Differential adaptation of the yeast Candida anglica to fermented food

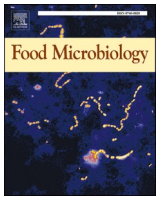
Frédéric Bigey, Xavière Menatong Tene, Marc Wessner, Martine Pradal, Jean-Marc Aury, Corinne Cruaud, Cécile Neuvéglise

Food Microbiology 123 (2024) 104584

Contents lists available at ScienceDirect

Food Microbiology

journal homepage: www.elsevier.com/locate/fm



Differential adaptation of the yeast *Candida anglica* to fermented food

Frédéric Bigey [a], Xavière Menatong Tene [a], Marc Wessner [b], Martine Pradal [a], Jean-Marc Aury [b], Corinne Cruaud [c], Cécile Neuvéglise [a,*]

[a] SPO, Univ Montpellier, INRAE, Institut Agro, Montpellier, France
[b] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, 91057, France
[c] Genoscope, Institut François Jacob, CEA, Université Paris-Saclay, Evry, 91057, France

ARTICLE INFO

Keywords:
Genome
CUG-Ser1 clade
Cider
Cheese
LINE non-LTR retrotransposon
Genetic diversity
Phenotypic diversity

ABSTRACT

A single strain of *Candida anglica,* isolated from cider, is available in international yeast collections. We present here seven new strains isolated from French PDO cheeses. For one of the cheese strains, we achieved a high-quality genome assembly of 13.7 Mb with eight near-complete telomere-to-telomere chromosomes. The genomes of two additional cheese strains and of the cider strain were also assembled and annotated, resulting in a core genome of 5966 coding sequences. Phylogenetic analysis showed that the seven cheese strains clustered together, away from the cider strain. Mating-type locus analysis revealed the presence of a *MAT*a locus in the cider strain but a *MAT*alpha locus in all cheese strains. The presence of LINE retrotransposons at identical genome position in the cheese strains, and two different karyotypic profiles resulting from chromosomal rearrangements were observed. Together, these findings are consistent with clonal propagation of the cheese strains. Phenotypic trait variations were observed within the cheese population under stress conditions whereas the cider strain was found to have a much greater capacity for growth in all conditions tested.

1. Introduction

Candida anglica NRRL Y-27079^T (= CBS 4262^T), a yeast strain isolated from cider, is the only representative of this species in international culture collections (Kurtzman et al., 2001). The presence of this yeast species and its potential role in cider is poorly documented. As for many other yeasts assigned to genus *Candida* (*Candida* encompasses distantly related species that do not form a monophyletic group), the taxonomic status of this species has not been clearly established. *C. anglica* belongs to the major clade CUG-Ser1 or Serinales and to the Debaryomycetaceae family (Groenewald et al., 2023). Some of its closest relatives are still considered to be *Candida* species, whereas others have been reassigned to the genus *Kurtzmaniella,* for which *K. cleridarum* is the reference species (Lachance and Starmer, 2008) (Fig. S1). In 2014, Kurtzman and Robnett suggested that *Candida quercitrusa* and *Candida natalensis* belong to the genus *Kurtzmaniella* (Kurtzman and Robnett, 2014). These species were transferred to genus *Kurtzmaniella* as new combinations in 2019, together with *Candida fragi* and the new species *Kurtzmaniella hittingeri* (Lopes et al., 2019). This genus, therefore, currently contains five species. Daniel et al. proposed the reclassification of other *Candida* spp. including *C. anglica*, to genus *Kurtzmaniella,* but the lack of genomic data suggests it would be prudent to wait for additional sequences before taking such a step (Daniel et al., 2014; Lopes et al., 2019). However, in some studies authors have already chosen to use the name *Kurtzmaniella* for *Candida zeylanoides* and *Candida santamariae* (Belleggia et al., 2020).

Only two genomes thought to belong to *C. anglica* are available to date. The genome deposited under GenBank assembly accession number GCA_019775655.1 is in fact *Candida santamariae*. During the writing of this article, a genome assembly of the type strain of *C. anglica*, NRRL Y-27079, was released as part of the Y1000+ Project (genome accession number: JANIVX000000000) (Opulente et al., 2023).

We present here seven new strains of *C. anglica* isolated from French PDO (protected designation of origin) cheeses as part of the MetaPDOcheese project (data not published). Metabarcoding analysis of the cheese microbiota in the MetaPDOcheese project showed that the read abundance of *C. anglica* reached more than 28% in some cheese samples, suggesting that this species might play a role in cheese ripening, despite never having been identified in cheese before. As such, it is not deliberately inoculated and its potential role requires further investigation. Additional knowledge is therefore needed regarding the functional characteristics and genetic diversity of this yeast, potentially reflecting

* Corresponding author. SPO, Univ Montpellier, INRAE, Institut Agro, 34060, Montpellier, France.
E-mail address: cecile.neuveglise@inrae.fr (C. Neuvéglise).

https://doi.org/10.1016/j.fm.2024.104584
Received 3 March 2024; Received in revised form 22 May 2024; Accepted 7 June 2024
Available online 7 June 2024

adaptive evolution to different fermented foods. The aim of this study was to generate a reference genome consisting of a high-quality assembly of near-complete telomere-to-telomere chromosomes, with curated genome annotation. Another goal was to investigate the genetic and phenotypic diversity of the eight available *C. anglica* strains. To this end, we sequenced the genomes of all strains and performed a population genomics analysis. We also generated assemblies for three strains to study the evolution of chromosome structure. Phenotypic traits were studied through quantitative measurements of growth under different conditions.

2. Materials and methods

2.1. Strains and culture media

Yeast strains were maintained on YPD medium (10 g/L yeast extract, 20 g/L Bacto peptone, 20 g/L glucose, 20 g/L agar). All the strains used in this study are listed in Table S1. Cheese strains of *C. anglica* have been deposited at the CIRM-Levures yeast collection (https://cirm-levures.bio-aware.com).

2.2. Genome sequencing and annotation

2.2.1. DNA extraction

Strains were cultured for at least 36 h in 10 mL YPD medium at 28 °C, with shaking at 220 rpm. Two methods were used for DNA extraction: one for sequencing with Illumina technology, and the other for long-read sequencing with Nanopore technology. The first method was based on an in-house protocol involving mechanical and chemical lysis, as previously described (Saubin et al., 2019). The second method was adapted from a protocol developed at Genoscope (Denis et al., 2018). Briefly, it involves cell wall lysis with zymolyase to generate spheroplasts, which are then lysed chemically in the presence of SDS. Proteins are then removed by precipitation with potassium acetate. Finally, the DNA is precipitated with isopropanol, washed with ethanol and resuspended in TE (10 mM Tris-HCl, 1 mM EDTA).

2.2.2. RNA extraction

Strains were cultured overnight in 10 mL YPD medium at 28 °C, with shaking at 220 rpm. Cell concentration was then estimated by spectrophotometry and 10^9 cells were harvested. Total RNA was isolated by the TRIzol® method (Chomczynski and Sacchi, 1987) as previously described (Duc et al., 2017).

2.2.3. Illumina library preparation and sequencing

Purification with 1.2x AMPure XP beads (Beckman Coulter Genomics, Danvers, MA, USA) was performed as a first step in the elimination of small fragments protocol.

Libraries were then prepared with NEBNext DNA Modules Products (New England Biolabs, Ipswich, MA, USA) with an 'on beads' protocol developed by Genoscope, starting with 250 ng genomic DNA. After adapter ligation, the ligated product was amplified by 12 cycles of PCR with the Kapa Hifi Hotstart NGS library Amplification kit (Roche, Basel, Switzerland), followed by purification with 0.6x AMPure XP beads (Beckman). Libraries were sequenced with an Illumina NovaSeq 6000 instrument (Illumina, San Diego, CA, USA), in paired-end mode, generating 150-base reads.

After Illumina sequencing, an in-house quality-control process was applied to the reads that passed the Illumina quality filters. The first step discarded low-quality nucleotides ($Q < 20$) from both ends of the reads. The Illumina sequencing adapter and primer sequences were then removed from the reads, and any reads shorter than 30 nucleotides after trimming were discarded. These trimming steps were achieved with in-house software based on the FastX package (Engelen and Aury). In the final step, read pairs mapping to the phage phiX genome were discarded with the SOAP aligner (Li et al., 2008), together with the PhiX174 enterobacterial phage reference sequence (GenBank: NC_001422.1). Short Illumina reads were subjected to post-processing to remove low-quality data, as previously described (Alberti et al., 2017; Aury et al., 2008).

2.2.4. ONT library preparation and sequencing

Genomic DNA (9 μg) was first purified with the Short Read Eliminator XL Kit (Pacific Biosciences, Menlo Park, CA, USA). The library was prepared according to the "1D Native barcoding genomic DNA (with EXP-NBD104 and SQK-LSK109)" protocol provided by Oxford Nanopore Technologies (Oxford Nanopore Technologies Ltd, Oxford, UK), with 2 μg purified genomic DNA. The sample (pooled with five other barcoded samples) was sequenced with a R9.4.1 MinION flow cell. Reads were basecalled with Guppy version 4.3.4.

2.2.5. Illumina RNASeq library preparation and sequencing

The library was prepared with the NEBNext Ultra II Directional RNA Library Prep for Illumina (New England BioLabs) according to the manufacturer's protocol, with 100 ng total RNA. The samples were sequenced with an Illumina NovaSeq 6000 instrument (Illumina, San Diego, CA, USA), in paired-end mode, generating 150-base reads.

2.2.6. Long read-based genome assembly for reference strain 29B2s

Raw Nanopore reads (Table S2A) were assembled with NECAT, using a genome size of 20 Mb and default parameters (Chen et al., 2021). The NECAT output was polished once with Racon (Vaser et al., 2017) for Nanopore reads, then once with Medaka (model r941_prom_high_g4011) for Nanopore reads, and twice with Hapo-G 1.3.4 (Aury and Istace, 2021) for Illumina short reads. We obtained an assembly of eight contigs corresponding to eight chromosomes of the *C. anglica* reference genome. Contig N50 (the contig size above which 50% of the total length of the sequence assembly is included) was 1.88 Mb.

