Genome Sequence Resource of *Colletotrichum horii*, an Important Pathogenic Fungus Threatening Persimmon Production

Hanyue Fan, Yongkuan Li, Sadaruddin Chachar, Yong Yang, and Changfei Guan[†]

State Key Laboratory of Crop Stress Biology for Arid Areas, College of Horticulture, Northwest A&F University, Yangling, Shaanxi 712100, China

Funding

This research was supported by National Key Research and Development Program of China (2019YFD1000600) and Special Project of Science and Technology Innovation Plan of Shaanxi Academy of Forestry (SXLK2020-0212).

Genome Announcement

Colletotrichum horii causes anthracnose disease on persimmon. It is an important fungus that seriously threatens persimmon production worldwide. To date, the genome of *C. horii* has not been reported, although various genomic sequences are available for the *C. gloeosporioides* species complex to which the pathogen belongs. Here, we isolated *C. horii* FJ-1 from persimmon in China and presented the high-quality and completed genome of *C. horii* with a total size of 74.32 Mb by using Nanopore sequencing technology. The genome of *C. horii* FJ-1 was assembled into 42 contigs with an N₅₀ of 3.1 Mb. In total, 15,529 genes were predicted, among which 358 genes encoded the candidate effector proteins. Our results provide the first genome resource for determining the pathogenic mechanism of *C. horii* to plant and identifying potential resistant candidates for persimmon-breeding programs.

The genus *Colletotrichum* comprises nine major clades (Cannon et al. 2012) and includes 252 species and 15 complexes (Talhinhas and Baroncelli 2021). *Colletotrichum* spp. are important plant pathogens that can cause anthracnose in economic crops such as persimmon (Zhang and Xu 2003), banana (Chillet et al. 2006), mango (Li et al. 2019), chili (Than et al. 2008), tomato (Barksdale 1972), and strawberry (Xiao et al. 2004). The genus has been voted as one of the top 10 fungal plant pathogens in molecular plant pathology (Dean et al. 2012).

Persimmon (*Diospyros kaki* Thunb.), one of the most important fruit trees in the world, is primarily cultivated in East Asia (Kanzaki and Nara 2016; Luo and Wang 2008). Persimmon anthracnose is a serious disease occurring in many countries where persimmon is cultivated. Based on molecular and morphological analyses, the pathogen of persimmon anthracnose was renamed *C. horii* by Weir and Johnston (2010) and belongs to the *C. gloeosporioides* species complex. Anthracnose symptoms first appear in the spring as spots; then, the spots develop into dark lesions. Under favorable conditions, adjacent lesions may coalesce, increasing in size until the entire twig is infected (Xie et al. 2010). As a destructive disease of persimmon trees, *C. horii* can cause leaf defoliation, fruit rot, and even the death of the whole plant (Zhang and Xu 2005). In some growing areas of persimmon in China, the anthracnose could reduce the yield of persimmon by more than 50% (Deng et al. 2019). However, due to a lack of the reference genome, the molecular pathogenic mechanism of *C. horii* to the plant remains unknown. Hence, the genome analysis of *C. horii* is a significant step to understand its pathogenic mechanism and persimmon–*C. horii* interactions, thereby providing a reference for the study of effective management strategies.

The strain FJ-1 was isolated from diseased persimmon branches (*D. kaki* L. ‘Fuping Jianshi’) from the National Field Genebank for Persimmon, Yangling, Shaanxi, China

[†]Corresponding author: C. Guan; guanchangfei@nwafu.edu.cn

*The e-Xtra logo stands for “electronic extra” and indicates that supplementary figures and supplementary tables are published online.

The author(s) declare no conflict of interest.

Accepted for publication 2 January 2022.

Table 1. Statistical features of the *Colletotrichum horii* FJ-1 genome

Features	Statistics
Genome size (bp)	74,324,508
Genome coverage	92.84x
Contig number	42
Contig N ₅₀ (bp)	3,135,412
Maximum contig length (bp)	8,827,220
GC content (%)	46.79
Complete BUSCOs (%)	97.93
Gene number	15,529
Average gene length (bp)	1,863.38
Repetitive sequence (%)	21.30
Ribosomal RNA	101
Transfer RNA	413
Pseudogene	2
Signal peptide	2,111
Secreted proteins	1,613
Candidate effectors	358

