scientific reports

Check for updates

OPEN

Whole genome of petroleum hydrocarbon degrading *Rhodococcus indonesiensis* isolated from Nacharam, Hyderabad, India

Syed Arshi Uz Zaman[1,2], Khushboo Sharma[3], Anuraj Nayarisseri[3]✉, Kamal A. Khazanehdari[2] & Rajabrata Bhuyan[1]✉

Petroleum pollution poses a critical environmental concern. Bioremediation has gained prominence as an eco-friendly approach for mitigating hydrocarbon pollution. This study reports the isolation and comprehensive characterization of a novel petroleum-degrading bacterium, *Rhodococcus indonesiensis* SARSHI1. Whole-genome sequencing was performed using a hybrid approach, integrating Oxford Nanopore Technologies (PromethION) and Illumina (NovaSeq 6000) platforms. The complete genome spans 5.7 Mbp and an additional plasmid of 159,118 bp, together encoding 5,150 coding sequences. Structural annotation identified 5220 genes, including 5094 protein-coding genes, one non-coding RNA, one CRISPR array, 56 pseudogenes, and 243 hypothetical proteins. Sequencing yielded 13,900,477 Illumina and 2,539,063 ONT reads, with 13,169,190 and 1,567,736 retained after quality processing, respectively. The assembly achieved 100% completeness with a coding density of 91.4%. Functional annotation revealed key hydrocarbon-degradation genes *alkB*, *ahyA*, and *almA* for long-chain alkanes, and *bph*, *ben*, and *xylC* for aromatics. Additionally, the presence of genes conferring multiple antibiotic resistances and those involved in secondary metabolite synthesis highlighted the strain's remarkable metabolic adaptability. The complete genome and plasmid sequences have been deposited in GenBank under accession numbers CP180630 and CP180631, respectively. The raw reads are available in the NCBI Sequence Read Archive (SRA) under accession numbers SRX27520007 (Illumina) and SRX27520006 (ONT).

Keywords Bioremediation, Petroleum hydrocarbon-degrading bacteria, Whole genome sequencing, Hybrid assembly, Oxford nanopore and illumina, Petroleum hydrocarbon-degrading genes

Petroleum-derived products are integral to modern economies and daily life. Global crude oil consumption has risen sharply since the nineteenth century, with projections estimating a demand of 121.5 million barrels per day (BPD) by 2050[1]. India, a major consumer, recorded 137.6 million metric tonnes of petroleum consumption between April and October 2023, reflecting a 3% year-on-year increase[2]. Despite their utility, petroleum extraction, refining, and transportation pose substantial environmental and health risks, primarily due to accidental spills and industrial emissions. In India, petroleum refineries represent major pollution sources, contributing substantially to greenhouse gas emissions and ecological degradation[3]. Persistent petroleum hydrocarbons (PHs) disrupt both aquatic and terrestrial ecosystems[4]. In aquatic habitats, PHs accumulate in sediments, interfere with trophic interactions, and biomagnify in organisms, compromising food webs[5,6]. In soils, high PH concentrations (> 20,000 mg/kg) reduce microbial diversity and impair biogeochemical cycles, whereas moderate contamination (4000–20,000 mg/kg) can enhance microbial coexistence[7]. Human exposure to PHs produces concentration- and duration-dependent toxic effects, ranging from dermatological and respiratory disorders to carcinogenicity and immunosuppression[8]. Notably, a study on Indian petroleum workers reported significant DNA damage even at low exposure levels, with benzene and its metabolites identified as major contributors[9].

[1]Department of Bioscience and Biotechnology, Banasthali Vidyapith, Rajasthan 304022, India. [2]Molecular Biology and Genomics Centre, Zabeel 2, 597, Dubai, UAE. [3]In Silico Research Laboratory, Eminent Biosciences, 91, Sector-A, Mahalakshmi Nagar, Indore, Madhya Pradesh 452010, India. ✉email: anuraj@eminentbio.com; rajabrata001@gmail.com

Conventional petroleum remediation strategies are often inefficient for large-scale spills and may cause secondary ecological disturbances[10,11]. Bioremediation, a sustainable alternative, exploits microbial metabolic pathways to degrade hydrocarbons into non-toxic by-products. Degradation efficiency is governed by environmental parameters such as temperature, pH, oxygen availability, nutrient concentration, and hydrocarbon bioavailability. Bacterial genera, owing to their metabolic versatility and adaptability, are particularly effective degraders, with members of *Pseudomonas*, *Bacillus*, *Burkholderia*, *Gordonia*, and *Rhodococcus* capable of utilizing hydrocarbons as their sole carbon and energy sources[12,13]. For instance, a study reported the isolation of *B. thuringiensis*, *B. pumilus*, and *R. hoagii* from the rhizosphere of *Panicum aquaticum* Poir in oil-contaminated soils[14]. Similarly, *Rhodococcus* strains from oil-contaminated Antarctic soils degrade hydrocarbons at sub-zero temperatures (− 2 °C)[15].

Rhodococcus species have garnered significant interest due to their metabolic plasticity and resilience under extreme environmental conditions, such as high salinity, low temperatures, and nutrient scarcity[16]. These bacteria exhibit an extraordinary ability to degrade a wide spectrum of petroleum hydrocarbons and xenobiotic compounds[17]. However, despite their recognized potential, genomic insights into the enzymatic pathways and regulatory networks governing hydrocarbon degradation remain limited. Recent advancements in whole-genome sequencing (WGS) have facilitated the identification of key genetic determinants and metabolic pathways involved in hydrocarbon degradation, enhancing bacterial applications for large-scale environmental remediation[18]. WGS enables comprehensive genomic analysis, pinpointing critical genetic elements and protein-coding sequences involved in petroleum biodegradation[19,20]. However, traditional next-generation sequencing (NGS), relying primarily on short-read sequencing, encounters limitations in resolving complex genomic architectures, particularly repetitive sequences, structural variations, and large insertions often present in bacterial genomes[21].

Hybrid sequencing, which integrates short-read and long-read sequencing technologies, effectively addresses these challenges. While short-read sequencing provides high precision in detecting small-scale genetic variations, long-read sequencing resolves large genomic rearrangements, repetitive regions, and complex structural variations[22]. This synergistic approach offers an unprecedentedly detailed view of bacterial genomes, facilitating the identification of novel genes, enzymatic pathways, and functional networks essential for hydrocarbon metabolism[23]. This approach provides insights into the metabolic networks and regulatory pathways that enable bacterial resilience in polluted environments[24].

In the present study, we have isolated a novel petroleum-degrading bacterial strain and performed an in-depth genomic characterization using hybrid sequencing (Fig. 1). By integrating advanced sequencing platforms, we achieved high-resolution insights into the strain's genetic framework, elucidating the metabolic pathways and enzymatic systems involved in hydrocarbon degradation. These findings provide a deeper understanding of the molecular mechanisms underlying the strain's potential application in environmental bioremediation strategies. Our research highlights the transformative role of genomic approaches in discovering novel bacterial resources for sustainable and eco-friendly remediation of petroleum-contaminated environments.

Methodology

Isolation of petroleum hydrocarbon-degrading bacteria

Oil-contaminated soil samples were collected from polluted sites near service and gas stations in Nacharam, Hyderabad, by excavating surface layers (10–12 inches). The collected samples were homogenized and stored at 4 °C before enrichment in nutrient agar (NA) containing crude oil. The soil samples were inoculated into minimal salt medium (MSM) supplemented with 2% (w/v) crude oil, while an additional flask containing nutrient media with crude oil served as a control. The flasks were incubated at 30 °C for 24 h, and bacterial growth was monitored by assessing turbidity. The cultures were serially diluted and spread onto MSM plates containing 2% (w/v) crude oil to isolate petroleum-degrading bacterial strains. The plates were incubated at 30 °C for 24–48 h[17,25]. Among the isolates, only SARSHI1 exhibited significant hydrocarbon degradation. The biochemical and microbial characterization of SARSHI1 was performed following Bergey's Manual of Determinative Bacteriology[26–29]. All experiments in the study were performed in triplicate, with results expressed as mean ± standard deviation (n = 3).

The biodegradation efficiency of SARSHI1 was further evaluated using crude oil concentrations of 5%, 10%, and 15% (w/v) over a seven-day incubation period. Following incubation, cultures were centrifuged at 3000–5000 rpm for 5 min to pellet bacterial cells. The supernatant was discarded, and the pellet was washed with 0.9% physiological saline. Bacterial growth dynamics were evaluated by measuring optical density at 600 nm (OD_{600}) at 24-h intervals[30–32].

Growth Kinetics and substrate utilization

Following crude oil degradation screening, the isolate was tested for its ability to utilize individual aliphatic and aromatic hydrocarbons as sole carbon sources to assess substrate specificity, with minor modifications to established methods[33,34]. A glycerol stock culture was revived in nutrient broth and incubated at 30 °C with shaking at 140 rpm for 24 h. The actively growing culture (1% v/v) was inoculated into MSM containing *n*-alkanes (C_{16}–C_{38}) and polycyclic aromatic hydrocarbons (naphthalene, phenanthrene, anthracene, and pyrene) at a final concentration of 500 mg L^{-1}. Although these concentrations exceed those naturally present in crude oil, they were chosen within the tolerance range reported for *Rhodococcus* sp.[34,35]. To improve dispersion and reduce optical interference, 0.01% (v/v) Tween 80 was added. Cultures were incubated at 30 °C with shaking at 140 rpm for seven days, and growth was monitored spectrophotometrically at 600 nm every 24 h.

While colony-forming unit (CFU) counts are widely used to evaluate the growth of hydrocarbon-degrading bacteria, OD_{600} measurements provide a reliable and reproducible estimate of biomass in cultures containing hydrophobic substrates. All hydrocarbons used were of analytical grade (≥ 99% purity). To examine degradation

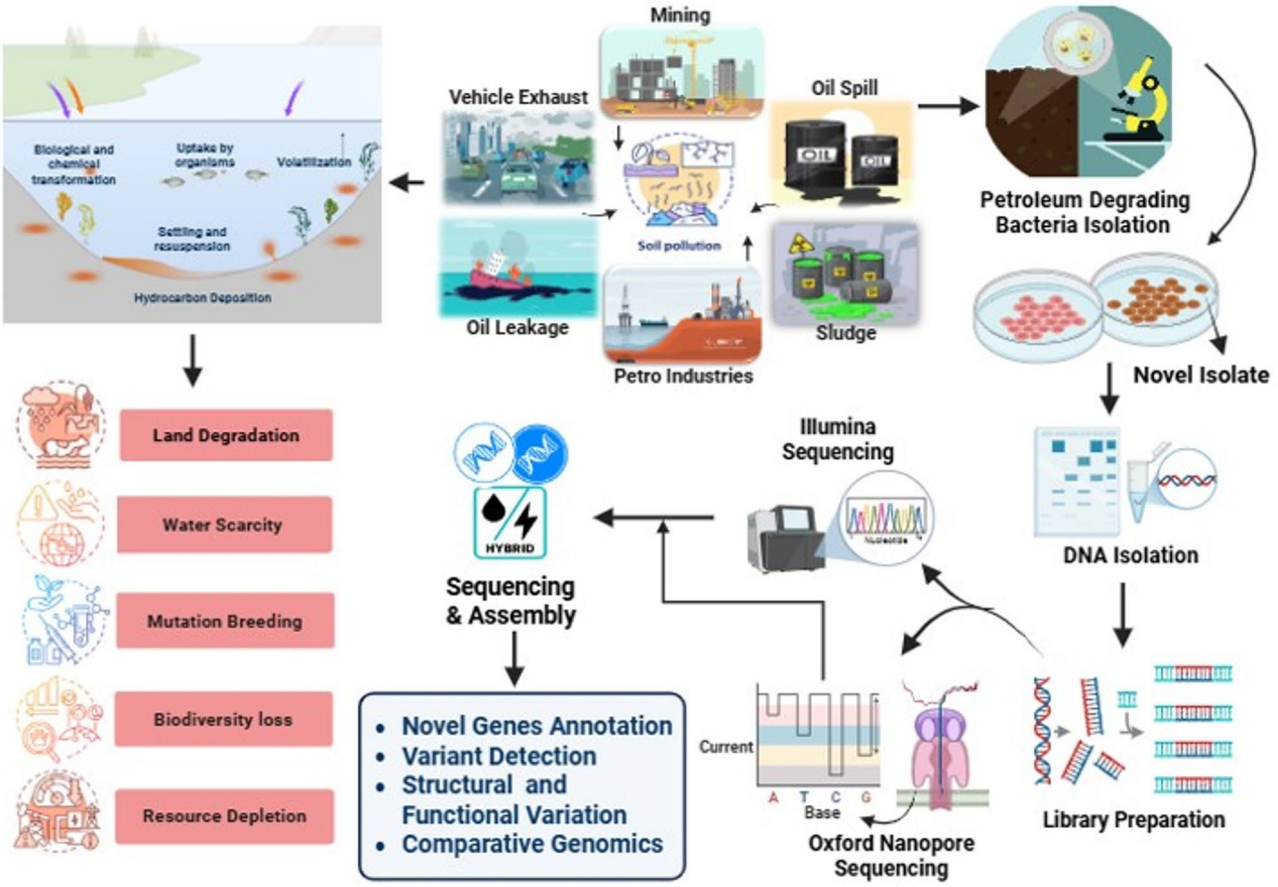


Fig. 1. Petroleum pollution: isolation and characterization of a novel petroleum-degrading bacterial strain using hybrid whole genome sequencing.

of high-molecular-weight (HMW) n-alkanes, cultures pre-grown for 24 h on a moderate-chain substrate (C_{24}) were sequentially transferred (1% v/v) into MSM containing progressively longer-chain alkanes (C_{28}–C_{38}) at 500 mg L^{-1}. Growth kinetics were recorded to evaluate changes in lag phase and biomass yield, and adapted cultures were then exposed to higher hydrocarbon concentrations (750–1000 mg L^{-1}) to assess tolerance. This sequential adaptation approach was designed to investigate potential acclimation responses of SARSHI1 to recalcitrant long-chain alkanes.

Enzyme assays

Dioxygenase activity

The activities of catechol 1,2-dioxygenase (C12O; EC 1.13.11.1) and catechol 2,3-dioxygenase (C23O; EC 1.13.11.2) in SARSHI1 were quantified to evaluate the induction of ring-cleavage pathways during aromatic hydrocarbon degradation, following the method described by Kumari et al.[36], with minor modifications. The strain was cultivated in MSM supplemented with naphthalene (0.5 g L^{-1}) as the sole carbon and energy source at 30 °C with shaking (150 rpm) for up to 15 days. Naphthalene served as the inducing compound, generating catechol intermediates via salicylate metabolism to stimulate dioxygenase expression. Samples were collected at defined time intervals. Cells were harvested by centrifugation (8000×*g*, 10 min, 4 °C), washed twice with 50 mM phosphate buffer (pH 7.0), and disrupted by sonication to obtain the cell-free crude extract, which served as the enzyme source.

The C12O assay was performed in a 1 mL reaction mixture containing 50 mM sodium phosphate buffer (pH 7.0), 1 mM EDTA, 0.1 mM catechol, and 0.02–0.06 mg of crude protein. The formation of *cis, cis*-muconic acid was monitored by measuring the increase in absorbance at 260 nm on a UV–visible spectrophotometer. Specific activity was calculated using the molar extinction coefficient (ε = 25,600 M^{-1} cm^{-1}) and expressed as µmol min^{-1} mg^{-1} protein. The C23O assay was performed similarly in 50 mM sodium phosphate buffer (pH 7.5) containing 0.1 mM catechol and 0.02–0.06 mg of crude protein. The formation of 2-hydroxymuconic semialdehyde was monitored at 375 nm, and enzyme activity was calculated using ε = 33,400 M^{-1} cm^{-1}, expressed as µmol min^{-1} mg^{-1} protein. Abiotic and biotic controls were included to ensure assay specificity[36,37].

