Data in Brief 66 (2026) 112613

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Draft genome sequence data of *Colletotrichum camelliae* isolated from *Camellia japonica* in the United States

Kenneth R. Leep[a], Renee S. Arias[b], Warren E. Copes[a], John Dobbs[c], Siva P. Kumpatla[a,*]

[a] United States Department of Agriculture (USDA), Agricultural Research Service (ARS), Thad Cochran Southern Horticultural Laboratory, 810 Hwy 26 W, Poplarville, MS, USA
[b] United States Department of Agriculture (USDA), Agricultural Research Service (ARS), National Peanut Research Laboratory (NPRL), 1011 Forrester dr. S.E., Dawson, GA, USA
[c] United States Department of Agriculture (USDA), Agricultural Research Service (ARS), Foreign Disease/Weed Science Research Unit, 1301 Ditto Ave, Fort Detrick, MD, USA

ARTICLE INFO

Article history:
Received 16 December 2025
Accepted 16 February 2026
Available online 19 February 2026

Dataset link: BioProject: PRJNA1273064 (Original data)

Keywords:
Anthracnose
Colletotrichum camelliae
Ornamental
Camellia japonica
Whole genome sequence

ABSTRACT

Colletotrichum camelliae Massee, a member of *Colletotrichum gloeosporioides* species complex (CGSC), is an economically important fungal pathogen that causes significant damage to *Camellia sinensis* (tea) and other *Camellia* species that include ornamental crops. Although the occurrence of *C. camelliae* in the United States has been documented, to date most of the reports and all available genome sequences of this pathogen are from strains isolated outside the United States. Here, we report the isolation of four *C. camelliae* isolates (GCC14, PS22-1, PS23, and PS24) from symptomatic *Camellia japonica* plants in George County, Mississippi. Whole genome sequencing of all four isolates was performed as paired-end reads of 150 bp using NovaSeq X Plus. Clean reads were *de novo* assembled, mapped to *C. camelliae* reference genome sequence, LS-19, and genome-wide variants were analyzed. Sequencing data were submitted to NCBI GenBank BioProject PRJNA1273064, and the following accession numbers were assigned: SRR34108282 (GCC14), SRR33953793 (PS22-1), SRR33955217 (PS23) and SRR33957315 & SRR33956925 (PS24). The genome sequence

* Corresponding author.
E-mail address: siva.kumpatla@usda.gov (S.P. Kumpatla).

https://doi.org/10.1016/j.dib.2026.112613

data reported here for *C. camelliae* are valuable for developing diagnostic assays for the identification, tracking, and management of this pathogen on *Camellia* species that include tea and ornamental crops. In addition, comparative genomics, phylogenetics, and evolutionary studies using these and other available genomes can aid in deriving insights on the diversity and pathogenicity of this pathogen, CGSC, and the *Colletotrichum* genus.

Specifications Table

Subject	Genomics
Specific subject area	Fungal plant pathogen genomics
Data format	Genome sequences, raw data, *de novo* assembly, filtered and trimmed reads, reference mapping data
Type of data	Tables, Figures
Data collection	*Colletotrichum camelliae* was isolated from symptomatic *Camellia japonica* plants in George County, Mississippi, USA, in 2023. The whole genome was sequenced on Illumina NovaSeq X Plus platform at the UC Davis Genome Center, California. Data were processed using CLC Genomics Workbench v25.0.1. Clean reads >120 nucleotides were *de novo* assembled, and contigs were used to generate alignments for phylogenetic identification. Clean reads were also mapped to the reference genome LS-19 and variants (SNPs, MNPs, and Indels) were identified.
Data source location	*C. camelliae* was collected from George County, Mississippi, USA, 30° 51' 54.5436" N, 88° 32' 43.2054" W. Sequencing data were placed in public repository: NCBI.
Data accessibility	BioProject PRJNA1273064, SRA numbers SRR34108282 (GCC14), SRR33953793 (PS22-1), SRR33955217 (PS23) and SRR33957315 & SRR33956925 (PS24). Repository name: National Center for Biotechnology Information (NCBI). Direct URL to data: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1273064/

