Data in Brief 54 (2024) 110547

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Draft genome sequence data on *Bacillus safensis* U41 isolated from soils of Santiniketan, India

Binoy Kumar Show[a], Andrew B. Ross[b], Raju Biswas[c], Shibani Chaudhury[a], Srinivasan Balachandran[a,*]

[a] *Department of Environmental Studies, Siksha-Bhavana, Visva-Bharati, Santiniketan 731235, West Bengal, India*
[b] *School of Chemical and Process Engineering, University of Leeds, Leeds, LS2 9JT, United Kingdom*
[c] *Department of Botany, Siksha-Bhavana, Visva-Bharati, Santiniketan 731235, West Bengal, India*

ARTICLE INFO

Article history:
Received 10 April 2024
Revised 4 May 2024
Accepted 16 May 2024
Available online 21 May 2024

Dataset link: Microbe sample from Bacillus sp. U41 (Original data)

Keywords:
Cellulase
Enzyme
Lignocellulose
Bacillus safensis U41

ABSTRACT

The draft genome sequence of an isolate of *Bacillus safensis* U41 from the soils of Santiniketan (23040′12″ N and 87039′52″ E) is reported here. *Bacillus safensis* is a bacterium that produces cellulases, which is essential for the breakdown of plant biomass. As such, it is a valuable source of digestive enzymes from plant biomass, especially cellulases. The genomic DNA was extracted from a single colony using a QIAgen Blood and Tissue kit (QIAgen Inc., Canada). Sequencing was performed via Illumina HiSeq X using 2 × 150 paired-end chemistry, generating 7,352,576 reads with sequence coverage of 509x. The assembly produced 20 contigs over 200 base pairs (bp) in length, with an N50 value of 901304 and an L50 of 2. The genome size was 3,732,407 bp, and the average GC content was 41.43 %. Genome annotation and gene predictions were performed using Prokka v.1.14.6, which identified 3783 coding sequences, 64 tRNA genes, and 3 rRNA genes.

[*] Corresponding author.
E-mail address: s.balachandran@visva-bharati.ac.in (S. Balachandran).

https://doi.org/10.1016/j.dib.2024.110547

Specifications Table

Subject	Microbiology • Applied Microbiology
Specific subject area	Omics: Genomics
Type of data	Table, Figure Raw, Analyzed, Filtered, Deposited
Data collection	The genomic DNA was extracted from a single colony using a QIAgen Blood and Tissue kit (QIAgen Inc., Canada). Sequencing was performed by Illumina HiSeq X. After FastQC check, trimming and size selection were performed by Cutadapt v2.9, the de novo assembly was performed by Unicycler v0.4.4, QUAST v.5.1.0 for assembly summary statistics, assembly pipeline using BayesHammer and assembly performed with SPAdes v3.13.0, Prokka v.1.14.6 used for genome annotation and gene predictions, and the PANZER webserver used for functional annotation of the proteins.
Data source location	The U41 strain was isolated from soils of Santiniketan, West Bengal, India (23.67760N, 87.68530 E).
Data accessibility	Repository name: NCBI (National Center for Biotechnology Information) GenBank Nucleotide database Data identification number: BioProject accession number PRJNA890884 Direct URL to data: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA890884 https://www.ncbi.nlm.nih.gov/sra/SRR22254258

1. Value of the Data

- The *B. safensis* U41 draft genome sequence may be useful for research on the taxonomy and ecology of bacteria, especially regarding taxonomic identification and dispersion.
- The researchers working in the fields of environmental microbiology, environmental biotechnology, genomics, and renewable energy may find value in the information provided in this article.
- This genomic sequence data of *B. safensis* strain U41 may be useful to researchers wishing to perform comparative genomic analysis between strains and environments.

2. Background

Bacillus safensis is a common Gram-positive, aerobic, spore-forming Bacillus found in soil [1]. Enzyme-based biodegradation can be a choice for developing appropriate methods for adequately utilizing biomass in a productive formulation. The plant biomasses used as substrates in bioenergy production comprise lignin, cellulose, and hemicelluloses. Cellulose shapes the principal portion of lignocellulose, surrounded by a hemicellulose matrix and the exterior by a lignin layer [2]. Cellulose consists of more than 100–140,000 d-glucose units and is condensed by β-1,4-glycosidic bonds or linkages and forms a straight-chain polymer. The occurrence of multiple hydroxyl groups on the glucose forms hydrogen bonds that hold the chain and make it steadier. Cellulose is hydrophilic, but its sizeable polymeric structure renders it less soluble in water [3,4]. In this work, the draft genome of cellulase-producing *B. safensis* U41 strain has been sequenced and analysed.