Alkane hydroxylase (alkB) activity

To evaluate *alkB*-mediated terminal oxidation activity, the conversion of *n*-hexadecane to 1-hexadecanol was analyzed by gas chromatography (GC). The 300 µL of actively growing culture ($OD_{600} \approx 1.0$) was inoculated into sterile 100 mL glass tubes containing 10 mL of MSM supplemented with 10 mM *n*-hexadecane (≥ 99% purity, Sigma-Aldrich) as the sole carbon source. Uninoculated tubes containing the same medium and substrate served as abiotic controls. All cultures were incubated at 30 °C and 150 rpm, and samples were collected at 0, 24, 48, 72, and 96 h for extraction and GC analysis. After incubation, cultures were acidified to pH 2.0 using 1 N HCl and extracted twice with equal volumes of *n*-hexane. The combined organic layers were dried over anhydrous sodium sulfate and concentrated to 1 mL under a gentle nitrogen stream. Extracts were stored at − 20 °C until GC analysis.

Authentic standards of *n*-hexadecane and 1-hexadecanol (≥ 99% purity, Sigma-Aldrich) were processed under identical conditions for calibration. GC analysis was performed on a Hewlett–Packard 5890 Series II gas chromatograph equipped with a flame ionization detector (FID) and an HP-5 capillary column (30 m × 0.25 µm film thickness). Helium was used as the carrier gas at a constant flow rate of 1 mL min^{-1}. The oven temperature was programmed from 40 °C (5 min hold) to 280 °C at 10 °C min^{-1}, with a final hold of 10 min. The injector and detector were maintained at 260 °C and 280 °C, respectively. Injections (1 µL) were made in (splitless) mode. Formation of 1-hexadecanol was confirmed by matching retention times with standards and quantified using peak-area ratios normalized to an internal standard (*n*-tetradecane, 10 µM)[38,39].

Biodegradation of crude oil

To evaluate crude oil degradation by SARSHI1, 1 mL of actively growing culture ($OD_{600} \approx 1.0$) was inoculated into 50 mL of sterile MSM containing 5% (w/v) crude oil. The pH was adjusted to 7.0 before sterilization. Flasks were incubated at 30 °C with shaking at 150 rpm for 15 days, including uninoculated controls. Samples were collected every 24 h to measure residual crude oil. After each collection, cultures were centrifuged at 5000×*g* for 5 min to pellet cells. The supernatant (residual hydrocarbons) was extracted with an equal volume of *n*-hexane. The organic layer was separated, dried over anhydrous sodium sulfate, and concentrated to 1 mL under a gentle nitrogen stream. The extracts were filtered through 0.22 µm PTFE filters and stored at − 20 °C until GC–MS analysis.

Analysis was conducted using an Agilent 7890B gas chromatograph (Agilent Technologies, Santa Clara, CA, USA), equipped with an HP-5MS capillary column (30 m × 0.25 mm i.d., 0.25 µm film thickness). Helium served as the carrier gas at a steady flow rate of 1 mL/min. The oven temperature program was set from 50 °C (held for 2 min) to 310 °C at 10 °C per minute, with a final hold of 10 min. Injector and detector temperatures were maintained at 260 °C and 300 °C, respectively. The mass spectrometer operated in electron impact (EI) mode at 70 eV, scanning m/z 50–550. Residual hydrocarbons were quantified by comparing peak areas with authentic standards using standard calibration methods[28,40].

Genomic DNA extraction, qualitative and quantitative analysis

Genomic DNA was extracted from a pure bacterial culture of isolate SARSHI1 using the QIAamp DNA Mini Kit (Qiagen) following the manufacturer's protocol. The purity of the extracted DNA was assessed spectrophotometrically (NanoDrop One spectrophotometer, Thermo Fisher Scientific) by measuring absorbance ratios at A260/280 and A260/230. The values of ~ 1.8 and ~ 2.0, respectively, are considered indicative of high-quality DNA with minimal protein or organic contaminant interference. To further verify the integrity of the genomic DNA, 50 ng of the extracted DNA was resolved on a 1% agarose gel with a 1 kb ladder. The absence of smearing or degradation bands on the gel confirmed the integrity of the extracted genomic DNA, indicating its suitability for downstream analysis[28,41].

Molecular characterization using 16S rRNA sequencing

The isolated DNA was amplified using universal 16S primers, 5′-GGATGAGCCCGCGGCCTA-3′ (forward) and 5′-CGGTGTGTACAAGGCCCGG-3′ (reverse). PCR amplification was performed using a 50 µL reaction mixture in a thermal cycler, comprising 10X polymerase assay buffer, 1 µL of genomic DNA template, 1 µL of Taq DNA polymerase (3 U/mL), 4 µL of each dNTP (2.5 mM), 2 µL each of universal forward and reverse primers, 1X gel loading buffer, and nuclease-free water to a final volume of 50 µL[29]. The PCR conditions included an initial denaturation at 95 °C for 5 min, followed by 30 cycles comprising denaturation at 95 °C for 40 s, primer annealing at 65 °C for 1 min, and extension at 72 °C for 2 min, with a final extension at 65 °C for 1 min[42]. A single discrete PCR amplicon band was observed on agarose gel electrophoresis[43,44]. The amplified DNA fragments were purified using the GeneJet Gel Extraction PCR Purification Kit (https://www.thermofisher.com/order/catalog/product/K0691)[45]. The amplicon was sequenced using Sanger dideoxy sequencing (SeqStudio Genetic Analyzer, Applied Biosystems, LeGene Biosciences Pvt Ltd, Indore, India)[45,46]. The forward and reverse chromatogram files were assembled and analyzed using DNA Baser v5.15[47]. The sequence nucleotide composition, molecular weight, and GC content were determined using EMBOSS software[48–50].

Phylogenetic analysis and rRNA structure prediction

The sequence was analyzed using BLASTN against the 16S ribosomal RNA database (Bacteria and Archaea), applying a similarity threshold of > 95%. The top 20 sequences showing the highest similarity to the query were retrieved in a FASTA format for the phylogenetic analysis[51,52]. These sequences were imported into MEGA XI and aligned using the MUSCLE algorithm with up to 16 iterations. Phylogenetic relationships were initially inferred using the UPGMA clustering method. The alignment was then exported in MEGA format, and a phylogenetic tree was constructed using the neighbour-joining method with 1000 bootstrap replicates. The evolutionary relationships were evaluated using the Maximum Composite Likelihood model[53–55]. The rRNA secondary

structure was predicted using UNAfold, while the minimum free energy (MFE) conformation was determined with the Mfold server, providing insights into the thermodynamic stability of the predicted structures[56–58].

Preparation of 2X150 WGS library

The paired-end sequencing library was prepared from the extracted DNA using the Illumina TruSeq Nano DNA Library Prep Kit (TruSeq, DNA Library Prep Kits, 2012). 100 ng of DNA was fragmented using the Covaris M220 system, resulting in an average fragment size of 350 bp. Covaris shearing generates double-stranded DNA fragments with 3′ or 5′ overhangs, which are end-repaired to produce blunt ends. The repaired DNA fragments were ligated with adapters, followed by size selection using AMPure XP beads (Beckman Coulter, Inc.) to ensure uniformity. The size-selected fragments were PCR-amplified with an index-specific primer kit to incorporate indexing adapters. These adapters facilitated the hybridization of DNA fragments onto the flow cell for sequencing[28]. The PCR-enriched library was assessed for quality and fragment distribution using the Agilent 4200 TapeStation system with High Sensitivity D1000 Screen Tape, following the manufacturer's protocol[28,59,60].

Cluster generation and sequencing

The paired-end (PE) Illumina sequencing library was quantified using Qubit, evaluated for fragment size with an Agilent TapeStation, and subsequently loaded onto the NovaSeq 6000 for cluster generation and high-throughput sequencing. PE sequencing captures DNA fragments in both directions, enhancing accuracy and facilitating the detection of structural variants and repeats. Cluster generation involves hybridizing the prepared library to adapter-bound oligos on the flow cell, enabling selective forward-strand cleavage after reverse-strand synthesis. This process ensures bidirectional sequencing with high-quality data and improved genome assembly[28,60,61].

Preparation of nanopore library

The Nanopore sequencing library was prepared from the quality-controlled genomic DNA sample using the Ligation Sequencing gDNA Native Barcoding Kit. Initially, 800 ng of the DNA sample was subjected to DNA repair, followed by purification using AMPureXP beads (Beckman Coulter, Inc.) and adaptor ligation. The samples were pooled according to the kit protocol. The pooled library was further purified using AMPureXP beads (Beckman Coulter, Inc.) to maintain high-quality preparation. The concentration of the purified pooled library was quantified using a Qubit fluorometer. Finally, the purified library was loaded onto an Oxford Nanopore PromethION P2 Solo flow cell for sequencing [Kit:SQK-LSK114][62]. Nanopore sequencing enables real-time selective sequencing of single DNA molecules by dynamically reversing the voltage across individual nanopores.

Genome assembly and quality control

The Raw sequencing reads from Illumina and Oxford Nanopore platforms underwent rigorous quality control and preprocessing. Quality control of Illumina paired-end reads was performed using FastQC (v0.11.9) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), with reads achieving a Phred quality score > 30 prioritized for downstream analysis[63]. The Oxford Nanopore reads were evaluated by read length distribution and per-base quality scores. The low-quality bases were trimmed, and adapters were removed using Trimmomatic v0.39 (https://github.com/timflutre/trimmomatic)[64]. A sliding window approach (window size = 4, minimum average quality score = 20) was applied to refine read quality. The bases at the ends of reads with quality scores below 3 were removed, and reads shorter than 36 bp were discarded to ensure high-quality data. The error correction was performed using SPAdes v3.15(https://ablab.github.io/spades/), employing a k-mer-based strategy (k = 21–51)[65]. The Oxford nanopore reads were filtered using Filtlong (https://github.com/rrwick/Filtlong), retaining only reads exceeding 1,000 bp with a mean quality score of at least 10[66], followed by adapter removal using Porechop (https://github.com/rrwick/Porechop)[67].

Hybrid genome assembly was performed using Unicycler (https://github.com/rrwick/Unicycler) in bold mode (–mode bold) to maximize scaffold continuity. Short-read error correction was performed with SPAdes (k = 21, 33, 55), while long reads were assembled using miniasm, followed by two rounds of polishing with Racon. A depth filter (-depth_filter 2.5 × median depth) was applied to exclude low-confidence contigs, and sequences shorter than 1000 bp were removed to retain only high-quality contigs. Structural variations and repeat regions were resolved through automatic bridging (-bridging auto), leveraging long reads to enhance assembly completeness[68]. Genome assembly quality was evaluated using QUAST v5.3 (https://github.com/ablab/quast), considering key metrics such as N50, total genome size, and misassembly rates[69]. Contiguity was assessed based on total assembly length and N50, while genome completeness was determined from the genome fraction corresponding to reference-mapped base coverage. Additionally, genome circularity and structural coherence were examined using Bandage (https://github.com/rrwick/Bandage) by employing hierarchical layouts to detect potential assembly discrepancies[70].

Functional annotation

Comprehensive genome annotation was performed using integrated bioinformatics tools to ensure structural and functional characterization. The genome was initially annotated using BAKTA v1.10.3 (https://github.com/oschwengers/bakta)[71], followed by NCBI Prokaryotic Genome Annotation Pipeline (PGAP) (https://github.com/ncbi/pgap) and Blast2GO (https://www.blast2go.com/). PGAP annotation was performed using the best-placed reference methodology, while gene prediction was conducted with GeneMarkS-2 + v6.9 (https://github.com/karlgem/GeneMarkS-2), (https://genemark.bme.gatech.edu/). eggNOG mapper (https://github.com/eggnogdb/eggnog-mapper) was utilized for functional annotation, incorporating orthology-based assignments, gene ontology (GO) terms, KEGG pathway mapping, and Clusters of Orthologous Groups (COG) classifications[72]. Genome assembly completeness and contamination were assessed using CheckM v1.0.2 (htt

ps://github.com/Ecogenomics/CheckM) and CheckM2 (https://github.com/chklovski/CheckM2)[73,74]. CheckM asses genome integrity using lineage-specific marker genes, while CheckM2 uses machine-learning models for improved accuracy, combining gene identification, DIAMOND-based annotation, and predictions of genome completeness and contamination.

Pathway analysis of annotated genes was performed using Blast2GO to elucidate their putative biological functions. Metabolic profiling was conducted with DRAM (Distilled and Refined Annotation of Metabolism) (https://github.com/WrightonLabCSU/DRAM) using default parameters[75]. Protein domains and conserved motifs were identified via InterProScan (https://github.com/ebi-pf-team/interproscan). Assembly and annotation accuracy were evaluated using BUSCO v5 (https://github.com/metashot/busco) against the *actinobacteria*_phylum_odb10 dataset, which comprises 292 conserved orthologs from 893 *Actinobacteria* genomes. Genome integrity was quantified in (-genome mode) based on the proportions of complete, fragmented, and missing genes[76].

Comparative genomics and taxonomic analysis

Average Nucleotide Identity (ANI) analysis was performed using FastANI (https://github.com/ParBLiSS/FastANI) with default parameters[77]. The SARSHI1 genome was compared bidirectionally with *Rhodococcus indonesiensis* CSLK01-03. FastANI fragment-based k-mer mapping strategy enables efficient and accurate ANI estimation, yielding key metrics based on fragment matches and total comparisons. To further refine taxonomic classification, digital DNA-DNA hybridization (dDDH) analysis was performed against the closest reference genome (GCA_030360185.1) using Genome-to-Genome Distance Calculator (GGDC 3.0) (https://ggdc.dsmz.de/). The analysis leverages three distinct formulas: length-normalized intergenomic distance**,** high-scoring segment pair-based alignment**,** anda sum-of-bits approach to calculate distance. Bootstrapping was applied, and species delineation was performed against the 70% dDDH threshold, a critical microbial taxonomic criterion to determine confidence intervals[78,79].

Whole-genome-based taxonomic classification was performed using the Type (Strain) Genome Server (TYGS) (https://tygs.dsmz.de/). The ten closest type (strains)were selected using the MASH distance algorithm and 16S rRNA gene-based comparisons**.** The intergenomic distances were computed under the trimming algorithm with 100 replicates, using the Genome BLAST Distance Phylogeny (GBDP) approach. The dDDH values were determined, followed by phylogenomic tree construction using the FASTME algorithm with subtree pruning and regrafting (SPR). Phylogenetic trees were visualized using PhyD3, with species clustering based on the 70% dDDH threshold and subspecies delineation at 79% dDDH[80].

Annotation of hydrocarbon degradation genes

The hydrocarbon degradation genes were annotated using Calgary approach to ANnoTating HYDrocarbon genes (CANT-HYD) (https://github.com/dgittins/CANT-HYD-HydrocarbonBiodegradation) analysis. The genomic sequences were processed with HMMER 3.4 (http://hmmer.org/), employing the *hmmscan* algorithm to search against the curated HMM database, using an E-value threshold of < 1e−5. Hidden Markov Models (HMMs) were derived from multiple sequence alignments (MSAs) of known enzymes, including alkane 1-monooxygenase (*alkB*), nitrate and sulfite reductases (*AhyA*), and flavin-containing monooxygenases (*AlmA*_GroupI). The KEGG and UniProt databases were used to map the identified genes to aerobic hydrocarbon degradation pathways[81].

The HADEG (Hydrocarbon Aerobic Degradation Enzymes and Genes) database was used to identify aerobic hydrocarbon degradation genes. The alignment was performed in BLASTp mode using DIAMOND v2.0 (https://github.com/bbuchfink/diamond) against the HADEG database. The orthologous gene clusters were identified and classified using Proteinortho6 (https://gitlab.com/paulklemm_PHD/proteinortho)[82]. The hydrocarbon monooxygenase genes were annotated using DIAMOND-based sequence searches against the Hydrocarbon Monooxygenase Gene Database (HMDB)[83]. A homology-based functional annotation was performed using DIAMOND with an E-value threshold of ≤ 1e−5 and a minimum identity threshold of 25% to ensure high-confidence hits. Further, the results were validated based on sequence conservation, domain architecture, and functional relevance with well-characterized hydrocarbon monooxygenases and rubredoxin electron transfer systems.

The antiSMASH (Antibiotics & Secondary Metabolite Analysis Shell) (https://github.com/antismash) analysis (enabling *ClusterBlast*) was performed to identify Biosynthetic gene clusters (BGCs)[84]. The *Procluster Region* Analysis approach was used with (*relaxed* strictness) to ensure the comprehensive detection and annotation of secondary metabolite gene clusters.

