AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

Chromosome-Level Genome Assembly of the Yeast *Candida verbasci*

Broňa Brejová,[a] Viktória Hodorová,[b] Hana Lichancová,[b] Eunika Peričková,[a] Veronika Anna Šoucová,[a] Matthias Sipiczki,[c] Tomáš Vinař,[d] Jozef Nosek[b]

[a]Department of Computer Science, Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Bratislava, Slovak Republic
[b]Department of Biochemistry, Faculty of Natural Sciences, Comenius University in Bratislava, Bratislava, Slovak Republic
[c]Department of Genetics and Applied Microbiology, University of Debrecen, Debrecen, Hungary
[d]Department of Applied Informatics, Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Bratislava, Slovak Republic

ABSTRACT *Candida verbasci* is an anamorphic ascomycetous yeast. We report the genome sequence of its type strain, 11-1055 (CBS 12699). The nuclear genome assembly consists of seven chromosome-sized contigs with a total size of 12.1 Mbp and has a relatively low G+C content (28.1%).

Candida verbasci is classified into the clade *Lodderomyces/Candida* of the family *Debaryomycetaceae* (subphylum *Saccharomycotina*), which includes important human pathogens (e.g., *Candida albicans*, *Candida parapsilosis*), along with the yeasts isolated from various insects (e.g., *Candida corydali*) (1–4). *C. verbasci* was originally identified in a microbial community associated with *Verbascum* flowers in Tbilisi, Georgia. As this yeast is not osmotolerant, it has been suggested that its presence in flowers resulted from insect transmission rather than natural propagation in sugar-rich nectar (1).

We analyzed the DNA of the type strain, 11-1055, using a combination of long- and short-read sequencing technologies. Total cellular DNA was isolated using a standard protocol (5) from an overnight culture grown in yeast extract-peptone-dextrose (YPD) medium (1% [wt/vol] yeast extract, 2% [wt/vol] peptone, 2% [wt/vol] glucose) at 28°C. In total, 1.9 Gbp (~157× coverage) was obtained in 197,400 long reads (mean, 9,744.8 nucleotides [nt]; longest read, 145,767 nt; N_{50}, 18,226 nt) using a MinION Mk1B device with an R9.4.1 flow cell and a rapid barcoding sequencing kit (SQK-RBK004; Oxford Nanopore Technologies). A paired-end (2 × 151-nt) TruSeq PCR-free DNA library was also sequenced using the NovaSeq 6000 platform at Macrogen (South Korea), yielding 82,821,710 short reads (12.5 Gbp; coverage, ~1,028×).

The genome sequence was assembled from Nanopore data using Flye v.2.8.3-b1695 (6). The assembly consisted of seven contigs and one scaffold. The only gap in the scaffold, immediately following the rDNA cluster, was filled manually with the corresponding 250-bp portion of an assembly created using miniasm v.0.3-r179 (7), with read overlaps found using Minimap v.2.13-r852-dirty (8), and polished using two iterations of Racon v.1.3.1 (9). The entire assembly was polished using three iterations of Pilon v.1.21 (10), with Illumina reads aligned using BWA-MEM v.0.7.17-r1188 (11), with a single rDNA repeat polished separately and used to replace copies in the assembly to avoid problems with ambiguous mapping of reads.

The final assembly contains one mitochondrial and seven nuclear contigs, with overall G+C contents of 28.7% and 28.1%, respectively. The nuclear contigs, terminated at both ends with telomeric arrays $(CAACAAACACTTGAGGTAAGGATG)_n$, correspond in length to the electrophoretic karyotype (Fig. 1A and B) and likely represent full-length chromosomes. To annotate the protein-coding genes, *C. albicans* and

Editor Jason E. Stajich, University of California, Riverside

Address correspondence to Broňa Brejová, brejova@dcs.fmph.uniba.sk, or Jozef Nosek, jozef.nosek@uniba.sk.

The authors declare no conflict of interest.

