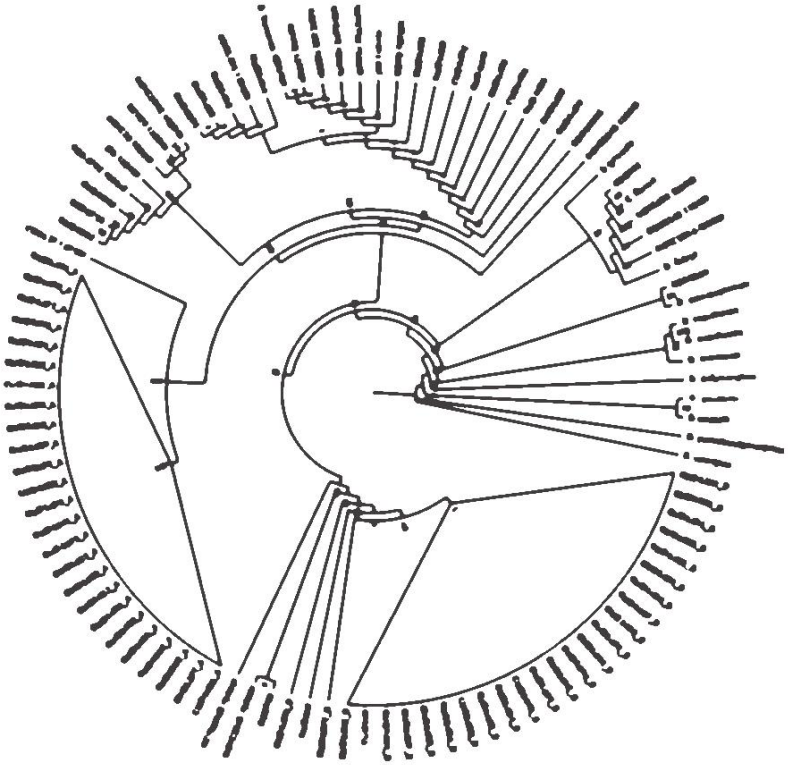
UNIVERSITY OF GOTHENBURG

DEPARTMENT OF CHEMISTRY AND MOLECULAR BIOLOGY

# Population genomics of the marine yeast *Debaryomyces hansenii*



**Mercè Montoliu Nerín**

Degree project for Master of Science (120 HEC) with a major in Molecular Biology with Specialization in Genomics and Systems Biology
2016, 60 HEC project
Second Cycle

# Index

## Abstract

In the present study, the genomes of 17 strains previously classified as *Debaryomyces hansenii* were sequenced. The analysis of their genomes through alignment of the sequences to reference genomes, variant calling, genome annotation and phylogenies suggested that five of the strains are, in fact, different species. Two being hybrids with other species, two being reclassified as *Candida flareri* and one unknown species. The species considered *D. hansenii* were grouped in two different clades, seven in the clade of the reference strain CBS767, and five in a different clade. These results are essential to understand the complexity of the *Debaryomyces* species classification and emphasize the need for genomic studies in yeast.

# Introduction

*Debaryomyces hansenii* is an ascomycetous yeast belonging to the order of Saccharomycetales. It is characterized by being salt-tolerant and high pH-tolerant (Prista *et al.,* 1997). This marine yeast can tolerate salinity levels up to 24% (4.11 M) of NaCl (Norkrans, 1966), representing one of the most halotolerant species of yeasts and becoming a model organism for the study of salt tolerance mechanisms in eukaryotic cells (Prista *et al.,* 2005). In saline environments *D. hansenii* accumulates large amounts of Na+ without being intoxicated (Neves *et al.,* 1997). Furthermore, sodium improves its growth performance, even under stress conditions (Prista *et al.,* 1997).

The physiology and biochemistry of salt-tolerance has been well studied in *D. hansenii* (Adler, 1986). Previous studies indicate production and retention of compatible solutes like glycerol in response to high salt concentrations (Gustafsson and Norkrans, 1976; Adler *et al*., 1985; André *et al*., 1988; Thomé and Trench, 1999; Thomé, 2005). Other studies have focused on ion transport (Norkrans, 1966) and the role of the cell wall (Thomé, 2007).

The genetics behind the pH and salt tolerance of *D. hansenii* are still not perfectly known. A few genes involved in osmoadaptation have been isolated (Thomé, 2004; Prista *et al*., 2005; Chao *et al*., 2009). Two genome references have been published: CBS 767, with a total genome size of 12.1 Mb, 7 nuclear chromosomes and mitochondria with a median of 6290 proteins and 6658 genes (Dujon *et al.,* 2004);

MTCC 234, with an estimated genome size of 11.4 Mb and 541 scaffolds (Kumar *et al.*, 2012). CBS 767 is the strain that has been used the most as a reference for all the recent research on *D. hansenii*, in studies like the one from Arroyo Gonzalez *et al.* (2009), that tested 6320 genes using a genome-wide expression array and discovered that only 109 were differentially expressed in presence of salt. One important characteristic of *D. hansenii* with importance for genomic studies is that it uses an alternative codon table, where the CUG codon codes for Serine instead of Leucine, as in *Candida albicans* (Moura et al., 2007).

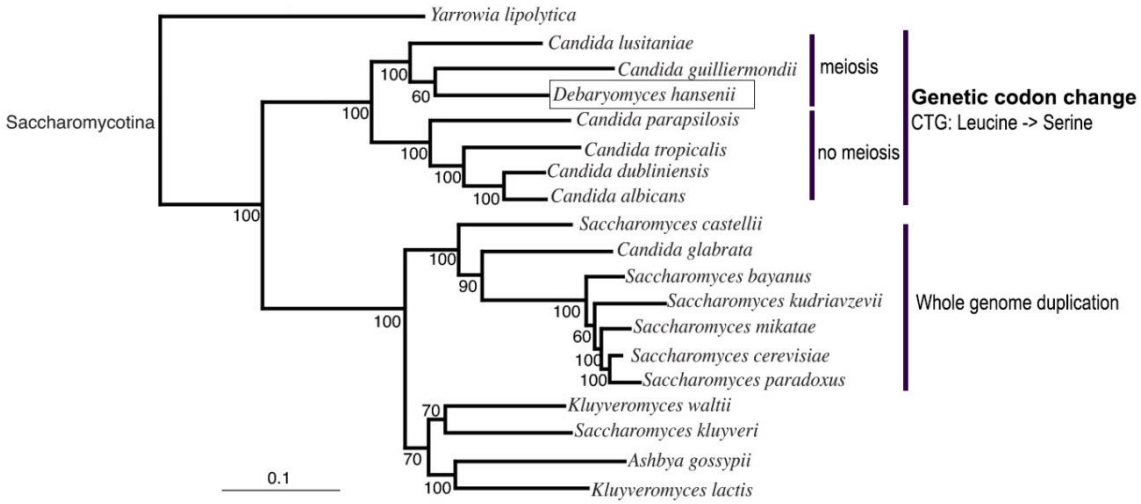



**Figure 1. Phylogeny of *Saccharomycotina*.** Two important features in two clades are made, the whole genome duplication of the clade of *S. cerevisiae*, and the genetic codon change of the clade of *D. hansenii*, where most of *Candida* species are. Inside the second clade a new distinction can be made, between the species that can have meiosis and the ones without. Modified part of a phylogenetic tree from Fitzpatrick *et al*. 2006.

The characteristics of *D. hansenii* make it possible for this specie to thrive in a wide variety of natural habitats and various manufactured food products, especially cheese, in which it participates in the maturation of the products, or sometimes occur as a contaminant (Jacques *et al.*, 2009) In addition, *D. hansenii* has also been implicated as a potential pathogen under its anamorphic form *Candida famata* var. *famata* (Jacques and Casaregola, 2008).

The taxonomic classification of the species related to *D. hansenii* has always been subject to debate. Originally two species were described: *D. hansenii* (Zopf) Lodder et Kreger van Rij and *Debaryomyces fabryi* Ota. Subdivision of the species *D. hansenii* (Zopf) Lodder et Kreger van Rij into two varieties *D. hansenii* var. *hansenii* and *D.*

*hansenii* var. *fabryi* was proposed on the basis of the electrophoretic patterns of glucose-6-phosphate dehydrogenase and on the maximum growth temperature (Nakase and Suzuki, 1985). Since then, multiple studies with different molecular techniques have tested which of the two taxonomic classifications is the most accurate (Prillinger *et al.*, 1999; Corredor *et al*., 2000; Daniel and Meyer, 2003; Romero *et al.*, 2005; Quiros *et al.*, 2006; Martorell *et al.* 2005; Tsui *et al.*, 2008; Jacques *et al.*, 2009). Most of the tests indicated that they were indeed two distinct species and they were finally classified as *D. hansenii* and *D. fabryi*.

Phylogenetic analysis using conserved spliceosomal intron sequence comparison has shown that what was known as *D. hansenii* is actually a complex of species with at least four members: *D. hansenii*, *Debaryomyces tyrocola*, *D. fabryi* and *Candida flareri* (previously *Candida famata* var. *flareri*, also named as *Debaryomyces subglobosus* (Groenewald *et al.*, 2008)) (Jacques *et al*., 2009). A later study of the group of Jacques *et al*. (2010), analysing the polymorphism of nuclear mitochondrial DNA insertions (NUMT), revealed the existence of at least three populations (clades A to C) in *D. hansenii*, with the first one containing CBS 767, the strain of the published reference genome (Dujon *et al.*, 2004), and the last one containing one strain of *C. famata* var. *famata* (CBS 1795). NUMT analysis also revealed the existence of both haploid and diploid strains, these last ones resulting from crosses between different *D. hansenii* clades (Jacques *et al*., 2010). The same group continued with their research on *Debaryomyces* sp. by taking strains in Artic glaciers and using the gene sequence of ACT1 to classify them (Jacques *et al*., 2015), the same sequence that was previously used in Jacques *et al.*, 2009. The coding sequences are available in GeneBank and were retrieved and included in the present study.

Ongoing work is carried out in order to on find quick and easy ways to differentiate between the different species, like using the PAD1 gene in a simple and affordable PCR method (Wrent *et al.* 2015). In the present study, whole genome sequencing of 17 strains stored in CBS and NCYC collections and classified under the specie name of *Debaryomyces hansenii* (Table 1) were carried out. Among these strains, the importance of two strains, J26 and J63, should be mentioned. They were isolated from Swedish waters and candidates for future phenotypic and genomic investigation by this research group. The strain NCYC2572 should be the same as CBS767, but originating from a different collection, so it was taken initially as a reference for the current project. DBH9

could be a second reference in the project, as the only difference compared to CBS767 is a point mutation on the DhHIS4 gene (Minhas *et al*., 2012).

### Aim

Because of its physiological characteristics, *D. hansenii* provides an excellent model system to examine evolutionary mechanisms for adaptation to both low and high salinities. To initiate a first population genomics study on this marine yeast, 17 strains presumed to be *D. hansenii* were obtained from culture collections representing a wide geographical spread and coming from various sources. The aim was to establish a description of the wide genotype space in this species, which would be the base for future studies on isolates from the salinity gradient along the Swedish coast, as well as for the experimental selection over hundreds of generation of specific traits and their genetics with the long term goal of linking genotype and phenotype.

# Material and methods

### Strains

The strains of *D. hansenii* used in this project were taken from NCYC (National Collection of Yeast Cultures, UK) and CBS (CBS-KNAW culture collection, Netherlands) collections. They were picked in different locations around the world, coming from different environments, and in different years, but all of them registered as *D. hansenii* because of their phenotype.

Two strains (J63 and J26) were sampled from Swedish seawater by the University of Gothenburg. Two strains (J16 and J52) were *Saccharomyces cerevisiae* included as a control for the methods.

| | Strain | Alternative strain names | Classification | Origin | Year |
|---|---|---|---|---|---|
| **1001** | NCYC2572 | CBS767, ATCC36239, CCRC21394, DBVPG6050, IFO0083, JCM1990, JCM2102, KCTC7645, MUCL30242, NRRLY-7426, NRRLY-10976, UCD74-86 | *D. hansenii* var. *hansenii* | Carlsberg laboratories, habitat associated with fermentation | 1994 |
| **1002** | NCYC8 | NCTC2059 | *D. hansenii* | Throat of patient with angina | 1925 |
| **1003** | NCYC9 | NCTC2048 | *D. hansenii* | Dutch cheese prepared in Russia | 1924 |
| **1004** | NCYC10 | NCTC2056 | *D. hansenii* | Unknown | 1925 |
| **1005** | NCYC103 | NCTC1681 | *D. hansenii* | Carlsberg laboratories, habitat | 1923 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | associated with fermentation | |
| **1006** | NCYC459 | | *D. hansenii* | Soil, New Zealand | 1955 |
| **1007** | NCYC475 | CBS811, JCM1439, NRRLY-1454, UCD75-11 | *D. hansenii* var. *hansenii* | Fermenting Kentucky tobacco | 1956 |
| **1008** | NCYC792 | NCMB1230 43 | *D. hansenii* | Sea Water | 1974 |
| **1009** | NCYC3045 | | *D. hansenii* | Dried salted black olives from Thassos, Greece | 2002 |
| **1010** | NCYC3364 | | *D. hansenii* var. *fabryi* | Capping machine of soft drinks factory in Brazil | 2006 |
| **1011** | CBS117 | | *D. hansenii* | Rennet from New Zealand | 1985 |
| **1012** | CBS5140 | | *D. hansenii* | Skin of a man, Hungary | |
| **1013** | CBS1101 | IFO0027, IFO0093 | *D. hansenii* | Salt pork | 1946 |
| **1014** | CBS1792 | | *D. hansenii* | Chilled beef from Brisbane, Australia | |
| **1015** | J63 | | *D. hansenii* | Seawater, Sweden | |
| **1016** | J26 | | *D. hansenii* | Seawater, Sweden | |
| **1017** | J16 | | *S. cerevisiae* | - | - |
| **1018** | J52 | | *S. cerevisiae* | - | - |
| **1019** | DBH9 | | *D. hansenii* | CBS767 carrying a mutation in DhHIS4 gene. | |

**Table 1. Strains used in this study,** with their identification number, strain name, alternative names, current classification and from where and when they were isolated.

