymes which are involved in the biosynthesis of carotenoids, e.g. phytoene desaturase and phytoene synthase. But in contrast to the *Cellulophaga* species *F. johnsoniae* also produces flexirubin. The genes which are involved in the flexirubin synthesis of *F. johnsoniae* were identified by McBride *et al.* (2009) [47].

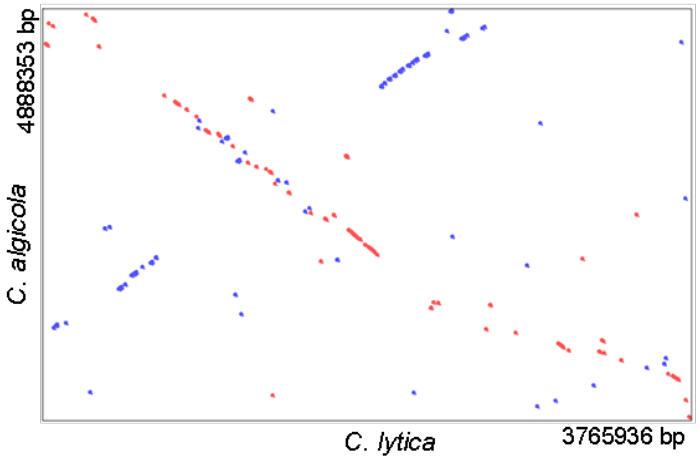

Figure 4. Synteny dot blot based on the genome sequences of *C. lytica* and *C. algicola*. Blue dots represent regions of similarity found on parallel strands and red dots show regions of similarity found on anti-parallel strands.

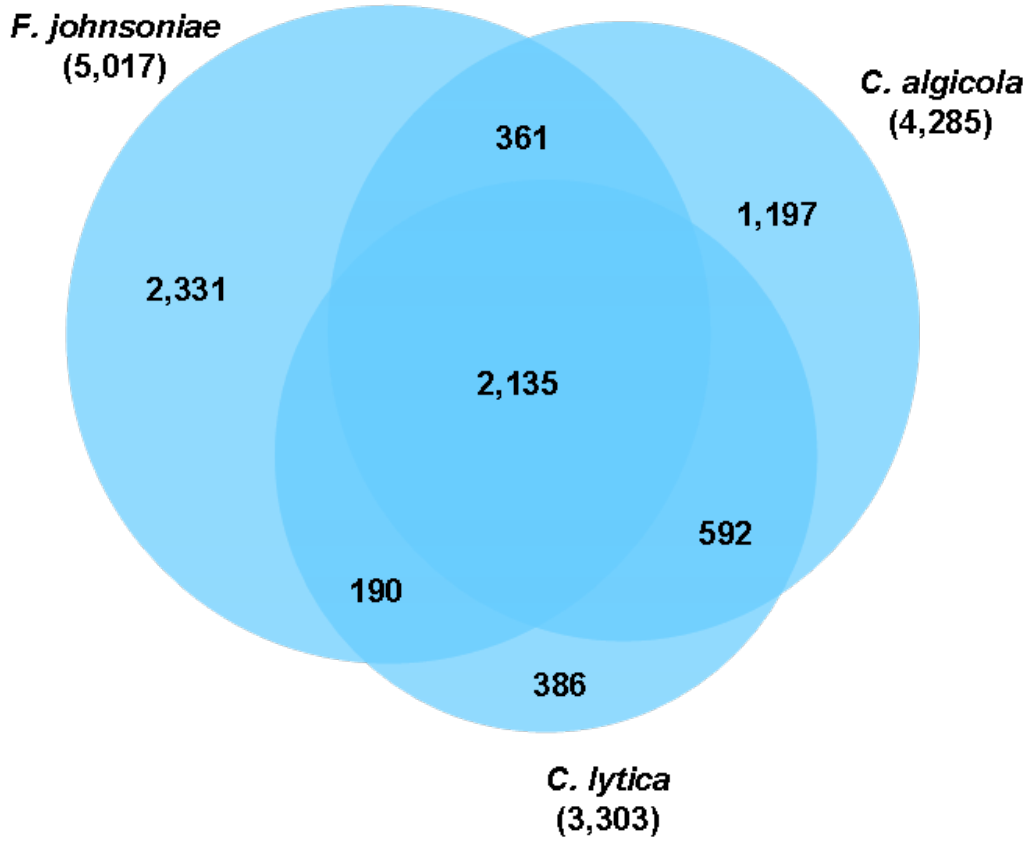

Figure 5. Venn-diagram depicting the intersections of protein sets (total numbers in parentheses) of *C. lytica*, *C. algicola* and *F. johnsoniae*.