Results

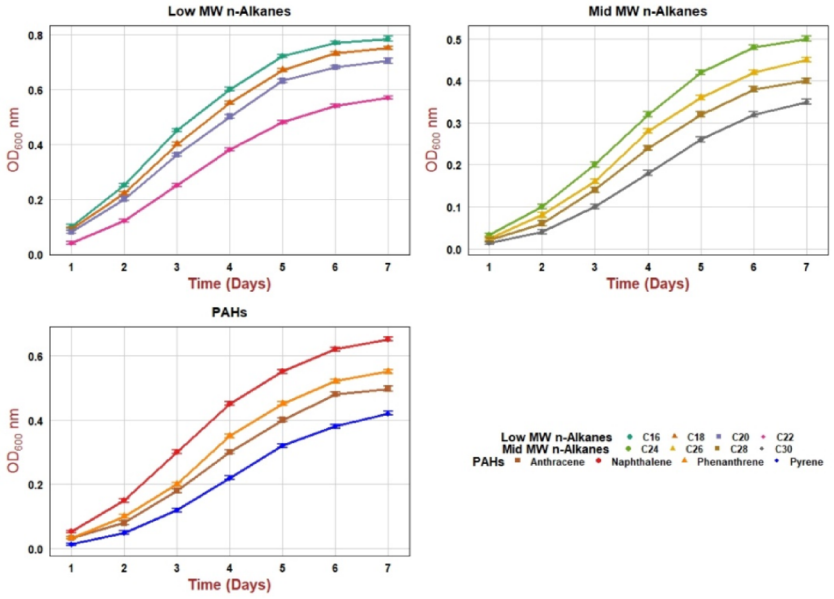
Isolation and screening of bacteria

SARSHI1 was characterized as Gram-positive, oxidase-negative, and non-motile, with extracellular proteolytic activity. Biochemical assays indicated negative tests for indole and methyl red, and a positive Voges-Proskauer, suggesting the absence of mixed-acid fermentation and the production of neutral end products from glucose (Fig. S1, Supplementary Information 1)[29]. Growth dynamics demonstrated that the strain could utilize crude oil as the sole carbon source, with a pronounced lag phase at 15% (w/v) crude oil, reflecting delayed adaptation to elevated hydrocarbon levels (Fig. S2, Supplementary Information 1).

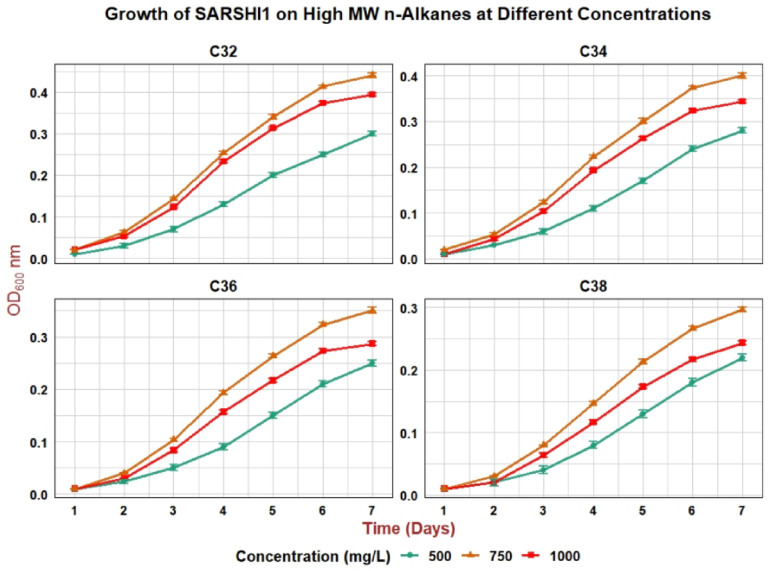
Growth kinetics on aliphatic and aromatic hydrocarbons

SARSHI1 displayed substrate-dependent growth on even-numbered *n*-alkanes (C_{16}–C_{30}) and polycyclic aromatic hydrocarbons (PAH) (naphthalene, phenanthrene, anthracene, pyrene) (Fig. 2A). The strain demonstrated rapid growth on short-chain alkanes (C_{16}–C_{20}), moderate growth on medium-chain alkanes (C_{22}–

C_{30}), suggesting a gradual decline in assimilation efficiency with increasing carbon chain length. Among the PAHs tested, naphthalene and phenanthrene supported moderate biomass accumulation (OD_{600} = 0.55–0.66), whereas anthracene and pyrene were less effectively metabolized (OD_{600} = 0.41–0.51). A characteristic lag phase was observed during the first 24 h across all treatments, followed by an exponential growth phase between days 2 and 5 and stabilization by day 7.



A: Growth Kinetics of SARSHI1 on different substrates.



B: Growth on High molecular weight *n*-alkanes.

Fig. 2. Growth kinetics of *Rhodococcus indonesiensis* SARSHI1 on aliphatic and aromatic hydrocarbons. Data represent the mean ± standard deviation of replicates (n = 3). (**A**) Growth on individual even-numbered *n*-alkanes (C_{16}–C_{30}) and polycyclic aromatic hydrocarbons (naphthalene, phenanthrene, anthracene, pyrene) was monitored as OD_{600} over a 7-day incubation period. (**B**) Growth profiles on high-molecular-weight *n*-alkanes (C_{32}–C_{38}) at different substrate concentrations (500, 750, and 1000 mg L^{-1}). Adaptation to 750 mg L^{-1} enhanced growth, whereas further increase to 1000 mg L^{-1} did not proportionally elevate OD_{600}, suggesting substrate-associated limitations.

Sequential adaptation significantly improved the growth of SARSHI1 on C_{32}–C_{38} *n*-alkanes (Fig. 2B). Cultures initially exposed to 500 mg L^{-1} showed modest growth (OD_{600} = 0.21–0.30). When adapted to 750 mg L^{-1}, OD_{600} values increased by about 20–30% (0.34–0.45), indicating better hydrocarbon utilization and tolerance. However, increasing the concentration to 1000 mg L^{-1} did not yield a proportional increase in growth. In several cases, OD_{600} values plateaued (0.28–0.40), indicating possible substrate inhibition or diffusion limitations at higher concentrations.

Enzyme assay and GC–MS (Gas chromatography–mass spectrometry) analysis

The activities of catechol 1,2-dioxygenase (C12O) and catechol 2,3-dioxygenase (C23O) for SARSHI1 were monitored over 15 days, alongside abiotic and biotic controls. C12O activity increased steadily, reaching a peak of approximately 4500 µmol min^{-1} mg^{-1} protein on day 10, before declining slightly to 3600 µmol min^{-1} mg^{-1} protein by day 15. Similarly, C23O activity peaked at 2100 µmol min^{-1} mg^{-1} protein on day 10, decreasing to 1700 µmol min^{-1} mg^{-1} protein by day 15 (Fig. 3A). Activities in controls remained negligible (< 65 µmol min^{-1} mg^{-1}), confirming the specificity of enzyme induction by naphthalene. Overall, C12O activity was consistently higher than C23O, indicating a predominance of the ortho-cleavage pathway during catechol metabolism.

GC–MS analysis of culture extracts following incubation with *n*-hexadecane showed a prominent peak corresponding to *n*-hexadecane ($C_{16}H_{34}$) and a distinct secondary peak representing hexadecanol ($C_{16}H_{34}O$) (Fig. 3B). The emergence of this oxidized product confirms the oxidation of *n*-hexadecane to its corresponding alcohol, indicating active alkane oxidation in *Rhodococcus indonesiensis* SARSHI1. This transformation reflects the catalytic activity of alkane hydroxylase (*alkB*), which mediates the terminal hydroxylation of *n*-alkanes. The relative increase in the alcohol peak and the concomitant decrease in the substrate peak area further verify the functional expression of the *alkB* gene. This terminal oxidation represents the initial step of aerobic alkane degradation, corroborating the enzymatic role predicted from genomic analysis.

The GC–MS profile provides direct evidence of *Rhodococcus indonesiensis* SARSHI1's metabolic potential to degrade complex hydrocarbon mixtures (Fig. 4). In the control chromatogram, a homologous series of n-alkanes from C_{10} to C_{38} was clearly visible, along with distinct peaks corresponding to the isoprenoid hydrocarbons pristane (Pr) and phytane (Ph). After incubation with *R. indonesiensis* SARSHI1, a remarkable reduction in total hydrocarbon abundance was observed. The complete disappearance of short-chain alkanes ($\leq C_{15}$) and a significant decline in mid-chain components (C_{16}–C_{21}) indicate their rapid utilization as primary carbon sources. The partial depletion of long-chain alkanes (C_{23}–C_{30}) suggests the activation of multiple alkane-oxidizing enzymes, which collectively broaden the strain's substrate range. Moreover, the substantial reduction of pristane and phytane (> 70%) reflects the strain's capability to oxidize branched hydrocarbons. A minor fraction of very long-chain hydrocarbons ($\geq C_{33}$) persisted after degradation, likely due to diffusion limitations or restricted enzyme accessibility to high-molecular-weight substrates.

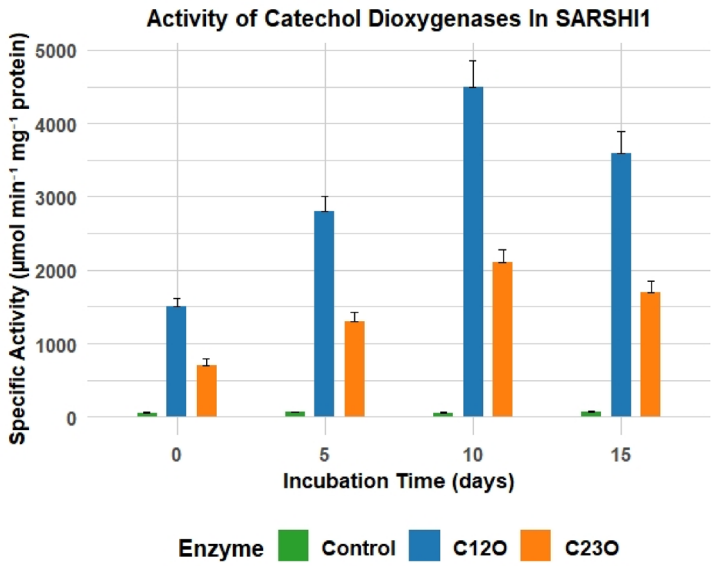
Quantitative and qualitative evaluation of extracted DNA

The quality of extracted DNA was evaluated by agarose gel electrophoresis. As shown in Fig. S3 (Supplementary Information 1), Lane M (1 kb DNA ladder), while Lane 1 displays a single, high-molecular-weight DNA band without visible smearing, confirming the integrity and purity of the genomic DNA. The absence of degraded fragments indicates minimal DNA fragmentation and contamination. Spectrophotometric quantification using NanoDrop revealed a DNA concentration of 54.2 ng/µL. The $A_{260/}A_{280}$ratio of 1.83 falls within the standard range for pure DNA, suggesting negligible protein contamination. The $A_{260/230}$ ratio of 1.71 indicates reasonable purity, although values approaching 2.0 are generally indicative of minimal interference from organic impurities such as phenol or other solvents. Overall, the sample met the quality control (QC) criteria for downstream molecular analyses.

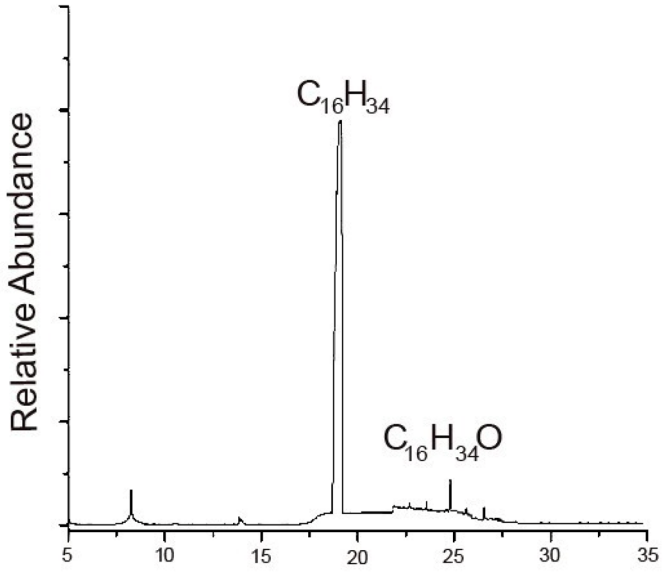
Molecular characterization by ribotyping

The forward and reverse ABI trace files generated by Sanger sequencing were assembled using DNA Baser and exported in FASTA format. Sequence identity was determined using NCBI BLAST (Basic Local Alignment Search Tool) with an e-value threshold of 0.0 and a coverage cut-off of 200%. Based on morphological, biochemical, and molecular analyses, the strain SARSHI1 was identified as a *Rhodococcus indonesiensis*. The 16S rRNA sequence of *Rhodococcus indonesiensis* SARSHI1 was deposited in the NCBI GenBank database under the accession number PV034287. The sequence exhibited a GC content of 60.69%, a molecular weight of 439,033 Daltons, and a total length of 720 bp. Comparative analysis showed that strain SARSHI1 shared 98.47% identity with partial 16S rRNA sequences of *Rhodococcus ruber* strains JC435 and DSM 43,338, and 98.33% similarity with *Rhodococcus aetherivorans* strains DSM 44,752 and 10bc312-the latter isolated from a petrochemical sludge bioreactor. The strain also exhibited 97.22% identity with *Rhodococcus nanhaiensis* strain SCSIO10187.

The phylogenetic tree (Fig. S4, Supplementary Information 1) was constructed using the Neighbour-Joining with the Maximum Composite Likelihood model. Evolutionary parameters were estimated under the Kimura 2-parameter model, while substitution patterns were derived according to the Tamura-Nei model. Summary statistics and maximum likelihood (ML) scores for the tree topology are provided in Table S1 (Supplementary Information 1). The average disparity index (0.946) and average composition distance (1.048) indicated moderate sequence divergence, with a mean pairwise distance of 4.31. A detailed structural analysis of the rRNA sequence was conducted to identify conserved functional motifs essential for ribosomal activity. Secondary structure prediction of the 16S rRNA sequence was performed to identify conserved functional motifs critical for ribosomal activity. Folding analysis using UNAFOLD (window size = 12; temperature = 37 °C; ionic concentration = 1 M NaCl) yielded a predicted free energy of $\Delta G = -256.00$ kcal mol^{-1} (Fig. S5, Supplementary Information 1). Functional annotations highlighted conserved domains, and thermodynamic properties were



A: Activity of Catechol dioxygenases



B: GC Chromatogram showing Alkane oxidation.

Fig. 3. Enzymatic activities and alkane oxidation by *Rhodococcus indonesiensis* SARSHI1. (**A**) Activities of catechol 1,2-dioxygenase (C12O; blue) and catechol 2,3-dioxygenase (C23O; orange) were monitored in cell-free crude extracts at 0, 5, 10, and 15 days. Abiotic and biotic controls (green) exhibited negligible activity, confirming substrate-induced enzyme expression. Data represent mean specific activities (µmol min^{-1} mg^{-1} protein) from three independent replicates, with error bars indicating standard deviations. (**B**) GC–MS chromatogram showing alkane oxidation. The prominent peak at approximately 20 min corresponds to the parent alkane (*n*-hexadecane, $C_{16}H_{34}$), while a secondary peak at ~25 min represents the corresponding primary alcohol (hexadecanol, $C_{16}H_{34}O$). The formation of hexadecanol indicates terminal oxidation of the alkane, consistent with alkane hydroxylase (*alkB*) activity.