The strains will be named using their strain name and the identification number given in the above table (from 1001 to 1019) for a better understanding of the results.

## Sequencing

The DNA extraction was done using the Bioline Isolate II genomic DNA Kit (http://goo.gl/Q0TuA1 (Last visited 2016-04-12)).

From the same DNA sample, two libraries were created for sequencing using Illumina HiSeq 2500 with a 2x101 setup in HighOutput mode on SciLifeLab Stockholm and stored at UPPMAX Next Generation sequence Cluster Storage.

## Assembly

Assembly was done by Mahesh Panchal, bioinformatician at the Department of Medical Biochemistry and Microbiology, Genomics; Manfred Grabherr. Uppsala University. Assembly expert on BILS (Bioinformatics Infrastructure for Life Sciences). Assemblies were done with several assemblers (Spades (Bankevich *et al*. 2012), MaSuRCA (Zimin

*et al*. 2013), SOAPdenovo (Luo *et al*. 2012), ABySS (Simpson *et al.* 2009)) and the best result for each dataset was chosen for further analysis.

Assemblers producing the best assembly for each strain:

- Spades: NCYC2572 (1001), NCYC8 (1002), NCYC9 (1003), NCYC10 (1004), NCYC103 (1005), NCYC459 (1006), NCYC475 (1007), NCYC792 (1008), CBS5140 (1012), CBS1101 (1013), CBS1792 (1014), J63 (1015), J26 (1016), DBH9 (1019).
- MaSuRCA: NCYC3045 (1009), NCYC3364 (1010), CBS117 (1011), J16 (1017), J52 (1018).

An improvement of the assemblies was tried using SOAPdenovo with the suggested k-mer sizes from PreQC (https://goo.gl/Uh8MVU (Last visited 2016-05-20)).

Usage example:

```
sga preprocess --pe-mode 1 1006_AH_uniq1.fastq 1006_AH_uniq2.fastq > 1006_AH.fastq
sga index -a ropebwt --no-reverse -t 8 1006_AH.fastq
sga preqc -t 8 1006_AH.fastq > 1006_AH.preqc
sga-preqc-report.py 1006_AH.preqc /home/tomasl/preqc_examples/*.preqc
```

## Reads

Reads quality check

FastQC on all the reads was done by SciLifeLab. Fastqc reports can be found on GitHub: (https://goo.gl/Vi8QdJ) for each of the libraries.

Read coverage

Read coverage was calculated using a script stored in my personal GitHub repository: https://goo.gl/HzbPGF.

To calculate it, the fastq read length information and number of reads from FastQC, and the genome size from the already made assemblies was used. Estimated genome sizes are given in Annex table 1.

PCR duplicates removal and merge of libraries

PCR duplicates was handled using a Perl script from Linnéa Smeds (2010): filterPCRdupl_v1.01.pl, available on https://goo.gl/L3rXxe (Last visited: 2016-04-07).

Whole pipeline that includes filterPCRdupl and cat libraries into one: https://goo.gl/X7XwJB.

## Contamination check and species classification

Kraken (v.0.10.5-beta) is an ultrafast metagenomic sequence classification pipeline using exact alignments to a reference database of known sequences (Wood and Salzberg, 2014). Manual page: http://goo.gl/XNeFdP (Last visited 2016-04-07).

Usage example:

Build standard database (Containing Bacteria, Virus and Plasmid refSeq): `kraken-build --standard --db DBs_name`

Build custom database (Containing Bacteria, Virus, plasmid, fungi and human refSeq):

`kraken-build --download-taxonomy --db DBc_name`

`kraken-build --download-library bacteria --db DBc_name` (bacteria, viruses, plasmids)

Download RefSeq from NCBI and add genomes: `kraken-build --add-to-library Refseq.fa --db DBc_name`

Reads: `kraken --threads NUMBER --db DB_name --fastq-input reads.fastq > results.kraken`

Alignments: `kraken --threads NUMBER --db DB_name --fasta-input reads.fasta > results.kraken`

Output translation: `kraken-translate --db DB_name results.kraken > labeled_results.labels`

Output report: `kraken-report --db DB_name results.kraken > report_results.report`

Whole pipeline from building standard and custom databases and to run it for all the reads and assemblies is available here: https://goo.gl/N4LQ2M.

## Alignment

Bowtie2 (v.2.2.8) was used to align reads to reference genomes of *D.hansenii* and closely related species (Langmead and Salzberg, 2012). Manual page of Bowtie2: http://goo.gl/w4Saql (Last visited: 2016-04-07).

Usage example:

```
bowtie2-build reference_genome.fna Ref_genome_name

bowtie2 -x Ref_genome_name -1 forward_reads.fastq -2
reverse_reads.fastq -S aligned_output.sam 2>&1 | tee -i
Screen_output_saved.txt
```

Reads to reference genome CBS 767 (*D. hansenii*)

Alignment of reads to the reference genome of *D. hansenii*. Results showed that some strains might not be what we were expecting, probably being a different species. That is why in further steps the alignment was done to other species.

Genome downloaded from: http://goo.gl/oo9rcz (Last visited: 2016-04-07)

Pipeline: https://goo.gl/DX1XdW.

Reads to reference genome S288c (*S. cerevisiae*)

This alignment was done to check that the two control strains were, in fact, *S. cerevisiae* as expected.

Genome downloaded from: http://goo.gl/A1TDk0 (Last visited: 2016-04-07)

Pipeline: https://goo.gl/wOQ3nQ.

Reads to reference genome ATCC 6260 (*Meyerozyma guilliermondii*)

*D. hansennii* has been often misidentified with *M. guilliermondii* in the past. Mostly with *D. hansenii* strains that were human fungal pathogens (Desnos-Ollivier *et al.*, 2008).

Genome downloaded from: http://goo.gl/i4qar1 (Last visited: 2016-04-07)

Pipeline: https://goo.gl/diwQ97.

Reads to reference genome CBS 789 (*D. fabryi*)

The recent publication of the genome of *D. fabryi* allowed the alignment of our strains to this reference specie close to *D. hansenii* and often misidentified with it (Tafer *et al.*, 2016).

Genome downloaded from: http://goo.gl/ceeGIS (Last visited: 2016-04-07)

Pipeline: https://goo.gl/UdLQwV.

Pairwise alignment of all reads to all assemblies

Knowing that some strains are probably different species or different clades of the same specie the alignment of the reads against all the assemblies of all our strains made it possible to indicate which ones could belong to the same specie and which couldn't.

All reads from both libraries were merged for every strain and thereafter aligned to the assembly for each strain.

Pipeline: https://goo.gl/Fyk573.

## Variant calling

A variant calling analysis was performed to get a better understanding of the differences between strains supposed to be from the same species, and how variable they are with regards to single nucleotide polymorphisms (SNPs).

Bowtie2 with merged libraries

Alignment of reads of the selected strains having their libraries merged previously to increase coverage, against CBS767 using Bowtie2.

Samtools (v.0.1.18) - Manual page: http://goo.gl/2hxJhW (Last visited 2016-04-13)

- view: to output bam format.
- sort: sort alignments by leftmost coordinates.
- index: index of bam files.

Usage example:

```
bowtie2-build reference_genome.fna Ref_genome_name

bowtie2 -x Ref_genome_name -1 forward_reads.fastq -2
reverse_reads.fastq -S alignment.sam

samtools view -Sb alignment.sam | samtools sort - alignment_sorted

samtools index alignment_sorted.bam
```

Whole pipeline: https://goo.gl/b6dE08.

Pre-Vcalling and Vcalling

The alignments obtained in the previous step were prepared for the variant caller programs by renaming the read groups in a way so that the program would be able to identify them as coming from the same sample. Finally, with samtools those reads were locally realigned and the bam files indexed again. The variant calling was performed by Frebayes and VarScan (for more details see below) both individually for each strain and also by running a cohort analysis.

Picard tools (v. 2.0.1), AddOrReplaceReadGroups: To perform a cohort variant analysis of all the strains, correct identification of individual samples is needed. Complete manual page: https://goo.gl/V0rsuF (Last visited: 2016-04-13).

Samtools (v. 0.1.18):

- calmd: Local realignment using the genome reference.
- index: Index of bam files.

- mpileup: to be used for VarScan cohort analysis of all the bam files at the same time.

Freebayes (v. 0.9.16): call variants without any filtering. - Complete manual page: http://goo.gl/5Ldvi4 (Last visited 2016-04-13).

- Cohort analysis: input as text file with all the bam files pathways. With and without calling indels.
- Individual analysis: input bam files individually each time. With and without calling indels.

VarScan (v. 2.3.9): call variants reducing the pre-set filtering. - Complete manual page: http://goo.gl/P1eAGq (Last visited 2016-04-13).

- indels
- snps

Usage example:

```
java -jar ~/picard/dist/picard.jar AddOrReplaceReadGroups
INPUT=alignment_sorted.bam OUTPUT=alignment_sorted_RG.bam RGID=name
RGLB=name RGPL=illumina RGPU=name RGSM=name

samtools calmd -Arb alignment_sorted_RG.bam Ref_genome.fasta >
alignment_sorted_RG_BAQ.bam

samtools index alignment_sorted_RG_BAQ.bam

echo -e
'alignment_sorted_RG_BAQ_1.bam\nalignment_sorted_RG_BAQ_2.bam…' >
bam.txt

freebayes --fasta-reference Ref_genome.fasta --ploidy 1 --bam-list
bam.txt > all_strains_variants_freebayes.vcf

freebayes --fasta-reference Ref_genome.fasta --ploidy 1 --bam
alignment_sorted_RG_BAQ_1.bam >
Individual_variant_analysis_freebayes_1.vcf

freebayes --fasta-reference Ref_genome.fasta --ploidy 1 --no-indels --
bam-list bam.txt > all_strains_variants_freebayes.vcf

freebayes --fasta-reference Ref_genome.fasta --ploidy 1 --no-indels --
bam alignment_sorted_RG_BAQ_1.bam >
Individual_variant_analysis_freebayes_1.vcf

samtools mpileup -f Ref_genome.fasta alignment_sorted_RG_BAQ_1.bam
alignment_sorted_RG_BAQ_2.bam … > bam_pileup.mpileup

java -jar VarScan.v2.3.9.jar mpileup2indel --min-reads2 10
bam_pileup.mpileup > all_strains_varscan_indels.vcf

java -jar VarScan.v2.3.9.jar mpileup2snp --min-coverage 5 --min-var-
freq 0.01 bam_pileup.mpileup > all_strains_varscan_snp.vcf
```

Whole pipeline: https://goo.gl/4L2bfS.

Post-Vcalling

snpEff (v. 4.2.) is a program to annotate the variants called (Cingolani *et al.* 2012). When installed, the database for *D. hansenii* is already there, but the chromosome names differ from the reference genome from NCBI and the codon table is set as standard instead of the alternative table for yeast. It was re-installed, with these changed configurations:

Check chromosomes: `java -jar snpEff/snpEff.jar -v GCA_000006445.2.29`

Change codon table on snpEff.config file:

```
GCA_000006445.2.29.genome : Debaryomyces_hansenii_cbs767
GCA_000006445.2.29.reference : ftp.ensemblgenomes.org
        GCA_000006445.2.29.CR382133.codonTable :
Alternative_Yeast_Nuclear
        GCA_000006445.2.29.B.codonTable : Alternative_Yeast_Nuclear
        GCA_000006445.2.29.C.codonTable : Alternative_Yeast_Nuclear
        GCA_000006445.2.29.CR382136.codonTable :
Alternative_Yeast_Nuclear
        GCA_000006445.2.29.E.codonTable : Alternative_Yeast_Nuclear
        GCA_000006445.2.29.F.codonTable : Alternative_Yeast_Nuclear
        GCA_000006445.2.29.G.codonTable : Alternative_Yeast_Nuclear
```

Changes chromosomes names were done in all the obtained vcf files according to the snpEff nomenclature using sed command.

Variants were annotated using snpEff trying to avoid upstream, downstream, intron and intergenic regions.

Usage example:

```
sed -e 's/CR382133.2/CR382133/g' -e 's/CR382134.2/B/g' -e
's/CR382135.2/C/g' -e 's/CR382136.2/CR382136/g' -e 's/CR382137.2/E/g'
-e 's/CR382138.2/F/g' -e 's/CR382139.2/G/g'
all_strains_variants_freebayes.vcf >
all_strains_variants_freebayes_ch.vcf

java -jar snpEff/snpEff.jar -v GCA_000006445.2.29 -no-downstream -no-
intergenic -no-intron -no-upstream -no-utr -stats stats.html
all_strains_variants_freebayes_ch.vcf >
all_strains_variants_fb_snpEff.vcf
```

Whole pipeline: https://goo.gl/1cI62p.

Filtering of variants on annotated vcf files. Multiple tests to get the best possible results using vcflib and SnpSift.

vcflib: Tools to manage vcf files, used to remove all the variants that were called, but not supported as different to the reference genome for all the strains (AF=0). Complete manual page: https://goo.gl/MFYjRO (Last visited 2016-04-13).

snpSift: Distributed together with snpEff, used to filter and manipulate vcf files. Used to remove the calls that in some strains were not supported as different to the reference genome, also to make files for each of the strains to count variants for each one, analysis of transitions/transversions rate for each one and filter using different quality thresholds. Command line and tests: https://goo.gl/6RyRSF.

## Genome annotation

The assemblies were annotated using maker (v. 2.31.8). Complete manual page: http://goo.gl/RwDxQE (Last visited 2016-04-13). The command: `/usr/local/maker/bin/maker -CTL` creates three control files in the current folder, where all the values will be set for each specific annotation run. Control files for each strain stored in: https://goo.gl/476rSr. In order to run the program, the command `maker` would be used in the folder with the three control files.

