# Chromosome level genome assembly and annotation of *Hanseniaspora mollemarum* CBS 18055 strain

Authors: Nicole Xanthe Bennis, Harshitha Indudhar Hanji, Marcel van de Broek, Jean-Marc Darana

Department of Biotechnology, Delft University of Technology, Van der Maasweg 9, 2629 HZ, Delft, the Netherlands

Corresponding author: Jean-Marc G. Daran,

Contact information: .

[j.g.daran@tudelft.nl](mailto:j.g.daran@tudelft.nl).

Department of Biotechnology, Delft University of Technology, Van der Maasweg 9, 2629 HZ, Delft, the Netherlands

## General introduction:

*Hanseniaspora* species gained attention due to the ability of these species to ferment simple sugars and to actively contribute to the development of bouquet aromas in wine and cider fermentations. We present a chromosome-level assembly of an isolate of *H. mollemarum* that would enhance its potential applications.

## Purpose:

The *Hanseniaspora* genus comprises over 20 species including the recently described *Hanseniaspora gamundiae* (1), *H. smithiae* (2), *H. terricola* (3) and *H. mollemarum*. The latter species was isolated in a soil sample from a Dutch garden (CBS 15034) (4). Following, new strains from Great-Britain and Hungary have been deposited at the Westerdijk fungal biodiversity institute (<https://wi.knaw.nl/>). However, the sole available genome derives from the CBS 15034 strain (Bioproject\_PRJNA631240). To match the interest of the beverage industry for *Hanseniaspora* strains, it is critical to gain both physiological and genomic knowledge.. Here, we report the genome sequence of a new strain of *Hanseniaspora mollemarum*.

## Description of the data:

The *H. mollemarum* CBS 18055 was isolated from an apple tree (*Malus domestica*) located in the South of France (43°01’24.7”N, 0°58’30.4”E). The sample was grown at 4°C on YPD medium (10 g.L-1 Yeast-Extract, 20 g.L-1 Bacto-Peptone, 20 g.L-1 glucose) supplemented with 3% ethanol. Single colonies were obtained by re-streaking the culture on solid YPD medium supplemented with 25 µg.mL-1chloramphenicol and 50 µg.mL-1 ampicillin incubated at room-temperature. This process was repeated thrice. Final isolates were stored in 30% (v/v) glycerol at -80°C.

Sanger sequencing of the ITS and the D1/D2 region of the rDNA of CBS 18055 amplified with primers ITS1/ITS4 (5) and NL1/NL4 (6) respectively revealed a 99.98 and 100% identity to that of *H. mollemarum* CBS 15034, respectively.

For genome sequencing, total DNA was extracted using the Genomic-tip 100/G kit protocol (Qiagen, Hilden, Germany). Shotgun library was prepared using Oxford\_Nanopore Technologies' SQK-LSK109 kit (ONT, Oxford, United Kingdom) and sequenced on a MinION MK1B device with R10 flow cell. Raw FAST5 signal files were base-called using GPU Guppy (ONT, version 4.5.4) in high-accuracy mode. After filtering for length (>1kb), 94060 reads were obtained, yielding 1.04 Gb (N50: 19089 bp), representing ~112-fold coverage of a *H. mollemarum* genome. Canu version 2.0 with settings genomeSize=12m, useGrid=0 and nanopore-raw was used for *de novo* assembly yielding eight contigs (7). For error correcting the assembly, a 150-bp read length TruSeq PCR-Free Illumina library (Illumina, San Diego, CA) with a 350-bp insert-size was sequenced on a NovaSeq 6000 (Illumina) by Novogene (Cambridge, United Kingdom) yielding 18,811,684 reads for a total of 2.82 Gigabases (307-fold coverage). Reads quality was assessed with FastQC v0.11.5. The genome was polished by mapping the untrimmed Illumina reads with Burrows-Wheeler Aligner (BWA version 0.7.15-r1142-dirty; default parameters used) (8, 9) to the assembly and further processed with SAMtools (version 1.3.1) (10, 11) and polished once by Pilon (version 1.18) with settings fix all (12).

This yielded a genome assembly of 9.18 Megabase (N50 = 1.33-Mb) with a GC content of 34,96%, comprising eight contigs. Contig 1-7 (CHR2-to-8) showed telomeric sequences at both ends while contig 8 (CHR1) had telomere sequences on one chromosome arm indicating that this contig also represented (part of) a linear chromosome (Table 1). Canu being notoriously known to have difficulty in assembling circular contigs, the fastq files were assembled using UniCycler Hybrid assembler (13). A circular contig of 32.9 kbp was identified as mitochondrial chromosome (Table 1). Annotation of the polished assembly was performed using Funannotate v1.8.15 (<https://github.com/nextgenusfs/funannotate>) (14) as previously described in Add (15).

## File description:

Hmollemarum\_CBS18055.annotations.txt (23 MB) 3086022b21525684ce25c840bdb5310d: Text file presenting the annotation of all Open reading frame identified with FUNannotate.

Hmollemarum\_CBS18055.fasta (9 MB) 938ff592da55e742d0d6303c9927d472 : MultiFASTA files of the 8 assembled chromosomes (7 nuclear and 1 mitochondrial).

Hmollemarum\_CBS18055.gff3 (2 MB): General feature format file describing the genome annotation.

In bioinformatics, the general feature format (gene-finding format, generic feature format, GFF) is a file format used for describing genes and other features of DNA, RNA and protein sequences.