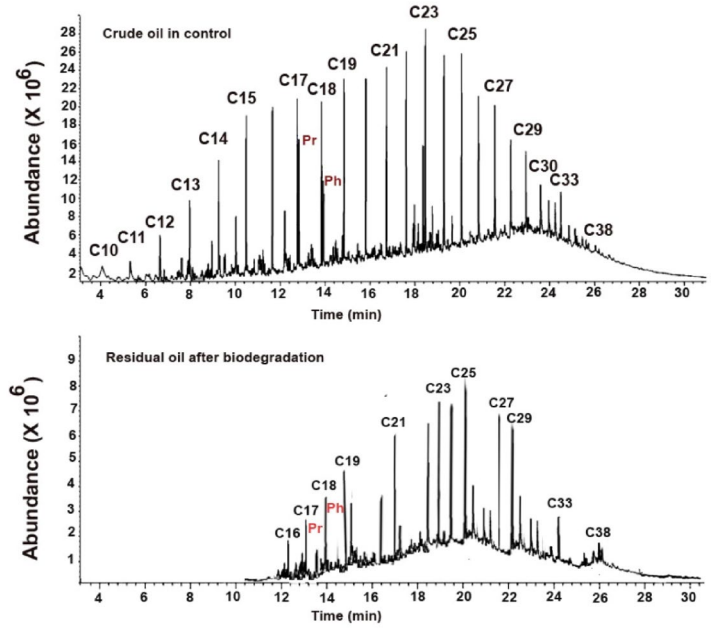
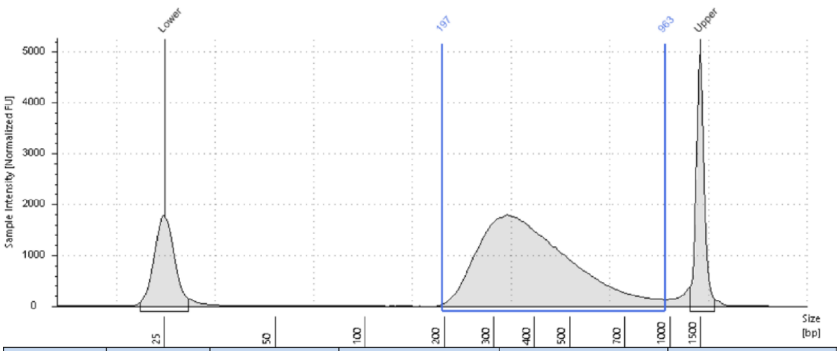


Fig. 4. GC–MS chromatograms illustrating crude oil degradation by *Rhodococcus indonesiensis* SARSHI1. The control sample exhibits a typical unimodal distribution of n-alkanes, predominantly between C_{23} and C_{25}, accompanied by isoprenoids, specifically pristane (Pr) and phytane (Ph). The residual sample (Day 15) reveals a significant depletion of low- and mid-chain n-alkanes, along with a pronounced reduction in Pr and Ph peaks. The partial retention of high-molecular-weight hydrocarbons ($\geq C_{25}$) suggests selective and progressive oxidation of hydrocarbon components during biodegradation.



From [bp]	To [bp]	Average Size [bp]	Conc. [ng/µl]	Region Molarity [nmol/l]	% of Total
197	963	409	26.7	110	95.85

Fig. 5. Library Profile of sample SARSHI1 on Agilent TapeStation.

summarized in Table S2 (Supplementary Information 1). Entropy analysis using RNAfold is illustrated in a hill plot (Fig. S6, Supplementary Information 1).

Library quality assessment and analytics of sequencing data

The electropherogram shown in Fig. 5 illustrates the fragment size distribution of a nucleic acid sample analyzed using TapeStation. The sample displayed a primary fragment size range of 197–963 bp, with an average fragment length of 409 bp. The nucleic acid concentration was 26.7 ng μL^{-1}, and the region molarity of 110 nmol L^{-1} indicated sufficient yield for downstream applications. The detected fragment population accounted for 95.85% of the total signal, reflecting minimal degradation or contamination. The presence of well-defined lower and upper markers further confirmed the high integrity, supporting its suitability for sequencing.

Adapter sequences containing more than 5% ambiguous nucleotides ("N"), and low-quality reads with over 10% of bases having a Phred quality score below 25 were systematically removed. A stringent quality control was used, with sliding window trimming of 10 bp, wherein bases with average quality within the window fell below (Phred < 25) were discarded. Additionally, the leading and trailing bases with quality scores below 25 were

removed to ensure sequence accuracy and integrity. Post-trimming, only reads exceeding 100 nucleotides in length were retained, resulting in approximately 12.89 million high-quality reads. Oxford Nanopore Technology (ONT) sequencing achieved an estimated genome coverage of ~ 1400 ×, reflecting the depth and reliability. Read statistics from both Illumina and ONT platforms are summarized in Table 1. The read lengths ranged from 68 to 936,949 bp, demonstrating the Nanopore platform's ability to generate both short and ultra-long reads.

Hybrid de-novo genome assembly

Hybrid genome assembly combining high-quality Illumina paired-end and ONT reads was performed using Unicycler v0.48 (Fig. 6). The assembly successfully reconstructed one chromosome and one plasmid, achieving an N50 value of 5,536,171 bp. Assembly quality and robustness were validated using QUAST and Bandage, with comprehensive metrics summarized in Table S1 (Supplementary Information 2). The final assembly comprised only two contigs, reflecting minimal fragmentation and high structural continuity. The total genome length was 5,695,289 bp, with the largest contig spanning 5,536,171 bp, suggesting that the majority of the genome was assembled into a single dominant contig. The N50 and N90 values of 5,536,171 bp further confirmed the assembly's contiguity and integrity. The GC content (70.08%) was consistent with known bacterial genomes, supporting the assembly's reliability. Moreover, the L50 and L90 values of 1 indicated that a single contig contained at least 50% and 90% of the genome.

The absence of Ns per 100 kbp (0.00) confirmed a gap-free assembly, while the auN value of 5,385,944 pinpoints genome completeness and continuity (Fig. S1 Supplementary Information 2). The raw sequencing data for SARSHI1 *Rhodococcus indonesiensis* have been deposited in the Sequence Read Archive (SRA) under Accession **SRX27520007** (Illumina data) and Accession **SRX27520006** (ONT data), as part of BioProject **PRJNA1217105** and BioSample **SAMN46479200.** The genome assembly visualization using Bandage revealed a nearly circular contig representing the complete genome, and a small circular structure corresponding to the plasmid (Fig. S2, Supplementary Information 2). A minor discontinuity was observed, likely associated with an unresolved repeat region. The polygonal and slightly irregular shape of the contig, rather than a perfectly smooth circle, likely reflects local variations in sequencing depth or unresolved repetitive elements.

Genome annotation and functional characterization

The circular genomic map of *Rhodococcus indonesiensis* SARSHI1 (5.7 Mbp), illustrated in Fig. 7, provides a comprehensive overview of its structural and functional organization. The circular representation facilitates the interpretation of gene distribution, replication dynamics, and potential evolutionary characteristics. The outermost ring displays protein-coding sequences (CDS: grey), along with key genetic elements including transfer RNAs (tRNA: green), ribosomal RNAs (rRNA: red), and non-coding RNAs (ncRNA: shades of blue). The presence of CRISPR loci (dark green) indicates active defense mechanisms against phage and plasmid invasion, reflecting adaptive immunity. The second ring depicts GC content variation, highlighting regions potentially associated with high transcriptional activity or horizontal gene transfer. The innermost rings represent GC skew, where positive skew (green) corresponds to the leading strand and negative skew (red) to the lagging strand, reflecting the organism's replication asymmetry and directional bias.

The complete genome of *Rhodococcus indonesiensis* strain SARSHI1 was annotated using the Prokaryotic Genome Annotation Pipeline (PGAP), revealing two contigs corresponding to a chromosome and a plasmid. The complete genome sequence has been deposited in the GenBank database under accession numbers CP180630 (complete genome) and CP180631 (plasmid). A total of 5220 genes were identified, including 5150 coding sequences, of which 5,094 were protein-coding. Additionally, 70 RNA genes were annotated, comprising 12 rRNAs (four copies each of 5S, 16S, and 23S), 55 tRNAs, and one ncRNA. A total of 56 pseudogenes were detected, characterized by frameshifts (28), incomplete sequences (38), internal stop codons (4), or multiple disruptions (9), with no ambiguous residues. Functional annotation revealed genes involved in hydrocarbon metabolism, as well as those associated with antibiotic biosynthesis, antimicrobial resistance, and plasmid-maintenance systems (pRiA4b ORF-3 family and related replication-associated genes).

Annotation using Bakta

The complete genome of *Rhodococcus indonesiensis* strain SARSHI1 comprises a single chromosomal replication origin (oriC) and lacks plasmid replication (oriV) or transfer (oriT) origins, indicating the absence of conjugative elements. The high GC content (~ 70.08%) aligns with its classification within *Actinobacteria.* The chromosome encodes a robust translational and regulatory network, including 18 ncRNA regions and a CRISPR array, suggesting adaptive immunity against phages or foreign DNA invasion. A total of 5,169 coding sequences (CDSs) were identified, including 243 hypothetical proteins and a predicted small open reading frame (sORF) potentially involved in regulatory or metabolic processes. Most CDSs exhibit high sequence identity,

	Raw reads		Processed reads		
Sample name	No. of reads	Total no. of bases	HQ reads	Total no. of bases	Data in GB
Illumina PE short reads					
SARSHI1	13,900, 477	4,182,022,022	13,169,190	3,904,771,108	3.9
ONT reads					
SARSHI1	2,539,063	4,826,079,667	1,567,736	4,112,035,435	4.1

Table1. Read statistics.

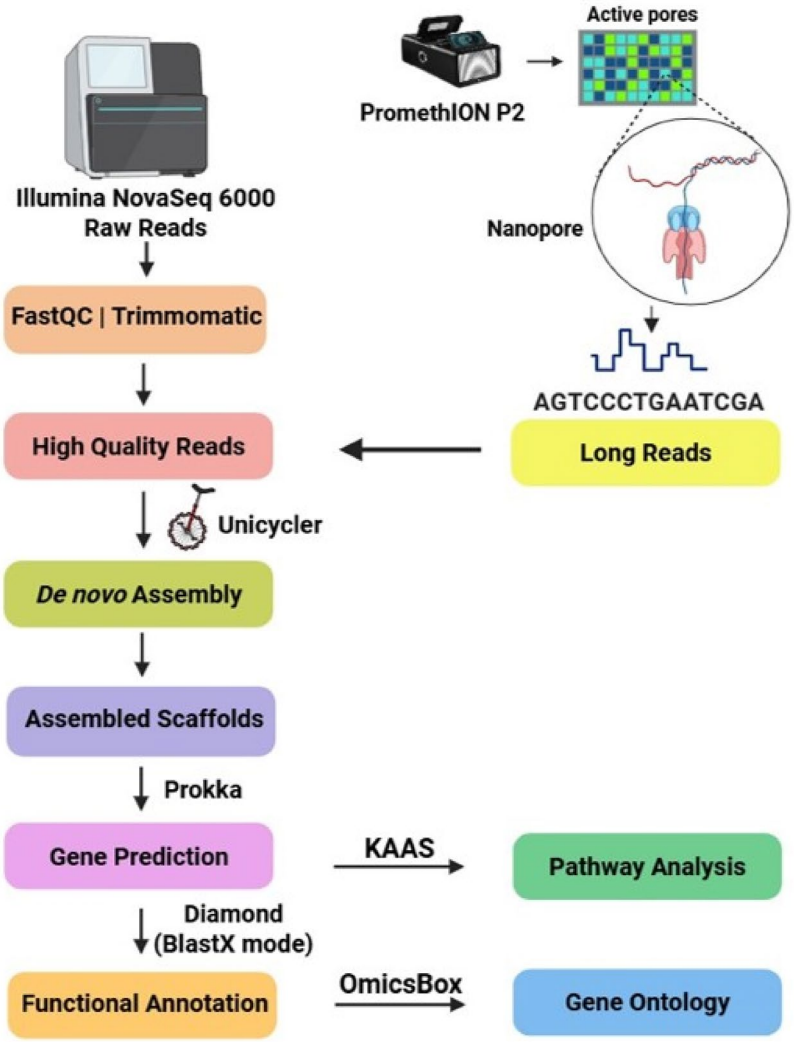


Fig. 6. Schematic representation of the Genome Sequencing and Annotation Workflow. Raw reads were generated using Illumina NovaSeq 6000 and Oxford Nanopore PromethION P2 platforms, quality-checked, and processed to obtain high-quality reads. De novo assembly was performed with Unicycler, followed by scaffold assembly, gene prediction (Prokka), and functional annotation using Diamond (BlastX mode), KAAS for pathway analysis.

low e-values, and high bit scores, whereas a subset remains uncharacterized, suggesting the presence of novel or hypothetical genes (Supplementary Information 3). Functional annotation using eggNOG-mapper revealed genes encoding metal-dependent hydrolases, ATP-binding proteins, and hydrocarbon-degrading enzymes, such as alkane hydroxylases, monooxygenases, aromatic hydroxylases, and ring-cleaving dioxygenases, indicating the potential for complete mineralization of aliphatic and aromatic hydrocarbons. Additionally, genes associated with oxidative stress response and membrane transporters suggest adaptive mechanisms for survival under hydrocarbon-rich environments (Supplementary Information 4).

The second assembled contig (contig_2) represents a plasmid comprising 134 CDSs and 2 pseudogenes. The absence of tRNAs, rRNAs, ncRNAs, CRISPR arrays, and replication or transfer origins confirms its extrachromosomal nature. Annotated elements include mobile genetic elements (IS256 family, TnsA-like transposases, and site-specific integrases), plasmid stability and inheritance systems (RelE/ParE toxin-antitoxin module, HigA antitoxin, and the partitioning proteins ParA and ParB), plasmid-encoded transporters (ABC and MFS families), and components of the type IV secretion systems (VirB4, TraM, and TrwC relaxase). Further annotation using BLAST2GO revealed homologs of pRiA4b ORF-3 family proteins and Mu transposase C-terminal domain domains, underscoring the plasmid's roles in horizontal gene transfer and potential dissemination of antibiotic resistance determinants (Supplementary Information 4). Collectively, these genomic features highlight the adaptive mechanisms and extrachromosomal gene dynamics of SARSHI1, underpinning its metabolic versatility and resilience in hydrocarbon-rich or environmentally stressed habitats.

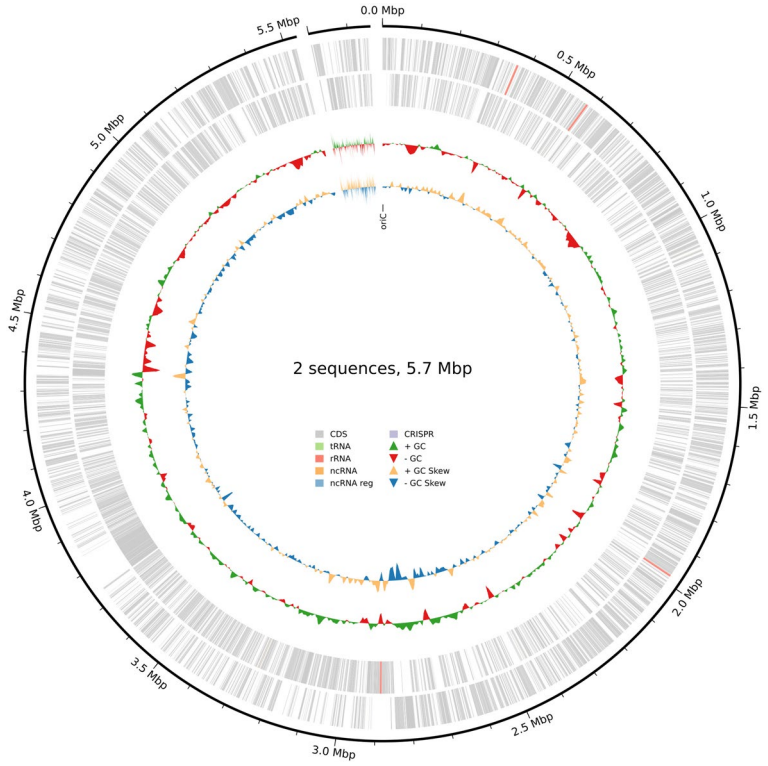


Fig. 7. Genome Map of *Rhodococcus indonesiensis* SARSHI1.

