Scientific Data

https://doi.org/10.1038/s41597-026-06723-4

Article in Press

Complete genome sequence of *Sphingomonas sp.* gentR, a high-level gentamicin-resistant bacterium

Received: 22 September 2025

Accepted: 26 January 2026

Cite this article as: Liu, Y., Jiang, L., Zhang, J. *et al.* Complete genome sequence of *Sphingomonas sp.* gentR, a high-level gentamicin-resistant bacterium. *Sci Data* (2026). https://doi.org/10.1038/s41597-026-06723-4

Yi Liu, Lijing Jiang, Jinhua Zhang, Qiufen Li & Baosheng Liu

Title

Complete genome sequence of *Sphingomonas sp.* gentR, a high-level gentamicin-resistant bacterium

Authors

Yi Liu[1], Lijing Jiang[1], Jinhua Zhang[1], Qiufen Li[1, 2], Baosheng Liu[1, 2]

Affiliations

1. Jiangxi Provincial Key Laboratory for Animal Health, College of Animal Science and Technology, Jiangxi Agricultural University, Nanchang 330045, Jiangxi, China；

2. Institute of Veterinary Drug, Jiangxi Agricultural University, Nanchang 330045, Jiangxi, China

Corresponding author(s): Baosheng Liu (liubaosh@jxau.edu.cn),

Qiufen Li (fans2425@jxau.edu.cn)

Abstract

We present the complete genome sequence of *Sphingomonas* sp. gentR, a strain exhibiting high-level resistance to gentamicin (MIC = 40 mg/mL). The genome was assembled from hybrid Illumina and Nanopore sequencing data into a gap-free sequence of 4.0 Mbp, comprising one chromosome and two plasmids. A total of 3,692 coding sequences were predicted, with comprehensive functional annotation revealing genes associated with antibiotic resistance, stress adaptation, and metabolic diversity. Three confirmed resistance genes—*ANT(2")-Ia*, *ANT(3")-IIa*, and *Sul1*—were co-localized within a genomic island on plasmid B. This dataset provides insight into the genetic basis of high-level aminoglycoside resistance in *Sphingomonas* and serves as a valuable resource for studying horizontal gene transfer, environmental adaptation, and bioremediation potential. The genome sequence is publicly available under GenBank accessions CP144670–CP144672 and China National Genomics Data Center (accession number GWHDOHA00000000).

Background & Summary

Sphingomonas is a genus of Gram-negative, catalase-positive, non-sporulating rod-shaped bacteria[1], characterized by the presence of unique sphingoglycolipids instead of lipopolysaccharides typically found in the cell walls of Gram-negative bacteria[2]. Accumulating evidence indicates that *Sphingomonas* species thrive under oligotrophic conditions and are ubiquitously distributed[3], even in extreme environments such as desert sand, glacial ice, deep terrestrial subsurface sediments, and spacecraft that have left Earth[4]. The beneficial traits of *Sphingomonas*, including plant growth promotion[5], gellan gum production[6], and degradation of environmental polycyclic aromatic hydrocarbon contaminants[7], have attracted significant research interest worldwide[8]. However, to date, little attention has been paid to the antimicrobial resistance (AMR) of *Sphingomonas*. In particular, the high-level AMR of *Sphingomonas* raises critical concerns regarding the potential transfer of resistance genes to pathogenic bacteria in animals, plants, and humans, as well as the potential application of highly resistant *Sphingomonas* strains for bioremediation in antimicrobial-rich environments.

Sphingomonas sp. gentR is a high-level gentamicin-resistant strain isolated from a gentamicin working solution (0.05 mg/mL) stored at 4 °C in the laboratory (Table 1). This strain was identified as an unclassified *Sphingomonas* species through 16S rDNA sequence alignment. The minimal inhibitory concentration of gentamicin against *S.* gentR was determined to be 40 mg/mL using the broth microdilution method[9]. Although *Sphingomonas* species are known to exhibit natural resistance to streptomycin[10], they are

generally susceptible to other aminoglycosides in most assays[11]. To our knowledge, this is the first report of high-level gentamicin resistance in a *Sphingomonas* species.

Item	Description
Bacterium isolation date	2019/8/12
Geographic location	Jiangxi Agricultural University, China
Isolation source	Gentamicin working solution (0.05 mg/mL)
Culture media	Tryptic soy broth or tryptic soy agar
Culture condition	37℃, 120rpm (with broth)
Colonial morphology	Yellow, round, and shiny
Gram stain	Gram-negative
Taxonomy	*Sphingomonas* sp. gentR
Gentamicin resistance	The minimum inhibitory concentration is 40 mg/mL.

Table 1. The profile of *Sphingomonas* sp. gentR

Cells of *S.* gentR were grown in tryptic soy broth and harvested during the log phase. Genomic DNA was extracted following the standard protocol provided by Oxford Nanopore Technologies. The genome was sequenced using a combination of the Illumina NovaSeq 6000 and Nanopore PromethION platforms. After quality filtering of sequencing reads, genome assembly was performed using a hybrid assembly strategy. The complete, gap-free genome was used for structural interpretation, component prediction, and functional annotation (Fig. 1).

The complete genome sequence of *S.* gentR presented here will facilitate the prediction and investigation of high-level gentamicin resistance mechanisms in *Sphingomonas* from a genomic perspective and aid in efforts to prevent the transmission of high-level gentamicin resistance genes via *Sphingomonas*. Additionally, this genome sequence provides a valuable resource for further studies on stress resistance, degradation of exogenous compounds, and synthetic capabilities of *Sphingomonas*.

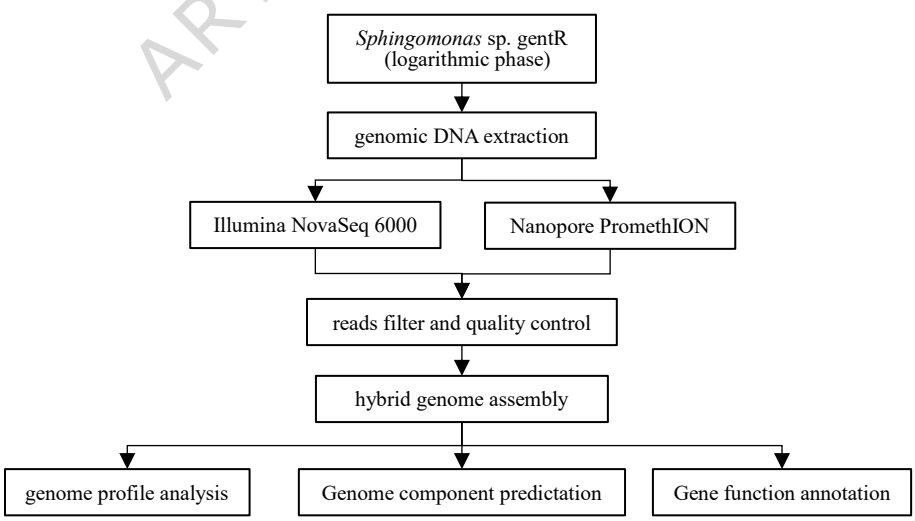


Fig. 1 Overview of the procedures for this study

Methods

Bacterial growth and genomic DNA extraction

Sphingomonas sp. gentR was inoculated at 5% (v/v) into tryptic soy broth and incubated overnight at 37 °C with

shaking at 120 rpm. Cells were harvested by centrifugation at 10,000 × *g* for 10 min, and genomic DNA was extracted from the pellet using the standard protocol from Oxford Nanopore Technologies (ONT). DNA concentration and quality were assessed using a NanoDrop One spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA), a Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, USA), and 0.7% (w/v) agarose gel electrophoresis.

