Genome Assembly and Transcriptome of *Colletotrichum sublineola* CsGL1, a New Resource to Study Anthracnose Disease in Sorghum

Patricia Baldrich,¹ Timothy Chaya,² Jeffrey L. Caplan,² and Blake C. Meyers^{1,3,†}

¹ Donald Danforth Plant Science Center, St. Louis, MO 63132, U.S.A.

² Department Plant and Soil Sciences, University of Delaware, Newark, DE 19716, U.S.A.

³ University of Missouri–Columbia, Division of Plant Sciences, 52 Agriculture Lab, Columbia, MO 65211, U.S.A.

Abstract

Colletotrichum species are globally distributed and well known as members of a destructive phytopathogenic genus, causing the anthracnose disease in a wide variety of crops and fruits. *Colletotrichum sublineola* is the causal agent of the anthracnose disease in sorghum, causing losses of up to 50% in yield. Here, we used PacBio sequencing combined with RNA-seq to generate a chromosome-level assembly and annotation of the *Colletotrichum sublineola* strain CsGL1.

Genome Announcement

Sorghum [*Sorghum bicolor* (L.) Moench] is a cereal native to northeastern Africa and is cultivated around the world for human nutrition, animal feed, and biofuel production. One of the most important diseases that affect sorghum is anthracnose, caused by the ascomycete fungus *Colletotrichum sublineola*. In warm and humid regions, *C. sublineola* can cause yield losses of up to 50% in susceptible sorghum cultivars (Tesso et al. 2012) by affecting all aerial parts of the plant. The genome assembled reported here is from strain CsGL1, which was collected and isolated from Kansas (U.S.A.) and has been shown to cause disease in sorghum (Xu 2019).

The two existing drafts of *C. sublineola* consists of 1,625- and 548-sequence scaffolds with total lengths of 46.75 and 64.86 Mbp, respectively (Table 1) (Baroncelli et al. 2014). In order to improve the assembly, we cultured strain GL1 on Mathur's media at 25°C, 12-h day and night cycle. We collected spores by washing four plates, resulting in a total of 1.45 g of spores, and extracted high-molecular weight genomic DNA (gDNA) using a published method (Schwessinger and Rathjen 2017). We assessed the quality of the gDNA with a Qubit 3.0, created gDNA libraries using SMRTbell Express template prep kit 2.0 (PacBio), and sequenced the resulting libraries using a PacBio Sequel II system at the University of Delaware Sequencing and Genotyping Center. We obtained an approximate coverage of 320x, considerably reducing the risk of error rate. We assembled the reads using Canu v2.0 (Koren et al. 2017) to obtain 78 contigs with a distribution between approximately 6,300 to 14 Kb, with a total genome size of 67,994,409 bp (approximately 68 Mb) (Table 1). To assess the completeness of the genome, we used Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.1.0 (Simão et al. 2015) and determined that our assembly was

Funding

This work is supported by the U.S. Department of Energy, Office of Science, Biological Systems Science Division, Biomolecular Characterization and Imaging Science program, under award number DE-SC0020348. Access to the Sequencing and genotyping Center was supported by a grant from the National Institutes of Health National Institute of General Medical Sciences (P20 GM103446) and the State of Delaware.

[†]Corresponding author: B. Meyers; bmeyers@danforthcenter.org

The author(s) declare no conflict of interest.

Accepted for publication 12 July 2021.

Copyright © 2021 The Author(s). This is an open access article distributed under the CC BY-NC-ND 4.0 International license.

Table 1. Summary of the new assembly compared with the two previous assemblies of other strains of *Colletotrichum sublineola*^a

Variables	Strains		
	CsGL1	TX430BB	CgSI1
Genome Assembly size (bp)	67,994,409	46,755,813	64,848,627
Number of contigs	78	1,625	548
Contig N ₅₀ (bp)	2,691,276	70,717	339,276
Contig N ₉₀ (bp)	855,749	13,454	66,235
Smallest contig (bp)	14,431		2,009
Largest contig (bp)	6,303,522	423,147	2,214,490
GC Content (%)	46.04	52.70	45.35
BUSCO completeness	99.20%		

^a Assembled using Canu v2.0.

