BMC Genomic Data

DATA NOTE Open Access

Whole-genome sequencing of global forest pathogen *Diplodia sapinea* causing pine shoot blight

QuanChao Wang[1], FeiFei Liu[2], HuaChao Xu[1*] and XuDong Zhou[1*]

Abstract

Objective The pathogenic fungus *Diplodia sapinea* is of significant importance due to its primary role inducing tip dieback on various *Pinus* species which are widely distributed throughout the world. The objective of this study is to further provide comprehensive and specific resources for genome assembly and sequence annotation of this important forest pathogen from China, thereby establishing a robust foundation for future studies on its systematics, population genetics, genomics and global movement.

Data description A high-quality genome of *D. sapinea* strain ZXD319 was sequenced utilizing the Nanopore PromethION and BGI DNBSEQ-T7 platforms. The assembled genome spans a total length of 36.81 Mb, comprising 14 contigs, with a GC content of 56.80% and an N50 value of 2,972,533 bp. It encompasses 11,200 protein-coding genes and 252 noncoding RNAs. The predicted genes were annotated against multiple public databases, and 1,611 potential virulence genes were identified through the Pathogen Host Interactions (PHI) database. Furthermore, the genome comparative analysis of *D. sapinea* and related species revealed 11,568 gene clusters and 3,436 single-copy clusters. Phylogenetic analysis indicated a close evolutionary relationship between *D. sapinea* with *D. corticola* and *D. seriata*. The genomic data presented herein serve as a valuable resource for future studies on this globally important pathogen.

Keywords Genome, Gene annotation, Comparative genome, *Diplodia* shoot blight, Fungi

Objective

Diplodia sapinea referred as *Diplodia pinea* or *Sphaeropsis sapinea*, represents a widespread pathogen on pines, posing a substantial threat to coniferous species [1]. As an opportunistic and necrotrophic pathogen, *D. sapinea* typically resides as a benign endophytic fungus within healthy pine trees, exhibiting no overt signs of disease [2, 3]. However, under conditions of environmental stress such as extreme climatic events like drought and elevated temperatures, or physical injuries, for instance, hail damage, the fungus transfers from its latent endophytic phase to an active pathogenic state [4–6]. This transition is frequently associated with a range of severe disease

*Correspondence:
HuaChao Xu
xhcinsect@zafc.edu.cn
XuDong Zhou
xudong.zhou@zafu.edu.cn
[1]National Key Laboratory for Development and Utilization of Forest Food Resources, Zhejiang A & F University, Hangzhou, Zhejiang 311300, China
[2]Research Institute of Fast-growing Trees (RIFT), Chinese Academy of Forestry (CAF), Zhanjiang, Guangdong 524000, China

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	Table S1, Genome assembly and annotation features of *Diploida sapinea* isolates	Document file (.docx)	Figshare, https://doi.org/10.6084/m9.figshare.29094050.v1 [14]
Data file 2	Figure S1, Venn diagram	Document file (.pdf)	Figshare, https://doi.org/10.6084/m9.figshare.28169636.v1 [16]
Data file 3	Figure S2, Phylogenetics analyses	Document file (.pdf)	Figshare, https://doi.org/10.6084/m9.figshare.28169651.v1 [17]
Data set 1	Genome assembly of *Diplodia sapinea* strain ZXD319	Fasta file (.fasta)	NCBI Nucleotide, https://identifiers.org/nucleotide:JBLWNE010000000 [18]

manifestations, including branch wilt, a marked reduction in growth rates, bark ulceration, anomalous color alterations, root histopathological changes, and potentially the demise of the entire tree [7, 8]. It is noteworthy that *D. sapinea* exhibits a broad host spectrum, encompassing species in the genera of *Abies*, *Larix*, *Picea*, *Thuja*, *Pseudotsuga*, and *Pinus*, thereby underscoring the considerable peril it represents to forest ecosystems and the timber sector [9–11].

For *D. sapinea*, genomic data of five isolates (CBS 117911, CBS 138184, KE8364, KE8391, KE8634) isolated from the pines are accessible in the NCBI database [12, 13]. However, there exist limitations to use them as reference genomes asthey were assembled from fragmented genome assemblies. In this study, we employed the long-read sequencing approach by Oxford Nanopore Technologies (ONT) in combination with a DNA fragment library to achieve a high-quality genome assembly. This innovative methodology has yielded a novel, high-quality genomic resource for *D. sapinea*. This substantially enhances our understanding of the pathogen's biological characteristics and provides valuable insights into elucidating its pathogenic mechanisms.