Functional annotation using Blast2GO

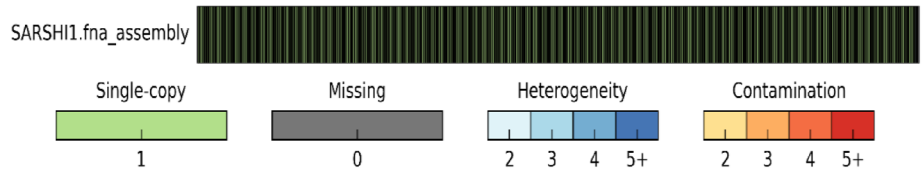
The assembly metrics, with an N50 of 5.5 Mbp, confirm a high-quality and contiguous genome assembly. Sequence alignment across multiple coverage thresholds exhibited strong consistency, particularly at 100% coverage. The highest number of alignment hits was observed in *Rhodococcus ruber* BKS 20–38 (1929 hits) and *Rhodococcus indonesiensis* (1905 hits), indicating a close phylogenetic and functional relationship with these taxa (Fig. S1, Supplementary Information 5). A minor discrepancy between the total sequences identified by BLAST (5144) and those assigned GO terms (4190), suggests that a subset of sequences remains functionally unclassified. Gene Ontology (GO) analysis revealed that the Molecular Function (MF) category was dominated by catalytic and electron transfer activities, while detoxification and localization processes prevailed within the Biological Process (BP) category (Fig. S2, Supplementary Information 5). Most sequences (803) were annotated with a single GO term, while a few exhibited more than ten, indicating multifunctionality. Enzymatic classification emphasized the predominance of oxidoreductases, reflecting strong oxidative metabolic potential (Fig. S3, Supplementary Information 5). Representative sequences exhibited high similarity to characterized proteins. For instance, sequence ENBDON_00001 (548 amino acids) showed 91.89% similarity to NR database proteins, with 20 BLAST hits, E-value (0.00E + 00), and 100% query coverage. The genome harbours extensive gene families associated with hydrocarbon metabolism, stress tolerance, and metal resistance (Supplementary Information 6).

Genomic insights into metabolism and functional annotation

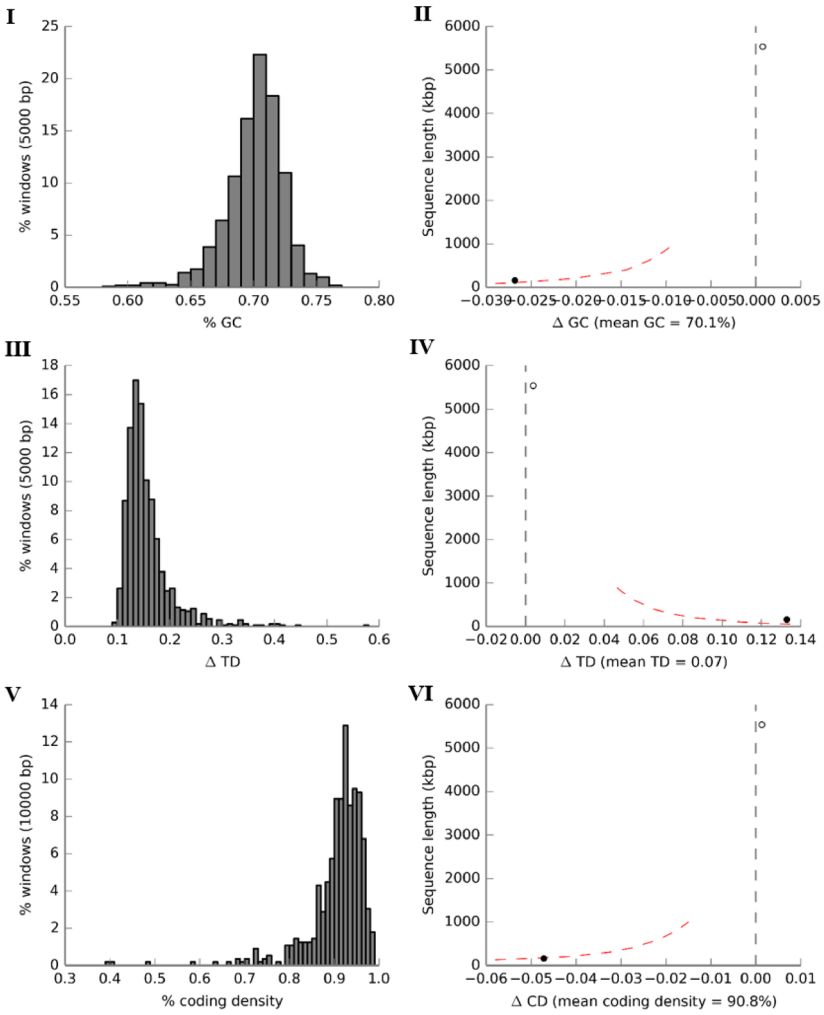
Genome quality metrics and functional overview

The assembled genome belongs to the order *Actinomycetales,* exhibiting 100% completeness with no detectable contamination. The presence of 570 single-copy marker genes with minimal duplication confirms the high-quality nature of the assembly. Most marker genes are present in single copies, with ≥ 90% amino acid identity (AAI), underscoring completeness and genome integrity (Fig. 8A). The GC content, tetranucleotide frequency deviation (ΔTD), and coding density (CD) distributions showed uniform patterns, reflecting a homogeneous and uncontaminated genome bin. The GC content histogram exhibited a unimodal peak centered at ~ 70%, characteristic of *Rhodococcus* species. Correspondingly, the ΔGC versus contig-length plot revealed no significant outliers, confirming compositional uniformity across contigs. The ΔTD analysis showed a narrow distribution (mean ΔTD = 0.07), indicative of stable tetranucleotide usage and minimal sequence heterogeneity. Coding density values clustered around 90.8%, typical of metabolically active *Actinobacteria*, further affirming the structural and functional integrity of the assembly (Fig. 8B). Collectively, these quality metrics demonstrate that the assembled genome is complete, structurally coherent, and well-suited for downstream comparative and functional analyses.

CheckM2 further corroborated the high genome quality, confirming 100% completeness with no missing sequences. The assembly exhibited a coding density of 91.4% indicative of the predominance of functional



A: CheckM-based bin quality assessment showing completeness, contamination, and heterogeneity.



B: Compositional Feature Analysis of the SARSHI1 Genome based on GC content, Tetranucleotide deviation, and Coding density.

protein-coding regions. The final assembly comprised two contigs, with the largest measuring 5.5 Mbp and an N50 value of 5,536,171 bp, reflecting high contiguity and minimal fragmentation. In total, 5170 protein-coding genes were predicted, with an average gene length of 333.85 bp, demonstrating a functionally rich and compact genomic organization. BUSCO analysis further validated the completeness of the genome, identifying 100% of the expected BUSCO genes, including 98.97% single-copy complete genes (n = 289) and 1.03% duplicated genes (n = 3) (Fig. S1, Supplementary Information 7). The recovery of all expected orthologous genes confirms a highly contiguous and comprehensive genome assembly with negligible redundancy.

The integrated Gene Ontology (GO) annotation framework provided a structured representation of gene products across three categories- molecular function (MF), biological process (BP), and cellular component (CC). This hierarchical structure revealed the interconnectedness of gene functions, with deeper nodes

◀**Fig. 8**. Compositional assessment of the SARSHI1 genome. (**A**) Genome quality metrics: Completeness, contamination, and strain heterogeneity were evaluated based on single-copy marker gene copy numbers. Green bars indicate single-copy markers, grey bars represent missing markers, and multi-copy markers are shaded according to amino acid identity (AAI)—blue for ≥ 90% (heterogeneity) and red for < 90% (contamination). Summary metrics show: single-copy = 1, missing = 0, heterogeneity = 0, and contamination = 0, confirming a high-quality, non-redundant assembly. (**B**) Compositional profile: Genome-wide compositional uniformity was assessed using GC content, tetranucleotide deviation (TD), and coding density (CD). Plots (I–II) show GC content distribution across 5 kbp windows and its relationship with contig length (ΔGC = 70.1%; red dashed lines denote expected variation limits). Plots (III–IV) illustrate tetranucleotide deviation, including the frequency distribution (mean ΔTD = 0.07) and ΔTD versus contig length. Plots (V–VI) depict coding density distributions (mean CD = 90.8%) and ΔCD versus contig length. Red dashed curves indicate the expected compositional variation thresholds.

representing more specific molecular roles. The CC ontology highlighted strong structural integrity and enrichment in intracellular enzymatic processes (Fig. S2, Supplementary Information 7). In the MF ontology, binding and catalytic activities dominated, particularly oxidoreductase (GO:0016491), hydrolase (GO:0016787), and transferase (GO:00167 functions, reflecting robust oxidative metabolism (Fig. S3, Supplementary Information 7). The BP ontology showed enrichment of genes involved in stress response, signal transduction, and adaptive metabolic regulation, underscoring environmental resilience. The predominance of oxidoreductase and hydrolase activities further substantiates robust hydrocarbon metabolic potential (Fig. S4, Supplementary Information 7).

Central metabolism and adaptation mechanism

The SARSHI1 genome encodes a comprehensive repertoire of metabolic pathways supporting diverse biochemical transformations. It possesses a complete glycolytic pathway (Embden-Meyerhof) and tricarboxylic acid (TCA) cycle, along with an integrated pentose phosphate pathway. However, alternative carbon fixation mechanisms, including the reductive Acetyl-CoA (Wood-Ljungdahl) and the 3-hydroxypropionate bi-cycle pathways, are incomplete, implying limited autotrophic potential. Genes encoding core components of the electron transport chain (ETC)- Complex I (NADH dehydrogenase), Complex III (cytochrome *bc1*), and Complex IV (cytochrome oxidases)- support aerobic respiration (Fig. S5A, Supplementary Information 7). The absence of specific subunits in certain bacteria restricts particular electron transfer reactions but allows the organism to exploit alternative electron acceptors and pathways encoded in its genome[85–87].

The genome also encodes an extensive nitrogen metabolism network, encompassing nitrate/nitrite reduction, nitrogen fixation, and ammonia oxidation, underscoring its potential role in nitrogen cycling. However, the lack of genes for the conversion of nitrous oxide to dinitrogen suggests incomplete denitrification. Additionally, the presence of arsenate and mercury reduction systems indicates potential for heavy metal detoxification and bioremediation. Conversely, the genome lacks photosynthetic machinery (Photosystem I/II) and complete sulfur metabolism pathways, reflecting constraints in sulfur cycling (Fig. S5B, Supplementary Information 7). Collectively, these features indicate that SARSHI1 is a metabolically versatile organism, capable of active carbon, nitrogen, and electron transport processes, yet with limitations in sulfur metabolism and complex carbohydrate degradation.

A diverse suite of aromatic hydrocarbon degradation pathways was identified, including xylene (ko00622), toluene (ko00623), nitrotoluene (ko00633), naphthalene (ko00626), aminobenzoate (ko00627), ethylbenzene (ko00642), styrene (ko00643), chloroalkane/chloroalkene (ko00625), and polycyclic aromatic hydrocarbon (ko00624) degradation. The presence of 3-phenylpropionate/trans-cinnamate dioxygenase ferredoxin reductase (K00529) further highlights its robust aromatic metabolic potential (Supplementary Information 8).

The alkane degradation pathway (ko00071) includes key enzymes such as alkane hydroxylases (*alkB1_2, alkM*; K00496), 3-hydroxyacyl-CoA dehydrogenase (EC:1.1.1.35), and short-chain acyl-CoA dehydrogenase (K00248), indicating an efficient system for aliphatic hydrocarbon catabolism. Partial adenosyl cobalamin (ADO-CBL) biosynthesis genes and dye-decolorizing peroxidases (K15733, K00485) indicate potential for xenobiotic degradation and detoxification (Supplementary Information 9).

Antibiotic biosynthesis and antimicrobial resistance

KEGG annotation unveiled a comprehensive metabolic framework governing antibiotic biosynthesis and antimicrobial resistance in SARSHI1. The genome encodes biosynthetic pathways (vancomycin (ko01055), type II polyketide products (KO01057), ansamycins (KO01051), streptomycin (KO00521)), resistance determinants, including beta-lactam resistance (ko01501), vancomycin resistance (KO01502), and CAMP resistance (KO01503). The drug metabolic enzymes, such as dimethylaniline monooxygenase (N-oxide forming)/hypotaurine monooxygenase (K00485), further delineate adaptive mechanisms that enhance survival in contaminated environments (Supplementary Information 9).

Comprehensive taxonomic analysis

ANI estimates were highly consistent across both bidirectional comparisons, confirming genomic similarity between SARSHI1 and *Rhodococcus indonesiensis* CSLK01-03. A minor variation (98.50% vs. 98.61%) was attributed to query–reference alignment difference. Matched fragments (1,578 out of 1,898 and 1,598 out of 1,706), respectively, further support a substantial proportion of shared genomic content. dDDH analysis against the closest reference genome (GCA_030360185.1) further validated species-level relatedness, with all

three formulas exceeding the 70% delineation threshold (Table S1, Supplementary Information 10). Formula 1 yielded 89.2% (distance 0.0851), Formula 2 88.2% (distance 0.0141), and Formula 3 91.7% (C.I. 89.2–93.7%), with a 99.61% probability of conspecificity. The minimal GC content difference (0.07%) corroborates taxonomic congruence. The Type Strain Genome Server (TYGS) analysis identified *Rhodococcus indonesiensis* CSLK01-03 as the closest relative, with dDDH values of 89.2% (d_0), 88.2% (d_4), and 91.7% (d_6) (Figs. S1 and S2, Supplementary Information 10). In contrast, *R. electrodiphilus* LMG 29,881 and *R. ruber* strains exhibited lower dDDH values (82.8%, 81.1%, and 80.2%, respectively), while distant species like *R. phenolicus* and *R. yananensis* displayed < 30% similarity. Collectively, these results confirm that SARSHI1 belongs to Species Cluster 9 within *Rhodococcus indonesiensis*, consistent with its 5.7 Mbp genome size and 70.08% GC content (Tables S2 and S3, Supplementary Information 10).

Protein domain and motif analysis

InterProScan identified functional domains in 4884 of 5170 proteins, with 3393 sequences annotated with GO terms. Comparative domain analyses using SMART, SuperFamily, and PANTHER databases revealed conserved motifs associated with hydrocarbon metabolism (Figs. S3–S5, Supplementary Information 10). Prominent repeat motifs included Hexapeptide, Pentapeptide, Ankyrin, and Tetratricopeptide repeats, along with pyrroloquinoline quinone and WD40-like β-propeller domains- both central to hydrocarbon degradation. Key enzymes of the meta-cleavage pathway, such as 4-hydroxy-2-oxovalerate aldolase (MF_01656) and acetaldehyde dehydrogenase (MF_01657), were identified, mediating aromatic ring-cleavage critical for hydrocarbon catabolism. HAMAP annotation confirmed core enzyme families responsible for hydrocarbon degradation and assisted metabolic processes. Approximately 87.17% of sequences were categorized as "others," reflecting extensive functional diversity, while uncharacterized DUF308 and DUF349 domains suggest potentially novel functions (Figs. 9 and S6, Supplementary Information 10). The efficient processing of aldehyde intermediates is a hallmark of hydrocarbon-degrading bacterial species, highlighting the metabolic specialization and biochemical versatility of SARSHI1.

Annotation of hydrocarbon degradation genes and pathways

The CANT_HYD analysis revealed a diverse repertoire of monooxygenases, dioxygenases, dehydrogenases, and reductases, underscoring the metabolic versatility of SARSHI1 in both aerobic and anaerobic hydrocarbon degradation. Prominent gene families included *AlkB, AhyA, AlmA_GroupI, LadAB, NdoBC, BmoXXY,* and *TmoAE*. Among these, the *AlkB*-encoded alkane 1-monooxygenases were the most abundant, with high-confidence homologs ENBDON_05321 (E-value: 6.60E-219, Score: 725.9) and ENBDON_05322 (E-value: 2.10E-190, Score: 631.9), catalysing initial alkane oxidation. Co-occurrence of rubredoxin systems (ENBDON_04732) and *AlmA*_GroupI FAD-containing monooxygenases (ENBDON_05147) further supports metabolic adaptability. Additional propane monooxygenase components (*PrmA, PrmC, sBmoX, TmoA_BmoA, TomA1, TomA3*) define a complete propane oxidation pathway, while *LadA_alpha/ LadA_beta* enzymes enable long-chain alkane metabolism.