Genome sequencing

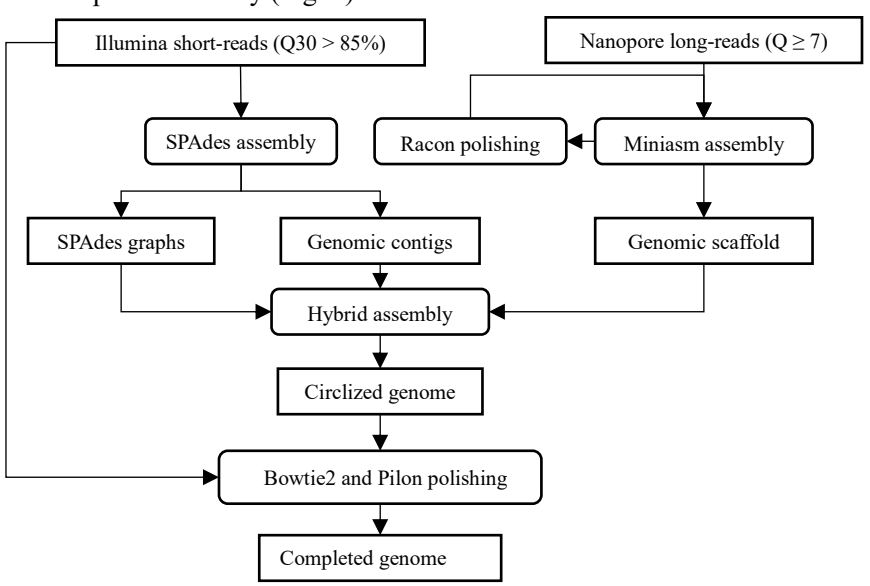
Genome sequencing was performed using both Illumina NovaSeq 6000 and Nanopore PromethION platforms at Wuhan Benagen Technology Company Limited (Wuhan, China).

Illumina Sequencing. For short-read sequencing on the NovaSeq 6000 platform, qualified genomic DNA sample was randomly fragmented using a Covaris ultrasonicator (Covaris, USA). Subsequently, the NEBNext® Ultra™ II DNA Library Prep Kit for Illumina (NEB, USA) was employed to perform end repair, 5′ phosphorylation and dA-tailing, adapter ligation, purification, and PCR amplification, thereby constructing the DNA sequencing library with an average insert size of 300 bp. Library quantification was conducted using an Agilent 2100 Bioanalyzer (Agilent DNA 1000 reagents; Agilent, Santa Clara, CA, USA) and real-time quantitative PCR (RT-qPCR). Qualified libraries were amplified on an Illumina cBOT instrument for cluster generation (NovaSeq 6000 PE cluster kit; Illumina). The clustered flow cell was sequenced on a NovaSeq 6000 sequencer (NovaSeq 6000 S4 Reagent Kit; Illumina) with 150 bp paired-end reads.

Oxford Nanopore PromethION Sequencing. For long-read sequencing, libraries were prepared using the SQK-LSK109 ligation kit according to the manufacturer's protocol. The purified library was loaded onto primed R9.4.1 Spot-On Flow Cells and sequenced on a PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK). Base calling was performed using Oxford Nanopore GUPPY software (v0.3.0).

Hybrid genome assembly

Quality-filtered reads from both platforms were used for hybrid genome assembly with Unicycler (v0.4.8, SI1) running in normal bridging mode[12]. High-accuracy Illumina reads (Q30 > 85%) were used to construct a high-quality genome skeleton (contigs), which was then scaffolded using long reads from Nanopore sequencing. The assembly was further polished with Pilon[13] (https://github.com/broadinstitute/pilon) and Bowtie2[14] (v2.4.2) using Illumina data to improve accuracy (Fig. 2).



Note: The rectangular boxes indicate the data format; the rounded square boxes show the data processing methods.

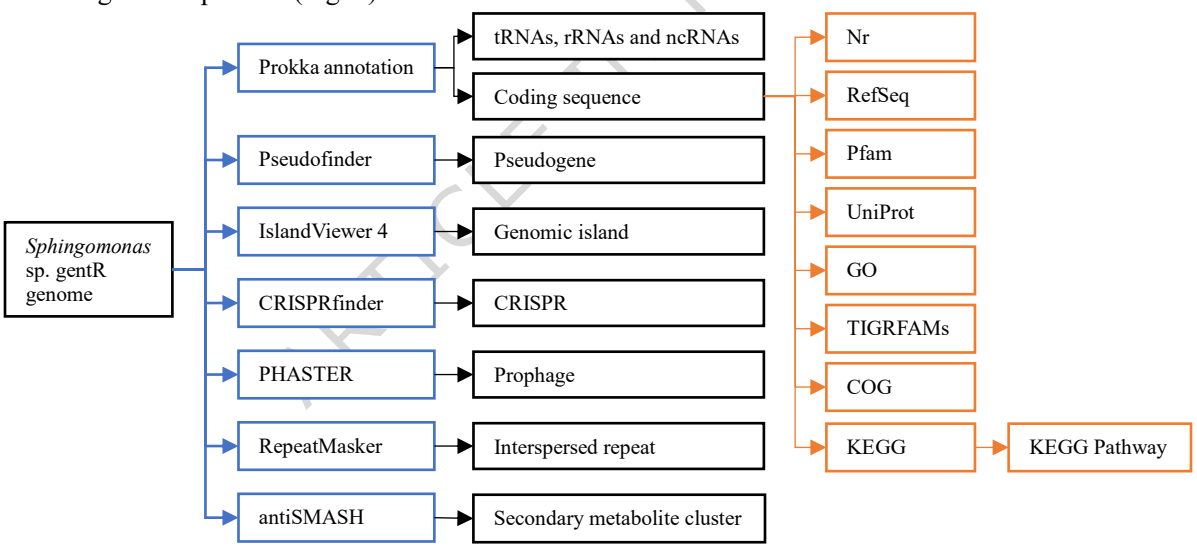
Fig. 2 Workflow of the hybrid genome assembly

Genome component prediction

Gene prediction was performed using Prokka (v1.13)[15], which employs Prodigal (v2.6), Aragorn (v1.2), RNAmmer (v1.2), and cmscan (v1.1) to predict coding sequences, tRNAs, rRNAs, and ncRNAs, respectively (Fig. 3). Pseudogenes were identified using Pseudofinder[16]. Genomic features including CRISPR arrays, genomic islands, prophages, interspersed repeats, and secondary metabolite gene clusters were predicted using CRISPRfinder[17] (https://crispr.i2bc.paris-saclay.fr/), IslandViewer 4[18] (http://www.pathogenomics.sfu.ca/islandviewer/), PHASTER[19] (http://phaster.ca/), RepeatMasker (http://repeatmasker.org), and antiSMASH[20](v5.2.0), respectively. The genome was also annotated using the National Center for Biotechnology Information (NCBI) Prokaryotic Genome Annotation Pipeline (PGAP) after submission to GenBank (Table 2).

Functional gene annotation

Coding sequences (CDS) were functionally annotated using nine databases: UniProt[21], Pfam[22], RefSeq[23], Nr (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz), TIGRFAMs[24], GO[25], KEGG[26], COG[27], and KEGG Pathway (https://www.genome.jp/kegg/). Results from Pfam, GO, COG, and KEGG were visualized to summarize functional genomic profiles (Fig. 3).



Note: The blue boxes show the softwires used in the annotation; The black boxes show the genomic components of the genome annotation; The orange boxes show the databases utilized in the coding sequence annotation.

Fig. 3 Workflow of the genome annotation

Data Records

Genome sequencing

After quality filtering, a total of 1.4 Gbp and 1.0 Gbp of clean data were obtained from Illumina and Oxford Nanopore sequencing, respectively, with average sequencing depths of 360.2× and 255.86× (Table 2).

	Illumina	Oxford Nanopore Technologies
Sequencing system	NovaSeq 6000	PromethION
Read length	2 × 150 bp	From 3,018bp to 201,469 bp
Run number	450	--

	Illumina	Oxford Nanopore Technologies
Flow cell (ID)	HVNH2DSXY	R9.4.1 (PAG32540)
Sequencing Kit	NovaSeq 6000 S4 Reagent Kit	SQK-LSK109
Indices / Barcodes	ATGGCTGA+CCATGGAA	NB19
Number of reads	9,646,748	60,926
Average length (bp)	150	16,413.4
Total bases (bp)	1,447,012,200	1,000,005,752
Q30 (%)	92.57	--
G+C (mol%)	66.12	--
N50 (bp)	150	34,488
Sequencing depth	360.2 ×	255.86 ×
File organization	JNDY-SC_1.clean.fq.gz JNDY-SC_2.clean.fq.gz	JNDY-SC.fq.gz

Table 2. A summary of Illumina and Oxford Nanopore sequencing

The complete genome assembly (gap-free) comprises 4,009,209 bp, including one chromosome (3,798,193 bp, 66.17% GC) and two plasmids (132,630 bp, 63.43% GC, and 78,386 bp, 62.31%). A circular genome map was generated using Proksee[28] (Fig. 4).

