AMERICAN SOCIETY FOR MICROBIOLOGY
Microbiology® Resource Announcements

Draft Genome Sequence of a Diploid and Hybrid *Candida* Strain, *Candida sanyaensis* UCD423, Isolated from Compost in Ireland

Adam Ryan,[a] Eoin Ó Cinnéide,[b] Sean A. Bergin,[a] Ghozlan Alhajeri,[a] Hawraa Almotawaa,[a] Isabelle Daly,[a] Sophia Heneghan,[a] Kellie Horan,[a] Roslyn Kavanagh,[a] Christopher Keane,[a] Aaron Martin,[a] Ada McDonagh,[a] Julia O'Leary,[a] Matthieu Osborne,[a] Emma Watson,[a] Kevin P. Byrne,[b] Kenneth H. Wolfe,[b] Geraldine Butler[a]

[a]School of Biomolecular and Biomedical Science, Conway Institute, University College Dublin, Dublin, Ireland
[b]School of Medicine, Conway Institute, University College Dublin, Dublin, Ireland

ABSTRACT *Candida sanyaensis* is a CUG-Ser1 clade yeast that is associated with soil. Assembly of short-read and long-read data shows that *C. sanyaensis* has a diploid and hybrid genome, with approximately 97% identity between the haplotypes. The haploid genome size is approximately 15.4 Mb.

The yeast *Candida sanyaensis* is a member of the subphylum Saccharomycotina and the phylum Ascomycota and is closely related to *Candida sojae* and *Candida tropicalis* (1). The species was originally isolated from soil samples from Hainan Island in south China and Taiwan. *C. sanyaensis* UCD423 was isolated from a wormery in Dublin by two passages of compost material in 9 ml liquid yeast extract-peptone-dextrose (YPD) medium containing chloramphenicol (30 μg/ml) and ampicillin (100 μg/ml) and culture on YPD plates at room temperature, similar to a method reported previously (2). The species was identified from the internal transcribed spacer (ITS) sequence, which is 99% identical to that of *C. sanyaensis* (1).

For Illumina sequencing, total genomic DNA was extracted from a YPD culture and purified by extraction with phenol-chloroform-isoamyl alcohol. Libraries were generated from 1 μg genomic DNA and sequenced by BGI Tech Solutions Co. (Hong Kong), as described by Morio et al. (3). A total of 150 bases were sequenced from each end with an Illumina HiSeq 4000 instrument, yielding 6.5 million spots. For long-read sequencing, genomic DNA was extracted using the Qiagen Genomic-tip 100G kit. The sequencing library was generated from 400 ng of DNA using a rapid barcoding kit (SQK-RBK004) from Oxford Nanopore Technologies (ONT), following the manufacturer's instructions. This library was mixed with three libraries from unrelated projects, purified with AMPure XP magnetic beads (Beckman Coulter), and eluted in 12 μl Tris-EDTA (TE) buffer, of which 10 μl was used for sequencing on a FLO-MIN106 flow cell primed with kit EXP-FLP002 on a MinION 1B sequencer using MinKNOW software v4.1.22 (ONT). Base calling (using the fast model [dna_r9.4.1_450bps_fast.cfg]) and demultiplexing were performed using Guppy v4.2.2 (ONT). The total number of MinION reads was 561,441, with a read N_{50} of 11,647 bp.

Low-quality reads (Q scores of $<$15) were removed from the Illumina data using Skewer v0.2.2 (4), and reads were assembled using SPAdes v3.14.0 (5) with default parameters. The MinION data were filtered using NanoFilt v2.7.1 (6), removing reads with quality scores of $<$10 and lengths of $<$10 kb. Reads were assembled using Canu v2.0 (7) with default parameters for a haploid assembly or with the options corOutCoverage = 200 and batOptions = -dg 3 -db 3 -dr 1 -ca 500 -cp 50 for a diploid assembly, assuming a genome size of 18 Mb. Assembly statistics were visualized using QUAST (8) (Table 1). Very different assemblies were obtained depending on the method and parameters used (Table 1).

