



OPEN ACCESS

EDITED BY

Sara Calhoun,
Berkeley Lab (DOE), United States

REVIEWED BY

Shakira Ghazanfar,
National Institute for Genomics and Advanced
Biotechnology (NIGAB), Pakistan
Nayla Munawar,
United Arab Emirates University,
United Arab Emirates

*CORRESPONDENCE

Ngoc Bich Pham
✉ pbngoc@ib.ac.vn

RECEIVED 29 August 2025

REVISED 28 December 2025

ACCEPTED 06 January 2026

PUBLISHED 02 February 2026

CITATION

Nguyen DH, Doan NT and Pham NB (2026)
Complete genome of *Penicillium herquei*
HGN12.1C isolated from *Dysosma*
difformis in Vietnam.
Front. Fungal Biol. 7:1695025.
doi: 10.3389/ffunb.2026.1695025

COPYRIGHT

© 2026 Nguyen, Doan and Pham. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Complete genome of *Penicillium herquei* HGN12.1C isolated from *Dysosma difformis* in Vietnam

Duong Huy Nguyen¹, Nhung Thi Doan¹ and Ngoc Bich Pham^{1,2*}

¹Institute of Biology (IB), Vietnam Academy of Science and Technology (VAST), Hanoi, Vietnam,

²Graduate University of Science and Technology (GUST), Vietnam Academy of Science and
Technology (VAST), Hanoi, Vietnam

KEYWORDS

Penicillium herquei, *Dysosma difformis*, endophytic fungus, genome assembly and
annotation, podophyllotoxin

Introduction

Penicillium herquei HGN12.1C is an endophytic fungus isolated from *Dysosma difformis*, a native medicinal plant found in the mountainous provinces of northern Vietnam. Previous studies have demonstrated that *Penicillium herquei* HGN12.1C (Nguyen et al., 2024), along with several other endophytic fungi such as *Trametes*, *Purpureocillium*, *Aspergillus*, and *Ganoderma* (Tran et al., 2022), as well as the strain *Fusarium proliferatum* TQN5T (Nguyen et al., 2023), are capable of synthesizing podophyllotoxin (PTOX) and its derivatives.

PTOX is an important natural compound with numerous medical applications, including effective antimicrobial, antifungal, and anticancer properties (Ardalani et al., 2017; Yu et al., 2017). PTOX has been found in various plant species such as *Podophyllum hexandrum* (Anand et al., 2022), *Callitris intratropica* (Wanner et al., 2015), and *Juniperus horizontalis* Moench (Cantrell et al., 2014). However, studies on PTOX in fungi remain largely unexplored, particularly due to the lack of high-quality genome sequence data, which would provide valuable resources for comparative taxonomy as well as for investigating the biosynthetic pathways and mechanisms of bioactive compounds such as PTOX. Therefore, sequencing and exploring the genome of the fungal strain *Penicillium herquei* HGN12.1C is considered a crucial step for genomic analysis, comparative studies, and the identification of genes involved in the biosynthesis of PTOX and other bioactive compounds.

Here, we present the complete genome dataset of *Penicillium herquei* HGN12.1C to provide a valuable resource for comparative genomics and investigations of biosynthetic pathways.

Value of data

Previous studies have characterized the biochemical features and biosynthetic pathways of podophyllotoxin in plants. In this work, we provide a high-quality genome of an endophytic fungus capable of biosynthesizing PTOX. The genome has been assembled and

annotated with high-quality, enabling further exploration of taxonomic, genetic, and metabolic characteristics, as well as genes related to the biosynthetic pathways of secondary metabolites, especially PTOX, in fungi.

Dataset description

The complete genome consists of nine contigs with a total length of 34,960,691 bp, an average read depth of 195×, and a GC content of 46.38% (Figure 1). It has an N50 value of 6,172,464 bp, an L50 of three, and the largest contig measures 8,398,835 bp. Among these, one contig (62,425 bp) was identified as the mitochondrial genome (Supplementary Table S1, Supplementary Figure S1). The nuclear genome comprises eight contigs, consisting of seven chromosome-level assemblies and one smaller unplaced scaffold (Table 1).