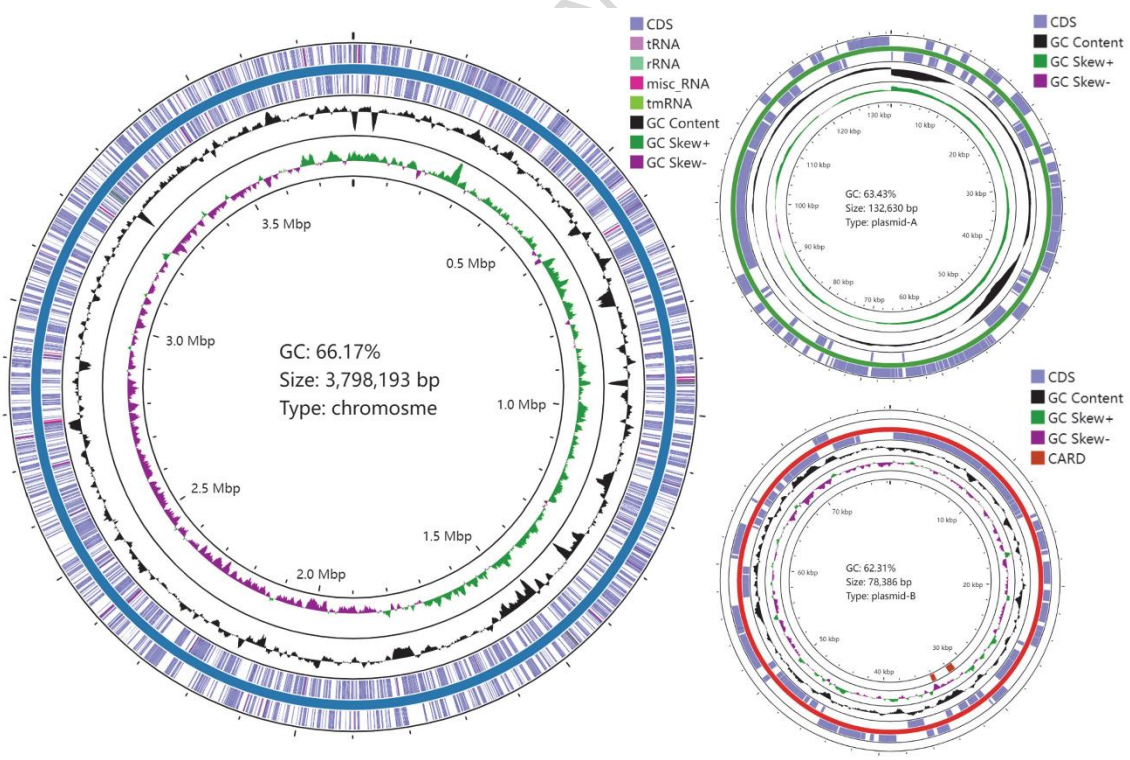


Fig. 4 The complete genome of *Sphingomonas* sp. gentR

The genome sequence data have been deposited in the Genome Warehouse[29]of the China National Genomics Data Center[30]under accession number GWHDOHA00000000[31], and in the NCBI Genome database under BioProject PRJNA1072271 with accession numbers CP144670, CP144671, and CP144672[32].

Genome components prediction

Prokka predicted a total of 3,784 genes, including 3,692 CDS and 92 RNA genes (tRNAs, rRNAs, tmRNA and ncRNAs). NCBI PGAP[33] annotation yielded slightly different results: 3,787 total genes, 3,682 CDS, and 75 RNA genes (tRNAs, rRNAs and ncRNAs). Pseudofinder identified approximately eight times more pseudogenes than PGAP. Various genomic functional components, including genomic islands, CRISPR arrays, prophages, interspersed repeats, and secondary metabolite gene clusters, were widely distributed throughout the genome (Table 3).

Item	Prokka	PGAP
Gene	3,784	3,787
Coding sequence (CDS)	3,692	3,682
tRNA	63	60
rRNA	12	12
tmRNA	1	-
ncRNA	16	3
Pseudogene	235*	30
Other genomic components		
Genomic islands	10	
CRISPR arrays	5	
Prophage	1	
Interspersed repeats	740	
Secondary metabolite clusters	4	

* Predicted by Pseudofinder.

Table 3. Genome component profile of *Sphingomonas* sp. gentR

Functional gene annotation

A total of 99.19% of CDS were annotated in at least one of the nine databases used. Both Nr and RefSeq annotated more than 98% CDS[34]. A summary of annotation results is provided in Table 4.

Item	Count	Percentage
CDS	3692	100%
Annotated	3662	99.19%
Nr	3651	98.89%
RefSeq	3628	98.27%
Pfam	3115	84.37%
UniProt	2021	54.74%
GO	1962	53.14%
TIGRFAMs	1929	52.25%
KEGG	1863	50.46%
COG	1341	36.32%
KEGG Pathway	1081	28.57%

Table 4. Functional annotation of coding genes in *Sphingomonas* sp. gentR

Pfam annotated 3115 genes in this genome. The most abundant genes were allocated to TonB-dependent receptor-related domains and two-component systems domains (Fig. 5).

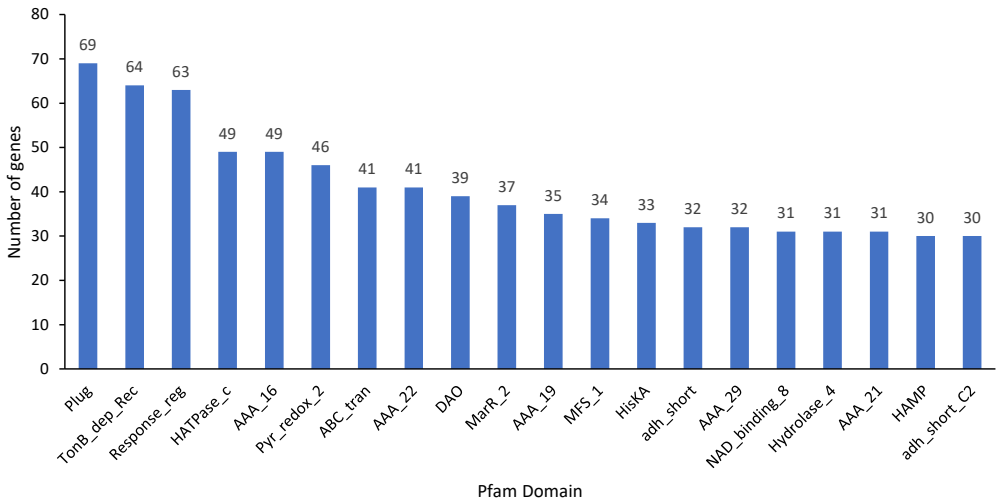


Fig. 5 Top20 domains annotated in Pfam database

Abbreviations: Plug: TonB-dependent receptor plug domain; TonB_dep_Rec: TonB-dependent receptor; Response_reg: Response regulator receiver domain; HATPase_c: Histidine kinase-like ATPase domain; AAA_16: AAA ATPase domain；Pyr_redox_2: Pyridine nucleotide-disulphide oxidoreductase; ABC_tran: ABC transporter; AAA_22: AAA domain; DAO: FAD dependent oxidoreductase; MarR_2: MarR family; AAA_19: Part of AAA domain; MFS_1: Major Facilitator Superfamily; HisKA: His Kinase A (phospho-acceptor) domain; adh_short: short chain dehydrogenase; AAA_29: P-loop containing region of AAA domain; Hydrolase_4 : Serine aminopeptidase; AAA_21: AAA domain, putative AbiEii toxin, Type IV TA system; NAD_binding_8: NAD(P)-binding Rossmann-like domain; HAMP: HAMP domain; adh_short_C2: Enoyl-(Acyl carrier protein) reductase.