Genes encoding Rieske-type oxygenases (*NdoB, NdoC*) and benzoate 1,2-dioxygenases suggest the presence of naphthalene and benzene degradation routes. The vanillate O-demethylase and nitrate/sulfite reductase systems indicate capability for anaerobic hydrocarbon oxidation. Genes such as *DszC*, *EbdA*,and *CmdA* involved in bio-desulfurization highlight potential applications for petroleum bioremediation, while F420-dependent oxidoreductases emphasize electron shuttle-based oxidation (Supplementary Information 11).

HADEG analysis identified extensive gene sets for aromatic and aliphatic hydrocarbon degradation, including benzoate (*benA*, *benB*, *benC*, *benD*), biphenyl (*bphF*, *bphI*, *bphX2*), and toluene (*xylC*) pathways. Genes for catechol (*catA*, *catC*, *pcaJ*), protocatechuate (*pcaG*, *pcaH*, *pcaR*), and gentisate pathway (*nagK*,

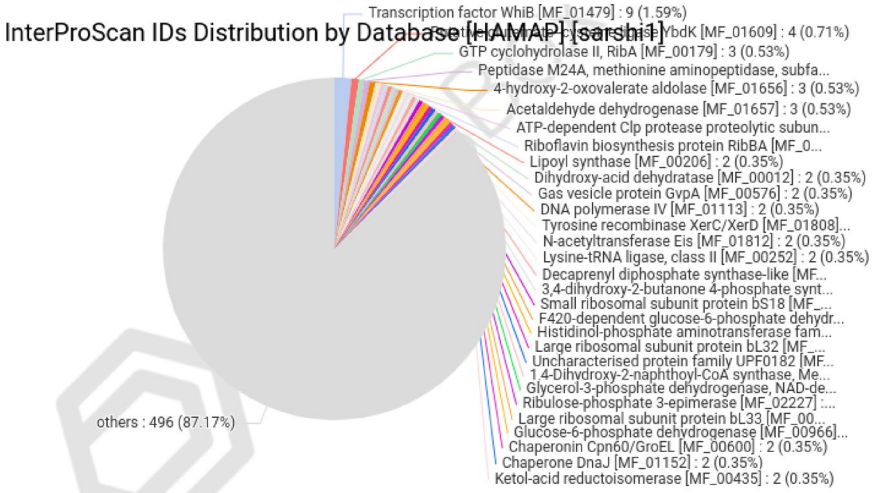


Fig. 9. Functional annotation of the SARSHI1 genome against the HAMAP database. Pie chart showing the distribution of protein families (InterProScan IDs) annotated through the HAMAP database.

nagL, *xlnE*) signify the complete aromatic mineralization potential. The (*alkB*, *alkG_rubA3_rdx*) and Baeyer–Villiger monooxygenases (BVMO, Q9I3H5) further reinforce aliphatic oxidation capacity. Gene distribution suggests that terminal oxidation (via *almA* and *ladA*) predominates over sub-terminal routes (Fig. 10). The Finnerty pathway (*AeAB_ahpC*, *AeAB_ahpF*) supports ω-oxidation, a key adaptive mechanism in *Rhodococcus* species[88–91]. Additionally, *ssuD* contributes to sulfur oxidation, complementing hydrocarbon metabolism (Supplementary Information 12 and Fig. S1; Supplementary Information 13).

Hydrocarbon Monooxygenase Database (HMDB) analysis identified hydrocarbon monooxygenases essential for alkane oxidation. High sequence identity was observed for *alkB* and rubredoxin-dependent hydroxylation system genes (*prmA–D*). For example, ENBDON_05051 showed 100% identity with *prmA* (B5D5P6), while ENBDON_05050–05048 displayed 92.79–95.66% identity. Low E-values (≤0.0) confirm high functional confidence. The presence of multiple gene copies and redundant enzymatic networks highlights a genetically reinforced system, enabling efficient oxidation, ring-cleavage, and anaerobic respiration under diverse environmental conditions (Table S1, Supplementary Information 13).

Aerobic degradation pathway of aliphatic and aromatic hydrocarbons

Microbial degradation of these hydrocarbons occurs in stages, starting with uptake through various mechanisms, including biosurfactant production and specialized transport systems, ultimately channelling intermediates into the tricarboxylic acid (TCA) cycle[92]. Aliphatic hydrocarbons are primarily degraded through terminal and subterminal oxidation. Terminal oxidation, mediated by alkane hydroxylases such as *alkB*, *almA*, and *ladA*, hydroxylates terminal carbon atoms to yield primary alcohols, which are subsequently oxidized to carboxylic acids before entering the β-oxidation pathway. In contrast, short-chain alkanes undergo subterminal oxidation to generate secondary alcohols[93]. Additionally, Baeyer–Villiger monooxygenases (BVMOs) contribute to the oxidation of ketones and cyclic intermediates. Gram-positive bacteria such as *Rhodococcus* utilize an ω-oxidation (Finnerty pathway) for metabolizing long-chain alkanes, yielding dicarboxylic acids as key intermediates.

Aromatic hydrocarbons are degraded through distinct dioxygenases. Benzene and biphenyl derivatives undergo initial hydroxylation by dioxygenases encoded by the *ben* (*benABC*) and *bph* (*bphA*, *bphB*, *bphC*, *bphD*) gene clusters, forming dihydroxylated intermediates such as catechol, protocatechuate, and gentisate[94]. The *xylC* gene encodes benzyl alcohol dehydrogenase, which oxidizes benzyl alcohol derivatives from toluene and xylene

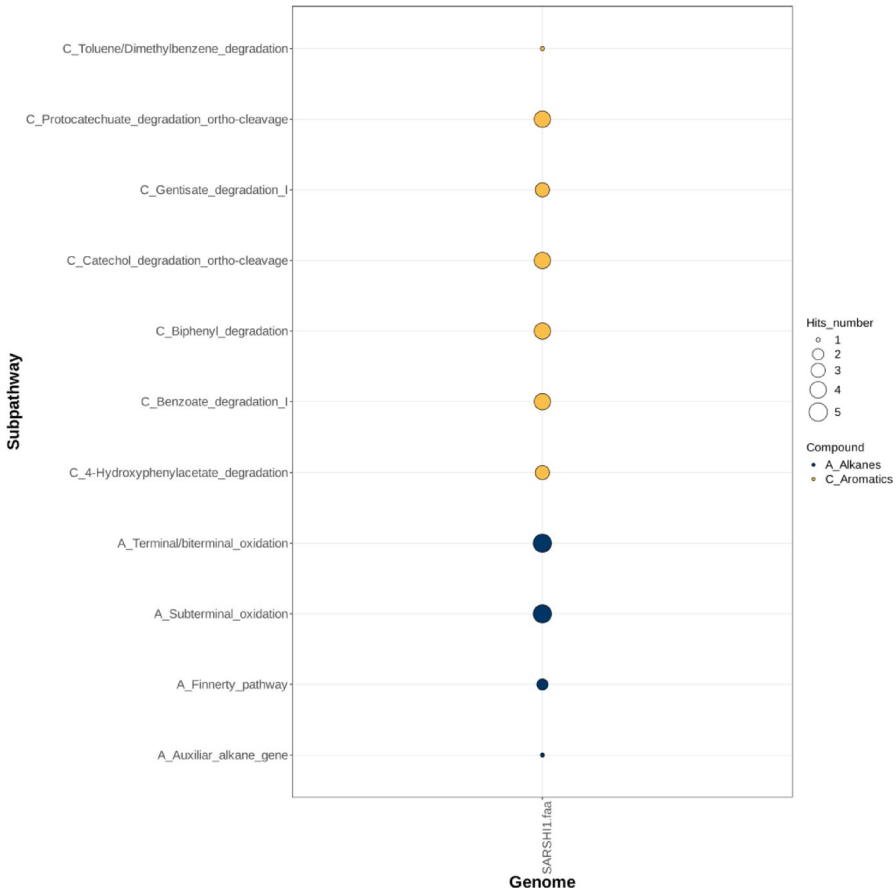


Fig. 10. HADEG-based predicted hydrocarbon degradation pathways in *Rhodococcus indonesiensis* SARSHI1. Bubble plot showing the presence and abundance of key genes involved in the degradation of aromatic (orange) and aliphatic (blue) compounds. Bubble size corresponds to the number of hits for each gene in the genome.

into their corresponding aldehydes. These dihydroxylated intermediates subsequently undergo ring cleavage via two principal mechanisms: ortho-cleavage through the β-ketoadipate pathway (via *catABC*, *pcaG*, *pcaH*, *pcaC*) and meta-cleavage (*xylE*), before their final assimilation into the TCA cycle (Fig. 11). Alternatively, the gentisate pathway involves the hydroxylation of aromatic compounds into gentisate, followed by ring cleavage mediated by gentisate 1,2-dioxygenase, ultimately directing intermediates into central metabolism(Tables S1 and S2, Supplementary Information 14). The combination of such functional aromatic and aliphatic degradation capabilities is rare among hydrocarbon-degrading bacteria, underscoring the broad-spectrum metabolic versatility of the strain.

Secondary metabolite analysis

The antiSMASH analysis identified 19 distinct BGCs, underscoring the genomic potential for diverse secondary metabolite biosynthesis (Fig. 12). These clusters encompass non-ribosomal peptide synthetases (NRPS), polyketide synthases (PKS), terpenes, ectoine, redox cofactors, and other metabolites, reflecting the organism's metabolic versatility, antibiotic biosynthetic potential, and stress adaptation capacity. Multiple NRPS and PKS clusters (Regions 1.1, 1.5, 1.10, 1.11, 1.13, 1.15, and 1.18) suggest the ability to synthesize structurally diverse bioactive compounds, with Region 1.11 (PKS) showing association with stenothricin, a known antimicrobial agent. Region 1.2 encodes a β-lactone BGC exhibiting a unique gene organization distinct from *Rhodococcus ruber*, suggesting a novel variant. Region 1.6 contains genes with < 60% similarity to characterized lasso peptide BGCs, indicating potential for the biosynthesis of previously undescribed structural variants. Four terpene clusters (Regions 1.7, 1.9, 1.14, and 1.17) may contribute to antimicrobial or signaling activities. Notably, Regions 1.9, 1.14, and 1.17 share < 50% similarity with known terpene clusters, implying divergent biosynthetic pathways. Region 1.16 encodes genes associated with redox cofactor biosynthesis (homology 0.40–0.49), while Region 1.19 exhibits low similarity (0.35–0.41) to known clusters, possibly representing a butyrolactone-like compound. Collectively, the low sequence homology across several BGCs underscores the metabolic plasticity of SARSHI1 and its potential to produce novel bioactive secondary metabolites.

Discussion

Oil, a non-renewable resource, underpins global economic stability but poses serious environmental and health hazards due to contamination by toxic petroleum hydrocarbons such as PAHs, resins, and asphaltenes[95]. These compounds bioaccumulate across trophic levels, thus threatening ecosystems and human health. Conventional remediation methods, including incineration, solvent extraction, and chemical leaching, mitigate visible spills but are restrained by high costs and secondary pollution[96]. Petroleum degradation rates depend on hydrocarbon composition, concentration, environmental factors, and microbial diversity. For instance, the half-life of low-molecular-weight PAHs ranges from 1.5–5.5 weeks in mildly contaminated soils to 2.5–52 weeks in heavily polluted environments[97]. Microbial bioremediation has emerged as an effective and sustainable approach. Advancements in multi-omics technologies have provided deeper insights into microbial metabolic pathways and the genetic adaptations that enable bacteria to degrade petroleum hydrocarbons efficiently. Multiple studies have identified diverse hydrocarbon-degrading bacteria with remarkable metabolic potential. Hossain et al. isolated 26 bacterial strains capable of utilizing polycyclic aromatic hydrocarbons and petroleum hydrocarbons as carbon sources[26]. Delegan et al. reported that *Rhodococcus opacus* S8 harbors genes for alkane degradation, surfactant biosynthesis, and cold adaptation, enabling hexadecane degradation under oxygen-limited conditions, facilitated by the formation of bacterial microconglomerates[98].

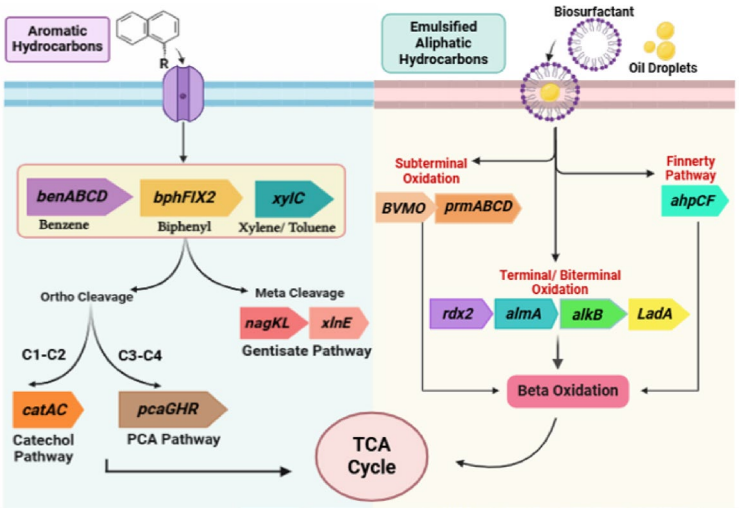
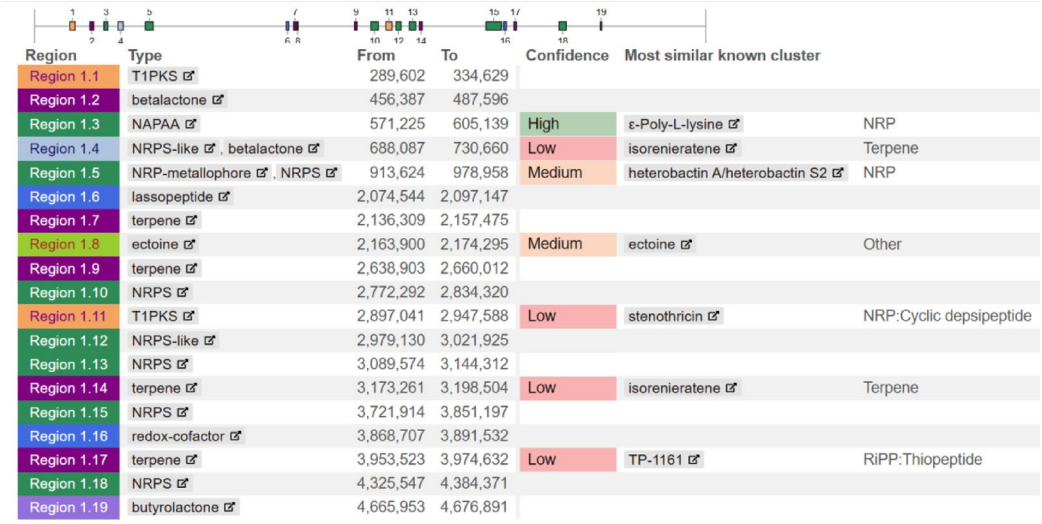


Fig. 11. Schematic overview of aerobic biodegradation of aromatic and aliphatic hydrocarbons in *Rhodococcus indonesiensis* SARSHI1. The diagram illustrates the mechanisms of hydrocarbon uptake, the key genes involved in degradation, and the pathways by which the bacterium metabolizes a range of aromatic and aliphatic hydrocarbons, ultimately channelling intermediates into central metabolism.



