Data in Brief 66 (2026) 112684

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Complete genome dataset of *Flavobacterium* sp. strain PL002 isolated from Antarctic *Porphyra* algae

Jennifer Charles Labo[a], Paris Lavin[b], Hui Yin Fan[c], Mohd Faizal Abu Bakar[d], Nur Athirah Yusof[a,*]

[a] *Biotechnology Research Institute, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia*
[b] *Departamento de Biotecnologia, Facultad de Ciencias del Mar y Recursos Biologicos, Universidad de Antofagasta, 601 Avenida Angamos, Antofagasta 1270300, Chile*
[c] *Food Security Research Laboratory, Faculty of Food Science and Nutrition, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia*
[d] *Advanced Genomics and Bioinformatics, Malaysia Genome and Vaccine Institute, National Institutes of Biotechnology Malaysia, Jalan Bangi, 43000 Kajang, Selangor, Malaysia*

ARTICLE INFO

Article history:
Received 16 December 2025
Revised 10 March 2026
Accepted 11 March 2026
Available online 16 March 2026

Keywords:
Long-read technology
Circular assembly
Psychrophilic adaptation
Algal symbiosis
Marine bacteriology
Chromosome closure
Comparative genomics

ABSTRACT

Flavobacterium sp. strain PL002 is a psychrotolerant marine bacterium isolated from Antarctic *Porphyra* algae. This article reports the complete genome dataset of strain PL002 generated using long-read sequencing technology to improve a previously published Illumina-based draft genome, which consisted of 170 contigs. High-molecular-weight genomic DNA was extracted from pure culture and subjected to single-molecule real-time sequencing. The resulting long-read sequences were assembled de novo to produce a single circular chromosome of 4475,065 bp with a GC content of 33 %. Genome closure recovered an additional 175,100 bp relative to the draft assembly and resolved repetitive regions, including complete ribosomal RNA (rRNA) operons. Structural annotation and gene prediction identified a total of 3916 genes, including 3795 protein-coding sequences, 24 rRNA genes, and 65 transfer RNA (tRNA) genes. The complete genome sequence and raw long-read data are publicly available in the NCBI database under BioProject PRJNA1337565 and SRA accession SRR35731852. The dataset is suitable for

* Corresponding author.
E-mail address: nrathirah.yusof@ums.edu.my (N.A. Yusof).

https://doi.org/10.1016/j.dib.2026.112684

reuse in comparative genomics, phylogenomic analysis, functional annotation, and studies of microbial adaptation in cold marine environments.

Specifications Table

Subject	Biology
Specific subject area	Microbial genomics and long-read genome sequencing.
Type of data	Table Raw, Assembled, Annotated
Data collection	Pure culture of *Flavobacterium* sp. strain PL002 was grown in Reasonar's 2A (R2A) broth at its optimal growth temperature of 15 °C until mid-log phase. Genomic DNA was extracted using the DNeasy® Blood and Tissue Kit (Qiagen, Hilden, Germany). Long-read sequencing was performed using the PacBio Sequel platform (Single Molecule Real-Time sequencing). Raw sequencing data were quality assessed using PacBio SMRT Link tools and assembled de novo using Flye v.29.6. Genome annotation and functional assignment were performed using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) v.6.10
Data source location	Antarctic *Porphyra* algae
Data accessibility	*Please note: All raw data associated with this study have been deposited in a publicly accessible repository prior to submission* Repository name: National Center for Biotechnology Information (NCBI) Data identification numbers: BioProject PRJNA1337565; Complete genome assembly (RefSeq) GCF_054165825.1; Draft genome assembly (RefSeq, previously published): GCF_014858825.1; Sequence Read Archive (SRA): SRR35731852; GenBank: JBSEEC000000000.1 Direct URLs to data: BioProject record: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1337565/ Complete genome assembly: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_054165825.1/ Draft genome assembly: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_014858825.1/ Sequence Read Archive (SRA) : https://www.ncbi.nlm.nih.gov/sra/?term=SRR35731852 GenBank : https://www.ncbi.nlm.nih.gov/nuccore/JBSEEC000000000.1/ The repository provides public and unrestricted access to all data associated with this study.
Related research article	C.P. Teoh, P. Lavin, N. Najimudin, P.C. Lee, L. Iancu, C. Purcarea, C.M.V. Wong, Draft genome sequence of *Flavobacterium* sp. strain PL002, isolated from Antarctic Porphyra algae, Microbiol. Resour. Announc. 10 (2021) e00063–21. https://doi.org/10.1128/mra.00063-21.