1. Value of the Data

- This is the first report of whole genome sequencing of *Colletotrichum camelliae* isolated from *Camellia japonica* ornamental plants in the United States, and the first-time a *C. camelliae* genome sequence is reported from the United States.
- The ornamental horticultural industry in the United States accounts for one-third of the total value of all specialty crops. Timely identification of existing and emerging pathogens and management of diseases caused by them are critical for sustaining the productivity and economic impact of this important category of plants.
- The four *C. camelliae* genomes reported here show many genetic variants when compared to the reference genome sequence of LS-19 isolated from *Camellia sinensis* (Tea) in China. Many single- and multi-nucleotide polymorphisms (SNPs, MNPs) and insertions/deletions (Indels) were observed throughout the genomes of all four isolates; most variants were SNPs.
- Since the sequences reported here are the first *C. camelliae* genomes from the United States, they are a valuable resource to understand the unique and differentiating features of these isolates compared to *C. camelliae* strains reported from other locations, predominantly from tropical and subtropical countries around the world.
- Considering the phenotypic plasticity of *Colletotrichum* [1], the availability of the genomes reported here will enable unambiguous identification and study of the potential spread of this pathogen in the United States.

- Comparative analysis of the *C. camelliae* genomes reported here enables the development of molecular markers and diagnostic tools for accurate identification, tracking, and management of this pathogen. It will also facilitate population structure analysis, refinement of phylogenetic trees, and identification and differentiation of potential new strains of this pathogen.
- The *C. camelliae* data reported here can be used in comparative genomics to identify genes that are targets for fungicides and to identify potential mutations associated with fungicide resistance.
- The genome data provided here are of substantial importance for studies aimed at understanding genome complexity, architecture, and composition of pathogenicity-related genes while providing insights about the genetic basis of polyphagous (generalists) and narrow host range (specialists) pathogens. *C. camelliae* is a specialist and the four isolates reported here (GCC14, PS22-1, PS23, and PS24) have genome sizes within the range (57.84 – 67.7 Mb) reported from outside the United States; when comparatively analyzed with the genomes of other *Colletotrichum* species, these genomes could be very valuable for investigating the compositional differences between generalists and specialists.

2. Background

The genus *Colletotrichum* (Family Glomerellaceae, Class Sordariomycetes, Division Ascomycota), voted as one of the top 10 most important genera of phytopathogenic fungi, contains economically important pathogens that cause anthracnose in a wide range of host plants worldwide [2]. *Colletotrichum gloeosporioides* species complex (CGSC), the most species-rich complex among the 16 species complexes of *Colletotrichum,* has a broad host range with its members causing damage to diverse agronomic and horticultural crops [2–4].

Colletotrichum camelliae Massee, a member of CGSC, is one of the most important fungal pathogens infecting *Camellia sinensis* (Tea) in almost all tea growing regions in the world. Compared to polyphagous *C. gloeosporioides, C. fructicola,* and other species of CGSC, *C. camelliae* has a narrow host range, primarily infecting *C. sinensis* and other *Camellia* species that include ornamentals such as *Camellia sasanqua, Camellia japonica*, and *Camellia oleifera* [5–7].

The ornamental horticultural industry in the United States accounts for one-third of the total value of all specialty crops [8]. Therefore, to maintain consistent productivity, it is important to monitor and manage existing diseases as well as to identify, track, and manage any new and emerging pathogens. To date, *C. camelliae* has been reported in the United States on *C. sinensis* and an ornamental *C. sasanqua* [6,9]. A recent study also reported a single isolate of *C. camelliae* on *C. japonica*, an important ornamental within *Camellia* genus [10].

Timely and accurate identification of symptomatic plants and associated fungal species and development of molecular diagnostics is crucial for understanding pathogenicity, investigating host range, and implementing effective management strategies [11]. Although approaches based on single-gene, multilocus, or internal transcribed spacer (ITS) have been widely used for the genetic identification and differentiation of *Colletotrichum* species, they were found to be insufficient in several cases [1]. Leveraging whole genome sequences and performing comparative analysis using other genomes enable efficient species differentiation, phylogenetic analysis, and evolutionary studies.

Only three *C. camelliae* genome sequences, all derived from isolates obtained in China, are currently available in the NCBI database. LS-19 (reference genome) and LT-3-1 were isolated from *C. sinensis* and CcLH18 was isolated from *C. oleifera*. Genome sequences for isolates from the United States are currently not available. Here, we report the isolation of *C. camelliae* from symptomatic *C. japonica* plants in Mississippi and provide the whole genome sequence of these isolates.

