Draft Genome Sequence of Clinical Isolate USM039K of the Pathogenic Yeast *Candida parapsilosis*

Dina Yamin,ᵃ Wan Khairunnisa Wan Juhari,ᵃ,ᵇ NurWaliyuddin Hanis Zainal Abidin,ᶜ Shuhaila Mat-Sharani,ᵈ Azian Harunᵃ,ᵉ

ᵃDepartment of Medical Microbiology and Parasitology, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian, Kelantan, Malaysia
ᵇMalaysian Node of the Human Variome Project, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian, Kelantan, Malaysia
ᶜHuman Identification/DNA Unit, School of Health Sciences, Universiti Sains Malaysia, Kubang Kerian, Kelantan, Malaysia
ᵈBiomedicine Program, School of Health Sciences, Universiti Sains Malaysia, Kubang Kerian, Kelantan, Malaysia
ᵉHospital Universiti Sains Malaysia, Kota Bharu, Kelantan, Malaysia

ABSTRACT Here, we announce the draft genome sequence of a *Candida parapsilosis* clinical isolate (USM039K) recovered from a patient with catheter-related bloodstream infection (CRBSI). The genome size is 12,860,016 bp long, with 188 scaffolds, a G+C content of 38.65%, and 5,467 genes.

This report is to present the draft genome sequence of a clinical isolate of *Candida parapsilosis*. *C. parapsilosis* is a yeast classified under cellular organisms, superkingdom of Eukaryota, kingdom of Fungi, subkingdom of Dikarya, phylum of Ascomycota, subfamily of Saccharomycotina, class of Saccharomycetes, order of Saccharomycetales, family of Debaryomycetaceae, and genus of *Candida*. This species is known as a causative agent for candidemia (1). The draft genome sequence of this species will enhance the current knowledge of the genetic variation among *C. parapsilosis* isolates. The organism was isolated from a patient with catheter-related bloodstream infection (CRBSI) in Hospital Universiti Sains Malaysia (USM), Kelantan, Malaysia. Ethical approvals were obtained from the Human Research Ethics of Universiti Sains Malaysia (JEPeM-USM-16040162).

The isolate was cultivated overnight on a Sabouraud dextrose agar plate with yeast cultivation medium and incubated at 37℃. Pure colonies were collected during the stationary phase. DNA extraction was done using phenol-chloroform extraction followed by ethanol precipitation (2). Species identity was confirmed by the sequences of the internal transcribed spacer (ITS) region of ribosomal DNA (3). Sample quantity and quality were assessed prior to sequencing using the Illumina NovaSeq6000 system.

Briefly, DNA fragmentation was performed by sonication to 350 bp. These fragments were end polished, A tailed, and ligated with the Illumina sequencing adapter prior to amplification by PCR. Library preparation was accomplished for the qualified DNA using the NEBNext Ultra DNA library prep kit (New England BioLabs [NEB], USA) according to the manufacturer's protocol. DNA libraries were purified using the AMPure XP system, analyzed for size distribution using an Agilent 2100 bioanalyzer, and quantified by real-time PCR. Then, the qualified libraries were sequenced to an average sequencing depth of $250\times$ using the Illumina NovaSeq 150PE protocol generating 2×150-bp paired-end reads. The raw output was transformed into raw reads by CASAVA base calling and converted into FASTQ file format, which produced 22,252,138 raw paired-end reads. The quality of the FASTQ file was assessed using FastQC software. The adapter and low-quality sequences were discarded using Trimmomatic (v0.36) by performing sliding window trimming with a minimum average quality of 20 and a minimum sequence read length of 20 bases (4).

De novo genome assembly for the *C. parapsilosis* USM039K strain and annotation analyses were conducted using the Galaxy platform with default parameters or those otherwise stated (5). The assembly was generated by SPAdes (v3.12.0) (6), and the quality was assessed using

Quality Assessment Tool (QUAST) (v5.0.2) (7). The draft genome resulted in a total of 188 scaffolds ($\geq$1,000 bp) with a total length of 12,860,016 bp, a G+C content of 38.65%, an N_{50} value of 173,876 bp, and an L_{50} value of 23. The longest scaffold in the draft genome is 533,946 bp. The assemblies produced 98% of aligned sequences with the reference genome (CDC317; GCA_000182765.2). Interspersed and low-complexity repetitive sequences were masked using RepeatMasker v4.0.9 (http://repeatmasker.org) with combined Dfam_3.0 and RepBase-20181026 databases according to the fungal species. The results showed that 287,166 bp (2.23%) was masked (8). The gene prediction of final scaffolds was performed using MAKER (v2.31.10) (9), AUGUSTUS v3.3.3 (10), and SNAP software (11). The species training was *Candida albicans*. In total, 5,467 genes were predicted, consisting of 4,655 (83.5%) single-exon genes and 923 (16.5%) multiexon genes in this newly sequenced strain.

Data availability. This whole-genome shotgun project has been deposited and is available at DDBJ/ENA/GenBank (BioProject number PRJNA610714 and assembly accession number JADCQT000000000). The version described in this paper is the first version, JADCQT010000000. Raw reads are available in the NCBI Sequence Read Archive under accession number SRR11249109.

REFERENCES

1. Tóth R, Nosek J, Mora-Montes HM, Gabaldon T, Bliss JM, Nosanchuk JD, Turner SA, Butler G, Vágvölgyi C, Gácser A. 2019. Candida parapsilosis: from genes to the bedside. Clin Microbiol Rev 32:e00111-18. https://doi.org/10.1128/CMR.00111-18.
2. Sambrook J, Russell D. 2001. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
3. White TJ, Bruns T, Lee SJWT, Taylor J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. PCR protocols: a guide to methods and applications 18(1):315–322.
4. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
5. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 46:W537–W544. https://doi.org/10.1093/nar/gky379.
6. 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.
7. 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.
8. Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013–2015. http://www.repeatmasker.org.
9. Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. Curr Protoc Bioinformatics 48:4.11.1–4.11.39. https://doi.org/10.1002/0471250953.bi0411s48.
10. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34:W435–W439. https://doi.org/10.1093/nar/gkl200.
11. Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59. https://doi.org/10.1186/1471-2105-5-59.