1. Value of the Data

- This dataset represents the first complete and gapless genome sequence of *Flavobacterium* sp. strain PL002, as the previously available genome from this isolate and other related *Flavobacterium* species deposited in NCBI GenBank are largely represented by draft and fragmented assemblies.
- The complete whole-genome sequence of *Flavobacterium* sp. strain PL002 provides valuable insights into molecular mechanisms involved in cold adaptation, particularly in psychrotolerant marine bacteria isolated from Antarctic environments.

- This genome enables comparative genomic analyses between *Flavobacterium* strains and related species, contributing to a better understanding of genome evolution, adaptation strategies, and functional diversity within the genus.
- The availability of a complete genome facilitates the identification of metabolic pathways, cold-active enzymes, and functional genes with potential biotechnological and industrial applications.
- The dataset serves as a useful genomic resource for researchers and students in the fields of microbiology, genomics, marine biology, and biotechnology, particularly those studying microorganisms adapted to extreme cold environments.

2. Background

This dataset was generated as a follow-up to the previously published draft genome sequence of *Flavobacterium* sp. strain PL002 reported by Teoh et al. (2021), which was based on Illumina short-read sequencing and assembled into multiple contigs [1]. Strain PL002 is a psychrotolerant marine bacterium isolated from Antarctic *Porphyra* algae and forms part of the cold-adapted microbial communities associated with polar macroalgae. Previous studies have reported that strain PL002 exhibits several cold-adaptive and biotechnologically relevant characteristics, including the production of cold-active enzymes, extracellular polysaccharides, and antifreeze-related activity that support survival under subzero and freeze-thaw conditions [2]. These functional traits suggest the presence of specialized genetic determinants involved in cold stress tolerance, membrane adaptation, and extracellular matrix production. However, short-read sequencing is limited in resolving bacterial genomic features such as repetitive region and multiple ribosomal RNA (rRNA) operons, which contributed to the fragmentation of the Illumina-based draft genome of strain PL002. Such limitations restrict comprehensive analysis of gene clusters potentially associated with cold adaptation, polysaccharide biosynthesis, secretion systems, and stress-response pathways that underpin the previously reported phenotypic characteristics of this strain. To address these limitations, long-read sequencing technology was employed to generate continuous reads spanning repetitive and structurally complex regions. High-molecular-weight genomic DNA was extracted from pure culture and used to construct long-read sequencing libraries, followed by de novo assembly and genome annotation. This data article extends the earlier report by providing a complete, circular genome assembly together with the associated raw sequencing reads and annotation files in public repositories, thereby supplying an improved genomic resource for studies of cold-adapted marine *Flavobacterium* species. The dataset supplies the underlying genomic resources for strain PL002 and documents the methodological update from short-read to long-read sequencing.

3. Data Description

This article describes the complete genome dataset of *Flavobacterium* sp. strain PL002 generated using long-read sequencing technology. The dataset is publicly available in the NCBI repository under BioProject PRJNA1337565 and includes the assembled genome, raw sequencing reads, and genome annotation files. The bacterial strain was isolated from Antarctic *Porphyra* algae, and a previously published genome based on Illumina short-read sequencing consisted of 170 contigs [1]. The finalized genome assembly comprises a single circular chromosome of 4475,065 bp with a GC content of 33 %. The complete genome sequence is available in GenBank under accession JBSEEC000000000.1. This improvement enables accurate reconstruction of bacterial genome structure, including complete rRNA operons and repetitive regions facilitating studies of rRNA operon organization and improving taxonomic resolution in microbial community analyses [3,4]. Genome annotation was performed using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP)., generating predicted protein-coding genes, ribosomal RNA genes, and transfer RNA genes provided as downloadable annotation files . A comparison between the former draft genome and the current complete genome assembly is provided in Table 1.