3. Data Description

The article presents the whole genome sequencing information of *Bacillus safensis* U41 and its cellulases gene, which is essential for the breakdown of plant biomass.

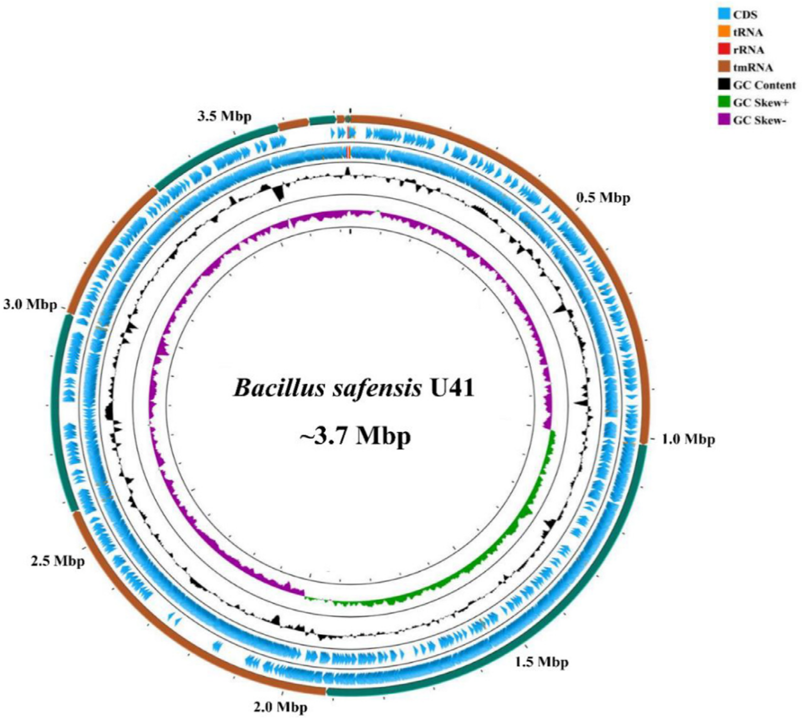

Fig. 1. Graphical presentation of strain U41 genome (~3.7 mb) performed with CG view server. The 6 concentric circles represent the following (from outermost to innermost): circles 2 contigs; circle 3 and 4, protein-coding genes on forward and reverse strands; circle 5 (black): G + C content; circle 6: G + C skew; circle 1: DNA base position (Mbp).

The assembly produced 20 contigs over 200 base pairs (bp) in length, with an N50 value of 901,304 and an L50 of 2. The genome shows a near-complete with low contamination (99.90% completeness and 0.20 % contamination), and sequencing coverage of 509x. The genome size was 3,732,407 bp, and the average GC content was 41.43 % (Fig. 1). Genome annotation produced 3783 CDS (coding sequences), 64 tRNA genes, and 3 rRNA genes (Table 1).

Strain U41 shared more than 97 % similarity of the 16S rRNA gene sequence with several members of the family *Bacillaceae*, which includes 15 species of *Bacillus*. (Supplementary Table S1). Among them, *B. safensis* subsp. *safensis* is the closest phylogenetic relative of strain U41 and shares 99.93 % sequence similarity. The maximum-likelihood tree of the 16S rRNA gene sequence of U41, with all its close relatives, clustered it on the same branch with *B. safensis* species (Fig. 2). A genome-based phylogenetic tree shows the higher bootstrap value (90 %) with the nearest

Table 1
General genomic features of *Bacillus safensis* U41.

Feature	Value
Genome size (bp)	3732,407
G + C content (%)	41.43
No. of contigs	20
N50	901,304
L50	2
CDS (coding sequences)	3783
tRNA	64
rRNA	3