Region	Type	From	To	Confidence	Most similar known cluster	
Region 1.1	T1PKS	289,602	334,629			
Region 1.2	betalactone	456,387	487,596			
Region 1.3	NAPAA	571,225	605,139	High	ε-Poly-L-lysine	NRP
Region 1.4	NRPS-like , betalactone	688,087	730,660	Low	isorenieratene	Terpene
Region 1.5	NRP-metallophore , NRPS	913,624	978,958	Medium	heterobactin A/heterobactin S2	NRP
Region 1.6	lassopeptide	2,074,544	2,097,147			
Region 1.7	terpene	2,136,309	2,157,475			
Region 1.8	ectoine	2,163,900	2,174,295	Medium	ectoine	Other
Region 1.9	terpene	2,638,903	2,660,012			
Region 1.10	NRPS	2,772,292	2,834,320			
Region 1.11	T1PKS	2,897,041	2,947,588	Low	stenothricin	NRP:Cyclic depsipeptide
Region 1.12	NRPS-like	2,979,130	3,021,925			
Region 1.13	NRPS	3,089,574	3,144,312			
Region 1.14	terpene	3,173,261	3,198,504	Low	isorenieratene	Terpene
Region 1.15	NRPS	3,721,914	3,851,197			
Region 1.16	redox-cofactor	3,868,707	3,891,532			
Region 1.17	terpene	3,953,523	3,974,632	Low	TP-1161	RiPP:Thiopeptide
Region 1.18	NRPS	4,325,547	4,384,371			
Region 1.19	butyrolactone	4,665,953	4,676,891			

Fig. 12. antiSMASH analysis of *Rhodococcus indonesiensis* SARSHI1. Identified biosynthetic gene clusters are shown with their genomic positions, predicted cluster types, and the degree of similarity to known reference clusters.

In the present study, a novel petroleum hydrocarbon-degrading bacterium, *Rhodococcus indonesiensis* SARSHI1, was isolated and subjected to comprehensive physiological, enzymatic, and genomic characterization. The high-quality hybrid whole-genome assembly generated through Illumina and ONT platforms yielded a complete circular genome (5.7 Mbp, GC content 70.08%) comprising a chromosome and plasmid. To the best of our knowledge, this represents the first complete genome sequence reported for *R. indonesiensis*. The genome demonstrates a predominance of single-copy protein-coding genes with minimal duplication, signifying assembly completeness and high genomic integrity. Bidirectional ANI and dDDH values exceeded the 70% species delineation threshold, confirming species-level relatedness. The marginal GC content difference (0.07%) and phylogenomic clustering via TYGS firmly placed SARSHI1 within the *R. indonesiensis* clade, underscoring its phylogenetic distinctness.

SARSHI1 exhibited distinct substrate-dependent growth across various petroleum hydrocarbons, demonstrating broad metabolic adaptability. The strain effectively utilizes even-numbered *n*-alkanes and lighter PAHs like naphthalene and phenanthrene, while higher molecular weight hydrocarbons support slower growth, likely due to structural complexity and lower bioavailability. Sequential adaptation markedly enhanced its ability to utilize long-chain alkanes (C_{32}–C_{38}) and tolerate higher hydrocarbon concentrations (up to 750 mg L^{-1}), suggesting an inducible enzymatic system. Growth inhibition at 1000 mg L^{-1} suggests substrate toxicity or mass transfer limitations. The specific induction of catechol 1,2-dioxygenase (C12O) and catechol 2,3-dioxygenase (C23O) confirms their role in aromatic degradation. The dominance of C12O activity (~4500 µmol min^{-1} mg^{-1} protein) over C23O (~2100 µmol min^{-1} mg^{-1}) indicates preferential utilization of the ortho-cleavage pathway. The decline in activity after day 10 suggests substrate depletion or feedback regulation, marking a transition from active degradation to a stationary phase of growth.

GC–MS and enzymatic analyses collectively confirmed the efficient alkane oxidation and crude oil degradation. The appearance of hexadecanol ($C_{16}H_{34}O$) from *n*-hexadecane verified the activity of alkane hydroxylase (*alkB*), indicating terminal oxidation of alkanes as the initial step in degradation. In the crude oil degradation profile, complete mineralization of low-molecular-weight alkanes (C_{10}–C_{15}) and a three-seven fold decrease in the abundance of mid- to long-chain n-alkanes (C_{16}–C_{30}) demonstrated extensive substrate utilization. The pronounced reduction of pristane and phytane peaks further indicated the strain's ability to degrade branched hydrocarbons. The observed oxidation, mineralization sequence, supported by high *alkB* activity and the formation of detectable alcohol intermediates, highlights the efficiency of terminal oxidation as a primary route for alkane catabolism in this strain.

These patterns align with the genomic evidence, detection of multiple alkane hydroxylase systems (*alkB, alkG_rubA3_rdx*), monooxygenases (*AhyA, AlmA_GroupI, LadAB*), and coexistence of terminal and subterminal oxidation systems, along with the Finnerty pathway (*ahpC, ahpF*), suggesting that SARSHI1 predominantly employs terminal oxidation for aliphatic hydrocarbon breakdown, providing metabolic redundancy across substrate ranges. Beyond aliphatic metabolism, the genome encodes complete pathways for xylene, toluene, aminobenzoate, chloroalkane, and PAH degradation (*benABCD, bphFIX2, NdoBC, BmoXY, TmoAE*), underscoring its potential for comprehensive hydrocarbon mineralization. The presence of genes encoding ortho-cleavage enzymes (*catA, catC, pcaJ*), and the protocatechuate branch (*pcaG, pcaH, pcaR*) of the β-ketoadipate pathway confirms the dominance of the ortho-cleavage route for aromatic degradation, an established hallmark of efficient PAH-degrading *Rhodococcus* species. The simultaneous presence of the gentisate pathway with the aforementioned ring-cleavage systems enhances substrate versatility, allowing efficient processing of structurally diverse aromatic intermediates.

Additionally, genes associated with antibiotic biosynthesis and resistance (β-lactam, vancomycin, and xenobiotic-metabolizing enzymes) suggest adaptive advantages in contaminated competitive environments. The identified BGCs may contribute to stress tolerance, interspecies interactions, and long-term environmental persistence. The combined evidence reveals that *Rhodococcus indonesiensis* SARSHI1 is well equipped for efficient hydrocarbon transformation through diverse oxidative and ring-cleavage systems. Its adaptive utilization of both aliphatic and aromatic hydrocarbons highlights coordinated regulation between metabolic and stress-response pathways. The coexistence of metal- and antibiotic-resistance genes further suggests ecological resilience and competitiveness in contaminated environments. The complete genome not only elucidates the molecular mechanisms underlying hydrocarbon metabolism but also provides a valuable framework for metabolic engineering and biotechnological exploitation in environmental remediation. Together, these traits positioned SARSHI1 as a promising candidate for bioremediation and biotechnological exploitation.

Conclusion

The study presents the first complete genome sequence of *Rhodococcus indonesiensis* SARSHI1, harbouring a versatile enzymatic arsenal and a highly adaptable metabolic network for hydrocarbon degradation. Its capacity to efficiently utilize low- to mid-chain n-alkanes, adapt to high-molecular-weight hydrocarbons, and preferentially employ the catechol ortho-cleavage pathway underscores its metabolic flexibility. The integration of multiple oxidative systems, stress-tolerance genes, and regulatory networks further enhances its ability to thrive in hydrocarbon-rich and variable environments. The presence of antibiotic resistance determinants, metal-resistance genes, and biosynthetic gene clusters for secondary metabolites suggests ecological competitiveness and resilience under polluted conditions. Collectively, these features position SARSHI1 as a potent biocatalyst for the biodegradation of complex petroleum hydrocarbons. The complete genome sequence is available in GenBank under accession numbers CP180630 (chromosome) and CP180631 (plasmid). The raw sequencing reads have been submitted to the Sequence Read Archive (SRA), NCBI, under accession numbers SRX27520007 (Illumina) and SRX27520006 (ONT).

Data availability

This whole-genome project has been deposited in GenBank under the Accession: CP180630 and CP180631 (plasmid). The read sequences have been deposited under BioProject Accession: PRJNA1217105, BioSample Accession: SAMN46479200, and SRA Accession: SRX27520007 and SRX27520006. Moreover, the queries can be directed to the corresponding author for any clarifications about the study if needed. WGS URL: https://www.ncbi.nlm.nih.gov/nuccore/CP180630 Plasmid URL: https://www.ncbi.nlm.nih.gov/nuccore/CP180631 BioProject URL: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1217105 BioSample URL: https://www.ncbi.nlm.nih.gov/biosample/SAMN46479200 SRA URL (Illumina): https://www.ncbi.nlm.nih.gov/sra/?term=SRX27520007 SRA URL (ONT): https://www.ncbi.nlm.nih.gov/sra/?term=SRX27520006.

Received: 26 March 2025; Accepted: 13 November 2025
Published online: 02 December 2025

References

1. Imam, A., Kanaujia, P. K., Ray, A. & Suman, S. K. Removal of petroleum contaminants through bioremediation with integrated concepts of resource recovery: A review. *Indian J. Microbiol.* **61**(3), 250–261. https://doi.org/10.1007/s12088-021-00928-4 (2021).
2. Narayan Thorat, B. & Kumar Sonwani, R. Current technologies and future perspectives for the treatment of complex petroleum refinery wastewater: A review. *Bioresour. Technol.* **355**, 127263. https://doi.org/10.1016/j.biortech.2022.127263 (2022).
3. Kuppusamy, S., Maddela, N. R., Megharaj, M. & Venkateswarlu, K. Ecological impacts of total petroleum hydrocarbons. In *Total Petroleum Hydrocarbons* 95–138 (Springer, 2020). https://doi.org/10.1007/978-3-030-24035-6_5.
4. Wei, Y., Ding, D., Qu, K., Sun, J. & Cui, Z. Ecological risk assessment of heavy metal pollutants and total petroleum hydrocarbons in sediments of the Bohai Sea, China. *Mar. Pollut. Bull.* **184**, 114218. https://doi.org/10.1016/j.marpolbul.2022.114218 (2022).
5. Kuppusamy, S., Maddela, N. R., Megharaj, M. & Venkateswarlu, K. Fate of total petroleum hydrocarbons in the environment. In *Total Petroleum Hydrocarbons* 57–77 (Springer, Cham, 2020). https://doi.org/10.1007/978-3-030-24035-6_3.
6. Pal, D. & Sen, S. Emerging petroleum pollutants and their adverse effects on the environment. In *Impact of petroleum waste on environmental pollution and its sustainable management through circular economy, environmental science and engineering* (eds Behera, I. D. & Das, A. P.) 103–137 (Springer, 2023).
7. Gao, H., Wu, M., Liu, H., Xu, Y. & Liu, Z. Effect of petroleum hydrocarbon pollution levels on the soil microecosystem and ecological function. *Environ. Pollut.* **293**, 118511. https://doi.org/10.1016/j.envpol.2021.118511 (2022).
8. Adipah, S. Introduction of petroleum hydrocarbons contaminants and its human effects. *J. Environ. Sci. Public Health* **3**(1), 1–9 (2019).
9. Pandey, A. K. et al. Multipronged evaluation of genotoxicity in Indian petrol-pump workers. *Environ. Mol. Mutagen* **49**(9), 695–707. https://doi.org/10.1002/em.20419 (2008).
10. Abdel-Shafy, H. I. & Mansour, M. S. M. A review on polycyclic aromatic hydrocarbons: Source, environmental impact, effect on human health and remediation. *Egypt. J. Pet.* **25**(1), 107–123. https://doi.org/10.1016/j.ejpe.2015.03.011 (2016).
11. Ossai, I. C., Ahmed, A., Hassan, A. & Hamid, F. S. Remediation of soil and water contaminated with petroleum hydrocarbon: A review. *Environ. Technol. Innov.* **17**, 100526. https://doi.org/10.1016/j.eti.2019.100526 (2020).
12. Das, N. & Chandran, P. Microbial degradation of petroleum hydrocarbon contaminants: An overview. *Biotechnol. Res. Int.* **2011**, 1–13. https://doi.org/10.4061/2011/941810 (2011).
13. Varjani, S., Pandey, A. & Upasani, V. N. Petroleum sludge polluted soil remediation: Integrated approach involving novel bacterial consortium and nutrient application. *Sci. Total Environ.* **763**, 142934. https://doi.org/10.1016/j.scitotenv.2020.142934 (2021).
14. Viesser, J. A. et al. Petroleum-tolerant rhizospheric bacteria: Isolation, characterization and bioremediation potential. *Sci. Rep.* **10**(1), 2060. https://doi.org/10.1038/s41598-020-59029-9 (2020).
15. Bej, A. K., Saul, D. & Aislabie, J. Cold-tolerant alkane-degrading Rhodococcus species from Antarctica. *Polar Biol.* **23**(2), 100–105. https://doi.org/10.1007/s003000050014 (2000).
16. Kuyukina, M. S. & Ivshina, I. B. Bioremediation of contaminated environments using Rhodococcus. In *Biology of Rhodococcus* Vol. 16 (ed. Alvarez, H. M.) 231–270 (Springer, 2019). https://doi.org/10.1007/978-3-030-11461-9_9.