The dataset includes BUSCO assessment results showing 98.4% completeness (1,291 of 1,312 core genes detected). Repetitive sequences identified using standard fungal libraries accounted for 20,550 bp, representing 0.06% of the total assembled genome size. It should be noted that this value likely underestimates the total repetitive content, as *de novo* repeat identification was not performed.

The dataset includes annotation files generated by the Funannotate pipeline integrating AUGUSTUS, GlimmerHMM, and SNAP predictors, with final gene counts summarized in Table 1. After removing overlapping and redundant genes, 11,054 protein-coding genes were identified. Combined with 249 tRNA genes and 19 rRNA genes, the *Penicillium herquei* HGN12.1C genome was annotated with a total of 11,322 genes (Table 1).

The complete mitochondrial genome of *Penicillium herquei* strain HGN12.1C consists of a circular DNA molecule of 29,635 bp in length (Supplementary Figure S1). The mitogenome encodes 44

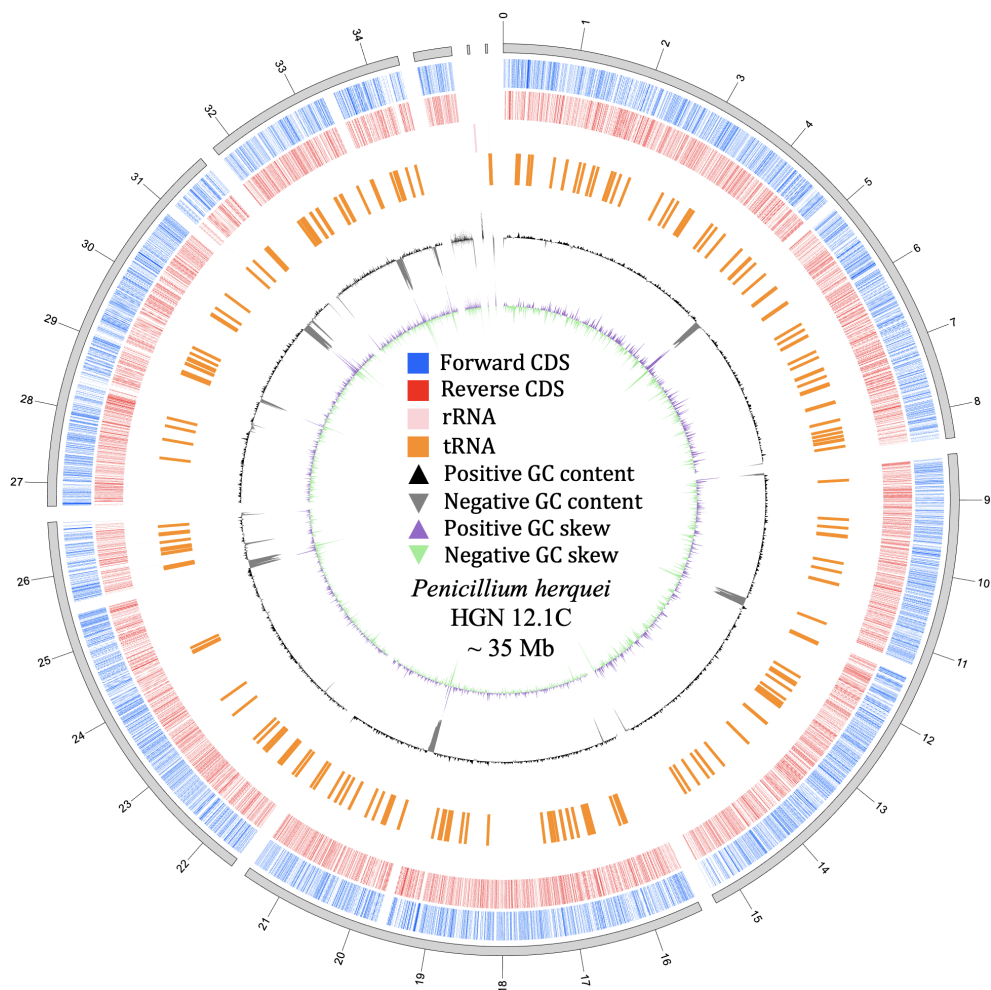


FIGURE 1

Structural overview of the *Penicillium herquei* HGN12.1C genome assembly. The eight nuclear chromosome-level contigs and one mitochondrial scaffold are concatenated for visualization purposes. The outermost ring indicates the genomic position scale in megabases (Mb). Concentric data tracks from the outside inward represent: (i) protein-coding genes (CDSs) on the forward strand (blue) and reverse strand (red); (ii) ribosomal RNA (rRNA) genes (pink); (iii) transfer RNA (tRNA) genes (orange); (iv) GC content deviation from the mean (black: positive; gray: negative); and (v) GC skew (purple: positive; green: negative).

