Contents lists available at ScienceDirect

Data in Brief

Data Article

Draft genome sequence data of *Lysinibacillus fusiformis* strain GM, isolated from potato phyllosphere as a potential probiotic

Daria S. Pudova [a,*], Marat T. Lutfullin [a], Elena I. Shagimardanova [a],
Guzel F. Hadieva [a], Leyla Shigapova [a], Anna A. Toymentseva [a],
Daniil A. Kabanov [a], Ayslu M. Mardanova [a], Semen G. Vologin [b],
Margarita R. Sharipova [a]

[a] *Institute of Fundamental Medicine and Biology, Kazan (Volga region) Federal University, Kazan, Russia*
[b] *TatSRIA - Subdivision of FIC KazanSC of RAS, Kazan, Russia*

ARTICLE INFO

ABSTRACT

Here we present the morphological and physiological properties of isolated *Lysinibacillus fusiformis* strain GM, its draft genome sequence as well as annotation and analysis of its genome. Initial analysis of MALDI-TOF mass spectrometry, 16S rRNA gene analysis and *in silico* DNA-DNA hybridization revealed that the strain belongs to the species *Lysinibacillus fusiformis*. The 4,678,122 bp draft genome consist of 17 scaffolds encoding 4588 proteins and 137 RNAs. Annotation of the genome sequence revealed cellulase and protease encoding genes, genes of adhesion proteins and putative genes responsible for the biosynthesis of antimicrobial metabolites. The Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession number NTMQ00000000.1 (https://www.ncbi.nlm.nih.gov/nuccore/NZ_NTMQ00000000.1).

* Corresponding author.
  E-mail addresses: dasha171711@gmail.com (D.S. Pudova), lutfullin.marat2012@yandex.ru (M.T. Lutfullin),
rjuka@mail.ru (E.I. Shagimardanova), g.h95@mail.ru (G.F. Hadieva), shi-leyla@yandex.ru (L. Shigapova),
tojmencevaaa@mail.ru (A.A. Toymentseva), dac100@mail.ru (D.A. Kabanov), mardanovaayslu@mail.ru (A.M. Mardanova),
semen_vologin@mail.ru (S.G. Vologin), marsharipova@gmail.com (M.R. Sharipova).

Specifications table

Subject area	Biology
More specific subject area	Genomics, applied microbiology, probiotic
Type of data	Table, figure, text files
How data was acquired	SEM, phylogeny, Illumina MiSeq
Data format	Analyzed
Experimental factors	Strain GM was isolated from the phyllosphere of potato leaves for morphological and phylogenetic analysis. Bacterial genomic DNA was extracted and sequenced by Illumina MiSeq. Sequenced genome was used for assembly, annotation and search of genes which can be useful for using of strain as a probiotic.
Experimental features	Description of morphological and physiological properties of isolated strain GM. It's phylogenetic position and genome sequence annotation.
Data source location	The sample was collected in Kazan, Republic of Tatarstan, Russia
Data accessibility	Data of whole genome shotgun sequencing project uploaded to NCBI database under accession number NTMQ00000000.1 (https://www.ncbi.nlm.nih.gov/nuccore/NZ_NTMQ00000000.1).

Value of the data

- Data can be used for further experiments towards studying the *Lysinibacillus fusiformis* GM strain as a perspective probiotic.
- The resulting data can significantly contribute to the development of new feed additives for poultry industry.
- Data allow broadening the understanding of the poorly studied genus of *Lysinibacillus*.

1. Data

Morphologically isolated cells of the GM strain were found to be Gram-positive *Bacilli*, with white creamish, round colonies after the first day of culturing at 37 °C (Fig. 1a). Scanning electron microscopy (SEM) of isolated culture showed the presence of rod-shaped cells that are approximately 0.6 μm in width and 2.0–2.75 μm in length (Fig. 1b, c). Spores appeared after 13 to 14 h of growth in LB medium. After 36 h of growth, spores comprised about 22%. The SEM analysis showed that the spores were oval, 0.6 μm in width and 1 μm in length. (Fig. 1b, c).