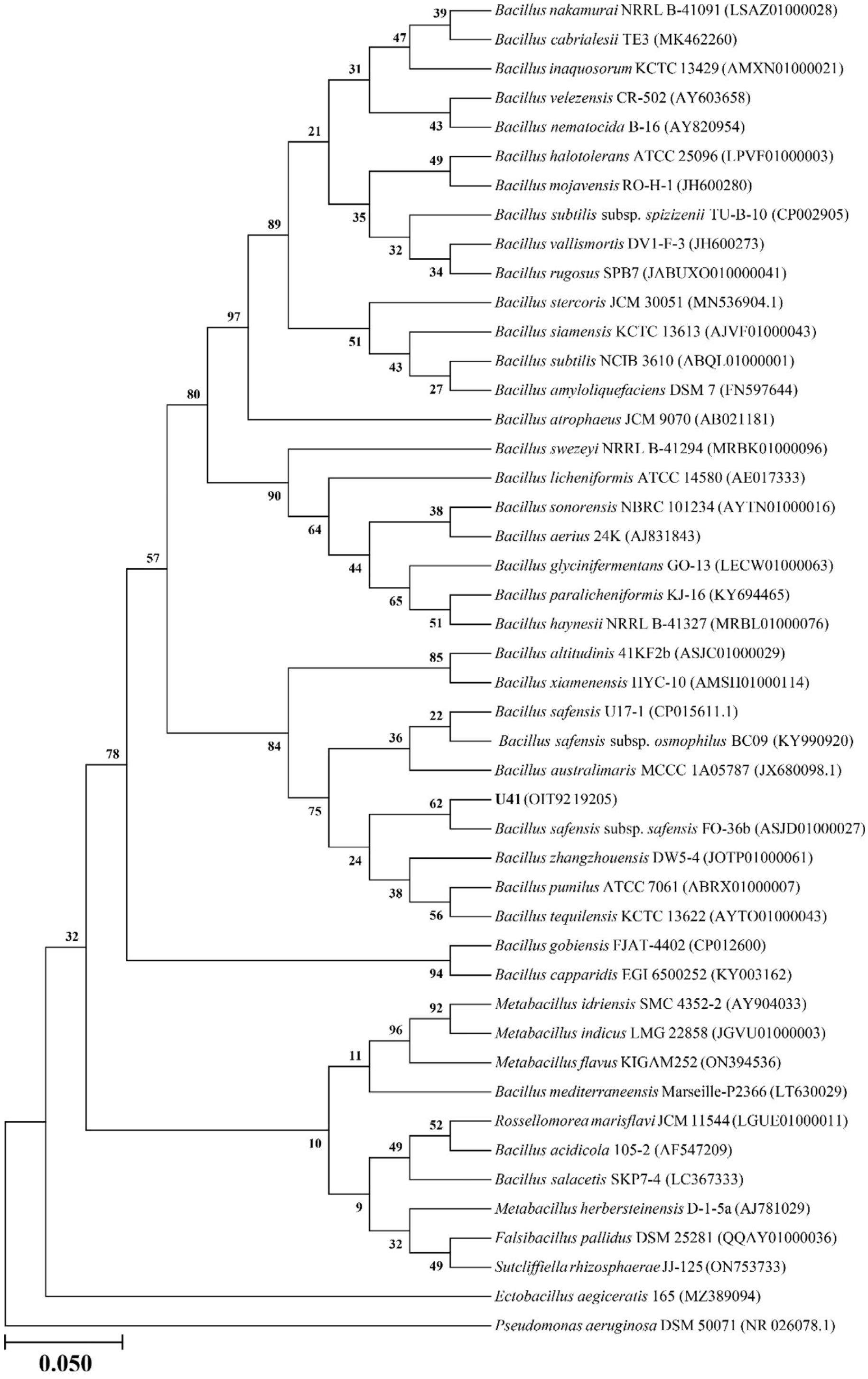
39
Bacillus nakamurai NRRL B-41091 (LSAZ01000028)
47
Bacillus cabrialesii TE3 (MK462260)
31
Bacillus inaquosorum KCTC 13429 (AMXN01000021)
Bacillus velezensis CR-502 (AY603658)
21
43
Bacillus nematocida B-16 (AY820954)
49
Bacillus halotolerans ATCC 25096 (LPVF01000003)
Bacillus mojavensis RO-H-1 (JH600280)
89
35
Bacillus subtilis subsp. *spizizenii* TU-B-10 (CP002905)
32
Bacillus vallismortis DV1-F-3 (JH600273)
34
Bacillus rugosus SPB7 (JABUXO010000041)
97
Bacillus stercoris JCM 30051 (MN536904.1)
51
Bacillus siamensis KCTC 13613 (AJVF01000043)
43
Bacillus subtilis NCIB 3610 (ABQL01000001)
80
27
Bacillus amyloliquefaciens DSM 7 (FN597644)
Bacillus atrophaeus JCM 9070 (AB021181)
Bacillus swezeyi NRRL B-41294 (MRBK01000096)
90
Bacillus licheniformis ATCC 14580 (AE017333)
38
64
Bacillus sonorensis NBRC 101234 (AYTN01000016)
Bacillus aerius 24K (AJ831843)
57
44
Bacillus glycinifermentans GO-13 (LECW01000063)
65
Bacillus paralicheniformis KJ-16 (KY694465)
51
Bacillus haynesii NRRL B-41327 (MRBL01000076)
85
Bacillus altitudinis 41KF2b (ASJC01000029)
Bacillus xiamenensis HYC-10 (AMSH01000114)
22
Bacillus safensis U17-1 (CP015611.1)
84
36
Bacillus safensis subsp. *osmophilus* BC09 (KY990920)
78
Bacillus australimaris MCCC 1A05787 (JX680098.1)
75
62
U41 (OIT92 19205)
Bacillus safensis subsp. *safensis* FO-36b (ASJD01000027)
24
Bacillus zhangzhouensis DW5-4 (JOTP01000061)
38
Bacillus pumilus ATCC 7061 (ABRX01000007)
56
Bacillus tequilensis KCTC 13622 (AYTO01000043)
32
Bacillus gobiensis FJAT-4402 (CP012600)
94
Bacillus capparidis EGI 6500252 (KY003162)
92
Metabacillus idriensis SMC 4352-2 (AY904033)
96
Metabacillus indicus LMG 22858 (JGVU01000003)
11
Metabacillus flavus KIGAM252 (ON394536)
Bacillus mediterraneensis Marseille-P2366 (LT630029)
52
Rossellomorea marisflavi JCM 11544 (LGUE01000011)
10
49
Bacillus acidicola 105-2 (AF547209)
Bacillus salacetis SKP7-4 (LC367333)
9
Metabacillus herbersteinensis D-1-5a (AJ781029)
32
Falsibacillus pallidus DSM 25281 (QQAY01000036)
49
Sutcliffiella rhizosphaerae JJ-125 (ON753733)
Ectobacillus aegiceratis 165 (MZ389094)
Pseudomonas aeruginosa DSM 50071 (NR 026078.1)
0.050