Table 1
Comparison of genome assembly and annotation statistics between the draft Illumina short-read assembly and the complete PacBio long-read assembly of *Flavobacterium* sp. strain PL002. Assembly quality metrics include contiguity statistics (number of contigs, N50, and L50) and genome completeness assessed using BUSCO with the flavobacteriia_odb10 lineage dataset. Genome annotation was performed using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP).

Feature	Draft genome (Illumina short-read)	Complete genome (PacBio long-read)
Assembly statistics		
Assembly accession	GCF_014858825.1	GCF_054165825.1
Assembly status	Draft	Complete
Genome size (bp)	4299,965 bp	4475,065 bp
GC content (%)	33	33
Number of contigs	180	1
Contig N50	86.4 kb	4.5 Mb
Contig L50	19	1
BUSCO Completeness	100	100
Annotation statistics		
Total genes	3803	3916
Protein-coding genes	3701	3795
CDSs	3745	3824
Pseudogenes	44	29
rRNA genes (total)	4	24
tRNA genes	51	65
ncRNA genes	3	3
CRISPR arrays	1	1

In the previously published draft genome report of *Flavobacterium* sp. strain PL002, analyses were limited by an Illumina-based fragmented assembly and gene-based phylogenetic inference. In this study, the availability of a complete, closed genome enabled a genome-scale phylogenomic analysis incorporating closely related *Flavobacterium* type strain genomes. The phylogenomic tree (Fig. 1), reconstructed using conserved single-copy core genes, placed strain PL002 as a well-supported and distinct lineage within the *Flavobacterium* clade, clustering closest to *Flavobacterium* ovolatum KCTC 52,693 and *Flavobacterium* faecale KCTC 32,457. Consistent with this placement, ANI analysis using the NCBI ANI pipeline showed that the highest genome-wide similarity was to *Flavobacterium* faecale (RefSeq accession GCA_003076455.1) with an ANIb value of 86.18 %, below the accepted species threshold of ~95–96 %. These analyses provide an improved evolutionary and comparative genomic framework for strain PL002 relative to the earlier draft genome.

4. Experimental Design, Materials and Methods

Flavobacterium sp. strain PL002 was maintained as a pure culture and grown in Reasonar's 2A broth at 15 °C with shaking at 180 rpm. Cells were harvested during the exponential growth phase by centrifugation at 8000 × *g* for 10 min at 4 °C. Genomic DNA was extracted using the DNeasy® Blood and Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions for Gram-negative bacteria with minor modifications to improve DNA yield and integrity [4,5]. DNA concentration and purity were measured using a NanoDrop spectrophotometer (Thermo Fisher Scientific) and a Qubit fluorometer (Invitrogen). DNA integrity and fragment size were assessed by agarose gel electrophoresis, and only high-molecular-weight DNA without visible degradation and with A260/280 ratios between 1.8 and 2.0 was used for library preparation.

Long-read sequencing was performed using the PacBio Sequel platform (Pacific Biosciences, Menlo Park, CA, USA). Genomic DNA was sheared to an average fragment size of approximately 20 kb and converted into SMRTbell templates using the SMRTbell Template Prep Kit (Pacific Biosciences). DNA polymerase binding was carried out using the DNA/Polymerase Binding Kit P6

