Draft Genome Sequence of the Yeast *Saccharomyces cerevisiae* GUJ105 From Gujarat, India

Chandan Badapanda, Rajesh Detroja, Ankita Rathore

Xcelris Labs Limited, Ahmedabad, Gujarat, India

Here, we report the draft genome sequence of *Saccharomyces cerevisiae* strain GUJ105, isolated clinically. The size of the genome is approximately 11.5 Mb and contains 5,447 protein-coding genes.

Address correspondence to Chandan Badapanda, chandan.badapanda@xcelrislabs.com.

*S*accharomyces cerevisiae is the first eukaryote for which a genome was sequenced (1). *S. cerevisiae* has been studied extensively for decades as a model organism for the following reasons: (i) it is used as a host for recombinant protein production (2), (ii) it is tolerant of a wide range of physiological stresses, such as low pH, high ethanol, and osmotic pressure (3, 4), (iii) it is used in the baking industry and in fermentations of food-grade alcohol production (5, 6), (iv) it is used as a model for biofuel production (7), (v) it is used as a model system for evolutionary biology (8), and (vi) it is used as a model for population and quantitative genetics. *S. cerevisiae* is found in multiple environments, one of which is the human body as an opportunistic pathogen (9, 10).

Here, we report the draft genome sequence of *S. cerevisiae*, derived from a clinical sample from Gujarat, India. The genomic DNA was sequenced using an Illumina HiSeq2500 platform. A total of 4.72 million paired-end reads having 2×150 bp chemistry with an estimated coverage of $62\times$ were produced. The adapter sequences were trimmed from the reads before assembly using Trimmomatic (v0.30) (11). *De novo* assembly was carried out with the assembly tool Velvet (12) in conjunction with the VelvetOptimiser (k-mer 67) (http://bioinformatics.net.au/software.velvetoptimiser.shtml).

The resulting assembly has 201 scaffolds with a total length of 11,563,971 bases (11.5 MB) (the largest scaffold having a length of 552,249 bp), an N_{50} of 276,334, and a G+C content of 37.95% and 100% of completeness, as estimated by mapping the orthologous genes (KOG databases) using CEGMA (13). Using *S. cerevisae*, the gene finder Augustus (14) identified 5,447 protein-coding genes. An automatic annotation using the BLASTp algorithm (15) revealed 5,447 predicted proteins with significant similarity (*E* value cutoff = 10^{-5}) to sequences deposited in the nonredundant (nr) protein database from NCBI. Using tRNAscan-SE version 1.3 software (16), we found 266 tRNA genes scattered across the scaffolds.

In our future study, we will carry out a comparative genomics study of strain GUJ105 with the existing clinical strains for better understanding of the basic mechanisms of adaptation to its natural habitat, as well as of its pathogenicity and virulence.

Accession number(s). Data related to the draft genome project of *Saccharomyces cerevisiae* GUJ105 have been deposited at GenBank under the accession number MCNH00000000.

REFERENCES

1. **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**:546–567. http://dx.doi.org/10.1126/science.274.5287.546.
2. **Mattanovich D, Branduardi P, Dato L, Gasser B, Sauer M, Porro D.** 2012. Recombinant protein production in yeasts. Methods Mol Biol **824**:329–358. http://dx.doi.org/10.1007/978-1-61779-433-9_17.
3. **Talemi SR, Tiger CF, Andersson M, Babazadeh R, Welkenhuysen N, Klipp E, Hohmann S, Schaber J.** 2016. Systems level analysis of the yeast osmo-stat. Sci Rep **6**:30950. http://dx.doi.org/10.1038/srep30950.
4. **Ostergaard S, Olsson L, Nielsen J.** 2000. Metabolic engineering of *Saccharomyces cerevisiae*. Microbiol Mol Biol Rev **64**:34–50. http://dx.doi.org/10.1128/MMBR.64.1.34-50.2000.
5. **Kuyper M, Hartog MM, Toirkens MJ, Almering MJ, Winkler AA, van Dijken JP, Pronk JT.** 2005. Metabolic engineering of a xylose-isomerase-expressing *Saccharomyces cerevisiae* strain for rapid anaerobic xylose fermentation. FEMS Yeast Res **5**:399–409. http://dx.doi.org/10.1016/j.femsyr.2004.09.010.
6. **Hahn-Hägerdal B, Karhumaa K, Fonseca C, Spencer-Martins I, Gorwa-Grauslund MF.** 2007. Towards industrial pentose-fermenting yeast strains. Appl Microbiol Biotechnol **74**:937–953. http://dx.doi.org/10.1007/s00253-006-0827-2.
7. **Olson DG, McBride JE, Shaw AJ, Lynd LR.** 2012. Recent progress in consolidated bioprocessing. Curr Opin Biotechnol **23**:396–405. http://dx.doi.org/10.1016/j.copbio.2011.11.026.
8. **Wolters JF, Chiu K, Fiumera HL.** 2015. Population structure of mitochondrial genomes in *Saccharomyces cerevisiae*. BMC Genomics **16**:451. http://dx.doi.org/10.1186/s12864-015-1664-4.
9. **Strope PK, Skelly DA, Kozmin SG, Mahadevan G, Stone EA, Magwene PM, Dietrich FS, McCusker JH.** 2015. The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. Genome Res **25**:762–774.
10. **Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, Cao Z, Gu Z, Bruno D, Miranda M, Nguyen M, Wilhelmy J, Komp C, Tamse R, Wang X, Jia P, Luedi P, Oefner PJ, David L, Dietrich FS, Li Y, Davis RW, Steinmetz LM.** 2007. Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. Proc Natl Acad Sci U S A **104**:12825–12830. http://dx.doi.org/10.1073/pnas.0701291104.
11. **Bolger AM, Lohse M, Usadel B.** 2014. Trimmomatic: a flexible trimmer

for Illumina sequence data. Bioinformatics **30:**2114–2120. http://dx.doi.org/10.1093/bioinformatics/btu170.

12. **Zerbino DR, Birney E**. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18:**821–829. http://dx.doi.org/10.1101/gr.074492.107.

13. **Parra G, Bradnam K, Korf I**. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23:**1061–1067. http://dx.doi.org/10.1093/bioinformatics/btm071.

14. **Hoff KJ, Stanke M**. 2013. WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Res **41:** W123–W128. http://dx.doi.org/10.1093/nar/gkt418.

15. **Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ**. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:**3389–3402. http://dx.doi.org/10.1093/nar/25.17.3389.

16. **Lowe TM, Eddy SR**. 1997. Eddy SR.tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25:**955–964. http://dx.doi.org/10.1093/nar/25.5.0955.