## Phylogeny

From the annotation transcripts a unique file was created with every sequence well identified by its strain number.

CD-HIT-EST was used to cluster the data, it takes a fasta file and for each nr representative sequence it generates a cluster of high identity sequences (Li and Godzik, 2006).

```
/home/tomasl/bin/cd-hit-v4.6.1-2012-08-27/cd-hit-est -i
/data02/merce/cd-hit/deha_all_transcripts.fasta -o deha_all_90 -c 0.9
-d 0 -M 5000 -T 8 -g 1 -r 1
```

The clusters presenting all the strains represented were selected for further analysis and the 100 longest selected for further phylogeny, not yet presented in this study.

Phylogeny based in the complete sequence of ACT1, its Exon 2 and GPD1.

All the assemblies were merged in a unique FASTA file using the command cat. A database containing all the assemblies was done through BLASTdb. User manual: http://goo.gl/btJyMm (Last visited 2016-05-18).

```
makeblastdb -in Allassemblies.fasta -dbtype nucl -parse_seqids
```

BLASTn run to find the sequences in the assemblies of the strains.

```
blastn -db /data02/merce/genes/DeHa_Allassemblies.fasta -query /data02/merce/genes/ACT1.fasta -out /data02/merce/genes/Deha_act1.txt

blastn -db /data02/merce/genes/DeHa_Allassemblies.fasta -query /data02/merce/genes/GPD1.fasta -out /data02/merce/genes/Deha_gpd1.txt
```

Extraction of the matching sequences from the FASTA file was done with the program of BLAST, blastdbcmd.

Usage example:

```
blastdbcmd -db /data02/merce/genes/DeHa_Allassemblies.fasta -entry 1001_NODE_47_length_73754_cov_48.5551_ID_13632 -range 41367-42849 -strand plus -out act1_1001.fasta
```

Procedure repeated to extract only the exon 2 for comparison with other species. The output FASTA files cat into one for the phylogeny.

All the commands used can be found here: https://goo.gl/I0N7ym.

Links to the data used from the research of Jacques *et al*., 2009 and Jacques *et al*., 2014: https://goo.gl/uULgFR.

Alignment of the sequences using MUSCLE v3.8.31 by Robert C. Edgar (Edgar, 2004).

```
muscle -in ACT1_Exon2_for_phylogeny_nocomasnospacesnocolnolines.fasta -out Muscle_ACT1_Exon2_for_phylogeny.fasta

muscle –in GPD1_idgood.fasta –out Muscle_GPD1.fasta
```

The output was given in FASTA format, so a convertor was needed to work with the files in PHYLIP. In this case, Alter was used, a program for format conversion (Glez-Peña *et al.* 2010).

JModelTest-2.1.10 was used to select the best model for the phylogeny. It is a tool to carry out statistical selection of best-fit models of nucleotide substitution. It implements five different model selection strategies: hierarchical and dynamical likelihood ratio tests (hLRT and dLRT), Akaike and Bayesian information criteria (AIC and BIC), and a decision theory method (DT). It also provides estimates of model selection uncertainty, parameter importance and model-averaged parameter estimates, including model-averaged tree topologies.

PhyML is a phylogeny software based on the maximum-likelihood principle (Guindon *et al*. 2010). The ideal model for the samples for the Exon 2 was GTR+G, and for the

complete sequence of ACT1 was GTR+I and the recommended settings from JModelTest were input into PhyML. For GPD1 the HYK+I+G model was used. All the trees were created using an initial neighbor-joining tree model and 100 bootstrap replicates. The visualization of the trees was done in FigTree v1.4.2, and in the same program trees were edited for a better presentation and understanding.

Phylogeny based on the SNP database

To create a phylogeny from the VCF files from the variant calling the PhyloSNP was run. It takes SNP data files to generate phylogenetic trees (Faisona *et al*., 2014). It never was able to be finalized and the issue is still unknown. Software page: https://goo.gl/8Oeicp (Last visited 2016-04-13). Because of the problems, the run was interrupted.

# Results

## Assembly quality and contamination check

*D. hansenii*, quality assessment report for each assembly: https://goo.gl/RCwqlE.
No big issues were found except for the strains NCYC10 (1004) and CBS117 (1011), for which the genomes were very fragmented and the N50 values low. Attempts at producing a better assembly with the currently available data was unrewarding, probably due to the low coverage (Annex table 2), and could, most likely, only be improved by generating new libraries.

Sequences were checked for contamination and no high levels of contamination were found in any of the samples (< 1% reads belonging to virus, bacteria, archaea or foreign plasmids). Detailed results for every sample on GitHub:
Raw-reads: https://goo.gl/eKNgVp.
Assemblies: https://goo.gl/HWXFt1.

## Sequence alignment

The sequences were aligned to the reference genome of *D. hansenii* CBS767 to check how well they mapped. Surprisingly, not all the strains mapped as expected to the reference genome, indicating that some of them may not be *D. hansenii* as assumed by earlier phenotypic classification. As NCYC2572 (1001) is supposed to be the same strain as CBS767 but stored in a different collection, we used that strain to set a threshold of what to consider as *D. hansenii*, even if its alignment rate was not optimal (79% overall alignment rate of the reads to the reference genome). According to that, there were 12 strains of *D. hansenii*, presenting an alignment rate between 77% and 90%. Among the rest of the strains, three were clearly not *D. hansenii*, NCYC459 (1006) and CBS5140 (1012) with an overall alignment rate of the reads no higher than 9%; and NCYC3364 (1010) with an alignment rate of 5,7%. Two other strains, NCYC3045 (1009) and CBS117 (1011), were mapping to the reference genome with an alignment rate of 69% and 51% respectively, indicating that they have a large part of their genome belonging to *D. hansenii*, but they could be hybrids with another specie. The strains J16 (1017) and J52 (1018), identified previously as *S. cerevisiae*, presented an overall alignment rate of 0,09% to the *D. hansenii* reference genome.

To check that J16 (1017) and J52 (1018) were, in fact, *S. cerevisiae*, the strains were mapped to the reference genome S228c, results can be found on the Annex table 3. A different species, *M. guilliermondii*, has been previously misidentified as *D. hansenii* in the past, in most of the cases with strains that were human pathogenic (Desnos-Ollivier *et al*., 2008). Some of the strains in this study were isolated from humans and belong to the unclassified strains. The mapping was done, but no clear match was found (Annex table 3). The genome of *D. fabryii*, one of the closest species to *D. hansenii* and famous for being a subject of debate on how to classify these two species, was published recently (Tafer *et al*., 2016). The available reference genome was used for mapping reads from strains in the present study to check for a presence of *D. fabryi* and none was found. Still, the strains NCYC459 (1006), CBS5140 (1012) and NCYC3364 (1010) were clearly closer to *D. fabryi* than to *D. hansenii*.

With the aim of understanding how closely related the studied strains are to each other, a pairwise alignment was done using the raw reads of all the strains against each one of their assemblies. Some strains had larger parts of their genome in common than to the others, forming different groups of strains. On one side we had NCYC2572 (1001) together with NCYC9 (1003), CBS1792 (1014) and J26 (1016), on another group there were the most closely related strains to the reference genome, NCYC8 (1002), NCYC103 (1005), NCYC475 (1007), NCYC792 (1008), CBS1101 (1013), J63 (1015), DBH9 (1019). The strain NCYC10 (1004) was impossible to classify with confidence, probably because of the bad quality of the data for that strain, in which the assembly is very fragmented. Among the unclassified species, the two probable hybrids were different between each other; and the strain NCYC3045 (1009) was closer to *D. hansenii* than CBS117 (1011). The strains NCYC459 (1006) and CBS5140 (1012) were clearly the same species as they map to each other almost as well as one strain to itself. The strain NCYC3364 (1010) was an individual unknown species.

## Pairwise alignment between strains

| | 1001 | 1002 | 1003 | 1004 | 1005 | 1006 | 1007 | 1008 | 1009 | 1010 | 1011 | 1012 | 1013 | 1014 | 1015 | 1016 | 1019 | CBS767 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1001** | 98.16 | 81.90 | 96.89 | 90.76 | 81.84 | 11.86 | 81.14 | 83.16 | 90.45 | 6.24 | 64.80 | 10.92 | 83.79 | 95.40 | 82.48 | 97.20 | 82.27 | 79.16 |
| **1002** | 81.94 | 97.65 | 82.04 | 93.38 | 96.86 | 12.84 | 94.58 | 95.10 | 78.21 | 6.79 | 62.61 | 12.78 | 96.54 | 82.51 | 95.94 | 82.16 | 97.36 | 90.41 |
| **1003** | 97.22 | 82.04 | 98.33 | 90.52 | 82.37 | 12.44 | 81.51 | 83.31 | 88.87 | 6.45 | 62.63 | 12.14 | 83.50 | 95.99 | 82.82 | 97.18 | 82.56 | 77.84 |
| **1004** | 86.41 | 91.04 | 86.67 | 96.41 | 90.39 | 12.01 | 89.94 | 89.88 | 82.46 | 6.61 | 62.27 | 11.78 | 91.46 | 87.17 | 90.24 | 87.02 | 91.13 | 85.99 |
| **1005** | 80.93 | 97.45 | 81.29 | 93.78 | 97.62 | 14.59 | 95.10 | 95.68 | 76.19 | 7.16 | 60.41 | 13.73 | 97.01 | 81.35 | 96.45 | 81.31 | 97.73 | 88.30 |
| **1006** | 12.86 | 12.17 | 12.91 | 12.77 | 12.31 | 98.16 | 12.26 | 12.93 | 12.24 | 18.38 | 11.62 | 97.77 | 12.34 | 12.88 | 12.96 | 12.94 | 12.30 | 8.70 |
| **1007** | 81.02 | 97.07 | 81.35 | 94.45 | 96.20 | 14.83 | 97.86 | 95.90 | 76.85 | 6.75 | 62.36 | 13.69 | 97.22 | 81.55 | 96.44 | 81.51 | 96.93 | 88.46 |
| **1008** | 81.81 | 94.95 | 82.04 | 93.26 | 94.75 | 13.41 | 94.01 | 97.33 | 78.34 | 6.66 | 63.68 | 12.82 | 96.07 | 82.13 | 97.01 | 82.73 | 95.02 | 89.50 |
| **1009** | 78.70 | 72.89 | 78.70 | 78.51 | 72.97 | 12.33 | 72.05 | 73.43 | 91.56 | 6.57 | 76.37 | 12.31 | 74.19 | 79.32 | 73.60 | 79.06 | 73.19 | 68.96 |
| **1010** | 9.75 | 9.39 | 9.76 | 9.67 | 9.50 | 20.05 | 9.39 | 9.92 | 7.72 | 93.78 | 7.79 | 20.11 | 9.64 | 9.60 | 9.74 | 10.28 | 9.57 | 5.72 |
| **1011** | 57.88 | 57.70 | 57.90 | 59.27 | 57.77 | 14.33 | 57.00 | 58.18 | 71.38 | 6.83 | 85.94 | 14.39 | 58.54 | 58.07 | 58.24 | 58.20 | 58.02 | 51.20 |
| **1012** | 13.08 | 12.51 | 13.11 | 13.01 | 12.61 | 97.15 | 12.57 | 13.25 | 12.55 | 18.88 | 12.00 | 97.97 | 12.64 | 13.04 | 13.28 | 13.19 | 12.62 | 8.97 |
| **1013** | 80.95 | 95.36 | 81.35 | 92.99 | 95.12 | 13.84 | 94.33 | 95.30 | 76.04 | 6.75 | 60.52 | 13.54 | 98.17 | 81.12 | 95.98 | 82.04 | 95.47 | 87.29 |
| **1014** | 93.14 | 81.79 | 95.66 | 92.39 | 81.46 | 12.93 | 81.55 | 81.59 | 91.12 | 6.19 | 63.64 | 11.83 | 85.00 | 97.96 | 81.70 | 95.91 | 82.01 | 78.54 |
| **1015** | 81.34 | 94.52 | 81.55 | 92.90 | 94.47 | 13.40 | 93.44 | 96.76 | 78.30 | 6.69 | 63.43 | 12.66 | 96.60 | 81.39 | 97.96 | 82.84 | 94.87 | 89.16 |
| **1016** | 95.37 | 80.83 | 95.17 | 89.18 | 81.02 | 12.24 | 80.23 | 82.53 | 87.45 | 6.68 | 62.06 | 12.18 | 82.55 | 93.91 | 82.11 | 98.29 | 81.18 | 76.46 |
| **1019** | 80.97 | 97.48 | 81.13 | 93.71 | 97.58 | 13.33 | 94.73 | 95.39 | 76.04 | 7.19 | 58.93 | 13.32 | 96.92 | 81.48 | 96.38 | 81.25 | 98.18 | 88.65 |

**Table 2. Pairwise alignment between strains Reads (lines) against each of the assemblies (columns).** The numbers correspond to percentages and reflect the overall alignment rate of the reads to the genome assembly. The groups of patterns that can be observed are marked in different colours that correlate with the different ranges of alignment rates to the reference genome CBS767 (last column). The closest group to the reference (dark grey) would be formed by NCYC8 (1002), NCYC103 (1005), NCYC475 (1007), NCYC792 (1008), CBS1101 (1013), J63 (1015), DBH9 (1019). The next group (light grey) would have the strains NCYC2572 (1001) together with NCYC9 (1003), CBS1792 (1014) and J26 (1016). The strain NCYC10 (1004) is not clear is it would be in the first group or in the second by only looking at the alignment (white). NCYC3045 (1009) and CBS117 (1011) are the diploid strains, but with distinct patterns to each other (orange and green). The strains NCYC459 (1006) and CBS5140 (1012) present the same range of alignment rates (light blue), and NCYC3364 (1010) belongs to another individual group (darker blue). Data obtained using Bowtie2.

## Species classification

A customized database was created for kraken (Wood and Salzberg, 2014) including all published reference sequences of fungi, plus the bacteria, virus, archaea and human sequences that were already used for the standard contamination check.