TABLE 1 Summary of genomic features and assembly statistics for strain HGN12.1C.

Genome assembly features	Values
Total assembly size (bp)	34,960,691
Total assembly size ($\geq 1,000$ bp)	34,960,691
Total assembly size ($\geq 50,000$ bp)	34,903,660
Number of scaffolds ($\geq 1,000$ bp)	8*
Number of scaffolds ($\geq 50,000$ bp)	7
Number of organelles	1
Largest scaffold (bp)	8,398,835
Genome size (bp)	34,960,691
Mitochondrial size (bp)	29,635
GC-content (%)	46.38
N50 length chromosome (bp)	6,172,464
L50 chromosome count	3
Genome coverage	195×
Completeness evaluation (%)	
Completeness	98.4
Complete and single-copy BUSCOs	97.9
Complete and duplicated BUSCOs	0.5
Fragmented BUSCOs	0.8
Missing BUSCOs	0.8
Repeat annotation	
Simple repeat (bp)	4,164
DNA transposon (bp)	16,386
Total percentage of repeat sequences	0.06
Total repeat length (bp)	20,550
Genomic annotation	
<i>Predictors</i>	AUGUSTUS SNAP GlimmerHMM
Predicted raw genes	10,716 8,830 11,689
Predicted genes (after EVM)	11,054
Protein coding genes	11,054
tRNA	249
rRNA (5S, 5.8S, 18S, 28S)	19 (3, 5, 7, 4)
Annotated genes (total)	11,322

*The assembly consists of seven chromosome-level contigs and one unplaced scaffold.

genes, comprising 15 protein-coding genes (PCGs), two ribosomal RNA (rRNA) genes, and 27 transfer RNA (tRNA) genes (Supplementary Table S1). The overall base composition is 36.9% A, 38.0% T, 11.6% C, and 13.4% G, resulting in a low GC content of 25.0%, which is consistent with other *Penicillium* species.

The 15 PCGs include seven NADH dehydrogenase subunits (*nad1–6*, *nad4L*), one cytochrome *c* reductase subunit (*cob*), three cytochrome *c* oxidase subunits (*cox1–3*), three ATP synthase subunits (*atp6*, *atp8*, *atp9*), and one ribosomal protein (*rps5*). The total length of these PCGs is approximately 16,500 bp, accounting for 55.7% of the entire mitogenome. The two rRNA genes include the small subunit (*rns*) and the large subunit (*rnl*). Notably, the *rnl* gene is fragmented into exons by the insertion of the *rps5* gene and introns. The set of 27 tRNA genes ranges in length from 71 to 86 bp and is sufficient to decode all 20 standard amino acids. The complete mitogenome sequence of *Penicillium herquei* strain HGN12.1C has been submitted to the GenBank database under accession number CP098395.1.

The dataset enables researchers to investigate fungal genomics, comparative analyses, and secondary metabolite biosynthesis pathways. All data files are publicly accessible and can be downloaded from the respective repositories for further analysis and validation.

Materials and methods

Fungal strain culture and isolation

The endophytic fungal strain was isolated from *D. difformis* collected in Ha Giang Province, Vietnam, and was preserved at -20° C for long-term storage at the Vietnam Academy of Science and Technology (VAST)—Culture Collection of Microorganisms (VCCM code: VCCM44283). For experimental procedures, the fungus was subcultured and maintained at 25° C. The isolation and cultivation were carried out according to the method previously described by Tran et al. (2022).

