



Draft Genome Sequence of *Candida railenensis* Strain CLIB 1423, Isolated from Papaya Fruit in French Guiana

Hugo Devillers, Cécile Grondin, Angèle Thiriet, Jean-Luc Legras

► To cite this version:

Hugo Devillers, Cécile Grondin, Angèle Thiriet, Jean-Luc Legras. Draft Genome Sequence of *Candida railenensis* Strain CLIB 1423, Isolated from Papaya Fruit in French Guiana. Microbiology Resource Announcements, In press, 10.1128/MRA.00554-22 . hal-03901576

HAL Id: hal-03901576

<https://hal.inrae.fr/hal-03901576>

Submitted on 15 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.


L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



Draft Genome Sequence of *Candida railenensis* Strain CLIB 1423, Isolated from Papaya Fruit in French Guiana

 Hugo Devillers,^a Cécile Grondin,^a Angèle Thiriet,^a Jean-Luc Legras^a

^aSPO, Univ Montpellier, INRAE, Institut Agro, Montpellier, France

ABSTRACT Here, we report the draft genome sequence and annotation of the yeast *Candida railenensis* strain CLIB 1423. The assembly consists of 57 nuclear scaffolds and 1 complete mitochondrial chromosome, for a total of 13.8 Mb (N_{50} , 0.54 Mb; L_{50} , 9). The annotation contains 6,013 coding DNA sequences (CDSs) (BUSCO completeness, 99.6%).

Candida railenensis is an ascomycetous yeast from the family *Debaryomycetaceae*. It is closely related to the genus *Kurtzmaniella* (1). Only a few species from this clade of yeasts have been sequenced yet, while a recent study (2) stressed the critical need for new sequenced genomes to decipher the taxonomic relationships in the *Kurtzmaniella* species complex.

C. railenensis strain CLIB 1423 was isolated from a papaya fruit in French Guiana in 2010 during an extensive collection of yeasts performed by the Biological Resource Center CIRM—Levures (3). Cells were grown for 72 h at 28°C on liquid yeast extract-peptone-dextrose (YPD) medium. Genomic DNA was extracted following a phenol-chloroform extraction protocol. The quality of the genomic DNA was evaluated using spectrophotometry and electrophoresis on a 0.7% agarose gel. The prepared DNA was then sent to the iGenSeq sequencing platform (Brain Institute, Paris) for library preparation using the Nextera XT kit. The libraries were sequenced using the Illumina NovaSeq 6000 platform with the SP reagent kit (300 cycles), yielding 6,062,601 read pairs (paired-end [PE], 2 × 150-bp format; coverage, 132×).

The reads were trimmed using Fastp v0.23.1 (4) using default parameters, except that the minimal read length was set to 50 nucleotides (18% of reads were discarded). The cleaned reads were assembled using SPAdes v3.15.3 (5) with default parameters, considering the following kmers: 21, 33, 55, and 77. With a minimal length of 5 kb, 58 scaffolds were selected as the final assembly, representing a cumulative length of 13,800,509 bases, an average GC content of 39.16%, and a N_{50} value of 539,109 bases. This assembly consists of 57 linear nuclear scaffolds and 1 complete circular mitochondrial chromosome.

Structural annotation of the coding genes (CDSs) was performed using Augustus v3.4.0 (6) with the pretrained *Debaryomyces hansenii* data set as the reference, without additional data, limiting the intron length to 4 kb. This yielded 5,767 putative CDSs. Manual curation of these genes was performed with particular attention paid to genes containing introns, to detect spurious gene fusion. As a result, 6,013 CDSs were reported. The completeness of this annotation was evaluated using BUSCO v5.2.2 (7) with the *saccharomycetes_odb10* lineage data set as the reference (2,137 proteins). Before curation, the BUSCO score was 98.6% (98.3% single copy). After curation, the BUSCO score reached 99.6% complete (99.3% single copy). Functional annotation of these 6,013 CDSs was performed using an annotation transfer tool that we are developing (<https://github.com/hdevillers/go-fannot>), which combines functional annotation retrieved from homology against the curated UniProt database (8) and motif detection using InterProScan (9).

Editor Jason E. Stajich, University of California, Riverside

Copyright © 2022 Devillers et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](#).

Address correspondence to Hugo Devillers, hugo.devillers@inrae.fr.