The purpose of this research was to obtain the draft genome sequences of four *C. camelliae* isolates, GCC14, PS22-1, PS23, and PS24, from the United States, perform comparative analysis, and derive information on their relatedness with *C. camelliae* genomes from other locations. The availability of these genome sequences not only enables development of diagnostics for ac-

Table 1
Read quality information for novel *C. camelliae* isolates, raw and post-trim.

Isolate	Number of reads (Raw)	Number of reads (Post-trim)	Amount of nucleotides (Raw)	Amount of nucleotides (Post-trim)
GCC14	33,540,392	33,449,960	5,031,058,800	5,008,387,405
PS22-1	26,459,460	26,401,141	3,968,919,000	3,947,684,168
PS23	34,223,466	34,175,147	5,133,519,900	5,115,698,575
PS24	35,855,472	35,706,859	5,378,320,800	5,342,198,071

Table 2
Adapters used by Novogene sequencing.

Adapter	Sequence
P5	AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT
P7	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG

Table 3
Genomic features and *de novo* assembly statistics (Quast report) for the four isolates (GCC14, PS22-1, PS23, and PS24) of *C. camelliae*.

Statistic	GCC14	PS22-1	PS23	PS24
Sequencing Coverage	70.23 ×	63.76 ×	81.20 ×	86.66 ×
Scaffolds, count	563	493	442	615
Contigs, count	629	607	501	710
Largest contig (bp)	1,485,726	1,234,331	1,253,778	1,317,036
Total genomic length (bp)	63,197,770	61,466,282	62,553,133	61,295,723
N50 (bp)	281,246	362,368	338,263	313,323
N90 (bp)	88,206	86,543	99,102	63,284
L50 (bp)	65	48	60	60
L90 (bp)	215	182	185	228
GC%	50.16	50.34	50.15	50.38

Table 4
Busco Statistics (using Glomerellales database).

Isolate	Complete BUSCOs / %	Single BUSCOs / %	Duplicated BUSCOs / %	Fragmented BUSCOs / %	Missing BUSCOs / %	n
GCC14	6058 / 99.3 %	6044 / 99.0 %	14 / 0.2 %	5 / 0.1 %	40 / 0.7 %	6103
PS22	6061 / 99.3 %	6049 / 99.1 %	12 / 0.2 %	5 / 0.1 %	37 / 0.6 %	6103
PS23	6063 / 99.3 %	6051 / 99.1 %	12 / 0.2 %	5 / 0.1 %	35 / 0.6 %	6103
PS24	6061 / 99.3 %	6049 / 99.1 %	12 / 0.2 %	5 / 0.1 %	37 / 0.6 %	6103

curate identification, differentiation, and tracking of *C. camelliae* in the United States, but also contributes to global research efforts in phylogenetic and evolutionary studies of CGSC and *Colletotrichum* genus.

3. Data Description

Draft genome sequences for isolates GCC14, PS22-1, PS23 and PS24 are available at the National Center for Biotechnology Information (NCBI) under the BioProject Number PRJNA1273064 with the accession numbers SRR34108282 (GCC14), SRR33953793 (PS22-1), SRR33955217 (PS23), and SRR33957315 and SRR33956925 (PS24). Basic read quality data is presented in Table 1 and sequences of adapters used by Novogene are provided in Table 2. Data evaluation performed by Quast and BUSCO analysis is presented in Tables 3 and 4, respectively.