GO annotation assigned functions to 1,962 genes (51.85% of total genes). The most enriched cellular components were cytoplasm, plasma membrane, and integral membrane components (Fig. 6). The most common molecular functions were ATP binding, DNA binding, and metal ion binding.

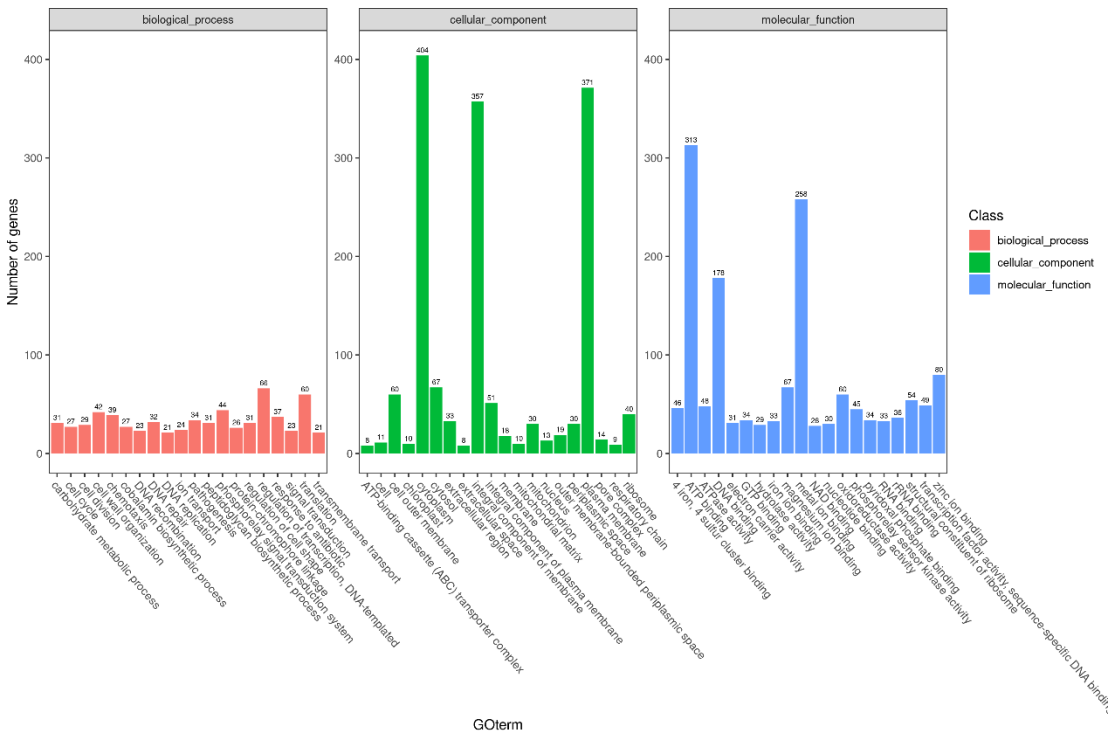


Fig. 6 GO annotation of the genomic genes

COG annotation classified 1,341 genes (35.44% of total genes) into 22 functional categories. The largest groups were: C (energy production and conversion), E (amino acid transport and metabolism), and J (translation, ribosomal structure, and biogenesis), together accounting for 10.7% of all CDS (Fig. 7).

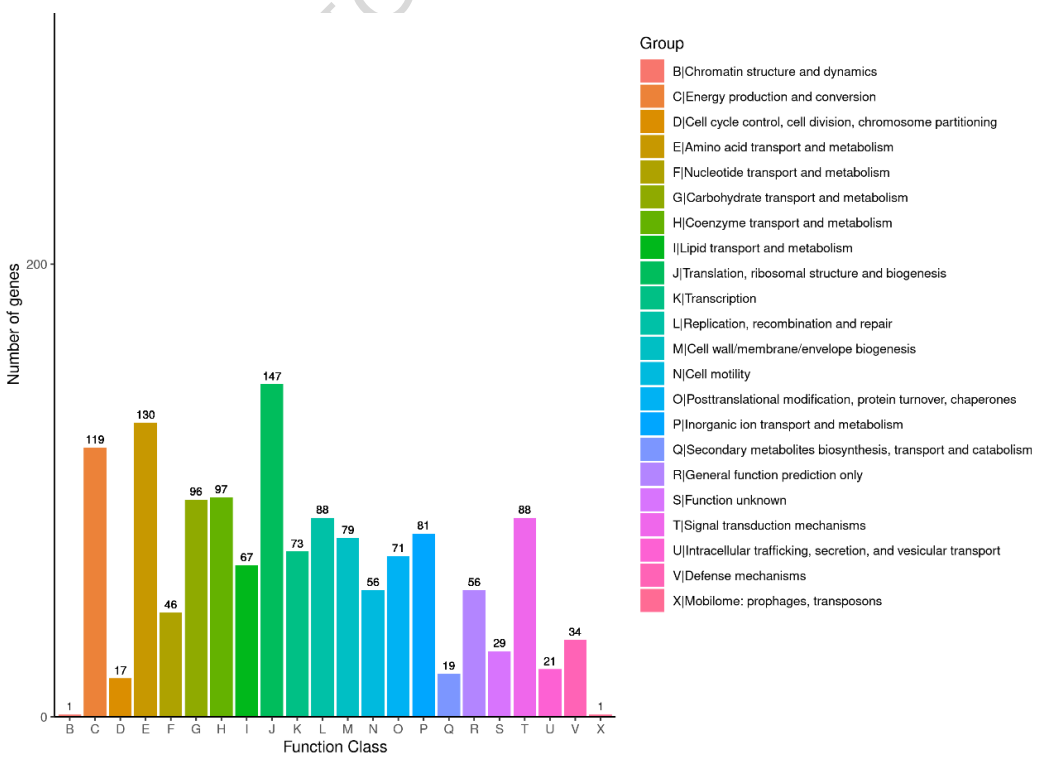


Fig. 7 COG annotation of the genomic genes

KEGG annotation identified 1,863 genes, with 1,081 assigned to KEGG pathways (Fig. 8). Metabolism-related genes were the most abundant (82.7% of annotated genes).