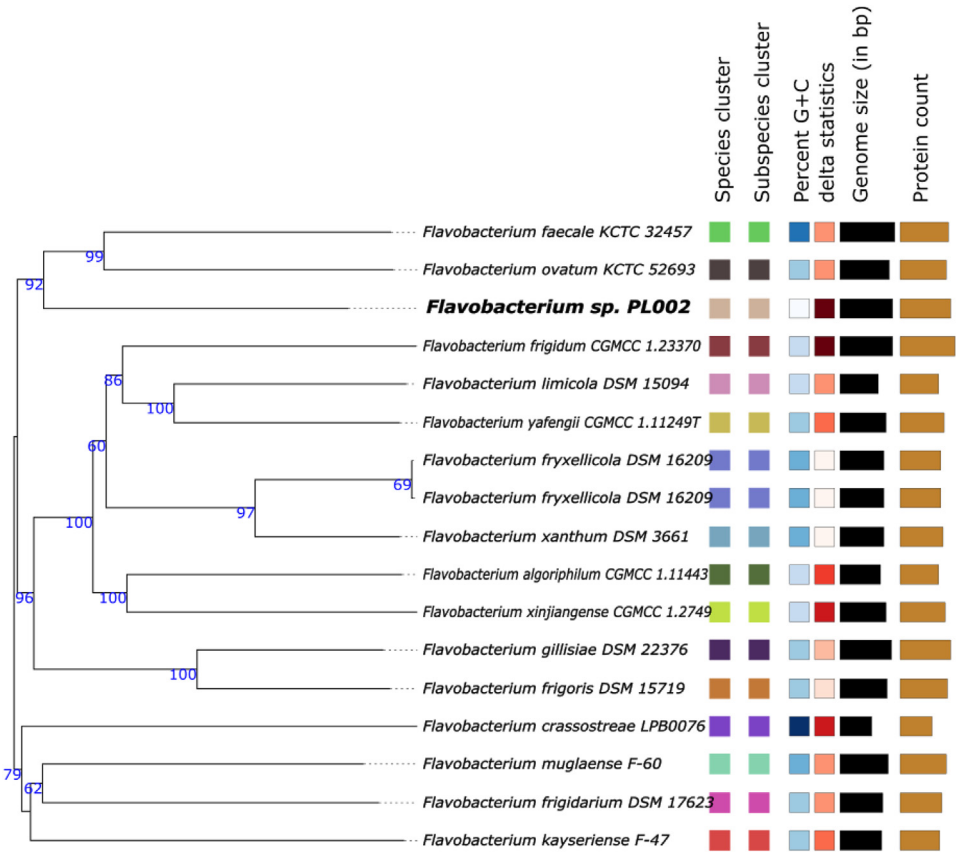

Fig. 1. Phylogenomic position of *Flavobacterium* sp. strain PL002 within the genus *Flavobacterium*. The phylogenomic tree was generated using the Type Strain Genome Server (TYGS) based on the Genome BLAST Distance Phylogeny (GBDP) approach. Whole-genome sequences of *Flavobacterium* sp. PL002 and representative *Flavobacterium* type strains were compared to calculate pairwise intergenomic distances using GBDP. The resulting distance matrix was used to infer the phylogenomic tree using the FastME algorithm implemented in TYGS. Branch support values shown at the nodes represent pseudo-bootstrap support percentages calculated from 100 replicates. *Flavobacterium* sp. PL002 is highlighted in bold to indicate its phylogenomic placement among closely related taxa. Additional genome features provided by TYGS are displayed alongside the tree, including species cluster, subspecies cluster, GC content (%), delta statistics, genome size (bp), and predicted protein count. The GC % column represents the genomic GC content of each strain and is visualized using a color gradient, where lighter shades indicate relatively lower GC content and darker shades indicate relatively higher GC content among the genomes included in the analysis.

v2 according to the manufacturer's protocol. Sequencing was conducted using Single Molecule Real-Time (SMRT) technology, and raw data were generated in BAM format.

Raw long-read sequencing data were subjected to quality filtering prior to genome assembly to remove low-quality and short reads. Subreads with lengths below 1000 bp and reads failing platform-specific quality thresholds were excluded. Quality assessment was performed using standard PacBio SMRT Link summary statistics, and only high-quality reads were retained for downstream de novo assembly. This filtering step ensured reliable genome reconstruction while minimizing sequencing artefacts.