Acknowledgements

We would like to gratefully acknowledge the help of Maren Schröder (DSMZ) for growing *C. lytica* cultures. 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, UT-Battelle and Oak Ridge National Laboratory under contract DE-AC05-00OR22725, as well as German Research Foundation (DFG) INST 599/1-2.

References

1. Johansen JE, Nielsen P, Sjøholm C. Description of *Cellulophaga baltica* gen. nov., sp. nov. and *Cellulophaga fucicola* gen. nov., sp. nov. and reclassification of [*Cytophaga*] *lytica* to *Cellulophaga lytica* gen. nov., comb. nov. *Int J Syst Bacteriol* 1999; **49**:1231-1240. PubMed doi:10.1099/00207713-49-3-1231

2. Euzéby JP. List of bacterial names with standing in nomenclature: A folder available on the Internet. *Int J Syst Bacteriol* 1997; **47**:590-592. PubMed doi:10.1099/00207713-47-2-590

3. Bowman JP. Description of *Cellulophaga algicola* sp. nov., isolated from the surfaces of Antarctic algae, and reclassification of *Cytophaga uliginosa* (ZoBell and Upham 1944) Reichenbach 1989 as *Cellulophaga uliginosa* comb. nov. *Int J Syst Evol Microbiol* 2000; **50**:1861-1868. PubMed

4. Nedashkovskaya OI, Suzuki M, Lysenko AM, Snauwaert C, Vancanneyt M, Swings J, Vysotskii MV, Mikhailov VV. *Cellulophaga pacifica* sp. nov. *Int J Syst Evol Microbiol* 2004; **54**:609-613. PubMed doi:10.1099/ijs.0.02737-0

5. Kahng HY, Chung BS, Lee DH, Jung JS, Park JH, Joen CO. *Cellulophaga tyrosinoxydans* sp. nov., a tyrosinase producing bacterium isolated from seawater. *Int J Syst Evol Microbiol* 2009; **59**:654-657. PubMed doi:10.1099/ijs.0.003210-0

6. Skerman VBD, McGowan V, Sneath PHA, eds. Approved Lists of Bacterial Names. [Approved

Lists of Bacterial Names in IJSEM Online - Approved Lists of Bacterial Names Amended edition]. *Int J Syst Bacteriol* 1980; **30**:225-420. doi:10.1099/00207713-30-1-225