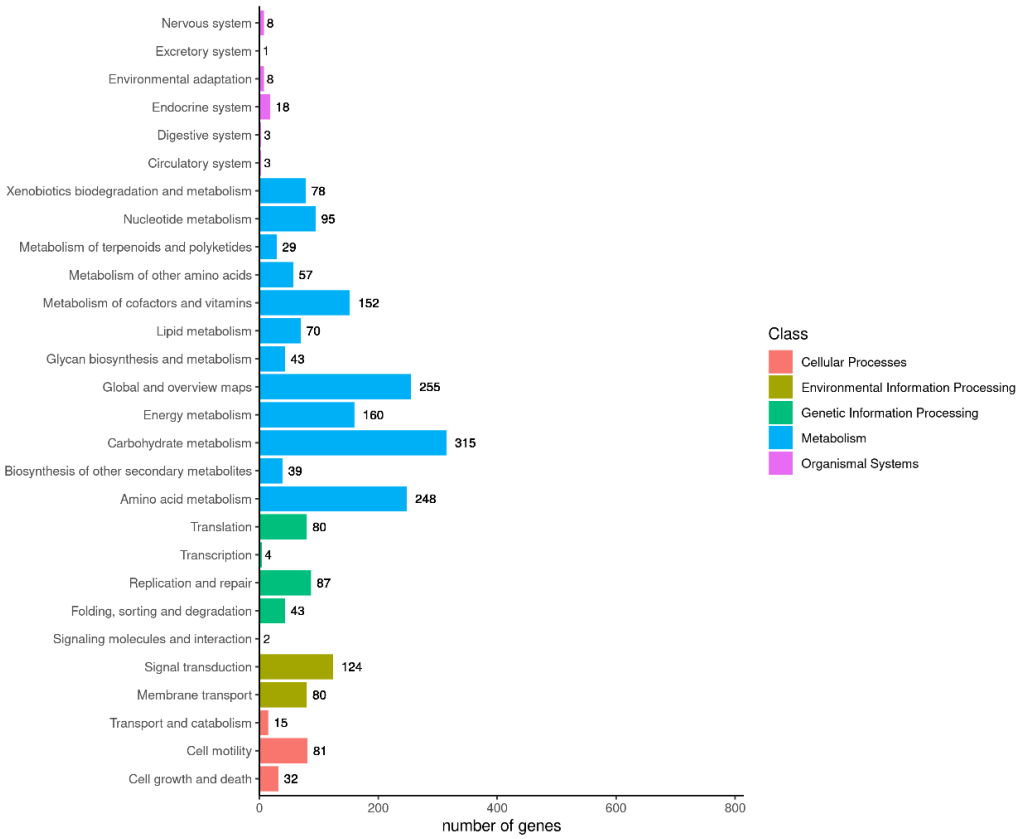


Fig. 8 KEGG pathway annotation of the genomic genes

Genome annotation of antibiotic resistance

Screening against the Antibiotic Resistance Genes Database (ARDB)[35] identified four genes which confer resistance to aminoglycosides (*ANT(2")-Ia*, *ANT(3")-IIa*), bacitracin (*BacA*), and sulfonamide (*Sul1*).

According to the Comprehensive Antibiotic Resistance Database (CARD)[36], 3,361 genes were predicted to be associated with antibiotic resistance. Among these, 52 genes had ≥60% sequence identity to known resistance genes (Table S1). Specifically, 22 genes were involved in antibiotic inactivation, 15 in efflux, and 15 in target alteration, protection, or replacement (Fig. 9). These genes were associated with resistance to aminoglycosides, carbapenems, cephalosporins, fluoroquinolones, tetracyclines, peptides, sulfonamides, lincosamides, and macrolides. Notably, only *Sul1*, *ANT(2")-Ia*, and *ANT(3")-IIa* showed >99% identity and were co-localized within a genomic island on plasmid B (Fig. 10).

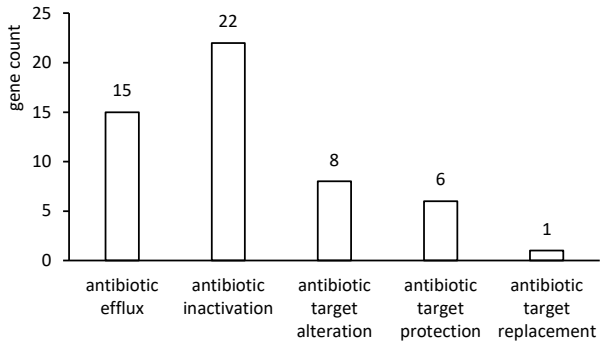
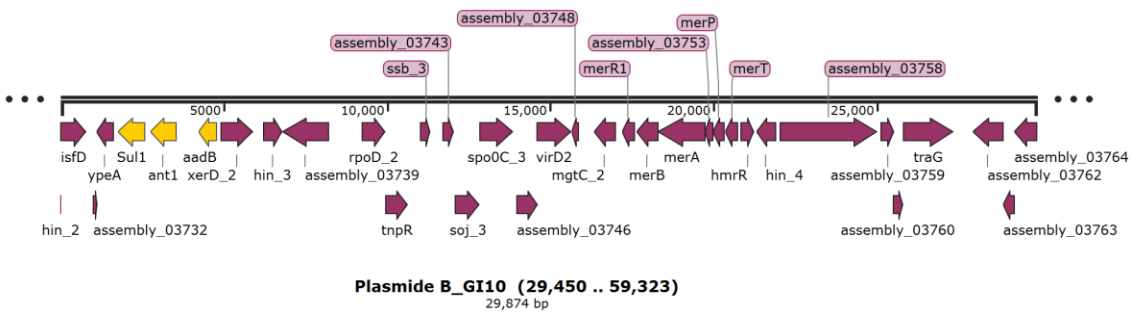


Fig. 9 Antibiotic resistance associated genes (identity⩾60%) predicted in CARD



A total of 29,874 base pairs with 65 CDS. *Sul1*, *ANT(2″)-Ia*, and *ANT(3″)-IIa* are marked in yellow.

Fig. 10 Genomic island framework carrying antimicrobial resistance genes

Data Overview

Genomic annotation of *S. sp. gentR* using the ARDB database identified four antimicrobial resistance genes (ARGs)—*ANT(2″)-Ia*, *ANT(3″)-IIa*, *BacA*, and *Sul1*—which confer resistance to gentamicin, streptomycin, bacitracin, and sulfonamides, respectively[9]. Comparative analysis with the CARD database further revealed a broad repertoire of resistance-related genes, including those involved in antibiotic efflux, inactivation, and target alteration/protection/replacement, potentially mediating resistance to macrolides, fluoroquinolones, aminoglycosides, tetracyclines, phenicols, and penicillins. Additionally, Pfam annotations indicated an enrichment of domains associated with two-component regulatory systems (Response_reg, HATPase_c, HisKA, and HAMP) and multidrug resistance regulators (MarR_2)[37], suggesting diverse and sophisticated resistance mechanisms in *S.* sp. gentR.

Sphingomonas species exhibit robust environmental adaptability[38]. In *S.* sp. gentR, COG annotation revealed a high abundance of genes related to category C (energy production and conversion), E (amino acid transport and metabolism), and J (translation, ribosomal structure, and biogenesis). Pfam domain analysis highlighted a predominance of genes encoding TonB-dependent receptor domains (Plug, TonB_dep_Rec), oxidoreductase-related domains (Pyr_redox_2, DAO, adh_short), and transporter family domains (ABC_tran and MFS_1). GO terms prominently featured cytoplasmic and membrane localization (cytoplasm, integral component of plasma membrane, plasma membrane) and functional categories such as ATP binding, DNA binding, and metal ion binding. KEGG pathway analysis further indicated that 82.7% of annotated genes were involved in metabolic processes. Collectively, these annotations illustrate the genetic repertoire supporting the environmental resilience and survival capacity of *S.* sp. gentR.

Technical Validation

Genomic validation and taxonomic classification

Prior to genome sequencing, the strain was confirmed to exhibit high-level gentamicin resistance and was preliminarily classified as *Sphingomonas* sp. based on 16S rRNA gene alignment. To further resolve its taxonomic position, average nucleotide identity (ANI) analysis was conducted using the EZBioCloud ANI calculator[39] (http://www.ezbiocloud.net/tools/ani). The genome of *S.* sp. gentR was compared with those of 12 strains selected from the top 100 16S rRNA homologs and four publicly available genomes of *Sphingomonas yabuuchiae* retrieved from NCBI. Five strains showed ANI values exceeding 95% (Table 5). Digital DNA–DNA hybridization (dDDH) was subsequently performed between *S.* sp. gentR and these five high-ANI strains via the Genome-to-Genome Distance Calculator (GGDC 3.0)[40](https://ggdc.dsmz.de/ggdc.php). With the exception of strain Xoc002 (dDDH < 70%), the remaining four strains displayed dDDH values above the species delineation threshold of 70%[41]. Strain LK11, an endophytic plant growth-promoting bacterium[42], reached a dDDH value of 99.9%, while the three *S. yabuuchiae* strains showed values around 73% (Table 6). Together with the initial 16S rRNA gene-based assignment, these results robustly confirm that strain gentR belongs to the genus *Sphingomonas*. Moreover, the genomic comparisons strongly indicate that strain gentR is likely conspecific with *Sphingomonas yabuuchiae*.