2.2.7. Genome annotation with RNA-seq and manual curation

Repeats in the genome assembly were masked with Tandem Repeat Finder (Benson, 1999) for tandem repeats and RepeatMasker (Smit et al.) for simple repeats and known repeats included in RepBase with the parameter "species Saccharomycetes" (Bao et al., 2015).

Genes were predicted from several proteomes: *Candida railenensis*, *Debaryomyces hansenii*, *Diutina rugosa, Kazachtania naganishii* and *Kazachstania africana*, all downloaded from NCBI. Proteomes were aligned against the genome in a two-step strategy. First, BLAT version 36 with default parameters (Kent, 2002) was used for rapid localisation of the regions putatively corresponding to these proteins in the genome. The best match and matches with a score ≥90% of the best match score were retained. Alignments were then refined with Genewise version 2.2.0 with default parameters (Birney et al., 2004), as this approach is more accurate for detecting intron/exon boundaries. Alignments were retained if >75% of the length of the protein could be aligned with the genome.

We also used short-read RNA-Seq data for the detection of expressed and/or specific genes. Short reads were mapped onto the genome assembly with HISAT2 version 2.2.1 with default parameters (Kim et al., 2019). The bam file was sorted and Stringtie version 2.2.1 (Pertea et al., 2016) was used, with the following parameters: p 16 -v -m 150. Only the most expressed transcript at each genomic locus was retained.

All the protein and RNA-Seq alignments were combined with Gmove (Dubarry et al.), an easy-to-use predictor that does not require a pre-calibration step. Briefly, putative exons and introns extracted from alignments were used to build a graph, on which the nodes and edges represented exons and introns, respectively. Gmove extracts all paths from the graph and searches for open reading frames consistent with the protein evidence. Genes with >50% untranslated regions and a coding sequence (CDS) length shorter than 300 bp, were excluded. This pipeline identified 6041 predicted genes with a mean of 1.08 exons per gene.

Expert manual curation of gene models was performed with the

RNA-seq data, making it possible to check and correct intron splice sites and to identify missing genes in the four assembled genomes. To this end, RNA-seq sequencing reads were mapped onto the reference genome with HISAT2 v2.2.1 (Kim et al., 2019). Manual structural annotation changes were made with the Artemis genome browser (Carver et al., 2012).

Functional annotation was performed with go-FAnnoT in-house software (https://github.com/hdevillers/go-fannot) that assigns annotations from homologies found in the curated UniProt database (release 2023_01).

In addition to protein-coding genes, tRNA genes were predicted with tRNAscan-SE v2.0.9 (Chan et al., 2021). BLAST was also used to search for the presence of a complete rDNA unit including 18S, 5.8S, 26S and 5S rRNA genes, together with ncRNA genes (Zhang et al., 2000). *Debaryomyces hansenii* sequences were used as the reference sequences for this analysis. The completeness of this annotation was evaluated with BUSCO v5.2.2 (Manni et al., 2021a,b), using the saccharomycetes_odb10 lineage data set as the reference (2137 proteins).

2.2.8. Genome assemblies for other C. anglica strains

Genome assemblies for three other *C. anglica* strains (28E1s, 29E1s and NRRL Y-27079) were generated using Illumina reads with SPAdes assembler version 3.13.1 (parameter: k 21,33,55,77,99,127) (Prjibelski et al., 2020). *Ab initio* gene detection was performed with MAKER2 (Holt and Yandell, 2011). Manual curation was performed as described for strain 29B2s.

2.3. Analysis of chromosomal rearrangements

SynChro was used to reconstruct synteny blocks from pairwise comparisons of the annotated genomes (Drillon et al., 2014). MUMmer4 package was used to generate global alignments of 29B2s, 28E1s and NRRL Y-27079 genomes (Marcais et al., 2018). The alignments were further filtered using the delta-filter utility (MUMmer4) with parameters -l 5000 and -m. Alignment coordinates were extracted with the show-coords utility to generate a tabular file. Circular representation of chromosomal rearrangements was performed in R using the package circlize version 0.4.16 (Gu et al., 2014).

2.4. Analysis of genetic diversity

2.4.1. Read mapping, detection of SNPs and indels, and phylogeny construction

The Illumina reads were mapped with BWA version 0.7.17 (Li and Durbin, 2009) on the genome assembly of strain 29B2s (Table S2B). We used Genome Analysis Toolkit (GATK) v4.1.7.0 for variant calling (HaplotypeCaller parameters: sample-ploidy 1; –emit-ref-confidence BP_RESOLUTION), with hard filtering according to GATK best practice (O'Connor and van der Auwera, 2020). This genotyping pipeline generated a VCF file of 35,348 variants containing 29,371 SNPs and 5977 indels identified in the eight yeast samples. The set of 29,371 SNPs was further filtered by removing SNP positions with missing genotype scores above 0.75, and genotypes with quality and read depths below 30 were masked with vcftools version v0.1.15 (parameters: minGQ 30, –minDP 30, –max-missing 0.75) (Danecek et al., 2011). The resulting filtered data set contained 28,413 SNPs.

For phylogenetic analysis, the VCF SNP file was converted into Phylip with a custom Perl script. A phylogenetic tree was constructed with IQ-TREE version 1.6.12 (Nguyen et al., 2015), based on a maximum-likelihood approach and with the best-fit model automatically selected by ModelFinder (parameter -m MFP). Branch support was assessed using an ultra-rapid bootstrap approximation (parameter -bb 1000).

Read mapping depth was analysed with genomecov from the BEDTools package v2.29.0 (parameter: d) (Quinlan and Hall, 2010). For each genome position, depth was divided by the genome-wide mean depth, and the resulting value underwent $\log_2$ transformation. Plots were obtained by the shifted average method (window: 5000; step: 500) in R (R Core Team, 2023). The impact of each variant was determined with SnpEff version 5.1d (Cingolani et al., 2012). Variants were classified as 'high impact' for variants predicted to have a high disruptive effect on the protein, likely causing protein truncation, loss of function (e.g. frame shift, stop loss or gain, start loss), 'moderate impact' for missense variants, intronic insertions or deletions, and 'low impact' mainly for synonymous variants.

2.4.2. Genome average nucleotide identity (ANI)

ANI was calculated using OrthoANIu algorithm (Yoon et al., 2017).

2.4.3. Phylogenomic tree

The phylogenomic tree of 26 yeast species (Table S1B) was constructed using the supermatrix method. To identify single-copy genes within input sequences for each species, the BUSCO software v5.3.1 (Manni et al., 2021a,b) was employed with the parameters –mode genome –lineage_dataset saccharomycetes_odb10. The BUSCO proteins corresponding to the single-copy genes in all species were retrieved. The protein families were individually aligned using Muscle v5.1 (Edgar, 2004) with default parameters. Poorly aligned regions were removed by trimming the alignments with trimAl v1.2rev59 using parameter -automated1 (Capella-Gutierrez et al., 2009). The resulting alignments were concatenated into a supermatrix. Finally, a maximum-likelihood tree was inferred using IQ-TREE v2.0.7 with parameters -B 1000 (1000 ultra-rapid bootstraps replicates) -alrt 1000 (1000 bootstrap replicates for SH-aLRT) -mset LG (restrict ModelFinder to test only LG models) (Nguyen et al., 2015).

2.5. Phenotypic characterization

Drop tests were performed on solid media prepared with the following carbon sources: seven sugars (glucose, D-galactose, maltose, lactose, sucrose, fructose and xylose), one organic acid (DL-lactate), a polyol (glycerol), six hydrophobic substrates (oleic acid, tributyrin, linoleic acid, octanoic acid, caproic acid, caprylic acid), ethanol and yeast extract. Carbon sources were added to the base medium YNB (1.7 g/L yeast nitrogen base, 50 mM phosphate buffer pH 6.8, 0.67 g/L NH_4Cl, and 20 g/L agar) at a final concentration of 1%, except for ethanol, which was added at final concentrations of 5% and 10%. For four carbon sources, the pH was adjusted to 4.5 with citric acid solution. Lipids were emulsified by sonication of a 20% mixture in water with 0.625% Tween 80 and added to YNB at a final concentration of 1%. Additional YNB plates were prepared, with NaCl concentrations of 1%–10% or hygromycin at a concentration of 100 or 200 mg/L. YPD medium (10 g/L yeast extract, 20 g/L Bacto peptone, 10 g/L glucose, and 20 g/L agar) was supplemented with sodium dodecyl sulfate (SDS, at a concentration of 0.01% or 0.02%), ethanol (5% or 10%) or 20% glycerol. For the detection of protease activity, a skimmed milk agar medium was prepared as previously described (Abdelmoteleb et al., 2017). All plates were incubated at 28 °C, except for YPD plates, which were incubated at 28 °C and 37 °C.

Strains were cultured overnight in 10 mL liquid YPD medium at 28 °C with shaking at 180 rpm. Cell concentrations were estimated by flow cytometry in a C6 Accuri (Ann Arbor, MI, United States) spectrophotometer. Three dilutions were then used to inoculate the plates: 10^7, 10^6 and 10^5 cells per mL. We transferred 200 μL of each dilution to a 96-well microplate, and a 96-pin replicator (Boekel Scientific) was used to inoculate the agar plates. Photographs were taken at 24, 48 and 72 h and at 6 and 8 days.

2.6. Pulsed-field gel electrophoresis

Yeast karyotyping was achieved by contour-clamped homogeneous electric field (CHEF) gel electrophoresis. Plugs of yeast chromosomes

were prepared as described elsewhere (Vezinhet et al., 1990). The CHEF-DR III pulsed-field gel electrophoresis system (Bio-Rad, Hercules, CA, United States) was set to 5.2 V/cm with pulses of 90–120 s for 12 h and then to 4 V/cm with pulses of 120–360 s for 24 h, and samples were run on 1% Seakem® Gold Agarose gels (Lonza, Rockland, ME, USA) in 0.5 x TAE buffer at 12 °C. The chromosomes of *Saccharomyces cerevisiae* CLIB 112 (=YNN295) and *Lachancea kluyveri* CBS 3082 were used as size markers. The agarose gels were stained with ethidium bromide (0.5 mg/mL) and washed with water before visualization under UV.

