Draft Genome Sequence of *Pseudozyma brasiliensis* sp. nov. Strain GHG001, a High Producer of Endo-1,4-Xylanase Isolated from an Insect Pest of Sugarcane

Juliana Velasco de Castro Oliveira,[a] Renato Augusto Corrêa dos Santos,[a] Thuanny A. Borges,[a] Diego Mauricio Riaño-Pachón,[a] Gustavo Henrique Goldman[a,b]

Laboratório Nacional de Ciência e Tecnologia do Bioetanol (CTBE), Centro Nacional de Pesquisa em Energia e Materiais (CNPEM), Campinas, São Paulo, Brazil[a]; Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, São Paulo, Brazil[b]

J.V.D.C.O. and R.A.C.D.S. contributed equally to this work.

Here, we present the nuclear and mitochondrial genome sequences of *Pseudozyma brasiliensis* sp. nov. strain GHG001. *P. brasiliensis* sp. nov. is the closest relative of *Pseudozyma vetiver*. *P. brasiliensis* sp. nov. is capable of growing on xylose or xylan as a sole carbon source and has great biotechnological potential.

Address correspondence to Diego Mauricio Riaño-Pachón, diego.riano@bioetanol.org.br, or Gustavo Henrique Goldman, ggoldman@usp.br.

*P*seudozyma brasiliensis* sp. nov. strain GHG001 is a yeast-like species that belongs to the order *Ustilaginales*. This strain was isolated from the intestinal tract of a *Chrysomelidae* larva associated with sugarcane roots in plantations in Ribeirão Preto, São Paulo, Brazil, following an enrichment protocol for microorganisms that use xylose as a sole carbon source. Based on the phylogenetic analysis of the ribosomal operon, we suggest that GHG001 represents a novel species that we named *P. brasiliensis* sp. nov.; its closest relative is *Pseudozyma vetiver* (1). GHG001 can grow well in xylose or xylan as its sole carbon source, where it produces high levels of endo-1,4-xylanase from the glycoside hydrolase (GH) family GH11 (2), the members of which show higher specific activity than other eukaryotic xylanases. Xylanases are essential for breaking down hemicellulose of plant cell walls, and they are routinely added to enzyme cocktails for the saccharification of pretreated biomass and second-generation ethanol production. Xylanases have further commercial applications, such as in bread making, the manufacture of food, drinks, and textiles, bleaching of cellulose pulp, and xylitol production (3).

Here, we present the genome sequence of *P. brasiliensis* sp. nov. strain GHG001. This genome was sequenced on the Illumina HiSeq2000 system, generating 73,703,379 paired-end reads of 100 bp (insert size, 250 bp). The reads were preprocessed with the Fastx-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). The genome size was estimated to be 22.09 Mbp based on *k*-mer count statistics (4), with an estimated coverage of 585×. The reads were randomly subsampled to a genome coverage of approximately 100×, and this subset was assembled using VelvetOptimiser and Velvet (5, 6). The remaining reads were used to extend the contigs and perform scaffolding using SSPACE Basic (7). The resulting assembly has 45 scaffolds, with a total length of 17,323,620 bp and an N_{50} of 720,612 bp. The average G+C content of the genome is 56.3%, which is similar to those of *Pseudozyma hubeiensis* SY62 (8)

and *Pseudozyma antarctica* T-34 (9). We evaluated the completeness of the gene space using CEGMA (10), which revealed that the current assembly is 97.98% complete. The scaffolds were masked for repeats using RepeatMasker, and gene prediction was carried out with GeneMark (11), Augustus (12), and STAP (http://korflab.ucdavis.edu/software.html), using MAKER (13). Gene finders were trained with the CEGMA-produced gene models. A total of 5,768 protein-encoding genes were identified, which is similar to the gene content of other *Pseudozyma* spp. A search against the NCBI nr database revealed 2,361 protein-encoding genes with strong sequence similarity hits to proteins in that database, providing a preliminary landscape of the genomic and metabolic capabilities of *P. brasiliensis*. Ribosomal genes were identified with RNAmmer (14), and the rRNA operon repeats (small subunit [SSU], internal transcribed spacer 1 [ITS1], 5.8 S, ITS2, and long subunit [LSU]) were collapsed into a single scaffold (PSEUBRA_SCAF27). One hundred nineteen tRNA genes were identified with tRNAscan-SE version 1.3.1 (15). The scaffold PSEUBRA_SCAF26 contains the mitochondrial genome.

Nucleotide sequence accession numbers. This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession no. AWXO00000000. The version described in this paper is version AWXO01000000.

ACKNOWLEDGMENTS

REFERENCES

1. **Chamnanpa T, Limtong P, Srisuk N, Limtong S.** 2013. *Pseudozyma vetiver* sp. nov., a novel anamorphic ustilaginomycetous yeast species iso-

lated from the phylloplane in Thailand. Antonie Van Leeuwenhoek **104:** 637–644.

2. **Paës G, Berrin JG, Beaugrand J.** 2012. GH11 xylanases: structure/function/properties relationships and applications. Biotechnol. Adv. **30:** 564–592.

3. **Polizeli ML, Rizzatti AC, Monti R, Terenzi HF, Jorge JA, Amorim DS.** 2005. Xylanases from fungi: properties and industrial applications. Appl. Microbiol. Biotechnol. **67:**577–591.

4. **Marçais G, Kingsford C.** 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics **27:**764–770.

5. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. **18:**821–829.

6. **Zerbino DR.** 2010. Using the Velvet *de novo* assembler for short-read sequencing technologies. Curr. Protoc. Bioinformatics **31:**11.5.1–11.5.12. doi:10.1002/0471250953.bi1105s31.

7. **Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W.** 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics **27:** 578–579.

8. **Konishi M, Hatada Y, Horiuchi J.** 2013. Draft genome sequence of the basidiomycetous yeast-like fungus *Pseudozyma hubeiensis* SY62, which produces an abundant amount of the biosurfactant mannosylerythritol lipids. Genome Announc. **1**(4)**:**e00409-13. doi:10.1128/genomeA.00409-13.

9. **Morita T, Koike H, Koyama Y, Hagiwara H, Ito E, Fukuoka T, Imura T, Machida M, Kitamoto D.** 2013. Genome sequence of the basidiomycetous yeast *Pseudozyma antarctica* T-34, a producer of the glycolipid biosurfactants mannosylerythritol lipids. Genome Announc. **1**(2)**:**e00064-13. doi:10.1128/genomeA.00064-13.

10. **Parra G, Bradnam K, Korf I.** 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23:**1061–1067.

11. **Borodovsky M, Lomsadze A, Ivanov N, Mills R.** 2003. Eukaryotic gene prediction using GeneMark.hmm. Curr. Protoc. Bioinformatics **35:** 4.6.1–4.6.10. doi:10.1002/0471250953.bi0406s35.

12. **Stanke M, Schöffmann O, Morgenstern B, Waack S.** 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics **7:**62. doi:10.1186/1471-2105-7-62.

13. **Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M.** 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. **18:**188–196.

14. **Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW.** 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. **35:**3100–3108.

15. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:** 955–964.