Data description

The *D. sapinea* strain ZXD319 was isolated from infected dead branches of *P. thunbergii* in Shandong, China, and selected for genome sequencing (DataFile 1; Table 1) [14]. The fungus was purified using the single hypha purification, and confirmed as *D. sapinea* by sequencing the internal transcribed spacer (ITS), translation elongation factor 1-alpha (*tef1-α*) and beta-tubulin (*tub2*) gene regions. A substantial amount of fresh mycelium was obtained from 2% malt extract agar (MEA) (20 g malt extract and 20 g agar per liter of water). The mycelium was immediately frozen in liquid nitrogen and stored at −80 °C in the laboratory. High-quality genomic DNA was extracted using an improved CTAB (cetyltrimethylammonium bromide) method, and DNA integrity and purity were evaluated by 0.8% agarose gel electrophoresis [15].

Whole-genome sequencing was performed at Shanghai Ouyi Biomedical Technology Co., Ltd., utilizing both short-read platform andlong-read Oxford Nanopore Technologies (ONT). The BGI DNBSEQ-T7 platform was employed for paired-end 150 bp (PE150) sequencing, and the fastp v. 0.23.2 software was utilized for quality control and deduplication of the short-read sequencing data, effectively removing low-quality and spliced sequences [19]. For the long-read sequencing, the PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK) was used for full-length sequencing. An initial genome size estimation was performed using the GCE v. 1.0.0 for K-mer analysis [20]. Following this, the long-read ONT data were assembled using the NextDenovo [21], and error correction was achieved by aligning the short-read sequencing data with the preliminary assembly using NextPolish [22]. Ultimately, the purge_haplotigs v. 1.0.4 tool was applied to eliminate redundancy from the genome post-initial assembly and correction, identifying and removing redundant hybrid contigs based on sequence similarity and the ratio of redundant segments to the total contig length [23]. The NCBI NT library (https://ftp.ncbi.nlm.nih.gov/blast/db/) was queried using Blastn v. 2.11.0 + for comparative analysis [24]. Short-read and long-read datasets were aligned to the assembled genome using bwa v. 0.7.12-r1039 [25] and minimap2 v. 2.24-r1122 [26], respectively, to calculate alignment rates and sequencing depth. Genome assembly quality was assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO) [27]. Concurrently, the Smudgeplot v. 0.2.3 dev was utilized to assess the species' ploidy and heterozygosity structure [28].

The genome size of the *D. sapinea* strain ZXD319 was estimated to be 38.29 Mb using K-mer (K = 17) analysis, with a heterozygosity rate of 0.21. A total of 11,242,863,900 bp of DNBseq short-reads are generated, achieving a sequencing depth of 294×(DateFile 1; Table 1) [14]. Additionally, 12,389,890,601 bp of ONT long-reads are obtained, resulting in a sequencing depth of 336×. The high-quality assembled genome is 36.81 Mb, with an N50 of 2,972,533 bp and 14 contigs (Data File 1; Table 1) [14]. Smudgeplot analysis indicated that ZXD319 had a haplotype proportion of 96%, suggesting it as the most likely haplotype. The completeness of the ZXD319 genome assembly was assessed using BUSCO based on the dothideomycetes_odb10 database (https://busco-data.ezlab.org/v5/data/lineages/), yielding a high completeness score of 99.5% (DateFile 1; Table 1) [14].

Maker2 v. 2.31.10 was used for de novo gene prediction, identifying a total of 11,200 protein-coding genes with an average length of 1,953 bp (DataFile 1; Table 1) [14, 29]. Additionally, 135 transfer RNAs were predicted by tRNAscan-SE v. 1.3.1, 56 ribosomal RNAs by Blastn v. 2.11.0+ [24], and 61 small nuclear RNAs by Rfam v. 14.8 [30], totaling 252 non-coding RNAs. Furthermore, tandem repeats were analyzed by the TRF v.4.09 [31], and interspersed repeats were predicted using RepeatMasker (http://www.repeatmasker.org/). And the assembled genome contains 289,879 base pairs (0.77%) of tandem repeats and 1,094,124 base pairs (2.87%) of interspersed repeats.