Fig. 2. Construction of maximum likelihood phylogenetic tree elicited from the 16S rRNA gene sequences of *Bacillus safensis* U41 and the type strains of Bacillus sp. Bootstrap values (>50% are expressed as percentages of 1000 replications) are shown at branching points.

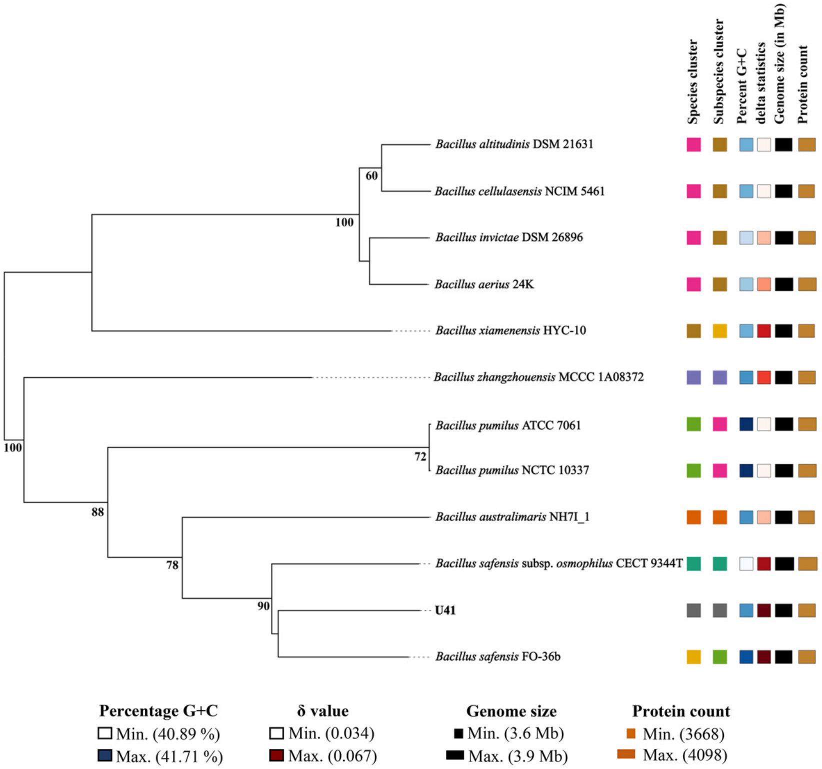

Fig. 3. TYGS tree designing based on whole genome sequence inferred with FastME 2.1.6.1 from GBDP distances calculated from genome sequences. The branch lengths are scaled in terms of GBDP distance formula d5. The numbers above branches are GBDP pseudo-bootstrap support values > 60% from 100 replications, with an average branch support of 78.3%. The tree was rooted at the midpoint.

relative species of *B. safensis* (Fig. 3). The digital DNA-DNA hybridization (dDDH) values of strain U41 are 90.8 %, with its closest relative, *B. safensis* species.