Strain GM showed a negative reaction to amylase, pectinase, phytase, but positive with protease and cellulase. It showed resistance to gentamycin, levomycetin, erythromycin, kanamycin and azithromycin, whereas it was sensitive to benzylpenicillin, ampicillin and tetracycline. Analysis of Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) revealed that the studied isolate is a representative of the genus *Lysinibacillus*. 16S rRNA gene analysis and *in silico* DNA-DNA hybridization (*is*DDH) were used for more accurate identification of isolates. The phylogenetic relationship of *L. fusiformis* GM to other species within the genus *Lysinibacillus* is visualized in a 16S rRNA based tree (Fig. 2).

Strain GM showed 97% ANI and 77.2% GGDC score to the *L. fusiformis* RB-21 and *L. fusiformis* SW-B9 strains, which corresponded to the recommended thresholds (95% for ANI and 70% for GGDC) for the identification of the species [1,2]. Morphological and phylogenetic analysis confirmed that strain GM is *L. fusiformis*, belonging to the phylum *Firmicutes* and class *Bacilli*. The genome sequence of *L. fusiformis* GM assembled in 42 contigs (> 200 bp), which are combined into 17 scaffolds with a calculated genome size of 4,678,122 bp and GC content of 37.43 mol %. The N50 size of the resulted contigs was 2,538,659 bp. The general statistics of genome are shown in Table 1. The genome sequence of *L. fusiformis* GM (Laboratory of
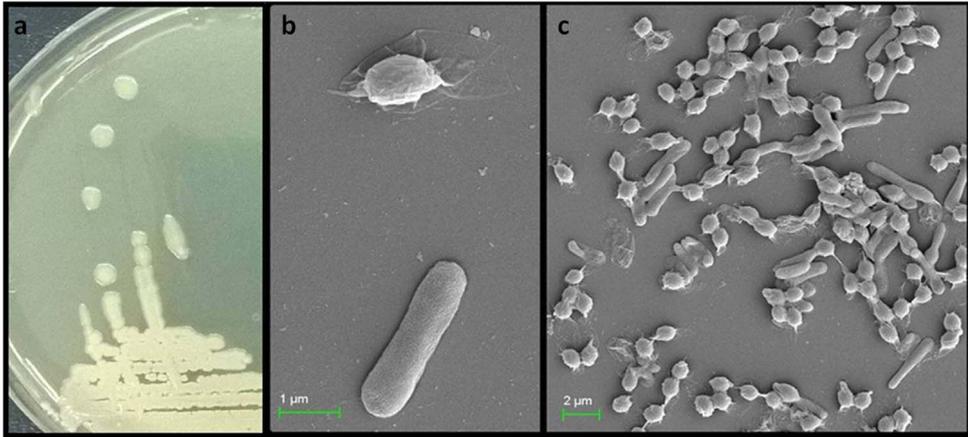
Fig. 1. Morphology of colonies (a) and scanning electron microscopic (SEM) analysis of cells and spores of _L. fusiformis_ GM strain (b, c).
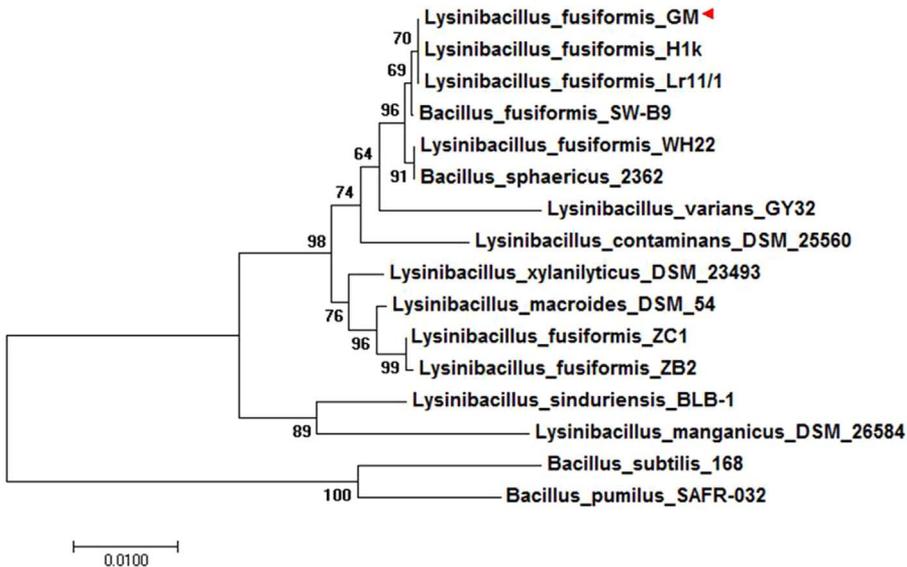
Fig. 2. Phylogenetic tree showing the position of _L.fusiformis_ GM relative to other species within the genus _Lysinibacillus_ including _Bacillus_ strains as an out-group (GenBank accession numbers for all represented 16S rRNA sequences are available in Additional file 1). The phylogenetic tree is based on 16S rRNA gene alignments and was obtained by MEGA 7.0.14 software. Phylogenetic tree was generated using the Maximum likelihood (ML) algorithm with 1000 bootstrap iterations.