(34°17'42.80" N, 108°04'8.21" E). Genomic DNA was extracted from 7-day-old cultures grown on potato dextrose agar (PDA) using the E.Z.N.A. fungal DNA kit (Omega Bio-Tek). Based on the concatenated dataset of *actin*, *chalcone synthase*, *glyceraldehyde 3-phosphate dehydrogenase*, and β -*tubulin-2* gene sequences of *C. horii* and those of other *C. gloeosporioides* species complex members (Talhinhas and Baroncelli 2021), a maximum-likelihood phylogenetic tree was constructed with MEGA v7.0 (Kumar et al. 2016), and the bootstrap value was set as 1,000. The phylogenetic tree showed that the isolate FJ-1 clustered with the *C. horii* strain with high bootstrap support (Supplementary Fig. S1). The genome of *C. horii* strain FJ-1 was sequenced using Oxford Nanopore Technologies and sequencing was performed at the Beijing Biomarker Technologies Co., Ltd. (Beijing, China). The purity, concentration, and integrity of genomic DNA were examined by Nanodrop, Qubit, and 0.35% agarose gel electrophoresis, respectively. The BluePippin automatic nucleic acid recovery system was used to recover large fragments of DNA. The library was constructed with an SQK-LSK109 kit and sequenced with a Nanopore III sequencer. After filtering out low-quality and short-length subreads (<2,000 bp), the remaining subreads were assembled to contigs without gaps by using NECAT v0.0.1 (<https://github.com/xiaochuanle/NECAT>) software.

In total, 6.9 GB of raw data (92.84x genome coverage) was obtained, and a complete genome was assembled with a total length of 74.32 Mb. The genome contained 42 contigs, with an N₅₀ of 3.1 Mb, a maximum length of 8.8 Mb, and a GC content of 46.79% (Table 1; Supplementary Tables S1 and S2). The completeness of the genome assembly was assessed using BUSCO v4.0 (Manni et al. 2021), resulting in a coverage rate of 97.93% (Supplementary Fig. S2). With the help of LTR_FINDER v1.05 (Xu and Wang 2007), MITE-Hunter (v 2010-09-29) (Han and Wessler 2010), RepeatScout v1.0.5 (Price et al. 2005), and PILER-DF v2.4 (Edgar and Myers 2005), the repetitive sequence database of the fungal genome was constructed based on the principles of structural prediction and de novo prediction. The database was classified by PASTECClassifier v2.0 (Wicker et al. 2007) and then merged with Repbase v14.02 (Jurka et al. 2005) database as the final repetitive sequence database. Based on the constructed repetitive sequence database, the repetitive sequence of this fungus was predicted using RepeatMasker v4.0.6 (Smit et al. 2013-2015) software, and a repetitive sequence of 15.83 Mb with a repetitive sequence ratio of 21.30% was obtained. Genscan v1.0 (Burge and Karlin 1997), Augustus v2.4 (Stanke and Waack 2003), GlimmerHMM v3.0.4 (Majoros et al. 2004), GeneID v1.4 (Blanco et al. 2007), and SNAP (v 2006-07-28) (Korf 2004) were used to predict the gene structure from scratch. GeMoMa v1.3.1 (Keilwagen et al. 2016) was used for prediction based on homologous protein, and EVM v1.1.1 (Haas et al. 2008) was adopted to integrate the prediction results obtained with the above two methods. Consequently, 15,529 genes were obtained, with an average length of 1,863.38 bp. For the prediction of noncoding RNAs, tRNAscan-SE v1.3 software (Lowe and Eddy 1997) was used to predict transfer RNA in the genome, and Infernal1.1 (Nawrocki and Eddy 2013) was used to predict ribosomal RNA in the genome based on the Rfam (Nawrocki et al. 2014) database. By comparing the predicted protein sequences with those

in the Swiss-Prot database and using GenBlastA v1.0.4 software (She et al. 2009), we searched the genome for homologous gene sequences (possible genes). Then, GeneWise v2.4.1 (Birney et al. 2004) was used to search for early termination codons and frameshift mutations in the gene sequences; thus, two pseudogenes were obtained (Table 1).

