Genomics Data 8 (2016) 67–69

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

Genome-wide functional annotation of *Phomopsis longicolla* isolate MSPL 10-6

Omar Darwish [a,1], Shuxian Li [b,*,1], Benjamin Matthews [c], Nadim Alkharouf [a,*]

[a] Department of Computer and Information Sciences, Towson University, MD 21252, USA
[b] United States Department of Agriculture, Agricultural Research Service (USDA-ARS), Crop Genetics Research Unit, Stoneville, MS 38776, USA
[c] USDA-ARS, Beltsville Agriculture Research Center, Beltsville, MD 21075, USA

ARTICLE INFO

Article history:
Received 6 March 2016
Received in revised form 16 March 2016
Accepted 21 March 2016
Available online 6 April 2016

Keywords:
Phomopsis longicolla
Phomopsis seed decay
Soybean pathogen
Functional annotation

ABSTRACT

Phomopsis seed decay of soybean is caused primarily by the seed-borne fungal pathogen *Phomopsis longicolla* (syn. *Diaporthe longicolla*). This disease severely decreases soybean seed quality, reduces seedling vigor and stand establishment, and suppresses yield. It is one of the most economically important soybean diseases. In this study we annotated the entire genome of *P. longicolla* isolate MSPL 10-6, which was isolated from field-grown soybean seed in Mississippi, USA. This study represents the first reported genome-wide functional annotation of a seed borne fungal pathogen in the *Diaporthe–Phomopsis* complex. The *P. longicolla* genome annotation will enable research into the genetic basis of fungal infection of soybean seed and provide information for the study of soybean–fungal interactions. The genome annotation will also be a valuable resource for the research and agricultural communities. It will aid in the development of new control strategies for this pathogen. The annotations can be found from: http://bioinformatics.towson.edu/phomopsis_longicolla/download.html. NCBI accession number is: AYRD00000000.

Specifications	
Organism/cell line/tissue	*Phomopsis longicolla*
Strain	MSPL 10-6
Sequencer	Illumina HiSeq 2500 sequencer
Data format	Analyzed, gff file
Experimental factors	n/a
Experimental features	Genome annotation
Consent	n/a
Sample source location	n/a

1. Direct link to deposited data

Annotations can be found here: http://bioinformatics.towson.edu/phomopsis_longicolla/download.html.

NCBI deposited data can be found here: http://www.ncbi.nlm.nih.gov/nuccore/AYRD00000000.

2. Experimental design, materials and methods

2.1. DNA extraction, library construction, and sequencing

P. longicolla is the primary cause of Phomopsis seed decay in soybean [1,2]. An isolate of *P. longicolla*, MSPL10-6, was isolated from field-grown soybean seed in Stoneville, Mississippi, USA in 2010 using the standard seed plating procedure [3]. DNA extraction, library construction, and sequencing of *P. longicolla* MSPL10-6 was previously described in [4].

2.2. Gene prediction

The genome sequence of *P. longicolla* MSPL10-6, was previously de-novo assembled into 108 scaffolds of 500 bases or larger [4]. Gene prediction analysis was done on these scaffolds, using a combination of homology searching and de novo prediction using Augustus web server [5] with complete gene option enabled and default for the rest of the parameters. *Fusarium graminearum* was selected as the reference species due to the relatively close phylogenetic relationship to *P. longicolla* MSPL10-6.

The analysis yielded a total of 16,597 genes (average length = 1704.37 bp, total length = 28,287,360 bp, total coding length = 24,840,981 bp). Out of which 4334 genes were found to consist of a single exon (average length = 1219.1 bp). The total number of exons

[*] Corresponding authors.
E-mail addresses: shuxian.li@ars.usda.gov (S. Li), nalkharouf@towson.edu (N. Alkharouf).
[1] These authors contributed equally to the present manuscript.

http://dx.doi.org/10.1016/j.gdata.2016.03.006
2213-5960/

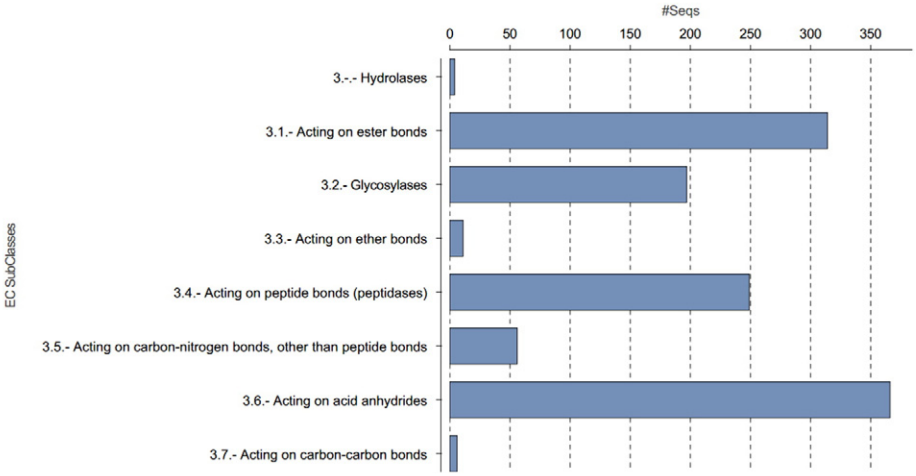

Fig. 1. Enzyme code level 3 distribution.

