Genome Sequence of *Phoma sorghina* var. *saccharum* That Causes Sugarcane Twisted Leaf Disease in China

Yixue Bao,^{1,2} Wei Yao,^{1,2} Zhenzhen Duan,² Charles A. Powell,³ Baoshan Chen,¹ and Muqing Zhang^{1,3,†}

¹ State Key Lab for Conservation and Utilization of Sub-tropical Agri-Biological Resources, Guangxi University, 530005, China

² Guangxi Key Lab for Sugarcane Biology, Guangxi University, 530005, China

³ Indian River Research and Education Center, IFAS, University of Florida, 2199 South Rock Rd., FL 34934, U.S.A.

Abstract

Phoma sorghina var. *saccharum* is a fungal pathogen that causes sugarcane twisted leaf disease in China. Here, we report complete genome assemblies of the *Phoma sorghina* var. *saccharum* isolate BS2-1, generated using single-molecule real-time sequencing. We present a high-quality genome sequence of a *Phoma* isolate that was assembled into 22 contigs with an N₅₀ length of 1.92 Mb, a total length of 33.12 Mb, and a GC content of 52.12%. A total of 7,870 genes were annotated, using a combination of gene prediction tools, including 281 noncoding RNAs, 515 genes encoding carbohydrate-active enzymes, 2,440 genes associated with pathogen-host interactions, and 583 genes encoding secreted proteins. The complete genome sequence will be useful for understanding host-pathogen interaction and for improving disease management strategies.

Funding

This work was supported by China Agricultural Research System grant 170109, Scientific Research and Technology Development Program of Guangxi grants Guike AA17202042-7, and Guike AD17129002, and the Innovation Project of Guangxi Graduate Education grant YCBZ 2018014.

Keywords

Phoma sorghina var. *saccharum*, sugarcane twisted leaf disease

Genome Announcement

The genus *Phoma* is one of the most abundant fungal genera, and its members are ubiquitous in the environment and occupy numerous ecological niches. Many *Phoma* species and related coelomycetes are plant pathogens that can cause leaf and stem spots. The diverse host ranges associated with the genus *Phoma* include corn, grain, citrus, sorghum, and sugarcane (de Gruyter et al. 2009). Sugarcane twisted leaf disease caused by *Phoma sorghina* var. *saccharum* results in a 5 to 10% loss in yield, which will pose a potential threat to cane production in China. The field symptoms appear as yellowing on midribs and leaf margins, which spread further to the entire leaf, along with curling and twisting of crown leaves (Lin et al. 2014).

Four sequences of the *Phoma* sp. genome are available, three of which are draft genomes produced using short reads of Illumina, including *Phoma herbarum* (BCGR00000000.1), *Phoma koolunga* (RWYX00000000.1), and *Phoma* sp. RAV-16-625 (SAUA00000000.1). The other complete genome of *Phoma* sp., XZ068 (RBKR00000000.1), is sequenced using long

Yixue Bao, Wei Yao, and Zhenzhen Duan contributed equally to this manuscript.

†Corresponding author: M. Zhang; zmuqing@163.com

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

The author(s) declare no conflict of interest.

Accepted for publication 26 May 2020.

Table 1. Summary of *Phoma sorghina* genome assembly and comparison with other *Phoma* sp.

Variable	<i>Phoma sorghina</i>	<i>Phoma herbarum</i>	<i>Phoma koolunga</i>	<i>Phoma</i> sp.	<i>Phoma</i> sp.
Intraspecific name	BS2-1	JCM 15942	PIN-15-415	RAV-16-625	XZ068
Total sequence length (bp)	33,124,632	39,263,577	50,537,364	33,436,265	40,199,292
Number of contigs	22	498	21,399	383	33
Contig N ₅₀ (bp)	1,920,476	559,926	30,743	195,204	1,497,047
% GC content	52.12%	51.70%	45.70%	51.90%	50.40%
Genome coverage	97x	462x	140x	94x	140x
Assembly method	Canu v. 1.5	ALLPATHS-LG v. 52488	SPAdes v. 3.11.1	SPAdes v. 3.11.1	Celera Assembler v. 8.3
Sequencing technology	PacBio RS II; Illumina	HiSeq 2500	Illumina NextSeq	Illumina NextSeq	PacBio; Illumina
WGS Project	VXJJ000000000	BCGR000000000	RWYX000000000	SAUA000000000	RBKR000000000
Reference	This study	Unpublished	Unpublished	Unpublished	Zhai et al. 2019

reads of PacBio and was assembled by the Celera Assembler (v.8.3) to predict complete gene models and clusters for secondary metabolites (Zhai et al. 2019).