Furthermore, functional annotation of the genes by the PANNZER webserver revealed that this isolate might harbor multiple genes associated with cellulase-producing enzymes. The isolate U41 contains 28 genes involved in cellulase production (Table 2).

Table 2
Genes involved in cellulase production.

Enzyme	Subtype	No. of gene
Cellulase	Endoglucanase	2
	Exoglucanase	1
	Cellulase	4
	beta-glucosidases	18
	Alpha- glucosidases	1
	oligo-1,6-glucosidases	1
	beta-galactosidases	1
Hemicellulase	Xylanase	8
	Arabinose	6
	Mannose	2
Pectinase	Pectinase	2
	Pectinesterase	3
Ligninase	Laccase, Polyphenol oxidase	2

4. Experimental Design, Materials and Methods

Bacillus safensis strain U41 was isolated from soils in Santiniketan, West Bengal, India (23.67760 N, 87.68530 E). After the standard dilution method, the strain was obtained by culture on Bushnell Haas agar medium supplemented with 1 % (w/v) CMC (carboxymethylcellulose) medium (BH−CMC). The resulting single colony was sub-cultured by repeated streaking onto BH−CMC agar plates followed by incubation for 48 h at 37 °C [5].

4.1. DNA extraction from bacterial culture

Genomic DNA extraction was performed using the QIAgen Blood and Tissue kit (QIAgen Inc., Canada) (Qiagen, Germany) per the manufacturer's guidelines. Quality ($OD_{260/280}$ ratio) and concentration of the isolated DNA were measured using a microplate reader (Agilent BioTek Epoch 2, USA).

4.2. Whole genome sequencing and assembly

The whole genome of strain U41 was sequenced using Illumina HiSeq X (2 × 150 paired-end) sequencing technology and generated 7352,576 raw reads. First, raw sequences were visualized in the updated version of FastQC, version 0.11.9 [6], followed by trimming and size selection (>200 bp) in Cutadapt v2.9 [7]. After trimming, 7347,811 paired clean reads were assembled with Unicycler v0.4.4, a hybrid bacterial genomes assembly pipeline with the default setting. The assembly pipeline was optimized and involved an error correction of sequenced reads with BayesHammer [8] and assembly with SPAdes, v3.13.0 [9], with a k-mer value of up to 99. Quality and summary statistics of the assembled genome were generated using the CheckM v1.1.6 and QUAST v.5.1.0. The assembled genome was visualized as a circular map using the CGView server (https://www.cgview.ca) [10] (Fig. 1). Annotation and gene prediction from the assembled genome was performed using prokka v. 1.14.6 (rapid prokaryotic genome annotation).

4.3. Phylogenetic tree construction

The 16S rRNA gene sequence was extracted from the whole genome and similarity search using the latest EzBioCloud server (www.ezbiocloud.net) with validated type strains [11]. All available almost complete 16S rRNA gene sequences from closely related type taxa were obtained from the NCBI database (www.ncbi.nlm.nih.gov) and aligned using the MUSCLE algorithm in the MEGA X program [12]. Gaps and missing data were fixed with the complete deletion option, and nucleotide substitution was done using the Tamura 3-parameter with the Gamma distribution evolutionarily invariable model (T92+G+I) [13]. The maximum-likelihood trees were reconstructed using bootstrap values based on 1000 replications in Mega version X (Fig. 2). Whole-genome-based taxonomic analysis, namely genome-to-genome distances (GGDs) and digital DNA-DNA hybridization (dDDH) were performed using the Type Strain Genome Server (TYGS) (https://www.tygs.dsmz.De) [14] (Fig. 3). The whole-genome-based phylogenetic tree was constructed using FastME [15] from the genome blast distance phylogeny (GBDP). The trees were rooted at the midpoint [16].

Limitations

Not applicable.

Ethics Statement

The authors have read and follow the ethical requirements for publication in Data in Brief and confirming that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Data Availability

Microbe sample from Bacillus sp. U41 (Original data) (NCBI).

