Genome Sequence Resource of *Bacillus* sp. RRD69, a Beneficial Bacterial Endophyte isolated from Switchgrass Plants

Zhenzhen Zhao[1], Norbert Bokros[2], Seth DeBolt[2], Piao Yang[1], Ye Xia[1, *]

[1]Department of Plant Pathology, The Ohio State University, Columbus, Ohio, 43210, USA

[2]Department of Horticulture, University of Kentucky, Lexington, Kentucky, 40546, USA

[*]Correspondence author: Ye Xia; Email: xia.374@osu.edu.

ABSTRACT: We report here the genome sequence of *Bacillus* sp. RRD69, a plant growth-promoting bacterial endophyte isolated from switchgrass plants grown on a reclaimed coal-mining site in Kentucky. RRD69 is predicted to contain 3,758 protein-coding genes with a genome size of 3.715 Mbp and a 41.41% GC content.

Keywords: endophytes, *Bacillus*, switchgrass, sequencing, genome information

Rhizosphere and endosphere localized bacterial communities are inextricably linked with improved crop performance. The *Bacillus* species, a group of plant-growth promoting bacteria (PGPB), are considered among the most well studied, effective, and commercially important PGPBs, with multiple members developed as bio-fertilizers and bio-control agents in agricultural production (Radhakrishnan et al., 2017; Hashem et al., 2019). Switchgrass (*Panicum virgatum* L.), a widely planted forage crop and ground cover plant, attracts more attention as an important biofuel crop (Xia et al., 2013). We report here the genome information of the bacterial endophyte *Bacillus* sp. RRD69 associated with switchgrass plants, which was reported to display beneficial potential for increasing the growth and weight of switchgrass plants (Xia et al., 2013).

In our study, the switchgrass samples were collected from a reclaimed coal-mining site in Kentucky (Xia et al., 2013). The roots of the collected plant samples were cut into 3-5 cm long segments, surface sterilized with the 20% bleach solution for 15 min, and rinsed with sterilized water 5 times. The surface-sterilized segments were further cut into smaller segments with the sizes of 1-1.5 cm and placed on plates with the tryptic soy agar (TSA) medium (Sigma, USA). These plates were incubated in a 26°C incubator for 3-5 days. Individual isolates emerging from those segments were separately isolated and cultured for further DNA extraction. The 16S rDNA amplifications with primers 27f (5'-GAGTTTGATCCTGGCTCA-3') and 1498r (5'-ACGGCTACCTTGTTACGACTT-3') were carried out. The amplified PCR products were further sequenced and analyzed by BLASTn searches in the NCBI databases (Devulder et al., 2003; Mignard and Flandrois, 2006). The top hits as the most possible taxonomic species were identified. Among all the isolates, the isolate of *Bacillus* sp. RRD69 was identified (Xia et al., 2013). This isolate was further cultured in tryptic soy broth (TSB) media on a rotary shaker overnight at 26℃. The bacterial cells were subject to the cetyltrimethylammonium bromide (CTAB) genomic DNA extraction (Wilson, 2001) and genome sequencing (Xia et al., 2013).

A PacBio SMRTbell library was constructed and sequenced with the PacBio RS platform. A total of 512,579 raw reads were initially generated. After quality filtering, trimming, and assembling using the Hierarchical Genome Assembly Process (HGAP, v.2.3.0) with the default settings (Chin et al., 2013), the final draft assembly contains 5 contigs in 5 scaffolds, with a total size of 3.715 Mbp and an N50 value of 975.148 kb. The input read coverage was 191.9X, and the total GC content was 41.41%. Prodigal (v.2.5) and GenePRIMP pipeline were used for the subsequent genome annotation (Hyatt et al., 2010; Pati et al., 2010). Genome annotation predicted a total of 3,884 genes, which includes 3,758 predicted protein-coding genes. These protein-coding

