BMC Genomic Data

DATA NOTE Open Access

Unveiling complete genome sequence of *Streptomyces griseoincarnatus* AR2 from Katpana cold desert: the roof of the world surrounded by Karakoram mountain range

Arisha Rameen Bhatti[1], Noor Hassan[1*], Hazrat Ali[1], Noor ul Huda[1], Habib Ur Rahman[1], Javeria Anwar[1], Anam Saqib[1], Mohsin Khan[2] and Aiman Umar[3]

Abstract

Objectives The genus *Streptomyces* encompasses a wide range of Gram-positive bacteria, holding significant biotechnological potential. *Streptomyces* species have been recognized for prolific production of multiple bioactive secondary metabolites including pigments, manifesting therapeutic potential and industrial applications. However, bioactive potential of secondary metabolites produced by *Streptomyces griseoincarnatus* remains relatively unexplored. In this study, we isolated a *S. griseoincarnatus* AR2 and generated the complete genome for exploration of biosynthetic gene clusters, hence revealing secondary metabolite potential of species.

Data description *S. griseoincarnatus* AR2 was isolated from Katpana Cold Desert in Skardu Valley, Gilgit-Baltistan, Pakistan. The complete genome assembly of AR2 comprised of 7,307,931 bp, one contig, G + C content of 72.39%, presenting an N50 of 7,307,931 bp and an L50 of 1. However, functional annotation via RAST server revealed 6647 coding sequences (CDS), 85 RNAs and 418 subsystems in AR2 genome. Phylogenetic analysis using Type Strain Genome Server (TYGS) revealed alignment of species AR2 with reference sequence of *S. griseoincarnatus* (AC# CP139576.1), with bacterial orthoANI value of 98.19% and dDDH value in range between 80.9 and 86.1%. Whereas, antiSMASH v8.0.1 highlighted the presence of 23 putative secondary metabolite biosynthetic gene clusters in the genome. The respective genome sequence of AR2 provides a promising resource for exploration of biosynthetic and metabolic pathways, secondary metabolites pathway, advancing application in pharmaceutical and industrial research.

Keywords *Streptomyces griseoincarnatus*, Genome assembly, Secondary metabolites, Biosynthetic gene clusters

*Correspondence:
Noor Hassan
noorhassan672@gmail.com
[1]Fermentation Biotechnology for Industrial Products Group, Industrial Biotechnology Division, National Institute for Biotechnology and Genetic Engineering–College, Pakistan Institute of Engineering and Applied Sciences (NIBGE–C, PIEAS), Faisalabad, Pakistan
[2]Department of Biological Sciences, Ohio University, Athens, USA
[3]Department of Food and Nutrition, Faculty of Sciences, Shaheed Benazir Bhutto Women University Peshawar, Peshawar, Pakistan

Objective

The genus *Streptomyces*, a largest group of actinomycetes, was first designated in 1943, comprising of 700 reported species. *Streptomyces* species are free living, gram positive, obligate aerobe, filamentous and non-motile bacteria, majorly present in terrestrial environment forming branched mycelium and spores on aerial hyphae [1, 2]. *Streptomyces* harbors a linear genome, various biosynthetic gene clusters (BGCs) and considerable metabolic diversity, producing a range of valuable secondary metabolites. Unlike other bacterial genus, BGCs in *Streptomyces* as modular polyketide synthases (PKS) and non-ribosomal peptide synthetases (NRPS) encode majority of antibiotics including streptomycin, tetracycline, chloramphenicol, and rifamycin [3]. Additionally, *Streptomyces* encode genes for bioactive metabolites as antitumor, anticancer, antifungal and immunosuppressive agents. The exceptional biosynthetic potential of *Streptomyces* species renders a marked position in drug discovery, synthetic biology and industrial research [3, 4].

Among various *Streptomyces* species, *S. griseoincarnatus*, a mesophilic bacteria forming aerial mycelium presents a limited research on biosynthesis of secondary metabolites and genomic exploration [5]. In current study, complete genome of *S. griseoincarnatus* AR2, isolated from Katpana dessert, exhibiting significant secondary metabolite production, was sequenced and assembled. The complete genome sequencing through hybrid Illumina and ONT (Oxford Nanopore Technologies) platform, predicted 23 putative region of biosynthetic gene clusters for secondary metabolite production. Significantly, genomic data generated in this study provides a valuable foundation for further research on exploration of biosynthetic and metabolic pathways, regulatory mechanisms, pathway engineering for sustainable production of novel bioactive compounds and promising biotechnological exploitation of *S. griseoincarnatus* AR2.