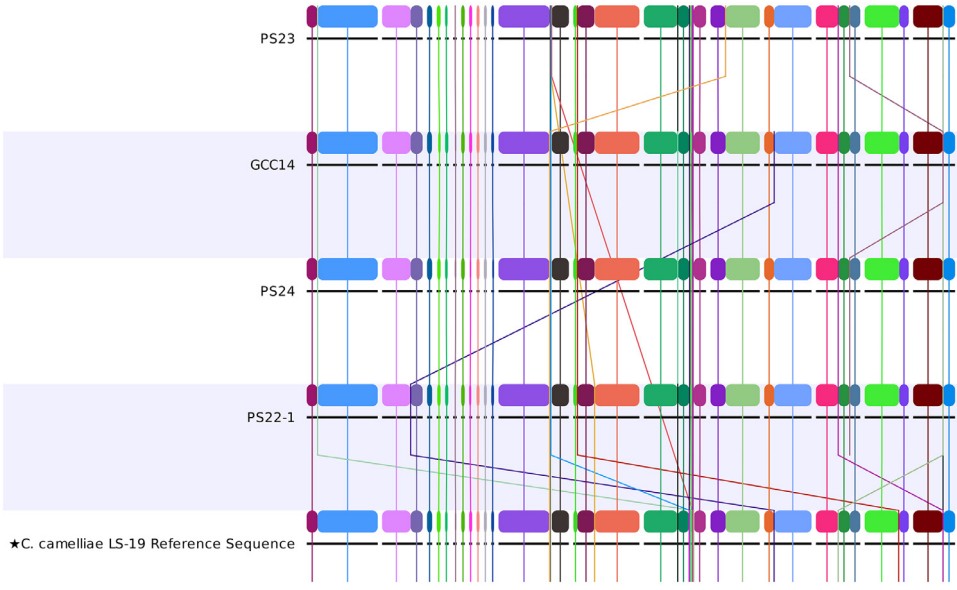

Fig. 1. Whole genome alignment of *Colletotrichum camelliae* isolates PS23, GCC14, PS24, and PS22-1 to reference genome LS-19.

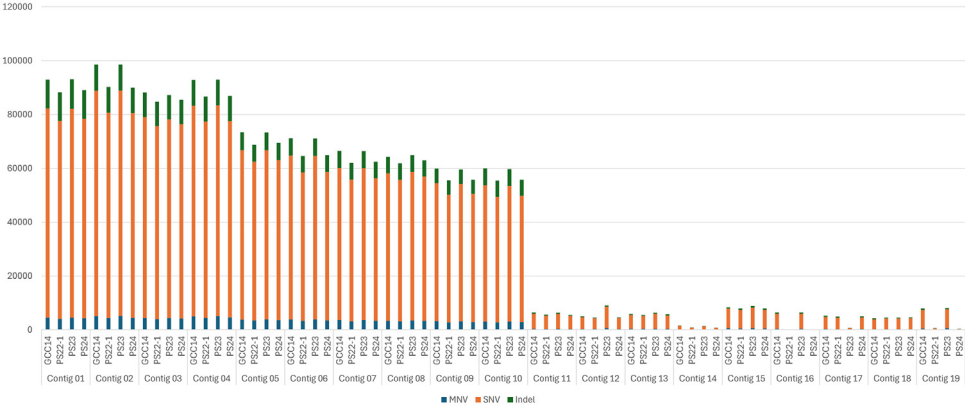

Fig. 2. Contig-wise distribution of variants. SNPs: single nucleotide polymorphisms; MNPs: multiple nucleotide polymorphisms; and Indels: total insertions and deletions.

The whole genome sequences of these isolates were aligned to the structurally annotated reference genome sequence of isolate LS-19 (Biosample ID: SAMN17126018). The alignment results are presented in Fig. 1. Genome-wide variant analysis was performed based on the comparison of sequencing data of the isolates and the LS-19 reference genome sequence using CLC Genomics Workbench and is presented in Fig. 2. Contigs 1 to 10 shown in Fig. 2 correspond to ten core chromosomes expected in the genome of *C. camelliae*, a member of CGSC [12]. There are 9 smaller contigs (contigs 11-19). Additional work needs to be done to determine whether one or more of these smaller contigs merge with the larger contigs or correspond to mini-chromosomes commonly observed in several species of the CGSC [12].

