

Draft genome sequence of *Kluyveromyces marxianus* MCHA

Gabriel Ibarra-Rivera,¹ Luis Alfonso Muñoz-Miranda,² Alejandro Pereira-Santana,³ Melchor Arellano-Plaza,¹ Anne Gschaedler-Mathis,¹ Manuel R. Kirchmayr,¹ Luis Joel Figueroa-Yañez¹

AUTHOR AFFILIATIONS See affiliation list on p. 2.

ABSTRACT The complete genome sequence of *Kluyveromyces marxianus* strain MCHA was obtained using Illumina and Oxford Nanopore sequencing technologies. The genome comprises nine scaffolds, corresponding to the eight nuclear chromosomes and the mitochondrial genome.

KEYWORDS *Kluyveromyces marxianus*, genome sequence

Kluyveromyces marxianus is an emerging nonconventional food-grade yeast isolated from diverse environments, such as fermented products, industrial waste, and plants (1). In fermented foods and beverages, it contributes to aroma formation through the production of volatile organic compounds (VOCs) (2). Its rapid growth rate, thermotolerance, and metabolic versatility have positioned *K. marxianus* as a promising organism for biotechnological applications (3). We report the draft genome sequence of *K. marxianus* strain MCHA isolated from *Agave cupreata* juice fermentations in a mezcal factory located in Piedras de Lumbre, Morelia, Michoacán, Mexico. The strain was obtained by serial dilution plating on YPD medium (10 g L⁻¹ yeast extract, 20 g L⁻¹ peptone, and 20 g L⁻¹ glucose; pH 6.0) and incubated at 30°C for 24–48 h with shaking at 200 rpm. Taxonomic identification was confirmed by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry.

Genomic DNA was extracted from fresh yeast cultures grown in YPD medium at an optical density of 1.5 using the Wizard Genome DNA Purification Kit (Promega, USA) following the manufacturer's instructions. DNA was sent to Novogene Co., Ltd. (Beijing, China) for library preparation and sequencing. Illumina libraries were prepared using the TruSeq DNA PCR-Free Library Preparation Kit (Illumina, USA) according to the manufacturer's instructions. Sequencing was performed on an Illumina HiSeq 2000 platform (2 × 100 bp), generating 6,309,056 paired-end reads. Read quality was assessed using FastQC v0.11.9 (4), and adapter sequences and low-quality bases were removed using Trimmomatic v0.39 (5).

For Oxford Nanopore sequencing, genomic DNA was processed without prior mechanical shearing. Libraries were prepared using the Rapid Barcoding Kit (SQK-RBK004; Oxford Nanopore Technologies, UK) according to the manufacturer's protocol, without additional size selection. Sequencing was performed on a MinION Mk1B device using a FLO-MIN106 (R9.4.1) flow cell. Basecalling was performed using Guppy v6.3.8 (Oxford Nanopore Technologies) with GPU support, and adapter trimming and demultiplexing were performed with Guppy's internal scripts. Low-quality and contaminant reads were filtered using NanoFilt v2.8.0 (6) and Filtlong v0.2.1 (7).

High-quality long reads ($n = 67,953$; total bases = 252.5 Mb) were assembled using Flye v2.9.1 (8) with the `--nano-raw` option and an estimated genome size of 11 Mb. The assembly was polished using Minimap2 v2.24 (9) and four rounds of Racon v1.5.0 (10), followed by Medaka v1.6.0 (11). A second round of polishing was performed using Illumina data with Pilon v1.23 (`--fix all option`) (12). Alignment of the Illumina reads to the assembly was carried out using Bowtie 2 (13).

Editor André O. Hudson, Rochester Institute of Technology, Rochester, New York, USA

Address correspondence to Luis Joel Figueroa-Yañez, lfigueroa@ciatej.mx.

The author declares no conflict of interest.

Received 18 February 2026

Accepted 20 March 2026

Published 20 April 2026

Copyright © 2026 Ibarra-Rivera et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

TABLE 1 Genome sequencing statistics.