Biosynthesis and Bioengineering of enzymes, Kazan Federal University, Republic of Tatarstan, Russia) was completed on September 5, 2017 and has been deposited to GenBank as the Whole Genome Shotgun project under the accession number NTMQ00000000.1.

The genome of GM strain contained 4,725 total genes including 4,588 protein-coding genes (CDSs), 25 rRNA (12 5S rRNA, 9 16S rRNA and 4 23S rRNA), 83 tRNA. Among all protein-coding genes 857 CDSs were annotated as hypothetical proteins. According to the KEGG pathway database, 1,222 protein-coding genes were connected to KEGG pathways and 2,240 protein-coding genes were connected to KEGG orthology. A total of 3056 protein-coding genes were allocated to COG (Clusters of Orthologous Groups) clusters.

Table 1
Statistics of genome assembly and annotation.

Attribute	Value
Genome size (bp)	4,678,122
DNA G+C (%)	37.43
DNA contigs (> 200 bp)	42
DNA scaffolds	17
Total genes	4843
Protein coding genes	4588
RNA genes	137
Pseudo genes	118
Genes assigned to COGs	3056

We identified a gene encoding a cellulase (glycosylhydrolase family 5 (pfam00150)), responsible for the cleavage of cellulose in the external environment. Cellulase converts the highly recalcitrant cellulose to fermentable mono- and oligosaccharides that can be easily assimilated in the body, thereby improving utilization of dietary carbohydrate and enhancing digestion [3]. In addition to genes responsible for cellulolytic activity, the genes of other hydrolytic enzymes – proteinases – were identified. The IMG annotation identified the gene of extracellular proteinase–subtilisin (pfam00082). Proteases, hydrolyzing peptide bonds, increase the digestion of macromolecules in animal feed and improve feed intake by increasing nutrient absorption in host animals [4]. The genes of antimicrobial substances were identified. Annotation of IMG ER system was predicted six gene clusters linked to the biosynthesis of potentially interesting metabolites. One of the identified gene clusters was predicted to be involved in the bacteriocin biosynthesis. Bacteriocins are ribosomally synthesized antimicrobial polypeptides that have been employed against many Gram-positive and Gram-negative pathogens, including Gram-positive spore-forming *Clostridium* (*C. perfringens*), Gram-negative *Campylobacter* (*C. jejuni*) and *Salmonella* (*S. enterica*) [5,6]. Analysis of bacteriocins genes was carried out using the BAGEL3 database [7]. It was shown that the GM strain possesses a cluster of genes, responsible for the synthesis of bacteriocin, relating to lanthipeptides. Cluster was identified on 4 scaffolds and has genes of lanthipeptide dehydratase domain (Lant_dehyd_N, pfam04737) and so-called SpaB_C (SpaB C-terminal) domain (SpaB_C, pfam14028). Probiotic strains that carry as many acquired anti-biotic resistance genes as possible are more preferable in biotechnology [8]. The strain was resistant to three antibiotics (erythromycin, gentamicin, and kanamycin), which are included in the six key antibiotics (chloramphenicol, erythromycin, gentamicin, tetracycline, streptomycin, and kanamycin) and are recommended by the European Food Safety Authority (EFSA) [9]. In addition, we searched for the presence of antibiotic resistance genes and related efflux pumps in the chromosome. Annotation of RAST server revealed genes of resistance to vancomycin (VanR (pfam00072, pfam00486), VanS (pfam00512, pfam02518), VanW (pfam04294)), streptothricin (SatA (pfam00583)), erythromycin (Erythromycin esterase (pfam05139)), fosfomycin (FosB (pfam00903), beta-lactams (BL (pfam13354), BLc (pfam00144)) and five genes of multidrug resistance efflux pumps (MatE (pfam01554), ACR_tran (pfam00873)). Using the Pfam database we identified the genes of adhere and stress-responsive proteins: fibronectin-binding protein A N-terminus (FbpA) domain (pfam05833), S-layer proteins (pfam00395), HisKA domain (pfam00512), HSP70 (pfam00012) domain, two chaperone protein DnaJ (pfam00226) and three proteins corresponding to universal stress proteins (pfam00582).