The authors declare no conflict of interest.

Received 31 May 2022

Accepted 2 November 2022

In addition to protein coding genes, 222 tRNA genes were predicted using tRNAscan-SE v2.0.9 (10). A complete rRNA unit was identified in scaffold 43, as well as three isolated 5S rRNA features. The long and short subunit rRNAs of the mitochondria were annotated. Last, 6 small nuclear/nucleolar RNA features (U1 to U6) were reported in the annotation. All rRNA and noncoding RNA (ncRNA) features were identified using BLASTn (11) with reference features from the curated annotations of *D. hansenii* (GenBank accession number [GCF_000006445.2](https://www.ncbi.nlm.nih.gov/nuccore/GCF_000006445.2)) and *Saccharomyces cerevisiae* ([GCF_000146045.2](https://www.ncbi.nlm.nih.gov/nuccore/GCF_000146045.2)).

Data availability. The data are available under the BioProject accession number [PRJEB51943](https://www.ncbi.nlm.nih.gov/bioproject/PRJEB51943). The BioSample accession number is [SAMEA13788477](https://www.ncbi.nlm.nih.gov/biosample/SAMEA13788477). The raw sequencing data are available under the SRA accession number [ERR9433495](https://www.ncbi.nlm.nih.gov/sra/ERR9433495). The GenBank assembly accession number is [GCA_935541525.1](https://www.ncbi.nlm.nih.gov/nuccore/GCA_935541525.1). The complete phenol-chloroform extraction protocol used is available at <https://www.protocols.io/view/dna-extraction-protocol-5qpvorx47v4o/v1>.

ACKNOWLEDGMENTS

We thank Sylvain Santoni for his valuable advice regarding the DNA extraction protocol. This genome was sequenced as part of the French National Research Institute for Agriculture, Food, and Environment DISC/CNOC (INRAE) CIRM 2020 strain sequencing project.

REFERENCES

- Lachance M-A, Starmer WTY. 2008. *Kurtzmaniella* gen. nov. and description of the heterothallic, haplontic yeast species *Kurtzmaniella cleridarum* sp. nov., the teleomorph of *Candida cleridarum*. *Int J Syst Evol Microbiol* 58:520–524. <https://doi.org/10.1099/ijse.0.65460-0>.
- Lopes MR, Santos ARO, Moreira JD, Santa-Brígida R, Martins MB, Pinto FO, Valente P, Morais PB, Jacques N, Grondin C, Casaregola S, Lachance M-A, Rosa CAY. 2019. *Kurtzmaniella hittingeri* f.a. sp. nov., isolated from rotting wood and fruits, and transfer of three *Candida* species to the genus *Kurtzmaniella* as new combinations. *Int J Syst Evol Microbiol* 69:1504–1508. <https://doi.org/10.1099/ijsem.0.003337>.
- Jacques N, Casaregola S. 2019. Large biodiversity of yeasts in French Guiana and the description of *Suhomyces coccinellae* f.a. sp. nov. and *Suhomyces faveliae* f.a. sp. nov. *Int J Syst Evol Microbiol* 69:1634–1649. <https://doi.org/10.1099/ijsem.0.003369>.
- Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>.
- Prijbelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes de novo assembler. *Curr Protoc Bioinforma* 70:e102. <https://doi.org/10.1002/cpbi.102>.
- Keller O, Kollmar M, Stanke M, Waack S. 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* 27:757–763. <https://doi.org/10.1093/bioinformatics/btr010>.
- Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness, p 227–245. In Kollmar M (ed), *Gene prediction: methods and protocols*. Springer, New York, NY.
- UniProt Consortium. 2021. UniProt: the universal protein knowledge base in 2021. *Nucleic Acids Res* 49:D480–D489. <https://doi.org/10.1093/nar/gkaa1100>.
- Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, Richardson L, Salazar GA, Williams L, Bork P, Bridge A, Gough J, Haft DH, Letunic I, Marchler-Bauer A, Mi H, Natale DA, Necci M, Orengo CA, Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Bateman A, Finn RD. 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* 49:D344–D354. <https://doi.org/10.1093/nar/gkaa977>.
- Chan PP, Lowe TM. 2019. tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol Biol* 1962:1–14. https://doi.org/10.1007/978-1-4939-9173-0_1.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. <https://doi.org/10.1186/1471-2105-10-421>.