genes were used to search through the NCBI nonredundant database, UniProt, TIGRFam, Pfam, Kyoto Encyclopedia of Genes and Genomes (KEGG), Clusters of Orthologous Genes (COG), PANTHER, and InterPro databases (Kanehisa and Goto, 2000; Haft et al., 2003; Tatusov et al., 2003; Thomas et al., 2003; Pruitt et al., 2007; Finn et al., 2014; Consortium, 2019). For the remaining 126 genes, 26 rRNA genes, 73 tRNA genes, and 27 noncoding RNAs were identified by using the tRNAScan-SE tool (v.2.0.5), ribosomal RNA genes models from SILVA (Pruesse et al., 2007), and Rfam profiles via INFERNAL (v.1.1.3), respectively (Finn et al., 2014). CheckM (v.1.0.18) was used to estimate the completeness of the genome at 99.6% with a predicted contamination level of only 0.6% (Parks et al., 2015; Arkin et al., 2018). Using the PANTHER hidden Markov model (HMM) scoring tool panther- Score (v.2.2), the protein sequences were further mapped against the PANTHER HMM database (v.15) to functionally annotate the genes and query for the significantly overrepresented genes (Mi et al., 2019). A circular representation of selected annotations and genome characteristics generated using the Circos software is depicted in Figure 1 (Krzywinski et al., 2009). Additional gene prediction analysis and manual functional annotation were performed within the Integrated Microbial Genomes (IMG) platform developed by the Joint Genome Institute (Walnut Creek, CA) (Markowitz et al., 2012).

A previous study has reported that multiple endophytic bacteria associated with switchgrass, including several *Bacillus* strains, displayed plant growth-promoting activity (Xia et al., 2013). And the genome sequencing of one *Bacillus* strain, *Bacillus* sp. YF23 was reported (Xia et al., 2019). The goal of this study is to address the genome information of another beneficial endophyte, *Bacillus* sp. RRD69. The comparative analysis of *Bacillus* sp. RRD69 with *Bacillus* sp. YF23 is stated in Table S1. With the emerging *Bacillus* PGPBs genomic resources, we will have an improved understanding of *Bacillus* PGPBs and their associations with plants.

Data Availability. The whole-genome sequence of this bacterium has been deposited at DDBJ/EMBL/GenBank under the BioProject accession no. PRJNA322990. The associated NCBI BioSample and NCBI SRA accession numbers are SAMN05216491 and SRP088163. The associated sequence data, as well as the further information on sample preparation, genome assembly, and annotation can be found at JGI with IMG taxon id number 2681812863 and NCBI ID 1855345. The version described in this paper is the first version. The repository of scripts used to construct Figure. 1 can be found: https://github.com/nbo245/rrd69.

ACKNOWLEDGMENTS

The genome sequencing and data annotation were carried out in the U.S. Department of Energy (DOE) Joint Genome Institute (JGI), a DOE Office of Science User Facility. The project was supported under the contract DE-AC02-05CH11231. This project was also partially supported by the Hatch Project from USDA-NIFA-OHO01392 at The Ohio State University and OARDC SEED GRANT OHOA1615.

References:

Arkin, A.P., Cottingham, R.W., Henry, C.S., Harris, N.L., Stevens, R.L., Maslov, S., Dehal, P., Ware, D., Perez, F., and Canon, S. 2018. KBase: the United States department of energy systems biology knowledgebase. Nature biotechnology 36:566.

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., and Eichler, E.E. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature methods 10:563-569.

Consortium, U. 2019. UniProt: a worldwide hub of protein knowledge. Nucleic acids research 47:D506-D515.

Devulder, G., Perriere, G., Baty, F., and Flandrois, J.P. 2003. BIBI, a bioinformatics bacterial identification tool. J Clin Microbiol 41:1785-1787.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., and Mistry, J. 2014. Pfam: the protein families database. Nucleic acids research 42:D222-D230.

Haft, D.H., Selengut, J.D., and White, O. 2003. The TIGRFAMs database of protein families. Nucleic acids research 31:371-373.

Hashem, A., Tabassum, B., and Fathi Abd Allah, E. 2019. Bacillus subtilis: A plant-growth promoting rhizobacterium that also impacts biotic stress. Saudi J Biol Sci 26:1291-1297.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics 11:119.

Kanehisa, M., and Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research 28:27-30.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. 2009. Circos: an information aesthetic for comparative genomics. Genome research 19:1639-1645.