Genome B	OrthoANIu value (%)	Genome A* length (bp)	Genome B length (bp)	Average aligned length (bp)	Genome A coverage (%)	Genome B coverage (%)	NCBI Accession
Sphingomonas sp. LK11	99.86	4,007,580	3,936,180	2,677,028	66.8	68.01	CP013916.1
*Sphingomonas yabuuchiae*_refseq	96.83	4,007,580	4,156,500	2,491,970	62.18	59.95	GCA_017052455.1
Sphingomonas yabuuchiae strain DSM 14562	96.79	4,007,580	4,162,620	2,446,090	61.04	58.76	GCA_014199595.1
Sphingomonas yabuuchiae strain JCM 11416	96.76	4,007,580	4,767,480	2,560,453	63.89	53.71	GCA_042661065.1
Sphingomonas sp. Xoc002	95.46	4,007,580	3,870,900	2,286,055	57.04	59.06	CP191175.1
Sphingomonas yabuuchiae strain NS355	92.50	4,007,580	3,806,640	2,155,878	53.79	56.63	GCA_001477495.1
Sphingomonas parapaucimobilis strain YK209	91.93	4,007,580	344,760 #	228,384	5.7	66.24	CP155754.1
Sphingomonas sanguinis strain NP2-R2	91.10	4,007,580	4,274,820	2,150,970	53.67	50.32	CP079203.1
Sphingomonas yabuuchiae strain SPH4	89.27	4,007,580	3,836,220	2,205,403	55.03	57.49	CP158795.1
Sphingomonas pseudosanguinis strain S81-1-1	86.20	4,007,580	3,531,240	1,912,922	47.73	54.17	CP189888.1
Sphingomonas paucimobilis strain ZJSH1	85.52	4,007,580	3,987,180	1,836,726	45.83	46.07	CP070367.1
Sphingomonas paucimobilis strain FDAARGOS_908	85.43	4,007,580	3,911,700	1,866,677	46.58	47.72	CP065670.1
Sphingomonas paucimobilis strain FDAARGOS_881	85.42	4,007,580	4,059,600	1,908,705	47.63	47.02	CP065713.1

Sphingomonas paucimobilis strain AIMST S2	85.39	4,007,580	4,004,520	1,878,775	46.88	46.92	CP035765.1
Sphingomonas paucimobilis	85.36	4,007,580	3,916,800	1,887,424	47.1	48.19	AP023323.1
Chromobacterium piscinae strain AK003	68.11	4,007,580	4,949,040	162,342	4.05	3.28	CP197095.1

Note: * OrthoANIu was calculated between *Sphingomonas* sp. gentR (genome A) and genome B; # Sequence of chromosome.

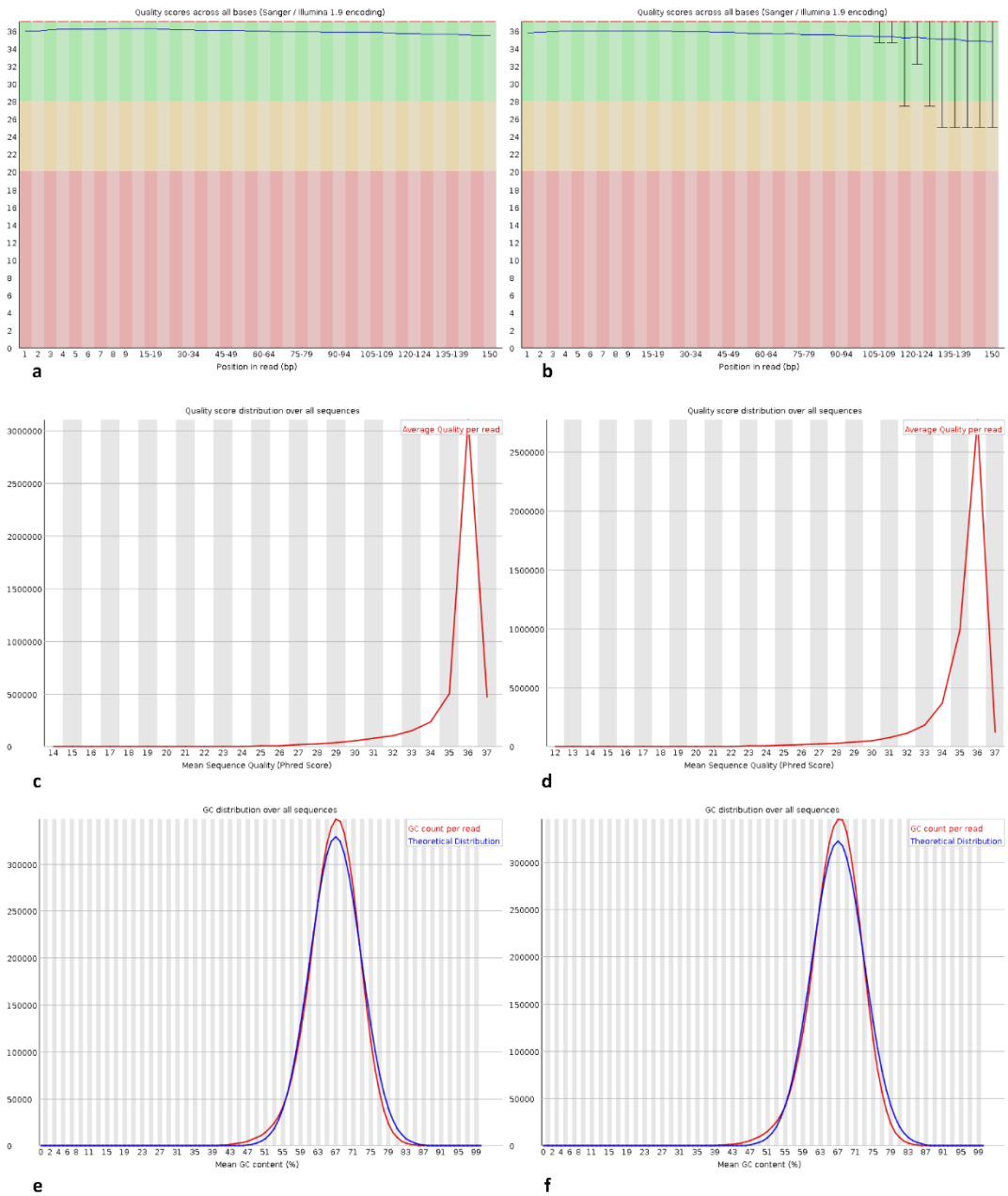
Table 5. ANI analysis results of *Sphingomonas* sp. gentR

Reference genome	DDH%	Model C.I.	Distance	Prob.% (DDH >= 70%)
*Sphingomonas yabuuchiae*_refseq	73.3	[70.3 - 76.2%]	0.0316	83.63
Sphingomonas sp. LK11	99.9	[99.9 - 100%]	0.0002	98.28
Sphingomonas yabuuchiae strain DSM 14562	73.3	[70.3 - 76.1%]	0.0316	83.59
Sphingomonas yabuuchiae strain JCM 11416	72.9	[69.9 - 75.7%]	0.0322	83.03
Sphingomonas sp. Xoc002	63.7	[60.8 - 66.5%]	0.0454	63.85

Table 6. dDDH analysis results of *Sphingomonas* sp. gentR

Quality control and processing of sequencing data

Raw reads generated by the Illumina NovaSeq 6000 platform were processed using SOAPnuke[43] (v2.1.4) to remove low-quality sequences, including adapter contaminants, reads containing ambiguous nucleotides (N's), and those with more than 50% of bases having a Phred quality score ≤5. After filtering, 9,646,748 clean reads (1,450,637,700 bp) were retained from the original 9,670,918 raw reads (1,447,012,200 bp). The resulting paired-end clean read files (JNDY-M2_1.clean.fq.gz and JNDY-M2_2.clean.fq.gz) were subsequently evaluated with FastQC (v0.11.9), which confirmed high sequencing quality based on per-base sequence quality (score >34), per-sequence quality scores (>35), and per-sequence GC content consistent with the theoretical distribution (Fig. 11).



Note: a, c e and b, d, f show the results of per base sequence quality, per sequence quality scores and per sequence GC content for JNDY-M2_1.clean.fq.gz and JNDY-M2_1.clean.fq.gz respectively.

Fig. 11 FastQC quality assessment results of Illumina paired-end reads

For Nanopore PromethION data, base calling was performed using GUPPY, which automatically filtered out reads with a quality score below 7. A total of 60,926 qualified reads were obtained, with lengths ranging from 3,018 bp to 201,469 bp (Table 2). The length distribution of these reads was visualized using R (v4.4.3) (Fig. 12).