Here, we reported the genome sequence of *Phoma sorghina* var. *saccharum* BS2-1, which was isolated from symptomatic sugarcane twisted leaves collected in Guangxi, China. The BS2-1 strain was purified by single-spore isolation and was identified using morphological observations and phylogenetic analysis of 28S rDNA (large subunit, LR0R/LR7), 18S rDNA (small subunit, NS1/NS4), internal transcribed spacer regions 1 and 2, 5.8S nuclear ribosomal DNA (ITS1/ITS4), β -tubulin (TUB2Fd/TUB4Rd), and the alpha translation elongation factor (Efd/EF1-2218R) (accession numbers KF171356 to KF171360, KJ669181 to KJ669188). High-quality genomic DNA was extracted from 7-day-old mycelia grown in vitro on potato dextrose water medium, using a modified Fungal DNA Midi Kit (Omega Bio-Tek, Inc.). The genome of the BS2-1 isolate was sequenced using both the Illumina HiSeq4000 and the PacBio RSII Sequel platforms according to manufacturer protocol. The paired-end 300-bp library produced 22,730,042 short reads that resulted in 3.4 Gb of data and 103x coverage of the genome. To improve the assembly, a long-read library of approximately 20 Kb of fragments was constructed, using the single-molecule real-time Bell template prep kit. Sequencing was performed in three PacBio RSII cells with P6-C4 chemistry. Self-corrections of subreads were performed using an error correction model in the Falcon package v1.8.7 (Beckett et al. 2014). The EC-tool was used to correct the raw data, and high-accuracy subreads that included 3,200,762,563 bp (approximately 97x coverage) were generated and were de novo assembled using Canu v1.5 (Koren et al. 2017). The entire genome consisted of 22 contigs totaling 33.12 Mb of sequence, with 52.12% GC content. The N₅₀ of the genome was 1.92 Mb. In comparison, the genome of *Phoma* sp. XZ068 (RBKR000000000.1) is 40.20 Mb with 50.4% GC content, consisting of 33 contigs with an N₅₀ value of 1.4 Mb and L₅₀ of 11 (Table 1). Twenty-two contigs ranged in length from 28.7 Kb to 4.02 Mb. The presence of telomeric repeats (TTAGGG) at the ends of a contig was indicative of a complete chromosome (Seidl et al. 2015). Eleven contigs had telomeric repeat sequences on both ends and nine contigs having telomeric repeats at the 3' end. Therefore, the genome contained at least 11 complete chromosomes. The genome also had one integrated circular mitochondrial DNA that was approximately 66.7 Kb. Ten contigs without telomeric repeats on both ends might represent dispensable chromosomes and require further characterization (Supplementary Table S1). The quality of the assembled genome was assessed using benchmarking universal single-copy orthologs (BUSCO v3.0.2) with the lineage dataset fungi_odb9. A total of 286 of 290 complete and single-copy BUSCOs were identified in our BS2-1 genome; of the remaining four, one was fragmented and three were absent (Simão et al. 2015).

Protein coding genes were identified in the assembled genome, using both de novo and homology methods available with Augustus v2.4 (Stanke and Waack 2003), GlimmerHMM v3.0.4 (Majoros et al. 2004), SNAP (Korf 2004), and GeMoMa (Birney et al. 2004). Consensus gene structures were integrated using Evidence Modeler v1.1.1 (Haas et al. 2008). All gene models were annotated according to the best alignment match obtained from Kyoto Encyclopedia of Genes and Genomes (KEGG), Swiss-Prot, and nonredundant protein (NR) databases using BLASTP in BLAST+ package v2.2.6 (Boeckmann et al. 2003; Deng et al. 2006; Kanehisa et al. 2004). A total of 7,870 protein-encoded genes were annotated in the assembled genome. Among the annotated genes that had homologs in other species, 2,767 (35.16%) genes were mapped to the KEGG database, 5,239 (66.57%) had similarity to