Data description

The *S. griseoincarnatus* AR2 was isolated from the samples, collected (collection date: 2024-07-20) from Katpana Cold Desert in Skardu Valley (35.310522°N and 75.590747°E), Gilgit-Baltistan, Pakistan. Pure culture was obtained by incubating the collected samples in ISP4 medium (g/L soluble starch 10, K_2HPO_4, $MgSO_4.7H_2O$, NaCl, $(NH_4)_2SO_4$, $CaCO_3$, and trace elements e.g. $FeSO_4.7H_2O$, $MnCl_2.7H_2O$, and $ZnSO_4.7H_2O$) at 28 °C for 14 days and morphologically identified the AR2 as distinct filamentous structure with branched organization and spore formation [6]. The extraction of genomic DNA was carried out according to DNA sample preparation guidelines of MicrobesNG, Birmingham, UK. DNA extraction of isolate AR2 was performed using GeneJET Genomic DNA purification Kit (Thermo Scientific). After delivery to MicrobesNG, extracted DNA was stored and resuspended in DNA shield (Zymo Research, USA), followed by cell lysis using TE buffer containing lysozyme (MPBio, USA), metapolyzyme (Sigma Aldrich, USA), and RNase A (ITW Reagents, Spain). Extracted DNA was quantified using Quant-iT dsDNA HS (ThermoFisher Scientific) assay in an Eppendorf AF2200 plate reader (Eppendorf UK Ltd, United Kingdom) and diluted for library preparation.

Genomic DNA libraries were prepared using Hybrid Sequencing Platform (Illumina and Oxford nanopore), integrating both long-read and short-read with target coverage of 30X Illumina and 50X ONT. Nextera XT Library Prep Kit (Illumina, USA) was used for sequencing of Illumina short-reads library sequencing with paired-end (2 × 250 bp) configuration. However, library preparation for long read sequencing was performed via Ligation Sequencing Kit protocol as SQK-RBK114.96 kit (ONT, UK), employing 200–400 ng of genomic DNA. Barcoded samples were pooled and sequenced on GridION (ONT, UK) platform using FLO-MIN114 (R10.4.1) flow cells with base-calling model r10.4.1_e8.2_400bps_hac_v4.3.0. Hamilton Microlab STAR automated liquid handling system was used for quantification, automation and library preparation steps.

Initial quality assessment of raw Illumina and ONT reads was carried out using FastQC [7] and Nanofilt v2.8.0 [8] for long ONT reads. Adapter sequences of Illumina and ONT reads were trimmed using Trimmomatic version 0.40 [9] and Porechop-v0.2.3 [10]. Moreover, SPAdes version 4.2.0 was employed for de novo hybrid genome assembly. High quality Illumina reads and nanopore long reads were generated using R10.4.1 chemistry, assembled into contings using Unicycler version 0.4.8 for complete assemblies [11]. Furthermore, the hybrid assembly with both Illumina and ONT data was polished by running four rounds of Racon with the ONT data, followed by Pilon and Racon with the Illumina data [12]. Hybrid quality was assessed through QUAST tool integrated in Galaxy database [13]. The completeness of the genome assembly has been tested with BUSCO v5.8.0 [14]. Functional annotation of AR2 genome was studied through PATRIC database [15] and genetic circuits for major subsystems were comprehensively studied through RAST server [16].

Furthermore, the analysis of AR2 genomic assembly revealed 7,307,931 bp length with 72.39% GC content, single contig, an N50 of 7,307,931 and an L50 of 1, 99.7% completeness index, an estimated coverage of 29.2× for Illumina and 50.0× for ONT (Table 1, Data File 1 and 2). However, functional annotation via RAST server identified 6647 coding sequences, 85 RNAs, 418 subsystems in AR2 genome assembly, including predominant subsystems with corresponding gene counts; carbohydrates