The numbers vary depending if it is done on the assemblies (Table 3) or on the raw reads (Annex table 4), as the amount of sequences and their length is different, and kraken only outputs the best match for every contig or read. In most cases the highest match is *D. hansenii*, even for some of the strains not considered to be *D. hansenii*. This is due to the fact that *D. hansenii* is the closest taxonomic group present in the reference database, and that is an important part of the analysis carried out by kraken. All of them presented matches with other fungi also, that can potentially give an idea of the amount of genetic exchange between species. A more detailed view of all the results, with the exact species names is available here: https://goo.gl/CMNe19. The species more distant from *D. hansenii* present a high percentage of their contigs as unclassified, meaning that no reference sequence could match. Note: this analysis was run before the genome reference of *D. fabryi* was published - And that *D. fabryi* would probably become the first match of the unidentified species.

| | Fungi | Bacteria | Viruses | Archaea | Others | Unclassified | Deha | Sc |
|---|---|---|---|---|---|---|---|---|
| **1001** | 80.36 | 0.48 | - | - | - | 18.57 | 74.29 | - |
| **1002** | 89.18 | 0.25 | - | - | 0.37 | 8.58 | 81.59 | - |
| **1003** | 70.17 | 1.26 | 0.14 | - | - | 27.59 | 64.85 | - |
| **1004** | 96.37 | 0.25 | 0.03 | - | 0.01 | 3.22 | 95.83 | - |
| **1005** | 45.51 | 6.14 | 0.43 | - | 0.05 | 47.34 | 41.59 | - |
| **1006** | 30.19 | 4.22 | 0.51 | - | - | 63.24 | 25.80 | - |
| **1007** | 79.24 | 1.69 | - | - | 0.21 | 18.23 | 73.61 | - |
| **1008** | 88.96 | 0.39 | - | - | 0.10 | 9.28 | 79.20 | - |
| **1009** | 96.60 | - | - | - | - | 3.40 | 95.43 | - |
| **1010** | 73.48 | 0.51 | - | - | 0.76 | 25.00 | 66.41 | - |
| **1011** | 95.72 | - | - | 0.02 | 0.07 | 4.19 | 95.19 | 00.02 |
| **1012** | 40.53 | 0.88 | - | - | 0.22 | 55.95 | 33.48 | 00.22 |
| **1013** | 61.90 | 3.78 | 0.40 | - | 0.07 | 32.84 | 55.02 | - |
| **1014** | 72.41 | 0.33 | - | - | 0.11 | 26.05 | 66.67 | - |
| **1015** | 88.51 | 0.26 | - | - | 0.35 | 9.74 | 77.82 | - |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **1016** | 73.44 | 0.69 | - | - | 0.28 | 24.62 | 65.15 | - |
| **1017** | 98.91 | 0.10 | - | - | - | 0.99 | 00.00 | 98.12 |
| **1018** | 97.89 | - | - | - | - | 2.11 | 00.00 | 97.51 |
| **1019** | 65.59 | 3.03 | 0.20 | - | - | 30.11 | 61.00 | 00.20 |

**Table 3. Classification of contigs made by kraken using all refSeq available of Fungi, Bacteria, Archaea, Virus, plasmids and human.** The numbers represent percentages of the number of contigs matching that classification from the total number of contigs. Deha: abbreviation for *D. hansenii*. Sc: abbreviation for *S. cerevisiae*. Translation of strain numbers: NCYC2572 (1001), NCYC8 (1002), NCYC9 (1003), NCYC10 (1004), NCYC103 (1005), NCYC459 (1006), NCYC475 (1007), NCYC792 (1008), NCYC3045 (1009), NCYC3364 (1010), CBS117 (1011), CBS5140 (1012), CBS1101 (1013), CBS1792 (1014), J63 (1015), J26 (1016), DBH9 (1019).

## Variant calling

In order to better understand the differences between the strains closer to *D. hansenii*, SNP calling was performed in those twelve organisms, comparing them to the reference genome CBS767. Big differences were found between them. Four strains presented a number of variants around 280.000 or higher, NCYC2572 (1001), NCYC9 (1003), CBS1792 (1014) and J26 (1016). The strain NCYC10 (1004) had around 50.000 variants compared to the reference. In four other strains between 10.000 and 12.000 variants were found, NCYC475 (1007), NCYC792 (1008), CBS1101 (1013) and J63 (1015). Furthermore, three strains had less than 1.000 variants, NCYC8 (1002), NCYC103 (1005) and DBH9 (1019), being in this last one 332, the smallest amount.

An interesting pattern was observed when looking at how the variants were spread over the chromosomes. On the strain with less variants, DBH9 (1019) the variants were located mostly on the telomeres of the chromosomes. The reason might be that these are highly repetitive sequences and the sequencing process does not have the same accuracy than in other regions of the chromosome. On the other hand, the strain NCYC2572 (1001), from the group with the highest variation, presents its variants all over the chromosome, with no clear distinction by regions. In this case there is no technical issue, but biological, it is clearly a highly different genome from the reference. An example of this for the chromosome CR382134.2 can be seen on figure 2. To go through all the chromosomes and the different strains: https://goo.gl/D224Em.

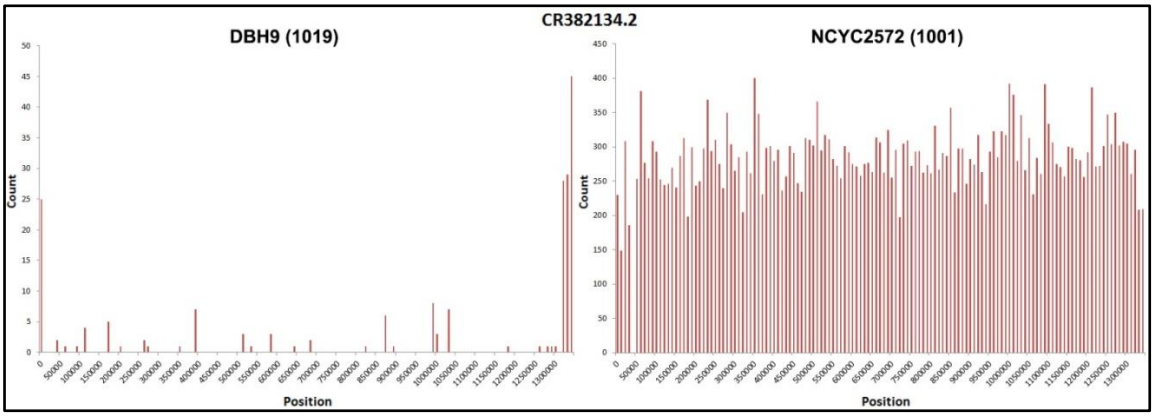



**Figure 2.** Comparison of the number of variants in each position of the Chromosome CR382134.2 for the strains DBH9 (1019) (left) and NCYC2572 (1001) (right). The Y axis represents the count of variants and the X axis the position in the chromosome in number of bases.

## Annotation

Genome annotation was performed on the assemblies and a number of transcripts were produced. For strains closer to *D. hansenii* only the transcriptome of the reference CBS767 was used as a guide for the program, but for the strains that remain unclassified *D. hansenii* and other Candida transcriptomes were used. Most of the results correlate with the number of transcripts for the reference genome CBS767, being around 6000. As a test, the analysis was also performed on the published reference genome CBS767 and only differed from the number of published transcripts by 36 transcripts.

| A | CBS767 | 1001 | 1002 | 1003 | 1004 | 1005 | 1007 | 1008 | 1013 | 1014 | 1015 | 1016 | 1019 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Maker** | 6326 | 6064 | 6401 | 6097 | 7171 | 6327 | 6219 | 6326 | 6421 | 6096 | 6375 | 6153 | 6380 |
| **Ref** | 6290 | | | | | | | | | | | | 5313 |

| B | 1006 | 1009 | 1010 | 1011 | 1012 |
|---|---|---|---|---|---|
| **D. hansenii** | 5934 | 10752 | 6099 | 10343 | 5830 |
| **D. hansenii + candida ssp.** | 5924 | 10769 | 6113 | 10455 | 5823 |

**Table 4. Number of transcripts predicted by Maker.** On the table A only the strains closer to *D. hansenii* are represented, including the reference genome itself (CBS767). On table B unclassified strains are shown, in which two annotations were made, using only *D. hansenii* transcriptome as guide, and *D. hansenii* + published candida transcriptomes.

To study more profoundly the strains considered to be *D. hansenii*, the predicted gene models from the annotation were clustered. There were a total of 2505 clusters with the 12 strains represented. Finally, the 100 longest clusters will be, in further research, selected among those to run a different phylogeny on every of those and obtain a consensus tree at the end.

## Phylogeny

An initial attempt to create a phylogeny from the SNP database using the program PhyloSNP (Faisona *et al*., 2014) failed. Instead, two alternative approaches were taken:

## ACT1 phylogeny

The first approach implied working with the well-studied gene *ACT1*, used in previous studies of *D.hansenii* and other *Debaryomyces* species (Jacques *et al.*, 2009 and Jacques *et al.*, 2015). The second exon of the transcript is known to be a good sequence to distinguish between the different species of *Debaryomyces* and, furthermore the intron sequence between the exon 1 and exon 2 it is ideal for identification of different types of *D. hansenii* inside the species.

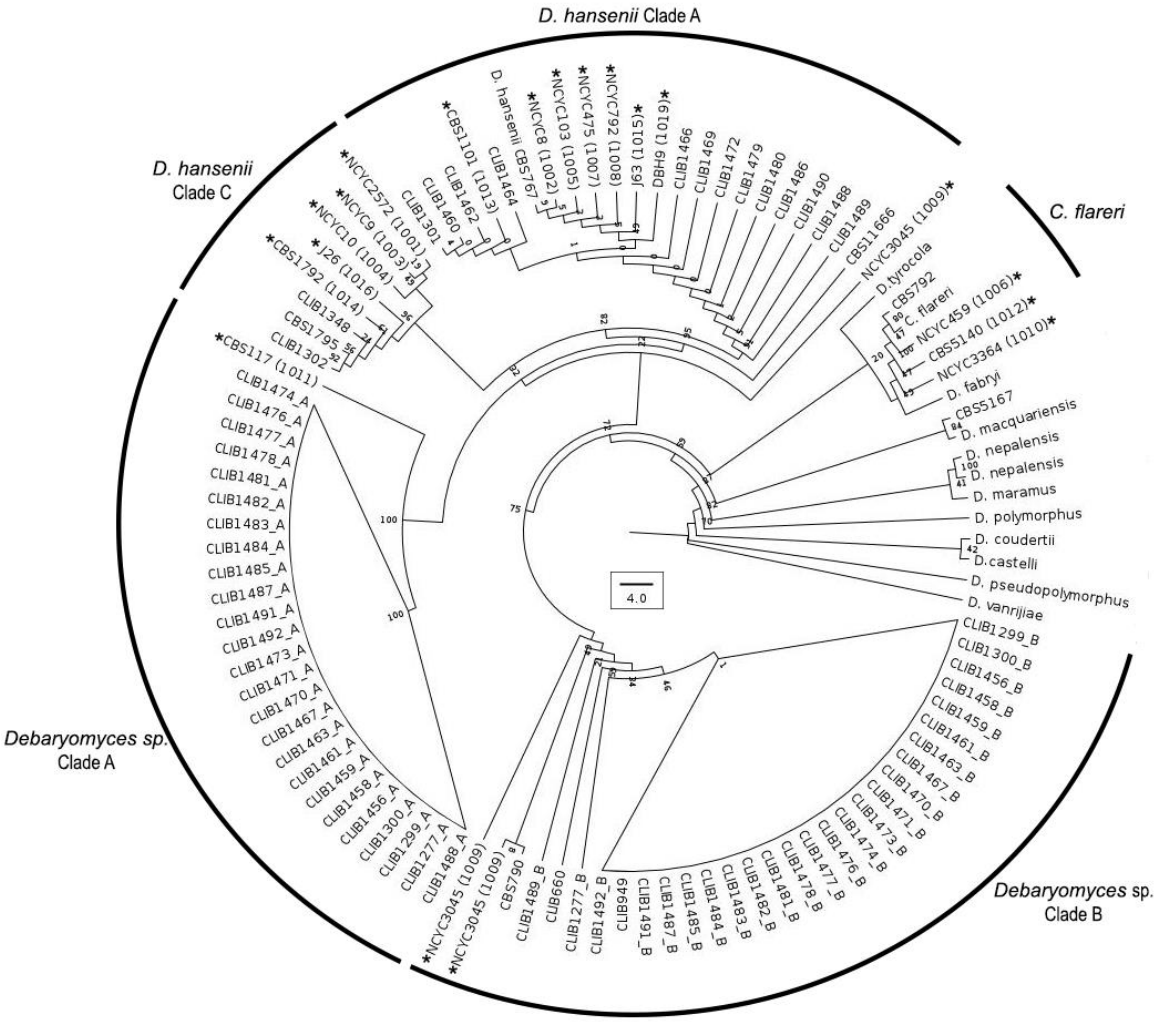



**Figure 3. Gene tree of the exon 2 of *ACT1* from *Debaryomyces* isolates.** Marked sequences (*) are the ones from this study. The rest of sequences come from the research of Jacques *et al*., 2009 and Jacques *et al*., 2015. Bootstrap values (%) based on 100 replicates are indicated on the nodes for main groups and clades.