proteins in the Swiss-Prot database, and 7,851 (99.76%) were mapped to the NR database. Homeoboxes and homeodomains, which are associated with leaf curling (Johannesson et al. 2001), were identified in the annotated genes (Supplementary Table S2). A total of 515 annotated genes were related to carbohydrate-active enzymes, as indicated by a CAZyme database search (Cantarel et al. 2009), 115 were associated with transport, as indicated by a Transporter Classification Database search, 2,440 genes were predicted to be involved in interactions between the pathogen and host, as evidenced by a Pathogen-Host Interactions search, and 583 genes were predicted to produce a secreted protein (Saier et al. 2006; Winnenburg et al. 2006). BLASTn and tRNAscan-SE were used to identify 281 noncoding RNAs, including 168 transfer RNAs and 113 ribosomal RNAs (Altschul et al. 1990; Lowe and Eddy 1997). Several gene clusters were identified in search of the antiSMASH fungal v4.0.0 server, including two nonribosomal peptide synthetases (NRPS), three terpenes, eight NRPS-like, and 11 polyketide synthases. Notably, BS2-1 had a specific and complete gene cluster for phomopsins, a gene family involved in toxic metabolite production, that are being pursued as potential leads for the development of antitumor drugs.

In summary, we reported a highly contiguous genome assembly for *Phoma sorghina* var. *saccharum*, the causative agent of sugarcane twisted leaf disease. Eleven chromosomes could be assembled from telomere to telomere. This whole-genome sequencing data will facilitate exploration of the host-pathogen interaction of this fungus and provide a resource for genes that could be targeted for therapeutic development. The genome sequence described in this article has been deposited in GenBank under the accession number VXJJ000000000 (BioProject: PRJNA565496; BioSample: SAMN12748852).

Author-Recommended Internet Resource

antiSMASH fungal v4.0.0 server: <https://fungismash.secondarymetabolites.org>

Literature Cited

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.
- Beckett, S. J., Boulton, C. A., and Williams, H. T. 2014. FALCON: A software package for analysis of nestedness in bipartite networks. *F1000 Res.* 3:185.
- Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and Genomewise. *Genome Res.* 14:988-995.
- 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.
- 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 (Database): D233-D238.
- de Gruyter, J., Aveskamp, M. M., Woudenberg, J. H., Verkley, G. J., Groenewald, J. Z., and Crous, P. W. 2009. Molecular phylogeny of *Phoma* and allied anamorph genera: Towards a reclassification of the *Phoma* complex. *Mycol. Res.* 113:508-519.
- 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.* 32:71-72.
- 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.
- Johannesson, H., Wang, Y., and Engström, P. 2001. DNA-binding and dimerization preferences of *Arabidopsis* homeodomain-leucine zipper transcription factors in vitro. *Plant Mol. Biol.* 45:63-73.
- 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.
- 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.
- Korf, I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* 5:59.
- Lin, Z. Y., Que, Y. X., Deng, Z. H., Xu, S. Q., Rao, G. P., and Zhang, M. Q. 2014. First report of *Phoma* sp. causing twisting and curling of crown leaves of sugarcane in the mainland of China. *Plant Dis.* 98:850.
- 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.
- 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.
- Saier, M. H., Jr., 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.
- Seidl, M. F., Faino, L., Shi-Kunne, X., van den Berg, G. C., Bolton, M. D., and Thomma, B. P. 2015. The genome of the saprophytic fungus *Verticillium tricorpus* reveals a complex effector repertoire resembling that of its pathogenic relatives. *Mol. Plant-Microbe Interact* 28:362-373.
- 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.
- Stanke, M., and Waack, S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19 (Suppl 2):ii215-ii225.
- 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.
- Zhai, Y., Li, Y., Zhang, J., Zhang, Y., Ren, F., Zhang, X., Liu, G., Liu, X., and Che, Y. 2019. Identification of the gene cluster for bistropolone-humulene meroterpenoid biosynthesis in *Phoma* sp. *Fungal Genet. Biol.* 129:7-15.