Using the diploid parameters with Canu resulted in an assembly that largely kept two haplotypes separate (Fig. 1). The first haplotype (haplotype A) is represented by

Citation Ryan A, Ó Cinnéide E, Bergin SA, Alhajeri G, Almotawaa H, Daly I, Heneghan S, Horan K, Kavanagh R, Keane C, Martin A, McDonagh A, O'Leary J, Osborne M, Watson E, Byrne KP, Wolfe KH, Butler G. 2021. Draft genome sequence of a diploid and hybrid *Candida* strain, *Candida sanyaensis* UCD423, isolated from compost in Ireland. Microbiol Resour Announc 10:e00761-21. https://doi.org/10.1128/MRA.00761-21.

Editor Jason E. Stajich, University of California, Riverside

Copyright © 2021 Ryan et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Geraldine Butler, gbutler@ucd.ie.

Received 28 July 2021
Accepted 25 August 2021
Published 23 September 2021

Downloaded from https://journals.asm.org/journal/mra on 05 October 2021 by 137.43.140.170.

TABLE 1 Assembly statistics for *C. sanyaensis* (for scaffolds of >500 bp)

	Data for assembly with:		
Parameter	**SPAdes**	**Canu (haploid)**	**Canu (diploid)**
No. of contigs[a]	9,833	144	207
Total length (bp)[b]	17,851,344	24,222,500	31,277,210
Size of largest contig (bp)	177,256	2,922,747	2,944,088
N_{50} (bp)[b]	2,983	1,214,081	281,063
L_{50}	1,342	6	12
GC content (%)	31.1	31.6	31.5

[a] The SPAdes assembly of the short-read data alone is highly fragmented. The haploid Canu assembly of the long-read data has the smallest number of contigs.

[b] For the long-read data alone, the haploid assembly has a smaller genome size and a greater N_{50} value than the diploid assembly.

the largest 10 scaffolds, ranging from 2.94 Mb to 369 kb (Fig. 1). The second haplotype (haplotype B) is fragmented into much smaller contigs, each of which matches a contig from the first haplotype (Fig. 1). Therefore, *C. sanyaensis* has a diploid genome and the haplotypes differ by ~3.3%, suggesting that the genome results from hybridization between related but not identical parents (9). The final diploid assembly of the MinION data was error corrected by incorporating the Illumina data, using nine rounds with Pilon v1.23 (10).