2.7. Flow cytometry

Yeast cells were prepared for flow cytometry analysis as previously described (Saubin et al., 2019), with a protocol including cell culture, fixation in ethanol, RNase A and proteinase K treatments, sonication and labelling with SYTOX R green (Invitrogen). DNA was quantified on a C6 Accuri (Ann Arbor, MI, United States) spectrophotometer with an excitation wavelength of 488 nm and an emission wavelength of 530 nm. Acquisition was performed on 30,000 events observed with gating on forward scatter/side scatter signals. The flow rate was set to about 2,000 events per second (medium flow, 35 mL/min; core, 16 mm).

3. Results and discussion

3.1. Sequencing, assembly and annotation of the Candida anglica reference genome

We collected 29 *Candida anglica* isolates in seven cheeses from three French PDOs (PDOs 28, 29 and 35) produced in the Auvergne-Rhône-Alpes region. We investigated only one strain per cheese, i.e., one strain per producer (28E, 29A, 29B, 29E and 35D) or two strains for producer 28C in the case of seasonal cheeses (cheeses 28C1 and 28C2). The isolates were collected from the cheese surface ("s" at the end of the strain name) or the core ("c") (Table S1). Flow cytometry analysis revealed that these seven strains were all haploid, like the type strain of *C. anglica,* NRRL Y-27079, which was isolated from cider (Table S1). Strain 29B2s was sequenced with a hybrid sequencing strategy coupling short Illumina reads and long Nanopore reads. The 13.7 Mb gap-free assembly consisted of eight contigs ranging from 1.3 Mb (chromosome A) to 2.5 Mb (chromosome H), each corresponding to one of the eight chromosomes. The same telomere repeats TGTATGGG as in *Candida zeylanoides* (McLaughlin et al., 2024) were found at 13 of the 16 contig extremities, suggesting that five of the contigs corresponded to telomere-to-telomere chromosomes, with the three remaining contigs lacking the left telomere (Table 1). Pulsed-field gel electrophoresis confirmed the presence of eight chromosomes of sizes compatible with the contigs (Fig. 1A). Chromosomal GC content ranged from 36.7% (chromosome H) to 40.2% (chromosome F).

Manual curation resulted in the identification of 6033 CDSs in the nuclear genome. BUSCO score was 98.6% before curation and increased to 99.8% after curation (Table S3). In addition to protein-coding genes, 243 tRNA genes were predicted. Three complete rRNA units, including the 5S rRNA gene, were annotated on chromosome H. Finally, one copy of the five small nuclear RNA genes (U1, U2, U4, U5 and U6) was included in the annotation. Ten relics of long interspersed nuclear elements (LINE) of the L1 family, were detected. The presence of LINE relics was also confirmed in the assembled genomes of strains 28E1s, 29E1s and Y-27079 (see below).

The full circular mitochondrial genome (chromosome X) of 34,910 bp (GC%: 22.6%) was also assembled and annotated (Table 1), revealing the presence of 21 CDS and 26 tRNA genes, all but *cox3,* encoding cytochrome *c* oxidase subunit 3, in the same orientation. The long and short subunits of the rRNA were also annotated. Three putative sequences of intron-encoded DNA homing endonucleases with a LAGLIDADG domain were found in *cox1* (cytochrome *c* oxidase subunit 1). Three putative intron-encoded endonucleases were predicted in the *cob* (cytochrome *b*) gene, one of which included a LAGLIDADG domain. The *nd5* gene (NADH-ubiquinone oxidoreductase chain 5) was split in two, whereas the *nd5* gene of *Debaryomyces hansenii* (accession number NC_010166) was not (Sacerdot et al., 2008). The 5′ part was located between *nd4L* and *atp9* and the 3' part was located upstream from *nd4*. We therefore annotated *nd5* as a pseudogene.

3.2. Nucleotide diversity and phylogeny

The short Illumina sequencing reads obtained for the eight *C. anglica* strains were mapped onto the reference assembly (strain 29B2s). In total, 35,348 variants were identified, including 29,371 SNPs and 5977 indels. This number of variants is smaller than that reported for 182 *Candida albicans* isolates, in which 589,255 SNPs were found (Ropars et al., 2018). This difference could be the result of a sampling bias due to the small number of strains analysed. We therefore compared our results with those of *Penicillium camemberti* var. *camemberti* for which only 13 individuals were sampled (11,484 SNPs) (Ropars et al., 2020). This clonal propagated fungus is inoculated at the beginning of the cheese-making process, which is not the case for *C. anglica*. We found 34,725 singletons (variants observed in only one genome): 29,082 SNPs and 5643 indels. The vast majority of these singletons, 34,410 (99%), were found only in the cider strain Y-27079, suggesting that this strain has diverged genetically from the cheese group. Despite the observed divergence of cider strain Y-27079, an ANI value of 99.6% was obtained with strain 29B2s, confirming that these two strains are conspecific. The genetic relationships between strains were investigated by generating a phylogenetic tree based on 28,413 biallelic SNPs (Fig. 2A). The cheese strains clustered together on this tree, far from the cider strain Y-27079. Indeed, the cheese strains displayed between 22 and 275 pairwise SNP differences (Fig. 2B), whereas the cider strain differed from the strains of the cheese group by more than 27,000 SNPs. The topology of the tree is consistent with the observed karyotypes: 29A1s and 28E1s have identical karyotypes and form a group separate from the other cheese strains (Profile P2, Fig. 1A).

Despite the low number of strains studied, the extremely low

Table 1
Assembly and annotation statistics for the 29B2s nuclear and mitochondrial genomes.

Chromosome	Size (bp)	GC%	Telomeres	CDS	CDS with intron	Pseudogenes	misc_RNA	Mobile elements	ncRNA	rRNA	tRNA
A	1288108	38.7	R	558	51	6	0	0	0	0	12
B	1342065	39.8	R	564	39	5	0	0	1	0	21
C	1393309	39.8	L, R	592	48	8	0	1	0	0	18
D	1418013	37.8	L, R	658	50	4	3	1	1	0	26
E	1879034	38.1	L, R	808	56	4	0	2	2	0	40
F	1904301	40.2	L, R	818	53	6	0	0	0	0	30
G	1981346	39.2	R	876	62	11	0	4	1	0	36
H	2513812	36.7	L, R	1159	92	8	0	2	0	12	60
Total nuclear	**13719988**			**6033**	**451**		**3**	**10**	**5**	**12**	**243**
X (mito)	34910	22.6		21	7	4	0	0	0	2	26

L: left telomere, R: right telomere.

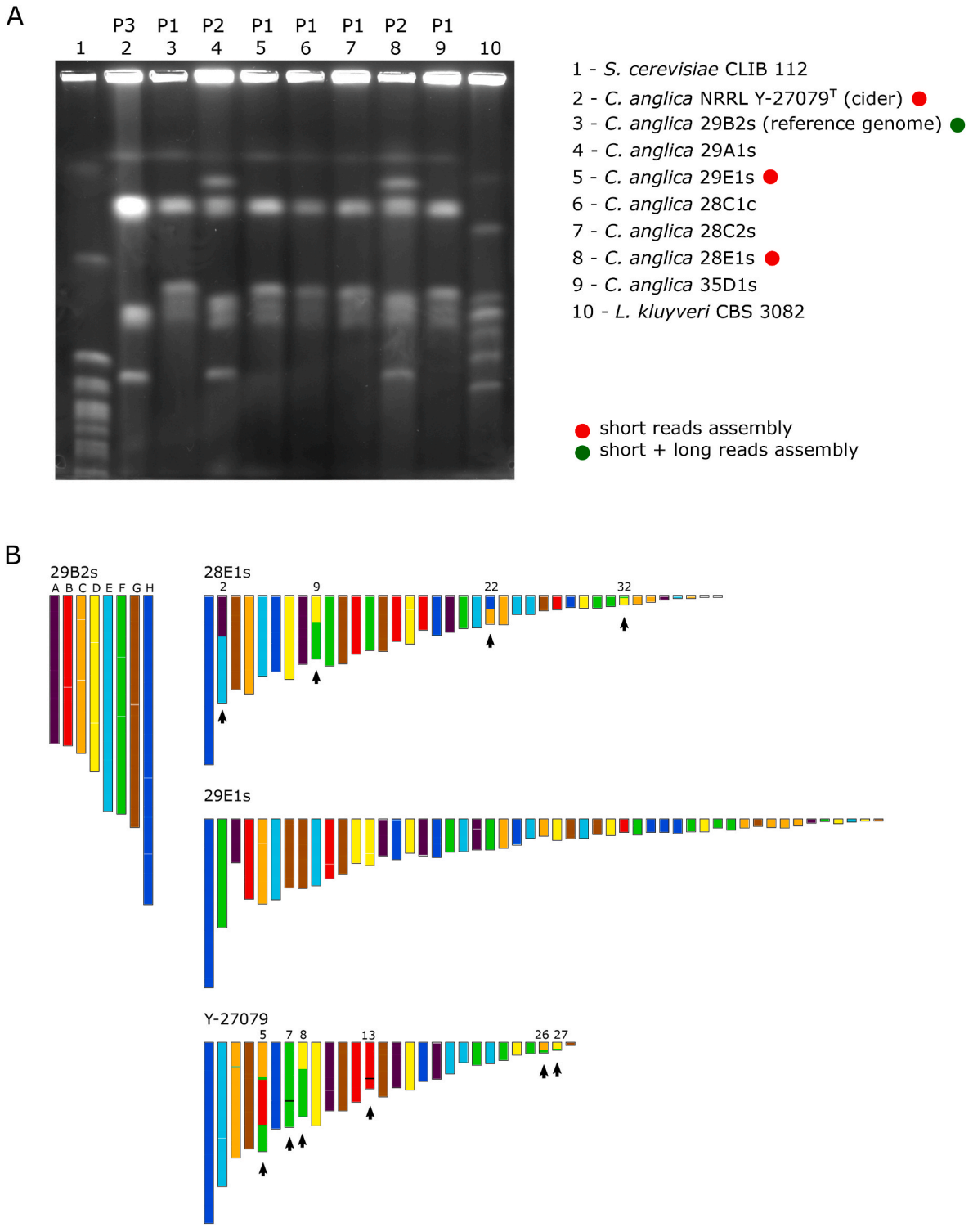
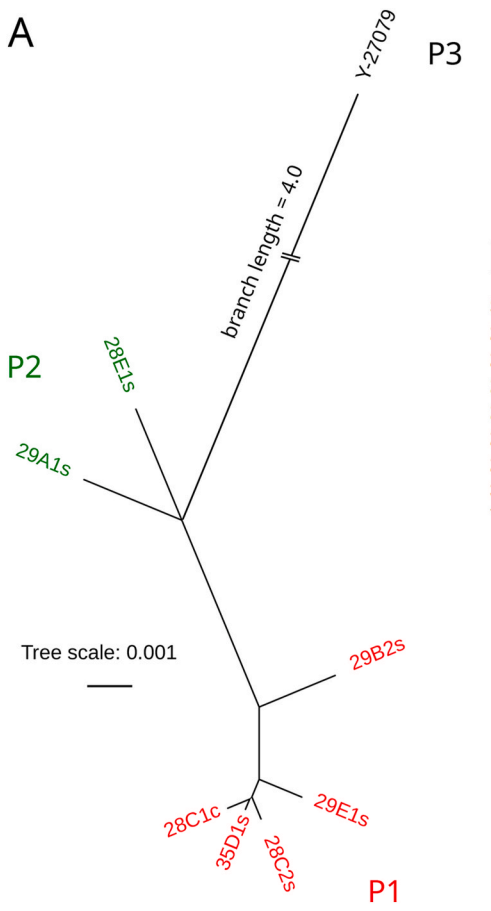