7. 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
8. Porter MF. An algorithm for suffix stripping. *Program: electronic library and information systems* 1980; **14**:130-137. doi:10.1108/eb046814
9. 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
10. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000; **17**:540-552. PubMed
11. 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
12. 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
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. Abt B, Lu M, Misra M, Han C, Nolan M, Lucas S, Hammon N, Deshpande S, Cheng JF, Tapia R, *et al*. Complete genome sequence of *Cellulophaga algicola* type strain (IC166^T). *Stand Genomic Sci* 2011; **4**:72-80. PubMed doi:10.4056/sigs.1543845
15. 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
16. 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
17. 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.
18. Ludwig W, Euzeby J, Whitman WG. Draft taxonomic outline of the *Bacteroidetes, Planctomycetes, Chlamydiae, Spirochaetes, Fibrobacteres, Fusobacteria, Acidobacteria, Verrucomicrobia, Dictyoglomi*, and *Gemmatimonadetes*. http://www.bergeys.org/outlines/Bergeys_Vol_4_Outline.pdf. Taxonomic Outline 2008;
19. Garrity GM, Holt J. Taxonomic outline of the *Archaea* and *Bacteria*. *In: Bergey's Manual of Systematic Bacteriology*, 2nd ed. vol. 1. *The Archaea, deeply branching and phototrophic bacteria*. Garrity GM, Boone DR, Castenholz RW (*eds*). 2001; 155-166.
20. Bernardet JF, Nakagawa Y, Holmes B. Proposed minimal standards for describing new taxa of the family *Flavobacteriaceae* and emended description of the family. *Int J Syst Evol Microbiol* 2002; **52**:1049-1070. PubMed doi:10.1099/ijs.0.02136-0
21. List Editor. Validation of the publication of new names and new combinations previously effectively published outside the IJSB. List No. 41. *Int J Syst Bacteriol* 1992; **42**:327-328. doi:10.1099/00207713-42-2-327
22. Reichenbach H. Order 1. Cytophagales Leadbetter 1974, 99AL. *In*: Holt JG (ed), Bergey's Manual of Systematic Bacteriology, First Edition, Volume 3, The Williams and Wilkins Co., Baltimore, 1989, p. 2011-2013.
23. Bernardet JF, Segers P, Vancanneyt M, Berthe F, Kersters K, Vandamme P. Cutting a Gordian knot: emended classification and description of the genus *Flavobacterium*, emended description of the family *Flavobacteriaceae*, and proposal of *Flavobacterium hydatis* nom. nov. (Basonym, *Cytophaga aquatilis* Strohl and Tait 1978). *Int J Syst Bacteriol* 1996; **46**:128-148. doi:10.1099/00207713-46-1-128
24. Lewin RA. A classification of flexibacteria. *J Gen Microbiol* 1969; **58**:189-206. PubMed
25. Reichenbach H. Genus I. *Cytophaga* Winogradsky 1929, 577, (AL) emend. *In:* Staley JT, Bryant MP, Pfenning N, Holt JG (*eds*). Bergey's manual of systematic bacteriology. Vol. 3. Balti-

more, Md. Williams & Wilkins, 1989, pp. 2015-2050.

26. BAuA. Classification of bacteria and archaea in risk groups. *TRBA* 2005; **466**:84.

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. Lewin RA, Lounsbery DM. Isolation, Cultivation and Characterization of *Flexibacteria*. *J Gen Microbiol* 1969; **58**:145-170. PubMed

29. 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

30. 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

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

32. 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

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

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

35. 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

36. Han C, Chain P. 2006. Finishing repeat regions automatically with Dupfinisher. *In*: Proceeding of the 2006 international conference on bioinformatics & computational biology. Arabina HR, Valafar H (eds), CSREA Press. June 26-29, 2006: 141-146.

37. 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.

38. 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

39. 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

40. 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

41. Sandmann G. Carotenoid biosynthesis and biotechnological application. *Arch Biochem Biophys* 2001; **385**:4-12. PubMed doi:10.1006/abbi.2000.2170

42. Sakai T, Ishizuka K, Kato I. Isolation and characterization of fucoidan-degrading marine bacterium. *Mar Biotechnol* 2003; **5**:409-416. PubMed doi:10.1007/s10126-002-0118-6

43. Mavromatis K, Abt B, Brambilla E, Lapidus A, Copeland A, Desphande S, Nolan M, Lucas S, Tice H, Cheng JF. Complete genome sequence of *Coraliomargarita akajimensis* type strain (04OKA010-24 ^{T}). *Stand Genomic Sci* 2010; **2**:290-299. PubMed doi:10.4056/sigs.952166

44. Liu Y, Harrison PM, Kunin V, Gerstein M. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol* 2004; **5**:R64. PubMed doi:10.1186/gb-2004-5-9-r64

45. 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

46. 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

47. McBride MJ, Xie G, Martens EC, Lapidus A, Henrissat B, Rhodes RG, Goltsman E, Wang W, Xu J, Hunnicutt DW. Novel features of the polysaccharide-digesting gliding bacterium *Flavobacterium johnsoniae* as revealed by genome sequence analysis. *Appl Environ Microbiol* 2009; **75**:6864-6875. PubMed doi:10.1128/AEM.01495-09