2. Experimental design, materials, and methods

2.1. Morphological and physiological analysis

Morphology of isolated GM strain was determined by light and scanning electron microscopy (SEM). The activity of proteolytic enzymes was tested for degradation of casein [10]. Secretion of enzymes for cellulose degradation was investigated using a nutrient medium supplemented with

carboxy-methyl cellulose (CMC), which was further stained with Congo red solution [11]. Drug sensitivity test carried out disc diffusion method using a disc of standard antibiotics.

2.2. Phylogeny analysis

The initial phylogeny of the isolate GM was studied using a Matrix-assisted laser desorption/ ionization time-of-flight mass spectrometry (MALDI-TOF MS), 16S rRNA and isDDH analysis. Phylogenetic analysis of the strain GM was performed using MEGA 7.0.14 software with 16S rRNA gene sequences of 14 *Lysinibacillus* strains and 2 *Bacillus* strains as an out-group (genes for all these species are available in NCBI database). Sequences were aligned using ClustalW algorithm. The remaining alignment sites (1382 bp) were selected for the subsequent analysis. The isDDH analysis was performed against other available *L. fusiformis* strains for Average Nucleotide Identity (ANI) and Genome-to-Genome Distance Calculate (GGDC). ANI values was calculated using JSpecies v1.2.1 [12] web server based on the MUMmer algorithm (ANIm) and GGDC values was calculated using Genome-to-Genome Distance Calculator v2.1 [2] based on the BLAST+ tool. The strain was deposited in the museum of the laboratory "Biosynthesis and Bioengineering of Enzymes" (Kazan Federal University, Russia).

2.3. Genomic DNA preparation

Before genomic DNA extraction *L. fusiformis* strain GM was inoculated in 50 ml of LB medium and grown overnight at 37 °C with rocking rate of 200 rpm. 50 mL were centrifuged at 5000x*g* for 10 min at 4 °C and genomic DNA was extracted using a standard phenol–chloroform extraction method [13]. The quality of the final DNA sample were evaluated by gel electrophoresis (1% agarose gel) and DNA concentration was estimated by using a NanoDrop 2000c Spectrophotometer (Thermo Scientific). In total, 6100 ng/μL of genomic DNA was received and sent for the sequencing.

2.4. Genome sequencing and assembly

The genomic DNA of *L. fusiformis* GM was sequenced using an Illumina MiSeq platform. The quality of the reads was checked in FastQC v0.11.3 program [14]. Trimming and filtration of reads was carried out in Trimmomatic v3.32 [15]. Purified reads were assembled de novo using SPAdes v3.8.1 assembler [16], followed by scaffolding in MeDuSa v1.6 [17]. The statistics of assembled genome were calculated in the QUAST v2.3 program [18].

2.5. Genome annotation

Genes and pseudogenes of *L. fusiformis* GM were identified using the NCBI Prokaryotic Genomes Automatic Annotation Pipeline. Subsequently, the data analysis was performed using IMG ER v4.570 system, which is an integrated pipeline (Prodigal V2.6.3, COG, Ffam, TIGRfam, KEGG database et al.) including the automatic identification and characterization of all the potential genes, proteins, and enzymes [19]. The raw genomic data of *L. fusiformis* GM was processed following the standard protocol of IMG ER system. The genome sequence was also uploaded into the RAST Web server for automated assignment of protein-coding genes into the COGs [20].

Transparency document. Supporting information

Transparency data associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2018.11.107.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2018.11.107.

References

[1] N. Han, Q. Yujun, W. Zhang, ANItools web: a web tool for fast genome comparison within multiple bacterial strains, Database: J. Biol. Databases Curation (2016), https://doi.org/10.1093/database/baw084.