Fig. 1. (A) Karyotypes of the 8 strains of *C. anglica*. The chromosomes of *Saccharomyces cerevisiae* and *Lachancea kluyveri* were used as markers; (B) Chromosome painting showing orthologues in syntenic blocks, as computed by SynChro software. Scaffolds of strains 28E1s and Y-27079 with a chromosomal rearrangement compared to 29B2s are marked with a black arrow and their number is above the scaffolds.

nucleotide diversity of the cheese group is surprising given that the seven strains were isolated from six different dairies. Two strains (28C1c and 28C2s) obtained from the same dairy in different seasons showed only 32 SNPs, suggesting that these strains could be maintained throughout the year. On the other hand, one strain (35D1s) from a different PDO showed only 22 and 34 SNPs compared to 28C1c and 28C2s, respectively. The six dairies are all located in the same French region, Auvergne-Rhône-Alpes, and the strains may therefore have spread and been exchanged between producers via equipment, starters, or ripening cultures. All of the cheese strains had a *MAT*alpha locus, transposable elements in identical genome position, and two different karyotypic profiles resulting from chromosomal rearrangements (Fig. 1A). Together, these findings are consistent with clonal propagation of these cheese strains, as opposed to sexual reproduction.



B

Strain	28C1c	28C2s	28E1s	29A1s	29B2s	29E1s	35D1s	Y-27079
28C1c	0	32	275	262	124	56	22	27577
28C2s		0	273	264	122	56	34	27579
28E1s			0	140	249	275	267	27493
29A1s				0	242	264	254	27477
29B2s					0	125	116	26931
29E1s						0	48	27445
35D1s							0	27581
Y-27079								0

Fig. 2. (A) Maximum likelihood phylogenetic tree for *C. anglica* inferred from 28,413 biallelic SNPs in IQ-TREE (best model according to BIC: TVM + F). The branch length for Y-27079 was too long to be shown in the figure, so the branch was shortened. P1 (red), P2 (green) and P3 (black) refer to the three karyotype profiles that are shown in Fig. 1A. (B) Pairwise SNP differences between the strains isolated from cheese and cider. The strain names are written in the same colours as the karyotype profiles. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.3. Comparative analysis of four Candida anglica genomes

Analysis of the karyotypes of the eight *C. anglica* strains revealed three different karyotype profiles (Fig. 1A). Profile P1 is common to the cheese strains 29B2s (reference genome), 29E1s, 28C1c, 28C2s and 35D1s. The estimated size of the chromosomal bands corresponds to the sizes of the chromosomes in the reference genome (Table 1). Profile P2 was observed for strains 29A1s and 28E1s, with at least two chromosomes that are clearly different in size from that of their counterparts in the reference genome. Finally, the cider strain (Y-27079) had a unique karyotype (profile P3). These observations revealed that chromosomal rearrangement events had taken place in some strains. To investigate the type and position of these rearrangements, we searched for large segmental duplications but did not find any. We then assembled three *C. anglica* strain genomes (28E1s, 29E1s and Y-27079) from Illumina short reads and reconstructed blocks of synteny between these genome assemblies and the reference genome (Fig. 1B–Table S4A). As suggested by its karyotype, no recombination was observed in the genome of the 29E1s strain, whereas the genomes of the 28E1s and Y-27079 strains did display recombination events (Table S5, Fig. 3). Four translocations were detected in strain 28E1s: one translocation involving telomeres between chromosomes A and E of the reference genome, two translocations between chromosomes D and F and one translocation between chromosomes C and H. Complex rearrangements were observed for strain Y-27079. The two translocations between chromosomes D and F found in 28E1s were also observed, consistent with the topology of the phylogenetic tree of the strains (Figs. 2A and 3). Two additional translocations were detected, one internal to chromosome F and one between chromosomes C and F (Fig. 3). Finally, scaffold 5 of Y-27079 contains parts of chromosomes C, F and B (Figs. 1B and 3).

We analysed the breakpoints of the synteny blocks to determine the genes affected by the rearrangements and the ancestral loci. In all but one case, the rearrangements occurred in intergenic regions with no structural impact on the flanking genes. The exception was the CAAN3_05S07096 gene encoding a 1777-amino acid protein from the cider strain Y-27079, for which it was possible to reconstruct the evolutionary scenario. This gene, encoding an E3 ubiquitin ligase and a putative helicase, was split in two, resulting in two pseudogenes on chromosomes B (gene B04192) and F (gene F02366) of the reference genome 29B2s. The same organisation was observed for strains 28E1s and 29E1s, suggesting that this rearrangement occurred in the ancestor of all the cheese strains.

Our analysis of the orthologs present in synteny blocks led to the identification of a core genome of 5966 CDS (Fig. 4; Table S6). By contrast, 32, 2, 7 and 13 CDS were found specifically in strains 29B2s, 28E1s, 29E1s and Y-27079, respectively. Most of these genes are pseudogenes or dubious genes, generally with no homologues and no known functions. Some of the others are duplicated genes. An example of segmental duplication is provided in Fig. S2A. The *GAP1* gene, encoding a general amino-acid permease, is duplicated once in 28E1s and twice in 29B2s, whereas Y-27079 has only a single copy of this gene. Another striking example is provided by the region of chromosome C of 29B2s (Fig. S2B). This region has accumulated at least three independent segmental duplication events from three different chromosomes. Four of the six duplicated genes are pseudogenes, two of which are truncated.

Only three of the 13 genes specific to the cider strain Y-27079 have a

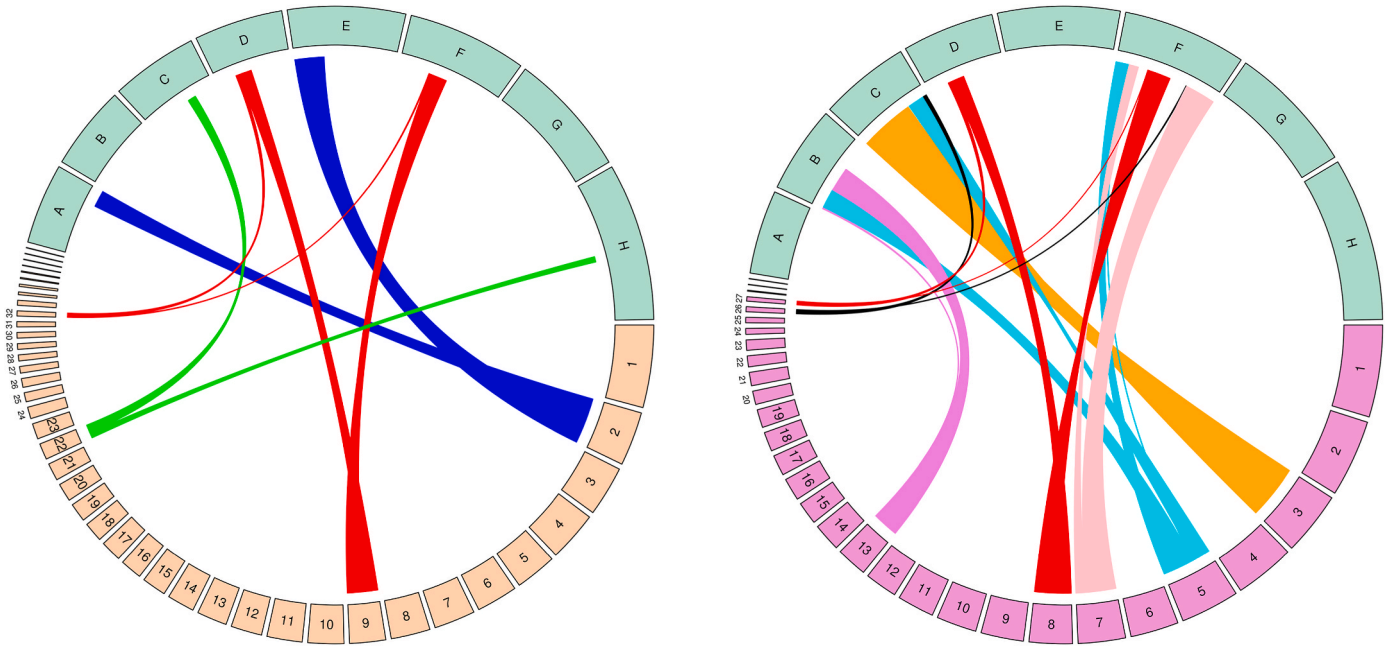


Fig. 3. Chromosomal rearrangements among *C. anglica* strains. Circular plots represent chromosome alignments of 29B2s with 28E1s (left) and Y-27079 (right). The eight chromosomes of the 29B2s genome are represented to scale by green rectangles with their names inside. Those of 28E1s and Y-27079 are shown in orange and pink, respectively. Alignments exceeding 5 kb are depicted as coloured lines. The red lines correspond to chromosomal rearrangements identified in 28E1s and Y-27079. Only genetic links between chromosomes and scaffolds resulting from rearrangement events are shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