99.2% complete, 97.6% single copy, and 1.6% duplicated, with one (0.4%) fragmented BUSCO and one (0.4%) missing BUSCO (using eukaryote lineage).

To better predict protein-coding genes, we extracted RNA for sequencing from two independent sets of plates, following a previously described method (Schumann et al. 2013). Spores were isolated using sterile water and excess water was then removed, using a sterile Buchner funnel and Wattman filter paper. We obtained 2.65 and 1.47 g of spores, respectively, for each isolation. We prepared two RNA-seq libraries, using PolyA enrichment Perkin Elmer NEXTflex Rapid directional RNA-seq V1, and sequenced them on an Illumina NextSeq 500 instrument, with a paired-end, 155-cycle run, resulting in 49,505,281 and 55,860,815 reads for each replicate. We performed the annotation using an RNA-seq-based prediction. First, we removed the adapters from the RNA-seq reads, using Cutadapt v2.9 (Martin 2011), and obtained 49,453,459 and 55,802,633 clean reads; we retained approximately 99.9% of the reads. Next, we assembled the reads using Trinity software v2.11 (Grabherr et al. 2011), predicting a total of 31,251 genes and 66,994 transcripts, with a global GC content of 54.40%. Next, we used these transcripts to identify candidate coding regions using Funannotate v1.7.0 and we obtained 10,898 genes corresponding to 11,399 transcripts as potential coding genes. We assessed the completeness of the transcriptome by running BUSCO v4.1.0 (Simão et al. 2015) and obtained 93.7% complete, 92.5% single copy, and 1.2% duplicated BUSCOs, with three (1.2%) fragmented BUSCO and 13 (5.1%) missing BUSCOs (using eukaryote lineage).

The genome described here provides an improved version of a *Colletotrichum sublineola* strain isolated from sweet sorghum in field conditions. This reference genome will provide new insights into the biology of this pathogen, the interaction with its hosts, and the evolution of the *Colletotrichum* family.

Data Availability

Long-read sequencing data and raw RNA-seq data can be found in the National Center for Biotechnology Information Short Read Archive under Bioproject number PRJNA659716.

Acknowledgments

We are grateful to T. Mengiste (Purdue University, West Lafayette, IN, U.S.A.) for providing *Colletotrichum sublineola* CsGL1 and to the University of Delaware DNA Sequencing & Genotyping Center for their assistance and expertise.

Author-Recommended Internet Resources

Funannotate v1.7.0: <https://github.com/nextgenusfs/funannotate>

SRA Bioproject PRJNA659716: <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA659716>

Literature Cited

Baroncelli, R., Sanz-Martín, J. M., Rech, G. E., Sukno, S. A., and Thon, M. R. 2014. Draft genome sequence of *Colletotrichum sublineola*, a destructive pathogen of cultivated sorghum. *Genome Announc.* 2.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. 2011. Full-length

- transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29:644-652.
- Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. 2017. Canu: Scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* 27:722-736.
- Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17:10-12.
- Schumann, U., Smith, N. A., and Wang, M.-B. 2013. A fast and efficient method for preparation of high-quality RNA from fungal mycelia. *BMC Res. Notes* 6:71.
- Schwessinger, B., and Rathjen, J. P. 2017. Extraction of high molecular weight DNA from fungal rust spores for long read sequencing. *Methods Mol. Biol.* 1659:49-57.
- 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.
- Tesso, T., Perumal, R., Little, C. R., Adeyanju, A., Radwan, G. L., Prom, L. K., and Magill, C. W. 2012. Sorghum pathology and biotechnology—A fungal disease perspective: Part II. Anthracnose, stalk rot, and downy mildew. *Eur. J. Plant Sci. Biotechnol.* 6:31-34.
- Xu, X. 2019. Identification and mapping of anthracnose resistance genes in sorghum [*Sorghum bicolor* (L.) Moench]. Presented at Sorghum in the 21st Century: Food, Feed and Fuel in a Rapidly Changing World, Cape Town, South Africa. United States Agency of International Development and Kansas State University, Manhattan, KS, U.S.A.