Data Article

Data on genome sequencing, assembly, annotation and genomic analysis of *Rhodococcus rhodochrous* strain SPC17 isolated from Lonar Lake

Satish Kumar a,b,*, Dhiraj Paul a,*, Yogesh Shouche a,***, Mangesh Suryavanshi c

a National Centre for Cell Science, Savitribai Phule Pune University Campus, Pune, MH, India
b ICAR-National Institute of Abiotic Stress Management, Baramati, Pune, MH, India
c Yenepoya Research Centre, Yenepoya Deemed to be University, Mangalore, 575018, Karnataka, India

**Abstract**

The bacterial isolates of genus *Rhodococcus* are best known for their significant biodegradation abilities. Here, we report the data related to draft genome sequencing of *Rhodococcus rhodochrous* strain SPC17 isolated from sediments of Lonar Lake. The de novo assembly of 1598096 Illumina's paired-end sequencing reads resulted in 51 contigs for an overall genome assembly size of 4.98Mb. A total of 4546 genes were predicted using the National Center for Biotechnology Information- Prokaryotic Genome Annotation Pipeline (NCBI-PGAP). RAST server-based annotation of the *Rhodococcus* strain SPC17 genome resulted in a total of 295 subsystems with 25% subsystem coverage. The data on the draft genome shotgun project are accessible at NCBI-GenBank under the accession number WUUR00000000. Our data resource will

*Corresponding author. National Centre for Microbial resource, National Centre for Cell Science, Savitribai Phule Pune University Campus, Pune, 411007, Maharashtra State, India.

**Corresponding author. National Centre for Cell Science, Savitribai Phule Pune University Campus, Pune, 411007, Maharashtra State, India.

***Corresponding author. National Centre for Cell Science, Savitribai Phule Pune University Campus, Pune, 411007, Maharashtra State, India.

E-mail addresses: satish.kumar10@icar.gov.in (S. Kumar), dhiraj.paul@nccs.res.in (D. Paul), yogesh@nccs.res.in (Y. Shouche).
facilitate further molecular and genomic studies of diverse hydrocarbon catabolizing genes present in *Rhodococcus rhodochrous* strain SPC17.

© 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

### 1. Data description

Bacterial isolates belonging to genus *Rhodococcus* are best known for their remarkable catabolic versatility and their ability to degrade a wide variety of aliphatic, aromatic, and polyaromatic hydrocarbons [1–3]. Taxonomically genus *Rhodococcus* belongs to the *Nocardiaceae* family under phylum Actinobacteria. The immense catabolic diversity and remarkable biotransformation capabilities of *Rhodococcus* have been largely attributed to the presence of a diverse array of catabolic genes encoding for several enzymatic classes like monoxygenases, dioxygenases, and hydroxylases [4]. In this data article, we present a description of genome sequencing data and the draft genome of *Rhodococcus*
The DNA was extracted from the pure culture of *R. rhodochrous* strain SPC17 and the genome sequencing was performed using the Illumina NovaSeq 6000 platform. A total of 1598096 paired-end sequencing reads were generated using Illumina’s 2 × 150 bp sequencing chemistry. The *de novo* genome assembly of 1598096 Illumina’s paired-end sequencing reads resulted in 51 contigs for an overall genome assembly size of 4.98Mb (4,985,784 bases). The quality parameters for the genome assembly such as N50 value (197398 bases), coarse consistency (97.6%), and fine consistency (96.1%) indicated the good quality of the assembled genome as evaluated using CheckM [6]. The genome of the *Rhodococcus* strain SPC17 has a high GC content of 68.39%. Using NCBI-PGAP (Prokaryotic Genome Annotation Pipeline) for genome annotation, it was determined that the genome contains 4484 coding sequences (CDSs), 4 rRNAs genes (1 copy of 5S, 1 copy16S, and 2 copy of 23S, respectively), 55 tRNAs, 3 ncRNAs (non-coding RNAs), and 60 pseudogenes. The genomic features of *Rhodococcus* strain SPC17 are represented in the circular map of the genome as shown in Fig. 1.

