AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements

De Novo Assembly of *Candida sojae* and *Candida boidinii* Genomes, Unexplored Xylose-Consuming Yeasts with Potential for Renewable Biochemical Production

Guilherme Borelli,[a] Juliana José,[a] Paulo José Pereira Lima Teixeira,[b] Leandro Vieira dos Santos,[a,c] Gonçalo Amarante Guimarães Pereira[a,c]

Departamento de Genética, Evolução e Bioagentes, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, SP, Brazil[a]; Dangl Lab, Department of Biology, University of North Carolina, Chapel Hill, North Carolina, USA[b]; GranBio/Biocelere Agroindustrial, Campinas, SP, Brazil[c]

***Candida boidinii* and *Candida sojae* yeasts were isolated from energy cane bagasse and plague-insects. Both have fast xylose uptake rate and produce great amounts of xylitol, which are interesting features for food and 2G ethanol industries. Because they lack published genomes, we have sequenced and assembled them, offering new possibilities for gene prospection.**

Received 10 November 2015 **Accepted** 18 November 2015 **Published** 14 January 2016

Citation Borelli G, José J, Teixeira PJPL, dos Santos LV, Pereira GAG. 2016. *De novo* assembly of *Candida sojae* and *Candida boidinii* genomes, unexplored xylose-consuming yeasts with potential for renewable biochemical production. Genome Announc 4(1):e01551-15. doi:10.1128/genomeA.01551-15.

Address correspondence to Gonçalo Amarante Guimarães Pereira, goncalo@unicamp.br.

Candida boidinii and *Candida sojae* are both xylose-consuming yeasts from the *Saccharomycetales* order of *Ascomycota*. *C. boidinii* was also found as a methylotrophic yeast (1), while *C. sojae* lacks further studies of its metabolic pathways. *C. boidinii* was first identified from a wash of tree barks (2), and recently, we have isolated it from sugarcane bagasse. *C. sojae* was first isolated from a liquid fraction of water-soluble substances of defatted soybean flakes, in Japan (3), and we have isolated it from *Diatraea saccharalis* (*Lepidoptera*), a plague-insect in an energy-cane cultivar.

From the yeast strains we have isolated, these *Candida* isolates were shown as the best strains in xylose consumption (*C. boidinii* and *C. sojae*) and in xylitol production (*C. sojae*; unpublished results), a sugar-alcohol interesting for food industry as a sweetener (4). In addition, genes related to xylose uptake and in xylitol production are target of interest in the 2G ethanol industry, since xylose is a sugar present in abundance in lignocellulosic substrates and xylitol is an intermediate of ethanol production from xylose (5). There was no exploration of biotechnological potential of these yeasts and both had no published genome. Thus, we have extracted their genomes and sequenced, assembled, and analyzed them, searching for genes from xylose metabolism (5).

Samples were sent to the University of North Carolina at the High-Throughput Sequencing Facility (HTSF) for genome sequencing using an Illumina HiSeq2000, which produced a library with insert sizes around 400 nucleotides (nt) and near to 116 million paired reads for *C. boidinii* and 109 million for *C. sojae*, with 100 nt each. Considering their reduced genome sizes, we conducted the assembly using the SPAdes version 3.5 pipeline (6).

Initial read coverage (900×) was reduced by randomly sorting reads into one-third. The default SPAdes pipeline worked very well for *C. boidinii*, but a first assembly round for *C. sojae* suggested that we were dealing with a polypoid genome. We proceeded with the *C. sojae* assembly using the dipSPAdes module, which resulted in a high-quality assembly for a consensus haploid genome. The *C. boidinii* final assembly comprised approximately 19 Mb organized into 428 scaffolds, with an N_{50} of 49 sequences and a GC content of 30.7%. The *C. sojae* final assembly comprised approximately 12 Mb organized into 511 scaffolds, with an N_{50} of 65 sequences, and a GC content of 32.4%. Final read coverage for the assembled genome was 144× for *C. boidinii* and 48× for *C. sojae*.

Gene predictions were performed using Augustus version 3.2.1 (7) with the previously available training set for *Candida tropicalis*. For *C. boidinii*, 5,978 genes with an average sequence length of 548 amino acids were identified, while for *C. sojae* these numbers were 52,31 genes with an average length of 466 amino acids.

Nucleotide sequence accession numbers. The *Candida sojae* whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession number LMTL00000000. The version described in this paper is the first version, LMTL01000000.

The *Candida boidinii* whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession number LMZO00000000. The version described in this paper is the first version, LMZO01000000.

ACKNOWLEDGMENTS

We acknowledge Piotr Mieczkowski at the University of North Carolina for his sequencing services.

This work was financed by CAPES Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil and CNPq National Council for Scientific and Technological Development within the Ministry of Science and Technology of Brazil, and performed at the State University of Campinas (Unicamp).

FUNDING INFORMATION

MCTI | Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) provided funding to Guilherme Borelli under grant number 131745/2015-8. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) provided funding to Guilherme Borelli. The São Paulo Research Foundation (FAPESP) provided funding to Juliana José under grant number 2014/90638-0.

REFERENCES

1. **Vongsuvanlert V, Tani Y**. 1988. Purification and characterization of xylose isomerase of a methanol yeast, *Candida boidinii*, which is involved in sorbitol production from glucose. Agric Biol Chem **52:**1817–1824. http://dx.doi.org/10.1271/bbb1961.52.1817.
2. **Ramírez C**. 1953. Estudio sobre nuevas especies de levaduras aisladas de diferentes sustratos. Microbiol Española **6:**249–253.
3. **Nakase T, Suzuki M, Takashima M, Miyakawa Y, Kagaya K, Fukazawa Y, Komagata K**. 1994. *Candida sojae*, a new species of yeast isolated from an extraction process of water-soluble substances of defatted soybean flakes. J Gen Appl Microbiol **40:**161–169. http://dx.doi.org/10.2323/jgam.40.161.
4. **Grembecka M**. 2015. Sugar alcohols—their role in the modern world of sweeteners: a review. Eur Food Res Technol **241:**1–14. http://dx.doi.org/10.1007/s00217-015-2437-7.
5. **Gírio FM, Fonseca C, Carvalheiro F, Duarte LC, Marques S, Bogel-Łukasik R**. 2010. Hemicelluloses for fuel ethanol: a review. Bioresour Technol **101:**4775–4800. http://dx.doi.org/10.1016/j.biortech.2010.01.088.
6. **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. http://dx.doi.org/10.1089/cmb.2012.0021.
7. **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. http://dx.doi.org/10.1186/1471-2105-7-62.