AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

High-Quality Genome Reconstruction of *Candida albicans* CHN1 Using Nanopore and Illumina Sequencing and Hybrid Assembly

Shipra Garg,[a] Piyush Ranjan,[b] John R. Erb-Downward,[b] Gary B. Huffnagle[a,c,d]

[a]Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, Michigan, USA
[b]Division of Pulmonary and Critical Care Medicine, University of Michigan, Ann Arbor, Michigan, USA
[c]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, USA
[d]Mary H. Weiser Food Allergy Center, University of Michigan, Ann Arbor, Michigan, USA

ABSTRACT We report an improved, nearly closed, high-quality draft genome reconstruction of the *Candida albicans* CHN1 strain (ATCC MYA-4779), a human isolate, using Illumina and Nanopore sequencing. Covering six complete and two partial nuclear chromosomes along with a partial mitochondrial genome, this assembly is 14,787,852 bases in size, with 5,935 genes.

The *Candida albicans* CHN1 strain is a human isolate that has been shown to stably colonize cefoperazone-pretreated mice (1–3). DNA sequencing of this isolate was performed using the long-read Nanopore MinION and high-accuracy Illumina MiSeq platforms, with a hybrid assembly being used for genome reconstruction. Yeast cells were grown in Sabouraud dextrose broth (Difco, Detroit, MI) at 37°C. The Qiagen DNeasy blood and tissue kit was used to isolate DNA for Illumina sequencing. Library preparation using the PrepX DNA library kit (number 640101; TaKaRa) was followed by bead size selection for inserts of 220 bp. Sequencing on the MiSeq platform was performed with 500 cycles at the University of Michigan DNA Sequencing Core. A total of 6,941,089 raw read pairs (2 × 251 bp) were subjected to adapter trimming using TrimGalore v0.5.0 (https://github.com/FelixKrueger/TrimGalore) and Cutadapt v1.18 (4) with a minimum Phred score of 30. The resulting 6,820,861 paired-end reads were then merged together using FLASH v1.2.11 (5) with a minimum overlap of 25 nucleotides, leading to 6,634,784 single-end reads. High-molecular-weight DNA for Nanopore sequencing was isolated using the Zymolyase-based extraction protocol recommended by Oxford Nanopore Technologies (ONT) for yeast DNA, with modifications from the Qiagen DNeasy blood and tissue kit. Library preparation was performed using the rapid sequencing kit SQK-RAD004 from ONT, and no size selection or shearing was applied. Nanopore sequencing was performed locally on the MinION platform. A total of 485,750 raw reads, with an estimated N_{50} of 23,280 bases, were obtained after base calling with Guppy v3.2.9. Reads were trimmed and filtered to remove adapter contamination using SNIKT v0.1.1 (https://github.com/piyuranjan/SNIKT) and seqtk v1.3 (https://github.com/lh3/seqtk). NanoPlot v1.28.2 (6) was used for quality checking (380,747 single-end reads, with a read N_{50} of 20,290 bases).

De novo assembly of the Nanopore reads was performed using Flye v2.7 (7) with three iterations of error correction. The Illumina high-accuracy short reads were used to perform another independent error correction (polishing) of the contigs using BWA v0.7.17 (8) and Pilon v1.23 (9). The resulting assembly contained 23 contigs, with a largest contig length of 3,195,262 bases, a total throughput of 14,787,852 bases, an assembly N_{50} value of 1,642,426 bases, and a GC content of 33.62%, as assessed by QUAST v5.0.2 (10). BUSCO v4.0.6 (11) was used with the saccharomycetes_odb10 data set to estimate the genome completeness of the assembly as 98.1% (of 2,137 benchmarking universal

Citation Garg S, Ranjan P, Erb-Downward JR, Huffnagle GB. 2021. High-quality genome reconstruction of *Candida albicans* CHN1 using Nanopore and Illumina sequencing and hybrid assembly. Microbiol Resour Announc 10:e00299-21. https://doi.org/10.1128/MRA.00299-21.

Editor Jason E. Stajich, University of California, Riverside

Address correspondence to Gary B. Huffnagle, ghuff@umich.edu.

Received 23 March 2021
Accepted 17 May 2021
Published 10 June 2021

single-copy ortholog [BUSCO] groups) with 35 missing core single-copy genes (CSCGs) from this lineage at the class level. As a comparison, *C. albicans* strain SC5314 (NCBI assembly number ASM18296v3 and RefSeq accession number GCF_000182965.3) has a completeness of 98.8% with 22 missing CSCGs, suggesting that some of these genes may no longer constitute the core *C. albicans* genome. The assembly contigs were then aligned with the strain SC5314 genome using MUMmer v4.0.0beta2 (12) and analyzed using Mauve v20150226 (13) for estimation of chromosomal completeness, organization, and synteny. Chromosomes 1 to 6 were recovered in full in this assembly. Chromosomes 7 and R were recovered in two and three long contigs, respectively. The mitochondrial genome was recovered in three contigs. Several small fragments that overlapped genomic regions in the SC5314 and CHN1 synteny comparisons were also retrieved. This genome is known to be diploid (14), and overlaps may represent sections of chromosomes with alternate alleles.

Data availability. This whole-genome shotgun project has been deposited in DDBJ/ENA/GenBank under the accession number JAFFGX000000000. The version described in this paper is version JAFFGX010000000. The *C. albicans* CHN1 GenBank Assembly accession number is GCA_017309835.1. The raw reads are available under accession numbers SRX9854709 (Illumina) and SRX9854710 (Nanopore) within BioProject PRJNA692229.

ACKNOWLEDGMENTS

We are grateful to the University of Michigan DNA Sequencing Core and Robert Dickson for support of the Illumina and Nanopore sequencing, respectively. We thank Nicole Falkowski for technical assistance.

Funding for this project was provided by NIAID/NIH grant 5R01 AI138348 and the Nina and Jerry D. Luptak Endowment from the Mary H. Weiser Food Allergy Center at the University of Michigan.

REFERENCES

1. Noverr MC, Noggle RM, Toews GB, Huffnagle GB. 2004. Role of antibiotics and fungal microbiota in driving pulmonary allergic responses. Infect Immun 72:4996–5003. https://doi.org/10.1128/IAI.72.9.4996-5003.2004.
2. Noverr MC, Falkowski NR, McDonald RA, McKenzie AN, Huffnagle GB. 2005. Development of allergic airway disease in mice following antibiotic therapy and fungal microbiota increase: role of host genetics, antigen, and interleukin-13. Infect Immun 73:30–38. https://doi.org/10.1128/IAI.73.1.30-38.2005.
3. Downward JRE, Falkowski NR, Mason KL, Muraglia R, Huffnagle GB. 2013. Modulation of post-antibiotic bacterial community reassembly and host response by *Candida albicans*. Sci Rep 3:2191. https://doi.org/10.1038/srep02191.
4. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10. https://doi.org/10.14806/ej.17.1.200.
5. Magoč T, Magoč M, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963. https://doi.org/10.1093/bioinformatics/btr507.
6. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 34:2666–2669. https://doi.org/10.1093/bioinformatics/bty149.
7. 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.
8. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997. https://arxiv.org/abs/1303.3997.
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. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.
11. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
12. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol 14:e1005944. https://doi.org/10.1371/journal.pcbi.1005944.
13. Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403. https://doi.org/10.1101/gr.2289704.
14. Odds FC, Brown AJP, Gow NAR. 2004. *Candida albicans* genome sequence: a platform for genomics in the absence of genetics. Genome Biol 5:230–237. https://doi.org/10.1186/gb-2004-5-7-230.