Table 1 Overview of data files/data sets

Label	Name of data file/ data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	The circular maps of the AR2 genome	Figure file (.png)	Figshare, (https://doi.org/10.6084/m9.figshare.29828582) [24]
Data file 2	Genomic feature / characteristics of AR2	Table file (.docx)	Figshare, (https://doi.org/10.6084/m9.figshare.29828585) [25]
Data file 3	Subsystems predicted in AR2 genome by using RAST Server	Figure file (.png)	Figshare, (https://doi.org/10.6084/m9.figshare.29828600) [26]
Data file 4	Phylogram representing relationship with closely related ancestors by PhyML tool	Figure file (.png)	Figshare, (https://doi.org/10.6084/m9.figshare.30256135) [27]
Data file 5	Pairwise comparisons of AR2 genomes vs. type strain genomes	Table file (.docx)	Figshare, (https://doi.org/10.6084/m9.figshare.30256096) [28]
Data file 6	Summary of Average Nucleotide Index of AR2 with the reference genome *S. griseoincarnatus*	Table file (.docx)	Figshare, (https://doi.org/10.6084/m9.figshare.29828603) [29]
Dataset 7	Illumina reads of the *S. griseoincarnatus* AR2	SRA file (.sra)	NCBI Sequence Read Archive, (http://identifiers.org/insdc.sra:SRR34999931) [30]
Dataset 8	GenBank accession number of the *S. griseoincarnatus* AR2	Fasta (.fasta)	GenBank Accession Number, (http://identifiers.org/nucleotide:JBRKBJ000000000) [31]

(562), membrane transport (88), cofactors, vitamins, prosthetic groups, pigments (258), motility and chemotaxis (8), RNA metabolism (118), virulence, disease and defense (65), metabolism of aromatic compounds (36), regulation and cell signaling (66), stress response (173), respiration (151), phages, prophages, transposable elements, plasmids (4), cell wall and capsule (156), fatty acids, lipids, and isoprenoids (191) (Data File 3).

The 16S rRNA evolutionary marker-based taxonomic classification of species AR2 was performed through ContEst16S tool [17]. The same tool was also used to determine level of contamination of the genome. Similarity search of query sequence on BLASTn at NCBI [18] and multiple sequence alignment through MAFFT tool with closely related sequences from NCBI [19], identified AR2 as *S. griseoincarnatus*, indicated by 16 S rRNA gene fragment-based BLAST outcomes with 100% coverage. The resulted aligned sequences were used for phylogram construction using PhyML tool available at NGphyloeny [20], confirming genomic similarity with *S. griseoincarnatus* (Data File 4). In addition, the analysis of16S rRNA gene fragments implied that the genome was not contaminated, therefore confirming the fact that there was no significant occurrence of extraneous DNA in the composite. Moreover, ANI index calculated using ANI calculator [21], revealed ANI index of 98.19% with reference sequence of *S. griseoincarnatus* (AC# CP139576.1). The dDDH value was evaluated through genome-to-genome distance calculator (GGDC) tool, built in TYGS for species-level identification [22] and further validation of *S. griseoincarnatus* AR2 as dDDH study based on d0, d4 and d6 formula. The dDDH value (< 70) for AR2 was recorded, ranging between 80.9 and 86.1% (Data File 5). OrhtoANIu value of species AR2 was calculated as 98.18%, exhibiting genomic length of 7,307,280 bp and 68.3% coverage, aligning with reference genome *S. griseoincarnatus* (AC# CP139576.1) of 7,522,500 bp and 66.40% coverage (Data File 6). The genome sequence of *S. griseoincarnatus* AR2 was successfully submitted and available under GenBank accession number (JBRKBJ000000000, Table 1, Datasets 7, 8).

The antiSMASH pipeline, v8.0.1(Bacterial version) evaluated presence of BGCs in AR2 genome assembly. The annotated genome was uploaded on antiSMASH web server [23], with activated detection modules including ClusterBlast, SubClusterBlast, ActiveSiteFinder, and KnownClusterBlast. Computational analysis of BGCs associated with de-novo microbial secondary metabolites production predicted 23 putative regions, particularly 6 regions of hypothetical group, 12 regions encompassing NRPS/ T1PKS, T2PKS, ectoine, NI siderophore, lassopeptide, terpenes and lanthipeptide-class-iii with highest similarity confidence. Whereas, 2 regions of RiPP and terpene exhibiting medium value and 3 regions of hydrogen cyanide, RiPP and NRPS/T1PKS manifesting low confidence value.

Limitations

Although the current study provides valuable insights into the bioactive secondary metabolites potential of *S. griseoincarnatus* AR2, there are still rooms for complete identification of secondary metabolites. The antiSMASH pipeline, v8.0.1 (Bacterial version) predicted several BGCs in species AR2 genome but many were identified as unknown. There is need to be confirmed the real biological activity and production of these metabolites.