The phylogeny for exon 2 of the *ACT1* gene from different species show clear clades already described in Jacques *et al.*, 2009 and Jacques *et al.*, 2015 (Fig. 3). From the glacier strains there is a clade A and a clade B, in most of the cases belonging to the two alleles found in diploid strains. The strain CBS117 (1011) seems to group with the clade A of the unknown *Debaryomyces* sp. and two replicates of the gene in NCYC3045 (1009) are grouped together with the clade B, while another of its replicates seems closer to the main *D. hansenii* clade. As said in the introduction, in two publications (Jacques *et al*., 2009 and Jacques *et al*., 2010) three clades were described inside *D. hansenii*. The strains NCYC8 (1002), NCYC103 (1005), NCYC475 (1007), NCYC792 (1008), CBS1101 (1013), J63 (1015) and DBH9 (1019) were grouped in the Clade A of *D. hansenii*, together with the reference CBS767. On the other side, the strains NCYC2572 (1001), NCYC9 (1003), NCYC10 (1004), CBS1792 (1014) and J26 (1016) would be considered to be in the cluster C of *Debaryomyces hansenii*. The unclassified strains NCYC459 (1006) and CBS5140 (1012) are clearly grouped with *C. flareri*. These two strains are also, as expected, close to *D. fabryi*, the same as NCYC3364 (1010), but this last one still remains unclassified.

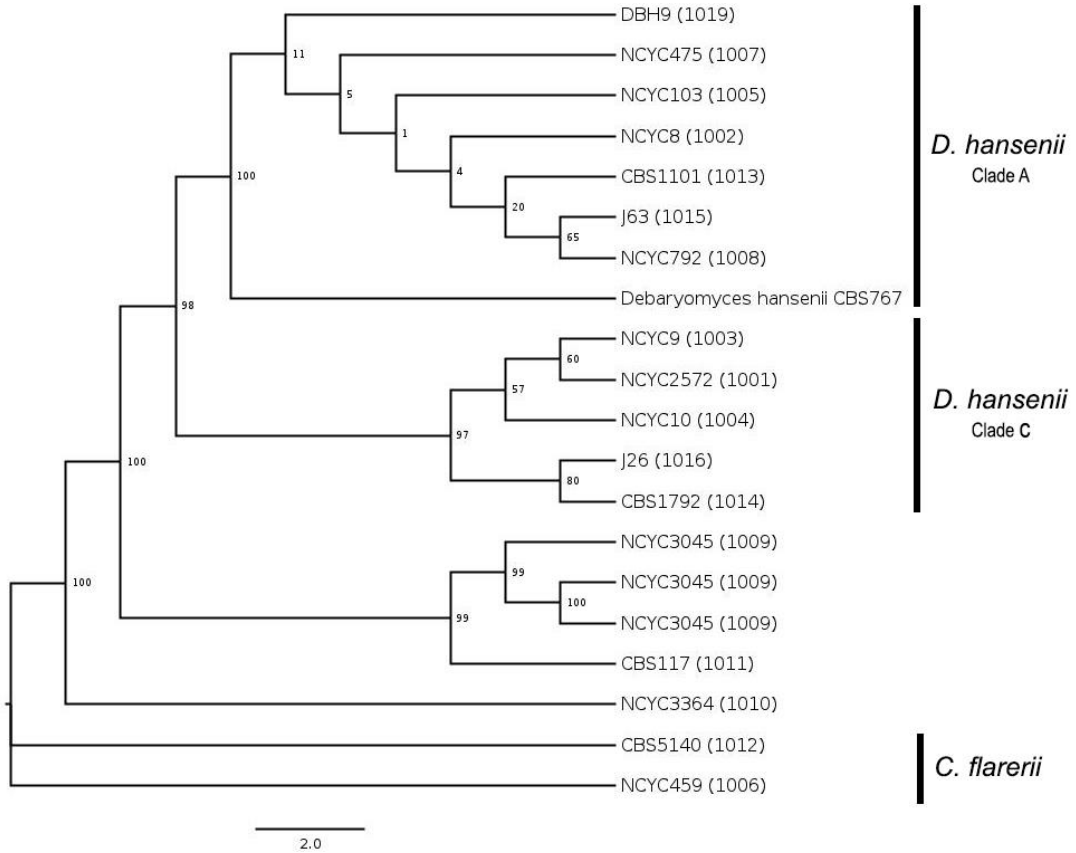



**Figure 4. Gene tree of the complete sequence of ACT1** from the strains, including the intron sequence for differentiation. Bootstrap values (%) based on 100 replicates are indicated on the nodes for main groups and clades.

The phylogeny for the complete sequence of *ACT1* allowed a more detailed view of the 17 strains from the current study (Fig. 4). Again, the two clades of *D. hansenii* were clearly discernible, being NCYC8 (1002), NCYC103 (1005), NCYC475 (1007), NCYC792 (1008), CBS1101 (1013), J63 (1015) and DBH9 (1019) in clade A and NCYC2572 (1001), NCYC9 (1003), NCYC10 (1004), CBS1792 (1014) and J26 (1016) in clade C. The clade formed by CBS5140 (1012) and NCYC459 (1006) was identified as *C. flareri* supported by the previous phylogeny (Fig. 3). NCYC3045 (1009) and CBS117 (1011) are in the same clade, being the two only diploid strains, but still unclassified. The same happens with NCYC3364 (1010), it is a different species but remains unknown.

**GPD1 phylogeny**

The *DhGPD1* is one of the genes that have been studied more in detail in *D. hansenii*. It codes for the protein GPD1 (glycerol 3-phosphate dehydrogenase NAD+), important because of its relation to the ability of the specie to survive in high concentrations of salt by producing glycerol inside the cell (Thomé, 2004; Thomé, 2005). A phylogeny of this gene has been done on the strains of the current research.

It was possible to find only one coding sequence for *DhGPD1* in each of the strains, as expected from previous research (Thomé, 2004), except for the diploid strain CBS117 (1011), in which two copies of the gene were found in its genome.

The phylogenetic tree on figure 5 shows again two big clades of *D. hansenii* strains, being the clade A the one with the reference CBS767. The strains NCYC103 (1005), NCYC475 (1007), NCYC792 (1008), CBS1101 (1013), NCYC8 (1002) and DBH9 (1019) had the exact same sequence as CBS767, and surprisingly also NCYC10 (1004), that from the ACT1 analysis it was grouped on the clade C. The strain J63 (1015) presented a SNP in its sequence that resulted in an aminoacid change (Fig. 7). The strains NCYC2572 (1001), NCYC9 (1003), CBS1792 (1014) and J26 (1016) were grouped in clade C and they presented 25 SNPs, most of them synonymous mutations (Fig. 7). The clade formed by CBS5140 (1012) and NCYC459 (1006) grouped together again as seen previously (Fig. 3 and 4) and NCYC3364 (1010) was individually separated from the rest. NCYC3045 (1009) and one of the alleles of CBS117 (1011) grouped together, but the second allele was differentiated from them.

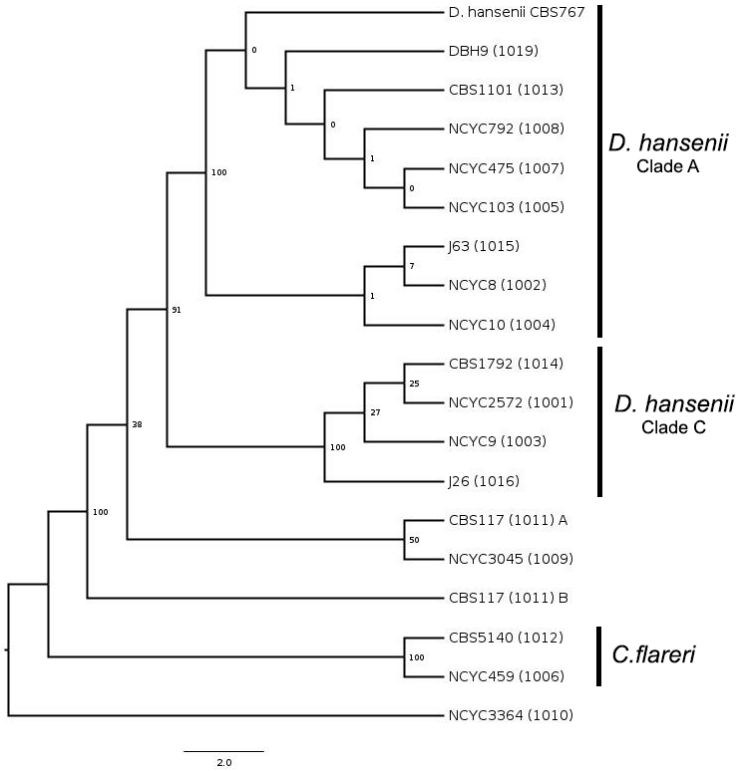



**Figure 5. Gene tree for *DhGPD1*.** Bootstrap values (%) based on 100 replicates are indicated on the nodes for main groups and clades.

# Discussion

The definition of species among micro-organisms is difficult even with the present amounts of sequence data (Jacques *et al.*, 2009). DNA sequence analysis has transformed the way in which yeasts are identified (Kurtzman *et al.* 2011). This is clearly illustrated in the present study by just running a first test mapping of the reads obtained in the sequencing to the reference genome. The results show clearly different species, not previously differentiated when using morphological characters and physiology to distinguish between the strains (Table 1).

## Non-*D. hansenii* strains

Three strains were clearly not *D. hansenii*, NCYC459 (1006), NCYC3364 (1010) and CBS5140 (1012). Through pairwise comparison it is seen that NCYC459 (1006) and CBS5140 (1012) are the same species, also in the phylogenies cluster together. They are strains coming from different parts of the globe and different environments, NCYC459 (1006) was isolated from soil in New Zealand by M.E. di Menna, and CBS5140 (1012) was picked from the skin of a man in Hungary by J. Galgóczy. This could suggest that it may be a species with high plasticity to different kinds of environment, as *D. hansenii*

is. In the research of Jacques *et al*., 2009, CBS5140 (2012) was reclassified as *C. flareri*, but in the yeast collection continues to be labelled as *D. hansenii*.

The strain NCYC3364 (1010) could not be classified. It is close to *D. fabryi* and *C. flareri*, but it is clearly a different species.

These cases are a clear example of the difficulties when working with some yeast species and support the need to add more information to the databases. Genome references from these strains could be generated from PacBio data, which would improve the coverage of the reads and a good assembly could be performed. On one side, we would have the genome reference of *C. flareri*, and additionally, a new species to be named.

## Diploid strains

NCYC3045 (1009) and CBS117 (1011) are clearly diploid *Debaryomyces* species because of their estimated genome size, doubled from the usual *D. hansenii* genome size (Annex Fig. 1). The diploidy could have been caused by hybridization with another species. These diploid strains still maintain a high percentage of their genome close to the *D. hansenii* genome, and from the results it was still impossible to elucidate the origin of the rest of their genomes. What was clear is that they are not the same hybrids, both strains share that a part of their genome comes from *D. hansenii*, but the rest is clearly different.

In the phylogeny using ACT1, three ACT1 representatives were found in NCYC3045 (1009). Two of the alleles were grouped together with the *Debaryomyces* unknown species clade B described on Jacques *et al*., 2015, isolated from artic glaciers. The third allele was grouped close to *D. hansenii* strains. On the strain CBS117 (1011) only one ACT1 was found and this one was grouped with the *Debaryomyces* unknown species clade A described in the mentioned research with arctic glacier isolates.

## *D. hansenii* strains

### Reference

NCYC2572 (1001) was considered the reference for this project when it started. It was stored in a different collection but it is supposed to be the same as CBS767. All the data suggests now that it is not CBS767 and it is not even as close to it as the other strains

are. The alignment results showed that it didn't map ideally to CBS767, and the high number of SNPS from the variant calling analysis and, in combination with phylogeny, made sure that it is distantly related to the reference strain. Instead, DBH9 (1019), is the closest strain to CBS767, as it was originally the reference strain itself, but a point mutation was performed on it (Minhas *et al*., 2012). DBH9 (1019) would be designated as the main reference for further research on these strains of *D. hansenii*. Furthermore, the strains NCYC2572 (1001) and CBS767 would be taken again from their respective collections and re-sequenced to reassure the findings and initiate the reclassification on the collection.

***D. hansenii* classification**

From the mapping of the strains to the reference genome CBS767 it was already possible to notice a differentiation between strains. One of the groups presented an overall alignment rate of the reads lower than 80%, while the other group presented levels higher than 87%. Those groups matched exactly with the different patterns seen when running the pairwise alignment between strains.

After the variant calling analysis the groups were more remarkably differentiated. One of the groups was closer to *D. hansenii* CBS767, the one represented by DBH9 (1019) that presented less than 1000 variants in most of the cases. The other group had extremely high levels of variants (>260.000), represented in this case by the other initial reference NCYC2572 (1001). In these analyses the only strain that remained uncertain to be put in any of the groups was NCYC10 (1004).

The phylogeny using the intron sequence of the ACT1 gene allowed the classification of *D. hansenii* into three clades inside the specie (Jacques *et al*., 2009). The clade A, with the reference strain CBS767; a clade B for which not many representatives have been found; and a clade C, in which is included *C. famata* var. *famata*, the anamorph of *D. hansenii*. The phylogeny using the complete sequence of the gene made possible the classification of the strains of this research. The two groups were clearly differentiated as in previous results. The strains NCYC8 (1002), NCYC103 (1005), NCYC475 (1007), NCYC792 (1008), CBS1101 (1013), J63 (1015) and DBH9 (1019) were grouped in clade A. The strains NCYC2572 (1001), NCYC9 (1003), NCYC10 (1004) and J26 (1016) in clade C. In this case, the strain NCYC10 (1004) was grouped together with the strains of clade C. The last clade would then be considered *C. famata*, but the

International Botanical Congress in Melbourne in July 2011 made a change in the International Code of Nomenclature for algae, fungi, and plants and adopted the principle "one fungus, one name" (McNeil *et* al., 2012); the system of permitting separate names to be used for anamorphs then ended (Cannon and Kirk, 2000). This means that all legitimate names proposed for a species, regardless of what stage they are typified by, can serve as the correct name for that species. In result, we would maintain the name of *D. hansenii* to this clade.

An important question that comes up from the results is if the strains of the clade C should still be considered *D. hansenii* taking into account the high levels of variance that the clade C has in reference to strains in the clade A. Until the present, all the investigations have used single gene techniques to differentiate them (Jacques *et al*., 2009; Jacques *et al*., 2010; Jacques *et al*., 2015; Wrent *et al*., 2015). Now that the complete genome has been compared, the differences are much more pronounced than expected.