17. Chen, X. et al. In situ bioremediation of petroleum hydrocarbon–contaminated soil: isolation and application of a Rhodococcus strain. *Int. Microbiol.* **26**(2), 411–421. https://doi.org/10.1007/s10123-022-00305-1 (2022).
18. Nazari, M. T. et al. Rhodococcus: A promising genus of actinomycetes for the bioremediation of organic and inorganic contaminants. *J. Environ. Manag.* **323**, 116220. https://doi.org/10.1016/j.jenvman.2022.116220 (2022).
19. Kaur, M. et al. Navigating the genomic landscape: A deep dive into clinical genetics with deep learning. In *Deep Learning in Genetics and Genomics* 185–224 (Elsevier, 2025). https://doi.org/10.1016/B978-0-443-27574-6.00006-0.
20. Nayarisseri, A. et al. Impact of next-generation whole-exome sequencing in molecular diagnostics. *Drug Invention Today* **5**(4), 327–334. https://doi.org/10.1016/j.dit.2013.07.005 (2013).
21. Tilgner, H. et al. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat Biotechnol* **33**(7), 736–742. https://doi.org/10.1038/nbt.3242 (2015).
22. Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D. & Ragoussis, J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci Rep* **6**(1), 31602. https://doi.org/10.1038/srep31602 (2016).
23. De Maio, N. et al. "Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb. Genom.* https://doi.org/10.1099/mgen.0.000294 (2019).
24. Y. Lu *et al.*, "Hybrid Clustering of Long and Short-read for Improved Metagenome Assembly," Jan. 26, 2021. https://doi.org/10.1101/2021.01.25.428115.
25. Das, D. et al. Complete genome sequence analysis of Pseudomonas aeruginosa N002 reveals its genetic adaptation for crude oil degradation. *Genomics* **105**(3), 182–190. https://doi.org/10.1016/j.ygeno.2014.12.006 (2015).
26. Hossain, M. S., Iken, B. & Iyer, R. Whole genome analysis of 26 bacterial strains reveals aromatic and hydrocarbon degrading enzymes from diverse environmental soil samples. *Sci. Rep.* **14**(1), 30685. https://doi.org/10.1038/s41598-024-78564-3 (2024).
27. Nayarisseri, A., Singh, P. & Singh, S. K. Screening, isolation and characterization of biosurfactant-producing Bacillus tequilensis strain ANSKLAB04 from brackish river water. *Int. J. Environ. Sci. Technol.* **16**(11), 7103–7112. https://doi.org/10.1007/s13762-018-2089-9 (2019).
28. Nayarisseri, A. et al. Whole genome sequence of petroleum hydrocarbon degrading novel strain *Microbacter* sp. EMBS2025 isolated from Chilika Lake, Odisha, India. *Sci. Rep.* **15**(1), 27961. https://doi.org/10.1038/s41598-025-13545-8 (2025).
29. Zaman, S. A. U., Bhrdwaj, A., Nayarisseri, A., Khazanehdari, K. A. & Bhuyan, R. Isolation and characterization of novel hydrocarbon-degrading bacteria from oil polluted soil near Nacharam, Hyderabad, India. *Sci. Rep.* **15**(1), 17219. https://doi.org/10.1038/s41598-025-01081-4 (2025).
30. Mohan, M., Kozhithodi, S. & Nayarisseri, A. Screening, purification and characterization of protease inhibitor from Capsicum frutescens. *Bioinformation* **14**(06), 285–293. https://doi.org/10.6026/97320630014285 (2018).
31. Nayarisseri, A. et al. Identification and characterization of neutral protease producing Paenibacillus polymyxa species EMBS024 BY 16S rRNA gene sequencing. *Int. J. Micro. Res.* **4**(5), 236–239. https://doi.org/10.9735/0975-5276.4.5.236-239 (2012).
32. Ravi, A. et al. Characterization of petroleum degrading bacteria and its optimization conditions on effective utilization of petroleum hydrocarbons. *Microbiol. Res.* **265**, 127184. https://doi.org/10.1016/j.micres.2022.127184 (2022).
33. Xiang, W. et al. Degradation of long-chain n-alkanes by a novel thermal-tolerant *Rhodococcus* strain. *Arch Microbiol.* **204**(5), 259. https://doi.org/10.1007/s00203-022-02872-3 (2022).
34. Yang, H.-Y., Jia, R.-B., Chen, B. & Li, L. Degradation of recalcitrant aliphatic and aromatic hydrocarbons by a dioxin-degrader *Rhodococcus* sp. strain p52. *Environ Sci Pollut Res* **21**(18), 11086–11093. https://doi.org/10.1007/s11356-014-3027-0 (2014).
35. Zampolli, J., Collina, E., Lasagni, M. & Di Gennaro, P. Biodegradation of variable-chain-length n-alkanes in *Rhodococcus* opacus R7 and the involvement of an alkane hydroxylase system in the metabolism. *AMB Expr.* **4**(1), 73. https://doi.org/10.1186/s13568-014-0073-4 (2014).
36. Kumari, B., Singh, S. N. & Singh, D. P. Characterization of two biosurfactant producing strains in crude oil degradation. *Process Biochem.* **47**(12), 2463–2471. https://doi.org/10.1016/j.procbio.2012.10.010 (2012).
37. Kim, D. et al. Monocyclic aromatic hydrocarbon degradation by *Rhodococcus* sp. Strain DK17. *Appl. Environ. Microbiol.* **68**(7), 3270–3278. https://doi.org/10.1128/AEM.68.7.3270-3278.2002 (2002).
38. Lo Piccolo, L., De Pasquale, C., Fodale, R., Puglia, A. M. & Quatrini, P. Involvement of an Alkane hydroxylase system of *Gordonia* sp. strain SoCg in degradation of solid *n*-Alkanes. *Appl. Environ. Microbiol.* **77**(4), 1204–1213. https://doi.org/10.1128/AEM.02180-10 (2011).
39. Whyte, L. G. et al. Biodegradation of variable-chain-length alkanes at low temperatures by a psychrotrophic *Rhodococcus* sp. *Appl. Environ. Microbiol.* **64**(7), 2578–2584. https://doi.org/10.1128/AEM.64.7.2578-2584.1998 (1998).
40. Li, C. et al. Biodegradation of crude oil by a newly isolated strain *Rhodococcus* sp. JZX-01. *Appl. Biochem. Biotechnol.* **171**(7), 1715–1725. https://doi.org/10.1007/s12010-013-0451-4 (2013).
41. Qiagen, Q. QIAamp DNA mini kit (Catalog No. 51304). Qiagen. Available at: 2020, [Online] (2020). Available: https://www.qiagen.com/us/products/dna-analysis/dna-purification/qiamp-dna-mini-kit
42. Date, L. E. *Thermo Scientific CloneJET PCR Cloning Kit* (2017).
43. Amareshwari, P. et al. Isolation and characterization of a novel chlorpyrifos degrading flavobacterium species EMBS0145 by 16S rRNA gene sequencing. *Interdiscip. Sci. Comput. Life Sci.* **7**(1), 1–6. https://doi.org/10.1007/s12539-012-0207-9 (2015).
44. Anuraj, N. S. et al. Identification and characterization of proteases and amylases producing *Bacillus licheniformis* strain EMBS026 by 16S rRNA gene sequencing. *Int. J. Micro. Res.* **4**(5), 231–235. https://doi.org/10.9735/0975-5276.4.5.231-235 (2012).
45. Chandok, H. et al. Screening, isolation and identification of probiotic producing lactobacillus acidophilus strains EMBS081 & EMBS082 by 16S rRNA gene sequencing. *Interdiscip. Sci. Comput. Life Sci.* **7**(3), 3. https://doi.org/10.1007/s12539-015-0002-5 (2015).
46. Nayarisseri, A., Suppahia, A., Nadh, A. G. & Nair, A. S. Identification and characterization of a pesticide degrading flavobacterium species EMBS0145 by 16S rRNA gene sequencing. *Interdiscip. Sci. Comput. Life Sci.* **7**(2), 2. https://doi.org/10.1007/s12539-015-0016-z (2015).
47. DNA Baser v5.15() 2022, [Online]. Available: DNA Baser v5.15 (2022). SciVance Technologies, www.DnaBaser.com
48. Khan, A., Nayarisseri, A. & Singh, S. K. Characterization and optimization of azo dyes degrading microbes isolated from textile effluent. *Sci. Rep.* **15**(1), 11241. https://doi.org/10.1038/s41598-025-95359-2 (2025).
49. Nayarisseri, A., Bhrdwaj, A., Sharma, K., Khan, A. & Singh, S. K. Identification and characterization of biosurfactant-producing novel *Paenibacillus dendritiformis* strain ANSKLAB02 from brackish river water. *CBIOT* **14**(1), 52–76. https://doi.org/10.2174/0122115501370139250207053902 (2025).
50. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European molecular biology open software suite. *Trends Genet.* **16**(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2 (2000).
51. Bhatia, M., Girdhar, A., Tiwari, A. & Nayarisseri, A. Implications of a novel Pseudomonas species on low density polyethylene biodegradation: An in vitro to in silico approach. *Springerplus* **3**(1), 1. https://doi.org/10.1186/2193-1801-3-497 (2014).
52. Bhrdwaj, A., Nayarisseri, A. & Singh, S. K. Characterization of natural therapeutic compounds producing novel bacterial strains isolated from Hyderabad, India. *Sci. Rep.* **15**(1), 25681. https://doi.org/10.1038/s41598-025-11615-5 (2025).
53. Nadh, A. G. et al. Identification of azo dye degrading sphingomonas strain EMBS022 and EMBS023 using 16S rRNA gene sequencing. *CBIO* **10**(5), 5. https://doi.org/10.2174/1574893610666151008012312 (2015).
54. Pyde, A. N. et al. Identification and characterization of foodborne pathogen Listeria monocytogenes strain Pyde1 and Pyde2 using 16S rRNA gene sequencing. *J. Pharm. Res.* **6**(7), 736–741. https://doi.org/10.1016/j.jopr.2013.07.009 (2013).

55. Shah, K. P. et al. Screening, isolation and identification of polygalacturonase producing *Bacillus tequilensis* strain EMBS083 using 16S rRNA gene sequencing. *Eur. J. Biol. Sci.* **5**(1), 09–13 (2013).
56. Nayarisseri, A., Singh, P. & Singh, S. K. Screening, isolation and characterization of biosurfactant producing *Bacillus subtilis* strain ANSKLAB03. *Bioinformation* **14**(06), 304–314. https://doi.org/10.6026/97320630014304 (2018).
57. Nayarisseri, A., Khandelwal, R. & Singh, S. K. Identification and characterization of lipopeptide biosurfactant producing *Microbacterium* sp isolated from brackish river water. *CTMC* **20**(24), 2221–2234. https://doi.org/10.2174/1568026620666620062814471 6 (2020).
58. Sharma, K., Nayarisseri, A. & Singh, S. K. Biodegradation of plasticizers by novel strains of bacteria isolated from plastic waste near Juhu Beach, Mumbai, India. *Sci. Rep.* **14**(1), 30824. https://doi.org/10.1038/s41598-024-81239-8 (2024).
59. Agilent Technologies. 4200 TapeStation system 2015, [Online]. Available: https://www.agilent.com/en/product/automated-electrophoresis/tapestation-systems/tapestation-instruments/4200-tapestation-system-228263
60. Nayarisseri, A. & Singh, S. K. Genome analysis of biosurfactant producing bacterium, *Bacillus tequilensis*. *PLoS ONE* **18**(6), e0285994. https://doi.org/10.1371/journal.pone.0285994 (2023).
61. Pradhan, S. K. et al. Illumina MiSeq based assessment of bacterial community structure and diversity along the heavy metal concentration gradient in Sukinda chromite mine area soils, India. *Ecol. Genet. Genom.* **15**, 100054. https://doi.org/10.1016/j.egg.2020.100054 (2020).
62. Versmessen, N. et al. Average nucleotide identity and digital DNA–DNA hybridization analysis following PromethION nanopore-based whole genome sequencing allows for accurate prokaryotic typing. *Diagnostics* **14**(16), 1800. https://doi.org/10.3390/diagnostics14161800 (2024).
63. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. 2010, [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
64. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 (2014).
65. Bankevich, A. et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**(5), 455–477. https://doi.org/10.1089/cmb.2012.0021 (2012).
66. Wick, R. *Filtlong: Read Trimming and Filtering Tool for Long Reads*. GitHub Repository (2021) [Online]. Available: https://github.com/rrwick/Filtlong
67. Wick, R. Porechop: *Adapter Trimmer for Oxford Nanopore Reads. GitHub Repository* (2021) [Online]. Available: https://github.com/rrwick/Porechop.
68. 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**(6), e1005595. https://doi.org/10.1371/journal.pcbi.1005595 (2017).
69. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086 (2013).
70. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: Interactive visualization of *de novo* genome assemblies. *Bioinformatics* **31**(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383 (2015).
71. Schwengers, O. et al. Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification: Find out more about Bakta, the motivation, challenges and applications, here. *Microb. Genom.* https://doi.org/10.1099/mgen.0.000685 (2021).
72. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**(12), 5825–5829. https://doi.org/10.1093/molbev/msab293 (2021).
73. 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**(7), 1043–1055. https://doi.org/10.1101/gr.186072.114 (2015).
74. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**(8), 1203–1212. https://doi.org/10.1038/s41592-023-01940-w (2023).
75. Shaffer, M. et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**(16), 8883–8900. https://doi.org/10.1093/nar/gkaa621 (2020).
76. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351 (2015).
77. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**(1), 5114. https://doi.org/10.1038/s41467-018-07641-9 (2018).
78. Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P. & Göker, M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinform.* **14**(1), 60. https://doi.org/10.1186/1471-2105-14-60 (2013).
79. Meier-Kolthoff, J. P., Carbasse, J. S., Peinado-Olarte, R. L. & Göker, M. TYGS and LPSN: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. *Nucleic Acids Res.* **50**(D1), D801–D807. https://doi.org/10.1093/nar/gkab902 (2022).
80. Meier-Kolthoff, J. P. & Göker, M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nat. Commun.* **10**(1), 2182. https://doi.org/10.1038/s41467-019-10210-3 (2019).
81. Khot, V. et al. CANT-HYD: A curated database of phylogeny-derived hidden markov models for annotation of marker genes involved in hydrocarbon degradation. *Front. Microbiol.* **12**, 764058. https://doi.org/10.3389/fmicb.2021.764058 (2022).
82. Rojas-Vargas, J., Castelán-Sánchez, H. G. & Pardo-López, L. HADEG: A curated hydrocarbon aerobic degradation enzymes and genes database. *Bioinformatics* https://doi.org/10.1101/2022.08.30.505856 (2022).
83. Wang, S. et al. HMDB: A curated database of genes involved in hydrocarbon monooxygenation reaction with homologous genes as background. *J. Hazard. Mater.* **460**, 132397. https://doi.org/10.1016/j.jhazmat.2023.132397 (2023).
84. Blin, K. et al. antiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**(W1), W81–W87. https://doi.org/10.1093/nar/gkz310 (2019).
85. White, D. C. & Sinclair, P. R. Branched electron-transport systems in bacteria. In *Advances in Microbial Physiology* Vol. 5 173–211 (Elsevier, 1971). https://doi.org/10.1016/S0065-2911(08)60407-5.
86. Wang, S. et al. Electron transport chains in organohalide-respiring bacteria and bioremediation implications. *Biotechnol. Adv.* **36**(4), 1194–1206. https://doi.org/10.1016/j.biotechadv.2018.03.018 (2018).
87. Kracke, F. et al. Microbial electron transport and energy conservation—the foundation for optimizing bioelectrochemical systems. *Front. Microbiol.* https://doi.org/10.3389/fmicb.2015.00575 (2015).
88. Larkin, M. J., Kulakov, L. A. & Allen, C. C. Biodegradation and *Rhodococcus* – Masters of catabolic versatility. *Curr. Opin. Biotechnol.* **16**(3), 282–290. https://doi.org/10.1016/j.copbio.2005.04.007 (2005).
89. Park, C. & Park, W. Survival and energy producing strategies of Alkane degraders under extreme conditions and their biotechnological potential. *Front. Microbiol.* **9**, 1081. https://doi.org/10.3389/fmicb.2018.01081 (2018).
90. Abbasian, F., Lockington, R., Mallavarapu, M. & Naidu, R. A comprehensive review of aliphatic hydrocarbon biodegradation by bacteria. *Appl. Biochem. Biotechnol.* **176**(3), 670–699. https://doi.org/10.1007/s12010-015-1603-5 (2015).
91. Cappelletti, M., Fedi, S. & Zannoni, D. Degradation of alkanes in *Rhodococcus*. In *Biology of Rhodococcus* Vol. 16 (ed. Alvarez, H. M.) 137–171 (Springer, 2019). https://doi.org/10.1007/978-3-030-11461-9_6.

92. Binazadeh, M., Karimi, I. A. & Li, Z. Fast biodegradation of long chain n-alkanes and crude oil at high concentrations with *Rhodococcus* sp. Moj-3449. *Enzyme Microb. Technol.* **45**(3), 195–202. https://doi.org/10.1016/j.enzmictec.2009.06.001 (2009).
93. Kawagoe, T., Kubota, K., Araki, K. S. & Kubo, M. Analysis of the Alkane hydroxylase gene and long-chain cyclic alkane degradation in *Rhodococcus*. *AIDS Patient Care STDs* **09**(03), 151–163. https://doi.org/10.4236/aim.2019.93012 (2019).
94. Krivoruchko, A., Kuyukina, M., Peshkur, T., Cunningham, C. J. & Ivshina, I. *Rhodococcus* strains from the specialized collection of Alkanotrophs for biodegradation of aromatic compounds. *Molecules* **28**(5), 2393. https://doi.org/10.3390/molecules28052393 (2023).
95. Cerqueda-García, D., García-Maldonado, J. Q., Aguirre-Macedo, L. & García-Cruz, U. A succession of marine bacterial communities in batch reactor experiments during the degradation of five different petroleum types. *Mar. Pollut. Bull.* **150**, 110775. https://doi.org/10.1016/j.marpolbul.2019.110775 (2020).
96. Fathepure, B. Z. Recent studies in microbial degradation of petroleum hydrocarbons in hypersaline environments. *Front. Microbiol.* https://doi.org/10.3389/fmicb.2014.00173 (2014).
97. Roslund, M. I. et al. Endocrine disruption and commensal bacteria alteration associated with gaseous and soil PAH contamination among daycare children. *Environ. Int.* **130**, 104894. https://doi.org/10.1016/j.envint.2019.06.004 (2019).
98. Delegan, Y. et al. Complete genome analysis of *Rhodococcus opacus* S8 capable of degrading alkanes and producing biosurfactant reveals its genetic adaptation for crude oil decomposition. *Microorganisms* **10**(6), 1172. https://doi.org/10.3390/microorganisms10061172 (2022).

Acknowledgements

The authors are thankful to Eminent Biosciences and LeGene Biosciences Pvt Ltd, Indore, India for 16S rRNA sequencing, Whole Genome Sequencing, and De novo assembly of the bacterium.

Author contributions

S.A.U.Z.: Contributed to the conceptualization, Investigation, Methodology, Sample Collection, NGS Data analysis, Validation, Visualization and original draft writing of the manuscript. S.A.U.Z. also participated in reviewing and editing the manuscript. K.S.: Participated in composition of the initial draft and were also involved in reviewing and editing the manuscript. A.N.: Involved in Conceptualization, Investigation, Methodology, Project administration, Supervision, Writing—review and editing. K.M.K. and R.B.: Conducted the investigation, provided supervision, and contributed to the review and editing of the manuscript.

Funding

We do not receive any funding for this study.

Declarations

Competing of interest

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-28934-2.

Correspondence and requests for materials should be addressed to A.N. or R.B.