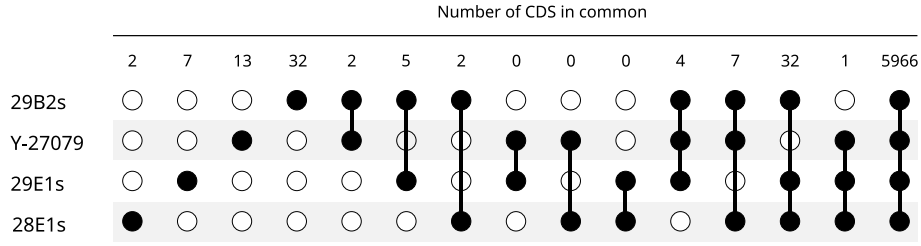


Fig. 4. Number of orthologues common to the genomes of *C. anglica* 29B2s (CAAN4), 28E1s (CAAN1), 29E1s (CAAN2) and NRRL Y-27079 (CAAN3).

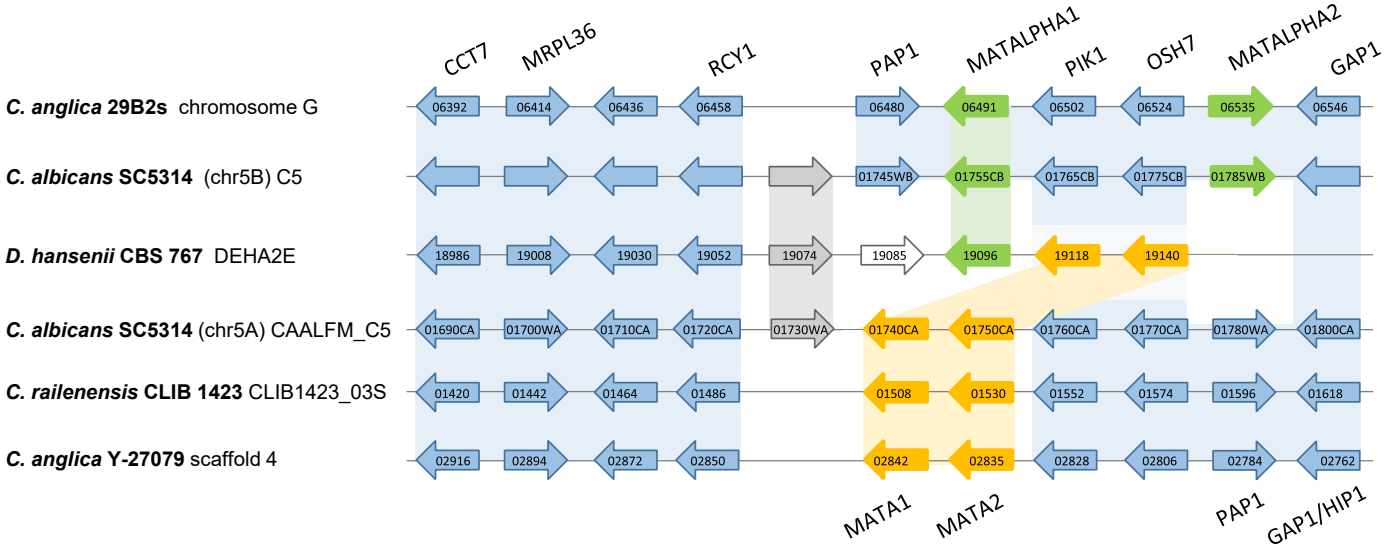


Fig. 5. Comparison of the structures of the *MAT* loci of *C. anglica*, *C. albicans*, *C. railenensis* and *Debaryomyces hansenii*. *MAT*a1 and *MAT*a2 genes are represented by yellow arrows, and *MAT*alpha1and *MAT*alpha2 by green arrows. Flanking genes that are conserved in synteny across strains are shown in blue. The names of the strains and species are shown on the left. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

predicted function. One is predicted to encode a ferric/cupric reductase (*CFL1*). The others are mating type locus genes. Y-27079 has a *MAT*a locus, like *Candida albicans* (haplotype A) and *Candida railenensis*, which belong to the same Debaryomycetaceae family as *C. anglica* (Fig. 5, Fig. S1) (Butler et al., 2004; Devillers et al., 2022). Conversely, the cheese strains harbour a *MAT*alpha locus like that of *Candida albicans* (haplotype B). By comparison, *Debaryomyces hansenii* carries a combination of MAT loci, with *MAT*alpha1, *MAT*a1 and *MAT*a2 in a syntenic region (Butler et al., 2009). *MAT*alpha1 is a pseudogene in 29B2s, and *MAT*alpha2 is also a pseudogene in all the cheese yeast genomes assembled here.

Another difference between the cheese strains and the cider strain concerned the number and position of the transposable elements. The largest copy, a 3594 bpL1-like element, was found in all strains. This copy contains a coding sequence of 1135 aa interrupted by a stop codon and without a methionine residue at the start. A non-LTR retrotransposon reverse transcriptase domain (located between amino acids 424 and 693) and an endonuclease domain (L1-EN) of the non-LTR retrotransposon LINE-1 (between amino acids 30 and 211) were detected, organised similarly to Ylli from *Yarrowia lipolytica* and the Zorro elements of *C. albicans* (Casaregola et al., 2002; Goodwin et al., 2001). Seven other degenerate copies were found in a conserved position in all strains. In addition, the cheese strains have two copies in common, whereas the cider strain has three specific copies. In all but two cases, tRNA genes were found in the vicinity of these elements (Fig. S3). These findings suggest that the L1-like element remained active after the genetic divergence of the cider strain, but is inactive in all the cheese strains, in which all the relics are located in the same position. Given that the cheese strains also showed a very low level of nucleotide divergence, the same *MAT*alpha locus, and two different karyotypic profiles resulting from chromosomal rearrangements, these findings are consistent with clonal propagation, as opposed to sexual reproduction.

3.4. Phenotypic diversity

Phenotypic analyses were performed by observing growth on agarose plates with different carbon sources, at different pH or temperatures, and under the influence of different stresses. Lipase and protease activities were also assessed. In total, 37 different sets of growth conditions were used.

We tested 11 carbon sources in YNB-based medium, of which maltose, lactose, sucrose, xylose and ethanol, and, to a lesser extent lactate, were the carbon sources least assimilated by the eight strains (Table 2; Fig. 6A). The addition of citric acid to the medium decreased the pH to 4.5, resulting in better growth of the strains on these carbon sources. For the remaining five carbon sources (glucose, fructose, galactose, yeast extract and glycerol), strain Y-27079, which was isolated from cider, formed much larger colonies than the seven strains isolated from cheese, whereas the colonies of all strains were of similar size on YPD medium. On YNB-glucose at pH 4.5, Y-27079 grew worse than at pH 6.8, but similar to the cheese strains.

Six hydrophobic substrates were tested in YNB-based medium (tributyrin, oleic acid, linoleic acid, octanoic acid, caproic acid, caprylic acid). Growth was weak on oleic acid for all strains (Table 2).

Five stress conditions were tested: temperature, ethanol resistance, osmotic stress (NaCl and glycerol), cell wall stress (SDS) and antibiotic stress (hygromycin). The strains did not grow on YPD when the incubation temperature was increased to 37 °C. The presence of 5% ethanol

Table 2
Physiological characteristics of *C. anglica* strains.