Markowitz, V.M., Chen, I.-M.A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., and Williams, P. 2012. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic acids research 40:D115-D122.

Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., and Thomas, P.D. 2019. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v. 14.0). Nature protocols 14:703-721.

Mignard, S., and Flandrois, J.P. 2006. 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. J Microbiol Methods 67:574-581.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research 25:1043-1055.

Pati, A., Ivanova, N.N., Mikhailova, N., Ovchinnikova, G., Hooper, S.D., Lykidis, A., and Kyrpides, N.C. 2010. GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. Nature methods 7:455-457.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., and Glockner, F.O. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35:7188-7196.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research 35:D61-D65.

Radhakrishnan, R., Hashem, A., and Abd Allah, E.F. 2017. Bacillus: A Biological Tool for Crop Improvement through Bio-Molecular Changes in Adverse Environments. Front Physiol 8:667.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., and Nikolskaya, A.N. 2003. The COG database: an updated version includes eukaryotes. BMC bioinformatics 4:1-14.

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome research 13:2129-2141.

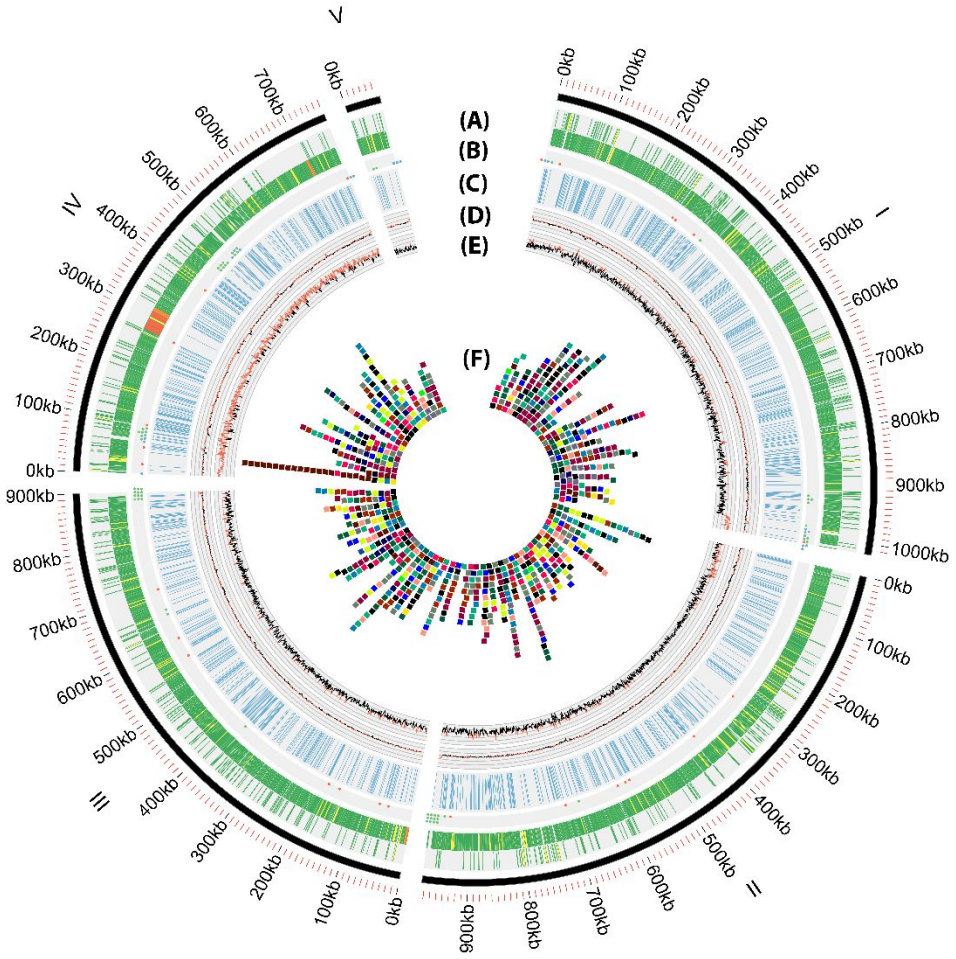
Wilson, K.J.C.p.i.m.b. 2001. Preparation of genomic DNA from bacteria 56:2.4. 1-2.4. 5.