Received 10 January 2023
Accepted 6 February 2023
Published 22 February 2023

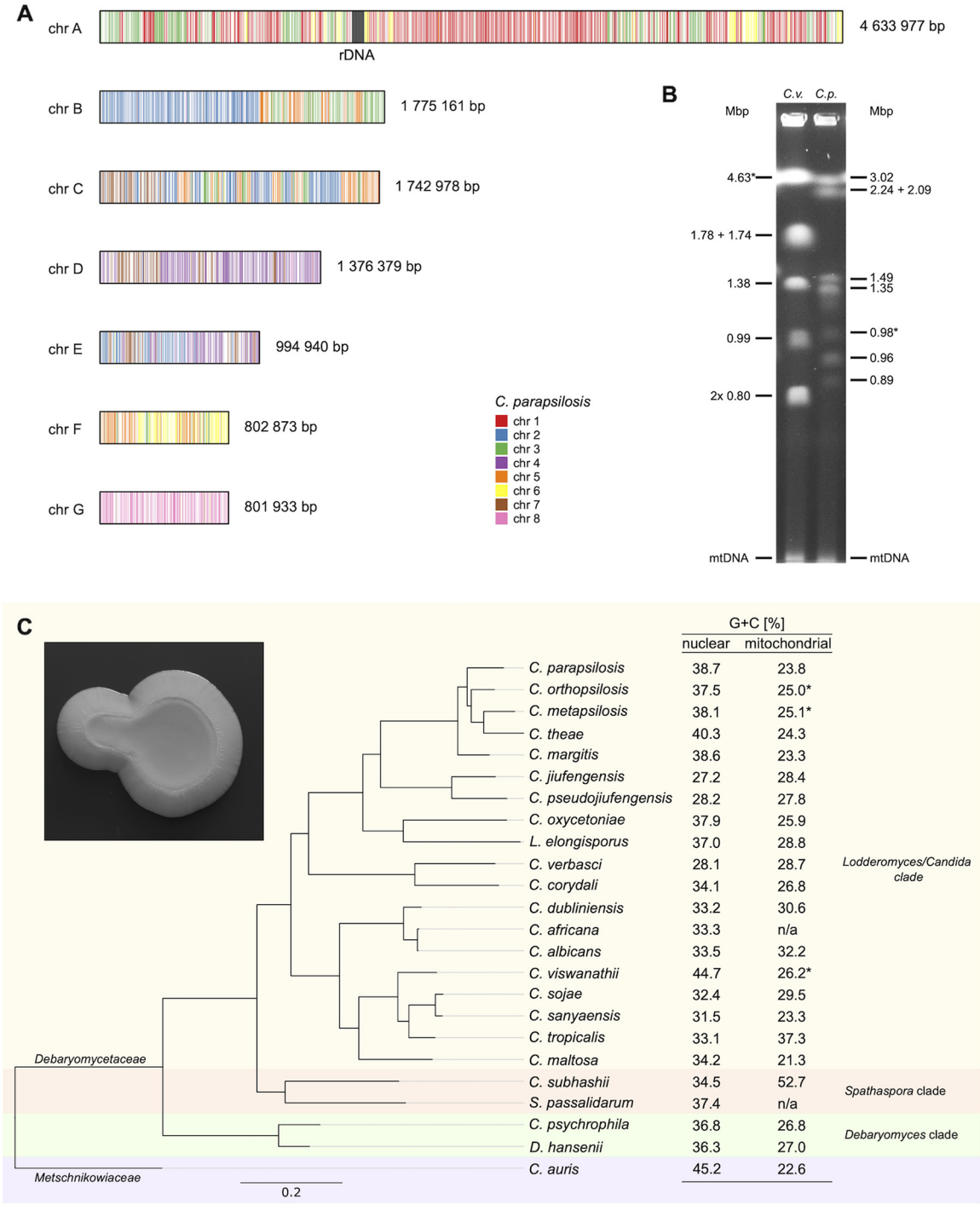

FIG 1 (A) Nuclear chromosomes of *C. verbasci*, colored based on their homology with *C. parapsilosis* strain CLIB214 (17). Local alignments between the two genomes were computed using LAST v.830 (lastal with option -E1e-10, followed by last-split) (18) and visualized using the ggplot2 v.3.3.6 library in R (19). (B) Electrophoretic karyotype of *C. verbasci* (*C.v.*), analyzed using a CHEF Mapper XA Chiller system (Bio-Rad), with a 120° angle between the electric fields, at the following settings: 120-s pulses for 20 h, followed by 240-s pulses for 28 h at 4 V/cm in a 0.8% agarose gel and 0.5× TBE buffer (45 mM Tris-borate, 1 mM EDTA) at 14°C. The chromosomes of *C. parapsilosis* strain CLIB214 (*C.p.*) were used as size markers (for more details, see reference 17). Asterisks indicate the chromosomes containing rDNA repeats. mtDNA, mitochondrial DNA. (C) Phylogenetic tree of the *Lodderomyces* clade species, constructed using the BUSCO_phylogenomics pipeline (20) from 2,120 single-copy orthologs present in at least 50% of the genome assemblies (options: –psc 50 –supermatrix) identified using BUSCO v.5.1.2 (21). The G+C contents of the nuclear and mitochondrial DNAs are shown. Asterisks indicate that the nuclear and mitochondrial DNA sequences were obtained from different strains of the species; n/a indicates that the sequence of mitochondrial DNA is not available. The inset illustrates a colony of *C. verbasci* grown for 7 days on a YPD plate.

C. corydali proteomes were downloaded from UniProt (12) and Shen et al. (13, 14), respectively, and aligned with the assembly using exonerate v.2.4.0 (15). Protein alignments were used as hints in Augustus v.3.2.3 (16), in the first iteration with parameters for *C. albicans* and in the second iteration with parameters trained on the predictions matching *Candida* proteins with at least 70% identity. In total, 5,313 protein-coding genes were annotated in the nuclear genome of *C. verbasci*.