**J63 and J26**

The strains J63 (1015) and J26 (1016), important for the future of the ongoing research, were classified into different clades. J63 (1015) is in the clade A, closer to CBS767, and J26 (1016) in the *D. hansenii* clade C. Both strains were isolated from the same spot, which indicates the importance for the study of populations from specific areas and how variation is maintained in populations.

## Phenomics

A phenomic study of the strains was performed, parallel to the present genomic study, by Omairah Ashraf. It tested generation times for the strains in different concentrations of salt (Fig. 6). In this study no differentiation of the *D. hansenii* strains could be observed. In all strains, from clade A and C, the generation time decreased as the level of salt increased. This indicated that *D. hansenii* strains are halophilic in addition to halotolerant, only in the highest concentrations of salt their growth was compromised, but still they were growing at a fast rate.

The two diploid strains were also halophilic, presenting a very similar pattern as the *D. hansenii* strains. This suggests that probably they share with *D. hansenii* the genes that confers them halophilic phenotype.

The *C. flareri* strains had a low generation time in the different concentrations of salt, but growth was faster without salt in the growth media. In this case these are halotolerant strains, but not halophilic. The same pattern is observed in the unknown specie NCYC3364 (1010).

*S. cerevisiae* strains were used as a control, to compare the previous results to not halophilic nor high halotolerant strain.

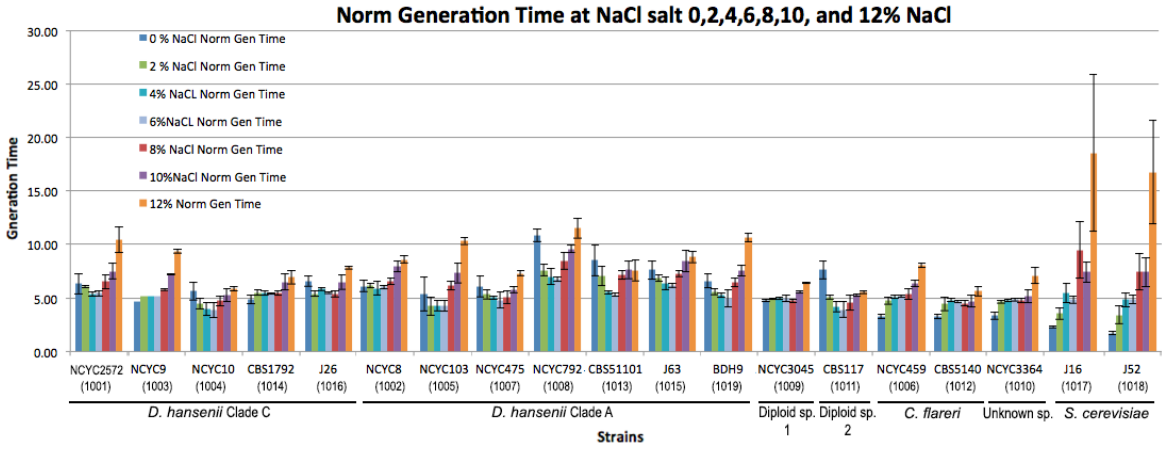



**Figure 6. Generation time at different concentrations of NaCl for each of the strains.** The strains have been ordered according to the findings in the results of this research to visualize better if differentiations between the different groups of strains are found. Original graphic made by Omairah Ashraf to show the results of the phenomic study.

## GPD1

The complete mechanism of *D. hansenii* to handle high concentrations of salt and even presenting an halophilic phenotype is not perfectly known still, but the importance of the enzyme GPD1 is significant. It synthesizes Glycerol (Thomé, 2004; Thomé, 2005) and the retention of this solute inside the cell confers *D. hansenii* the ability to tolerate a saline media (Gustafsson and Norkrans, 1976). The study of this transcript and the pathway associated could help understanding the process of salt-tolerance on eukaryotic cells. The protein GPD1 has two main domains, NAD-dependent glycerol-3-phosphate dehydrogenase N-terminus and NAD-dependent glycerol-3-phosphate dehydrogenase C-terminus (pfam.xfam.org (Pfam) – Last visited: 2016-05-23). The mutations observed of different aminoacids of the domains in some of the strains could be an explanation for their different phenotypes to variable concentrations of salt (Fig. 7). We can assure that none of them loose the tolerance to salt so the observed mutations might not be affecting that aspect of the GPD1 performance. We could see a division between

halophilic strains and only halotolerant strains, and it could be investigated further if the mutations are provoking the different phenotypes to be noticeable.

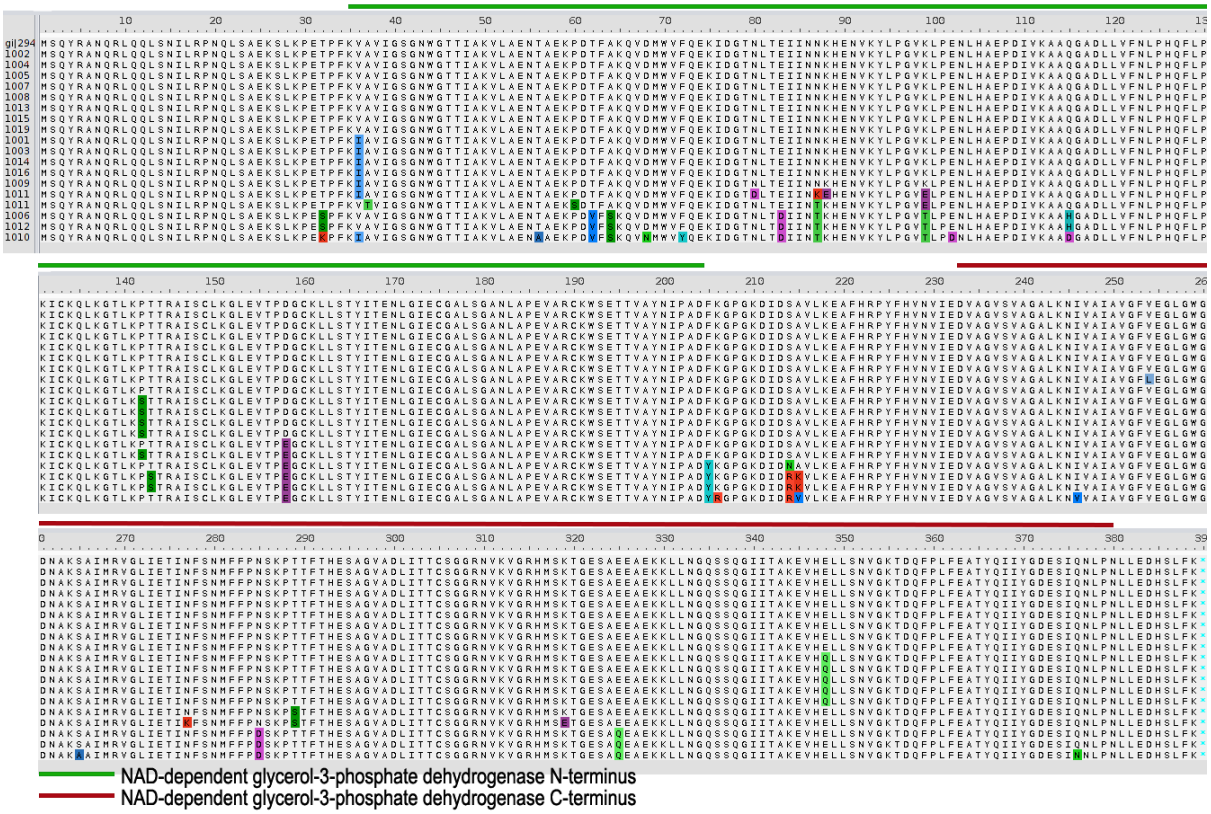



**Figure 7. Alignment of the protein sequence of GPD1** on Aliview with the unmatched aminoacids remarked. With green and red both domains are identified. Order of the strains: CBS767, NCYC8 (1002), NCYC10 (1004), NCYC103 (1005), NCYC475 (1007), NCYC792 (1008), CBS1101 (1013), J63 (1015), DBH9 (1019), NCYC2572 (1001), NCYC9 (1003), CBS1792 (1014), J26 (1016), NCYC3045 (1009), CBS117 (1011), NCYC459 (1006), CBS5140 (1012) and NCYC3364 (1010).

# Conclusions

From the 17 strains classified as *D. hansenii* on the collections:

**12 strains were considered *D. hansenii* and divided into two clades.**

- Clade A formed by NCYC8 (1002), NCYC103 (1005), NCYC475 (1007), NCYC792 (1008), CBS1101 (1013), J63 (1015) and DBH9 (1019), together with the reference strain CBS767.
- Clade C formed by NCYC2572 (1001), NCYC9 (1003), CBS1792 (1014) and J26 (1016).

- The strain NCYC10 (1004) was not possible to be located undoubtedly in any of the groups.

**Two strains were classified as *C. flareri*, NCYC459 (1006) and CBS5140 (1012).**

**Three species remain unknown.**

- The strain NCYC3364 (1010) is close related to *C. flareri* and *D. fabryi*, but still unidentified.
- The strains NCYC3045 (1009) and CBS117 (1011) are two different diploid species, with part of its genome from *D. hansenii*.

Reference strains

**DBH9 (1019) is a better reference for future studies than NCYC2572 (1001) because of its closeness to the reference genome CBS767.** NCYC 2572 (1001) belongs to a different clade of *D. hansenii* than the reference genome CBS767, contrary to what is specified on the collection description.

# Future perspectives

The strains NCYC 2572 (1001) and CBS767 will be taken from its respective collections and sequenced in order to reclassify them as different strains.

The strains NCYC459 (1006) and CBS5140 (1012) will be re-sequenced using a PacBio approach to obtain the *C. flareri* reference genome.

The strain NCYC3364 (1010) will be re-sequenced using PacBio technologies in order to obtain a new close related *Debaryomyces* species reference genome.

The strains NCYC3045 (1009) and CBS117 (1011) could be re-sequenced to study more in detail their genome to be able to identify similarities and differences to the *D. hansenii* genome. The genes of the shared sequences could be studied to understand the mechanisms of halotolerance as they present a phenotype like *D. hansenii*.

As the strains J63 (1015) and J26 (1016) will be used for future phenotypic studies in the research group, a good reference assembly should be made by re-sequencing using PacBio techniques. At the same time, we would have a reference genome representing each of the *D. hansenii* clades.

The aim of obtaining the reference assemblies is headed to a deeper study of the genetics of *D. hansenii* in relation to its phenotype.

In a long time future perspective it is aimed for a wide population study of strains isolated from different locations of the Swedish coast.

# References


Adler, L., Blomberg, A., Nilsson, A. (1985). Glycerol metabolism and osmoregulation in salt-tolerant yeast *Debaryomyces hansenii*. *Journal of Bacteriology* 162, 300-306.

Adler, L. (1986). Physiological and Biochemical Characteristics of the yeast *Debaryomyces hansenii* in relation to salinity. *The Biology of Marine Fungi, Cambridge University Press* 9, 81-89.

André, L., Nilsson, A., Adler, L. (1988). The role of glycerol in osmotolerance of the yeast *Debaryomyces hansenii*. *Journal of General Microbiology* 134, 669-677.

Arroyo Gonzalez, N., Vazquez, A., Ortiz Zuazaga, O., Sen, A., Luna Olvera, H., Peña de Ortiz, S., Govind, N.S. (2009). Genome-wide expression profiling of the osmoadaptation response of *Debaryomyces hansenii*. *Yeast* 26, 111-124. doi: 10.1002/yea.1656.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* 19(5), 455–477. doi:10.1089/cmb.2012.0021.

Cannon, P.F., Kirk, P.M. (2000). The philosophies and practicalities of amalgamating anamorph and teleomorph concepts. *Studies in Mycology* 45, 19–25.

Chao, H., Yen, Y., Ku, M.S.B. (2009). Characterization of salt-induced *DhAHP*, a gene coding for alkyl hydroperoxide reductase, from the extremely halophilic yeast *Debaryomyces hansenii*. *BMC Microbiology* 9, 182. doi: 10.1186/1471-2180-9-182.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6(2), 80-92. doi: 10.4161/fly.19695.

Corredor, M., Davila, A.M., Gaillardin, C., Casaregola, S. (2000). DNA probes specific for the yeast species *Debaryomyces hansenii*: useful tools for rapid identification. *FEMS Microbiology Letters* 193, 171-177.

Darriba, D., Taboada, G.L., Doallo, R., Posada, D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* 9(8), 772.

Daniel, H.M., Meyer, W. (2003). Evaluation of ribosomal RNA and actin gene sequences for the identification of ascomycetous yeasts. *International Journal Food Microbiology* 86, 61-78. doi: 10.1016/S0168-1605(03)00248-4.

Desnos-Ollivier, M., Ragon, M., Robert, V., Raoux, D., Gantier, J-C., Dromer, F. (2008). Debaryomyces hansenii (Candida famata), a rare human fungal pathogen often misidentified as Pichia guilliermondii (Candida guilliermondii). *Journal of clinical microbiology* 46:10, 3237-3242. doi:10.1128/JCM.01451-08.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., de Montigny, J., Marck, C., Neuveglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J-M., Beyne, E., Bleykasten, C., Boisrame, A., Boyer, J., Cattolico, L., Confanioleri, F., de Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J-M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G-F., Straub, Marie-L., Suleau, A., Swennen, D., Tekaia, F., Wesolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P., Souciet, J-L. (2004). Genome evolution in yeasts. *Nature* 430, 35-44. doi: 10.1038/nature02579.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. 32(5), 1792-97. doi: 10.1093/nar/gkh340.

Faisona, W.J., Rostovtsevb, A., Castro-Nallarc, E., Crandallc, K.A., Chumakovb, K., Simonyanb, V., Mazumadera, R. (2014). Whole genome single-nucleotide variation profile-based phylogenetic tree building methods for analysis of viral, bacterial and human genomes. *Genomics* 104, 1-7. doi:10.1016/j.ygeno.2014.06.001.