The RAST server-based annotation of the *Rhodococcus* strain SPC17 genome resulted in a total of 295 subsystems with 25% subsystem coverage (Fig. 2). The subsystem category distribution of the genes assigned to different subsystems indicated the highest genes assigned to metabolism of amino acids and derivatives (354 genes), followed by the metabolism of carbohydrates (246 genes), cofactors, vitamins, prosthetic group, and pigments (190 genes), fatty acids, lipids, and isoprenoids (175 genes) and protein metabolism (173 genes). RAST based annotation classified 77 genes in subsystem category ‘metabolism of aromatic compounds’ and 39 genes in subsystem category ‘stress response’. The 77 genes assigned to subsystem category ‘metabolism of aromatic compounds’ included genes related to Quinate degradation (2 genes), Biphenyl Degradation (10 genes), Benzoate degradation (6 genes), and

**Fig. 1.** Graphical circular view of the genome map of *Rhodococcus rhodochrous* strain SPC17. The tracks in the figure are displayed as concentric rings, from outermost to innermost tracks representing the 1-Reference position in the genome, 2-Position and order of the 51 assembled contigs, 3- CDS- forward strands, 4-CDS- reverse strand, 5-Non-coding features, 6-AMR genes (Anti-microbial Resistance), 7-Genes for Transporters, 8- GC content and 9- GC skew respectively.
p-Hydroxybenzoate degradation (1 gene), Catechol branch of beta-ketoadipate pathway (12 genes), Salicylate and gentisate catabolism (6 genes), Protocatechuate branch of beta-ketoadipate pathway (14 genes), Homogentisate pathway of aromatic compound degradation (21 genes), and Gentisate degradation (5 genes). The 12 genes assigned to branched ketoadipate pathway in Rhodococcus strain SPC17 implies its ability to catabolize the benzoate and phthalate compounds. Total 39 genes were assigned to subsystem category 'Stress response' and the analysis of the genes for this subsystem category revealed the presence of genes betA (encoding for Choline dehydrogenase (EC 1.1.99.1), betT (encoding a high-affinity choline uptake protein BetT) and opuD (encoding for glycine betaine transporter OpuD). This implies the ability of the isolate to synthesize and uptake ‘Glycine betaine (N,N,N-trimethylglycine)’ which may be the main osmolyte synthesized by this bacterium under osmotic stress conditions usually present in the soda lake environment. The details of the RAST-based annotation are given in supplementary sheet-S1. The genomic data reported here will pave the way for further study of the different catabolic genes implicated in hydrocarbon metabolism in this bacterium.

2. Experimental design, materials, and methods

2.1. Bacterial strain isolation and identification

The bacterial isolate SPC17 was isolated from the sediment sample of Lonar Lake on Standard Plate Count (SPC) microbiological medium as described by Watve and coworkers [7]. The strain SPC17 was identified as Rhodococcus rhodochrous based on the mass spectra of the cell extracts of overnight grown pure cultures following the method described by Pandey and coworkers [8]. The mass spectra were acquired in a linear positive ion extraction mode within a mass range from 2000 to 20,000 Da. using AUTOFLEX TOF/TOF Mass spectrometer (Bruker Daltonik GmbH, Germany). The MALDI Biotyper software 3.0 (Bruker Daltonik) was used to visualize and match the mass spectra which resulted in the identity of the pure culture as Rhodococcus rhodochrous with the biotyper score value 2.2.

2.2. Genome sequencing, assembly, and annotation

For genome sequencing, bacterial DNA was extracted from an overnight grown pure culture of Rhodococcus rhodochrous strain SPC17 using MoBio microbial DNA extraction kit (MoBio Laboratories). The DNA quality was checked on 1.0% agarose gel. The library preparation for Illumina sequencing was performed using the TruSeq DNA Library Preparation Kit (Illuminalnc., USA) according to the

![Fig. 2. An overview of the subsystem categories assigned to the genes predicted in the genome of Rhodococcus rhodochrous SPC17. The whole-genome sequence of the strain SPC17 was annotated using the RAST server.](image-url)
manufacturers instructions. The library insert sizes selected were 300 bases for Illumina sequencing.
The genome was sequenced using the Illumina’s Novaseq platform with $2 \times 150$ paired-end chemistry.
The quality of the raw reads was analyzed by FastQC software [9]. Trimming of the raw reads was done by
Trimmomatic 0.32 software [10] and only the good quality reads were assembled using de novo assembler
SPAdes 3.10.0 [11]. The annotation was carried out with the National Center for Biotechnology Information–
Prokaryotic Genome Annotation Pipeline (NCBI PGAP) [12] through the NCBI Genome submission portal (Genome Submit at http://ncbi.nlm.nih.gov). A genome circular map was
created using the ‘circular viewer’ functionality implemented in the PATRIC web server [13]. The
mapping of genes to subsystems and metabolic reconstruction of the high-quality genome annotations
were obtained using Rapid Annotation using the Subsystem Technology (http://rast.nmpdr.org) [14].