Annotation

Functional gene annotation was performed using Non-Redundant Protein sequence database (NR; 10,969 genes, 97.94%), Swiss-Prot database (5,672 genes, 50.64%), TrEMBL database (10,953 genes, 97.79%), Clusters of orthologous groups for eukaryotic complete genomes (KOG: 5,057 genes, 45.15%), InterProScan database (8,580 genes, 76.61%), Gene Ontology (GO: 6,103 genes, 54.49%), Kyoto Encyclopedia of Genes and Genomes (KEGG; 10,063 genes, 89.85%) and Pfam database (7,841 genes, 70.01%) (DateFile 1; Table 1) [14]. Additional annotations were carried out using the Pathogen Host Interactions (PHI) database [32] and the Carbohydrate-Active Enzymes (CAZy) database [33]. A total of 1,611 PHI-related genes are identified in the genomes, and 2,335 genes per genome are annotated from the CAZy database, including 1,189 encoding glycoside hydrolases (GHs), 572 encoding glycosyltransferases (GTs), 309 auxiliary activities (AAs), and 207 carbohydrate-binding modules (CBMs), 106 encoding carbohydrate esterases (CEs), 61 encoding polysaccharide lyases (PLs). Meanwhile, an average of 304 putative secretory proteins are identified in the genome using SignalP v. 4.0 [34] and TMHMM v. 2.0 [35].

The comparative genomics of the orthologous gene clusters were identified between *D. sapinea* ZXD319 and seven related species (*Botryosphaeria dothide*, *D. corticola*, *D. seriata*, *Lasiodiplodia theobromae*, *Macrophomina phaseolina*, *Neofusicoccum parvum*, *Aplosporella prunicola* as outgroup). To ensure protein quality, proteins shorter than 30 amino acids were excluded from the analysis. The Diamond comparison software was configured with an E-value threshold of ≤ 1e-3, while the MCL clustering inflation parameter was set to 1.5, and the OrthoFinder algorithm was employed [36]. The analysis reveals that these eight fungi collectively possessed a total of 93,076 genes, which form 11,568 gene clusters. Of these, 3,436 genes are identified as single-copy clusters, while 4,587 genes are present in all eight fungal species (DataFile 2; Table 1) [16].

Based on the commonly shared single-copy orthologous genes within gene family clustering, phylogenetic analysis was conducted using the RAxML v. 8 and the maximum likelihood method to construct evolutionary trees for these eight species, with *Aplosporella prunicola* as the outgroup [37]. The results indicate a close relationship between *D. sapinea* with *D. corticola*, and *D. seriata*, followed by *L. theobromae* (DataFile 3; Table 1) [17].

In this study, we successfully generated a high-quality draft genome of the pine pathogen *D. sapinea*, which is of great significance to promote our understanding of its biological characteristics and pathogenic mechanisms, as well as to develop effective control strategies.

Limitation

The whole genome was sequenced utilizing the Nanopore PromethION for long-read sequencing, complemented by PE150 paired-end sequencing on the BGI DNBSEQ-T7 platform to obtain high-quality genomic data. However, these datasets have not yet reached chromosomal-level assembly. To achieve a high-quality, chromosome-scale assembly suitable for comparative genomics, additional assembly and sequencing technologies, such as Hi-C, are required.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-025-01328-z.

Supplementary Material 1: Fig S1. Orthologous cluster identification among *Diplodia sapinea* (ZXD319; sequenced in this study) and other seven closely related species (*Botryosphaeria dothide, D. corticola, D. seriata, Lasiodiplodia theobromae, Macrophomina phaseolina, Neofusicoccum parvum*, and *Aplosporella prunicola*).

Supplementary Material 2: Fig. S2. Maximum likelihood phylogenetic tree encoded by eight Botryosphaeriaceae fungal genomes. *Aplosporella prunicola* (GCF_010093885.1) represents the outgroup.

Supplementary Material 3.

Acknowledgements

Not applicable.

Authors' contributions

XuDong Zhou and HuaChao Xu conceived the experiments; QuanChao Wang and FeiFei Liu completed experiments and wrote the manuscript. All authors edited and approved the final manuscript.

