Data in Brief 37 (2021) 107240

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Whole genome sequence data of *Bacillus australimaris* strain B28A, isolated from Marine Water in India

Wael Ali Mohammed Hadi[a,*], Boby T. Edwin[b], Ananthakrishnan Jayakumaran Nair[a]

[a] *Department of Biotechnology, University of Kerala, Thiruvananthapuram, Kerala, India*
[b] *CEPCI Laboratory & Research Institute, Kollam, Kerala, India*

ARTICLE INFO

Article history:
Received 27 April 2021
Revised 10 June 2021
Accepted 18 June 2021
Available online 21 June 2021

Keywords:
WGS
Genome
Data
Bacillus Marine-bacteria
Enzymes

ABSTRACT

Bacillus genus members are dominant in the Eastern Arabian Sea and are known for producing many industrial enzymes. Bacillus australimaris B28A, isolated from seawater, had an enzymatic activity. Here, the whole genome sequence of Bacillus australimaris B28A is reported. The 3,766,107-bp genome, with a GC content of 41.6%, comprised 3936 protein-coding genes, seven ribosomal RNA, and 75 transfer RNA. Several bioactive secondary metabolite genes in the genome, including surfactin, lichenysin, bacillibactin, bacilysin, paenilamicin, fengycin, and carotenoid, were identified using antiSMASH. The 1396 proteins were predicted using RAST, including asparaginase enzyme: an anticancer enzyme. Sequence data have been deposited in the DDBJ/ENA/GenBank database under the accession number JAGQFH000000000. The version described in this paper is JAGQFH000000000.1. The BioProject ID in the GenBank database is PRJNA670955. The raw data is publicly available at "https://www.ncbi.nlm.nih.gov/sra/SRR14203888".

* Corresponding author.
E-mail address: hwael.bty@keralauniversity.ac.in (W.A.M. Hadi).

https://doi.org/10.1016/j.dib.2021.107240

Specifications Table

Subject	Microbiology
Specific subject area	Marine Biology and genomics
Type of data	Genome sequence data in FASTA and FASTQ format
How data were acquired	Genome sequencing with Illumina HiSeq 4000 at MACROGEN Inc. South Korea
Data format	Raw, Filtered, and Analysed genome sequences.
Parameters for data collection	Bacterial genomic DNA was extracted from a pure culture of Bacillus australimaris B28A
Description of data collection	Genome features (Table 1), genome map (Fig. 1a), Whole genome sequencing, raw, assembly, and annotation
Data source location	*Bacillus australimaris* B28 was obtained from marine water in the city of Thiruvananthapuram, Kerala, India. (GPS: 8.554185N 76.849609E). The isolated sample 16srRNA accession number is MT010836. The BioSample ID is SAMN18571126
Data accessibility	This whole-genome sequencing data have been deposited at DDBJ/EMBL/GenBank under the accession number JAGQFH000000000 (https://www.ncbi.nlm.nih.gov/nuccore/JAGQFH000000000). The sequence data have been deposited in the NCBI SRA database under the accession number SRR14203888. (https://www.ncbi.nlm.nih.gov/sra/SRR14203888) The BioProject ID in GenBank is : PRJNA670955

Value of the Data

- The complete genome sequence of *B. australimaris* B28A provides the essential information and insight for industrial biotechnology applications in enzymes.
- The genome of *B. australimaris* B28A and data are valuable for comparative genomic studies and protein biosynthesis applications.
- *B. australimaris* B28A genome data accelerate the knowledge for all microbial research communities and biotechnology applications.

1. Data Description

Marine bacteria are diverse and potential as a resource for industrial enzyme applications [1]. Firmicutes are one of the most dominant groups in the Eastern Arabian Sea [2]. *Bacillus australimaris* B28A was previously isolated from seawater of costal of Thiruvananthapuram, in with deposit 16SrRNA accession number is MT010836. The strain had an industrial enzymatic activity, with a potential l-asparaginase activity [3]. The B. australimaris B28 genome raw data was comprised a total number of bases sequenced of 1,452,683,420 bps, with total reads in Illumina paired-end sequencing of 9,620,420 and GC content of 42.79%. The raw data ratio of bases with a Phred quality score of over 20 (Q20) is 89.40%, and over 30 (Q30) is 80.01%, respectively. The filtered data was resulted in a total read base of 695,777,974 bps, with total reads 5,130,218, and GC content of 41.96%: the quality with Q20 is 98.40%, and Q30 94.92%. The B. australimaris B28 annotated genome was sequenced to 3,766,107 bps and comprised 58 contigs and a G+C content of 41.60% (Fig. 1A, Table 1). There were 3936 coding sequences (CDS), seven ribosomal RNA, and 75 transfer RNA. Numerous secondary metabolite genes in the genome include surfactin, lichenysin, bacillibactin, bacilysin, paenilamicin, fengycin, and carotenoid predicted by antiSMASH [4,5]. The annotated gene using RAST [6–8] categorized identified proteins in twenty-seven main clusters and 318 subsystems of 1396 genes, including asparaginases enzyme. L-asparaginase is recognized as potential anticancer medicine [3]. Subsystems annotation included in (Supplementary 1).