Author contributions

Conceived and designed the experiments: DP, YS, SK. Wet lab experiment: SK, DP, MVS, Data
analysis, and interpretation: SK, DP, Manuscript preparation: SK, DP. Contributed reagents/materials/
analysis tools: YS, MVS. All authors read and approved the final manuscript.

Acknowledgments

Author DP acknowledges the Department of Science and Technology (DST), Govt. of India for
awarding DST Young Scientist fellowship and providing funding for the work (SERB project ID: YSS/
2015/000513). Author SK acknowledges the Indian Council of Agricultural Research for sponsorship
during the study program. The help of Dr. Swapnil Prakash Doijad Biomedizinisches Forschungszentrum Seltersberg towards genome sequencing is thankfully acknowledged.

Conflict of Interest

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2020.105336.

References

[1] M.J. Larkin, L.A. Kulakov, C.C. Allen, Biodegradation and Rhodococcus—masters of catabolic versatility, Curr. Opin. Biotech. 16 (2005) 282–290.
[2] L. Martinkova, B. Uhnikova, M. Patek, J. Neavera, V. Kaen, Biodegradation potential of the genus Rhodococcus, Environ. Int. 35 (2009) 162–177.
[3] J. Zampolli, Z. Zeaiter, A. Di Canito, P. Di Gennaro, Genome analysis and omics approaches provide new insights into the biodegradation potential of Rhodococcus, Appl. Microbiol. Biotechnol. 103 (2019) 1069–1080.
[4] M.P. McLeod, R.L. Warren, W.W. Hsiao, N. Araki, M. Myhre, C. Fernandes, D. Miyazawa, W. Wong, A.L. Lillquist, D. Wang, M. Dosanjh, The complete genome of Rhodococcus sp. RHA1 provides insights into a catabolic powerhouse, Proc. Natl. Acad. Sci. U.S.A. 17 (103) (2006) 15582–15587.
[5] D. Paul, S.V. Kumbhare, S.S. Mhatre, S.P. Chowdhury, S.A. Shetty, N.P. Marathe, S. Bhute, Y.S. Shouche, Exploration of microbial diversity and community structure of Lonar Lake: the only hypersaline meteorite crater lake within basalt rock, Front. Microbiol. 22 (6) (2016) 1553.
[6] D.H. Parks, M. Imelfort, C.T. Skennerton, P. Hugenholtz, G.W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes, Genome Res. 1 (7) (2015) 1043–1055, 25.
[7] M. Wayne, V. Shejwal, C. Sonawane, M. Rahalkar, A. Matapurkar, Y. Shouche, M. Patole, N. Phadnis, A. Chumphenkar, K. Damle, S. Karandikar, The ‘K’selected oligophilic bacteria: a key to uncultured diversity? Curr. Sci. 25 (2000) 1535–1542.
[8] A. Pandey, R. Jain, A. Sharma, K. Dhakar, G.S. Kaira, P. Rahi, A. Dhyan, N. Pandey, P. Adhikari, Y.S. Shouche, 16S rRNA gene sequencing and MALDI-TOF mass spectrometry based comparative assessment and bioprospection of psychrotolerant bacteria isolated from high altitudes under mountain ecosystem, SN Appl. Sci. 1 (1) (2019) 278.
[9] S. Andrews, Babraham bioinformatics-FastQC a quality control tool for high throughput sequence data, URL: https://www.
bioinformatics.babraham.ac.uk/projects/fastqc, 2010.
[10] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 1 (30) (2014) 2114–2120.
[11] A. Bankevich, et al., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (2012) 455–477.

[12] T. Tatusova, M. DiCuccio, A. Badretdin, V. Chetvernin, E.P. Nawrocki, L. Zaslavsky, A. Lomsadze, K.D. Pruitt, M. Borodovsky, J. Ostell, NCBI prokaryotic genome annotation pipeline, Nucl. Acids Res. 24 (44) (2016) 6614–6624.

[13] A.R. Wattam, et al., Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center, Nucl. Acids Res. 45 (2016) 535–542.

[14] R. Overbeek, R. Olson, G.D. Pusch, G.J. Olsen, J.J. Davis, T. Disz, R.A. Edwards, S. Gerdes, B. Parrello, M. Shukla, V. Vonstein, The SEED and the Rapid annotation of microbial genomes using subsystems Technology (RAST), Nucl. Acids Res. 42 (2014) D206–D214.