CRediT Author Statement

Binoy Kumar Show: Conceptualization, Data curation, Formal analysis, Methodology, Validation, Writing – original draft; **Andrew B. Ross:** Conceptualization, Investigation, Project administration, Supervision, Writing – review & editing; **Raju Biswas:** Data curation, Formal analysis, Resources, Writing – original draft; **Shibani Chaudhury:** Conceptualization, Investigation, Project administration, Supervision, Writing – review & editing; **Srinivasan Balachandran:** Conceptualization, Investigation, Project administration, Supervision, Writing – review & editing.

Acknowledgements

Authors sincerely acknowledge BBSRC, United Kingdom, for financial support through the research grant BB/S011439/1. We also thankfully acknowledge the microbiology laboratory of Dr. Bomba Dam, Assistant Professor, Department of Botany, Institute of Science, Visva-Bharati (A Central University), Shantiniketan, India, for providing the infrastructure for bioinformatics analysis.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary Materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2024.110547.

References

[1] A. Lateef, I.A. Adelere, E.B. Gueguim-Kana, The biology and potential biotechnological applications of *Bacillus safensis*, Biologia 70 (2015) 411–419, doi:10.1515/biolog-2015-0062.

[2] J.K. Saini, R. Saini, L. Tewari, Lignocellulosic agriculture wastes as biomass feedstocks for second-generation bioethanol production: concepts and recent developments, 3 Biotech. 5 (2015) 337–353, doi:10.1007/s13205-014-0246-5.

[3] K. Karimi, M.J. Taherzadeh, A critical review of analytical methods in pretreatment of lignocelluloses: composition, imaging, and crystallinity, Biores. Technol. 200 (2016) 1008–1018, doi:10.1016/j.biortech.2015.11.022.

[4] B.K. Show, S. Banerjee, A. Banerjee, R. GhoshThakur, A.K. Hazra, N.C. Mandal, A.B. Ross, S. Balachandran, S. Chaudhury, Insect gut bacteria: a promising tool for enhanced biogas production, Rev. Environ. Sci. Biotechnol. 21 (1) (2022) 1–25, doi:10.1007/s11157-021-09607-8.

[5] D. Sinha, S. Banerjee, S. Mandal, A. Basu, A. Banerjee, S. Balachandran, N.C. Mandal, S. Chaudhury, Enhanced biogas production from Lantana camara via bioaugmentation of cellulolytic bacteria, Biores. Technol 340 (2021) 125652, doi:10.1016/j.biortech.2021.125652.

[6] S. Andrews, FastQC: A Quality Control Tool for High Throughput Sequence Data [Online], 2010 Available online at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[7] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet J. 17 (1) (2011) 10–12, doi:10.14806/ej.17.1.200.

[8] S.I. Nikolenko, A.I. Korobeynikov, M.A. Alekseyev, BayesHammer: bayesian clustering for error correction in single-cell sequencing, BMC Genomics 14 (1) (2013) 1–11, doi:10.1186/1471-2164-14-S1-S7.

[9] J.R. Grant, P. Stothard, The CGView Server: a comparative genomics tool for circular genomes, Nucleic Acids Res. 36 (suppl_2) (2008) W181–W184, doi:10.1093/nar/gkn179.

[10] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A.V. Pyshkin, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (5) (2012) 455–477, doi:10.1089/cmb.2012.0021.

[11] S.H. Yoon, S.M. Ha, S. Kwon, J. Lim, Y. Kim, H. Seo, J. Chun, Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies, Int. J. Syst. Evol. 67 (5) (2017) 1613, doi:10.1099/ijsem.0.001755.

[12] S. Kumar, M. Nei, J. Dudley, K. Tamura, MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences, Brief. Bioinform. 9 (4) (2008) 299–306, doi:10.1093/bib/bbn017.

[13] M. Hasegawa, H. Kishino, T.A. Yano, Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, J. Mol. Evol. 22 (1985) 160–174, doi:10.1007/BF02101694.

[14] J.P. Meier-Kolthoff, M. Göker, TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy, Nat. Commun. 10 (1) (2019) 2182, doi:10.1038/s41467-019-10210-3.

[15] A. Eriksson, A. Manica, The doubly conditioned frequency spectrum does not distinguish between ancient population structure and hybridization, Mol. Biol. Evol. 31 (6) (2014) 1618–1621, doi:10.1093/molbev/msu103.

[16] J.S. Farris, Estimating phylogenetic trees from distance matrices, Am. Nat. 106 (951) (1972) 645–668 https://www.jstor.org/stable/2459725.