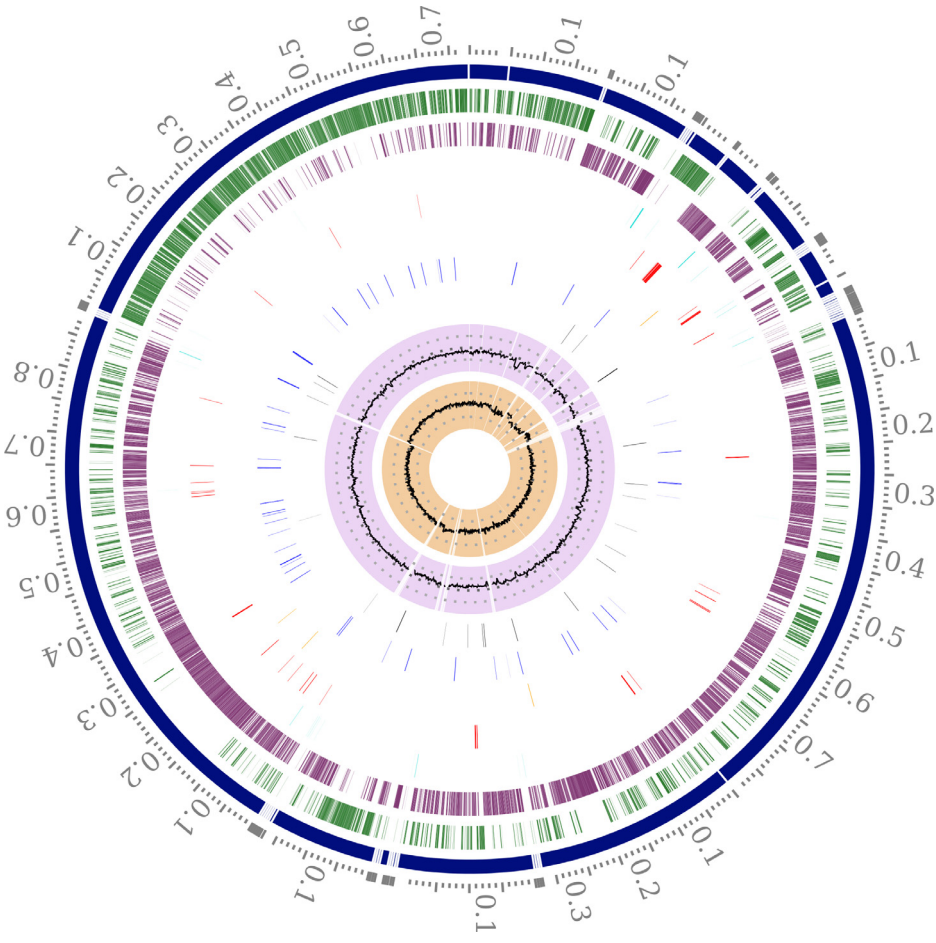

Fig. 1. The Circular genome map and Cluster analysis of B. australimaris B28A visualized by the Patric genome analysis server [6,9]. The outer grey layer in the genome map indicates the position label in megabase pair (Mbp) followed by contigs' length, representing a blue color. The next discontinuous layer refers coding region of genes (CDs) as green for the forward strand and purple for the reverse strand. Aqua color represents Non-CDs Features, and red color refers to antimicrobial resistance genes (AMR), respectively. The next layer is orange, which refers to virulence factors (VF) genes, followed by the blue-colored layer referred to as transporters. The internal last three layers represent drug targets, GC content, and GC skew; black, black with a light pink shade, and black with peach shade color.

Table 1
Genome feature of *Bacillus australimaris* B28.

Genome feature	Value
Genome size	3,766,107
G+C content	41.6
Number of rRNA genes	7
Number of tRNA genes	75
Number of ORFs	3936
Number of contigs	58
N50	339274
L50	3

2. Experimental Design, Materials and Methods

B. australimaris B28A was previously isolated and identified with 16SrRNA with GenBank accession number MT010836. Genomic DNA was extracted using the Monarch® Genomic DNA Purification Kit (New England Biolabs.). Illumina HiSeq 4000 paired-end (2 × 151 bp) sequencing of B. australimaris B28A was performed by Macrogen Inc. (Seoul, Korea). The library was processed using the Nextera XT DNA Library Preparation Kit (96 samples) (Illumina, Inc., San Diego, CA, USA). Total sequencing reads 5,130,218 of 4,447,316 were mapped. After mapping, Sambamba [10] and SAMTools [11] were respectively used to remove duplicated reads and identify variants. The reads were assembled into 58 contigs, a GC content of 41.60% using GCE Assembler (version 1.2; https://cge.cbs.dtu.dk/services/Assembler/) [12]. The assembled data was annotated using RAST rapid annotation using subsystem technology version 2.0 [6–8]. Bacterial secondary metabolite biosynthesis gene clusters were identified and annotated by antiSMASH version 5.0 and 6.0 using assembled fasta file output [4,5].