4. Experimental Design, Materials and Methods

4.1. Fungal isolation

Plant samples from *Camellia japonica* 'Gunsmoke' and 'Pink Serenade' that exhibited foliar anthracnose symptoms and stem dieback were collected in George County, Mississippi, USA. Samples were surface sterilized in 10 % (v/v) sodium hypochlorite solution (1:10 dilution of commercial bleach, approximately 5–9 % NaOCl), rinsed in ultra-pure sterile reverse osmosis water, and the healthy and diseased tissue margin of foliar and stem canker symptoms were placed on potato dextrose agar (PDA). Fungal isolates were recovered and pure isolates obtained by passaging mycelial tips on PDA. Resulting isolates were grown at 24 °C for 3-5 days under constant light and morphologically identified to the genus level (*Colletotrichum*) through microscopic examination of conidia and conidiophores from acervuli. Mycelial tips were extracted and grown in potato dextrose broth on a shaker for an additional 3-5 days at 24 °C in low light conditions. The resulting mycelial globules were removed and vacuum filtered to eliminate the supernatant broth. Globules were split open, and agar blocks extracted using a fire-sterilized scalpel. Globules were placed into sterile 5 ml centrifuge tubes and frozen overnight at -80 °C and subsequently placed in a freeze dryer for approximately 48 h. The lyophilized tissue was placed in a 2 ml bead beater vial approximately ¼ full of 1 mm zirconium silica beads, with 1 ml DEPC-treated, molecular biology grade water added. The vials were placed on a Biospec Products Mini-Beadbeater and beaten at 2600 rpm for 3 min at room temperature (24 ± 3 °C). (BioSpec Products, Inc, Bartlesville, OK, USA).

Isolates were deposited in the Agricultural Research Service Culture Collection at the Northern Regional Research Laboratory (NRRL) located in Peoria, IL, USA, and assigned NRRL numbers 64918 (GCC14), 64914 (PS22-1), 64916 (PS23) and 64993 (PS24).

4.2. Genomic DNA preparation

DNA extraction for all samples was initially performed using the Dynabeads DNA Direct Universal Kit (Fisher Scientific, USA) according to the manufacturer's instructions; isolate PS24 was successfully extracted at high quality using this method, but other isolates failed quality checks. As such, DNA was re-extracted from GCC14, PS22-1 and PS23 using the Wizard Genomic DNA Purification Kit (Promega Corp, USA), again following the manufacturer's instructions. The concentration and purity of DNA was measured using a NanoDrop 1000 Spectrophotometer (Thermo Scientific, USA).

4.3. Genome sequencing and assembly

Microbial whole genome libraries (350 bp insert size) were prepared using an ABclonal Rapid Plus DNA Library Prep kit (ABclonal, Woburn, MA, USA). Sequencing was performed as paired-end reads of 150 base pairs using the Illumina NovaSeq X Plus platform by Novogene, at the UC Davis Genome Center in Davis, CA, USA. To generate clean reads, adapters P5 and P7 (see Table 2 for adapter sequences) were removed from the reads, and low-quality reads, short reads (<120 bp) and homopolymers were trimmed using CLC Genomics Workbench v25.0.3 (Qiagen, Aarhus, Denmark) [13]. The resultant clean reads were assembled in CLC Genomics Workbench and mapped to the reference genome of *C. camelliae* (Strain ID: LS-19; Biosample ID: SAMN17126018).

Genome completeness was assessed using BUSCO v6.0.0 [14] using the Glomerellales database (glomerellales_odb12, rev 7/2025). Table 4 shows BUSCO scores for all categories across all four isolates. Furthermore, genomes were quality checked using Quast v5.3.0 [15] and results of this quality analysis are presented in Table 3.

Table 5
Identification and removal of contaminants using Kraken2.

Isolate	Total Raw Reads	Reads Trimmed	Contaminant Taxids (Count)	Reads Remaining after Trim and Clean
GCC14	33,540,392	136,422	76775 (405), 9606 (9,834), 1747 (1,486), 1282 (200)	33,392,045
PS22	26,459,460	159,643	9606 (597)	26,299,220
PS23	34,223,466	108,737	9606 (1,628), 1747 (242)	34,112,859
PS24	35,855,472	235,562	9606 (798), 1747 (453), 76775 (129)	35,618,530

Genomes were also scanned for contaminants using Kraken2 v2.14. Any reads identified to the species level containing more than 100 reads were extracted from the raw data before *de novo* assembly was performed. Table 5 presents a summary of the contaminants found and removed from the raw read sequences of the isolates.

Limitations

Not applicable.

Ethics Statement

The current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Credit Author Statement

Kenneth R. Leep: Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft, Writing – review & editing; **Renee S. Arias**: Conceptualization, Methodology, Resources, Formal analysis, Writing – review & editing; **Warren E. Copes**: Conceptualization, Funding acquisition, Resources, Writing – review & editing; **John Dobbs**: Formal analysis; **Siva P. Kumpatla**: Conceptualization, Methodology, Funding acquisition, Writing – original draft, Writing – review & editing.

Data Availability

BioProject: PRJNA1273064 (Original data) (NCBI).