Genome function annotation includes general database annotation and proprietary database annotation. The general annotation results were obtained by comparing the predicted gene sequences with functional databases such as EuKaryotic Orthologous Groups (KOG) (Tatusov et al. 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2004), Swiss-Prot (Boeckmann et al. 2003), TrEMBL (Boeckmann et al. 2003), and nonredundant (NR) (Deng et al. 2006). Among all coding genes, 15,449 (99%) genes were annotated by the NR database, 7,218 (46%) were classified using the gene ontology database, 3,620 (23%) were assigned to the KEGG database, and 7,080 (45%) were mapped to the KOG database (Supplementary Figs. S3, S4, and S5; Supplementary Table S3). The functional annotation of 1,034 carbohydrate-active enzyme (CAZy) genes was obtained using HMMER v2.2.1 software (Eddy 1998) in the CAZy database (Cantarel et al. 2009) (Supplementary Fig. S5). In total, 141 transport proteins and 4,964 pathogen–host interaction proteins were annotated by the TCDB (Saier et al. 2006) and PHI (Winnenburg et al. 2006) databases, respectively. Based on the protein sequences of all predicted genes, 2,111 proteins containing signal peptides were predicted using SignalP 4.0 software (Petersen et al. 2011). The remaining 1,613 proteins after removing the protein containing a transmembrane helix from the predicted protein containing a signal peptide are secretory proteins. Then, we predicted 358 fungal candidate effector proteins after further analysis of secreted proteins by EffectorP v1.0 software (Sperschneider et al. 2015).

Additionally, we compared the FJ-1 genome assembly statistics with those of seven strains (Cg56, ICMP18580, HC540, Cg1H19, Cg01, hn1205-1, and CF75) belonging to *C. gloeosporioides* complexes in the NCBI database (Supplementary Table S4). The genome assembly statistics of *C. horii* strain FJ-1 were comparable with those of other *C. gloeosporioides* species complex.

Taken together, these findings provide a significant discovery for the genome analysis of *C. horii*, which can help us to better understand its pathogenic mechanism and persimmon–*C. horii* interactions. In turn, such understanding can contribute to the effective control methods of this disease. This Whole-Genome Shotgun project has been deposited at DNA Data Bank of Japan/European Nucleotide Archive/GenBank under the accession JAHZMN000000000 (BioProject: PRJNA748574, BioSample: SAMN20339690). The version described in this article is version JAHZMN010000000.

Literature Cited

- Barksdale, T. H. 1972. Resistance in tomato to six anthracnose fungi. *Phytopathology* 62:660-663.
- Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and Genomewise. *Genome Res.* 14:988-995.
- Blanco, E., Parra, G., and Guigó, R. 2007. Using geneid to identify genes. *Curr. Protoc. Bioinf.* 4:4.3.1-4.3.28.
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31:365-370.
- Burge, C., and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268:78-94.
- Cannon, P. F., Damm, U., Johnston, P. R., and Weir, B. S. 2012. *Colletotrichum*—Current status and future direction. *Stud. Mycol.* 73:181-213.
- Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. 2009. The Carbohydrate-Active EnZymes database (CAZy): An expert resource for Glycogenomics. *Nucleic Acids Res.* 37:D233-D238.
- Chillet, M., Hubert, O., Rives, M. J., and de Lapeyre de Bellaire, L. 2006. Effects of the physiological age of bananas on their susceptibility to wound anthracnose due to *Colletotrichum musae*. *Plant Dis.* 90:1181-1185.
- Dean, R., Van Kan, J. A. L., Pretorius, Z. A., Hammond-Kosack, K. E., Di Pietro, A., Spanu, P. D., Rudd, J. J., Dickman, M., Kahmann, R., Ellis, J., and Foster, G. D. 2012. The Top 10 fungal pathogens in molecular plant pathology. *Mol. Plant Pathol.* 13:414-430.
- Deng, Q. E., Ding, X. Y., Li, J. A., Cui, L. K., and Xu, J. Q. 2019. Morphological characteristics and genetic diversity of *Colletotrichum horii* infecting persimmon tree in China. *Eur. J. Plant Pathol.* 156:437-449.
- Deng, Y. Y., Li, J. Q., Wu, S. F., Zhu, Y. P., Chen, Y. W., and He, F. C. 2006. Integrated nr database in protein annotation system and its localization. *Comput. Eng.* 32:71-74.
- Eddy, S. R. 1998. Profile hidden Markov models. *Bioinformatics* 14:755-763.
- Edgar, R. C., and Myers, E. W. 2005. PILER: Identification and classification of genomic repeats. *Bioinformatics* 21:i152-i158.
- Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., and Wortman, J. R. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9:R7.
- Han, Y., and Wessler, S. R. 2010. MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* 38:e199.
- Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110:462-467.
- Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32:D277-D280.
- Kanzaki, S., and Nara, N. J. 2016. The origin and cultivar development of Japanese persimmon (*Diospyros kaki* Thunb.). *J. Jpn. Soc. Food Sci. Technol.* 63:328-330.

- Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Jan, G., and Frank, H. 2016. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* 44:e89.
- Korf, I. 2004. Gene finding in novel genomes. *BMC Bioinf.* 5:59.
- Kumar, S., Stecher, G., and Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33:1870-1874.
- Li, Q. L., Bu, J. Y., Shu, J., Yu, Z. H., Tang, L. H., Huang, S. P., Guo, T. X., Mo, J. Y., Luo, S. M., Solangi, G. S., and Hsiang, T. 2019. *Colletotrichum* species associated with mango in southern China. *Sci. Rep.* 9:18891.
- Lowe, T. M., and Eddy, S. R. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25:955-964.
- Luo, Z., and Wang, R. 2008. Persimmon in China: Domestication and traditional utilizations of genetic resources. *Adv. Hortic. Sci.* 22:239-243.
- 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.
- Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. 2021. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* 38:4647-4654.
- Nawrocki, E. P., Burge, S. W., Bateman, A., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and Finn, R. D. 2014. Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Res.* 43:D130-D137.
- Nawrocki, E. P., and Eddy, S. R. 2013. Infernal 1.1: 100-Fold faster RNA homology searches. *Bioinformatics* 29:2933-2935.
- Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. 2011. SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* 8: 785-786.
- Price, A. L., Jones, N. C., and Pevzner, P. A. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* 21:i351-i358.
- Saier, M. H., Tran, C. V., and Barabote, R. D. 2006. TCDB: The Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res.* 34:D181-D186.
- She, R., Chu, J. S. C., Wang, K., Pei, J., and Chen, N. S. 2009. GenBlastA: Enabling BLAST to identify homologous gene sequences. *Genome Res.* 19: 143-149.
- Smit, A. F. A., Hubley, R., and Green, P. 2013-2015. RepeatMasker Open-4.0. <http://www.repeatmasker.org>
- Sperschneider, J., Gardiner, D. M., Dodds, P. N., Tini, F., Covarelli, L., Singh, K. B., Manners, J. M., and Taylor, J. M. 2015. EffectorP: Predicting fungal effector proteins from secretomes using machine learning. *New Phytol.* 210: 743-761.
- Stanke, M., and Waack, S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19:ii215-ii225.
- Talhinhas, P., and Baroncelli, R. 2021. *Colletotrichum* species and complexes: Geographic distribution, host range and conservation status. *Fungal Divers.* 110:109-198.
- Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. 2000. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28:33-36.
- Than, P. P., Prihastuti, H., Phoulivong, S., Taylor, P. W. J., and Hyde, K. D. 2008. Chili anthracnose disease caused by *Colletotrichum* species. *J. Zhejiang Univ. Sci. B* 9:764-778.
- Weir, B. S., and Johnston, P. R. 2010. Characterisation and neo-typification of *Gloeosporium kaki* Hori as *Colletotrichum horii* nom. nov. *Mycotaxon* 111: 209-219.
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., and Schulman, A. H. 2007. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8:973-982.
- Winnenburg, R., Baldwin, T. K., Urban, M., Rawlings, C., Köhler, J., and Hammond-Kosack, K. E. 2006. PHI-base: A new database for pathogen host interactions. *Nucleic Acids Res.* 34:D459-D464.
- Xiao, C. L., Mackenzie, S. J., and Legard, D. E. 2004. Genetic and pathogenic analyses of *Colletotrichum gloeosporioides* from strawberry and noncultivated hosts. *Phytopathology* 94:446-453.
- Xie, L., Zhang, J. Z., Cai, L., and Hyde, K. D. 2010. Biology of *Colletotrichum horii*, the causal agent of persimmon anthracnose. *Mycology* 1:242-253.
- Xu, Z., and Wang, H. 2007. LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35:W265-W268.
- Zhang, J. Z., and Xu, T. 2003. Various stages and amount of *Colletotrichum gloeosporioides* on overwintering twigs of persimmon. *J. Plant Prot.* 30:437-438.
- Zhang, J. Z., and Xu, T. 2005. Cytological characteristics of the infection in different species, varieties and organs of persimmon by *Colletotrichum gloeosporioides*. *Mycosystema* 24:116-122.