Parameter	Value
Oxford Nanopore sequencing	
Read number	67,953
Total bases (nt)	252,503,966
Illumina sequencing	
Read number	6,309,056
Total bases (nt)	946,358,400
Genome assembly	
Total length	10,881,469 pb
Number of scaffolds	9
Longest scaffold	1,734,893 pb
N50	1,410,736 pb
GC content (%)	40.03
Genome coverage	21.35×
Protein-coding genes	4,787
tRNA genes	187
Genome assembly assessment	
Complete BUSCOs	98.3%
Fragmented BUSCOs	0.4%
Missing BUSCOs	1.3%

The final assembly was scaffolded with RagTag (14) using the reference genome of *K. marxianus* DMKU3-1042 ([GCA_001417885.1](https://doi.org/10.1101/001417)). Assembly quality was evaluated with QUAST v4.1 (15), and completeness was assessed using BUSCO v5.8.2 (16) with the *Saccharomyces* database. Gene prediction and functional annotation were performed with Funannotate v1.8.14 (17) using *Kluyveromyces lactis* as the training species. The final genome assembly consisted of nine scaffolds with high completeness based on BUSCO analysis. Default parameters were used for all software unless otherwise specified. Detailed sequencing, assembly, and annotation metrics are summarized in Table 1.

ACKNOWLEDGMENTS

The first author, G. Ibarra-Rivera (CVU 784110), thanks SECIHTI (formerly CONAHCYT) for the doctoral scholarship granted.

AUTHOR AFFILIATIONS

¹Industrial Biotechnology Unit, Center for Research and Assistance in Technology and Design of the State of Jalisco, A.C. (CIATEJ), Guadalajara, Jalisco, Mexico

²Center for Research on Infectious Diseases, Department of Microbiology and Pathology, University Center for Health Sciences, University of Guadalajara, Guadalajara, México

³SECIHTI-Research and Assistance Center in Technology and Design of the State of Jalisco, Parque Científico Tecnológico de Yucatán, Mérida, A.C. (CIATEJ), Mérida, Yucatán, Mexico

AUTHOR ORCIDs

Gabriel Ibarra-Rivera  <http://orcid.org/0000-0002-5119-2538>

Luis Joel Figueroa-Yañez  <http://orcid.org/0000-0002-7283-8035>

AUTHOR CONTRIBUTIONS

Gabriel Ibarra-Rivera, Conceptualization, Data curation, Formal analysis, Writing – original draft | Luis Alfonso Muñoz-Miranda, Data curation, Writing – review and editing | Alejandro Pereira-Santana, Data curation, Writing – review and editing | Melchor Arellano-Plaza, Data curation, Writing – review and editing | Anne Gschaedler-Mathis,

Project administration, Resources, Writing – review and editing | Manuel R. Kirchmayr, Project administration, Writing – review and editing.

DATA AVAILABILITY

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under accession [JBUZOE000000000](#). The version described in this paper is [JBUZOE010000000](#). The raw sequencing reads are available at the NCBI Sequence Read Archive under accession numbers [SRR34364681](#) and [SRR34364682](#). The version described in this paper is the first version.

REFERENCES

1. Bilal M, Ji L, Xu Y, Xu S, Lin Y, Iqbal HMN, Cheng H. 2022. Bioprospecting *Kluyveromyces marxianus* as a robust host for industrial biotechnology. *Front Bioeng Biotechnol* 10:851768. <https://doi.org/10.3389/fbioe.2022.851768>
2. Perpetuini G, Tittarelli F, Perla C, Tofalo R. 2022. Influence of different aggregation states on volatile organic compounds released by dairy *Kluyveromyces marxianus* strains. *Foods* 11:2910. <https://doi.org/10.3390/foods11182910>
3. Reina-Posso D, Gonzales-Zubieta FA. 2025. Expanding horizons: the untapped potential of *Kluyveromyces marxianus* in biotechnological applications. *Fermentation* 11:98. <https://doi.org/10.3390/fermentation11020098>
4. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
5. 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>
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. Wick RR, Menzel P. 2017. Filtlong. Available from: <https://github.com/rrwick/Filtlong>. Retrieved 19 Aug 2025.
8. 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>
9. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
10. 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>
11. Oxford Nanopore Technologies. 2023. Medaka: sequence correction tool for Oxford Nanopore reads (version 1.6.0) [software]. Available from: <http://github.com/nanoporetech/medaka>
12. 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>
13. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>
14. Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol* 23:258. <https://doi.org/10.1186/s13059-022-02823-7>
15. 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>
16. Manni M, Berkeley MR, Seppey M, Zdobnov EM. 2021. BUSCO: assessing genomic data quality and beyond. *Curr Protoc* 1:e323. <https://doi.org/10.1002/cpz1.323>
17. Palmer JM, Stajich JE. 2022. Computer software. Funannotate (Version 1.8.13) [Computer Software]. Available from: <https://github.com/nextgenusfs/funannotate>