Fitzpatrick, D.A., Logue, M.E., Stajich, J.E., Butler, G. (2006). A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evolutionary Biology* 6:99. doi: 10.1186/1471-2148-6-99.

Glez-Peña, D., Gómez Blanco, D., Reboiro-Jato, M., Fdez-Riverola, F., Posada, D. (2010) ALTER: program-oriented format conversion of DNA and protein alignments. Nucleic Acids Research. Web server issue. 0305-1048. doi:10.1093/nar/gkq321.

Groenewald N., Daniel H.-M., Robert V., Poot G.A., Smith M.T. (2008). Polyphasic re-examination of *Debaryomyces hansenii* strains and reinstatement of *D. hansenii*, *D. fabryi* and *D. subglobosus. Persoonia* 21, 17-27. doi: 10.3767/003158508X336576.

Guindon, S. and Gascuel, O. (2003). A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Systematic Biology* 52: 696-704.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3):307-21.

Gustafsson, L., Norkrans, B. (1976). On the mechanism of salt tolerance. Production of glycerol and heat during growth of *Debaryomyces hansenii*. *Archives of Microbiology* 110, 177-183.

Jacques, N., Casaregola, S. (2008). Safety assessment of dairy microorganisms: The hemiascomycetous yeasts. *International Journal of Food Microbiology* 126, 321-326. doi: 10.1016/j.ijfoodmicro.2007.08.020.

Jacques, N., Mallet, S., Casaregola, S. (2009). Delimitation of the species of the *Debaryomyces hansenii* complex by intron sequence analysis. *International Journal of Systematic and Evolutionary Microbiology* 59, 1242-1251. doi: 10.1099/ijs.0.004325-0.

Jacques, N., Sacerdot, C., Derkaoui, M., Dujon, B., Ozier-Kalogeropoulos, O., Casaregola, S. (2010). Population polymorphism of Nuclear Mitochondrial DNA insertions reveals widespread diploidy associated with loss of heterozygosity in *Debaryomyces hansenii*. *Eukaryotic Cell* 9, 449-459. doi: 10.1128/EC.00263-09.

Jacques, N., Zenouche, A., Gunde-Cinerman, N., Casaregola, S. (2015). Increased diversity in the genus *Debaryomyces* from Artic glacier samples. *Antonie van Leeuwenhoek* 107, 487-501. doi:10.1007/s10482-014-0345-7.

Kumar, S., Randhawa, A., Ganesan, K., Raghava, G.P.S., Mondal, A.K. (2012). Draft Genome Sequence of Salt-Tolerant Yeast Debaryomyces hansenii var. Hansenii MTCC 234. *Eukaryotic Cell* 11:7, 961-962. doi: 10.1128/EC.00137-12.

Kurtzman, C.P, Fell, W.F., Boekhout, T. (2011). Gene sequence analysis and other DNA-based methods for yeast species recognition. *The yeasts, a taxonomy study* 1, 10: 137-144.

Langmead, B., Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9, 357-359. doi: 10.1038/nmeth.1923.

Li, W., Godzik, W. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-9. doi: 10.1093/bioinformatics/btl158.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T-W., Wang, J. (2002). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1:18. doi: 10.1186/2047-217X-1-18.

Martorell, P., Fernandez-Espinar, M.T., Querol, A. (2005). Sequence-based identification of species belonging to the genus *Debaryomyces*. *FEMS Yeast Research* 5, 1157-1165. doi: 10.1016/j.femsyr.2005.05.002.

McNeill, J., Barrie, F.R., Buck, W.R., Demoulin, V., Greuter, W., Hawksworth, D.L., Herendeen, P.S., Knapp, S., Marhold, K.., Prado, J., Prud'homme Van Reine, W.F., Smith, G.F., Wiersema, J.H., Turland, N.J. (2012). Preface, International Code of Nomenclature for algae, fungi, and plants (Melbourne Code) adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011, Regnum Vegetabile 154, A.R.G. Gantner Verlag KG, ISBN 978-3-87429-425-6.

Minhas, A., Sharma, A., Kaur, H., Rawal, Y., Ganesan, K., Mondal, A. K. (2012). Conserved Ser/Arg-rich Motif in PPZ Orthologs from Fungi Is Important for Its Role in Cation Tolerance. *The Journal of Biological Chemistry*, 287(10), 7301–7312. doi:10.1074/jbc.M111.299438.

Moura, G.R., Lousado, J.P., Pinheiro, M., Carreto, L., Silva, R.M., Oliveira, J.L., Santos, M.AS. (2007). Codon-triplet context unveils unique features of the *Candida albicans* protein coding genome. *BMC Genomics* 8, 444. doi: 10.1186/1471-2164-8-444.

Nakase, T., Suzuki, M. (1985). Taxonomic studies on *Debaryomyces hansenii* (Zopf) Lodder et Kreger-Van Rij and related species. II. Practical discrimination and nomenclature. *Journal of General Applied Microbiology* 31, 71-86.

Neves, M.L., Oliveira, R.P., Lucas, C.M. (1997). Metabolic flux response to salt-induced stress in the halotolerant yeast *Debaryomyces hansenii*. *Microbiology* 143, 1133-1139.

Norkrans, B. (1966). Studies on marine occurring yeasts: Growth related to pH, NaCl concentration and temperature. *Archiv für Mikrobiologie* 54, 374-392.

Prillinger, H., Molnar, O., Eliskases-Lechner, F., Lopandic, K. (1999). Phenotypic and genotypic identification of yeasts from cheese. *Antonie van Leeuwenhoek* 75, 267-283. doi: 10.1023/A:1001889917533.

Prista, C., Almagro, A., Loureiro-Dias, M.C., Ramos, J. (1997). Physiological basis for the high salt tolerance of Debaryomyces hansenii. *Applied and Environmental Microbiology* 63, 4005-4009.

Prista, C., Loureiro-Dias, M.C., Montiel, V., García, R., Ramos, J. (2005). Mechanisms underlying the halotolerant way of *Debaryomyces hansenii*. *FEMS Yeast Reasearch* 5, 693-701. doi: 10.1016/j.femsyr.2004.12.009.

Quiros, M., Martorell, P., Valderrama, M.J., Querol, A., Peinado, J.M., de Siloniz, M.I. (2006). PCR-RFLP analysis of the IGS region of rDNA: a useful tool for the practical discrimination between species of the genus *Debaryomyces*. *Antonie van Leeuwenhoek* 90, 211-219. doi: 10.1007/s10482-006-9076-8.

Romero, P., Patino, B., Quiros, M., Gonzalez-Jaen, M.T., Valderrama, M.J., de Siloniz, M.I., Peinado, J.M. (2005). Differential detection of *Debaryomyces hansenii* isolated from intermediate-moisture foods by PCR-RFLP of the IGS region of rDNA. *FEMS Yeast Research* 5, 455-461. doi: 10.1016/j.femsyr.2004.09.002.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research* 19: 1117-1123. doi: 10.1101/gr.089532.108.

Tafer, H., Sterflinger, K., Lopandi, K. (2016). Draft genome of Debaryomyces fabryi CBS 789T, isolated from a human interdigital mycotic lesion. *Genome Announcements* 4(1), e01580-15. doi:10.1128/genomeA.01580-15.

Thomé, P.E., Trench, R.K. (1999). Osmoregulation and the genetic induction of glycerol-3-phosphate dehydrogenase by NaCl in the euryhaline yeast *Debaryomyces hansenii*. *Marine Biotechnology* 1, 230-238. doi: 10.1007/PL00011772.

Thomé, P.E. (2004). Isolation of a GPD gene from *Debaryomyces hansenii* encoding a glycerol 3-phosphate dehydrogenase NAD+. *Yeast.* 21, 119-126. doi: 10.1002/yea.1070.

Thomé, P.E. (2005). Heterologous expression of glycerol 3-phosphate dehydrogenase gene DhGPD1 from the osmotolerant yeast *Debaryomyces hansenii* in *Saccharomyces cerevisiae*. *Current Microbiology* 51, 87-90. doi: 10.1007/s00284-005-4446-4.

Thomé, P.E. (2007). Cell wall involvement in the glycerol response to high osmolarity in the halotolerant yeast *Debaryomyces hansenii*. *Antonie Van Leeuwenhoek* 91, 229-235. doi: 10.1007/s10482-006-9112-8.

Tsui, C.K., Daniel, H.M., Robert, V., Meyer, W. (2008). Re-examining the phylogeny of clinically relevant *Candida* species and allied genera based on multigene analyses. *FEMS Yeast Research* 8, 651-659. doi: 10.1111/j.1567-1364.2007.00342.x.

Wood, D.E., Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 15, R46. doi: 10.1186/gb-2014-15-3-r46.

Wrent, P., Rivas, E-M., Gil de Prado, E., Peinado, J.M., de Silóniz, M-I. (2015). Development of species-specific primers for rapid identification of *Debaryomyces hansenii*. *International Journal of Food Microbiology* 193, 109-113. doi: 10.1016/j.ijfoodmicro.2014.10.011.

Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., Yorke, J.A. (2013). The MaSuRCA genome assembler. *Bioinformatics* 29 (21), 2669-2677. doi:10.1093/bioinformatics/btt476.

# Annex

### Genome size

| Strain | Genome size |
|---|---|
| 1001 | 12131848 |
| 1002 | 12231799 |
| 1003 | 12131016 |
| 1004 | 17653984 |
| 1005 | 12478626 |
| 1006 | 11829865 |
| 1007 | 12048106 |
| 1008 | 12269922 |
| 1009 | 22266991* |
| 1010 | 12693060 |
| 1011 | 20547860* |
| 1012 | 11600942 |
| 1013 | 12574314 |
| 1014 | 12145151 |
| 1015 | 12352049 |
| 1016 | 12346731 |
| 1017 | 12522328 |
| 1018 | 12537811 |
| 1019 | 12311012 |

**Table 1. Estimated genome sizes from the assemblies.** It has been marked (*) the sequences that presents a double sized genome. Translation of strain numbers: NCYC2572 (1001), NCYC8 (1002), NCYC9 (1003), NCYC10 (1004), NCYC103 (1005), NCYC459 (1006), NCYC475 (1007), NCYC792 (1008), NCYC3045 (1009), NCYC3364 (1010), CBS117 (1011), CBS5140 (1012), CBS1101 (1013), CBS1792 (1014), J63 (1015), J26 (1016), DBH9 (1019).

### Read coverage

| Strain | Coverage |
|---|---|
| 1001 | 49 X |
| 1002 | 24 X |
| 1003 | 46 X |
| 1004 | 27 X |
| 1005 | 41 X |
| 1006 | 40 X |
| 1007 | 48 X |
| 1008 | 26 X |
| 1009 | 21 X |
| 1010 | 25 X |
| 1011 | 8 X |
| 1012 | 46 X |
| 1013 | 36 X |
| 1014 | 47 X |
| 1015 | 41 X |
| 1016 | 53 X |
| 1017 | 37 X |
| 1018 | 51 X |
| 1019 | 38 X |

**Table 2. Read coverage for each of the strains.** Translation of strain numbers: NCYC2572 (1001), NCYC8 (1002), NCYC9 (1003), NCYC10 (1004), NCYC103 (1005), NCYC459 (1006), NCYC475 (1007), NCYC792 (1008), NCYC3045 (1009), NCYC3364 (1010), CBS117 (1011), CBS5140 (1012), CBS1101 (1013), CBS1792 (1014), J63 (1015), J26 (1016), DBH9 (1019).

**Alignment of reads to genome references**

| | *D. hansenii* | | *S. cerevisiae* | | *M. guilliermondii* | | *D. fabryi* | |
|---|---|---|---|---|---|---|---|---|
| **Strain/Libraries** | **AH** | **BC** | **AH** | **BC** | **AH** | **BC** | **AC** | **BC** |
| 1001 | 79.02 | 78.83 | 1.53 | 1.49 | 2.16 | 2.12 | 13.02 | 13.05 |
| 1002 | 90.26 | 89.92 | 2.25 | 2.20 | 3.22 | 3.20 | 14.07 | 14.04 |
| 1003 | 77.46 | 77.38 | 1.77 | 1.73 | 2.73 | 2.69 | | |
| 1004 | 85.48 | 85.85 | 1.17 | 1.14 | 2.05 | 2.05 | | |
| 1005 | 87.30 | 87.29 | 1.42 | 1.39 | 3.12 | 3.09 | | |
| 1006 | 8.68 | 8.76 | 1.74 | 1.70 | 2.54 | 2.51 | 38.44 | 38.51 |
| 1007 | 87.61 | 87.48 | 2.45 | 2.39 | 3.99 | 3.93 | | |
| 1008 | 89.23 | 89.00 | 2.61 | 2.55 | 3.61 | 3.54 | | |
| 1009 | 68.99 | 68.94 | 1.70 | 1.69 | 2.48 | 2.47 | 12.85 | 12.87 |
| 1010 | 5.71 | 5.73 | 2.06 | 1.96 | 2.88 | 2.77 | 22.33 | 22.30 |
| 1011 | 51.41 | 51.11 | 2.55 | 2.52 | 3.83 | 3.80 | 13.39 | 13.41 |
| 1012 | 8.98 | 8.96 | 2.12 | 2.08 | 2.94 | 2.91 | 39.79 | 39.65 |
| 1013 | 86.48 | 86.45 | 2.47 | 2.42 | 3.92 | 3.88 | | |
| 1014 | 78.13 | 77.92 | 1.65 | 1.62 | 2.35 | 2.33 | | |
| 1015 | 88.65 | 88.47 | 2.48 | 2.40 | 3.54 | 3.46 | | |
| 1016 | 75.98 | 75.87 | 2.10 | 2.07 | 3.06 | 3.03 | | |
| 1017 | 0.08 | 0.09 | 95.75 | 95.59 | 4.12 | 4.18 | | |
| 1018 | 0.08 | 0.08 | 96.10 | 95.91 | 2.81 | 2.80 | | |
| 1019 | 87.51 | 87.94 | 1.09 | 1.08 | 2.61 | 2.62 | | |

**Table 3. Bowtie2 alignments against the genomes of *D. hansenii, S. cerevisiae, M. guilliermondii* and *D. fabryi*.** The libraries names correspond to the identification codes when sequenced: BC47HDACXX (BC) and AH9BY4ADXX (AC). Translation of strain numbers: NCYC2572 (1001), NCYC8 (1002), NCYC9 (1003), NCYC10 (1004), NCYC103 (1005), NCYC459 (1006), NCYC475 (1007), NCYC792 (1008), NCYC3045 (1009), NCYC3364 (1010), CBS117 (1011), CBS5140 (1012), CBS1101 (1013), CBS1792 (1014), J63 (1015), J26 (1016), DBH9 (1019).

