Draft Genome Sequence of *Saccharomyces cerevisiae* Barra Grande (BG-1), a Brazilian Industrial Bioethanol-Producing Strain

Natalia Coutouné, Aline Tieppo Nogueira Mulato, Diego Mauricio Riaño-Pachón,*
Juliana Velasco de Castro Oliveira

Brazil Laboratório Nacional de Ciência e Tecnologia do Bioetanol (CTBE), Centro Nacional de Pesquisa em Energia e Materiais (CNPEM), São Paulo, Brazil

ABSTRACT Here, we present the draft genome sequence of *Saccharomyces cerevisiae* BG-1, a Brazilian industrial strain widely used for bioethanol production from sugarcane. The 11.7-Mb genome sequence consists of 216 scaffolds and harbors 5,607 predicted protein-coding genes.

The Brazilian process of bioethanol production is characterized by industrial fermentations in large tanks (0.5 to 3 million liters) and high yeast cell densities (10 to 15% wt/vol). A complete fermentation cycle is performed within a very short period (6 to 12 h), achieving ethanol concentrations of 7 to 11% (vol/vol). In addition, after a treatment with sulfuric acid (pH 1.8 to 2.5) the yeast cells are returned to tanks to start a new fermentation batch (1). Accordingly, Brazilian industrial strains must be able to deal with several stressing agents, and the genomes of the strains currently published (2, 3) show significant adaptive genetic variability (4).

Saccharomyces cerevisiae strain Barra Grande (BG-1) is a heterothallic strain, isolated from a Brazilian distillery in 1989-1990 (4). BG-1 has tolerance to high temperature, does not flocculate, ferments with low foam production, and is adapted to the Brazilian fermentation process. Its genome was sequenced on the Illumina HiSeq 2500 system, using a paired-end library with an insert size of 300 bp, and three mate-pair libraries with insert sizes of 3 to 4 kb, 5 to 7 kb, and 8 to 11 kb, producing 4,187,211, 3,605,199, and 4,578,254 sequences, respectively. The reads were processed with the FastQC (5), NextClip (only the mate-pair reads) (6), and Trimmomatic (7) tools. Genome assembly was carried out with SPAdes version 3.7.1 (8), Pilon version 1.16 (9), and Redundans (10). Assessment of ploidy level was carried out with ploidyNGS (11). BG-1 is a diploid organism. The haploid nuclear genome is assembled in 215 scaffolds, and additional scaffold contains the full 2-μm plasmid sequence. The nuclear genome has a total length of 11,691,159 bp and an N_{50} of 337,041 bp, with a coverage of 234×. The 2-μm plasmid has a length of 5,623 bp, with a coverage of 12,679×. The average G+C content of the nuclear genome is 38.25%, which is similar to *S. cerevisiae* S288C (~38%) (12). Gene prediction and annotation was carried out with the Funannotate pipeline (13). A total of 5,607 putative protein-coding genes were identified, and 89.38% have clear homologs in *S. cerevisiae* S288C (above 98% similarity). We found 280 genes coding for tRNAs, and the rRNA operon is collapsed into a single scaffold. The BG-1 draft genome sequence provides a source of information for elucidating the genetic mechanisms underlying the fermentation process and stress tolerance.

Accession number(s). This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession number MSHP00000000. The version described in this project is the first version, MSHP01000000.

ACKNOWLEDGMENTS

REFERENCES

1. Amorim HV, Lopes ML, de Castro Oliveira JV, Buckeridge MS, Goldman GH. 2011. Scientific challenges of bioethanol production in Brazil. Appl Microbiol Biotechnol 91:1267–1275. https://doi.org/10.1007/s00253-011 -3437-6.

2. Argueso JL, Carazzolle MF, Mieczkowski PA, Duarte FM, Netto OV, Missawa SK, Galzerani F, Costa GGL, Vidal RO, Noronha MF, Dominska M, Andrietta MGS, Andrietta SR, Cunha AF, Gomes LH, Tavares FCA, Alcarde AR, Dietrich FS, McCusker JH, Petes TD, Pereira GAG. 2009. Genome structure of a Saccharomyces cerevisiae strain widely used in bioethanol production. Genome Res 19:2258–2270. https://doi.org/10 .1101/gr.091777.109.

3. Babrzadeh F, Jalili R, Wang C, Shokralla S, Pierce S, Robinson-Mosher A, Nyren P, Shafer RW, Basso LC, de Amorim HV, de Oliveira AJ, Davis RW, Ronaghi M, Gharizadeh B, Stambuk BU. 2012. Whole-genome sequencing of the efficient industrial fuel-ethanol fermentative Saccharomyces cerevisiae strain CAT-1. Mol Genet Genomics 287:485–494. https://doi .org/10.1007/s00438-012-0695-7.

4. Stambuk BU, Dunn B, Alves SL, Duval EH, Sherlock G. 2009. Industrial fuel ethanol yeasts contain adaptive copy number changes in genes involved in vitamin B_1 and B6 biosynthesis. Genome Res 19:2271–2278. https://doi.org/10.1101/gr.094276.109.

5. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/ fastqc.

6. Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M. 2014. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. Bioinformatics 30:566–568. https://doi.org/10.1093/bioinformatics/ btt702.

7. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10 .1093/bioinformatics/btu170.

8. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.

9. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963. https://doi.org/10.1371/journal .pone.0112963.

10. Pryszcz LP, Gabaldón T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res 44:e113. https://doi .org/10.1093/nar/gkw294.

11. dos Santos RAC, Goldman GH, Riano-Pachon DM. 2016. ploidyNGS: visually exploring ploidy with next generation sequencing data. bioRxiv. https://doi.org/10.1101/086488.

12. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. 1996. Life with 6000 genes. Science 274:546, 563–567.

13. Palmer JM. 2016. Funannotate: pipeline for genome annotation. http:// www.github.com/nextgenusfs/funannotate.