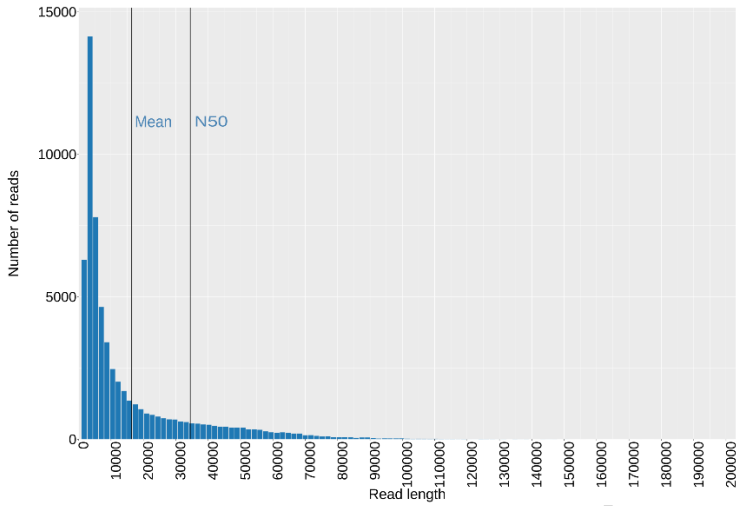
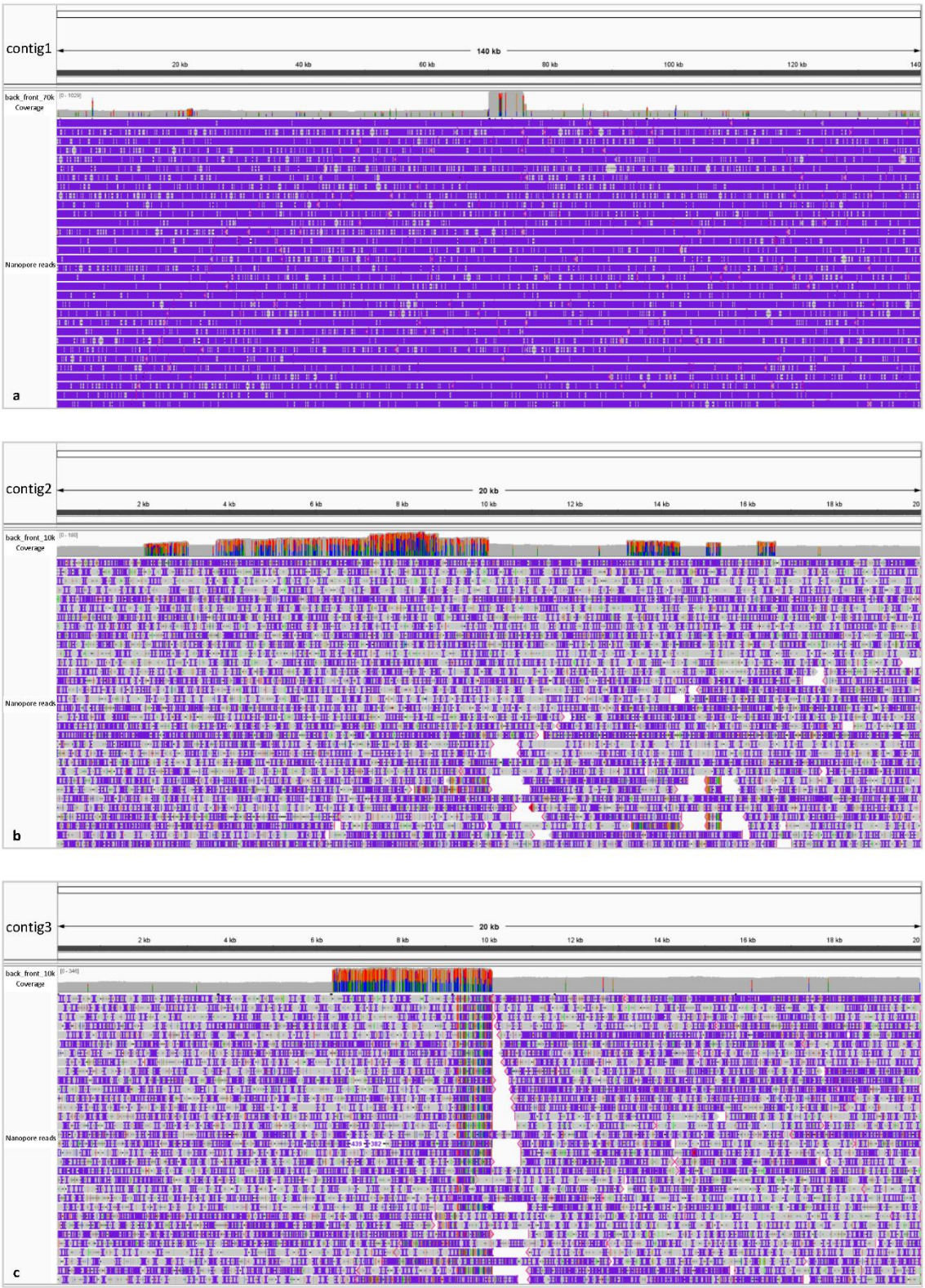


Fig. 12 Histogram of the length distribution of Nanopore long-reads

Genome assembly and quality assessment

The hybrid assembly generated three contigs with sizes of 3,798,193 bp, 132,630 bp, and 78,386 bp (Fig. 4). Circularization was automatically performed by Unicycler through bridging, and the starting base sequence of each replicon was determined[12]. Quality evaluation using CheckM2 (v1.0.2)[44] indicated a completeness of 100% (PGAP reports 99.74%) and a contamination rate of 0.76% (PGAP reports 2%). Mapping analysis was carried out by aligning Illumina reads with BWA (v0.7.18)[45] and Nanopore reads with minimap2 (v2.28) against the final assembly, yielding mapping rates of 99.15% and 98.10%, respectively. To verify the circularity of the three contigs, 70 kb (for contig1), 10 kb (for contig2), and 10 kb (for contig3) of sequence from both ends of each contig were extracted and aligned against the Nanopore long reads using minimap2 (parameters: -ax map-ont). The results demonstrated complete coverage of the terminal regions by long reads across all three contigs, confirming their circular nature. Visualization of the alignments was performed using Integrative Genomics Viewer (IGV, v2.14) (Fig. 13).



Note: a, b and c show the alignment results of contig1, contig2 and contig3 against nanopore reads respectively.

Fig. 13 Circularization verification of three contigs with alignment against nanopore long reads

Genome annotation

Genome annotation was performed using both Prokka and NCBI PGAP to ensure accuracy. Functional annotation incorporated multiple major databases for comprehensive gene function analysis.

Data Availability

The sequencing data and assembled genome sequence generated in this study have been deposited in publicly accessible repositories. The details are as follows:

Raw Sequencing Reads

The datasets have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession number PRJNA1072271 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1072271). The data are organized under BioSample SAMN39740430 and SRA experiment accession SRR36181637[46]. The repository contains the following files:

- JNDY-SC.fq.gz – Compressed FASTQ file containing Nanopore long reads.
- JNDY-SC_1.clean.fq.gz – Compressed FASTQ file containing paired-end Illumina short reads (read 1).
- JNDY-SC_2.clean.fq.gz – Compressed FASTQ file containing paired-end Illumina short reads (read 2).

Assembled Genome Sequence

The complete, annotated genome assembly of *Sphingomonas* sp. gentR has been deposited in the NCBI GenBank database under the same BioProject PRJNA1072271 and BioSample SAMN39740430. The assembly is available as a FASTA file (assembly.fna) and is organized into three records:

- CP144670: Chromosome sequence.
- CP144671: Plasmid A sequence.
- CP144672: Plasmid B sequence.

Genome Warehouse Access

The complete genome sequence is also available in the Genome Warehouse of the China National Genomics Data Center under accession number GWHDOHA00000000 (https://ngdc.cncb.ac.cn/gwh).

These datasets are freely accessible and can be used to explore the genomic basis of high-level gentamicin resistance and other functional traits in *Sphingomonas* sp. gentR.