The de novo genome assembly was conducted using Flye v. 2.9.6, a long-read assembler designed for PacBio SMRT sequencing data [6,7]. The assembly was performed using the –pacbio-raw option with an estimated genome size of approximately 4 Mb, while all other parameters were kept at default settings (options: flye –pacbio-raw reads.fastq –genome-size 4 m). The as-

sembly process generated a single, circular chromosome, confirmed by the overlapping of its terminal ends. The final assembly has a coverage depth of 55.0x based on PacBio Sequel sequencing data. All assembly statistics, including genome size and GC content, were derived from this final assembly output.

Genome annotation was performed using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) version 6.1, employing the latest NCBI reference databases and metadata standards. The complete genome assembly of *Flavobacterium* sp. strain PL002 was submitted to NCBI and annotated for protein-coding genes, ribosomal RNA genes, and transfer RNA genes using PGAP v 6.10. To ensure comparability of gene statistics between the draft and complete genome assemblies of *Flavobacterium* sp. strain PL002, annotation data were obtained directly from the NCBI annotation records associated with each publicly available genome accession. Gene counts reported in Table 1 were extracted from the corresponding NCBI RefSeq annotation files, which were generated using standardized NCBI Prokaryotic Genome Annotation Pipeline procedures and curated metadata. This approach ensures that genome statistics are based on accession-linked, publicly available annotations and avoids inconsistencies arising from the use of heterogeneous annotation pipelines.

Genome-based phylogenomic analysis was performed using the Type (Strain) Genome Server (TYGS) [8]. The complete genome sequence of strain PL002 was uploaded to TYGS and compared against reference genomes, including all closely related *Flavobacterium* type strains available in the database. Intergenomic distances were calculated using the Genome BLAST Distance Phylogeny (GBDP) method [9]. which computes distances based on high-scoring segment pairs obtained from whole-genome BLAST comparisons. Phylogenomic tree reconstruction was conducted using the FastME algorithm and branch support was assessed by bootstrap analysis [10]. Default TYGS parameters were applied throughout the analysis. The resulting phylogenomic tree was visualized and annotated to display bootstrap support values and genome-based metadata for comparative interpretation.

Average nucleotide identity (ANI) was calculated using the NCBI ANI analysis pipeline [11]. The complete genome sequence of *Flavobacterium* sp. strain PL002 was compared against publicly available *Flavobacterium* type strain genomes. ANI values were computed using the BLAST-based ANI method (ANIb), and the highest ANI value obtained against the closest related type strain genome was reported [12,13]

Limitations

The dataset derives from genomic DNA obtained from a single laboratory culture of *Flavobacterium* sp. strain PL002; consequently, it represents one isolate under one growth condition. Long-read sequencing was conducted from a single DNA extraction and a single SMRTbell library without technical replication. Data were generated on the PacBio Sequel platform in HiFi (CCS) mode; while the circular-consensus reads are high-accuracy, residual platform-specific biases may persist, and uncorrected subreads, where present, retain characteristic error profiles. No complementary short-read sequencing was produced for hybrid correction in this study. Environmental metadata associated with the original sampling (e.g., in situ physicochemical parameters) were not recorded. Functional annotation reflects the specific database releases and pipeline versions available at the time of processing and may change upon re-annotation with alternative tools or updated resources. For investigators requiring short-read coverage, the short-read sequences from the initial study can be retrieved from the NCBI Sequence Read Archive, enabling optional hybrid analyses.

Ethics Statement

This work did not involve any animals or human subjects. The manuscript represents the author's original work, which has not been published elsewhere.