DNA extraction and genome sequencing

In this study, fungal genomic DNA was extracted using the E.Z.N.A.[®] Fungal DNA Mini Kit (OMEGA, USA) following the manufacturer's instructions. The quantity and quality of the extracted DNA were assessed using a Qubit[™] dsDNA HS Assay Kit (on a Qubit 3.0 Fluorometer) and by running 0.8% agarose gel electrophoresis, respectively. Library preparation was performed with the SMRTbell Express Template Prep Kit 2.0 (PacBio), followed by polymerase attachment and purification using the Sequel Binding and Internal Ctrl Kit 3.0 (PacBio). Finally, the fungal genome was sequenced using the PacBio Sequel platform with the Sequel SMRT Cell 1M v3 Tray chip and the Sequel Sequencing Kit 3.0 (PacBio).

Preprocessing and genome assembly

Read quality control and *de novo* genome assembly were performed using FastQC (<http://www.bioinformatics>).

babraham.ac.uk/projects/fastqc/) and the Hierarchical Genome-Assembly Process (HGAP) version 4 (Chin et al., 2013). Genome completeness was evaluated using BUSCO version 5.8.0 (Simão et al., 2015) based on 1,312 single-copy ortholog groups from the fungi_odb10 database, while assembly parameters were assessed using QUAST version 5.3.0 (Gurevich et al., 2013). Repeat sequences in the fungal genome were detected using RepeatMasker version 4.2.1 (<https://www.repeatmasker.org/>) (Tarailo-Graovac and Chen, 2009).

Genome annotation and functional analyses

Genome annotation and gene prediction were performed using the Funannotate pipeline version 1.8.17 (Palmer and Stajich, 2020), which integrates multiple gene predictors, including AUGUSTUS version 3.5.0 (Stanke and Morgenstern, 2005), GlimmerHMM version 3.0.4 (Majoros et al., 2004), and the semi-HMM-based Nucleic Acid Parser (SNAP version 2.68.5+ubuntu24.04.1) (Korf, 2004). Within this pipeline, AUGUSTUS was trained using BUSCO genes, with the *Aspergillus nidulans* species model used as a closely related reference to predict gene structures with high accuracy. Additionally, AUGUSTUS gene prediction models were refined using high-quality protein sequence datasets from two closely related *Penicillium* species, *Penicillium chrysogenum* Wisconsin 54-1255 (GCF_000226395.1) and *Penicillium rubens* Wisconsin 54-1255 (GCA_000226395.1), which served as training references to enhance the accuracy of protein-coding gene prediction and annotation specific to the *Penicillium* genus. GlimmerHMM was employed to identify genes with simple exon–intron structures and to detect short genes, while SNAP was used to predict genes with complex structures and alternative splicing patterns, trained using the BUSCO Dikarya lineage dataset. DIAMOND version 2.1.11 was used to filter low-quality gene models, and Evidence Modeler version 2.1.0 (EVM) (Haas et al., 2008) was applied to remove redundant genes and integrate predictions from the three gene predictors into a comprehensive genome annotation. Final gene predictions were based on consensus results to maximize accuracy and completeness. Transfer RNA and ribosomal RNA genes were annotated using tRNAscan-SE version 2.0.12 (<https://github.com/UCSC-LoweLab/tRNAscan-SE>) and Barrnap version 0.9 (<https://github.com/tseemann/barrnap>), respectively.

The mitochondrial genome assembly was annotated locally using the MITOS2 pipeline (Donath et al., 2019) within a Conda environment. To ensure high-specificity annotation, analyses were conducted using the RefSeq 89f reference database (fungi-specific) and the Mold, Protozoan, and Coelenterate Mitochondrial Code (NCBI Translation Table 4). Annotation outputs were manually curated to ensure correct nomenclature for ribosomal RNAs (rns and rnl) and to resolve gene fragmentation boundaries. The circular physical map of the mitochondrial genome was generated using OrganellarGenomeDRAW (OGDRAW) (Greiner et al., 2019).

The complete genome of *Penicillium herquei* HGN12.1C was visualized as a Circos plot using the pyCircize package. (<https://moshi4.github.io/pyCircize/>).

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: NCBI BioProject PRJNA837748, BioSample SAMN54322392.

Author contributions

DN: Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. ND: Data curation, Formal analysis, Investigation, Methodology, Writing – original draft. NP: Conceptualization, Data curation, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research is funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 108.06-2025.109.