[2] A.F. Auch, H.-P. Klenk, M. Göker, Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs, Stand. Genom. Sci. (2010), https://doi.org/10.4056/sigs.541628.

[3] A.K. Manhar, Y. Bashir, D. Saikia, D. Nath, K. Gupta, B.K. Konwar, R. Kumar, et al., Cellulolytic potential of probiotic Bacillus Subtilis AMS6 isolated from traditional fermented soybean (Churpi): an in-vitro study with regards to application as an animal feed additive, Microbiol. Res. (2016), https://doi.org/10.1016/j.micres.2016.03.004.

[4] A.K. Ray, K. Ghosh, E. Ringø, Enzyme-producing bacteria isolated from fish gut: a review, Aquacult. Nutr. (2012), https://doi.org/10.1111/j.1365-2095.2012.00943.x.

[5] A.B. Lagha, B. Haas, M. Gottschalk, et al., Antimicrobial potential of bacteriocins in poultry and swine production, Veter-. Res. (2017), https://doi.org/10.1186/s13567-017-0425-6.

[6] L. Timbermont, L. De Smet, F. Van Nieuwerburgh, et al., Perfrin, a novel bacteriocin associated with netB positive Clostridium perfringens strains from broilers with necrotic enteritis, Veter-. Res. (2014), https://doi.org/10.1186/1297-9716-45-40.

[7] A.J. Van Heel, A. de Jong, M. Montalbán-López, J. Kok, O.P. Kuipers, BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides, Nucleic Acids Res. (2013), https://doi.org/10.1093/nar/gkt391.

[8] M. Gueimonde, B. Sanchez, C.G. de Los Reyes-Gavilán, A. Margolles, Antibiotic resistance in probiotic bacteria, Front. Microbiol. (2013), https://doi.org/10.3389/fmicb.2013.00202.

[9] S. Mingmongkolchai, W. Panbangred, in vitro evaluation of candidate *Bacillus* spp. for animal feed, J. Gen. Appl. Microbiol. (2017), https://doi.org/10.2323/jgam.2016.09.002.

[10] G. Berg, Plant-dependent genotypic and phenotypic diversity of antagonistic rhizobacteria isolated from different Verticillium host plants, Appl. Environ. Microbiol. (2002), https://doi.org/10.1128/AEM.68.7.3328-3338.2002.

[11] E. Baldan, S. Nigris, C. Romualdi, S. D'Alessandro, A. Clocchiatti, et al., Beneficial bacteria isolated from grapevine inner tissues shape Arabidopsis thaliana roots, PLoS One (2015), https://doi.org/10.1371/journal.pone.0140252.

[12] M. Richter, R. Rosselló-Móra, F.O. Glöckner, J. Peplies, JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison, Bioinformatics (2016), https://doi.org/10.1093/bioinformatics/btv681.

[13] S. Köchl, H. Niederstätter, W. Parson, DNA extraction and quantification of forensic samples using the phenol-chloroform method and real-time PCR, Methods Mol. Biol. 297 (2005) 13–29. https://doi.org/10.1385/1-59259-867-6:013.

[14] S. Andrews, FastQC: a quality control tool for high throughput sequence data ⟨http://www.bioinformatics.babraham.ac.uk/projects/fastqc⟩. (Accessed 26 Apr 2010), 2010.

[15] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics (2014), https://doi.org/10.1093/bioinformatics/btu170.

[16] A. Bankevich, S. Nurk, D. Antipov, et al., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. (2012), https://doi.org/10.1089/cmb.2012.0021.

[17] E. Bosi, B. Donati, M. Galardini, S. Brunetti, et al., MeDuSa: a multi-draft based scaffolder, Bioinformatics (2015), https://doi.org/10.1093/bioinformatics/btv171.

[18] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics (2013), https://doi.org/10.1093/bioinformatics/btt086.

[19] V.M. Markowitz, I.-M.A. Chen, K. Palaniappan, et al., IMG: the integrated microbial genomes database and comparative analysis system, Nucleic Acids Res. (2012), https://doi.org/10.1093/nar/gkr1044.

[20] R.K. Aziz, D. Bartels, A.A. Best, et al., The RAST server: rapid annotations using subsystems technology, BMC Genom. (2008), https://doi.org/10.1186/1471-2164-9-75.