Abbreviations

RAST	Rapid Annotation using Subsystem Technology
antiSMASH	Antibiotics and Secondary Metabolite Analysis Shell
dDDH	Digital DNA-DNA Hybridization
NCBI	National Centre for Biotechnology
QUAST	Quality Assessment Tool for Genome Assemblies
PATRIC	Pathosystems Resource Integration Center
MAFFT	Multiple Alignment using Fast Fourier Transform
BLAST	Basic Local Alignment Search Tool
ANI	Average Nucleotide Identity

Author contributions

ARB, NH, AU and HA: performed species isolation, cultivation and DNA extraction. NH, AS, MK and HR: performed the genome analysis. ARB and JA prepared the manuscript draft. NH: supervised the project, designed the

experiments, and edited the manuscript. The authors read and approved the final manuscript.

Funding
No funding.

Data availability
The genomic data described in this Data note can be freely and openly accessed on GenBank of NCBI under the accession no | [JBRKBJ000000000](http://identifiers.org/nucleotide: JBRKBJ000000000) | [31]. In addition, the SRA accession of the AR2 species is | [SRR34999931](http://identifiers.org/insdc.sra:SRR34999931) | [28], and the BioProject number is | [PRJNA1298538](https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1298538) | as well as BioSample number is | [SAMN50278237](https://www.ncbi.nlm.nih.gov/biosample/?term=SAMN50278237) |. Associated Data Files are available on Figshare: The circular maps of the AR2 genome [24]; Genomic feature / characteristics of AR2 [25]; Subsystems predicted in AR2 genome by using RAST Server [26], Phylogram [27], Pairwise comparisons of species AR2 genomes vs. type strain genomes [28] and Average Nucleotide Index of AR2 [29].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 1 October 2025 / Accepted: 9 December 2025
Published online: 22 December 2025