Acknowledgments

The authors gratefully acknowledge the Key Laboratory of Vaccine and Applied DNA Technology, Institute of Biology, Vietnam Academy of Science and Technology (VAST) for providing facilities and support for data analysis in this study.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial

intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Anand, U., Biswas, P., Kumar, V., Ray, D., Ray, P., Loake, V. I. P., et al. (2022). *Podophyllum hexandrum* and its active constituents: Novel radioprotectants. *Biomed. Pharmacother.* 146, 112555. doi: 10.1016/j.biopha.2021.112555
- Ardalani, H., Avan, A., and Ghayour-Mobarhan, M. (2017). Podophyllotoxin: a novel potential natural anticancer agent. *Avicenna J. Phytomedicine* 7, 285–294. doi: 10.22038/ajp.2024.25338
- Cantrell, C. L., Zheljazkov, V. D., Carvalho, C. R., Astatkie, T., Jeliakova, E. A., and Rosa, L. H. (2014). Dual extraction of essential oil and podophyllotoxin from *Creeping Juniper* (*Juniperus horizontalis*). *PLoS One* 9, e106057. doi: 10.1371/journal.pone.0106057
- Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569. doi: 10.1038/nmeth.2474
- Donath, A., Jühling, F., Al-Arab, M., Bernhart, S. H., Reinhardt, F., Stadler, P. F., et al. (2019). Improved annotation of protein-coding genes boundaries in metazoan mitochondrial genomes. *Nucleic. Acids Res.* 47, 10543–10552. doi: 10.1093/nar/gkz833
- Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic. Acids Res.* 47, W59–W64. doi: 10.1093/nar/gkz238
- Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi: 10.1093/bioinformatics/btt086
- Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 9, R7. doi: 10.1186/gb-2008-9-1-r7
- Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinf.* 5, 59. doi: 10.1186/1471-2105-5-59
- Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20, 2878–2879. doi: 10.1093/bioinformatics/bth315
- Nguyen, D. H., Tran, Q. H., Le, L. T., Nguyen, H. H. T., Tran, H. T., Do, T. P., et al. (2024). Genomic characterization and identification of candidate genes for putative podophyllotoxin biosynthesis pathway in *Penicillium herquei* HGN12.1C. *Microb. Biotechnol.* 17, e70007. doi: 10.1111/1751-7915.70007
- Nguyen, G. T., Nguyen, H. T. H., Tran, H. T., Tran, H. T., Ho, A. N., Tran, Q. H., et al. (2023). Enhanced podophyllotoxin production of endophyte *Fusarium proliferatum* TQN5T by host extract and phenylalanine. *Appl. Microbiol. Biotechnol.* 107, 5367–5378. doi: 10.1007/s00253-023-12659-1
- Palmer, J. M., and Stajich, J. (2020). doi: 10.5281/zenodo.4054262
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351
- Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic. Acids Res.* 33, W465–W467. doi: 10.1093/nar/gki458
- Tarailo-Graovac, M., and Chen, N. (2009). Using repeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma* 25, 4.10.1–4.10.14. doi: 10.1002/0471250953.bi0410s25
- Tran, T. H., Nguyen, G. T., Nguyen, H. T. H., Tran, T. H., Tran, H. Q., Tran, H. Q., et al. (2022). Isolation and cytotoxic potency of endophytic fungi associated with *Dyosma difformis*, a study for the novel resources of podophyllotoxin. *Mycobiology* 50, 389–398. doi: 10.1080/12298093.2022.2126166
- Wanner, J., Jirovetz, L., and Schmidt, E. (2015). Callitris intratropica R.T. Baker & H.G. Smith as a novel rich source of deoxypodophyllotoxin. *Curr. Bioact. Compd.* 11, 73–77. doi: 10.2174/157340721102150820124412
- Yu, X., Che, Z., and Xu, H. (2017). Recent advances in the chemistry and biology of podophyllotoxins. *Chem. – Eur. J.* 23, 4467–4526. doi: 10.1002/chem.201602472

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/ffunb.2026.1695025/full#supplementary-material>

SUPPLEMENTARY TABLE 1

Gene content and organization of the *Penicillium herquei* HGN12.1C mitochondrial genome

SUPPLEMENTARY FIGURE 1

Circular physical map of the mitochondrial genome of *Penicillium herquei* strain HGN12.1C. The map shows the distribution of 44 genes, including protein-coding genes (PCGs), ribosomal RNAs (rRNAs), and transfer RNAs (tRNAs). The inner ring represents the GC content graph.