Ethics Statement

N/A.

CRediT Author Statement

Wael Ali Mohammed Hadi: Investigation, Writing, original draft, Formal analysis, Methodology, Visualization, Project administration; **Boby T. Edwin:** Supervision; **A. Jayakumaran Nair:** Writing - reviewing & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Acknowledgments

The first author acknowledges the Indian Council for Cultural Relations (ICCR) for the scholarship.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107240.

References

[1] T.H. Cheng, N. Ismail, N. Kamaruding, J. Saidin, M. Danish-Daniel, Industrial enzymes-producing marine bacteria from marine resources, Biotechnol. Rep. 27 (2020) e00482, doi:10.1016/j.btre.2020.e00482.
[2] V. Sachithanandam, N. Saravanane, K. Chandrasekar, P. Karthick, P. Lalitha, S. Sai Elangovan, M. Sudhakar, Microbial diversity from the continental shelf regions of the Eastern Arabian Sea: a metagenomic approach, Saudi J. Biol. Sci. 27 (2020) 2065–2075, doi:10.1016/j.sjbs.2020.06.011.
[3] N. Chakravarty, Priyanka, J. Singh, R.P. Singh, A potential type-II L-asparaginase from marine isolate Bacillus australimaris NJB19: statistical optimization, in silico analysis and structural modeling, Int. J. Biol. Macromol. 174 (2021) 527–539, doi:10.1016/j.ijbiomac.2021.01.130.

[4] K. Blin, S. Shaw, K. Steinke, R. Villebro, N. Ziemert, S.Y. Lee, M.H. Medema, T. Weber, AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline, Nucleic Acids Res. 47 (2019) W81–W87, doi:10.1093/nar/gkz310.
[5] K. Blin, S. Shaw, A.M. Kloosterman, Z. Charlop-Powers, G.P. van Wezel, M.H. Medema, T. Weber, antiSMASH 6.0: improving cluster detection and comparison capabilities, Nucleic Acids Res. (2021), doi:10.1093/nar/gkab335.
[6] T. Brettin, J.J. Davis, T. Disz, R.A. Edwards, S. Gerdes, G.J. Olsen, R. Olson, R. Overbeek, B. Parrello, G.D. Pusch, M. Shukla, J.A. Thomason, R. Stevens, V. Vonstein, A.R. Wattam, F. Xia, RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes, Sci. Rep. 5 (2015) 8365, doi:10.1038/srep08365.
[7] R. Overbeek, R. Olson, G.D. Pusch, G.J. Olsen, J.J. Davis, T. Disz, R.A. Edwards, S. Gerdes, B. Parrello, M. Shukla, V. Vonstein, A.R. Wattam, F. Xia, R. Stevens, The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST), Nucleic Acids Res. 42 (2014) D206–D214, doi:10.1093/nar/gkt1226.
[8] R.K. Aziz, D. Bartels, A.A. Best, M. DeJongh, T. Disz, R.A. Edwards, K. Formsma, S. Gerdes, E.M. Glass, M. Kubal, F. Meyer, G.J. Olsen, R. Olson, A.L. Osterman, R.A. Overbeek, L.K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G.D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, O. Zagnitko, The RAST server: rapid annotations using subsystems technology, BMC Genom. 9 (2008) 75, doi:10.1186/1471-2164-9-75.
[9] J.J. Davis, A.R. Wattam, R.K. Aziz, T. Brettin, R. Butler, R.M. Butler, P. Chlenski, N. Conrad, A. Dickerman, E.M. Dietrich, J.L. Gabbard, S. Gerdes, A. Guard, R.W. Kenyon, D. Machi, C. Mao, D. Murphy-Olson, M. Nguyen, E.K. Nordberg, G.J. Olsen, R.D. Olson, J.C. Overbeek, R. Overbeek, B. Parrello, G.D. Pusch, M. Shukla, C. Thomas, M. VanOeffelen, V. Vonstein, A.S. Warren, F. Xia, D. Xie, H. Yoo, R. Stevens, The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities, Nucleic Acids Res. 48 (2019) D606–D612, doi:10.1093/nar/gkz943.
[10] A. Tarasov, A.J. Vilella, E. Cuppen, I.J. Nijman, P. Prins, Sambamba: fast processing of NGS alignment formats, Bioinformatics 31 (2015) 2032–2034, doi:10.1093/bioinformatics/btv098.
[11] P. Danecek, J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M.O. Pollard, A. Whitwham, T. Keane, S.A. McCarthy, R.M. Davies, H. Li, Twelve years of SAMtools and BCFtools, Gigascience 10 (2021) 1–4, doi:10.1093/gigascience/giab008.
[12] M.V. Larsen, S. Cosentino, S. Rasmussen, C. Friis, H. Hasman, R.L. Marvig, L. Jelsbak, T. Sicheritz-Ponten, D.W. Ussery, F.M. Aarestrup, O. Lund, Multilocus sequence typing of total-genome-sequenced bacteria, J. Clin. Microbiol. 50 (2012) 1355–1361, doi:10.1128/JCM.06094-11.