Xia, Y., Greissworth, E., Mucci, C., Williams, M.A., and De Bolt, S. 2013. Characterization of culturable bacterial endophytes of switchgrass (P anicum virgatum L.) and their capacity to influence plant growth. Gcb Bioenergy 5:674-682.

Xia, Y., DeBolt, S., Ma, Q., McDermaid, A., Wang, C., Shapiro, N., Woyke, T., and Kyrpides, N.C. 2019. Improved Draft Genome Sequence of Bacillus sp. Strain YF23, Which Has Plant Growth-Promoting Activity. Microbiol Resour Announc 8.

Figure legend

Figure 1. Circular representation of the *Bacillus* sp. RRD69 genome using Circos. The circles, from outside to inside, denote protein-coding genes colored by size (A), RNA genes (B), transmembrane helix regions (C), GC content along with a 1-kb window, with red lines indicating the regions above 41.4% genome average and black lines indicating the regions below the genome average (D), GC skew, with red lines indicating a skew greater than zero and black lines indicating a skew less than zero (E), and genes annotated into distinct PANTHER protein classes (F). The repository for the storage of scripts used to construct the figure can be found at https://github.com/nbo245/rrd69.



Protein Coding Genes (Ring A):

	# Seqs
0bp < CDS < 2Kb	3617
2Kb < CDS > 5Kb	135
5Kb < CDS	6

RNA Genes (Ring B):

	# Seqs
rRNA genes	26
tRNA genes	73
Other RNA genes	27

RNA Genes (Ring F):

Category Name	Protein Class	# Seqs
Amino Acid Transporter	PC00046	27
ABC Transporter	PC00003	105
Cysteine Protease	PC00081	5
Dehydrogenase	PC00092	75
DNA Binding Protein	PC00009	13
Esterase	PC00097	3
Hydrolase	PC00121	49
Ligase	PC00142	51
Lyase	PC00144	39
Metalloprotease	PC00153	42
Oxidoreductase	PC00176	68
Phospholipase	PC00186	3
Primary Active Transp.	PC00068	26
Ribosomal Protein	PC00202	45
Sec. Carrier Transp.	PC00258	26
Serine Protease	PC00203	37
Transferase	PC00220	58
Transporter	PC00227	80
Winged Helix/Forkhead	PC00246	73

Supporting Material

Genome Sequence Resource of *Bacillus* sp. RRD69, a Beneficial Bacterial Endophyte isolated from Switchgrass Plants

Zhenzhen Zhao[1], Norbert Bokros[2], Seth DeBolt[2], Piao Yang[1], Ye Xia[1,*]

Table S1. Summary of the comparison of genome information between *Bacillus* sp. YF23 and *Bacillus* sp. RRD69.

	***Bacillus* sp. RRD69**	***Bacillus* sp. YF23**
Genome size	3.72 Mbp	5.82 Mbp
Protein coding genes	3758 (96.76%)	5740 (96.75%)
RNA genes	126 (3.24%)	193 (3.25%)
rRNA genes	26 (0.67%)	44 (0.74%)
tRNA genes	73 (1.88%)	116 (1.96%)
Other RNA genes	27 (0.7%)	33 (0.56%)
Biosynthetic Gene Clusters	8	14
Genes in Biosynthetic Clusters	244 (6.28%)	411 (6.93%)
Protein coding genes coding signal peptides	182 (4.69%)	268 (4.52%)
Protein coding genes coding transmembrane proteins	1069 (27.52%)	1713 (28.87%)
COG Categories		
Amino acid transport and metabolism	309 (10.24%)	401 (9.81%)
Transcription	277 (9.18%)	377 (9.22%)
Defense mechanisms	78 (2.59%)	188 (2.89%)

Author contributions:

Conceptualization, Z.Z., S.D., and Y.X.; software, N.B.; writing—original draft preparation, Z.Z.; writing—review and editing, Z.Z., N.B., S.D., P.Y. and Y.X.; project administration, S.D. and Y.X.; and funding acquisition, S.D. and Y.X. All authors have read and agreed to the published version of the manuscript.