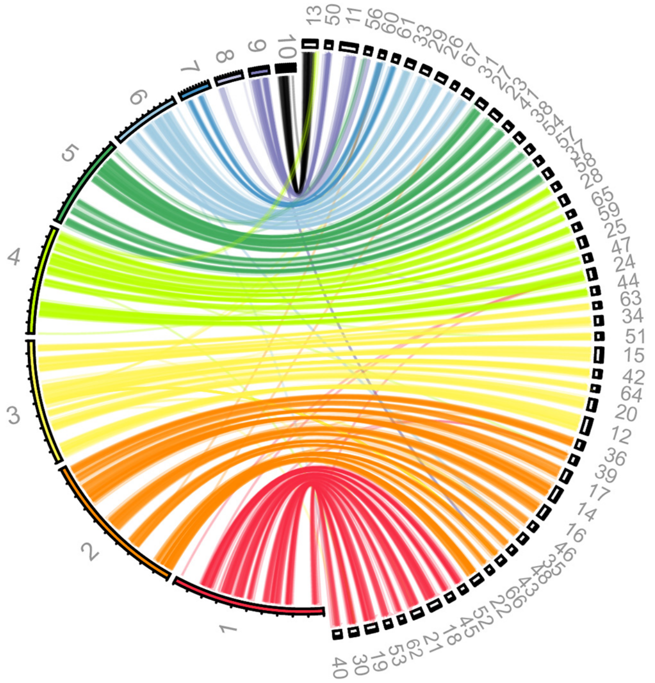

FIG 1 *C. sanyaensis* UCD423 has a diploid genome. Similarity between the two haplotypes of *C. sanyaensis* UCD423 was visualized with Circos (11) and Circoletto (12) as described by O'Brien et al. (13), using the diploid assembly after polishing with Pilon. The 10 scaffolds from haplotype A are shown on the inner ring on the left. Scaffolds of ≥100 kb from haplotype B are shown on the outer ring on the right. Smaller scaffolds are omitted for clarity. Sequences with similarity were identified by BLASTN, and alignments are plotted as links between the two haplotypes. The matches are colored with respect to the scaffold in the first haplotype. Scaffold 8 contains the ribosomal DNA (rDNA) and has few matches with scaffolds from the second haplotype outside the rDNA region, even when scaffolds of <100 kb are included. Assembly of the two haplotypes might have collapsed around the rDNA locus, possibly because of loss of heterozygosity. Matches to scaffold 56 suggest that scaffold 5 and scaffold 8 should be joined. Similarly, matches to scaffold 13 suggest that scaffold 4 and scaffold 10 should be joined. The total length of a single haplotype is approximately 15.4 Mb. The average sequence identity between the haplotypes is 96.72%, as calculated from scaffolds of >100 kb using the average nucleotide identity (ANI) calculator described by Rodriguez and Konstantinidis (14) with default parameters.

Downloaded from https://journals.asm.org/journal/mra on 05 October 2021 by 137.43.140.170.

Data availability. This whole-genome shotgun project has been deposited in DDBJ/ENA/GenBank under accession number CAJVQF00000000. The raw reads from Illumina sequencing are available under SRA accession number ERR6313261 and those from MinION sequencing under SRA accession number ERR6310792. These data are also available under project PRJEB46370. The ITS sequence is at accession number MZ507576.

ACKNOWLEDGMENTS

This work was supported by undergraduate teaching resources from University College Dublin. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Thanks go to Rebeca Clavero for isolating the yeast.

REFERENCES

1. Hui FL, Niu QH, Ke T, Li YX, Lee CF. 2013. *Candida sanyaensis* sp. nov., an ascomycetous yeast species isolated from soil. Antonie Van Leeuwenhoek 103:47–52. https://doi.org/10.1007/s10482-012-9785-0.
2. Sylvester K, Wang QM, James B, Mendez R, Hulfachor AB, Hittinger CT. 2015. Temperature and host preferences drive the diversification of *Saccharomyces* and other yeasts: a survey and the discovery of eight new yeast species. FEMS Yeast Res 15:fov002. https://doi.org/10.1093/femsyr/fov002.
3. Morio F, O'Brien CE, Butler G. 2020. Draft genome sequence of the yeast *Kazachstania telluris* CBS 16338 isolated from forest soil in Ireland. Mycopathologia 185:587–590. https://doi.org/10.1007/s11046-020-00449-6.
4. Jiang H, Lei R, Ding SW, Zhu S. 2014. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics 15:182. https://doi.org/10.1186/1471-2105-15-182.
5. 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. https://doi.org/10.1089/cmb.2012.0021.
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. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res 27:722–736. https://doi.org/10.1101/gr.215087.116.
8. 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.
9. Gabaldon T. 2020. Hybridization and the origin of new yeast lineages. FEMS Yeast Res 20:foaa040. https://doi.org/10.1093/femsyr/foaa040.
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. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645. https://doi.org/10.1101/gr.092759.109.
12. Darzentas N. 2010. Circoletto: visualizing sequence similarity with Circos. Bioinformatics 26:2620–2621. https://doi.org/10.1093/bioinformatics/btq484.
13. O'Brien CE, Zhai B, Ola M, Ó Cinnéide E, O'Connor Í, Rolling T, Miranda E, Babady NE, Hohl TM, Butler G. 2021. Identification of a novel *Candida metapsilosis* isolate suggests ongoing hybridization. bioRxiv 2021.07.15.452539. https://doi.org/10.1101/2021.07.15.452539.
14. Rodriguez-R LM, Konstantinidis KT. 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ 4:e1900v1. https://doi.org/10.7287/peerj.preprints.1900v1.

Downloaded from https://journals.asm.org/journal/mra on 05 October 2021 by 137.43.140.170.