References

1. Williams ST. Genus *Streptomyces* waksman and henrici 1943. In: Williams ST, Sharpe ME, Holt JG, editors. Bergey's manual of systematic bacteriology. 1989;4:2452-92.
2. Anderson AS, Wellington EM. The taxonomy of *Streptomyces* and related genera. Int J Syst Evol Microbiol. 2001;51(3):797–814.
3. Chater KF. Recent advances in Understanding *Streptomyces*. F1000Res. 2016;5:2795.
4. Belknap KC, Park CJ, Barth BM, Andam CP. Genome mining of biosynthetic and chemotherapeutic gene clusters in *Streptomyces* bacteria. Sci Rep. 2020;10(1):2003.
5. Reimer LC, Sardà Carbasse J, Koblitz J, Ebeling C, Podstawka A, Overmann J. BacDive in 2022: the knowledge base for standardized bacterial and archaeal data. Nucleic Acids Res. 2022;50(D1):D741–6.
6. Pridham TG, Hesseltine CW, Benedict RG. A guide for the classification of *streptomycetes* according to selected groups; placement of strains in morphological sections. Appl Microbiol. 1958;6(1):52–79.
7. Wingett SW, Andrews S. FastQ screen: A tool for multi-genome mapping and quality control. F1000Res. 2018;7:1338.
8. Older CE, Richardson BM, Wood M, Waldbieser GC, Ware C, Griffin MJ, Ott BD. Evaluating nanopore sequencing for microbial community characterization in catfish pond water. J World Aquac Soc. 2023. https://doi.org/10.1111/jwas.13002.
9. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
10. Xu F, Ge C, Li S, Tang S, Wu X, Luo H, Deng X, Zhang G, Stevenson A, Baker RC. Evaluation of nanopore sequencing technology to identify *Salmonella enterica* choleraesuis var. Kunzendorf and Orion var. 15+, 34. Int J Food Microbiol. 2021;346:109167.
11. Tatapudi MR, Naveena Lavanya Latha J, Girija AS. Whole genome sequencing and analysis of Pan-Drug resistant Acinetobacter baumannii using a hybrid de Novo approach (Illumina and Nanopore). 2024; https://doi.org/10.2139/ssrn.4978816
12. Breckell GL, Silander OK. Do You Want to Build a Genome? Benchmarking Hybrid Bacterial Genome Assembly Methods. bioRxiv. 2021. https://www.biorxiv.org/content/10.1101/2021.11.07.467652v1.
13. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-5.
14. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38:4647–54.
15. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, Machi D, Mao C, Nordberg EK, Olson R, Overbeek R, Pusch GD, Shukla M, Schulman J, Stevens RL, Sullivan DE, Vonstein V, Warren A, Will R, Wilson MJ, Yoo HS, Zhang C, Zhang Y, Sobral BW. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 2014;42:D581–91.
16. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R. The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res. 2014;42:D206–14.
17. Lee I, Chalita M, Ha SM, Na SI, Yoon SH, Chun J. ContEst16S: an algorithm that identifies contaminated prokaryotic genomes using 16S RNA gene sequences. Int J Syst Evol Microbiol. 2017;67(6):2053–7.
18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
19. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res. 2002;30(14):3059–66.
20. Guindon S, Lethiec F, Duroux P, Gascuel O. PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. 2005;33:W557–9.
21. Yoon SH, Ha SM, Lim J, Kwon S, Chun J. A large-scale evaluation of algorithms to calculate average nucleotide identity. Antonie Van Leeuwenhoek. 2017;110(10):1281–6.
22. Meier-Kolthoff JP, Carbasse JS, Peinado-Olarte RL, Göker M. TYGS and LPSN: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. Nucleic Acids Res. 2022;50(D1):D801–7.
23. Blin K, Shaw S, Vader L, Szenei J, Reitz ZL, Augustijn HE, Cediel-Becerra JDD, de Crécy-Lagard V, Koetsier RA, Williams SE, Cruz-Morales P, Wongwas S, Segurado Luchsinger AE, Biermann F, Korenskaia A, Zdouc MM, Meijer D, Terlouw BR, van der Hooft JJJ, Ziemert N, Helfrich EJN, Masschelein J, Corre C, Chevrette MG, van Wezel GP, Medema MH, Weber T. AntiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. Nucleic Acids Res. 2025;53(W1):W32–8.
24. Bhatti AR, Hassan N, Ali H, Huda N, Rahman H, Anwar J, Saqib A, Khan M, Umar A. Unveiling complete genome sequence of *Streptomyces Griseoincarnatus* AR2 from Katpana cold desert: the roof of the world surrounded by Karakoram mountain range. Figshare. 2025. https://doi.org/10.6084/m9.figshare.29828582
25. Bhatti AR, Hassan N, Ali H, Huda N, Rahman H, Anwar J, Saqib A, Khan M, Umar A. Unveiling complete genome sequence of *Streptomyces Griseoincarnatus* AR2 from Katpana cold desert: the roof of the world surrounded by Karakoram mountain range. Figshare. 2025. https://doi.org/10.6084/m9.figshare.29828585
26. Bhatti AR, Hassan N, Ali H, Huda N, Rahman H, Anwar J, Saqib A, Khan M, Umar A. Unveiling complete genome sequence of *Streptomyces Griseoincarnatus* AR2 from Katpana cold desert: the roof of the world surrounded by Karakoram mountain range. Figshare. 2025. https://doi.org/10.6084/m9.figshare.29828600
27. Bhatti AR, Hassan N, Ali H, Huda N, Rahman H, Anwar J, Saqib A, Khan M, Umar A. Unveiling complete genome sequence of *Streptomyces Griseoincarnatus* AR2 from Katpana cold desert: the roof of the world surrounded by Karakoram mountain range. Figshare. 2025. https://doi.org/10.6084/m9.figshare.30256135
28. Bhatti AR, Hassan N, Ali H, Huda N, Rahman H, Anwar J, Saqib A, Khan M, Umar A. Unveiling complete genome sequence of *Streptomyces Griseoincarnatus* AR2 from Katpana cold desert: the roof of the world surrounded by Karakoram mountain range. Figshare. 2025. https://doi.org/10.6084/m9.figshare.30256096
29. Bhatti AR, Hassan N, Ali H, Huda N, Rahman H, Anwar J, Saqib A, Khan M, Umar A. Unveiling complete genome sequence of *Streptomyces Griseoincarnatus* AR2 from Katpana cold desert: the roof of the world surrounded by

Karakoram mountain range. Figshare. 2025. https://doi.org/10.6084/m9.figshare.29828603
30. Bhatti AR, Hassan N, Ali H, Huda N, Rahman H, Anwar J, Saqib A, Khan M, Umar A. Unveiling complete genome sequence of *Streptomyces Griseoincarnatus* AR2 from Katpana cold desert: the roof of the world surrounded by Karakoram mountain range. 2025. http://identifiers.org/insdc.sra:SRR34999931
31. Bhatti AR, Hassan N, Ali H, Huda N, Rahman H, Anwar J, Saqib A, Khan M, Umar A. Unveiling complete genome sequence of *Streptomyces Griseoincarnatus* AR2 from Katpana cold desert: the roof of the world surrounded by Karakoram mountain range. 2025. http://identifiers.org/nucleotide:JBRKBJ000000000

Publisher's note