Funding

This study was supported by The National Key R&D Program of China (2022YFD1401000) and the Launching Funds for Talents of Zhejiang A&F University (2020FR036).

Data availability

The data described in this data note were deposited under the National Center for Biotechnology Information (NCBI) underBioProject JBLWNE010000000. Associated Datafiles are available on Figshare: Table S1, Genome assembly and annotation features of Diplodia sapinea (https://doi.org/10.6084/m9.figshare.29094050.v1); Figure S1, Venndiagram (https://doi.org/10.6084/m9.figshare.28169636.v1); Figure S2, Phylogenetics analyses (https://doi.org/10.6084/m9.figshare.28169651.v1).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 9 January 2025 / Accepted: 15 May 2025
Published online: 27 May 2025

References

1. Phillips AJL, Alves A, Abdollahzadeh J, Slippers B, Wingfield MJ, Groenewald JZ, Crous PW. The botryosphaeriaceae: genera and species known from culture. Stud Mycol. 2013;76(1):51–167. https://doi.org/10.3114/sim0021.
2. Blumenstein K, Bußkamp J, Langer GJ, Langer EJ, Terhonen E. The Diplodia tip blight pathogen *Sphaeropsis sapinea* is the most common fungus in Scots Pines' mycobiome, irrespective of health status—a case study from Germany. J Fungi. 2021;7(8):607. https://doi.org/10.3390/jof7080607.
3. Terhonen EL, Babalola J, Kasanen R, Jalkanen R, Blumenstein K. *Sphaeropsis sapinea* found as symptomless endophyte in Finland. Silva Fenn. 2021;55(1). https://doi.org/10.14214/sf.10420.
4. Stanosz GR, Blodgett JT, Smith DR, Kruger EL. Water stress and *Sphaeropsis sapinea* as a latent pathogen of red pine seedlings. New Phytol. 2001;149(3):531–8. https://doi.org/10.1046/j.1469-8137.2001.00052.x.
5. Smith H, Wingfied MJ, Coutinho TA. The role of latent *Sphaeropsis sapinea* infections in post-hail associated die-back of Pinus patula. For Ecol Manag. 2002;164(1–3):177–84. https://doi.org/10.1016/S0378-1127(01)00610-7.
6. Brodde L, Adamson K, Julio Camarero J, Castaño C, Drenkhan R, Lehtijärvi A, Luchi N, Migliorini D, Sánchez-Miranda Á, Stenlid J. Diplodia tip blight on its way to the North: drivers of disease emergence in Northern Europe. Front Plant Sci. 2019;9:1818. https://doi.org/10.3389/fpls.2018.01818.
7. Blodgett JT, Bonello P. The aggressiveness of *Sphaeropsis sapinea* on Austrian pine varies with isolate group and site of infection. Pathol. 2003;33(1):15–9. https://doi.org/10.1046/j.1439-0329.2003.00303.x.
8. Caballol M, Ridley M, Colangelo M, Valeriano C, Camarero JJ, Oliva J. Tree mortality caused by Diplodia shoot blight on *Pinus sylvestris* and other mediterranean Pines. Ecol Manage. 2022;505:119935. https://doi.org/10.1016/j.foreco.2021.119935.
9. Palmer MA, Stewart EL, Wingfield MJ. Variation among isolates of *Sphaeropsis sapinea* in the North central united States. Phytopathology. 1987;77(6):944–8. https://doi.org/10.1038/nmeth.1701.
10. Whitehill JGA, Lehman JS, Bonello P. *Ips pini* (Curculionidae: Scolytinae) is a vector of the fungal pathogen, *Sphaeropsis sapinea* (Coelomycetes), to Austrian Pines, *Pinus nigra*. (Pinaceae) Environ Entomol. 2014;36:114–20. https://doi.org/10.1603/0046-225X(2007)36[114:IPCSIA]2.0.CO;2.
11. Xie JY, Li H, Wan Y, Li DW, Zhu LH. Shoot blight of *Pinus bungeana* caused by *Diplodia sapinea* newly reported in China. Crop Prot. 2024;181:106711. https://doi.org/10.1016/j.cropro.2024.106711.
12. Yu CM, Diao YF, Lu Q, Zhao JP, Cui SN, Xiong X, Lu AN, Zhang XY, Liu HX. Comparative genomics reveals evolutionary traits, mating strategies, and pathogenicity-related genes variation of Botryosphaeriaceae. Front Microbiol. 2022;13:800981. https://doi.org/10.3389/fmicb.2022.800981.