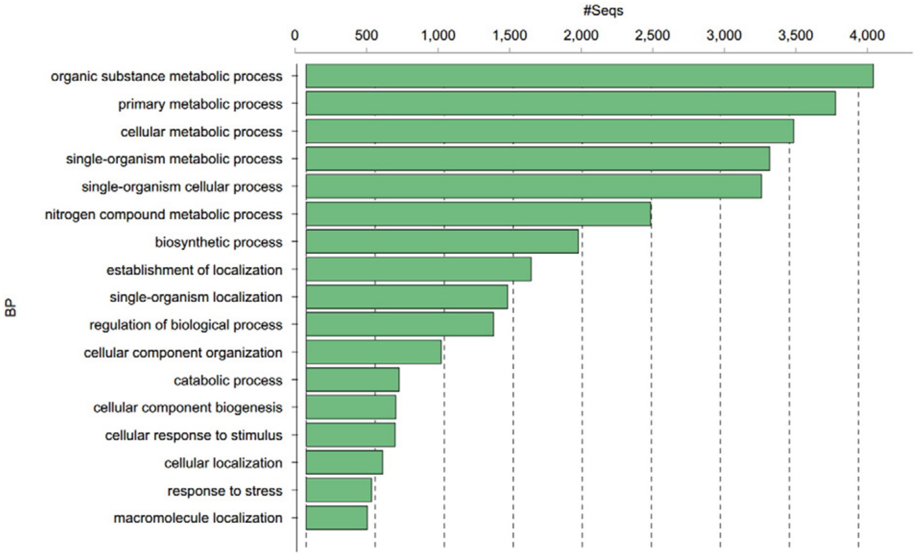

Fig. 2. GO distribution in level 3 (top 50) for biological process (BP).

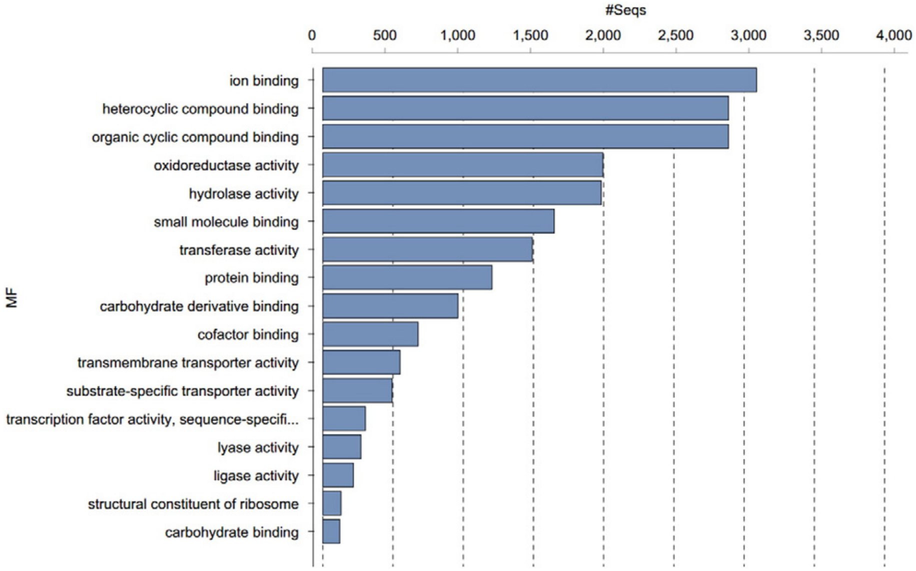

Fig. 3. GO distribution in level 3 (top 50) for molecular function (MF).

in all predicted genes was found to be 47,213 exons (average length = 361.76 bp, total length = 4,435,952 bp). The gene prediction statistics are summarized in Supplemental Table 1.

2.3. Gene functional annotation

Predicted genes were functionally annotated using Blast2GO [6]. The gene models were blasted (blastx) [7] against the NCBI non-redundant protein database. Then domain finding searches were done using InterProScan [8]. Enzyme codes and GO ontologies were then assigned to the gene models [9].

From a total of 16,596 genes, 9.64% failed to obtain significant hits with Blast. 18.01% of them returned significant sequence alignments, but cannot be linked to any Gene Ontology entries. Overall functional labels were assigned to 59.45% of the predicted genes. Enzyme codes were assigned to 15.45% of the genes. The enzyme code (EC) distributions (level 3) are summarized in Fig. 1. The GO distributions (level 3) are summarized in Fig. 2 (biological process) and Fig. 3 (molecular function).

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2016.03.006.

Acknowledgments

This work was partially supported by the USDA-ARS projects 6066-21220-012-00D.

The mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the United States Department of Agriculture. USDA is an equal opportunity provider and employer.

References

[1] T.W. Hobbs, A.F. Schmitthenner, G.A. Kuter, A new *Phomopsis* species from soybean. Mycologia 77 (1985) 535–544.

[2] S. Li, G.L. Hartman, D. Boykin, Aggressiveness of *Phomopsis longicolla* and other *Phomopsis* spp. on soybean. Plant Dis. 94 (2010) 1035–1040.

[3] S. Li, *Phomopsis* seed decay of soybean. in: A. Sudaric (Ed.), Soybean — Molecular Aspects of Breeding, Intech Publisher, Vienna, Austria. 2011, pp. 277–292.

[4] S. Li, O. Darwish, N.W. Alkharouf, B. Matthews, P. Ji, L.L. Domier, N. Zhang, B.H. Bluhm, Draft genome sequence of *Phomopsis longicolla* isolate MSPL 10-6. Genom. Data 3 (2014) 55–56.

[5] K.J. Hoff, M. Stanke, WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Res. 41 (2013) (Web Server issue): W123-8.

[6] Götz, et al., High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36 (10) (2008) 3420–3435.

[7] Altschul, et al., Basic local alignment search tool. J. Mol. Biol. 215 (1990) 403–410.

[8] E.M. Zdobnov, R. Apweiler, InterProScan — an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17 (9) (2001) 847–848.

[9] The Gene Ontology Consortium. Gene Ontology Consortium: going forward., Nucl Acids Res 43 (2015) Database issue D1049–D1056.