Acknowledgments

This work was supported by United States Department of Agriculture, Agricultural Research Service, Research Project (6062-30500-001).

Declaration of Competing Interest

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

References

[1] A. Menicucci, S. Iacono, M. Ramos, C. Fiorenzani, N.A. Peres, L.W. Timmer, A. Prodi, R. Baroncelli, Can whole genome sequencing resolve taxonomic ambiguities in fungi? The case study of *Colletotrichum* associated with ferns, Front. Fungal Biol. 6 (2025) 1540469, doi:10.3389/ffunb.2025.1540469.

[2] R. Dean, J.A.L. Van Kan, Z.A. Pretorius, K.E. Hammond-Kosack, A. Di Pietro, P.D. Spanu, J.J. Rudd, M. Dickman, R. Kahmann, J. Ellis, G.D. Foster, The top 10 fungal pathogens in molecular plant pathology, Mol. Plant Pathol. 13 (2012) 414–430, doi:10.1111/j.1364-3703.2011.00783.x.

[3] F. Liu, Z.Y. Ma, L.W. Hou, Y.Z. Diao, W.P. Wu, U. Damm, S. Song, L. Cai, Updating species diversity of *Colletotrichum*, with a phylogenomic overview, Stud. Mycol. 101 (2022) 1–56, doi:10.3114/sim.2022.101.01.

[4] Z. Ma, F. Liu, C.K.M. Tsui, L. Cai, Phylogenomics and adaptive evolution of the *Colletotrichum gloeosporioides* species complex, Commun. Biol. 8 (2025) 593, doi:10.1038/s42003-025-08024-9.

[5] B.S. Weir, P.R. Johnson, U. Damm, The *Colletotrichum gloeosporioides* species complex, Stud. Mycol. 73 (2012) 115–180, doi:10.3114/sim0011.

[6] D.F. Farr, A.Y. Rossman, Fungal databases, U.S. National Fungus Collections, ARS, USDA. https://nt.ars-grin.gov/fungaldatabases/, accessed December 6, 2025.

[7] Y. Wang, J.-Y. Chen, X. Xu, U. Cheng, L. Zheng, J. Huang, D.-W. Li, Identification and characterization of *Colletotrichum* species associated with anthracnose disease of *Camellia oleifera* in China, Plant Disease 104 (2020) 474–482, doi:10.1094/PDIS-11-18-1955-RE.

[8] X. Wei, H. Khachatryan, A. Hodges, C. Hall, M. Palma, A. Torres, R. Brumfield, Exploring market choices in the US ornamental horticulture industry, Agribus 39 (2023) 65–109, doi:10.1002/agr.21769.

[9] F. Liu, B.S. Weir, U. Damm, P.W. Crous, Y. Wang, B. Liu, M. Wang, M. Zhang, L. Cai, Unravelling *Colletotrichum* species associated with *Camellia*: employing ApMat and GS loci to resolve species in the *C. gloeosporioides* complex, Persoonia 35 (2015) 63–86, doi:10.3767/003158515X687597.

[10] J.M. Orrock, B. Rathinasabapathi, B.S. Richter, Anthracnose in U.S. Tea: pathogen characterization and susceptibility among six tea accessions, Plant Disease 104 (2020) 1055–1059, doi:10.1094/PDIS-07-19-1518-RE.

[11] C.S. Bhunjun, A.J.L. Phillips, R.S. Jayawardena, I. Promputtha, K.D. Hyde, Importance of molecular data to identify fungal plant pathogens and guidelines for pathogenicity testing based on Koch's postulates, Pathogens 10 (2021) 1096, doi:10.3390/pathogens10091096.

[12] H. Wang, R. Huang, J. Ren, L. Tang, S. Huang, X. Chen, J. Fan, B. Li, Q. Wang, T. Hsiang, H. Liu, Q. Li, The evolution of mini-chromosomes in the fungal genus *Colletotrichum*, mBio 14 (2023), doi:10.1128/mbio.00629-23.

[13] QIAGEN Bioinformatics, 2025. QIAGEN CLC Genomics Workbench, Version 25.0.3, QIAGEN, Aarhus, Denmark digitalinsights.qiagen.com.

[14] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (2015) 3210–3212, doi:10.1093/bioinformatics/btv351.

[15] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics 29 (2013) 1072–1075, doi:10.1093/bioinformatics/btt086.