CRediT Author Statement

Fan Hui Yin and Paris Leonardo Lavin: Conceptualization, Methodology; **Jennifer Charles Labo and Paris Leonardo Lavin:** Data curation; **Jennifer Charles Labo:** Writing – Original draft preparation; **Fan Hui Yin:** Funding acquisition, Resources; **Nur Athirah Yusof and Fan Hui Yin:** Supervision; **Jennifer Charles Labo and Nur Athirah Yusof:** Writing - Reviewing and Editing; **Paris Leonardo Lavin and Mohd Faizal Abu Bakar:** Investigation, Software, Validation.

Data Availability

PACBIO sequencing of antarctic Flavobacterium (Original data) (NCBI)

Acknowledgements

The funding support from the Yayasan Penyelidikan Antartika Sultan Mizan (YPASM), grant number 'LPS2309', is gratefully acknowledged. This work was also partially supported by Instituto Antartico Chileno (INACH), grant N° RT_20–19 and the MINEDUC-UA project, code ANT 22991.

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.

References

[1] C.P. Teoh, P. Lavin, N. Najimudin, P.C. Lee, L. Iancu, C. Purcarea, C.M.V. Wong, Draft genome sequence of *Flavobacterium* sp. strain PL002, isolated from Antarctic *Porphyra* algae, Microbiol. Resour. Announc. 10 (2021) e00063-21, doi:10.1128/mra.00063-21.
[2] G. Necula-Petrareanu, P. Lavin, V.I. Paun, G.R. Gheorghita, A. Vasilescu, C. Purcarea, Highly stable, cold-active aldehyde dehydrogenase from the marine Antarctic *Flavobacterium* sp. PL002, Fermentation 8 (2022) 7, doi:10.3390/fermentation8010007.
[3] T.E. Brewer, M. Albertsen, A. Edwards, R.H. Kirkegaard, E.P. Rocha, N. Fierer, Unlinked rRNA genes are widespread among bacteria and archaea, ISME J. 14 (2) (2020) 597–608.
[4] D. Seol, J.S. Lim, S. Sung, Y.H. Lee, M. Jeong, S. Cho, … H. Kim, Microbial identification using rRNA operon region: database and tool for metataxonomics with long-read sequence, Microbiol. Spectr. 10 (2) (2022) e02017-21.
[5] B.E.K. Staloch, H. Niero, R.C. de Freitas, P. Ballone, F. Rodrigues-Costa, D.B.B. Trivella, A. Dessen, M.A.C. da Silva, A.O. de Souza Lima, Draft genome sequence of *Psychrobacter nivimaris* LAMA 639 and its biotechnological potential, Data Br. 41 (2022) 107927, doi:10.1016/j.dib.2022.107927.
[6] X.J. Ching, N. Najimudin, Y.K. Cheah, C.M.V.L. Wong, Complete genome sequence data of tropical thermophilic bacterium *Parageobacillus caldoxylosilyticus* ER4B, Data Br. 40 (2022) 107764, doi:10.1016/j.dib.2021.107764.
[7] M. Kolmogorov, J. Yuan, Y. Lin, P.A. Pevzner, Assembly of long, error-prone reads using repeat graphs, Nat. Biotechnol. 37 (5) (2019) 540–546.
[8] 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.
[9] J.P. Meier-Kolthoff, A.F. Auch, H.P. Klenk, M. Göker, Genome sequence-based species delimitation with confidence intervals and improved distance functions, BMC Bioinform. 14 (1) (2013) 60.
[10] Lefort, V., Desper, R., & Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny infere
[11] Ciufo, S., Kannan, S., Sharma, S., Badretdin, A., Clark, K., Turner, S., … & DiCuccio, M. (2018). Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes
[12] J. Goris, K.T. Konstantinidis, J.A. Klappenbach, T. Coenye, P. Vandamme, J.M. Tiedje, DNA–DNA hybridization values and their relationship to whole-genome sequence similarities, Int. J. Syst. Evol. Microbiol. 57 (1) (2007) 81–91.
[13] K.T. Konstantinidis, J.M. Tiedje, Genomic insights that advance the species definition for prokaryotes, Proc. Natl. Acad. Sci. 102 (7) (2005) 256.