Code Availability

All software tools used in this study have been properly cited or accompanied by relevant website links. Unless otherwise specified in the manuscript, all data analyses were performed using default parameters as described in the respective software manuals. The running codes for genome assembly and annotation using Unicycler and prokka, respectively, have been uploaded as a supplementary information (SI1) to Figshare.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (Grant No. 32360880, 31360553).

Author contributions

Jinhua Zhang and Baosheng Liu conceived the project. Yi Liu, Lijing Jiang and Qiufen Li collected the samples, performed the genome assembly, gene annotation and other bioinformatics analysis. Yi Liu and Lijing Jiang wrote the manuscript. Qiufen Li and Baosheng Liu revised the manuscript.

Yi Liu and Lijing Jiang contributed equally to this work.

Competing interests

The authors declare no competing interests.

References

1. Chen, H. *et al.* Reclassification and emended description of *Caulobacter leidyi* as *Sphingomonas leidyi* comb. nov., and emendation of the genus *Sphingomonas*. *Int. J. Syst. Evol. Microbiol.* **62,** 2835-2843 (2012).
2. Yabuuchi, E. *et al.* Proposals of *Sphingomonas paucimobilis* gen. nov. and comb. nov., *Sphingomonas parapaucimobilis* sp. Nov., *Sphingomonas yanoikuyae* sp. nov., *Sphingomonas adhaesiva* sp. nov., *Sphingomonas capsulata* comb. nov., and two genospecies of the genus *Sphingomonas*. *Microbiol. Immunol.* **34,** 99-119 (1990).
3. White, D. C., Sutton, S. D. & Ringelberg, D. B. The genus *sphingomonas*: physiology and ecology. *Curr. Opin. Biotechnol.* **7**, 301-306 (1996).
4. Li, Y. *et al. Sphingomonas yabuuchiae* sp. nov. and *Brevundimonas nasdae* sp. nov., isolated from the Russian space laboratory Mir. *Int. J. Syst. Evol. Microbiol.* **54,** 819-825 (2004).
5. Kampfer, P. *et al. Flavobacterium plantiphilum* sp. nov., *Flavobacterium rhizophilum* sp. nov., *Flavobacterium rhizosphaerae* sp. nov., *Chryseobacterium terrae* sp. nov., and *Sphingomonas plantiphila* sp. nov. isolated from salty soil showing plant growth promoting potential. *Syst. Appl. Microbiol.* **48,** 126588 (2025).
6. Liu, H. *et al.* Fed-batch fermentation strategy for efficient welan gum production by *Sphingomonas* sp. FM01. *J. Sci. Food Agric.* **105,** 926-936 (2025).
7. Sanchez-Arroyo, A., Plaza-Vinuesa, L., de Las Rivas, B., Mancheno, J. M. & Munoz, R. Analysis of the subtype I amidohydrolase responsible for Ochratoxin A degradation in the *Sphingomonas* genus. *Int. J. Biol. Macromol.* **306,** 141720 (2025).
8. Asaf, S., Numan, M., Khan, A. L. & Al-Harrasi, A. *Sphingomonas*: from diversity and genomics to functional role in environmental remediation and plant growth. *Crit. Rev. Biotechnol.* **40,** 138-152 (2020).
9. Jiang, L. *et al.* Identification of a high-level gentamicin-resistant *Sphingomonas* strain and its antimicrobial susceptibility test. *Biol. Disaster Sci.* (in Chinese) **48,** 181-190 (2025).
10. Vanbroekhoven, K. *et al*. Streptomycin as a selective agent to facilitate recovery and isolation of introduced and indigenous *Sphingomonas* from environmental samples. *Environ. Microbiol.* **6,** 1123-1136 (2004).
11. Park, H. K. *et al*. *Sphingomonas aeria* sp. Nov. from indoor air of a pharmaceutical environment. *Antonie Van Leeuwenhoek.* **107,** 47-53 (2015).
12. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13,** e1005595 (2017).
13. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963, (2014).
14. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. **9,** 357-359 (2012).
15. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068-2069 (2014).
16. Syberg-Olsen, M. J., Garber, A. I., Keeling, P. J., McCutcheon, J. P. & Husnik, F. Pseudofinder: detection of pseudogenes in prokaryotic genomes. *Mol. Biol. Evol.* **39,** msac153 (2022).
17. Grissa, I., Vergnaud, G. & Pourcel, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic. Acids. Res.* **35,** W52-W57 (2007).

18. Bertelli, C. *et al*. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic. Acids. Res.* **45,** W30-W35 (2017).
19. Arndt, D. *et al*. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic. Acids. Res.* **44,** W16-W21 (2016).
20. Blin, K. *et al*. AntiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. *Nucleic. Acids. Res.* **53,** W32-W38 (2025).
21. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic. Acids. Res.* **53,** D609-D617 (2025).
22. Mistry, J. *et al*. Pfam: the protein families database in 2021. *Nucleic. Acids. Res.* **49,** D412-D419 (2021).
23. Goldfarb, T. *et al*. NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic. Acids. Res.* **53,** D243-D257 (2025).
24. Haft, D. H. *et al*. TIGRFAMs and genome properties in 2013. *Nucleic. Acids. Res.* **41,** D387-D395 (2012).
25. Aleksander, S. A. *et al*. The gene ontology knowledgebase in 2023. *Genetics* **224,** iyad031 (2023).
26. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic. Acids. Res.* **53,** D672-D677 (2025).
27. Galperin, M. Y. *et al*. COG database update 2024. *Nucleic. Acids. Res.* **53,** D356-D363 (2025).
28. Grant, J. R. *et al*. Proksee: in-depth characterization and visualization of bacterial genomes. *Nucleic. Acids. Res.* **51,** W484-W492 (2023).
29. Chen, M. *et al*. Genome warehouse: a public repository housing genome-scale data. *Genom. Proteomics Bioinformatics* **19,** 584-589 (2021).
30. Bai, X. *et al*. Database resources of the national genomics data center, China national center for bioinformation in 2024. *Nucleic. Acids. Res.* **52,** D18-D32 (2024).
31. NGDC https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA016435 (2024).
32. NCBI GenBank http://identifiers.org/insdc.gca:GCA_036596345.1 (2024).
33. Tatusova, T. *et al*. NCBI prokaryotic genome annotation pipeline. Nucleic. Acids. Res. 44, 6614-6624 (2016).
34. Li, W. *et al*. RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic. Acids. Res.* **49,** D1020-D1028 (2021).
35. Liu, B. & Pop, M. ARDB--antibiotic resistance genes database. *Nucleic. Acids. Res.* **37,** D443-D447 (2009).
36. Alcock, B. P. *et al*. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic. Acids. Res.*, gkz935 (2019).
37. Alvarez, A. F. & Georgellis, D. Environmental adaptation and diversification of bacterial two-component systems. *Curr. Opin. Microbiol.* **76,** 102399 (2023).
38. Wang, X. et al. SRNA molecules participate in hyperosmotic stress response regulation in *sphingomonas melonis* TY. *Appl. Environ. Microbiol.* **90**, e0215823 (2024).
39. Yoon, S., Ha, S., Lim, J., Kwon, S. & Chun, J. A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie. Van. Leeuwenhoek*. 110, 1281-1286 (2017).
40. Meier-Kolthoff, J. P., Auch, A. F., Klenk, H. & Göker, M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14 (2013).
41. Madhaiyan, M., Saravanan, V. S., Wirth, J. S. & Whitman, W. B. Reclassification of *Sphingomonas aeria* as a later heterotypic synonym of *Sphingomonas carotinifaciens* based on whole-genome sequence analysis. *Int. J. Syst. Evol. Microbiol*. 70, 2355-2358 (2020).

42. Ali, A. et al. Biotransformation of benzoin by *Sphingomonas* sp. LK11 and ameliorative effects on growth of *Cucumis sativus*. *Arch. Microbiol*. 201, 591-601 (2019).
43. Chen, Y. et al. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* 7, 1-6 (2018).
44. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 25, 1043-1055 (2015).
45. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* 25, 1754-1760 (2009).
46. NCBI Sequence Read Archive http://identifiers.org/dbest:SRP648589 (2025)