The chromosome-level genome assembly of *C. verbasci* will be instrumental in further functional analyses, as well as in comparative studies of pathogenic and nonpathogenic species from the clade *Lodderomyces* (Fig. 1C).

Data availability. The genome assembly and Illumina and Nanopore reads have been deposited in the European Nucleotide Archive (ENA) and GenBank under accession numbers CANTUO000000000, ERR10286776, and ERR10286775, respectively. The version described in this paper is the first version, CANTUO010000000. The annotated mitochondrial DNA sequence was also deposited at GenBank (accession number OK589855). The results are also available in the genome browser at http://genome.compbio.fmph.uniba.sk/.

ACKNOWLEDGMENTS

This research was supported by grants from the Slovak Research and Development Agency (APVV 18-0239, 19-0068, and 20-0166) and the Slovak Grant Agency (VEGA 1/0463/20, 1/0538/22, and 1/0234/23). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

REFERENCES

1. Sipiczki M. 2013. Detection of yeast species also occurring in substrates associated with animals and identification of a novel dimorphic species in Verbascum flowers from Georgia. Antonie Van Leeuwenhoek 103:567–576. https://doi.org/10.1007/s10482-012-9841-9.
2. Nguyen NH, Suh SO, Blackwell M. 2007. Five novel Candida species in insect-associated yeast clades isolated from Neuroptera and other insects. Mycologia 99:842–858. https://doi.org/10.3852/mycologia.99.6.842.
3. Ji Z-H, Jia JH, Bai F-Y. 2009. Four novel Candida species in the Candida albicans/Lodderomyces elongisporus clade isolated from the gut of flower beetles. Antonie Van Leeuwenhoek 95:23–32. https://doi.org/10.1007/s10482-008-9282-7.
4. Liu X-J, Yi Z-H, Ren Y-C, Li Y, Hui F-L. 2016. Five novel species in the Lodderomyces clade associated with insects. Int J Syst Evol Microbiol 66:4881–4889. https://doi.org/10.1099/ijsem.0.001446.
5. Hodorova V, Lichancova H, Bujna D, Nebohacova M, Tomaska L, Brejova B, Vinar T, Nosek J. 4 June 2018. De novo sequencing and high-quality assembly of yeast genomes using a MinION device. London Calling, 2018, London, UK. https://nanoporetech.com/resource-centre/de-novo-sequencing-and-high-quality-assembly-yeast-genomes-using-minion-device.
6. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37:540–546. https://doi.org/10.1038/s41587-019-0072-8.
7. Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32:2103–2110. https://doi.org/10.1093/bioinformatics/btw152.
8. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
9. Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res 27:737–746. https://doi.org/10.1101/gr.214270.116.
10. 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.
11. Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595. https://doi.org/10.1093/bioinformatics/btp698.
12. UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49:D480–D489. https://doi.org/10.1093/nar/gkaa1100.
13. Shen XX, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, Haase MAB, Wisecaver JH, Wang M, Doering DT, Boudouris JT, Schneider RM, Langdon QK, Ohkuma M, Endoh R, Takashima M, Manabe RI, Čadež N, Libkind D, Rosa CA, DeVirgilio J, Hulfachor AB, Groenewald M, Kurtzman CP, Hittinger CT, Rokas A. 2018. Tempo and mode of genome evolution in the budding yeast subphylum. Cell 175:1533–1545.e20. https://doi.org/10.1016/j.cell.2018.10.023.
14. Shen X-X. 2018. Tempo and mode of genome evolution in the budding yeast subphylum. Figshare. Data set. https://doi.org/10.6084/m9.figshare.5854692.v1. Accessed 20 December 2022.
15. Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31. https://doi.org/10.1186/1471-2105-6-31.
16. 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. https://doi.org/10.1186/1471-2105-7-62.
17. Cillingová A, Tóth R, Mojáková A, Zeman I, Vrzoňová R, Siváková B, Baráth P, Neboháčová M, Klepcová Z, Brázdovič F, Lichancová H, Hodorová V, Brejová B, Vinař T, Mutalová S, Vozáriková V, Mutti G, Tomáška Ľ, Gácser A, Gabaldón T, Nosek J. 2022. Transcriptome and proteome profiling reveals complex adaptations of Candida parapsilosis cells assimilating hydroxyaromatic carbon sources. PLoS Genet 18:e1009815. https://doi.org/10.1371/journal.pgen.1009815.
18. Frith MC, Kawaguchi R. 2015. Split-alignment of genomes finds orthologies more accurately. Genome Biol 16:106. https://doi.org/10.1186/s13059-015-0670-9.
19. Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York, NY.
20. McGowan J. 2020. BUSCO_phylogenomics. https://doi.org/10.5281/zenodo.4320788.
21. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol 38:4647–4654. https://doi.org/10.1093/molbev/msab199.