## Specie identification using kraken (Wood and Salzberg, 2014) on the strains reads.

| Strain | Fungi | Bacteria | Virus | Archaea | Others | Unclassified | Deha | Sc | Human |
|---|---|---|---|---|---|---|---|---|---|
| 1001 AH 1 | 82.38 | 0.01 | 0.00 | 0.00 | 0.01 | 17.59 | 78.50 | 0.01 | 0.01 |
| 1001 AH 2 | 81.22 | 0.01 | 0.00 | 0.00 | 0.01 | 18.74 | 77.38 | 0.01 | 0.01 |
| 1001 BC 1 | 82.08 | 0.01 | 0.00 | 0.00 | 0.01 | 17.90 | 78.25 | 0.01 | 0.01 |
| 1001 BC 2 | 81.33 | 0.01 | 0.00 | 0.00 | 0.02 | 18.64 | 77.53 | 0.01 | 0.02 |
| 1002 AH 1 | 98.31 | 0.01 | - | - | 0.01 | 1.97 | 93.71 | 0.01 | 0.01 |
| 1002 AH 2 | 97.78 | 0.00 | - | - | 0.01 | 2.19 | 93.20 | 0.01 | 0.01 |
| 1002 BC 1 | 98.30 | 0.01 | - | - | 0.01 | 1.68 | 93.71 | 0.01 | 0.01 |
| 1002 BC 2 | 97.37 | 0.00 | - | - | 0.02 | 2.60 | 92.82 | 0.01 | 0.02 |
| 1003 AH 1 | 82.81 | 0.02 | 0.00 | 0.00 | 0.01 | 17.16 | 78.91 | 0.00 | 0.01 |
| 1003 AH 2 | 80.93 | 0.02 | 0.00 | 0.00 | 0.01 | 19.02 | 77.08 | 0.01 | 0.01 |
| 1003 BC 1 | 82.55 | 0.02 | 0.00 | 0.00 | 0.01 | 17.41 | 78.71 | 0.00 | 0.01 |
| 1003 BC 2 | 80.84 | 0.02 | 0.00 | 0.00 | 0.02 | 19.11 | 77.05 | 0.01 | 0.02 |
| 1004 AH 1 | 91.91 | 0.04 | 0.00 | 0.00 | 0.00 | 8.04 | 89.12 | 0.00 | 0.00 |
| 1004 AH 2 | 88.13 | 0.04 | 0.00 | 0.00 | 0.01 | 11.81 | 85.43 | 0.00 | 0.01 |
| 1004 BC 1 | 91.77 | 0.05 | 0.00 | 0.00 | 0.01 | 8.18 | 89.03 | 0.00 | 0.01 |
| 1004 BC 2 | 89.33 | 0.04 | 0.00 | 0.00 | 0.02 | 10.60 | 86.66 | 0.00 | 0.02 |
| 1005 AH 1 | 98.48 | 0.06 | 0.00 | - | 0.00 | 1.45 | 93.36 | 0.00 | 0.00 |
| 1005 AH 2 | 97.04 | 0.06 | 0.00 | - | 0.01 | 2.89 | 91.97 | 0.00 | 0.01 |
| 1005 BC 1 | 98.43 | 0.06 | 0.00 | - | 0.00 | 1.50 | 93.46 | 0.00 | 0.00 |
| 1005 BC 2 | 96.92 | 0.06 | 0.00 | - | 0.01 | 3.00 | 92.09 | 0.00 | 0.01 |
| 1006 AH 1 | 17.32 | 0.05 | 0.00 | - | 0.02 | 82.60 | 12.74 | 0.02 | 0.02 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1006 AH 2 | 16.73 | 0.04 | 0.00 | - | 0.03 | 83.18 | 12.21 | 0.02 | 0.03 |
| 1006 BC 1 | 17.22 | 0.05 | 0.00 | - | 0.02 | 82.69 | 12.69 | 0.02 | 0.02 |
| 1006 BC 2 | 16.73 | 0.04 | 0.00 | - | 0.03 | 83.18 | 12.30 | 0.02 | 0.03 |
| 1007 AH 1 | 98.29 | 0.02 | 0.00 | - | 0.00 | 1.68 | 91.13 | 0.00 | 0.00 |
| 1007 AH 2 | 97.08 | 0.02 | 0.00 | 0.00 | 0.01 | 2.88 | 89.97 | 0.00 | 0.01 |
| 1007 BC 1 | 98.23 | 0.02 | 0.00 | 0.00 | 0.00 | 1.74 | 91.22 | 0.00 | 0.00 |
| 1007 BC 2 | 97.03 | 0.02 | 0.00 | 0.00 | 0.01 | 2.93 | 90.12 | 0.00 | 0.01 |
| 1008 AH 1 | 97.60 | 0.01 | - | - | 0.01 | 2.38 | 91.44 | 0.00 | 0.01 |
| 1008 AH 2 | 96.60 | 0.00 | 0.00 | 0.00 | 0.01 | 3.37 | 90.49 | 0.00 | 0.01 |
| 1008 BC 1 | 97.54 | 0.01 | -Â | 0.00 | 0.01 | 2.44 | 91.52 | 0.00 | 0.01 |
| 1008 BC 2 | 96.50 | 0.00 | - | - | 0.02 | 3.47 | 90.59 | 0.00 | 0.02 |
| 1009 AH 1 | 72.97 | 0.01 | 0.00 | 0.00 | 0.02 | 26.99 | 69.45 | 0.01 | 0.02 |
| 1009 AH 2 | 71.69 | 0.01 | 0.00 | 0.00 | 0.03 | 28.26 | 68.21 | 0.01 | 0.03 |
| 1009 BC 1 | 72.75 | 0.02 | 0.00 | 0.00 | 0.02 | 27.20 | 69.22 | 0.01 | 0.02 |
| 1009 BC 2 | 71.70 | 0.01 | 0.00 | 0.00 | 0.03 | 28.24 | 68.23 | 0.01 | 0.03 |
| 1010 AH 1 | 14.33 | 0.03 | 0.00 | 0.00 | 0.02 | 85.61 | 9.20 | 0.03 | 0.02 |
| 1010 AH 2 | 13.97 | 0.02 | 0.00 | 0.00 | 0.02 | 85.96 | 8.94 | 0.03 | 0.02 |
| 1010 BC 1 | 14.12 | 0.03 | 0.00 | 0.00 | 0.02 | 85.83 | 9.15 | 0.03 | 0.02 |
| 1010 BC 2 | 13.79 | 0.02 | 0.00 | 0.00 | 0.03 | 86.14 | 8.91 | 0.02 | 0.03 |
| 1011 AH 1 | 58.50 | 0.02 | 0.00 | 0.00 | 0.02 | 41.44 | 53.28 | 0.01 | 0.02 |
| 1011 AH 2 | 58.00 | 0.01 | 0.00 | 0.00 | 0.03 | 41.94 | 52.80 | 0.02 | 0.03 |
| 1011 BC 1 | 58.30 | 0.02 | 0.00 | 0.00 | 0.03 | 41.64 | 53.10 | 0.02 | 0.03 |
| 1011 BC 2 | 57.72 | 0.01 | 0.00 | 0.00 | 0.04 | 42.21 | 52.57 | 0.02 | 0.04 |
| 1012 AH 1 | 17.31 | 0.02 | 0.00 | - | 0.02 | 82.64 | 12.17 | 0.02 | 0.02 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1012 AH 2 | 16.85 | 0.02 | 0.00 | 0.00 | 0.02 | 83.08 | 11.82 | 0.02 | 0.02 |
| 1012 BC 1 | 17.20 | 0.02 | 0.00 | - | 0.02 | 82.74 | 12.13 | 0.02 | 0.02 |
| 1012 BC 2 | 16.82 | 0.02 | 0.00 | 0.00 | 0.03 | 83.12 | 11.83 | 0.02 | 0.03 |
| 1013 AH 1 | 96.50 | 0.04 | 0.00 | 0.00 | 0.00 | 3.45 | 90.88 | 0.00 | 0.00 |
| 1013 AH 2 | 95.14 | 0.04 | 0.00 | 0.00 | 0.01 | 4.80 | 89.58 | 0.00 | 0.01 |
| 1013 BC 1 | 96.44 | 0.05 | 0.00 | 0.00 | 0.00 | 3.51 | 90.90 | 0.00 | 0.00 |
| 1013 BC 2 | 95.09 | 0.04 | 0.00 | 0.00 | 0.01 | 4.85 | 89.65 | 0.00 | 0.01 |
| 1014 AH 1 | 81.84 | 0.01 | 0.00 | 0.00 | 0.01 | 18.13 | 78.14 | 0.00 | 0.01 |
| 1014 AH 2 | 80.64 | 0.01 | 0.00 | - | 0.01 | 19.33 | 76.97 | 0.00 | 0.01 |
| 1014 BC 1 | 81.65 | 0.01 | 0.00 | 0.00 | 0.01 | 18.32 | 77.98 | 0.00 | 0.01 |
| 1014 BC 2 | 80.39 | 0.01 | 0.00 | - | 0.02 | 19.57 | 76.75 | 0.00 | 0.02 |
| 1015 AH 1 | 96.65 | 0.01 | - | - | 0.01 | 3.33 | 90.17 | 0.00 | 0.01 |
| 1015 AH 2 | 95.87 | 0.01 | 0.00 | 0.00 | 0.02 | 4.10 | 89.43 | 0.00 | 0.02 |
| 1015 BC 1 | 96.60 | 0.01 | 0.00 | 0.00 | 0.01 | 3.38 | 90.32 | 0.00 | 0.01 |
| 1015 BC 2 | 95.65 | 0.01 | 0.00 | 0.00 | 0.02 | 4.31 | 89.48 | 0.00 | 0.02 |
| 1016 AH 1 | 81.29 | 0.01 | 0.00 | 0.00 | 0.01 | 18.68 | 76.99 | 0.01 | 0.01 |
| 1016 AH 2 | 79.98 | 0.01 | 0.00 | - | 0.01 | 19.99 | 75.72 | 0.00 | 0.01 |
| 1016 BC 1 | 81.05 | 0.01 | 0.00 | 0.00 | 0.01 | 18.92 | 76.80 | 0.01 | 0.01 |
| 1016 BC 2 | 79.84 | 0.01 | 0.00 | 0.00 | 0.02 | 20.12 | 75.64 | 0.01 | 0.02 |
| 1017 AH 1 | 98.04 | 0.03 | 0.00 | - | 0.02 | 1.89 | 0.05 | 93.14 | 0.02 |
| 1017 AH 2 | 96.71 | 0.03 | 0.00 | 0.00 | 0.03 | 3.21 | 0.04 | 91.83 | 0.03 |
| 1017 BC 1 | 97.99 | 0.05 | 0.00 | - | 0.02 | 1.93 | 0.05 | 93.03 | 0.02 |
| 1017 BC 2 | 97.02 | 0.02 | 0.00 | 0.00 | 0.04 | 2.90 | 0.04 | 92.09 | 0.04 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1018 AH 1 | 98.12 | 0.01 | - | - | 0.01 | 1.84 | 0.04 | 94.68 | 0.01 |
| 1018 AH 2 | 97.06 | 0.01 | 0.00 | - | 0.02 | 2.90 | 0.03 | 93.62 | 0.02 |
| 1018 BC 1 | 98.08 | 0.01 | 0.00 | - | 0.02 | 1.88 | 0.04 | 94.65 | 0.02 |
| 1018 BC 2 | 97.04 | 0.01 | 0.00 | - | 0.03 | 2.91 | 0.03 | 93.61 | 0.03 |
| 1019 AH 1 | 98.19 | 0.03 | 0.00 | - | 0.00 | 1.78 | 95.54 | 0.00 | 0.00 |
| 1019 AH 2 | 95.70 | 0.03 | 0.00 | 0.00 | 0.01 | 4.26 | 93.10 | 0.00 | 0.01 |
| 1019 BC 1 | 98.14 | 0.04 | 0.00 | 0.00 | 0.00 | 1.82 | 95.51 | 0.00 | 0.00 |
| 1019 BC 2 | 96.64 | 0.03 | 0.00 | - | 0.01 | 3.31 | 94.04 | 0.00 | 0.01 |

**Table 4. Classification of contigs on the raw reads, made by kraken using all refSeq available of Fungi, Bacteria, Archaea, Virus, plasmids and human.** The numbers represent percentages of the number of contigs matching that classification from the total number of contigs. Deha: abbreviation for *D. hansenii*. Sc: abbreviation for *S. cerevisiae*. The libraries names correspond to the identification codes when sequenced: BC47HDACXX (BC) and AH9BY4ADXX (AC) and the numbers to the forward (1) and reverse (2) reads. Translation of strain numbers: NCYC2572 (1001), NCYC8 (1002), NCYC9 (1003), NCYC10 (1004), NCYC103 (1005), NCYC459 (1006), NCYC475 (1007), NCYC792 (1008), NCYC3045 (1009), NCYC3364 (1010), CBS117 (1011), CBS5140 (1012), CBS1101 (1013), CBS1792 (1014), J63 (1015), J26 (1016), DBH9 (1019).