Medium	Test	28C1c	28C2s	28E1s	29A1s	29B2s	29E1s	35D1s	Y-27079
YNB glucose (1%)	Control	2	2	2	2	2	2	2	3.5
YNB glucose (1%) pH 4.5	pH	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5
YNB maltose (1%)	Carbon source	w	w	w	w	w	w	w	w
YNB maltose (1%) pH 4.5	pH	1	1	1	1	1	1	1	1.5
YNB lactose (1%)	Carbon source	w	w	w	w	w	w	w	w
YNB lactose (1%) pH 4.5	pH	1	1	1	1	1	1	1	1.5
YNB lactate (1%)	Carbon source	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1
YNB lactate (1%) pH 4.5	pH	1	1	1	1	1	1	1	1.5
YNB galactose (1%)	Carbon source	2	2	2	2	2	2	2	3
YNB sucrose (1%)	Carbon source	w	w	w	w	w	w	w	w
YNB fructose (1%)	Carbon source	1.5	1.5	1.5	1.5	1.5	1.5	1.5	3
YNB xylose (1%)	Carbon source	w/-	w/-	w/-	w/-	w/-	w/-	w/-	w/-
YNB glycerol (1%)	Carbon source	1	1	1	1	1	1	1	3.5
YNB yeast extract (1%)	Carbon source	2	2	2	2	2	2	2	2.5
YNB ethanol (5%)	Alcohol as C source	w/-	w/-	w/-	w/-	w/-	w/-	w/-	w/-
YNB ethanol (10%)	Alcohol as C source	w/-	w/-	w/-	w/-	w/-	w/-	w/-	w/-
YPD	Control	3	3	3	3	3	3	3	3
YPD 37 °C	Temperature	–	–	–	–	–	–	–	–
YPD ethanol (5%)	Ethanol stress	3	3	3	3	3	3	3	3
YPD ethanol (10%)	Ethanol stress	1.5	1.5	2	2	1.5	1.5	1.5	3
YNB glucose (1%) NaCl (1%)	Osmotic stress	2	2	2	2	2	2	2	3
YNB glucose (1%) NaCl (2%)	Osmotic stress	2	2	2	2	2	2	2	3
YNB glucose (1%) NaCl (4%)	Osmotic stress	2	2	2	2	2	2	2	2
YNB glucose (1%) NaCl (8%)	Osmotic stress	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1
YNB glucose (1%) NaCl (10%)	Osmotic stress	w	–	w	–	w/-	w	w	w/-
YPD glycerol (20%)	Osmotic stress	2	2	2	2	2	2	2	2
YPD SDS (0.01%)	Cell wall stress	2	2	2	2	–	2	2	3
YPD SDS (0.02%)	Cell wall stress	–	–	–	–	–	–	–	3
YNB glucose (1%) hygro. (100 mg/L)	Antibiotic stress	1	–	2	2	1	1	1	2.5
YNB glucose (1%) hygro. (200 mg/L)	Antibiotic stress	–	–	1	1	–	–	–	1
YNB tributyrin (1%)	Hydrophobic substrate, lipase activity	–	–	–	–	–	–	–	–
YNB oleic acid (1%)	Hydrophobic substrate	w	w	w	w	w	w	w	w
YNB linoleic acid (1%)	Hydrophobic substrate	–	–	–	–	–	–	–	–
YNB octanoic acid (1%)	Hydrophobic substrate	–	–	–	–	–	–	–	–
YNB caproic acid (1%)	Hydrophobic substrate	–	–	–	–	–	–	–	–
YNB caprylic acid (1%)	Hydrophobic substrate	–	–	–	–	–	–	–	–
Skimmed milk medium	Protease activity	w	w	w	w	w	w	w	w

w: weak growth, -: no growth, 0.5–3.5: colony size (arbitrary unit).

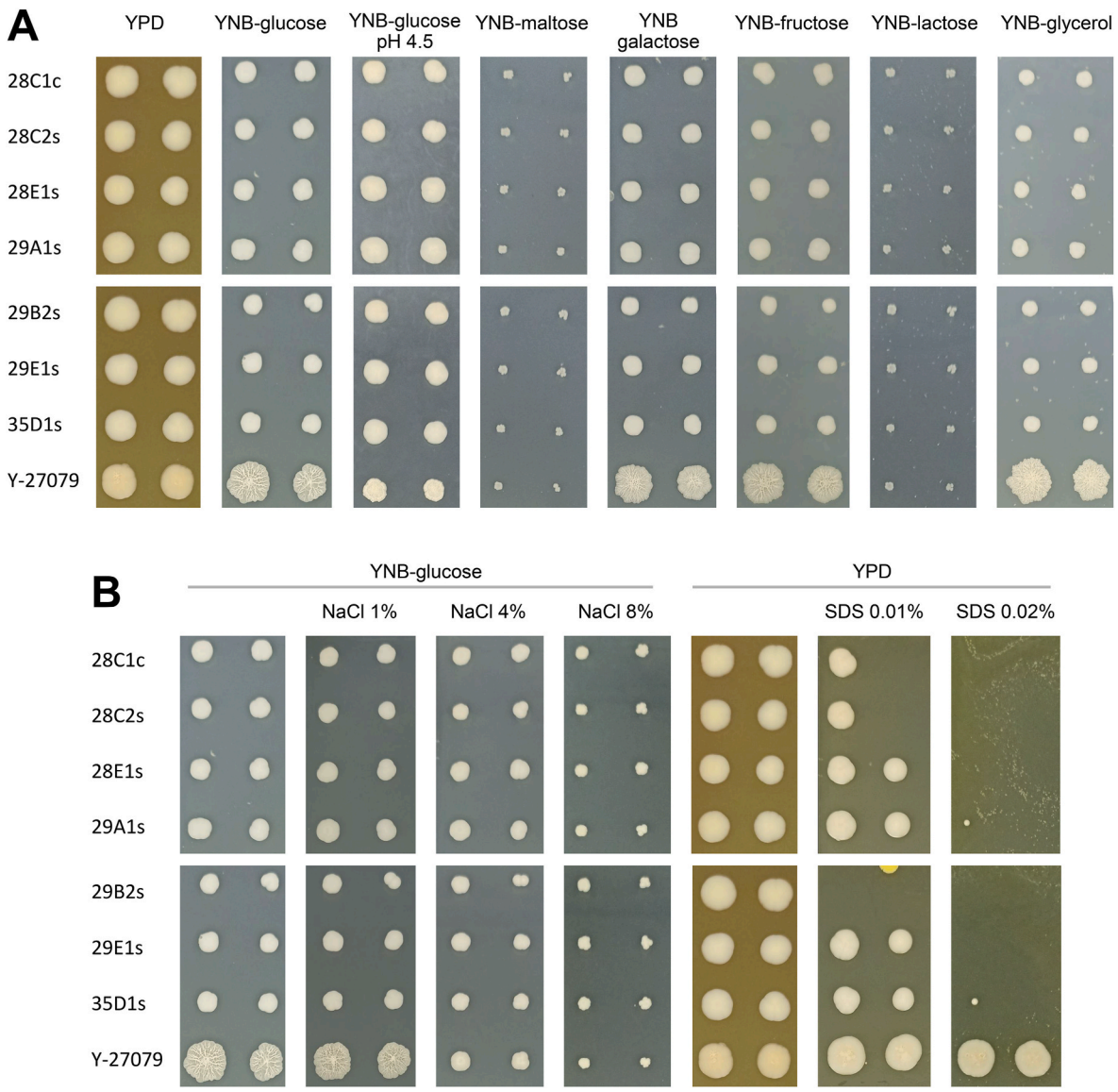


Fig. 6. Growth of *C. anglica* strains after 8 days at 28 °C in different culture conditions. For each medium, two 10-fold dilutions of cells were used. (A) Evaluation of carbon source utilization; (B) Growth in stressful conditions.

in YPD had no effect on the growth of any of the strains. By contrast, when ethanol was added to the medium at a concentration of 10%, Y-27079 was unaffected but the colonies of the cheese strains were much smaller, indicating a greater sensitivity to ethanol (Table 2). The ethanol resistance of Y-27079 may have been acquired through adaptive evolution in response to the ethanol present in cider. All the strains grew similarly on 20% glycerol, but the colonies on this medium were smaller than those obtained on YPD. High concentrations of NaCl, up to 10%, were used as a second osmotic stress. Colony size was unaffected by the addition to YPD of up to 2% NaCl. In the presence of 4% NaCl, Y-27079 colonies were smaller than those of the cheese strains, suggesting a lower salt tolerance of the strain isolated from cider. This finding was confirmed by the results obtained when 8% NaCl was added to the medium (Fig. 6B). However, when NaCl concentration reached 10%, all the strains grew poorly. NaCl resistance may be a signature of adaptation to cheese environments, which are generally very salty. SDS was used to induce cell wall stress. All the strains except 29B2s grew on medium containing 0.01% SDS, but Y-27079 formed much larger colonies than the cheese strains. At 0.02%, only the cider strain was able to grow as well as on YPD without SDS. Antibiotic stress was achieved by adding hygromycin at concentrations of 100 and 200 mg/L. The strains differed in their ability to grow: 28C2s was unable to grow in the presence of hygromycin. Four cheese strains grew poorly when hygromycin was present at 100 mg/L and did not grow at all at the higher concentration. The other two cheese strains and Y-27079 were able to grow in the presence of 200 mg/L hygromycin.

Enzymes secreted by fungi in cheese break down cheese proteins and fats, contributing to flavour and texture. For this reason, we investigated lipase and protease activities on YNB plus tributyrin, and on skimmed milk medium. Very weak growth was observed, with no halo around the colonies, demonstrating the absence of such extracellular activities. However, this is not surprising as many yeasts do not secrete such enzymes.

3.5. *Genotype-phenotype relationships*

Given the small number of strains included in this study and the clonal population structure of cheese strains, we were unable to perform genome-wide association studies (GWAS). We therefore tried to link phenotypic traits with nucleotide variants in coding sequences (CDS). The impact of such variants was classified by SnpEff software as “low”, “moderate”, or “high” with respect to strain 29B2s (Table S7).

First, we focused on two phenotypic traits that varied within the cheese group: the sensitivities of the strains to SDS and hygromycin. Strain 29B2s was found to be sensitive to SDS whereas the other cheese strains grew in the presence of 0.01% SDS. We investigated the genetic basis of this trait by focusing on strain 35D1s, which is the most closely related strain to 29B2s, from which it differs by 253 SNPs and indels. The nine mutations classified as having a high impact included a mutation of the *CTT1* gene (CAAN4_C07228). SDS can induce the oxidative stress response in yeast cells due to its ability to disrupt cell membranes and organelles (Sirisattha et al., 2004). SDS exposure has also been shown to lead to an increase in the expression of genes involved in antioxidant defence, such as the superoxide dismutase and catalase T gene *CTT1* (Cao et al., 2020). In 29B2s, *CTT1* is interrupted by a stop codon four amino acids downstream from the initiator methionine residue. However, this strain carries a second intact copy of *CTT1* (CAAN4_D03928), which invalidates this hypothesis (Fig. S2B). The other eight mutations with a high impact affected genes of unknown function or with no clear connection to SDS-mediated damage. The mutations affecting this trait may therefore be considered to be "moderate-impact" variants (e.g., non-synonymous mutations) according to the snpEff software or they may be located in promoter or intergenic regions.