13. van der Nest MA, Bihon W, De Vos L, Naidoo K, Roodt D, Rubagotti E, Slippers B, Steenkamp ET, Wilken PM, Wilson A. Draft genome sequences of *Diplodia sapinea*, *Ceratocystis manginecans*, and *Ceratocystis moniliformis*. IMA Fungus. 2014;5:135–140. https://doi.org/10.5598/imafungus.2014.05.01.13.
14. Wang QC, Liu FF, Xu HC, Zhou XD. Data file 1-Table S1, genome assembly and annotation features of Diplodia sapinea isolate. Figshare. 2025. https://doi.org/10.6084/m9.figshare.29094050.v1
15. Möller E, Bahnweg G, Sandermann H, Geiger H. A simple and efficient protocol for isolation of high molecular weight DNA from filamentous fungi, fruit bodies, and infected plant tissues. Nucleic Acids Res. 1992;20(22):6115. https://doi.org/10.1093/nar/20.22.6115.
16. Wang QC, Liu FF, Xu HC, Zhou XD. Data file 2-Figure S1, Venn diagram. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28169636.v1.
17. Wang QC, Liu FF, Xu HC, Zhou XD. Data file 3- Figure S2, phylogenetics analyses. Figshare. 2025. https://doi.org/10.6084/m9.figshare.28169651.v1.
18. Wang QC, Liu FF, Xu HC, Zhou XD. Dataset 1- Genome assembly of *Diplodia sapinea* strain ZXD319. NCBI. 2025. https://identifiers.org/nucleotide:JBLWNE010000000.
19. Chen SF, Zhou YQ, Chen YR, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90. https://doi.org/10.1093/bioinformatics/bty560.
20. Liu BH, Shi YJ, Yuan JY, Hu XS, Zhang H, Li N, Li ZY, Chen YX, Mu DS, Fan W. Estimation of genomic characteristics by analyzingk-mer frequency in de novo genome project. arXiv 2013;1308.2012. https://arxiv.org/abs/1308.2012.
21. Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, Li J, Sandoval JR, Cooper DN, Ye K. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. Genome Biol. 2024;25(1):107. https://doi.org/10.1186/s13059-024-03252-4.
22. Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics. 2020;36(7):2253–5. https://doi.org/10.1093/bioinformatics/btz891.
23. Roach MJ, Schmidt SA, Borneman AR. Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19:1–10. https://doi.org/10.1186/s12859-018-2485-7.
24. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
25. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324.
26. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100. https://doi.org/10.1093/bioinformatics/bty191.
27. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2. https://doi.org/10.1093/bioinformatics/btv351.
28. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020;11(1):1432. https://doi.org/10.1038/s41467-020-14998-3.
29. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:1–14. http://www.biomedcentral.com/1471-2105/12/491.
30. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. 2005;33(suppl1):D121–4. https://doi.org/10.1093/nar/gki081.
31. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80. https://doi.org/10.1093/nar/27.2.573.
32. Urban M, Pant R, Raghunath A, Irvine AG, Pedro H, Hammond-Kosack KE. The Pathogen-Host interactions database (PHI-base): additions and future developments. Nucleic Acids Res. 2015;43(D1):D645–55. https://doi.org/10.1093/nar/gku1165.
33. Jia FF, Zhang LJ, Pang XH, Gu XX, Abdelazez A, Liang Y, Sun SR, Meng XC. Complete genome sequence of bacteriocin-producing Lactobacillus plantarum KLDS1. 0391, a probiotic strain with Gastrointestinal tract resistance and adhesion to the intestinal epithelial cells. Genomics. 2017;109(5–6):432–7. https://doi.org/10.1016/j.ygeno.2017.06.008.
34. Petersen TN, Brunak S, Von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8(10):785–6. https://doi.org/10.1038/nmeth.1701.
35. Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305(3):567–80. https://doi.org/10.1006/jmbi.2000.4315.
36. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:1–14. https://doi.org/10.1186/s13059-015-0721-2.
37. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:1–14. https://doi.org/10.1186/s13059-019-1832-y.

Publisher's Note