Genome Sequences of Industrially Relevant *Saccharomyces cerevisiae* Strain M3707, Isolated from a Sample of Distillers Yeast and Four Haploid Derivatives

Steven D. Brown,[a,b] Dawn M. Klingeman,[a,b] Courtney M. Johnson,[a,b] Alicia Clum,[c] Andrea Aerts,[c] Asaf Salamov,[c] Aditi Sharma,[c] Matthew Zane,[c] Kerrie Barry,[c] Igor V. Grigoriev,[c] Brian H. Davison,[a,b] Lee R. Lynd,[b,d,e] Paul Gilna,[a,b] Heidi Hau,[b,d] David A. Hogsett,[b,d] Allan C. Froehlich[b,d]

Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA[a]; BioEnergy Science Center, Oak Ridge, Tennessee, USA[b]; US Department of Energy Joint Genome Institute, Walnut Creek, California, USA[c]; Mascoma Corporation, Lebanon, New Hampshire, USA[d]; Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire, USA[e]

Saccharomyces cerevisiae strain M3707 was isolated from a sample of commercial distillers yeast, and its genome sequence together with the genome sequences for the four derived haploid strains M3836, M3837, M3838, and M3839 has been determined. Yeasts have potential for consolidated bioprocessing (CBP) for biofuel production, and access to these genome sequences will facilitate their development.

Address correspondence to Steven D. Brown, brownsd@ornl.gov, or Allan C. Froehlich, afroehlich@mascoma.com.

*S*accharomyces cerevisiae* is the first eukaryote for which a genome was completely sequenced (1). It has been studied intensely for decades as a model organism, is generally recognized as safe (GRAS), is used as a host for recombinant protein production (2), and is tolerant of a wide range of physiological stresses, such as low pH, high ethanol, and high osmotic stress (3, 4, 5). *S. cerevisiae* is used in the baking industry and fermentations for food-grade alcohol production, its pentose utilization capabilities have been developed (3, 6), and yeasts have potential for consolidated bioprocessing (CBP) for biofuel production (7, 8). The 12-Mb *S. cerevisiae* S288c genome encodes approximately 6,000 open reading frames, and the organism can maintain stable haploid or diploid states (9). To better understand *S. cerevisiae* and its potential for CBP development, we generated the genome sequences for the diploid strain M3707 isolated from a sample of distillers yeast and four haploid derivatives named M3836, M3837, M3838, and M3839.

The genomes of *S. cerevisiae* strain M3707 and strains M3836, M3837, M3838, and M3839 were sequenced using the Illumina (10) HiSeq DNA sequencing platform according to the manufacturer's instructions. Long mate-pair and paired-end (PE) libraries were used to sequence genomic DNA for each strain, with insert sizes of ~4 kb and ~270 bp, respectively. In addition, cDNA derived from RNA that represented 17 different growth conditions was sequenced from a ~270-bp insert size PE library to aid gene model predictions. Filtered reads were assembled with AllPaths-LG (release version R41554) (11). Each strain has an estimated genome size of approximately 11.5 Mb, with a mean genome sequence coverage of ~149× across all strains. The genome sequences assembled into between 84 and 120 contigs and 43 scaffolds for each strain. Approximately 6,000 genes were predicted in each of the five strains using the Joint Genome Institute (JGI) annotation pipeline (12) that contained a set of gene predictors (13, 14, 15).

The genome sequences of the new strains and strain 288c are very similar, with 99.05 to 99.1% similarity between the genomes as measured by average nucleotide identity (ANI) (16). Near-perfect orthologs (≥98% coverage and ≥98% percent identity) are found for 90.4% of M3707 and 288c genes, and 97.4% of these are syntenic. Interestingly, over half of the predicted gene models are identical copies; the rest are affected by insertions and deletions. In comparison to 288c, M3707 lacks around 3.5% of sequence (~325 kb) containing 46 genes, while it carries an additional 1.7% (~150 kb) of sequence containing 33 genes. Most of the unique genes encode proteins of unknown function. However, an interesting exception is a gene containing a melibiose domain, which is conserved among the five strains that were sequenced as part of this study. *S. cerevisiae* strain M3707 and strains M3836, M3837, M3838, and M3839 are industrially relevant, and their genome sequences will provide a useful foundation for future metabolic engineering, systems biology, and comparative genomics studies to develop CBP *S. cerevisiae* strains.

Nucleotide sequence accession numbers. The genome sequences and annotations are available via the JGI fungal genomics resource MycCosm (http://jgi.doe.gov/fungi [17]) and have been deposited in DDBJ/EMBL/GenBank under the accession no. AMQB00000000, AMQC00000000, AMQD00000000, AMQE00000000, and AMQF00000000 for strains M3707, M3836, M3837, M3838, and M3839, respectively.

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.

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.

3. **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.

4. **Sherman F.** 2002. Getting started with yeast. Methods Enzymol. **194:**3–21.

5. **Ostergaard S, Olsson L, Nielsen J.** 2000. Metabolic engineering of *Saccharomyces cerevisiae*. Microbiol. Mol. Biol. Rev. **64:**34–50.

6. **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.

7. **Ilmén M, den Haan R, Brevnova E, McBride J, Wiswall E, Froehlich A, Koivula A, Voutilainen SP, Siika-aho M, la Grange DC, Thorngren N, Ahlgren S, Mellon M, Deleault K, Rajgarhia V, van Zyl WH, Penttilä M.** 2011. High level secretion of cellobiohydrolases by *Saccharomyces cerevisiae*. Biotechnol. Biofuels **4:**30.

8. **Olson DG, McBride JE, Shaw AJ, Lynd LR.** 2012. Recent progress in consolidated bioprocessing. Curr. Opin. Biotechnol. **23:**396–405.

9. **Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED.** 2012. *Saccharomyces* Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. **40:**D700–D705.

10. **Bennett S.** 2004. Solexa Ltd. Pharmacogenomics **5:**433–438.

11. **Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB.** 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U. S. A. **108:** 1513–1518.

12. **Grigoriev IV, Martinez DA, Salamov AA.** 2006. Fungal genomic annotation, p 123–142. *In* Aurora DK, Berka RM, Singh G (ed), Applied mycology and biotechnology, vol 6, Bioinformatics. Elsevier Press, Philadelphia, PA.

13. **Salamov AA, Solovyev VV.** 2000. Ab initio gene finding in *Drosophila* genomic DNA. Genome Res. **10:**516–522.

14. **Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M.** 2008. Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. Genome Res. **18:**1979–1990.

15. **Birney E, Clamp M, Durbin R.** 2004. GeneWise and Genomewise. Genome Res. **14:**988–995.

16. **Richter M, Rosselló-Móra R.** 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U. S. A. **106:** 19126–19131.

17. **Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm RA, Otillar R, Poliakov A, Ratnere I, Riley R, Smirnova T, Rokhsar D, Dubchak I.** 2012. The genome portal of the Department of Energy Joint Genome Institute. Nucleic Acids Res. **40:**D26–D32.