A comparison of the 28C2s (sensitive to hygromycin) and 28E1s (resistant to hygromycin) strains led to the identification of 13 "high-impact" variants, one of which introduces a frameshift into the homologue of *MNN4* in the sensitive strain. Mnn4p is a putative positive regulator of the mannosylphosphate transferase Mnn6p. Its inactivation can lead to defective glycosylation resulting in marked sensitivity to the aminoglycoside hygromycin B in *S. cerevisiae* (Dean, 1995). This candidate gene will be tested experimentally in future studies on *C. anglica* strains.

Then, we investigated the variability between cheese strains and the cider strain for other traits, such as sugar assimilation and NaCl resistance (see below). Unfortunately, a large number of SNPs and indels were found between the two groups. Indeed, a large number of variants with an impact on CDSs were detected in comparisons of strain Y-27079 and strain 29B2s (8754 low-impact, 5257 moderate-impact and 330 high-impact variants).

The cider strain generally grew much better than the cheese strains on the sugars tested, particularly for glucose, galactose, and fructose. There are many possible reasons for this difference in growth profile. We first investigated the possible presence of hexokinase mutations. Two genes in 29B2s and Y-27079 are putative homologues of the hexokinase *HXK2* (CAAN4_H08460 and CAAN3_01S08438, respectively) and glucokinase *GLK1* (CAAN4_B08042 and CAAN3_12S04984, respectively) genes. The *GLK1* sequences of the two strains were found to be identical, but we identified an SNP in *HXK2*, resulting in the replacement of a serine residue (29B2s) with a threonine residue (Y-27079) at codon 50. As this amino acid is not located within the active site of the enzyme, we think it is likely to have little effect on Glk1p activity.

We then focused specifically on genes involved in galactose metabolism (Table S8). We used gene sequences from *S. cerevisiae* and *C. albicans* to search for *GAL* genes involved in galactose metabolism. We were unable to identify *GAL3*, but *GAL1*, the ohnologue of *GAL3* with galactokinase activity, was found to have a homologue in *C. anglica*. *GAL80*, a regulator of *GAL* gene transcription in *S. cerevisiae*, is restricted to the Saccharomycetaceae family, and indeed, no homologue was detected in *C. anglica*. Instead, two other transcription factors, Rtg1p and Rtg3p, were detected in *C. anglica,* which are known to regulate the transcription of *GAL* genes in *C. albicans* (Harrison et al., 2022). A 15.6 kb region containing four genes (*GAL2/HGT1, GAL1*, *GAL7*, and *GAL10*) was found in 29B2s and Y-27079 (Fig. S4) and presented a chromosomal organisation similar to that in *C. albicans*. This region is almost identical to that in 29B2s and Y-27079, including CDSs and intergenic regions, with a single non-synonymous mutation in *GAL2/HGT1* (M134I). Likewise, the sequences of the *GAL4* and *RTG1* genes were identical in the two strains. Two synonymous mutations were detected in *RTG3*.

The last strategy we used to investigate the genetic basis of phenotypic traits involved comparing the presence of pseudogenes between 29B2s and Y-27079. Only one of the 73 pseudogenes of strain Y-27079 had a putative function related to salt tolerance. Indeed, we identified a homologue of *ENA2* (CAAN3_11S05204), encoding an enzyme that catalyses the hydrolysis of ATP coupled with the transport of sodium or lithium ions, thereby mediating salt tolerance. This gene, which is not a pseudogene in 29B2s (CAAN4_G14510), 28E1s (CAAN1_11S00936) or 29E1s (CAAN2_07S00936), may underlie the greater salt tolerance of the cheese strains than of the cider strain.

4. Conclusions

Only one strain of the yeast *C. anglica,* from cider, had been isolated until this study. We present here seven new strains isolated from the core and rind of French PDO cheeses from different producers. This study provides new insight into this food-associated species, together with a high-quality reference genome sequence and annotation for a relatively little-known clade of the Serinales. This genome sequence may also help to clarify the taxonomy of genus *Kurtzmaniella* and closely related taxa.

The seven cheese strains were very closely related genetically and there are several lines of evidence suggesting that the propagation of *C. anglica* is probably clonal. As a consequence, these strains behaved identically, except for a few phenotypes. Conversely, the cider strain is genetically and phenotypically more divergent. The study of the relationships between genotypes and phenotypes identified candidate genes for involvement in the variation of phenotypic traits, potentially demonstrating adaptation to cheese and cider environments. The role of *C. anglica* strains in cheese-making and ripening remains unknown and should be studied further. It remains unknown, for example, whether this species is associated with anthropogenic environments or could be isolated from natural environments.

Funding

The MetaPDOcheese project was funded by the *Centre National Interprofessionnel de l'Economie Laitière* (CNIEL), France. This work was supported by the Genoscope, the *Commissariat à l'Énergie Atomique et aux Énergies Alternatives* (CEA) and the *France Génomique national infrastructure*, funded as part of the "Investissements d'Avenir" program managed by the Agence Nationale pour la Recherche (contract ANR-10-INBS-09).

CRediT authorship contribution statement

Frédéric Bigey: Writing – review & editing, Writing – original draft, Visualization, Formal analysis, Data curation. **Xavière Menatong Tene:** Writing – review & editing, Resources, Investigation. **Marc Wessner:** Methodology, Formal analysis. **Martine Pradal:** Resources, Investigation. **Jean-Marc Aury:** Writing – review & editing, Supervision, Project administration, Methodology, Formal analysis. **Corinne Cruaud:** Writing – review & editing, Methodology. **Cécile Neuvéglise:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that might be considered to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the USDA-ARS Culture Collection (NRRL) for providing the type strain of *Candida anglica*, NRRL Y-27079.

We thank Hugo Devillers (INRAE) for technical assistance in bioinformatics and Kenneth Wolfe (University College Dublin) for his help with studies of the *MAT* locus.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fm.2024.104584.

References

Abdelmoteleb, A., Troncoso-Rojas, R., Gonzalez-Soto, T., Gonzalez-Mendoza, D., 2017. Antifungical activity of Autochthonous Bacillus subtilis isolated from Prosopis juliflora against Phytopathogenic fungi. MYCOBIOLOGY 45, 385–391.

Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., Albini, G., Aury, J. M., Belser, C., Bertrand, A., Cruaud, C., Da Silva, C., Dossat, C., Gavory, F., Gas, S., Guy, J., Haquelle, M., Jacoby, E., Jaillon, O., Lemainque, A., Pelletier, E., Samson, G., Wessner, M., Genoscope Technical, T., Acinas, S.G., Royo-Llonch, M., Cornejo-Castillo, F.M., Logares, R., Fernandez-Gomez, B., Bowler, C., Cochrane, G., Amid, C., Hoopen, P.T., De Vargas, C., Grimsley, N., Desgranges, E., Kandels-Lewis, S., Ogata, H., Poulton, N., Sieracki, M.E., Stepanauskas, R., Sullivan, M.B., Brum, J.R., Duhaime, M.B., Poulos, B.T., Hurwitz, B.L., Tara Oceans Consortium, C., Pesant, S., Karsenti, E., Wincker, P., 2017. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. Sci. Data 4, 170093.

Aury, J.M., Cruaud, C., Barbe, V., Rogier, O., Mangenot, S., Samson, G., Poulain, J., Anthouard, V., Scarpelli, C., Artiguenave, F., Wincker, P., 2008. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. BMC Genom. 9, 603.

Aury, J.M., Istace, B., 2021. Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. NAR Genom Bioinform 3, lqab034.

Bao, W., Kojima, K.K., Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6, 11.

Belleggia, L., Milanovic, V., Ferrocino, I., Cocolin, L., Haouet, M.N., Scuota, S., Maoloni, A., Garofalo, C., Cardinali, F., Aquilanti, L., Mozzon, M., Foligni, R., Pasquini, M., Trombetta, M.F., Clementi, F., Osimani, A., 2020. Is there any still undisclosed biodiversity in Ciauscolo salami? A new glance into the microbiota of an artisan production as revealed by high-throughput sequencing. Meat Sci. 165, 108128.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580.

Birney, E., Clamp, M., Durbin, R., 2004. GeneWise and genomewise. Genome Res. 14, 988–995.

Butler, G., Kenny, C., Fagan, A., Kurischko, C., Gaillardin, C., Wolfe, K.H., 2004. Evolution of the MAT locus and its Ho endonuclease in yeast species. Proc. Natl. Acad. Sci. U. S. A. 101, 1632–1637.

Butler, G., Rasmussen, M.D., Lin, M.F., Santos, M.A., Sakthikumar, S., Munro, C.A., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J.L., Agrafioti, I., Arnaud, M.B., Bates, S., Brown, A.J., Brunke, S., Costanzo, M.C., Fitzpatrick, D.A., de Groot, P.W., Harris, D., Hoyer, L.L., Hube, B., Klis, F.M., Kodira, C., Lennard, N., Logue, M.E., Martin, R., Neiman, A.M., Nikolaou, E., Quail, M.A., Quinn, J., Santos, M.C., Schmitzberger, F.F., Sherlock, G., Shah, P., Silverstein, K.A., Skrzypek, M.S., Soll, D., Staggs, R., Stansfield, I., Stumpf, M.P., Sudbery, P.E., Srikantha, T., Zeng, Q., Berman, J., Berriman, M., Heitman, J., Gow, N.A., Lorenz, M.C., Birren, B.W., Kellis, M., Cuomo, C.A., 2009. Evolution of pathogenicity and sexual reproduction in eight Candida genomes. Nature 459, 657–662.

Cao, C., Cao, Z., Yu, P., Zhao, Y., 2020. Genome-wide identification for genes involved in sodium dodecyl sulfate toxicity in Saccharomyces cerevisiae. BMC Microbiol. 20, 34.

Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973.

Carver, T., Harris, S.R., Berriman, M., Parkhill, J., McQuillan, J.A., 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics 28, 464–469.

Casaregola, S., Neuveglise, C., Bon, E., Gaillardin, C., 2002. Ylli, a non-LTR retrotransposon L1 family in the dimorphic yeast Yarrowia lipolytica. Mol. Biol. Evol. 19, 664–677.

Chan, P.P., Lin, B.Y., Mak, A.J., Lowe, T.M., 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 49, 9077–9096.

Chen, Y., Nie, F., Xie, S.Q., Zheng, Y.F., Dai, Q., Bray, T., Wang, Y.X., Xing, J.F., Huang, Z.J., Wang, D.P., He, L.J., Luo, F., Wang, J.X., Liu, Y.Z., Xiao, C.L., 2021. Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat. Commun. 12, 60.

Chomczynski, P., Sacchi, N., 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem. 162, 156–159.

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 (Austin) 6, 80–92.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., Genomes Project Analysis, G., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158.

Daniel, H.M., Lachance, M.A., Kurtzman, C.P., 2014. On the reclassification of species assigned to Candida and other anamorphic ascomycetous yeast genera based on phylogenetic circumscription. Antonie Leeuwenhoek 106, 67–84.

Dean, N., 1995. Yeast glycosylation mutants are sensitive to aminoglycosides. Proc. Natl. Acad. Sci. U. S. A. 92, 1287–1291.

Denis, E., Sanchez, S., Mairey, B., Beluche, O., Cruaud, C., Lemainque, A., Wincker, P., Barbe, V., 2018. Extracting high molecular weight genomic DNA from Saccharomyces cerevisiae. Protocolexchange 1–6.

Devillers, H., Grondin, C., Thiriet, A., Legras, J.L., 2022. Draft genome sequence of Candida railenensis strain CLIB 1423, isolated from Papaya Fruit in French Guiana. Microbiol Resour Announc 11, e0055422.

Drillon, G., Carbone, A., Fischer, G., 2014. SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. PLoS One 9, e92621.

Dubarry, M., Noel, B., Rukwavu, T., Farhat, S., Da Silva, C., Seeleuthner, Y., Lebeurrier, M., Aury, J.M., Gmove a tool for Eukaryotic Gene Predictions using Various Evidences. https://f1000research.com/posters/5-681.

Duc, C., Pradal, M., Sanchez, I., Noble, J., Tesniere, C., Blondin, B., 2017. A set of nutrient limitations trigger yeast cell death in a nitrogen-dependent manner during wine alcoholic fermentation. PLoS One 12, e0184838.

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinf. 5, 113.

Engelen, S., Aury, J.M., Fastxtend. https://www.genoscope.cns.fr/fastxtend/.

Goodwin, T.J., Ormandy, J.E., Poulter, R.T., 2001. L1-like non-LTR retrotransposons in the yeast Candida albicans. Curr. Genet. 39, 83–91.

Groenewald, M., Hittinger, C.T., Bensch, K., Opulente, D.A., Shen, X.X., Li, Y., Liu, C., LaBella, A.L., Zhou, X., Limtong, S., Jindamorakot, S., Gonçalves, P., Robert, V., Wolfe, K.H., Rosa, C.A., Boekhout, T., Čadež, N., Péter, G., Sampaio, J.P., Lachance, M.A., Yurkov, A.M., Daniel, H.M., Takashima, M., Boundy-Mills, K., Libkind, D., Aoki, K., Sugita, T., Rokas, A., 2023. A genome-informed higher rank classification of the biotechnologically important fungal subphylum Saccharomycotina. Stud. Mycol. 105, 1–22.

Gu, Z., Gu, L., Eils, R., Schlesner, M., Brors, B., 2014. Circlize Implements and enhances circular visualization in R. Bioinformatics 30, 2811–2812.

Harrison, M.C., LaBella, A.L., Hittinger, C.T., Rokas, A., 2022. The evolution of the GALactose utilization pathway in budding yeasts. Trends Genet. 38, 97–106.

Holt, C., Yandell, M., 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinf. 12, 491.

Kent, W.J., 2002. BLAT–the BLAST-like alignment tool. Genome Res. 12, 656–664.

Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L., 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37, 907–915.

Kurtzman, C.P., Robnett, C.J., 2014. Description of Kuraishia piskuri f.a., sp. nov., a new methanol assimilating yeast and transfer of phylogenetically related Candida species to the genera Kuraishia and Nakazawaea as new combinations. FEMS Yeast Res. 14, 1028–1036.

Kurtzman, C.P., Robnett, C.J., Yarrow, D., 2001. Three new species of Candida from apple cider: C. anglica, C. cidri and C. pomicola. Antonie Leeuwenhoek 80, 237–244.

Lachance, M.A., Starmer, W.T., 2008. Kurtzmaniella gen. nov. and description of the heterothallic, haplontic yeast species Kurtzmaniella cleridarum sp. nov., the teleomorph of Candida cleridarum. Int. J. Syst. Evol. Microbiol. 58, 520–524.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

Li, R., Li, Y., Kristiansen, K., Wang, J., 2008. SOAP: short oligonucleotide alignment program. Bioinformatics 24, 713–714.

Lopes, M.R., Santos, A.R.O., Moreira, J.D., Santa-Brigida, R., Martins, M.B., Pinto, F.O., Valente, P., Morais, P.B., Jacques, N., Grondin, C., Casaregola, S., Lachance, M.A., Rosa, C.A., 2019. Kurtzmaniella hittingeri f.a., sp. nov., isolated from rotting wood and fruits, and transfer of three Candida species to the genus Kurtzmaniella as new combinations. Int. J. Syst. Evol. Microbiol. 69, 1504–1508.

Manni, M., Berkeley, M.R., Seppey, M., Simao, F.A., Zdobnov, E.M., 2021a. BUSCO Update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol. Biol. Evol. 38, 4647–4654.

Manni, M., Berkeley, M.R., Seppey, M., Zdobnov, E.M., 2021b. BUSCO: assessing genomic data quality and beyond. Curr Protoc 1, e323.

Marcais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., Zimin, A., 2018. MUMmer4: a fast and versatile genome alignment system. PLoS Comput. Biol. 14, e1005944.

McLaughlin, R.W., Hession, C., Bergin, S., Cosgrove, A., Dowd, A., Garvey, N., Litovskich, G., Osaigbovo, E., Popa, D., Thuku, C., Butler, G., Wolfe, K.H., Byrne, K. P., 2024. Genome sequences of two isolates of the yeast Candida zeylanoides: UCD849 from soil in Ireland, and AWD from an African wild dog. Microbiol Resour Announc, e0108123.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274.

O'Connor, B.D., van der Auwera, G., 2020. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media, Incorporated.

Opulente, D.A., Leavitt LaBella, A., Harrison, M.C., Wolters, J.F., Liu, C., Li, Y., Kominek, J., Steenwyk, J.L., Stoneman, H.R., VanDenAvond, J., Miller, C.R., Langdon, Q.K., Silva, M., Goncalves, C., Ubbelohde, E.J., Li, Y., Buh, K.V., Jarzyna, M., Haase, M.A.B., Rosa, C.A., Cadez, N., Libkind, D., DeVirgilio, J.H., Beth Hulfachor, A., Kurtzman, C.P., Sampaio, J.P., Goncalves, P., Zhou, X., Shen, X.X.,

Groenewald, M., Rokas, A., Hittinger, C.T., 2023. Genomic and ecological factors shaping specialism and generalism across an entire subphylum. bioRxiv.
Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., Salzberg, S.L., 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. 11, 1650–1667.
Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., Korobeynikov, A., 2020. Using SPAdes de novo assembler. Curr Protoc Bioinformatics 70, e102.
Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.
R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org.
Ropars, J., Didiot, E., Rodriguez de la Vega, R.C., Bennetot, B., Coton, M., Poirier, E., Coton, E., Snirc, A., Le Prieur, S., Giraud, T., 2020. Domestication of the emblematic white cheese-making fungus Penicillium camemberti and its diversification into two varieties. Curr. Biol. 30, 4441–4453 e4444.
Ropars, J., Maufrais, C., Diogo, D., Marcet-Houben, M., Perin, A., Sertour, N., Mosca, K., Permal, E., Laval, G., Bouchier, C., Ma, L., Schwartz, K., Voelz, K., May, R.C., Poulain, J., Battail, C., Wincker, P., Borman, A.M., Chowdhary, A., Fan, S., Kim, S.H., Le Pape, P., Romeo, O., Shin, J.H., Gabaldon, T., Sherlock, G., Bougnoux, M.E., d'Enfert, C., 2018. Gene flow contributes to diversification of the major fungal pathogen Candida albicans. Nat. Commun. 9, 2253.
Sacerdot, C., Casaregola, S., Lafontaine, I., Tekaia, F., Dujon, B., Ozier-Kalogeropoulos, O., 2008. Promiscuous DNA in the nuclear genomes of hemiascomycetous yeasts. FEMS Yeast Res. 8, 846–857.
Saubin, M., Devillers, H., Proust, L., Brier, C., Grondin, C., Pradal, M., Legras, J.L., Neuveglise, C., 2019. Investigation of genetic relationships between Hanseniaspora species found in Grape Musts revealed interspecific hybrids with dynamic genome structures. Front. Microbiol. 10, 2960.
Sirisattha, S., Momose, Y., Kitagawa, E., Iwahashi, H., 2004. Toxicity of anionic detergents determined by Saccharomyces cerevisiae microarray analysis. Water Res. 38, 61–70.
Smit, A.F.A., Hubley, R., Green, P., RepeatMasker. http://repeatmasker.org/.
Vaser, R., Sovic, I., Nagarajan, N., Sikic, M., 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27, 737–746.
Vezinhet, F., Blondin, B., Hallet, J.N., 1990. Chromosomal DNA patterns and mitochondrial DNA polymorphism as tools for identification of enological strains of Saccharomyces cerevisiae. Appplied Microbiology and Biotechnology 32, 568–571.
Yoon, S.H., Ha, S.M., Lim, J., Kwon, S., Chun, J., 2017. A large-scale evaluation of algorithms to calculate average nucleotide identity. Antonie Leeuwenhoek 110, 1281–1286.
Zhang, Z., Schwartz, S., Wagner, L., Miller, W., 2000. A greedy algorithm for aligning DNA sequences. J. Comput. Biol. 7, 203–214.